


Development of a single nucleotide polymorphism array for population genomic studies in four European pine species

Annika Perry¹  | Witold Wachowiak²  | Alison Downing³ | Richard Talbot³ | Stephen Cavers¹ 

¹UK Centre for Ecology & Hydrology
Edinburgh, Penicuik, UK

²Institute of Environmental Biology, Faculty
of Biology, Adam Mickiewicz University,
Poznań, Poland

³Edinburgh Genomics, Ashworth
Laboratories, University of Edinburgh,
Edinburgh, UK

Correspondence

Witold Wachowiak, Institute of
Environmental Biology, Faculty of Biology,
Adam Mickiewicz University, Uniwersytetu
Poznańskiego 6, 61-614 Poznań, Poland.
Email: witwac@amu.edu.pl

Funding information

Narodowe Centrum Nauki, Grant/Award
Number: UMO-2017/27/B/NZ9/00159;
Natural Environment Research Council,
Grant/Award Number: NE/K012177/1;
PROTREE project, Grant/Award Number:
BB/L012243/1; BBSRC; DEFRA; ESRC;
Forestry Commission; Scottish Government

Abstract

Pines are some of the most ecologically and economically important tree species in the world, and many have enormous natural distributions or have been extensively planted. However, a lack of rapid genotyping capability is hampering progress in understanding the molecular basis of genetic variation in these species. Here, we deliver an efficient tool for genotyping thousands of single nucleotide polymorphism (SNP) markers across the genome that can be applied to genetic studies in pines. Polymorphisms from resequenced candidate genes and transcriptome sequences of *P. sylvestris*, *P. mugo*, *P. uncinata*, *P. uliginosa* and *P. radiata* were used to design a 49,829 SNP array (Axiom_PineGAP, Thermo Fisher). Over a third (34.68%) of the unigenes identified from the *P. sylvestris* transcriptome were represented on the array, which was used to screen samples of four pine species. The conversion rate for the array on all samples was 42% ($N = 20,795$ SNPs) and was similar for SNPs sourced from resequenced candidate gene and transcriptome sequences. The broad representation of gene ontology terms by unigenes containing converted SNPs reflected their coverage across the full transcriptome. Over a quarter of successfully converted SNPs were polymorphic among all species, and the data were successful in discriminating among the species and some individual populations. The SNP array provides a valuable new tool to advance genetic studies in these species and demonstrates the effectiveness of the technology for rapid genotyping in species with large and complex genomes.

KEYWORDS

divergence, genotyping, natural selection, polymorphism, SNP array, speciation

1 | INTRODUCTION

Due to their high ecological and economic value, tree species have been intensively studied to evaluate their genetic diversity,

evolutionary and population history as well as applied areas including conservation and tree breeding (De La Torre et al., 2014; Pyhäjärvi, Kujala, & Savolainen, 2019). Until recently, in nonmodel species, the scope of research has been limited by the lack of accessible genomic

Annika Perry and Witold Wachowiak contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

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

resources that could be used for high-resolution genotyping for population genetic and genomic studies. This is particularly true for species with large and complex genomes, for which genome assembly is particularly challenging (Prunier, Verta, & MacKay, 2016; Zimin et al., 2017).

The draft conifer genomes that have been published so far (e.g. *Picea abies*: Nystedt et al., 2013; *Pinus taeda*: Zimin et al., 2014, 2017; *Pinus lambertiana*: Stevens et al., 2016; *Pseudotsuga menziesii*: Neale et al., 2017; *Abies alba*: Mosca et al., 2019), highlight the complexities involved in working with species within the order: their genomes exceed 20 gigabase pairs and contain transposons and other repetitive content including gene families and pseudogenes. However, developments in high-throughput sequencing methods have allowed insights into genomes of virtually any organism. Whole transcriptome sequencing is a relatively straightforward and informative alternative method (as compared to whole genome sequencing) as it effectively subsamples the genome by sequencing the coding regions alone: reducing both the vast size and complexity of the task. Previous research, involving transcriptome sequencing of *Pinus sylvestris* and several closely related taxa from the *Pinus mugo* complex, has led to the discovery of thousands of polymorphic regions for pines (Wachowiak, Trivedi, Perry, & Cavers, 2015).

Increasing numbers of examples of SNP-based genotyping methods for humans, animals, crops, model plant species and some forest trees (Kathiresan et al., 2009; Kranis et al., 2013; Lepoittevin et al., 2015; Singh et al., 2015) demonstrate their potential to significantly advance genomic studies. Arrays containing thousands of loci provide much higher coverage and resolution of the genome compared with traditional sequencing and PCR-based genotyping methods. However, their usefulness is dependent on the inclusion of markers, which have been carefully chosen to avoid ascertainment bias, either via their distribution across the genome or their frequency within and among populations. A further consideration is the fact that arrays are necessarily confined to genotyping at a preselected subset of all available loci and will only genotype at this selected subset, as well as fail to uncover additional variation—the SNP selection process should therefore be carefully undertaken to optimize the SNP set for the question being asked. However, such tools can provide powerful and cost-efficient alternatives to other genotyping methods, such as genotyping by sequencing and exome capture, which are also computationally expensive and more demanding of bioinformatic expertise by comparison. Furthermore, genotyping arrays can be applied for repeated screening of a common set of SNPs across different experiments and the genomic coverage they offer appears to be ample to allow trait prediction methods such as those employed in ‘breeding without breeding’ and genomic selection approaches (El-Kassaby & Lstibůrek, 2009; Grattapaglia & Resende, 2011) and may be useful in evolutionary studies.

