Association mapping identified novel candidate loci affecting wood formation in Norway spruce

John Baison[1], Amarylis Vidalis[3], Linghua Zhou[1], Zhi-Qiang Chen[1], Zitong Li[9], Mikko J. Sillanpää[10], Carolina Bernhardsson[1,2], Douglas Scoffield[5], Nils Forsberg[1], Lars Olsson[8], Bo Karlsson[11], Harry Wu[1], Pär K. Ingvarsson[4], Sven-Olof Lundqvist[8,12], Totte Niittylä[1]*, M Rosario García-Gil[1]*

*Double last authorship.

Corresponding authors: m.rosario.garcia@slu.se; john.baison@slu.se

Telephone Number: +46 (0) 907868413

1. Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Science, Umeå, Sweden
2. Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden
3. Section of Population Epigenetics and Epigenomics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, München, Germany
4. Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden
5. Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden
6. Department of Ecology and Genetics: Evolutionary Biology, Uppsala University, Uppsala, Sweden
7. Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Science, Uppsala, Sweden.
8. RISE Bioeconomy, Drottning Kristinas väg 61, SE-114 86 Stockholm, Sweden

9. Ecological Genetics Research Unit, Department of Biosciences, University of Helsinki, P.O. Box 65, FI-00014 Helsinki, Finland
10. Department of Mathematical Sciences, Biocenter Oulu, University of Oulu, Finland
11. Skogforsk, Ekebo 2250 SE-268 90 Svalöv, Sweden
12. IIC, Rosenlundsgatan 48B, SE-118 63 Stockholm, Sweden

Total Word Count: 6500

Introduction Word Count: 953

Materials and methods Word Count: 1930

Results Word Count: 1565

Discussion Word Count: 1917

Conclusion Word Count: 91

Acknowledgements Word Count: 48

Figures

Fig. 1 Outline of the association mapping approach (Colour) Page 9

Fig 2. Phenotype trajectories (Colour) Page 13

Fig 3. PCA plot of all the 517 mother trees (Colour) Page 20

Fig 4. Decay of linkage disequilibrium (LD) Page 22

Fig 5. Frequencies of significant markers plotted against linkage map Page 33

Tables

Table 1 List of the phenotypes, their abbreviations and measurement unit Page 8

Table 2 Phenotypes, Latent Traits, SNP, SNP feature, frequency and PVE Page 24

Supplementary Data

Fig.S1 Phenotype Trajectories (Colour) Page 1

Table S1 ConGenIE BLAST search of contigs with significant QTLs Page 10

Methods S1 PVE evaluation of a QTL Page 33

Methods S2 Association mapping script Page 33

Methods S3 Stabilization selection script Page 34

Summary

- Norway spruce (*Picea abies*) is an important boreal forest tree species of significant ecological and economic importance. Hence there is a strong imperative to dissect the genetics controlling important wood quality traits in the species.
- We performed a functional genome-wide association mapping of 17 wood traits in Norway spruce using 178101 single-nucleotide polymorphisms (SNPs) generated from exome genotyping of 517 mother trees. The wood traits were defined using functional modelling of wood properties across annual growth rings.
- Association mapping was performed using a multilocus LASSO penalized regression method and we detected a total of 51 significant SNPs from 39 candidate genes that are involved in wood formation.
- Our study represents the first functional multi-locus genome-wide association mapping (AM) in Norway spruce. The results advance our understanding of the genetics influencing wood traits, identify novel candidate genes for further functional studies and support current Norway spruce breeding efforts.

Key Words: Genome-wide association mapping, Sequence capture, Functional mapping, Single nucleotide polymorphisms, Candidate genes, Norway spruce

Introduction

Norway spruce (*Picea abies* (L.) Karst.) is a dominant boreal softwood species of significant economic and ecological importance (Hannrup *et al.*, 2004). Long-term Norway spruce breeding programmes for improvement of growth and survival were initiated in the 1940s and recently, wood quality has become one of the priority traits (Bertaud & Holmbom, 2004; Hannrup *et al.*, 2004). Norway spruce breeding in Sweden complete one cycle in about 20 years and such long generation times make improvements in growth and wood quality very slow. Among wood quality traits, wood density is considered a key indicator of stability, strength and stiffness of sawn timber (Hauksson *et al.*, 2001). Several studies of wood quality observed that fast growth conflicts with high quality wood, as shown by the negative genetic correlation between wood volume growth and density in Norway spruce (Olesen, 1977; Dutilleul *et al.*, 1998; Chen *et al.*, 2014). In order to combine fast growth and desirable wood properties through breeding, and to shorten the breeding cycle, it is therefore imperative to design effective early selection methods and breeding strategies. In an effort to design optimal breeding and selection strategies for reducing or breaking negative genetic correlations between traits it is essential to identify alleles that are responsible for generating favourable or unfavourable genetic correlations (Hallingbäck *et al.*, 2014).

When DNA markers were first introduced in 1980s, tree breeders were provided a possibility to correlate phenotypes with polymorphic DNA markers and to conduct selection using genotypes instead of phenotypes (Lande & Thompson, 1990). Groover et al. (1994) first identified quantitative trait loci (QTL) for wood density variation in loblolly pine using linkage analyses based on segregating family pedigrees. However, maker-aided selection (MAS) based on results from QTL analyses was never implemented in practical tree breeding due to the so-called Beavis effect (e.g. inflated estimates of allelic effects and underestimation

of QTL number for economically important traits) (Beavis, 1998), inconsistent associations among different families and the low transferability of markers (Strauss *et al.*, 1992). Association Mapping (AM) is a more powerful QTL detection method that was introduced to tree genetics using a candidate gene approach (Thumma *et al.*, 2010). AM overcomes the limited resolution of family-based QTL mapping by relying on historical recombination in the mapping population (Neale & Savolainen, 2004; Thavamanikumar *et al.*, 2013; Huang & Han, 2014). The effectiveness of AM relies on genome-wide levels of LD, which decays rapidly within coding regions in conifer species, however, it may be extensive in certain non-coding regions (Moritsuka *et al*., 2012). Fast-decaying LD, coupled with complex polygenic nature for both growth and wood quality traits (Hall *et al.*, 2016) implies that a large number of genomic regions need to be investigated to identify significant QTL (Beaulieu *et al*., 2011).

The availability of a draft genome sequence for Norway spruce (Nystedt *et al.*, 2013) has opened new possibilities for the development of genetic markers to conduct both AM at the genome-wide level (genome-wide association, GWAS) and genomic selection (GS). Several reduced representation-based approaches such as sequence capture and transcriptome sequencing (Hirsch *et al.*, 2014) have been developed as complexity-reduction methods suited for studying large genomes, such as the 20Gb Norway spruce genome. These approaches reduce the sequence space by decreasing the repetitive sequence content of the genome. In this study we employed a solution-based sequence capture method.

Several AM studies have been performed in trees and have identified genetic loci linked to, for instance, wood properties in *Populus trichocarpa (Porth et al., 2013)*, adaptive traits in *Pinus contorta* (Parchman *et al.*, 2012) and to wood quality traits in *Eucalyptus* (Porth *et al.*, 2013; Resende *et al.*, 2012). Such studies aimed at dissecting the genetic basis of wood properties can benefit from the application of mathematical functions that account for year-to-year variation across annual growth rings, cambial age and distance from pith (Li *et*

125 *al.*, 2014). Mathematical modelling allows the incorporation of phenotypic growth trends that
126 increase the precision and resolution of QTL detection through the integration of the
127 phenotype information over multiple time points and reduction of residual variance (Ma *et al.*,
128 2002). Such functional mapping analysis can be conducted using a multistage approach
129 (Heuven & Janss, 2010). First, the phenotype trends of each individual are modelled using
130 curve-fitting methods and the parameters describing the curve are then considered as latent
131 traits. The latent traits are then used in an independent association analyses to search for
132 genomic regions affecting the trait and to estimate genetic marker effects (Li *et al.*, 2014).
133 In this study, we applied a functional genome-wide association mapping (AM)
134 approach to identify genomic regions contributing to wood quality traits in Norway spruce
135 [*Picea abies* (L.) Karst.]. Estimated breeding values (EBVs) were calculated for growth and
136 wood quality traits at the resolution of annual growth rings and were then used to extract
137 latent traits from fitting quadratic splines, Fig. 1a. We applied quadratic splines since
138 traditional analyses that utilise a single point data across annual growth rings may confound
139 the analyses by averaging across a full sample. Such averaging may obscure mechanisms
140 acting at specific time points during wood formation and will make identification of
141 underlying genes more difficult. In this study, we have refined our data in order to cover
142 within ring features in earlywood (EW) and latewood (LW), as well as in a more weather
143 influenced part in between named transitionwood (TW). This study has also performed the
144 first analysis of number of cells per ring calculated from SilviScan data. Penalized LASSO
145 regression (Tibshirani, 1996) and the stabilizing selection probability method of
146 (Meinshausen & Bühlmann, 2010) were then used, Fig. 1c, to detect significant associations
147 between latent traits derived from EBVs and 178101 SNP markers covering the Norway
148 spruce genome, Fig. 1b.
149

Materials and Methods

Plant material and phenotype data

Plant material and phenotype data used in this study have previously been described in Chen et al. (2014). In brief, two progeny trails were established in 1990 in Southern Sweden (S21F9021146 aka F1146 (trial1) and S21F9021147 aka F1147 (trial2)). These trials were composed of 1373 and 1375 open pollinated families, respectively, and form the basis of our analyses. We selected 517 families in 112 sampling stands to use in the investigation of wood properties. At each site, increment wood cores of 12 mm were collected from six trees of the selected families at breast height (1.3 m) (6 progeny × 2 sites = 12 progenies in total). A total of 5618 trees, 2973 and 2645 trees from the F1146 and F1147 trails respectively, were analysed. The pith to bark profiling of the wood physical attributes was analysed using the SilviScan technology (Evans 1994, 2006) at Innventia, Stockholm, Sweden, where the initial data evaluations were performed using customized methods. These methods focus on the identification and dating of all annual rings and their compartments of earlywood, transitionwood and latewood. For the current study, Innventia also calculated three additional traits, number of cells per ring (NC), wood percentage (WP), and a trait named Mass Index (MI), introduced to express the relative amount of biomass, all derived from the SilviScan data. MI was then used to identify trees with an uncommon positive correlation between density and growth, that is more biomass. The traits included in the current study are listed in Table 1.

