Journal: Molecular Ecology Resources

Title: A genome-wide SNP genotyping resource for tropical pine tree species

Running title: A genotyping resource for tropical pine trees

Authors:

Colin Jackson[1]*, Nanette Christie[2]*, Melissa Reynolds[2], Christopher Marais[2], Yokateme Tii-kuzu[2], Madison Caballero[3], Tamanique Kampman[2], Erik A. Visser[2], Sanushka Naidoo[2], Dominic Kain[4], Ross W. Whetten[1], Fikret Isik[1], Jill Wegrzyn[3], Gary Hodge[1], Juan Jose Acosta[1#], Alexander A. Myburg[2#]

1. Department of Forestry & Environmental Resources, North Carolina State University, Raleigh, NC, USA. 2. Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. 3. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. 4. HQPlantations Pty Ltd, North Lakes, LD, 4509, Australia.

* These authors contributed equally. # Co-corresponding authors

Corresponding authors:

Juan Jose Acosta – jjacosta@ncsu.edu

Alexander A. Myburg – zander.myburg@fabi.up.ac.za

Abstract:

We performed gene and genome targeted SNP discovery towards the development of a genome-wide, multi-species genotyping array for tropical pines. Pooled RNA-seq data from shoots of seedlings from five tropical pine species was used to identify transcript-based SNPs resulting in 1.3 million candidate Affymetrix SNP probe sets. In addition, we used a custom 40K probe set to perform capture-seq in pooled DNA from 81 provenances representing the natural ranges of six tropical pine species in Mexico and Central America resulting in 563K candidate SNP probe sets. Altogether, 300K RNA-seq (72%) and 120K capture-seq (28%) derived SNP probe sets were tiled on a 420K screening array that was used to genotype 576 trees representing the 81 provenances and commercial breeding material. Based on the screening array results, 50K SNPs were selected for commercial SNP array production (Axiom 384 format, Thermo Fisher Scientific) including 20K polymorphic SNPs for *P. patula, P. tecunumanii, P. oocarpa* and *P. caribaea*, 15K for *P. greggii* and *P. maximinoi*, 13K for *P. elliottii* and 8K for *P. pseudostrobus*. We included 9.7K ancestry informative SNPs that will be valuable for species and hybrid discrimination. Of the 50K SNP markers, 25% are polymorphic in only one species, while 75% are shared by two or more species. The Pitro50K SNP chip will be useful for population genomics and molecular breeding in this group of pine species that, together with their hybrids, represent the majority of fast-growing tropical and subtropical pine plantations globally.

Keywords (four to six):

Tropical pines, SNP, genotyping array, Pitro50K, molecular breeding

1 Introduction

The genus *Pinus* L. contains more than 100 species that are almost exclusively distributed throughout the northern hemisphere (Price, Liston, & Strauss, 1998). Across this large geographic range, there are two primary pockets of biodiversity, one in China, and another in Mexico and Central America. Among the many species found in Mexico and Central America, a considerable number are considered to be threatened or endangered in their natural habitat. Conservation efforts made to address the threat to these species led to the establishment of *ex situ* conservation parks and commercial landrace populations across South America and southern Africa (Dvorak, Gutierrez, et al., 2000). Eight of these tropical and subtropical pine species (*P. oocarpa, P. patula, P. tecunumanii, P. greggii, P. caribaea, P. elliottii, P. maximinoi* and *P. pseudostrobus*) were included in this study due to their commercial and ecological importance. Furthermore, *P. tecunumanii* and *P. oocarpa* are valued for their wood properties, *P. caribaea* and *P. maximinoi* for their rapid growth, and *P. patula* and *P. greggii* for their frost tolerance and ability to be planted at higher elevations (Dvorak, Gutierrez, et al., 2000). The ability to deploy these species in a variety of commercial settings, coupled with their ability to readily hybridize and produce viable offspring, has generated interest from breeders aiming to develop hybrid breeding populations for specific niche environments (Gwaze, 1999; Hongwane et al., 2018).

In its current state, while effective, traditional tree improvement for *Pinus* species is a slow, laborious and expensive process. Much of the financial and labor costs associated with traditional breeding are tied up in progeny test establishment, maintenance, and measurement where accurate selections on mature expressed traits take up to six years to be made (McKeand,

1988, Isik and McKeand 2019). The significant time and cost associated with these traditional approaches have shifted the focus of the tree breeding community to the implementation of genome-assisted breeding technologies. Genomic characterization of pine trees has long been an area of interest that has evolved as new technologies have become available, starting with isozyme technologies shifting to random amplified polymorphic DNA (RAPD) and simple sequence repeat (SSR) marker technology (Conkle, 1979; Devey, Beil, Smith, Neale, & Moran, 1996; Devey, Fiddler, Liu, Knapp, & Neale, 1994; Echt & May-Marquardt, 1997; Rudin & Ekberg, 1978).

More recently, next generation sequencing (NGS) methods have revolutionized the field of genetics by facilitating the development of high-throughput, low-cost genotyping assays. Single nucleotide polymorphisms (SNPs) have become the DNA marker of choice for animal and plant genomic studies due to their prevalence throughout the genome, low cost of genotyping and being codominant in nature. Prior to NGS technologies becoming available in pines, SNPs were primarily discovered and validated through expressed sequence tag (EST) libraries (Chancerel et al., 2011; Dantec et al., 2004; Eckert et al., 2009). Now, much of the SNP discovery has shifted to reduced representation sequencing approaches such as RNA sequencing (RNA-seq) and targeted capture sequencing including exome capture (Durán et al., 2019; Liu et al., 2016; Lu et al., 2016; Neves, Davis, Barbazuk, & Kirst, 2013; Suren et al., 2016; Telfer et al., 2019). This focus on reduced representation methods stems from the massive size (> 20 Gbp) and repetitive nature of pine genomes, which makes whole genome sequencing costly and computationally inefficient for SNP discovery. *Pinus taeda* L., for example, is one of the best characterized pine genomes, but still comprises over 2.8 million contigs (Zimin et al., 2017) with 80% estimated repetitive content

(Wegrzyn et al., 2014). Due to these challenges, the use of molecular marker technology in pine tree breeding is still in its infancy compared to livestock and other commercial crop species (Isik, 2014).

High density arrays of over 100K markers and medium density arrays of over 10K SNPs have been developed and successfully deployed for numerous food crop species including maize, wheat and soybean (Ganal et al., 2011; Song et al., 2013; Wang et al., 2014). To date, SNP arrays have also been developed for a number of forest trees such as *Populus trichocarpa*, *Pseudostrobus menziesii*, *Picea abies L.,* and *Eucalyptus sp.* (Azaiez et al., 2018; Geraldes et al., 2013; Grattapaglia et al., 2011; Howe et al., 2020; Silva-Junior, Faria, & Grattapaglia, 2015). Limited SNP array resources have been developed for pine trees, with arrays being successfully designed for *P. pinaster* (Chancerel et al., 2013; Plomion et al., 2016), *P. monticola* Douglas ex D. Don (Liu, Sniezko, Sturrock, & Chen, 2014), *Pinus taeda* L. (Caballero et al., 2021) and four species of European pines (Perry et al., 2020). The lack of low-cost, high-throughput, and reproducible genotyping resources has created a bottleneck in the advancement of pine tree genetics research. There is now opportunity to greatly advance research and commercial breeding programs using new SNP genotyping technologies. We aimed to fulfil this need by creating a multi-species genotyping array that is informative across eight species of tropical pines, their hybrids, and other closely related species in the *Pinus* section Trifoliae subsections Australes and Ponderosae. We report the development of the Pitro50K SNP array with 49,674 markers that provide ~15K usable markers for each of the targeted pine species and their hybrids. The validated SNP set will provide a powerful high-throughput genotyping tool for tree breeders and researchers across the globe.

2 Materials and Methods

2.1 Biological Material and Sequencing

RNA sequencing data was generated for *P. patula* and *P. tecunumanii* in a previous study (Visser, Wegrzyn, Myburg, & Naidoo, 2018) and subsequently also for *P. greggii*, *P. maximinoi* and *P. oocarpa* (Kampman et al., unpublished). The RNA-seq libraries were generated from juvenile shoot tissue of 4-6 month old seedlings, from each species inoculated and mock-inoculated with *Fusarium circinatum* (isolate FCC3579) and harvested at three days and seven days post inoculation (dpi). RNA sequencing for *P. patula* and *P. tecunumanii* was performed using Illumina HiSeq2500 as described in Visser et al. (2018) using 300 bp insert libraries and PE125 or PE150 reads. RNA sequencing for *P. oocarpa*, *P. maximinoi* and *P. greggii* was similarly performed using 300 bp insert libraries and PE150 reads for all samples.

Targeted DNA capture sequencing (capture-seq) was performed on a total of 81 pooled samples from the five species mentioned above plus *P. caribaea* (**Figure 1**). We distinguished between subdivisions for three species: *P. greggii* (North and South), *P. tecunumanii* (high and low elevation) and *P. patula* (var. *patula* and var. *longipedunculata*). Each pooled sample contained four to eight trees selected from different families representing a single provenance for that species (**Supplementary Table 1**). Trees selected for sampling originated from first generation testing material (i.e. seed collected from wild populations) established in Camcore (NC State University, Raleigh NC) field trials across South Africa. A total of 567 trees were sampled from 81 provenances, selected to cover the natural range of the six species in Mexico and Central America (**Figure 1b, Supplementary Table 1**).

