



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](https://doi.org/10.1101/gr.276358.121)

P<P	Published online September 15, 2022 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

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

3 Tal J. Shalev¹, Omnia Gamal El-Dien^{1,2}, Macaire M.S. Yuen¹, Shu Shengqiang³, Shaun D. Jackman⁴,
4 René L. Warren⁴, Lauren Coombe⁴, Lise van der Merwe⁵, Ada Stewart⁶, Lori B. Boston⁶, Christopher
5 Plott⁶, Jerry Jenkins⁶, Guifen He³, Juying Yan³, Mi Yan³, Jie Guo³, Jesse W. Breinholt^{7,8}, Leandro G.
6 Neves⁷, Jane Grimwood⁶, Loren H. Rieseberg⁹, Jeremy Schmutz^{3,6}, Inanc Birol⁴, Matias Kirst¹⁰, Alvin D.
7 Yanchuk⁵, Carol Ritland¹, John H. Russell⁵, Joerg Bohlmann¹

8

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

18

19 Corresponding authors: Tal J. Shalev and Joerg Bohlmann

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

21 **Phone: 1-604-822-9673**

22 **Email:** tal.shalev@mssl.ubc.ca; bohlmann@mssl.ubc.ca

23

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

25 **Running title:** Genomics of western redcedar

26 **Abstract:** We assembled the 9.8 Gbp genome of western redcedar (WRC, *Thuja plicata*), an ecologically
27 and economically important conifer species of the Cupressaceae. The genome assembly, derived from a
28 uniquely inbred tree produced through five generations of self-fertilization (selfing), was determined to be
29 86% complete by BUSCO analysis – one of the most complete genome assemblies for a conifer.
30 Population genomic analysis revealed WRC to be one of the most genetically depauperate wild plant
31 species, with an effective population size of approximately 300 and no significant genetic differentiation
32 across its geographic range. Nucleotide diversity, π , is low for a continuous tree species, with many loci
33 exhibiting zero diversity, and the ratio of π at zero- to four-fold degenerate sites is relatively high (\sim
34 0.33), suggestive of weak purifying selection. Using an array of genetic lines derived from up to five
35 generations of selfing, we explored the relationship between genetic diversity and mating system. While
36 overall heterozygosity was found to decline faster than expected during selfing, heterozygosity persisted
37 at many loci, and nearly 100 loci were found to deviate from expectations of genetic drift, suggestive of
38 associative overdominance. Non-reference alleles at such loci often harbor deleterious mutations and are
39 rare in natural populations, implying that balanced polymorphisms are maintained by linkage to dominant
40 beneficial alleles. This may account for how WRC remains responsive to natural and artificial selection,
41 despite low genetic diversity.

42

43 **Introduction**

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

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

62 Western redcedar (WRC, *Thuja plicata*) is an ecologically, economically, and culturally
63 important species in the Cupressaceae. Endemic to the Pacific Northwest of North America and ranging
64 from Northern California to Southern Alaska, WRC is a stress-tolerant, slow-growing tree prized for its
65 durable, lightweight and rot-resistant wood (Grime 1977). WRC is one of only five extant *Thuja* species
66 and is estimated to have diverged from its North American sister species *Thuja occidentalis* around 26
67 MYA (Li and Xiang 2005). Genetic studies have indicated low diversity in WRC (Copes 1981; Glaubitz

68 et al. 2000; O’Connell et al. 2008). Microsatellite data suggest that all WRC originated from an isolated
69 refugium near the southern end of its current distribution and radiated north and inland following the last
70 glacial period (O’Connell et al. 2008). Current climate models predict that its range will increase over the
71 next century, particularly in the interior of British Columbia (BC) (Gray and Hamann 2013), thus making
72 it a priority for genome analysis and expediting of traditional breeding cycles via genomic selection (GS).

73 Uniquely among conifers, WRC employs a mixed mating system of outcrossing and self-
74 fertilization (selfing), with a mean outcrossing rate of around 70% (El-Kassaby et al. 1994; O’Connell et
75 al. 2001, 2004), and appears to suffer very little inbreeding depression for fitness growth traits (Wang and
76 Russell 2006; Russell and Ferguson 2008). Mating systems in plants, particularly rates of selfing (s) and
77 its complement, outcrossing ($1 - s$), are of interest to evolutionary biologists due to their implications for
78 genetic diversity and fitness and have been investigated extensively over the past century (Stebbins 1957;
79 Lande and Schemske 1985; Barrett and Eckert 1990; Barrett et al. 2003; Wright et al. 2013). Though
80 inbreeding depression resulting from selfing can lead to negative fitness impacts, a benefit of selfing may
81 include reproductive assurance (Fisher 1941; Baker 1955), which, in the absence of strong inbreeding
82 depression, can allow self-compatible populations to expand their geographic range faster than obligate
83 outcrossers (Lande and Schemske 1985). Research on inbreeding in plants has mostly focused on mating
84 strategies in angiosperms (Barrett and Eckert 1990; Jarne and Charlesworth 1993; Vogler and Kalisz
85 2001; Barrett et al. 2003; Kalisz et al. 2004; Wright et al. 2013). Characterization of mixed mating
86 systems in conifers has been limited, largely by their long generation times and generally high self-
87 incompatibility (Sorensen 1982; Bishir and Namkoong 1987; Remington and O’Malley 2000; Williams et
88 al. 2003; Williams 2008). The exceptional ability of WRC to maintain such a mating system has allowed
89 for successful selfing for up to five generations in experimental trials, making WRC a potential model for
90 the study of inbreeding in conifers and more broadly in gymnosperms (Russell and Ferguson 2008).

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

94

95 **Results**

96 *The WRC genome assembly represents a highly complete conifer genome*

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

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

110 We assembled 7.95 Gbp of the estimated 9.8 Gbp genome to produce a draft assembly with an
111 N50 of 2.31 Mbp, the largest scaffold being 16.3 Mbp. This assembly comprises 67,895 scaffolds > 1 kbp
112 (**Table 1; Table S3**). Benchmarking Universal Single Copy Ortholog (BUSCO) analysis (Simão et al.
113 2015) determined the genome assembly to be 86% complete in the gene space. This is one of the highest
114 completeness estimates for a conifer genome (**Table 2; Table S4A**).

115 **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

116

117 Genome annotation using evidence from Iso-Seq full-length cDNAs resulted in the identification
118 of 39,659 gene models supported by the alignment of unique primary transcripts (**Dataset S1**), and an
119 additional 26,150 alternative transcripts. A total of 25,984 gene models had Pfam protein family
120 annotation, 31,537 had transcriptome support over their full length (100%), and 19,506 had peptide
121 homology coverage support 90% or greater (**Table S5; Dataset S2**). Intron length ranged from 20 bp to
122 148.3 kbp, which is consistent with estimates from other conifers, with maximum lengths ranging from
123 68 kbp (Norway spruce) (Nystedt et al. 2013) up to 579 kbp (sugar pine) (Stevens et al. 2016). Scott et al.
124 (2020) reported a maximum intron length of 1.4 Mb in the highly contiguous giant sequoia genome
125 assembly. Repeat elements comprised 60% of the WRC genome, which is low compared to other
126 conifers. Repeats comprised 79% of the sugar pine (Stevens et al. 2016) and giant sequoia genomes (Scott
127 et al. 2020). Single copy orthologs (SCOs) were detected by orthogroup comparison to the giant sequoia
128 gene set, yielding 11,937 SCOs (**Dataset S3**).

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

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

Taxon	<i>Thuja plicata</i>	<i>Sequoiadendron giganteum</i> (Scott et al. 2020)	<i>Picea glauca</i> (Warren et al. 2015)	<i>Picea abies</i> (Nystedt et al. 2013)	<i>Pinus lambertiana</i> (Stevens et al. 2016)	<i>Pinus taeda</i> (Zimin et al. 2014)	<i>Pseudotsuga menziesii</i> (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

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

141

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

143 We estimated nucleotide diversity, short-range linkage disequilibrium (LD), population structure, genetic
 144 differentiation, and effective population size (N_e) in $n = 112$ unrelated trees from across the geographic
 145 range of WRC (range-wide population; RWP) (**Table S7**). Trees were grouped into three subpopulations:
 146 Northern-Coastal ($n = 77$), Central ($n = 26$), and Southern-Interior ($n = 9$) (**Figure 1A**). Using a panel of
 147 single nucleotide polymorphisms (SNPs) that were genotyped via targeted sequence capture approach, we
 148 identified 2,454,925 variant and invariant sites, which were filtered separately and resulted in sets of
 149 18,371 SNPs (**Dataset S5**) and 2,186,998 invariant sites (see **Materials and Methods**). Total mean SNP
 150 depth was $34.3\times$.

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

156

157 *Linkage Disequilibrium*

158 Decay of linkage disequilibrium (LD), the non-random association of alleles at different loci in a
159 population, can inform on how likely different loci are to be assorted together during recombination. We
160 assessed short-range LD as represented by the squared correlation coefficient r^2 in the RWP, at a minor
161 allele frequency (MAF) threshold of 0.05 to avoid bias due to rare alleles ($n = 16,202$ SNPs). The mean of
162 all pairwise r^2 estimates was 0.299 with a median of 0.151. The half-decay value (the distance in which r^2
163 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
164 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
165 apart (**Figure 2B**). These estimates for LD decay are several orders of magnitude greater than those found
166 in other conifers, as well as many other tree species, where LD has been reported to decay rapidly within
167 tens to a few thousand bp (Krutovsky and Neale 2005; Heuertz et al. 2006; Pyhäjärvi et al. 2011; Pavy et
168 al. 2012; Fahrenkrog et al. 2017).

