


# 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

Carolina Bernhardsson<sup>1,2</sup> | Yanjun Zan<sup>3</sup> | Zhiqiang Chen<sup>3</sup> | Pär K. Ingvarsson<sup>4</sup>  | Harry X. Wu<sup>3,5,6</sup> 

<sup>1</sup>Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

<sup>2</sup>Department of Organismal Biology, Uppsala University, Uppsala, Sweden

<sup>3</sup>Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Science, Umeå, Sweden

<sup>4</sup>Linnean Centre for Plant Biology, Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Science, Uppsala, Sweden

<sup>5</sup>Beijing Advanced Innovation Centre for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China

<sup>6</sup>Black Mountain Laboratory, CSIRO National Research Collection Australia, Canberra, ACT, Australia

## Correspondence

Harry X. Wu, Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Science, Umeå, Sweden.  
Email: harry.wu@slu.se

## Funding information

Stiftelsen för Strategisk Forskning, Grant/Award Number: RBP14-0040; Horizon2020-B4EST, Grant/Award Number: 773383

## Abstract

Norway spruce (*Picea abies* L. Karst) is one of the most important forest tree species with significant economic and ecological impact in Europe. For decades, genomic and genetic studies on Norway spruce have been challenging due to the large and repetitive genome (19.6 Gb with more than 70% being repetitive). To accelerate genomic studies, including population genetics, genome-wide association studies (GWAS) and genomic selection (GS), in Norway spruce and related species, we here report on the design and performance of a 50K single nucleotide polymorphism (SNP) genotyping array for Norway spruce. The array is developed based on whole genome resequencing (WGS), making it the first WGS-based SNP array in any conifer species so far. After identifying SNPs using genome resequencing data from 29 trees collected in northern Europe, we adopted a two-step approach to design the array. First, we built a 450K screening array and used this to genotype a population of 480 trees sampled from both natural and breeding populations across the Norway spruce distribution range. These samples were then used to select high-confidence probes that were put on the final 50K array. The SNPs selected are distributed over 45,552 scaffolds from the *P. abies* version 1.0 genome assembly and target 19,954 unique gene models with an even coverage of the 12 linkage groups in Norway spruce. We show that the array has a 99.5% probe specificity, >98% Mendelian allelic inheritance concordance, an average sample call rate of 96.30% and an SNP call rate of 98.90% in family trios and haploid tissues. We also observed that 23,797 probes (50%) could be identified with high confidence in three other spruce species (white spruce [*Picea glauca*], black spruce [*P. mariana*] and Sitka spruce [*P. sitchensis*]). The high-quality genotyping array

Carolina Bernhardsson and Yanjun Zan contributed equally.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd

will be a valuable resource for genetic and genomic studies in Norway spruce as well as in other conifer species of the same genus.

#### KEYWORDS

genetic diversity, genome resequencing, genomic selection, Norway spruce, SNP array

## 1 | INTRODUCTION

Forests occupy one-third of the global land mass, covering more than four billion hectares of the planet and play key roles in water, oxygen and nutrient cycles as well as in carbon sequestration (FAO, 2016). The coniferous forest biome makes up one-third of the world's forests, representing 80% of the Earth's biomass (Neale & Kremer, 2011). Conifers also include some of the most important tree species used for plantation establishment, wood production and tree improvement programmes (FAO, 2015). Of the 264 million hectares covered by planted forests (6.6% of the total world forests), more than 50% consist of conifer species. The importance of conifers has motivated large investments into fundamental research on the basic and applied biology of trees (Plomion et al. 2016) and has driven the development of the most advanced tree breeding programmes in the world (Isik & McKeand, 2019; Wu et al., 2016). Projected climate changes in the 21st century are likely to have profound impacts on the functioning of Earth's ecosystems, including most conifer species (Garcia et al., 2014). Their commercial importance and the threats of climate change effects on conifers make it important to study biodiversity, the genetic basis of climate adaptation, and the genomic basis of productivity. Conifers are ideal species for such tasks due to their large geographical distribution and rich genetic diversity (Neale & Wheeler, 2019). To understand the genomic basis of climate adaptation and to accelerate tree breeding programmes in conifers, genetic markers have been used to dissect the genetic basis of adaptive and commercial traits and to explore marker-assisted selection. Traditionally, random DNA markers such as RFLPs, RAPDs, simple sequence repeats (SSRs) and single nucleotide polymorphism (SNP) markers derived from candidate gene approaches have been used for association studies (Thavamanikumar et al., 2013). Due to the limited number of markers available in such studies, the large number of quantitative trait loci (QTLs) underlying quantitative trait variation (Hall et al., 2016), and the rapid linkage disequilibrium (LD) decay in forest trees (Neale & Savolainen, 2004), the dissection of QTLs underlying quantitative trait variation has had limited success in conifers. Consequently, marker-assisted selection has not been implemented in tree breeding (Isik, 2014). However, the recent development of genomic selection (GS), which utilizes large numbers of genome-wide markers to predict complex phenotypes, has the potential to shorten the breeding cycles, increase selection intensity and improve the accuracy of breeding values (Grattapaglia et al., 2018). However, one of the main limiting factors in implementing GS in conifers is the lack of affordable, reliable and abundant genome-wide markers.

Several SNP arrays have recently been developed in conifers for use in genome-wide association studies (GWAS) and GS. These have mostly been based on candidate gene sequencing but have also utilized data from microarrays or RNA sequencing and are generally limited to a few thousand SNPs (Bartholome et al., 2016; Beaulieu et al., 2014; Resende et al., 2012; Zapata-Valenzuela et al., 2013) to several tens of thousands of SNPs (Howe et al., 2020; Perry et al., 2020). Two high-density SNP arrays relying on the Infinium iSelect technology were designed for the conifer species white spruce (*Picea glauca*), containing 7,338 and 9,559 SNPs, respectively, using in silico SNP prediction through the alignment of transcript sequences and candidate genes (Pavy et al., 2013). A 9K Illumina Infinium SNP array was developed for maritime pine (*Pinus pinaster*) by bundling markers from SNPs discovered in candidate gene sequencing and from 454 sequencing reads of RNA derived from multiple tissues from three provincial parents (Plomion et al., 2016). A similar Infinium SNP array was developed from in silico SNP resources and exome capture sequencing for black spruce (*Picea mariana*) (Pavy et al., 2016). Recently, an Axiom SNP genotyping array with 55K SNPs was developed for Douglas-fir (*Pseudotsuga menziesii*) from transcriptome sequencing (Perry et al., 2020). For Norway spruce, high-quality SNPs have been developed based on large-scale sequence capture and have been employed for both GWAS and GS (Azaiez et al. 2018; Baison et al., 2019; Chen et al., 2018; Vidalis et al. 2018). Various SNP arrays have also been available for poplar and other broadleaved tree species that have been used in association genetics and GS studies (Gerald et al., 2013). One of the most successful SNP arrays in hardwood tree species is the EUChip60K, which was based on resequencing of 240 trees from 12 species (Silva-Junior et al., 2015) and has been used to genotype many thousands of *Eucalyptus* trees for GS and GWAS (Grattapaglia et al., 2018).

Conifers, and particularly the commercially important pine and spruce species, have large genomes spanning 20 to 30 Gb. Developing genome-wide SNP arrays, covering both intragenic and intergenic regions, was until recently still a significant challenge due to the lack of high-quality reference genomes. The particular challenge with genotyping conifer genomes stems from their large and complex genomes that contain a high fraction of repetitive elements and abundant polymorphisms, which yields many opportunities for spurious binding of probes or primers. However, recent genome sequencing of several conifer species (Neale et al., 2014, 2017; Nystedt et al., 2013; Stevens et al., 2016; Warren et al., 2015) has made it possible to develop genome-wide marker panels using whole genome resequenced trees for GWAS, population genetics studies and GS. In this paper, we report the development, evaluation and

transferability of a highly efficient Norway spruce 50K SNP array using whole genome resequencing, probably for the first time in conifers.

## 2 | MATERIALS AND METHODS

### 2.1 | Plant materials

We used three steps to design and validate the final genotyping array. First, we used whole genome resequenced data based on 35 Norway spruce samples, previously described in Bernhardsson et al. (2020) and Wang et al. (2020), for the initial SNP selection. Second, we screened the selected SNPs in 480 Norway spruce samples collected from two field trials, one consisting of 258 trees from a provenance trial of a species range-wide collection established in Hungary and 222 trees derived from a Swedish breeding population trial established by Skogforsk (Table 1). All 480 samples were screened using a pilot screening array consisting of ~450K SNPs and these data formed the basis for the final SNP selection. Among the 480 trees, nine individuals were replicated twice each to serve as internal controls. Finally, to evaluate the final 50K array we genotyped three sets of samples. First, we genotyped a collection of 28 haploid megagametophytes collected from cones of the reference genome individual Z4006 (Nystedt et al., 2013). Second, a set of Norway spruce full-sibling trios collected from four families (48 trees in total) were genotyped to assess possible Mendelian segregation errors. Finally, we genotyped 49 white spruce (*Picea glauca*), 61 black spruce (*Picea mariana*) and 50 Sitka spruce (*Picea sitchensis*) samples planted in Sweden to assess the between-species transferability of the final array. Detailed information regarding sampling origins and sample metadata are available in Tables S1–S5.

For the haploid megagametophytes, seeds were soaked in 1% H<sub>2</sub>O<sub>2</sub> for 16 hr and germinated in a Petri dish on top of moistened

filter paper at room temperature (~21°C). When embryos reached ~5 mm in length, seed coats were removed and megagametophytes were separated from embryos using sterile razor blades and manually ground in liquid N<sub>2</sub> in 1.5-ml tubes using plastic pestles. The diploid samples used for screening the pilot array and for validating genotyping rates and for assessing transferability were collected during early summer 2018 and DNA was extracted from either newly flushed needles or from cambium samples. DNA was extracted using a NucleoSpin Plant II DNA Kit (Macherey-Nagel) following the default protocol. Based on NanoDrop 2000 (Thermo Fisher Scientific) measurements, the DNA yield was highly variable among samples, ranging from 303 to 1,116 ng (mean ± SD = 465 ± 201 ng). The extracted DNA samples were shipped to the Microarray Research Services Laboratory at Thermo Fisher Scientific on dry ice and were requantified using PICOGREEN.

### 2.2 | Construction of the pilot screening array

The 35 whole genome resequenced Norway spruce samples were originally collected from Russia (one), Romania (one), Poland (one), Belarus (one), Sweden (22), Norway (five) and Finland (four) (described in more detail in Bernhardsson et al., 2020 and Wang et al., 2020). The WGS samples were used to find and extract candidate genome sequences for probe design of the screening array. In short, the mapping and genotype calling of samples were performed as follows. The raw sequencing reads were mapped against the full version 1.0 assembly of Norway spruce (Nystedt et al., 2013) using BWA MEM version 0.7.15 (Li & Durbin, 2009), with default parameters, and the BAM files were subsequently subset by SAMTOOLS version 1.5 (Li et al., 2009) to only include scaffolds >1 kb. The reduced assembly and bam files (containing 1,970,460 out of ~10 million scaffolds and 9.4 Gb out of 12.5 Gb of the full version 1.0 genome assembly) were then split into 20 subsets, each containing ~100,000 scaffolds.