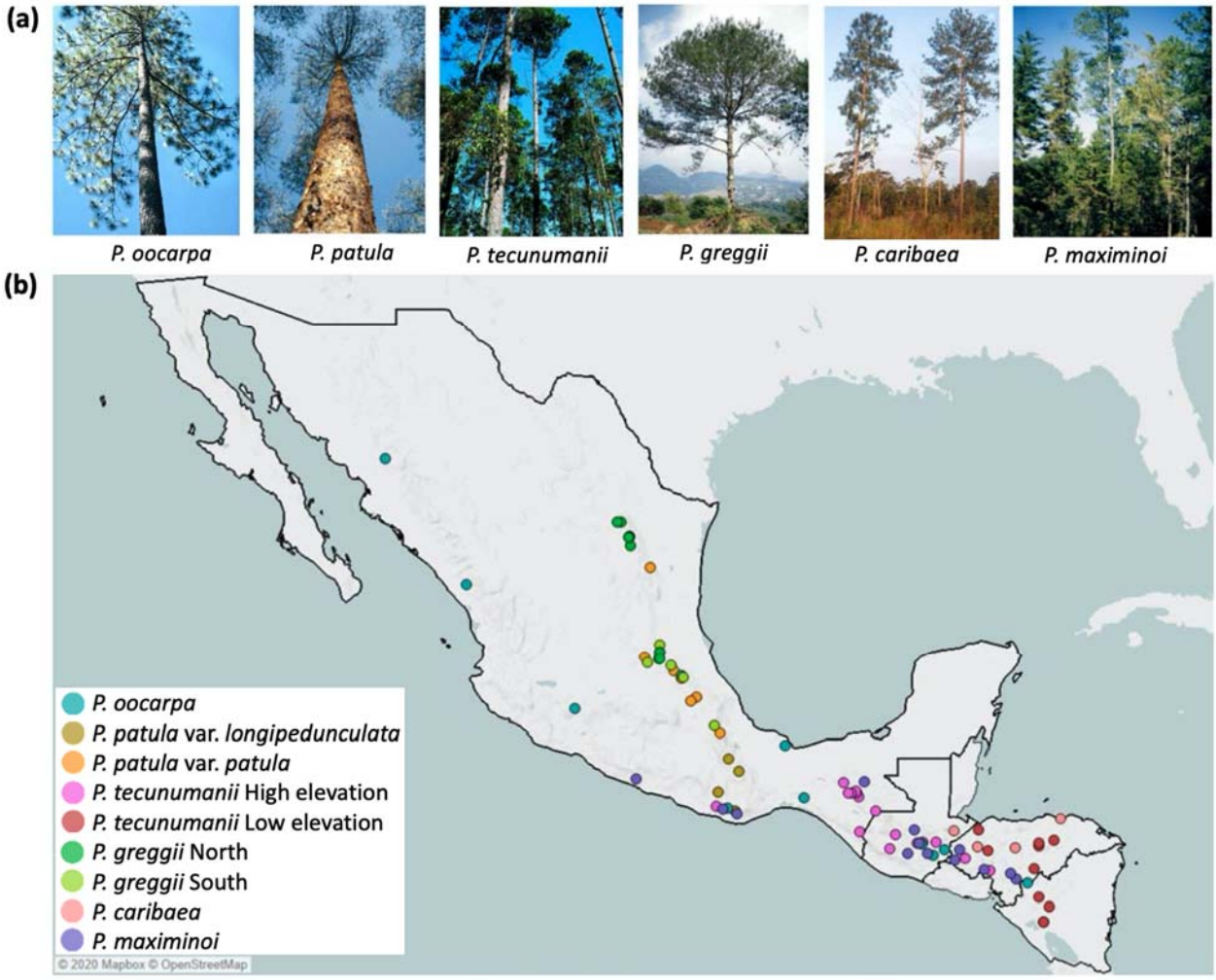

Figure 1. The natural range of tropical pine tree species in Mexico and Central America. (a) *Pinus oocarpa*, *P. patula*, *P. tecunumanii*, *P. greggii* and *P. caribaea* are members of the Australes subsection, whereas *P. maximinoi* belongs to the Ponderosae subsection. (b) The geographic location of the 81 provenances in Mexican and Central America representing the six tropical pine species that were sampled for targeted DNA sequencing. Samples from these provenances were later also genotyped using the screening array, in order to perform marker validation and selection of 50K SNPs for commercial array development.

Genomic DNA was extracted by the Forest Molecular Genetics (FMG) Programme's DNA Marker Platform (University of Pretoria, South Africa). DNA was isolated from 50 mg fresh needle tissue using the NucleoSpin® Plant II DNA extraction kit (Machery-Nagel, Germany) according to the manufacturer's specifications. One hundred nanograms of DNA per sample was submitted to

RAPiD Genomics (Gainesville, FL, USA) for capture-seq. DNA was mechanically sheared to an average size of 400 bp using a Covaris E210 focused ultrasonicator (Covaris, Woburn, MA). Libraries were constructed by repairing the ends of the sheared fragments using the End-It DNA End-Repair kit (Epicentre Biotechnologies, Madison, WI), producing blunt end fragments. Ligation of a single adenine residue to the 3' end of the blunt end fragment was performed using 15-U Klenow fragment (New England Biolabs Inc., Ipswich, MA) and deoxyadenosine triphosphate (dATP) (Promega, Madison, WI). Barcoded adapters that are suited for Illumina Sequencing were ligated to the libraries and the ligated fragments were PCR-amplified using standard cycling protocols (Mamanova et al., 2010).

A custom set of 40K target capture probes were developed by RAPiD Genomics to facilitate capture-seq for SNP discovery. Of these 40K probes, 30K were designed from single copy regions of the v2.01 *P. taeda* genome assembly (Zimin et al., 2017) and 10K were designed from the *P. patula* and *P. tecunumanii* transcriptome assemblies (Visser et al., 2018; Visser, Wegrzyn, Steenkmap, Myburg, & Naidoo, 2015). In total, 16 barcoded libraries were pooled for hybridization to the target capture probes. Enrichment was performed using the Select XT Target Enrichment System by Agilent (Palo Alto, CA, USA). Post enrichment, samples were amplified for an additional 6-14 cycles. All samples were sequenced using Illumina HiSeq 3000 with paired-end 150 bp reads. De-multiplexing was then performed using Illumina's BCLtofastq.

2.2 Variant Calling and Filtering

Paired-end RNA-seq and capture-seq data were subjected to variant calling via custom bioinformatics pipelines (**Figure 2a, b**). Both datasets underwent quality control and were

trimmed, with a minimum base quality score of 30 and read length of 50 required, using Sickle v1.33 (Johsi and Fass 2011). Trimmed reads originating from separate RNA-seq libraries (per species) were aligned back to the respective transcriptome assemblies using Burrows-Wheeler Aligner's (BWA) v0.7.15 (Li & Durbin, 2009) mem routine under default parameters. Trimmed reads originating from capture-seq were aligned using the same BWA method, but to a modified version of the v2.01 *P. taeda* reference genome assembly ("tropical pine reference genome"; modification procedures outlined in **Supplementary Method 1**). Variant calling for RNA-seq and capture-seq datasets were performed using Freebayes (Garrison & Marth, 2012) v1.1.0 and v1.20-2g29c4002, respectively. Within the RNA-seq dataset, variants were called jointly by species using each species' respective transcriptome assembly as reference under the following parameters: min-mapping-quality = 0, min-base-quality = 15, min-alternate-fraction = 0.05, min-alternate-count = 3, min-coverage = 10, and pooled continuous. Capture-seq reads that did not align within 500 bp of a capture probe location were removed prior to variant calling, as were reads that did not map in proper pairs using Samtools v1.7 (Li et al., 2009). Variants were called jointly using all 81 alignment files as input under the following parameters: min-alternative-fraction = 0.01, min-alternative-count = 2, min-coverage = 8, min-mapping-quality = 1, min-base-quality = 20, report-genotype-likelihood-max, pooled-discrete, and cnv_map. Output from both datasets were subsequently filtered to only include putative SNPs using vcflib v1.0.0-rcl (Garrison, 2016). Capture-seq derived SNPs were subjected to two separate additional filtering processes based on coverage and alternative allele fraction thresholds (**Supplementary Method 2**).

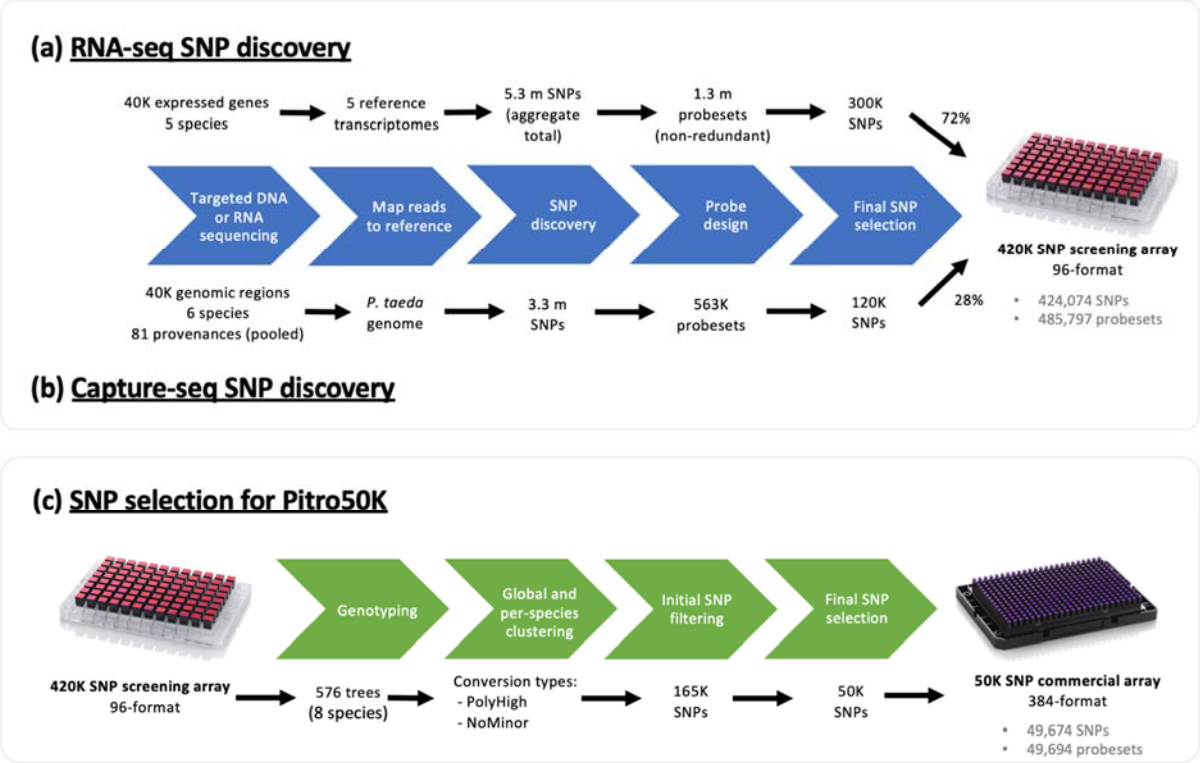

Figure 2. Gene and genome targeted SNP discovery towards the development of a commercial array for tropical pines. (a) RNA sequencing (RNA-seq) data of *P. oocarpa*, *P. patula*, *P. tecunumanii*, *P. greggii* and *P. maximinoi* and (b) targeted DNA capture sequencing (capture-seq) data of *P. oocarpa*, *P. patula*, *P. tecunumanii*, *P. greggii*, *P. caribaea* and *P. maximinoi* was subjected to SNP discovery and probe design. A total of 300K RNA-seq (72%) and 120K capture-seq (28%) derived SNPs were tiled on a 420K screening array (Axiom 96-format). (c) Based on the screening array results, after genotyping 576 trees from eight tropical pine species, 50K SNPs were selected for development of a commercial SNP array for routine, high-throughput, low cost, genome-wide profiling of pine trees.

