QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data

Tree Genetics & Genomes (2017) 13:33
DOI 10.1007/s11295-017-1118-z

ORIGINAL ARTICLE

QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data

P. Fuentes-Utrilla[1] · C. Goswami[1] · J. E. Cottrell[2] · R. Pong-Wong[1] · A. Law[1] · S. W. A'Hara[2] · S. J. Lee[2] · J. A. Woolliams[1]

Received: 12 May 2016 / Revised: 19 December 2016 / Accepted: 24 December 2016

Abstract Sitka spruce *(Picea sitchensis* (Bong.) Carr) is the most common commercial plantation species in Britain and a breeding programme based on traditional lines has been in operation since the early 1960s. Rotation lengths of 40-years have led breeders to adopt a process of indirect selection at younger ages based on traits well correlated with final selection, but still the generation interval is unlikely to reduce much below twenty years. Recent successful developments with genomic selection in animal breeding have led tree breeders to consider the application of this technology. In this study a RAD sequence assay was developed as a means of investigating the potential of molecular breeding in a non-model species. DNA was extracted from nearly 500 clonally replicated trees growing in a single full-sibling family at one site in Britain. The technique proved successful in identifying 132 QTLs for 5-year bud-burst and 2 QTLs for 6-year height. In addition, the accuracy of predicting phenotypes by genomic selection was strikingly high at 0.62 and 0.59 respectively. Sensitivity analysis with 200 offspring found only a slight fall in correlation values (0.54 and 0.38) although when the training population reduced to 50 offspring predictive values fell further (0.33 and 0.25). This proved an encouraging first investigation into the potential use of genomic selection in the breeding of Sitka spruce. The authors investigate how problems associated with effective population size and linkage disequilibrium can be avoided and suggest a practical way of incorporating genomic selection into a dynamic breeding programme.

Keywords Sitka spruce · genome selection · RADseq · molecular breeding · height · bud-burst

Communicated by D. Grattapaglia

Electronic supplementary material The online version of this article (doi:10.1007/s11295-017-1118-z) contains supplementary material, which is available to authorized users.

✉ S. J. Lee
steve.lee@forestry.gsi.gov.uk

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, EH25 9RG Midlothian, UK

[2] Forest Research, Northern Research Station, Roslin, EH25 9SY Midlothian, UK

Introduction

Sitka spruce (*Picea sitchensis* (Bong.) Carr) is native to a narrow range of coastline stretching nearly 3,000 km along the seaboard of the Pacific North West from mid-Alaska to northern California. The species plays an important role in plantation forestry in northern Europe (Hermann 1987), and is currently the most widely planted conifer in Great Britain and Ireland, where it occupies over one million hectares of land. It also makes a commercial contribution to forestry in Denmark, France and more recently, Sweden (Lee et al. 2013). Both within and beyond its native range, it is mainly used for construction timber and wood pulp (Bousquet et al. 2007). Great Britain has an active Sitka spruce breeding programme in which the main objective has been to increase the end-of-rotation value to the construction grade timber industry by selecting parents combining good growth rate, with improved stem straightness, branching qualities and wood stiffness (Lee and Connolly 2010).

Final selection goals for Sitka spruce in genetic trials are increases in final rotation volume and the proportion of quality construction grade timber. In an attempt to accelerate the selection process in genetic trials, breeders have adopted indirect selection. This involves selection at a young age on the basis

Published online: 04 February 2017

of traits which are well correlated with the final selection goals. For example 6-year height in Sitka spruce is a surrogate for final rotation volume, and pin penetration of a Pilodyn gun at 12-15 years correlates well with mid-rotation whole-tree wood density which is a good indicator of timber strength (Lee et al. 2002a, b). Indirect selection has met with some success since the start of the programme in 1963 and has allowed good progress to be made in the re-selection of superior parent trees based on early progeny test data (backward selection) to construct the first generation breeding population. In some species, early indirect selections in progeny trials (forward selection) along with development of techniques such as grafting scions from those selections onto the upper crown areas of established, mature trees (known as top-grafting; Goading et al, 1999) and chemical treatment of subsequent grafts have advanced the age of flowering which further reduces the generation interval

Tree breeders are always looking for ways to reduce operational costs and generation intervals. Molecular tools offer a potential solution to reduce the cost and time required to complete these selection cycles and during the last decade there has been considerable interest and some notable progress in their development for forest tree species. In addition to reducing the length of the breeding cycle molecular approaches may provide the opportunity to increase selection intensity and reduce field testing effort (Grattapaglia, 2014). Early attempts to use genetic markers involved in association studies with phenotypic traits did not fulfil their promise in forest trees either when targeting candidate genes or the development of dense SNP panels in genome wide association studies (GWAS; Beaulieu et al. 2011). This is because (i) these approaches only explained a small percentage of the variation in the traits under investigation and (ii) associations identified did not transfer well across populations or environments (Pelgas et al. 2011; Ritland et al. 2011). This experience reflects those found in much larger genome-wide association studies involving domestic animals where exploitation of single markers has occurred (Houston et al. 2008) but is an exception (Meuwissen et al. 2016). For these reasons tree breeders found the GWAS techniques of little practical use although they did help to identify QTL, and causative variance which remains of interest to the scientific community.

Recently, emphasis has shifted towards the concept of genomic selection (GS) first proposed by Meuwissen et al. (2001) for use in animal breeding. GS techniques do not set out to validate markers associated with causative variants, but instead use all SNP markers simultaneously to maximise the accuracy of an estimated breeeding value. GS uses a 'training' population which is both genotyped using a large number of markers and phenotyped for the traits of interest. These data are then used to create a prediction model (G-BLUP) based on the construction of genomic relationships among individuals in the population, which can then be used to predict breeding values of individuals for which there is only SNP genotypic information. The benefit of this approach is that once a panel of SNP markers has been obtained, GS can be used for any trait i.e. it does not involve trait specific markers such as those employed in GWAS.

The GS approach is attractive as it has the potential to improve selection accuracy and facilitate greater selection intensity, whilst reducing the generation interval substantially (Grattapaglia and Resende 2011). A study of GS by Beaulieu et al. (2014) showed training sets of less than 2,000 individuals could provide prediction accuracies comparable to traditional field-based evaluations for open-pollinated white spruce (*P. glauca)* families.