Here, we describe the design and testing of a 49,829 SNP chip (Axiom_PineGAP array) for population genetic and molecular breeding studies in pines, with a focus on Scots pine (*Pinus sylvestris* L.), dwarf mountain pine (*P. mugo* T.), mountain pine (*P. uncinata* R.) and peat-bog pine (*P. uliginosa* N.). These taxa form a monophyletic group

within Pinaceae, but differ strongly in phenotype, geographical distribution and ecology (Wachowiak, Perry, Donnelly, & Cavers, 2018). This, together with their recent speciation history and high phenotypic divergence, mean they are especially suitable for comparative analyses of patterns of polymorphism and divergence and form a useful experimental system for studies of phenotypic trait variation at different ecological and evolutionary timescales (Wachowiak et al., 2015; Wachowiak, Zaborowska, et al., 2018). For the array design, we considered all available genomic resources for these species, including SNPs derived from transcriptomes and resequencing of candidate genes from previous population genetic studies. The aim was to develop a new large scale molecular tool for population genomic analysis of pines and to test its ability to differentiate species by genotyping a sample collection from each of the focal species.

2 | MATERIALS AND METHODS

2.1 | Design of the array

An initial set of 260,262 SNPs were provided to Affymetrix, derived from transcriptome and candidate gene sequences and markers from previous population genetic studies. Each probe consisted of the target SNP plus two (forward and reverse) 35-nt flanking sequences. Filtering to select the final set included blast results to reference genome, Affymetrix recommendations, SNP frequency in a discovery panel and representation of transcripts. Other sequences in the genome similar to the probe, indels and the repetitive nature of large pine genomes may lead to spurious probe or primer binding and consequently reduce the number of converted SNPs. Therefore, for each of the two flanking sequences per SNP, Affymetrix calculated a repetitive variable and a p-convert score. SNPs with more than 300 hits of the flanking sixteen nucleotides between the SNP sequence and reference *P. taeda* genome (v1.0 assembly available at the time of array design, Zimin et al., 2014) were classified as repetitive (T) and removed from the list. Similarly, we rejected a further set of SNPs based on the values of the p-convert score (probability of probe success based on the thermodynamics of the probe and the number of 16-nt matches to the reference genome), classified as either ‘not possible’ or ‘not recommended’ ($0 \leq \text{p-convert} < 0.4$). We avoided inclusion of any singletons. The final array comprised 49,829 SNPs from across the source data sets (Table S1). The vast majority ($N = 49,052$) were obtained from transcriptome sequencing of four pine species: *Pinus sylvestris*, *P. mugo*, *P. uncinata* and *P. uliginosa* (Wachowiak et al., 2015). These included SNPs, which were common to all species and also SNPs fixed in one species and polymorphic within and among others. A further set of SNPs ($N = 578$) were included from candidate genes ($N = 279$), which had been resequenced in previous population genetic studies of the pine species (Kujala & Savolainen, 2012; Mosca, Eckert, Di Pierro, et al., 2012; Palmé, Wright, & Savolainen, 2008; Wachowiak, Balk, & Savolainen, 2009; Wachowiak, Zaborowska, et al., 2018). Variation in mitochondrial DNA (mtDNA) was targeted using a set of SNPs ($N = 14$), which had been discovered by Donnelly et al. (2017). A

TABLE 1 Conversion rates for single nucleotide polymorphism (SNP) array using thresholds set by high-quality samples ($N = 529$)

Source of SNPs	N	Successfully converted (%)			Not converted (%)			Ref
		MHR	NMH	PHR	CRBT	Other	OTV	
Candidate genes	578	10.6	11.4	23.2	8.5	45.9	0.5	1–6
<i>Pinus radiata</i> transcriptome	185	40.5	3.2	0.00	1.6	53.5	1.1	7
mtDNA	14	21.4	0.00	14.3	0.0	57.1	7.1	8
<i>Pinus sylvestris</i> transcriptome	49,052	10.0	9.3	22.4	7.3	49.5	1.6	9
Total	49,829	10.1	9.3	22.3	7.2	49.5	1.6	

Note: N, count of SNPs. Conversion types: MHR, monomorphic high resolution; NMH, no minor homologue; PHR, polymorphic high resolution; CRBT, call rate below threshold (96%); OTV, off target variant. References: 1, Garcia-Gil, Mikkonen, and Savolainen (2003); 2, Wachowiak et al. (2009); 3, Palmé et al. (2008); 4, Kujala and Savolainen (2012); 5, Mosca, Eckert, Liechty, et al. (2012); 6, Wachowiak, Zaborowska, et al. (2018); 7, ENA accession numbers ERS1034542-53; 8, Donnelly et al. (2017); 9, Wachowiak et al. (2015).

set of SNPs putatively associated with susceptibility to Dothistroma needle blight (discovered in *Pinus radiata*, ENA accession numbers ERS1034542-53) were also included ($N = 185$). SNPs derived from transcriptome sequences (Wachowiak et al., 2015) were assessed to determine how well they represented variation across *Pinus* unigenes. During transcriptome assembly, contigs were aligned into 40,968 unigenes, of which 40,798 were considered to be high quality (not containing putative retrotransposon sequences). The proportion of unigenes included on the array and within the successfully converted SNP set were determined. The representation of different gene ontology (GO) classifications on the array and within the converted SNP set was also assessed in comparison with the transcriptome. After a new assembly of the *P. taeda* reference genome became available (GCA_000404065.3_Ptaeda2.0, Zimin et al., 2017), we used the SNP sequences (71-mers) included on the array as blast queries. A blast hit was counted where there was a minimum 95% match between the 71-nt SNP probe and the reference genome. A high proportion of hits to the *P. taeda* genome in SNPs which were not successfully converted may indicate that these failed due to, for example: nontarget hybridization; undetected SNPs in the flanking region; and the presence of a third allele. In contrast, a low proportion of hits to the *P. taeda* genome in nonconverted SNPs may indicate that the target sequence spanned an intron, for example, or that they were simply missing from the genome assembly.