2.3 Probe Design and Marker Selection for Screening Array

Probe design was performed by extracting 35 bp of sequence from the reference genome from either side of each SNP. However, probe sets were only retained if they were free of any secondary SNPs within the 35 bp window and the 71-nt probe sequence did not occur anywhere else in the genome. RNA-seq and capture-seq derived candidate SNPs that passed technical specifications for use on the Axiom array, together with their extracted probe sequences were submitted to Thermo Fisher for *in silico* scoring. Scoring of the probe sets was performed against

all reference transcriptomes as well as the “tropical pine reference genome”. Probe sets were scored as “recommended”, “at best neutral”, or “not recommended” by Thermo Fisher based on their internal p-convert metric coupled with other quality control metrics. In order for a marker to be considered “recommended”, it needed a p-convert value larger than six, no adjacent SNP markers within 24 bases and a non-repetitive probe sequence. A marker was considered “not recommended” if the probe sequence was duplicated or not unique within the reference, if the marker had a p-convert value smaller than 0.4, if the marker had an adjacent marker within 21 bases, or if three or more adjacent markers were within 24 bases of the target SNP. All other markers were considered neutral. Selection for inclusion on the screening array emphasized probe sets that scored recommended or neutral against the reference assemblies, along with high coverage across scaffolds and favorable depth, alternative allele fraction and homology count metrics (**Supplementary Method 3**). Additionally, A/T and G/C SNPs were excluded due to the additional space requirements needed to tile and genotype these markers (two probe sets per SNP). After initial selection of candidate SNPs and probe sets (424,074 SNP markers), the remaining space on the screening array was filled with opposite strand probe sets for 61,192 transcriptome derived SNPs to increase the chance of developing successful SNP assays for these markers. Note that from this point forward, the term ‘SNP’ will be used to denote a unique SNP probe set.

2.4 Marker Validation on Screening Array

The screening array was used to genotype a total of 576 tropical pine tree samples from eight species*: P. oocarpa* (12%), *P. patula* (13%), *P. tecunumanii* (26%), *P. greggii* (11%), *P.*

caribaea (11%), *P. elliottii* (6%) and hybrid (*P. elliottii x P. caribaea*; 3%) from the Australes subsection as well as *P. maximinoi* (13%) and *P. pseudostrobus* (3%) from the Ponderosae subsection. Within these, 28 megagametophyte samples from six species were included in the study for the SNP screening phase of the project (**Table 1**). The Axiom Analysis Suite (V3.1 User Guide) was used to perform global genotype clustering (per SNP, across all samples) and separate per-species clustering (per SNP, across the samples from a specific species) to guide genotype calling. Clustering runs included all samples where the best practice sample quality control criteria (recommended by Thermo Fisher) applied. Recommended cut-offs included a dish quality control (dQC) greater than or equal to 0.82 and a call rate greater than 97% across all SNPs for a sample to qualify for further use. For two of the single species *P. greggii* and *P. caribaea* runs, the quality control criteria was modified to include more samples in the genotype calling process, i.e. samples were subjected to a sample call rate greater than or equal to 84% and 90%, respectively.

Samples that passed QC filtering were subjected to genotype calling and conversion classification for each clustering run (global and species specific). Six conversion types were used: polymorphic high resolution (polyHigh), monomorphic high resolution (monoHigh), no minor homozygote (noMinor), call rate below threshold (CRBT), off target variant (OTV), and Other. Of these categories polyHigh, monoHigh and noMinor were considered "recommended" for downstream analysis. The remaining categories were considered "not recommended". The genotyping results for the per-species clustering runs were concatenated into one dataset of similar dimensions than the output for the global clustering run (a matrix with dimensions: number of SNPs x number of samples). However, we recorded a single conversion type per SNP

for the global clustering analysis, but a separate conversion type per SNP for each of the per-species clustering runs.

Table 1. Breakdown of samples that were submitted for genotyping with the 420K screening array at subsection, species and subdivision levels.

Subsection name	Total samples	Species name	Total samples	Subdivision name	Total samples	Mega-gametophyte samples	Quality control: pass[†]	Quality control: fail[†]
Australes	479	*P. oocarpa*	71	*P. oocarpa*	71	4	69	2
		P. patula	76	*P. patula* var. *patula*	52	2	52	0
				P. patula var. *longipedunculata*	24	0	24	0
		P. tecunumanii	149	*P. tecunumanii* High Elevation	89	4	88	1
				P. tecunumanii Low Elevation	60	6	60	0
		P. greggii	65	*P. greggii* South	35	2	29	6
				P. greggii North	30	0	30	0
		P. caribaea	64	*P. caribaea*	64	5	64	0
		P. elliottii	35	*P. elliottii*	35	0	35	0
		P. elliottii x *P. caribaea*	19	*P. elliottii* x *P. caribaea*	19	0	19	0
Ponderosae	97	*P. maximinoi*	77	*P. maximinoi*	77	5	76	1
		P. pseudostrobus	20	*P. pseudostrobus*	20	0	20	0
Total	**576**		**576**		**576**	**28**	**566**	**10**

[†]Default sample quality control (QC) thresholds were applied in the Axiom Analysis Suite for clustering per species, except for *P. caribaea* and *P. greggii*

DQC: ≥ 0.82 (default)

QC call rate: ≥ 97 (default); ≥ 90 (*P. caribaea*); ≥ 84 (*P. greggii*)

Percent of passing samples: ≥ 95 (default); ≥ 90 (P. *caribaea*); ≥ 90 (*P. greggii*)

Average call rate for passing samples: ≥ 98.5 (default); ≥ 90 (*P. caribaea*); ≥ 93 (*P. greggii*)

2.5 Commercial Array Marker Selection

To select markers for tiling on the commercial 50K array, validated markers from the screening array were filtered from three different starting points, each with its own aim (**Figure 3**). Set A used global clustering, where the aim was to select a set of markers that would be useful for future joint analyses across species and phylogenetic classes. To be considered for selection in Set A, markers had to be classified as polyHigh and had to be homozygous in all 28 megagametophyte samples. Subsection-level markers (Set A1) included markers with a subsection-level minor allele frequency (MAF) greater than 0.05 and that were informative in both subsections (Australes and Ponderosae; **Supplementary Figure 1**). Set A2 included stable assay markers that were shared across species in a joint analysis scenario. Markers were selected for this category if they had a species-level MAF greater than 0.05 in three or more species.

Set B used the single species clustering runs where the aim was to target markers that show utility across species when analysed individually. SNPs were considered for selection in Set B, if they (i) had a MAF greater than or equal to 0.05 in the individual per-species clustering runs, (ii) converted as either polyHigh or noMinorHom in the respective species and (iii) were homozygous in all megagametophyte samples in that species. Within Set B, Set B1 included markers that met Set B criteria in three or more of the species; Set B2 included markers that met Set B criteria in two species (markers shared by two species); and Set B3 included markers that met Set B criteria in only one species.