169

170 *Population structure and genetic differentiation*

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

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

192

193 *Nucleotide diversity in WRC*

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

200 To assess the efficacy of purifying selection, we estimated π_0/π_4 , the ratio of π in 0-fold to 4-fold
201 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
202 0.325. The site frequency spectrum (SFS) for 4-fold SNPs appeared to decay slower than the SFS for 0-

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

205

206 *Effective population size (N_e)*

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

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

218

219 *Persistent heterozygosity during complete selfing highlights genomic regions under selection*

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

226 Under selfing in diploids, heterozygosity is expected to decline by 50% in each generation purely
227 through genetic drift. Mean heterozygosity declined slower than expected (**Figure 4A; Table S12; Table**
228 **S13A**), while median observed heterozygosity was significantly lower than expected beginning in
229 generation S3 ($p < 0.05$, pairwise Sign test; **Table S13B**). Mean heterozygosity at FS was 0.296, while in
230 the RWP it was 0.219 (**Table S12**). In comparison, Chen et al. (2013) found mean heterozygosities of
231 0.33 and 0.36 in lodgepole pine (*Pinus contorta*) and white spruce, respectively.

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

239 Following the expectation of a 50% decline in heterozygosity per generation under complete
240 selfing and assuming a model of only genetic drift, we anticipated that 25% of SLs would become fixed
241 for one allele at any given locus and 25% for the other in each generation. Thus, by generation S4,
242 46.875% of SLs should fix for each allele, and 6.25% should remain heterozygous. We identified 83
243 SNPs that deviated from expected proportions of fixation at a false discovery rate threshold of 0.05
244 (hereafter: outlier SNPs) (**Dataset S9**). Of these, 15 fixed for the reference allele and 2 fixed for the
245 alternate allele more often than would be expected under drift alone; meanwhile, all outlier SNPs had a
246 higher proportion of heterozygous alleles by S4 than expected under drift. Outlier SNPs were present on
247 all putative LGs in the genome, and mean depth was similar to the mean depth of the total SNP set (29.3×
248 vs 34.3×, respectively). VEP predicted effects for 67 outlier SNPs, 14 (16.9%) of which were in coding
249 regions (**Dataset S10**). Gene Ontology (GO) annotation was available for 30 genes containing outlier

250 SNPs; ten GO categories were over-represented in outliers when compared to the entire SNP set ($p <$
251 0.05 , Fisher's Exact Test) (**Table S14**).

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

258

259 **Discussion**

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

264

265 *The WRC genome*

266 Sequence assembly of large and repetitive conifer genomes is becoming more feasible, with new
267 technologies such as Single Molecule Real Time (SMRT) long-read sequencing or linked-reads (Zimin et
268 al. 2017). WRC is one of only two conifers outside of the Pinaceae whose genome sequence has been
269 published. Recently, Scott et al. (2020) reported the first genome sequence in Cupressaceae, giant
270 sequoia, with a near-chromosome-scale assembly of 8.125 Gbp of the estimated 9 Gbp genome using a
271 combination of Oxford Nanopore long reads and Illumina short reads together with a Dovetail HiRise
272 Chicago and Hi-C statistical scaffolding and assembly approach. Though future chromosome-level

273 assembly would be of value to improve contiguity, BUSCO completeness scores (86.0% and 84.3% for
274 WRC and giant sequoia, respectively) and the very high completeness of the annotated gene set suggest
275 that the WRC assembly is currently of very high quality for the gene space. Previous studies using flow
276 cytometry estimated WRC's genome size at 12.0 – 12.5 Gbp (Ohri and Khoshoo 1986; Hizume et al.
277 2001). The WRC genome assembled at 7.95 Gbp. This discrepancy may be partially explained by
278 filtering of k -mers with very high depth of coverage in GenomeScope to remove organelle-derived reads,
279 which may also remove other heterochromatic sequences such as centromeres and telomeres; however, a
280 recent study in maize (*Zea mays*) found that selfing over several generations can reduce genome size by
281 up to 7.9% (Roessler et al. 2019), which may suggest that we are observing genome loss in WRC as well
282 given the genome assembly source. Flow cytometry to assess genome loss during selfing would be a
283 valuable future endeavour.

284

285 *Genetic diversity in WRC*

286 We estimated π to be 0.0027 in SCOs and 0.0020 across all sequenced space. These estimates are lower
287 than many other plant species using comparable methods, for example, Norway spruce ($\pi = 0.0049 -$
288 0.0063) (Wang et al. 2020), weedy broomcorn millet ($\pi = 0.14$; *Panicum miliaceum*) (Li et al. 2021), and
289 most *Populus* species ($\pi = 0.0041 - 0.011$) (Liu et al. 2022). We found lower π estimates only in highly
290 cultivated plants, such as soybean ($\pi = 0.0015$; *Glycine* spp.) (Bayer et al. 2022), or rare, isolated species,
291 such as *Populus qionghaoensis* ($\pi = 0.0014$), which is restricted to a single small island and has an
292 estimated N_e of ~ 500 (Liu et al. 2022). Additionally, our probe selection strategy for genotyping targeted
293 regions of high variability due to the very low levels of polymorphism in initial sequencing runs. This
294 may have led to inflated estimates of π , suggesting that genome-wide diversity may be even lower.

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

297 following a bottleneck event (Brandvain and Wright 2016; Chen et al. 2019). The low N_e (~ 300) and
298 general lack of population structure, genetic differentiation, or nucleotide divergence between geographic
299 subpopulations despite a relatively wide geographic range and continuous population suggest that much
300 of the variation in WRC was likely eliminated due to bottlenecks following the last glacial period, a
301 pattern confirmed by our Stairway Plot results and affirming the conclusions of previous studies (Copes
302 1981; Glaubitz et al. 2000; O'Connell et al. 2008). Mating system likely plays a role as well in WRC's
303 low diversity. It has been argued that selfing species should generally have a lower nucleotide diversity
304 due to a reduction of the effective recombination rate (Buckler IV and Thornsberry 2002). Thus, the
305 exceptionally slow rate of LD decay observed in our RWP is further evidence of a recent population
306 bottleneck or long-term effects of inbreeding (Golding and Strobeck 1980; Zhang et al. 2004; Slatkin
307 2008). In future studies, more extensive sampling, in particular for the Southern-Interior region, could
308 help in gaining more accurate estimates of genetic differentiation across populations of WRC.

309

310 *Selfing in WRC*

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

319 Of greater interest, however, is that the majority of loci deviating from expectations of drift
320 during selfing remained heterozygous, suggestive of balancing selection or associative overdominance at

321 these loci, with high levels of LD promoting genetic hitch-hiking near loci under selection to remain
322 heterozygous. The presence of missense variants in outlier loci coupled with the general rarity of
323 missense mutations in natural populations offers further support for associative overdominance as an
324 explanation for the retention of heterozygosity at these loci. This is congruent with relatively high π_0/π_4 in
325 the RWP, suggesting strong positive selection is maintaining current allele frequencies. No outlier loci
326 remained heterozygous in all lines, which suggests that these loci do not harbour strongly deleterious or
327 lethal mutations. Further, all three genotypes exist for many of these loci in the RWP.

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

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

342

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

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

365 WRC's apparent adaptability and potential for range expansion make it an important forest tree in
366 a time of changing climate and environments (Gray and Hamann 2013). Low genetic diversity and unique
367 mating system need to be considered as WRC breeding adopts strategies of GS that largely rely on
368 controlling relatedness in the population (Ritland et al. 2020). Recombination rate is another important

369 consideration, as GS relies on the presence of LD between SNPs and causal regions for traits, in addition
370 to relatedness between individuals (Meuwissen et al. 2001). WRC's high LD may be an advantage for
371 finding linked SNPs, but may also increase the risk of unintentional selection for correlated traits not
372 under selection. This could be mitigated by whole genome sequencing across breeding populations,
373 similar to GS approaches in livestock breeding (Raymond et al. 2018; Georges et al. 2019).

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

378

379 **Methods and Materials**

380 *Plant materials*

381 The WRC RWP represented $n = 112$ individuals originating from across the geographic range growing at
382 the Cowichan Lake Research Station (CLRS) at Mesachie Lake, BC, Canada. Trees were separated into
383 three geographic subpopulations, Northern-Coastal, Central, and Southern-Interior, based on UPGMA
384 clustering of genetic distances (O'Connell et al. 2008). SLs were produced over 12 years (1995 – 2007)
385 using an accelerated breeding approach (Russell and Ferguson 2008) at CLRS. Briefly, 15 pairs (30
386 individuals) of unrelated parents from across coastal BC and Vancouver Island (**Table S15**) were crossed
387 to create 15 FS families and ensure an initial inbreeding coefficient of $F = 0$. Each FS line was then selfed
388 for up to five generations (S1 – S5) with one generation every two years, facilitated by GA₃ hormone
389 treatment. A single S5 seedling of SL 23 (2323-211-S5) was used for genome sequencing. We selected
390 189 individuals from the 15 FS selfing families for genotyping for the SL analysis (**Table S1**).

391

392 *Genome sequencing and assembly*

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

396 Genome sequencing was executed using three types of libraries: short fragment paired-end, large
397 fragment mate-pair, and linked-reads from large molecules using 10× Genomics Chromium. Depth of k -
398 mer coverage profiles were computed for multiple values of k using ntCard v1.0.1 (Mohamadi et al. 2017)
399 (**Figure S8**). The largest value of k providing a k -mer coverage of at least 15 was selected, based on an
400 estimated coverage of > 99.9%, yielding $k = 128$ (Lander and Waterman 1988). We analyzed and
401 visualized k -mer profiles using GenomeScope v1.0.0 (Vurture et al. 2017). Paired-end reads were
402 assembled using ABySS v2.1.4 (parameters: $k=128$; $kc=3$) and scaffolded using the mate-pair reads with
403 ABySS-Scaffold (Jackman et al. 2017) (**Figure S9**). Linked-reads were aligned and misassemblies were
404 identified and corrected with Tigmint v1.1.2 (Jackman et al. 2018). The assembly was scaffolded using
405 the linked-reads with ARCS v1.0.5 (Yeo et al. 2018) ($-c 2$; $-m 4-20000$) and ABySS-Scaffold ($-n 5-7$; $-s$
406 $5000-20000$). Molecule size of the linked read libraries was estimated using ChromeQC v1.0.4
407 (<https://bcgsc.github.io/chromeqc>). Detailed DNA extraction, sequencing, and assembly methods can be
408 found in **Supplemental Methods**.