A prerequisite for both GWAS and GS approaches is the availability of large numbers of SNP markers. For livestock, extensive international sequencing initiatives have facilitated large-scale SNP discovery, and the availability of these markers has enabled the development of a range of SNP panels ranging in size from 60k in pigs, to in excess of 700k for cattle and sheep (Van Raden et al. 2013). Such panels have also been developed in crop plants and a 60k SNP array is currently available for rice and maize (Gupta et al. 2008). In contrast, conifer sequencing has been challenging due to the large size and highly repetitive nature of their genomes. The recently published draft genomes of white spruce (Birol et al. 2013) and Norway spruce (*Picea abies;* Nystedt et al. 2013) are each estimated to be ~20Gb long compared to ~3Gb for the human genome (Venter et al. 2001), and 485Mb for poplar (Tuskan et al. 2006). There has been only limited sequencing effort for Sitka spruce and this has hampered the development of the necessary genomic tools for implementing GS.

In a species where an assembled genome sequence is not yet available, reduced representation libraries such as those employed in RAD sequencing (Restriction-site Associated DNA Sequencing or RADseq; Davey et al. 2011) offer a relatively cheap alternative method for identifying the large numbers of SNPs necessary for both GWAS and GS approaches (Andrews et al. 2016). To date RADseq has been applied to tree species such as Eucalyptus and Norway spruce as well as perennial plants such as grass (Grattapaglia et al. 2011; Slavov et al. 2014).

Even with appropriately designed SNP libraries of sufficient size to cope with large conifer genomes there are further challenges for implementation. Neale and Savolainen (2004) suggest that due to their relatively large effective population size (*Ne*), linkage disequilibrium (LD) between loci in some conifer species will only extend over relatively short distances compared to domesticated livestock species. This has led to the conclusion that GS is only likely to be successful in populations in which *Ne* is much reduced such as highly selected breeding sub-groups or seed orchards. (Thavamanikumar et al. 2013, Beaulieu et al. 2014). However, one advantage that forest trees do have over livestock is that very large full-

sib families can be generated through controlled pollination followed by the collection of large quantities of cones and seed. In a single full-sib family, LD extends for long distances in contrast to open-pollinated populations and this could be exploited in the development of operational approaches to GS targeted towards selection of individuals within family.

Initial experiments based on forest tree species were conducted in *Pinus taeda* (Resende et al. 2012a and 2012b) and in several Eucalyptus species (Resende et al. 2012c) to test the performance of GS in estimating breeding values for a range of selection traits in forest trees using populations with restricted effective population size, several thousand markers and large training populations. For example, using a population of 61 full-sib crosses based on 32 parents, training populations of either 800 or 951 individuals and ~4,853 SNP markers Resende et al. (2012a and b) obtained prediction accuracies of between 0.17-0.74 for nine selection traits in *Pinus taeda*. Although prediction accuracies dropped sharply when models were applied to a new, unrelated population the location of the genomic regions for the traits was consistent suggesting that the loci responsible were conserved across the two populations. One of the perceived benefits of GS in forest trees is the ability to practice multiple trait selection. This is because GS can be used to estimate individual breeding values for each selection trait which could then be combined into a single overall selection index if required.

This study investigates the potential for genomic approaches in Sitka spruce by:

i. Exploring the feasability of using RAD sequencing technology to develop a SNP panel of practical utility in molecular breeding;
ii. Applying the GWAS approach to identifying potential Quantitative Trait Loci (QTL) for 6-year height and 5-year bud burst;
iii. Estimating the accuracy of within-family selection using GS methodology, and;
iv. Discussing how these genomic approaches might be applied in a non-model species.

Materials and Methods

Sample collection

In spring 2005, Forest Research (FR) established large Sitka spruce field trials consisting of the same 1,500 offspring from each of three full-sib families clonally replicated across three climatically contrasting sites in Britain. In what follows, the term offspring will represent a genotype, and a ramet will represent a clonal copy of an offspring. The three full-sib families were based on crosses involving six unrelated parents from the Forest Research Sitka spruce breeding population. Each site was partitioned into four complete randomised blocks and each individual offspring is represented by four ramets at a single site, one ramet per block; 12 ramets in total across all three sites. This study concentrates on one of these full-sib families at a single site located in south-west (SW) England (lattitude 50.59N; longitude 4.06W; 140m above sea level; accumulated temperature above 5^{o}C (AT5) 1,769).

Trait assessments

The four ramets of each of the 1,500 offspring at the SW England site were measured for (i) timing of bud-burst on a 1 to 8 scale according to Krutzsch (1973) at the start of their fifth growing season, and (ii) height (cm) after six growing seasons. For bud-burst, all ramets in the trial were assessed on three occasions over a three week period and the occasion which provided the greatest variance of scores among offspring was used in the analysis. Mortality on the site was low (0.2% or 120 trees at five years; 0.3% or 130 trees at six years), with none of the offspring having more than one loss amongst its four representative ramets.

RADseq in Sitka spruce

DNA was extracted from the needles of one representative ramet of each genotype. The needles (100mg) from each sample were finely chopped and placed in a 2ml Eppendorf tube containing two stainless-steel ball bearings (3mm). The samples were frozen in liquid nitrogen, ground to a very fine powder using a Reitch mixer-mill and stored at -80^{o}C. DNA was extracted using the Qiagen DNeasy Plant mini-kit. The Qiagen protocol was modified in a number of different ways to maximise DNA yield. For the lysis step, 600μl lysis buffer were used and the incubation period was extended to 45 mins. For the neutralisation step, 195μl neutralisation buffer was used and the period on ice increased to 20 mins. The elution incubation was increased to 15 mins and the elution-product was re-applied to the column and spun through a second time. The quality and concentration of DNA extractions were checked using a PicoGreen spectrophotometer (Invitrogen) and only those extracts which contained at least 2.5μg of DNA were taken forward for RAD analysis.

Primary digestion: selection of restriction enzyme

The first step of a RADseq study is the selection of the most appropriate restriction enzyme(s) since this determines the number of genetic markers obtained. All genotyping projects operate within a restricted budget and therefore selection of the appropriate restriction enzyme(s) involves a necessary compromise between the number of markers genotyped, the number of individuals multiplexed and the depth of coverage

required per locus per genotype. A pilot study to inform the choice of restriction enzyme(s) was therefore carried out in which the DNA of two parents and 20 offspring from the full-sib family was digested using the following four restriction enzymes; two 8-base pair (bp) (SbfI and SgrAI) and two 6-bp (PstI and XmaI). Using the methods described by Etter et al. (2011) RADseq libraries were prepared for each enzyme using the size range selection of 300 to 700-bp. To get a better coverage of the parents, we used a ratio of five times the amount of parental DNA relative to offspring DNA in order to achieve a 5-fold increase in the number of Illumina reads for the parental samples compared to those for each offspring. The RADseq libraries were sequenced in High Output lanes on the Illumina HiSeq 2000 instrument. Libraries from the four enzymes were sequenced in separate lanes, with an additional lane for the library relating to PstI due to lower than expected number of reads observed in the first lane.