Sample origin	Swedish breeding population trial	Hungarian provenance trial	Total
Russian-Baltic (Rus_Bal) <sup>2</sup>	9	10	19
Alpine (ALP) <sup>3</sup>	63	86 (84)	149 (147)
Central Europe (CEU) <sup>4</sup>	9	115 (109)	124 (118)
Northern Poland (NPL)	8	13	21
Carpathian (ROM) <sup>5</sup>	1	16	17
Fennoscandia (NFE) <sup>1</sup>	41 (38)	1	42 (39)
Southern/Central Scandinavia (C_Sc)	87	17 (16)	104 (103)
Unknown (U)	4	—	4
Total	222 (219)	258 (249)	480 (468)

**TABLE 1** Sample origin of the 480 genotypes used for screening the pilot array; numbers in parentheses show the number of samples from each origin and trial that passed the QC thresholds

Note: Sample origin: 1. Fennoscandia contains samples from Finland and northern Sweden; Southern Scandinavia from Central/Southern Sweden and Central/Southern Norway; 2. Russian-Baltic from Russia, Belarus, Estonia, Latvia and Lithuania; 3. Alpine from Denmark, Germany, Switzerland, France and Italy; 4. Central Europe from Slovakia, Czech Republic, Southern Poland, Hungary and Austria; 5. Carpathian from Romania and Bulgaria.

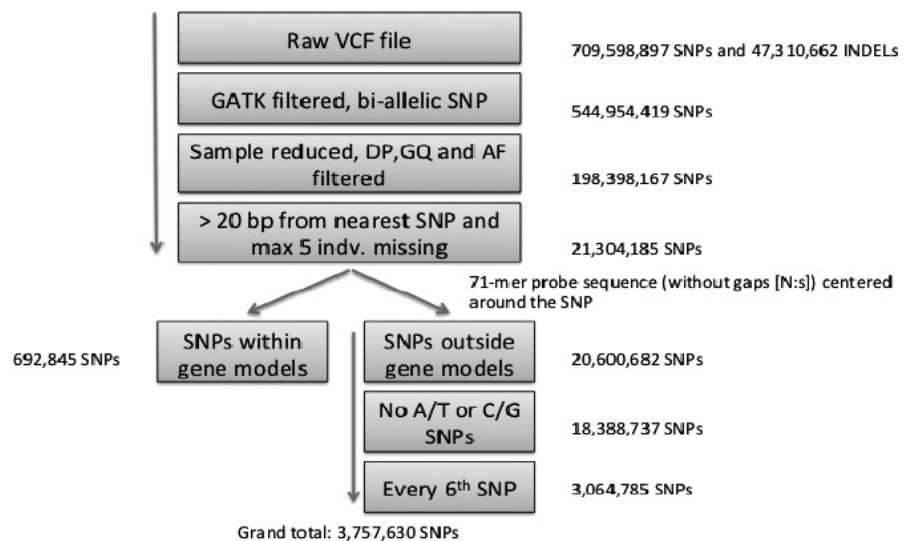
All subset BAM files were then marked for optical duplicates using *PICARD* version 2.0.1 (<https://broadinstitute.github.io/picard/>) and aligned around indels using *GATK* version 3.7 (McKenna et al., 2010). Per-individual variants were called using *GATK* *HAPLOTYPECALLER* in *v.vcf* format (DePristo et al., 2011; Van der Auwera et al., 2013) before a joint genotype call over all 35 individuals was conducted separately on the 20 genomic subsets using *GATK* *GENOTYPEGVCF* (DePristo et al., 2011; Van der Auwera et al., 2013).

The combined raw VCF-file (containing more than 709 million SNPs and 43 million indels, Figure 1) across the 20 genomic subsets was filtered in several steps. First, only bi-allelic SNPs > 5 bp away from an indel and that followed the filtering criteria based on *GATK*'s "best practice" (<https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set>) were kept (Bernhardsson et al., 2020). Since six of the WGS samples had a quite low sequence coverage (average coverage ~ 6x) and thereby also a lower confidence in SNP calls, the VCF-files were subset to only include 29 samples derived from Norway, Finland and Sweden (Fennoscandia), which all had high coverage (15–20 × for called sites on average). Since the Norway spruce genome is highly repetitive (~70% of the 1K scaffold assembly contains repeat sequences (Nystedt et al., 2013), we filtered individual calls for depth, accepting a range between 6x and 30x per individual with a genotype quality (GQ) > 15. Only SNPs with an alternative allele frequency (AF) between 0.05 and 0.95 and with a maximum of 30% missing data were kept at this filtering step. To fulfil Affymetrix's filtering criteria ([https://tools.thermofisher.com/content/sfs/brochures/snp\\_template\\_for\\_axiom\\_mydesign\\_custom\\_arrays\\_v2.zip](https://tools.thermofisher.com/content/sfs/brochures/snp_template_for_axiom_mydesign_custom_arrays_v2.zip)), we then extracted 71-mer probe sequences for SNPs with >20 bp to nearest SNP and where a maximum of five individuals showed missing data. If no gaps (Ns) were found in the probe sequences that we extracted from the assembly, the SNP was considered a good candidate for in silico probe evaluation. A final down-sampling was made of all candidate probes to fit the recommended number of probes used for testing (3,757,630 probe sequences). During this filtering, all SNPs positioned within gene models (hereafter called intragenic SNPs)

were kept, while SNPs outside of gene models (hereafter called 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 sixth SNP was kept. When ranking the proposed markers, all intragenic markers were considered as "important" while all intergenic SNPs were assigned a "standard" importance. This resulted in a total of 3,757,630 SNPs which were sent to ThermoFisher's bioinformatics service for in silico Axiom testing (Figure 1).

For quality control of the array, 8,000 36-mer probe sequences (so called DQC sequences, following ThermoFisher's guidelines) were extracted from monomorphic regions (based on the unfiltered VCF-file for all 35 samples) of a hard-masked version of the Norway spruce assembly. These DQC sequences were evenly distributed between the two strands (+/-) and also between A/T and C/G sites as the probe's ligation position (position 31 in the sequence). In total, 2,000 of these DQCs will be incorporated into the array for control or every run to control for signal variation across the array at sites in the genome known not to vary among individuals.

To select 450K SNPs for the pilot screening array, in silico tests of 3,757,630 SNPs were conducted by Affymetrix. A pConvert score (ranging from 0 to 1) was produced for each SNP by the test. This score reflects the relative probability of probe success and is based on the thermodynamics of the probe sequence itself as well as the number of 16-nt hits found in the reference genome (Affymetrix used the Norway spruce reference genome version 1.0, Nystedt et al., 2013). The probes were first divided into two blocks, "not possible" and "buildable," where the "not possible" probes are given a pConvert score of 0. For the "buildable" probes, the scores are subsequently translated into three recommendation levels, where a pConvert score of 0.6–1 is "recommended", 0.4–0.6 "neutral" and 0–0.4 "not recommended." Among the 3,757,623 SNPs (after removing seven duplicates), 761,311 markers were recommended that had no interfering polymorphisms located within 24 bases on either side of the marker. These recommended markers contained



**FIGURE 1** Schematic illustration of the variant filtering pipeline for extracting candidate probe sequences for the Axiom in silico testing at ThermoFisher. Each of the filtering steps described in the text is presented in a grey boxes with the number of surviving SNPs labelled beside

all the intragenic 259,994 markers selected plus the highest ranked and recommended intergenic SNPs (190,499), resulting in a total of 450,493 SNPs that was used for design of the pilot screening array.

### 2.3 | Genotype calling of Axiom screening array

In total, 480 Norway spruce samples from two trials (Table 1) were genotyped using the pilot screening array. Genotype calling of the 450K pilot Axiom screening array was performed using the Axiom analysis suite (version 4.0, available for download at <https://www.thermofisher.com/se/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html>), following best practice with default parameters (an SNP call rate cutoff [cutoff]  $\geq 0.97$  and a sample call using a Dish-QC threshold [axiom\_dishqc\_DQC]  $\geq 0.82$ ) (Affymetrix, 2016). The sample call rate is defined as the average SNP call rate across all SNPs for a sample. The called genotypes were then used to classify the 450,493 SNPs into six categories of SNP performance (Table 2) (Affymetrix, 2016). A VCF file with allelic calls for all 450K SNPs, coded as A, T, C or G, was exported from the Axiom analysis suite and used for all downstream analyses.

For the species transferability validation with white, black and Sitka spruce species, genotype calling was made using the best practice pipeline with a few modifications. It was not possible to use the Dish-QC value (axiom\_dishqc\_DQC  $> 0.82$ ) and sample call rate (qc\_call\_rate)  $\geq 0.97$ ) as a proportion of the probes were not expected to be transferable to these species. To obtain summary statistics for the probes and call genotypes to evaluate transferability in spruce species, a modified sample Dish-QC value (0.75) and sample call rate

(0.75) were used with the remaining setup being identical to the best practice pipeline.

### 2.4 | Selection of the 50K SNP array from the pilot screening array

Although PolyHighResolution (PHR) SNPs, NoMinorHom (NH) SNPs and MonoHighResolution (MHR) SNPs were all recommended by the Axiom analysis suite for consideration in downstream analyses, we selected the final 50K array only from the PHR SNPs for stringency. Three filtering steps were performed on the PHR SNPs to obtain the final 50K probes. SNPs with MAF lower than 0.05 in either of the two populations were excluded. SNPs with pairwise LD  $\geq 0.8$  (linkage disequilibrium measured as  $r^2$ ) were pruned to reduce the number of nonindependent SNPs. This was achieved by first calculating all pairwise  $r^2$  values using `vcftools` (version 0.1.13) (Danecek et al., 2011). To minimize the computing time due to constant I/O operation, only SNP pairs with  $r^2$  values  $> 0.6$  were output by using `"vcftools -vcf INPUT.vcf --geno-r2 --min-r2 0.6 -out OUTPUT."` An "igraph" object was subsequently built using the output from `vcftools` by connecting all SNP pairs with LD  $\geq 0.8$ . Then, independent SNPs were extracted by selecting the maximum number of independent SNPs from each cluster. This was achieved by first building networks that connect all SNPs with LD  $\geq 0.8$ . We selected the hub SNPs and removed the radial SNPs in these networks to minimize the number of selected SNPs while maximizing information retained. Second, selecting hubs and removing the radial loci from the network one at a time will result in the collapse of old networks. We therefore rebuilt the network from the remaining SNPs and then repeated steps 1 and 2 until no networks with more than two SNPs were found. Third, we randomly selected one SNP from the remaining SNPs pairs from step

	Number of SNPs <sup>b</sup>	Average heterozygosity <sup>c</sup>	Average MAF <sup>d</sup>	Average missingness <sup>e</sup>
Full screening array	450,493 (100%)	0.17 (0.00–0.94)	0.13 (0.00–0.50)	0.04 (0.00–0.94)
PHR* SNPs <sup>a</sup>	176,800 (39.3%)	0.24 (0.00–0.87)	0.17 (0.00–0.50)	0.01 (0.00–0.03)
NH* SNPs	69,455 (15.4%)	0.06 (0.00–0.50)	0.03 (0.00–0.25)	0.01 (0.00–0.03)
MHR* SNPs	12,820 (2.9%)	0.00 (–)	0.00 (–)	0.00 (0.00–0.03)
CRBT SNPs	49,901 (11.1%)	0.28 (0.00–0.85)	0.22 (0.00–0.50)	0.06 (0.03–0.94)
OTV SNPs	3,404 (0.8%)	0.16 (0.00–0.94)	0.10 (0.00–0.50)	0.03 (0.00–0.19)
O SNPs	138,113 (30.7%)	0.17 (0.00–0.89)	0.12 (0.00–0.50)	0.10 (0.00–0.94)

<sup>a</sup>Clusters recommended by ThermoFisher.