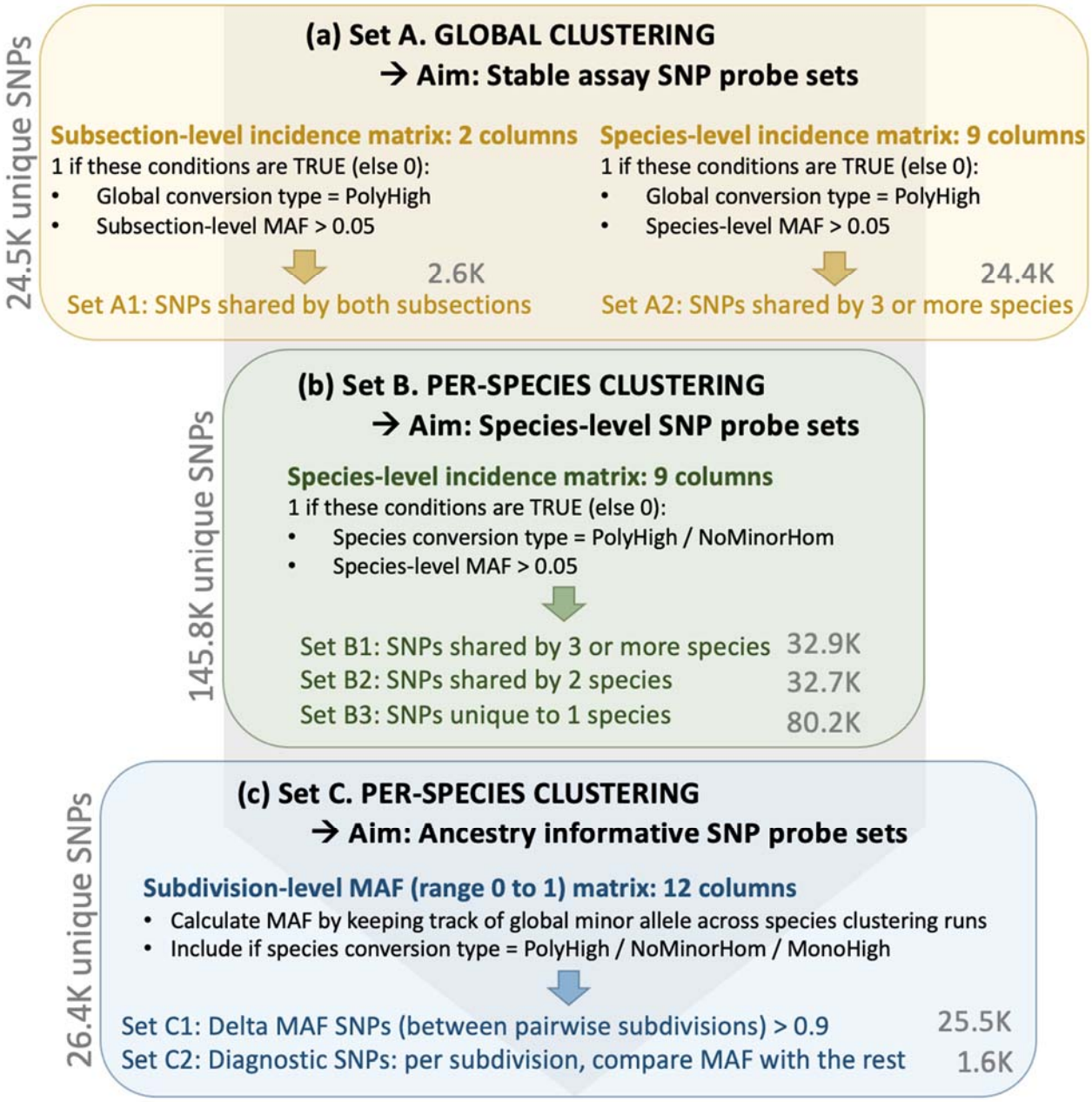

Figure 3. Initial filtering of screening array SNP probe sets. Flow diagram outlining how the SNP probe sets on the screening array were filtered using three different starting points, each with its own aim: (a) Set A represent stable assay SNP probe sets, (b) Set B represent species-level SNP probe sets and (c) Set C represent ancestry informative SNP probe sets. This filtering strategy gave rise to seven different subsets (Sets A1, A2, B1, B2, B3, C1 and C2), altogether resulting in a set of 153K unique SNP probe sets. The final 50K SNP probe sets were selected from this filtered set.

Set C used the single species clustering analyses to identify markers that were potentially ancestry informative and those that could discriminate between species or subdivisions. Markers were included in Set C if they (i) were classified as recommended (polyHigh, monoHigh or noMinor) for one or more species, while (ii) being homozygous in all megagametophyte samples for those species. We also included MonoHigh SNPs (unlike sets A and B), to be able to capture opposite MonoHigh SNPs. Furthermore, for this set we calculated the MAF by tracking the global minor allele across all species clustering runs, and then used this to calculate the MAF on a scale from 0 to 1 for each marker, per species. The calculated MAF was used to create two subsets within set C. Set C1 included high delta MAF markers where the change in MAF for a given marker between any two species was greater than 0.9 (all pairwise species combinations were considered; see **Supplementary Figure 2**). Set C2 was selected as a set of potentially diagnostic markers – markers with a large difference in MAF when comparing one species or subdivision to all others (or subsets of species). For this analysis, the MAF was calculated at the subdivision level, to allow the inclusion of SNPs with MAF values at opposite ends of the spectrum, also distinguishing subdivisions. **Supplementary Table 2** gives the different thresholds that were used to produce the final list in Set C2.

The total number of unique SNPs in these sets (153K) exceeded the space available on the final commercial Pitro50K array. These markers were scored by Thermo Fisher and assessed based on their p-convert value, as described previously. To achieve a target of at least 15K polymorphic markers for each species, markers were selected to maximize the overlap between sets. The bulk of markers from Set A2 was present within Set B1 and were selected for inclusion on the array.

Additional markers were selected from Sets B2 and B3 to increase the number of polymorphic markers within *P. greggii* and *P. maximinoi*. Markers from these sets were only selected for inclusion if they had a MAF of 0.2 or greater to enrich for more common variants within their populations. Markers from Set C1 were not added as >8K were already included in the markers selected from Sets A and B, but all the additional diagnostic markers from Set C2 were added. Furthermore, 115 SNPs were selected for tiling an additional two times on the array to act as a technical control to assess repeatability.

2.6 Assessment of Selected Commercial Array Markers

Post selection and scoring, we performed descriptive analysis, MAF spectra characterization, allelic concordance analysis and principal component analysis (PCA) using R version 3.5.1 (R core team, 2018) using genotypes from the screening array for selected markers. SNP & Variation Suite v8.8.3 was used to calculate the principal components (using **Supplementary Table 6** as input) and the R package *plotly* (Sievert, 2020) was used to visualize the PCA results in three dimensions, as well as to create stacked bar plots for visualizing conversion type categories and SNP sharing between species. We used the non-negative matrix factorization algorithm (sNMF; Frichot et al., 2014) from the LEA R package (Frichot and Francois, 2015) for population structure estimation. The number of ancestral populations (K) when the cross-entropy curve exhibited a plateau, appeared to be at K=10 or K=11 (**Supplementary Figure 3**). We used twelve repetitions for each value of K, which was tested from one to fifteen. We displayed the Q-matrix for K=2 up to K=11 clusters using a bar plot representation. The samples were ordered within each species and subdivision per provenance by latitude (from South to

North). Due to concerns about variable performance of the SNP assays in different genetic backgrounds potentially leading to species-specific genotype clustering patterns, we investigated allelic concordance between global and per-species clustering analyses. For each SNP, the number of samples with exactly the same genotype calls in the global vs the per-species clustering analysis were calculated.

3 Results

3.1 SNP Discovery and Selection for Screening Array

A total of 5 million paired-end capture-seq reads comprising 753 million bases on average was produced for each of six tropical pine species (**Figure 1a**). After trimming reads to a minimum base quality score and minimum read length, 4.3 million trimmed reads on average consisting of 615 million bases were used for downstream analysis (**Supplementary Table 3**). Alignment of the trimmed reads to the tropical pine genome yielded very high alignment percentages of >98% for each species. Filtering of the reads to include only those that mapped in proper pairs and within 500 bases on either side of an area targeted for sequencing (site of a capture probe) reduced the number of reads to be used in variant calling by ~57% across all species.

RNA sequencing generated on average 5.7 billion raw bases from 39.6 million paired-end reads for each of five tropical pine species (the species in **Figure 1a**, except *P. caribaea*). After quality control, trimming reads for base quality and read length, an average of 4.9 billion bases from 34.9 million reads were retained for further analysis. Variant calling for both RNA-seq and capture-seq data generated a list of variants including complex events, multi-

nucleotide polymorphisms, SNPs, and indels. Variant calling was not restricted to SNPs alone per recommended best practices for variant calling using Freebayes (Garrison & Marth, 2012) in order to increase detection power. The resulting variant list was then filtered to include only SNPs for downstream analysis.

Within the RNA-seq data, the number of variants called ranged from 686K to 1.4 million SNPs per species post filtering, aggregating to a total of 5.3 million SNPs (**Figure 2a**). *P. patula* had the highest number of SNPs discovered within the species while *P. greggii* had the lowest number of SNPs generated from sequencing data. From these SNPs, between 175K and 301K SNP probe sets were successfully designed per species (**Supplementary Table 4**), resulting in a non-redundant set of 1.3 million probe sets. Variant calling within the capture-seq data yielded a list of 3.3 million raw SNPs (**Figure 2b**). The additional filtering steps (**Supplementary Method 2**) created two lists of 418K (top-down: SNPs shared among most species) and 1.3 million SNPs (bottom-up: SNPs detected in at least one provenance). A total of 403K SNPs were shared between the two lists. Sets were merged to create a list of 1.3 million SNPs for probe design, which yielded a total of 563K successful probe sets.

Altogether, a total of 424,074 unique SNPs across species were selected, including 300K RNA-seq and 120K capture-seq derived SNPs, for tiling on the screening array (**Figure 2**). The opposite strand probes of 61,192 of the selected markers were also tiled, resulting in a total of 485,797 probe sets targeting the 424,074 SNPs.

3.2 SNP Evaluation and Selection for Commercial Array

Evaluation of the screening array probe sets via genotyping of 576 samples across nine species resulted in a high success rate for samples. Using global clustering, only 22 samples did not pass the recommended best practice dQC and call rate thresholds recommended by Thermo Fisher. With the per-species clustering, only 10 samples did not reach the quality control thresholds (**Table 1**). Across clustering analyses, 43-66% of the 485K SNP probe sets were "recommended" (that were in polyHigh, monoHigh, noMinor clusters; **Table 2**). Across all species, monoHigh and Other probe sets were the two most common conversion types, while OTV was the least common (**Figure 4a**). Ultimately, we filtered these SNP probe sets to select a set of 50K informative markers for tiling on the commercial 384-format Axiom array. On average, 78% of the recommended SNP probe sets were RNA-seq derived and 22% were capture-seq derived (**Table 2**). Filtering the recommended probe sets into the sets outlined in **Figure 3** resulted in a set of 153K unique SNPs that formed the basis for our selection. Note that here (and in the remaining part of the text) SNPs or markers refer to specific SNP probe sets, unless distinguished explicitly. Set A (stable assay SNPs), Set B (species-level SNPs) and Set C (ancestry informative SNPs) contained lists of 24,540, 145,778 and 26,370 markers, respectively, that were not mutually exclusive (**Supplementary Figure 4**).

Table 2. The number of recommended probe sets, out of the 485,797 SNP probe sets on the screening array, resulting from global clustering and the nine per-species clustering runs.

