Identifying and testing marker-trait associations for growth and phenology in three pine species

Annika Perry[1*], Witold Wachowiak[2], Joan Beaton[3], Glenn Iason[3], Joan Cottrell[4], Stephen Cavers[1]

1. UK Centre for Ecology & Hydrology Edinburgh, Penicuik, Midlothian, EH26 0QB, UK
2. Institute of Environmental Biology, Faculty of Biology, Adam Mickiewicz University in Poznań, Poland
3. James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK
4. Northern Research Station, Forest Research, Roslin, EH25 9SY, UK

* Corresponding author: annt@ceh.ac.uk

Abstract

Identifying the additive genetic variation underlying complex traits is important for species of economic and/or ecological value. In particular, where DNA markers can be associated with trait variation they can be used to develop models to predict phenotypes as the basis of future selection and conservation programmes. Here, SNPs associated with growth (height and annual increment) and phenology (budburst and bud set) were identified in three closely related pine species including *Pinus sylvestris* (Scots pine). A genotyping array was used to screen 20,795 SNPs from coding regions for their association with trait variation using mixed linear model (MLM) and multilocus mixed model (MLMM) approaches: 113 SNPs located at 111 loci were significantly associated with the traits, with the majority associated with either budburst or growth increment in *P. sylvestris*. Common SNPs (MAF > 0.05) identified as significantly associated with bud set were found in genes putatively involved in only growth

and development, whereas SNPs associated with growth and budburst were located in genes putatively involved in growth and development, response to environment and, to a lesser extent, reproduction. Predicted values estimated using the model for growth had highly significant correlations with phenotypes quantified in a *P. sylvestris* common environment experiment established at two sites in Scotland (YA and GS), but only at one of the sites (YA, height at 2020: r = 0.376, $p < 0.001$). Predicted values estimated with the model for budburst were found to be weakly but significantly correlated with duration of budburst at one of the field sites (GS, duration at 2018: r = 0.242, $p = 0.012$) and negatively associated with timing of budburst at the other (YA, stage six: r = -0.216, $p = 0.033$). Genomic prediction using the model for growth was more successful than random selection as a method of selecting tall trees at both sites. This study provides tentative support for the development of prediction models for traits that are of interest to both foresters and conservationists, while highlighting the need for caution when applying them to trees growing in different environments.

Keywords

Marker-trait association; predictive model; genetic variation; local adaptation, common garden trial; quantitative traits; SNP array; Scots pine; *Pinus mugo* complex

1. Introduction

A primary goal of association genetics studies of trees is to accelerate research, in what are typically very long-lived organisms, by developing a capability to predict phenotype from genotype. However, phenotypic traits are mostly complex, i.e. quantitative and controlled by many genes (Goddard and Hayes, 2009), and may vary in expression and heritability depending on the environment in which they are assessed (Schlichting, 1986). Therefore, a combination of a high number of markers applied to a large number of samples and well-

46 assessed phenotypes, ideally from multiple environments, are required to develop robust
47 predictive models. The popularity and power of genetic association studies continues to grow
48 thanks to improvements in the scale, quality and cost of high-throughput sequencing and
49 genotyping. In particular, the accessibility of cost-effective high-throughput genotyping has
50 benefited those studying nonmodel organisms for which genome assembly is challenging due
51 to genome size and/or complexity (Prunier et al., 2016, Zimin et al., 2017).

52 Pines are among the most important commercial forest tree species in the world (Kanninen,
53 2010), and have high ecological value in forests across the northern hemisphere.
54 Understanding the genetic architecture of key adaptive traits such as growth, form, disease
55 resistance and phenology is of interest to a range of stakeholders that include the forestry
56 industry and conservationists. Due to their large size and complexity, pine genomes are
57 particularly challenging to assemble, and this has only been achieved for loblolly pine (*Pinus*
58 *taeda*; Zimin et al., 2014) and sugar pine (*Pinus lambertiana*; Stevens et al., 2016), which are
59 among the largest genomes ever sequenced and assembled. However, thousands of
60 polymorphic regions potentially suitable for use in high-throughput genotyping for
61 association studies in pine have been discovered using high-throughput sequencing methods
62 including whole transcriptome studies (Blanca et al., 2012, Chancerel et al., 2011, Durán et
63 al., 2019, Geraldes et al., 2011, Liu et al., 2014, Parchman et al., 2010, Trick et al., 2009,
64 Wachowiak et al., 2015).

65 Using genome-wide DNA markers and their estimated effects to predict breeding values was
66 first proposed by (Meuwissen et al., 2001) who found that selection based on this method
67 could significantly increase the rate of genetic gain in subsequent generations. Since then, a
68 large number of studies have focussed on the development of prediction models and their

accuracy in predicting phenotypes, with a focus on wood or fruit quality in tree species (Kumar et al., 2012, Minamikawa et al., 2017, Muranty et al., 2015, Beaulieu et al., 2014, Isik et al., 2016, Resende et al., 2012a, Resende et al., 2012b, Thistlethwaite et al., 2017), although there are also efforts to develop genomic selection methods for disease resistance traits (Westbrook et al., 2020, Stocks et al., 2019). Association studies and tests of the strength of genomic prediction using the associated single nucleotide polymorphism (SNP) markers have been performed in a small number of pine species for a few traits including serotiny (*Pinus pinaster*, Budde et al., 2014), circumference, height and stem straightness (*Pinus pinaster*, Bartholomé et al., 2016), oleoresin flow (*Pinus taeda*, Westbrook et al., 2013) and growth and wood quality traits in *P. sylvestris* (Calleja-Rodriguez et al., 2020). Genomic prediction aims to increase the efficiency of breeding programmes, shorten the breeding cycle length, improve timber yield and quality and reduce loss of trees due to pests and diseases in commercial forestry as well as screen natural populations for their adaptive potential to future threats such as climate change and disease (Isabel et al., 2020, Capblancq et al., 2020). However, multiple trials are necessary to identify, test and validate the SNPs associated with each trait before genomic prediction can be applied with any confidence due to the potentially confounding effect of phenotypic plasticity. Another overlooked aspect is the difficulty of applying these approaches to species which do not have well-established breeding populations, and the comparative 'messiness' of natural populations from which much seed for much planting is derived (Herbert et al., 1999). These natural populations are likely to lack strong selective pressures of the kind which are imposed on breeding populations by selection for valuable traits, and the lack of pedigree information makes SNP-trait association significantly harder.

92 In our study we focused on a set of three closely related species from the *Pinus* genus, *P.*
93 *sylvestris*, *P. mugo* and *P. uncinata*, that differ strongly in phenotype, geographical
94 distribution and ecology. These species have previously been studied in several biometric,
95 quantitative trait and population genetic investigations (Boratyńska and Boratyński, 2007,
96 Lewandowski et al., 2000, Wachowiak et al., 2013, Wachowiak et al., 2018a, Wachowiak et
97 al., 2018b) and a SNP array has been developed for them, based on candidate gene and
98 transcriptome sequencing (Perry et al., 2020). They form a monophyletic group within
99 Pinaceae (Grotkopp et al., 2004), having diverged within the last 5 million years and adapted
100 to different environments (Wachowiak et al., 2013, Wachowiak et al., 2011, Wachowiak et
101 al., 2018a). Despite morphological, geographical and ecological differentiation, the three taxa
102 show high genetic similarity in biochemical and molecular markers, have the same number of
103 chromosomes (2n = 24), show weak reproductive barriers, and share many ancestral
104 polymorphisms segregated in the pine genome (Lewandowski et al., 2000; Wachowiak et al.,
105 2013). Consequently, these taxa form a valuable experimental system in which to undertake
106 comparative analysis for molecular signatures of selection, ecological divergence and local
107 adaptation at inter- and intraspecific levels.

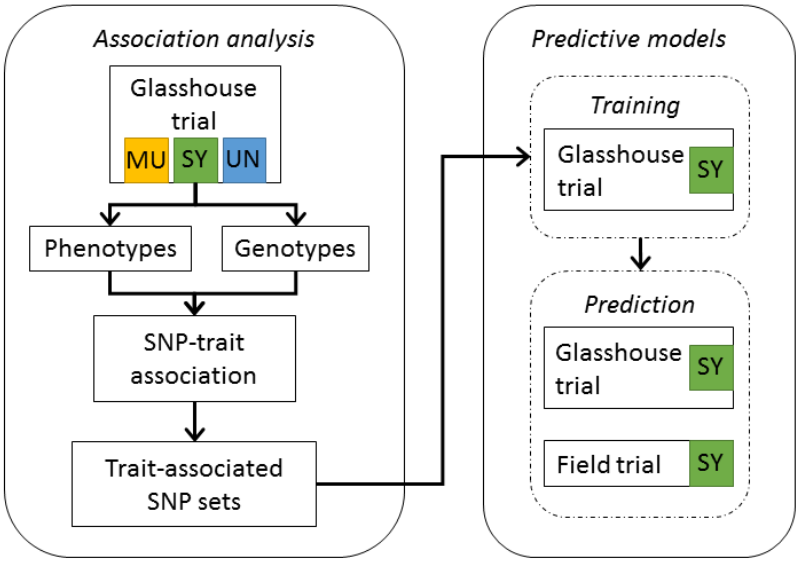
108 In this study, using data from common garden experiments and genotypes from a large new
109 multispecies SNP array, we first identified SNPs associated with growth and phenology in the
110 three species. Then, to evaluate their potential use as a tool for genomic prediction, we tested
111 models making use of these SNPs to estimate predicted values for traits in an independent
112 field trial. Finally, we discuss the potential, and some of the limitations, for making use of
113 genomic prediction in breeding programmes for *P. sylvestris*.

114

2. Methods

Experimental design and analyses performed in the study are summarised in Figure 1.

Figure 1. Plant material, datasets and analyses used in the study. MU: *P. mugo*; SY: *P. sylvestris*; UN: *P. uncinata*.



2.1. Plant material and phenotype assessments

Collection of plant material, experimental design and phenotype assessments are described by (Wachowiak et al., 2018a). Briefly, open-pollinated seeds of the three pine species were collected from three to five families per population from twenty-eight natural populations in Europe covering the extent of each species' range, including thirteen populations of *P. sylvestris* (SY), nine *P. mugo* (MU), and six *P. uncinata* (UN). Seeds were sown on trays of compost in spring 2010. After germination, a provenance–progeny trial was established in an unheated glasshouse at the UK Centre for Ecology and Hydrology, Edinburgh, UK (latitude 55.861261, longitude -3.207819). Seedlings were grown under natural light with automatic watering applied during the growing season. The trial was divided into 25 randomized blocks

130 with up to five families per population of which the first 18 blocks were analysed by
131 Wachowiak et al. (2018a).

