An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce *(Picea abies*)

Carolina Bernhardsson[1,2,3,*], Amaryllis Vidalis[1,4], Xi Wang[1,3], Douglas G. Scofield[1,5,6], Bastian Shiffthaler[7], John Basion[2], Nathaniel R. Street[7], M Rosario García Gil[2], Pär K. Ingvarsson[1,3,*]

[1] Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[2] Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Science, Umeå, Sweden

[3] Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Science, Uppsala, Sweden.

[4] Department of Population Epigenetics and Epigenomics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising,, Germany

[5] Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden

[6] Department of Ecology and Genetics: Evolutionary Biology, Uppsala University, Uppsala, Sweden

[7] Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

*Authors for correspondence: carolina.bernhardsson@umu.se, par.ingvarsson@slu.se

Abstract

Norway spruce (*Picea abies* (L.) Karst.) is a conifer species with large economic and ecological importance. As with most conifers, the *P. abies* genome is very large (~20 Gbp) and contains high levels of repetitive DNA. The current genome assembly (v1.0) covers approximately 60% of the total genome size, but is highly fragmented consisting of more than 10 million scaffolds. Even though 66,632 protein coding gene models are annotated, the fragmented nature of the assembly means that there is currently little information available on how these genes are physically distributed over the 12 *P. abies* chromosomes. By creating an ultra-dense genetic linkage map, we can anchor and order scaffolds at the pseudo-chromosomal level in *P. abies*, which complements the fine-scale information available in the assembly contigs. Our ultra dense haploid consensus genetic map consists of 15,005 markers from 14,336 scaffolds and where 17,079 gene models (25.6% of protein coding gene annotations) have been anchored to the 12 linkage groups (pseudo-chromosomes). Three independent component maps, as well as comparisons to earlier published *Picea* maps are used to evaluate the accuracy and marker order of the linkage groups. We can demonstrate that approximately 3.8% of the scaffolds and 1.6% of the gene models covered by the consensus map are likely wrongly assembled as they

contain genetic markers that map to different regions or linkage groups of the *P. abies* linkage map. We also evaluate the utility of the genetic map for the conifer research community by using an independent data set of unrelated individuals to assess genome-wide variation in genetic diversity using the genomic regions anchored to chromosomes. The results show that our map is dense enough to allow detailed evolutionary analysis across the *P. abies* genome.

Introduction

Genetic linkage maps have been used to order genetic markers and link phenotypic traits to genomic regions and chromosomes by calculating recombination in crosses for over a century (Sturtevant 1913a; Sturtevant 1913b). With the recent development of Next Generation Sequencing technologies (NGS), large numbers of markers can now be scored at a relatively low cost and within a reasonable time, which has enabled the possibility to create high-density genetic maps consisting of thousands of markers that consequently can achieve very high resolutions. These genetic maps enable a complementary approach to the local fine-scale genomic information that is available in the scaffolds of a genome assembly, since a genetic map adds information on genome organization over larger scales (chromosome level) (Fierst 2015). By grouping markers into linkage groups (potential chromosomes), and subsequently ordering them within each linkage group, it is possible to anchor underlying scaffolds to putative chromosomes, here after referred to as pseudo-chromosomes, and align them with high precision (Fierst 2015). If several genetic markers, derived from a single scaffold, are placed on the map, information on their relative placement in the genetic map can be used to orient the scaffold, but also to evaluate scaffolding decisions made in the genome assembly and hence locate and resolve possible assembly errors (Drost et al. 2009; Bartholomé et al. 2015). For

instance, when two markers originating from a single scaffold are mapped to different linkage groups or to different regions within a linkage group, the contigs making up the scaffold have probably been wrongly joined during the assembly process. On the other hand, if markers are placed close to each other on the genetic map this indicates that the scaffolding decision likely was correct.

Norway Spruce (*Picea abies*) is one of the most important conifer species in Europe, both ecologically and economically. With a natural distribution ranging from the west coast of Norway to the Ural mountains and across the Alps, Carpathians and the Balkans in central Europe, it composes, together with *Pinus sylvestris*, the majority of the continuous boreal forests of the Northern hemisphere. For these reasons it is often considered as a key stone species for the region (Farjon 1990). *P. abies* has a genome size of ~20 Gbp that is characterized by very high amounts of repetitive sequences. Like most conifers, *P. abies* has a karyotype consisting of 2n=24 and where chromosomes are all uniformly sized (Sax and Sax 1933). Due to the large and complex genome of conifers, this ecologically and economically important group of plants was, until recently, lacking species with available reference genomes. In 2013 the first draft assembly of the Norway spruce genome was published (Nystedt et al. 2013). Despite extensive whole-genome shotgun sequencing derived from both haploid and diploid tissues, the *P. abies* genome assembly is still highly fragmented due to the complex nature and size of the genome. The current *P. abies* genome assembly (v1.0) consists of 10.3 million scaffolds that are longer than 500 bp and contains 70,736 annotated gene models of which 66,632 are protein coding. Despite the large size of the genome assembly, it still only covers about two thirds of the total genome size (12 Gbp out of the 20 Gbp *P. abies* genome) (Nystedt et al. 2013; De La Torre et al. 2014).

97 In this paper, we use probe capture sequencing to identify segregating SNP
98 markers in an open-pollinated half-sib family. These are used to create an ultra-dense
99 haploid genetic map consisting of 21,056 markers derived from 14,336 gene bearing
100 scaffolds in the Norway spruce (*Picea abies*) genome assembly. Our aim was to 1)
101 anchor and order these scaffolds in an effort to assign as many gene models as
102 possible to pseudo-chromosomes, and 2) to evaluate the accuracy of the *Picea abies*
103 genome assembly v1.0. To evaluate the accuracy of the map itself, we have also
104 performed scaffold order comparisons with previously published genetic maps for *P.*
105 *abies* and *Picea glauca*. Finally we evaluate the utility of the genetic map by
106 performing genome-wide analyses of genetic diversity for the genomic regions
107 anchored in the map in a sample of c. 500 unrelated *P. abies* trees.

108 **Material and Methods**

109 *DNA extraction and exome sequencing*

110 In the autumn of 2013, seeds were collected from cones of 30 putative ramets of
111 Z4006, the individual from which the reference genome for *Picea abies* was obtained
112 (Nystedt et al. 2013), and seeds from five of these ramets were used for the
113 construction of the genetic map. Megagametophytes were dissected from 2,000 seeds
114 by removing the diploid seed coat surrounding the haploid megagametophyte tissue.
115 DNA extraction from megagametophytes was performed using a Qiagen Plant Mini
116 Kit except that the AP1 buffer was replaced by the PL2 buffer from a Macherey-
117 Nagel NucleoSpin Plant II kit. Each extracted sample was measured for DNA quality
118 using a Qubit® ds DNA Broad Range (BR) Assay Kit, and all samples with a total
119 amount of DNA >354 ng were kept. The remaining 1,997 samples were sent to
120 RAPiD Genomics© (Gainesville, Florida, USA) in September 2014 for exome
121 capture sequencing using 31,277 haploid probes that had been specifically designed

122 for *P. abies* based on the v1.0 genome assembly (for further detail of the probes, see
123 Vidalis et al. 2018).

124 The exome capture sequence data was delivered from RAPiD Genomics© in
125 October 2015. The raw reads were mapped against the complete *P. abies* reference
126 genome v.1.0 using BWA-MEM v.0.7.12 (Li and Durbin 2009). Following read
127 mapping the genome was subset to only contain the probe bearing scaffolds (a total of
128 18,461 scaffolds) using Samtools v.1.2 (Li and Durbin 2009; Li et al. 2009). Mark
129 duplicates and local realignment around indels was performed using Picard
130 (http://broadinstitute.github.io/picard/) and GATK (McKenna et al. 2010; DePristo et
131 al. 2011). Genotyping was performed using GATK Haplotypecaller (version 3.4-46,
132 (DePristo et al. 2011; Van der Auwera et al. 2013) with a diploid ploidy setting and
133 gVCF output format. We used a diploid ploidy setting to detect possible sample
134 contamination from diploid tissue for the haploid samples. CombineGVCFs was then
135 run on batches of ~200 gVCFs to hierarchically merge them into a single gVCF and a
136 final SNP call was performed using GenotypeGVCFs jointly on the 10 combined
137 gVCF files, using default read mapping filters, a standard minimum confidence
138 threshold for emitting (stand-emit-conf) of 10, and a standard minimum confidence
139 threshold for calling (stand_call_conf) of 20. See Vidalis et al. (2018) for a full
140 description of the pipeline used for calling variants.