<sup>b</sup>Number of SNPs with the percentage of SNPs in parentheses.

<sup>c</sup>Average heterozygosity for SNPs with the range of heterozygosity in parentheses.

<sup>d</sup>Average minor allele frequency (MAF) for SNPs with the range of MAF in parentheses.

<sup>e</sup>Average missingness per SNP with the range of missingness in parentheses.

**TABLE 2** SNP metrics for the different cluster categories: full screening array, PolyHighResolution (PHR), NominorHom (NH), MonoHighResolution (MHR), CallRateBelowThreshold (CRBT), OffTargetVariant (OTV) and Other (O) markers

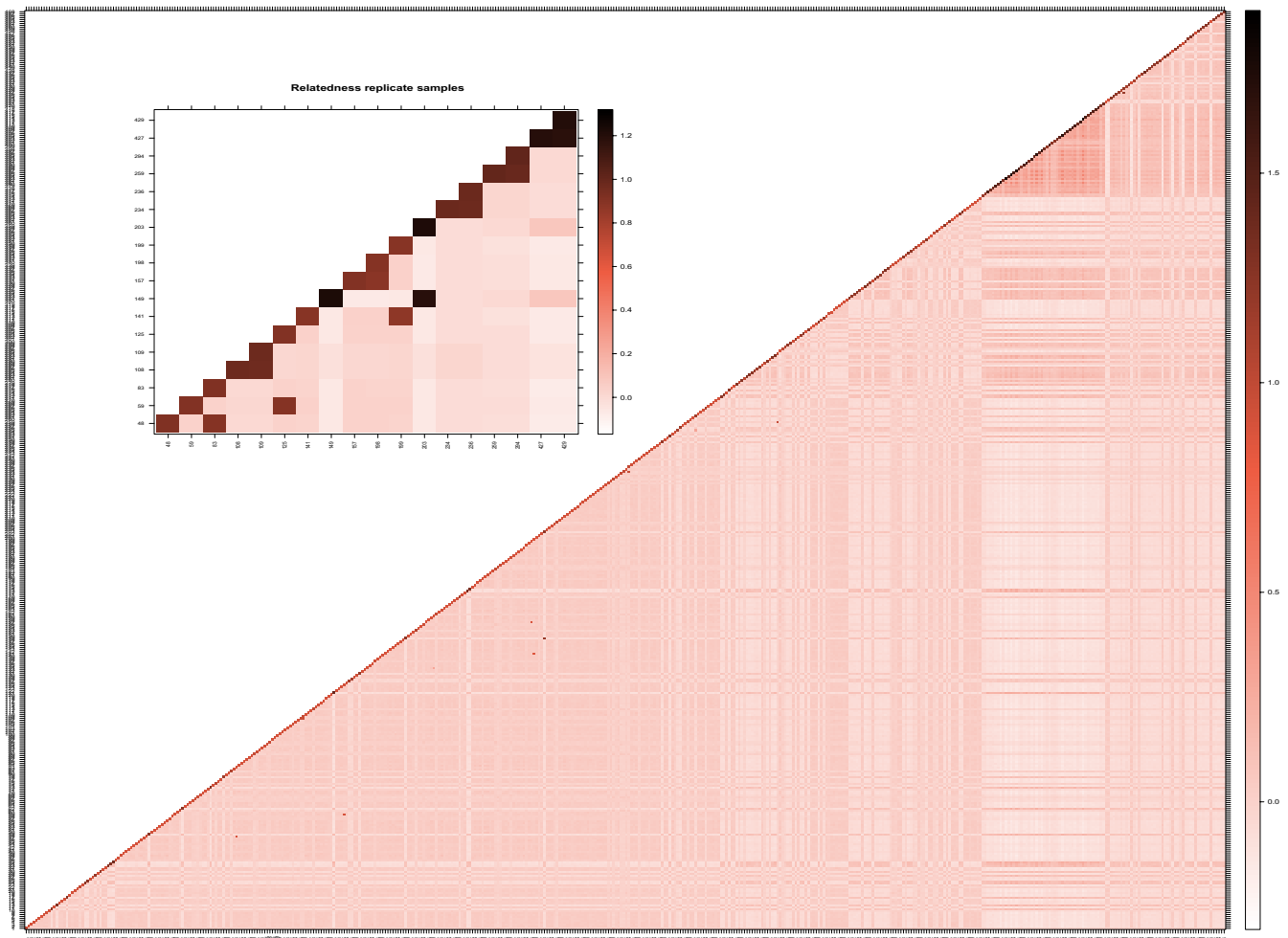
3. Fourth, the hub SNPs from steps 1 and 2 and SNPs from step 3 were kept for downstream analysis in our study. All these analyses were performed using customized R scripts using the “igraph” package (available at <https://github.com/yanjunzan/script/tree/master/umeaArray>). Ultimately, SNPs with low average congruence scores ( $< 0.95$ , measured as the mean congruency across nine pairs of replicates), and SNPs with heterozygosity levels  $> 0.6$ , were removed.

To select the final SNPs for the array, we attempted to cover as many of genomic regions as possible by first selecting one SNP per scaffold. If an intragenic SNP within the scaffold was available, that SNP was prioritized, otherwise an intergenic SNP was randomly selected. Meanwhile, G/C and A/T SNPs were avoided whenever possible. To tag as many unique gene models as possible, an additional 160 SNPs were selected to incorporate 160 gene models not yet covered under the preceding procedure. We also included an additional 125 SNPs that were flanking known associations from Baison et al. (2019), Elfstrand et al. (2020) or preliminary associations from GWAS on bud flush, bud set and wood quality traits (our unpublished data). Finally, an additional 1,608 SNPs were randomly selected to bring the total number of selected SNPs up to 47,445, which could fit on the 50K Axiom array together with ~2,000 control probes to

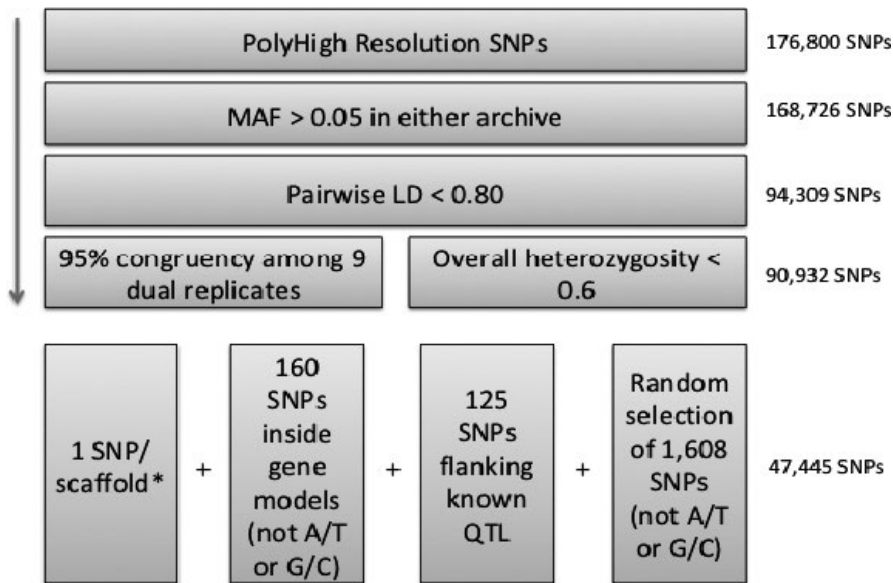
account for background noise during imaging analysis. A final investigation, to confirm that the selected SNPs were evenly distributed across the Norway spruce genome, was performed by comparing the targeted scaffolds to available genetic maps (Bernhardsson et al., 2019 and our unpublished data) by counting the number of SNPs and scaffolds positioned on different linkage groups (LGs).

## 2.5 | Evaluation and validation of the 50K genotyping array

To evaluate the performance of the 50K genotyping array, we selected and genotyped another three sets of samples. First, four full-sib Norway spruce families consisting of two parents and between 12 and 14 offspring were genotyped to estimate the Mendelian inheritance (MI) error rate. The MI error rate was calculated as the proportion of family trios that violate the Mendelian inheritance rule. For example, under Mendelian inheritance only AB genotypes should be observed in the offspring when the parents are homozygous AA and BB, respectively. Similarly, when parents are homozygous AA and heterozygous AB their offspring should contain the two genotypes



**FIGURE 2** Visualization of the additive relatedness matrix estimated across all 468 samples. The relatedness matrix was calculated with the A.mat function in the R package “rrBLUP” using all PolyHigh resolution SNPs (176,800). Inset: zoom of the nine replicated samples



Grand total: 47,445 SNPs covering 45,552 scaffolds and 19,794 gene models

\* If possible, the SNPs representing each scaffold were chosen not to be A/T or G/C substitutions

AA and AB. Second, 28 haploid megagametophytes were genotyped to evaluate the probe specificity and examine whether probes were binding to different paralogues. For a 100% probe specificity, all genotyped megagametophytes should be homozygous. Therefore, a specificity error rate was calculated for each probe as the proportion of megagametophytes showing a heterozygous call. Third, 160 samples from three other spruce species (white spruce, black spruce and Sitka spruce) were genotyped to evaluate the transferability of our spruce array to other spruce species.

## 2.6 | Principal component analysis and population structure

The population structure of the screening array samples was visualized using a principal component analysis (PCA). First, the realized additive relationship matrix (Figure 2) was constructed using the "A.mat" function from the *rrBLUP* R package (Endelman, 2011) and then a scaled and centred PCA was performed using the 459 nonreplicated samples with the "prcomp" function in R (R Core Team, 2015). This was done by using either all PHR SNPs or the final 50K selected SNPs (Figure 5). The goal was to assess whether the estimated population structure was similar between the 50K and the all PHR SNP (177K) sets.

## 2.7 | Further assessment on ascertainment bias, population structure and genetic diversity

Allele frequency distribution for the selected ~50K array and PHR ~ 177K SNPs were compared to evaluate the selection bias in

terms of MAF. In addition, we compared the MAF and heterozygosity between the range-wide provenance trial collection and Skogforsk's breeding population samples to determine how well the Swedish breeding population captured range-wide genetic diversity. These parameters were calculated for the 50K selected SNPs within each population. Using the estimates above, we also assessed the difference in diversity and how population structure was captured using intergenic or intragenic SNPs. All analyses were implemented with customized R/python scripts that are available on github <https://github.com/yanjunzan/script/tree/master/umeaArray>.

