INVESTIGATION

Genomics of Clinal Local Adaptation in *Pinus sylvestris* Under Continuous Environmental and Spatial Genetic Setting

Jaakko S. Tyrmi,*,† Jaana Vuosku,* Juan J. Acosta,‡ Zhen Li,§, Lieven Sterck,§,** Maria T. Cervera,†† Outi Savolainen,*,† and Tanja Pyhäjärvi*,†,1**

*Department of Ecology and Genetics, University of Oulu, FI-90014 Oulu, Finland, †Biocenter Oulu, University of Oulu, FI-90014 Oulu, Finland, ‡Camcore, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, §Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium, **VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium, and ††Centro de Investigación Forestal (CIFOR), Instituto Nacional de Investigaciones Agrarias (INIA), 28040 Madrid, Spain

ORCID IDs: 0000-0002-4757-6563 (J.S.T.); 0000-0001-9701-5759 (J.V.); 0000-0002-9429-5166 (J.J.A.); 0000-0001-8920-9270 (Z.L.); 0000-0001-7116-4000 (L.S.); 0000-0001-6797-2347 (M.T.C.); 0000-0001-9851-7945 (O.S.); 0000-0001-6958-5172 (T.P.)

ABSTRACT Understanding the consequences of local adaptation at the genomic diversity is a central goal in evolutionary genetics of natural populations. In species with large continuous geographical distributions the phenotypic signal of local adaptation is frequently clear, but the genetic basis often remains elusive. We examined the patterns of genetic diversity in *Pinus sylvestris*, a keystone species in many Eurasian ecosystems with a huge distribution range and decades of forestry research showing that it is locally adapted to the vast range of environmental conditions. Making *P. sylvestris* an even more attractive subject of local adaptation study, population structure has been shown to be weak previously and in this study. However, little is known about the molecular genetic basis of adaptation, as the massive size of gymnosperm genomes has prevented large scale genomic surveys. We generated a both geographically and genomically extensive dataset using a targeted sequencing approach. By applying divergence-based and landscape genomics methods we identified several loci contributing to local adaptation, but only few with large allele frequency changes across latitude. We also discovered a very large (ca. 300 Mbp) putative inversion potentially under selection, which to our knowledge is the first such discovery in conifers. Our results call for more detailed analysis of structural variation in relation to genomic basis of local adaptation, emphasize the lack of large effect loci contributing to local adaptation in the coding regions and thus point out the need for more attention toward multi-locus analysis of polygenic adaptation.

KEYWORDS
adaptation
population genetics – empirical
landscape genetics
gymnosperms
Pinus sylvestris
Local adaptation
Targeted DNA Sequencing
Structural Variation

Populations of species with vast continuous distributions can inhabit very different environments. These populations are often locally adapted, defined as each population having higher fitness than any introduced population at its home site (Kawecki and Ebert 2004), preferably demonstrated by performing reciprocal transplant experiment (Savolainen *et al.* 2013) as has been done with many plant species, such as the *Arabidopsis* genus (*e.g.*, Leinonen *et al.* 2011; Ågren and Schemske 2012; Hämälä *et al.* 2018). Local adaptation can also be inferred from patterns of phenotypic variation or environmental correlation, as has been shown for example, in *Drosophila melanogaster* (Adrion *et al.* 2015), humans (Fan *et al.* 2016) and also forest trees (Giertych 1991; Mimura and Aitken 2007; Savolainen *et al.* 2007; Alberto *et al.* 2013b; Prunier *et al.* 2016; Gárate-Escamilla *et al.* 2019; Pyhäjärvi *et al.* 2020).

Local adaptation with a polygenic basis has received more attention lately, because a great deal of adaptive variation is quantitative with multiple underlying loci (Buckler *et al.* 2009; Rockman 2012;

doi: https://doi.org/10.1534/g3.120.401285
Manuscript received February 25, 2020; accepted for publication June 1, 2020; published Early Online June 16, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.12403448.
[1]Corresponding author: Department of Ecology and Genetics, University of Oulu, FI-90014 Oulu, Finland. E-mail: tanja.pyhajarvi@oulu.fi

Berg and Coop 2014; Yeaman 2015; Hall *et al.* 2016; Boyle *et al.* 2017), also in forest trees (Lind *et al.* 2018). A well-known model of polygenic adaptation of a single population in a new environment is presented by the Fisher/Orr model (Fisher 1918; Orr 1998) which predicts an exponential distribution of QTL effects (see Barton & Keightley 2002). However, local adaptation arises due to differential selection in different populations in variable environments, possibly connected by gene flow. This kind of selection often results in phenotypic clines (Huxley 1938). Several theoretical predictions for the underlying genetic architecture of clines have been proposed. While differential selection along environmental gradients in continuous populations on single locus governed traits is expected to result in allele frequency clines (*e.g.*, Slatkin 1973), for polygenic models the expectations are more complex.

Barton (1999) has examined a model with polygenic architecture where a subset of loci will have successive, sharp allele frequency clines along the environmental gradient and maintain the phenotypic mean close to the optimum. The underlying loci are, perhaps unrealistically, expected to have similar effect size on the trait. Such sharp allele frequency clines would be seen as F_{ST} outliers, although the majority of loci governing the underlying traits may remain undetected as most alleles are expected to stay near fixation throughout the range in this model.

Latta (1998, 2003) and LeCorre and Kremer (2003, 2012; Kremer and Le Corre 2012) show that in a high gene flow and strong selection scenario of a polygenic trait, the contribution of covariance between loci becomes more important than between population allele frequency differentiation. Also, in a simulation study Yeaman (2015) showed that local adaptation is indeed a possible outcome even when only small effect alleles are present given that there is enough standing genetic variation. Most importantly in this model, the contributions of individual loci may be transient making detection of the contributing loci more difficult. Nonetheless, allele frequency clines have been observed in many empirical studies (*e.g.*, Schmidt *et al.* 2008; Adrion *et al.* 2015; Wang *et al.* 2018).

Even with genome-wide datasets of tens or hundreds of thousands of loci sampled across localities, identifying the loci underlying adaptive clinal variation remain a challenge. The majority of methods for uncovering adaptive loci are based on the island model of population structure (Lewontin and Krakauer 1973; Foll and Gaggiotti 2008; Excoffier *et al.* 2009; Vitalis *et al.* 2014) and do not fully utilize the spatial information on the clinal genetic variation. Environmental association analysis (Coop *et al.* 2010) and simple regression models can be used to identify clinal trends (Ma *et al.* 2010; Chen *et al.* 2012; Kujala *et al.* 2017).

It is also important to consider putative effects of recombination across adaptive loci on the genetic architecture of local adaptation. The effect of gene flow is expected to override the effect of weak differential selection on a particular locus. However, physical linkage between multiple small effect alleles makes them behave like a single large effect allele, as described by Yeaman and Whitlock (2011). They show that local adaptation under gene flow may favor genetic architecture where recombination is reduced between loci contributing to local adaptation, which may be caused by physical proximity, transposable element action, translocations or inversions (Kirkpatrick and Barton 2006). This will result in increased linkage disequilibrium (LD), and thus the examination of unusual LD patterns may be a fruitful approach in discovering the genetic architecture of local adaptation.

Pinus sylvestris (Scots pine) is a widely distributed conifer species with nearly three centuries of history as a study subject of forestry, ecology and adaptive variation (Pyhäjärvi *et al.* 2020). The range of *P. sylvestris* spans a huge distribution area in Eurasia from southern Spain to northern Scandinavia and eastern Russia. Its distribution is mostly continuous displaying only limited population structure in the nuclear genome, with the exception of some of the more isolated populations for instance in Spain and Italy (Karhu *et al.* 1996; Pyhäjärvi *et al.* 2007; Kujala and Savolainen 2012). However, differentiation within the main range can be seen in mitochondrial haplotype structure, providing information about the recent colonization routes of *P. sylvestris* (Cheddadi *et al.* 2006; Naydenov *et al.* 2007; Pyhäjärvi *et al.* 2008).

In *P. sylvestris* multiple latitudinal phenotypic clines have been repeatedly observed in traits important for abiotic adaptation such as cold tolerance (Eiche 1966; Aho 1994; Hurme *et al.* 1997, 2000) and the timing of growth start and cessation (Mikola 1982; Beuker 1994; Karhu *et al.* 1996). These traits vary latitudinally with environmental conditions, such as temperature, day length, UV radiation intensity and seasonality. Common garden experiments have shown that these traits have a considerable genetic basis suggesting local adaptation (Giertych 1991; Savolainen *et al.* 2007; Alberto *et al.* 2013a; Pyhäjärvi *et al.* 2020). When searching for genomic basis of local adaptation, demographic effects may lead to spurious signals if the underlying population structure remains unaccounted for (Hoban *et al.* 2016). Lack of genome-wide structure, together with highly differentiated phenotypic variation, makes *P. sylvestris* an ideal species for investigating the genetic basis of local adaptation in a large genome. Furthermore, the low level of LD (Wachowiak *et al.* 2009) and lack of any known hybridization with other species should aid in detecting non-equilibrium patterns in the genome. The genetic basis of the adaptation remains largely unknown even though some details have been uncovered in previous studies using data from few candidate genes. Only few F_{ST} outliers have been found, but several cases of latitudinal allele frequency clines and variants associated to timing of bud set have been uncovered (Kujala and Savolainen 2012; Kujala *et al.* 2017). Similar observations of allele frequency clines have been made in other tree species as well, such as *Populus* (Ma *et al.* 2010; Evans *et al.* 2014) and *Picea* (Holliday *et al.* 2010; Chen *et al.* 2012). However, many of these important traits likely have polygenic architecture, possibly complicating efforts in detecting the underlying genetic variation (Lind *et al.* 2018).

Clinal variation, genetic differences across the range and the effect of natural selection in *P. sylvestris* are obvious at the phenotypic level. In this study we create the first genome-wide dataset of *P. sylvestris* to examine patterns of genetic diversity and to search for genomic signature of local adaptation. Similarly to e.g. Yeaman *et al.* (2016), the use of exome capture instead of whole genome sequencing makes it feasible to sequence large number of samples in this species with a genome size of 23.6 Gbp (Zonneveld 2012), or roughly 7 times larger than the human genome. The use of exome capture allows the examination of a significant portion of the coding sequence for genetic diversity, long range LD patterns and the detection of large structural variants. However, this kind of data are not well suited – especially in *P. sylvestris* with limited genomic resources – for sliding window type of analysis or comparing features of coding, non-coding and intronic areas. In this study we use this novel data set sampled from wide geographical area to answer the following questions regarding the manifestation of the phenotypic patterns at the molecular level of variation: 1) Is the commonly applied discrete island population model properly describing the distribution of genetic diversity, or is a model incorporating continuous isolation-by-distance more suitable for a widely distributed and wind dispersed

species such as *P. sylvestris*? 2) Is the strong local adaptation at the phenotypic level reflected in the genetic diversity as corresponding allele frequency clines, high differentiation among populations or increased LD in vicinity of selected sites in *P. sylvestris*?