141

142 *SNP filtration and megagametophyte relationships*

143 Sites with insertions/deletions (indels), low quality flag, > 20% missing data, minor
144 allele frequency (MAF) < 0.4 as well as all sites outside the extended probe regions
145 (120 bp probes ±100 bp) were filtered out using vcftools (Danecek et al. 2011). A
146 final filtration step was set so that only markers confirmed as heterozygous in the

147 maternal genotype Z4006 were kept. All heterozygous calls in the haploid samples
148 were then recoded as missing and samples with > 40% missing data were also filtered
149 out to avoid samples with possible contamination of diploid tissue or with poor
150 sequencing quality. This resulted in a final data set of 1,559 samples containing a
151 total of 14,794 SNPs.

152 All 1,559 samples were used in a principal component analysis (PCA) to
153 evaluate the relationship among samples. The reference allele was coded as "0" while
154 the alternative allele was coded as "1", and all remaining missing data were re-coded
155 to the average value for that marker (i.e. the allele frequency of the alternate allele).
156 The first two axes of the PCA explained a total of 17% of the variation (10% and 7 %,
157 respectively for PC1 and PC2) while remaining axes individually explained 0.6-1%.
158 The samples grouped into three distinct clusters which all were oriented differently
159 along the PC1-PC2 axes, with a 4th group connecting the clusters in the center of the
160 plot (Figure S1). The PCA analysis indicate that our data are more heterogeneous than
161 what is expected for a single open-pollinated family, likely indicating that samples
162 came from more than one maternal trees (i.e.,ramets from different genotypes).
163 Samples were therefore split into clusters representing putatively different maternal
164 families using strict cutoffs: Cluster 1 (321 samples) - PC2 >5; Cluster 2 (279
165 samples) – PC1 >0 and PC2< -5; and Cluster 3 (858 samples) - PC1 < -2 (Figure S1).
166 To confirm that these clusters represent single segregating families, PCAs were
167 conducted on all clusters separately. For all three clusters, all axes explained roughly
168 the same amount of variation and all the samples grouped into a single cloud without
169 any detectable outliers (data not shown).

170 Since we detected multiple maternal families in the data set, a second SNP
171 filtration step was performed using vcftools (Danecek et al. 2011) and R (R Core

172 Team 2013) separately on the three clusters, keeping only samples with < 10%
173 heterozygous calls. SNPs within the extended probe regions (see above) having <
174 20% missing data (all calls not homozygous reference or homozygous alternative 1
175 treated as missing) and with a MAF > 0.4 were kept as informative markers
176 (supplementary file: Informative markers). For each unique probe in the three data
177 sets, only the most balanced marker (highest MAF and lowest amount of missing
178 data) was kept for map creation and named with an ID based on scaffold and probe
179 position. This resulted in 9,073 markers from 7,101 scaffolds for Cluster 1 (314
180 samples), 11,648 markers from 8,738 scaffolds for Cluster 2 (270 samples) and
181 19,006 markers from 13,301 scaffolds for Cluster 3 (842 samples) for a total of
182 21,056 markers from 14,336 scaffolds across all three clusters (Table 1). In total,
183 these scaffolds cover 0.34 Gb of the *P. abies* genome and contain 17,079 protein
184 coding gene models.

185 **Table 1:** Overview of the three component maps and the total number of
186 markers available in the consensus map. Cluster: Name of each family group
187 that was identified in the principal component analysis. Samples: Number of
188 megagametophytes in each cluster. Marker pre drop/ Markers post drop:
189 Number of markers in each component map before and after markers were
190 dropped if markers from the same scaffold were located within 15 cM from
191 each other in the first round of component map construction. Scaffolds:
192 Number of scaffolds represented in each component map.

Cluster	Samples	Markers pre-drop/post-drop	Scaffolds
Cluster 1	314	9,073 / 7,179	7,101
Cluster 2	270	11,647 / 8,821	8,738

Cluster 3	842	19,006 / 13,479	13,301
Total	1,426	21,056 / 15,005	14,336

193

194 *Component and consensus maps*

195 Genetic linkage maps were created with the R-package BatchMap (Schifthaler et al.
196 2017), a parallel implementation of the R-package Onemap (Margarido, Souza, and
197 Garcia 2007). All markers were recoded using the D1.11 cross-type (Wu et al. 2002)
198 and grouped into LGs with LOD = 8 and a maximum recombination fraction = 0.35.
199 LGs were then ordered using the RECORD algorithm (Van Os et al. 2005) with 40
200 times counting, parallelized over 20 cores, and mapped using the Kosambi mapping
201 function and the map batches approach (Schifthaler et al. 2017) over four parallel
202 cores. To reduce the noise in the maps, markers from the same scaffold that mapped
203 within 15 cM from each other, were dropped so that only one marker was used to
204 represent the scaffold in the final map. However, if any markers from the same
205 scaffold mapped more than 15 cM apart, all markers from that scaffold were kept.
206 This approach was motivated by the fact that sequence data from markers < 15 cM
207 apart did not show any evidence for recombination when using a visual inspection of
208 the data and that this inconsistency in marker ordering is probably due to a lack of
209 resolution in the mapping populations together with the usage of a heuristic ordering
210 approach (Mollinari et al. 2009). Finally, a heat map with pairwise recombination
211 fraction (lower triangular) and phase LOD score (upper triangular) for the ordered
212 markers was created to evaluate the ordering accuracy (data not shown).

213 To evaluate correspondence between LGs in maps from different clusters the
214 number of unique scaffolds shared between cluster LGs were counted (Figure S2). A

215 consensus map over all three clusters was then created for each chromosome with the
216 R-package LPmerge (Endelman and Plomion 2014) with clusters ordered according
217 to marker numbers, a maximum interval setting ranging from one to 10 and map
218 weights proportional to sample size. The consensus map with the lowest mean root
219 mean square error (RMSE), was then set as the best consensus map for each
220 chromosome. Order correlations between component maps and the consensus maps
221 were estimated with Kendall's tau (Table 2 and Figure S3a-l). For visual
222 representation of the consensus map and the characteristics of the anchored genomic
223 scaffolds we created a Circos plot using the R-package omicCircos (Hu et al. 2014),
224 available from Bioconductor (https://bioconductor.org/biocLite.R).

225

226 *Accuracy of the reference P. abies genome assembly and distribution of*
227 *recombination hot spots/cold spots*

228 To evaluate the accuracy of the *P. abies* reference genome v1.0, scaffolds carrying at
229 least two markers (here after called multi-marker scaffolds) were used to determine
230 whether markers were positioned in the same region of an LG, on different regions
231 from a single LG or on different LGs. In the consensus map, we considered markers
232 to be positioned in the same region on an LG if all markers from a scaffold mapped
233 within a 5 cM interval of each other. If any marker from the scaffold was positioned
234 further apart, the scaffold was tagged as a likely wrongly assembled scaffold. The
235 same considerations were made for scaffolds with markers positioned on different
236 LGs.

237 To analyze the distribution of recombination hot spots/cold spots, a sliding
238 window analysis using a window size of 5 cM was performed along the LGs of the
239 consensus map. In each window, the total physical length of all unique scaffolds

240 located within the window as well as the number of scaffolds and corresponding gene
241 models, was counted.

242

243 *Comparative analyses of Picea linkage maps*

244 To evaluate the consistency of our genetic map with earlier maps from *P. abies* we
245 compared our haploid consensus map to the *P. abies* linkage map from Lind et al.
246 2014. The Lind et al (2014) map was created using genetic markers generated using a
247 *P. glauca* SNP array (Pavy et al. 2013). The SNP array sequences from the *P. glauca*
248 array were blasted (tblastn) against the *P. abies* v1.0 genome assembly and reciprocal
249 best hits with >95% identity were extracted and assigned to the corresponding
250 scaffold in the *P. abies* genome. We performed similar analyses to also compare the
251 synteny between our consensus map and the *P. glauca* composite map from Pavy et al.
252 2017. Again, array sequences from the *P. glauca* SNP array (Pavy et al. 2013) were
253 blasted against the *P. abies* 1.0 genome and reciprocal best hits were assigned the
254 corresponding map positions from *P. abies* and *P. glauca*. In order to evaluate which
255 LGs that correspond to the same chromosome, we assessed the number of shared
256 scaffolds between our consensus map, the Lind et al. 2014 and Pavy et al. 2017 maps.
257 Consistency of scaffold order where then evaluated using a visual comparison (Figure
258 3 and 4) and by calculating correlations of marker order using Kendall's tau.

259

260 *Population genetic analysis of the consensus genetic map*

261 In order to independently evaluate the utility of the consensus map for downstream
262 research, we used a subset of the data from Baison et al. (2018) to estimate patterns of
263 nucleotide diversity across the Norway spruce genome. The data from Baison et al.

