the plant journal
SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) **117**, 944–955 doi: 10.1111/tpj.16535

RESOURCE

Whole-genome resequencing facilitates the development of a 50K single nucleotide polymorphism genotyping array for Scots pine (*Pinus sylvestris* L.) and its transferability to other pine species

Maximiliano Estravis Barcala[1], **Tom van der Valk**[2,†,‡,§], **Zhiqiang Chen**[1], **Tomas Funda**[1], **Rajiv Chaudhary**[1], **Adam Klingberg**[1,3], **Irena Fundova**[1,¶], **Mari Suontama**[3], **Henrik Hallingbäck**[3], **Carolina Bernhardsson**[4,5], **Björn Nystedt**[2], **Pär K. Ingvarsson**[5], **Ellen Sherwood**[6,7], **Nathaniel Street**[8], **Ulf Gyllensten**[9], **Ove Nilsson**[1] and **Harry X. Wu**[1,*]

[1]*Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre (UPSC), Swedish University of Agricultural Sciences, Umeå, Sweden,*
[2]*Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden,*
[3]*Skogforsk, Sävar, Uppsala, Sweden,*
[4]*Department of Organismal Biology, Human Evolution, Uppsala University, Uppsala, Sweden,*
[5]*Department of Plant Biology, Linnean Centre for Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden,*
[6]*Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden,*
[7]*Department of Gene Technology, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden,*
[8]*Department of Plant Physiology, Umeå Plant Science Centre (UPSC), Umeå University, Umeå, Sweden,*
[9]*Department of Immunology, Genetics, and Pathology, Biomedical Center, Science for Life Laboratory, Uppsala University, Uppsala, Sweden*

Received 19 September 2023; revised 25 October 2023; accepted 26 October 2023; published online 10 November 2023.

*For correspondence (e-mail harry.wu@slu.se).
[†]Present address: Centre for Palaeogenetics, Stockholm, Sweden
[‡]Present address: Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden
[§]Present address: Science for Life Laboratory, Stockholm, Sweden
[¶]Present address: Department of Forest Genetics and Regeneration, Norwegian Institute for Bioeconomy Research (NIBIO), Ås, Norway

SUMMARY

Scots pine (*Pinus sylvestris* L.) is one of the most widespread and economically important conifer species in the world. Applications like genomic selection and association studies, which could help accelerate breeding cycles, are challenging in Scots pine because of its large and repetitive genome. For this reason, genotyping tools for conifer species, and in particular for Scots pine, are commonly based on transcribed regions of the genome. In this article, we present the Axiom PsyI50K array, the first single nucleotide polymorphism (SNP) genotyping array for Scots pine based on whole-genome resequencing, that represents both genic and intergenic regions. This array was designed following a two-step procedure: first, 192 trees were sequenced, and a 430K SNP screening array was constructed. Then, 480 samples, including haploid megagametophytes, full-sib family trios, breeding population, and range-wide individuals from across Eurasia were genotyped with the screening array. The best 50K SNPs were selected based on quality, replicability, distribution across the draft genome assembly, balance between genic and intergenic regions, and genotype–environment and genotype–phenotype associations. Of the final 49 877 probes tiled in the array, 20 372 (40.84%) occur inside gene models, while the rest lie in intergenic regions. We also show that the PsyI50K array can yield enough high-confidence SNPs for genetic studies in pine species from North America and Eurasia. This new genotyping tool will be a valuable resource for high-throughput fundamental and applied research of Scots pine and other pine species.

Keywords: genome resequencing, SNP array, *Pinus sylvestris*, pines, genomic selection, genome-wide association studies.

INTRODUCTION

Conifer species account for 60% of the worldwide industrial wood harvest, even though they represent just 17% of Earth's forest area (Cooper, 2003). However, this distribution is not homogenous. Large regions of the Boreal climate zone in North America and Eurasia consist almost exclusively of conifer forests, which are important for both ecological and economic purposes (Farjon, 2018). Moreover, boreal forests store approximately 22% of the world's carbon (Wu & Nilsson, 2023). It is thus of no surprise that the world's most advanced tree breeding programs are developed for conifer species (Isik & McKeand, 2019; McKeand et al., 2021; Wu et al., 2016). In the context of global warming and climate change, it is more important than ever to study the genetic basis of local adaptation and productivity. Molecular markers have been in use for several decades to assist selection and dissect the genetic nature of both adaptive and productive (commercial) traits. Traditionally, restriction fragment length polymorphisms, amplified fragment length polymorphism and simple sequence repeats, and now more commonly single nucleotide polymorphisms (SNPs) have been used for genotype–phenotype association studies in forest trees (Thavamanikumar et al., 2013). However, conifers are mostly outcrossing species, with large effective population sizes and a rapid decay of linkage disequilibrium (Neale & Savolainen, 2004). Therefore, the relatively low number of markers used in previous studies has had very limited success in the dissection of quantitative trait variation (Hall et al., 2016).

With the development of genomic selection (GS), there is potential to shorten breeding cycles, increase selection intensity, and improve the accuracy of breeding value estimates (Grattapaglia, 2022). However, this method requires a high number of reliable genome-wide markers, a target especially hard to meet in conifers, which have large and highly repetitive genomes (De La Torre et al., 2014; Niu et al., 2022). SNP genotyping arrays with tens of thousands of markers have been available in the last decade for broadleaf tree species (Geraldes et al., 2013; Silva-Junior et al., 2015), which have small genomes compared to conifers, and they have been used to genotype large numbers of individuals for GS.

Scots pine (*Pinus sylvestris* L.) is the most widespread Pinaceae species in the world, and the second foremost species for wood production in Sweden after Norway spruce (*Picea abies* (L.) Karst.). Breeding of Scots pine in Sweden started with the selection of plus trees in the 1940s and 1950s, and a further expansion of the breeding populations occurred in the 1980s (Haapanen et al., 2015). Until recently, conifer SNP arrays have been limited to a couple thousand SNPs, and are currently based on RNA sequencing, exome capture, or candidate gene sequencing (Chancerel et al., 2013; Graham et al., 2022; Howe et al., 2020; Kastally et al., 2022; Pavy et al., 2013; Perry et al., 2020; Plomion et al., 2016; reviewed in Ahmar et al., 2021; Grattapaglia, 2022). Markers in such SNP arrays thus only represent coding sequences, leaving out all the intergenic space, which constitutes the bigger part of the genome and is an integral part of population genetics studies. To our knowledge, only the recent Norway spruce Piab50K array (Bernhardsson et al., 2021) utilized whole-genome resequencing for its development, thus covering both genic and intergenic regions, using a similar approach to the one we took in the present study.

Recently, the PiSy50K genotyping array was developed for Scots pine based on transcriptome sequencing, exome capture sequencing, and candidate genes (Kastally et al., 2022). Similarly, a SNP array for the genotyping of four European pine species, among them Scots pine, was developed from resequenced candidate genes and transcriptomic sequences (Perry et al., 2020). In both cases, intergenic regions were not represented, and this important part of the genome is therefore rendered invisible for genotyping.

In this paper, we present the first high-throughput SNP genotyping array for Scots pine (and related species) developed in two steps: whole-genome resequencing and further screening and filtering by genotyping individuals from breeding populations and range-wide provenances. Importantly, both the genic and intergenic regions are represented in the array's probes and this resource will thus be useful for fundamental and applied studies in the fields of population genetics, evolution, GS, and breeding.

