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 Schiffthaler[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] Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, 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 Genetics, 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] Umeå Plant Science Centre, Department of Plant Physiology, 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 of substanital economic and ecological importance. In common with most conifers, the *P. abies* genome is very large (~20 Gbp) and contains a high fraction of repetitive DNA. The current *P. abies* genome assembly (v1.0) covers approximately 60% of the total genome size but is highly fragmented, consisting of >10 million scaffolds. The genome annotation contains 66,632 gene models that are at least partially validated (www.congenie.org), however, 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 anchored and ordered scaffolds into linkage groups, which complements the fine-scale information available in assembly contigs. Our ultra-dense haploid consensus genetic map consists of 21,056 markers derived from 14,336 scaffolds that contain 17,079 gene models (25.6% of the validated gene models) that we have anchored to the 12 linkage groups. We used data from three independent component maps, as well as comparisons with previously published *Picea* maps to evaluate the accuracy and marker ordering of the linkage groups. We demonstrate that approximately 3.8% of the anchored scaffolds and 1.6% of the gene models covered by the consensus map have likely assembly errors as they contain genetic markers that map to different regions within or between linkage groups. We further 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 linkage groups. The results show that our map is sufficiently dense to enable detailed evolutionary analyses across the *P. abies* genome.

Introduction

For over a century genetic linkage maps have been used to order genetic markers and link phenotypic traits to genomic regions and chromosomes by calculating recombination events in crosses (Sturtevant 1913a; Sturtevant 1913b). With the advent 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 generation of high-density genetic maps consisting of thousands of markers that, in combination with a sufficiently large mapping population, can achieve unprecedented mapping resolution even in non-model systems and in species with large genomes. Genetic maps represent a complementary approach to the local, fine-scale genomic information that is available in scaffolds from a genome assembly, with a genetic map providing information on genome organization over larger scales (up to whole-chromosome level) (Fierst 2015). By grouping markers into linkage groups and subsequently ordering them within each linkage group, it is possible to anchor underlying scaffolds containing those markers to putative chromosomes with high precision (Fierst 2015). If several genetic markers, derived from a single genomic scaffold, are placed on the map, information on their relative placement in the genetic map can be used to orient the scaffold and to evaluate scaffolding decisions made in the genome assembly and hence to locate and resolve possible assembly errors (Drost et al. 2009; Bartholomé et al. 2015). For instance, when two

markers originating from a single scaffold map to different linkage groups or to different regions within a linkage group, the contigs comprising the scaffold are candidates for having been wrongly joined during the assembly process. On the other hand, if markers from the same scaffold map close to each other this increases the likelihood that the scaffolding decisions were correct.

Norway Spruce (*Picea abies*) is one of the most important conifer species in Europe, both from an ecological and economic perspective. The natural distribution range of *P. abies* extends from the west coast of Norway to the Ural mountains and across the Alps, Carpathians and the Balkans in central Europe. *P. abies* composes, together with *Pinus sylvestris*, the majority of the continuous boreal forests of the Northern hemisphere where it is considered a keystone species (Farjon 1990). *P. abies* has a genome size of ~20 Gbp that is characterized by a very high fraction of repetitive sequences. Like most conifers, *P. abies* has a karyotype consisting of $2n$=24 and with chromosomes that are all uniformly sized (Sax and Sax 1933). Due to the large and complex genome of conifers, this important group of plants was, until recently, lacking species with available reference genomes. In 2013 the first draft assembly of the *P. abies* 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 >500 bp and contains 70,736 annotated gene models of which 66,632 are at least partially validated by supporting evidence (ESTs or UniProt proteins) (Nystedt et al. 2013; De La Torre et al. 2014). Although the current genome assembly only covers about two thirds of the total genome size (12 Gbp out of the 20 Gbp *P. abies* genome), it is expected to contain the majority of expressed genes.

In this paper, we used sequence capture to identify segregating SNP markers in megagametophytes from three open-pollinated mother trees. These markers were used to create an ultra-dense haploid genetic map consisting of 21,056 probe-markers derived from 14,336 gene-bearing scaffolds in the *P. abies* genome assembly. Our aim with creating the genetic map was to 1) anchor, and where possible, order scaffolds to assign as many gene models as possible to linkage groups, and 2) to evaluate the accuracy of the *P. abies* genome assembly v1.0 on the basis of anchored scaffolds. To evaluate the accuracy of the map itself, we compared scaffold order to previously published genetic maps for *P. abies* and the closely related *Picea glauca*. Finally, we evaluated utility of the genetic map for population genomic studies by performing genome-wide analyses of genetic diversity for the genomic regions anchored in the map using a sample of c. 500 unrelated *P. abies* trees.

Material and Methods

DNA extraction and sequence capture

In the autumn of 2013, seeds were collected for linkage map construction from five of 30 putative ramets of Z4006, the genotype used to generate the reference genome for *Picea abies* (Nystedt et al. 2013). Megagametophytes were dissected from 2,000 seeds by removing the diploid seed coat surrounding the haploid megagametophyte tissue. DNA extraction from megagametophytes was performed using a Qiagen Plant Mini Kit. Each extracted sample was measured for DNA quality using a Qubit® ds DNA Broad Range (BR) Assay Kit, and all samples with a total amount of DNA >354 ng were kept. The remaining 1,997 samples were sent to RAPiD Genomics© (Gainesville, Florida, USA) in September 2014 for sequence capture using 31,277 capture probes that had been specifically designed to target 19,268 partially-validated gene models from the *P. abies* genome assembly. Where possible, probes were

122 designed to flank regions of known contig joins in the v1.0 genome assembly (for
123 further detail of the probe design, see Vidalis et al. 2018).

124 The capture data was sequenced by RAPiD Genomics© on an Illumina HiSeq
125 2000 using 1x75 bp sequencing and was delivered in October 2015. The raw reads
126 were mapped against the complete *P. abies* reference genome v.1.0 using BWA-
127 MEM v.0.7.12 and default settings (Li and Durbin 2009). Following read mapping,
128 the genome was subset to only contain the probe-bearing scaffolds (a total of 18,461
129 scaffolds) using Samtools v.1.2 (Li and Durbin 2009; Li et al. 2009). Duplicates were
130 marked and local realignment around insertion/deletions (indels) was performed using
131 Picard (http://broadinstitute.github.io/picard/) and GATK
132 (https://software.broadinstitute.org/gatk/) (McKenna et al. 2010; DePristo et al. 2011).
133 Genotyping was performed using GATK Haplotypecaller (version 3.4-46, (DePristo
134 et al. 2011; Van der Auwera et al. 2013) with a diploid ploidy setting and gVCF
135 output format. We used a diploid ploidy setting to increase the likelihood of detecting
136 possible sample contamination from diploid tissue for the haploid megagametophyte
137 samples. CombineGVCFs was then run on batches of ~200 gVCFs to hierarchically
138 merge them into a single gVCF and a final SNP call was performed using
139 GenotypeGVCFs jointly on the 10 combined gVCF files, using default read mapping
140 filters, a standard minimum confidence threshold for emitting (stand-emit-conf) of 10,
141 and a standard minimum confidence threshold for calling (stand_call_conf) of 20. See
142 Vidalis et al. (2018) and the script "per_sample_gvcf.sh" (available at
143 https://github.com/parkingvarsson/HaploidSpruceMap) for a full description of the
144 pipeline used for calling variants.

145

146 *SNP filtration and megagametophyte relationships*

147 After SNP filtering, we performed a principle component analysis (PCA) to evaluate
148 the relationship among samples (see Supplementary file for details on the PCA
149 analysis and subsequent filtering steps). Based on the PCA and a hierarchical
150 clustering approach, we divided samples into three clusters representing putative
151 maternal families (Supplementary, Figure S1-3) that were then analyzed
152 independently. In the end we obtained 9,073 probe-markers from 7,101 scaffolds for
153 Cluster 1 (314 samples), 11,648 probe-markers from 8,738 scaffolds for Cluster 2
154 (270 samples) and 19,006 probe-markers from 13,301 scaffolds for Cluster 3 (842
155 samples) with a total of 21,056 probe-markers from 14,336 scaffolds across all three
156 clusters (Table 1). In total, these scaffolds cover 0.34 Gbp of the *P. abies* genome and
157 contain 17,079 partially validated gene models.
158
159 **Table 1:** Overview of the three component maps and the total number of probe-markers
160 available in the consensus map. Cluster: Name of each putative maternal family that
161 was identified in the principal component analysis. Samples: Number of
162 megagametophytes in each cluster. Markers: Number of probe-markers in each
163 component map with number of unique segregating bins within brackets (one marker
164 for each bin was used to anchor the bin markers to the genetic map). Scaffolds: Number
165 of scaffolds represented in each component map.