264 (2018) originally contained 517 individuals sequenced with 40,018 probes designed
265 for diploid spruce samples (Vidalis et al 2018). We extracted data for the probes that
266 were anchored in our genetic map and further hard filtered the data by only
267 considering bi-allelic SNPs within the extended probe regions (120bp probes ±100bp)
268 with a QD >5, MQ >50 and a overall DP between 3000 and 16000. Samples showing
269 >25% missing data were also removed from further analysis. We used the data to
270 calculate nucleotide diversity (π), the number of segregating sites and Tajima's D. We
271 used the R package vcfR (Knaus and Grünwald 2017) to read the VCF-file into R and
272 then used in-house developed scripts to perform all calculations. We assigned probes
273 to LGs and map positions by assigning them the coordinates of the physically closest
274 (in bp) probe. We also calculated pairwise linkage disequilibrium (LD) between
275 markers within probes using vcftools (Danecek et al. 2011) and imported the results
276 into R where they were used to calculate Zn scores (Kelly 1997) per probe using an
277 in-house developed script. Finally we ran sliding window analyses along the pseudo-
278 chromosomes for the different summary statistics using 10cM windows that were
279 moved in 1 cM incremental steps.

280

281 **Results**

282 A *P. abies* consensus linkage map was generated from three haploid component maps
283 containing a total of 15,005 unique markers from 14,336 gene containing scaffolds
284 from the *P. abies* genome assembly v1.0. The consensus map anchors 0.34 Gbp of the
285 *P. abies* 1.0 assembly, corresponding to only 1.7% of the complete *P. abies* genome
286 or 2.8% of the assembled genome. However, these scaffolds anchor 25.6% of all
287 predicted protein coding genes in *P. abies* and the the anchored scaffolds harbor
288 31.7%, 20.6% and 25.8% of the High-, Medium- and Low confidence gene models

289 from Nystedt et al (2013), respectively. The consensus map has a total length of 3,326
290 centiMorgan (cM), distributed over 12 linkage groups (LGs), which corresponds to
291 the known haploid chromosome number of Norway spruce (Sax and Sax 1933), and
292 with an average marker distance of 0.22 cM/marker (Table 2, Figure 1: track a).

293 Correlations of marker order between the three component maps and the
294 consensus map ranged from 0.96 to 0.998, while the correlations between marker
295 orders between individual component maps ranged from 0.943 to 0.993 (Table S1 and
296 Figure S3). LG XI, which display the largest discrepancy in marker order between
297 component maps, has a 200 marker region in the distal end of the chromosome where
298 the resolution is too low to identify a correct order and where the whole region is
299 positioned at 0 cM (Figure S3k), explain the lower order correlations for this LG.

300

301 **Table 2:** Marker density and size of each component genetic map created
302 from the three clusters as well as for the consensus map. LG: Linkage group.
303 Cluster 1-3: Component maps for cluster 1-3 with number of markers
304 assigned and map size (in cM) for each of the LGs. Consensus: Number of
305 markers and map size of the LGs in the consensus map.

LG	Cluster 1		Cluster 2		Cluster 3		Consensus	
	Markers	Length (cM)	Markers	Length (cM)	Markers	Length (cM)	Markers	Length (cM)
I	768	403.2	867	440.3	1,373	358.0	1,520	359.3
II	570	273.8	669	294.2	1,042	265.6	1,172	265.6
III	682	321.0	813	388.7	1,232	304.4	1,379	304.4
IV	602	315.1	718	353.0	1,078	271.5	1,199	271.5

V	593	278.4	815	401.2	1,160	309.6	1,305	299.7
VI	510	257.8	685	275.2	1,017	241.3	1,142	241.3
VII	532	324.0	688	395.5	1,141	275.9	1,245	275.9
VIII	613	325.2	710	361.6	1,048	279.6	1,158	279.5
IX	623	300.8	610	314.0	1,122	247.3	1,244	247.3
X	504	267.4	745	356.8	1,118	234.7	1,229	265.9
XI	553	216.0	774	304.7	1,040	205.3	1,167	205.2
XII	629	310.4	727	387.2	1,108	289.3	1,245	310.7
Total	7,179	3,592.9	8,821	4,262.5	13,479	3,282.4	15,005	3,326.3

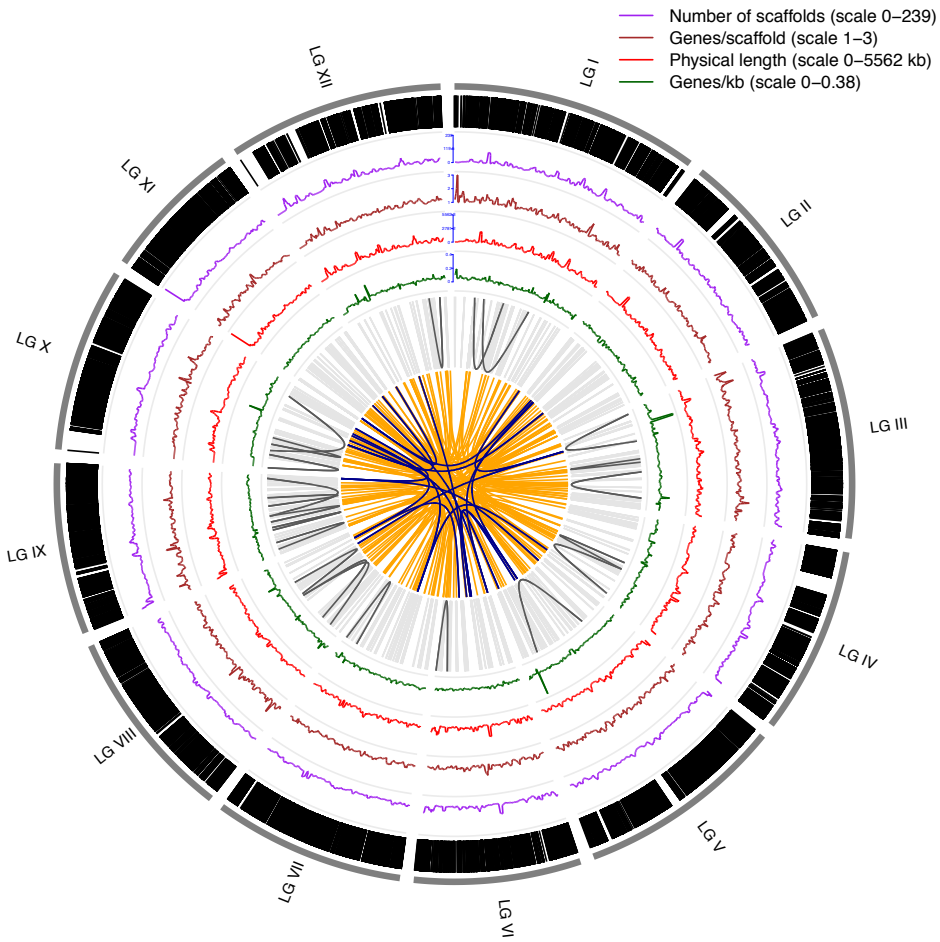


Figure 1: Circos plot of the consensus map. A) Marker distribution over the 12 linkage groups (LG I-LG XII). Each black vertical line represents a marker (15,005 in total) in the map and is displayed according to the marker positions in cM. Track B-E visualizes a sliding window of size 5 cM, with 1 cM incremental steps, along the linkage groups. B) Number of scaffolds, scaling 0-239. C) Number of gene models/scaffold, scaling 1-3. D) Physical length of scaffolds, scaling 0-5,562 kb. E) Number of gene models/kb, scaling 0-0.38. Track F-G visualizes multi marker scaffolds, where each line is a pairwise position comparison of markers from the same scaffold. F) Position comparisons of markers from the same scaffold that are located on

319 the same LG. Light grey lines indicate markers that are located < 5cM from
320 each other while dark grey lines indicate markers located > 5cM apart. G)
321 Position comparisons of markers from the same scaffold that are located on
322 different LGs. Orange lines indicated markers from the same scaffold split
323 over 2 LGs, while dark blue lines indicated markers split over 3 LGs.

