Genomic prediction of growth traits in *Pinus taeda* using genome-wide sequence-based DArT-seq markers

Ananda Virginia de Aguiar[1], Dione Mendes Alves Teixeira-Freitas[2], Janeo Eustaquio de Almeida Filho[3], Valderês Aparecida de Sousa[4], Marcos Deon Villela Resende[5], Orzenil Bonfim Silva-Junior[6], Dario Grattapaglia[7]

[1,4,5]Embrapa Florestas - Estrada da Ribeira, Km 111 - 83411-000 - Colombo, PR (ananda.aguiar@embrapa.br)
[2,6,7]Embrapa Recursos Genéticos e Biotecnologia Empresa Brasileira de Pesquisa Agropecuária – Brasília, Brazil
[3] Departamento de Genética, Universidade Federal de Viçosa

Background

Commercial loblolly pine (*Pinus taeda* L.) plantations in Brazil supply raw material for a large array of wood based industries and businesses.The increasing local pine wood demand has encouraged development of seed production programs by private companies as well by public institutions. Continuous supplies of genetically improved seed have enhanced productivity and turned intensive pine silviculture into a sustainable and attractive activity. However, pine breeding cycles of 15 to 25 years are long even in tropical and subtropical regions of Brazil. Furthermore, breeding decisions are traditionally made using only growth traits in most programs due to the challenges in measuring wood quality traits in large progeny trials. This limits the genetic gains that could be obtained if additional wood quality traits were contemplated.

With the advent of more accessible DNA marker technologies it is now economically and logistically possible to consider applying molecular breeding to pine breeding programs in order to select indirectly for multiple traits. Despite of the advances made in QTL mapping and association genetics in loblolly pine [1, 2], no effective application to tree breeding has been achieved to date. Genomic Selection (GS) promises to bridge this gap by predicting phenotypic performance based on a genome-wide panel of markers whose effects on the phenotype are estimated simultaneously in a large and representative 'training' population of individuals without applying rigorous significance tests [3]. This approach has been proposed as a way to shorten tree breeding cycles and allow indirect selection while increasing selection intensity for all traits simultaneously, consequently enhancing genetic gains per unit time [4]. The power demonstration of this approach was pioneered in loblolly pine [5] and *Eucalyptus* [6] soon followed by other studies in loblolly pine [7] and other conifers [8]. All these studies have shown positive prospects for the adoption of GS in tree breeding.

In loblolly pine, however, due to the complexity of its genome, there are still some regarding the development of a standard large-scale marker genotyping platform. Although SNP chips have been developed [9] and used for GS experiments [5] these were limited in genomic coverage to genic regions and not readily accessible to the breeder´s community. Sequence based genotyping methods based on genome complexity reduction have been successfully used in loblolly pine. Genotyping by exome sequence capture was used to build a linkage map covering 2,841 genes,. A linkage map with 2,469 DArT-seq markers [10] was used to anchor a portion of the loblolly pine genome assembly [11], and the standard GBS protocol was used to generate ~17,000 markers in a small set of *Pinus contorta* accessions [12]. In this work we used 16,355 sequence-based DArT-Seq markers covering both genic and non-genic regions to genotype a genomic selection in a training population of 960 loblolly pine trees and to develop predictive models for growth traits.

Material and methods

Population, genotyping and phenotyping. Phenotype and genotype data were obtained for 960 trees of a progeny trial established in Ponta Grossa (PR) (25°25'S 49°15'W) involving 35 maternal open pollinated families of elite trees from a clonal seed orchard. This population constitutes the breeding population of EMBRAPA Forestry. Assessments included total height starting at age 2, and every year thereafter until age 6 and DBH (stem diameter at breast height) from age 4 to 6. DBH and height data were used to estimate tree volume. DNA extractions were carried out from pine needles and total genomic DNA sent to DArT Pty (Yarralumla, Australia) for DArT-seq genotyping. DArT-seq in *P. taeda* was carried out by complexity reduction using double digest with PstI_ad/TaqI/HhaI_ad. PstI adapter was tagged with 96 different barcodes enabling multiplexed sequencing into a single lane. FASTQ files were quality filtered at ≥ 90 % confidence and for the presence of barcode sequences. All 960 DNA samples were genotyped in duplicate in separate lanes providing fully replicated sequencing data with minimal lane effect. Sequences were aligned against a pseudo-reference consensus sequence created by assembling all the sequence reads generated for all the DNA samples at each locus. The output files from the alignment generated using BWA were processed using an analytical pipeline developed by DArT Pty to produce tables for the dominant PAV markers and co-dominant SNPs.