132 Phenology (traits assessed: BS, timing of bud set, BB, timing of budburst) and growth (traits
133 assessed: H, total height; I, annual increment - the increase in height from one year to the
134 next) were recorded for every seedling to evaluate within- and between-species variation
135 (species means for trees sampled in this study recorded in Table S1). Nested analyses of
136 variance (ANOVA) were performed in Minitab 17 (Minitab Statistical Software, 2010) with
137 species, and population nested within species, as fixed effects, families nested within
138 population as a random effect, and block as a random effect. Response variables were growth
139 and phenology traits. Analyses were also performed for each species separately. Families with
140 a single individual were removed from the analyses (four families from MU: family 26 from
141 population M5; family 9 from population M12; and families 5 and 9 from population M8). To
142 assess the proportion of variation that is under genetic control, the narrow sense heritability
143 (h^2) for each trait was estimated. Narrow-sense heritability, which is the proportion of total
144 phenotypic variance (V_P) explained by additive genetic effects (V_A; Falconer and Mackay,
145 1996), was estimated using among family, block and residual variance (V_{fam}, V_{block} and V_{res},
146 respectively) from data pooled across populations:

$$h^2 = \frac{V_A}{V_P} = \frac{RV_{fam}}{V_{fam} + V_{block} + V_{res}}$$

147 where R is the relatedness of individuals within families (individuals within a family are
148 assumed to be half-siblings as they are from a single mother tree with an unknown paternal
149 contribution).

Bud set was scored when a visible apical bud with clearly developed scales was formed at the tip of a stem in each seedling and was measured as the number of days since the date on which the first plant that set a terminal bud was observed (in the first year of growth: BS2010). Budburst was scored when new needles emerged around the tip of the apical bud in the main stem and was measured as the number of days since the date on which the first plant to burst bud was observed (in the second and third years, BB2011, BB2012). Phenology observations were conducted twice a week. The height of young pines was measured from the second to fourth year of the pine growth (H2011, H2012, H2013). The annual increment was estimated for growth between 2011-12 (I2012) and 2012-13 (I2013).

An independent multi-site, field-based provenance-progeny trial of *P. sylvestris* was also phenotyped and used to test results from the glasshouse trial described above. Seeds from eight families from each of 21 native Scottish *P. sylvestris* populations were collected in March 2007 and germinated at the James Hutton Institute, Aberdeen (latitude 57.133214, longitude -2.158764) in June 2007. Germinated seedlings were grown either in a glasshouse with automatic watering or in pots outside (with additional watering when necessary) until the trees were moved to one of three transplant sites. A subset of trees from two of these sites were genotyped as part of this study: a site in the Borders of Scotland (Yair, YA: latitude 55.603625, longitude -2.893025) was planted in October 2012; a site in Aberdeenshire (Glensaugh, GS: latitude 56.893567, longitude -2.535736) was planted in spring 2012. Trees transplanted to YA were initially grown in a glasshouse whereas trees transplanted to GS were started in pots outside. The two transplantation sites also generally experience different climates, with the YA site typically warmer and drier than the GS site (Table S2) and with a longer growing season.

Trees were planted in four randomised blocks at 3 m x 3 m spacing. A guard row of Scots pine trees was planted around the periphery of the blocks. Each block comprised one individual from each of eight families per 21 populations (168 trees). Budburst and height have been assessed annually since 2015. Height was measured in the winter before the growing season began from 2015 to 2020. Height was also measured before the start of the second growing season in March 2008. The annual increment was estimated as the increase in growth from one year to the next. Each tree was assessed for budburst stage annually from 2015 until 2019 at weekly intervals from early spring until budburst was complete. Seven distinct stages of budburst were defined (Table S3). The number of days for each tree to reach each stage of budburst, starting from the day the first tree was observed at each stage, was recorded. When trees progressed through budburst stages rapidly, skipping a stage between assessments, a mean value was taken between the two. The duration of the core stages of budburst (time taken to progress from stage 4 to stage 6) was also estimated. Although the method used to record budburst was not identical in the glasshouse and multi-site field trials, the observation of needles as described by (Wachowiak et al., 2018a) is equivalent to stages 5 and 6 in the multi-site field trial. To better understand the relationship between budburst timing and duration among years and stages, Pearson's correlation coefficient and significance values between the two were estimated using a package 'Hmisc' (Harrell Jr, 2020) in R (R Core Team, 2020), for each site separately.

2.2. Genotyping array

The design of the array, genotyping and SNP calling are as described by (Perry et al., 2020). Briefly, an array comprising 49,829 single nucleotide polymorphisms (SNPs) was used to genotype 1,920 DNA samples (from needles of four pine species: the species included here

196 plus *Pinus uliginosa*) according to the Affymetrix Axiom Assay protocol on a GeneTitan and
197 following genotyping, genotype calls were performed using Axiom Analysis Suite as
198 recommended by the manufacturer. A subset of trees from the experimental glasshouse trial
199 described in the previous section were genotyped including twelve populations of *P. sylvestris*
200 (N = 461) and five populations of *P. mugo* (N = 145) and *P. uncinata* (N = 201). Up to 10
201 trees were genotyped per family (except for population SY33 which was genotyped up to a
202 maximum of 14 trees per family). Five families were genotyped per population with the
203 exception of SY44 (N families = 4), SY30 (N families = 3) and MU5 (N families = 3).
204 Samples were filtered to remove all those with a call rate < 80 % (N = 45).

205 The multi-site field trial of *P. sylvestris* was also partially genotyped: 100 trees from YA and
206 108 trees from GS, each comprising the same five populations (Beinn Eighe, BE; Glen Affric,
207 GA; Glen Loy, GL; Glen Tanar, GT; Rhidorroch, RD) with 19-22 individuals per population
208 for each site. There were 7-8 families genotyped for each population with 1-3 half-siblings in
209 each family at each site. These datasets are henceforth referred to as YA-SY and GS-SY.

210 **2.3.Population genetic structure, kinship and statistical power**

211 Population genetic structure was assessed visually by constructing a neighbour joining (NJ)
212 tree in the R package 'ape' (Paradis and Schliep, 2019) based on a distance matrix generated
213 in TASSEL version 5.2.39 (Bradbury et al., 2007) using all samples with call rate > 80 % in
214 all species (N = 762). SNPs with call rate < 80 % (N = 48) were excluded. Pairwise kinship
215 (centred identity by state) was estimated for each species independently including only
216 samples with call rate > 80 % (MU: N = 115; SY: N = 456; UN: N = 191) using all
217 polymorphic markers in TASSEL. The skewness of the distribution within each species'
218 matrix was calculated using the D'Agostino skewness test in the R package 'fBasics' (Wuertz

et al., 2020). The statistical power of each species' dataset (MU; SY; UN), the *P. mugo* complex (MU-UN), and the full dataset including all species (MU-SY-UN) to detect true associations between SNPs and adaptive traits was estimated using the method reported by (Wang and Xu, 2019) under the following assumptions: nominal type 1 error (false positive) = 0.05; QTL size = 0.05. Statistical power was estimated at different levels of polygenic effect (λ): from 0.1 (where polygenic variance is 10 % of phenotypic variance) to 10 (where polygenic variance is 10 x phenotypic variance). Allele frequencies of all SNPs subsequently found to be significantly associated with the adaptive traits in the MU-UN dataset were checked in each species separately (MU and UN) to assess the contribution of each species to associated genetic variation.

2.4. Genetic associations and putative functions

Identification of SNPs potentially associated with phenology (traits: budburst and bud set) and growth (traits: height and increment) was conducted for each trait in each year. Association with SNPs was tested in each species separately (MU; SY; UN) as well as in all species together (MU-SY-UN) and in the *P. mugo* complex (MU-UN). A mixed linear model (MLM) with a covariance (kinship: centred identity by state) matrix and a matrix derived from principal component (PCA) scores, to allow for population stratification, among individuals was fitted to each locus independently in TASSEL (version 5.2.39). The proportion of true null hypotheses was estimated using a false discovery rate (FDR) approach, retaining SNPs associated with traits with adjusted *p* values < 0.05.

A multi-locus mixed model (MLMM) approach, with 10 steps, was used to identify whether any loci have large effects (Segura et al., 2012). Highly significant SNPs (based on estimations of genetic variance, $p < 0.001$) were included in a forward-backward stepwise

242 approach, one by one, as cofactors in the model. The multiple Bonferroni criterion, defined as
243 the largest model whose cofactors all have a *p*-factor below a Bonferroni-corrected threshold
244 of 0.05, was used to indicate the best model.

245 SNPs were divided into two classes on the basis of their minor allele frequency (MAF): MAF
246 > 0.05: common; MAF < 0.05: rare. As it is likely that the majority of traits are controlled by
247 many genes of very small effect it is important to consider every SNP identified. The narrow
248 sense heritability of each trait and the proportion of common SNPs identified as associated
249 with each trait were examined to determine whether traits associated with high frequency
250 SNPs are also associated with high levels of narrow sense heritability, e.g. due to their
251 prevalence across the populations in question. Each SNP found to be significant was also
252 examined to compare the putative function of the genes on which they are located with the
253 trait in question. To do this, the full unigene sequence for each SNP was BLASTed against
254 the uniprotkb_viridiplantae database, the result with the highest score (minimum e-value 1E-
255 50) for each unigene was retained, and the putative function determined by a literature survey
256 using the search term '*protein name* function plant'. Where the protein was uncharacterised,
257 the protein domain and/or family was recorded and the most likely function inferred. Where
258 putative functions could be determined the genes were grouped according to their role in the
259 following phenotypic responses: 'Response to environment' (including abiotic and biotic
260 stress response), 'Growth and development' (including cell division, differentiation and
261 senescence); 'Reproduction' (including flowering time and seed yield). Although many
262 cellular processes (e.g. metabolism, signalling pathways, DNA binding, transcription,
263 translation) were also identified as putative functions, these were assumed to be underlying
264 control and expression of phenotypic functions and were not assigned a function.