2.2 | Source of samples and DNA extraction

The technology demanded an initial commitment of 5 x 384 samples for development of the array. This full sample set of 1920 was genotyped, and all successful genotypes were used in optimization of SNP calling filters (see next paragraph). Here, we report full details of array testing using a subset of 87 samples, which included four species of pine (*Pinus sylvestris*: SY; *P. mugo*: MU; *P. uncinata*: UN; *P. uliginosa*: UL; Tables S2–S4). DNA was extracted from needles using

a Qiagen DNeasy 96 kit following the manufacturer's instructions. Needles were dried on silica gel prior to extraction and DNA was quantified using a Qubit spectrophotometer to ensure a minimum standardized concentration of 35 ng/μl. The quality of genomic DNA was also checked visually for fragmentation on 1% agarose gel.

2.3 | Genotyping and SNP calling

Genotyping was done at Edinburgh Genomics following DNA amplification, fragmentation, chip hybridization, single-base extension through DNA ligation and signal amplification performed according to the Affymetrix Axiom[®] Assay protocol. Genotyping was performed in 384-well format on a GeneTitan according to the manufacturer's procedure. Genotype calls were performed using Axiom Analysis Suite software as recommended by the manufacturer (Thermo Fisher). In order to both maximize the number of samples included in subsequent analyses and minimize the distortive effect of poor quality samples on genotype calls, three separate analyses were performed (Table S5). Samples were assigned to an analysis group based on their call rate (CR) and dish QC (DQC: a metric provided by Thermo Fisher which is generated by measuring signals at multiple sites in the genome known not to vary among individuals), using the following thresholds: DQC 'high' ≥ 0.82 ; DQC 'low' < 0.82 ; CR 'high' ≥ 96 ; CR 'low' < 96 . Analyses: (a) DQC high + CR high; (b) DQC high + CR low; (c) DQC low + CR low. High-quality samples ($N = 529$), with high CR and DQC, were used to set thresholds for allele calls. Posteriors for allele calls were subsequently used as priors for analyses 2 ($N = 753$) and 3 ($N = 251$).

2.4 | Testing of the array and statistical analyses

To test the array, a subset of genotyped samples of each species ($N = 87$) were analysed, including five populations per species (three populations for UL). Each population comprised 3–5 individuals,

each from different mother trees (Tables S2 and S4). SNP type (polymorphic, monomorphic, CR < 80%) was compared among species to identify the proportion of SNPs, which could be analysed further. Principal component analysis for all samples across all species was performed in TASSEL (Bradbury et al., 2007). Separate analyses were performed using all available SNPs ($N = 20,795$) and only common (mean allelic frequency, MAF > 0.1) SNPs, which were polymorphic among all species ($N = 2,358$). The first two principal components from each analysis were plotted to qualitatively assess the extent of variation within and among species.

3 | RESULTS

3.1 | Conversion rates for SNPs on the array

In total, 49,237 (over 98%) SNPs on the array represented genomic regions from pine transcriptome sequencing, and the remaining 1.19% included SNPs from candidate genes and mitochondrial regions (Table 1; Table S1). Unique hits to the GCA_000404065.3_Ptaeda2.0 reference genome were found for the majority of the assayed probes ($N = 39,863$). Similarly, almost 80% of probes representing SNPs, which were not successfully converted had a unique hit to the reference genome ($N = 23,106$). Sixty-two probes developed based on the nucleotide sequence of candidate genes could not be found in the *P. taeda* genome. The conversion rate for the array on all genotyped

samples, using thresholds set by high-quality samples, was 41.73% ($N = 20,795$ SNPs; Table 1), and it was similar for SNPs derived from different sources (although lower conversion rates were observed for SNPs from *mtDNA* compared with SNPs from published genes).

Over a third (34.68%) of the unigenes identified from the *P. sylvestris* transcriptome were represented on the array with an average of 3.47 SNPs per unigene (range: 1–40 SNPs per unigene). Nearly a quarter (23.67%) of the unigenes identified from the *P. sylvestris* transcriptome included SNPs which were successfully converted, with an average 2.12 SNPs per unigene (range: 1–19 SNPs per unigene). The representation of GO terms in unigenes containing converted SNPs was more comparable to that of their coverage across the transcriptome than on the full array (Figure 1), for example: metabolic process, regulation of biological process and response to stimulus were overrepresented in the full array while cell death, cell communication and membrane were underrepresented. The mean percentage difference in GO term representation among the transcriptome and full array was 1.27% (range: 0.03%–6.84%) whereas the mean percentage difference in representation of GO terms among the transcriptome and converted SNPs was 0.11% (range 0.01%–0.33%).