MATERIALS AND METHODS

Plant material and genotyping

Seeds from 109 *P. sylvestris* samples from 12 populations spanning 31 degrees of latitude were used in generating the dataset for this study (Figure 1, Table 1). The main sampling area included two latitudinal gradients, one from northern Finland to Poland and another north-south gradient in western Russia, to increase power of genome scans (Lotterhos and Whitlock 2015). A total of 120 samples were initially genotyped, of which one was later removed due to sampling the same tree twice and additional 10 were removed due to low sequencing coverage. Haploid genomic DNA was extracted from megagametophyte tissue by using E.Z.N.A. SP Plant DNA kit (Omega Biotek). DNA was fragmented to an average length of 200 nucleotides with Bioruptor ultrasonicator (Diagenode). Libraries were prepared by using NEBNext DNA Library Prep Master Mix Set for Illumina and NEBNext Multiplex Oligos for Illumina E7600S (New England BioLabs) to multiplex libraries of four samples. Targeted capture was performed for each pool according to MycroArray MYbaits protocol v.2.3.1.

Bait design for targeted sequence capture

Bait design was based on a set of *P. sylvestris* transcripts described previously (Li *et al.* 2017). Briefly, transcriptomes of *P. sylvestris* were assembled from 454 read data derived from different developmental stages using the Newbler software (v2.8.1). Those were then integrated with public transcriptomes from PlantGDB-assembled Unique Transcripts (based on GenBank release 187) and a public set of EST assemblies. This initial set contained 121,538 transcripts on which 36,106 open reading frames (ORFs) were predicted by TransDecoder (r20131117). The ORFs were then mapped against the repeat masked *P. taeda* reference genome version 1.01 (Neale *et al.* 2014) with gmap (Wu and Nacu 2010) in order to obtain exon sequences. An ORF was omitted if it could be mapped equally well to several locations of the reference suggesting a paralogous sequence. In total 10,330 ORFs encompassing an area of 12,221,835 bp were selected as targets for initial bait design.

MycroArray MYbaits (Ann Arbor, MI) service was used to create an initial set of 100 base long baits with 2x tiling resulting in a total of 176,334 baits. Four pilot experiments including target capture and sequencing were then conducted to determine bait performance. The putative position of each bait in the genome was determined by aligning the bait sequences to the unmasked *P. taeda* reference genome v. 1.01 (Neale *et al.* 2014) with blastn. A well working bait was defined as having a unique high-quality hit to omit possible paralogous sequences, at least 75 out of 100 bases aligning to omit baits on exon-intron boundaries and less than 4% mismatches to ensure successful alignment. To analyze bait ability to capture target areas 2 × 100 bp paired-end sequencing reads were generated in total for 32 *P. sylvestris* megagametophyte and needle samples with Illumina Hiseq 2500 instrument and 2 × 150 bp paired-end reads with MiSeq instrument. Baits failing to capture any sequence were omitted from the final bait set. After filtering, 60,000 high quality bait sequences were selected as the final bait set (File S1) that was used for examining the 109 samples used in this study. Sequencing was performed using Illumina HiSeq 2500 instrument at Institute of Molecular Medicine Finland (FIMM), by multiplexing four randomly selected samples to each lane, with 100-bp paired-end reads

Genotype calling workflow

Raw reads from Illumina sequencing were aligned to unmasked *P. taeda* reference genome version 1.01 with bowtie2 version 1.1.1 (Langmead and Salzberg 2012) using parameters to include only properly paired alignments (–no-mixed) omit discordant alignments (–no-discordant), and to omit results with no proper alignment (–no-unal). The resulting SAM files were modified with Picard toolkit (http://broadinstitute.github.io/picard/) and SAMtools (Li *et al.* 2009) by converting SAM files to BAM format with SamFormatConverter, sorting with SortSam, removing duplicate sequences with MarkDuplicates, defining read groups with AddOrReplaceReadGroups and indexing with SAMtools index. Examination of alignments revealed that despite omitting targets at known paralogous areas in the bait design stage, many baits captured paralogous sequences from unknown areas not covered in the *P. taeda* reference genome. This was causing issues in read alignment and often leading to spurious SNP calls. The process of detecting the issue and circumventing incorrect SNP calls is described in supplementary methods (File S2). In short, the SNP calling was performed twice using freebayes, first to detect problematic areas identified as heterozygous SNP calls not expected when sequencing haploid DNA, and second time to call SNPs only in problem-free areas.

The technical quality was evaluated by generating a fastqc quality report for raw reads, SAMtools flagstat report for alignment success, along with visual inspection of alignments with SAMtools tview and Integrative Genomics Viewer (Thorvaldsdóttir *et al.* 2013). Based on these reports, 10 samples were removed due to low technical quality. Variant calls were filtered with VCFtools (Danecek *et al.* 2011) to remove sites with quality score below 30 and read depth lower than 5. The entire variant position was removed if it contained non-SNP variants, non-biallelic variants, or had more than 33% missing data. The final high-quality dataset contained 81,301 SNPs (File S3). The genotype calling workflow was parallelized using workflow management software STAPLER (Tyrmi 2018).

Diversity and population structure

To estimate the levels of genetic diversity, pairwise nucleotide diversity (Nei and Li 1979) was calculated with a modified version of python script provided in Garner *et al.* (2016) (File S3). The size of available genome used for analysis was 3.8 Mbp. To calculate Tajima's D and pairwise ST the SNP data set was filtered with vcftools–thin parameter to remove variants closer than 10 kbp from each other to reduce correlation between sites due to physical proximity. After filtering, a set of 4,874 SNPs were available.

Tajima's D estimates (Tajima 1989) were calculated for the whole dataset and also for each population separately using ∂a∂i (Gutenkunst *et al.* 2009). The allele frequency spectrum was generated using ∂a∂i with first down-projecting the sample size to 86 to account for missing data. Hudson *et al.* (1992) pairwise F_{ST} values were then calculated for each population pair by using the equation presented – and recommended over the Weir and Cockerham estimate (Weir and Cockerham 1984) – in Bhatia *et al.* (2013) for a two-population, bi-allelic scenario. An unbiased genome-wide estimate of F_{ST} for each population pair was obtained by calculating the nominator and denominator of equation 10 presented in Bhatia *et al.* (2013) separately for each site, averaged over all sites after which the division was performed.

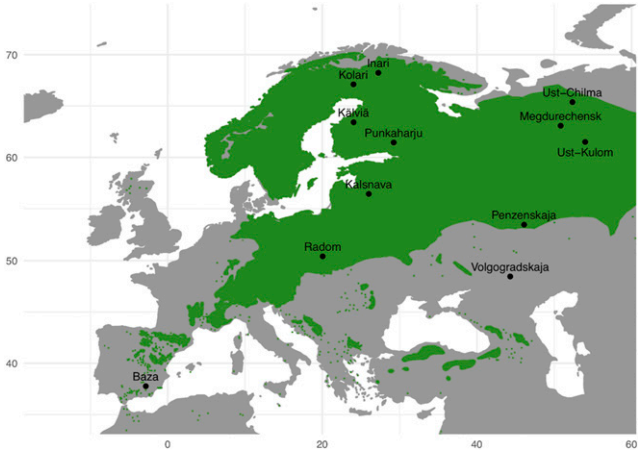

Figure 1 Map of sampling locations with *P. sylvestris* distribution is marked in green color.

Population structure was also examined with principal component analysis (PCA) (McVean 2009) using the prcomp R package. Sample size was evened between different populations to six as uneven sample sizes may distort the PCA projections. For the analysis singleton variants were removed as recommended for instance by Galinsky *et al.* (2016).

Population structure was further analyzed by using STRUCTURE software (Pritchard *et al.* 2000). As STRUCTURE is computationally demanding the SNP set was stringently filtered to obtain a smaller high-quality data set of 4,197 bi-allelic SNPs. At least 10 kbp distance between variants, minor allele frequency of 0.3 and maximum proportion of missing data per site of 0.2 were used. Burn-in length of 250,000 and run length of 50,000 steps were used. K values from 1 to 10 were tested with three replicate runs for each value of K. The software package Clumpak (Kopelman *et al.* 2015) was used to visualize the results and in determining the most likely value of K by using the method of Evanno *et al.* (2005).

The R-package conStruct (Bradburd *et al.* 2018) was used for spatial analysis of population structure. It allows explicit testing for presence of isolation-by-distance, often found in continuous populations and thus reduces the probability of overestimating the number of potential clusters. Two models, non-spatial which is similar to the model the ADMIXTURE software uses (Alexander *et al.* 2009), and a spatial model which accounts for isolation-by-distance patterns, were tested. For both models K-values from 1 to 5 were examined using 50,000 iterations in each. To test whether the spatial or non-spatial model better explains the genetic variation and to compare results between different K-values the cross-validation pipeline provided with conStruct was used with 50 iterations up to K value of 8 (Figure S2).

The level of LD in the dataset was estimated by nonlinear regression of r^2 on between-site distance in base pairs (Hill and Robertson 1968). The expected relationship is presented in the following equation, $E(r^2) = \left[\frac{10+\rho d}{(2+\rho d)(11+\rho d)}\right]\left[1+\frac{(3+\rho d)(12+12\rho d+\rho^2 d^2)}{n(2+\rho d)(11+\rho d)}\right]$, where n is the sample size, d is a distance between adjacent sites in base pairs, $\rho = 4N_e c$, population recombination rate between adjacent sites and c is the recombination rate (Hill and Weir 1988). r^2 was calculated between all variants located within the same scaffold over all populations as the populations are nearly panmictic according to the conStruct analysis. Singletons were omitted from this analysis.

To detect loci forming allele frequency clines along the sampled latitudinal gradients, possibly indicating that they are under varying selective pressure along the gradient, a test of generalized linear mixed effect models was fitted for all loci using R package lme4. The first model was created with glmer function for each SNP by setting genotypes as a response variable, population information as a fixed effect and latitude as a random effect. The second model was created similarly but with latitude omitted. The two models were then compared to each other by calculating a p-value with ANOVA to infer whether or not latitude contributes to the model. In addition to latitude we also performed the analysis using 21 other environmental variables downloaded from WorldClim (Hijmans *et al.* 2005). The Baza population was omitted from all selection scan analyses as it was shown to be the only population clearly differentiated from the others in every analysis of population structure.