## 3 | RESULTS

### 3.1 | Construction of the 450K pilot screening array

A total of 3,757,630 SNPs including all intragenic SNPs (692,845) and every sixth of the non-A/T or C/G intergenic SNPs (3,064,785) were selected from the original >709 million SNPs, by the multiple filtering processes (Figure 1). These SNPs were sent to ThermoFisher for in silico probe evaluation and selection. After evaluation, all recommended intragenic SNPs (259,994) and the best ranked intergenic SNPs (190,499) were chosen for construction of the 450K pilot screening array.

### 3.2 | Screening of the 450K pilot array and selection of the final 50K Axiom array

A total of 468 samples (97.5% of the total 480) passed the quality control for genotype calling and were considered successfully

**FIGURE 3** Schematic illustration of the probe selection pipeline from the 450K screening array to the final 50K array

genotyped by the 450K screening array (Table 1). Based on the pairwise additive relationship, the nine replicated samples could be fully identified (Figure 2), which gave an average estimated genotype reproducibility of 99.8% over all 450K pilot array SNPs.

Based on hybridization performance and called genotypes, the SNPs were grouped into six categories. The pilot screening array SNPs were composed of all six categories (Table 2), with the largest number of SNPs (39.3%) belonging to the PHR SNPs. Average heterozygosity for all 450K SNPs was 0.17, with MAF of 0.13 and missingness of 0.04. The PHR SNPs displayed higher levels of both heterozygosity (0.24) and MAF (0.17), and showed a lower level of missingness (0.01) compared to the remaining SNPs. The other two recommended SNP categories, MHR and NH, showed very low levels of genetic variation among the 468 samples (Table 2). PHR SNPs were therefore the only category considered for the final 50K array.

In order to select the final ~50K SNPs, the ~177K PHR SNPs were filtered to only keep independent SNPs while tagging as many unique contigs and gene models as possible. This resulted in a final selection of 47,445 SNPs, covering 45,552 scaffolds and 19,794 gene models (Figure 3). To evaluate the genomic distribution of the selected ~50K SNPs, targeted scaffolds were compared to available genetic linkage maps (Bernhardsson et al., 2019 and our unpublished data), and the number of scaffolds positioned on the genetic maps, as well as the number of selected SNPs on that scaffold, were recorded for

each linkage group. In total, 16,659 (35.2%) of the SNPs and 15,103 (33.3%) of the scaffolds could be positioned on the 12 LGs (Table 3), showing that the SNPs selected for the array have a genome-wide distribution. In total, 345 of these scaffolds, harbouring 482 SNPs, appear to be split across several LGs, indicating potential assembly errors (Table 3) (Bernhardsson et al., 2019).

Highly fragmented genome assemblies that are lacking large fractions of the genome due to high genomic repetitiveness can suffer from collapsed read mappings, which in turn may result in spurious SNP calls. Such false SNPs will show strong deviations from Hardy-Weinberg equilibrium (HWE) because they will have an excess of heterozygous calls due to the misalignment of reads from multiple genomic regions (Bernhardsson et al., 2020; McKinney et al., 2017). To analyse how the selected ~50K SNPs behave in comparison to the whole ~450K screening array and the ~177K PHR SNPs in terms of HWE, the MAF of each SNP was plotted against its observed heterozygosity (Figure 4). While the full ~450K screening array contains numerous SNPs with either too low or too high heterozygosity relative to their MAF, the majority of PHR SNPs and the selected ~50K SNPs follow the expected pattern under HWE. The selected SNPs also spanned the entire range of MAFs of the PHR SNPs, except at MAF < 0.05 because these were deliberately filtered out due to low polymorphism rates.

PCA indicates that the final 50K SNP set captures the same population structure as the PHR 177K SNP set for both the trees from

**TABLE 3** Distribution of the ~50,000 final array markers positioned on scaffolds previously mapped to genetic linkage groups (LGs) (Bernhardsson et al., 2019 and our unpublished data)

LG <sup>a</sup>	Number of markers (scaffolds) <sup>b</sup>	Percentage of mapped markers (scaffolds) <sup>c</sup>	Percentage of total number of markers (scaffolds) <sup>d</sup>
LG 1	1,539 (1,403)	9.2% (9.3%)	3.2% (3.1%)
LG 2	1,342 (1,212)	8.1% (8.0%)	2.8% (2.7%)
LG 3	1,392 (1,271)	8.4% (8.4%)	2.9% (2.8%)
LG 4	1,306 (1,187)	7.8% (7.9%)	2.8% (2.6%)
LG 5	1,360 (1,221)	8.2% (8.1%)	2.9% (2.7%)
LG 6	1,260 (1,148)	7.6% (7.6%)	2.7% (2.5%)
LG 7	1,450 (1,327)	8.7% (8.8%)	3.1% (2.9%)
LG 8	1,364 (1,260)	8.2% (8.3%)	2.9% (2.8%)
LG 9	1,312 (1,187)	7.9% (7.9%)	2.8% (2.6%)
LG 10	1,303 (1,198)	7.8% (7.9%)	2.7% (2.6%)
LG 11	1,186 (1,089)	7.1% (7.2%)	2.5% (2.4%)
LG 12	1,363 (1,255)	8.2% (8.3%)	2.9% (2.8%)
Scaffold split over several LGs	482 (345)	2.9% (2.3%)	1.0% (0.8%)
Total	16,659 (15,103)	100% (100%)	35.2% (33.3%)

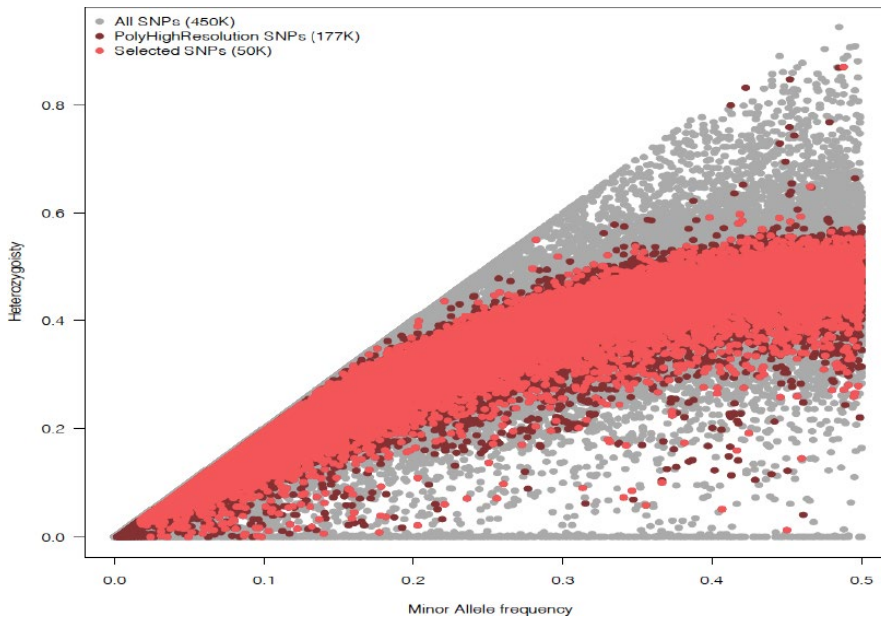
<sup>a</sup>The linkage group (LG) that the marker scaffolds were mapped to in the genetic maps. Markers positioned on scaffolds shown to be split over several LGs in the genetic maps are presented as a separate category.

<sup>b</sup>Number of markers positioned on scaffolds mapped to a certain LG. Number of unique scaffolds that are mapped to a certain LG is presented in parentheses.

<sup>c</sup>Percentage of mapped markers (16,659 in total) that are positioned on scaffolds mapped to a certain LG. Percentage of unique scaffolds (15,103 in total) is presented in parentheses.

<sup>d</sup>Percentage of markers (47,445 in total) that are mapped to a certain LG. Percentage of unique scaffolds (45,552 in total) is presented in parentheses.





**FIGURE 4** Scatter plot of the minor allele frequency and heterozygosity for the final SNP selection (50K, right red) in comparison to all screened SNPs (450K, grey) and all PolyHigh resolution SNPs (177K, dark red)

the range-wide provenance trial and the trees from the Swedish breeding population. The two clusters representing the two trials form a classical “horseshoe shape” (Figure 5) that is characteristic of samples where genetic similarity decays with geographical distance (Novembre & Stephens, 2008). The trees from two trials (Skogforsk and Hungary) showed a partly overlapping population structure even though the majority of the Skogforsk breeding population, which contains more samples with a Northern origin, occupy the right cluster while the Hungarian trial, which contains more samples with an Alpine or a central Europe origin, occupy the left cluster (left panel in Figure 5; Table 1). The patterns were clearer when looking at origins of all samples rather than to which trial they belonged. Samples in the right cluster had a northwest–northeast origin (with samples from Fennoscandia [FNE], Southern/Central Scandinavia [C\_Sc], Russian-baltic [Rus\_Bal] and Northern Poland [NPL]) while the left cluster had a more southwest–southeast origin (with samples from the Alpine region [ALP], central Europe [CEU] and Carpathians [ROM]). The four samples with unknown origin grouped in the middle of the FNE samples (right panel in Figure 5). Four of the documented ALP samples were positioned in between the two clusters, which might indicate a hybrid origin, and a small proportion of the samples did not group according to their documented origin, which might indicate sample mix-ups when the population trials were established and the sample origins were documented.

### 3.3 | Evaluation and validation of the 50K array

Twenty-eight Norway spruce haploid megagametophytes (Table S3), 48 samples from four full sib families consisting of the two parents and between 12 and 14 offspring and 160 samples from white, black and Sitka spruce (Table S4) were used for validation of the final 50K SNP array. Because this array was specifically designed for Norway spruce, joint genotype calling for all samples/species using the Axiom best