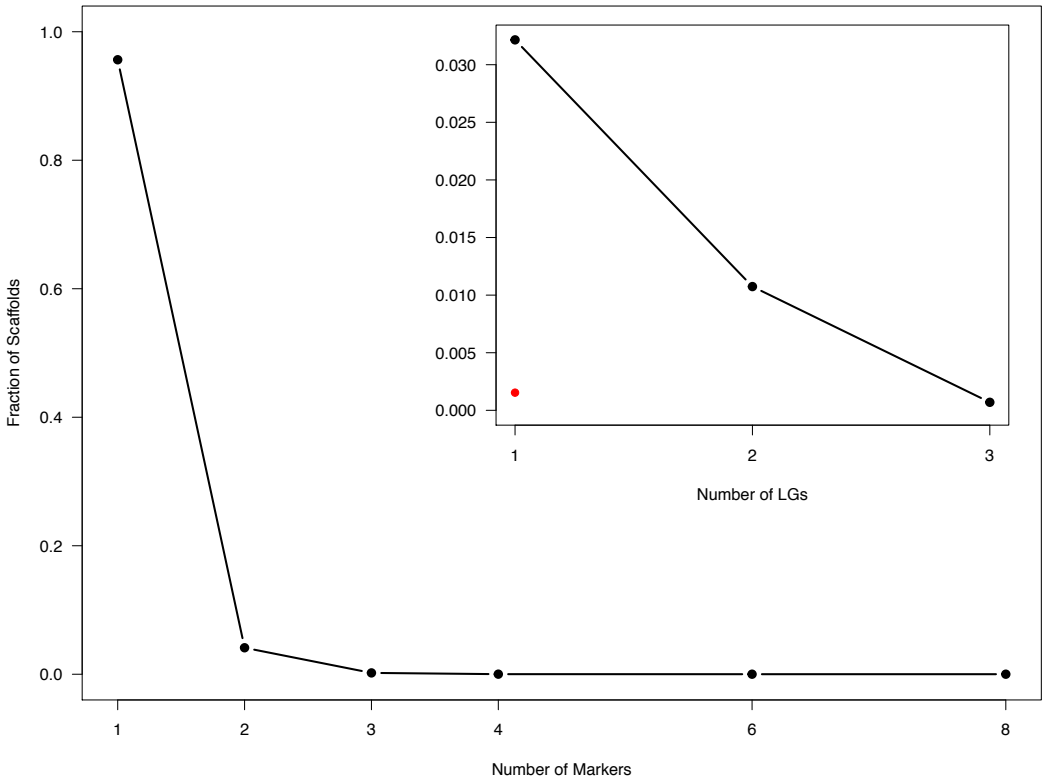
324

325 *Evaluation of the P. abies genome assembly v1.0*

326 The average physical size of the scaffolds anchored per LG is 29 Mbp (26.1 - 35.3
327 Mbp). All chromosomes show variation in marker density along the linkage groups,
328 but number of markers, scaffolds, gene models and physical size are all highly
329 correlated (Figure 1: track b-e). However, a few regions show higher recombination
330 rates than the rest of the genome, where short physical length (in Kbp) co-occur with
331 high gene density (number of gene models/Kbp) (Figure 1: track d and e). The
332 average gene density is 0.05 genes/Kbp (0.047 – 0.059 per LG) with a standard
333 deviation of 0.02 (0.01 – 0.04 per LG). 1.41% of the windows have > 0.1 genes/Kbp
334 and 0.24% have > 0.2 genes/Kbp. The highest gene density can be seen in regions on
335 LGIII and LGV with 0.37 genes/Kbp. These regions contain one and two scaffolds,
336 respectively, are present in one or two of the three component maps and contain one
337 gene model each.

338 4,859 scaffolds (33.9%) had more than one unique marker combined over all
339 three component maps before marker pruning. Of these, 625 scaffolds (4.36%) had
340 multiple markers also in the consensus map, either due to suspicious grouping and/or
341 ordering in the component maps or that different markers were represented in
342 different component maps. 186 of these multi-marker scaffolds show a split over
343 several LGs (inter-split scaffolds) or over different parts of the same LG (intra-split

344 scaffolds). 22 scaffolds (0.15% of mapped scaffolds and 0.45% of original multi-
345 marker scaffolds) have markers positioned > 5 cM apart on the same LG and 164
346 scaffolds (1.14% of mapped scaffolds and 3.38% of original multi-marker scaffolds)
347 have markers mapped to 2 or 3 different LGs (Figure 2 and Table S2). All LGs harbor
348 inter-split scaffolds, while 10 LGs (LGII and LGXI are the exceptions) harbors intra-
349 split scaffolds (Figure 1: track f and g).

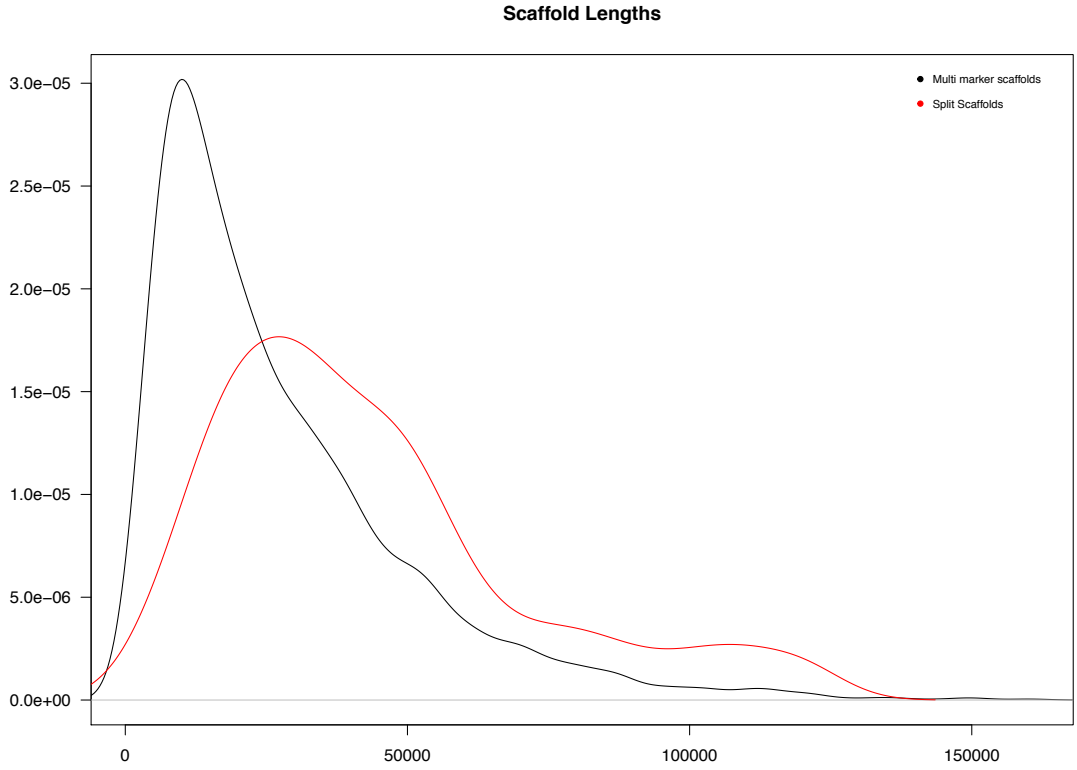


350

351 **Figure 2:** Fraction of scaffolds that are being represented by 1-8 unique
352 markers in the consensus map. Insert: Fraction of total number of scaffolds
353 that have multiple markers (2-8) that are distributed over 1-3 linkage
354 groups (inter-split scaffolds). Red dot indicate the fraction of scaffolds with
355 multiple markers which are positioned > 5cM apart on the same linkage
356 group (intra-split scaffolds).

357

358 The scaffolds covered by the map range in length from 0.22 to 208.1 Kbp with a
359 median of 17.1 Kbp, while multi-marker scaffolds range from 0.39 to 161.5 Kbp
360 (median of 21 Kbp) in length. The 186 scaffolds that are split within or across LGs
361 range in size from 2.5 to 121.6 Kbp, with a median length of 36.9 Kbp. Split scaffolds
362 are significantly longer than the multi-marker scaffolds in general (t = -7.76, df =
363 194.54, p-value = 4.77e-13; Figure 3), suggesting that longer scaffolds more often are
364 prone to assembly errors compared to shorter scaffolds. Split scaffolds are mostly
365 harboring high- and medium confidence gene models (Table 3). A visual inspection
366 of the split scaffolds shows that 75 and 10 of the inter-split and intra-split scaffolds,
367 respectively, have the predicted split(s) between different gene models on the same
368 scaffold where as 88 of the inter-split scaffolds and 12 of the intra-split scaffolds have
369 the predicted split within a single gene model (Table S3). In addition, 21 inter-split
370 scaffolds show an even more complicated picture, where an interior regions of the
371 gene model (most often containing an intron > 5kb) map to another chromosome
372 where as the 5' and 3' regions of the gene model map to the same chromosome
373 location (Table S3). Of the 17,079 gene models that are anchored to the consensus
374 genetic map, 330 are positioned on inter- or intra-split scaffolds (5.4% of those gene
375 models that are positioned on originally multi-marker scaffolds) and 100 show a split
376 within gene models (1.6% of gene models from multi-marker scaffolds) (Table 3).



377

378 **Figure 3:** Kernel density estimate of scaffold lengths for all multi-marker
379 scaffolds (black line) and for scaffolds showing a split within or across LGs
380 (red line). The split scaffolds are significantly longer than the multi-marker
381 scaffolds in general (t = -7.76, df = 194.54, p-value = 4.77e-13).