Genomic predictions. BLUP/REML was the mixed linear model used to analyze phenotype data. All phenotypes were corrected for fixed effects and deregressed for parent effects and genomic values (GEBV) predicted using Random Regression-Best Linear Unbiased Predictor using the RR-

BLUP R package [13]. The linear mixed model was adjusted for the estimation of marker effects. A Jack-knife cross-validation method [14] was used to estimate the prediction accuracy.

Results and discussion

A total of 16,355 DArT-seq markers were obtained and used in the analyses. Using all markers, predictive accuracies from cross validation varied between 0.51 at age 2 up to 0.63 at age 6 for DBH, height and volume traits. Such predictive accuracies were obtained relating individuals in the training and validation subsets. Maximal accuracies ranging between 0.78 and 0.93 were obtained when only 20 % (3,271) of the top markers ranked by effect were used. This result indicates that the predictive models are likely capturing relatedness. Furthermore, these results suggest that fewer markers could be tentatively used for the practice of GS, if the candidate selection population consists of progeny of the training population. However, while such marker selection might work well in such conditions, models developed from a selected set of markers based on their effect might suffer from over-fitting and therefore display a reduction in prediction accuracy in later generations [15]. In *Picea glauca*, removing uninformative loci from predictive models also seemed to have a positive effect on cross-validation, with a slight increase in accuracy [16]. Like in our study, this increase could be partly accounted because these marker subsets were not identified in an independent dataset [15] upwardly biasing accuracy estimates. In fact, the linear regression coefficients between the observed and predicted values were greater than 1, indicating that the marker effect was biased. In conclusion, these results show that DArT-seq markers provide a useful platform genotyping of loblolly pine providing large numbers of high quality markers. In line with our previous reports in loblolly pine and eucalyptus [5, 6], these results further support that genomic prediction might soon become an operational tool in the practice of tree breeding. However caution should be taken when developing prediction models to avoid over fitting that might mislead the perspectives of potential gain of this method especially as one intends to predict in generations removed from the training set.

References

1. Neale DB: **Genomics to tree breeding and forest health**. *Curr Opin Genet Dev* 2007, **17**(6):539-544.
2. Neale DB, Kremer A: **Forest tree genomics: growing resources and applications**. *Nature Reviews Genetics* 2011, **12**(2):111-122.
3. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**(4):1819-1829.

4. Grattapaglia D: **Breeding forest trees by Genomic Selection: current progress and the way forward. Chapter 26.** In: *Advances in Genomics of Plant Genetic Resources.* Edited by Tuberosa R, Graner, A., Frison, E. New York: Springer; 2014: 652-682.
5. Resende MFR, Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M: **Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments**. *New Phytologist* 2012, **193**(3):617-624.
6. Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA *et al*: **Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees**. *New Phytologist* 2012, **194**(1):116-128.
7. Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R: **SNP markers trace familial linkages in a cloned population of *Pinus taeda* - prospects for genomic selection**. *Tree Genetics & Genomes* 2012:1-12.
8. Beaulieu J, Doerksen T, Clement S, Mackay J, Bousquet J: **Accuracy of genomic selection models in a large population of open-pollinated families in white spruce**. *Heredity* 2014, **doi:10.1038/hdy.2014.36**.
9. Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, Neale DB: **Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* l., *Pinaceae*)**. *Genetics* 2010, **185**(3):969-982.
10. Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A: **Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus***. *BMC Proceedings* 2011, **5**(Suppl 7):P54.
11. Alves-Freitas D, Silva-Junior OB, Kilian A, Grattapaglia D: **Anchoring 323 Mbp of the Pinus taeda Genome Assembly v1.01 to a Single-Tree Sub-Centimorgan Genetic Map of 2,469 DArT-Seq Markers and Microsatellites**. In: *Plant Animal Genomex Conference XXII.* vol. P493. San Diego; 2014.
12. Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA: **Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform**. *Tree Genetics & Genomes* 2013, **9**(6):1537-1544.
13. Endelman JB: **Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP**. *Plant Genome-Us* 2011, **4**(3):250-255.
14. Quenouille MH: **Approximate Tests of Correlation in Time-Series**. *J Roy Stat Soc B* 1949, **11**(1):68-84.
15. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: **Pitfalls of predicting complex traits from SNPs**. *Nature Reviews Genetics* 2013, **14**(7):507-515.
16. Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J: **Genomic selection accuracies within and between environments and small breeding groups in white spruce**. *BMC Genomics* 2014, **15**.