3.2 | Testing of the array

Over a quarter of successfully converted SNPs were polymorphic among all species ($N = 5,665$, Table 2; Table S6), of which over

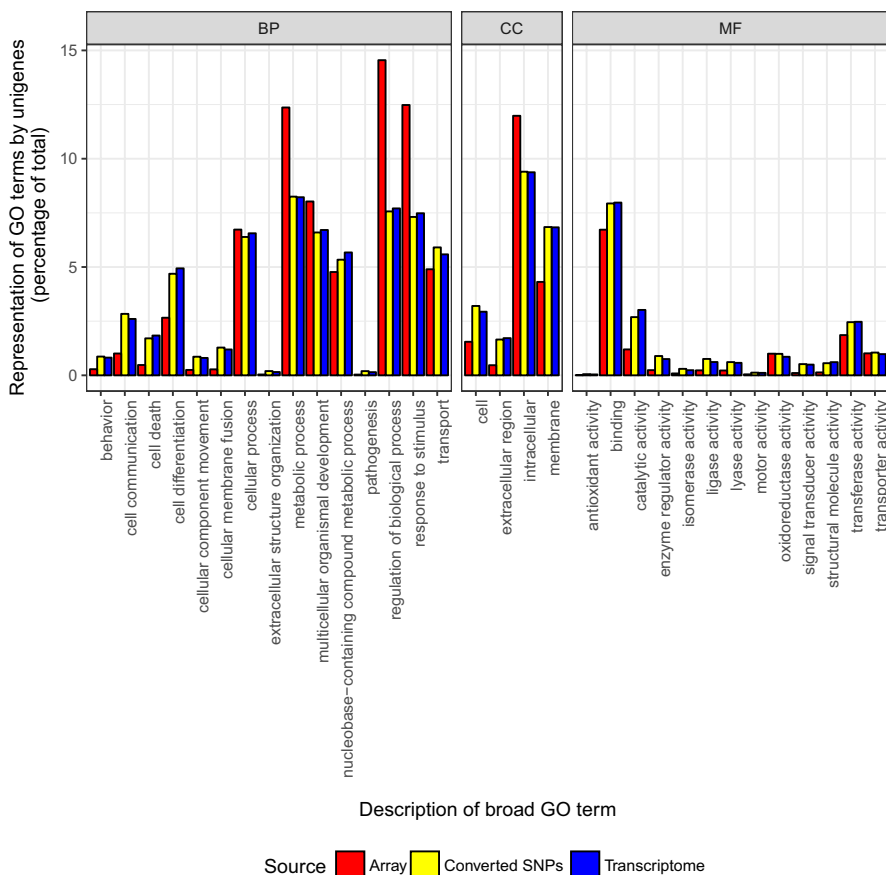
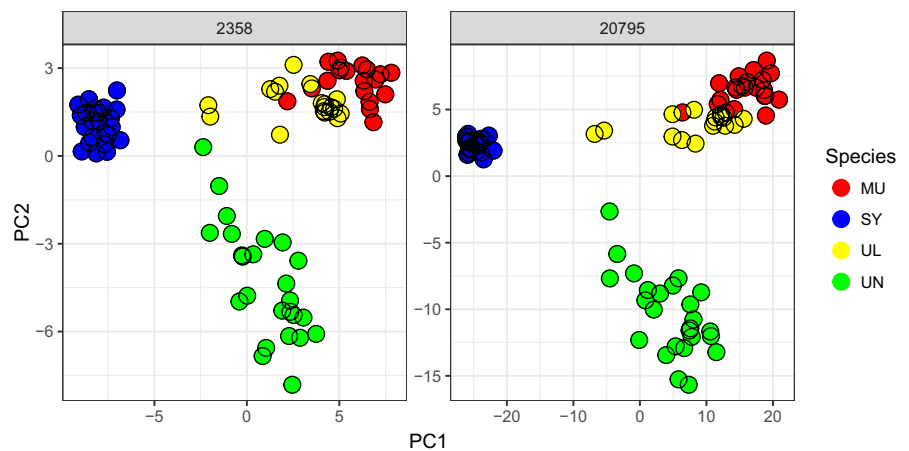


FIGURE 1 Gene ontology (GO) classification of unigenes and their relative representation (count of unigenes assigned to each GO term as a percentage of total unigene count from each source) across the *Pinus sylvestris* transcriptome (Wachowiak et al., 2015), the full array and successfully converted single nucleotide polymorphisms. Classifications: BP, biological processes; CC, cell component; MF, molecular function [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Single nucleotide polymorphism type (CR < 80, call rate < 80%; Mono, monomorphic; Poly, polymorphic) among four pine species (MU, *P. mugo*; SY, *P. sylvestris*; UN, *P. uncinata*; UL, *P. uliginosa*)

SY	UN	MU: CR < 80			MU: Mono			MU: Poly		
		UL: CR < 80	UL: Mono	UL: Poly	UL: CR < 80	UL: Mono	UL: Poly	UL: CR < 80	UL: Mono	UL: Poly
CR < 80	CR < 80									3
	Mono					2		1		1
	Poly	2	1	31		1	1	6	5	68
Mono	CR < 80	15	14	89		6	3	12	6	78
	Mono	25	184	156	6	4,950	91	14	431	330
	Poly	113	90	1,179	2	198	80	116	302	2,482
Poly	CR < 80	14	7	126				1	14	7
	Mono		42	64	2	174	54	2	111	155
	Poly	102	144	1,939	2	120	324	114	407	5,665

FIGURE 2 Principal components 1 (x-axes) and 2 (y-axes) from PCAs based on diversity at 2,358 polymorphic common single nucleotide polymorphism (SNPs; MAF > 0.1, left panel) and all 20,795 SNPs (right panel). Species: MU, *Pinus mugo*; SY, *P. sylvestris*; UL, *P. uliginosa*; UN, *P. uncinata* [Colour figure can be viewed at wileyonlinelibrary.com]