175

176 Table 1 List of the phenotypes, their abbreviations and measurement unit.

Phenotype	Abbreviation	Unit
Ring wood density	WD	kg m^{-3}
Early wood density	EWD	kg m^{-3}
Transition wood density	TWD	kg m^{-3}
Latewood density	LWD	kg m^{-3}
Ring width	RW	µm
Early wood ring width	ERW	µm
Transition wood ring width	TRW	µm
Latewood ring width	LRW	µm
Ring number of cells	NC	
Early wood number of cells	ENC	
Transition wood number of cells	TNC	
Latewood number of cells	LNC	
Early wood percentage	EP	%
Transition wood percentage	TP	%
Latewood percentage	LP	%
Early/Latewood percentage	EP/LP	%
Modulus of elasticity	MOE	GPa
Mass Index (Density x Growth)	MI	

177

178

179

180

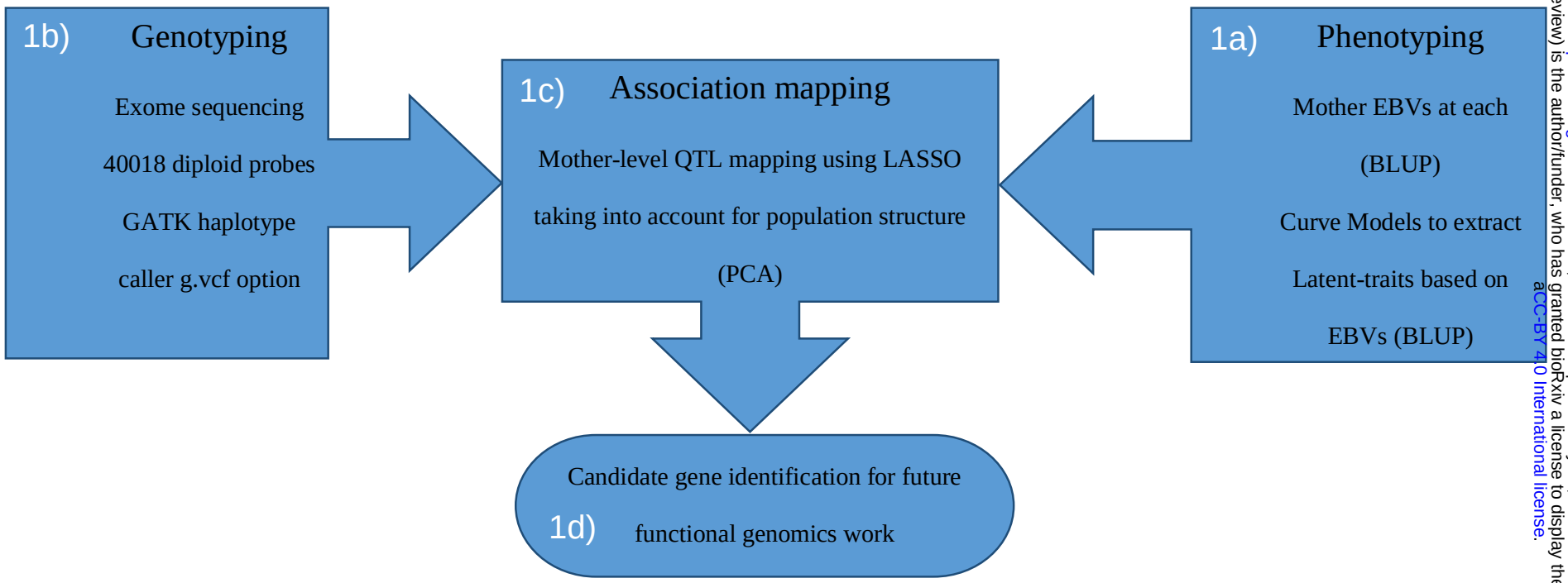


Fig 1. Outline of the association mapping approach: 1a) Mother Estimated breeding values (EBVs) were determined using a univariate, bivariate or multivariate mixed linear model based on the different fitness of the model with the resultant values adjusted with the mean. The adjusted EBVs were plotted against cambial age (annual ring number) to produce time trajectories for each trait. A quadratic spline curve model was then applied to the EBVs to estimate latent-traits. 1b) Sequence capture on the 517 from 40018 diploid probes resulted in 178101 single-nucleotide

polymorphisms (SNPs). 1c) Association mapping was the performed using a multi-locus LASSO penalized regression method with PCA components acting as covariates accounting for the population structure in our mother trees. 1d) Finally, a candidate gene identification process for contigs with significant SNPs was conducted in ConGenIE and public sequence databases.

194

195 *Statistical analysis*

196 EBVs were calculated for each cambial age (annual ring) separately and used for statistical
197 modelling to derive latent traits. The variance and covariance components were estimated
198 using ASREML 4.0 (Gilmour *et al.*, 2014) as described in Chen et al., (2014). In brief, the
199 EBVs at each cambial age were estimated using a univariate, bivariate or multivariate mixed
200 linear models. The fit of different models were evaluated using the Akaike Information
201 Criteria (AIC) and the optimal model was selected based on a compromise of model fit and
202 complexity. Breeding values were then centred in order to obtain within genotype trends. A
203 univariate linear mixed model for joint-site analysis was implemented as:

204 The following univariate linear mixed model for joint-site analysis was fitted as:

205 $$Y_{ijkl} = u + S_i + B_{j(i)} + F_k + SF_{ik} + e_{ijkl} \quad [1]$$

206 where Y_{ijkl} is the observation on the *l*th tree from the *k*th family in jth block within the
207 *i*th site, *u* is the general mean, S_i and $B_{j(i)}$ are the fixed effects of the *i*th site and the *j*th block
208 within the ith site, respectively, F_k and SF_{ik} are the random effects of the *k*th family and the
209 random interactive effect of the *i*th site and *k*th family, respectively, e_{ijkl} is the random
210 residual effect. Multivariate mixed linear models were used to estimate BV for different
211 phenotype traits if the model fitted better than bivariate or univariate based on AIC.

212 A number of trees were observed that broke the negative correlation usually observed
213 between density and growth. These trees exhibited both high density and fast growth, thus
214 larger biomass. In order to identify putative genes involved in this favourable combination of
215 traits, we defined a new trait termed Mass Index (MI), that we subsequently used in the
216 association mapping. The MI was defined as follows:

217 *Mass index = (Individual average density/population average density) * (individual*
218 *cross-sectional area / population average cross-sectional area).*

219 The index was then treated as a new dynamic trait in the AM analyses where
220 individuals with an index > 1 indicate a wood mass per length unit than the population
221 average in the cross-section at breast height. The index was calculated for all progeny and
222 were used to calculate BVs for the 517 mother trees.

223 The EBVs were plotted against cambial age (annual ring number) to produce time
224 trajectories for each trait (Fig. 2 and Fig. S1) and used to estimate latent curve parameters. At
225 the first stage, all the trajectories versus cambial age were fitted with a quadratic spline with
226 multiple knots in order to describe the dynamics of the EBVs across age. In this study, this
227 was done with the values of four parameters obtained from the spline fitting: the intercept, the
228 slope and two knot parameters (K1 and K2). The intercept and slopes were used to evaluate
229 the mean and rate of change for the trait across the annual rings, respectively. K1 and K2
230 represent inflection points in the cambial age trajectories where the development of the EBVs
231 enters new phases. These two points (K1 and K2) are therefore supposed to have biological
232 significance, warranting a closer analysis of the genes imparting these shifts in the EBVs
233 dynamics. The four latent traits show lower correlations compared to the direct measurements
234 on the original scales and they also have constant variances, thereby reducing the need to
235 account for residual dependencies in the model (Li *et al.*, 2014).

236 The general definition of a quadratic spline with multiple knots is as follows:

237 $$\beta(t) = b_0 + b_1 t + b_2 t^2 + + b_3 (t-t_t)^2_+ + b_4 (t-t_2)^2_+ + \ldots + b_{2+k} (t-t_k)^2, \quad (1)$$

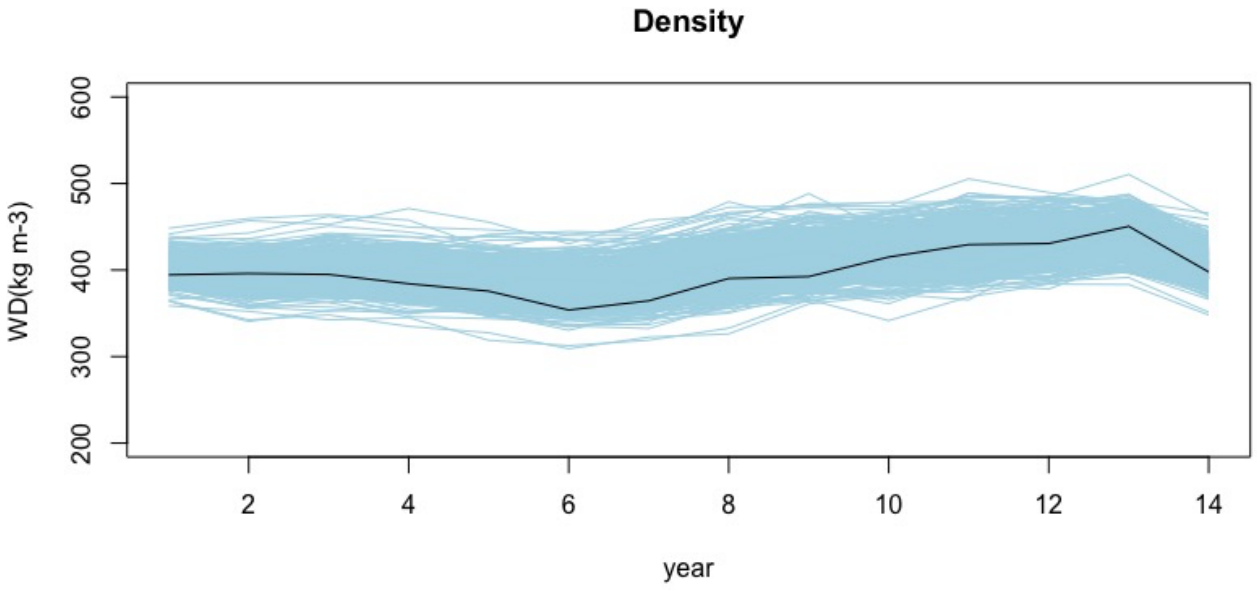
238 which is continuous and where t_i (i=1,…,k; $t_1 < t_2 \ldots < t_k$) are defined as knots, and $(t - t_i)^2_+ = (t$
239 $- t_i)^2$ if $t > t_i$ (t_i>0; i=1,…,k), and otherwise is equal to zero. The number of knots has to be
240 properly defined in order to provide an accurate description of the data under investigation, as
241 well as functional starting points for the search of their locations (Li *et al.*, 2015). In our case,

242 since the growth pattern of wood property traits were not complex, we choose two knots of
243 the time interval.

244 Hence, the quadratic spline model to describe the growth trajectory of individual *i*
245 applied in this study was defined as:

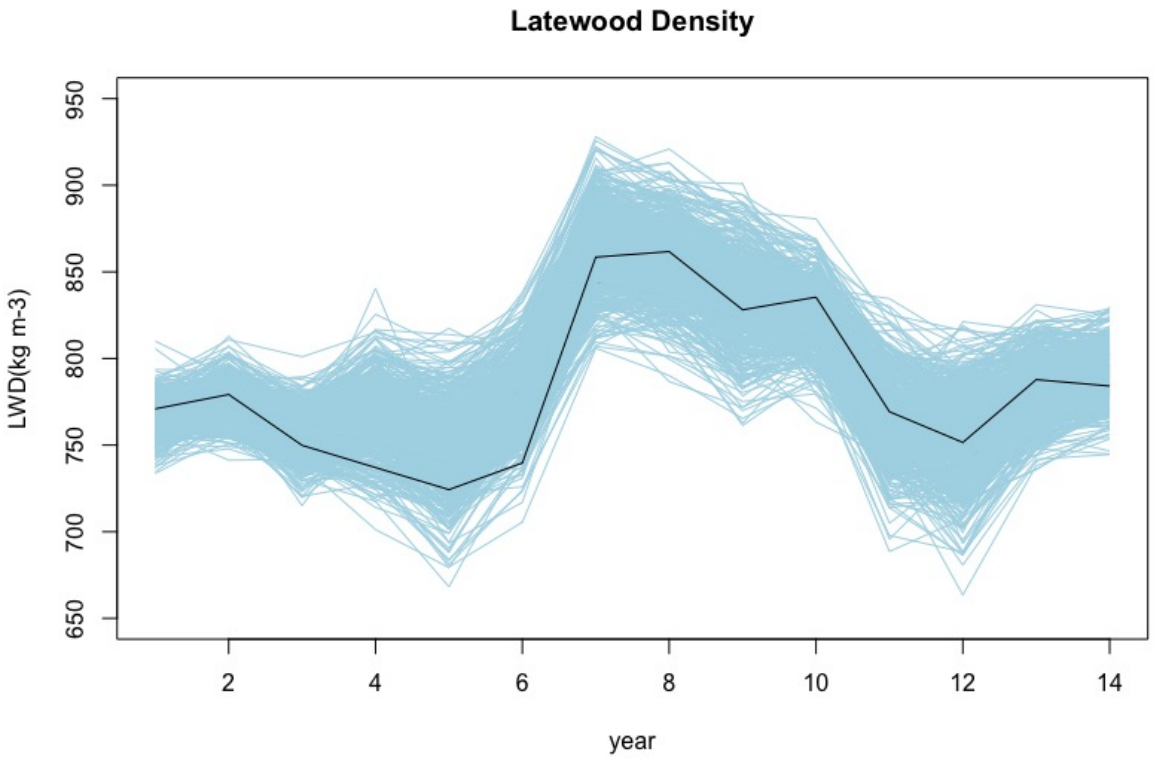
246 $$y_i(t) = \beta_0 + \beta_1 t + \beta_2 (t - t_1)_+ + \beta_3 (t - t_2)_+ + \varepsilon_i(t), \qquad \varepsilon_i(t) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2). \qquad (2)$$