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

413

414 *Genome annotation*

415 For PacBio Iso-Seq, full-length cDNAs were synthesized from total RNA. We then generated
416 transcript assemblies from 1.4B 2×150 and 50M 2×100 stranded paired-end Illumina RNA-seq reads
417 using PERTRAN (Shengqiang et al. 2013), 18M PacBio Iso-Seq Circular Consensus Sequences (CCS),
418 and previous RNA-seq assemblies (Shalev et al. 2018) ([NCBI PRJNA704616](https://.ncbi.nlm.nih.gov/submit/PRJNA704616)). We determined gene loci
419 by transcript assembly alignments and EXONERATE v2.4.0 (Slater and Birney 2005) alignments of
420 proteins from *Arabidopsis thaliana*, *Glycine max*, *Populus trichocarpa*, *Oryza sativa*, *Vitis vinifera*,
421 *Aquilegia coerulea*, *Solanum lycopersicum*, *Amborella trichopoda*, *Physcomitrella patens*, *Selaginella*
422 *moellendorffii*, *Sphagnum magellanicum*, UniProt Pinales and Cupressales, and Swiss-Prot proteomes to
423 the repeat-soft-masked WRC genome using RepeatMasker v4.0.8 (Smit et al. 2015) with up to 2 kbp
424 extension on both ends. Gene models were predicted using FGENESH+ v3.1.1 (Salamov and Solovyev
425 2000), FGENESH_EST v2.6, and EXONERATE and PASA (Haas et al. 2003) assembly ORFs. The best-
426 scored predictions for each locus were selected and improved by PASA, adding untranslated regions
427 (UTRs), splicing correction, and alternative transcripts. All software was run using default parameters.
428 Detailed RNA extraction, sequencing, and genome annotation methods can be found in **Supplemental**
429 **Methods**.

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

437

438 *SNP genotyping*

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

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

456

457 *SNP filtering and annotation*

458 Variant sites were filtered using VCFtools v0.1.17 (Danecek et al. 2011) with the following flags: --max-
459 missing 0.95; --minQ 30; --min-meanDP 15; --max-meanDP 60. SNPs with an allele balance > 0.2 and $<$
460 0.8 or < 0.01 were retained to eliminate incorrectly called heterozygotes using vcfFilter in vcflib v1.0.1
461 (Garrison 2016). To eliminate paralogous loci, we excluded: SNPs with a read-ratio deviation score D
462 (McKinney et al. 2017) > 5 and < -5 , SNPs with a heterozygosity greater than 0.55, and SNPs with excess

463 heterozygosity and deviations from HWE in the RWP at a p -value cutoff of 0.05 and $1e-5$, respectively.

464 We also excluded SNPs with negative inbreeding coefficients (F_{IS}), using the formula:

$$465 \quad F_{IS} = 1 - \frac{H_O}{H_E \times (1 - 0.17)}$$

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

477

478 *Linkage disequilibrium*

479 Pairwise LD (r^2) was estimated in the RWP using PLINK v1.9 (Chang et al. 2015). LD was calculated for
 480 all scaffolds containing at least two SNPs (--r2; --ld-window-r2 0; --ld-window-kb 999999; --ld-window
 481 999999; --maf 0.05). Under drift-recombination equilibrium, our expectation of LD decay over distance
 482 will be:

$$483 \quad E(r^2) = \frac{1}{(1 + C)}$$

484 Where C is the product of the population recombination parameter $\rho = 4N_e r$ and the distance in bp, N_e is
 485 the effective population size and r is the recombination rate per bp (Sved 1971). Adjusting for population
 486 size n and a low level of mutation, decay of LD was estimated as a factor of n , and C (Hill and Weir
 487 1988; Remington et al. 2001; Marroni et al. 2011; Fahrenkrog et al. 2017).

$$488 \quad E(r^2) = \left[\frac{10 + C}{(2 + C)(11 + C)} \right] \left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)} \right]$$

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

490

491 *Population structure and genetic differentiation*

492 For STRUCTURE (Pritchard et al. 2000; Falush et al. 2007) analysis, SNPs were further filtered to
 493 include only SNPs with an r^2 threshold of 0.1 using a 2.17 Mb window, following the outcome of our LD
 494 analysis, a MAC threshold of 3 was used to remove singletons, and only SCO and intergenic SNPs were
 495 retained ($n = 4,765$). For all other genetic differentiation analyses, all SCO and intergenic SNPs were
 496 used ($n = 13,427$). Population structure of the RWP was estimated using STRUCTURE v2.3.4 and a
 497 DAPC using the adegenet package (v2.1.1) in R (Jombart 2008; Jombart et al. 2010; Jombart and Ahmed
 498 2011). A hierarchical F_{ST} test as implemented in the hierfstat package (v0.04-22) in R (Goudet 2005) was
 499 used to assess genetic differentiation between subpopulations in the RWP and a Mantel test was executed
 500 using mantel.randtest for 9,999 permutations in ade4 v1.7-15 in R (Dray and Dufour 2007) to assess
 501 isolation by distance. PCA was performed on the genotype matrix of each subpopulation to visualise
 502 genetic distance between individuals using ade4.

503 STRUCTURE software was run using 10,000 MCMC repetitions with 10,000 repetitions of burn-
 504 in and 10 iterations of each K. Results were analyzed using methods of Evanno et al. (2005) and
 505 Puechmaille (2016) of cluster and admixture estimation and selection; fastStructure v1.0 (Raj et al. 2014)
 506 was also used with 10-fold cross-validation. For DAPC, find.clusters was used to select the optimal

507 number of clusters based on Bayesian Information Criterion (BIC), and 10-fold cross-validation with
508 1,000 replicates was performed with xvaldapc to select the number of PCs and discriminant functions to
509 retain.

510

511 *Nucleotide diversity*

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

518

519 *Effective population size (N_e)*

520 We estimated N_e using the LD model estimation method under random mating as implemented in
521 NeEstimator v2.1 (Do et al. 2014). This method uses background LD shared among samples to estimate
522 N_e ; thus, SNPs with as little LD as possible are required (Waples and Do 2008; Gilbert and Whitlock
523 2015). Due to the high LD observed in WRC, we first generated putative linkage groups (LGs) for the
524 WRC genome by aligning all genomic scaffolds containing SNPs to the giant sequoia genome (Scott et al.
525 2020) using BLAST+. Scaffolds were assigned to their most likely LG based on bitscore. We then used
526 the nucmer command from MUMmer v4 (Marçais et al. 2018) to determine the most likely alignment
527 region for each scaffold in each LG. We retained SNPs estimated to be at least 2.17 Mbp apart. A MAF
528 threshold of 0.05 was established to eliminate bias that may be introduced by rare alleles, and a 95%
529 nonparametric JackKnife confidence interval was taken for the estimated value, as recommended by
530 Waples and Do (2008) and Gilbert and Whitlock (2015).

531 Stairway Plot 2 (Liu and Fu 2020) was used on the folded SFS from intergenic and 4-fold
532 degenerate positions to further assess N_e changes over time. We used the following parameters: nseq =
533 222; L = 238,557; pct_training = 0.67; nrand = 55, 110, 165, 220; ninput = 200; mu = 3.74e-9;
534 year_per_generation = 50.

535

536 *Genotype correction for SLs*

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

545

546 *Change in heterozygosity and inbreeding coefficients over time*

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

554
$$F_{\text{UNI}} = \frac{x_{ij}^2 - (1 + 2p_i)x_{ij} + 2p_i^2}{2p_i(1 - p_i)}$$

555 Where x is the number of copies of the reference allele and p is the population-wide allele frequency at
 556 that locus. Calculations of F_{UNI} do not consider LD; thus, we used SNPs filtered for LD and MAC ($n =$
 557 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,
 558 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
 559 corrected genotypes in SLs and uncorrected genotypes in the RWP.

560

561 *Isolating SNPs significantly deviating from expectations of drift*

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

576

577 *Data access*

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

581 The SNP data generated in this study has been submitted to the Zenodo data repository under DOI
582 [10.5281/zenodo.6562381](https://doi.org/10.5281/zenodo.6562381), and is available in the **Supplemental Datasets**.

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

588

589 **Acknowledgements**

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

601

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

612

613 **Competing Interest Statement:** The authors declare no competing interests.

614 **References**

- 615 Acosta JJ, Fahrenkrog AM, Neves LG, Resende MFR, Dervinis C, Davis JM, Holliday JA, Kirst M.
616 2019. Exome resequencing reveals evolutionary history, genomic diversity, and targets of selection
617 in the conifers *Pinus taeda* and *Pinus elliottii*. *Genome Biol Evol* **11** doi: 10.1093/gbe/evz016.
- 618 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J*
619 *Mol Biol* **215**: 403–410. doi: 10.1016/S0022-2836(05)80360-2.
- 620 Antos JA, Filipescu CN, Negrave RW. 2016. Ecology of western redcedar (*Thuja plicata*): Implications
621 for management of a high-value multiple-use resource. *For Ecol Manage* **375**: 211–222. doi:
622 10.1016/j.foreco.2016.05.043.
- 623 Baker HG. 1955. Self-Compatibility and Establishment After “Long-Distance” Dispersal. *Evolution* **9**:
624 347–348. doi: 10.2307/2405656.
- 625 Barrett SCH, Eckert CG. 1990. Variation and Evolution of Mating Systems in Seed Plants. In *Biological*
626 *Approaches and Evolutionary Trends in Plants*, pp. 229–254.
- 627 Barrett SCH, Richards AJ, Bayliss MW, Charlesworth D, Abbott RJ. 2003. Mating strategies in flowering
628 plants: The outcrossing-selfing paradigm and beyond. *Philos Trans R Soc B Biol Sci* **358**: 991–1004.
629 doi: 10.1098/rstb.2003.1301.
- 630 Bayer PE, Valliyodan B, Hu H, Marsh JI, Yuan Y, Vuong TD, Patil G, Song Q, Batley J, Varshney RK,
631 et al. 2022. Sequencing the USDA core soybean collection reveals gene loss during domestication
632 and breeding. *Plant Genome* **15**: e20109. doi: 10.1002/tpg2.20109.
- 633 Bierne N, Tsitrone A, David P. 2000. An inbreeding model of associative overdominance during a
634 population bottleneck. *Genetics* **155**: 1981–1990. doi: 10.1093/genetics/155.4.1981.
- 635 Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MM, Keeling CI, Brand D,
636 Vandervalk BP, et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from