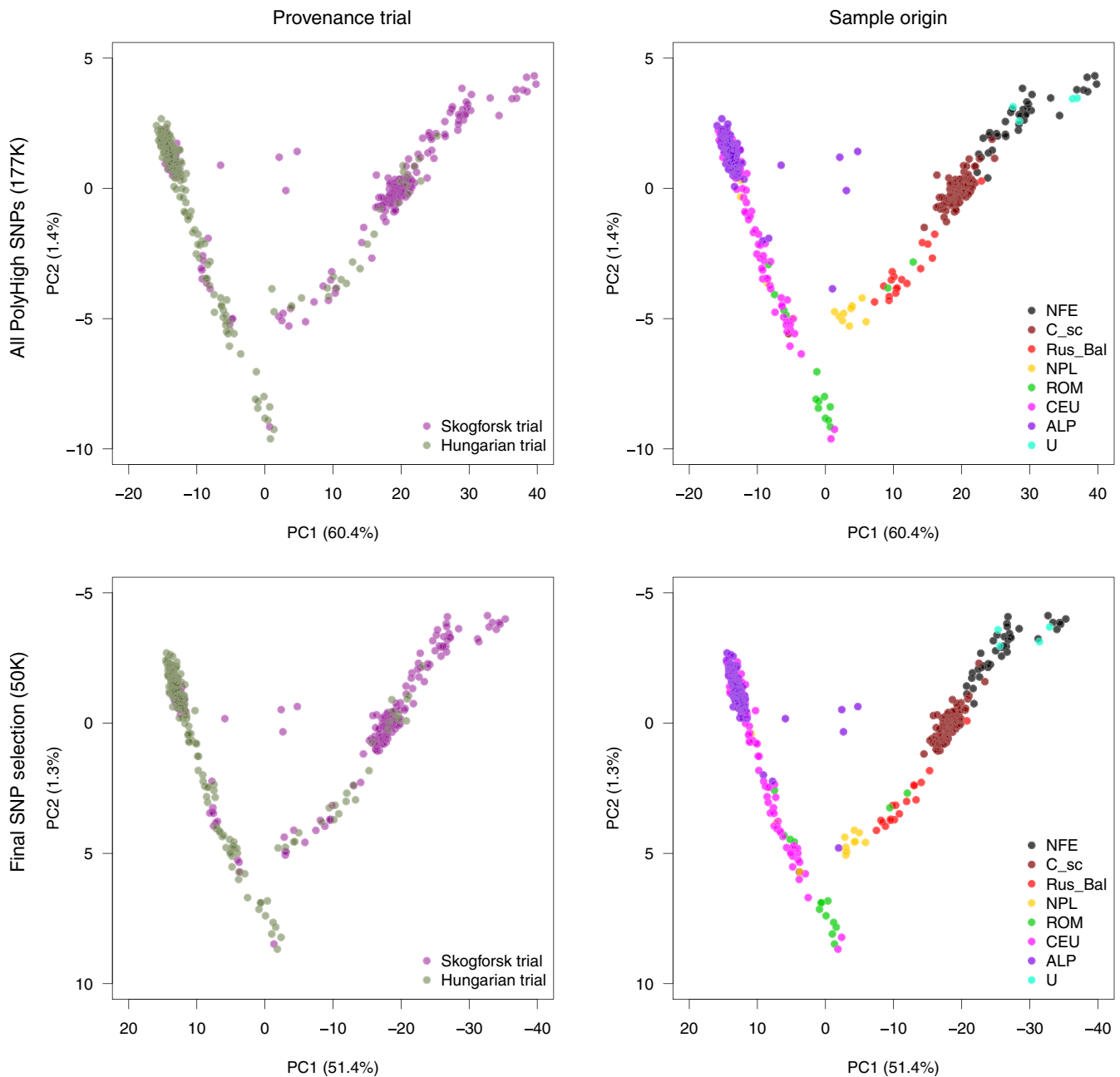
practice was not possible due to the variable probe performance in the three other species. Therefore, two independent genotyping calls were performed, one for all Norway spruce samples following the best practice in the Axiom analysis suite and a second run for other the spruce species which employed slightly lower sample QC values. A few samples, including four offspring, four haploid megagametophytes and one black spruce, were removed from the downstream analyses because they failed the sample QC. The overall performance of this array was then evaluated using sample and probe (SNP) call rate, probe specificities and MI error rates estimated from the remaining samples.

#### 3.3.1 | Sample and SNP call rate and probe specificity

The average sample call rate was 98.90% (minimum 97.67% and maximum 99.43%, Figure 6a). Out of the 47,445 probes, 45,541 (96%) were classified in the three high-confidence categories (PHR, MHR, NH) with an averaged call rate of 99.11% (minimum 85.77% and maximum 100.00% Figure 6b). The remaining 1,904 SNPs, classified as OTV or Other, were not recommended for reasons described above (Table S1). The averaged probe specificity, calculated as the proportion of samples with homozygous calls among 24 haploid megagametophytes, was 99.5% (Figure 6c; Table S5). The high specificity and call rate illustrate that the designed array is of high quality.

#### 3.3.2 | Mendelian inheritance (MI) error rate

Among 45,541 high-confidence probes, 6,438 were fixed for alternative alleles ( $P1 = AA$ ,  $P2 = aa$ ) in at least one family and 36,256 were fixed for the same allele ( $P1 = AA$ ,  $P2 = AA$ ) in at least one family. Unfortunately, those two sets of probes completely overlap with each other, resulting in 36,256 probes which could be evaluated for



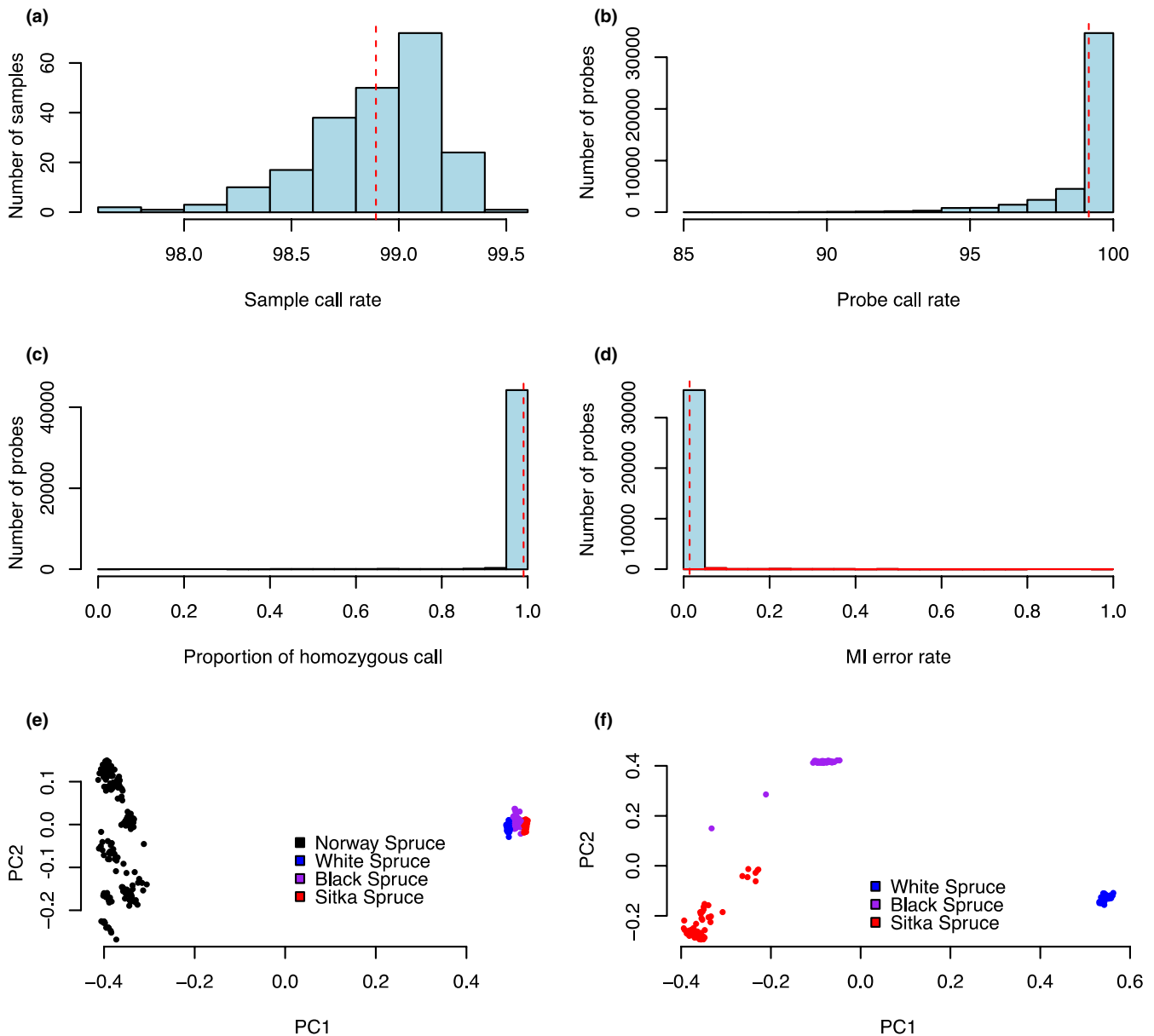
**FIGURE 5** Population structure estimated using a principal component analysis on the relatedness matrix calculated based on all 177K PolyHigh resolution SNPs (top row) and from the final 50K SNP selection (bottom row). Left-hand panels are coloured based on which provenance trial the samples originate from while the right-hand panels are coloured based on documented sample origin. Replicated samples have been removed from the analysis. NFE—Fennoscandia contains samples from Finland and northern Sweden; C-sc—Southern Scandinavia from Central/Southern Sweden and Central/Southern Norway; Rus\_Bal—Russian-Baltic from Russia, Belarus, Estonia, Latvia and Lithuania; NPL—Northern Poland; ROM—Carpathian from Romania and Bulgaria; CEU—Central Europe from Slovakia, Czech Republic, Southern Poland, Hungary and Austria; ALP—Alpine from Denmark, Germany, Switzerland, France and Italy; U—unknown

Mendelian segregation errors (see Materials and Methods). Overall, there were very low rates of Mendelian segregation errors, with 97.8% of the probes having MI error rates of <5% (Figure 6d).

After QC for probe call rate, specificity and MI error rate from samples of family trios and haploid megagametophytes, 1,645, 1,298 and 797 probes may not meet quality standards, yielding at least 42,598 (90%) high-quality probes on the array that are available for genotyping analyses with high confidence (Table S5).

### 3.3.3 | Array ascertainment bias

The MAF values of SNPs were divided into 25 bins (2% intervals) and the frequency distributions were compared between the 50K array and the full MAF distribution of the ~177K PHR SNPs. The results show that the final array captured on average 2.7% of the SNPs from each MAF bin with relatively even coverage from 2.2% to 2.9% except for MAF < 5% that was excluded intentionally



**FIGURE 6** Summary of the array evaluation metrics. (a) Histogram of the sample call rate for Norway spruce. The dashed red line indicates the averaged call rate. (b) Histogram of the probe call rate for Norway spruce. The dashed red line indicates the averaged call rate. (c) Histogram of the proportion of homozygous calls for 45,541 probes estimated using 24 haploid tissues. The dashed red line indicates the averaged proportion of homozygous calls. (d) Histogram of the Mendelian inheritance (MI) error rate for 36,256 probes estimated using 48 family trios. (e) Principal component analysis for all four spruce species. (f) Principal component analysis for the three non-Norway spruce species

when selecting SNPs from the ~177K PHR SNPs (Figure S1a). This indicates that there was no obvious bias in the selection of SNPs based on MAF.

### 3.3.4 | Comparison of genetic diversity between range-wide collection and breeding populations

When comparing the distribution of MAF and heterozygosity between the range-wide provenance trial and the Skogforsk breeding population, we noticed a slight enrichment of low-frequency alleles

in the provenance trial (mean MAF is 0.16 and 0.18 for the provenance trial and Skogforsk population, respectively; Figure S1b,c) and a slightly lower heterozygosity (0.23 for the provenance trial and 0.27 for the Skogforsk population; Figure S1d,e). In addition, there were 66 SNPs that were fixed in the provenance trial but which were all segregating in the breeding population. The array was designed based on variants segregating in a resequencing panel consisting of trees sampled from the Nordic countries, and the 66 nonvariable SNPs observed in the range-wide provenance population could therefore indicate a slight ascertainment bias in the SNPs included on the array.