RESULTS

Screening of the 430K SNP array and selection of the final 50K array

A total of 478 samples (99.58%) passed the quality controls outlined in the Best Practice Analysis Workflow (see "Materials and Methods" section). For these, 430 735 SNPs were genotyped and further classified according to the SNPs' quality. Table 1 shows metrics for the full screening array and for each of the six categories produced by SNPolisher. PolyHighResolution (PHR) SNPs were the largest category, with 48.98% of all genotyped SNPs. For downstream analyses, Affymetrix recommends SNPs that fall into the PHR, NoMinorHom (NH), and MonoHighResolution (MHR) categories. However, PHR SNPs in our study show a higher

Table 1 SNP metrics for the six categories produced by the SNPolisher *ps-classification* function, plus the full screening array

	Number of SNPs	Average heterozygosity	Average MAF	Average missingness
Full screening array	430 735 (100%)	0.2274 (0–0.8849)	0.1769 (0–0.5)	0.0472 (0–0.9540)
PHR SNPs (Recommended)	210 972 (48.98%)	0.2629 (0–0.8473)	0.1993 (0.0021–0.5)	0.0065 (0–0.0293)
NH SNPs (Recommended)	26 372 (6.21%)	0.1087 (0.0021–0.4686)	0.0549 (0.001–0.2409)	0.0094 (0–0.0293)
MHR SNPs (Recommended)	14 756 (4.43%)	0 (–)	0 (–)	0.0009 (0–0.0293)
CRBT SNPs	55 109 (12.79%)	0.2808 (0–0.8849)	0.2317 (0–0.5)	0.0610 (0.0314–0.9540)
OTV SNPs	3256 (0.76%)	0.1650 (0.0021–0.8766)	0.1094 (0.001–0.5)	0.0245 (0–0.1925)
O SNPs	119 910 (27.83%)	0.1966 (0–0.8180)	0.1632 (0–0.5)	0.1274 (0–0.9540)

For the number of SNPs, the percentage is indicated in parentheses. For the other metrics, the mean is followed by the minimum and the maximum in parentheses. CRBT, CallRateBelowThreshold; MHR, MonoHighResolution; NH, NoMinorHom; O, Other; OTV, OffTargetVariant; PHR, PolyHighResolution; SNP, single nucleotide polymorphism.

heterozygosity and minor allele frequency (MAF) than their NH and MHR counterparts (and also higher than the overall array averages). For these reasons, together with their sufficient number, only PHRs were considered for downstream analyses towards the selection of the final 50K array.

The PHR SNPs were further filtered as summarized in Figure 3. Due to the high number of SNPs in the screening array, many of them were found to be tightly linked to one another, and so we kept only one SNP from each high linkage network (pairwise LD ≥0.8). Moreover, we eliminated SNPs that had different genotypes in any of the 10 pairs of replicates. Finally, we kept only those SNPs that were homozygous in all 30 megagametophytes. These steps led us to a subset of ~76K pre-selected SNPs.

To reach the final selection of 50K SNPs, we kept all the pre-selected SNPs, which were inside of gene models according to the latest *P. sylvestris* draft genome annotation (in preparation). On top of that, we kept SNPs that we found to be associated with phenotypic traits or geographical provenance of the samples via genome-wide association studies (GWAS) analyses, as we believed that these SNPs would be of particular interest to breeders. These traits were (with the number of SNPs associated in parentheses): branch quality (5), diameter (7), height (12), knot number (25), vitality (17), stem volume (9), latitude (19), longitude (55), and altitude (16). Finally, random sampling was performed in every contig with pre-selected SNPs.

The sequences of the final 50K probes were submitted to Thermo Fisher for tiling. Out of the 50 000

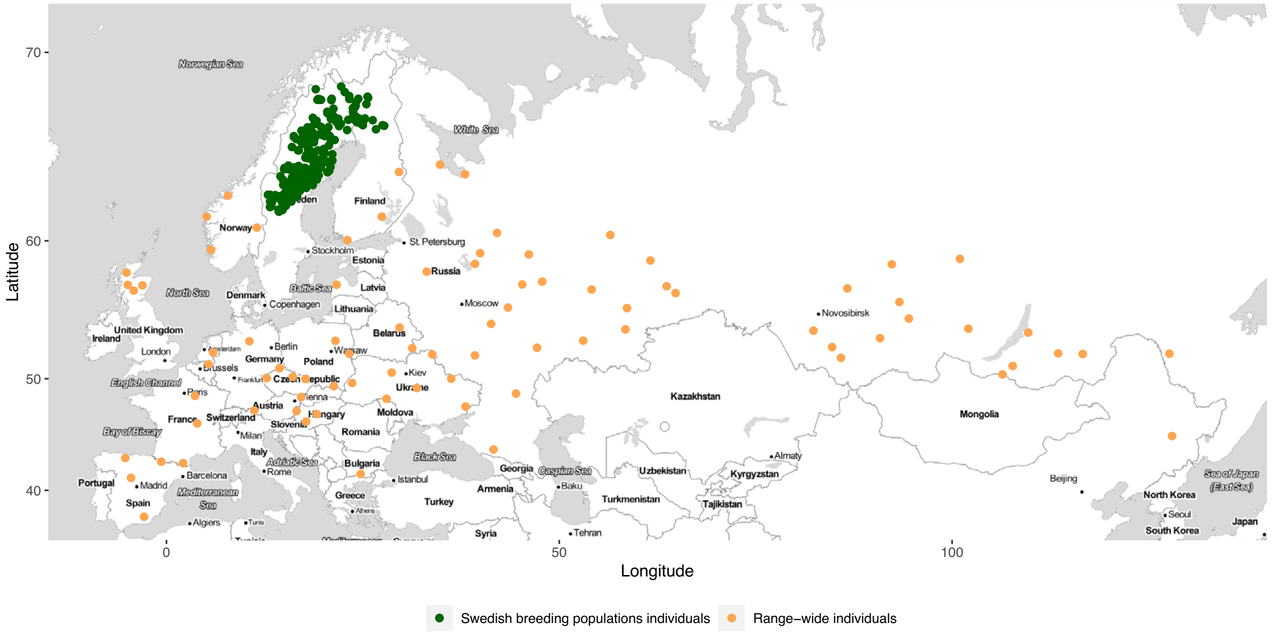

Figure 1. Geographic location of the 385 breeding-population (301; green) and range-wide (84; orange) Scots pine individuals genotyped using the screening array.

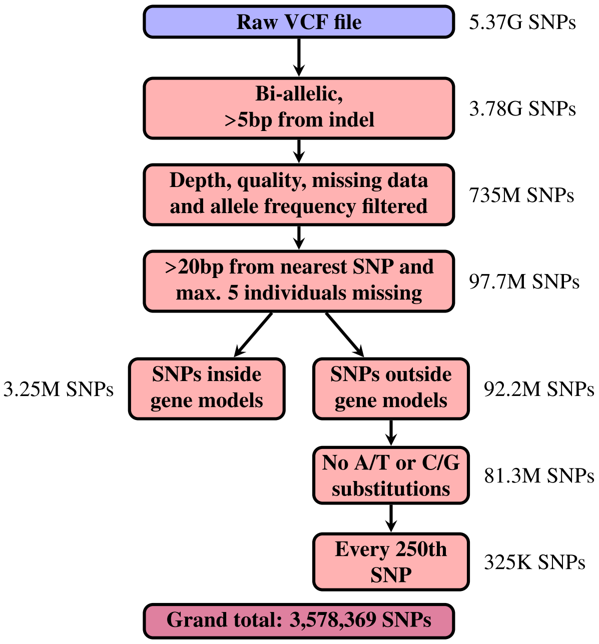

Figure 2. Schematic workflow of the probe selection pipeline from resequencing and single nucleotide polymorphism calling to the *in silico* testing in Thermo Fisher for the screening array construction.

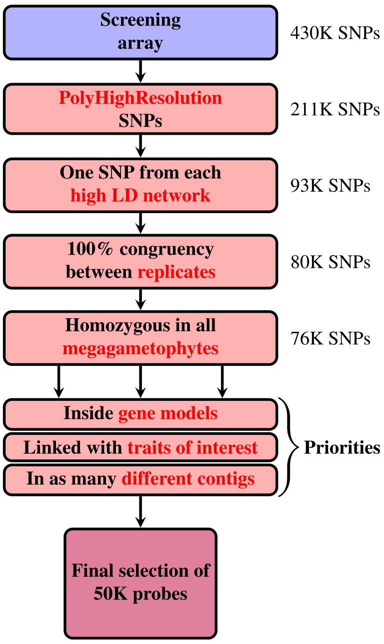

Figure 3. Schematic workflow of the probe selection pipeline from the 430K screening array to the final 50K selected probes.

probes, only 123 could not be tiled on the array, leaving us with the final number of 49 877 genome-wide SNPs. Of these, 20 372 (40.84%) are inside of 9002 gene models (the rest are in intergenic regions), and 156 are associated with the traits of interest (44 of them inside gene models, the remaining 112 intergenic). Of the 9002 gene models, 6147 (68.28%) are shared with those present in the PiSy50K array (Kastally et al., 2022), while 2855 (31.72%) are unique to our new array Psyl50K (Figure 4).