Second digestion RAD (SD-RADseq)

Of the four enzymes tested the 6-bp PstI enzyme (restriction site CTGCAG) came closest to providing our target number of mappable markers but it exceed it by around 24% (results not shown). In order to reduce the number of markers further a novel complexity-reducing step was developed in which the products of the primary digestion with PstI were subjected to a second digestion with an additional enzyme. Since the smallest DNA fragments in the library were 300-bp long, the additional enzyme was selected on the basis that it would cut 24% of the markers within the first 300-bp beyond the restriction site in order to remove such fragments from the library. In order to be conservative, the length was lowered to 250-bp.

The techniques employed to achieve this reduction via the choice of the most appropriate second restriction enzyme involved extracting the paired-end reads associated with each marker across all individuals in the library, and assembling them using IDBA-UD (Peng et al. 2012), with a minimum contig size of 700-bp. We then checked the frequency of cutting sites within the first 250-bp for all commercially available restriction enzymes using the application *restrict* from the EMBOSS suite (Rice et al. 2000) excluding those with cutting sites in any of the RAD adapters used. The enzyme 'AlwI' (restriction site GGATC) was chosen since it showed presence of cutting sites within the first 250-bp in 24.6% of the paired-end contigs (see 'Results').

To test the reduction in total number of markers using the AlwI enzyme, a new RADseq library was created based on just the two parents. A second digestion (SD) RADseq library was prepared by digesting the new PstI RADseq library with the AlwI restriction enzyme for 30 mins at 37°C, followed by a heat inactivation step of 10 mins at 65°C. To reduce the sequencing costs, the second digestion library was sequenced in an Illumina MiSeq run (50-bp single end (SE)). We evaluated the effectiveness of the AlwI SD-RADseq library as follows: first, we obtained our reference 50-bp set of undigested segregating markers by running the RADseq analysis (as explained below) on the reads from the PstI RADseq library trimmed to 50-bp; second, we obtained the AlwI SD-RADSeq markers of the parents running the RADseq analysis with the same parameters; finally, we mapped the observed RADSeq markers of the SD-RADSeq digested parents to the catalogue of markers from the undigested parents and counted the number of undigested markers hit.

Subsequently, using this secondary digestion process, RADseq was performed on 622 randomly chosen progeny from the full-sib family. A total of 48 offspring were sequenced per Illumina HiSeq 2000 lane (High Output, SBS chemistry v1). Costs per sample were further reduced by sequencing to 50-bp single-end reads rather than the 100bp paired-end used for the pilot libraries.

Processing of RADseq data

RADseq reads for each sample within libraries were demultiplexed using the software RADtools v1.2.4 (Baxter et al. 2011) with parameter –fuzzy_MIDs (this allows one base mismatch in the barcode). Prior to further analysis, Illumina adapters were removed from the reads using *scythe* v0.994 (Buffalo 2014), and reads were filtered with a minimum quality threshold of Q20 using *Sickle* v1.33 (Joshi and Fass, 2011). Reads from the pilot study libraries were trimmed to 96-bp to remove the last few cycles where read quality drops in Illumina longer reads, but left untrimmed at 50-bp in SD-RADseq libraries (cycle quality remains high in short read lengths). *De novo* clustering of RADseq markers and sample genotyping were carried out using the *Stacks* software v0.9996 (Catchen et al. 2011). First, RADseq reads for each sample were grouped in 'stacks' of reads (roughly corresponding to markers) using the *ustacks* module, with a maximum of two SNPs between tags ("alleles" within a stack (parameter -M = 2), deleveraging (-*d*) enabled, and a minimum number of reads per stack (-*m*) of two for the pilot study. A value of m = 2 maximises the number of stacks at the expense of grouping PCR and sequencing errors into stacks, but for the pilot we wanted to obtain a rough estimate to the total number of markers for each enzyme. For the final SD-RADseq libraries we used a value of m = 2 for the offspring but m = 12 for the parents (to remove low coverage markers originated from PCR and sequencing errors in the samples used as references for the mapping). A catalogue of markers was constructed from the stacks observed for the parents using the module *cstacks* with a maximum number of mismatches between sample tags or alleles of zero (-n = 0); this is the recommended value by Catchen et al. (2011) for a F1 pseudo-test cross.

Genotypes were calculated comparing the markers in each sample with the alleles in the catalogue using the *genotypes* module, selecting markers appearing in a minimum of four progeny samples (*-r* = 4) and with a minimum stack depth of five reads (*-t* = 5), and the map type option *–CP* for "cross-pollination". Genotypes were exported in a JoinMap 3 format (Van Ooijen and Voorrips, 2001) for linkage mapping analysis.

Quality control and linkage maps

A set of 34,347 markers were detected by the 'Stacks' software. Following filtering out those markers which showed evidence of non-Mendelian segregation and those which were missing in more than 300 individuals, this number was reduced to 8,397. Custom software implementing the method of minimum recombinations (Olson and Boehnke 1990) was used to obtain the linkage groups and maps for each group. Linkage groups were confirmed by repeating the analysis using JoinMap 3 (Van Ooijen and Voorrips, 2001) in which 8,132 markers were larranged into 12 linkage groups.

Statistical analysis

The statistical analysis of the height and bud-burst data was based on the mean performance of the ramets representing each offspring. Since only a single family was genotyped, with data from one site, and each offspring had one ramet per block, the genomic analysis was free of nuisance factors and family stratification. In the single non-genomic model described below, all ramets were included.

Genetic variation in height & bud-burst

Variance was estimated using two different approaches, either with or without information from genomics. Ignoring the genomics data, a mixed linear model was fitted to 5,987 ramets (5 year bud-burst) or 5,982 ramets (6 year height) of the form:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (1)$$

Here, $\mathbf{y}$ is the vector of observations for each offspring ramet; $\boldsymbol{\beta}$ is the vector of nuisance fixed effects representing the mean (1 d.f.) and blocks (3 d.f.); $\mathbf{u}$ is the vector of random Multi-Variate Normal (MVN) effects for each of the 1,500 offspring; $\mathbf{X}$, $\mathbf{Z}$ are design matrices relating observations to effects; and $\mathbf{e}$ is a vector of MVN residuals for each offspring ramet. It was assumed $\mathbf{u} \sim \text{MVN}(0, \sigma_M^2\ \mathbf{I}_O)$ and $\mathbf{e} \sim \text{MVN}(0, \sigma_E^2\ \mathbf{I}_R)$ where $\mathbf{I}_O$ is the identity matrix for offspring, and $\mathbf{I}_R$ is the identity matrix for ramets. The components were estimated using ASReml 3 (Gilmour et al. 2009). The variance σ_M^2 includes all genetic variance found within full-sib families which, in the absence of selection is expected to be ½ of the additive genetic variance (σ^2_A) and ¾ of the dominance variance (σ^2_D) plus other fractions of the epistatic variances. For each trait, the phenotypic variance was estimated as $\sigma^2_W = \sigma_M^2 + \sigma_E^2$ and a broad heritability *within* family as $H_{(1)}^2 = \sigma_M^2/\sigma^2_W$. A further estimate was derived of the fraction of the genetic variance in means of 4 ramets, $H_{(2)}^2$ which replaced σ_E^2 by $\sigma_E^2/4$.