- 637 whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497. doi:
638 10.1093/bioinformatics/btt178.
- 639 Bishir J, Namkoong G. 1987. Unsound seeds in conifers: estimation of numbers of lethal alleles and of
640 magnitudes of effects associated with the maternal parent. *Silvae Genet* **36**: 180–185.
- 641 Brandvain Y, Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. *Trends Genet*
642 **32** doi: 10.1016/j.tig.2016.01.004.
- 643 Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage
644 disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A* **101**: 15255–15260. doi:
645 10.1073/pnas.0404231101.
- 646 Buckler IV ES, Thornsberry JM. 2002. Plant molecular diversity and applications to genomics. *Curr Opin*
647 *Plant Biol* **5**: 107–111. doi: 10.1016/S1369-5266(02)00238-8.
- 648 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:
649 architecture and applications. *BMC Bioinformatics* **10** doi: 10.1186/1471-2105-10-421.
- 650 Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK:
651 Rising to the challenge of larger and richer datasets. *Gigascience* **4** doi: 10.1186/s13742-015-0047-
652 8.
- 653 Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. 2013. Mining conifers' mega-genome
654 using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP
655 discovery platform. *Tree Genet Genomes* **9**: 1537–1544. doi: 10.1007/s11295-013-0657-1.
- 656 Chen J, Glémin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant
657 and animal species. *Mol Biol Evol* **34**: 1417–1428. doi: 10.1093/molbev/msx088.
- 658 Chen J, Källman T, Gyllenstrand N, Lascoux M. 2010. New insights on the speciation history and
659 nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity (Edinb)* **104**: 3–14.

- 660 doi: 10.1038/hdy.2009.88.
- 661 Chen J, Li L, Milesi P, Jansson G, Berlin M, Karlsson B, Aleksic J, Vendramin GG, Lascoux M. 2019.
662 Genomic data provide new insights on the demographic history and the extent of recent material
663 transfers in Norway spruce. *Evol Appl* **12**: 1539–1551. doi: 10.1111/eva.12801.
- 664 Cherry ML. 1995. Genetic variation in western red cedar (*Thuja plicata* Donn) seedlings. The University
665 of British Columbia.
- 666 Christenhusz MJM, Reveal JL, Farjon A, Gardner MF, Mill RR, Chase MW. 2011. A new classification
667 and linear sequence of extant gymnosperms. *Phytotaxa* **19**: 55–70. doi: 10.11646/phytotaxa.19.1.3.
- 668 Copes DL. 1981. Isoenzyme uniformity in western red cedar seedlings from Oregon and Washington.
669 *Can J For Res* **11**: 451–453. doi: 10.1139/x81-060.
- 670 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth
671 GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.
672 doi: 10.1093/bioinformatics/btr330.
- 673 De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J,
674 Nilsson O, Ritland K, et al. 2014. Insights into conifer giga-genomes. *Plant Physiol* **166**: 1724–
675 1732. doi: 10.1104/pp.114.248708.
- 676 De Miguel M, Bartholomé J, Ehrenmann F, Murat F, Moriguchi Y, Uchiyama K, Ueno S, Tsumura Y,
677 Lagraulet H, De Maria N, et al. 2015. Evidence of intense chromosomal shuffling during conifer
678 evolution. *Genome Biol Evol* **7**: 2799–2809. doi: 10.1093/gbe/evv185.
- 679 Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: Re-
680 implementation of software for the estimation of contemporary effective population size (N_e) from
681 genetic data. *Mol Ecol Resour* **14**: 209–214. doi: 10.1111/1755-0998.12157.
- 682 Dray S, Dufour AB. 2007. The ade4 package: Implementing the duality diagram for ecologists. *J Stat*

- 683 *Softw* **22**: 1–20. doi: 10.18637/jss.v022.i04.
- 684 El-Kassaby YA. 1999. Phenotypic plasticity in western redcedar. *For Genet* **6**: 235–240. citeulike-article-
685 id:7424246.
- 686 El-Kassaby YA, Russell J, Ritland K. 1994. Mixed mating in an experimental population of western red
687 cedar, *Thuja plicata*. *J Hered* **85**: 227–231. doi: 10.1093/oxfordjournals.jhered.a111441.
- 688 Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics.
689 *Genome Biol* **20** doi: 10.1186/s13059-019-1832-y.
- 690 Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package
691 rrBLUP. *Plant Genome* **4**: 250–255. doi: 10.3835/plantgenome2011.08.0024.
- 692 Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software
693 STRUCTURE: A simulation study. *Mol Ecol* **14**: 2611–2620. doi: 10.1111/j.1365-
694 294X.2005.02553.x.
- 695 Fahrenkrog AM, Neves LG, Resende MFR, Dervinis C, Davenport R, Barbazuk WB, Kirst M. 2017.
696 Population genomics of the eastern cottonwood (*Populus deltoides*). *Ecol Evol* **7**: 9426–9440. doi:
697 10.1002/ece3.3466.
- 698 Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype
699 data: Dominant markers and null alleles. *Mol Ecol Notes* **7**: 574–578. doi: 10.1111/j.1471-
700 8286.2007.01758.x.
- 701 Farjon A. 2018. Conifers of the World. *Kew Bull* **73** doi: 10.1007/s12225-018-9738-5.
- 702 Fisher RA. 1941. Average Excess and Average Effect of a Gene Substitution. *Ann Eugen* **11**: 53–63. doi:
703 10.1111/j.1469-1809.1941.tb02272.x.
- 704 Garrison E. 2016. A C++ library for parsing and manipulating VCF files. *GitHub*.

- 705 <https://github.com/vcflib/vcflib>.
- 706 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr*
707 **1207.3907**. <http://arxiv.org/abs/1207.3907> (Accessed April 9, 2021).
- 708 Georges M, Charlier C, Hayes B. 2019. Harnessing genomic information for livestock improvement. *Nat*
709 *Rev Genet* **20**: 135–156. doi: 10.1038/s41576-018-0082-2.
- 710 Gilbert KJ, Whitlock MC. 2015. Evaluating methods for estimating local effective population size with
711 and without migration. *Evolution* **69**: 2154–2166. doi: 10.1111/evo.12713.
- 712 Glaubitz JC, El-Kassaby YA, Carlson JE. 2000. Nuclear restriction fragment length polymorphism
713 analysis of genetic diversity in western redcedar. *Can J For Res* **30**: 379–389. doi: 10.1007/s00216-
714 012-6081-9.
- 715 Golding GB, Strobeck C. 1980. Linkage disequilibrium in a finite population that is partially selfing.
716 *Genetics* **94**: 777–789. doi: 10.1093/genetics/94.3.777.
- 717 Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol Ecol*
718 *Notes* **5**: 184–186. doi: 10.1111/j.1471-8286.2004.00828.x.
- 719 Gray LK, Hamann A. 2013. Tracking suitable habitat for tree populations under climate change in
720 western North America. *Clim Change* **117**: 289–303. doi: 10.1007/s10584-012-0548-8.
- 721 Grime J. 1977. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to
722 Ecological and Evolutionary Theory. *Am Nat* **111**: 1169–1194. doi: 10.1086/283244.
- 723 Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch
724 DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript
725 alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666. doi: 10.1093/nar/gkg770.
- 726 Hedrick PW, Hellsten U, Grattapaglia D. 2016. Examining the cause of high inbreeding depression:

- 727 Analysis of whole-genome sequence data in 28 selfed progeny of *Eucalyptus grandis*. *New Phytol*
728 **209**: 600–611. doi: 10.1111/nph.13639.
- 729 Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N.
730 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of
731 Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105. doi:
732 10.1534/genetics.106.065102.
- 733 Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations.
734 *Theor Popul Biol* **33**: 54–78. doi: 10.1016/0040-5809(88)90004-4.
- 735 Hizume M, Kondo T, Shibata F, Ishizuka R. 2001. Flow cytometric determination of genome size in the
736 Taxodiaceae, Cupressaceae *sensu stricto* and Sciadopityaceae. *Cytologia (Tokyo)* **66**: 307–311. doi:
737 10.1508/cytologia.66.307.
- 738 Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J,
739 Jones SJM, et al. 2018. Tigmint: Correcting assembly errors using linked reads from large
740 molecules. *BMC Bioinformatics* **19** doi: 10.1186/s12859-018-2425-6.
- 741 Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L,
742 Warren RL, et al. 2017. ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom
743 filter. *Genome Res* **27**: 768–777. doi: 10.1101/gr.214346.116.
- 744 Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017. The $K =$
745 2 conundrum. *Mol Ecol* **26** doi: 10.1111/mec.14187.
- 746 Jarne P, Charlesworth D. 1993. The evolution of the selfing rate in functionally hermaphrodite plants and
747 animals. *Annu Rev Ecol Syst* **24**: 441–466. doi: 10.1146/annurev.es.24.110193.002301.
- 748 Jombart T. 2008. Adegnet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*
749 **24**: 1403–1405. doi: 10.1093/bioinformatics/btn129.

- 750 Jombart T, Ahmed I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data.
751 *Bioinformatics* **27**: 3070–3071. doi: 10.1093/bioinformatics/btr521.
- 752 Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: A new method
753 for the analysis of genetically structured populations. *BMC Genet* **11** doi: 10.1186/1471-2156-11-94.
- 754 Kalisz S, Vogler DW, Hanley KM. 2004. Context-dependent autonomous self-fertilization yields
755 reproductive assurance and mixed mating. *Nature* **430**: 884–887. doi: 10.1038/nature02776.
- 756 Korunes KL, Samuk K. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the
757 presence of missing data. *Mol Ecol Resour* **21** doi: 10.1111/1755-0998.13326.
- 758 Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB
759 v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for
760 evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**: D807–D811. doi:
761 10.1093/nar/gky1053.
- 762 Krutovsky KV, Neale DB. 2005. Nucleotide Diversity and Linkage Disequilibrium in Cold-Hardiness-
763 and Wood Quality-Related Candidate Genes in Douglas Fir. *Genetics* **171**: 2029–2041. doi:
764 10.1534/genetics.105.044420.
- 765 Lande R, Schemske DW. 1985. The Evolution of Self-Fertilization and Inbreeding Depression in Plants.
766 I. Genetic Models. *Evolution* **39**: 24–40. doi: 10.2307/2408514.
- 767 Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical
768 analysis. *Genomics* **2**: 231–239. doi: 10.1016/0888-7543(88)90007-9.
- 769 Li C, Liu M, Sun F, Zhao X, He M, Li T, Lu P, Xu Y. 2021. Genetic Divergence and Population
770 Structure in Weedy and Cultivated Broomcorn Millets (*Panicum miliaceum* L.) Revealed by
771 Specific-Locus Amplified Fragment Sequencing (SLAF-Seq). *Front Plant Sci* **12** doi:
772 10.3389/fpls.2021.688444.