2.5 Prediction models: construction and internal testing

Phenotypic prediction multiple linear regression models were constructed in R. A number of different models were constructed and compared using different sets of SNPs and different traits to train the model. Predictive models were constructed using SNPs identified as potentially associated with variation in phenology (trait: budburst: BB2011) and growth (traits: height and increment: H2013 and I2013). Because the SNPs used in these models were identified in both the SY and in the MU-SY-UN datasets, separate models were also constructed comprising just SNPs identified in the SY dataset. Predictive models were also constructed using the same number of randomly selected SNPs from all polymorphic loci with similar MAF to the SNPs from each prediction model. Additionally, predictive models were constructed using all available polymorphic SNPs for SY. All models (phenology and growth from SY and MU-SY-UN; phenology and growth from SY; random; all polymorphic SNPs) were run both with and without a MAF filter (retaining only SNPs which were common (MAF > 0.05) in the datasets from which the significant associations were originally identified). The prediction model for all polymorphic SNPs was constructed using ridge regression with the R package 'rrBLUP' (Endelman, 2011), rather than multiple linear regression, as recommended when the number of loci is greater than the number of samples. For all models, where necessary, family means were used to replace missing data. The predictive models were run using a training set comprising 60 % of *P. sylvestris* trees from the glasshouse trial, which had been used to identify associated SNPs, and were internally tested using the remaining 40 % of *P. sylvestris* glasshouse trees. Models were run using *P. sylvestris* trees and not *P. uncinata* or *P. mugo* as subsequent testing of the models was in this species alone. We used budburst and growth but not bud set data as subsequent model testing was applied to data from an independent trial for which only these traits were available.

289 Pearson's correlation coefficient and significance for correlations between predicted values
290 generated by the predictive models and observed values for both phenology and growth (both
291 H2013 and I2013 were tested to see which performed best for the growth predictive model)
292 were estimated using the R package 'Hmisc' (Harrell Jr, 2020).

293 SNPs used in each prediction model were assessed for their variation among *P. sylvestris*
294 populations using the R package 'hierfstat' (Goudet and Jombart, 2020). Basic statistics
295 including overall observed heterozygosity (H_O), mean gene diversities within populations
296 (H_S), inbreeding coefficient (F_{IS}) and population differentiation (F_{ST}) were estimated for each
297 set of SNPs described above.

298 **2.6 Independent testing of the models in two outdoor *P. sylvestris* trials**

299 SNPs identified as potentially associated with budburst and growth were tested using
300 genotype and phenotype data from an independent field trial of *P. sylvestris*, established at
301 contrasting sites (YA and GS) in 2012. Genotyped trees from the field trials were assigned
302 predicted values for both phenology and growth) using multiple linear regression models
303 constructed using either all available SNPs or only those found to be significantly associated
304 with the trait. The models were those that performed best (i.e. the strongest correlation) in the
305 internal test and with the full set of *P. sylvestris* trees from the glasshouse trial (call rate > 80
306 %, N = 456) as a training set. Observed values for growth (height and increment) and
307 budburst (number of days to reach budburst stages 4 to 6 from the first observation at each
308 site, and duration for each tree to progress from stage 4 to stage 6) at multiple years (2015-
309 2020 for increment, 2015-2019 for budburst) were compared with values generated by the
310 predictive models. Multiple years were used to ensure that annual variation caused by
311 seasonal differences could also be considered. Because height is a cumulative measure, only

the most recent (2020) and the measurements made prior to transplantation (2011) were compared with the predicted values. To assess the performance of the predictive models, the Pearson's correlation coefficient and significance values between predicted and observed values for phenology and growth were estimated for each site (GS and YA) separately using the R package 'Hmisc' (Harrell Jr, 2020). The use of two sites in independent testing also allowed comparison of the performance of the predictive models in different environments.

The effectiveness of using the predictive model as a genomic selection tool was also tested and compared with other selection methods. For each method, 10 trees were selected from each site: for genomic selection the 10 trees at each site with the highest values generated by the predictive model were chosen; for phenotype selection the 10 tallest trees at each site prior to the start of the second growing season (measured in March 2008) were chosen; for comparison, 10 trees were also randomly chosen from each site. The average height at 13 years (2020) of the 10 trees selected using each method was compared. The trees selected using each method were also compared to the 10 tallest trees at each site at age 13.

3. Results

3.1. Intra- and inter-specific trait variation

Bud set was, on average, earliest for MU and latest for SY with a mean difference of nearly 19 days between the two species (Table S1, Table S4). Bud set for UN occurred, on average, 8.28 days after MU and 10.63 days before SY. Budburst was similarly earliest for MU but was latest for UN in both years assessed although the mean difference between species was greater in 2012 (15.17 days) than in 2011 (5.42 days). For all years, on average, MU were the shortest trees and SY were the tallest with increment similarly greater in SY than in UN or MU. By 2013, SY trees were on average over double the height of the average MU tree, with

335 UN trees on average just over two-thirds the height of the average SY tree. Narrow sense
336 heritability estimates for all species were highest for height (h^2 = 0.72-1.19) and lowest for
337 budburst in 2012 (h^2 = 0.25) although standard errors for all estimates were very large due to
338 the small sample sizes (Table S8). Phenotypes for the independent multi-site field trial are
339 also provided in Table S5.

340 **3.2. Summary of genotyping array**

341 High quality genotypes (call rate > 80 %) were obtained for over 94 % of trees genotyped
342 within the trial (N = 762): MU, N = 115; SY, N = 456; UN, N = 191 (Table S6). There were
343 over 9,500 high quality (call rate > 80 %) polymorphic SNPs which were shared among the
344 three species (Table 1), with a further 1,352 SNPs which were polymorphic in at least two
345 species and monomorphic in a third. The set of successfully converted SNPs (N = 20,795)
346 reported by (Perry et al., 2020) was found to be identical to those successfully converted in
347 this study. Genotyped trees from YA and GS were all high quality (Table S7).

348

Table 1. Count of polymorphic, monomorphic and low call rate SNPs within and among species

SY (456/461)	UN (191/201)	MU (115/145)		
		CR<80 (4,884)	Mono (4,639)	Poly (11,272)
CR<80 (9)	CR<80 (288)			
	Mono (5,297)		1	
	Poly (15,210)	6		2
Mono (5,767)	CR<80	47	3	8
	Mono	208	3,446	507
	Poly	452	251	845
Poly (15,019)	CR<80	195		35
	Mono	120	723	292
	Poly	3,856	215	9,583

Species codes: SY, *P. sylvestris*; UN, *P. uncinata*; MU, *P. mugo*. Parentheses after species code indicates number of samples with call rate > 80 % compared with number of samples genotyped. SNP codes: CR<80, call rate < 80 %; Mono, monomorphic; Poly, polymorphic. Parentheses after SNP codes for each species indicates the total number of SNPs within each category for each species individually.

3.3. Population genetic structure, kinship and statistical power

The neighbour joining tree generated from the distance matrix indicated weak population structure as reported in previous studies (Wachowiak et al., 2013, Wachowiak et al., 2018b). The pairwise kinship distribution was strongly skewed toward positive kinship values for each species (D'Agostino's skewness test, MU: $z = 101.389$, $p < 2.2 \times 10^{-16}$; SY: $z = 446.904$, $p < 2.2 \times 10^{-16}$; UN: $z = 153.664$, $p < 2.2 \times 10^{-16}$), as expected given the use of half siblings. These results support the use of mixed model approaches and correction for population stratification prior to testing for genetic association.

The statistical power to detect true associations between SNPs and adaptive traits was found to be extremely low for both MU and UN even when the polygenic effect was assumed to be 10x the phenotypic variance (Table S9). This is likely to be due to the low sample numbers: the statistical power of SY was similarly low if the sample numbers were reduced to those of MU and UN, although the statistical power remained low even when the polygenic effect was increased. The SY dataset was found to have relatively high statistical power, and the joint

370 MU-UN dataset had lower power, but significantly more than for each species individually.
371 The statistical power of a dataset including all three pine species was found to be very high
372 regardless of the polygenic effect. For these reasons, the following datasets were analysed for
373 associations with traits: the *P. mugo* complex (MU-UN), *P. sylvestris* (SY) and all three pine
374 species (MU-SY-UN).

375 **3.4. Identification of loci associated with traits**

376 One hundred and thirteen SNPs were identified as associated with phenology and growth in
377 the three pine species (Table 2; Table S10). These included SNPs which were identified in
378 more than one species' datasets. There was very little overlap of identified SNPs among
379 different traits or among years within the same trait: four SNPs were associated with more
380 than one trait, of which only one (comp51128_c0_seq1_1529) was associated with both
381 phenology (trait: BB2011) and growth (trait: I2013). The vast majority of SNPs were
382 identified using the MLM approach (N SNPs = 108) rather than the MLMM approach (N
383 SNPs = 14) and there were many more common SNPs (MAF > 0.05) identified using the
384 former method (MLM: N = 36; MLMM: N = 1). There were nine SNPs identified as
385 significantly associated with traits in both MLM and MLMM. Significantly associated SNPs
386 were identified for all traits in all years except BB2012. The traits with most associated SNPs
387 were BB2011 (N = 54), I2013 (N = 34) and BS2010 (N = 18), whereas other years/traits all
388 had low numbers of associated SNPs (H2011, N = 3; H2012, N = 1; I2012, N = 1).

389

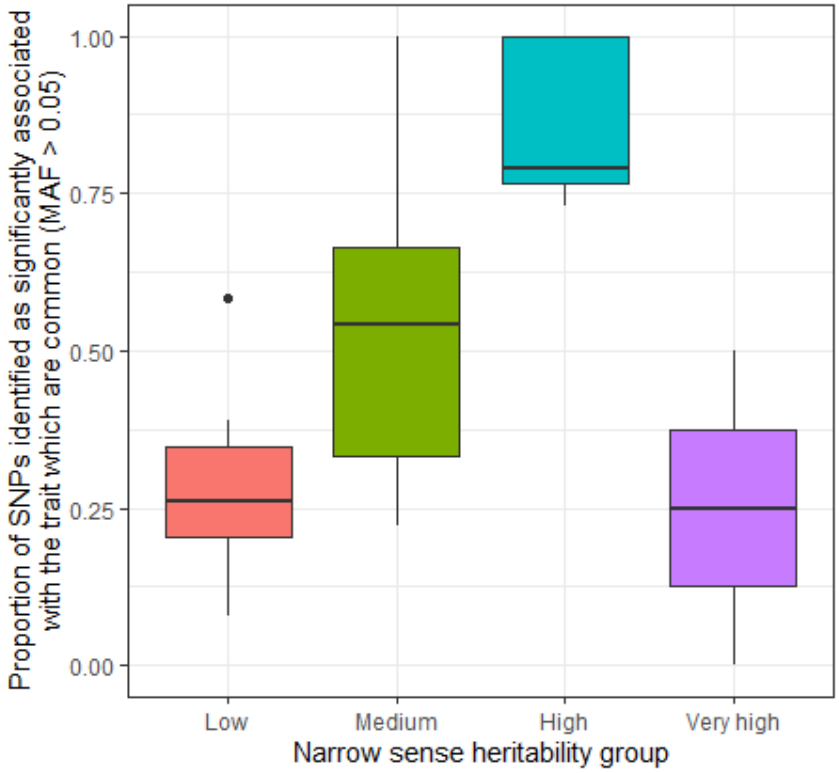
Table 2. SNPs associated with phenology and growth traits in the three pine species identified from a mixed linear model (MLM) in TASSEL and a multi-locus mixed model (MLMM) in R

		MLM		MLMM	
Trait	Species	Common	Rare	Common	Rare
Phenology					
BB2011	MU-UN	9	25		4
	SY	7	11		3
	MU-SY-UN	4	19		3
BS2010	SY	4	14		
Growth					
H2011	SY		1		
	MU-SY-UN	1	1		
H2012	MU-SY-UN	1			
H2013	SY	2			
	MU-SY-UN	4		1	
I2012	MU-SY-UN	1			
I2013	MU-UN	6	1		1
	SY	2	20		4
	MU-SY-UN	6	4	1	1

Species codes: MU, *P. mugo*; SY, *P. sylvestris*; UN, *P. uncinata*;. Trait codes: budburst (BB); bud set (BS); height (H); increment (I). Common: SNPs with MAF > 0.05; Rare: SNPs with MAF < 0.05

The MAF of SNPs identified for each trait/species' dataset were compared with their narrow sense heritability which was grouped into one of four categories: low, N = 8: $h^2 < 0.4$; medium, N = 7: $0.4 \leq h^2 < 0.6$; high, N = 7: $0.6 \leq h^2 < 1$; very high, N = 2: $h^2 > 1$ (Figure 2). As the narrow sense heritability increased from low to high, the proportion of SNPs which were common similarly increased, although this relationship did not extend to traits with very high h^2.