247 Then the intercept β_0, slope β_1, β_2, (Knot 1 (k1)) and β_3 (Knot 2 (k2)) are estimated by
248 standard least squares, and their estimates were considered as the latent trait in the subsequent
249 QTL analysis conducted in R-studio (Team, 2015). The latent traits were then analysed using
250 the LASSO model in order to identify SNPs showing significant associations to the traits.
251
252
253
254
255
256
257 A)

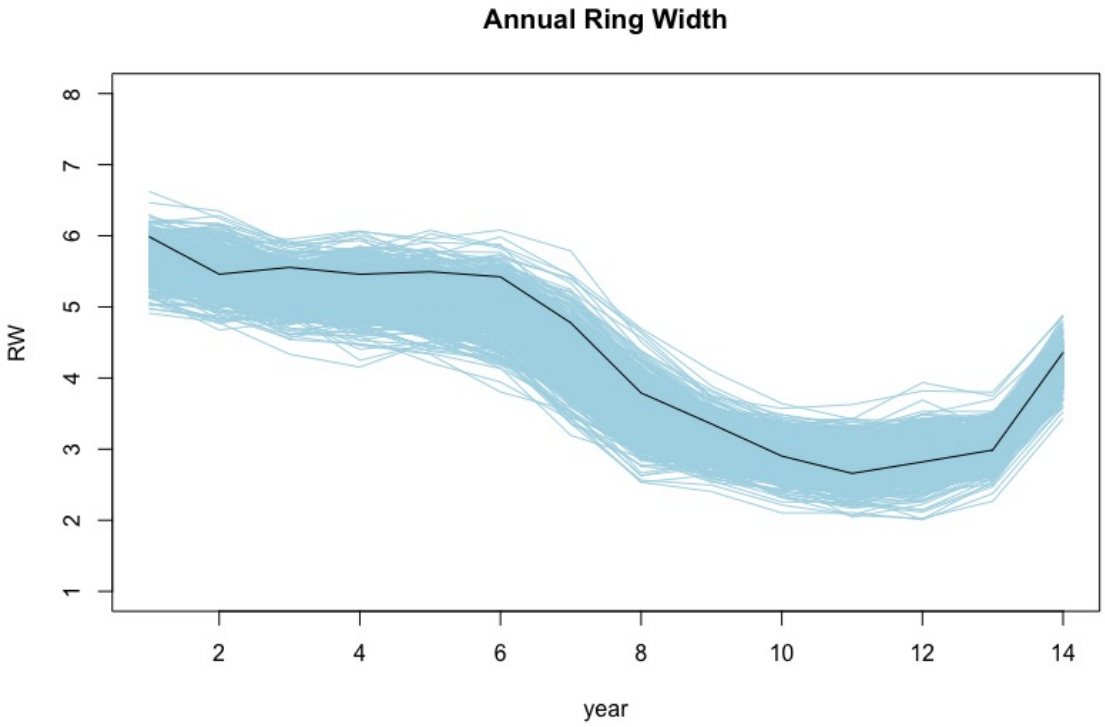


258

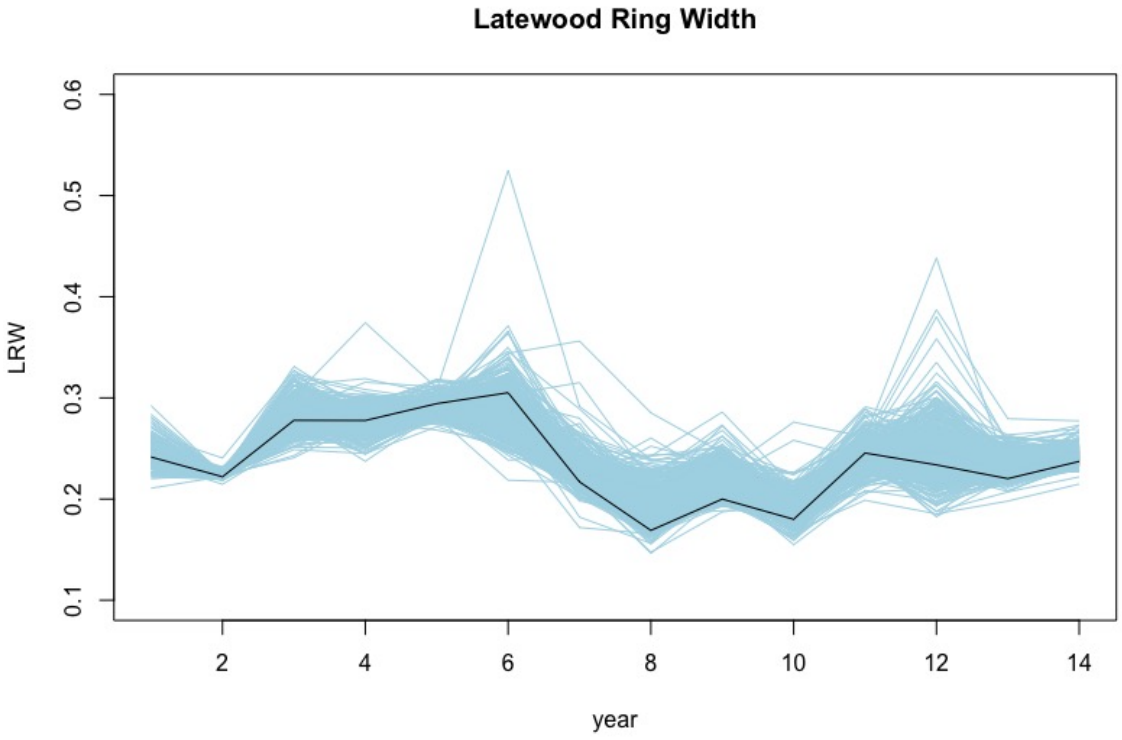
B)



C)



264 D)



265

266

267 Fig 2. EBV trajectories of four wood quality traits by time:(A) wood density, (B) latewood
268 density, (C) annual ring width and (D) latewood ring width. Individual trajectories for each
269 trait are shown in light blue lines and the black line represents the mean trajectory for the
270 phenotype. These individual trajectories were used to determine the four latent traits of each
271 tree, using quadratic splines with two knots.

272

273 *Sequence capture, genotyping and SNP annotation*

274 Total genomic DNA was extracted from 517 unrelated individuals using the Qiagen Plant
275 DNA extraction protocol with DNA quantification performed using the Qubit® ds DNA
276 Broad Range (BR) Assay Kit (Oregon, USA). Sequence capture was performed using the 40
277 018 diploid probes previously designed and evaluated for *P. abies* (Vidalis *et al.*, 2018) and
278 samples were sequenced to an average depth of 15x using an Illumina HiSeq 2500 (San

Diego, USA). Raw reads were mapped against the *P.abies* reference genome v1.0 using BWA-mem (Langmead & Salzberg, 2012; Li & Durbin, 2009). SAMTools v.1.2 (Li *et al.*, 2009) and Picard v.1.140 (McKenna *et al.*, 2010) were used for sorting and removal of PCR duplicates. Variant calling was performed using GATK HaplotypeCaller v.3.6 (McKenna *et al.*, 2010) in gVCF output format. Samples were then merged into batches of ~200 before all 517 samples were jointly called.

Variant Quality Score Recalibration (VQSR) method was performed in order to avoid the use of hard filtering for exome/sequence capture data. For the VQSR analysis two datasets were created, a training subset and input file. The training dataset was derived from a Norway spruce genetic mapping population with loci showing expected segregation patterns (Bernhardsson *et al.*, 2018) and assigned a prior value of 15.0. The input file was derived from the raw sequence data using GATK best practices with the following parameters: extended probe coordinates by +100 excluding INDELS, excluding LowQual sites, and keeping only bi-allelic sites. The following annotation parameters QualByDepth (QD), MappingQuality (MQ) and BaseQRankSum, with tranches 100, 99.9, 99.0 and 90.0 were then applied for the determination of the good versus bad variant annotation profiles. After obtaining the variant annotation profiles, the recalibration was then applied to filter the raw variants. Using VCFTools v.0.1.13 (Danecek *et al.*, 2011), SNP trimming and cleaning involved the removal of any SNP with a minor allele frequency (MAF) and "missingness" of < 0.05 and >20%, respectively.

The resultant SNPs were annotated using default parameters for snpEff 4 (Cingolani *et al.*, 2012). Ensembl general feature format (GTF, gene sets) information was utilized to build the *P. abies* snpEff database.

Genetic Structure

A principal component analysis (PCA) was performed on the sampled trees using SNPs derived from the sequence capture data. SNPs with missing values following VQSR were imputed using the nearest neighbour principle in TASSEL (Bradbury *et al.*, 2007). This approach was essential considering that PCA demands no missing data points. The covariate matrix derived from the PCA was then displayed by plotting principal component 1 scores against principal component 2 scores in Figure 2. The PCA plot was used to make inference about the population structure. The first two components of the PCA covariate matrix explaining most of the variation were then applied to the AM to account for population structure and correcting for any stratification within the study.

Linkage disequilibrium was calculated using VCFtools v.0.1.13 software using the squared correlation coefficient between genotypes (r2) within scaffolds using the "geno-r2". The trend-line of LD decay with physical distance was fitted using nonlinear regression (Hill & Weir, 1988) and the regression line was displayed using R (Team, 2015).

Trait Association Mapping

The LASSO model as described by Li *et al* (2014), Fig. 1c, was applied to all latent traits for the detection of QTLs.

The LASSO model:

$$\min_{(\alpha_0,\alpha_j)} \frac{1}{2n}\sum_{i=1}^{n}(y_i-\alpha_0-\sum_{j=1}^{p}x_{ij}\alpha_j)^2+\lambda\sum_{j=1}^{p}|\alpha_j|, \qquad (3)$$

where y_i is the phenotypic value of an individual i (i=1,…,n; n is the total number of individuals) for the latent trait β_0, β_1, β_2 or β_3, α_0 is the population mean parameter, x_{ij} is the genotypic value of individual i and marker j coded as 0, 1 and 2 for three marker genotypes AA, AB and BB, respectively, α_j is the effect of marker j (i=1,…,n; n is the total number of markers), and λ (>0) is a shrinkage tuning parameter. A fundamental idea of LASSO is to

328 utilize the penalty function to shrink the SNP effects toward zero, and only keep a small
329 number of important SNPs which are highly associated with the trait in the model.