### 3.3.5 | SNPs from intragenic and intergenic regions

We observed a minor, but statistically significant difference in both MAF (mean MAF is 0.169 and 0.176 for intergenic and intragenic SNPs, respectively;  $p = 1.0 \times 10^{-7}$  from *t* test) and heterozygosity (mean heterozygosity is 0.250 and 0.256 for intergenic and intragenic SNPs, respectively;  $p = 8.5 \times 10^{-9}$  from *t* test) in the screening data. However, these differences are only significant due to the large number of SNPs assessed and do not represent biologically significant differences. In line with this, the two sets of SNPs differ very little in the population structure they capture (Figure S1f–i).

### 3.3.6 | Transferability to other spruce species

Although the array was designed to target Norway spruce, half of the probes (23,797) were called with high confidence in three other spruce species (white, black and Sitka spruce). A PCA on all the samples clearly separated the four species into two major clusters (Figure 6e). As expected, the other three spruce species, which all belong to the North American clade of *Picea* (Clade II in Lockwood et al., 2013), were more genetically similar to each other than to Norway spruce. To evaluate whether these markers could be used to further distinguish the three North American species, a subsequent PCA with only the North American species was performed (Figure 6f). In this analysis, the three species were clearly separated into three major clusters with black spruce being closer to Sitka spruce than to white spruce, as expected, based on a published phylogeny for the genus *Picea* based on plastid, mitochondrial and nuclear sequences (Lockwood et al., 2013). These results demonstrate a potentially broader application of this array for more species within the same genus.

## 4 | DISCUSSION

Development of efficient genotyping resources for identifying alleles underlying local adaptation, trait variation and GS in conifers is a significant challenge due to their large and complex genomes (Neale & Wheeler, 2019). Dissection of the molecular basis of trait variation in forest trees began in the 1990s with the introduction of QTL mapping in controlled-cross pedigrees using random DNA markers (Neale & Kremer, 2011; Neale 2004; Strauss et al., 1992). Later, SNP markers from candidate genes were used to exploit population-wide LD to perform association mapping (AM). The AM approach was initially applied in *Eucalyptus* (Thumma et al., 2005) and has subsequently been used in many conifer tree species (Beaulieu et al., 2011; Dillon et al., 2010; Gonzalez-Martinez et al., 2007). However, neither QTL analysis using limited family pedigrees nor the candidate gene approach for AM resulted in the identification of useful markers for forest breeding. This is because QTLs were mapped with very large confidence intervals on chromosomes due to the limited number of markers used (Grattapaglia et al., 2018).

To increase the marker density for AM in conifer trees, access to a genome-wide SNP array would enable high-throughput and relatively cost-efficient genotyping. SNP arrays have already been developed for a number of spruce species and in other conifers based on transcriptome data (Howe et al., 2020; Perry et al., 2020; Plomion et al., 2016). However, transcriptome-based approaches, such as RNA sequencing, have thus far yielded relatively small arrays, covering <10,000 SNPs in most cases, and due to the nature of transcriptome data they also generally lack genomic information from intergenic regions (Bartholome et al., 2016; Pavy et al., 2013, 2016).

The Axiom 50K Norway spruce SNP genotyping array is a novel and efficient resource for population and quantitative genetics and for GS studies. The array contains known intragenic and intergenic SNPs that are evenly distributed across the Norway spruce genome. The three-step strategy we used, with probe development based on WGS samples, screening of a large number of preliminary SNPs using two large trials, a breeding population and a species-wide range collection, and final array evaluation using both haploid and within-family segregation analyses to assess SNP specificity and Mendelian segregation of SNPs proves that this array is highly efficient and robust.

In comparison to other genotyping techniques, such as WGS, genotyping-by-sequencing (GBS) and sequence capture, which are computationally and bioinformatically demanding and/or expensive to perform (Baisson et al., 2019; Pan et al., 2015; Wang et al., 2020), SNP arrays are less computationally demanding to analyse because the majority of the bioinformatics analyses were made when the chip was developed. GBS data often also include a large fraction of missing data which requires imputation and computational interpretation prior to subsequent analysis (Hussain et al., 2017). This makes our array very valuable for scientists and breeders with limited bioinformatic knowledge. The spruce genome, which is both very large (~19.6 Gb) and highly repetitive (~70% repeat content in scaffolds >1,000 bp), has made it difficult to develop a reference genome assembly of high quality. With only ~66% of the genome present in the currently available assembly (Nystedt et al., 2013), a large proportion of resequencing reads are redundant because they cannot be mapped to the assembly, which in practical terms increases the cost of sequencing per mapped base. However, there is also a risk that a proportion of the reads mapping to the reference would be misaligned if repetitive regions are collapsed in the assembly. This would increase the number of false variants in downstream analysis (Bernhardsson et al., 2020). This is another advantage of our Axiom 50K SNP genotyping array, as these risks were minimized by carefully selecting the probes to avoid such problematic genomic regions and subsequently evaluating the probe performance by specifically assessing probe specificity using haploid samples.

### 4.1 | Screening array design and performance

Resequencing data have not been employed for selection of SNPs for a genotyping array in any conifer species to date, but this practice

has been commonly used in many fruit trees and crops (Basil et al. 2015; Bianco et al. 2016; Singh et al. 2015; Marrano et al., 2019; Pandey et al. 2017; Roorkiwal et al. 2018; Wang et al. 2016) and is often combined with a large screening array (Montanari et al., 2019; Unterseer et al. 2014). Our screening array results indicate that prescreening of SNP aids in the design of a high-quality genotyping array in conifers. Although large parts of the current assembly suffer from collapsed genomic regions (Bernhardsson et al., 2020), we are able to select all 450K probes with the highest ThermoFisher's pConvert score category ( $0.6 \leq p\text{Convert} \leq 1.0$ ) from the ~3.76 million candidate SNPs which were obtained through filtering from the original >709 million SNPs. After using the screening array to genotype 480 trees, ~58% of the 450K screening probes yielded high-confidence SNPs that were recommended for inclusion on the final genotyping array by ThermoFisher (Table 2). In total, 39% of screening probes were also classified as PHR SNPs, making them high-quality candidates for the final array. With such a large number of PHR SNPs available to us, we were able to include only PHR SNPs on the final array.

## 4.2 | Genotyping array performance

We evaluated the 50K genotyping array for probe specificity (uniqueness on the genome), Mendelian segregation error and population structure between the genotyping array data and the full set of ~177K PHR SNPs.

Probe specificity is particularly important for conifer genomes which are known to harbour abundant paralogues, pseudogenes and repeats. The specificity of the 50K SNP array is 99.5%, indicating that the SNPs selected are highly reliable and that they target unique regions in the Norway spruce genome. The probability that a probe hybridizes to more than one region of the genome is thus very low, being about 0.5%. Benchmarking probe specificity with SNP arrays developed for conifers or forest trees is not possible as probe specificities have not been reported for other arrays.

The probes on the 50K array were evenly distributed throughout the Norway spruce genome and also evenly distributed between intra- and intergenic regions, offering a truly genome-wide coverage that will be highly valuable for several downstream applications. The final array validation also showed that the selected SNPs have low Mendelian inheritance (segregation) error rates, with 98% of the probes having MI error rates < 5%, similar to what was observed for the EUChip60k (Mendelian allelic inheritance concordance > 95%, Silva-Junior et al., 2015).

The final Axiom 50K array was as efficient as the full 177k PHR set in identifying true population structure in the 468 screening samples and it had a high precision in identifying the origin of four unknown samples (Figure 5). The Swedish breeding population was sampled from a total of 5,056 breeding trees. The population structure of these trees has previously been studied using 134,605 SNPs derived from ~40,000 sequence capture probes (Chen et al, unpublished data). The 50K genotyping array identified the same

population origin and structure for the 222 Swedish samples (e.g., seven geographical populations) that were obtained using either the 177K PHR SNP set or when using the large sequence capture SNP data set.

Across all 76 Norway spruce samples (28 megagametophytes and 48 family trio samples) genotyped using the 50K array for performance evaluation, as many as 45,000 SNPs (96%) were shown to belong to the three highest confidence categories (PHR, NH and MHR) with an average sample call rate of 98.9% and a SNP call rate of 99.11%. This is very high in comparison with results in Douglas-fir (88.2% sample call rate and 50.4% SNP call rate, Howe et al., 2020) and other tree SNP arrays which generally have failure rates on the order of 20% (Plomion et al., 2016). The sample and SNP call rates using our 50K array is comparable or even higher than the EUChip60K array data (average SNP call rate > 90% and sample call rates across all SNPs > 97%) even though our genome is about 30 times larger and substantially more complex than the *Eucalyptus* genome (Silva-Junior et al., 2015).

The reproducibility of a replicated sample is an important quality benchmark of array performance. The white spruce Infinium assays (PgAS1 of 13K SNPs and PgLM3 of 14K SNPs) estimated 99.5% and 99.9% reproducibility (Pavy et al., 2013) and the genotyping accuracy for duplicated trees in Douglas-fir was 99.3% (Howe et al., 2020). Our screening array of 450K SNPs of Norway spruce had a reproducibility of 99.8% for replicated samples across all SNPs and the selected 50K SNPs had 100% reproducibility among the replicated samples, similar to what was observed for the EUChip60K array (Silva-Junior et al., 2015).

## 4.3 | Array ascertainment bias

When designing an SNP array, the ascertainment procedures of the SNPs selected for inclusion on the array need to be carefully evaluated in future applications, such as population genetics and GWA studies. SNPs included on the array were selected to fulfil specific criteria, such as MAF, and therefore represents a biased subset compared to a random sample of SNPs. Such ascertainment bias causes systematic deviations of population genetic statistics from theoretical expectations and will inevitably be present when SNP array data are used for estimating population genetic parameters, such as genetic diversity, or when inferring population structure or the demographic history of a sample (Lachance & Tishkoff, 2013).

There are two kinds of ascertainment bias that need to be considered for SNP array data, depth and width. Ascertainment depth refers to the fact that only SNPs occurring with sufficient number in a sample population (e.g., minimum MAF) are included on the final array. Ascertainment width, on the other hand, is affected because markers are generally first identified in a small panel of individuals from part of the species' range. However, a comparison of MAF distributions between the 50K array and the full ~177K PHR SNPs revealed no significant bias in ascertainment SNP depth. When comparing the distribution of MAF and heterozygosity

between our range-wide provenance trial and Skogforsk breeding population samples, we noted a slight enrichment of low-frequency alleles and consequently a slightly lower heterozygosity in the provenance trial. However, as the 29 trees used to design the array all had a Nordic origin (Central/Southern Sweden and Fennoscandia), this probably reflects a slight bias in ascertainment width, as more alleles with a Northern origin were captured in the resequenced samples. This small bias may reflect the possible influence of hybridization between *Picea abies* with *P. obovata* in Fennoscandia (e.g., Tsuda et al., 2016). Hybridization between the two species is known to have influenced genetic diversity in Fennoscandian populations, with a gradient of increasing effects of hybridization closer to the Ural Mountains (Tsuda et al., 2016). The range of distribution of *P. abies* in Fennoscandia represents the most recent expansion of this species following the last glaciation. There is also evidence that central Fennoscandia could have slightly higher levels of genetic composition and diversity due to the meeting of the two expansion routes that colonized this region since the LGM (Lagercrantz & Ryman, 1990).