Clustering	Species name	Recommended Markers[†]	Percentage recommended out of total	RNA-Seq derived SNPs[‡]	Capture-Seq derived SNPs[‡]
Global	All samples	210,361	43%	76.8%	23.2%
Per-species	*P. oocarpa*	257,908	53%	78.9%	21.1%
	P. patula	280,803	58%	78.4%	21.6%
	P. tecunumanii	245,691	51%	78.2%	21.8%
	P. greggii	214,766	44%	77.7%	22.4%
	P. caribaea	242,855	50%	79.3%	20.7%
	P. elliottii	319,978	66%	77.3%	22.7%
	P. elliottii x *P. caribaea*	298,744	61%	78.0%	22.0%
	P. maximinoi	258,164	53%	78.7%	21.4%
	P. pseudostrobus	249,223	51%	80.0%	20.0%
Average		**257,849**	**53%**	**78%**	**22%**

[†]Recommended SNPs were assigned conversion types: polymorphic high resolution (PolyHigh), monomorphic high resolution (MonoHigh) or no minor homozygote (NoMinor).

[‡]Each SNP was initially discovered via RNA sequencing (RNA-seq) or targeted capture sequencing (Capture-seq) data.

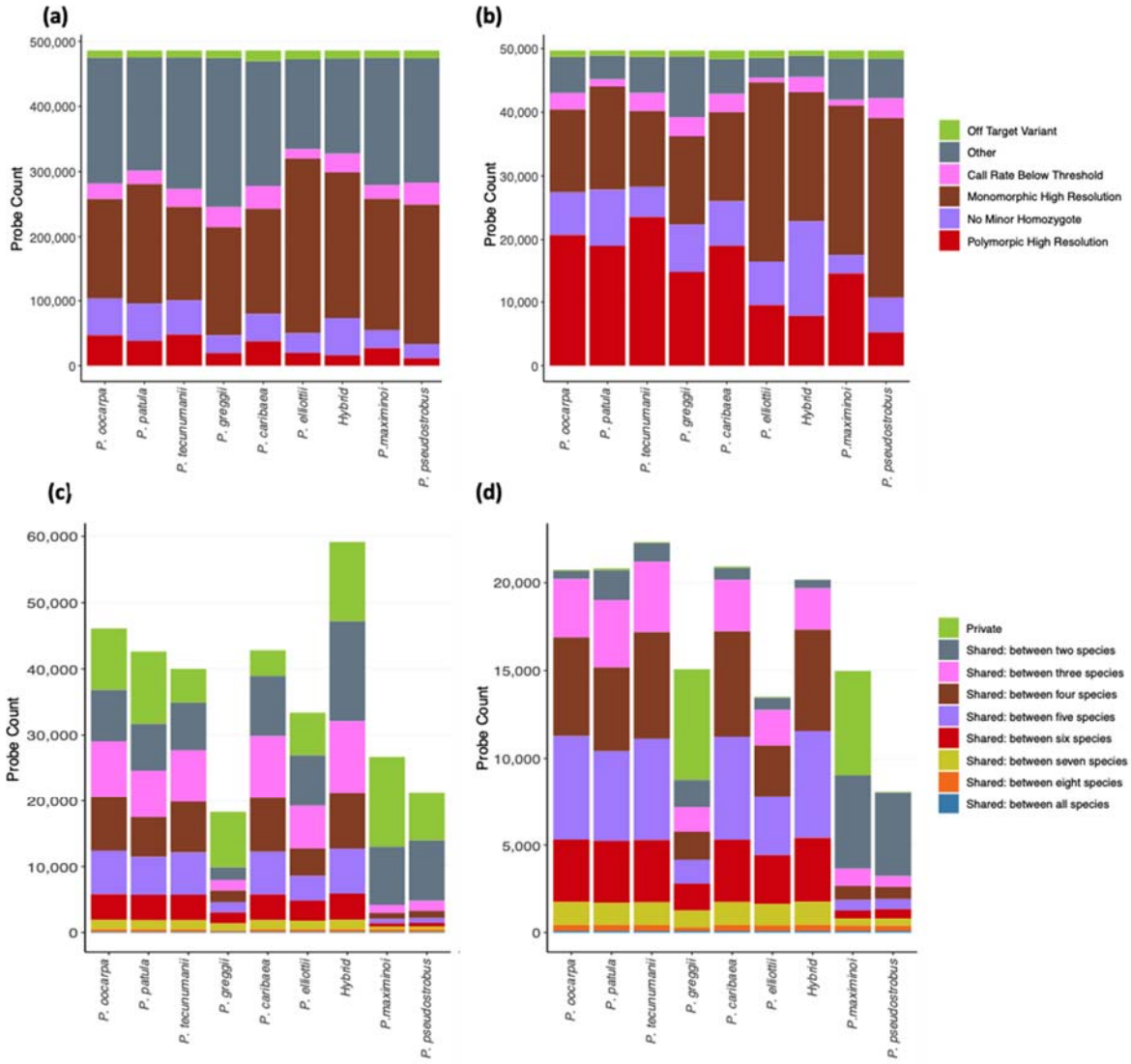

Figure 4. Comparison of conversion types and SNPs sharing between species, for SNP probe sets on the screening versus commercial array. (a) Stacked bar plot of conversion type categories after the nine single species clustering runs for all 485,797 SNP probe sets on the screening array and (b) for the subset of 49,694 SNP probe sets selected for the commercial array. (c) Stacked bar plot of the SNP probe sets on the screening array that were classified as polymorphic high resolution or no minor homozygote with a MAF >= 0.05. Per species, a breakdown is given of the proportions of SNPs shared across all species, 8, 7, 6, 5, 4, 3 and 2 species as well as of SNPs private to that species. (d) Stacked bar plot of the SNP probe sets on the commercial array giving a summary of the number of probe sets that are informative per species (*P. oocarpa* = 20,746; *P. patula* = 20,825; *P. tecunumanii* = 22,329; *P. greggii* = 15,082; *P. caribaea* = 20,933; *P. elliottii* = 13,510; P. elliottii x *P. caribaea* (Hybrid) = 20,189; *P. maximinoi* = 14,978; *P. pseudostrobus* = 8,098). Per species, a breakdown is given of the proportions of SNPs shared across all species, 8, 7, 6, 5, 4, 3 and 2 species as well as of SNPs private to that species.

Table 3. Selection strategy of 50K informative SNP probe sets for tiling on a commercial array: Pitro50K.

Tiling order category	Set[†]	Level	Description and species included[‡]	Number of probe sets selected[§]	Aggregate number
1 (Technical repeats)	Set A (90%) Set B (10%)	Global Species-level	PolyHigh (global clustering); PolyHigh and polymorphic (MAF>0.05) in 8 or 9 species (per-species clustering)	115	115
2	Set B	Species-level	SNPs shared by 3+ species: *P. patula*, *P. tecunumanii, P. greggii*, *P. caribaea*, *P. elliottii*, *P. maximinoi*	18,489	18,604
3	Set B	Species-level	SNPs shared by 2 species (*P. patula*, *P. tecunumanii, P. greggii*, *P. caribaea*, *P. elliottii*, *P. maximinoi*), always including *P. greggii* or *P. maximinoi*	3,795	22,399
4	Set B	Species-level	SNPs unique to *P. greggii* or *P. maximinoi* with MAF>0.2 if possible	17,723	40,122
5	Set C2	Subdivision level	Diagnostic SNPs (200 or less per subdivision)	966	41,088
6	Set B	Species-level	SNPs shared by 2 species: *P. patula*, *P. tecunumanii, P. caribaea*, *P. elliottii*	8,097	49,185
7	Set B	Species-level	SNPs informative in only one of the species: *P. patula*, *P. tecunumanii, P. caribaea*, *P. elliottii*	509	49,694

[†] Sets are defined in Figure 3.
[‡] For categories 2, 3, 6 and 7, SNPs may be also informative in *P. oocarpa, P. elliottii* x *P. caribaea* (hybrid) and *P. pseudostrobus.*
[§] SNPs were only assigned to a category if they were not yet included in a previous category.

Set A was used to select a set of 115 markers that were included as technical repeats on the commercial array. These markers were (i) PolyHigh in the global clustering analysis and (ii) PolyHigh and polymorphic in 8 or 9 out of the 9 species in the per-species clustering analysis (tiling

order 1 in **Table 3**). No other markers unique to Set A were included in the selection of 50K, as the bulk of SNPs in Set A (99%) were also part of Set B. Furthermore, we did not want to include Set A-specific markers that would not convert to informative SNPs in the single species clustering analysis.

For selection from Set B, we prioritized *P. patula, P. tecunumanii*, *P. greggii, P. caribaea, P. elliottii* and *P. maximinoi*, since these species are the most commercially important as both pure species and in hybrid programs. First, 18,489 markers were selected as being informative in three or more of the above-mentioned species (tiling order 2 in **Table 3**). *P. greggii* and *P. maximinoi* were under-represented in this set and we therefore selectively increased the number of markers for *P. greggii* and *P. maximinoi*. Markers selected for this purpose were required to be shared by two species, with at least one being *P. greggii* or *P. maximinoi* (from sets B1 and B2; tiling order 3 in **Table 3**), or unique to either *P. greggii* or *P. maximinoi*, with a MAF greater than 0.2 where possible, in order to emphasize more common polymorphisms (mostly from set B3; tiling order 4 in **Table 3**) even if species-specific. Together, this added an additional 21,518 markers to our selection.

We added all diagnostic markers from Set C2 (**Supplementary Table 2**) that were not yet selected (966 out of 1,564; tiling order 5 in **Table 3**). The remaining markers were selected from Set B, but shared by only one or two of the species we prioritized – not including *P. greggii* or *P. maximinoi* (tiling order 6 and 7 in **Table 3**).

We ended up with a total 49,694 unique probe sets (representing 49,674 SNP markers) that were used in the development of the commercial array which is named Pitro50K (**Figure 2c**).

For 20 markers, probe sets from both strands were tiled due to our specific selection of probes for diagnostic purposes (set C2). For these SNPs, probe sets on opposite strands were diagnostic for different species. Between 8K and 22K markers were informative for each of the species included in the study (**Figure 4d**). **Supplementary Table 5** gives detailed annotations of the 49,694 probe sets and also stipulates the list of species in which each marker is informative. **Supplementary Table 6** provides the forward strand base call genotypes (across all probe sets) per sample.