330 The stability selection probability (SSP) of each SNP being selected to the model was
331 applied as a way to control the false discovery rate and determine significant SNPs (Gao *et*
332 *al.*, 2014; Li & Sillanpää, 2015). For a marker to be declared significant, a SSP inclusion ratio
333 (Frequency) was used with an inclusion frequency of at least 0.52 for all traits. This frequency
334 inferred that the expected number of falsely selected markers was less than one (1), according
335 to the formula of Buhlmann *et al*, (2014). Population structure was accounted for in all
336 analyses by including the first two principal components based on the genotype data as
337 covariates into the model. An adaptive LASSO approach (Zou et al. 2006) was used to
338 determine the percentage of phenotypic variance (PVE) (H^2_{QT}) of all the QTLs (Methods S1).
339 These analysis were all performed in R (Team, 2015), with all the scripts provided in the
340 supplementary material.

341

342 *Candidate gene mining*

343 To assess homology of contigs with significant associations, a BLAST search was performed
344 against ConGenIE and public sequence databases, Fig. 1d. After the identification of
345 significant SNPs, the complete *P. abies* contigs that harboured the QTLs were then BLASTed
346 against the ConGenIE database and if no significant hit were detected the whole contig was
347 then extracted. The complete contigs in fasta format were then used to perform a nucleotide
348 BLAST (Blastn) search using the option for only highly similar sequences (megablast) in the
349 National Center for Biotechnology Information (NCBI) nucleotide collection database
350 (https://blast.ncbi.nlm.nih.gov/Blast.cgi?).

351

352

353

354

355 **Results**

356 *Norway spruce SNP identification and mapping population structure*

357 All of the 517 Norway spruce mother trees in the study were considered for variant detection
358 and an average of 1.5 million paired end reads were sequenced per individual for the 40019
359 exome capture probes. This resulted in the identification of 178101 high confidence SNPs. In
360 order to account for effects derived from population stratification we performed a PCA and
361 identified two separate main population groups as well as a number of individuals scattered in
362 between these two main groups. Nevertheless, the differences due to population structure
363 were small with the first two principal components cumulatively explaining only 2.18% of the
364 genetic variation observed (Fig. 3). LD was also determined between all the SNPs, within
365 contigs as well as within significant contigs only and LD decay across physical distance is
366 plotted in Fig 4.

367

368

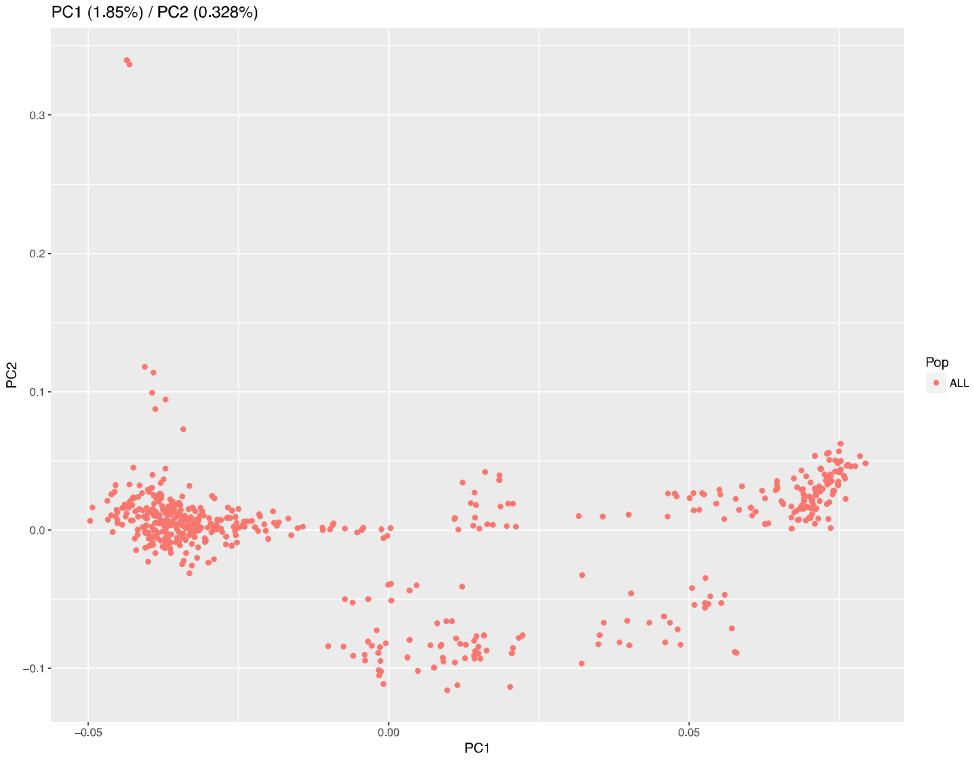
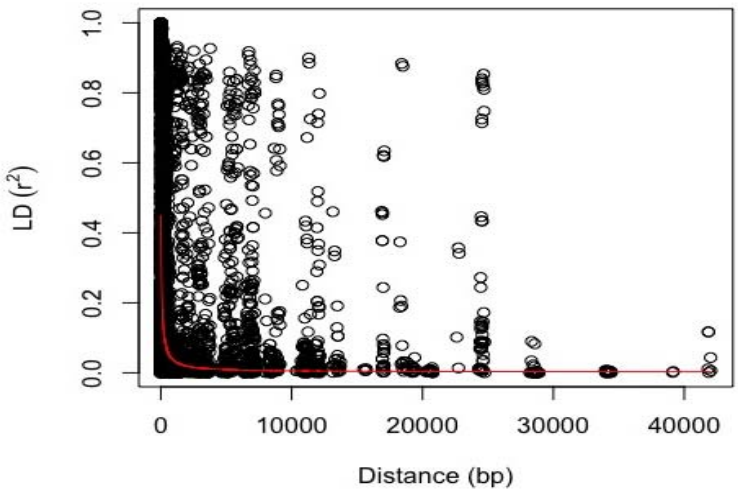


Fig 3. PCA plot of all the 517 mother trees. After VQSR and hard filtering of the SNPs, imputation using the nearest neighbor principle was performed in TASSEL. The PCA indicated a presence of two distinct populations within the 517 mother trees from the Norway spruce breeding program in Sweden. The inferred population structure was used for the correction of stratification within the AM analysis.

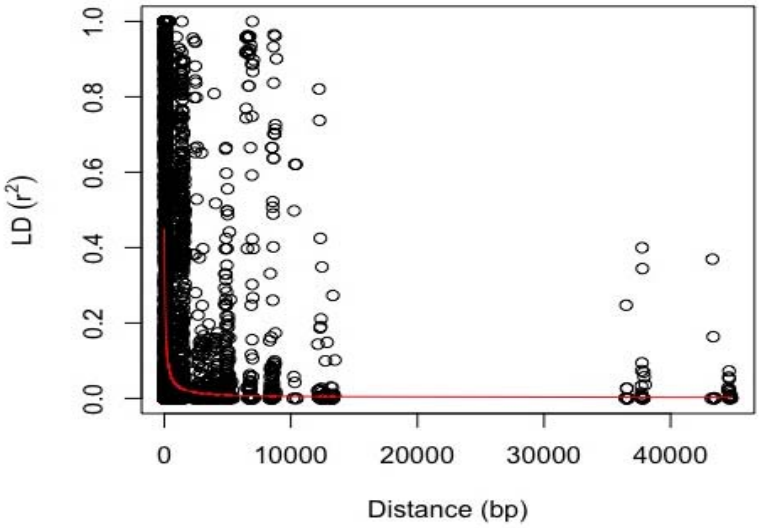
Significant SNPs affecting wood traits

Employing a SSP inclusion frequency of at least 0.52 on the intercept, slope and two knots (K1 and K2) as latent traits, we detected 51 significant QTL across 17 individual traits with the phenotypic variances explained QTL (H^2_{QTL}) ranging from 0.01 to 4.93% (Table 2). Several appreciable QTLs were identified with WD and RW having the highest number of associations, at a total of 13 and 14 QTLs, respectively. This was followed by EP/LP-ratio, which had six QTLs. WD, RW and EP/LP were the only three observed traits that have QTLs

384 detected in all four latent traits. For these three phenotypes, the majority of the QTLs were
385 detected when the average ring phenotype was used to derive the latent traits (Table 2). NC
386 associated with one QTL that was detected for the entire ring, whilst six QTLs were identified
387 when EW, TW and LW were analysed separately, with H^2_{QTL} ranging from 0.01-4.93%
388 (Table 2).
389
390
391
392
393
394
395
396
397 A)



398
399
400 B)



401

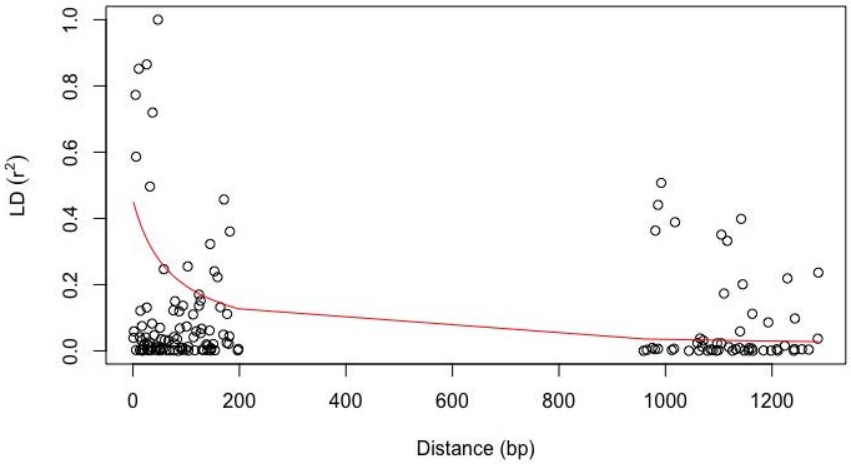
402

403

404

405

406 C)



407

408 Fig 4. (A) Decay of linkage disequilibrium (LD) across all the tagged genomic sequences, the

409 majority being exoms. Squared coefficients of allele frequency (r^2) are plotted against

410 distance in base pairs. The fitted curve (red) is representative of the trend of decay from the
411 178101 SNPs utilised in the association mapping (AM). (B) Decay of LD with distance in
412 base pairs between sites from across 41 contigs with significant associations. (C) Decay of
413 LD across contig MA_96191 that has a significant association for ratio of percentage
414 earlywood vs latewood on which two probes were captured.
415
416
417

418

419 Table 2 Phenotypes, Latent Traits, SNP, SNP feature, frequency and PVE

Phenotype	Latent Trait	QTL	*SNP	Allele	SNP Feature	Frequency	PVE
WD	Intercept	167610	MA_10435406_13733	A/G	Downstream variant	0.71	4.64%
	Slope	30469	MA_33109_11804	A/G	Upstream variant	0.72	4.50%
	K1	30469	MA_33109_11804	A/G	Upstream variant	0.551	4.15%
	K2	157442	MA_10432646_63090	G/A	Upstream variant	0.567	2.43%
EWD	Intercept	167610	MA_10435406_13733	A/G	Downstream variant	0.545	3.38%
	Slope	23798	MA_20321_44812	C/T	Upstream variant	0.53	0.69%
		70955	MA_118446_4316	T/A	Upstream variant	0.644	0.40%
TWD	Slope	131698	MA_10235390_3386	G/A	Stop gained	0.672	1.58%
		160208	MA_10433411_3386	T/C	Intron variant	0.595	3.41%
	K1	89044	MA_212523_6278	T/C	Upstream variant	0.534	3.34%
LWD	Slope	43797	MA_62987_13474	T/C	Missense variant	0.524	1.81%
		165481	MA_10434805_21408	C/T	Intron variant	0.588	1.21%

