Plant Science xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Genomic selection in maritime pine

Fikret Isik[a,1], Jérôme Bartholomé[a,b], Alfredo Farjat[c,1], Emilie Chancerel[a,b], Annie Raffin[a,b], Leopoldo Sanchez[d], Christophe Plomion[a,b], Laurent Bouffier[a,b,*]

[a] INRA, UMR1202, BIOGECO, Cestas F-33610, France
[b] Univ. Bordeaux, UMR1202, BIOGECO, Talence F-33170, France
[c] Department of Statistics, North Carolina State University, Raleigh, NC, USA
[d] INRA, UR0588, AGPF, 45075 Orléans, France

ARTICLE INFO

Article history:
Received 22 March 2015
Received in revised form 4 August 2015
Accepted 13 August 2015
Available online xxx

Keywords:
Linkage disequilibrium
Tree breeding
Genomic relationship
Bayesian regression
Pinus pinaster

ABSTRACT

A two-generation maritime pine (*Pinus pinaster* Ait.) breeding population ($n = 661$) was genotyped using 2500 SNP markers. The extent of linkage disequilibrium and utility of genomic selection for growth and stem straightness improvement were investigated. The overall intra-chromosomal linkage disequilibrium was $r^2 = 0.01$. Linkage disequilibrium corrected for genomic relationships derived from markers was smaller ($r_V^2 = 0.006$). Genomic BLUP, Bayesian ridge regression and Bayesian LASSO regression statistical models were used to obtain genomic estimated breeding values. Two validation methods (random sampling 50% of the population and 10% of the progeny generation as validation sets) were used with 100 replications. The average predictive ability across statistical models and validation methods was about 0.49 for stem sweep, and 0.47 and 0.43 for total height and tree diameter, respectively. The sensitivity analysis suggested that prior densities (variance explained by markers) had little or no discernible effect on posterior means (residual variance) in Bayesian prediction models. Sampling from the progeny generation for model validation increased the predictive ability of markers for tree diameter and stem sweep but not for total height. The results are promising despite low linkage disequilibrium and low marker coverage of the genome (~1.39 markers/cM).

1. Introduction

Genomic selection (GS) is considered a paradigm shift in animal and plant breeding [1] and has the potential to revolutionize breeding of forest trees. GS aims to trace all the quantitative trait loci (QTL) controlling phenotype to predict genetic merit of individuals [2,3]. GS relies on a large number of DNA markers that cover the whole genome to exploit the linkage disequilibrium (LD) between markers and any QTL. Theoretically, if the marker coverage is dense enough, all the QTL controlling a trait will be in LD with at least one marker [4]. Therefore, the success of GS depends on the effective population size and on the extent of LD between DNA markers and loci affecting complex traits [5]. In contrast to marker-assisted selection, prior information on the association between phenotypes and markers, the location of QTL on the genome and their relative effect upon the phenotype are not prerequisites for GS. Advances in high-throughput genotyping technologies [6–8] has made available a large number of DNA markers to animal and crop breeders [9–12]. As a result, the concept of GS has been widely used for cattle breeding since 2008 [13–16] and has been extended to other animal and plant breeding programs world-wide [11,17–20]. However, GS is in its infancy with forest trees.

There has been extensive coverage of statistical methods used in GS and they were classified into two groups [21]. In the first group, the i-th phenotypic outcome (y_i) is regressed on markers via the regression function $g\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a vector of marker covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients [22]. Bayesian shrinkage methods [2], ridge regression [23] or Bayesian LASSO regression [21] are statistical methods that fall into this category. Such models allow prediction of individual marker effects. The second approach uses genomic relationships derived from markers in a mixed model framework for prediction of genomic breeding values [24,25]. This method is frequently called Genomic Best Linear Unbiased Prediction (GBLUP) and is an appealing method for ease of computation because there is no need to predict the marker effects. The number of solutions from the model is reduced

* Corresponding author at: INRA, UMR1202, BIOGECO, Cestas F-33610, France.
E-mail address: bouffier@pierroton.inra.fr (L. Bouffier).
[1] Permanent address: Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA.

http://dx.doi.org/10.1016/j.plantsci.2015.08.006
0168-9452/

to the number of individuals. Empirical and simulation studies suggest that the statistical methods usually differ only marginally in the predictive accuracy of genomic estimated breeding values [2,26–28].

Forest trees, particularly conifers subjected to breeding, have long (15 years or more) cycles of breeding and testing [29]. Breeding trees is logistically difficult to implement because of their reproductive biology (late flowering), their large physical sizes and, notably, their late maturation for the phenotypic evaluation of economically important traits. Using markers for selection has long been promoted to reduce the cost and time of progeny testing [30]. Several proof-of-concept studies of genomic prediction in forest trees have been published in recent years based on small (<8 k) number of SNP markers [28,31–35]. Despite advances in developing genomic resources for forest trees [36–38] and promising results from proof-of-concept studies, no application of GS in tree breeding programs has been reported [30,39]. In addition, large physical genome sizes of conifers [40] may pose a challenge to achieving the necessary dense marker coverage of genomes. For example, the genome size of maritime pine (*Pinus pinaster* Ait.), is estimated to be 24.5 Gb [41]. The first whole-genome shotgun assembly of loblolly pine suggests a genome size of 20.1 Gb [42]. Since forest trees are still relatively undomesticated and characterized by large effective population sizes, the extent of LD is expected to be very low in these outcrossing species [42]. For example, in loblolly pine (*Pinus taeda* L.), the average short distance LD (physical scale) based on 19 candidate genes decayed to less than $r^2 = 0.2$ within about 1500 base pairs [43]. In maritime pine the pattern of long distance LD (genetic scale) was examined over 12 chromosomes using 194 unrelated individuals and 2600 SNP markers with an average map distance of 1.4 cM between markers [44]. Authors reported complete lack of long distance LD.