3.3 Assessment of Selected 50K Markers

Conversion types for all markers on the screening array and those selected on the commercial array were assessed by species. A clear shift can be seen between the two datasets pertaining to the proportion of recommended markers (PolyHigh, MonoHigh, NoMinor); only 49-66% of markers per species were recommended in the screening phase (**Figure 4a**), compared to >75% (for all species) in the set of 50K markers (**Figure 4b**). Similarly, the distribution of markers shared among species shifted between the screening array and the production array. Of the nine species, six had very few private markers and markers shared by two species, as expected given the selection process followed. However, for *P. greggii* and *P. maximinoi*, more than half of the markers were private or shared by two species (**Figure 4c, 4d)**. This shift in distribution indicates that our categorical selection process was successful in maximizing the proportion of shared SNPs and the total number of SNPs for those species with fewer shared SNPs.

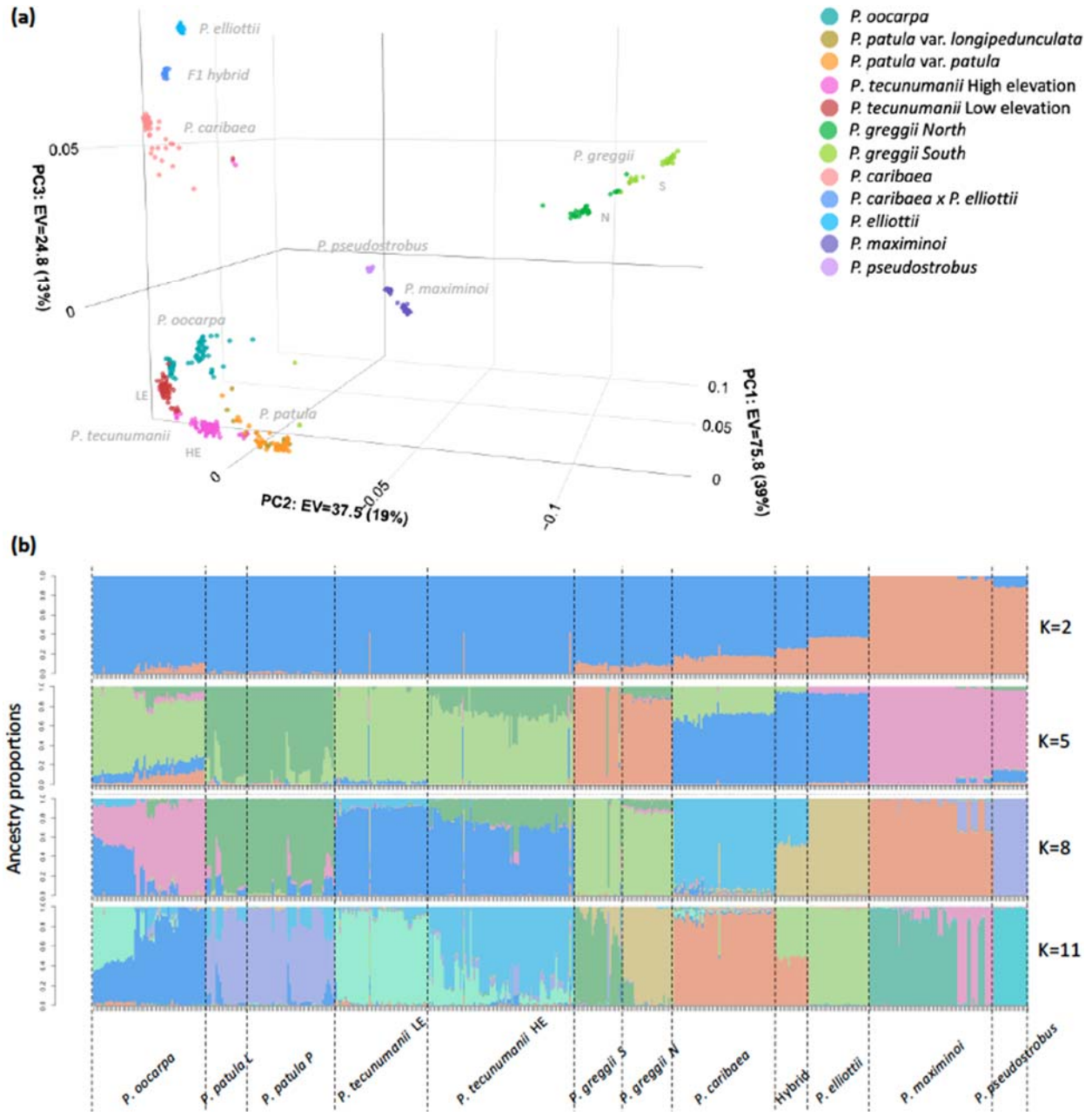

Figure 5. Principal component analysis and structure plot of genome assignment. (a) Principal component (PC) analysis of the selected 49,694 SNP probe sets across all 538 samples. Different colors distinguish between the species and subdivisions. The first PC explains 39%, the second PC 19% and the third PC explains 13% of the total variance (see the interactive 3D PCA plot: https://chart-studio.plotly.com/~nanettec/5/#/). (b) Structure plot of genome assignment for K=2, K=5, K=8 and K=11, where K is the number of ancestral populations. The samples are ordered by species and then by provenance from South to North (based on latitude).

Minor allele spectra analysis using the 0-1 scale (keeping the global minor allele constant across all per-species clustering runs) revealed a large number of very rare and very common alleles per species with good representation across the entire frequency scale. **Supplementary Figure 5** shows that SNPs with low MAF are more common throughout the dataset, however these are not the primary source of variation. Areas of enrichment for more common SNPs caused by our selection process can be seen around MAF = 0.2 for *P*. *greggii* and *P*. *maximinoi*.

Using the 50K markers, a PCA resolved most species and subdivisions into unique genetic groups (**Figure 5a**). *Pinus maximinoi* and *P. pseudostrobus* samples (in the Ponderosae subsection) behaved as expected and clustered distantly from the other Australes species. *P. greggii* and *P. tecunumanii* samples resolve into their subdivision categories (North/South and high/low elevation) well. However, there was complete overlap between the *P. patula* subspecies. The 50K markers also successfully partitioned *P. caribaea*, *P. elliottii*, and their hybrids into distinct clusters, with the F1 hybrid exactly intermediate. A PCA using the subset of 1.5K diagnostic markers from Set C2 (**Figure 3**), shows the same general trend but with tighter clustering and clearer resolution between clusters of species and subdivisions (**Supplementary Figure 6**).

In the structure plot, from K=5 onwards, it is possible to distinguish between *P. tecunumanii* high and low elevation, as well as between *P. greggii* North and South (**Figure 5b**). Similar to our conclusion with the PCA, it remains difficult to distinguish the two *P. patula* subspecies. For *P. oocarpa*, we can discriminate the South, Central and North provenances, as the samples in the structure plot are ordered according to latitude (South to North) of the sampled provenances (K=11; **Supplementary Figure 7, Supplementary Table 7**). The three *P. oocarpa*

groups can be distinguished geographically and in a PCA plot of the 50K markers including only *P. oocarpa, P. patula* and *P. tecunumanii* samples (**Supplementary Figure 8**).

Approximately half of the SNPs (20,024 or 54% of the SNPs included in the analysis) had an allelic concordance of more than 99% between global and per-species clustering analyses; whereas 5,617 SNPs (15% of the SNPs included in the analysis) had a concordance of 95% or less (corresponding to 500 out of the 521 samples) (**Supplementary Figure 9**). These results suggest that species genetic background does affect SNP clustering and genotype calling.

4 Discussion

4.1 SNP selection towards a genotyping resource for tropical pines

Elevation and associated environmental factors such as temperature and rainfall are important considerations for the conservation and genetic improvement of tropical pines, activities that will be supported by the new Pitro50K SNP array. Elevation generally increases from South to North in Mexico and Central America where the 81 provenances representing six tropical pine species were sampled (**Supplementary Figure 10**). In the tropical south, we included *P. caribaea* and *P. tecunumanii* low elevation, provenances typically associated with low cold tolerance. More towards the north we included *P. maximinoi* and *P. tecunumanii* high elevation provenances. We also included provenances of *P. patula* and *P. greggii* occurring at higher elevation, on mist mountains in the cold, humid highlands. Finally we included provenances of *P. oocarpa,* geographically the most wide-spread of the species and most likely the ancestral species for the Australes subsection (Dvorak, Gutierrez, et al., 2000). Sampling of individuals across the range allowed us to maximize the genetic diversity within our pool of samples. Given the number

of environmental factors acting across the populations for any given species, this sampling design likely reduced bias that may have occurred with a narrower sampling scheme.

The use of pooled samples of families within a provenance allowed us to capture more genetic diversity that would have not been possible with individual samples. The small number of individuals within some of the pools was largely dictated by availability of genetic material. While each pool had a small number of individuals, the use of pooled samples containing fewer than 20 individuals has been successfully implemented using targeted capture sequencing in conifers (Rellstab et al. 2019; Gepts et al. 2016). This emphasis on maximizing genetic diversity allowed for selection of SNPs that are likely common across multiple populations and therefore may be more informative across applications. Capture of this shared allelic diversity is supported by the high number of polymorphic markers shared among species in our final design.

One caveat to this conclusion is the enrichment of the *P*. *greggi* and *P*. *maximinoi* data sets with species specific markers. Given *P. maximinoi*'s status as the genetic outgroup in this study, it is understandable that it would share fewer markers with the other species and that we would have to supplement the SNP set with *P. maximinoi* specific markers. *Pinus greggii* produced the smallest number of SNP markers during the SNP discovery phase. It has been shown to have much less genetic diversity as a species than other members of the Australes subsection (Wehenkel et al., 2017). This evidence, taken together, supports that the *P*. *greggii* dataset likely reflects the species diversity and that we were justified in supplementing it with private markers to enrich SNP numbers for genetic analysis of this species.

Our PCA analysis results mimic what one would expect from a phylogenetic perspective, where individuals from species such as *P*. *oocarpa* and *P*. *tecunumanii* cluster more closely to one another than to other more distantly related species (Dvorak, Jordon, Hodge, & Romero, 2000) and *P*. *maximinoi* is the most genetically distant of the assayed species (Vargas-Mendoza et al., 2011). Additionally, the PCA analysis using only a subset of markers was able to identify three *P. tecunumanii* samples that may have been mislabelled, or may be cryptic F1 hybrids. These samples were removed prior to the filtering analysis to avoid spurious selection of uninformative markers. This result highlights the array's potential for DNA fingerprinting and hybrid discrimination using both the full array and a subset of diagnostic markers.