		171223	MA_10436058_4902	G/A	Intron variant	0.712	4.03%
RW	Intercept	11535	MA_10694_9101	A/C	Synonymous variant	0.545	1.95%
		112391	MA_879270_7373	C/T,A	Stop gained	0.532	1.45%
		112394	MA_879384_3894	C/A	Splice region variant	0.692	2.56%
	Slope	165481	MA_10434805_21408	C/T	Intron variant	0.521	2.66%
	K1	23808	MA_20322_28351	T/G	Synonymous variant	0.554	1.78%
		165481	MA_10434805_21408	C/T	Intron variant	0.533	0.18%
	K2	23808	MA_20322_28351	T/G	Synonymous variant	0.55	1.20%
		165481	MA_10434805_21408	C/T	Intron variant	0.615	1.79%
TRW	Slope	111057	MA_817099_1105	T/A	Missense variant	0.685	1.12%
	K1	33110	MA_38472_13803	T/A	Upstream gene variant	0.657	3.23%
		89295	MA_214776_1624	G/A	Upstream gene variant	0.688	4.51%
	K2	111057	MA_817099_1105	T/A	Missense variant	0.672	1.20%
LRW	Intercept	143628	MA_10428744_29330	C/T	Downstream variant	0.668	0.5%
	K2	164772	MA_10434624_20686	C/A	Downstream variant	0.571	0.06%

MOE	Slope	165481	MA_10434805_21408	C/T	Intron variant	0.602	1.00%
NC	K1	145839	MA_10429444_12692	G/C	Upstream variant	0.645	3.82%
ENC	Slope	98508	MA_402880_2045	A/C	Upstream variant	0.667	0.03%
		167610	MA_10435406_13733	A/G	Downstream variant	0.685	0.01%
TNC	Intercept	95870	MA_346723_2241	T/C	Upstream variant	0.667	3.78%
		126785	MA_9447489_687	A/C	Upstream gene variant	0.68	4.93%
LNC	Intercept	143628	MA_10428744_29330	C/T	Downstream variant	0.66	3.14%
	Slope	143628	MA_10428744_29330	C/T	Downstream variant	0.672	4.77%
EP	Intercept	16868	MA_15729_40331	G/T	Intron variant	0.609	3.32%
		91242	MA_246125_1213	G/A	Synonymous variant	0.594	3.41%
TP	Intercept	101203	MA_462319_4322	A/C	Upstream gene variant	0.594	1.16%
		132014	MA_10251995_2442	A/C	Upstream gene variant	0.601	3.22%
LP	K1	162397	MA_10434007_77578	C/T	Upstream gene variant	0.892	1.14%
EP/LP	Intercept	51657	MA_80954_29644	G/A	Downstream variant	0.63	0.81%
		60787	MA_98424_947	C/T	Intron variant	0.655	1.80%
		123639	MA_8790100_1384	A/C	Upstream variant	0.628	0.75%

	K1	59480/36496	MA_96191_7122	A/G	Synonymous	0.6	2.37%
	K2	117333	MA_1045136_4310	T/C	Missense variant	0.523	1.34%
	K3	72414	MA_122136_11653	A/T	Non-Coding	0.617	4.05%
Mass	Intercept	166235	MA_10435002_4986	G/A	Intergenic variant	0.533	0.65%
Index	Slope	61096	MA_99004_17108	G/A	Synonymous variant	0.66	0.01%
(Growth x		67181	MA_109804_10278	G/A	Missense variant	0.612	0.05%
Density)		1401	MA_1378_4718	C/A	Exon/stop gained	0.588	1.19%
		138744	MA_10427214_13968	G/T	Missense variant	0.58	1.80%
		162397	MA_10434007_77578	C/T	Upstream variant	0.627	1.44%
	K1	21924	MA_19222_1789	A/G	Upstream variant	0.71	1.82%

*SNP: The SNP name was composed of the contig (MA_number) and SNP position on contig. For example, the first SNP MA_1043540_13733 was located on contig MA_1043540 at position 13733 bp.

Several QTLs shared *within* each trait and *across* traits were observed in the analysis. WD, RW, TRW and LNC had one (30469), two (165481 and 23808), one (111057) and one (143628) QTL shared by two latent traits, respectively. One of the common QTL (30469) for WD had a frequency of 0.72 with an H^2_{QTL} of 4.50% for the slope trait, which indicates that it is highly significant for the phenotypes. Common QTLs *within* RW were observed for slope, K1 and K2 latent traits, with moderate frequencies ranging from 0.521 to 0.615 and influenced their respective traits to modest degree (H^2_{QTL} in ranges of 0.18-2.66%).

For QTLs common *across* the different latent traits, QTL 165481 was shared between LWD, RW and MOE; this is not surprising because of the close correlation between MOE and wood density, which in turn generally show negative correlation to RW. Intron variant MA_10434805g0010_165481 explained between 0.18-2.66% of the H^2_{QTL} observed in the respective traits. The SNP associated with this QTL also had high frequencies of 0.602 and 0.615 in MOE and RW explaining H^2_{QTL} of 1.00 and 2.66%, respectively. It was also observed that the SNP MA_10434805g0010_165481 is a common QTL *within* the wood traits related to Width (Table 2). SNP MA_10435406g0010_167610 was shared between WD, EWD and ENC. This SNP was characterized by having high frequencies in WD (0.71) and ENC (0.685), however it had a moderate frequency of 0.545 for EWD. This QTL was detected by the intercept latent trait for WD and EWD, and the slope latent trait in ENC (Table 2), with H^2_{QTL} ranging from 0.01-4.64%. The QTL had a high influence on the density related traits as it explained 4.64% (WD) and 3.38% (EWD).

Trees showing a positive correlation between growth and density had seven QTL specific for this observed phenomenon (MI) and had modest influence on the trait (H^2_{QTL} in the ranges 0.05-1.82%). Five of the QTL were detected using the slope as the latent trait with high frequencies ranging between 0.58 to 0.66 for SNP 61096 (Table 2).

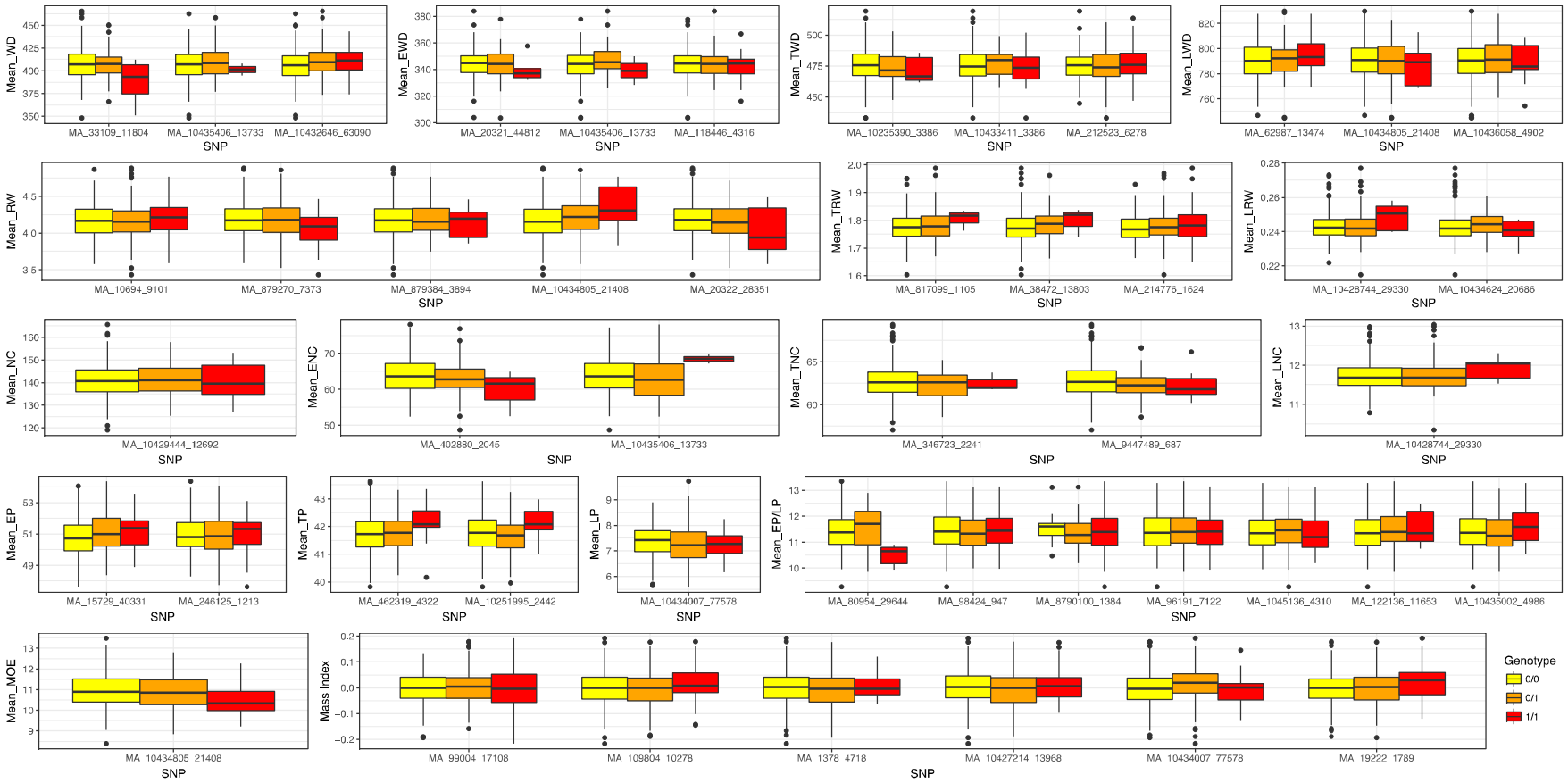


Fig 5. Box plot of the estimated genotypic effect on the phenotypes in the study. The significant SNPs associated and each one of the traits have been correlated to give the impact each genotype has on the average of the overall trait.

Genetic association with phenotypes

Sequence capture and the SNP-trait associations allowed the mining of candidate genes involved in spruce wood formation coupled with the identification of orthologous annotations and descriptions from *Populus* and *Arabidopsis*. This also allowed for the anchoring of significant markers on to the genetic linkage map for Norway spruce Fig 5.

RW, TRW, and LRW were associated with nine gene models. For RW five genes, endoglucanase 11-like, Alpha-dioxygenase 1 (DIOX1), Proliferating cell nuclear antigen (PCNA), B3-DNA-binding and E3 ubiquitin-protein ligase were identified. The SNP MA_879270g0010_112391, a splice region variant, explained 2.56% of the H^2_{QTL} and is associated with DIOX1. Marker MA_20322g0010_23808 for RW is associated with the protein domain for a plant specific B3-DNA binding protein, explaining 1.78% variation, with similar orthologous genes in *Arabidposis* and *Populus* (Table S1). The three putative genes associated with TRW are a Serine/threonine-protein kinase, a Homeodomain protein (HB2) and a Senescence-associated protein, and all have high H^2_{QTL} ranging from 2.14 to 4.50%. Contig MA_10434624 is homologos to a Pectin esterase and was associated with the downstream variant MA_10434624g0010_164772 for LRW. This may suggest a link between LRW and pectin modification. QTL associated with gene MA_214776g0010 for the TRW may be linked with serine/threonine-protein kinase gene (Os01g0689900), this occurrence of kinase-like related genes was also observed across TRW, NC, EP, EP/LP and EWD (Table S1).