Cluster	Samples	Markers	Scaffolds
Cluster 1	314	9,073 (3,924)	7,101
Cluster 2	270	11,647 (5,311)	8,738
Cluster 3	842	19,006 (11,479)	13,301
Total	1,426	21,056	14,336

166

167 *Component and consensus maps*

168 We created genetic linkage maps using the R-package BatchMap (Schiffthaler et al.
169 2017), a parallel implementation of the R-package Onemap (Margarido, Souza, and
170 Garcia 2007). All probe-markers were recoded using the D1.11 cross-type (Wu et al.
171 2002), tested for segregation distortion ($p < 0.05$ after Bonferroni correction)
172 (Supplementary, Figure S4) and grouped into marker bins. The probe-marker with
173 lowest amount of missing data in each bin was then used to represent the bin when
174 constructing the genetic map. Bin markers were grouped into LGs using LOD = 8 and
175 a maximum recombination fraction = 0.35. LGs were then ordered using the
176 RECORD algorithm (Van Os et al. 2005) with 16 times counting, parallelized over 16
177 cores, reordered in a 10 marker sliding window with 1 marker incremental steps using
178 the command 'ripple' and finally mapped using the Kosambi mapping function and
179 the 'map batches' approach (Schiffthaler et al. 2017) over four parallel cores. Finally,
180 heat maps with pairwise recombination fraction (lower triangular) and phase LOD
181 score (upper triangular) for the ordered markers were created to evaluate the ordering
182 accuracy of independent linkage groups (Supplementary, Figure S5 and S6A-L). We
183 observed 183 probe-marker bins showing signs of segregation distortion. These bins
184 were, however, randomly distributed over the linkage groups and did not appear to
185 affect marker ordering and map distance and were therefore retained in subsequent
186 analyses.

187 To evaluate correspondence between LGs in maps derived from the three PCA
188 clusters, the number of unique scaffolds shared between cluster LGs were counted
189 (Supplementary, Figure S5). We then created a consensus map for each linkage group
190 from the three independent component maps using the R-package LPmerge

(Endelman and Plomion 2014) with component maps ranked according to marker numbers (Cluster 3, Cluster 2, Cluster 1), a maximum interval setting ranging from one to 10 and map weights proportional to the size of the mapping population (Cluster 3= 0.5, Cluster 2 and Cluster 1 = 0.25). From all possible consensus maps generated by LP merge, for each linkage group we selected the map with the lowest mean root mean square error (RMSE) to serve as the consensus map (Endelman and Plomion 2014). Order correlations between individual component maps and the consensus maps (Table 2 and Supplementary, Figure S7A-L) as well as between the three component maps (Supplementary, Figure S8A-L) were estimated using Kendall's τ. For visual representation of the consensus map we created a Circos plot using the R-package omicCircos (Hu et al. 2014), available from Bioconductor (https://bioconductor.org/biocLite.R).

To evaluate the inflation of map distances due to possible genotyping errors, we performed 100 rounds of random subsampling of 100 probe-marker bins per LG and component map. The following marker ordering and genetic distance calculation were performed with 10 rounds of RECORD and the Kosambi mapping function.

Accuracy of the reference Picea abies genome assembly

To evaluate the accuracy of scaffolds from the v1.0 *P. abies* reference genome containing at least two probe-markers (here after called multi-marker scaffolds) we determine whether probe-markers from the same genomic scaffold mapped to the same region of an LG, on different regions within a single LG or on different LGs. In the consensus map, we considered markers to be positioned in the same region on an LG if all probe-markers from a scaffold mapped within a 5 cM interval of each other. If any marker from the scaffold was positioned further apart, the scaffold was tagged

as containing a putative assembly error. The same considerations were made for scaffolds with probe-markers positioned on different LGs.

Comparative analyses of Picea linkage maps

To evaluate the consistency of our genetic map with earlier maps from *P. abies* we compared our haploid consensus map to the *P. abies* linkage map from Lind et al. (2014). The Lind et al. map was created using genetic markers generated using an Illumina 3072 SNP Golden Gate Assay. We performed using `tblastn` sequence homology searches against the *P. abies* v1.0 genome assembly for the SNP array sequences of the makers mapped in the Lind et al. map and extracted reciprocal best hits with >95% identity, which were then assigned to the corresponding scaffold in the *P. abies* genome. We performed similar analyses to compare the synteny between our consensus map and the *P. glauca* composite map from Pavy et al. (2017). Again, we used `tblastn` sequence homology search comparisons of array sequences from the *P. glauca* SNP array (Pavy et al. 2013) with scaffolds from the *P. abies* v1.0 genome assembly to assign corresponding map positions between *P. abies* and *P. glauca*. In order to evaluate correspondence between LGs from the different genetic maps, we assessed the number of shared scaffolds between our consensus map, the Lind et al. and Pavy et al. maps. Consistency of scaffold ordering was then evaluated using visual comparisons (Figure 4 and 5) and by calculating correlations of marker orders using Kendall's τ.

238 *Population genetic analysis of the consensus genetic map*

239 In order to independently evaluate the utility of the consensus map for downstream
240 research, we used a subset of the data from Baison et al. (2018) to estimate patterns of
241 nucleotide diversity across the Norway spruce genome. The data from Baison et al.
242 originally contained 517 individuals sequenced with 40,018 probes designed for
243 diploid spruce samples (Vidalis et al 2018). We extracted data for all probes that we
244 had anchored in our genetic map from the VCF file containing the data from Baison et
245 al.. We further hard-filtered the resulting VCF file by only considering bi-allelic SNPs
246 within the extended probe regions (120 bp probes ±100 bp) with a QD >5, MQ >50
247 and a overall DP between 3000 and 16000. Samples containing >25% missing data
248 were removed from further analysis. We used the data to calculate nucleotide
249 diversity (π), the number of segregating sites and Tajima's D (Tajima 1989). We used
250 the R package vcfR (Knaus and Grünwald 2017) to read the VCF-file into R and then
251 used in-house developed scripts to perform all calculations (available at
252 https://github.com/parkingvarsson/HaploidSpruceMap). We assigned probes to LGs
253 and map positions by assigning them the coordinates of the physically closest (in bp)
254 probe. We also calculated pairwise linkage disequilibrium (LD) between markers
255 within probes using vcftools (Danecek et al. 2011) and imported the results into R
256 where they were used to calculate Z_{nS} scores (Kelly 1997) per probe using an in-house
257 developed script (available at https://github.com/parkingvarsson/HaploidSpruceMap).
258 Finally, we ran sliding window analyses along the linkage groups for the different
259 summary statistics using 10 cM windows that were moved in 1 cM incremental steps.

260

Results

We generated a *P. abies* consensus linkage map from three haploid component maps containing a total of 21,056 unique probe-markers from 14,336 scaffolds in the *P. abies* genome assembly v1.0. The consensus map anchored 0.34 Gbp of the *P. abies* 1.0 assembly, corresponding to 1.7% of the complete *P. abies* genome or 2.8% of the genome assembly. However, these scaffolds anchor 25.6% of all validated gene models with these anchored scaffolds containing 31.7%, 20.6% and 25.8% of the High-, Medium- and Low confidence gene models from Nystedt et al (2013), respectively. The consensus map had a total length of 3,556 centiMorgan (cM), distributed over 12 linkage groups (LGs), corresponding to the haploid chromosome number (Sax and Sax 1933), and with an average distance of 0.17 cM between probe-markers (Table 2, Figure 1A).

Correlations of probe-marker order between the three component maps and the consensus map ranged from 0.96 to 0.998, while the correlations between marker orders between individual component maps ranged from 0.943 to 0.993 (Supplementary, Figure S7 and S8). 183 probe-marker bins showed evidence of segregation distortion in Cluster 3, but these were randomly distributed over all linkage groups and we did not observe regions showing clusters of markers with segregation distortion or with conflicting marker orders between clusters (Supplementary, Figure S8). LG XI, which displayed the largest discrepancy in marker order between component maps, has a region at the distal end of the LG, covering 252 probe-markers, where the resolution was too low to identify the correct marker order and where the entire region was positioned at 36.115 cM (Supplementary, Figure S7K and S8K), explaining the lower correlations in marker order between individual maps for this LG.