A total of 3214 contigs of the draft assembly are represented in the final array, out of the 3318 contigs that had at least one PHR SNP (96.87%). This shows that the filtering and selection of the final SNPs did not affect the distribution of the SNPs across the draft genome assembly. These 3214 contigs represent 29.14% of all contigs (11028) of the draft assembly; however, short contigs are underrepresented compared to the whole assembly (Figure S4). We also mapped all SNPs to the Scots pine chromosome-scale assembly (in preparation). Psyl50K probes are evenly distributed across this end-to-end chromosome assembly (Figure 5). Moreover, 97.59% of Psyl50K's probes align to the assembly with more than 90% identity and 98.5% probe coverage. This represents an improvement in sequence fidelity compared to the other available SNP array for Scots pine (Kastally et al., 2022), in which 86.33% of the probes align to the genome assembly with the same parameters.

Evaluation of the 50K SNP array

A frequent problem in genome assembly is the collapse of repetitive regions, which can lead to stacked read mappings and thus spurious SNP calls. This is particularly worrying in the case of large and highly repetitive genomes such as that of *P. sylvestris*. This phenomenon leads to SNPs having either too high or too low heterozygosity relative to their MAF, as can be observed with the screening array at large (Figure 6, gray points). In contrast, the great majority of PHR SNPs (and the final selected SNPs, which are a subset of PHRs) have the expected pattern of heterozygosity under Hardy–Weinberg equilibrium (Figure 6, dark red and bright red points).

Moreover, the selected SNPs span the entire range of MAFs without any indication of ascertainment bias (Figure 7). The underrepresentation of extremely rare alleles (MAF < 0.02) in the final selected set is due to an intentional filtering of rare alleles when selecting PHR SNPs (orange bars in Figure 7) as these rare alleles most likely arise from low quality and/or spurious SNP calls. The MAF distribution pattern of genic and intergenic SNPs of the final array (one of the novel features of this tool) is shown in Figure S5.

Our 50K array was able to capture a longitudinal cline among the genotyped range-wide samples (Figure 1), from the Iberian Peninsula in the west (longitude 5° W) to far East Siberia and Northeastern China (longitude 128° E; Figure 8A). This cline was less sharply defined when considering all 211K PHR SNPs (Figure 8B), which was

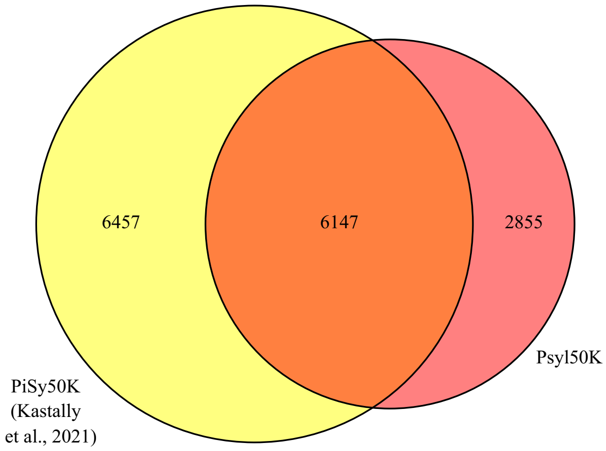

Figure 4. Venn diagram of Scots pine predicted gene models containing at least one single nucleotide polymorphism from the PiSy50K (Kastally et al., 2022) and Psyl50K (current study) arrays.

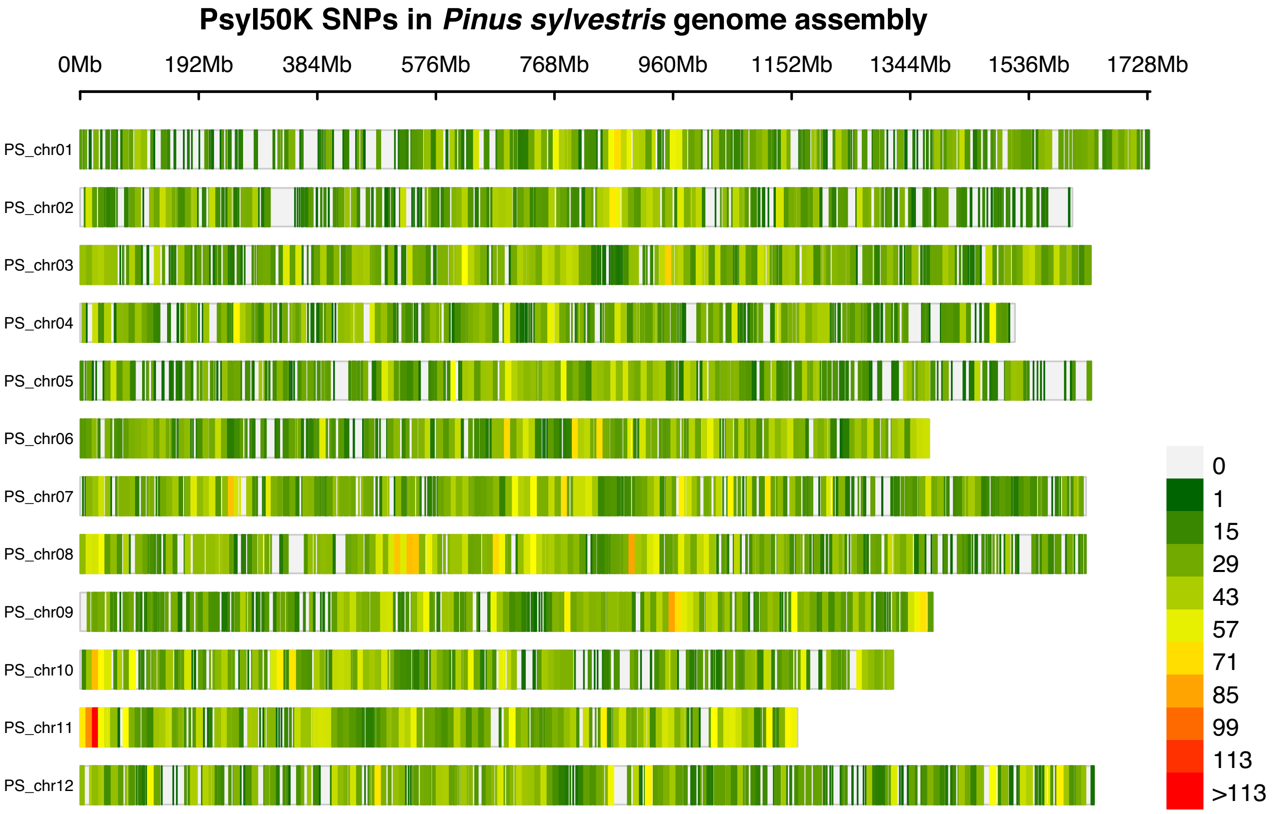

Figure 5. Probe density of the Psyl50K array in 10 Mb windows for each of the 12 chromosomes of the latest Scots pine genome assembly.

expected considering the SNP quality filtering and selection explained in the previous section. This result will be most useful to help study the evolutionary and demographic processes that this widely distributed species has experienced throughout its history.

Family relations were also perfectly captured by the 50K array, as shown in Figure 8C. Each of the six half-sib families (seven to eight offspring each) forms a tight cluster separated from the rest of the families. These clusters are also clearly defined when using all the PHR SNPs (Figure 8D).

Results: Transferability to other pine species

Individuals from six different pine species (*Pinus* sp.) were genotyped with the new Psyl50K array. By slightly