4.2 Recommendations for SNP marker genotype clustering and use in hybrids

The results from the global and species-specific clustering revealed that there were an additional 8.5K markers that scored "polymorphic high resolution" in the species-specific dataset (compared to the global, cross-species dataset) under the same genotyping parameters. This suggests that careful consideration is necessary when selecting samples to cluster together for genotype calling because sample genetic composition can drastically alter clustering and genotyping calls. As more pine tree samples are assayed in future experiments, a set of species-specific cluster position files can be compiled as references to help inform and increase the accuracy of genotype calls for each species. A similar approach has been demonstrated in *Eucalyptus*, where sample size and taxonomic composition of cluster files can be used to optimize the call rate, genotype concordance, and total number of SNPs successfully genotyped using a multi-species SNP chip (Silva-Junior et al., 2015).

Furthermore, it is understood that the lack of technical and biological replicates in the development of the 50K array limits our ability to assess repeatability of the novel markers used in the design. Additionally, due to the lack of family structure within our samples, it was not possible to assess Mendelian segregation. Due to the focus on developing a multi-species SNP chip, we opted to maximize the coverage of genetic diversity in the natural ranges of the species. Further studies will need to be performed to assess technical repeatability and Mendelian segregation patterns, as well as develop sets of reference samples to be used to inform cluster position files during genotype calling for the different species and hybrid combinations that will be analysed with the chip.

Two of the most important and exciting expected uses for this SNP array are the identification of hybrid individuals and its use for genomic selection in hybrid populations. Some interspecific tropical pine hybrids have been shown to have superior growth traits, wood quality, and disease resistance compared to their pure species counterparts (Dungey, 2001; Kanzler, Nel, & Ford, 2014). For these reasons, locating markers that are beneficial for traits of interest through genome wide association or whole genome regression methods to inform hybrid breeding schemes is of much interest. However, one must understand that population structure between the crossed species may cause spurious associations, because alleles that are fixed in both species will be in complete linkage disequilibrium with one another (Grattapaglia & Kirst, 2008). Therefore, knowing the behavior of these markers in each species involved in the hybrid cross is key to implementing both applications effectively.

4.3 Pitro50K a New Tool for Pine Genomics

This manuscript details the design and analysis of the Pitro50K genotyping array using targeted capture sequencing and species-specific transcriptomic SNP discovery. This array has been designed specifically for six species of tropical and subtropical pines along with their hybrid combinations. Additionally, we demonstrated utility in related species with the inclusion of *P. elliottii* and *P. pseudostrobus* in the SNP validation and selection processes. In total, 49,674 SNP markers were selected to comprise the array with at least 15K high quality polymorphic markers for each of the target species in our design. The emphasis on selecting a set of SNPs shared across multiple species coupled with a large range in allele frequencies not only makes this array a powerful tool for genomic selection studies, but also for assessment of population structures within and between species. One of the most immediately applicable and pressing uses for this technology is the development of high-quality genetic maps for these pine species with eventual culmination into consensus genetic maps and anchored genome assemblies to aid in genetic dissection of complex growth, development and defence related traits.

Additional to the array itself, this study accessed a number of separate genomic and transcriptomic resources that could be valuable to future studies. The sequencing data generated from this study could serve as a foundation for interrogation of DNA markers and technologies at the single-species level through the creation of sub-arrays, genotype-by-sequencing protocols or amplicon sequencing methodologies. A smaller “gold standard set” of markers identified over time from the use of the array could be more selectively targeted using such approaches. Ultimately, the Pitro50K array will provide the global community with a foundational genomic

resource that is highly transferable, reproducible and computationally friendly for the implementation of genome-wide genotyping in breeding and genetic research programs.

Acknowledgements

The authors thank Marja O'Neill for technical assistance, RAPiD Genomics for performing capture-sequencing of 81 tropical pine provenances, Camcore (NC State University, Raleigh, NC) and Forest Molecular Genetics (FMG) Programme industry members for providing genetic materials and funding for SNP discovery and screening and the North Carolina State University Cooperative Tree Improvement Program for student support. AAM and SN acknowledge funding support from the Forestry Sector Innovation Fund (FSIF), Department of Science and Innovation (DSI), Technology Innovation Agency (TIA), and National Research Foundation - Bioinformatics and Functional Genomics Programme (NRF-BFG Grant UID 97911) of South Africa. Pitro50K SNP array was designed under the Conifer SNP Consortium and Thermo Fisher agreement. FI and JJA acknowledge funding support from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) project (Grant # 2016-67013-24469).

Data Availability Statement

The pooled targeted capture sequencing data are available via NCBI SRA BioProject accession PRJNA742386. RNA-seq data are available via NCBI SRA BioProject accessions PRJNA416697 (*P. tecunumanii*), PRNJA416698 (*P. patula*), PRJNA685280 (*P. oocarpa*), PRJNA685281 (*P. greggii*) and PRJNA685282 (*P. maximinoi*). Metadata and probe set sequences

used for markers selected for the 50K commercial array are available as supporting information (**Supplementary Table 5**). Genotype dataset used for PCA and STRUCTURE analysis is available in supporting information (**Supplementary Table 6**).

Author Contributions

Research was conceived, planned and guided by AAM, JJA and GH; CJ and NC performed SNP discovery, selection, genotyping data analysis, and contributed equally in preparing this manuscript. CM assisted with SNP discovery from the capture-seq data. RNA-seq analysis and transcriptome assembly were performed by SN, EV, TK. JW and MC contributed computational resources and guidance. DK contributed genotyping samples and support. DNA extraction and initial marker analysis of the 81 provenances used for capture-seq were performed by MR and YT. Additional support and advice were provided by RW and FI throughout the SNP discovery and probe design process. All authors participated in manuscript review and revision.

References

Applied Biosystems (2017). *Axiom Analysis Suite 3.1 User Guide*. Available at https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html

Azaiez, A., Pavy, N., Gérardi, S., Laroche, J., Boyle, B., Gagnon, F., … 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*(1), 1–13.

https://doi.org/10.1186/s12864-018-5247-z

Caballero, M., Lauer, E., Bennett, J., Zaman, S., McEvoy, S., Acosta, J., … Isik, F. (2021). Toward genomic selection in *Pinus taeda*: Integrating resources to support array design in a complex conifer genome. *Applications in Plant Sciences, 9*(6). https://doi.org/10.1002/aps3.11439

Chancerel, E., Lamy, J. B., Lesur, I., Noirot, C., Klopp, C., Ehrenmann, F., … Plomion, C. (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*(1), 50. https://doi.org/10.1186/1741-7007-11-50

Chancerel, E., Lepoittevin, C., Le Provost, G., Lin, Y.-C., Jaramillo-Correa, J. P., Eckert, A. J., … Plomion, C. (2011). Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. *BMC Genomics*, *12*. https://doi.org/10.1186/1471-2164-12-368

Conkle, M. T. (1979). Isozyme Variation and Linkage in Six Conifer Species. *Isozymes of North American Forest Trees and Forest Insects*, 11–17.

Dantec, L. Le, Chagné, D., Pot, D., Cantin, O., Garnier-Géré, P., Bedon, F., … Plomion, C. (2004). Automated SNP detection in expressed sequence tags: Statistical considerations and application to maritime pine sequences. *Plant Molecular Biology*, *54*(3), 461–470. https://doi.org/10.1023/B:PLAN.0000036376.11710.6f

Devey, M. E., Beil, J. C., Smith, D. N., Neale, D. B., & Moran, G. F. (1996). A genetic linkage map for *Pinus radiata* based on RFLP, RAPD, and microsatellite markers. *Theoretical and Applied Genetics*, *92*(6), 673–679. https://doi.org/10.1007/BF00226088

Devey, M. E., Fiddler, T. A., Liu, B.-H., Knapp, S. J., & Neale, D. B. (1994). An RFLP linkage

map for Loblolly pine based on a three-generation outbred pedigree. *Theor Appl Genet*, *88*, 273–278.

Dungey, H. S. (2001). Pine hybrids - A review of their use performance and genetics. *Forest Ecology and Management*, *148*(1–3), 243–258. https://doi.org/10.1016/S0378-1127(00)00539-9

Durán, R., Rodriguez, V., Carrasco, A., Neale, D., Balocchi, C., & Valenzuela, S. (2019). SNP discovery in radiata pine using a de novo transcriptome assembly. *Trees - Structure and Function*, *33*(5), 1505–1511. https://doi.org/10.1007/s00468-019-01875-w

Dvorak, W. S., Gutierrez, E. A., Hodge, G. R., Romero, J. L., Stock, J., & Rivas, O. (2000). *Conservation & Testing of Tropical & Subtropcial Forest Tree Species by the CAMCORE Cooperative*. NCSU.