Variances were estimated for the genomic information by using a G-BLUPmodel. A genomic relationship matrix $\mathbf{G}$ was constructed from the SNP information following Amin et al. (2007), where the genomic relationships between animals i and j is given by:

$$g_{ij} = n^{-1}\sum_{k=1}^{n} (x_{ik}-2p_k)(x_{jk}-2p_k)/[2p_k(1-p_k)]$$

$$g_{ii} = 1 + n^{-1} \sum_{k=1}^{n} (H_{E,k}-H_{ik})/H_{E,k}$$

where x_{ik} is the genotype of the i^{th} individual at the k^{th} SNP when coded as 0, 1 and 2, for the reference allele homozygote, the heterozygote and alternative homozygote, respectively; p_k is the frequency of the reference allele, n is the number of SNPs used for estimating relationships, $H_{E,k}$ is the expected heterozygosity at locus k, and H_{ik} is the observed heterozygosity in animal i at locus k. Pairs of offspring had differing arrays of genotypes used to calculate relationships since with RADseq there is a randomness in the compliment of loci which achieved the necessary thresholds to be assigned. In this study, all offspring are full-sibs and so the expected pairwise relationships (based upon sampling of alleles) are identical, but the genotype data allow the actual similarity and dissimilarity in the true relationships to be quantified. The following mixed linear model was fitted to the means over the 4 blocks for the 622 offspring which were considered to have sufficient genotypic data:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (2)$$

Here, $\mathbf{y}$ is the vector of mean phenotype for each offspring; $\mathbf{1}$ is a vector 1's; μ is the population mean; $\mathbf{u}$ and $\mathbf{e}$ are vectors of genetic and residual effects respectively for each of the offspring included in the genomic analysis. Z is the design matrix for the genetic effects, which here ie equal to $\mathbf{I}_O$. It was assumed $\mathbf{u} \sim \text{MVN}(0, \sigma_G^2\ \mathbf{G})$ and $\mathbf{e} \sim \text{MVN}(0, \sigma_R^2\ \mathbf{I}_O)$. In this model σ_R^2 is the variance of the deviations averaged over the four blocks. The model was fitted using ASReml 3 (Gilmour et al.2009). For each trait, the phenotypic variance was estimated as $\sigma^2_T = \sigma_G^2 + \sigma_R^2$ and genomic heritability was calculated as $h^2 = \sigma_G^2/\sigma^2_T$. The variance σ_G^2 is an estimate of the additive genetic variance in the set of full-sibs contained within $\mathbf{G}$.

QTL detection in a single full-sib family for height and bud-burst

The balance of the design, and the use of a single family, removed the need to consider nuisance factors and cryptic family stratification in the analysis. Therefore, GWAS analyses were carried out on the 622 genotyped offspring with custom software using the GRAMMAR approach implemented in an in-house bespoke programme (Pong-Wong pers.comm). This involved 8,132 loci, but with different subsets of offspring available per locus. Significance was assessed from 10,000 permutations, where the phenotypes of the 622 offspring were randomly re-assigned to the sets of genotypes to establish 5% genome wide significance levels which is the only significance level reported below. Significant SNPs were assigned to the syntenic groups as described previously.

Genomic evaluations to predict phenotypes and breeding values

The potential accuracy of genomic evaluations within family for height and bud burst were examined using model (2) and five-fold cross-validation. The 622 genotyped offspring were divided independently of the phenotype and genetic information into five sets, each containing either 124 or 125 offspring. In five cycles of analyses, the phenotypes of each 124 or 125 tree set were masked in turn and model (2) was fitted to the remaining 497 or 498 phenotypes. In each cycle, the accuracy of predicting the masked phenotypes was calculated as the correlation of predicted value and phenotype, along with the bias measured by the regression of y on $\hat{y}$. A genomic predictor cannot be directly correlated to breeding values here since the true breeding value is unknown, therefore the predicted breeding value is correlated with the phenotype. The phenotype has added noise from the environment so in genomic predictions of breeding value the correlation cannot be one. The maximum value that may be expected (t) was calculated from heritability estimates after adjusting for variance within families and means of four ramets, using $t = \sqrt{[½h^2/(½h^2 + ¼(1-h^2))]} = \sqrt{[2h^2/(1 + h^2)]}$. Correlations with phenotype were divided by t to approximate the correlation with breeding value. The values of h^2 used were 0.8 (Hannertz et al. 1999) and 0.3 (Lee et al. 2002a) for bud-burst and height respectively.

Sensitivity analyses for prediction accuracy were investigated in two ways using a sub-set of the 250 offspring chosen to have the greatest number of genotyped markers and re-randomised into five equal sets. The predictive accuracies were re-assessed by cross-validation using (i) 200 offspring in five folds as training sets to predict the remaining 50, and similarly (ii) 50 as a training set to predict the remaining 200.

Results

Development of RADSeq markers

Evaluation of restriction enzymes for RADseq libraries in the pilot population

The four restriction enzymes tested on the parental trees and 20 offspring produced very different numbers of markers (Supplementary Table S1). In all cases, the number of observed markers was lower than expected assuming the restriction sites were randomly distributed and if we consider a genome size of 19.6Gb and GC content of 37.9% (values for the closely related species *P. abies*; Nystedt et al. 2013). All enzymes produced >100,000 markers although, as expected, the number of markers obtained using the two 8bp cutters (average number markers on the parents: SbfI = 161,627; SgrAI = 155,185) was much lower than that with the 6-bp cutters (PstI = 592,684; XmaI = 941,751). Nonetheless, the number of markers for mapping was much lower, ranging from ~2,000-3,000 for SbfI and SgrAI to ~32,000 for XmaI and ~56,000 for PstI. The larger reduction in mapping markers for XmaI compared to PstI is due to a much lower average coverage per marker in the samples despite having similar number of reads (Supplementary Table S2), due to (i) larger number of markers in the XmaI vs. PstI libraries, and (ii) markers with very high coverage probably located in high copy repetitive elements found in the genome (e.g. in parent 1 the maximum marker coverage is 169,611 reads for PstI and 777,918 reads for XmaI) (Supplementary Table S2).