To identify putative loci responsible for local adaptation we used the program pcadapt (Luu *et al.* 2017). It infers population structure with PCA and then identifies putative outliers with respect to how they are related to the population structure, making it well suited for examining datasets containing isolation-by-distance patterns. We used all SNPs with minor allele count over 10 and with Baza population omitted for generating PCA. The number of principal components to be used in the outlier analysis was chosen as two, by first producing a scree plot (Figure S3) with pcadapt and then applying Cattell's graphical rule. Pcadapt assigns a p-value for each SNP. The p-value distribution (Figure S4) is then used to obtain FDR estimates.

To further detect potential loci underlying local adaptation, we also used the Bayesian F_{ST}-outlier method bayescan (Foll and Gaggiotti 2008) that is based on identifying locus specific components affecting allele frequencies as a signal of selection. Bayescan was run with default parameters with the exception of setting prior odds for

Table 1 Study population location and summary statistic information

Population	Latitude	Longitude	π_4 ($\times 10^{-3}$)	π_0/π_4	Tajima's D	P ($\times 10^{-3}$)
Inari	68° 54′ N	27° 1′ E	3.90	0.404	−0.250	1.63
Kolari	67° 10′ N	24° 3′ E	3.71	0.393	−0.290	1.81
Kälviä	63° 51′ N	23° 27′ E	3.94	0.395	−0.259	1.49
Punkaharju	61° 45′ N	29° 23′ E	3.87	0.395	−0.283	2.27
Kalsnava	56° 43′ N	26° 1′ E	3.79	0.384	−0.299	1.21
Radom	50° 24′ N	20° 3′ E	3.85	0.400	−0.290	1.47
Ust-Chilma	65° 22′ N	52° 21′ E	3.94	0.396	−0.223	1.60
Megdurechensk	63° 4′ N	50° 49′ E	4.04	0.397	−0.283	1.81
Ust-Kulom	61° 30′ N	54° 0′ E	3.76	0.380	−0.301	1.19
Penzenskaja	53° 27′ N	46° 6′ E	4.12	0.398	−0.272	2.16
Volgogradskaja	47° 45′ N	44° 30′ E	4.19	0.400	−0.278	1.89
Baza	37° 46′ N	2° 49′ W	3.82	0.389	−0.156	1.11

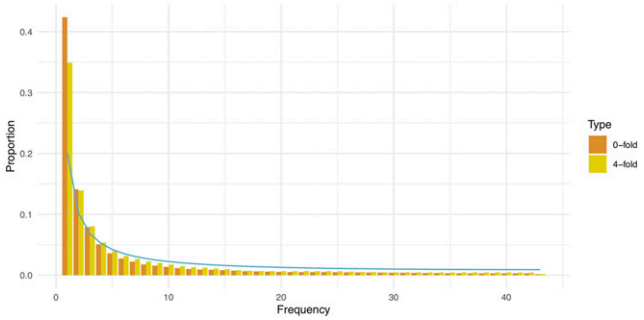

Figure 2 Minor allele frequency spectrum calculated over all samples. Spectrum is projected down to 87 samples to account for missing data. Blue line denotes expected spectrum shape calculated according to equation presented in *Nordborg et al. (2005)* figure 7.

the neutral model to 100 from the default of 10 to account for the large set of SNPs as recommended in the bayescan documentation.

Bayescan analysis also revealed the presence of a large haplotype structure in 11 samples with SNPs in complete LD in several scaffolds. To find all scaffolds included in the haplotype structure an r^2 value was calculated between one of the SNPs contained in the haplotype and all other SNPs in our dataset. Scaffolds containing one or more SNPs with r^2 value of 1.0 in this comparison were then assumed to be part of the haplotype structure. As it is possible that SNPs are in complete LD only by chance, we examined how likely it is that the whole haplotype structure would be due to chance. This was done by randomly choosing 10,000 sets of scaffolds with each set having similar properties to the ones containing the haplotypes and testing how often similar haplotype structure could be seen. More specifically, each randomized set contained similar number of scaffolds (59) with similar number of SNPs (25 per scaffold) and contained at least one SNP with equally high or higher minor allele frequency (11/109) as the scaffolds containing the haplotype. The permutation approach was also used to further study if the haplotype had been affected by selection by calculating within population π and between population F_{ST} and d_{XY} values for the permuted set. F_{ST} is a commonly used measure for differentiation, but it can be affected by reduced within-population nucleotide diversities. d_{XY} is a measure for estimating absolute levels of differentiation and is unaffected by this potential bias, although it is susceptible to bias stemming from unequal sample sizes (Nei and Li 1979; Cruickshank and Hahn 2014). F_{ST} values were calculated as described earlier, d_{XY} was calculated using the following equation: $d_{xy} = \frac{1}{n}\sum_{i=1}^{n} p_{i1}(1-p_{i2}) + p_{i2}(1-p_{i1})$. Results of the permutation analysis were then compared to observed values of the haplotype region.

Data availability

Figure S1 contains p-value distributions for linear regression analysis. Figure S2 contains cross-validation results for conStruct analysis. Figure S3 contains scree plot for pcadapt analysis. Figure S4 contains p-value distribution for pcadapt analysis. Figure S5 contains between-population covariance visualizations for conStruct analysis. Figure S6, S7 and S8 contain boxplot representation for between population F_{ST}, d_{XY} and pairwise within population π for the peculiar haplotype pattern region and for permuted datasets. Table S1 lists interesting outliers for linear regression analysis. Table S2 shows pairwise F_{ST}-values for all population pairs with the putative inversion removed. File S1 contains the final 60,000 bait set sequences in fasta format. File S2 contains additional method description for filtering paralogous variants. File S3 contains SNP-data in vcf-format used in the analysis of this publication. Custom script used in calculating π is provided in File S4. Raw Illumina sequences are available at NCBI SRA with accession number PRJNA592869. Supplemental material available at figshare: https://doi.org/10.25387/g3.12403448.

RESULTS

Nucleotide diversity

Pairwise synonymous nucleotide diversity averaged over populations was 0.0039 with different populations showing similar diversity (Table 1). π_N/π_S ratio was on average 0.394 and again similar levels can be seen in all populations indicating homogenous levels of negative selection across populations. Tajima's D value over all populations was -1.29 and it was also negative within every sampled population with Baza population having a less negative value than others (Table 1). This result is also reflected in the minor allele frequency spectrum calculated for all samples (Figure 2), which shows an excess of rare alleles compared to the standard neutral expectation.

Assessing population structure

We performed a pairwise F_{ST} (Hudson *et al.* 1992), STRUCTURE (Pritchard *et al.* 2000) and PCA (McVean 2009) analysis to evaluate the genetic relationships between populations. All analysis indicate that the Spanish Baza population is differentiated from other populations and in addition to that, very subtle population structure separates eastern and western samples from each other. In general, pairwise F_{ST} estimates show low level of differentiation between most

Table 2 Weighted genome-wide averages of pairwise F_{ST} estimates for all populations

Population	Inari	Kolari	Kälviä	Punkaharju	Kalsnava	Radom	Ust-Chilma	Megdurechensk	Ust-Kulom	Penzenskaja	Volgogradskaja
Kolari	0.013										
Kälviä	−0.001	0.016									
Punkaharju	0.000	0.017	0.005								
Kalsnava	0.019	0.019	0.023	0.020							
Radom	0.002	0.021	0.005	0.007	0.015						
Ust-Chilma	0.010	0.022	0.013	0.010	0.034	0.018					
Megdurechensk	0.008	0.021	0.011	0.009	0.029	0.016	0.001				
Ust-Kulom	0.064	0.045	0.063	0.064	0.044	0.052	0.054	0.052			
Penzenskaja	0.009	0.023	0.011	0.013	0.033	0.015	0.008	0.009	0.060		
Volgogradskaja	0.000	0.017	0.005	0.005	0.023	0.007	0.010	0.006	0.063	0.000	
Baza	0.068	0.082	0.073	0.071	0.077	0.065	0.083	0.077	0.117	0.081	0.072

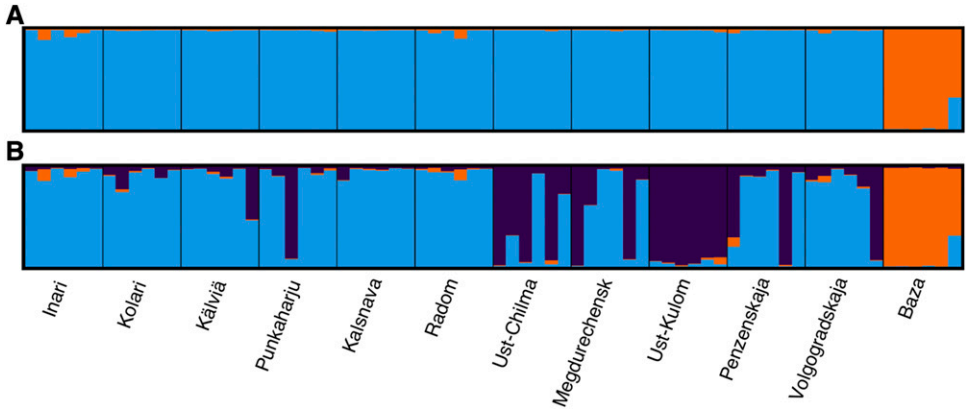

Figure 3 Visualization of STRUCTURE results using K values of 2 (A) and 3 (B).

sample populations (Table 2) with overall F_{ST} of 0.031. Contrasts with Baza are higher (average of 0.079).

The STRUCTURE results (Figure 3) analyzed using Evanno method indicate that the most likely value of K is three. Again, Baza population forms a distinct group compared to the main distribution. However, the rest of the range is divided into two groups, where the other contains all samples of Ust-Kulom population and parts of other, mainly eastern, populations. PCA analysis suggests that the genetic differentiation within and across populations in general is weak as each principal component explains just small fraction of total variance (Figure 4A), although some trends can be observed. The samples originating from Baza population are separated from the rest by the first principal component. The rest of the range is being clustered more closely together, with a trend separating the eastern and western samples from each other (Figure 4B, 4C).