40% were also common (MAF > 0.1) in all species ($N = 2,358$). Nearly a further quarter of all converted SNPs were monomorphic among all species ($N = 4,950$) or had a low call rate in at least one species ($N = 4,921$). About two thirds of successfully converted SNPs had call rates above 90% (Table S6). A total of 3,293 SNPs were fixed in one species and polymorphic in the other three, although the majority ($N = 2,482$) were fixed in *P. sylvestris* and polymorphic in *P. mugo*, *P. uncinata* and *P. uliginosa*. Numbers of polymorphic SNPs were highest in *P. uncinata* ($N = 13,494$) and *P. uliginosa* ($N = 13,031$) and lowest in *P. sylvestris* ($N = 9,701$) and *P. mugo* ($N = 10,441$); however, the mean proportion of heterozygous SNPs per sample was highest for *P. mugo* (0.244) compared with other species (*P. sylvestris* = 0.123; *P. uliginosa* = 0.216; *P. uncinata* = 0.206). Numbers of SNPs with CR < 80% were low for all species (N : *P. sylvestris* = 122; *P. uncinata* = 506; *P. uliginosa* = 562) except *P. mugo* ($N = 4,337$).

In order to test the SNPs, species identity and relationships among and within species were examined. The analysis separated the 87 individuals into three major groups (Figure 2): *P. sylvestris* and *P. uncinata* were well defined and separate, whereas *P. mugo* and *P. uliginosa* formed adjacent, partially overlapping clusters. A significant percentage of the variation among species and

populations, using all available SNPs and for a subset of common polymorphic SNPs, was explained by principal component axes 1 and 2 (20,795 SNPs: PC1 = 23.41%, PC2 = 4.24%; 2,358 SNPs: PC1 = 12.12%, PC2 = 3.68%). There was little difference between the distribution of species when including all SNPs or a subset of common polymorphic SNPs. PC1 primarily explained variation between *P. sylvestris* and the *P. mugo* complex and variation within the latter, while PC2 primarily explained variation between *P. uncinata* and the rest. *Pinus sylvestris* showed less variation overall than the other species. Separate principal components analyses for each species in isolation, using common (MAF > 0.1) polymorphic SNPs, showed evidence for differentiation among populations in all species (Figure S1).

4 | DISCUSSION

The design of this genotyping array and its successful testing in a small number of populations across four related pine species demonstrates that the method can be effective in organisms with large and complex genomes. The design and testing of the array were necessarily constrained by the availability of material—both plant tissue

and genetic references—and time. Ideally, the inclusion of parent-progeny pairs in a prescreening step for SNP discovery and selection would allow testing for Mendelian segregation (Chen et al., 2014). Furthermore, genotyping biological and experimental replicates would allow the array's value to be empirically tested by assessing the reproducibility of the markers. Even though our relatively small sample size, lack of suitable technical and biological replicates and predominance of previously untested novel markers do not allow for full validation of the array performance (Bianco et al., 2016), a significant number of high-quality SNPs were successfully genotyped, representing over 40% of the assayed loci. As expected, the conversion rate observed in the pine array was not as high as in model species or species with smaller, more extensively studied genomes such as crop species (>90% in alfalfa bean: Li et al., 2014; rice: Singh et al., 2015; cotton: Cai, Zhu, Zhang, & Guo, 2017; pigeon pea: Saxena et al., 2018; ~75% in maize: Unterseer et al., 2014 and groundnut: Pandey et al. 2017). The reported conversion rate in some animal species was lower (~60% in shellfish: Lapègue et al., 2014; chicken: Kranis et al., 2013). High-density SNP arrays developed in other tree species showed similar conversion rates (around 60%–70% in *Pinus taeda*: Eckert et al., 2009; *Pinus pinaster*: Plomion et al., 2016; *Pseudotsuga menziesii*: Howe et al., 2013, 2020; *Picea glauca*: Pavy et al., 2013) depending on the source of SNPs for assay development and the type of genotyping platform.

The moderate conversion rate for the pine array was probably due to a combination of factors. The use of transcriptome sequences for SNP discovery, which aimed to identify markers in coding regions, may have resulted in the inclusion of exon–intron boundaries or unknown polymorphisms within oligonucleotide binding regions. Duplication and homology within flanking regions are also likely for a proportion of the unconverted SNPs, despite measures to control for this during the array design by using the first assembly of *Pinus taeda* genome (Zimin et al., 2014) to identify loci within duplicated or paralogous regions. As about 80% of successfully converted SNPs had unique hits to the novel assembly of the *P. taeda* genome (Zimin et al., 2017), we assume that the conversion rate may be partially affected by nontarget hybridization. In addition, the predominance of Scots pine transcriptome in initial SNP selection and the subsequent sample set used to set thresholds for allele calls may have compromised the conversion rate to some extent, viewed across all species. However, the use of multiple species in designing and testing the array has the major advantage of increasing the potential applications of the array (for example, the same array may be used for both multiple and single species assays) and its subsequent relevance to a broader range of the scientific community. Although the conversion rate of SNPs derived from *P. radiata* (putative *Dothistroma*-associated SNPs) was limited and included no polymorphic SNPs, this species is not phylogenetically close to *P. sylvestris* (Gernandt, Lopez, Garcia, & Liston, 2005). Therefore, in future iterations of the array, other pines from the same section which are frequently studied due to their ecological and/or economic importance, such as *P. nigra*, *P. halepensis*, *P. pinaster* or *P. pinea*, may be better candidates if strong trait-linked SNPs in these species have been identified.