Second Digestion RADseq for complexity reduction of RADseq libraries

Of the total number of identified markers, ~80% exhibited expected Mendelian segregation (Supplementary Table S1). The number of segregating markers for the 8-bp cutters SbfI and SgrAI was too low to provide a high resolution linkage map in the nearly 20Gb genome of *P. sitchensis*. Although the frequent cutter (XmaI) produced a more appropriate number of segregating markers (~25,000), overall the XmaI library presented low coverage per marker (see above). A large sequencing effort would be required to get enough coverage for the segregating markers if XmaI were to be used to produce RADseq markers in a large progeny. The evidence indicated that the PstI cutter was the most appropriate enzyme for our purposes, but even by using this enzyme it would have been too costly to genotype our set of 622 offspring.

In order to genotype all progeny within our budget we estimated that we needed a further reduction of ~25% in the number of loci. *In silico* restriction of the paired-end assemblies of PstI RADseq markers produced a list of potential restriction enzymes (Supplementary Table S3). From that list,

AlwI showed an *in silico* reduction of markers of 24.6%. This enzyme is commercially available so we chose AlwI for our SD-RADseq approach. For the PstI undigested library of the pilot family (with 50bp-trimmed reads) we observed 37,902 mapping markers. From those, we observed 28,976 markers on the AlwI digested parents. This corresponds to a 23.5% reduction in the number of markers, very close to the expected 24.6%, indicating that our SD-RADseq approach was valid.

The list of number of reads and marker coverage for the AlwI SD PstI-RADseq libraries of the full family is shown in Supplementary Table S4. After processing the samples with *'Stacks'* we obtained a final number of mapping markers of 34,347 that were present in at least four individuals from the progeny. This number is larger than the 28,976 markers from our validation analysis above. The pilot family included only 20 offspring compared to 622 in the final dataset. Considering that only markers present in at least four offspring were selected, the larger number of markers observed in the full progeny set is likely to be the result of the larger number of samples.

Variances components for bud-burst and height

Table 1 shows the estimates of variance observed within and between offspring together with standard errors. It is clear from $H^2_{(1)}$ that the correlation among ramets of the same offspring is notably higher for bud-burst than for height. $H^2_{(2)}$ is the value that is most analogous to the estimates obtained from analyses using genomics since these correspond most closely to expectations from averaging over the four ramets for each offspring. It should be remembered that these estimates are from within a single full-sib family and therefore the heritabilities underestimate the fraction of variance shared by offsping as it excludes all genetic variation among full-sib families.

Estimates of variance using the genomic data are shown in Table 2. The matrix **G** used for estimation is built assuming additive SNP effects and so genetic components are additive genetic variance. The estimates given are statistically very significant and, as in Table 1, bud burst has a higher heritability than height. Comparison with Table 1 also suggests that the genetic variance detected using genomics is less than might be expected if the variance between offspring was solely additive genetic variance. If this were true, then $h^2_{(1)}$ would be comparable to $H^2_{(2)}$ derived from the ramets in Table 1. Comparisons of genetic variances using G-BLUP are difficult since the interpretation of σ_G^2 depends on assumptions of Hardy Weinberg equilibria within all marker loci, but one appropriate comparison is $\sigma^2_E/4$ and σ_R^2 in Tables 1 and 2 respectively, as these measure the variance *not* explained by the genetic models used for analysis. This shows that the genomic model (Table 2) has more unexplained variance, particularly for bud-burst, which may be due to the presence of non-additive genetic variance and other sources of variance among full sibs (e.g. any epigenetic effects) or the density of markers being insufficient to capture the full additive genetic variance.

Detection of Quantitative Trait Loci (QTL)

For height, only two significant SNP markers were identified as having genome-wide significance located on linkage groups 4 and 9 (see supplementary Figure S1). The difference between the homozygotes was predicted to be 35 cm (SE 0.076) and 36.8 cm (SE 0.075) respectively, representing 10.4% and 10.9% of the mean height of 335.8cm

For bud-burst, a much larger number of 132 SNP markers was identified as being of genome-wide significance. These occurred in five of the 12 linkage groups, distributed as shown in Table 3 (see also supplementary Figure S2). The distribution of the these SNPs within the linkage groups was clustered on the linkage maps although such correspondence would be expected due to the linkage disequilibrium information that was used to develop both the maps and detect the QTL. Examination of the clustering on the maps suggests 13 distinct QTL; with seven appearing on linkage group 10.

Accuracy of predicting phenotypes with GS

Table 4 shows the estimated genomic evaluation for the full data set (622 trees with 124 or 125 masked trees) and the two sensitivity analysis (200 trees with 50 masked, and 50 trees with 200 masked). The accuracy of prediction of the phenotypes within the family for the full dataset was moderately high and consistent across all five sets giving a mean of 0.58

Table 1 The total variance within offspring and between offspring ($\sigma_W^2 = \sigma_M^2 + \sigma_E^2$) and heritabilities for bud-burst at 5 years of age and height in metres at 6 years of age, from analyses using only ramet phenotypes

Trait	Total Variance	Heritabilities		Residual Variance	
	σ_W^2	$H^2_{(1)}$	$H^2_{(2)}$	σ_E^2	$\sigma_E^2/4$
Bud Burst	0.542 (0.014)	0.57 (0.012)	0.84 (0.007)	0.236 (0.005)	0.059 (0.001)
Height (m)	0.364 (0.007)	0.25 (0.014)	0.57 (0.018)	0.273 (0.005)	0.068 (0.001)

The heritabilities calculated are: observed broad-sense heritability, $H^2_{(1)} = \sigma_M^2/(\sigma_M^2 + \sigma_E^2)$; broad-sense heritability of offspring performance if averaged over 4 ramets/offspring, $H^2_{(2)} = \sigma_M^2/(\sigma_M^2 + \sigma_E^2/4)$. Standard errors are given in parentheses. Variance components are defined in the Materials & Methods.

Table 2 The total variance ($\sigma_T^2 = \sigma_G^2 + \sigma_R^2$) and estimates of heritability when using G-BLUP for bud-burst at 5 years of age and height in metres at 6 years of age

Trait	Total Variance σ_T^2	Residual Variance σ_R^2	Heritability $h^2_{(1)}$
Bud-burst	0.328 (0.019)	0.237 (0.018)	0.40 (0.012)
Height (m)	0.171 (0.010)	0.122 (0.001)	0.30 (0.017)

The heritabilities calculated are $h^2_{(1)} = \sigma_G^2/(\sigma_G^2 + \sigma_R^2)$. Standard errors are given in parentheses. Variance components are defined in the Materials & Methods.

for bud-burst and 0.40 for 6-year height. The higher value for bud-burst was not unexpected due to the apparently greater heritability. Using estimates of $h^2 = 0.80$ for bud-burst (in Norway spruce using the same assessment technique; Hannertz et al. 1999) and 0.3 for 6-year height (Lee et al. 2002a) estimated the accuracy of prediction of the breeding values within the family as 0.62 for bud-burst and 0.59 for height. The estimated accuracies were greater, 0.92 and 0.73, when the estimates of genomic heritability in Table 2 were used. Such values are strikingly high, but note that these are accuracies of predictions within a single full-sib family at a single site, and using means of four ramets for each of the 497/498 offspring in the training set.