ConStruct analysis was performed using the both non-spatial model and the spatial model (Figure 5), which incorporated information on the geographical distance between populations. A cross-validation test indicated that for the spatial and non-spatial models the predictive accuracy improved with more layers, but only modest improvement can be seen after K2 (Figure S2). At K2 the spatial model has a better fit than the non-spatial model. Even though the K2 model has two layers, the second layer contributes very little (1–2%) to populations other than Baza where it contributes 8% (Figure 5A). We also inspected the value of parameter α_D which controls the shape of the decay of covariance in the spatial model, with values close to 0 indicating no isolation-by-distance (equation

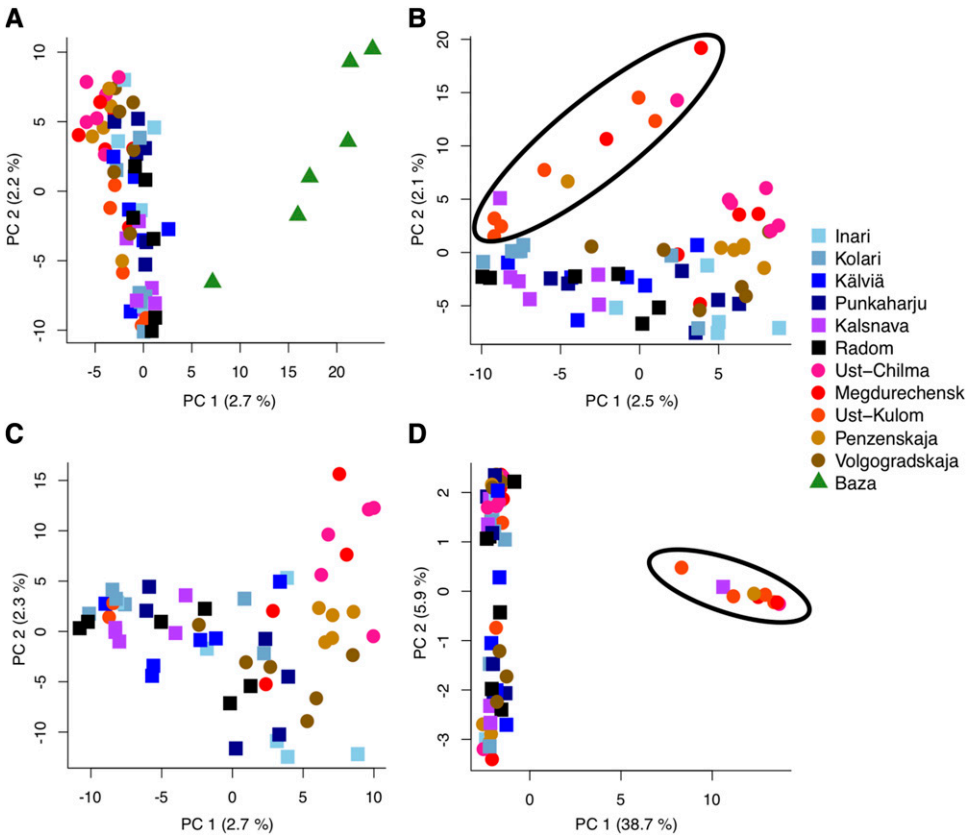

Figure 4 PCA projections of two first principal components of all samples (A), excluding Baza population samples (B), excluding Baza and the samples containing the putative inversion (C) and projection created using variants from putatively inverted area (D). In figures B and D the samples encompassed within black circle contain the putatively inverted haplotype. Circles represent the samples of the western cline, squares the samples of eastern cline and triangles the samples in isolated Baza population. Total variance explained by principal component is indicated within parentheses next to respective principal component axis header.

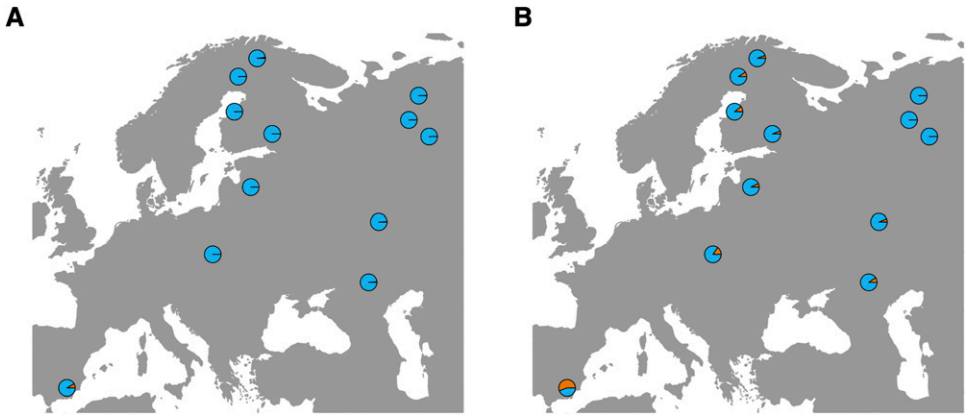

Figure 5 Admixture proportions for two layers estimated for different populations using conStruct spatial (A) and non-spatial (B) models.

3 in Bradburd *et al.*, 2018). The first layer with larger proportion produces a value of 0.0020 for α_D indicating that a very weak isolation-by-distance-pattern can be detected through most of the sampled distribution. Other layer parameters describing the isolation-by-distance control for the sill of covariance matrix in each layer (α_0= 0.0098), control the shape of the decay of covariance with distance in each layer (α_2 = 0.093), global variance due to shared ancestral frequency (γ=0.144) and population specific drift parameters (*i.e.*, nugget values) (0.051-0.058) (Figure S5A for K1 and Figure S5B for K2). Most of the covariance shown in the figure S5B (0.185) is explained by the covariance originating from the same ancestry (*i.e.*, layer). The within population covariance (dots) is slightly higher by 0.06. The contribution of IBD is very small but nonetheless explains the data better than a spatial model without the alpha parameters.

Identifying loci responsible for local adaptation

Comparison of linear regression models with or without latitude was used to identify clinal allele frequency patterns. At a p-value cutoff of 0.001 a total of 12 SNPs were outliers, although due to limitations in the annotation of the reference genome the putative gene of interest could be defined only in three cases. Outliers that could be annotated were the Early-responsive to dehydration stress 1 (ERD1) known to play a role in dehydration stress (Simpson *et al.* 2003), a putative pentatricopeptide repeat containing gene and an ATP−binding cassette transporter gene. With a more lenient p-value cut-off of 0.01 a total of 131 SNPs were outliers, including several genes with interesting function (Table S1). However, controlling for false discovery rate (FDR, Benjamini and Hochberg 1995) with q-values obtained from the p-value distribution suggests that a high proportion of top candidates are false positives as the minimum q-value for the dataset was 0.46. The sampling was designed to particularly detect latitudinal allele frequency clines, but we also downloaded all available environmental variables for each sampling site from WorldClim and performed the same analysis, yielding no outliers with low q-values as with the tests using latitude. The lowest q-value of 0.18 for a top outlier SNP was obtained with longitude. Interestingly, the top outliers seem to have identical allele frequencies population-wise with high minor allele frequency especially in the Ust-Kulom population. These SNPs, although still suggesting a high chance for type I error, were further examined by studying their LD patterns. Other environmental variables also yielded even p-value distributions (Figure S1) suggesting few true outliers.

Bayescan, an F_{ST} outlier detection method, was used to detect putative SNPs underlying local adaptation. Using 0.1 FDR level we obtained a single outlier locus, which is located in non-coding area of the *P. taeda* reference genome v. 1.01 in position 404,961 of tscaffold3905. TreeGenes database (Wegrzyn *et al.* 2008) *P. taeda* annotation also places the area into non-coding area, but blast search of the surrounding sequence against all known gymnosperm genes at ConGenIE (http://congenie.org/) (Sundell *et al.* 2015) revealed that the outlier locus appears to lie within a gene with an unknown function. The outlier locus has a distinct allele frequency pattern

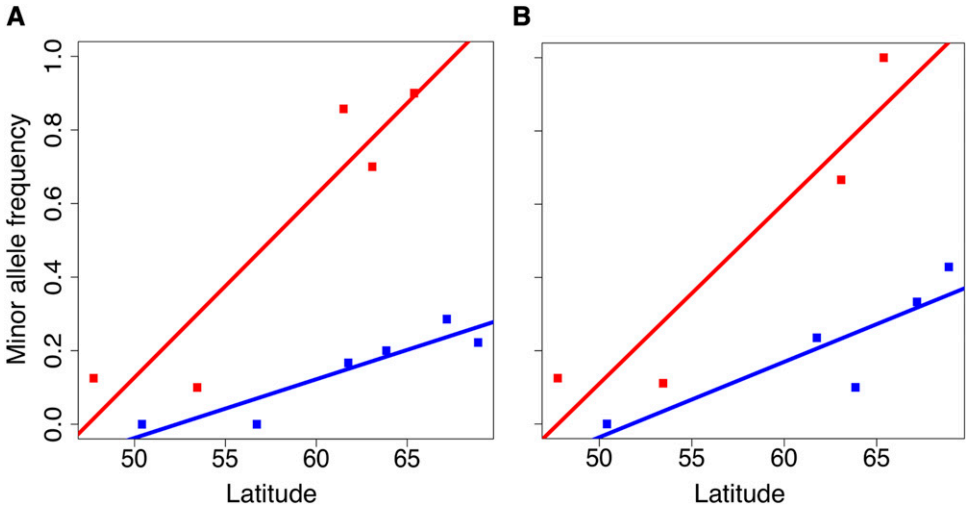

Figure 6 A) Bayescan outlier locus allele frequencies at sampling sites (Y-axis) across latitude of the sites (X-axis). Populations are marked with red (eastern) and blue (western) squares with respective least squares trend line. B) Allele frequency of the second highest scoring bayescan result.

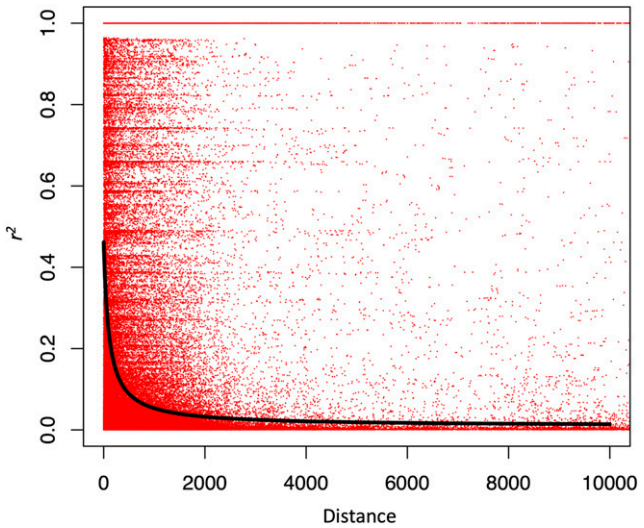

Figure 7 Linkage disequilibrium coefficients (r^2) based on all pairwise SNP comparisons for all samples. Black line shows the squared correlation of allele frequencies r^2 against physical distance between the SNPs (Hill and Weir 1988).

(Figure 6A) where the frequency of the alleles varies along latitude with steeper cline in the east. Interestingly, the second highest scoring variant, although above the 0.1 FDR limit, shows a very similar allele frequency cline pattern to the first outlier (Figure 6B). The variant is located within an intron of a Rubisco gene family member. The two top variants are not in LD nor do they appear to be located in the same scaffold of the *P. taeda* reference genome. The subsequent SNPs in p-value rank, although also above the FDR limit, are the same set of SNPs detected in linear regression analysis with identical allele frequencies.