NC, ENC, TNC and LNC are associated with a total of three putative genes and three protein domains. Of the three putative genes, two are associated with serine/kinase activity and one is involved in cysteine and methionine synthesis (Table 1 S1). All the SNPs

476 associated with these traits were either downstream or upstream of coding regions and may
477 thus act as modifiers of gene expression. The SNP MA_402880g0010_98508 (an upstream
478 gene variant) significantly associated with ENC located on gene MA_402880g0010 is
479 homologous to a *Populus* sphingolipid biosynthesis protein. SNP
480 MA_9447489g0010_126785 associated with TNC was located in the gene
481 MA_9447489g0010 which is homologous to a peptidase domain from *Arabidopsis* and
482 showed the highest H^2_{QTL} in the dataset (4.93%). This domain is similar to an orthologous
483 zinc carboxypeptidase enzyme of *Oryza sativa* (Zn-dependent exopeptidases superfamily
484 protein) (Table S1).

485 Wood percentage traits, EP, LP, TP and the ratio of EP/LP had significant associations
486 with ten SNPs. Four of the six significant SNP variants for EP/LP are modifiers with the other
487 two SNPs, being a synonymous (MA_96191g0010_59480) and missense
488 (MA_1045136g0010_117333) variant. The synonymous SNP MA_96191g0010_59480 was
489 associated with the gene model MA_96191g0010, which is homologous to a *P. sitchensis*
490 Glycosyltransferase (GT), similar to UDP-glucosyltransferase 73B2 (AT4G34135) from
491 *Arabidopsis*. Five protein domains were also detected, that were linked to phytochrome
492 kinase substrate 1, TIR/NBS/LRR and zein binding domains (Table S1). The significant SNP
493 MA_15729g0010_16868, an intron variant, that is associated with EP, is located in the gene
494 MA_15729g0010, which is homologous to a DNA-3-methyladenine glycosylase II enzyme.
495 The SNPs identified for TP and LP are all downstream gene variants (Table 2).

496 WD, EWD, TWD and LWD had a total of 12 significant associations. For the
497 associations with WD we identified the SNP MA_10435406g0010_167610 that is a 3′-gene
498 variant which explained the highest H^2_{QTL} observed (4.64%) and is located in a gene that is
499 homologous to a Phosphoadenosine phosphosulfate reductase gene cysH_2. This locus was
500 also detected for EWD and ENC explaining H^2_{QTL} of 3.38% and 0.01%, respectively. A

missense SNP, MA_33109g0010_30469, was associated with WD and located within the gene MA_33109g0010 homologous to an *Arabidopsis* senescence associated gene 24 (Table S1). The three significant SNPs identified for EWD were all modifiers, upstream and downstream gene variants. Of the three significant SNP associations for TWD, two, SNP MA_10235390_131698 (stop gained) and SNP MA_212523g0010_89044 (upstream gene variant), were identified within genes. The intron variant MA_10433411g0010_160208 associated with TWD and is found in the gene MA_10433411g0010 that is homologous to an *Arabidopsis* Transducin/WD40 repeat-like superfamily protein. Two of the three significant SNPs identified for LWD were intron variants (MA_10434805g0010_165481 and MA_10436058g0010_171223) with the third being a missense variant (MA_62987g0010_43797). The SNP MA_10434805g0010_165481 was found in the gene MA_10434805g0010, which is homologous to an *Arabidopsis* Proliferating Cell Nuclear Antigen Protein (PCNA). This SNP is also associated with RW and explained 1.21% and 2.66% $H^2_{QTL,}$ respectively.

The Mass Index trait, that is linked to a positive effect of wood volume growth and increased density (growth x density) yielded a total of seven associated SNPs, with two upstream gene variants, two missense variants one intergenic variant, one stop gained variant and one synonymous nucleotide replacement (Table 2). The slope latent trait had five genes with modest influence on the phenotype ranging from 0.01-1.80%. The genes were homologous to *Arabidopsis* GRAS transcription factor, Aluminium induced protein, Protein virilizer, ARM repeat superfamily protein and an uncharacterized protein. The SNP MA_1378g0010_1401 encodes for a premature stop codon (stop gained, a high impact variant) on gene MA_1378g0010, which is homologous to an *Arabidopsis* protein virilizer involved in mRNA splicing regulation. The gene homologous to the GRAS transcription factor was associated with the SNP MA_99004g0010_61096 (a synonymous variant). The

526 SNP MA_19222g0010_21921, an upstream gene variant, was located in the gene
527 MA_19222g0010 which is homologous to a *Picea sitchensis* ADP (NB-ARC domain) and
528 explained the highest H^2_{QTL} of 1.82% (Table S1).

529 Wood density traits were associated with a total of 12 genes, the largest number of
530 genes identified from the contigs. Percentage of wood was linked to ten putative genes, cell
531 width had nine putative genes and number of cells was associated with six genes. Two genes
532 were shared *across* multiple traits, PCNA was common *across* RW and LWD, and
533 phophoadenosine phosphosulfate reductase was shared *across* WD, EWD and ENC. Genes
534 with the Serine/threonine-protein phosphatase and TIR-NBS-LRR domains were also
535 identified *across* width, wood density and cell percentage traits.

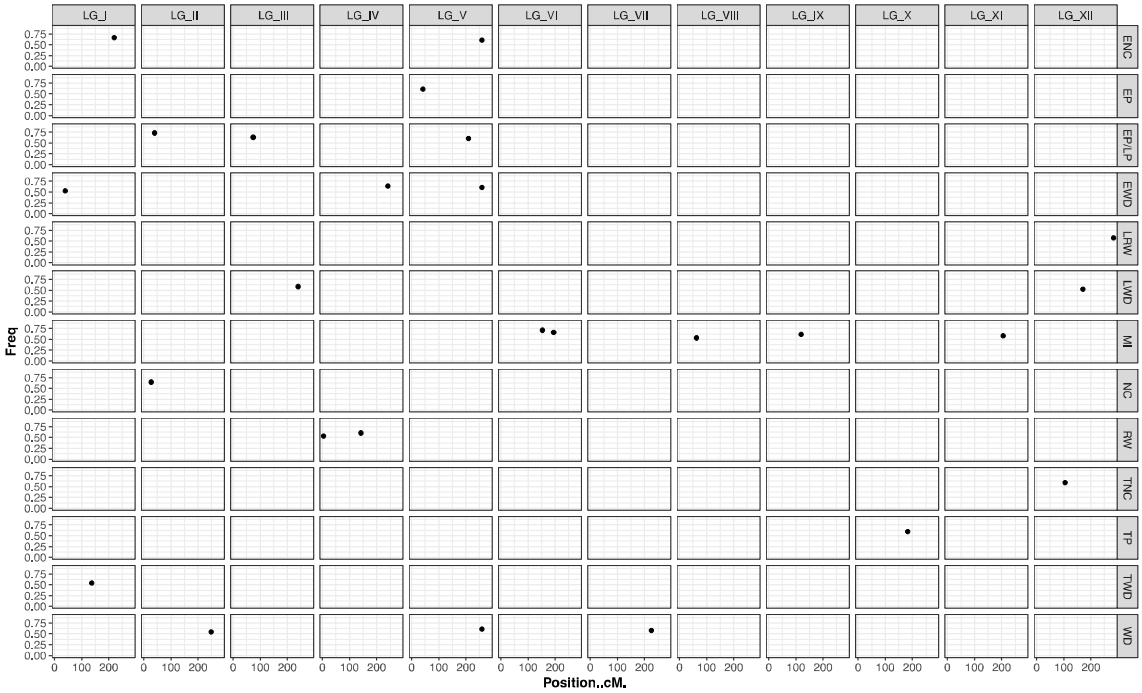


Fig 6. Frequencies of the significant markers selected using the multi-locus LASSO model for whole ring, earlywood and latewood associated with contigs plotted against their locations on a genetic linkage map derived from similar sequence captured probes. Significant associations for the traits were identified on the twelve linkage groups (LG) as follows: [**LG_I**: EWD, TWD and ENC], [**LG_II**: NC, EP/LP and WD], [**LG_III**: EP/LP and RW], [**LG_IV**: RW and EWD], [**LG_V**: EP, EP/LP, ENC, EWD and WD], [**LG**_VI: MI], [**LG_VII**: WD], [**LG_VIII**: MI], [**LG_IX**: MI], [**LG_X**: TP], [**LG_XI**: MI] and [**LG_XII**: TNC, LWD and LRW].

Discussion

We applied a functional mapping approach in a genome-wide association mapping context and identified 51 significant QTLs that were associated with wood formation in Norway spruce. Previous work utilizing a functional mapping analysis in forest trees have used a limited number of molecular markers (Li *et al.*, 2014; Ma *et al.*, 2002). Li *et al.*, applied this

552 analysis in a bi-parental Scots pine (*P. sylvestris* L.) cross using 319 markers. Hence, our
553 work represents a major advance in that we have been able to apply this approach at a
554 genome-wide scale (178101 SNPs) on unrelated mother trees, with a dynamic functional trait
555 dataset comprising 15-time points/annual growth rings (*i.e.*, cambial age). Latent traits
556 represent significant time points in the trait development allowing us to detect putative genes
557 at these critical junctures in wood formation.

558 The number of detected QTLs is relatively small compared with several recent studies
559 in *Populus* (Evans *et al.*, 2014; McKown *et al.*, 2014; Porth *et al.*, 2013). The sample size,
560 number of SNPs used and the stringency with which we accepted significant SNPs
561 contributed to the modest number of QTL. Previous functional mapping studies, (Li *et al.*,
562 2014) involving SNPs in conifers have used two levels of evaluating QTLs, whereby they
563 have suggestive and significant QTL. In our study, we only reported significant QTL. As
564 indicated in Hall *et al* (2016), there should be hundreds to thousands of QTL of moderate to
565 very small effect related to growth and wood quality traits in trees. Hence, a large population
566 and accurate phenotyping are required for a reliable identification of most QTLs (Korte &
567 Farlow, 2013). However, the sample size of our study allowed the detection of the
568 largest/most significant QTL. The study identified significant associations explaining
569 relatively small proportions of the phenotypic variance being observed, ranging from 0.01-
570 4.93%. This is in line with other studies of QTL for wood traits (González-Martínez *et al.*,
571 2007; Porth *et al.*, 2013).

572

573 *Genetic associations with potential to improve wood properties*

574 With all the SNPs, having been derived from known genomic positions, it was
575 possible to identify genes linked to the associated QTLs and infer their potential function in
576 wood formation.

577 The gene MA_10694g0010 is homologous to an enzyme involved in cell wall
578 biosynthesis, endoglucanase 11-like, and was associated with RW (intercept latent) (Table
579 S1). The association of this gene with the RW intercept implies that the gene influences the
580 trait throughout the growth period. This enzyme is a vital component of xylogenesis and is
581 involved in the active digestion of the primary cell wall (Goulao *et al.*, 2011). The
582 endoglucanase 11-like, was associated with a synonymous SNP MA_10694g0010_11535 for
583 (RW) suggesting an involvement in cell expansion and cell wall loosening during wood
584 formation. Endoglucanases have been proposed as enzymes involved in controlling cell wall
585 loosening (Cosgrove, 2005). Endoglucanase 11-like gene is part of the endo-1 family in
586 which the eno-1-4-β-glucanase *Korrigan* gene belongs. Characterisation of the *Korrigan* gene
587 in *P. glauca* has identified it to be functionally conserved and essential for cellulose synthesis
588 (Maloney *et al.*, 2012). Hence, MA_10694g0010 is a candidate for the remodelling of cell
589 walls that affects the mechanical and growth properties of wood cells, and consequently
590 annual ring width.