Figure 2. Variation in the proportion of common SNPs (MAF > 0.05) identified as significantly associated with each trait for different groupings of narrow sense heritability (h^2): low, < 0.4; medium, $0.4 \leq h^2 < 0.6$; high $0.6 \leq h^2 < 1$; very high $h^2 > 1$.



A higher number of associated SNPs were identified in SY (N = 62) than MU-UN (N = 43). Only one SNP (comp51128_c0_seq1_1529) was identified as significant in both datasets although it was associated with phenology (BB2011 for MU-UN) and growth (I2013 for SY): it was common (MAF > 0.05) in MU-UN but rare (MAF < 0.05) in SY (Table S10). A further 45 SNPs were found to be associated with traits when all species were combined within a single analysis, although 11 of these were also identified in SY and 23 were identified in MU-UN. No MAF filter was applied prior to screening SNPs for association with the traits of interest: 37 SNPs were common (MAF > 0.05) in at least one dataset. Allele frequencies for SNPs identified as significantly associated with adaptive traits in MU-UN were compared for UN and MU separately (Table S11). Diversity was much lower in UN than MU for the

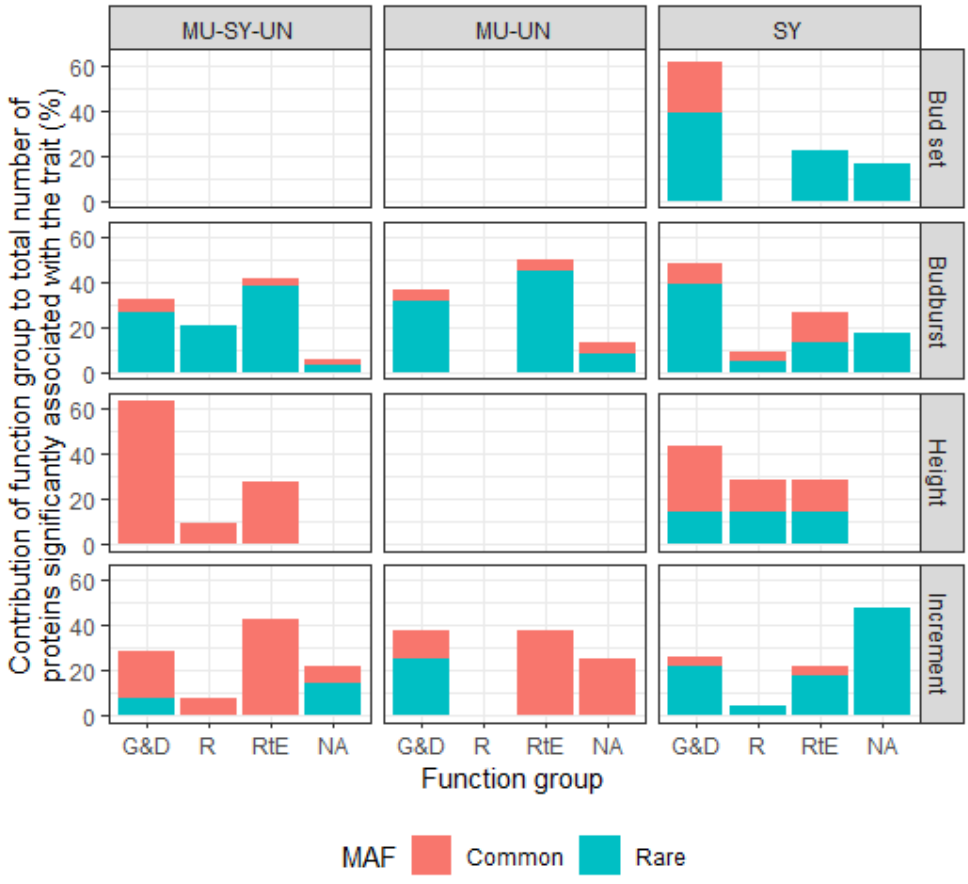
majority of SNPs: 23 of the 36 SNPs identified as associated with BB2011 were monomorphic in UN. In contrast, diversity in UN was much higher for SNPs identified as significantly associated with I2013 (Table S11). Similarly, the standard error for MU was more than twice that of UN for BB2011 (MU: 0.73; UN: 0.33) whereas the standard error for both species was similar for I2013 (MU: 0.47; UN: 0.32) (Table S1).

3.5. Putative function of genes containing SNPs associated with traits

One hundred and thirteen SNPs associated with phenology and growth in the three pine species were located at 111 gene loci (one unigene, comp48223_c0_seq1, contained three SNPs). One locus was originally identified in *Pinus radiata* (Doth_comp54682_c0_seq1_159), the remaining were identified following transcriptome sequencing in *P. sylvestris* and the taxa of the *P. mugo* complex (Perry et al., 2020: Table S10). The genetic sequences of the loci associated with each trait were found to be highly similar to proteins with a range of putative functions (Tables S12a-c). The majority of SNPs associated with bud set (all identified in SY) were found in genes that code for proteins putatively involved in growth and development (61.11 %) with a few (exclusively rare) SNPs found in proteins putatively involved in response to environment (22.22 %, Figure 3). In contrast, budburst had high numbers of associated SNPs (both rare and common) in genes that code for proteins putatively involved in response to environment and growth and development (mean % contribution of putative function groups to the total number of proteins containing SNPs significantly associated with budburst across species' datasets: 39.01 % and 39.09 % for growth and development and response to environment, respectively). Whereas the majority of SNPs associated with height were found in proteins putatively associated with growth and development, SNPs associated with increment were found in proteins putatively

associated with both growth and development and response to environment. There are some differences among species in the putative function of proteins containing significantly associated SNPs: the majority of SNPs in SY are found in proteins putatively associated with growth and development for all traits (Figure 3) whereas MU-SY-UN and MU-UN have higher proportions of SNPs in proteins putatively associated with response to environment as well as growth and development.

Figure 3. Contribution of putative function groups (G&D: growth and development; R: reproduction; RtE: response to environment) to the total number of proteins containing SNPs significantly associated with each trait (bud set, budburst, height and increment) as a percentage of the total number of proteins identified for each trait for each species' dataset (MU: *P. mugo*; SY: *P. sylvestris*; UN: *P. uncinata*). Proteins which were uncharacterised, for which no known function in plants was found or for which only cellular processes could be identified are categorised "NA". Total for each trait may be higher than 100 % as there may be more than one putative function assigned to a single protein. MAF: minor allele frequency (MAF > 0.05: common; MAF < 0.05: rare)



3.6. Predictive models for budburst and growth: internal testing in *P. sylvestris*

There was a large dropout in the number of SNPs which were suitable to include in subsequent predictive models: of the 38 SNPs identified as potentially associated with growth (H2013 and I2013) in the *P. sylvestris* and all species' datasets (SY and MU-SY-UN respectively), 24 were monomorphic in either (or both) the SY and the independent *P. sylvestris* datasets (YA-SY and GS-SY). Therefore, 14 SNPs were included in the model, of which seven were rare (MAF < 0.05) in the SY dataset (although only three of these were rare in the MU-SY-UN datasets in which they had been identified as associated with the traits). Of the 14 SNPs, five were identified in the SY dataset, four were identified in the MU-SY-UN dataset, three were identified in both the MU-UN and MU-SY-UN datasets and two were identified in both the SY and MU-SY-UN datasets. For the budburst predictive model, a total of twelve SNPs were used, of which five were rare in the SY dataset. The remaining 23 SNPs were monomorphic in at least one of the SY, YA-SY and GS-SY datasets. Of the 12 SNPs used in the model, one (which was rare) was identified in the MU-SY-UN dataset with the remaining eleven identified in the SY dataset.

The SNPs used to construct growth and budburst predictive models were found to have lower differentiation among populations (F_{ST}, Table S13) than the full set of polymorphic SNPs for SY (0.03 and 0.06, respectively). The inbreeding coefficient (F_{IS}) was 0.6 – 0.7 for the majority of SNP sets (Table S13) with slightly higher values observed in the SNP sets for the random budburst model with a MAF filter (F_{IS} = 0.8) and the growth model using SNPs identified in the SY dataset (F_{IS} = 0.9). Observed heterozygosity and gene diversity (H_O and H_S, respectively) were both lower in the sets of SNPs which were filtered to include only those which were common (MAF > 0.05) in the original dataset.

The performance of each predictive model (i.e. the strength and significance of the correlation of predicted values with the observed values for each trait) are summarised in Table 3. Models constructed using random SNPs were not successful in predicting values that were correlated with observed values for each trait. In all cases, the models constructed without a MAF filter (MAF > 0.05) always performed better than the equivalent models constructed using only common SNPs, although there was little difference in performance for those models constructed using all polymorphic SNPs. The predictive model for budburst constructed using SNPs identified in the SY dataset alone ($r = 0.40$, $p < 0.001$) performed better than the equivalent model using SNPs identified in both the SY and MU-SY-UN datasets ($r = 0.37$, $p < 0.001$) although there was only a single rare (MAF < 0.05) SNP which was present in the latter and not the former. For this reason, this predictive model for budburst (using SNPs identified in SY and with no MAF filter: final budburst model) was chosen to be tested independently.