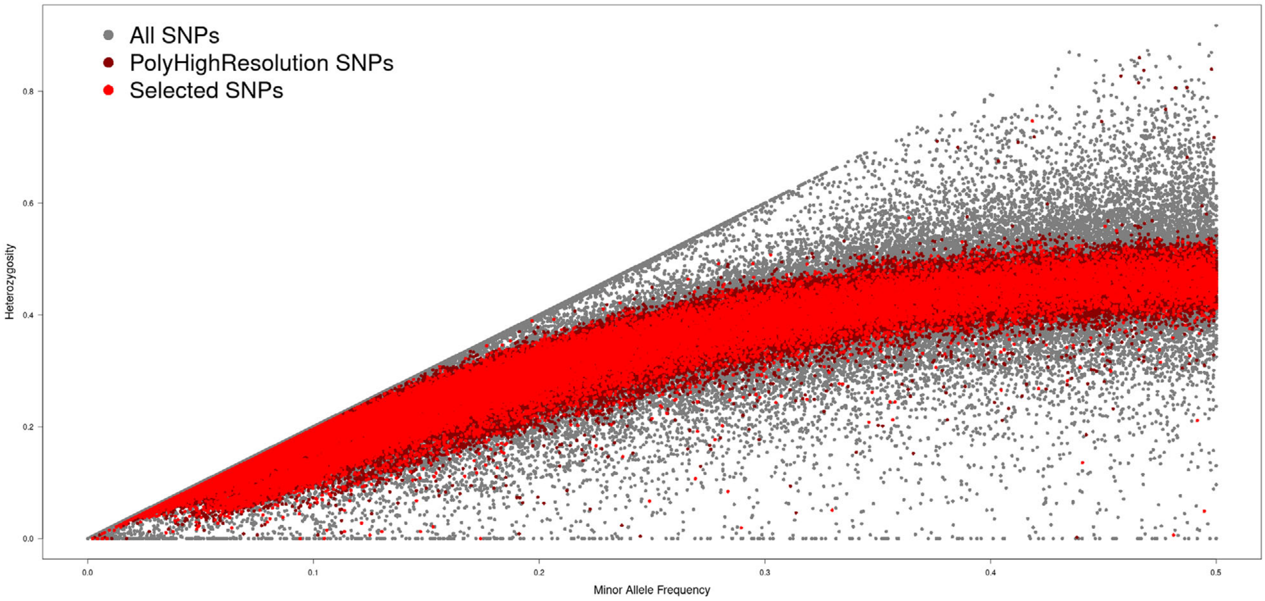

Figure 6. Scatter plot of the heterozygosity and the minor allele frequency of all the screening array single nucleotide polymorphisms (SNPs, 430K; gray), PHR SNPs (211K; dark red), and the final set of selected SNPs (50K; bright red).

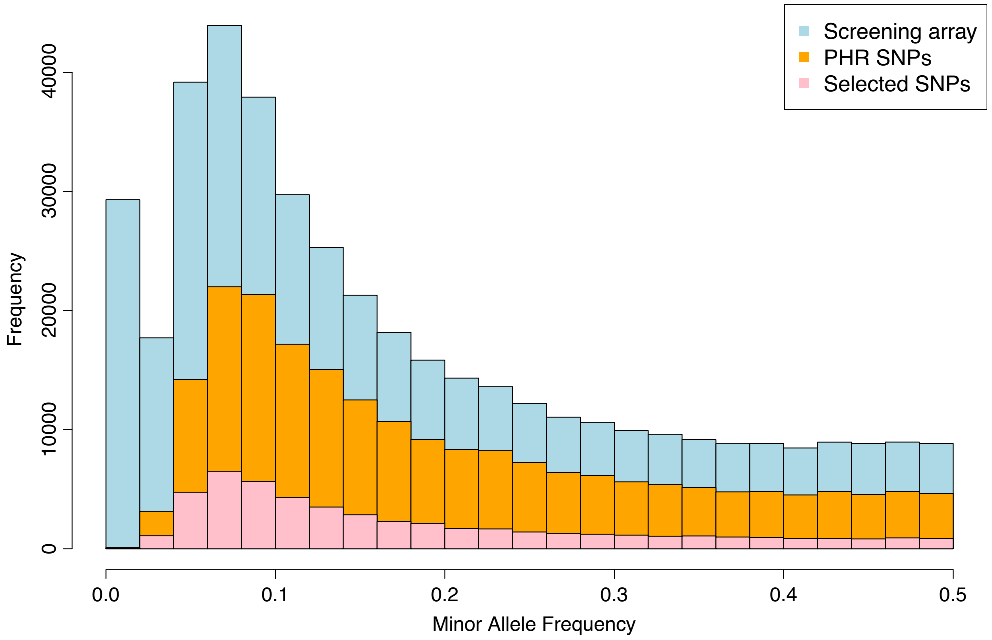

Figure 7. Distribution of minor allele frequency among all the screening array single nucleotide polymorphisms (SNPs, 430K; light blue), PolyHighResolution (PHR) SNPs (211K; orange), and the final selected SNPs (50K; pink).

lowering the quality thresholds (see "Materials and Methods" section), 93.94% of the samples belonging to the subgenus *Pinus* (*P. banksiana*, *P. contorta*, *P. nigra*, and *P. tabuliformis*) passed the quality control (124 out of 132 samples). More specifically, it was 98.31% (58 out of 59) of the samples from the section *Pinus* (to which *P. sylvestris* belongs) and 90.41% (66 out of 73) of the samples from the section *Trifoliae*. Moreover, a total of 10 841 SNPs were labeled as recommended in these four species by the Axiom Best Practices pipeline. With these quality thresholds, there was no *P. strobus* or *P. cembra* sample to pass the quality control. As previously explained, these two species belong to the subgenus *Strobus* and are thus relatively distant from *P. sylvestris*. These results show that

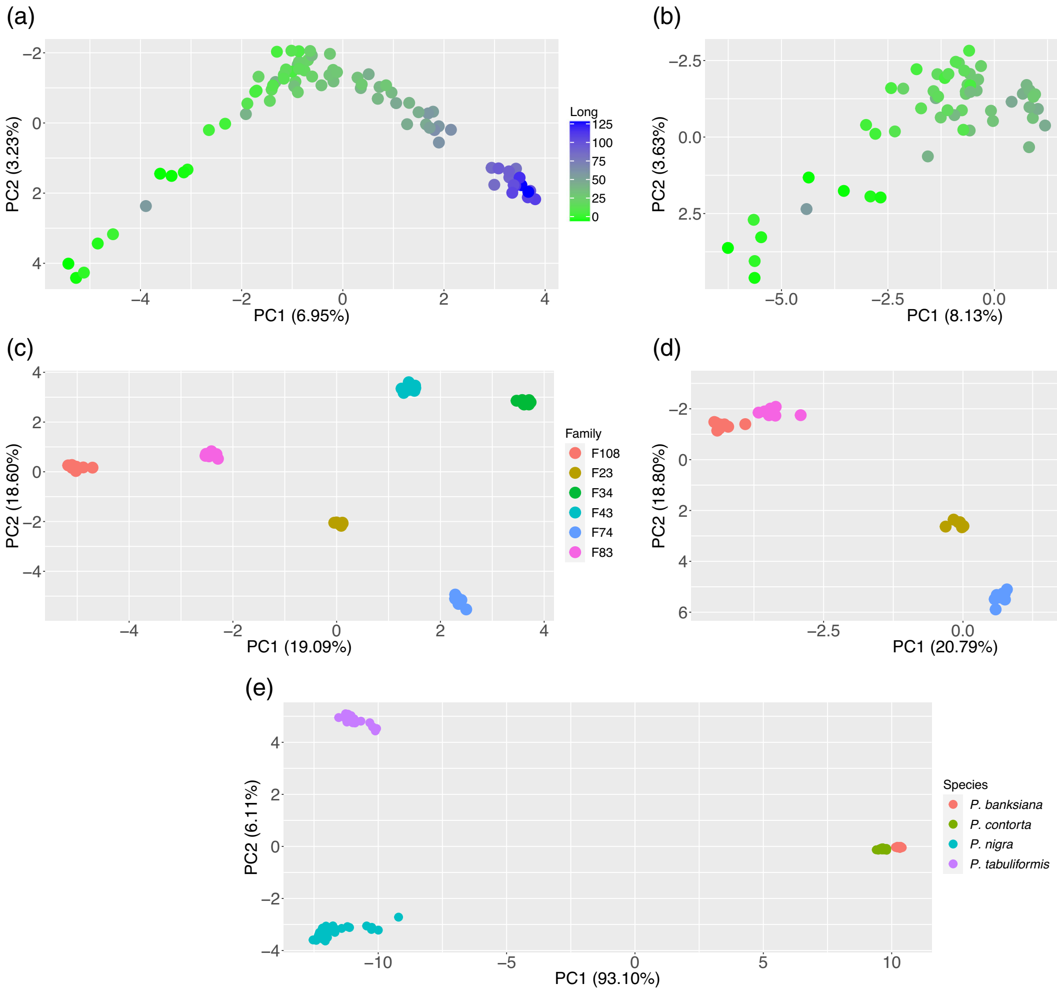

Figure 8. Evaluation of the Psyl50K array in *Pinus sylvestris* and its transferability to other pine species.
(a, b) Population structure of 84 range-wide *P. sylvestris* samples estimated using a principal component analysis (PCA) on the relatedness matrix calculated based on (a) the final 50K single nucleotide polymorphisms (SNPs) and (b) 211K PolyHighResolution (PHR) SNPs. The color scale represents the longitude of origin of the samples, from west (green) to east (blue) across Eurasia.
(c, d) Family structure estimated using a PCA on the relatedness matrix calculated based on (c) the final 50K SNPs and (d) 211K PHR SNPs. Each family consists of 7–8 *P. sylvestris* full-siblings.
(e) *Pinus* spp. clustering estimated using a PCA on the relatedness matrix calculated based on 10 841 recommended SNPs.

our new 50K SNP array loses sensitivity as species get further from *P. sylvestris*, the species for which the array was designed, which is expected from an evolutionary perspective.