382

383 **Table 3**: Overview of annotated gene models anchored to the genetic map.
384 Gene models: Annotated protein coding gene models with High-, Medium-
385 and Low confidence level (Nystedt et al. 2013). Mapped scaffolds: Number
386 of gene models positioned on scaffolds that are anchored to the genetic map
387 (Percentage of total number of gene models for each confidence level).
388 Multi-marker scaffolds: Number of gene models positioned on scaffolds with
389 multiple markers in the genetic map (Percentage of gene models on mapped
390 scaffolds). Inter-split scaffolds: Number of gene models positioned on the
391 164 scaffolds that are split between LGs in the genetic map (Percentage of

392 gene models on mapped scaffolds / Percentage of gene models on multi-
393 marker scaffolds). Intra-split scaffolds: Number of gene models positioned
394 on the 22 scaffolds that are split between different regions of the same LG
395 (Percentage of gene models on mapped scaffolds / Percentage of gene
396 models on multi-marker scaffolds). Split within gene models: Number of
397 gene models that have an internal split (Percentage of gene models on
398 mapped scaffolds / Percentage of gene models on multi-marker scaffolds).

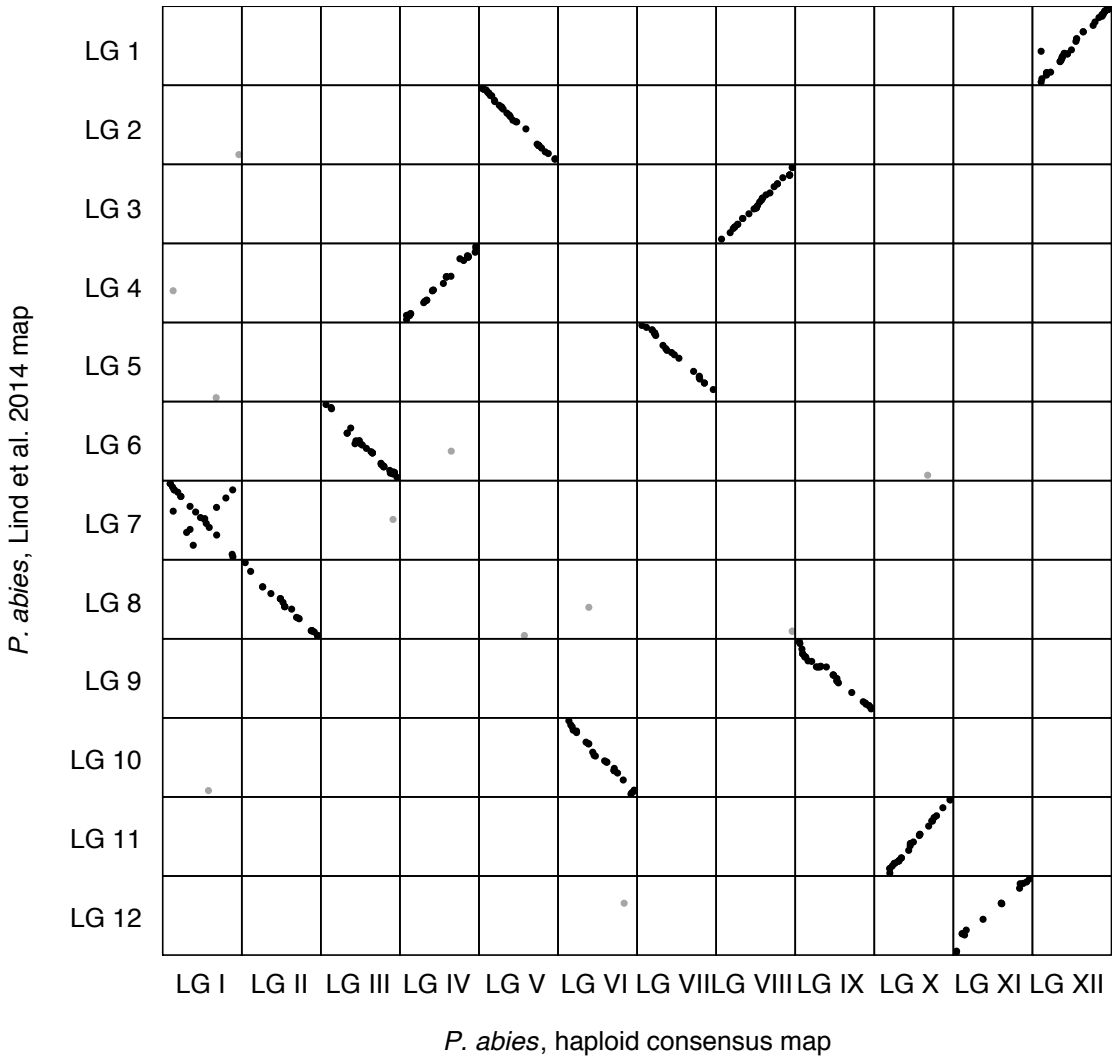
Gene models	**Mapped scaffolds**	**Multi-marker scaffolds**	**Inter-split scaffolds**	**Intra-split scaffolds**	**Split within gene models**
High confidence	8,379 (31.7%)	3,122 (37.3%)	145 (1.7% / 4.6%)	15 (0.18% / 0.48%)	58 (0.69% / 1.9%)
Medium confidence	6,624 (20.6%)	2,215 (33.4%)	114 (1.7% / 5.1%)	15 (0.23% / 0.68%)	29 (0.44% / 1.3%)
Low confidence	2,076 (25.8%)	762 (36.7%)	35 (1.7% / 4.6%)	6 (0.29% / 0.79%)	13 (0.63% / 1.7%)
Total	17,079 (25.6%)	6,099 (35.7%)	294 (1.7% / 4.8%)	36 (0.21% / 0.59%)	100 (0.59% / 1.6%)

399

400 *Comparative analyses to other Picea linkage maps*

401 In order to assess the accuracy and repeatability of the *P. abies* genetic maps we
402 compared our consensus map to a *P. abies* QTL map from Lind et al. (2014). This
403 map consists of 686 markers, genotyped in 247 offspring from a full sib family using
404 markers derived from a *P. glauca* SNP array. 353 comparisons between 298 markers

405 from Lind et al. (2014) and 288 scaffolds contained in our consensus map were
406 identified at a > 95 % identity threshold. Of these markers, 96.7% group to the same
407 chromosome in the two maps while the remaining 3.3% (11 out of 353) are
408 distributed across several linkage groups (Figure 4). Correlations of marker order
409 between the two *P. abies* maps ranged from 0.53 to 0.99 across the 12 LGs. The
410 comparison between the haploid consensus map LG I and LG 7 from Lind et.al
411 (2014), which has the lowest correlation of marker order, show inconsistencies of
412 marker order where several markers are arranged in the opposite order from the rest
413 of the markers. The remaining chromosomes show high synteny with a consistent
414 marker order between the two genetic maps.



415

416 **Figure 4:** Marker order comparison between the haploid consensus map
417 and the *P. abies* map from Lind et al. 2014. Consensus LG I - LG XII are
418 located on the x-axis from left to right. Lind et al. 2014 LG 1 - LG 12 are
419 located on the y-axis from top to bottom. Each dot represents a marker
420 comparison from the same scaffold, where black coloration displays the LG
421 where the majority of marker comparisons are mapped. Grey coloration
422 display markers mapping to a different LG compared to the majority of
423 markers.

424

425 Synteny between *P.* abies and *P. glauca* species was assessed by comparing
426 chromosome location and marker order between our *P. abies* consensus map and the
427 composite map of *P. glauca* from Pavy et al. (2017). 11,458 comparisons from 4,934
428 gene models in the composite map in *P. glauca* (Pavy et al. 2017) and 5,451 scaffolds
429 in the *P. abies* consensus map could be retrieved. 93.3% (10,733 out of 11,458 hits)
430 of these were found to be located on homologous chromosomes while the remaining
431 6.7% (725 comparisons) are distributed across the 12 linkage groups (Figure 5). The
432 correlations of marker order between the two maps were comparable to the
433 corresponding correlations between component maps in *P. abies* showing that synteny
434 is largely conserved between *P. abies* and *P glauca*.