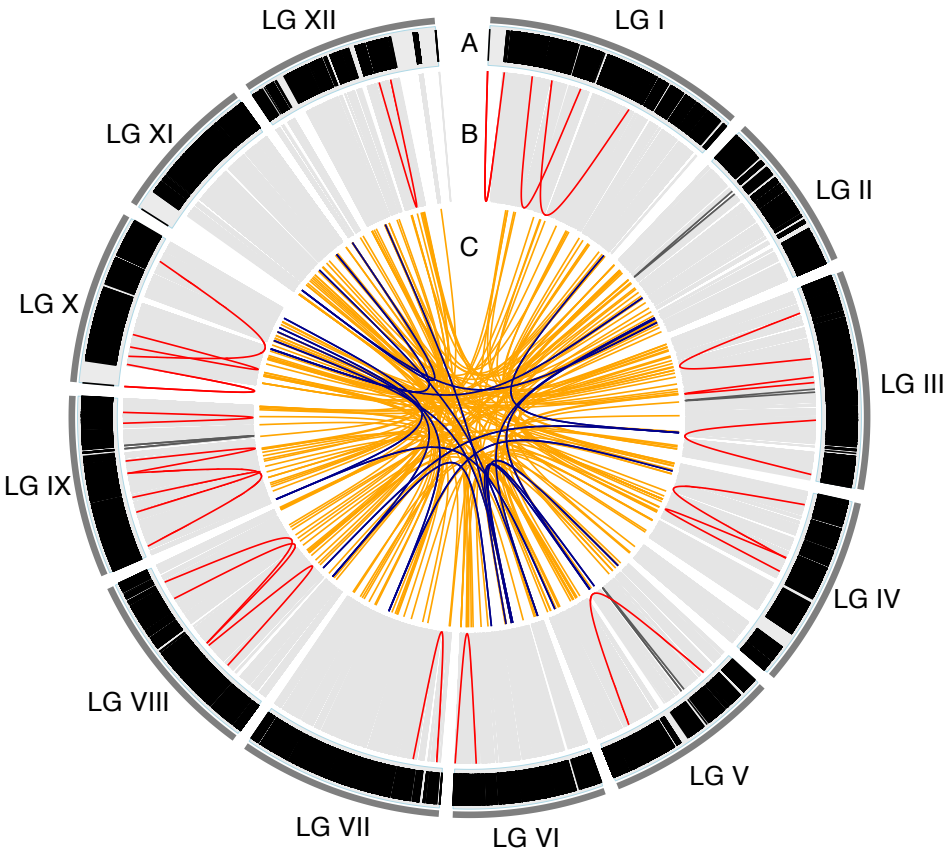
We used a random subsampling approach to evaluate potential inflation of map distances due to possible genotyping errors. From these analyses, total map size for Cluster 1 ranged between 2,166.8 and 2,450.0 cM with an average size of 2,294.2 cM and a standard deviation (SD) of 3.6- 5.8 cM per LG. Cluster 2 ranged between 2,304.2 and 2,663.6 cM with an average of 2,478.3 cM and a SD of 4.4 – 9.1 cM per LG, while Cluster 3 ranged between 1,855.4 and 2,093.2 cM with an average of 1,971.0 cM and a SD of 2.7 – 7.3 cM per LG. The estimated inflation was therefore predicted to be 0.15 – 0.31 cM per probe-marker bin across the three component maps (Table 3). This inflation per probe-marker bin roughly corresponded to the map resolution of the clusters (Cluster 1- 0.32 cM: Cluster 2 - 0.37 cM: Cluster 3 – 0.12 cM) and yielded an error estimate of ~1 genotype error per marker-bin or 11-17 genotype errors per sample.

Table 2: Marker density and size of each component genetic map created from the three clusters as well as for the consensus map. LG: Linkage group. Cluster 1-3: Component maps for cluster 1-3 with number of probe-markers (marker-bins) assigned, map size (in cM) and maximum gap in map (in cM) for each of the LGs. Consensus: Number of markers and map size of the LGs in the consensus map.

LG	Cluster 1			Cluster 2			Cluster 3			Consensus	
	Markers	Length (cM)	Max gap (cM)	Markers	Length (cM)	Max gap (cM)	Markers	Length (cM)	Max gap (cM)	Markers	Length (cM)
I	975 (421)	385.5	8.0	1,159 (553)	439.9	21.1	1,967 (1,185)	414.1	8.8	2,172	414.1
II	701 (305)	249.2	9.6	863 (366)	289.0	9.4	1,456 (864)	289.8	10.9	1,608	250.3
III	859 (394)	324.0	4.6	1,069 (479)	381.1	7.1	1,738 (1,075)	346.4	5.2	1,940	342.5
IV	771 (323)	298.7	14.5	970 (452)	350.9	8.6	1,531 (916)	303.0	27.0	1,704	303.0
V	761 (311)	273.2	8.9	1,116 (499)	395.6	9.5	1,649 (1,032)	342.6	15.1	1,865	275.0
VI	648 (292)	241.0	8.4	915 (399)	270.7	4.6	1,456 (894)	269.5	8.4	1,622	240.2
VII	682 (331)	314.0	8.4	923 (443)	380.8	13.4	1,625 (1,013)	321.9	7.9	1,769	321.0
VIII	775 (339)	307.0	5.6	943 (454)	367.26	9.8	1,465 (904)	315.6	6.6	1,609	305.9
IX	792 (332)	283.3	5.4	786 (364)	295.6	5.9	1,589 (911)	285.1	7.4	1,738	285.0
X	648 (289)	231.6	7.0	960 (454)	342.7	6.9	1,564 (917)	272.7	7.1	1,709	273.1
XI	677 (253)	200.6	3.7	1,025 (411)	269.2	4.0	1,440 (818)	233.6	3.0	1,608	233.4
XII	784 (334)	281.6	9.3	919 (437)	360.7	11.1	1,526 (950)	312.3	14.3	1,712	312.3
Total	9,073 (3,924)	3,389.4	14.5	11,648 (5,311)	4,143.4	21.1	19,006 (11,479)	3,706.7	27.0	21,056	3,555.8

Table 3: Estimated genetic length of each Linkage Group (LG) in the three component maps. LG: linkage group in the consensus map; Observed genetic length (cM): The genetic length of the LG calculated from all probe-marker bins (same as in table 2); Mean estimated genetic length (cM): the average length of the LG when using 100 random probe-marker bins in 100 map calculations; SD (cM): Standard deviation of the estimated length; Inflation/Marker bin: The difference between observed genetic length and the estimated length divided by the number of probe-marker bins in the linkage group.

LG	Cluster 1				Cluster 2				Cluster 3			
	Observed genetic length (cM)	Mean estimated genetic length (cM)	SD (cM)	Inflation / Marker bin (cM)	Observed genetic length (cM)	Mean estimated genetic length (cM)	SD (cM)	Inflation / Marker bin (cM)	Observed genetic length (cM)	Mean estimated genetic length (cM)	SD (cM)	Inflation / Marker bin (cM)
I	385.5	245.5	5.2	0.33	439.9	252.3	6.1	0.34	414.2	204.8	7.3	0.18
II	249.2	168.8	4.7	0.26	289.0	192.9	4.4	0.26	289.8	166.4	2.8	0.14
III	324.0	195.8	5.8	0.33	381.1	218.6	5.8	0.34	346.4	168.5	3.9	0.17
IV	298.7	204.7	5.0	0.29	350.9	215.6	5.7	0.30	303.0	167.0	3.5	0.15
V	273.2	195.7	4.6	0.25	395.6	218.4	9.0	0.36	342.6	180.0	5.1	0.16
VI	241.0	161.8	4.7	0.27	270.7	170.0	4.7	0.25	269.5	142.2	2.9	0.14
VII	314.0	223.6	5.3	0.27	380.8	248.7	6.3	0.30	321.9	175.9	3.7	0.14
VIII	307.0	203.6	4.9	0.31	367.26	226.7	5.7	0.31	315.6	179.2	4.3	0.15
IX	283.3	194.0	4.9	0.27	295.6	185.3	6.8	0.30	285.1	157.2	3.0	0.14
X	231.6	164.4	3.6	0.23	342.7	193.6	4.5	0.33	272.7	141.5	2.7	0.14
XI	200.6	141.8	4.7	0.23	269.2	147.0	4.8	0.30	233.6	119.6	3.0	0.14
XII	281.6	194.4	4.9	0.26	360.7	209.6	9.1	0.35	312.3	168.7	3.1	0.15
Total	3,389.4	2,294.2	-	0.28	4,143.4	2,478.3	-	0.31	3,706.7	1,971.0	-	0.15