Disadvantages of using multiple species include the effects of ascertainment bias and genomic divergence: for example, the sample set used to set thresholds for allele calls included only 0.57% *P. mugo* samples, which subsequently had extremely high levels of SNPs with low call rates compared with the other species. It is unknown whether this was a cause or effect: there were similarly low levels of *P. uliginosa* in the high-quality sample set but the level of SNPs with low call rates was not equivalently high, although this could be due to the genetic relatedness of *P. uliginosa* and *P. sylvestris* (Wachowiak et al., 2011).

Overall, the SNP conversion rate was similar for SNPs derived from different sources including transcriptomes, resequenced genes and mtDNA data variants. The array, and particularly the set of converted SNPs, represented the pine transcriptome (and therefore, putatively, the genome) well: unigenes identified in transcriptome sequencing were well represented, as were the range and relative proportion of associated gene functions. Although a published reference genome for the studied species is not yet available, it is known that there is rapid decay of linkage disequilibrium in conifers (Brown, Gill, Kuntz, Langley, & Neale, 2004; Wachowiak et al., 2009); therefore, the set of converted SNPs identified in this study are likely to represent unique mutations in individual genes. These findings support the use of SNP arrays in studies focussing on local adaptation and association genetics. Breeding studies are also increasingly looking to marker-assisted selection (Isik, 2014) to inform selection for trait improvement. These studies require well-phenotyped individuals and a large pool of SNPs from across the genome with which to identify putatively associated markers. A basic understanding of their putative function is also important to put results in context. *Pinus sylvestris*, which is both economically and ecologically important, is a particularly suitable candidate for these studies: quantitative traits, such as growth, phenology and disease resistance, are commonly measured in progeny-provenance trials of this species and there would be considerable commercial value in the identification of markers linked to key traits.

The converted SNPs were found to reflect results from previous studies on inter- and intraspecific relationships in the focal species which, despite morphological, geographical and ecological differentiation, show high genetic similarity based on biometric, biochemical (monoterpenes, isozymes) and molecular markers (Boratyńska & Boratyński, 2007; Lewandowski et al., 2000; Wachowiak, Boratyńska, & Cavers, 2013). The presence of a high number of shared SNPs among the species, as found in comparative transcriptome sequencing (Wachowiak et al., 2015) and candidate gene studies (Wachowiak, Zaborowska, et al., 2018), confirms their close genetic relationships and common ancestry. Considering their recent divergence (Wachowiak et al., 2011), a significant proportion of converted SNPs (~25%) were polymorphic among all four pine species, or fixed in one species and polymorphic in the others. The percentage of converted SNPs, which were polymorphic within a single species ranged from over a quarter (*P. sylvestris*) to nearly half (*P. uliginosa*), comparable with results published for *P. pinaster* (47%: Plomion et al., 2016). The amount and frequency of SNPs within and

among species reported here demonstrate the efficacy and potential applications of the array to genetic studies in pines focussing on evolutionary history and population genetics, while the number of polymorphic loci would be expected to increase, potentially improving resolution further, given a larger number of samples and populations.

Under multivariate analysis, three major clusters were identified (with similar resolution when including all available SNPs or only a subset of common polymorphic SNPs), which corresponded to *P. sylvestris*, *P. uncinata* and *P. mugo*/*P. uliginosa*. The position of the *P. uliginosa* samples, between *P. sylvestris* and *P. mugo* but more proximate to the latter, has been observed in previous studies (Wachowiak, Żukowska, Wójkiewicz, Cavers, & Litkowiec, 2016). Although the number of samples from individual populations was limited, the analysis showed the array's potential for discriminating populations at broad spatial scales. Within species, populations showed clear structure, while those from *P. uliginosa* appeared to be most clearly diverged, supporting results from a recent study using *mtDNA* markers (maternally inherited in pines; Łabiszak, Zaborowska, & Wachowiak, 2019). There was evidence for distinct population clusters in other species for highly diverged populations at the species' margins (including *P. sylvestris* in Scotland, *P. mugo* from the Abruzzi Mountains in Italy and *P. uncinata* from Castiello de Jaca in Spain).

The design and testing of the pine array demonstrate its potential application to a range of studies including, but not limited to population genetics, evolutionary history, conservation management, local adaptation, association genetics and tree breeding. The array is the highest resolution molecular tool developed to date for the focal pine species and has significant potential to further understanding of pine genetics and genomics.

ACKNOWLEDGMENTS

This work was financially supported primarily by two grants in the UK: GAPII (NE/K012177/1), funded by NERC, and PROTREE (BB/L012243/1), funded jointly by BBSRC, DEFRA, ESRC, the Forestry Commission, NERC and the Scottish Government, under the Tree Health and Plant Biosecurity Initiative. WW acknowledges financial support from Polish National Science Centre (UMO-2017/27/B/NZ9/00159).

AUTHOR CONTRIBUTIONS

The research was designed and planned by WW and SC; SNP selection for the array, data analysis and manuscript writing were performed by AP and WW; SNP genotyping was performed by AD and RT; Manuscript review and revision and final approval of manuscript were performed by all authors.