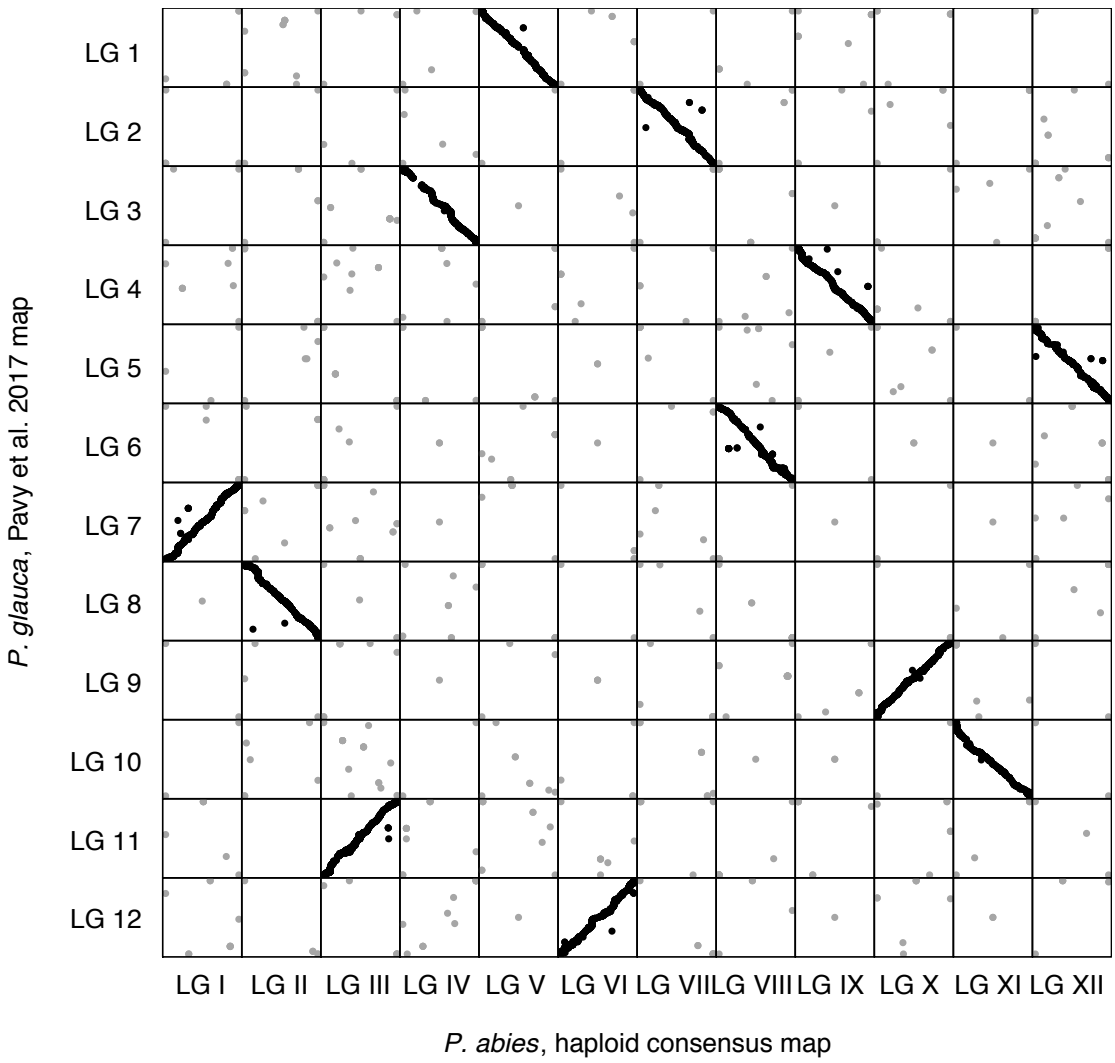


Figure 5: Marker order comparison between the haploid consensus map and the *P. glauca* map from Pavy et al. 2017. Consensus LG I - LG XII are located on the x-axis from left to right. Pavy et al. 2017 LG 1 - LG 12 are located on the y-axis from top to bottom. Each dot represents a marker comparison from the same scaffold, where black color display markers mapping to the same LG in the two species where as grey color indicate markers mapping to different LGs.

444 *Population genetic analyses based on the consensus map*

445 22,413 probes, covering 12,908 scaffolds, were used in the population genetic
446 analyses based on the consensus genetic map. On a per probe basis, we observe
447 substantial variation in all neutrality statistics, with the number of segregating sites
448 ranging from 0 - 77 (mean 15.9), nucleotide diversity (π) from 0 - 0.4 (0.005), Zns
449 from 0 - 1 (mean 0.04) and Tajima's D from -2.4 – 3.5 (mean -0.85). To study large-
450 scale trends and possible chromosomal differences we performed sliding window
451 analyses across the linkage groups for the different summary (Figure 6). One
452 interesting large-scale feature we observe is that SNP densities are often highest at the
453 distal or central regions of linkage groups, indicating the possible location of
454 centromers and telomeres where recombination rates are expected to be reduced (Gaut
455 et al 2007) and where we hence would expect higher densities of probes per cM
456 (Figure 6a). The large-scale analyses also reveal several instances where entire
457 chromosomal arms might be under different evolutionary regimes (Figure 6b-c).
458 Finally we can identify regions that appear to be evolving under the influence of
459 natural selection. For instance, several regions show higher than average levels of
460 nucleotide diversity and positive Tajima's D (eg. on LG IV, V and XII), suggesting
461 that they might harbor genes under balancing selection. Similarly, regions with low
462 nucleotide diversity, an excess of rare alleles and strong linkage disequilibrium (i.e.
463 negative Tajima's D and high Z_{ns} scores, e.g. on LG III) could indicate regions of
464 possible selective sweeps (Figure 6c-d).

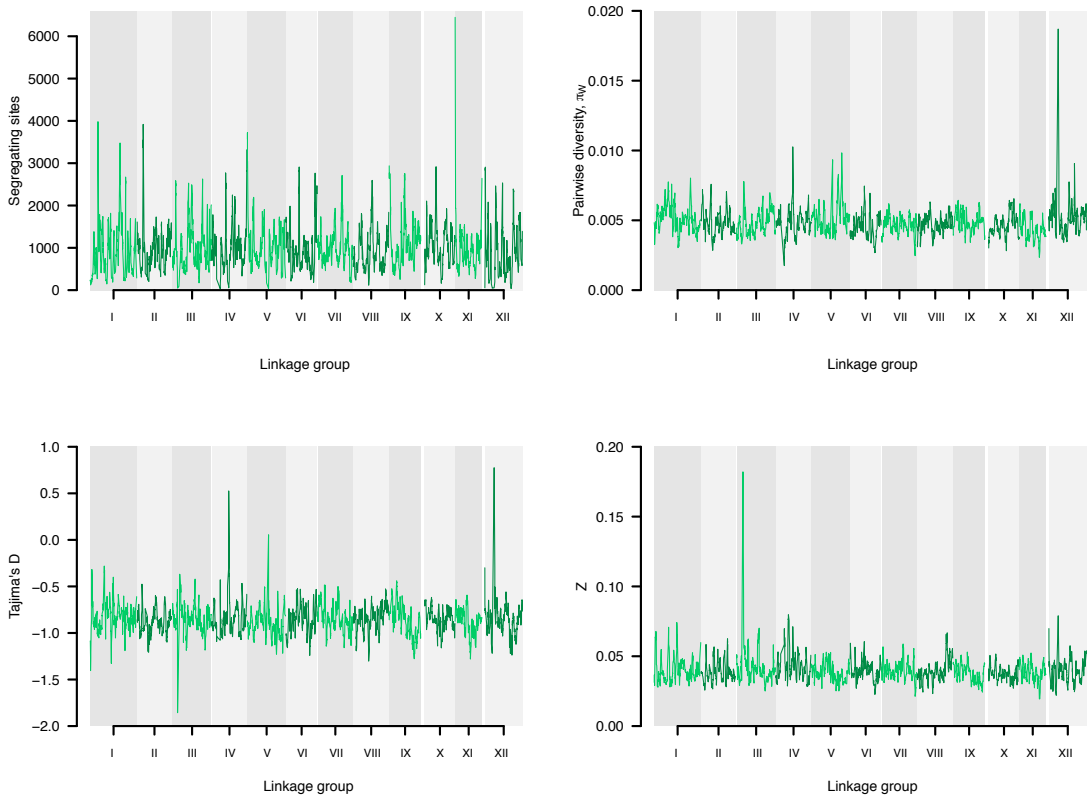


Figure 6. Sliding window analysis of neutrality statistics. Analyses were performed using 10 cM windows with 1 cM incremental steps along the consensus map linkage groups. A) Number of segregating sites. B) Pairwise nucleotide diversity (π). C) Tajima's D and D) Linkage disequilibrium Zn scores.

Discussion

This is, to our knowledge, the densest genetic linkage map ever created for a conifer species and possible even for any tree species. We have successfully used this genetic map to anchor 1.7% of the 20 Gbp *P. abies* genome, corresponding to 2.8% of the v1.0 genome assembly (Nystedt et al. 2013), to the 12 linkage groups that constitute the haploid chromosome number in spruces (Sax and Sax 1933). The Norway spruce genome has a very large proportion of gene-poor heterochromatin, so while the

479 fraction of the genome that we successfully anchor to the assembly may seem small,
480 these scaffolds cover 24% of gene bearing scaffolds and 25% of all protein coding
481 gene models from Nystedt et al. (2013).

482 The individual linkage groups from the three component maps (36 LGs from
483 three independent maps) consists of 648-1,967 markers before and 504-1,373 markers
484 after marker elimination and it is, therefore, not feasible to analyze the maps using an
485 exhaustive ordering algorithm (Mollinari et al. 2009). Instead, we decided to use
486 RECORD (Van Os et al. 2005) with 40 times counting, parallelized over 20 cores, for
487 each linkage group to find the most likely marker order. A heuristic approach, such as
488 RECORD, will undoubtedly introduce some errors in marker ordering, but analyses
489 from simulated data suggest that the distance between estimated and true marker
490 position is quite small (20-30 markers) for a data sets of similar size as ours
491 (Schiffthaler et al. 2017). However, reliable marker ordering require robust data and
492 the more genotyping errors and missing data that is present the harder it will be to
493 find the true order. This in turn will impact the final size of the map, where both
494 errors in marker order and genotyping results in inflation in the size of the map
495 (Cartwright et al. 2007).