GS success, however, not only depends upon the extent of LD at any given time but also on its dynamics over recombination cycles. Simulation studies suggested that response of GS will decline after each generation because LD weakens after recombination takes place [3,45]. Therefore, a very large number of markers are likely needed to cover the whole genome in conifers in order to develop reliable and stable prediction models across generations. In this study, we used a maritime pine breeding population to estimate the extent of long distance LD and develop genomic prediction models. This is the first genomic prediction proof-of-concept study for this species. The study is based on a breeding population from two successive generations of the breeding scheme. The objectives were two-fold: (i) carry out LD analysis for each linkage group while correcting for genetic relatedness in the population, and (ii) compare three statistical models, namely, genomic Best Linear Unbiased Prediction (GBLUP), Bayesian ridge regression, and Bayesian LASSO for their efficiency in genomic predictions of growth and stem sweep (a measure of tree stem straightness), two important traits of the maritime pine breeding program.

2. Material and methods

2.1. Breeding population and pseudo phenotypes

The maritime pine breeding program in southwestern France started in the 1960s with the phenotypic selection of 635 individuals (G0 population) from unimproved pine plantations [46]. Selected trees were grafted in clonal archives for breeding. Progeny from G0 trees were first obtained by collecting cones on selected trees in the forest (wind pollination with unknown male pollen) then by crosses between grafted copies, using different mating schemes. Progeny ($n \approx 100$) from crosses were tested in replicated field trials to select the next-generation population (G1 population). The breeding population will have completed three generations of breeding, testing and selection in 2020 (selection of G3 population). Stem sweep (distance between the tree stem and a vertical pole at 1.5 m above ground) was measured between age 7 and 12 years. Total height and tree diameter at 1.3 m above ground were measured between age 6 and 15 years. A meta-analysis consisting of 39 progeny trials, with more than 300,000 data points, was carried out to estimate breeding values (EBV) for stem sweep at age 8 years, total height and tree diameter at age 12 years using the Treeplan genetic evaluation system [47]. For the present study, 184 unrelated founders (G0 trees) and 477 G1 trees were genotyped (Fig. S1). Among the 477 G1 selections, 355 selections have both parents identified in the G0 population. The 122 remaining selections have only their mother identified in the G0 population as they were selected in open-pollinated progeny trials. In total, there were 191 maternal half-sib families in this G1 population. The number of individuals per half-sib family ranged from 1 to 13 with an average of 2.5 individuals per half-sib family. Inbreeding coefficients were equal to zero in the two-generation breeding population because it was comprised of unrelated founders and their offspring generation.

We used EBV as pseudo phenotypes in genomic predictions. All EBV were based on progeny test data and pedigree derived additive genetic relationships with high accuracies, ranging from 0.67 to 0.99. By definition, the accuracy of EBV is the correlation between the true breeding values and the EBV [48]. The accuracy r is estimated as $r = \sqrt{1 - \left(\frac{S^2}{1+F\sigma_A^2}\right)}$ where S is the standard error of the EBV, F is the coefficient of inbreeding and σ_A^2 is the additive genetic variance [49]. EBV for total height and tree diameter were highly correlated, whereas EBV of these two traits had weak or no correlation with EBV of stem sweep (Fig. 1). Although the range of accuracies was not large, using EBV as phenotypes in genomic prediction may introduce bias and heterogeneity [50]. We then compared EBV and the de-regressed breeding values (dEBV) as pseudo phenotypes to estimate the effect on reliability of genomic predictions. The dEBV for individual i was obtained as $\hat{u}_i^* = \hat{u}_i / r_i$, where $\hat{u}_i$ is the EBV and r_i is the accuracy of EBV [50]. The resulting de-regressed breeding values were then weighted according to $w_i = \left(1 - h^2\right) / \left[c + (1 - r_i)/r_i\right] h^2$, where h^2 is the heritability of the trait and c is the proportion of variance not accounted for by the markers (assumed to be 50%) [50,51].

2.2. Genotyping and LD analysis

We used a 12K Infinium SNP array (Illumina Inc., San Diego, CA, USA) described by [52] to genotype 661 trees. One-year old pine needles (diploid tissue) were harvested to extract DNA. The Infinium assay was used to recover 2600 informative markers from the G0 population. In this study the same markers were assayed in the G1 population. Missing genotypic data points (3265 or 0.19%) were imputed from the marginal allele distribution for each marker. In other words, missing genotypes were sampled from scored genotypes (0, 1 and 2) assuming the population was in Hardy–Weinberg equilibrium. Markers with minor allele frequency (MAF) below 5% were discarded (100 out of 2600). In total, 2500 markers were used for genomic prediction and model comparison. A high level of heterozygosity was found in the population with an average value of 0.39 (±0.02) over all individuals (Fig. S2).

LPmerge software [53] was used to produce a composite genetic linkage map based on five published [54–56] and two unpublished (kindly provided by MT Cervera) linkage maps. Markers were assigned to a genetic map position to analyze the extent of LD along 12 maritime pine chromosomes (Fig. S3). Intra-chromosomal

Please cite this article in press as: F. Isik, et al., Genomic selection in maritime pine, Plant Sci. (2015), http://dx.doi.org/10.1016/j.plantsci.2015.08.006