309

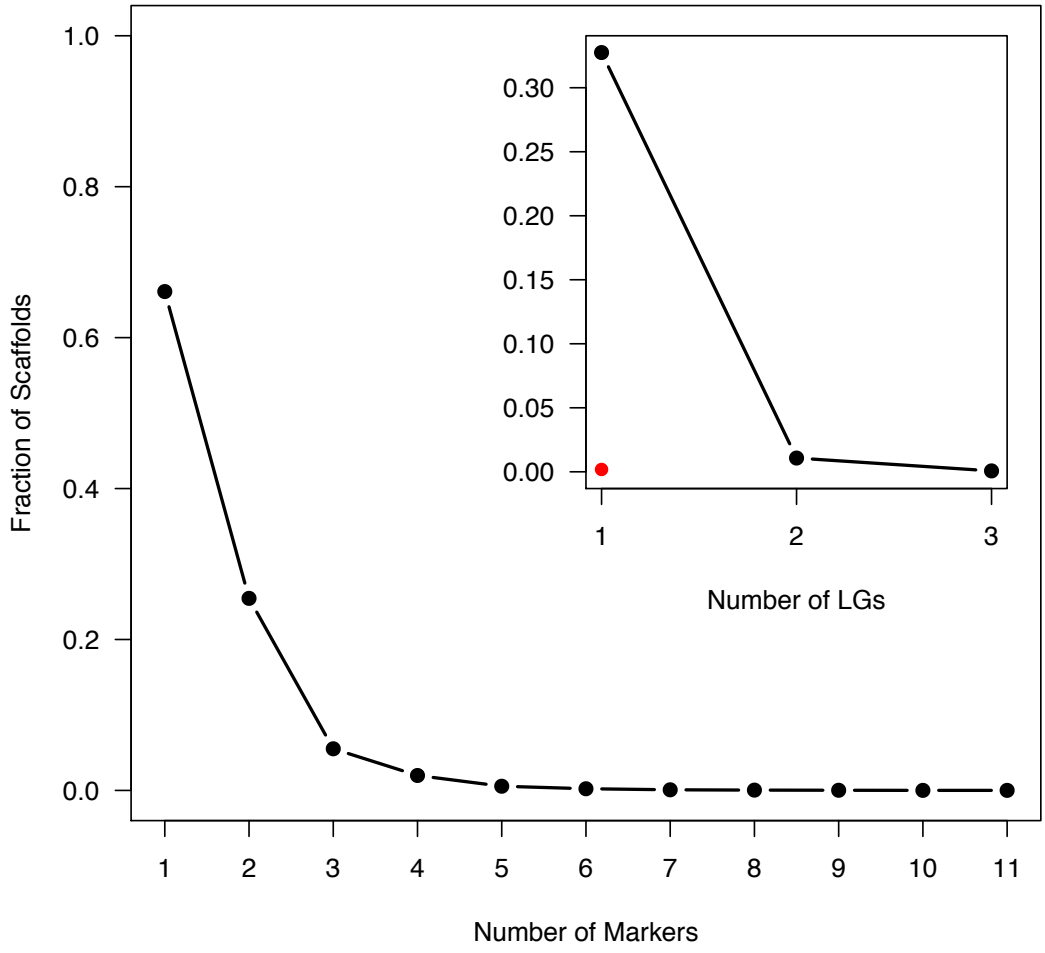
310 **Figure 1:** Circos plot of the consensus map. A) Marker distribution over the 12
311 linkage groups (LG I-LG XII). Each black vertical line represents a marker (21,056 in
312 total) in the map and is displayed according to the marker positions in cM. Track B-C
313 visualizes multi marker scaffolds, where each line is a pairwise position comparison of
314 probe-markers from the same scaffold. B) Position comparisons of probe-markers from
315 the same scaffold that are located on the same LG. Light grey lines indicate probe-
316 markers that are located < 5cM from each other, dark grey lines indicate probe-markers
317 located 5-10 cM apart and red lines indicate probe-markers >10 cM apart. C) Position
318 comparisons of probe-markers from the same scaffold that are mapping to different LGs.
319 Orange lines indicated probe-markers from the same scaffold split over 2 LGs, while
320 dark blue lines indicated probe-markers split over 3 LGs.

321

Evaluation of the Picea abies genome assembly v1.0

4,859 scaffolds (33.9%) contained more than one unique probe-marker combined over all three component maps. 185 of these multi-marker scaffolds contained markers that were located in more than one LG (*inter-split scaffolds*) or over different parts of the same LG (*intra-split scaffolds*). 26 scaffolds (0.18% of mapped scaffolds and 0.54% of multi-marker scaffolds) contained markers that were positioned on the same LG but at distances exceeding 5 cM in the consensus map. When exploring the individual component maps, it was apparent that for two of these scaffolds (MA_281725 on LG X and MA_10431182 on LG I) the probe-markers in the consensus map all came from different component maps. The consensus map thus contain a gap that we can not verify using any of the individual component maps (Figure 1 and Supplementary, Figure S9). Three other scaffolds (MA_9458 on LG IX, MA_10431315 on LG II and MA_10432328 on LG III) all have multiple probe-markers present in at least one component map and were these component maps do not support the split we observe in the consensus map (Supplementary, Figure S9). It thus appears that these splits are artifacts arising from the construction of the consensus map.

There were 164 scaffolds (1.14% of mapped scaffolds and 3.38% of multi-marker scaffolds) containing markers that were mapped to two or three different LGs (Figure 2 and Supplementary, Figure S10). All LGs contained inter-split scaffolds, while 10 LGs (LGII and LGXI are the exceptions) contained intra-split scaffolds supported by the component maps (Figure 1B-C and Supplementary, Figure S9).



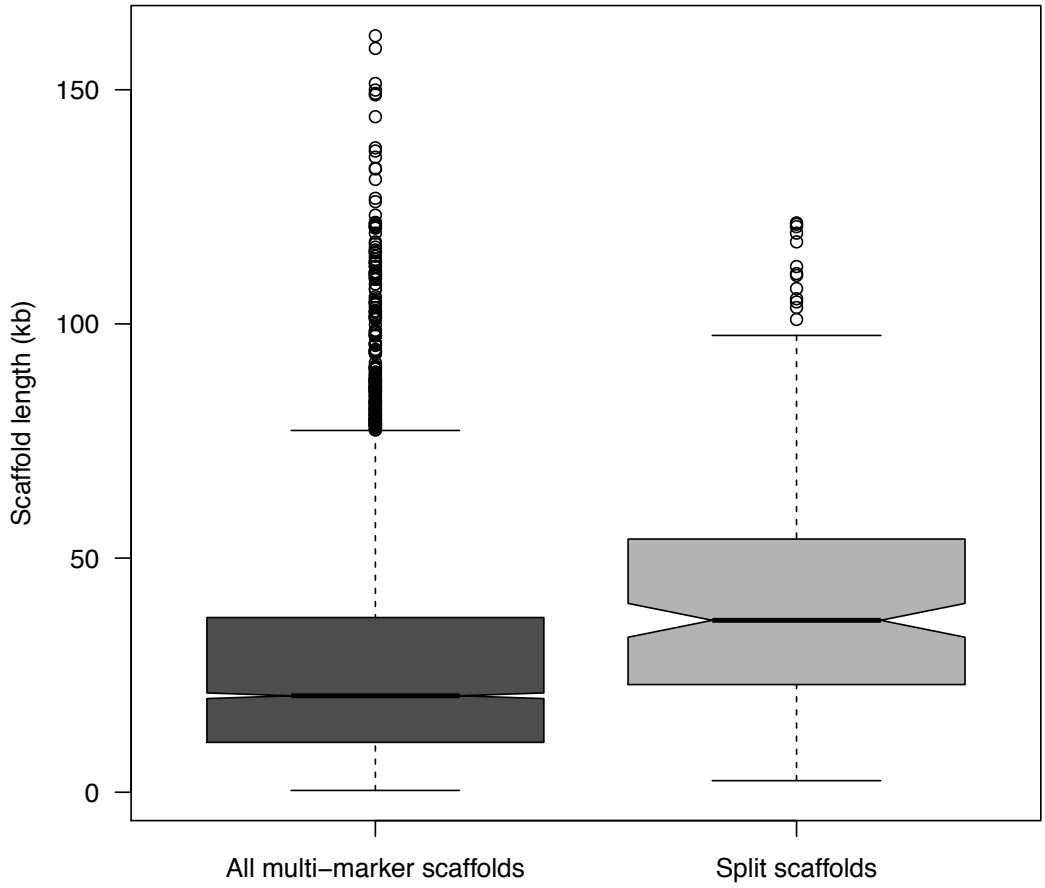
344

345 **Figure 2:** Fraction of scaffolds that are being represented by 1-11 unique
346 markers in the consensus map. Insert: Fraction of scaffolds that have multiple probe-
347 markers (2-11) that are distributed over 1-3 linkage groups (inter-split scaffolds). Red
348 dot indicate the fraction of scaffolds with multiple probe-markers which are positioned
349 > 5cM apart on the same linkage group (intra-split scaffolds).