496 By collecting our 2,000 megagametophytes from what were initially thought to
497 be five different ramets of Z4006 we accidentally sampled three unrelated families.
498 This error stemmed from a mix-up of genotypes due to wrong assignment of ramet ID
499 to the different ramets in the seed orchard. Unfortunately, we were not able to assess
500 which megagametophytes that were collected from the different putative ramets since
501 seeds were pooled prior to DNA extraction and the sampling errors were not detected
502 until after all sequencing was completed. We used a PCA to assign samples into three
503 independent clusters and used subsequent PCAs of the putative individual families to

504 verify the reliability of these clusters. However, we cannot completely rule out that a
505 small fraction of samples have been wrongly assigned to the three families and this
506 would further inflate map size by introducing excess recombination events. Another
507 potential confounding issue is tissue contamination. Norway spruce
508 megagametophytes are very small and are surrounded by a diploid seed coat that
509 needs to be removed before DNA extraction. If traces of the diploid seed coat remain
510 in the material used for DNA extractions, the haploid samples will be contaminated
511 with diploid material. To identify and eliminate this possibility, we called sequence
512 variants using a diploid model and any heterozygous SNP calls were subsequently
513 treated as missing data. Samples with a high proportion of heterozygous (> 10 %) or
514 missing calls (> 20%) were excluded from further analyses to reduce the possibilities
515 of genotyping error due to tissue contamination influencing downstream analyses.

516 Both sample- and tissue contaminations will affect the accuracy of the genetic
517 map, both with regards to marker order and map size. The smaller family sizes
518 resulting from dividing our original 2,000 samples into three independent families
519 yield lower resolution of the component maps. However, fortuitously enough it also
520 allows us to incorporate more markers into the consensus map since different markers
521 were segregating in the different mother trees from which the three families were
522 derived. Furthermore, it also allowed us to evaluate marker ordering across three
523 independently derived maps. Although our consensus map is 60-70% larger than
524 previously estimated *Picea* maps (3,326 cM vs. 1,889-2,083 cM), it also contain 2-22
525 times more markers than earlier maps (Pavy et al. 2012; Lind et al. 2014; Pavy et al.
526 2017). When comparing marker order between our three independent component
527 maps (cluster 1-3), we found overall high order of correlations (0.94-0.99, Table S1),
528 which is similar to what is observed between maps derived from simulated data

529 without genotyping errors but with 20% missing data (Schiffthaler et al. 2017). Also,
530 earlier *Picea* maps were all diploid F_1 crosses and even the densest composite map
531 only contained 2,300-2,800 markers per framework map (Table 1 - Pavy et al. 2017),
532 compared to our haploid component maps that contain between 7,179 and 13,479
533 markers each (Table 2).

534 The comparisons between our haploid consensus map and earlier maps in *Picea*
535 show an overall high correlation of marker order, which is in line with previous
536 studies suggesting highly conserved synteny within *Picea* and in conifers in general
537 (de Miguel et al. 2015; Pavy et al. 2017). LG I from our haploid consensus map and
538 LG 7 from Lind et al. 2014 show a inverted order for approximately half of the
539 markers that were compared (Figure 4). However, if this inversion is due to ordering
540 errors in one of the maps or represents true biological differences between the parents
541 used for the respective maps is not known at the moment, and further investigations
542 are needed to resolve this issue.

543 A small percentage of the marker comparisons in both the intra and interspecific
544 maps do not co-align to homologous LGs. These errors likely arise form the repetitive
545 nature of the Norway spruce genome (and conifer genomes in general) where regions
546 with high sequence similarity often can be found interspersed through out the genome.
547 If the true homologous region between different maps is missing or has been
548 collapsed in the Norway spruce genome assembly due to high sequence similarity,
549 pairwise sequence comparisons may end up assigning homology to regions that are
550 located on different chromosomes.

551 4% of the scaffolds carrying multiple makers show a pattern where different
552 markers are mapping to different regions either within or between chromosomes in
553 the consensus map. This likely indicates errors in scaffolding during the assembly of

554 the v1.0 *P. abies* genome (Nystedt et al. 2013). If this estimate represents the overall
555 picture of the Norway spruce genome assembly, as many as 400,000 of the ~10
556 million total scaffolds, and 2,400 of the ~60,000 gene containing scaffolds, may
557 suffer from assembly errors. Approximately half of these, 2% of the multi-marker
558 scaffolds (100/4,859), have splits that occur within a single gene model. It is likely
559 that many of these problematic scaffolds stem from incorrect scaffolding of exons
560 from paralogous genes with a high sequence similarity. Since the Norway spruce
561 genome contains a high proportion of repetitive content, that also includes a large
562 number of pseudo genes, this is perhaps not surprising. Additional work is needed to
563 disentangle these issues and to resolve any assembly errors. False scaffold joins in a
564 genome assembly is not a unique feature for *P. abies*, rather it appears to be a
565 frequent problem in the assembly process. For instance, dense genetic maps in both
566 *Eucalyptus* and *Crassostrea* have identified and resolved false scaffold joins, thereby
567 improving the genome assemblies in these species (Bartholomé et al. 2015;
568 Hedgecock et al. 2015). Our goal for the Norway spruce genetic map is not only to
569 identify incorrect scaffolding decisions in the v1.0 genome assembly, but to also help
570 improve future iterations of the genome.

571 Our populations genetic analyses based on the scaffolds anchored to the
572 consensus map shows the utility of having a dense, accurate genetic map and suggest
573 that the map will facilitate further analyses of genome-wide patterns in Norway
574 spruce. Assigning even a small fraction of the genome to linkage groups allows us to
575 analyze patterns of genetic diversity in approximately a quarter of all predicted genes
576 from Norway spruce. This allows for analyses of broad-scale patterns of variation
577 across the spruce genome and as the genome assembly is further improved it should
578 allow us physically anchor a larger fraction of the genome to chromosomes and

579 thereby allow for even more fine-scaled analyses of how different evolutionary forces
580 have interacted in shaping patterns of genetic diversity across the Norway spruce
581 genome.

582 **Acknowledgements**

583 This study was supported by Knut and Alice Wallenberg's foundation through
584 funding to the Norway spruce genome project. AV was partially supported by a grant
585 from the Stiftelsen Gunnar och Birgitta Nordins fond through the Kungl. Skogs- och
586 Lantbruksakademien (KSLA). All computations were performed on resources
587 provided by SciLifeLab and SNIC at the Uppsala Multidisciplinary Center for
588 Advanced Computational Science (UPPMAX) under project b2010042.

589

590 **Author contribution**

591 PKI and MRGG conceived the study. AV collected cones and extracted DNA. CB,
592 AV, DS and JB set up bioinformatics pipeline for analyzing sequence capture data.
593 AV and CB performed PCA and identified samples belonging to the three clusters.
594 CB, DS and BS created the genetic maps. CB and PKI performed intra- and
595 interspecific map comparisons. CB, XW and PKI performed population genetic
596 analysis. CB performed all remaining analyses and wrote first draft of manuscript. All
597 authors commented on the manuscript at various stages during the writing.

598

599 **Data availability**

600 BatchMap input files for the three clusters, component maps and consensus map files
601 are available from zenodo.org at https://doi.org/10.5281/zenodo.1209842. All scripts
602 needed to recreate the analyses described in the paper are publically available at

603 https://github.com/parkingvarsson/HaploidSpruceMap. Raw sequence data for all
604 samples included in this study are available though the European Nucleotide Archive
605 under accession number PRJEB25757.

606

607

608 **References**

609 Baison, J., Vidalis, A., Zhou, L., Chen, Z-Q., Li, Z, Sillanpää, M.J., Bernahrdsson, C.,
610 Scofield, D.G., Forsberg, N., Olsson, L., Karlsson, B., Wu, H., Ingvarsson, P.K.,
611 Lundqvist, S-O., Niittylä, T., Garcia Gil, M.R. 2018. Association mapping
612 identified novel candidate loci affecting wood formation in Norway spruce.
613 bioRxiv https://doi.org/10.1101/292847

614 Bartholomé, Jérôme, Eric Mandrou, André Mabiala, Jerry Jenkins, Ibouniyamine
615 Nabihoudine, Christophe Klopp, Jeremy Schmutz, Christophe Plomion, and
616 Jean-Marc Gion. 2015. High-Resolution Genetic Maps of Eucalyptus Improve
617 Eucalyptus Grandis Genome Assembly. *New Phytologist* 206: 1283–96.
618 doi:10.1111/nph.13150.