The four species belonging to the subgenus *Pinus* are separated into the corresponding two sections (*Pinus* and *Trifoliae*) with PC1, and further into each separate species (PC2 and PC3; Figure 8E; Figure S6). Analyzing the SNPs recommended for the two species in the cluster closest to *P. sylvestris* (section *Pinus*, species *P. nigra*, and *P. tabuliformis*), the two subspecies of *P. nigra* (Eckenwalder, 2009) can be distinguished: *P. nigra* subsp. *nigra* in the Balkan region, and *P. nigra* subsp. *salzmannii* in the western Mediterranean (Figure S7). Moreover, by analyzing *P. contorta* separately (17 007 recommended SNPs), the samples belonging to *P. contorta* subsp. *contorta* are separated from those belonging to *P. contorta* subsp. *latifolia* and *P. contorta* subsp. *murrayana* (Figure S8).

We were able to analyze samples from the subgenus *Strobus* (*P. cembra* and *P. strobus*) separately by lowering the DishQC threshold to 0.27. This way, a total of 9207 SNPs were recommended, and the two species could be clearly separated (Figure S9).

For both sets of the recommended SNPs, roughly half of them are localized within gene models whilst the other half lie in intergenic regions. For the subgenus *Pinus*, 5800 (53.50%) SNPs are genic and 5041 (46.50%) intergenic; for the subgenus *Strobus*, 4511 (49%) SNPs are genic and 4696 (51%) intergenic. For both subgenera, the MAF pattern in recommended genic and intergenic SNPs (Figure S10) is similar to the one for *P. sylvestris* (Figure S5). Thus, our new SNP array can be used in these species for the same types of studies as in *P. sylvestris*, taking advantage of both genic and intergenic SNPs.

DISCUSSION

The size and complexity of conifer genomes pose challenges for the dissection of the genetic basis of trait variation, local adaptation, and breeding potential. A relatively low number of markers translates into quantitative trait loci (QTLs) mapped in large genome areas and are thus of little use for GS in forest breeding (Grattapaglia et al., 2018).

Since the advent of SNP genotyping arrays with several thousand markers, high-throughput genotyping became achievable at a relatively low cost for a large number of samples. In forest tree species, SNP arrays have been developed in the last decade for economically important genera like *Populus*, *Eucalyptus*, and several conifers (Ahmar et al., 2021; Grattapaglia, 2022). However, most recently developed SNP arrays in conifer species were exclusively based on transcriptomic data (RNA-seq, exome capture sequencing, and candidate genes), thus representing only coding, genic regions (Chancerel et al., 2013; Graham et al., 2022; Howe et al., 2020; Kastally et al., 2022; Pavy et al., 2013; Perry et al., 2020; Plomion et al., 2016). A few studies have used genome-wide reduced-representation sequencing as a complement to transcriptomic sources for conifer SNP array development (Caballero et al., 2021; Jackson et al., 2022). A recently published SNP array for Norway spruce uses the same two-step procedure that we apply in this article: first, a screening array is developed from SNP calling based on whole-genome resequencing, and then the best SNPs are selected based on genotyping of a heterogeneous set of samples including breeding populations, range-wide individuals, family trios and megagametophytes, apart from biological replicates (Bernhardsson et al., 2021). Similarly, in this paper, we developed a SNP array for *P. sylvestris* from whole-genome resequencing where both the genic and intergenic regions are represented. The SNPs present in the final array have been filtered based on quality, replicability, association with important phenotypic traits, and distribution across the draft genome assembly.

An important advantage of SNP arrays compared to other sequencing-based genotyping methods like genotyping by sequencing (GBS) or whole-genome sequencing (WGS) is that very little bioinformatic preparatory processing is required from the end-user once the array is developed and available. Contrary to GBS, for example, SNP array data analysis does not necessarily require imputation (Hussain et al., 2017). For SNP arrays, the bulk of the analyses are performed when designing the array, which is an end-product in itself. It is thus straightforward for a breeder or forest industry researcher to analyze samples genotyped on an SNP array, compared to the skills necessary to perform cleaning, SNP calling, and interpretation of a GBS or WGS analysis. Besides, for species with large genomes that lack a chromosome-scale assembly, which is the case of most conifers, the risk of having false variant discovery is increased in mapping-based methods like GBS. This is due to the mapping of reads coming from different genome locations to the same, collapsed region in the assembly (Bernhardsson et al., 2020).

In this paper, we report the development of the first whole-genome resequencing-based SNP array for a pine species (*Pinus* sp.), the Axiom Psyl50K array. The high quality of the SNP calling in the 430K screening array allowed us to only use PHR (the highest quality category according to Affymetrix) as a first step towards the selection of the final 50K probes (Table 1; Figure 3). PHR SNPs on their own were enough to capture the longitudinal cline of range-wide samples (Figure 8a,b) and the family structure of the family trios (full-sibling families; Figure 8c,d). Also, PHR SNPs have a balanced heterozygosity and MAF spectrum (Figure 6).

After deciding to only use PHR SNPs for further filtering, we focused on two technical issues: congruency between replicates, and false heterozygous calls (Figure 3). Approximately 86% of the PHR SNPs had the same call for every pair of replicates (10 pairs in total), and these were kept. Of these, 95% were homozygous for all 30 megagametophytes, and the rest were eliminated from the filtering pipeline. With these two steps, we made sure to keep probes with a lower probability of having wrong calls due to plate processing differences or location of the samples inside the plate, among other possible causes. The megagametophyte samples were specially included as a technical control, since they are a haploid tissue, and thus any observed heterozygosity would be due to a genotyping error or a lower SNP calling sensitivity. This way, we arrived at approximately 76K pre-selected SNPs.

Since the final array can only accommodate 50K probes, the selection of the final subset of SNPs from the pre-selected ones consisted of making sure to lose no gene model and contig from the draft assembly containing pre-selected SNPs. Importantly, we retained every SNP, both genic and intergenic, that was found to be associated with traits of interest via GWAS, both using phenotypic (such as height, diameter, or stem volume) and geographic (latitude, longitude, and altitude) data.

The final 50K SNPs span the same gene models (a total of 9002), and 96.87% of the contigs with PHR SNPs, showing that the selection process interfered neither with the genic-to-intergenic balance nor with the genome-wide distribution of the SNPs, while at the same time filtering for high quality and prioritizing regions of interest. By mapping the probes to the latest genome draft assembly, we could assess that the SNPs were uniformly distributed across the chromosomes (Figure 5).

The balance between genic and intergenic regions is an important feature of the PsyI50K array, which distinguishes it from other SNP arrays developed for pine species. Furthermore, to our knowledge, there is only one other conifer SNP array spanning both types of regions, the one developed for Norway spruce (Bernhardsson et al., 2021). Selection pressure might be different in genic and intergenic regions, giving rise to contrasting variation patterns and possibly effects on phenotypic expression (Li et al., 2016; Schierding et al., 2016). Thus, it is important, from the evolutionary as well as breeding perspectives, to genotype both types of genomic regions, giving a clearer picture of the genome-wide variation landscape and facilitating the study of conserved evolution patterns in this important group of conifer trees.

In this paper, we also showed that the PsyI50K SNP array is useful for genotyping species other than *P. sylvestris*. Given the large number of markers and their genome-wide nature, it is expected that more distant species will have fewer of the *P. sylvestris* SNPs present or confidently called. *Pinus* is a large genus with more than 100 species (Jin et al., 2021), and in our study, we showed that more than 9K SNPs were confidently called and recommended by Axiom Best Practices not only for species from the subgenus *Pinus* which *P. sylvestris* belongs but also for the more distant subgenus *Strobus*. This number of SNPs has been shown to be sufficient for the estimation of predictive abilities (Grattapaglia, 2022), and, in our study, it was powerful enough to clearly separate species (Figure 8E; Figures S6 and S9) and even subspecies in the case of *P. nigra* (Figure S7) and *P. contorta* (Figure S8). We thus propose that the PsyI50K could be satisfactorily used for the genotyping of *Pinus* species, belonging to both the subgenus *Pinus* (about 70 species; sections *Pinus* and *Trifoliae* with 25 and 45 species, respectively) and *Strobus* (about 45 species). This versatility highlights the importance of PsyI50K as the only SNP array up to the present day for any pine species to represent both the genic and intergenic regions.