350

351 The scaffolds covered by the consensus map ranged in length from 0.22 to 208.1
352 Kbp with a median of 17.1 Kbp, while multi-marker scaffolds ranged from 0.39 to
353 161.5 Kbp (median of 21 Kbp). The 185 scaffolds that are split within or across LGs
354 ranged in size from 2.5 to 121.6 Kbp, with a median length of 36.9 Kbp. Split
355 scaffolds were significantly longer than multi-marker scaffolds in general (t = -7.7, df

356 = 193.4, p-value = 7.0e-13; Figure 3), suggesting that longer scaffolds are more likely
357 to contain assembly errors compared to shorter scaffolds. Split scaffolds mostly
358 contained high- and medium confidence gene models (Table 4). A visual inspection
359 of the split scaffolds revealed that for 75 and 10 of the inter-split and intra-split
360 scaffolds, respectively, the predicted position of the split(s) occurred between
361 different gene models on the same scaffold. Of greater concern, for 88 of the inter-
362 split scaffolds and 11 of the intra-split scaffolds the predicted position of the split was
363 located within a single gene model (Supplementary, Figure S9 and S10). In addition,
364 21 inter-split scaffolds showed an even more complicated picture, where an interior
365 region of the gene model (most often containing an intron > 5kb) mapped to another
366 chromosome whereas the 5' and 3' regions of the gene model mapped to the same
367 chromosome location (Supplementary, Figure S10). However, 84% (184 out of a total
368 of 219 splits) appear to occur between contig joins (where a sequence of N's appear in
369 the assembly) of the scaffold. Of the 17,079 gene models that were anchored to the
370 consensus genetic map, 330 were positioned on inter- or intra-split scaffolds (5.4% of
371 gene models that were positioned on multi-marker scaffolds) and 100 showed a split
372 within gene models (1.6% of gene models from multi-marker scaffolds) (Table 4).



373

374 **Figure 3:** Box plot of scaffold lengths for all multi-marker scaffolds (dark gray
375 box) and for scaffolds showing a split within or across LGs (light gray box). The split
376 scaffolds are significantly longer than the multi-marker scaffolds in general (t = -7.70,
377 df = 193.39, p-value = 7.00e-13).

378

379 **Table 4**: Overview of annotated gene models anchored to the genetic map. Gene
380 models: Annotated protein coding gene models with High-, Medium- and Low
381 confidence level (Nystedt et al. 2013). Mapped scaffolds: Number of gene models
382 positioned on scaffolds that are anchored to the genetic map (Percentage of total
383 number of gene models for each confidence level). Multi-marker scaffolds: Number of
384 gene models positioned on scaffolds with multiple markers in the genetic map
385 (Percentage of gene models on mapped scaffolds). Inter-split scaffolds: Number of
386 gene models positioned on the 164 scaffolds that are split between LGs in the genetic
387 map (Percentage of gene models on mapped scaffolds / Percentage of gene models on

388 multi-marker scaffolds). Intra-split scaffolds: Number of gene models positioned on the
389 22 scaffolds that are split between different regions of the same LG (Percentage of
390 gene models on mapped scaffolds / Percentage of gene models on multi-marker
391 scaffolds). Split within gene models: Number of gene models that have an internal split
392 (Percentage of gene models on mapped scaffolds / Percentage of gene models on multi-
393 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%)	16 (0.23% / 0.68%)	29 (0.44% / 1.3%)
Low confidence	2,076 (25.8%)	762 (36.7%)	35 (1.7% / 4.6%)	5 (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%)

394

395 *Comparative analyses to other Picea linkage maps*

396 In order to assess the accuracy and repeatability of the *P. abies* genetic maps we
397 compared our consensus map to the *P. abies* map presented in Lind et al. (2014). 353
398 comparisons between 298 markers from Lind et al. and 288 scaffolds contained in our
399 consensus map were identified at a > 95 % identity threshold. Of these markers,
400 96.7% grouped to the same LG in the two maps while the remaining 3.3% (11 out of
401 353) were distributed across several LGs (Figure 4). Correlations of marker order
402 between the two *P. abies* maps ranged from 0.53 to 0.99 across the 12 LGs. The

comparison between the haploid consensus map for LG I and LG 7 from Lind et.al, which had the lowest correlation of marker order, showed inconsistencies of marker order where a contiguous subset of markers were arranged in the opposite order from the rest of the markers for that LG. The remaining LGs showed high synteny, with consistent marker ordering between the two genetic maps.

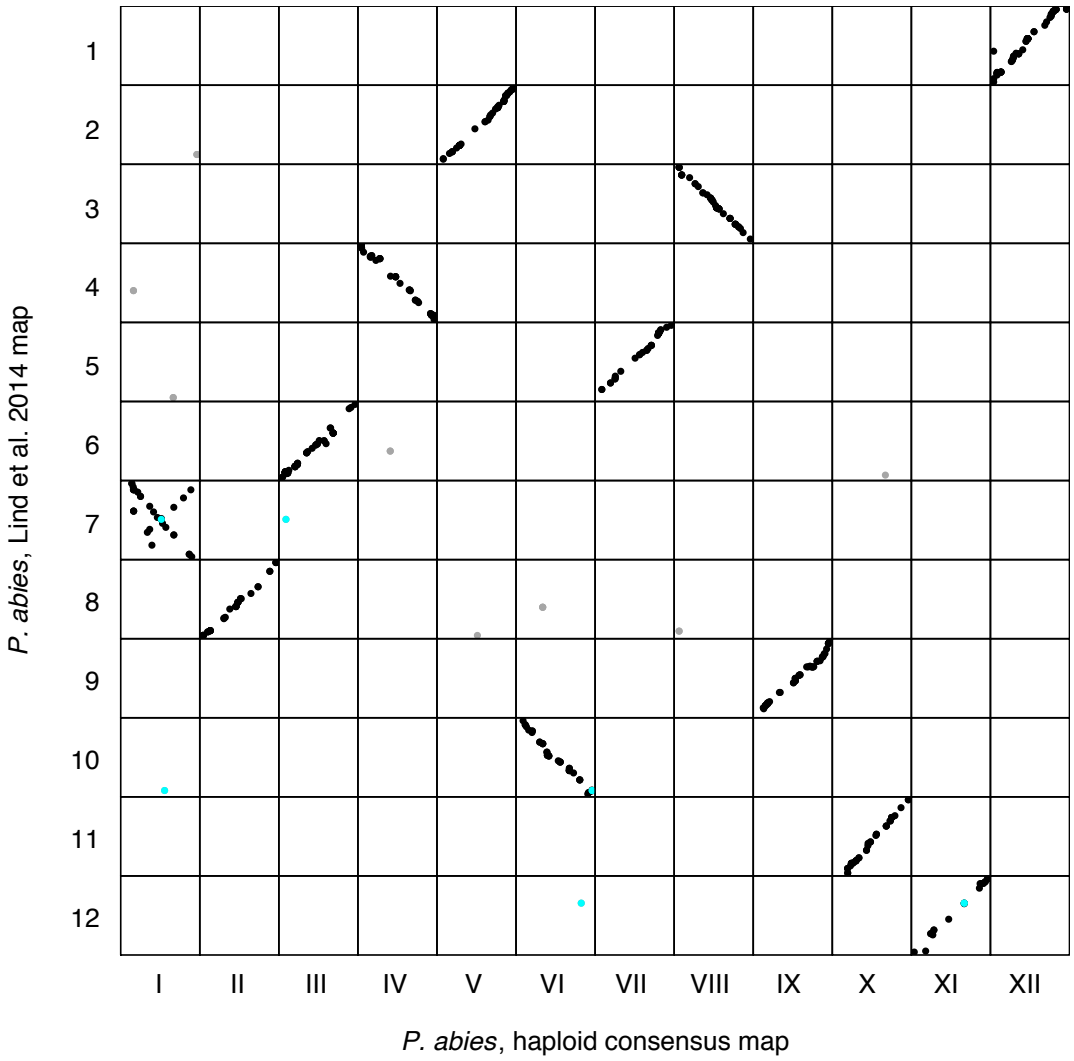


Figure 4: Marker order comparison between Linkage Groups (LGs) from the haploid consensus map presented here and the *Picea abies* map from Lind et al. (2014). Consensus LG I - LG XII are located on the x-axis from left to right. Lind et al. LG 1 - LG 12 are located on the y-axis from top to bottom. Each dot represents a marker comparison from the same scaffold, where black coloration represents the LG where the majority of marker comparisons are mapped. Grey coloration represents markers

415 mapping to a different LG compared to the majority of markers. Turquoise coloration
416 represents markers located on split scaffolds, which are indicative of assembly errors.