- 773 Li JH, Xiang QP. 2005. Phylogeny and biogeography of *Thuja* L. (Cupressaceae), an eastern Asian and
774 North American disjunct genus. *J Integr Plant Biol* **47**: 651–659. doi: 10.1111/j.1744-
775 7909.2005.00087.x.
- 776 Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome
777 duplications in conifers and other seed plants. *Sci Adv* **1** doi: 10.1126/sciadv.1501084.
- 778 Liu S, Zhang L, Sang Y, Lai Q, Zhang X, Jia C, Long Z, Wu J, Ma T, Mao K, et al. 2022. Demographic
779 History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to
780 Introgression across *Populus* Genome. *Mol Biol Evol* **39** doi: 10.1093/molbev/msac008.
- 781 Liu X, Fu Y-X. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra.
782 *Genome Biol* **21**: 280. doi: 10.1186/s13059-020-02196-9.
- 783 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and
784 versatile genome alignment system. *PLoS Comput Biol* **14** doi: 10.1371/journal.pcbi.1005944.
- 785 Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. 2011. Nucleotide
786 diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4)
787 gene. *Tree Genet Genomes* **7**: 1011–1023. doi: 10.1007/s11295-011-0391-5.
- 788 McKinney GJ, Waples RK, Seeb LW, Seeb JE. 2017. Paralogs are revealed by proportion of
789 heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural
790 populations. *Mol Ecol Resour* **17** doi: 10.1111/1755-0998.12613.
- 791 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The
792 Ensembl Variant Effect Predictor. *Genome Biol* **17** doi: 10.1186/s13059-016-0974-4.
- 793 Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide
794 dense marker maps. *Genetics* **157**: 1819–1829. doi: 11290733.
- 795 Mohamadi H, Khan H, Birol I. 2017. ntCard: A streaming algorithm for cardinality estimation in

- 796 genomics data. *Bioinformatics* **33**: 1324–1330. doi: 10.1093/bioinformatics/btw832.
- 797 Mukrimin M, Kovalchuk A, Neves LG, Jaber EHA, Haapanen M, Kirst M, Asiegbu FO. 2018. Genome-
798 wide exon-capture approach identifies genetic variants of Norway spruce genes associated with
799 susceptibility to *Heterobasidion parviporum* infection. *Front Plant Sci* **9** doi:
800 10.3389/fpls.2018.00793.
- 801 Neale DB, McGuire PE, Wheeler NC, Stevens KA, Crepeau MW, Cardeno C, Zimin AV, Puiu D, Pertea
802 GM, Sezen UU, et al. 2017. The Douglas-Fir genome sequence reveals specialization of the
803 photosynthetic apparatus in Pinaceae. *G3 Genes, Genomes, Genet* **7**: 3157–3167. doi:
804 10.1534/g3.117.300078.
- 805 Neale DB, Zimin AV, Zaman S, Scott AD, Shrestha B, Workman RE, Puiu D, Allen BJ, Moore ZJ,
806 Sekhwal MK, et al. 2022. Assembled and annotated 26.5 Gbp coast redwood genome: a resource for
807 estimating evolutionary adaptive potential and investigating hexaploid origin. *G3 Genes, Genomes,*
808 *Genet* **12** doi: 10.1093/G3JOURNAL/JKAB380.
- 809 Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction
810 endonucleases. *Proc Natl Acad Sci U S A* **76**: 5269–5273. doi: 10.1073/pnas.76.10.5269.
- 811 Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N,
812 Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer
813 genome evolution. *Nature* **497**: 579–584. doi: 10.1038/nature12211.
- 814 O’Connell LM, Ritland K, Thompson SL. 2008. Patterns of post-glacial colonization by western redcedar
815 (*Thuja plicata*, Cupressaceae) as revealed by microsatellite markers. *Botany* **86**: 194–203. doi:
816 10.1139/B07-124.
- 817 O’Connell LM, Russell J, Ritland K. 2004. Fine-scale estimation of outcrossing in western redcedar with
818 microsatellite assay of bulked DNA. *Heredity (Edinb)* **93**: 443–449. doi: 10.1038/sj.hdy.6800521.

- 819 O'Connell LM, Viard F, Russell J, Ritland K. 2001. The mating system in natural populations of western
820 redcedar (*Thuja plicata*). *Can J Bot* **79**: 753–756. doi: 10.1139/cjb-79-6-753.
- 821 Ohri D, Khoshoo TN. 1986. Genome size in gymnosperms. *Plant Syst Evol* **153**: 119–132. doi:
822 10.1007/BF00989421.
- 823 Pavy N, Namroud MC, Gagnon F, Isabel N, Bousquet J. 2012. The heterogeneous levels of linkage
824 disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity (Edinb)*
825 **108**: 273–284. doi: 10.1038/hdy.2011.72.
- 826 Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype
827 data. *Genetics* **155**: 945–959. doi: 10.1093/genetics/155.2.945.
- 828 Prunier J, Verta JP, Mackay JJ. 2016. Conifer genomics and adaptation: At the crossroads of genetic
829 diversity and genome function. *New Phytol* **209**: 44–62. doi: 10.1111/nph.13565.
- 830 Puechmaile SJ. 2016. The program STRUCTURE does not reliably recover the correct population
831 structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol*
832 *Resour* **16**: 608–627. doi: 10.1111/1755-0998.12512.
- 833 Pyhäjärvi T, Kujala ST, Savolainen O. 2011. Revisiting protein heterozygosity in plants-nucleotide
834 diversity in allozyme coding genes of conifer *Pinus sylvestris*. *Tree Genet Genomes* **7**: 385–397.
835 doi: 10.1007/s11295-010-0340-8.
- 836 R Core Team. 2021. R: A language and environment for statistical computing.
- 837 Raj A, Stephens M, Pritchard JK. 2014. FastSTRUCTURE: Variational inference of population structure
838 in large SNP data sets. *Genetics* **197**: 573–589. doi: 10.1534/genetics.114.164350.
- 839 Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. 2018. Utility of whole-
840 genome sequence data for across-breed genomic prediction. *Genet Sel Evol* **50** doi: 10.1186/s12711-
841 018-0396-8.

- 842 Remington DL, O'Malley DM. 2000. Whole-genome characterization of embryonic stage inbreeding
843 depression in a selfed loblolly pine family. *Genetics* **155**: 337–348. doi: 10.1093/genetics/155.1.337.
- 844 Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman
845 MM, Buckler IV ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the
846 maize genome. *Proc Natl Acad Sci U S A* **98**: 11479–11484. doi: 10.1073/pnas.201394398.
- 847 Ritland K, Miscampbell A, van Niejenhuis A, Brown P, Russell J. 2020. Selfing and correlated paternity
848 in relation to pollen management in western red cedar seed orchards. *Botany* **98**: 185–200. doi:
849 10.1139/cjb-2019-0123.
- 850 Roessler K, Muyle A, Diez CM, Gaut GRJ, Bousios A, Stitzer MC, Seymour DK, Doebley JF, Liu Q,
851 Gaut BS. 2019. The genome-wide dynamics of purging during selfing in maize. *Nat Plants* **5**: 980–
852 990. doi: 10.1038/s41477-019-0508-7.
- 853 Russell JH, Ferguson DC. 2008. Preliminary results from five generations of a western redcedar (*Thuja*
854 *plicata*) selection study with self-mating. *Tree Genet Genomes* **4**: 509–518. doi: 10.1007/s11295-
855 007-0127-8.
- 856 Russell JH, Kope HH, Ades P, Collinson H. 2007. Variation in cedar leaf blight (*Didymascella thujina*)
857 resistance of western redcedar (*Thuja plicata*). *Can J For Res* **37** doi: 10.1139/X07-034.
- 858 Salamov AA, Solovyev VV. 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* **10**:
859 516–522. doi: 10.1101/gr.10.4.516.
- 860 Scott AD, Zimin AV, Puiu D, Workman R, Britton M, Zaman S, Caballero M, Read AC, Bogdanove AJ,
861 Burns E, et al. 2020. A reference genome sequence for giant sequoia. *G3 Genes, Genomes, Genet*
862 **10**: 3907–3919. doi: 10.1534/g3.120.401612.
- 863 Shalev TJ, Yuen MMS, Gesell A, Yuen A, Russell JH, Bohlmann J. 2018. An annotated transcriptome of
864 highly inbred *Thuja plicata* (Cupressaceae) and its utility for gene discovery of terpenoid

- 865 biosynthesis and conifer defense. *Tree Genet Genomes* **14** doi: 10.1007/s11295-018-1248-y.
- 866 Shengqiang S, Goodstein D, Rokhsar D. 2013. PERTRAN: Genome-guided RNA-seq Read Assembler.
867 In *Cold Spring Harbor Lab Genome Informatics*, New York, NY.
- 868 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing
869 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:
870 3210–3212. doi: 10.1093/bioinformatics/btv351.
- 871 Slate J, David P, Dodds KG, Veenivliet BA, Glass BC, Broad TE, McEwan JC. 2004. Understanding the
872 relationship between the inbreeding coefficient and multilocus heterozygosity: Theoretical
873 expectations and empirical data. *Heredity (Edinb)* **93**: 255–265. doi: 10.1038/sj.hdy.6800485.
- 874 Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison.
875 *BMC Bioinformatics* **6** doi: 10.1186/1471-2105-6-31.
- 876 Slatkin M. 2008. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical
877 future. *Nat Rev Genet* **9**: 477–485. doi: 10.1038/nrg2361.
- 878 Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013-2015 . <http://www.repeatmasker.org>.
879 <http://www.repeatmasker.org>.
- 880 Sorensen FC. 1982. The Roles of Polyembryony and Embryo Viability in the Genetic System of Conifers.
881 *Evolution* **36**: 725–733. doi: 10.2307/2407885.
- 882 Stebbins GL. 1957. Self Fertilization and Population Variability in the Higher Plants. *Am Nat* **91**: 337–
883 354. doi: 10.1086/281999.
- 884 Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, Paul R, Gonzalez-Ibeas D,
885 Koriabine M, Holtz-Morris AE, et al. 2016. Sequence of the sugar pine megagenome. *Genetics* **204**:
886 1613–1626. doi: 10.1534/genetics.116.193227.