In summary, the PsyI50K SNP array presented in this article is a comprehensive, high-throughput genotyping tool for *P. sylvestris* and other pine species. Contrary to other arrays previously developed for pines, this one has been designed using genome-wide sequencing, which makes it the only array up to the present day for any pine species that includes both the genic and intergenic regions. We thus anticipate that PsyI50K will be useful for studies in diverse areas of both fundamental and applied research such as population genomics, GS, and linkage mapping in *P. sylvestris* and other economically and ecologically important species.

MATERIALS AND METHODS

Plant materials and DNA extraction

The first step towards constructing the 50K SNP array was to perform whole-genome resequencing on 192 Scots pine individuals, of which 172 were selected from three Swedish breeding populations (southern, central, and northern Sweden) and 20 represented natural provenances. The complete list of samples with their geographic origin is shown in Table S1 and Figure S1.

In the second step, a total of 480 individuals were genotyped with the 430K SNP screening array constructed from whole-genome resequencing. Of these, 301 trees originated from central and northern Swedish breeding populations and 74 were collected from a range-wide provenance trial in Arboretum Sofronka, Czechia (49°47′30″ N 13°23′18″ E) (Kaňák, 2016). Additional 10 trees representing native populations of Czechia (3), Finland (2), Scotland (2), and Spain (3) were included. The geographic origin of these 385 samples is shown in Figure 1 and detailed in Table S2.

The remaining 95 samples genotyped with the screening array consisted of 55 samples forming six family trios (mother, father, and seven to eight full-siblings each; five of the parents were already included as the Swedish breeding-population individuals described above) collected from a clone archive in Sävar and a progeny test in Vännäsby (both managed by Skogforsk, the Forestry Research Institute of Sweden); 30 haploid megagametophytes (see harvesting procedure below); and 10 replicated samples (6 and 4 from the range-wide and breeding-population sets, respectively).

To evaluate the array's transferability to other pine species, unrelated samples from each of jack pine (*Pinus banksiana* Lamb.), Swiss pine (*P. cembra* L.), lodgepole pine (*P. contorta* Dougl.), black pine (*P. nigra* Arnold), and eastern white pine (*P. strobus* L.) were collected from a genus-wide experiment with pine species in Arboretum Sofronka (Kaňák, 2016). Range-wide samples of southern Chinese pine (*P. tabuliformis* Carr.) were collected by the College of Biological Sciences and Technology, Beijing Forestry University, China. These six species were chosen to represent the different subdivisions of the genus *Pinus* (Jin et al., 2021). *P. strobus* and *P. cembra* belong to the subgenus *Strobus*, whereas the other four species belong to the subgenus *Pinus*, where *P. sylvestris* also belongs. Within the subgenus *Pinus*, *P. contorta* and *P. banksiana* belong to the section *Trifoliae*, whereas *P. nigra* and *P. tabuliformis* belong to the section *Pinus*, together with *P. sylvestris* itself. In summary, of the species genotyped in this study, *P. nigra* and *P. tabuliformis* are the closest to *P. sylvestris*, followed by *P. contorta* and *P. banksiana*, and finally *P. cembra* and *P. strobus* are the most distant. The geographic origin of these samples is shown in Figures S2 (Eurasia) and S3 (North America) and detailed in Table S3. The total number of individuals involved in the array's transferability evaluation was 192 (24 for *P. tabuliformis*, 30 each for *P. contorta*, *P. strobus*, and *P. cembra*, 35 for *P. nigra*, and 43 for *P. banksiana*).

To harvest haploid megagametophytes, seeds from mother Y3088 (from the Skogforsk Sävar collection) were soaked in 1% H_2O_2 for 24 h, washed repeatedly under MiliQ water and placed on top of moistened filter paper in a Petri dish at room temperature (20°C) for germination. When the embryos reached 1 cm in length, each megagametophyte was separated from the seed coat and the embryo using a sterile razor blade and was manually ground in liquid nitrogen using plastic pestles and 1.5-ml tubes.

All DNA extractions were carried out with the E.Z.N.A. SP Plant DNA Kit (Omega Bio-Tek, Norcross, GA, USA) following the manufacturer's instructions. Extracted DNA concentration was measured with NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA) before samples were sent to Thermo Fisher Scientific's Microarray Research Services Laboratory (Santa Clara, CA, USA) for genotyping, or to SciLifeLab (Stockholm, Sweden) for 10× whole-genome resequencing. For library preparation, the

Illumina TruSeq DNA PCR-free library preparation kit was used (Illumina, San Diego, CA, USA). Libraries were sequenced on the NovaSeq 6000 Sequencing System using the NovaSeq 6000 v1.0 reagents (Illumina, San Diego, CA, USA).

Resequencing read mapping and SNP calling

The 192 whole-genome resequenced samples were used to find and extract candidate genome sequences for probe design of the screening array. The raw sequencing reads were mapped against the draft genome assembly of Scots pine (in preparation) using bwa mem version 0.7.17 (Li & Durbin, 2009), with default parameters and marking shorter split hits as secondary. Reads around indels were realigned using GATK v3.7 IndelRealigner (Van der Auwera & O'Connor, 2020), and subsequently filtered for mapping quality of ≥30 using samtools version 1.17 (Li et al., 2009). Optical duplicates were removed using Picard version 2.10.3 MarkDuplicates (broadinstitute.github.io/picard). After mapping and filtering, the average genome coverage across all samples was 7.5× (range 5.3×–10×). SNPs were then jointly called for all samples using bcftools version 1.13 (Danecek et al., 2021) on default parameters (*bcftools mpileup* followed by *bcftools call*).

Construction of the pilot screening array

SNPs from the whole-genome resequenced samples were called separately for each contig in the draft assembly and then combined into a single VCF file, which contained 5.37 billion SNPs. A schematic workflow of SNP filtering is shown in Figure 2. First, only bi-allelic SNPs were kept and all SNPs within 5 bp from an indel were removed. Since the Scots pine genome is highly repetitive, only SNPs in regions with a depth of at least one-third and a maximum of three times the average depth across all individuals and a genotype quality score greater than 15 were included. Additionally, only SNPs with an alternative allele frequency between 0.05 and 0.95 and with a maximum of 25% missing data across all samples were kept at this filtering step. Then, 71-mer probe sequences were extracted for SNPs more than 20 bp away from the nearest neighbor SNP and where a maximum of five individuals showed missing data. Finally, all SNPs positioned within gene models were kept, while SNPs outside of gene models (i.e., intergenic SNPs) were filtered for not being A/T or C/G substitutions, as these require twice the number of probes per SNP in comparison to other SNP substitutions. Remaining intergenic SNPs were down-sampled so that every 250th SNP was kept. When ranking the proposed markers, all intragenic ones were considered as 'important' while all of their intergenic counterparts were assigned a 'standard' importance. This resulted in a total of 3 578 369 SNPs, which were sent to Thermo Fisher Bioinformatics Service (Santa Clara, CA, USA) for *in silico* testing. For quality control of the array, 8000 36-mer probe sequences (so-called DishQC sequences, following Thermo Fisher guidelines) were extracted from monomorphic regions (based on the unfiltered raw VCF file for all samples) of a hard-masked version of the draft genome assembly. These DishQC sequences were evenly distributed between the two strands (+/−) and between A/T and C/G sites as the probe's ligation position. In total, 2000 of these DishQCs were incorporated into the screening array for technical control. The final 430K probes for the screening array were selected by Thermo Fisher based on *in silico* tests of probe success probability.

Screening array SNP calling and filtering

After the raw CEL files were received from Thermo Fisher following genotyping with the screening array, Best Practice Analysis Workflow recommended for all Axiom Genotyping Arrays was carried out according to the manufacturer's instructions (Affymetrix, 2022) and with default parameters for *P. sylvestris* samples (DishQC threshold higher than 0.82, and SNP call rate higher than 0.97). In the case of samples from other *Pinus* species, the quality thresholds were slightly lowered in order to be able to call genotypes (DishQC higher than 0.70 and SNP call rate higher than 0.85).