Table 3. Pearson's correlation coefficient (r) and associated significance values for comparison of predicted and actual values for each trait both with and without a MAF filter when using prediction models constructed with SNPs significantly associated with each trait (Budburst; Growth), a random set of SNPs or all polymorphic SNPs.

Training trait	SNP set	MAF: No	MAF: Yes
Predictive models: Budburst			
BB2011	Budburst	0.37***	0.12
	Budburst (SY only)	0.40***	0.12
	Random	-0.04	-0.02
	All SNPs	0.57***	0.57***
Predictive models: Growth			
H2013	Growth	0.26***	0.25**
	Growth (SY only)	0.20**	0.19*
	Random	0.14	0.11
	All SNPs	0.49***	0.48***
I2013	Growth	0.19*	0.14
	Growth (SY only)	0.19*	0.09
	Random	0.02	-0.01
	All SNPs	0.35***	0.35***

MAF: No = no Minor Allele Frequency filter applied; Yes = only common (MAF > 0.05) SNPs included. MAF was calculated using the datasets from which the SNPs were originally identified as being associated with each trait. Significance values: *, p 0.01-0.05; **, p 0.001-0.01; ***, p < 0.001

Using H2013 as a training trait, the predictive model for growth performed more poorly using SNPs identified in the SY dataset than using SNPs identified in both the SY and MU-SY-UN datasets. However, with I2013 as a training trait in the same model, there was no difference in performance when the different SNP sets were used. There were highly significant positive correlations between observed and predicted values for H2013 when using the prediction models for growth whereas using I2013 as the training trait for the predictive model resulted in far lower levels of correlation between predicted and observed values. Therefore, the predictive model for growth using SNPs identified in both the SY and MU-SY-UN datasets with no MAF filter and using H2013 as a training trait (referred to as the final growth model) was chosen to be tested independently.

The effect of the trait used to train the model was also seen in comparisons of the performance of the models constructed using all polymorphic SNPs: for each trait, predicted

512 values were more closely correlated with the observed values in models using budburst than
513 in those using growth traits (H2013, I2013).

514 **3.7. Testing the prediction models in an independent *P. sylvestris* trial**

515 Prior to testing the final growth and budburst prediction models using data from the *P.*
516 *sylvestris* field trials, the relationship between duration and timing of budburst was examined
517 more closely. Timing (time taken to reach stages 4, 5 and 6) showed a significant negative
518 correlation with duration (time taken to progress from stage 4 to 6) of budburst at each year
519 assessed for stage 4, but the relationship was positively correlated for stage 6 (Table S14). In
520 contrast, the time to reach stage 6 showed a significant positive correlation with the duration
521 of budburst. Time to reach stage 5 was both positively (at GS) and negatively (YA) correlated
522 with the duration of budburst.

523 Predicted values were estimated using the final predictive models for budburst and growth as
524 well as models constructed using all available SNPs and compared with values observed in
525 the field. The field sites had shared populations and families but contrasting climates,
526 allowing the models to be independently tested on traits measured in different environments.
527 The predicted values for each trait were not significantly correlated with the observed values
528 when using models constructed with all available SNPs (Table 4). In contrast, a number of
529 significant correlations were observed when using models constructed with SNPs associated
530 with the traits in question. The predicted values for budburst were found to be significantly
531 positively correlated with the duration of budburst but only in GS in 2015 and 2018 (Table 4)
532 indicating a possible effect of annual environmental variation on the predictive power of the
533 model. They were also negatively associated with the time taken to reach stage 6 of budburst,

but only in YA in 2017 (although the values were also close to significance, p = 0.06, in 2018).

Table 4. Pearson's correlation coefficient (r) and associated significance values for comparison of predicted and observed values for each trait. Predicted values estimated by final predictive models for growth and budburst assessed for their performance in an internal test (Associated SNP models: budburst - SNPs identified only in SY, no MAF filter applied, N SNPs = 11; growth - SNPs identified in both SY and MU-SY-UN, no MAF filter applied, N SNPs = 14). Predictive models constructed using all available SNPs (no MAF filter applied, N SNPs = 15,019). Duration: time taken for each tree to progress from stage 4 to stage 6. Stage 6: time taken to reach stage 6 of budburst. Description of each budburst stage is given in Table 1.

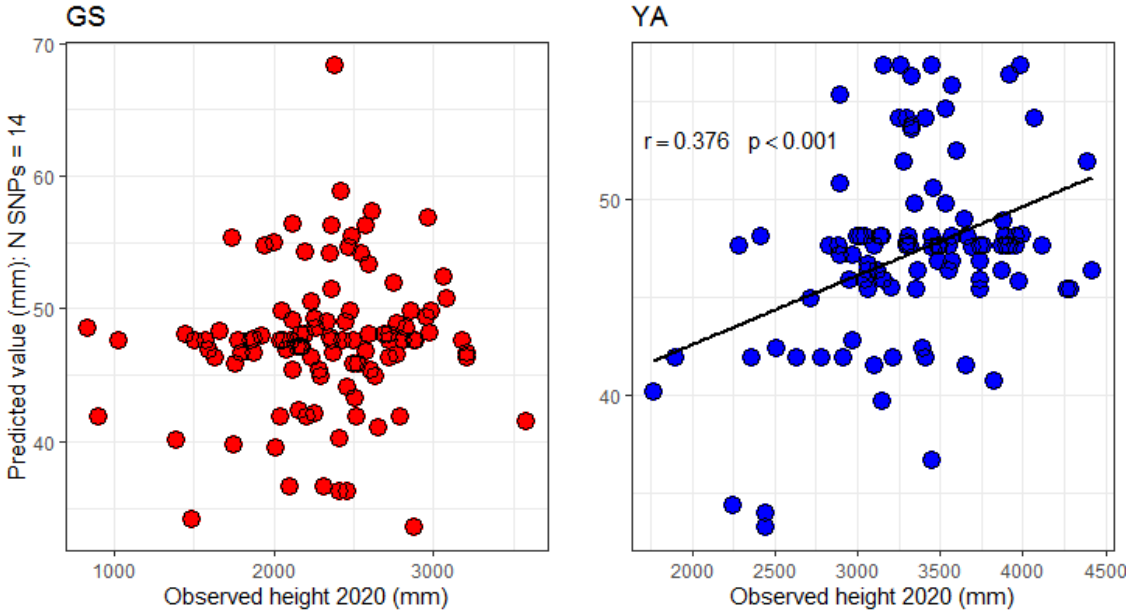
		Associated SNPs			All SNPs	
Observed trait	Year	GS	YA		GS	YA
Predictive model: Budburst (training trait: BB2011)						
Duration	2015	0.207 *	0.063		-0.005	0.088
	2016	0.105	0.136		-0.112	0.016
	2017	0.162	0.019		-0.131	-0.030
	2018	0.242 *	-0.087		-0.150	-0.105
	2019	0.152	0.075		0.099	0.188
Stage 6	2015	0.150	0.007		-0.167	0.167
	2016	0.094	-0.019		-0.101	0.004
	2017	0.120	-0.216	*	-0.093	-0.047
	2018	0.153	-0.187		-0.177	0.029
	2019	0.128	-0.058		0.111	0.134
Predictive model: Growth (training trait: H2013)						
Height	2008	-0.020	0.023		0.093	0.002
	2020	0.104	0.376	***	0.034	0.144
Increment	2015	0.022	NA		0.060	NA
	2016	0.173	0.299	**	0.149	0.158
	2017	0.121	0.312	**	-0.012	0.175
	2018	0.012	0.329	***	0.030	0.138
	2019	0.123	0.205	*	0.065	0.111
	2020	-0.001	0.262	**	-0.058	0.110

Significance values: *, p 0.01-0.05; **, p 0.001-0.01; ***, p < 0.001

The predicted values for growth were found to be significantly associated with observed increment measurements at YA in every year, but not in GS (Figure 4). The correlation between predicted values and observed height in 2008 (at age one) was not significant at either YA or GS, despite the strong correlation observed between predicted values and

552 observed height at age 13 at YA indicating that the cumulative effect of the trees growing in
553 the environment at YA contributed to the strength of the association.

554 Figure 4. Correlations of observed height measured in 2020 at age 13 against predicted values
555 using the final predictive model for growth at GS (correlation not significant) and YA.

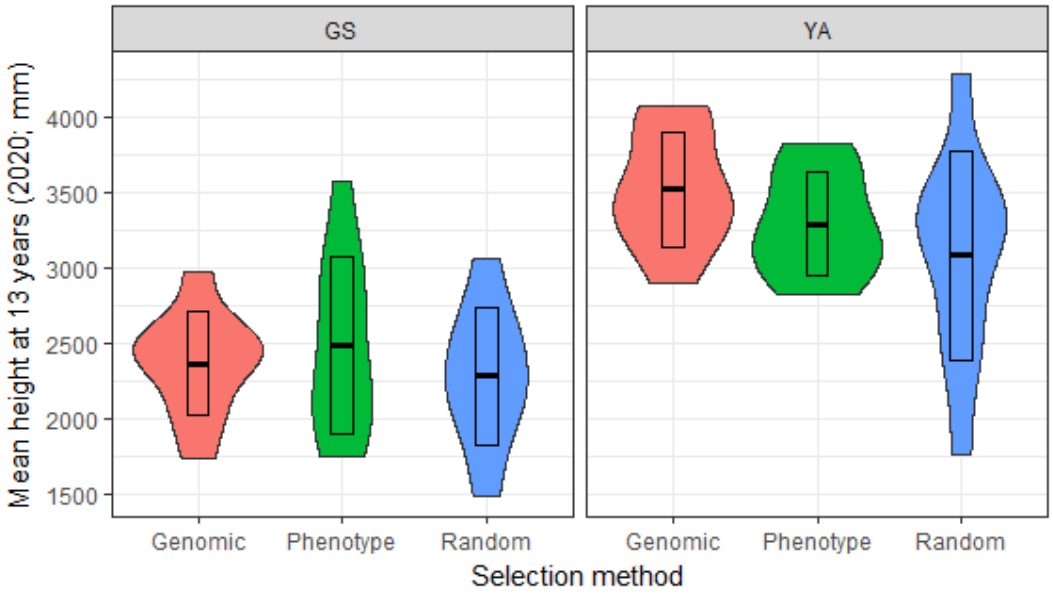


556

557 The effectiveness of the final predictive model for growth as a genomic selection tool was
558 tested by comparing different selection methods (Figure 5) in trees at both field trial sites (GS
559 and YA). The selected trees were from all five genotyped populations and included trees from
560 28 of the 40 families. The majority of families were only represented by a single tree,
561 although there were exceptions: two individuals were selected from single families in each of
562 the sites using the phenotype method; two individuals were selected from each of two families
563 in GS and from each of three families in YA using the genomic method; two individuals were
564 selected from each of three families in YA using the random method. Genomic selection was
565 the most successful method of selecting tall trees growing at YA: trees were on average 227
566 mm and 437 mm taller than trees selected using the phenotype and random methods,
567 respectively. The differences between selection methods were much smaller at GS: the mean

height of the 10 tallest trees selected using each of the selection methods was within a range of 203 mm. The coefficient of variation (CV) for trees chosen using the phenotype selection method was over 60 % greater than for those chosen using the genomic selection method at GS (23.68 and 14.63, respectively), indicating that trees chosen using the phenotype method were more variable for this trait at the site. Trees selected using the genomic and phenotype selection methods at YA had very similar CVs (10.84 and 10.31, respectively). Using the phenotype selection method, there were three trees at GS and none at YA that were among the ten tallest trees at each site. The genomic selection method identified two trees at YA and one at GS which were among the ten tallest trees at each site and the random selection method identified one tree at each of YA and GS which were among the ten tallest trees at each site.