- 887 Stewart WN. 1983. *Paleobotany and the evolution of plants*. 2nd ed. Cambridge University Press,
888 Cambridge, UK.
- 889 Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations.
890 *Theor Popul Biol* **2**: 125–141. doi: 10.1016/0040-5809(71)90011-6.
- 891 Telfer E, Graham N, Macdonald L, Li Y, Klápště J, Resende M, Neves LG, Dungey H, Wilcox P. 2019.
892 A high-density exome capture genotype-by-sequencing panel for forestry breeding in *Pinus radiata*.
893 *PLoS One* **14** doi: 10.1371/journal.pone.0222640.
- 894 Vidalis A, Scofield DG, Neves LG, Bernhardsson C, García-Gil MR, Ingvarsson PK. 2018. Design and
895 evaluation of a large sequence-capture probe set and associated SNPs for diploid and haploid
896 samples of Norway spruce (*Picea abies*). *bioRxiv* doi: 10.1101/291716.
- 897 Vogler DW, Kalisz S. 2001. Sex among the flowers: The distribution of plant mating systems. *Evolution*
898 **55**: 202–204. doi: 10.1111/j.0014-3820.2001.tb01285.x.
- 899 Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017.
900 GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202–
901 2204. doi: 10.1093/bioinformatics/btx153.
- 902 Wang T, Russell JH. 2006. Evaluation of selfing effects on western redcedar growth and yield in
903 operational plantations using the tree and stand simulator (TASS). *For Sci* **52**: 281–289. doi:
904 10.5849/forsci.15-042.
- 905 Wang X, Bernhardsson C, Ingvarsson PK. 2020. Demography and Natural Selection Have Shaped
906 Genetic Variation in the Widely Distributed Conifer Norway Spruce (*Picea abies*). *Genome Biol*
907 *Evol* **12**: 3803–3817. doi: 10.1093/gbe/evaa005.
- 908 Waples RS, Do C. 2008. LDNE: A program for estimating effective population size from data on linkage
909 disequilibrium. *Mol Ecol Resour* **8**: 753–756. doi: 10.1111/j.1755-0998.2007.02061.x.

- 910 Warren RL, Keeling CI, Yuen MM Saint, Raymond A, Taylor GA, Vandervalk BP, Mohamadi H,
911 Paulino D, Chiu R, Jackman SD, et al. 2015. Improved white spruce (*Picea glauca*) genome
912 assemblies and annotation of large gene families of conifer terpenoid and phenolic defense
913 metabolism. *Plant J* **83**: 189–212. doi: 10.1111/tpj.12886.
- 914 Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV,
915 Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and
916 phylogenomics. *Mol Biol Evol* **35**: 543–548. doi: 10.1093/molbev/msx319.
- 917 Williams CG. 2008. Selfed embryo death in *Pinus taeda*: A phenotypic profile. *New Phytol* **178**: 210–
918 222. doi: 10.1111/j.1469-8137.2007.02359.x.
- 919 Williams CG, Auckland LD, Reynolds MM, Leach KA. 2003. Overdominant lethals as part of the conifer
920 embryo lethal system. *Heredity (Edinb)* **91**: 584–592. doi: 10.1038/sj.hdy.6800354.
- 921 Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014.
922 Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding
923 Regions of *Capsella grandiflora*. *PLoS Genet* **10** doi: 10.1371/journal.pgen.1004622.
- 924 Wright S. 1922. Coefficients of Inbreeding and Relationship. *Am Nat* **56**: 330–338. doi: 10.1086/279872.
- 925 Wright S. 1931. Evolution in Mendelian Populations. *Genetics* **16**: 97–159. doi: 10.1007/BF02459575.
- 926 Wright SI, Kalisz S, Slotte T. 2013. Evolutionary consequences of self-fertilization in plants. *Proc R Soc*
927 *B Biol Sci* **280** doi: 10.1098/rspb.2013.0133.
- 928 Xin Z, Chen J. 2012. A high throughput DNA extraction method with high yield and quality. *Plant*
929 *Methods* **8** doi: 10.1186/1746-4811-8-26.
- 930 Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait
931 analysis. *Am J Hum Genet* **88**: 76–82. doi: 10.1016/j.ajhg.2010.11.011.

- 932 Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS: Scaffolding genome drafts with linked reads.
933 *Bioinformatics* **34**: 725–731. doi: 10.1093/bioinformatics/btx675.
- 934 Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE. 2004. Impact
935 of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium
936 maps. *Proc Natl Acad Sci U S A* **101**: 18075–18080. doi: 10.1073/pnas.0408251102.
- 937 Zhang YY, Fischer M, Colot V, Bossdorf O. 2013. Epigenetic variation creates potential for evolution of
938 plant phenotypic plasticity. *New Phytol* **197**: 314–322. doi: 10.1111/nph.12010.
- 939 Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M,
940 Wegrzyn JL, de Jong PJ, et al. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome.
941 *Genetics* **196**: 875–890. doi: 10.1534/genetics.113.159715.
- 942 Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg
943 SL. 2017. An improved assembly of the loblolly pine mega-genome using long-read single-
944 molecule sequencing. *Gigascience* **6** doi: 10.1093/gigascience/giw016.
- 945