The pcadapt scan (Luu *et al.* 2017), used for identifying loci under selection by searching for excess divergence along principal components of population structure, assigned a total of 489 SNPs as outliers when a q-value cutoff of 0.1 was used. This set also includes the set of SNPs with identical allele frequencies detected in Bayescan and the linear regression analysis.

Linkage disequilibrium patterns

Linkage disequilibrium patterns (Figure 7) suggest that LD decays quickly within the *P. sylvestris* genome as the r^2 values fall below 0.2 within 145 bp. However, outlier scans revealed that many SNPs in different scaffolds had identical allele frequencies within each population, with particularly high minor allele frequency in the Ust-Kulom population. More careful examination of the LD pattern for these SNPs revealed a distinct haplotype structure in 11 samples, of which five belong to the eastern Ust-Kulom, two to Megdurechensk, two to Penzenskaja, one to Ust-Chilma and one to the Latvian Kalsnava population. We did not find any putative technical explanations for the phenomenon, as stringent parameters were used in the read alignment and SNP calling, where only concordantly, uniquely aligned reads were retained. In addition, the read alignments were visually inspected using SAMtools tview, and no abnormalities were found.

In total 169 variants located in 59 different reference sequence scaffolds had identical allele frequency and LD pattern (Figure 8A). The possibility of detecting such haplotype structure by chance was explored using a permutation test showing that the probability is very low (p-value < 0.001). In the set of samples exhibiting the haplotype structure, the average nucleotide diversity within the 1-kbp area surrounding each outlier variant was only 0.0003, compared to value of 0.0034 within the same region observed between other samples. Average pairwise nucleotide diversity value calculated between the samples exhibiting the haplotype structure and other samples was 0.0093 for the same area consistent with higher than average F_{ST} values. Variants were found to be polymorphic only within the haplotypes, but no variant was polymorphic within both haplotypes, providing further proof that no recombination events between them have taken place.

Westbrook *et al.* (2015) generated a consensus genetic map for *P. taeda* and aligned many of their EST sequences corresponding to the marker data to the *P. taeda* reference genome v. 1.01, thus providing putative physical location for many scaffolds of the reference. There were 12 cases where the EST sequence alignment covers or is within a few kbp of the SNPs belonging to the *P. sylvestris* haplotype structure. In ten of these cases the Westbrook *et al.* (2015) data suggests that the scaffolds belong to linkage group one, but one scaffold seems to be part of linkage group 3 and another on of linkage group 10. The SNPs positioned in linkage group 1 are located between positions 51.12 cM and 94.54 cM. This 43.42 cM area is 23.48% of the total length of the first linkage group. Given that the 12 chromosomes of *P. sylvestris* seem to be similar in size, we can expect each chromosome to be roughly 2.0 Gbp long. We can then, naively, take the proportion the haplotype structure covering the first chromosome's genetic map and apply it to the expected physical size of the chromosome. This approach gives us an estimated size of 470 Mbp for the haplotype structure. Another way for estimating the haplotypes minimum size is to simply sum the lengths of the *P. taeda* reference scaffolds containing the haplotype structure, which gives a size estimate of 35.9 Mbp. As we have a capture target in 11.0% of *P. taeda* reference genome scaffolds, we estimate that the total area covered by the haplotype structure is roughly 35.9 / 0.11 = 326 Mbp.

Allele frequencies of the haplotype structure were highly similar to the detected population structure patterns. To investigate whether the haplotype drove the observed population structure, STRUCTURE, PCA and pairwise F_{ST} analysis were redone without the scaffolds containing the haplotype structure. No change was observed in the STRUCTURE results. However, the F_{ST} results appear to be affected such that lower values are now seen between Ust-Kulom and other populations (Table S2), but the values are still relatively high. It may be that several scaffolds that are part of the inversion are not identified as part of it, as they happen to not contain any informative SNPs for identifying the inversion. In the PCA results, the projection where Baza samples are omitted (Figure 4B), the samples containing the haplotype structure are separated by the second principal component and the other eastern samples are separated from western samples by the first principal component. Removing also the samples that contain the haplotype structure results in projection where Eastern and Western samples are more distinctly separated and occupy the opposing ends of the first principal component. When PCA was performed using only the areas identified as part of alternate haplotype, the samples containing the alternate haplotype structure form a distinct group with the separating principal component explaining large proportion of the variance (Figure 4D).

Permutation test was performed to test whether the patterns of between population F_{ST}, d_{XY} and within population π in the haplotype region are consistent with ongoing positive selection in a subset of populations. Pairwise F_{ST} and d_{XY} values were significantly higher in comparisons between Ust-Kulom and other populations

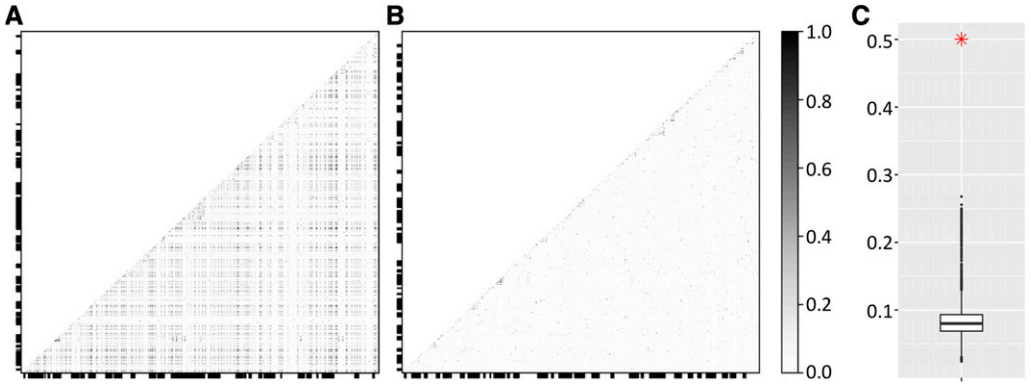

Figure 8 A) Heatmap visualization of allelic correlation coefficient (r^2) values below diagonal calculated between all SNPs identified as being part of an inversion, and all variants within their surrounding 1 kbp areas. Alternating thick and thin X and Y axis borders denotes variants belonging to the same scaffolds. B) Similar heatmap to A, but random variants with similar allele frequency to the inversion were selected along with their 1 kbp surrounding areas to visualize typical linkage disequilibrium patterns. Some scaffolds show LD within them, but between scaffolds mostly only low values r^2 can be seen. C) Means of r^2 values for 10,000 random 1kb areas (one of which is visualized in 1B heatmap) marked in black and the mean r^2 value of blocks containing the putative inversion haplotype marked with red asterisk.

within the haplotype area compared to the permuted data (Figures S6 and S7). Values of π in the Ust-Kulom population were very low between the five samples containing the haplotype structure (0.0006). However, the values between the five samples containing the haplotype and the two samples not containing the haplotype were particularly high (0.0061), as one might expect to observe between diverged haplotypes. This resulted in mean value of π within Ust-Kulom population to be on a similar level to other populations (Figure S8). In other populations where the haplotype pattern was detected in low frequency, the mean π values were particularly high in the affected area due to the high number of pairwise comparisons between the peculiar and normal haplotypes.

DISCUSSION

This first genome-wide analysis of *P. sylvestris* genetic diversity covering a large proportion of its distribution show estimates of genetic diversity and population structure largely in line with previous studies. Outlier scans used for uncovering loci contributing to local adaptation detected many candidate genes, of which many have been shown to be targets of selection in previous studies (Table S1) for example in *A. thaliana* (Knoth and Eulgem 2008) and Eucalyptus (Jordan *et al.* 2017) even though such coincidence does not verify the effect of natural selection in these loci. Furthermore, a previously unidentified large structural variation, possibly related to local adaptation, was uncovered.

Indications of population structure

The neutral nucleotide diversity found in this study is similar to what has been observed in previous studies of this species (Kujala and Savolainen 2012) and in several other conifers (Brown *et al.* 2004; Eckert *et al.* 2013; Grivet *et al.* 2017). Given the mostly continuous distribution of wind-pollinated *P. sylvestris* and the previous findings of near-absent population structure, only very negligible differences between populations were anticipated, with the exceptions of geographically isolated populations (Pyhäjärvi *et al.* 2007; Wachowiak *et al.* 2009; Kujala and Savolainen 2012). It has been suggested that putative *P. sylvestris* refugia during the last glaciation has existed in the Mediterranean area, northern Europe and also in the east, possibly in Ural Mountains (Naydenov *et al.* 2007). Therefore, weak genetic structure resulting from expansion from two distinct refugia, from east and west, was expected.

All approaches of population structure analysis uniformly indicated that the Baza population, geographically isolated from the main distribution, was most, although still weakly differentiated from the other populations. STRUCTURE and PCA analysis also gave some indication that the most likely number of groups is three. However, the models of these frequently used methods do not explicitly account for geographic isolation-by-distance, which can be assumed to exist within the distribution of many species, including *P. sylvestris*. Omission of this phenomenon from the models may cause these methods to spuriously assign populations to separate groups, when the genetic variation could in fact be explained by continuous isolation-by-distance (Bradburd *et al.* 2018). This also seems to have happened with our *P. sylvestris* analysis, where STRUCTURE suggested three distinct clusters, but the conStruct spatial model explains the vast majority of genetic covariance by within-sampling-location effect accompanied with weak isolation-by-distance pattern across populations.

Our results are in contrast with results from many other tree species, such as *Picea abies* where considerable structure has been detected despite many similarities in distribution, population size and reproductive biology (Chen *et al.* 2019). Several studies in *Populus* have also suggested the presence of distinct population structure (Keller *et al.* 2010; Evans *et al.* 2014; Geraldes *et al.* 2014). *P. sylvestris* rarely hybridizes with other species, and is not capable of clonal reproduction, but the exact connection between these characteristics and lack of major population structure is not understood. This lack of genome-wide structure is an advantage when investigating the genetic basis of adaptation, as such structure is a complicating factor in selection scans (Hoban *et al.* 2016).

Putative signs of local adaptation

As *P. sylvestris* is known to be locally adapted to various environmental conditions within its vast distribution (Savolainen *et al.* 2007), we anticipated to identify signs of natural selection in the genomic variation. We performed an F_{ST} -based outlier scan, which identified only single statistically significant outlier SNP with an allele frequency cline in western transect and a particularly strong cline in the eastern transect, with another non-significant variant exhibiting very similar allele frequency cline to the top outlier. The first outlier did not have any reliable annotation, but the second variant was located within an intron of a Rubisco gene, which has been suggested to have a role in ecological adaptation to different temperatures and CO_2 concentrations (Hermida-Carrera *et al.* 2017). A large number of outliers in F_{ST} based selection scan would have been unexpected considering earlier

findings (Kujala and Savolainen 2012) and most theory suggesting that the nature of the underlying genetic architecture is likely highly polygenic, but observing only single outlier is surprising. It is possible that higher number of populations and samples for bayescan analysis would have allowed more outliers to be uncovered or that most of the adaptive variation is in the non-coding regulatory regions not investigated here.