#### 4.4 | SNPs from intragenic and intergenic regions

SNPs from intergenic regions are important for detecting associations in GWAS and inclusion of a large number of SNPs from intergenic regions is expected to increase both GWAS power and the efficiency of GS. For evolutionary population genetic analyses, markers in intragenic and intergenic regions may generally differ in patterns of variation, selection signature and their effects on trait variation. Thousands of trait-associated SNPs have been identified in intergenic regions in humans, and half of the disease-associated SNPs in humans that thus far have been identified reside within intergenic regions (Li et al., 2016). SNPs in genic regions are also more likely to display signatures of both positive and negative selection than SNPs in nongenic regions, and intergenic SNPs are key components of the spatial and regulatory network for human growth (Coop et al. 2009; Helyar et al., 2011; Schierding et al., 2016).

It has also been shown that intergenic and intragenic regions behave differently in terms of population genetic summary statistics in Norway spruce (Wang et al., 2020) and intergenic regions appear to have a higher impact on adaptation in species with larger genomes (Mei et al., 2018). SNP arrays developed thus far in conifers have largely been based on candidate gene and/or transcriptome sequencing because markers on those SNP arrays are mainly situated in or close to genes they may not provide a representative view of genome-wide variation. In our array, we noted minor but statistically significant differences in both MAF and heterozygosity between intergenic and intragenic SNPs. This could indicate historical differences in the action of natural selection or the demographic history for different genomic regions in our screening populations. However, the two SNP sets differ very little in the pattern of population structure that they capture, suggesting that

such effects may be small. By combining both intergenic and intragenic SNPs on our genotyping array, we can therefore give a much clearer picture of the genomic landscape of variation in terms of population genetic variation, adaptation and possibly also phenotype associations.

#### 4.5 | Array transferability to other spruce species

The genus *Picea* consists of a total of 35 species (Farjón, 2001). We tested the transferability of the array to three other spruce species that are important in commercial plantation, breeding and production in the northern hemisphere, white spruce, black spruce and Sitka spruce. We found that about 50% of the SNPs (23,797) can be reliably transferred to the three species and genotyped with high confidence. This transferability is high and similar to the 57% transfer rate observed between white spruce and Norway spruce (e.g., 0.5 million probes derived from 23,684 genes of white spruce were mapped to 13 543 Norway spruce genes) by Azaiez et al. (2018). The transferability of our SNP array is higher than what was observed for a white spruce SNP array used to genotype Sitka spruce (22.4%), black spruce (17.6%) or Norway spruce (12.5%) (Pavy et al., 2013). Our array is also able to clearly separate Norway spruce (Clade I) from the more distantly related species from the North American clade (white, black and Sitka spruce from Clade II, Lockwood et al., 2013). *Picea obovata* and *P. omorika* are two species that are more closely related to *P. abies* (all in Clade I) than the three North America spruce species. Although these two species are not of great commercial importance, the latter species has been the focus of conservation efforts and the SNP array could therefore potentially be applied to perform more basic research in this species. However, we have not tested the conversion rates of the array for these two closely related species, but given the close relationship among these three species, we expect the array will have a high level of success rate when genotyping *P. obovata* and *P. omorika* samples.

The 50K SNP genomic resources presented and evaluated for Norway spruce in this study represent an unprecedented effort to deploy high-throughput SNP genotyping in conifers. The 50K SNP array is the largest genotyping chip ever produced for any spruce species and included SNPs from both intragenic and intergenic regions. We envisage that this array will make significant contributions to questions related to population genetics, comparative genomics, association genetics, genomic prediction and linkage mapping in Norway spruce as well as providing a template for designing future genotyping arrays in other spruce and conifer species.

#### ACKNOWLEDGEMENTS

This project was supported by the Swedish Foundation for Strategic Research (SSF) to H.X.W. (RBP14-0040) and Horizon2020 B4EST. The computation and data handling provided by the Swedish National Infrastructure for computing (SNIC) at Uppmax was partially funded by the Swedish Research Council (2016-07213). We would like to thank Tomas Funda, Lu Wang, Zhou Wei, Zuzana Binova and Linghua

Zhou for help with DNA extraction, and Bo Karlsson, Anders Fries, Éva Ujvari Jarmay, László Nagy, David Hall, Jingxiang Meng and Ruiqi Pian for their assistance in field sample collections. We also thank Fikret Isik for organizing and coordinating the Conifer SNP array Consortium.

#### AUTHOR CONTRIBUTIONS

H.X.W. designed and planned the project; C.B. conducted analyses for resequencing, pilot array and population structure. Y.Z. conducted SNP calling for pilot and genotyping array. Z.C. designed and sampled trees. P.K.I. designed the resequencing experiment. C.B., Y.Z. and H.X.W. wrote the manuscript.

Research interest: C.B. interested in bioinformatics; Y.Z. interested in bioinformatics; Z.C. interested in quantitative genetics; P.K.I. interested in population genetics; and H.X.W. interested in quantitative genetics and breeding.

#### DATA ACCESSIBILITY STATEMENT

Data from this project are archived in Figshare and accessible as: 1. Axiom 50K array for ~300 Norway spruce raw data and annotation file at <https://doi.org/10.6084/m9.figshare.12631358.v1>. 2. Axiom 450K SNP array for 480 Norway spruce raw data and Array annotation file at <https://doi.org/10.6084/m9.figshare.12630938.v1>.