591 The synonymous SNP MA_20322g0010_23808 is associated with RW located on
592 gene MA_20322g0010 which is homologous with a plant specific B3-DNA binding protein
593 domain, that is shared among various plant-specific transcription factors. This includes
594 transcription factors involved in auxin and abscisic acid responsive transcription (Yamasaki *et*
595 *al.*, 2004). Auxin is one of the central phytohormones in the control of plant growth and
596 development (Abel & Theologis, 1996), and also known to be involved in cell wall loosening
597 and elongation (Cosgrove, 2016). This suggests a possible functional role for
598 MA_20322g0010 in influencing RW.

599 An intron variant located in the MA_10434805g0020 gene, which is homologous to
600 PCNA was detected *across* several phenotypes (LWD, RW and MOE) associated with the
601 slope latent trait (Table 2). The detection of this gene *across* these phenotypes using the slope

latent trait implies that the gene affects the rate of change of these phenotypes. Thus, this would be a good gene to target for further studies. PCNA proteins function as integral enzymes in the regulatory pathways of cell cycle regulation and DNA metabolism (Maga & Hübscher, 2003). PCNA has been associated with chromatin remodelling, DNA repair, sister-chromatid cohesion and cell cycle control, which are all vital processes in plant growth (Strzalka & Ziemienowicz, 2010), but it has not been previously associated with wood formation traits.

In our study we detected a significant downstream SNP (MA_10434624g0010_164772) associated with LRW on gene MA_10434624g0020, homologous to pectinmethylesterases (PMEs), which are cell wall associated enzymes responsible for demethylation of polygalacturonans (Phan *et al.*, 2007). This enzyme has been shown to be linked with many developmental processes in plants, such as, cellular adhesion and stem elongation (Micheli, 2001). An association study in White spruce identified a significant nonsynonymous SNP coding for cysteine associated with earlywood and total wood cell wall thickness associated with pectinmethylesterase (Beaulieu *et al.*, 2011). Our study identified a PME SNP association in the latewood stage, supporting the importance of PMEs in wood cell development.

A SNP (MA_10435406g0010_167910) downstream on gene MA_10435406g0010 was detected *across* the traits ENC, WD and EWD. The association of this gene with the WD and EWD intercept implies that it an impact on the overall development of density throughout the growth period. Since density is correlated with number of cells, this association with the slope latent trait of ENC means the gene influences its rate of change. The gene is homologous to Phosphoadenosine phosphosulfate reductase (PAPS), which plays a central role in the reduction of sulphur in plants. An analysis of PAPS enzymes in *Arabidopsis* (Klein & Papenbrock, 2004) and *Populus* (Kopriva *et al.*, 2004) revealed that enzymes involved in

627 sulphate-conjugation, play an important role in plant growth and development (Klein &
628 Papenbrock, 2004). Reduced sulphur is utilized by the sulphate assimilation pathway for the
629 synthesis of essential amino acids cysteine and methionine (Kopriva & Koprivova, 2004).
630 Methione acts as a methyl donor in both lignin, hemicellulose and pectin biosynthesis
631 providing a possible mechanism of how PAPS could influence wood density and number of
632 cells.

633 When analyzing QTLs detected for traits linked to the percentage of cells (EP, LP and
634 EP/LP) we identified three putative candidate genes, DNA-3-methyladenine glycosylase II
635 enzyme, phytochrome kinase substrate 1 and glycosyltransferase. DNA-3-methyladenine
636 glycosylase II enzyme is responsible for carrying out base excision repairs (BER) in the
637 genome in order to maintain genomic integrity. This enzyme has the ability to initiate a broad
638 substrate recognition and provides a wide resistance to DNA damaging agents (Wyatt *et al.*,
639 1999). This DNA repair capacity can be expected to be essential for the process of cell
640 propagation and growth.

641 A synonymous SNP (MA_96191_59480) within the gene MA_96191g0010, which is
642 homologous to Glucosyltransferase in *P. sitchensis* was associated with EP/LP. Glycosyl
643 transferases operate by facilitating the catalytic sequential transfer of sugars from activated
644 donors to acceptor molecules that form region and stereospecific glycosidic linkages (Lairson
645 *et al.*, 2008). The Arabidopsis ortholog (UDP-glucosyltransferase 73B2) encodes for a
646 putative flavonol 7-O-glucosyltransferase involved in stress responses. In our study, this
647 significant association was associated with EP/LP, however a nonsynonymous variant in a
648 gene coding for a Glycosyl transferase in *Populus* was associated with fibre development and
649 elongation (Porth *et al.*, 2013). Therefore, gene MA_96191g0010 is a novel candidate for
650 further investigation of how flavonol metabolism may influence the proportion of early and
651 late wood in Norway spruce.

652 Two genes concerning wood formation, PAPS and PCNA, were also detected across
653 related traits density, growth number of cells and MOE. Significant SNP
654 (MA_10435406_167610) in the PAPS reductase gene is common *across* ENC, WD and
655 EWD, with SNP MA_10434805_165481 located in an intron for a gene encoding for PCNA
656 protein being detected *across* WD, RW and MOE (Table S1). The presence of these common
657 QTL suggests that these traits might be under the control of the same genes or genetic
658 pathways. Chen et al (2014) reported a significant positive genetic correlation between wood
659 density and MOE, which increased with tree age. However, wood volume growth and density
660 have a negative correlation (Chen *et al.*, 2014), with our study being the first to detect QTLs
661 for trees exhibiting a positive correlation for this phenomenon (MI). The common QTL
662 observed *across* WD, EWD and ENC indicates that the number of cells during the juvenile
663 wood development stages has a significant impact on the overall density. The seasonal
664 changes in EWD to LWD has been speculated to be due to a change in auxin levels leading to
665 the initiation of wall-thickening phase, which has a direct impact on the wood quality traits
666 such as MOE. This phase coincides with the cessation of height growth and where available
667 resources are used for cell-wall thickening (Sewell *et al.*, 2000), which may explain the
668 common QTL between LWD, RW and MOE, as part of the same feedback loop mechanism.
669 We identified two associations to homologous genes related to nucleic acid repair
670 functions, DICER-LIKE3 (DCL3) and DNA mismatch repair protein (MSH5), which are
671 concerned with RNA processing as well as DNA repair, respectively. These genes are
672 involved in ensuring the fidelity of DNA replication and to preserve genomic integrity (Hsieh
673 & Yamane, 2008). These genes are possibly associated with cambial cell division and endo-
674 reduplication during wood formation and can conceivably have effects on wood density.
675 An association for TWD with a SNP located upstream of gene MA_212523g0010, is
676 homologous to Kinesin-related protein 13 (gene-L484_021891). Kinesin-related proteins are

677 known to be involved in secondary wall deposition, which can impact wood density (Zhong *et*
678 *al.*, 2002), cell wall strength and oriented deposition of cellulose microfibrils.

679 Several receptor-like Kinases (TIR/NBS/LRR and Serine/threonine-protein
680 phosphatase) homologs were identified *across* traits (TRW, NC, EP, EP/LP and EWD) (Table
681 S1). These protein domains control a large range of processes including hormone perception
682 and plant development. Approximately 2.5% of the annotated genes in *Arabidopsis* genome
683 are RLK homologs (Shiu & Bleecker, 2001), where they among other functions play an
684 important role in the differentiation and separation of xylem and phloem cells (Fisher &
685 Turner, 2007). Similar to our study a synonymous SNP in a RLK gene was associated with
686 early wood proportion (EP) in White spruce (Beaulieu *et al.*, 2011), hence RLKs seem to be
687 involved in modifying a number of different wood properties from density to cell identity and
688 number.

689 Norway spruce trees that possess the ability of fast growth and high wood density are
690 very rare, but such trees and associated SNPs were discovered in our study. Trees combining
691 these traits are of high interest to forest industries and owners, and thus also in focus for
692 breeders. Of the seven genes significantly linked to this phenomenon of particular interest was
693 a synonymous SNP on MA_99004g0100 gene homologous to a transcription factor from the
694 GRAS family (Table S1). GRAS is an important class of plant-specific proteins derived from
695 three members: GIBBERELLIC-ACID INSENSITIVE (GAI), REPRESSOR of GAI (RGA)
696 and SCARECROW (SCR) (GRAS) (Hirsch & Oldroyd, 2009). GRAS genes are known to be
697 involved in the regulation of plant development through the regulation of gibberellic acid
698 (GA) and light signalling (Cenci & Rouard, 2017; Hirsch & Oldroyd, 2009). Furthermore GA
699 signalling has also been shown to stimulate wood formation in *Populus* (Mauriat & Moritz,
700 2009). Thus, the GRAS transcription factor identified here and the other six genes positively

associated with MI provide interesting genetic markers and tools to understand this phenomenon.

Conclusion

This work has dissected the genetic basis of wood properties in Norway spruce with use of functional association mapping. In total, we identified 51 Significant QTLs for wood properties and mining of candidate genes located in the vicinity of significant QTLs identified genes that could be directly or indirectly responsible for variations in the observed traits. Significant novelty in our results is provided by the identification of QTLs associated to both high wood density and fast growth, thus larger biomass. These genes are candidates for further functional verification in Norway spruce.

Acknowledgements

We acknowledge the support of the Bio4Energy research organization for wood property analyses and evaluations. All genetic data was obtained through funding from the Knut and Alice Wallenberg foundation. SSF project support of continuing this work. JB is supported though a postdoc position funded by the Kempe foundation.

References

Abel, S., & Theologis, A. (1996). Early Genes and Auxin Action. *Plant Physiology, 111*(1), 9.

Beaulieu, J., Doerksen, T., Boyle, B., Clément, S., Deslauriers, M., Beauseigle, S., Blais, S., Poulin, P.-L., Lenz, P., Caron, S., Rigault, P., Bicho, P., Bousquet, J., & MacKay, J. (2011). Association Genetics of Wood Physical Traits in the Conifer White Spruce and Relationships With Gene Expression. *Genetics, 188*(1), 197-214. doi: 10.1534/genetics.110.125781

Beavis, W. D. (1998). QTL analyses: power, precision, and accuracy. *Molecular dissection of complex traits, 1998*, 145-162.

Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D. G., Shiffthaler, B., Baison, J., Street, N. R., Garcia Gil, M. R., & Ingvarsson, P. K. (2018). An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (Picea abies). *bioRxiv*. doi: 10.1101/292151

Bertaud, F., & Holmbom, B. (2004). Chemical composition of earlywood and latewood in Norway spruce heartwood, sapwood and transition zone wood. *Wood Science and Technology, 38*(4), 245-256. doi: 10.1007/s00226-004-0241-9

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.

Cenci, A., & Rouard, M. (2017). Evolutionary Analyses of GRAS Transcription Factors in Angiosperms. *Frontiers in Plant Science, 8*, 273. doi: 10.3389/fpls.2017.00273

Chen, Z.-Q., Gil, M. R. G., Karlsson, B., Lundqvist, S.-O., Olsson, L., & Wu, H. X. (2014). Inheritance of growth and solid wood quality traits in a large Norway spruce population tested at two locations in southern Sweden. *Tree Genetics & Genomes, 10*(5), 1291-1303.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92.