DATA AVAILABILITY STATEMENT

The data sets supporting the results of this article are freely available through the NERC's Environmental Information Data Centre, as follows: (a) Scots pine SNP for Axiom array (Table S1): <https://doi.org/10.5285/cbaa464a-ac18-42bf-8518-c746d8d97270>. (b) Scots

pine SNP genotypes for Axiom array validation (Table S3): <https://doi.org/10.5285/7ee55609-d6b1-4693-8b36-2bf84fef76c2>. (c) Conversion rate, frequency and state of single nucleotide polymorphism loci for Scots pine Axiom microarray (Table S6): <https://doi.org/10.5285/0ba33e96-67cb-4650-b2bd-6ee13fa7de97>

ORCID

Annika Perry  <https://orcid.org/0000-0002-7889-7597>

Witold Wachowiak  <https://orcid.org/0000-0003-2898-3523>

Stephen Cavers  <https://orcid.org/0000-0003-2139-9236>

REFERENCES

- Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., ... Troglio, M. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant Journal*, *86*, 62–74. <https://doi.org/10.1111/tpj.13145>
- Boratyńska, K., & Boratyński, A. (2007). Taxonomic differences among closely related pines *Pinus sylvestris*, *P. mugo*, *P. uncinata*, *P. rotunda* and *P. uliginosa* as revealed in needle sclerenchyma cells. *Flora*, *202*(7), 555–569.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., & Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(42), 15255–15260. <https://doi.org/10.1073/pnas.0404231101>
- Cai, C., Zhu, G., Zhang, T., & Guo, W. (2017). High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics*, *18*(1), 654. <https://doi.org/10.1186/s12864-017-4062-2>
- Chen, N., Van Hout, C. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian inheritance to improve high-throughput SNP discovery. *Genetics*, *198*(3), 847–857. <https://doi.org/10.1534/genetics.114.169052>
- De La Torre, A. R., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J. M., ... Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, *166*(4), 1724–1732. <https://doi.org/10.1104/pp.114.248708>
- Donnelly, K., Cottrell, J., Ennos, R. A., Vendramin, G. G., A'Hara, S., King, S., ... Cavers, S. (2017). Reconstructing the plant mitochondrial genome for marker discovery: A case study using *Pinus*. *Molecular Ecology Resources*, *17*(5), 943–954. <https://doi.org/10.1111/1755-0998.12646>
- 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 & Genomes*, *5*(1), 225–234. <https://doi.org/10.1007/s11295-008-0183-8>
- El-Kassaby, Y. A., & Lstibůrek, M. (2009). Breeding without breeding. *Genetics Research*, *91*(2), 111–120. <https://doi.org/10.1017/S001667230900007X>
- García-Gil, M. R., Mikkonen, M., & Savolainen, O. (2003). Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Molecular Ecology*, *12*(5), 1195–1206.
- Gernandt, D. S., Lopez, G. G., Garcia, S. O., & Liston, A. (2005). Phylogeny and classification of *Pinus*. *Taxon*, *54*(1), 29–42.
- Grattapaglia, D., & Resende, M. (2011). Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, *7*, 241–255. <https://doi.org/10.1007/s11295-010-0328-4>