619 Cartwright, Dustin A, Michela Troggio, Riccardo Velasco, and Alexander Gutin.
620 2007. Genetic Mapping in the Presence of Genotyping Errors. *Genetics* 176:
621 2521–27. doi:10.1534/genetics.106.063982.

622 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E.
623 Handsaker, et al. 2011. The Variant Call Format and VCFtools. *Bioinformatics*
624 27: 2156–58. doi:10.1093/bioinformatics/btr330.

625 De La Torre, Amanda R., Inanc Birol, Jean Bousquet, Pär K. Ingvarsson, Stefan
626 Jansson, Steven J.M. Jones, Christopher I. Keeling, et al. 2014. Insights into

Conifer Giga-Genomes. *Plant Physiology* 166: 1724 – 1732. http://www.plantphysiol.org/content/166/4/1724.short.

de Miguel, Marina, Jérôme Bartholomé, François Ehrenmann, Florent Murat, Yoshinari Moriguchi, Kentaro Uchiyama, Saneyoshi Ueno, et al. 2015. Evidence of Intense Chromosomal Shuffling during Conifer Evolution. *Genome Biology and Evolution* 7: 2799–2809. doi:10.1093/gbe/evv185.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data. *Nat Genet* 43: 491–98. doi:10.1038/ng.806.

Drost, Derek R., Evandro Novaes, Carolina Boaventura-Novaes, Catherine I. Benedict, Ryan S. Brown, Tongming Yin, Gerald A. Tuskan, and Matias Kirst. 2009. A Microarray-Based Genotyping and Genetic Mapping Approach for Highly Heterozygous Outcrossing Species Enables Localization of a Large Fraction of the Unassembled *Populus Trichocarpa* Genome Sequence. *The Plant Journal* 58: 1054–67. doi:10.1111/j.1365-313X.2009.03828.x.

Endelman, Jeffrey B., and Christophe Plomion. 2014. LPmerge: An R Package for Merging Genetic Maps by Linear Programming. *Bioinformatics* 30: 1623–24. doi:10.1093/bioinformatics/btu091.

Farjon, A. 1990. Pinaceae. Drawings and Descriptions of the Genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and Picea. *Pinaceae. Drawings and Descriptions of the Genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and Picea.* Koeltz Scientific Books.

651 https://www.cabdirect.org/cabdirect/abstract/19920656698.

652 Fierst, Janna L. 2015. Using Linkage Maps to Correct and Scaffold de Novo Genome
653 Assemblies: Methods, Challenges, and Computational Tools. *Frontiers in*
654 *Genetics* 6: 220. doi:10.3389/fgene.2015.00220.

655 Gaut, Brandon S., Stephen I. Wright, Carène Rizzon, Jan Dvorak, and Lorinda K.
656 Anderson. 2007. Recombination: An Underappreciated Factor in the Evolution
657 of Plant Genomes. Nature Reviews Genetics 8: 77–84.

658 Hedgecock, Dennis, Grace Shin, Andrew Y Gracey, David Van Den Berg, and Manoj
659 P Samanta. 2015. Second-Generation Linkage Maps for the Pacific Oyster
660 Crassostrea Gigas Reveal Errors in Assembly of Genome Scaffolds. G3: Genes,
661 Genomes, Genetics: 5: 2007–19. doi:10.1534/g3.115.019570.

662 Hu, Ying, Chunhua Yan, Chih-Hao Hsu, Qing-Rong Chen, Kelvin Niu, George
663 Komatsoulis, and Daoud Meerzaman. 2014. OmicCircos: A Simple-to-Use R
664 Package for the Circular Visualization of Multidimensional Omics Data. *Cancer*
665 *Informatics* 13: 13. doi:10.4137/CIN.S13495.

666 Kelly, J. K. 1997. "A Test of Neutrality Based on Interlocus Associations." *Genetics*
667 146: 1197–1206.

668 Knaus, Brian J., and Niklaus J. Grünwald. 2017. vcfR : A Package to Manipulate and
669 Visualize Variant Call Format Data in R. *Molecular Ecology Resources* 17: 44–
670 53. doi:10.1111/1755-0998.12549.

671 Li, H., and R. Durbin. 2009. Fast and Accurate Short Read Alignment with Burrows-
672 Wheeler Transform. *Bioinformatics* 25: 1754–60.
673 doi:10.1093/bioinformatics/btp324.

674 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G.

675 Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009.
676 The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25: 2078–
677 79. doi:10.1093/bioinformatics/btp352.

678 Lind, Mårten, Thomas Källman, Jun Chen, Xiao-Fei Ma, Jean Bousquet, Michele
679 Morgante, Giusi Zaina, et al. 2014. A Picea Abies Linkage Map Based on SNP
680 Markers Identifies QTLs for Four Aspects of Resistance to Heterobasidion
681 Parviporum Infection. *PloS One* 9: e101049. doi:10.1371/journal.pone.0101049.

682 Margarido, G R A, A P Souza, and A A F Garcia. 2007. OneMap: Software for
683 Genetic Mapping in Outcrossing Species. *Hereditas* 144: 78–79.
684 doi:10.1111/j.2007.0018-0661.02000.x.

685 McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian
686 Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. The Genome
687 Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation
688 DNA Sequencing Data. *Genome Research* 20: 1297–1303.
689 doi:10.1101/gr.107524.110.

690 Mollinari, M, G R A Margarido, R Vencovsky, and A A F Garcia. 2009. Evaluation
691 of Algorithms Used to Order Markers on Genetic Maps. *Heredity* 103: 494–502.
692 doi:10.1038/hdy.2009.96.

693 Nystedt, Björn, Nathaniel R. Street, Anna Wetterbom, Andrea Zuccolo, Yao-Cheng
694 Lin, Douglas G. Scofield, Francesco Vezzi, et al. 2013. The Norway Spruce
695 Genome Sequence and Conifer Genome Evolution. *Nature* 497: 579–84.
696 doi:10.1038/nature12211.

697 Pavy, Nathalie, Astrid Deschênes, Sylvie Blais, Patricia Lavigne, Jean Beaulieu,
698 Nathalie Isabel, John Mackay, and Jean Bousquet. 2013. The Landscape of

Nucleotide Polymorphism among 13,500 Genes of the Conifer Picea Glauca, Relationships with Functions, and Comparison with Medicago Truncatula. *Genome Biology and Evolution* 5: 1910–25. doi:10.1093/gbe/evt143.

Pavy, Nathalie, Manuel Lamothe, Betty Pelgas, France Gagnon, Inanç Birol, Joerg Bohlmann, John Mackay, Nathalie Isabel, and Jean Bousquet. 2017. A High-Resolution Reference Genetic Map Positioning 8.8 K Genes for the Conifer White Spruce: Structural Genomics Implications and Correspondence with Physical Distance. *The Plant Journal* 90: 189–203. doi:10.1111/tpj.13478.

Pavy, Nathalie, Betty Pelgas, Jérôme Laroche, Philippe Rigault, Nathalie Isabel, and Jean Bousquet. 2012. A Spruce Gene Map Infers Ancient Plant Genome Reshuffling and Subsequent Slow Evolution in the Gymnosperm Lineage Leading to Extant Conifers. *BMC Biology* 10: 84. doi:10.1186/1741-7007-10-84.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. http://www.r-project.org.

Sax, Karl, and Hally Jolivettte Sax. 1933. Chromosome Number and Morphology in the Conifers. *Journal of the Arnold Arboretum* 14: 356-375.

Sturtevant, A. H. 1913a. The Linear Arrangement of Six Sex-Linked Factors in Drosophila, as Shown by Their Mode of Association. *Journal of Experimental Zoology* 14: 43–59. doi:10.1002/jez.1400140104.

———. 1913b. A Third Group of Linked Genes in Drosophila Ampelophila. *Science* 37: 990–92. doi:10.1126/science.37.965.990.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit

723 Best Practices Pipeline. In *Current Protocols in Bioinformatics*, 11.10.1-
724 11.10.33. Hoboken, NJ, USA: John Wiley & Sons, Inc.
725 doi:10.1002/0471250953.bi1110s43.

726 Van Os, Hans, Piet Stam, Richard G F Visser, and Herman J Van Eck. 2005.
727 RECORD: A Novel Method for Ordering Loci on a Genetic Linkage Map.
728 *Theoretical and Applied Genetics.* 112: 30–40. doi:10.1007/s00122-005-0097-x.

729 Vidalis, A. Scofield, D.G., Neves, L-G., Bernhardsson, C., García-Gil, M.R.,
730 Ingvarsson, P.K. 2018. Design and evaluation of a large sequence-capture
731 probe set and associated SNPs for diploid and haploid samples of Norway
732 spruce (*Picea abies*) *BioRxiv* doi: https://doi.org/10.1101/291716

733 Wu, Rongling, Chang-Xing Ma, Ian Painter, and Zhao-Bang Zeng. 2002.
734 Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases
735 in Outcrossing Species. *Theoretical Population Biology* 61: 349–63.
736 doi:10.1006/tpbi.2002.1577.

737