Figure 5. Height at 13 years (measured before the growing season started in 2020) of 10 trees at Yair (YA) and Glensaugh (GS) selected using different methods: Genomic: genomic selection using values from the final predictive model for growth (SNPs identified in both SY and MU-SY-UN, no MAF filter applied, N SNPs = 14); Phenotype: phenotype selection where trees were selected based on their height at one year (before the start of the second growing season, 2008); Random: trees were randomly selected from each site. Crossbars indicate means and standard deviations.



Discussion

This study is among the first to use a high throughput array to identify SNPs associated with growth and phenology traits in conifers. Although the traits examined here are likely to be important in an ecological context, *P. sylvestris* is of significant economic value and the approach has potential for selection of traits of interest to industry. The use of a high-throughput SNP array allowed nearly 50,000 SNPs to be simultaneously genotyped in a large number of trees, the vast majority (94 %) of which were of sufficiently high quality to be used in subsequent analyses. To increase the sample size of the datasets and the statistical power of our analyses, data from *P. mugo* and *P. uncinata*, which are both part of the *P. mugo* complex, were combined. The dropout rate for *P. mugo* was much higher, and the call rate much lower than for *P. sylvestris* and *P. uncinata*. It is likely that this is a consequence of the dominance of *P. sylvestris* in the sample set used to set allele calling thresholds, coupled with the genetic distance between the two species (Perry et al., 2020). Despite this, nearly a third of SNPs on the array were high quality in all three species and nearly half of all successfully converted SNPs were polymorphic in all three species – twice the number reported by Perry et al., (2020).

Our study applied the genotyped SNP dataset to test for associations with previously published, well-characterised phenotypes for three pine species (Wachowiak et al., 2018a), identifying 113 SNPs significantly associated with variation in growth and phenology over multiple years. The large amount of interspecific variation for each trait, summarised in this study using the subsampled genotyped trees, supports the use of three species to identify SNPs as the greater range of phenotypes provides greater scope to identify genetic differences underlying the variation, as well as the opportunity to compare SNPs identified in each

610 species. As shown in our previous study, the between-population variation in both phenology
611 and height was far less in *P. mugo* and *P. uncinata* than in *P. sylvestris*, reflecting the fact that
612 the latter was sampled from a much broader geographical distribution and across much wider
613 environmental gradients in photoperiod and temperature (Wachowiak et al., 2018a). Despite
614 the smaller environmental gradient represented by our *P. mugo* sampling, the number of SNPs
615 identified as significantly associated with phenology was similar to the number identified in
616 *P. sylvestris*. However, the number of SNPs identified as significantly associated with growth
617 traits was much lower.

618 Overall, the majority of SNPs identified in this study were rare. Of those that were common,
619 the numbers of SNPs identified as significantly associated with traits were similar among *P.*
620 *sylvestris* and the *P. mugo* complex for both phenology (11 and nine for *P. sylvestris* and the
621 *P. mugo* complex, respectively) and growth (four and six for *P. sylvestris* and the *P. mugo*
622 complex, respectively). Although one SNP was found to be associated with both phenology
623 and growth (the former in the *P. mugo* complex and the latter in *P. sylvestris*) it was
624 extremely rare in *P. sylvestris*. Most likely, this reflects the confounding effect observed when
625 a small number of individuals (in this case, two) have both a rare allele at this locus and are at
626 the tail-end of a trait distribution (the two individuals were ranked 366 and 412 out of 413 for
627 increment in 2013). This finding supports the use of MAF filtering, which is frequently
628 performed either before or during analysis, although here we report all SNPs to evaluate the
629 relative contribution of rare and common SNPs to each trait and to assess the predictive
630 power of models constructed using SNPs with and without MAF filtering. There were very
631 few instances of the same SNP being associated among traits, among species or among years,
632 which may indicate the involvement of different genes at different stages of development or
633 in response to varying environmental conditions, as well as the very small effect sizes of most

SNPs in polygenic traits (Korte and Farlow, 2013). Our earlier comparative genetic studies of a large set of SNPs located in nuclear genes similarly found almost no shared polymorphisms under selection between different taxa of the *P. mugo* complex (Wachowiak et al., 2018b). The majority of SNPs associated with height, which showed high levels of narrow sense heritability among species, were common, whereas the majority of SNPs associated with budburst (of which there were far more than for height), which showed low levels of narrow sense heritability among species, were rare. The positive relationship between narrow sense heritability, excepting those with very high heritability values, and the proportion of SNPs identified which were common suggest that fewer SNPs of larger effect are associated with traits with high narrow sense heritability, while a larger number of SNPs with smaller effects are associated with traits with low narrow sense heritability. Calleja-Rodriguez et al. (2020) found that predictive ability (estimated as the correlation between the genomic estimated breeding values and phenotypes) was significantly positively associated with narrow sense heritability in *P. sylvestris* and there are numerous studies in humans discussing the contribution of SNPs with different MAF to the heritability of traits (e.g. Park et al., 2011; Yang et al., 2010). However, these are concerned with the relative contribution of variants with different MAFs to overall heritability, rather than comparing traits with different heritability values and comparing the MAF of SNPs associated with each. The lack of a similar relationship for the two individuals with very high values for narrow sense heritability, which in fact exceed the maximum value (1) for this measure, could be a result of the use of kinship and distance matrices to account for population stratification which may also, inadvertently, prevent the identification of SNPs which have an extremely close relationship with family structure.

The quantitative nature of adaptive traits assumes they are polygenically controlled (Mackay, 2001). The majority of studies on genetic control of adaptive traits in conifers have each identified multiple QTLs or SNPs associated with variation in timing of bud set, budburst and growth (Bartholomé et al., 2016, Eckert et al., 2009, Holliday et al., 2010, Hurme et al., 2000, Jermstad et al., 2001, Jermstad et al., 2003, Plomion et al., 1996, Prunier et al., 2013), although there have also been a limited number of specific genes implicated in the control of adaptive traits in conifers. For example (Eckert et al., 2015) tested 475 SNPs and found six significant associations with height and budburst in sugar pine (*Pinus lambertiana*) and (Budde et al., 2014) identified 17 SNPs significantly associated with serotiny in maritime pine (*Pinus pinaster*) using an array with 251 SNPs from candidate genes. A study by (Bai et al., 2019) used specific-locus amplified fragment sequencing (SLAF-seq) to screen over 450,000 SNPs to identify around 30 SNPs associated with resin-yielding capacity and volume of wood in Masson's pine (*Pinus massoniana*). A high-throughput array was also used by (Westbrook et al., 2013) to identify 231 SNPs significantly associated with oleoresin flow in clonally replicated sites over multiple years, of which the vast majority were specific to individual sites. Loci related to budburst/set were identified in *Picea abies* and *Pinus sylvestris* (PaFTL2, (Avia et al., 2014) and PsFTL2, (Gyllenstrand et al., 2007), respectively). Other traits which have been significantly associated with SNPs in pines include wood properties (González-Martínez et al., 2007), stem quality (Xiong et al., 2016), and disease resistance (Quesada et al., 2010), all of which were done in *Pinus taeda*.

Phenological variation in *Pinus* spp. observed in common garden studies has been repeatedly shown to be significantly associated with the environment at the site of origin (Howe et al., 2003, Hurme et al., 1997, Repo et al., 2000, Salmela et al., 2011, Wachowiak et al., 2018a) with trees from northern European populations setting bud and flushing earlier than trees from

more southerly populations. Whereas environmental cues are expected to play an important role in initiating phenological processes (Dougherty et al., 1994) including budburst (Laube et al., 2014), bud set is thought to be endogenous in *Pinus* spp, with photoperiod and temperature having relatively minor effects (Cooke et al., 2012). In this study, we found a high proportion of common SNPs in genes putatively involved in environmental responses (including response to abiotic and biotic stress and environmental cues) for both budburst and growth, but not for bud set. Common SNPs associated with bud set were exclusively located in genes related to growth and development. At this stage, assigning unigenes in conifers is largely presumptive and relies on similarity to domains or families of proteins with a large and/or speculative range of functions, many of which are, as yet, unexplored or undefined. However, the divergence of assignment among SNPs associated with budburst and bud set, and its concurrence with physiological understanding of these functions, suggests the genes implicated may have a role in key adaptive traits. Furthermore, as it has previously been demonstrated that intragenic linkage disequilibrium decays rapidly in the investigated species (Wachowiak et al., 2009, Wachowiak et al., 2013), there is a higher likelihood that SNPs identified may be directly involved in variation of phenology and growth.

Although predictive models constructed using all available polymorphic SNPs were the most successful at predicting values in the internal validation set they had no predictive ability when tested in an independent set of trees, possibly reflecting the divergent geographic ranges and associated environments of tree populations used in both trials. In contrast, predictive models constructed using SNPs identified as significantly associated with budburst and growth in the training set were found to be successful at estimating values in both the internal glasshouse grown validation set and the independent field grown sets of trees although the predictive ability of the models varied spatially (among the sites) and temporally (among

705 years). The final predictive model comprised SNPs from all species' datasets (SY, MU-UN
706 and MU-SY-UN) indicating that the approach using all three species to identify SNPs was
707 justified. The final predictive model for growth generated values that were highly
708 significantly correlated with actual height and increment over multiple years, although only at
709 YA. In contrast, the predictive ability of the model for trees at GS was poor. Phenotypic
710 variation is a product of both heritable genetic variation and environmental variation.
711 Consequently, the extent to which predictions are accurate will depend on the interplay
712 between the underlying genetic control of the traits and a host of external cues and stresses.
713 These will affect the control and dynamics of a large number of processes that will, in turn,
714 affect the expression of the traits both directly and indirectly. Trees growing at the YA site are
715 much larger than at GS, indicating that there may be environmental limitations for growth at
716 GS. Trees which were grown in the glasshouse and were used to identify SNPs associated
717 with growth are unlikely to have many environmental limitations and this could be why the
718 predictive model works well only for the YA site.