946 **Figure Headings**

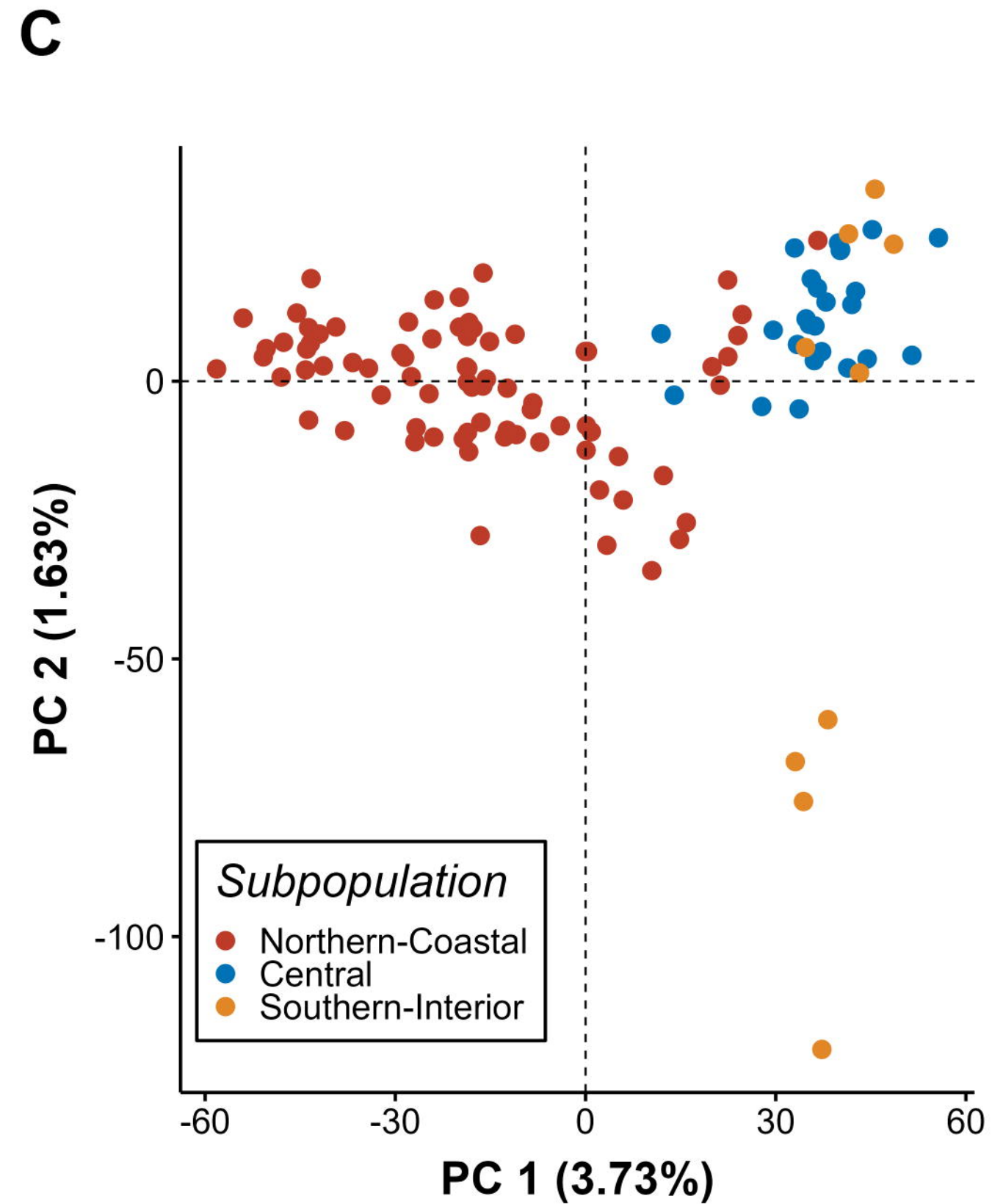
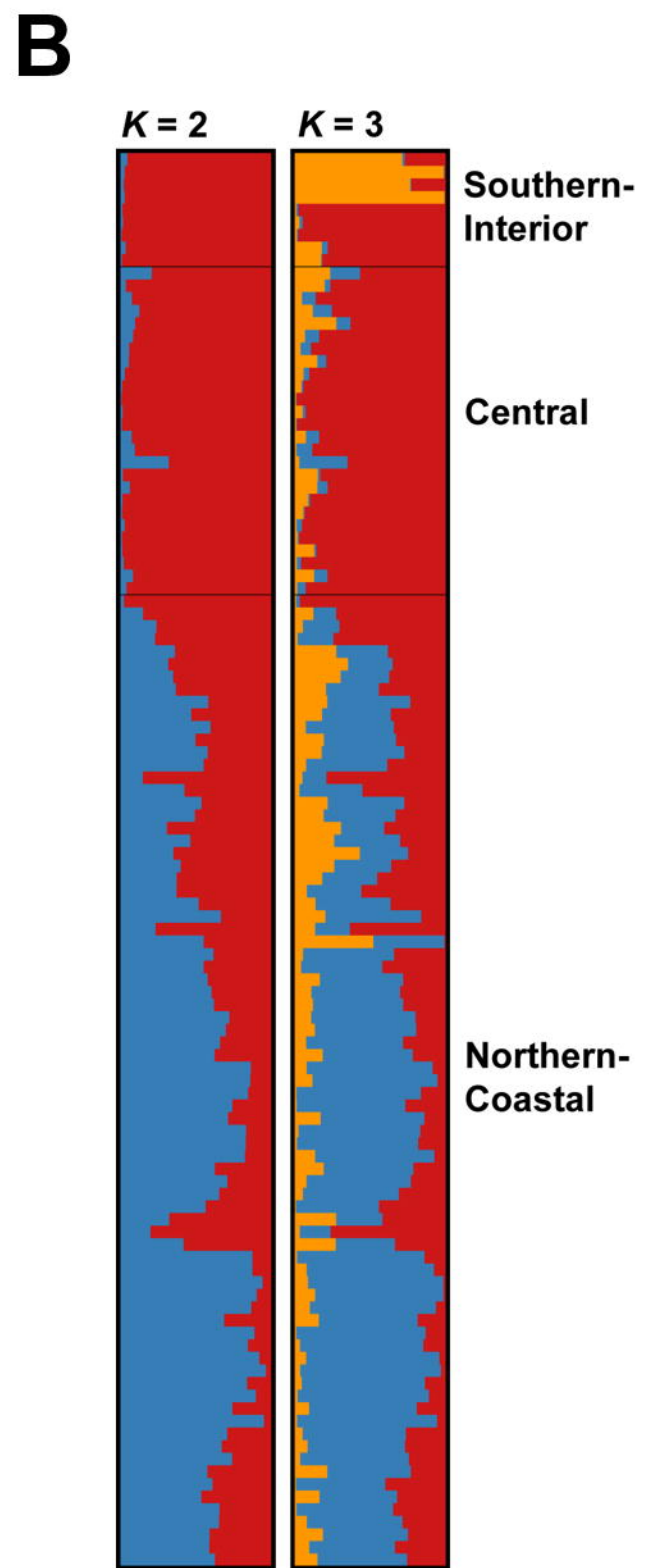
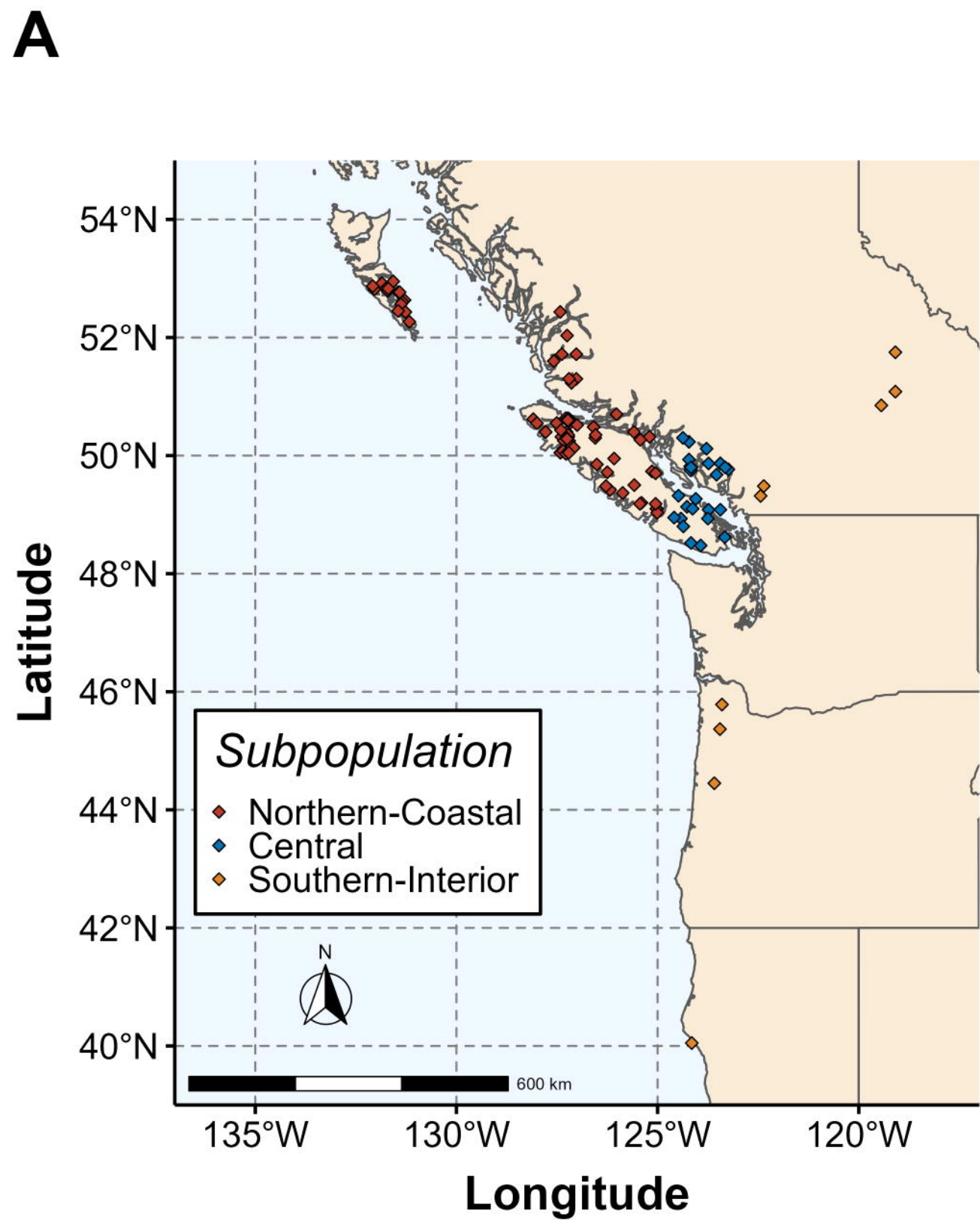
947 **Figure 1: Genetic structure is weak across the geographic range of WRC. A) Map of geographic**
 948 **origin for trees in the range-wide population (RWP) ($n = 112$).** Subpopulations were defined *a priori*
 949 based on analysis outcomes of O'Connell et al. (2008). Trees were separated into three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