417

418 Synteny between *P. abies* and *P. glauca* species was assessed by comparing LG
419 location and marker order between our *P. abies* consensus map and the composite
420 map of *P. glauca* from Pavy et al. (2017). 14,112 comparisons of 4,053 gene models
421 in the composite map in *P. glauca* (Pavy et al. 2017) and 4,310 scaffolds in the *P.*
422 *abies* consensus map were identified at a > 95% identity threshold. 92.7% (13,084 out
423 of 14,112 comparisons) of these were located on homologous LGs while the
424 remaining 7.3% (1,028 comparisons from 388 *P.abies* scaffolds) were distributed
425 across the 12 LGs (Figure 5). 8.2% of all comparisons from multi-probe scaffolds
426 were between non-homologous LGs while 44.3% of all comparisons from split
427 scaffolds were between non-homologous LGs. 31.9% of all non-homologous LG
428 comparisons involved split scaffolds. The correlations of marker order between the
429 two maps were comparable to the correlations we observed between individual
430 component maps in *P. abies* (0.96-0.99), showing that synteny is highly conserved
431 between *P. abies* and *P glauca*.

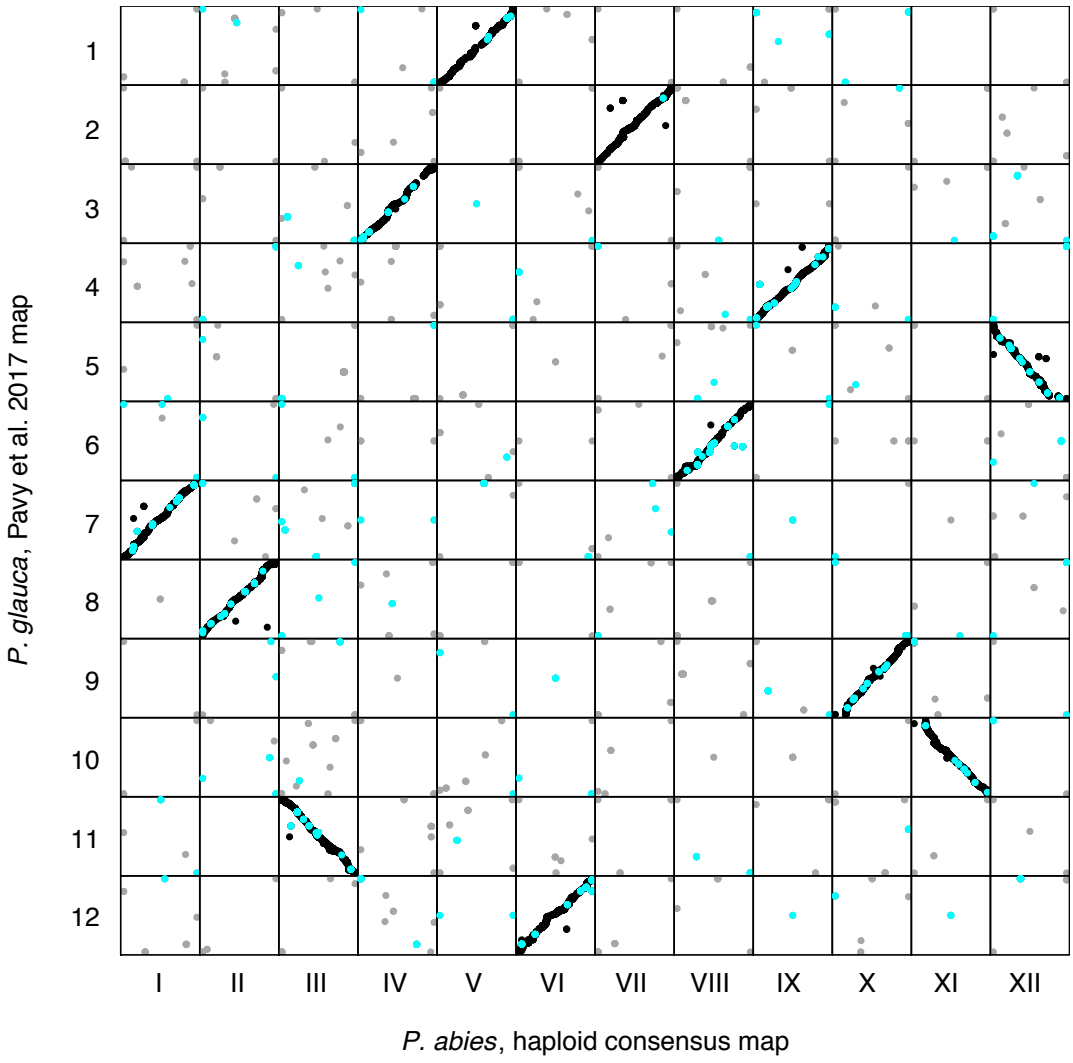


Figure 5: Marker order comparison of Linkage Groups (LGs) between the *Picea abies* haploid consensus map presented here and the *Picea glauca* map from Pavy et al. (2017). Consensus LG I - LG XII are located on the x-axis from left to right. Pavy et al. LG 1 - LG 12 are located on the y-axis from top to bottom. Each dot represents a marker comparison from the same scaffold, where black coloration represents markers mapping to the same LG in the two species, grey coloration represents markers mapping to different LGs. Turquoise coloration represents markers located on split scaffolds, indicating an assembly error.

Population genetic analyses based on the consensus map

22,413 probes, covering 12,908 scaffolds, were used in the population genetic analyses based on the consensus genetic map. On a per-probe basis, we observed substantial variation in all neutrality statistics, with the number of segregating sites ranging from 0 - 77 (mean 15.9), nucleotide diversity (π) from 0 - 0.4 (0.005), Zns from 0 - 1 (mean 0.04) and Tajima's D from -2.4 – 3.5 (mean -0.85). To study large-scale trends and possible chromosomal differences we performed sliding window analyses across the LGs for the different summaries (Figure 6). One interesting large-scale feature we observed was that SNP densities were often highest at the distal or central regions of LGs, indicating the possible location of centromeres and telomeres, for which recombination rates are expected to be reduced (Gaut et al. 2007) and where we hence would expect higher densities of probes per cM (Figure 6a). The large-scale analyses also revealed several instances where entire chromosomal arms might be experiencing different evolutionary patterns (Figure 6b-c). Finally, we identified regions that appear to be evolving under the influence of natural selection. For instance, several regions showed higher than average levels of nucleotide diversity and positive Tajima's D (e.g. on LG IV, V and XII), suggesting that they might harbor genes under balancing selection. Similarly, regions with low nucleotide diversity, an excess of rare alleles and strong linkage disequilibrium (i.e. negative Tajima's D and high Z_{ns} scores, e.g. on LG III) could indicate regions harboring possible selective sweeps (Figure 6c-d).