- Howe, G. T., Yu, J., Knaus, B., Cronn, R., Kolpak, S., Dolan, P., ... Dean, J. F. D. (2013). A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. *BMC Genomics*, *14*, 137. <https://doi.org/10.1186/1471-2164-14-137>
- Howe, G. T., Jayawickrama, K., Kolpak, S. E., Jennifer Kling, J., Trappe, M., Hipkins, V., ... McEvoy, S. (2020). An axiom SNP genotyping array for Douglas-fir. *BMC Genomics*, *21*, 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*, *45*(3), 379–401.
- Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., ... Altshuler, D. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, *41*(3), 334–341. <https://doi.org/10.1038/ng.327>
- Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., ... Burt, D. W. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics*, *14*(1), 59. <https://doi.org/10.1186/1471-2164-14-59>
- Kujala, S., & Savolainen, O. (2012). Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genetics & Genomes*, *8*(6), 1451–1467.
- Lapègue, S., Harrang, E., Heurtebise, S., Flahauw, E., Donnadiéu, C., Gayral, P., ... Klopp, C. (2014). Development of SNP-genotyping arrays in two shellfish species. *Molecular Ecology Resources*, *14*(4), 820–830. <https://doi.org/10.1111/1755-0998.12230>
- Lepoittevin, C., Bodénès, C., Chancerel, E., Villate, L., Lang, T., Lesur, I., ... Kremer, A. (2015). Single-nucleotide polymorphism discovery and validation in high-density SNP array for genetic analysis in European white oaks. *Molecular Ecology Resources*, *15*(6), 1446–1459. <https://doi.org/10.1111/1755-0998.12407>
- Lewandowski, A., Boratyński, A., & Mejnartowicz, L. (2000). Allozyme investigations on the genetic differentiation between closely related pines *Pinus sylvestris*, *P. mugo*, *P. uncinata* and *P. uliginosa* (Pinaceae). *Plant Systematics and Evolution*, *221*(1), 15–24. <https://doi.org/10.1007/bf01086377>
- Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., ... Brummer, E. C. (2014). Development of an Alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS ONE*, *9*(1), e84329. <https://doi.org/10.1371/journal.pone.0084329>
- Łabiszak, B., Zaborowska, J., & Wachowiak, W. (2019). Patterns of mtDNA variation reveal complex evolutionary history of relict and endangered peat bog pine (*Pinus uliginosa*). *AoB PLANTS*, *11*(2), plz015. <https://doi.org/10.1093/aobpla/plz015>
- Mosca, E., Cruz, F., Gómez-Garrido, J., Bianco, L., Rellstab, C., Brodbeck, S., ... Neale, D. B. (2019). A reference genome sequence for the European silver fir (*Abies alba* Mill.): A community-generated genomic resource. *G3: Genes, Genomes Genetics*, *9*(7), 2039–2049. <https://doi.org/10.1534/g3.119.400083>
- Mosca, E., Eckert, A. J., Di Pierro, E. A., Rocchini, D., La Porta, N., Belletti, P., & Neale, D. B. (2012). The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. *Molecular Ecology*, *21*(22), 5530–5545. <https://doi.org/10.1111/mec.12043>
- Mosca, E., Eckert, A. J., Liechty, J. D., Wegrzyn, J. L., La Porta, N., Vendramin, G. G., & Neale, D. B. (2012). Contrasting patterns of nucleotide diversity for four conifers of Alpine European forests. *Evolutionary Applications*, *5*(7), 762–775. <https://doi.org/10.1111/j.1752-4571.2012.00256.x>
- Neale, D. B., McGuire, P. E., Wheeler, N. C., Stevens, K. A., Crepeau, M. W., Cardeno, C., ... Wegrzyn, J. L. (2017). The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3: Genes, Genomes, Genetics*, *7*(9), 3157–3167. <https://doi.org/10.1534/g3.117.300078>
- Nystedt, B., Street, N., Wetterbom, A., Zuccolo, A., Lin, Y., Scofield, D., ... Sena, J. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*, 579–584.
- Palmé, A. E., Wright, M., & Savolainen, O. (2008). Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. *Molecular Biology and Evolution*, *25*(12), 2567–2577. <https://doi.org/10.1093/molbev/msn194>
- Pandey, M., Agarwal, G., Kale, S., Clevenger, J., Nayak, S. N., Sriswathi, M., ... Varshney, R. K. (2017). Development and evaluation of a high density genotyping 'Axiom_Arachis' array with 58 K SNPs for accelerating genetics and breeding in Groundnut. *Scientific Reports*, *7*, 40577. <https://doi.org/10.1038/srep40577>
- Pavy, N., Gagnon, F., Rigault, P., Blais, S., Deschênes, A., Boyle, B., ... Bousquet, J. (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources*, *13*(2), 324–336. <https://doi.org/10.1111/1755-0998.12062>
- Plomion, C., Bartholomé, J., Lesur, I., Boury, C., Rodríguez-Quilón, I., Lagrault, 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>
- Prunier, J., Verta, J.-P., & MacKay, J. J. (2016). Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. *New Phytologist*, *209*(1), 44–62. <https://doi.org/10.1111/nph.13565>
- Pyhäjärvi, T., Kujala, S. T., & Savolainen, O. (2019). 275 years of forestry meets genomics in *Pinus sylvestris*. *Evolutionary Applications*. <https://doi.org/10.1111/eva.12809>
- Saxena, R. K., Rathore, A., Bohra, A., Yadav, P., Das, R. R., Khan, A. W., ... Varshney, R. K. (2018). Development and application of high-density Axiom Cajanus SNP array with 56K SNPs to understand the genome architecture of released cultivars and founder genotypes. *The Plant Genome*, *11*. <https://doi.org/10.3835/plantgenome2018.01.0005>
- Singh, N., Jayaswal, P. K., Panda, K., Mandal, P., Kumar, V., Singh, B., ... Singh, N. K. (2015). Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. *Scientific Reports*, *5*, 11600.
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., ... Langley, C. H. (2016). Sequence of the Sugar pine megagenome. *Genetics*, *204*(4), 1613–1626. <https://doi.org/10.1534/genetics.116.193227>
- Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., ... Schön, C. C. (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>
- Wachowiak, W., Balk, P., & Savolainen, O. (2009). Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genetics & Genomes*, *5*(1), 117–132.
- Wachowiak, W., Boratyńska, K., & Cavers, S. (2013). Geographical patterns of nucleotide diversity and population differentiation in three closely related European pine species in the *Pinus mugo* complex. *Botanical Journal of the Linnean Society*, *172*(2), 225–238.
- Wachowiak, W., Palme, A. E., & Savolainen, O. (2011). Speciation history of three closely related pines *Pinus mugo* (T.), *P. uliginosa* (N.) and *P. sylvestris* (L.). *Molecular Ecology*, *20*(8), 1729–1743.
- Wachowiak, W., Perry, A., Donnelly, K., & Cavers, S. (2018). Early phenology and growth trait variation in closely related European pine species. *Ecology and Evolution*, *8*(1), 655–666. <https://doi.org/10.1002/ece3.3690>
- Wachowiak, W., Trivedi, U., Perry, A., & Cavers, S. (2015). Comparative transcriptomics of a complex of four European pine species. *BMC Genomics*, *16*(1), 234.
- Wachowiak, W., Zaborowska, J., Łabiszak, B., Perry, A., Zucca, G., González-Martínez, S., & Cavers, S. (2018). Molecular signatures of

- divergence and selection in closely related pine taxa. *Tree Genetics & Genomes*, 14(6), 83. <https://doi.org/10.1007/s11295-018-1296-3>
- Wachowiak, W., Żukowska, W. B., Wójkiewicz, B., Cavers, S., & Litkowiec, M. (2016). Hybridization in contact zone between temperate European pine species. *Tree Genetics & Genomes*, 12(3), 48. <https://doi.org/10.1007/s11295-016-1007-x>
- Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marçais, G., ... Langley, C. H. (2014). Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, 196(3), 875–890. <https://doi.org/10.1534/genetics.113.159715>
- Zimin, A., 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>

SUPPORTING INFORMATION

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

How to cite this article: Perry A, Wachowiak W, Downing A, Talbot R, Cavers S. Development of a single nucleotide polymorphism array for population genomic studies in four European pine species. *Mol Ecol Resour.* 2020;20:1697–1705. <https://doi.org/10.1111/1755-0998.13223>