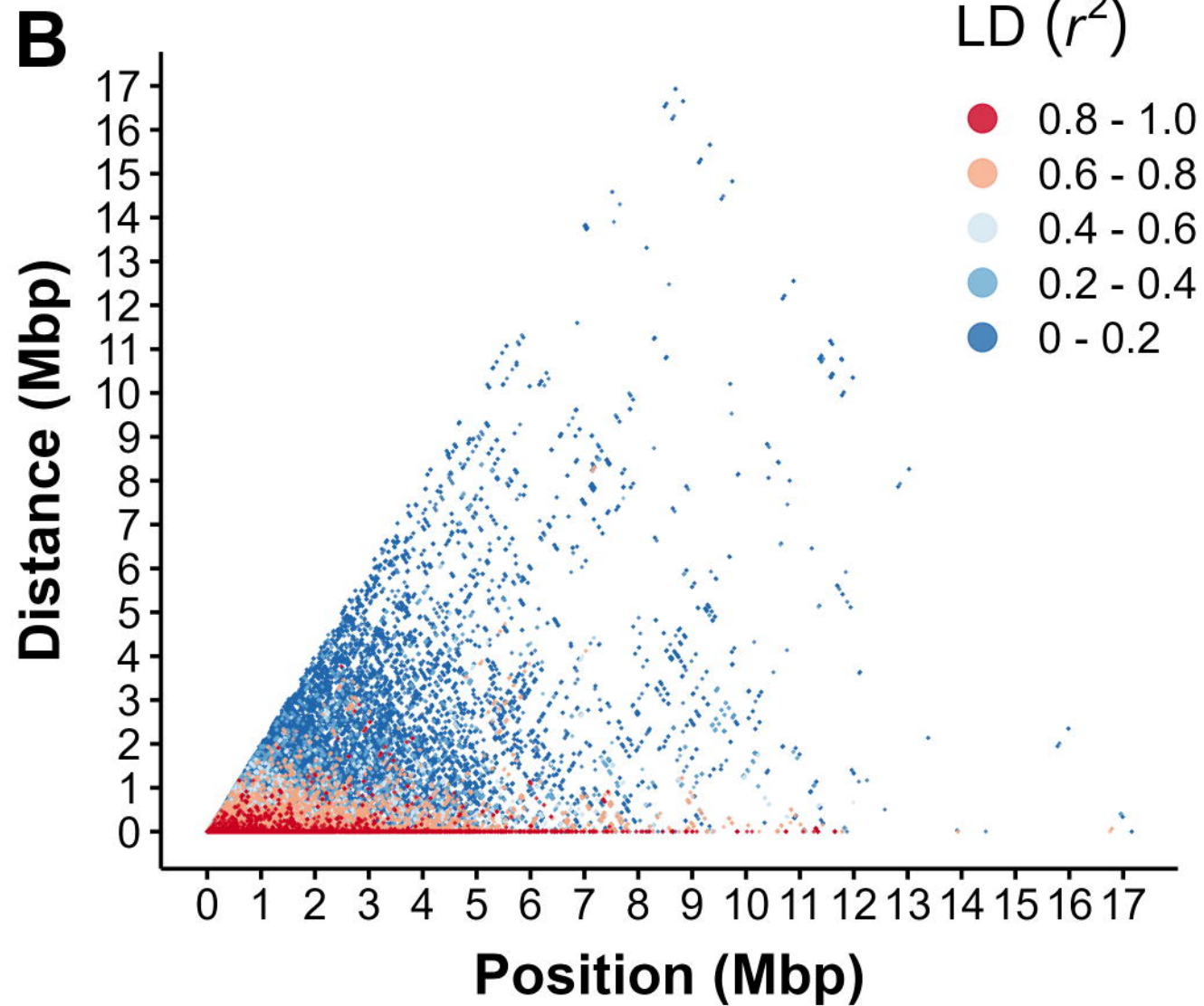
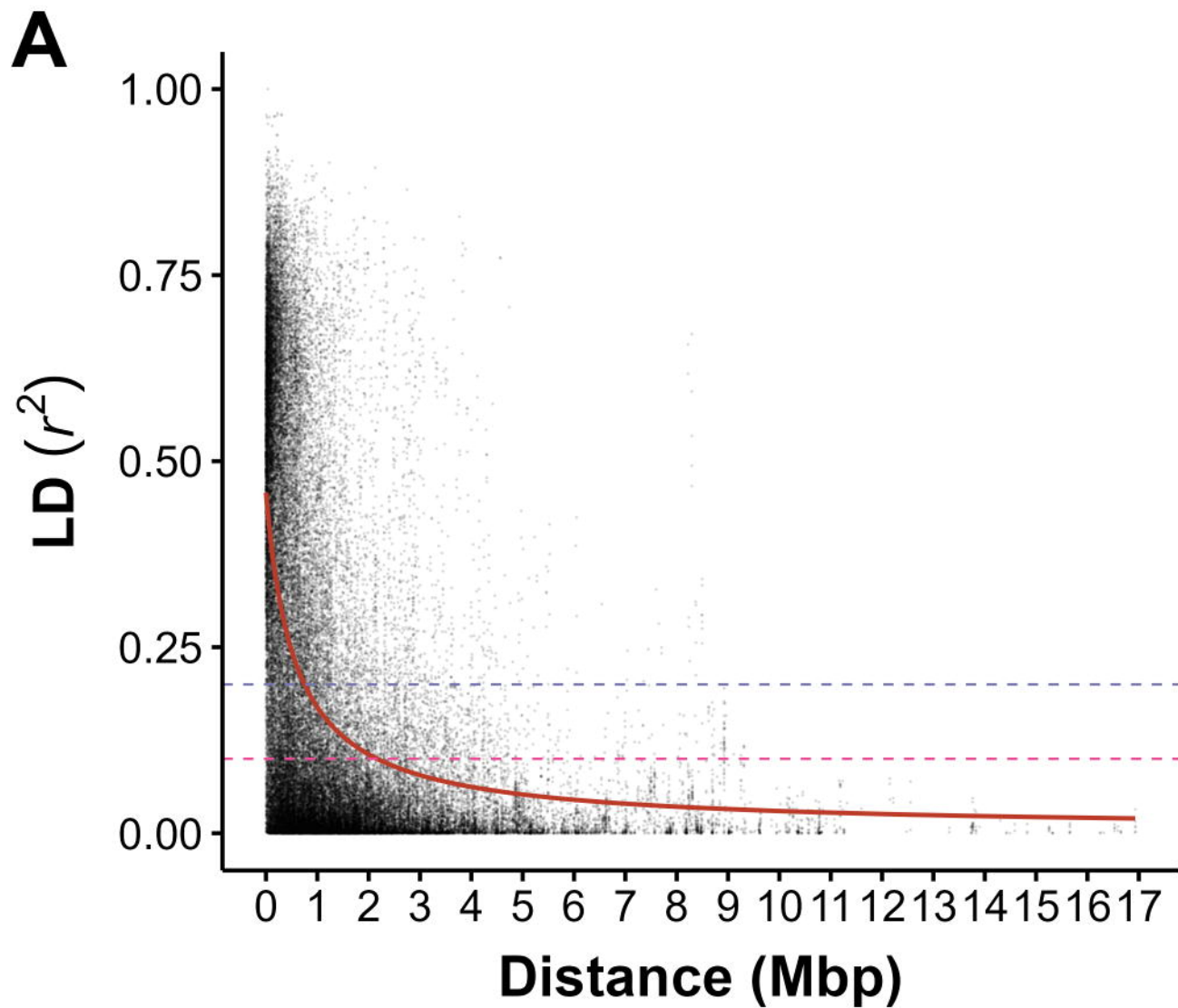
957 **Figure 2: Within-scaffold linkage disequilibrium (LD) in the range-wide population. A)** LD was
 958 assessed using SNPs with a minor allele frequency cutoff of 0.05 to reduce error associated with rare
 959 alleles ($n = 16,202$ SNPs). Decay was estimated using a non-linear model (red line); r^2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

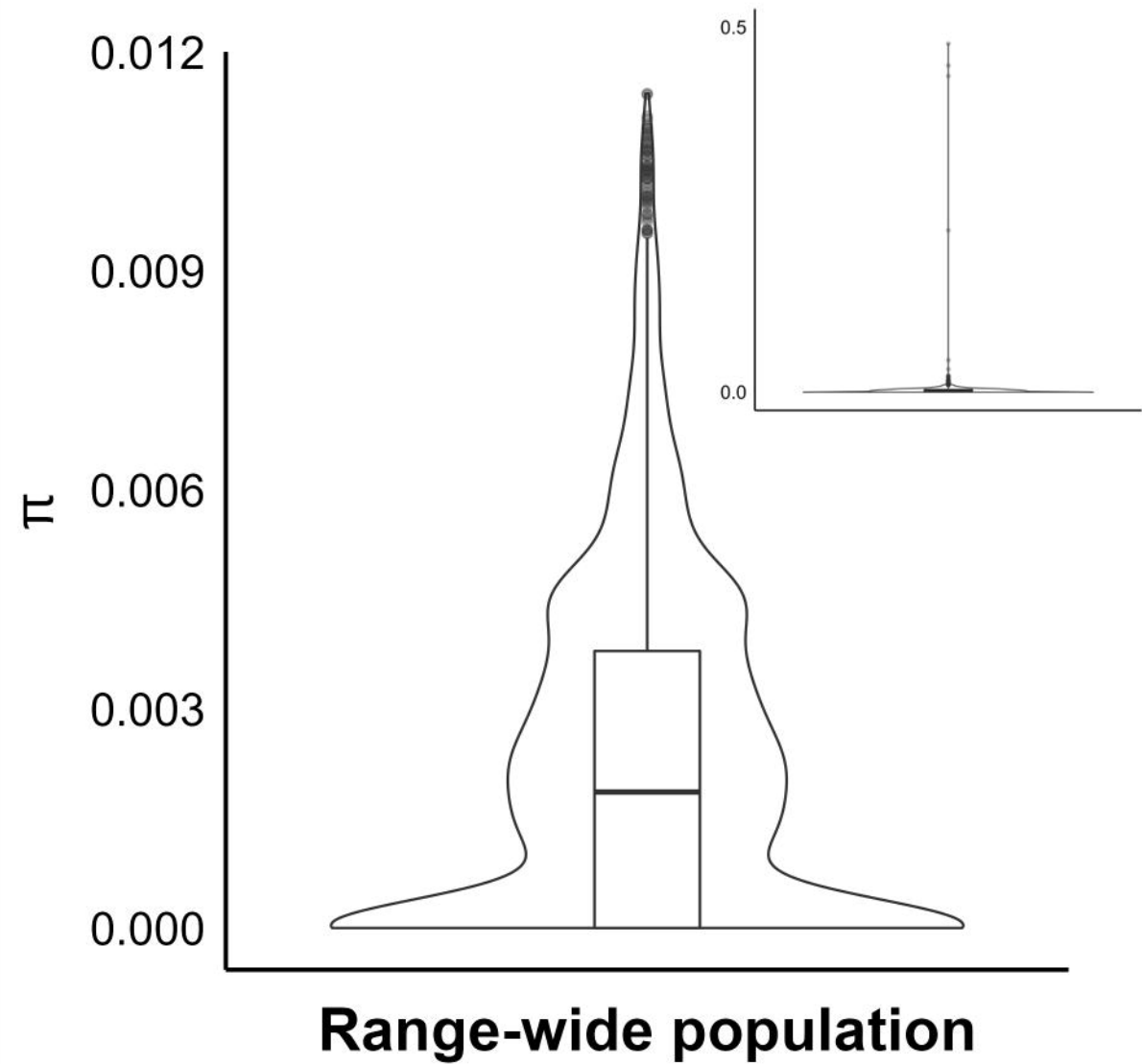
964 **Figure 3: A) Overall distribution of average π of the range-wide population (RWP) in SCOs.** We
 965 detected 1,411 SCOs with a π of zero, with an average π of 0.00272. **B) Overall distribution of average**
 966 **d_{XY} between each pair of geographic subpopulations.** No significant differences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

970 **Figure 4: Change in heterozygosity (H) and inbreeding coefficients (F) over five successive**
 971 **generations of complete selfing in WRC. A)** Observed vs. expected change in heterozygosity over five
 972 successive generations of complete selfing in $n = 28$ (FS – S4) and $n = 11$ (S5) different selfing lines
 973 (SLs), at $n = 18,371$ SNP loci, after manual error correction. Each line at each generation is represented
 974 by a single tree. Black points indicate boxplot outliers. Observed median heterozygosity declines faster
 975 than expected under theoretical expectations during complete selfing, despite many SNP loci remaining
 976 heterozygous across all generations. **B)** Inbreeding coefficients (F) for $n = 28$ samples (FS – S4) and $n =$
 977 11 samples (S5). Black points indicate boxplot outliers. Under complete selfing, F is expected to increase
 978 by a factor of $\frac{1}{2}(1+F)$ in the previous generation (red diamonds). F increases at a slower rate than
 979 expected in our SLs.

980





A**B**