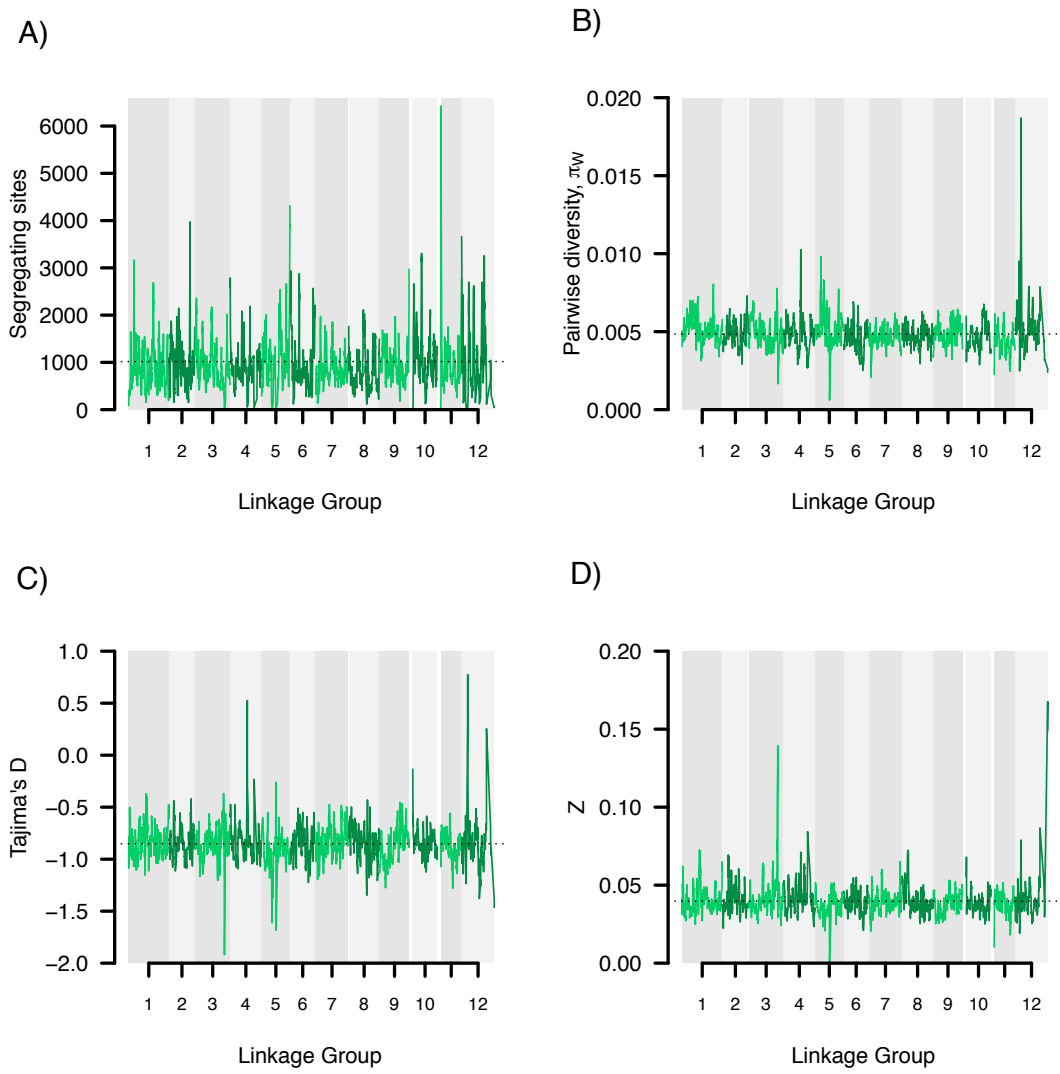


Figure 6. Sliding window analysis of neutrality statistics. Analyses were performed using 10 cM windows with 1 cM incremental steps along the consensus map linkage groups and visualized using coloring alternates between adjacent LGs. A) Number of segregating sites. Dashed horizontal line indicates the overall average of 1017. B) Pairwise nucleotide diversity (π). Dashed horizontal line indicates the overall average of 0.005. C) Tajima's D. Dashed horizontal line indicates the overall average of -0.852. D) Linkage disequilibrium Zn scores. Dashed horizontal line indicates the overall average of 0.040.

Discussion

This is, to our knowledge, the densest genetic linkage map ever created for a conifer species and possible for any tree species. We successfully used this genetic map to

476 anchor 1.7% of the 20 Gbp *P. abies* genome, corresponding to 2.8% of the v1.0
477 genome assembly (Nystedt et al. 2013), to 12 LGs, constituting the haploid
478 chromosome number (Sax and Sax 1933). The *P. abies* genome has a very large
479 proportion of gene-poor heterochromatin, so while the fraction of the genome that we
480 successfully anchored to the assembly is relatively small, those anchored scaffolds
481 cover 24% of all gene-containing assembly scaffolds and 25% of all partially
482 validated gene models from Nystedt et al. (2013).

483 The individual LGs from the three component maps (36 LGs from three
484 independent maps) consisted of 648-1,967 probe-markers and 305-1,185 probe-
485 marker bins and, as such, it was not feasible to analyze the maps using an exhaustive
486 ordering algorithm (Mollinari et al. 2009). Instead, we used RECORD (Van Os et al.
487 2005) with 16 times counting, parallelized over 16 cores and with reordering of
488 markers within 10 marker windows, for each LG to determine the most likely marker
489 order. An heuristic approach, such as RECORD, will undoubtedly introduce some
490 errors in marker ordering (Mollinari et al. 2009), but analyses from simulated data
491 suggested that the average distance between estimated and true marker position is
492 small (< 5 markers) for data sets of similar size to ours (Schiffthaler et al. 2017).
493 However, reliable marker ordering requires robust data and the more genotyping
494 errors and missing data that are present, the harder it will be to determine the true
495 order. This in turn will impact the final size of the map, where both errors in marker
496 order and genotyping results in inflation in the size of the map (Cartwright et al.
497 2007).

498 By collecting our 2,000 megagametophytes from what we initially thought were
499 five different ramets of Z4006, we accidentally sampled material from at least three
500 unrelated families. This error stemmed from a mix-up of genotypes due to wrong

assignment of ramet ID to the different ramets in the seed orchard. Unfortunately, we were not able to assess which megagametophytes were collected from the different putative ramets since the seed bags were pooled prior to DNA extraction and the sampling errors were not detected until after all sequencing was completed. We used a PCA and hierarchical clustering approach to assign samples into three independent clusters, representing three putative maternal families. We also used PCAs of the putative individual families to verify that these clusters were consistent with offspring derived from a single mother tree (Supplementary, Figure S3). However, we nevertheless cannot completely rule out that a small fraction of samples have been incorrectly assigned to the three families and this would lead to inflated map sizes by introducing an excess of recombination events. Another potential confounding issue is tissue contamination. *P. abies* megagametophytes are very small and are surrounded by a diploid seed coat that needs to be removed prior to DNA extraction. If traces of the diploid seed coat remain in the material used for DNA extractions, the haploid samples will be contaminated with diploid material. To identify and eliminate this possibility, we called sequence variants using a diploid model and any heterozygous SNP calls were subsequently treated as missing data. Samples with a high proportion of heterozygous (>10 %) or missing calls (>20%) were excluded from further analyses to reduce the possibilities of genotyping error due to tissue contamination influencing downstream analyses. We estimated map lengths from 100 rounds of subsampling of 100 random probe-marker bins per component LG and used this to demonstrate that individual maps showed size inflations of 0.15-0.31 cM per probe-marker bin. This inflation is on the same order as the map resolutions for the different clusters and, therefore, indicated an average of ~1 genotyping error per probe-marker bin or 11-17 genotyping errors per sample.

526 Both sample- and tissue contaminations can influence the accuracy of the genetic
527 map, both with regards to marker order and map size. The smaller family sizes
528 resulting from dividing our original 2,000 samples into three independent families
529 yielded lower resolution of the three component maps. Fortuitously enough, however,
530 this also enabled us to incorporate more markers into the consensus map since
531 different markers were segregating in the different mother trees from which the three
532 families were derived. Furthermore, it also allowed us to evaluate marker ordering
533 across three independently derived maps. Although our consensus map was 70-90%
534 (60-120% for the individual component maps) larger than previously estimated *Picea*
535 maps (3,556 cM vs. 1,889-2,083 cM), it also contained 2-31 times more markers than
536 earlier maps (Pavy et al. 2012; Lind et al. 2014; Pavy et al. 2017). When comparing
537 marker order between our three independent component maps (Cluster 1-3), we found
538 overall high correlations of marker order (0.94-0.99, Supplementary, Figure S8),
539 which is similar to what has previously been observed between estimated and true
540 positions in maps derived from simulated data without genotyping errors but with
541 20% missing data (Mollinari et al 2009; Schiffthaler et al. 2017). Also, earlier *Picea*
542 maps were all based on diploid F_1 crosses with even the densest composite map
543 containing only 2,300-2,800 markers per framework map (Table 1 - Pavy et al. 2017),
544 compared to our haploid component maps that contained between 3,924 and 11,479
545 probe-marker bins each (Table 2).

546 The comparisons between our haploid consensus map and earlier maps in *Picea*
547 showed an overall high correlation of marker order, which is in line with previous
548 studies suggesting highly conserved synteny within *Picea* and in conifers in general
549 (de Miguel et al. 2015; Pavy et al. 2017). LG I from our haploid consensus map and
550 LG 7 from Lind et al. (2014) showed an inverted order for approximately half of the

551 markers compared (Figure 4). Whether this inversion is due to ordering errors in one
552 of the maps or represents true biological differences between the parents used for the
553 respective maps is, however, not currently known and further investigations are
554 needed to resolve this issue.