All passing samples were called using the *apt-genotype-axiom* function from the Affymetrix Analysis Power Tools suite, which uses the BRLMM-P algorithm for SNP calling (Affymetrix, 2022). The corresponding genotypes were classified using the SNPolisher R package (Affymetrix, 2015), specifically the *ps-metrics* and *ps-classification* functions. Further filtering and selection of SNPs were performed with in-house R and Python scripts, together with vcftools version 0.1.17 (Danecek et al., 2011). SNP probes were mapped to the latest Scots pine genome assembly (in preparation) using gmap (Wu & Watanabe, 2005) with default parameters.

GWAS for SNP prioritization

To select SNPs to be tiled into the 50K genotyping array, three different sets of individuals were phenotyped. In all cases, the genotypes for 430K SNPs in the screening array for these individuals were known. A separate GWAS analysis was performed for each set. The first set consisted of 90 unrelated individuals from a Northern Swedish progeny trial (63.42° N, 16.67° E) which were part of the 192 trees that had been resequenced for the construction of the screening array (see above). For these 90 individuals, the following traits were measured at field tree age 10 (Supplementary Methods): height, knot number, branch quality, and vitality. Spatial-adjusted phenotypic value for those traits were used for GWAS. The second set consisted of 201 unrelated plus-tree individuals from northern Swedish breeding populations (originally from northern Sweden and northern Finland) which were part of the 480 trees used for the evaluation of the screening array. For these 201 individuals, breeding values of height, diameter at breast height, vitality, and stem volume were estimated based on 177 existing Scots pine field trials (Supplementary Methods). For these two sets, the latitude, longitude, and altitude (meters above sea level) of each tree's origin were also known. Finally, the third set consisted of 84 range-wide samples from the 480 trees used for the evaluation of the screening array (see above). For these individuals, the latitude and longitude of origin were known.

Once we had all the genotype and phenotype information for each set, an association test with a univariate linear mixed model was performed for each trait using GEMMA (Genome-wide Efficient Mixed-Model Association; Zhou & Stephens, 2012). Markers, which passed the *P*-value threshold of 10^{-5}, were prioritized, after the pre-selection based on SNP quality, towards the final selection of the 50K SNPs.

Population structure

To estimate and visualize the population structure of the genotyped *P. sylvestris* samples (and the species and subspecies discrimination for the other *Pinus* species), additive relationship matrices were constructed using the R *rrBLUP* package (Endelman, 2011) for each set of SNPs and samples. Then, a scaled and centered PCA was performed with the *prcomp* function in R.

AUTHOR CONTRIBUTIONS

MEB conducted all probe selection analyses for the final SNP array and its evaluation, population genetics, genotype–phenotype association, and transferability to other species. TvdV and CB conducted resequencing analyses and probe selection for the screening array. HXW, ZC, TF, HH, and MS designed sampling strategy for resequencing, screening array, and validation populations and carried out tree sampling. RC, TF, AK, IF, and MEB performed sample preparation and DNA extraction. ES prepared the resequencing libraries and led the sequencing facility tasks. HXW, BN, PKI, NS, UG, and ON designed the project and advised on re-sequencing strategy and analytical protocols. MEB and HW wrote the manuscript draft. All authors read and agreed to the last version of the manuscript.

ACKNOWLEDGEMENTS

We thank Jan Kaňák (Arboretum Sofronka, Plzeň, Czechia) for help with collecting range-wide samples of Scots pine, jack pine, lodgepole pine, eastern white pine, black pine and Swiss pine, Petr Novotný and his colleagues (Forestry and Game Management Research Institute, Jíloviště-Strnady, Czechia) for providing native Czech samples, Joan Cottrell (Forest Research, Northern Research Station, Roslin, UK) and Stephen Cavers (UK Centre for Ecology & Hydrology, Edinburgh, UK) for providing samples from Scottish populations, Ricardo Alia and Delphine Grivet (Forest Research Center, INIA-CSIC, Madrid, Spain) for providing Spanish samples, and Xiao-Ru Wang (Umeå University, Sweden) for providing northern Scandinavian samples for the screening array. Shihui Niu (Beijing Forestry University, China) provided *Pinus tabuliformis* samples. We also want to thank the Conifer SNP Consortium for their work with Thermo Fisher in developing Axiom SNP arrays for conifers. Computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Uppmax partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973 (SNIC projects: 2020/5-316, 2021/5-312, 2021/2-15, 2021/5-99, 2022/2-7, 2022/5-68, 2022/5-342; NAISS projects: 2023/2-17, 2023/5-89, 2023/5-323). The authors acknowledge support from the National Genomics Infrastructure (NGI) in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg (KAW) Foundation and the Swedish Research Council (VR), and from the National Bioinformatics Infrastructure Sweden (NBIS). H.H. was partly supported by FORMAS general grant no. 2019-00874. We gratefully acknowledge help from the Norway spruce and Scots pine genome sequencing consortium staff at Umeå University, the Swedish University of Agricultural Sciences, Uppsala University, and KTH Royal Institute of Technology. The research was mainly funded by the Knut and Alice Wallenberg Foundation in two linked projects to the Umeå Plant Science Centre (KAW 2018.0272) and SciLifeLab (KAW 2018.0273).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data from this project are archived in Figshare and are accessible as: Axiom 450K SNP array for 480 Scots pine individuals. Raw data and array annotation file (figshare.com). Axiom 50K SNP array for 192 individuals from six different pine species. Raw data and array annotation file (figshare.com).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Geographic origin of the 192 whole-genome resequenced Scots pine individuals used for the construction of the SNP screening array.

Figure S2. Geographic origin of the Eurasian *Pinus* spp. individuals genotyped with the PsyI50K SNP array for transferability evaluation.

Figure S3. Geographic origin of the North American *Pinus* spp. individuals genotyped with the PsyI50K SNP array for transferability evaluation.

Figure S4. Length (in base pairs) density distribution for all contigs in the *P. sylvestris* draft genome assembly (black) and contigs with SNPs selected for the PsyI50K array (red).

Figure S5. Density plot of minor allele frequency (MAF) among genic (solid blue line) and intergenic (dashed red light) SNPs present in the PsyI50K array.

Figure S6. *Pinus* spp. clustering estimated using a principal component analysis (PCA) on the relatedness matrix calculated based on 10 841 recommended SNPs. PC1 against PC3 is shown (compared to PC1 against PC2 in Main Text Figure 8e).

Figure S7. *Pinus nigra* and *P. tabuliformis* clustering estimated using a principal component analysis (PCA) on the relatedness matrix calculated based on 20 176 recommended SNPs for the *P. nigra*–*P. tabuliformis* cluster. *P. nigra* subspecies are highlighted.

Figure S8. *Pinus contorta* subspecies clustering estimated using a principal component analysis (PCA) on the relatedness matrix calculated based on 17 007 recommended SNPs for *P. contorta*.

Figure S9. *Pinus cembra* and *P. strobus* clustering estimated using a principal component analysis (PCA) on the relatedness matrix calculated based on 9207 recommended SNPs for the *P. cembra*–*P. strobus* cluster.

Figure S10. Density plot of Minor Allele Frequency (MAF) among genic (solid blue lines) and intergenic (dashed red lights) SNPs present in the recommended SNPs for the subgenus *Pinus* (left panel; 10 841 total SNPs) and subgenus *Strobus* (right panel; 9207 total SNPs).

Data S1. Supplementary Methods.

Table S1. Geographic origin of 192 Scots pine individuals whole-genome resequenced for SNP calling and screening array development.

Table S2. Geographic origin of 385 Scots pine breeding populations and range-wide individuals genotyped for final SNP array development.

Table S3. Geographic origin of 192 *Pinus* spp. individuals genotyped with the PsyI50K array for transferability evaluation.

REFERENCES

Affymetrix. (2015) *SNPolisher user guide (version 1.5.2)*. Available at: https://tools.thermofisher.com/content/sfs/manuals/SNPolisher_User_Guide.pdf [Accessed 21st August 2023].

Affymetrix. (2022) *Axiom genotyping solution analysis guide*. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf [Accessed 21st August 2023].

Ahmar, S., **Ballesta, P.**, **Ali, M.** & **Mora-Poblete, F.** (2021) Achievements and challenges of genomics-assisted breeding in forest trees: from marker-assisted selection to genome editing. *International Journal of Molecular Sciences*, **22**, 10583.