When 200 offspring were used to predict phenotypes of the other 50 trees in the sensitivity analysis, the correlation between predicted breeding value and phenotype fell only slightly, to 0.54 for bud burst and 0.38 for 6 year height (Table 4). However the reduction in correlation was much greater when only 50 offspring were used in the training set to predict the breeding values of the other 200, reducing to 0.33 for bud-burst and 0.25 for 6 year height. So whilst it is good to push the extremes when testing a model, the low size of the training popualtion relative to the much larger predicted breeding values population seemed to be of little value on this occasion.

Discussion

Development of a SNP panel of markers using RADseq

RADseq provided a successful and cost effective means of achieving the initial objective of the study, which was to develop a set of SNP markers targeted at Sitka spruce sufficient to study the possibilities of genomic selection for improved performance. The technique enabled the discovery of large numbers of SNP markers in preliminary trials using different restriction enzymes. However the major challenge was to reduce the number of SNPs identified by the restriction enzyme to a level that allowed a cost-effective study, i.e. one that gave adequate coverage per locus per individual and yet allowed a large numbers of individuals to be genotyped within the resources available. The flexibility of RADseq is evident in that the compromise was achieved by using the 'Pst1' enzyme incorporating an extra digestion with the 'Alw1' enzyme to reduce numbers of loci further.

A shortcoming of this study is that the segregating SNPs are restricted to a single family. It remains unknown to what degree other families would exhibit the same segregating SNPs using the same enzyme digestion protocol, and also to what extent these would be in common between families. At present it is unclear if RADseq is the way to proceed for routine genotyping as the cost of developing custom SNP chips using the discovered SNP markers continues to fall. The lack of a Sitka spruce whole genome sequence assembly remains a problem when using the SNPs generated in the development of a linkage map. Showing order within a linkage group was only satisfactorily addressed in this study using custom minimum-recombination software; consequently the quality of the map remains unknown.

Table 3 Distribution of the 122 genome wide significant markers for bud-burst at 5 years of age, across the 12 Sitka spruce linkage groups

Linkage Group	1	6	8	10	12	Total
Significant SNPs	7	6	4	99	16	132
Informal QTL	2	2	1	7	1	13

Identifying QTL

The study identified two significant QTLs for height located on two distinct linkage groups from the 12 available. In contrast, the number was much higher for bud-burst with 132 significant QTLs clustered on five of the linkage groups. The reason for the greater number of QTLs for bud-burst compared to height is unclear but differences in the heritability and the genetic architecture of the traits could be contributory factors. The ability to map the position of SNP markers will be improved once an assembled whole genome sequence for Sitka spruce becomes available in the future. Our results contrast with those of Pelgas et al. (2011) working with white spruce who found a total of 33 distinct QTLs for bud-burst and 52 for height growth across four saturated individual linkage maps representing two unrelated mapping populations. Corresponding numbers for the composite map were 11 and 10 QTL. The reasons for the greater number of QTLs in their study are unclear although it is worth noting that they adopted a low stringency in their QTL identification. Difference in the structure of the white spruce and Sitka spruce genomes is unlikely to be the main reason as the two species are very close taxonomically. Indeed we were able to match about 80% of our SNP containing RAD sequences to the publicly available assembly of white spruce provided by Birol et al. (2013).

Table 4 Estimated predictive accuracy of 5 year bud-burst and 6 year height phenotypes and breeding values (BV) derived from 5-fold cross-validation using different numbers of trees in the training set (*n*)

Trait	Training Set	Correlation with Phenotype			Correlation with BV	
	n	Min	Max	Mean	(a)[a]	(b)[b]
Bud Burst	497/498	0.50	0.65	0.58	0.62	0.92
	200	0.48	0.59	0.54	0.57	0.85
	50	0.17	0.34	0.33	0.35	0.52
Height	497/498	0.32	0.46	0.40	0.59	0.73
	200	0.27	0.50	0.38	0.56	0.69
	50	0.21	0.28	0.25	0.37	0.46

The correlations with phenotypes shown are the minimum, maximum and mean values obtained across the five validation sets. The correlations with BV are estimates obtained by scaling the mean correlation with phenotype by its upper bound either (a) by *t* derived from the heritabilities of Lee et al. (2002a) as shown in Materials and Methods, or (b) the square root of the heritabilities shown in Table 2.

[a] Values of *t* used were 0.94 and 0.68 for bud-burst and height respectively.

[b] Values used for scaling were 0.63 and 0.55 for bud-burst and height respectively.

Accuracy of selection using GS methodology

Genomic evaluations using G-BLUP do not depend on sequence or marker order, although an assumption is often made that the SNPs used to build relationship matrices are scattered randomly across the genome. The accuracies presented in this study are strictly within families and show that moderate to high predictions of breeding value within a single full-sib family are attainable with training sets consisting of only 50 offspring, albeit using four ramets per offspring which increased the genetic information in the training data. In this study the accuracy of the predictions of phenotype are unambiguous, and support good predictions of breeding value. However, more precise estimates of the accuracy of predicting a breeding value are indirect and less clear. Two routes of assessment were taken: firstly, using literature estimates of h^2 to overcome the problem that the clonal structure of the population generates only an estimate of broad-sense heritability (without using genomic data); secondly, by using the estimates of h^2 from the genomic data obtained, but where the results left open the question of whether or not the genomic data were sufficient to capture all the genetic variance segregating within the family. The putative accuracies of predicting breeding values were greater using the second option; however these estimates are likely to be optimistic as they will be inflated by any underestimate of the genetic variance. However, the results do indicate that larger training sets were capable of accurately predicting the genetic variance that was captured by the markers.

There are reasons why genetic variance may be missed in this study when using **G**. Firstly **G** was constructed in a simple fashion which used genotype assignments including some degree of error since the genotypes were assigned based on a minimum number of markers. If an assembled sequence had been available, specifically a linkage map, then it would be feasible to impute genotypes across all offspring (since coverage of parents was >60 reads per locus) with considerable accuracy, and so greatly reduce both the genotype errors and missing genotypes. It would be anticipated that a more accurate **G**-matrix would result in greater accuracy in prediction. Evidence for this was that the training sets using the 200 most reliably genotyped individuals gave accuracies of predicting phenotypes only very slightly lower than using the training sets of nearly 500 trees. The ability to impute from low density genotyping is important to opening up larger training sets.