555 A small percentage of the marker comparisons in both the intra- and inter-
556 specific maps did not co-align to homologous LGs. Some of these errors likely arose
557 form the repetitive nature of the *P. abies* genome (and conifer genomes in general),
558 where regions with high sequence similarity can often be found interspersed
559 throughout the genome. If the true homologous region between different maps is
560 missing or has been collapsed in the genome assembly due to high sequence similarity,
561 pairwise sequence comparisons may end up assigning homology to regions that are
562 located on different chromosomes. However, it might also be hat these errors
563 represent scaffold assembly errors for scaffolds containing only a single probe-marker
564 or where one region of the scaffold is not captured by the probes, therefore negating
565 evaluation. Approximately 72% of all non-homologues LG comparisons between *P.*
566 *abies* and *P. glauca* were from multi-markers scaffolds (of which 45% were from
567 probe-markers on split scaffolds in the consensus map (turquoise points in Figure 5).
568 The remaining 28% were comparisons with scaffolds that were only represented by a
569 single probe in the consensus map.

570 Four percent of the scaffolds containing multiple makers showed a pattern where
571 different markers mapped to different regions, either within or between LGs in the
572 consensus map. This indicates possible errors in scaffolding during the assembly of
573 the v1.0 *P. abies* genome (Nystedt et al. 2013). If this estimate represents the overall
574 picture for the entire assembly, as many as 400,000 of the ~10 million total scaffolds,
575 and 2,400 of the ~60,000 gene-containing scaffolds, may suffer from assembly errors.

576 Most worryingly, 2% of the multi-marker scaffolds (100/4,859) contained splits that
577 occurred within a single gene model. It is likely that many of these problematic
578 scaffolds stem from incorrect scaffolding of exons from paralogous genes with a high
579 sequence similarity. Since the *P. abies* genome contains a high proportion of
580 repetitive content, that also includes a large number of pseudo-genes, this is perhaps
581 not surprising. Additional work is needed to disentangle these issues and to resolve
582 any assembly errors. False scaffold joins in a genome assembly are not a unique
583 feature for *P. abies*, rather it appears to be a frequent problem in the assembly process.
584 For instance, dense genetic maps in both *Eucalyptus* and *Crassostrea* have identified
585 and resolved false scaffold joins, thereby improving the genome assemblies in these
586 species (Bartholomé et al. 2015; Hedgecock et al. 2015). Our goal for the *P. abies*
587 genetic map was not only to identify incorrect scaffolding decisions in the v1.0
588 genome assembly, but to also help improve future iterations of the genome. Long-
589 read sequencing technologies (e.g. Pacific Bioscience or Oxford Nanopore) could be
590 used to resolve these problematic scaffolds and help disentangle the reasons for their
591 ambiguous localization in the genetic map. A future reference genome for *P. a*bies,
592 based on long read technologies will also be able to utilize this genetic map in a much
593 more efficient way since the resulting assembled scaffolds will be substantially longer
594 and would hence enable anchoring a greater fraction of the genome to LGs, ultimately
595 to the point that chromosome-scale assemblies may be achieved.

596 Our population genetic analyses based on the scaffolds anchored to the consensus
597 map demonstrates the utility of having a dense, accurate genetic map and suggest that
598 the map will facilitate further analyses of genome-wide patterns of variation and
599 selection in *P. abies* in addition to facilitating comparative analyses among spruce
600 species. Assigning even a small fraction of the genome to LGs enabled us to analyze

patterns of genetic diversity in approximately a quarter of all predicted genes. This allowed for analyses of broad-scale patterns of variation across the genome and, as the genome assembly is further improved and an even greater proportion of the assembly if physically anchored to the genetic map, will allow for even more fine-scaled analyses of how different evolutionary forces have interacted in shaping patterns of genetic diversity across the *P. abies* genome.

Acknowledgements

This study was supported by Knut and Alice Wallenberg's foundation through funding to the Norway spruce genome project. AV was partially supported by a grant from the Stiftelsen Gunnar och Birgitta Nordins fond through the Kungl. Skogs- och Lantbruksakademien (KSLA). NRS was supported by the Trees and Crops for the Future (TC4F) project. All computations were performed on resources provided by SciLifeLab and SNIC at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project b2010042.

Author contribution

PKI and MRGG conceived the study. AV collected cones and extracted DNA. CB, AV, DS and JB set up bioinformatics pipeline for analyzing sequence capture data. AV and CB performed PCA and identified samples belonging to the three clusters. CB, DS and BS created the genetic maps. CB and PKI performed intra- and interspecific map comparisons. CB, XW and PKI performed population genetic analysis. CB performed all remaining analyses and wrote the first draft of the manuscript. NRS contributed to manuscript writing and development of the map

624 construction approach. All authors commented on the manuscript at various stages
625 during the writing.

626

627 **Data availability**

628 BatchMap input files for the three clusters, component maps and consensus map files
629 are available from zenodo.org at https://doi.org/10.5281/zenodo.1209841. All scripts
630 needed to recreate the analyses described in the paper are publically available at
631 https://github.com/parkingvarsson/HaploidSpruceMap. Raw sequence data for all
632 samples included in this study are available though the European Nucleotide Archive
633 under accession number PRJEB25757.

634

635

636 **References**

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

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

647 Cartwright, Dustin A, Michela Troggio, Riccardo Velasco, and Alexander Gutin.
648 2007. Genetic Mapping in the Presence of Genotyping Errors. *Genetics* 176:
649 2521–27. doi:10.1534/genetics.106.063982.

650 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E.
651 Handsaker, et al. 2011. The Variant Call Format and VCFtools. *Bioinformatics*
652 27: 2156–58. doi:10.1093/bioinformatics/btr330.

653 De La Torre, Amanda R., Inanc Birol, Jean Bousquet, Pär K. Ingvarsson, Stefan
654 Jansson, Steven J.M. Jones, Christopher I. Keeling, et al. 2014. Insights into
655 Conifer Giga-Genomes. *Plant Physiology* 166: 1724 – 1732.
656 http://www.plantphysiol.org/content/166/4/1724.short.

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

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

665 Drost, Derek R., Evandro Novaes, Carolina Boaventura-Novaes, Catherine I.
666 Benedict, Ryan S. Brown, Tongming Yin, Gerald A. Tuskan, and Matias Kirst.
667 2009. A Microarray-Based Genotyping and Genetic Mapping Approach for
668 Highly Heterozygous Outcrossing Species Enables Localization of a Large
669 Fraction of the Unassembled *Populus Trichocarpa* Genome Sequence. *The Plant*
670 *Journal* 58: 1054–67. doi:10.1111/j.1365-313X.2009.03828.x.

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

674 Farjon, A. 1990. Pinaceae. Drawings and Descriptions of the Genera Abies, Cedrus,
675 Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and
676 Picea. *Pinaceae. Drawings and Descriptions of the Genera Abies, Cedrus,*
677 *Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and*
678 *Picea.* Koeltz Scientific Books.
679 https://www.cabdirect.org/cabdirect/abstract/19920656698.

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

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

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

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

694 Kelly, J. K. 1997. “A Test of Neutrality Based on Interlocus Associations.” *Genetics*

146: 1197–1206.

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

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25: 2078–79. doi:10.1093/bioinformatics/btp352.

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

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

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

Mollinari, M, G R A Margarido, R Vencovsky, and A A F Garcia. 2009. Evaluation

719 of Algorithms Used to Order Markers on Genetic Maps. *Heredity* 103: 494–502.
720 doi:10.1038/hdy.2009.96.

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

725 Pavy, Nathalie, Astrid Deschênes, Sylvie Blais, Patricia Lavigne, Jean Beaulieu,
726 Nathalie Isabel, John Mackay, and Jean Bousquet. 2013. The Landscape of
727 Nucleotide Polymorphism among 13,500 Genes of the Conifer Picea Glauca,
728 Relationships with Functions, and Comparison with Medicago Truncatula.
729 *Genome Biology and Evolution* 5: 1910–25. doi:10.1093/gbe/evt143.

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

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

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

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

743 Schiffthaler B, Bernhardsson C, Ingvarsson PK, Street NR (2017) BatchMap: A
744 parallel implementation of the OneMap R package for fast computation of
745 F_1 linkage maps in outcrossing species. PLoS ONE 12(12): e0189256.
746 https://doi.org/10.1371/journal.pone.0189256

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

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

752 Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by
753 DNA polymorphism. Genetics 123: 585–595.

754 Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin,
755 Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. From
756 FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit
757 Best Practices Pipeline. In *Current Protocols in Bioinformatics*, 11.10.1-
758 11.10.33. Hoboken, NJ, USA: John Wiley & Sons, Inc.
759 doi:10.1002/0471250953.bi1110s43.

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

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

767 Wu, Rongling, Chang-Xing Ma, Ian Painter, and Zhao-Bang Zeng. 2002.

768 Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases

769 in Outcrossing Species. *Theoretical Population Biology* 61: 349–63.

770 doi:10.1006/tpbi.2002.1577.

771