We also applied the *pcadapt* method, which accounts for the possible population structure via principal component analysis and identifies outliers relative to this structure and is well suited for scenarios involving continuous population structure (Forester *et al.* 2018). The approach yielded 489 putative outliers. As expected, the outliers also included the SNPs identified as part of the haplotype structure discussed below. These SNPs were also top outliers in the linear regression analysis of allele frequencies, which also identified some other interesting genes. However, low q-values suggest low probability of true positive outliers in this test.

Several putative explanations exist for detecting low number of outliers in the bayescan and linear regression outlier analysis. First, as the targeted sequencing approach by definition will only allow examination of very small proportion of the genome, much of the adaptive variation cannot be detected. In species with large genomes, proportionately more adaptive variation is expected to be found outside coding region (Mei *et al.* 2018), which may also explain the lack of adaptive signal in the data obtained by exome capture. Second, as discussed above, it is possible that the eastern and western parts of our sampling have in fact originated from different refugia after the most recent glacial period, or several periods as the same refugia may have existed during many or most such periods. Therefore distinctive genetic adaptations may have evolved within each refugium, as suggested for instance by Naydenov *et al.* (2007), rendering in particular landscape genetics approaches ineffective and requiring larger amount of populations sampled from both east and west. Third, as the genetic basis of local adaptation has shown to be mostly polygenic, only a small proportion of all variants can be expected to be under strong enough selection for prolonged period to be detected. Detection of alleles with small effect may require considerably larger sample size. Powerful approaches exist for exploiting very large numbers of samples (Berg and Coop 2014; Field *et al.* 2016; Racimo *et al.* 2018), but even when applying such methods it may be challenging to control for population structure to avoid false positive signal (Berg *et al.* 2019).

Linkage disequilibrium patterns and putative large inversion

Earlier work has shown that LD decays very rapidly within the *P. sylvestris* genome (Dvornyk *et al.* 2002; Pyhäjärvi *et al.* 2007; Wachowiak *et al.* 2009) and also in other conifers such as *P. taeda* (Lu *et al.* 2016; Acosta *et al.* 2019). This study allows for examining longer range patterns of LD than before as in many cases multiple target sequences are positioned within the same scaffold. Our findings are in line with the previous studies showing that r^2 fall below 0.2 within 145 bp. An advantage of low the level of genome-wide LD observed previously, and in this work, is that variants detected in the outlier scans are probably very close to the causative polymorphism (Neale and Savolainen 2004).

Interestingly, the linear regression, bayescan and PCAdapt analysis revealed a large number of SNPs forming an unexpected haplotype structure. Permutation test shows that the LD pattern detected is not expected by chance alone. Analysis of pairwise nucleotide differences of the affected region shows that very low level of nucleotide diversity can be observed within the samples where the haplotype is present, but in the other samples the diversity level seems comparable to average genome-wide level. Simple estimates indicate that it is very likely to be several hundred million base pairs long.

Several biological explanations exist for detecting such haplotype pattern. First, they can be created by partial selective sweeps but considering that the haplotype structure spans 43 cM in the *P. taeda* genetic map, this explanation of sweep does not seem possible under the normal recombination rates. Nonetheless, this explanation cannot be completely dismissed. Extended haplotype patterns may also be caused by natural recombination rate variation along the genome, but this explanation is unlikely in this case, as the haplotype is strictly limited only to specific samples and geographically to the east and it seems unlikely that local recombination rate would be such different only in a subset of individuals. Also, the fact that LD is complete suggests that no recombination events have occurred between the haplotypes.

Considering all the observations we have made of the haplotype structure, an inversion contributing to local adaptation seems to be the most probable cause. Inversions, unlike the other putative explanations discussed above, can create large areas of restricted recombination as they prevent proper chromatid pairing (Andolfatto *et al.* 2001). Several specific methods do exist for identifying inversions (Kosugi *et al.* 2019), but to our knowledge none is applicable to our data, as they require a good quality reference genome, and preferably a whole genome sequence data set. When a complete reference genome for *P. sylvestris* becomes available, a whole genome sequencing combined with a method such as DELLY (Rausch *et al.* 2012) or GRIDSS (Cameron *et al.* 2017) could be used to confirm the existence of a structural variation. Alternatively, it might be possible to compare genetic maps generated from two crosses, with and without the inverted haplotype.

We could find no previous observations of long haplotypes in *P. sylvestris* literature, and only very few mentions of putative inversions in an earlier cytological study (Muratova 1997) in line with relatively strong synteny *e.g.*, beween *P. sylvestris* and *P. taeda* (Komulainen *et al.* 2003). Large inversions have been suggested as being targets of selection in many species (Wellenreuther and Bernatchez 2018), with the largest such inversions exceeding 200 Mbp between two *Helianthus* sister species (Barb *et al.* 2014). To our knowledge, the putative inversion detected in our study remains as the largest one to show any indication to contribute to local adaptation to date, although the relative size the inversion compared to whole genome size is certainly not as large as inversions in some other species. The samples containing the inversion polymorphism seem to be located broadly in the similar, but not identical, geographic area where STRUCTURE and PCA analysis indicate slight population structure border. This may suggest that the inversion event may have occurred within the speculated eastern refugium, but further investigation would be required to uncover the possible origin.

It is plausible that this area is an inversion undergoing rapid increase in frequency due to selection, as it was not only picked out by outlier scans, but also further evidence from the permutation analysis of pairwise F_{ST} and d_{XY} show patterns congruent with ongoing selection as extreme divergence of Ust-Kulom population (Figure S6-S7). However, further empirical studies are needed to confirm the phenotypic and fitness effects of the inversion and to examine whether the pattern is caused by recent locally adaptive allele or a global sweep that has not yet spread into lager geographic area. The ongoing spread of globally beneficial alleles may produce outliers in

allele frequency based analysis (Booker *et al.* 2019). The latter scenario is not as plausible, as globally beneficial allele would likely already have spread in wider geographic area via efficient pollen flow.

Possible role of inversions in local adaptation had been recognized in genetics research early on (Dobzhansky 1970), and the concept of 'supergenes' has since been further explored (Thompson and Jiggins 2014). Inversions may contribute to local adaptation if they encompass more locally adapted alleles than alternate haplotypes (Kirkpatrick and Barton 2006), if the inversion contains alleles with positive epistasis (Feldman *et al.* 1996) or if there is particularly low deleterious mutation load within the inversion (Nei *et al.* 1967). If fitness advantage arises, the inversion will rise in frequency within the geographic area it provides selective advantage until it reaches migration-selection balance. Empirical studies have uncovered inversions contributing to local adaptation for instance in *D. melanogaster* (Kapun and Flatt 2018), sticklebacks (Jones *et al.* 2012), yellow monkeyflower (Gould *et al.* 2018), teosinte (Pyhäjärvi *et al.* 2013), humans (Puig *et al.* 2015) and in many others (Wellenreuther and Bernatchez 2018).

Further studies are required to examine how the haplotype structure and the other detected outliers affect fitness and if they contribute to local adaptation. Regardless of the mechanistic reason for the haplotype pattern, the existence of such geographically restricted haplotype is significant, because they have not been reported in large conifer genomes before.

CONCLUSIONS

In this work we have examined the genetic diversity of *P. sylvestris* along a large portion of its range. Some patterns of population structure can be seen in a marginal population but within the continuous main range the isolation-by-distance explains well any differentiation detected, unlike in many other tree species. This mitigates the issues caused by structure in detecting signs of selection, but our results also show that while clear phenotypic signals of local adaptation have been detected, the molecular background remains largely elusive even if many well-established approaches were used here to detect the signature of selection. However, many interesting outliers were detected that have been shown to contribute to local adaptation in earlier studies. Furthermore, in this study we find a very large putative inversion, likely spanning an area equivalent to several *Arabidopsis thaliana* genomes. To our knowledge, this is the first time that a potentially non-neutral inversion has been shown to segregate in conifers, even though such occurrences can certainly be expected by theory and evolutionary important inversions have observed in wide set of plant species (Yeaman 2013; Huang and Rieseberg 2020).

ACKNOWLEDGMENTS

The authors thank Skogforsk for providing seeds for sequencing, Matias Kirst for helping with developing the exome capture protocol, Gideon Bradburd for help with conStruct, Chedly Kastally for help with genetic map data and members of the Plant Genetics Research group in the University of Oulu for many helpful comments and suggestions. We thank the CSC-IT Center for Science, Finland, for computational resources. This work was supported by European Comission 7th Framework Programme project ProCoGen (289841) to O.S., Biocenter Oulu, Emil Aaltosen Säätiö (160284 O), Oulun Läänin Talousseuran Maataloussäätiö to JT, Academy of Finland (287431 and 293819) to T.P. Z.L. was funded by a postdoctoral fellowship from the Special Research Fund of Ghent University (BOFPDO2018001701).

LITERATURE CITED

Acosta, J. J., A. M. Fahrenkrog, L. G. Neves, M. F. R. Resende, C. Dervinis *et al.*, 2019 Exome resequencing reveals evolutionary history, genomic diversity, and targets of selection in the conifers *Pinus taeda* and *Pinus elliottii*. Genome Biol. Evol. 11: 508–520. https://doi.org/10.1093/gbe/evz016

Adrion, J. R., M. W. Hahn, and B. S. Cooper, 2015 Revisiting classic clines in *Drosophila melanogaster* in the age of genomics. Trends Genet. 31: 434–444. https://doi.org/10.1016/j.tig.2015.05.006

Ågren, J., and D. W. Schemske, 2012 Reciprocal transplants demonstrate strong adaptive differentiation of the model organism *Arabidopsis thaliana* in its native range. New Phytol. 194: 1112–1122. https://doi.org/10.1111/j.1469-8137.2012.04112.x

Aho, M., 1994 frost hardening of one-year-old *Pinus sylvestris* (L.) seedlings: Effect of origin and parent trees. Scand. J. For. Res. 9: 17–24. https://doi.org/10.1080/02827589409382808

Alberto, F. J., S. N. Aitken, R. Alía, S. C. González-Martínez, H. Hänninen *et al.*, 2013a Potential for evolutionary responses to climate change - evidence from tree populations. Glob. Change Biol. 19: 1645–1661. https://doi.org/10.1111/gcb.12181

Alberto, F. J., J. Derory, C. Boury, J.-M. Frigerio, N. E. Zimmermann *et al.*, 2013b Imprints of natural selection along environmental gradients in phenology-related genes of *Quercus petraea*. Genetics 195: 495–512. https://doi.org/10.1534/genetics.113.153783