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nat Rev Mol Cell Biol, 6*(11), 850-861. doi: http://www.nature.com/nrm/journal/v6/n11/suppinfo/nrm1746_S1.html

Cosgrove, D. J. (2016). Catalysts of plant cell wall loosening. *F1000Research, 5*, F1000 Faculty Rev-1119. doi: 10.12688/f1000research.7180.1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Dutilleul, P., Herman, M., & Avella-Shaw, T. (1998). Growth rate effects on correlations among ring width, wood density, and mean tracheid length in Norway spruce (Picea abies). *Canadian Journal of Forest Research, 28*(1), 56-68.

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J.-G., Tuskan, G. A., & DiFazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat Genet, 46*(10), 1089-1096. doi: 10.1038/ng.3075

http://www.nature.com/ng/journal/v46/n10/abs/ng.3075.html#supplementary-information

Evans, R. (1994) Rapid Measurement of the Transverse Dimensions of Tracheids in Radial Wood Sections from Pinus Radiata, Holzforschung 48: 168–72.

767 Evans, R. (2006) Wood stiffness by X-ray diffractometry. In: Stokke DD, Groom HL (eds)
768 Characterization of the cellulosic cell wall. Wiley, Hoboken, pp. 138-146.
769 Fisher, K., & Turner, S. (2007). PXY, a receptor-like kinase essential for maintaining polarity
770 during plant vascular-tissue development. *Current Biology*, *17*(12), 1061-1066.
771 Gao, H., Wu, Y., Li, J., Li, H., Li, J., & Yang, R. (2014). Forward LASSO analysis for high-
772 order interactions in genome-wide association study. *Briefings in Bioinformatics*,
773 *15*(4), 552-561.
774 Gilmour, A., Gogel, B., Cullis, B., Welham, S., Thompson, R., Butler, D., Cherry, M.,
775 Collins, D., Dutkowski, G., & Harding, S. (2014). ASReml user guide. Release 4.1
776 structural specification. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*
777 *www. vsni. co. uk*.
778 González-Martínez, S. C., Wheeler, N. C., Ersoz, E., Nelson, C. D., & Neale, D. B. (2007).
779 Association genetics in Pinus taeda LI Wood property traits. *Genetics*, *175*(1), 399-
780 409.
781 Goulao, L. F., Vieira-Silva, S., & Jackson, P. A. (2011). Association of hemicellulose- and
782 pectin-modifying gene expression with Eucalyptus globulus secondary growth. *Plant*
783 *Physiology and Biochemistry*, *49*(8), 873-881. doi:
784 https://doi.org/10.1016/j.plaphy.2011.02.020
785 Hallingbäck, H. R., Sánchez, L., & Wu, H. X. (2014). Single versus subdivided population
786 strategies in breeding against an adverse genetic correlation. *Tree Genetics &*
787 *Genomes*, *10*(3), 605-617.
788 Hannrup, B., Cahalan, C., Chantre, G., Grabner, M., Karlsson, B., Bayon, I. L., Jones, G. L.,
789 Müller, U., Pereira, H., & Rodrigues, J. C. (2004). Genetic parameters of growth and
790 wood quality traits in Picea abies. *Scandinavian Journal of Forest Research*, *19*(1),
791 14-29.
792 Hauksson, J. B., Bergqvist, G., Bergsten, U., Sjöström, M., & Edlund, U. (2001). Prediction
793 of basic wood properties for Norway spruce. Interpretation of Near Infrared
794 Spectroscopy data using partial least squares regression. *Wood Science and*
795 *Technology*, *35*(6), 475-485. doi: 10.1007/s00226-001-0123-3
796 Heuven, H. C., & Janss, L. L. (2010). *Bayesian multi-QTL mapping for growth curve*
797 *parameters*. Paper presented at the BMC proceedings.
798 Hill, W., & Weir, B. (1988). Variances and covariances of squared linkage disequilibria in
799 finite populations. *Theoretical population biology*, *33*(1), 54-78.
800 Hirsch, C. D., Evans, J., Buell, C. R., & Hirsch, C. N. (2014). Reduced representation
801 approaches to interrogate genome diversity in large repetitive plant genomes.
802 *Briefings in Functional Genomics*, elt051.
803 Hirsch, S., & Oldroyd, G. E. D. (2009). GRAS-domain transcription factors that regulate
804 plant development. *Plant Signaling & Behavior*, *4*(8), 698-700.
805 Hsieh, P., & Yamane, K. (2008). DNA mismatch repair: molecular mechanism, cancer, and
806 ageing. *Mechanisms of ageing and development*, *129*(7), 391-407.
807 Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop
808 plants. *Annual review of plant biology*, *65*, 531-551.
809 Klein, M., & Papenbrock, J. (2004). The multi-protein family of Arabidopsis
810 sulphotransferases and their relatives in other plant species. *Journal of experimental*
811 *botany*, *55*(404), 1809-1820. doi: 10.1093/jxb/erh183
812 Kopriva, S., Hartmann, T., Massaro, G., Hönicke, P., & Rennenberg, H. (2004). Regulation of
813 sulfate assimilation by nitrogen and sulfur nutrition in poplar trees. *Trees*, *18*(3), 320-
814 326. doi: 10.1007/s00468-003-0309-4

Kopriva, S., & Koprivova, A. (2004). Plant adenosine 5′-phosphosulphate reductase: the past, the present, and the future. *Journal of experimental botany*, *55*(404), 1775-1783. doi: 10.1093/jxb/erh185

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, *9*(1), 1.

Lairson, L., Henrissat, B., Davies, G., & Withers, S. (2008). Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.*, *77*, 521-555.

Lande, R., & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, *124*(3), 743-756.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, Z., Hallingbäck, H. R., Abrahamsson, S., Fries, A., Gull, B. A., Sillanpää, M. J., & García-Gil, M. R. (2014). Functional multi-locus QTL mapping of temporal trends in Scots pine wood traits. *G3: Genes| Genomes| Genetics*, *4*(12), 2365-2379.

Li, Z., & Sillanpää, M. J. (2015). Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends in Plant Science*, *20*(12), 822-833. doi: http://dx.doi.org/10.1016/j.tplants.2015.08.012

Ma, C.-X., Casella, G., & Wu, R. (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics*, *161*(4), 1751-1762.

Maga, G., & Hübscher, U. (2003). Proliferating cell nuclear antigen (PCNA): a dancer with many partners. *Journal of Cell Science*, *116*(15), 3051.

Maloney, V. J., Samuels, A. L., & Mansfield, S. D. (2012). The endo-1, 4-β-glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytologist*, *193*(4), 1076-1087.

Mauriat, M., & Moritz, T. (2009). Analyses of GA20ox-and GID1-over-expressing aspen suggest that gibberellins play two distinct roles in wood formation. *The Plant Journal*, *58*(6), 989-1003.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., & Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.

McKown, A. D., Klápště, J., Guy, R. D., Geraldes, A., Porth, I., Hannemann, J., Friedmann, M., Muchero, W., Tuskan, G. A., & Ehlting, J. (2014). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. *New Phytologist*, *203*(2), 535-553.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417-473.

Micheli, F. (2001). Pectin methylesterases: cell wall enzymes with important roles in plant physiology. *Trends in Plant Science*, *6*(9), 414-419. doi: http://doi.org/10.1016/S1360-1385(01)02045-3

Moritsuka, E., Hisataka, Y., Tamura, M., Uchiyama, K., Watanabe, A., Tsumura, Y., & Tachida, H. (2012). Extended linkage disequilibrium in noncoding regions in a conifer, Cryptomeria japonica. *Genetics*, *190*(3), 1145-1148.

865 Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers.
866 *Trends in Plant Science, 9*(7), 325-330.
867 Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F.,
868 Delhomme, N., Giacomello, S., & Alexeyenko, A. (2013). The Norway spruce
869 genome sequence and conifer genome evolution. *Nature, 497*(7451), 579-584.
870 Olesen, P. (1977). The variation of the basic density level and tracheid width within the
871 juvenile and mature wood of Norway spruce. *For. Tree Improv, 12*, 1-22.
872 Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C.
873 (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine.
874 *Molecular Ecology, 21*(12), 2991-3005.
875 Phan, T. D., Bo, W., West, G., Lycett, G. W., & Tucker, G. A. (2007). Silencing of the Major
876 Salt-Dependent Isoform of Pectinesterase in Tomato Alters Fruit Softening. *Plant*
877 *Physiology, 144*(4), 1960-1967. doi: 10.1104/pp.107.096347
878 Porth, I., Klapšte, J., Skyba, O., Hannemann, J., McKown, A. D., Guy, R. D., DiFazio, S. P.,
879 Muchero, W., Ranjan, P., & Tuskan, G. A. (2013). Genome-wide association mapping
880 for wood characteristics in Populus identifies an array of candidate single nucleotide
881 polymorphisms. *New Phytologist, 200*(3), 710-726.
882 Resende, M. D., Resende, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar,
883 A. M., Abad, J. M., Takahashi, E. K., Rosado, A. M., & Faria, D. A. (2012). Genomic
884 selection for growth and wood quality in Eucalyptus: capturing the missing heritability
885 and accelerating breeding for complex traits in forest trees. *New Phytologist, 194*(1),
886 116-128.
887 Sewell, M., Bassoni, D., Megraw, R., Wheeler, N., & Neale, D. (2000). Identification of
888 QTLs influencing wood property traits in loblolly pine (Pinus taeda L.). I. Physical
889 wood properties. *TAG Theoretical and Applied Genetics, 101*(8), 1273-1281.
890 Shiu, S.-H., & Bleecker, A. B. (2001). Plant receptor-like kinase gene family: diversity,
891 function, and signaling. *Sci stke, 113*(113), re22.
892 Strauss, S., Lande, R., & Namkoong, G. (1992). Limitations of molecular-marker-aided
893 selection in forest tree breeding. *Canadian Journal of Forest Research, 22*(7), 1050-
894 1061.
895 Strzalka, W., & Ziemienowicz, A. (2010). Proliferating cell nuclear antigen (PCNA): a key
896 factor in DNA replication and cell cycle regulation. *Annals of botany, 107*(7), 1127-
897 1140.
898 Team, R. (2015). RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL*
899 *http://www. rstudio. com*.
900 Thavamanikumar, S., Southerton, S. G., Bossinger, G., & Thumma, B. R. (2013). Dissection
901 of complex traits in forest trees—opportunities for marker-assisted selection. *Tree*
902 *Genetics & Genomes, 9*(3), 627-639.
903 Thumma, B. R., Southerton, S. G., Bell, J. C., Owen, J. V., Henery, M. L., & Moran, G. F.
904 (2010). Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus
905 nitens. *Tree Genetics & Genomes, 6*(2), 305-317.
906 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*
907 *Statistical Society. Series B (Methodological)*, 267-288.
908 Vidalis, A., Scofield, D. G., Neves, L. G., Bernhardsson, C., García-Gil, M. R., & Ingvarsson,
909 P. (2018). Design and evaluation of a large sequence-capture probe set and associated
910 SNPs for diploid and haploid samples of Norway spruce (Picea abies). *bioRxiv*. doi:
911 10.1101/291716
912 Wyatt, M. D., Allan, J. M., Lau, A. Y., Ellenberger, T. E., & Samson, L. D. (1999). 3-
913 methyladenine DNA glycosylases: structure, function, and biological importance.
914 *Bioessays, 21*(8), 668-676.

Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Tomo, Y., Hayami, N., Terada, T., Shirouzu, M., Osanai, T., Tanaka, A., Seki, M., Shinozaki, K., & Yokoyama, S. (2004). Solution Structure of the B3 DNA Binding Domain of the Arabidopsis Cold-Responsive Transcription Factor RAV1. *The Plant Cell, 16*(12), 3448.

Zhong, R., Burk, D. H., Morrison, W. H., & Ye, Z.-H. (2002). A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. *The Plant Cell, 14*(12), 3101-3117.