Bernhardsson, C., **Wang, X.**, **Eklöf, H.** & **Ingvarsson, P.K.** (2020) Variant calling using whole genome resequencing and sequence capture for population and evolutionary genomic inferences in Norway spruce (*Picea abies*). In: Porth, I.M. & De la Torre, A.R. (Eds.) *The spruce genome*. New York: Springer International Publishing, pp. 9–36. Available from: https://doi.org/10.1007/978-3-030-21001-4_2

Bernhardsson, C., **Zan, Y.**, **Chen, Z.**, **Ingvarsson, P.K.** & **Wu, H.X.** (2021) Development of a highly efficient 50K single nucleotide polymorphism genotyping array for the large and complex genome of Norway spruce (*Picea abies* L. Karst) by whole genome resequencing and its transferability to other spruce species. *Molecular Ecology Resources*, **21**, 880–896.

Caballero, M., **Lauer, E.**, **Bennett, J.**, **Zaman, S.**, **McEvoy, S.**, **Acosta, J.** ***et al.*** (2021) Toward genomic selection in *Pinus taeda*: integrating resources to support array design in a complex conifer genome. *Applications in Plant Sciences*, **9**, e11439.

Chancerel, E., **Lamy, J.-B.**, **Lesur, I.**, **Noirot, C.**, **Klopp, C.**, **Ehrenmann, F.** ***et al.*** (2013) High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology*, **11**, 50.

Cooper, R.J. (2003) World markets for coniferous forest products: recent trends and future prospects. *Acta Horticulturae*, **615**, 349–353.

Danecek, P., **Auton, A.**, **Abecasis, G.**, **Albers, C.A.**, **Banks, E.**, **DePristo, M.A.** ***et al.*** (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Danecek, P., **Bonfield, J.K.**, **Liddle, J.**, **Marshall, J.**, **Ohan, V.**, **Pollard, M.O.** ***et al.*** (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, giab008.

De La Torre, A.R., **Birol, I.**, **Bousquet, J.**, **Ingvarsson, P.K.**, **Jansson, S.**, **Jones, S.J.** ***et al.*** (2014) Insights into conifer giga-genomes. *Plant Physiology*, **166**, 1724–1732.

Eckenwalder, J.E. (2009) *Conifers of the world: the complete reference*. Portland: Timber Press.

Endelman, J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4**, 250–255.

Farjon, A. (2018) The Kew review: conifers of the world. *Kew Bulletin*, **73**, 8.

Geraldes, A., **Difazio, S.**, **Slavov, G.**, **Ranjan, P.**, **Muchero, W.**, **Hannemann, J.** ***et al.*** (2013) A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Molecular Ecology Resources*, **13**, 306–323.

Graham, N., **Telfer, E.**, **Frickey, T.**, **Slavov, G.**, **Ismael, A.**, **Klápště, J.** ***et al.*** (2022) Development and validation of a 36K SNP Array for radiata pine (*Pinus radiata* D. Don). *Forests*, **13**, 176.

Grattapaglia, D. (2022) Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. *Forests*, **13**, 1554.

Grattapaglia, D., **Silva-Junior, O.B.**, **Resende, R.T.**, **Cappa, E.P.**, **Müller, B.S.**, **Tan, B.** ***et al.*** (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. *Frontiers in Plant Science*, **9**, 1693.

Haapanen, M., **Jansson, G.**, **Nielsen, U.B.**, **Steffenrem, A.** & **Stener, L.G.** (2015) *The status of tree breeding and its potential for improving biomass production–a review of breeding activities and genetic gains in Scandinavia and Finland*. Uppsala: Skogforsk https://www.skogforsk.se/contentassets/9d9c6eeaef374a2283b2716edd8d552e/the-status-of-tree-breeding-low.pdf [Accessed 20th September 2020].

Hall, D., **Hallingbäck, H.R.** & **Wu, H.X.** (2016) Estimation of number and size of QTL effects in forest tree traits. *Tree Genetics & Genomes*, **12**, 110.

Howe, G.T., **Jayawickrama, K.**, **Kolpak, S.E.**, **Kling, J.**, **Trappe, M.**, **Hipkins, V.** ***et al.*** (2020) An axiom SNP genotyping array for Douglas-fir. *BMC Genomics*, **21**, 9.

Hussain, W., **Baenziger, P.S.**, **Belamkar, V.**, **Guttieri, M.J.**, **Venegas, J.P.**, **Easterly, A.** ***et al.*** (2017) Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. *Scientific Reports*, **7**, 16394.

Isik, F. & **McKeand, S.E.** (2019) Fourth cycle breeding and testing strategy for Pinus taeda in the NC State University cooperative tree improvement program. *Tree Genetics & Genomes*, **15**, 70.

Jackson, C., **Christie, N.**, **Reynolds, S.M.**, **Marais, G.C.**, **Tii-kuzu, Y.**, **Caballero, M.** ***et al.*** (2022) A genome-wide SNP genotyping resource for tropical pine tree species. *Molecular Ecology Resources*, **22**, 695–710.

Jin, W.-T., **Gernandt, D.S.**, **Wehenkel, C.**, **Xia, X.-M.**, **Wei, X.-X.** & **Wang, X.-Q.** (2021) Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2022302118.

Kaňák, J. (2016) *Arboretum Sofronka 1956–2016*. Plzeň: RAMAP Plzeň.

Kastally, C., **Niskanen, A.K.**, **Perry, A.**, **Kujala, S.T.**, **Avia, K.**, **Cervantes, S.** ***et al.*** (2022) Taming the massive genome of scots pine with PiSy50k, a new genotyping array for conifer research. *The Plant Journal*, **109**, 1337–1350.

Li, H., **Achour, I.**, **Bastarache, L.**, **Berghout, J.**, **Gardeux, V.**, **Li, J.** ***et al.*** (2016) Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. *NPJ Genomic Medicine*, **1**, 16006.

Li, H. & **Durbin, R.** (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., **Handsaker, B.**, **Wysoker, A.**, **Fennell, T.**, **Ruan, J.**, **Homer, N.** ***et al.*** (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McKeand, S.E., **Payn, K.G.**, **Heine, A.J.** & **Abt, R.C.** (2021) Economic significance of continued improvement of loblolly pine genetics and its efficient deployment to landowners in the southern United States. *Journal of Forestry*, **119**, 62–72.

Neale, D.B. & **Savolainen, O.** (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330.

Niu, S., **Li, J.**, **Bo, W.**, **Yang, W.**, **Zuccolo, A.**, **Giacomello, S.** ***et al.*** (2022) The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell*, **185**, 204–217.

Pavy, N., **Gagnon, F.**, **Rigault, P.**, **Blais, S.**, **Deschênes, A.**, **Boyle, B.** ***et al.*** (2013) Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources*, **13**, 324–336.

Perry, A., **Wachowiak, W.**, **Downing, A.**, **Talbot, R.** & **Cavers, S.** (2020) Development of a single nucleotide polymorphism array for population genomic studies in four European pine species. *Molecular Ecology Resources*, **20**, 1697–1705.

Plomion, C., **Bartholomé, J.**, **Lesur, I.**, **Boury, C.**, **Rodríguez-Quilón, I.**, **Lagraulet, H.** ***et al.*** (2016) High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Molecular Ecology Resources*, **16**, 574–587.

Schierding, W., **Antony, J.**, **Cutfield, W.S.**, **Horsfield, J.A.** & **O'Sullivan, J.M.** (2016) Intergenic GWAS SNPs are key components of the spatial and regulatory network for human growth. *Human Molecular Genetics*, **25**, 3372–3382.

Silva-Junior, O.B., **Faria, D.A.** & **Grattapaglia, D.** (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 eucalyptus tree genomes across 12 species. *The New Phytologist*, **206**, 1527–1540.

Thavamanikumar, S., **Southerton, S.G.**, **Bossinger, G.** & **Thumma, B.R.** (2013) Dissection of complex traits in forest trees—opportunities for marker-assisted selection. *Tree Genetics & Genomes*, **9**, 627–639.

Van der Auwera, G.A. & **O'Connor, B.D.** (2020) *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. Sebastopol: O'Reilly Media.

Wu, H. & **Nilsson, O.** (2023) Threatened forests: As the northern forests suffer from the effects of climate change, genomics has great potential to help them adapt. *EMBO Reports*, **24**, e57106.

Wu, H.X., **Hallingbäck, H.R.** & **Sánchez, L.** (2016) Performance of seven tree breeding strategies under conditions of inbreeding depression. *G3: Genes Genomes, Genetics*, **6**, 529–540.

Wu, T.D. & **Watanabe, C.K.** (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Zhou, X. & **Stephens, M.** (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**, 821–824.