#### ORCID

Pär K. Ingvarsson  <https://orcid.org/0000-0001-9225-7521>

Harry X. Wu  <https://orcid.org/0000-0002-7072-4704>

#### REFERENCES

- Affymetrix (2016). *Axiom analysis suite user guide (version 2.0)*. Affymetrix Inc.
- Azaiez, A., Pavy, N., Gérardi, S., Laroche, J., Boyle, B., Gagnon, F., Mottet, M.-J., Beaulieu, J., & Bousquet, J. (2018). A catalog of annotated high-confidence SNPs from exome capture and sequencing reveals highly polymorphic genes in Norway spruce (*Picea abies*). *BMC Genomics*, *19*, 942. <https://doi.org/10.1186/s12864-018-5247-z>
- Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., Poncet, C., Micheletti, D., Kerschbamer, E., Di Piero, E. A., Larger, S., Pindo, M., Van de Weg, E., Davassi, A., Laurens, F., Velasco, R., Durel, C.-E., & Troggio, M. (2016). Development and validation of the Axiom@Apple480K SNP genotyping array. *Plant J*, *86*, 62–74. <https://doi.org/10.1111/tpj.13145>
- Bassil, N. V., Davis, T. M., & Zhang, H. et al. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria × ananassa*. *BMC Genomics*, *16*, 155. <https://doi.org/10.1186/s12864-015-1310-1>
- Baison, J., Vidalis, A., Zhou, L., Chen, Z. Q., Li, Z., Sillanpää, M. J., Bernhardsson, C., Scofield, D., Forsberg, N., Grahn, T., Olsson, L., Karlsson, B., Wu, H., Ingvarsson, P. K., Lundqvist, S. O., Niittylä, T., & García-Gil, M. R. (2019). Genome-Wide Association Study (GWAS) identified candidate loci affecting wood formation in Norway spruce. *The Plant Journal*, *100*, 83–100. <https://doi.org/10.1111/tpj.14429>
- Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., & Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, *17*, 604. <https://doi.org/10.1186/s12864-016-2879-8>
- Beaulieu, J., Doerksen, T., Boyle, B., Clément, S., Deslauriers, M., Beauseigle, S., Blais, S., Poulin, P.-L., Lenz, P., Caron, S., Rigault, P., Bicho, P., Bousquet, J., & MacKay, J. (2011). Association Genetics of wood physical traits in the conifer white spruce and relationships with gene expression. *Genetics*, *188*, 197–214. <https://doi.org/10.1534/genetics.110.125781>
- Beaulieu, J., Doerksen, T., Clément, S., MacKay, J., & Bousquet, J. (2014). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity*, *113*, 343–352. <https://doi.org/10.1038/hdy.2014.36>
- Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D. G., Schifftaler, B., Baison, J., Street, N. R., García-Gil, M. R., & Ingvarsson, P. K. (2019). An Ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (*Picea abies*). *G3: Genes Genomes, Genetics*, *9*, 1623–1632.
- Bernhardsson, C., Wang, X., Eklöf, H., & Ingvarsson, P. K. (2020). Variant calling using NGS and sequence capture data for population and evolutionary genomic inferences in Norway Spruce (*Picea abies*). In I. Porth, & A. de la Torre (eds), *The Spruce Genome*. Compendium of Plant Genomes. Springer, Cham. [https://doi.org/10.1007/978-3-030-21001-4\\_2](https://doi.org/10.1007/978-3-030-21001-4_2)
- Chen, Z.-Q., Baison, J., Pan, J., Karlsson, B. O., Andersson, B., Westin, J., García-Gil, M. R., & Wu, H. X. (2018). Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. *BMC Genomics*, *19*, 946. <https://doi.org/10.1186/s12864-018-5256-y>
- Coop, G., Pickrell, J. K., & Novembre, J. et al. (2009). The role of geography in human adaptation. *PLoS Genetics*, *5*, e1000500.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491–498. <https://doi.org/10.1038/ng.806>
- Dillon, S. K., Nolan, M., Li, W., Bell, C., Wu, H. X., & Southerton, S. G. (2010). Allelic variation in cell wall candidate genes affects solid wood properties in natural populations and land races of radiata pine. *Genetics*, *185*, 1477–1487.
- Elfstrand, M., Zhou, L., Baison, J., Olson, Å., Lundén, K., Karlsson, B. O., Wu, H. X., Stenlid, J., & García-Gil, M. R. (2020). Genotypic variation in Norway spruce correlates to fungal communities in vegetative buds. *Molecular Ecology*, *29*, 199–213. <https://doi.org/10.1111/mec.15314>
- Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*, 250–255. <https://doi.org/10.3835/plantgenome2011.08.0024>
- FAO (2015). *The global forest resources assessment*. FAO. Retrieved from <http://www.fao.org/3/a-i4808e.pdf>
- FAO (2016). *State of the World's Forests 2016. Forests and agriculture: Land-use challenges and opportunities*. FAO.
- Farjón, A. (2001). *World checklist and bibliography of conifers*, (2nd ed.). Royal Botanic Gardens.
- García, R. A., Cabeza, M., Rahbek, C., & Araujo, M. B. (2014). Multiple dimensions of climate change and their implications for biodiversity. *Science*, *344*, 1247579. <https://doi.org/10.1126/science.1247579>
- Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., Gunter, L. E., Wymore, A. M., Grassa, C. J., Farzaneh, N., Porth, I., McKown, A. D., Skyba, O., Li, E., Fujita, M., Klápště, J., Martin, J., Schackwitz, W., Pennacchio, C., ... Tuskan, G. A. (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.
- González-Martínez, S. C., Wheeler, N. C., Ersoz, E., Nelson, C. D., & Neale, D. B. (2007). Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics*, 175, 399–409.
- Grattapaglia, D., Silva-Junior, O. B., Resende, R. T., Cappa, E. P., Müller, B. S. F., Tan, B., Isik, F., Ratcliffe, B., & El-Kassaby, Y. A. (2018). Quantitative genetics and genomics converge to accelerate forest tree breeding. *Frontiers in Plant Science*, 9, 1693.
- Hall, D., Hallingbäck, H. R., & Wu, H. X. (2016). Estimation of number and distribution of QTL effects in forest tree traits. *Tree Genetics & Genomes*, 12, 1–17.
- Helyar, S. J., Hemmer-hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, 11, 123–136. <https://doi.org/10.1111/j.1755-0998.2010.02943.x>
- Howe, G. T., Jayawickrama, K., Kolpak, S. E., Kling, J., Trappe, M., Hipkins, V., Ye, T., Guida, S., Cronn, R., Cushman, S. A., & McEvoy, S. (2020). An Axiom SNP genotyping array for Douglas-fir. *BMC Genomics*, 21, 9. <https://doi.org/10.1186/s12864-019-6383-9>
- Hussain, W., Baenziger, P. S., Belamkar, V., Guttieri, M. J., Venegas, J. P., Easterly, A., Sallam, A., & Poland, J. (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. <https://doi.org/10.1038/s41598-017-16006-z>
- Isik, F. (2014). Genomic selection in forest tree breeding: The concept and an outlook to the future. *New Forests*, 45(3), 379–401. <https://doi.org/10.1007/s11056-014-9422-z>
- 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. <https://doi.org/10.1007/s11295-019-1377-y>
- Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, 35, 780–786. <https://doi.org/10.1002/bies.201300014>
- Lagercrantz, U., & Ryman, N. (1990). Genetic structure of Norway Spruce (*Picea abies*): Concordance of morphological and allozymic variation. *Evolution*, 44, 38–53.
- Li, H., Achour, I., Bastarache, L., Berghout, J., Gardeux, V., Li, J., Lee, Y., Pesce, L., Yang, X., Ramos, K. S., Foster, I., Denny, J. C., Moore, J. H., & Lussier, Y. A. (2016). Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. *NPJ Genomic Medicine*, 1, 16006. <https://doi.org/10.1038/npjgenmed.2016.6>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lockwood, J. D., Aleksić, J. M., Zou, J., Wang, J., Liu, J., & Renner, S. S. (2013). A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Molecular Phylogenetics and Evolution*, 69, 717–727. <https://doi.org/10.1016/j.ympev.2013.07.004>
- Marrano, A., Martínez-García, P. J., Bianco, L., Sideli, G. M., Di Pierro, E. A., Leslie, C. A., Stevens, K. A., Crepeau, M. W., Troglio, M., Langley, C. H., & Neale, D. B. (2019). A new genomic tool for walnut (*Juglans regia* L.): Development and validation of the high-density Axiom™ *J. regia* 700K SNP genotyping array. *Plant Biotechnology Journal*, 17, 1027–1036. <https://doi.org/10.1111/pbi.13034>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, 17, 656–669. <https://doi.org/10.1111/1755-0998.12613>
- Mei, W., Stetter, M. G., Gates, D. J., Stitzer, M. C., & Ross-Ibarra, J. (2018). Adaptation in plant genomes: Bigger is different. *American Journal of Botany*, 105, 16–19. <https://doi.org/10.1002/ajb2.1002>
- Montanari, S., Bianco, L., Allen, B. J., Martínez-García, P. J., Bassil, N. V., Postman, J., Knäbel, M., Kitson, B., Deng, C. H., Chagné, D., Crepeau, M. W., Langley, C. H., Evans, K., Dhingra, A., Troglio, M., & Neale, D. B. (2019). Development of a highly efficient Axiom™ 70 K SNP array for *Pyrus* and evaluation for high-density mapping and germplasm characterization. *BMC Genomics*, 20, 331. <https://doi.org/10.1186/s12864-019-5712-3>
- Neale, D. B., & Kremer, A. (2011). Forest tree genomics: Growing resources and applications. *Nature Review Genetics*, 12, 111–122. <https://doi.org/10.1038/nrg2931>
- Neale, D. B., McGuire, P. E., Wheeler, N. C., Stevens, K. A., Crepeau, M. W., Cardeno, C., Zimin, A. V., Puiu, D., Perteau, G. M., Sezen, U. U., Casola, C., Koralewski, T. E., Paul, R., Gonzalez-Ibeas, D., Zaman, S., Cronn, R., Yandell, M., Holt, C., Langley, C. H., ... Wegrzyn, J. L. (2017). The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. G3: *Genes Genomes, Genetics*, 7, 3157–3167. <https://doi.org/10.1534/g3.117.300078>
- Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends in Plant Science*, 9(7), 325–330. <https://doi.org/10.1016/j.tplants.2004.05.006>
- Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., Cardeno, C., Koriabine, M., Holtz-Morris, A. E., Liechty, J. D., Martínez-García, P. J., Vasquez-Gross, H. A., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marçais, G., ... Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15, R59. <https://doi.org/10.1186/gb-2014-15-3-r59>
- Neale, D. B., & Wheeler, N. C. (2019). *The conifers: Genomes, variation and evolution*. Springer International Publishing.
- Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40, 646–649. <https://doi.org/10.1038/ng.139>
- Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497, 579–584. <https://doi.org/10.1038/nature12211>
- Pan, J., Wang, B., Pei, Z.-Y., Zhao, W., Gao, J., Mao, J.-F., & Wang, X.-R. (2015). Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Molecular Ecology Resources*, 15, 711–722. <https://doi.org/10.1111/1755-0998.12342>
- Pandey, M., Agarwal, G., & Kale, S. et al. (2017). Development and Evaluation of a High Density Genotyping 'Axiom\_Arachis' Array with 58K SNPs for Accelerating Genetics and Breeding in Groundnut. *Sci Rep*, 7, 40577. <https://doi.org/10.1038/srep40577>
- Pavy, N., Gagnon, F., Deschênes, A., Boyle, B., Beaulieu, J., & Bousquet, J. (2016). Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: An example from black spruce (*Picea mariana*). *Molecular Ecology Resources*, 16, 588–598.



- Pavy, N., Gagnon, F., Rigault, P., Blais, S., Deschênes, A., Boyle, B., Pelgas, B., Deslauriers, M., Clément, S., Lavigne, P., Lamothe, M., Cooke, J. E., Jaramillo-Correa, J. P., Beaulieu, J., Isabel, N., Mackay, J., & Bousquet, J. (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and Nordic congeneric taxa. *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. <https://doi.org/10.1111/1755-0998.13223>
- Plomion, C., Bartholomé, J., Lesur, I., Boury, C., Rodríguez-Quilón, I., Lagrault, H., Ehrenmann, F., Bouffier, L., Gion, J. M., Grivet, D., de Miguel, M., de María, N., Cervera, M. T., Bagnoli, F., Isik, F., Vendramin, G. G., & González-Martínez, S. C. (2016). High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Molecular Ecology Resources*, *16*, 574–587.
- R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Resende, M. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*, *190*, 1503–1510. <https://doi.org/10.1534/genetics.111.137026>
- Roorkiwal, M., Jain, A., Kale, S. M., Doddamani, D., Chitkani, A., Thudi, M., & Varshney, R. K. (2018). Development and evaluation of high-density Axiom® CicerSNP Array for high-resolution genetic mapping and breeding applications in chickpea. *Plant Biotechnol J*, *16*, 890–901. <https://doi.org/10.1111/pbi.12836>
- 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. <https://doi.org/10.1093/hmg/ddw165>
- 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. *New Phytologist*, *206*, 1527–1540.
- Singh, N. et al. (2015). Single-copy gene based 50K SNP chip for genetic studies and molecular breeding in rice. *Sci. Rep.*, *5*, 11600. <https://doi.org/10.1038/srep11600>
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M., Holtz-Morris, A. E., Martínez-García, P. J., Sezen, U. U., Marçais, G., Jermstad, K., McGuire, P. E., Loopstra, C. A., Davis, J. M., Eckert, A., de Jong, P., ... Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, *204*, 1613–1626. <https://doi.org/10.1534/genetics.116.193227>
- Strauss, S. H., Lande, L., & Namkoong, G. (1992). Limitations of molecular-marker-aided selection in forest tree breeding. *Canadian Journal of Forest Research*, *22*, 1050–1061.
- Thavamani Kumar, 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. <https://doi.org/10.1007/s11295-013-0594-z>
- Thumma, B. R., Nolan, M. F., Evans, R., & Moran, G. F. (2005). Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp. *Genetics*, *171*, 1257–1265.
- Tsuda, Y., Chen, J., Stocks, M., Källman, T., Sønstebo, J. H., Parducci, L., Semerikov, V., Sperisen, C., Politov, D., Ronkainen, T., Välijanta, M., Vendramin, G. G., Tollefsrud, M. M., & Lascoux, M. (2016). The extent and meaning of hybridization and introgression between Siberian spruce (*Picea obovata*) and Norway spruce (*Picea abies*): Cryptic refugia as stepping stones to the west? *Molecular Ecology*, *25*, 2773–2789.
- Unterseer, S., Bauer, E., & Haberer, G. et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, *15*, 823. <https://doi.org/10.1186/1471-2164-15-823>
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1-11.10.33.
- Vidalis, A., Scofield, D. G., Neves, L. G., Bernhardtsson, C., García-Gil, M. R., & Ingvarsson, P. (2018). Design and evaluation of a large sequence-capture probe set and associated SNPs for diploid and haploid samples of Norway spruce (*Picea abies*). *bioRxiv*, <https://doi.org/10.1101/291716>
- Wang, X., Bernhardtsson, C., & Ingvarsson, P. K. (2020). Demography and natural selection have shaped genetic variation in the widely distributed conifer Norway spruce (*Picea abies*). *Genome Biology and Evolution*, *12*, 3803–3817. <https://doi.org/10.1093/gbe/evaa005>
- Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., & Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci Rep*, *6*, 20728. <https://doi.org/10.1038/srep20728>
- Warren, R. L., Keeling, C. I., Yuen, M. M., Raymond, A., Taylor, G. A., Vandervalk, B. P., Mohamadi, H., Paulino, D., Chiu, R., Jackman, S. D., Robertson, G., Yang, C., Boyle, B., Hoffmann, M., Weigel, D., Nelson, D. R., Ritland, C., Isabel, N., Jaquish, B., ... Bohlmann, J. (2015). Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *The Plant Journal*, *83*, 189–212.
- Wu, H. X., Hallingbäck, H. R., & Sánchez, L. (2016). Performance of seven tree breeding strategies under conditions of inbreeding depression. *Gene, Genome and Genetics*, *6*, 529–540. <https://doi.org/10.1534/g3.115.025767>
- Zapata-Valenzuela, J., Whetten, R. W., Neale, D., McKeand, S., & Isik, F. (2013). Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3*, *3*, 909–916. <https://doi.org/10.1534/g3.113.005975>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Bernhardtsson C, Zan Y, Chen Z, Ingvarsson PK, Wu HX. 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. *Mol Ecol Resour.* 2021;21:880–896. <https://doi.org/10.1111/1755-0998.13292>