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664. https://doi.org/10.1101/gr.094052.109

Andolfatto, P., F. Depaulis, and A. Navarro, 2001 Inversion polymorphisms and nucleotide variability in Drosophila. Genet. Res. 77: 1–8. https://doi.org/10.1017/S0016672301004955

Barb, J. G., J. E. Bowers, S. Renaut, J. I. Rey, S. J. Knapp *et al.*, 2014 Chromosomal evolution and patterns of introgression in Helianthus. Genetics 197: 969–979. https://doi.org/10.1534/genetics.114.165548

Barton, N. H., 1999 Clines in polygenic traits. Genet. Res. 74: 223–236. https://doi.org/10.1017/S001667239900422X

Barton, N. H., and P. D. Keightley, 2002 Understanding quantitative genetic variation. Nat. Rev. Genet. 3: 11–21. https://doi.org/10.1038/nrg700

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Berg, J. J., and G. Coop, 2014 A Population genetic signal of polygenic adaptation. PLoS Genet. 10: e1004412. https://doi.org/10.1371/journal.pgen.1004412

Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi *et al.*, 2019 Reduced signal for polygenic adaptation of height in UK Biobank. eLife 8: e39725. https://doi.org/10.7554/eLife.39725

Beuker, E., 1994 Adaptation to climatic changes of the timing of bud burst in populations of *Pinus sylvestris* L. and *Picea abies*. Tree Physiol. 14: 961–970. https://doi.org/10.1093/treephys/14.7-8-9.961

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting FST: The impact of rare variants. Genome Res. 23: 1514–1521. https://doi.org/10.1101/gr.154831.113

Booker, T. R., S. Yeaman, and M. C. Whitlock, 2019 Global adaptation confounds the search for local adaptation. bioRxiv. doi: 10.1101/742247 (Preprint posted August 21, 2019).

Boyle, E. A., Y. I. Li, and J. K. Pritchard, 2017 An expanded view of complex traits: From polygenic to omnigenic. Cell 169: 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Bradburd, G. S., G. M. Coop, and P. L. Ralph, 2018 Inferring continuous and discrete population genetic structure across space. Genetics 210: 33–52. https://doi.org/10.1534/genetics.118.301333

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA 101: 15255–15260. https://doi.org/10.1073/pnas.0404231101

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718. https://doi.org/10.1126/science.1174276

Cameron, D. L., J. Schröder, J. S. Penington, H. Do, R. Molania *et al.*, 2017 GRIDSS: Sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 27: 2050–2060. https://doi.org/10.1101/gr.222109.117

Cheddadi, R., G. G. Vendramin, T. Litt, L. François, M. Kageyama *et al.*, 2006 Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. Glob. Ecol. Biogeogr. 15: 271–282. https://doi.org/10.1111/j.1466-822X.2006.00226.x

Chen, J., T. Källman, X. Ma, N. Gyllenstrand, G. Zaina *et al.*, 2012 Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*). Genetics 191: 865–881. https://doi.org/10.1534/genetics.112.140749

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

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423. https://doi.org/10.1534/genetics.110.114819

Le Corre, V., and A. Kremer, 2003 Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. Genetics 164: 1205–1219.

Le Corre, V., and A. Kremer, 2012 The genetic differentiation at quantitative trait loci under local adaptation. Mol. Ecol. 21: 1548–1566. https://doi.org/10.1111/j.1365-294X.2012.05479.x

Cruickshank, T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23: 3133–3157. https://doi.org/10.1111/mec.12796

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dobzhansky, T., 1970 *Genetics of the evolutionary process*, Columbia University Press, New York.

Dvornyk, V., A. Sirviö, M. Mikkonen, and O. Savolainen, 2002 Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*. Mol. Biol. Evol. 19: 179–188. https://doi.org/10.1093/oxfordjournals.molbev.a004070

Eckert, A. J., A. D. Bower, K. D. Jermstad, J. L. Wegrzyn, B. J. Knaus *et al.*, 2013 Multilocus analyses reveal little evidence for lineage-wide adaptive evolution within major clades of soft pines (*Pinus subgenus* Strobus). Mol. Ecol. 22: 5635–5650. https://doi.org/10.1111/mec.12514

Eiche, V., 1966 Cold damage and plant mortality in experimental provenance plantations with Scots pine in northern Sweden. Stud. For. Suec. 36: 1–219.

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

Evans, L. M., G. T. Slavov, E. Rodgers-Melnick, J. Martin, P. Ranjan *et al.*, 2014 Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nat. Genet. 46: 1089–1096. https://doi.org/10.1038/ng.3075

Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298. https://doi.org/10.1038/hdy.2009.74

Fan, S., M. E. B. Hansen, Y. Lo, and S. A. Tishkoff, 2016 Going Global by Adapting Local: A Review of recent human adaptation. Science 354: 54–59. https://doi.org/10.1126/science.aaf5098

Feldman, M. W., S. P. Otto, and F. B. Christiansen, 1996 Population genetic perspectives on the evolution of recombination. Annu. Rev. Genet. 30: 261–295. https://doi.org/10.1146/annurev.genet.30.1.261

Field, Y., E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton *et al.*, 2016 Detection of human adaptation during the past 2000 years. Science 354: 760–764. https://doi.org/10.1126/science.aag0776

Fisher, R., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Proc. Roy. Soc. 52: 399–433.

Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977–993. https://doi.org/10.1534/genetics.108.092221

Forester, B. R., J. R. Lasky, H. H. Wagner, and D. L. Urban, 2018 Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Mol. Ecol. 27: 2215–2233. https://doi.org/10.1111/mec.14584

Galinsky, K. J., G. Bhatia, P. R. Loh, S. Georgiev, S. Mukherjee *et al.*, 2016 Fast principal-component analysis reveals convergent Evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. 98: 456–472. https://doi.org/10.1016/j.ajhg.2015.12.022

Gárate-Escamilla, H., A. Hampe, N. Vizcaíno-Palomar, T. M. Robson, and M. Benito Garzón, 2019 Range-wide variation in local adaptation and phenotypic plasticity of fitness-related traits in *Fagus sylvatica* and their implications under climate change. Glob. Ecol. Biogeogr. 28: 1336–1350. https://doi.org/10.1111/geb.12936

Garner, A. G., A. M. Kenney, L. Fishman, and A. L. Sweigart, 2016 Genetic loci with parent-of-origin effects cause hybrid seed lethality in crosses between Mimulus species. New Phytol. 211: 319–331. https://doi.org/10.1111/nph.13897

Geraldes, A., N. Farzaneh, C. J. Grassa, A. D. McKown, R. D. Guy *et al.*, 2014 Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. Evolution. 68: 3260–3280. https://doi.org/10.1111/evo.12497

Giertych, M., 1991 Provenance variation in growth and phenology, pp. 87–101 in *Developments in Plant Genetics and Breeding*, Elsevier, Netherlands.

Gould, B. A., Y. Chen, and D. B. Lowry, 2018 Gene regulatory divergence between locally adapted ecotypes in their native habitats. Mol. Ecol. 27: 4174–4188. https://doi.org/10.1111/mec.14852

Grivet, D., K. Avia, A. Vaattovaara, A. J. Eckert, D. B. Neale *et al.*, 2017 High rate of adaptive evolution in two widespread European pines. Mol. Ecol. 26: 6857–6870. https://doi.org/10.1111/mec.14402

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695https://doi.org/. 10.1371/journal.pgen.1000695

Hall, D., H. R. Hallingbäck, and H. X. Wu, 2016 Estimation of number and size of QTL effects in forest tree traits. Tree Genet. Genomes 12: 110. https://doi.org/10.1007/s11295-016-1073-0

Hämälä, T., T. M. Mattila, and O. Savolainen, 2018 Local adaptation and ecological differentiation under selection, migration, and drift in *Arabidopsis lyrata*. Evolution. 72: 1373–1386. https://doi.org/10.1111/evo.13502

Hermida-Carrera, C., M. A. Fares, Á. Fernández, E. Gil-Pelegrín, M. V. Kapralov *et al.*, 2017 Positively selected amino acid replacements within the RuBisCO enzyme of oak trees are associated with ecological adaptations. PLoS One 12: e0183970. https://doi.org/10.1371/journal.pone.0183970

Hijmans, R. J, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, 2005 Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25: 1965–1978.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. TAG Theor. Appl. Genet. 38: 226–231. https://doi.org/10.1007/BF01245622

Hill, W. G., and B. S. Weir, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33: 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd *et al.*, 2016 Finding the genomic asis of local adaptation: pitfalls, practical solutions, and future directions. Am. Nat. 188: 379–397. https://doi.org/10.1086/688018

Holliday, J., K. Ritland, and S. N. Aitken, 2010 Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol. 188: 501–514. https://doi.org/10.1111/j.1469-8137.2010.03380.x

Huang, K., and L. H. Rieseberg, 2020 Frequency, origins, and evolutionary role of chromosomal inversions in plants. Front. Plant Sci. 11: 296. https://doi.org/10.3389/fpls.2020.00296

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Hurme, P., T. Repo, O. Savolainen, and T. Pääkkönen, 1997 Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). Can. J. For. Res. 27: 716–723. https://doi.org/10.1139/x97-052

Hurme, P., M. Sillanpää, T. Repo, E. Arjas, and O. Savolainen, 2000 Genetic basis of climatic adaptation in Scots pine by Bayesian QTL analysis. Genetics 156: 1309–1322.