The predictions reported here are strictly within a single family that cannot be extrapolated to between families. Beaulieu et al. (2011) examined the possibility of marker transferability between families in white spruce. They found predictions within family to be more precise than between families. When the validation involved families not in the training group; the accuracies obtained were small, sometimes negative, and typically not statistically different from zero. It is not clear from their study how much between family predictions depended on the ability to predict some of the sibs within-family that were within the validation set. As with Sitka spruce, the large effective population size of white spruce would restrict the potential to predict across families. Therefore, the evidence to date would suggest the major benefit of genomic selection would be in its potential to predict breeding value within large full-sib families of trees.

The accuracy of our breeding value for height was 0.59 which compares to previous GS estimated accuracies of the same trait in *Pinus taeda* of 0.64-0.74 (Resende et al. 2012a) and 0.47-0.52 (Zapata-Valenzuela et al. 2012). It is not possible to attribute the underlying reasons for the differences in the accuracies since the family structure, size of the training populations and number of markers analysed differed between the studies. It must be remembered that our results provide the prediction accuracies obtained when using only a single full-sib cross; the simplest possible population structure. short-

coming of this study was the lack of resources to investigate marker transferability between even a sub-set of the other two full-sib families planted in 2005. However, since this is the first published study investigating the association of markers and any phenotypic characteristics in Sitka spruce, it is a worthwhile starting point for future comparisons.

Application in the breeding of Sitka spruce

As explained earlier, temperate-zone tree breeding can be time consuming and costly due to the long generation intervals and expression late in life of important commercial traits. Tree breeders try to circumvent these problems by employing indirect selection techniques measured early in life that are genetically well correlated with final selection goals. This reliance on progeny testing slows generation turnover and progress. Recent developments in breeding of dairy cattle have seen traditional progeny testing being replaced with very early age GS in a bid to reduce generation intervals whilst increasing selection intensity and at the same time, reducing overall operating costs. The slight reducion in accuracy per geneotype evaluation is more than compensated by the increase in annual genetic gain and financial benefits.

Critical analysis shows some important differences in the population structure of dairy cattle and Sitka spruce. The Sitka spruce breeding programme has completed just one cycle of selection and testing and compared to crops and animals, is undomesticated. This results in a large effective population size (*Ne*) and low linkage disequilibrium (LD) at the population level. A deliberate re-design of the breeding programmes to reduce the *Ne* of Sitka spruce can have benefits within a generation (van Heerwaarden, pers comm.) but the biological restrictions of the generation interval of Sitka spruce means the LD impact will not disappear in the next decade. In contrast, dairy cattle are highly domesticated, have had much lower *Ne* for generations and consequently LD extends over much longer distances. A new model is required for Sitka spruce and likely most other tree species.

An alternative approach for Sitka spruce breeders may be provided by the example of Atlantic salmon (*Salmo salar*) as described by Lillehammer *et al*. (2013). In common with tree breeding, salmon has a greater *Ne* than dairy cattle, and an ability to generate large numbers of individuals per family. Salmon breeders have turned this to their advantage in a form of full-sib testing and generation of familiy specific DNA-markers to enable GS within the family. If Sitka spruce breeders followed this model, it would involve creating a number of single-pair matings (full-sib families) through controlled pollination, and planting in the field of a restricted number of offspring (around 50 per family) appropriate to assessing the importance of GxE. Selection between the families would continue to be based on traditional assessment of phenotypes in the field. A set of family-specific DNA-markers could be generated by measuring and sequencing the initial few offspring once they have reached suitable phenotypic indirect-selection ages. Very intensive within family selection could then follow by use of the bespoke family-specific prediction equations to reduce hundreds, perhaps thousands of embryos from repeat pollination of the same parents, to just a few selected superior genotypes for either further field testing, multiplication and direct deployment or (following maturation) involvement if further breeding work with similar but unrelated early selections. The intensity and commitment to the two-stage selection process will likely decline with time as confidence increases in the accuracy of such early marker-based selection for a wide suite of traits including adaptability and disease resistance, negating the need for further field testing. Also with time, knowledge may grow in applying markers across unrelated families i.e. transferability of markers may become possible, but that is not currently envisaged. Following final selection, genotypes could be directly deployed to the field perhaps using advanced tissue culture techniques. Reduction of the the generation interval however will still not be possible until the next bottleneck preventing earlier generation turn-over in Sitka spruce which currently is the relatively late flowering age (around 15-years old) of the species.

The advantage for Sitka spruce breeding now, is that although accuracy of genotype prediction is reduced slightly, generation gain is likely to be more due to the much increased selection intensity and reducing generation interval. As found in cattle breeding there is the additional advantage that overall field trials costs are reduced. The proposed Sitka spruce model does still rely on field trials to a certain extent but the assumption is made that whilst field trial costs will at best stay constant, genotyping accuracies are likely to increase as marker density increases and genotyping costs reduce. See Isik (2014) for a likely application of GS in loblolly pine in which the generation interval is predicted to half whilst genetic gain per year is doubled.

Conclusion

This has been the first study to investigate the potential of molecular breeding in Sitka spruce, a non-model species for which there is currently no whole genome assembly. The study found that RADseq technology was successful in generating a large number of randomly located markers that could be developed into a SNP panel. Applying the GWAS approach to identify potential QTL proved encouraging for 5 year bud-burst with 132 significant SNPs identified, albeit clustered, but only two for 6 year height. The prediction of phenotypes using GS methodology resulted in encouraging accuracies and demonstrated potential for use of this technology, although it is challenging to extrapolate beyond the

limitation that the data were generated from a single full-sib family, clonally replicated over a single site. Nevertheless, family-specific predictions could be used to increase within-family selection intensity to achieve annual gain commensurate with traditional Sitka spruce breeding, with lower overall operational costs due to reduced field testing requirements.

Acknowledgements The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement No 211868 (Project Noveltree). John Woolliams and Andrew Law were supported by BBSRC Institute Strategic Grant funding (BBS/E/D/20211551, BBS/E/D/20211554) and Edinburgh University, Scotland. Steve Lee and Joan Cottrell were supported by core funding from the Forestry Commission (GB). Pablo Fuentes-Utrilla was funded by a fellowship from the "Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I-D+i 2008-2011" of the Ministry of Education, Spain. The authors would like to thank Edinburgh Genomics (formally GenePool) for the DNA sequencing work and the Technical Support Unit of Forest Research for the phenotypic assessments, and also helpful comments made by anonymous referees.