719 Ideally, a predictive model should be used in populations from very similar environments as
720 the population used to identify SNPs associated with traits and to construct the predictive
721 model (Resende et al., 2012b). For instance, a predictive model for serotiny constructed by
722 (Budde et al., 2014) also had variable success when applied to different populations of *Pinus*
723 *pinaster*.

724 Similarly, the predicted values for budburst were significantly (albeit only weakly) correlated
725 with the duration of budburst for two years at GS, but not at YA. The relationship between
726 bud burst timing and duration was found to vary as budburst progressed: trees which were
727 observed to reach the first few stages of budburst (where scales were open but needles not yet

visible) early in the season did not complete the whole budburst process sooner as might be expected. Instead, these trees took longer overall to complete budburst and it is clear that this relationship is not consistent among sites. However, it further demonstrates the influence of the environment on phenotypic variation and the caution that must be applied when interpreting or extrapolating results from differing environments.

There was a significant correlation between the predicted values for budburst and timing of budburst but only in one year, and at YA only. This was a negative relationship, such that trees which were predicted to complete budburst early in the season actually completed budburst late. Although this initially seems surprising, it does have a plausible biological explanation. The predictive model was constructed using SNPs which were identified as significantly associated with the timing of budburst in a set of trees from a common garden glasshouse experiment, whilst the validation data were collected from trees in a field trial. The environmental difference between the glasshouse and the field was clearly substantial, with possibly the most important deviation between the two being that temperatures in the glasshouse did not drop below freezing throughout the winter. The relationship between the chilling requirement (the accumulation of time spent below a certain temperature) and the initiation of budburst is complex: tree species and populations differ in their chilling requirement as well as in their forcing requirement (the accumulation of time spent above a certain temperature) after the chilling requirement is met (Körner, 2006). An increase in chill days (mean temperature < 5 °C) can significantly advance budburst timing in *P. sylvestris* (Laube et al., 2014). Heritable genetic variation in the timing of budburst is therefore likely to be strongly influenced by environmental cues including chilling and subsequent forcing. The contrast between the two environments means that trees requiring a greater number of chill days before the initiation of budburst will experience a delay in the glasshouse but burst their

752 buds earlier in the field, resulting in a negative relationship of the trait among the two
753 environments. Moreover, variation in the climate ensures that chilling and forcing conditions
754 vary among sites as well as annually. Although the mean number of annual chill days is
755 higher in GS than YA it is has fewer growing degree days which may delay the onset of
756 budburst in some families or populations. Another factor to consider is the different ages of
757 the trees used to identify SNPs significantly associated with traits, and the age of the trees in
758 the independent trial used to validate the predictive models. (Resende et al., 2012b) reported
759 that models generated using young *Pinus taeda* trees did not perform well at predicting
760 phenotypes for trees at age 6 years.

761 Predictive models potentially provide a tool with which to determine the phenotype of trees
762 without having to either grow them for a significant period or regularly assess them in the
763 field, saving both time and money. They therefore have several potential applications
764 including selecting for key traits in commercial breeding programmes and assessing native
765 forests for their response to abiotic and biotic stress. However, results from this study
766 demonstrate the extent to which values generated by predictive models can vary in the
767 strength of their correlation with the observed values depending on the environment in which
768 they are tested. In particular this is likely to affect predictive models which are trained in one
769 environment and then used to generate values for a different environment, but also for
770 environments which change over time: something which is likely to increase in severity and
771 likelihood in the near future given climate change predictions (Franklin et al., 2016).
772 However, the small-scale comparisons between different selection methods demonstrate the
773 potential for the predictive growth model to be used to select trees which are taller on average
774 than those selected randomly at the field trial sites and which show similar success as the
775 phenotype selection method at one of the sites (GS) but without the need to wait and

phenotype each tree individually. As we had only small sample sizes and a relatively small pool of trees from which to select, the approach will require further testing using a larger set of trees in a further set of experimental trials.

Conclusions

Despite its ecological and economic importance there have been no previous studies exploring association between SNPs and key adaptive traits in *P. sylvestris*. Our study demonstrates the potential usefulness of the high throughput array developed by (Perry et al., 2020) for identifying genes and SNPs with significant associations with phenology and growth traits. Development of a predictive model that has been validated in an independent trial is a demonstration of the application of the approach to breeding trials in the future. However, the study shows a strong influence of site environment on development of the traits. This may affect the ability of predictive models to generate values for populations departing from the environmental conditions in which the models were trained.

Author contributions

The research was designed and planned by AP, SC and WW; GI and SC designed the multi-site field trials and its data collection protocols, data collection were performed by WW, AP, JB, SC, JC and GI; data analysis and manuscript writing were performed by AP. Manuscript review and revision and final approval of the manuscript were performed by all authors.

References

AVIA, K., KÄRKKÄINEN, K., LAGERCRANTZ, U. & SAVOLAINEN, O. 2014. Association of FLOWERING LOCUS T/TERMINAL FLOWER 1-like gene FTL 2 expression with growth rhythm in Scots pine (*Pinus sylvestris*). *New Phytologist,* 204**,** 159-170.

BAI, Q., CAI, Y., HE, B., LIU, W., PAN, Q. & ZHANG, Q. 2019. Core set construction and association analysis of *Pinus massoniana* from Guangdong province in southern China using SLAF-seq. *Scientific reports,* 9**,** 1-13.

BARTHOLOMÉ, J., VAN HEERWAARDEN, J., ISIK, F., BOURY, C., VIDAL, M., PLOMION, C. & BOUFFIER, L. 2016. Performance of genomic prediction within and across generations in maritime pine. *BMC genomics,* 17**,** 1-14.

BEAULIEU, J., DOERKSEN, T., CLÉMENT, S., MACKAY, J. & BOUSQUET, J. 2014. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity,* 113**,** 343-352.

BLANCA, J., ESTERAS, C., ZIARSOLO, P., PÉREZ, D., COLLADO, C., DE PABLOS, R. R., BALLESTER, A., ROIG, C., CAÑIZARES, J. & PICÓ, B. 2012. Transcriptome sequencing for SNP discovery across *Cucumis melo*. *BMC genomics,* 13**,** 280.

BORATYŃSKA, K. & BORATYŃSKI, A. 2007. Taxonomic differences among closely related pines *Pinus sylvestris*, *P. mugo*, *P. uncinata*, *P. rotundata* and *P. uliginosa* as revealed in needle sclerenchyma cells. *Flora-Morphology, Distribution, Functional Ecology of Plants,* 202**,** 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**,** 2633-2635.

BUDDE, K. B., HEUERTZ, M., HERNÁNDEZ-SERRANO, A., PAUSAS, J. G., VENDRAMIN, G. G., VERDÚ, M. & GONZÁLEZ-MARTÍNEZ, S. C. 2014. In situ genetic association for serotiny, a fire-related trait, in Mediterranean maritime pine (*Pinus pinaster*). *New Phytologist,* 201**,** 230-241.

CALLEJA-RODRIGUEZ, A., PAN, J., FUNDA, T., CHEN, Z., BAISON, J., ISIK, F., ABRAHAMSSON, S., WU, H.2020. Evaluation of the efficient of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC Genomics*. 21, 1-17

CAPBLANCQ, T., FITZPATRICK, M. C., BAY, R. A., EXPOSITO-ALONSO, M. & KELLER, S. R. 2020. Genomic prediction of (mal) adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution, and Systematics,* 51.

CHANCEREL, E., LEPOITTEVIN, C., LE PROVOST, G., LIN, Y.-C., JARAMILLO-CORREA, J. P., ECKERT, A. J., WEGRZYN, J. L., ZELENIKA, D., BOLAND, A. & FRIGERIO, J.-M. 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**,** 1-14.

COOKE, J. E., ERIKSSON, M. E. & JUNTTILA, O. 2012. The dynamic nature of bud dormancy in trees: environmental control and molecular mechanisms. *Plant, cell & environment,* 35**,** 1707-1728.

DOUGHERTY, P. M., WHITEHEAD, D. & VOSE, J. M. 1994. Environmental influences on the phenology of pine. *Ecological Bulletins***,** 64-75.

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,* 33**,** 1505-1511.

ECKERT, A. J., MALONEY, P. E., VOGLER, D. R., JENSEN, C. E., MIX, A. D. & NEALE, D. B. 2015. Local adaptation at fine spatial scales: an example from sugar pine (*Pinus lambertiana*, Pinaceae). *Tree Genetics & Genomes,* 11**,** 42.

ECKERT, A. J., WEGRZYN, J. L., PANDE, B., JERMSTAD, K. D., LEE, J. M., LIECHTY, J. D., TEARSE, B. R., KRUTOVSKY, K. V. & NEALE, D. B. 2009. Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). *Genetics,* 183**,** 289-298.

ENDELMAN, J. B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome,* 4**,** 250-255.

FALCONER, D. S. & MACKAY, T. F. C. 1996. *Introduction to quantitative genetics,* Essex, UK., Longman.

FRANKLIN, J., SERRA-DIAZ, J. M., SYPHARD, A. D. & REGAN, H. M. 2016. Global change and terrestrial plant community dynamics. *Proceedings of the National Academy of Sciences,* 113**,** 3725-3734.

GERALDES, A., PANG, J., THIESSEN, N., CEZARD, T., MOORE, R., ZHAO, Y., TAM, A., WANG, S., FRIEDMANN, M. & BIROL, I. 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources,* 11**,** 81-92.

GODDARD, M. E. & HAYES, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics,* 10**,** 381-391.

GONZÁLEZ-MARTÍNEZ, S. C., WHEELER, N. C., ERSOZ, E., NELSON, C. D. & NEALE, D. B. 2007. Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics,* 175**,** 399-409.

GOUDET, J. & JOMBART, T. 2020. hierfstat: estimation and tests of hierarchical F-statistics. R package version 0.5-7 ed.

GROTKOPP, E., REJMÁNEK, M., SANDERSON, M. J. & ROST, T. L. 2004. Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses. *Evolution,* 58**,** 1705-1729.

GYLLENSTRAND, N., CLAPHAM, D., KÄLLMAN, T. & LAGERCRANTZ, U. 2007. A Norway spruce FLOWERING LOCUS T homolog is implicated in control of growth rhythm in conifers. *Plant physiology,* 144**,** 248-257.

HARRELL JR, F. E. & WITH CONTRIBUTIONS FROM C. DUPONT AND MANY OTHERS 2020. Hmisc: Harrell Miscellaneous.