Huxley, J., 1938 Clines: an auxiliary taxonomic principle. Nature 142: 219–220. https://doi.org/10.1038/142219a0

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61. https://doi.org/10.1038/nature10944

Jordan, R., A. A. Hoffmann, S. K. Dillon, and S. M. Prober, 2017 Evidence of genomic adaptation to climate in *Eucalyptus microcarpa*: implications for adaptive potential to projected climate change. Mol. Ecol. 26: 6002–6020. https://doi.org/10.1111/mec.14341

Kapun, M., and T. Flatt, 2018 The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. Mol. Ecol. 28: 1263–1282. https://doi.org/10.1111/mec.14871

Karhu, A., P. Hurme, M. Karjalainen, P. Karvonen, K. Kärkkäinen *et al.*, 1996 Do molecular markers reflect patterns of differentiation in adaptive traits of conifers? Theor. Appl. Genet. 93: 215–221. https://doi.org/10.1007/BF00225748

Kawecki, T. J., and D. Ebert, 2004 Conceptual issues in local adaptation. Ecol. Lett. 7: 1225–1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x

Keller, S. R., M. S. Olson, S. Salim, S. A. William, and T. Peter, 2010 Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*. Mol. Ecol. 19: 1212–1226. https://doi.org/10.1111/j.1365-294X.2010.04546.x

Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and Speciation. Genetics 173: 419–434. https://doi.org/10.1534/genetics.105.047985

Knoth, C., and T. Eulgem, 2008 The oomycete response gene *LURP1* is required for defense against *Hyaloperonospora parasitica* in *Arabidopsis thaliana*. Plant J. 55: 53–64. https://doi.org/10.1111/j.1365-313X.2008.03486.x

Komulainen, P., G. R. Brown, M. Mikkonen, A. Karhu, M. R. García-Gil *et al.*, 2003 Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*. Theor. Appl. Genet. 107: 667–678. https://doi.org/10.1007/s00122-003-1312-2

Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose, 2015 Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15: 1179–1191. https://doi.org/10.1111/1755-0998.12387

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo *et al.*, 2019 Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20: 117. https://doi.org/10.1186/s13059-019-1720-5

Kremer, A., and V. Le Corre, 2012 Decoupling of differentiation between traits and their underlying genes in response to divergent selection. Heredity 108: 375–385. https://doi.org/10.1038/hdy.2011.81

Kujala, S., T. Knürr, K. Kärkkäinen, D. B. Neale, M. J. Sillanpää *et al.*, 2017 Genetic heterogeneity underlying variation in a locally adaptive clinal trait in *Pinus sylvestris* revealed by a Bayesian multipopulation analysis. Heredity 118: 413–423. https://doi.org/10.1038/hdy.2016.115

Kujala, S., and O. Savolainen, 2012 Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): Signs of clinal adaptation? Tree Genet. Genomes 8: 1451–1467. https://doi.org/10.1007/s11295-012-0532-5

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Latta, R. G., 1998 Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. Am. Nat. 151: 283–292. https://doi.org/10.1086/286119

Latta, R. G., 2003 Gene flow, adaptive population divergence and comparative population structure across loci. New Phytol. 161: 51–58. https://doi.org/10.1046/j.1469-8137.2003.00920.x

Leinonen, P. H., D. L. Remington, and O. Savolainen, 2011 Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. Evolution 65: 90–107. https://doi.org/10.1111/j.1558-5646.2010.01119.x

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175–195.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Z., A. R. De La Torre, L. Sterck, F. M. Cánovas, C. Avila *et al.*, 2017 Single-copy genes as molecularmarkers for phylogenomic studies in seed plants. Genome Biol. Evol. 9: 1130–1147. https://doi.org/10.1093/gbe/evx070

Lind, B. M., M. Menon, C. E. Bolte, T. M. Faske, and A. J. Eckert, 2018 The genomics of local adaptation in trees: are we out of the woods yet? Tree Genet. Genomes. 14: 1–29. https://doi.org/10.1007/s11295-017-1224-y

Lotterhos, K. E., and M. C. Whitlock, 2015 The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol. Ecol. 24: 1031–1046. https://doi.org/10.1111/mec.13100

Lu, M., K. V. Krutovsky, C. D. Nelson, T. E. Koralewski, T. D. Byram *et al.*, 2016 Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics 17: 730. https://doi.org/10.1186/s12864-016-3081-8

Luu, K., E. Bazin, and M. G. B. Blum, 2017 pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. 17: 67–77. https://doi.org/10.1111/1755-0998.12592

Ma, X. F., D. Hall, K. R. St. Onge, S. Jansson, and P. K. Ingvarsson, 2010 Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the *Populus tremula* photoperiodic pathway. Genetics 186: 1033–1044. https://doi.org/10.1534/genetics.110.120873

McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686. https://doi.org/10.1371/journal.pgen.1000686

Mei, W., M. G. Stetter, D. J. Gates, M. C. Stitzer, and J. Ross-Ibarra, 2018 Adaptation in plant genomes: Bigger is different. Am. J. Bot. 105: 16–19. https://doi.org/10.1002/ajb2.1002

Mikola, J., 1982 Bud-set phenology as an indicator of climatic adaptation of Scots pine in Finland. Silva Fenn. 16: 178–184.

Mimura, M., and S. N. Aitken, 2007 Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. Heredity 99: 224–232. https://doi.org/10.1038/sj.hdy.6800987

Muratova, E. N., 1997 Cytogenetical study on Scots pine (Pinus sylvestris L.) in the Central Yakutia, Cytogenetic studies of forest trees and shrub species: Proc. First IUFRO Cytogenetics Working Part. 2–04.

Naydenov, K., S. Senneville, J. Beaulieu, F. Tremblay, and J. Bousquet, 2007 Glacial vicariance in Eurasia: Mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. BMC Evol. Biol. 7: 233. https://doi.org/10.1186/1471-2148-7-233

Neale, D. B., and O. Savolainen, 2004 Association genetics of complex traits in conifers. Trends Plant Sci. 9: 325–330. https://doi.org/10.1016/j.tplants.2004.05.006

Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu *et al.*, 2014 Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15: R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nei, M., K.-I. Kojima, and H. E. Schaffer, 1967 Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57: 741.

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

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3: e196. https://doi.org/10.1371/journal.pbio.0030196

Orr, H. A., 1998 The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution 52: 935–949. https://doi.org/10.1111/j.1558-5646.1998.tb01823.x

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

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

Puig, M., S. Casillas, S. Villatoro, and M. Cáceres, 2015 Human inversions and their functional consequences. Brief. Funct. Genomics 14: 369–379. https://doi.org/10.1093/bfgp/elv020

Pyhäjärvi, T., M. R. García-Gil, T. Knürr, M. Mikkonen, W. Wachowiak *et al.*, 2007b Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177: 1713–1724. https://doi.org/10.1534/genetics.107.077099

Pyhäjärvi, T., M. J. Salmela, and O. Savolainen, 2008 Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. Tree Genetics and Genomes 4: 247–254.

Pyhäjärvi, T., M. B. Hufford, S. Mezmouk, and J. Ross-Ibarra, 2013 Complex patterns of local adaptation in teosinte. Genome Biol. Evol. 5: 1594–1609. https://doi.org/10.1093/gbe/evt109

Pyhäjärvi, T., S. T. Kujala, and O. Savolainen, 2020 275 years of forestry meets genomics in *Pinus sylvestris*. Evol. Appl. 13: 11–30. https://doi.org/10.1111/eva.12809

Racimo, F., J. J. Berg, and J. K. Pickrell, 2018 Detecting polygenic adaptation in admixture graphs. Genetics 208: 1565–1584. https://doi.org/10.1534/genetics.117.300489

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes *et al.*, 2012 DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rockman, M. V., 2012 The QTN program and the alleles that matter for evolution: All that's gold does not glitter. Evolution. 66: 1–17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Savolainen, O., M. Lascoux, and J. Merilä, 2013 Ecological genomics of local adaptation. Nat. Rev. Genet. 14: 807–820. https://doi.org/10.1038/nrg3522

Savolainen, O., T. Pyhäjärvi, and T. Knürr, 2007 Gene flow and local adaptation in trees. Annu. Rev. Ecol. Evol. Syst. 38: 595–619. https://doi.org/10.1146/annurev.ecolsys.38.091206.095646

Schmidt, P. S., C.-T. Zhu, J. Das, M. Batavia, L. Yang *et al.*, 2008 An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 105: 16207–16211. https://doi.org/10.1073/pnas.0805485105

Simpson, S. D., K. Nakashima, Y. Narusaka, M. Seki, K. Shinozaki *et al.*, 2003 Two different novel cis-acting elements of *erd1*, a *clpA* homologous *Arabidopsis* gene function in induction by dehydration stress and dark-induced senescence. Plant J. 33: 259–270. https://doi.org/10.1046/j.1365-313X.2003.01624.x

Slatkin, M., 1973 Gene flow and selection in a cline. Genetics 75: 733–756.

Sundell, D., C. Mannapperuma, S. Netotea, N. Delhomme, Y. Lin *et al.*, 2015 The plant genome integrative explorer resource: PlantGenIE.org. New Phytol. 208: 1149–1156. https://doi.org/10.1111/nph.13557

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

Thompson, M. J., and C. D. Jiggins, 2014 Supergenes and their role in evolution. Heredity 113: 1–8. https://doi.org/10.1038/hdy.2014.20

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative enomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14: 178–192. https://doi.org/10.1093/bib/bbs017

Tyrmi, J. S., 2018 STAPLER: a simple tool for creating, managing and parallelizing common high-throughput sequencing workflows. bioRxiv. doi: 10.1101/445056 (Preprint posted October 16, 2018).https://doi.org/10.1101/445056

Vitalis, R., M. Gautier, K. J. Dawson, and M. A. Beaumont, 2014 Detecting and measuring selection from gene frequency data. Genetics 196: 799–817. https://doi.org/10.1534/genetics.113.152991

Wachowiak, W., P. A. Balk, and O. Savolainen, 2009 Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). Tree Genet. Genomes 5: 117–132. https://doi.org/10.1007/s11295-008-0188-3

Wang, J., J. Ding, B. Tan, K. M. Robinson, I. H. Michelson *et al.*, 2018 A major locus controls local adaptation and adaptive life history variation in a perennial plant. Genome Biol. 19: 72. https://doi.org/10.1186/s13059-018-1444-y

Wegrzyn, J. L., J. M. Lee, B. R. Tearse, and D. B. Neale, 2008 TreeGenes: A forest tree genome database. Int. J. Plant Genomics 2008: 1–7. https://doi.org/10.1155/2008/412875

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.

Wellenreuther, M., and L. Bernatchez, 2018 Eco-evolutionary genomics of chromosomal inversions. Trends Ecol. Evol. 33: 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Westbrook, J. W., L. G. Neves, M. Kirst, G. F. Peter, S. Chamala *et al.*, 2015 A consensus genetic map for *Pinus taeda* and *Pinus elliottii* and extent of linkage disequilibrium in two genotype-phenotype discovery populations of *Pinus taeda*. G3 (Bethesda) 5: 1685–1694. https://doi.org/10.1534/g3.115.019588

Wu, T. D., and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881. https://doi.org/10.1093/bioinformatics/btq057

Yeaman, S., 2013 Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proc. Natl. Acad. Sci. USA 110: E1743–E1751. https://doi.org/10.1073/pnas.1219381110

Yeaman, S., 2015 Local adaptation by alleles of small effect. Am. Nat. 186: S74–S89. https://doi.org/10.1086/682405

Yeaman, S., and M. C. Whitlock, 2011 The genetic architecture of adaptation under migration-selection balance. Evolution 65: 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau *et al.*, 2016 Convergent local adaptation to climate in distantly related conifers. Science 353: 1431–1433. https://doi.org/10.1126/science.aaf7812

Zonneveld, B. J. M., 2012 Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. Nord. J. Bot. 30: 490–502. https://doi.org/10.1111/j.1756-1051.2012.01516.x

Communicating editor: P. Ingvarsson