Dvorak, W. S., Jordon, A. P., Hodge, G. P., & Romero, J. L. (2000). Assessing evolutionary relationships of pines in the *Oocarpae* and *Australes* subsections using RAPD markers. *New Forests*, *20*(2), 163–192. https://doi.org/10.1023/A:1006763120982

Echt, C. S., & May-Marquardt, P. (1997). Survey of microsatellite DNA in pine. *Genome*, *40*(1), 9–17. https://doi.org/10.1139/g97-002

Eckert, A. J., Pande, B., Ersoz, E. S., Wright, M. H., Rashbrook, V. K., Nicolet, C. M., & Neale, D. B. (2009). High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes*, *5*(1), 225–234. https://doi.org/10.1007/s11295-008-0183-8

Frichot, E., Francois, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6, 925-929. https://doi.org.10.1111/2041-210X.12383

Frichot, E., Mathieu, F., Trouillon, T.,Bouchard, G., Francois, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *GENETICS*, 196(4), 973-983. https://doi.org/10.1534/genetics.113.160572

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., … Falque, M. (2011). A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE*, *6*(12), e28334. https://doi.org/10.1371/journal.pone.0028334

Garrison, E. (2016). Vcflib, a simple C++ library for parsing and manipulating VCF files. https://github.com/vcflib/vcflib

Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*

Gepts, P., Gao, D., Wang, S., Syring, J. V., Tennessen, J. A., Jennings, T. N., … Cronn, R., (2016). Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. *Frontiers in Plant Science, 7*(484), 1-15. https://doi.org/10.3389/fpls.2016.00484

Geraldes, A., DiFazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., … 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*(2), 306–323. https://doi.org/10.1111/1755-0998.12056

Golden Helix Inc. SNP & Variation Suite (Version 8.x) [Software]. Available at: https://www.goldenhelix..com

Grattapaglia, D., & Kirst, M. (2008, September 1). Eucalyptus applied genomics: From gene sequences to breeding tools. *New Phytologist*, Vol. 179, pp. 911–929.

https://doi.org/10.1111/j.1469-8137.2008.02503.x

Grattapaglia, D., Silva-Junior, O. B., Kirst, M., de Lima, B. M., Faria, D. A., & Pappas, G. J. (2011). High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: Assay success, polymorphism and transferability across species. *BMC Plant Biology*, *11*(1), 1–18. https://doi.org/10.1186/1471-2229-11-65

Gwaze, D. P. (1999). Performance of some F1 interspecific Pine hybrids in Zimbabwe. *Forest Genetics*, *6*(4), 283–289.

Hongwane, P., Mitchell, G., Kanzler, A., Verryn, S., Lopez, J., & Chirwa, P. (2018). Alternative pine hybrids and species to *Pinus patula* and *P. radiata* in South Africa and Swaziland. *Southern Forests*. https://doi.org/10.2989/20702620.2017.1393744

Howe, G. T., Jayawickrama, K., Kolpak, S. E., Kling, J., Trappe, M., Hipkins, V., … McEvoy, S. (2020). An Axiom SNP genotyping array for Douglas-fir. *BMC Genomics*, *21*(1), 9. https://doi.org/10.1186/s12864-019-6383-9

Isik, F. (2014). Genomic selection in forest tree breeding: The concept and an outlook to the future. *New Forests*, Vol. 45, pp. 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*(5), 70.

Joshi, N. A., Fass, J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at: https://github.com/najoshi/sickle

Kanzler, A., Nel, A., & Ford, C. (2014). Development and commercialisation of the *Pinus patula* x *P. tecunumanii* hybrid in response to the threat of *Fusarium circinatum*. *New Forests*, *45*, 417–437. https://doi.org/10.1007/s11056-014-9412-1

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, J. J., Schoettle, A. W., Sniezko, R. A., Sturrock, R. N., Zamany, A., Williams, H., … Kegley, A. (2016). Genetic mapping of *Pinus flexilis* major gene (Cr4) for resistance to white pine blister rust using transcriptome-based SNP genotyping. *BMC Genomics*, *17*(1), 753. https://doi.org/10.1186/s12864-016-3079-2

Liu, J. J., Sniezko, R. A., Sturrock, R. N., & Chen, H. (2014). Western white pine SNP discovery and high-throughput genotyping for breeding and conservation applications. *BMC Plant Biology*, *14*(1), 1–13. https://doi.org/10.1186/s12870-014-0380-6

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

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., … Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, *7*(2), 111–118. https://doi.org/10.1038/nmeth.1419

McKeand, S. E. (1988). Optimum age for family selection for growth in genetic tests of loblolly pine. *Forest Science*, *34*(2), 400–411. https://doi.org/10.1093/forestscience/34.2.400

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal*, *75*(1), 146–156. https://doi.org/10.1111/tpj.12193

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*, 1755-0998.13223. https://doi.org/10.1111/1755-0998.13223

Plomion, C., Bartholomé, J., Lesur, I., Boury, C., Rodríguez-Quilón, I., Lagraulet, H., … González-Martínez, S. C. (2016). High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Molecular Ecology Resources*, *16*(2), 574–587. https://doi.org/10.1111/1755-0998.12464

Price, R. A., Liston, A., & Strauss, S. H. (1998). Phylogeny and systematics of *Pinus*. In D. M. Richardson (Ed.), *Ecology and Biogeography of Pinus* (pp. 49–68). Cambridge: Cambridge University Press.

R Core Team (2018). R: A language and environment for stastical computing. R Foundation for Stastical Computing. https://www.R-project.org

Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S., Gugerli, F. (2019). Using transcriptome sequencing and polled exome capture to study local adaptation in the giga-genome of Pinus cembra. *Molecular Ecology Resources, 19*(2), 536-551. https://doi.org/10.1111/1755-0998.12986

Rudin, D., & Ekberg, I. (1978). Linkage studies in *Pinus sylvestris* L. using macro gametophye allozymes. *Silvae Genetica*, *27*,(1), 1–12.

Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. *Chapman and Hall/CRC.* ISBN 9781138331457, https://plotly-r.com

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*. https://doi.org/10.1111/nph.13322

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B.

(2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE*, *8*(1), e54985. https://doi.org/10.1371/journal.pone.0054985

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., … Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*(5), 1136–1146. https://doi.org/10.1111/1755-0998.12570

Telfer, E., Graham, N., Macdonald, L., Li, Y., Klápště, J., Resende, M., … Wilcox, P. (2019). A high-density exome capture genotype-by-sequencing panel for forestry breeding in *Pinus radiata*. *PLOS ONE*, *14*(9), e0222640. https://doi.org/10.1371/journal.pone.0222640

Vargas-Mendoza, C., Medina-Jaritz, N., Ibarra-Sanchez, C., Romero-Salas, E., Alcalde-Vazquez, R., & Rodriguez-Banderas, A. (2011). Phylogenetic analysis of Mexian pine species based on three loci from different genomes (Nuclear, Mitochondiral, and Chloroplast). In J. Agboola (Ed.), *Relevant Perspecites in Global Environmental Change* (pp. 139–154). InTech.

Visser, E. A., Wegrzyn, J. L., Myburg, A. A., & Naidoo, S. (2018). Defence transcriptome assembly and pathogenesis related gene family analysis in *Pinus tecunumanii* (low elevation). *BMC Genomics*, *19*(1), 1–13. https://doi.org/10.1186/s12864-018-5015-0

Visser, E. A., Wegrzyn, J. L., Steenkmap, E. T., Myburg, A. A., & Naidoo, S. (2015). Combined de novo and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. *BMC Genomics*, *16*(1), 1057. https://doi.org/10.1186/s12864-015-2277-7

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., … Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, *12*(6), 787–796. https://doi.org/10.1111/pbi.12183

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., … Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, *196*(3), 891–909. https://doi.org/10.1534/genetics.113.159996

Wehenkel, C., Mariscal-Lucero, S. del R., Jaramillo-Correa, J. P., López-Sánchez, C. A., Vargas-Hernández, J. J., & Sáenz-Romero, C. (2017). *Genetic Diversity and Conservation of Mexican Forest Trees*. https://doi.org/10.1007/978-3-319-66426-2_2

Zimin, A. V, Stevens, K. A., Crepeau, M. W., Puiu, D., Wegrzyn, J. L., Yorke, J. A., … Salzberg, S. L. (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience*, *6*(1), 1–4. https://doi.org/10.1093/gigascience/giw016

[dataset] University of Pretoria; 2017; Low elevation *Pinus tecunumanii* defence transcriptome; NCBI SRA BioProject accession: PRJNA416697.

[dataset] University of Pretoria; 2017; *Pinus patula* Transcriptome or Gene expression: PRJNA416697.

[dataset] University of Pretoria; 2020; *Pinus oocarpa* Transcriptome; NCBI SRA BioProject accession: PRJNA685280.

[dataset] University of Pretoria; 2020; *Pinus greggii* Transcriptome; NCBI SRA BioProject accession: PRJNA685281.

[dataset] University of Pretoria; 2020; *Pinus maximinoi* Transcriptome; NCBI SRA BioProject accession: PRJNA685282.

[dataset] University of Pretoria, Camcore (NC State University, Raleigh NC); 2018; Targeted capture sequencing of pooled samples from six tropical pine species across 81 provenances; NCBI SRA BioProject accession: PRJNA742386

Supporting Information

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

Supplementary Method 1. Altering of the *P. taeda* genome assembly (v2.01) to create the tropical pine reference genome.
Supplementary Method 2. Filtering of capture-seq SNP variants prior to probe design.
Supplementary Method 3. Filtering of SNP probe sets for inclusion on the screening array.
Supplementary Table 1. Geographic and climatic information for the 81 provenances representing the natural ranges of six tropical pine species in Mexico and Central America.
Supplementary Table 2. Summary of the diagnostic SNP probe sets included on Pitro50K.
Supplementary Table 3. Sequencing metrics generated from targeted capture sequencing.
Supplementary Table 4. Summary of the numbers of SNP markers identified and SNP probe sets successfully designed from RNA sequencing per species.
Supplementary Table 5. Annotation of each of the 49,694 SNP probe sets tiled on the Pitro50K genotyping array.
Supplementary Table 6. SNP & Variation Suite input file for the 49,694 SNP probe sets across 538 samples.
Supplementary Table 7. Sample metadata accompanying the structure plots.
Supplementary Figure 1. Stable assay SNPs at the subsection level using global clustering.
Supplementary Figure 2. SNPs with a large pair-wise minor allele frequency difference between species.
Supplementary Figure 3. Cross-entropy for pine species.
Supplementary Figure 4. Venn diagram of how the sets identified after initial filtering of SNP probe sets on the screening array overlap with the selected 50K SNP probe sets.
Supplementary Figure 5. Minor allele frequency distributions of the 50K SNP probe sets per species.
Supplementary Figure 6. Principal component analysis of the 1,564 diagnostic SNP probe sets across all 538 samples.
Supplementary Figure 7. High quality image of the structure plot for K=11.
Supplementary Figure 8. Principal component analysis and geographic location of *P. oocarpa, P. patula* and *P. tecunumanii* samples.
Supplementary Figure 9. Allelic concordance between genotypes originating from global clustering versus per-species clustering runs.
Supplementary Figure 10. Elevation increases as latitude changes from South to North in Mexico and Central America.