Data Archiving Statement The data are archived in the University of Edinburgh Datavault with the following doi: http://dx.doi.org//10.7488/cfb6f594-4dc8-4928-a790-90dbc8bec2aa

Author's Contribution P. Fuentes-Utrilla carried out the RADseq library preparation and analysis, developed the SD-RADseq method and contributed to the paper; C. Goswami carried out the data analysis and contributed to the paper; J.E. Cottrell co-conceived the study and co-wrote the paper; R. Pong-Wong was involved in the basic design of the RADseq work; A. Law gave guidance on data analysis; S.W. A'Hara was involved in the DNA extraction and contributed to paper; S.J. Lee co-conceived the study and co-wrote the paper; J. A. Woolliams also co-conceived the study, co-wrote the paper, and was the main guidance in interpretation of the data.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

References

Amin N, van Duijn CM, Aulchenko YS (2007) A genomic background based method for association analysis in related individuals. PLoS One 2, e1274

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG et al (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS ONE 6, e19315

Beaulieu J, Doerksen T, Boyle B, Clément S, Deslauriers M, Beauseigle S, Blais S, Poulin P-L, Lenz P, Caron S et al (2011) Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. Genetics 188:197–214

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014) Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics 15(1): 1048

Bousquet J, Isabel N, Betty Pelgas B, Cottrell J, Rungis D, Ritland K (2007) 'Spruce'. In: Genome Mapping and Molecular Breeding in Plants, Vol. 7 Forest Trees, pp93-114 Chitta R. Kole (Ed). Springer, Heidelberg, Berlin, New York, Tokyo

Birol I, Raymond A, Jackman S, Pleasance S, Coope R et al (2013) Assembling the 20Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497

Buffalo V (2014) Scythe - A Bayesian adapter trimmer (version 0.994 BETA). Available at https://github.com/vsbuffalo/scytheCatchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwa JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. G3 1:171–182

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH, De Koning D-J (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes|Genomes|Genetics 1(3):171–182

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA, Welch JJ (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE 6(4):e18561

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml User Guide Release 3.0. Hemel Hempstead VSN International Ltd

Goading GD, Bridgwater FE, Bramlett DL, Lowe W (1999) Top Grafting Loblolly Pine in the Western Gulf Region. Proceedings of the 25th Biennial Southern Forest Tree Improvement Conference, New Orleans, Louisiana, USA

Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. In "Genomics of Plant Genetic Resources Vol 1 pp 651-682. eds R. Tuberosa, A. Graner & E. Frison. DOI 10.1007/978-94-007-7572-5_26, Springer Science + Business Media Dordrecht 2014

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomics 7:241–255

Grattapaglia D, de Alencar S, Pappas G (2011) Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of *Eucaplytus. grandis* and *E. globulus*. BMC Proc 5:45

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101(1):5–18

Hannerz M, Sonesson J, Ekberg I (1999) Genetic correlations between growth and growth rhythm observed in a short-term test and performance in long-term field trials of Norway spruce. Can J For Res 29: 768–778

Hermann RK (1987) North American tree species in Europe: transplanted species offer good growth potential on suitable sites. J For 85:27–32

Houston RD, Haley CS, Hamilton A, Guy DR, Tinch AE, Taggart JB, McAndrew BJ, Bishop SC (2008) Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). Genetics 178(2):1109–1115

Isik F (2014) Genomic selection in forest tree breeding: the concept and the outlook to the future. New For 45:379–401

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

Krutzsch P (1973) Norway spruce development of buds. IUFRO S2.02.11. International Union of Forest Research Organization, Vienna

Lee SJ, Connolly T (2010) Finalizing the selection of parents for the Sitka spruce (*Picea sitchensis* (Bong.) Carr) breeding population in Britain using Mixed Model Analysis. Forestry 83:423–431

Lee SJ, Thompson D, Hansen JK (2013) Sitka spruce *Picea sitchensis* (Bong.) Carr) In: Forest Breeding in Europe. Managing Forest Ecosystems 25: 177-227

Lee SJ, Woolliams J, Samuel CJA, Malcolm DC (2002a) A study of population variation and inheritance in Sitka spruce: II. Age trends in genetic parameters and for vigour traits and optimum selection ages. Silvae Genet 51(2–3):55–64

Lee SJ, Woolliams J, Samuel CJA, Malcolm DC (2002b) A study of population variation and inheritance in Sitka spruce: III. Age trends in genetic parameters and optimum selection ages for wood density, and genetic correlations with vigour traits. Silvae Genet 51:143–151

Lillehammer M, Meuwissen THE, Sonesson AK (2013) A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genet Sel Evol 45:39

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Meuwissen THE, Hayes BJ, Goddard ME (2016) Genomic selection: a paradigm shift in animal breeding. Animal Front 6(1):6–14

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325–330

Nystedt B, Street N, Wetterbom A, Zuccolo A, Lin Y et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451):579–584. doi:10.1038/nature12211

Olson JM, Boehnke M (1990) Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. Am J Hum Genet 47(3): 470–82

Pelgas B, Bousquet J, Merimans K, Isabel N (2011) QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. BMC Genomics 12: 145

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428

Resende MDV, Munoz P, Acosta JJ, Peter GF et al (2012a) Accelerating the domestication of trees using genomic selection methods: accuracy of prediction models across ages and environments. New Phytol 193:617–624

Resende MFR, Munoz P, Resende MDV, Garrick DJ et al (2012b) Accuracy of genomic selection methods in a standard dataset of Loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD et al (2012c) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Ritland K, Krutovsky K, Tsumara Y, Pelgas B, Isabel N et al (2011) Genetic mapping in conifers. In: Plomion C, Bousquet J, Kole C (eds) Genetics, Genomics and Breeding of conifers. CRC Press, New York

Slavov GT, Nipper R, Farrar K, Allison GG, Bosch M et al (2014) Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. New Phytol 201:1227–1239

Thavamanikumar S, Southertonn SG, Bossinger G, Thumma BR (2013) Dissection of complex traits in forest trees-orrortunities for marker-assisted selection. Tree Genet Genomes 9:627–639

Tuskan G, DiFazio S, Jansson S, Bohlmann J, Grigoriev I et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Van Ooijen JW, Voorrips RE (2001). JoinMap ® 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, Netherlands

Van Raden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME et al (2013) Genomic imputation and evaluation using high density Holstein genotypes. J Dairy Sci 96:668–678

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291:1304–1351

Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J et al (2012) SNP markers trace familial linkages in a cloned population of Pinus taeda-prospects for genomic selection. Tree Genet Genomes 8: 1307–1318