HERBERT R., SAMUEL, S., PATTERSON, G. 1999. Using local stock for planting native trees and shrubs. Forestry Commission, Edinburgh. 10pp

HOLLIDAY, J. A., RITLAND, K. & AITKEN, S. N. 2010. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). *New Phytologist,* 188**,** 501-514.

HOWE, G. T., AITKEN, S. N., NEALE, D. B., JERMSTAD, K. D., WHEELER, N. C. & CHEN, T. H. 2003. From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany,* 81**,** 1247-1266.

HURME, P., REPO, T., SAVOLAINEN, O. & PÄÄKKÖNEN, T. 1997. Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). *Canadian Journal of Forest Research,* 27**,** 716-723.

HURME, P., SILLANPÄÄ, M. J., ARJAS, E., REPO, T. & SAVOLAINEN, O. 2000. Genetic basis of climatic adaptation in Scots pine by Bayesian quantitative trait locus analysis. *Genetics,* 156**,** 1309-1322.

ISABEL, N., HOLLIDAY, J. A. & AITKEN, S. N. 2020. Forest genomics: Advancing climate adaptation, forest health, productivity, and conservation. *Evolutionary Applications,* 13**,** 3-10.

ISIK, F., BARTHOLOMÉ, J., FARJAT, A., CHANCEREL, E., RAFFIN, A., SANCHEZ, L., PLOMION, C. & BOUFFIER, L. 2016. Genomic selection in maritime pine. *Plant Science,* 242**,** 108-119.

JERMSTAD, K., BASSONI, D., JECH, K., WHEELER, N. & NEALE, D. 2001. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. I. Timing of vegetative bud flush. *Theoretical and Applied Genetics,* 102**,** 1142-1151.

JERMSTAD, K. D., BASSONI, D. L., JECH, K. S., RITCHIE, G. A., WHEELER, N. C. & NEALE, D. B. 2003. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas fir. III. Quantitative trait loci-by-environment interactions. *Genetics,* 165**,** 1489-1506.

KANNINEN, M. 2010. *Plantation forests: global perspectives,* London

Washington, DC., Earthscan.

KÖRNER, C. 2006. Significance of temperature in plant life. *Plant growth and climate change***,** 48-69.

KORTE, A. & FARLOW, A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods,* 9**,** 1-9.

KUMAR, S., CHAGNÉ, D., BINK, M. C., VOLZ, R. K., WHITWORTH, C. & CARLISLE, C. 2012. Genomic selection for fruit quality traits in apple (Malus× domestica Borkh.). *PloS one,* 7**,** e36674.

LAUBE, J., SPARKS, T. H., ESTRELLA, N., HÖFLER, J., ANKERST, D. P. & MENZEL, A. 2014. Chilling outweighs photoperiod in preventing precocious spring development. *Global change biology,* 20**,** 170-182.

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**,** 15-24.

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**,** 380.

MACKAY, T. F. 2001. The genetic architecture of quantitative traits. *Annual review of genetics,* 35**,** 303-339.

MEUWISSEN, T., HAYES, B. & GODDARD, M. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157**,** 1819-1829.

MINAMIKAWA, M. F., NONAKA, K., KAMINUMA, E., KAJIYA-KANEGAE, H., ONOGI, A., GOTO, S., YOSHIOKA, T., IMAI, A., HAMADA, H. & HAYASHI, T. 2017. Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. *Scientific reports,* 7**,** 1-13.

MURANTY, H., TROGGIO, M., SADOK, I. B., AL RIFAÏ, M., AUWERKERKEN, A., BANCHI, E., VELASCO, R., STEVANATO, P., VAN DE WEG, W. E. & DI GUARDO, M. 2015. Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture research,* 2**,** 1-12.

PARADIS, E. & SCHLIEP, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35**,** 526-528.

PARCHMAN, T. L., GEIST, K. S., GRAHNEN, J. A., BENKMAN, C. W. & BUERKLE, C. A. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC genomics,* 11**,** 180.

PARK J-H., GAIL, M. H., WEINBERG, C. R., CARROLL, R. J., CHUNG, C., WANG, Z., CHANOCK, S. J., FRAUMENI, J. F., CHATTERJEE, N. 2011. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *PNAS*, 108, 18026-18031

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*.

PLOMION, C., DUREL, C.-E. & O'MALLEY, D. 1996. Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. *Theoretical and Applied Genetics,* 93**,** 849-858.

PRUNIER, J., PELGAS, B., GAGNON, F., DESPONTS, M., ISABEL, N., BEAULIEU, J. & BOUSQUET, J. 2013. The genomic architecture and association genetics of adaptive characters using a candidate SNP approach in boreal black spruce. *BMC genomics,* 14**,** 368.

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**,** 44-62.

QUESADA, T., GOPAL, V., CUMBIE, W. P., ECKERT, A. J., WEGRZYN, J. L., NEALE, D. B., GOLDFARB, B., HUBER, D. A., CASELLA, G. & DAVIS, J. M. 2010. Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). *Genetics,* 186**,** 677-686.

R CORE TEAM 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

REPO, T., ZHANG, G., RYYPPÖ, A., RIKALA, R. & VUORINEN, M. 2000. The relation between growth cessation and frost hardening in Scots pines of different origins. *Trees,* 14**,** 456-464.

RESENDE, M. D., RESENDE, M. F., SANSALONI, C. P., PETROLI, C. D., MISSIAGGIA, A. A., AGUIAR, A. M., ABAD, J. M., TAKAHASHI, E. K., ROSADO, A. M. & FARIA, D. A. 2012a. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist,* 194**,** 116-128.

RESENDE, M. F., MUÑOZ, P., RESENDE, M. D., GARRICK, D. J., FERNANDO, R. L., DAVIS, J. M., JOKELA, E. J., MARTIN, T. A., PETER, G. F. & KIRST, M. 2012b. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics,* 190**,** 1503-1510.

SALMELA, M. J., CAVERS, S., COTTRELL, J. E., IASON, G. R. & ENNOS, R. A. 2011. Seasonal patterns of photochemical capacity and spring phenology reveal genetic differentiation among native Scots pine (*Pinus sylvestris* L.) populations in Scotland. *Forest Ecology and Management,* 262**,** 1020-1029.

SCHLICHTING, C. D. 1986. The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics,* 17**,** 667-693.

SEGURA, V., VILHJÁLMSSON, B. J., PLATT, A., KORTE, A., SEREN, Ü., LONG, Q. & NORDBORG, M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics,* 44**,** 825.

STEVENS, K. A., WEGRZYN, J. L., ZIMIN, A., PUIU, D., CREPEAU, M., CARDENO, C., PAUL, R., GONZALEZ-IBEAS, D., KORIABINE, M. & HOLTZ-MORRIS, A. E. 2016. Sequence of the sugar pine megagenome. *Genetics,* 204**,** 1613-1626.

STOCKS, J. J., METHERINGHAM, C. L., PLUMB, W. J., LEE, S. J., KELLY, L. J., NICHOLS, R. A. & BUGGS, R. J. 2019. Genomic basis of European ash tree resistance to ash dieback fungus. *Nature Ecology & Evolution,* 3**,** 1686-1696.

THISTLETHWAITE, F. R., RATCLIFFE, B., KLÁPŠTĚ, J., PORTH, I., CHEN, C., STOEHR, M. U. & EL-KASSABY, Y. A. 2017. Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. *BMC genomics,* 18**,** 1-16.

TRICK, M., LONG, Y., MENG, J. & BANCROFT, I. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant biotechnology journal,* 7**,** 334-346.

WACHOWIAK, W., BALK, P. A. & 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**,** 117.

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**,** 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**,** 1729-1743.

WACHOWIAK, W., PERRY, A., DONNELLY, K. & CAVERS, S. 2018a. Early phenology and growth trait variation in closely related European pine species. *Ecology and evolution,* 8**,** 655-666.

WACHOWIAK, W., TRIVEDI, U., PERRY, A. & CAVERS, S. 2015. Comparative transcriptomics of a complex of four European pine species. *BMC genomics,* 16**,** 234.

WACHOWIAK, W., ZABOROWSKA, J., ŁABISZAK, B., PERRY, A., ZUCCA, G. M., GONZÁLEZ-MARTÍNEZ, S. C. & CAVERS, S. 2018b. Molecular signatures of divergence and selection in closely related pine taxa. *Tree genetics & genomes,* 14**,** 83.

WANG, M. & XU, S. 2019. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity,* 123**,** 287-306.

WESTBROOK, J. W., RESENDE JR, M. F., MUNOZ, P., WALKER, A. R., WEGRZYN, J. L., NELSON, C. D., NEALE, D. B., KIRST, M., HUBER, D. A. & GEZAN, S. A. 2013. Association genetics of oleoresin flow in loblolly pine: discovering genes and predicting phenotype for improved resistance to bark beetles and bioenergy potential. *New Phytologist,* 199**,** 89-100.

WESTBROOK, J. W., ZHANG, Q., MANDAL, M. K., JENKINS, E. V., BARTH, L. E., JENKINS, J. W., GRIMWOOD, J., SCHMUTZ, J. & HOLLIDAY, J. A. 2020. Optimizing genomic selection for blight resistance in American chestnut backcross populations: A trade-off with American chestnut ancestry implies resistance is polygenic. *Evolutionary applications,* 13**,** 31-47.

WUERTZ, D., SETZ, T. & CHALABI, Y. 2020. fBasics: Rmetrics - markets and basic statistics. R package version 3042.89.1 ed.

XIONG, J. S., MCKEAND, S. E., ISIK, F., WEGRZYN, J., NEALE, D. B., ZENG, Z.-B., E SILVA, L. D. C. & WHETTEN, R. W. 2016. Quantitative trait loci influencing forking defects in an outbred pedigree of loblolly pine. *BMC genetics,* 17**,** 138.

YANG, J., BENYAMIN, B., McEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. H., MONTGOMERY, G. W., 2010. Common SNPs explain a large proportion fo the heritability for human height. *Nature genetics*. 42, 565-569

ZIMIN, A., STEVENS, K. A., CREPEAU, M. W., HOLTZ-MORRIS, A., KORIABINE, M., MARÇAIS, G., PUIU, D., ROBERTS, M., WEGRZYN, J. L. & DE JONG, P. J. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics,* 196**,** 875-890.

ZIMIN, A. V., STEVENS, K. A., CREPEAU, M. W., PUIU, D., WEGRZYN, J. L., YORKE, J. A., LANGLEY, C. H., NEALE, D. B. & SALZBERG, S. L. 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience,* 6**,** giw016.

Data accessibility

Phenotypes, sampling locations and SNPs will be uploaded to the EIDC (https://eidc.ac.uk/)