A species-diagnostic SNP panel for discriminating lodgepole pine, jack pine, and their interspecific hybrids

Cullingham CI, Cooke JEK, Dang S, Coltman DW

Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

Corresponding author: Catherine Cullingham, cathy.cullingham@ualberta.ca, (780) 492-8368

Abstract

Accurate stock identification is important for forest management, yet this can be a challenge for tree species that hybridize naturally. Species discriminating molecular markers provide a means to identify stock with high accuracy. In Canada, lodgepole pine (*Pinus contorta* Dougl. ex Loud. var. *latifolia*) and jack pine (*P. banksiana* Lamb) form a large hybrid zone in Alberta and Northwest Territories; within this hybrid zone, the identification of parentals and hybrids is difficult due to an overlap in morphological characteristics. Pure and hybrid ancestry can be resolved using microsatellite markers, but these are difficult and costly to type. We have developed a panel of SNP markers using 454 transcriptome sequence data that are more cost effective, easier to score and have greater discriminating power for differentiating species than microsatellites. Our SNP panel provides accurate and cost efficient forest seed stock identification and will thereby facilitate reforestation and our pipeline can be applied to other hybrid systems globally.

Keywords

jack pine, lodgepole pine, *Pinus banksiana*, *Pinus contorta,* seed stock, SNP

Introduction

Significant effort is invested in the selection and improvement of seed stock for forest management in Canada and globally (Bucci and Vendramin 2000; Ying and Yanchuk 2006; Hamann et al. 2011). The proper identification of stock for future plantings is critical in this process, and is also important for conducting biological studies (e.g. Yang et al. 1999). Most stock identification for collecting seed is conducted using morphological characteristics (Boland et al. 2006; Ying and Yanchuk 2006; Sarmiento et al. 2011). However, there are many instances where forest tree species hybridize naturally, requiring the use of molecular and genetic methods to identify lineages (*Picea*, Bennuah et al. 2010; *Populus*, Hamzeh et al. 2007; Lexer et al. 2010; Meirmans et al. 2010; *Quercus* Burgarella et al. 2009; Ortego and Bonal 2010. Lodgepole pine (*Pinus contorta* Dougl. ex Loud. var. *latifolia*) and jack pine (*P. banksiana* Lamb) are two North American species of economic and ecological importance that hybridize readily, resulting in a hybrid zone where their ranges overlap in Alberta and the Northwest Territories. While these species are relatively easy to distinguish using morphological characters (Wheeler and Guries 1987), the hybrid zone presents a mosaic of morphological types where distinction between parental and hybrid is blurred (Zavarin et al. 1969; Pollack and Dancik 1985; Rweyongeza et al. 2007, Bleiker and Carroll 2011). To distinguish between these groups microsatellite markers have been developed and used to positively identify pure jack pine attacked by mountain pine beetle (*Dendroctonus ponderosae* Hopkins; Cullingham et al. 2011).

Simulation and empirical studies have suggested that microsatellites are more informative than SNPs based on the hypothesis that the number of alleles is a good predictor of assignment power (Kalinowski 2004; Winans et al. 2004, Narum et al. 2008). However, in their development of a measure of marker informativeness for admixture analysis, Rosenberg et al. (2003) found the potential for SNPs, if selected carefully, to have equivalent resolution to microsatellites. In the last decade, this approach has been employed and highly informative SNP marker sets have been developed to distinguish lineages for a number of species, including grape wine cultivars (*Vitus vinifera* L.; Cabezas et al. 2011), humans (*Homo sapiens*; Lao et al. 2008), and chum salmon (*Onocorhynchus keta*; Smith and Seeb 2008). As well, with the advent of next-generation sequencing access to thousands of SNPs allows us to identify multiple diagnostic or near-diagnostic markers. Therefore, through the careful selection of SNP loci, we can obtain high discriminating power to differentiate species class, with a small set of loci, thereby reducing cost and increasing efficiency. SNPs have additional technical advantages over microsatellites: they are more cost-effective to score where SNPs can cost a few cents to >$1 per genotype and microsatellites can cost upwards of $5 per genotype (Morin et al. 2004, Guichoux et al. 2011), have lower error rates (Hoffman and Amos 2005) and are transferrable across platforms and can therefore be used by different laboratories (Baric et al. 2008).

Our objective was to develop a robust, cost-effective SNP panel that can be used to effectively distinguish lodgepole pine, jack pine and their interspecific hybrids. Using 454 transcriptome sequence data generated for both lodgepole and jack pine, we selected potential SNPs *in silico* that showed fixed differences between lodgepole and jack pine, and following wet-lab validation, selected 26 that were either fully discriminating or had large frequency differences. We genotyped 921 lodgepole, jack and hybrid pine individuals that had been previously genotyped at 10 microsatellite loci (Cullingham et al. 2011, 2012). We then compared the assignment of each of these datasets using the empirical data and simulated data to assess the utility of the SNP panel. Across a hybrid zone the genome will have variable levels of introgression (e.g. Martinsen et al. 2001). Thus we looked at the amount of introgression among the markers to identify the most efficient panel for species discrimination.

Methods

Plant material and Sequencing

Tissue samples were prepared from xylem, bark, needles and roots of two year old seedlings subjected to six different treatments: control, water deficit by water withholding, or mechanical wounding at one, two, four or eight days prior to harvest. All plants were harvested on the same day, with tissues frozen immediately in liquid nitrogen. Samples were stored at -80 °C until processing. Total RNA extractions were carried out according to Chang et al. (1993), then treated with DNase I (New England Biolabs, Pickering, ON). RNA was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Mississauga, ON). RNA from 11 lodgepole and 11 jack pine were combined into separate pools for cDNA synthesis. cDNA Synthesis was carried out by Evrogen (Moscow, Russia) using the SMART approach (Zhu et al. 2001). cDNA were normalized using the DSN normalization method (Zhulidov et al. 2004), which included denaturation/association treatment by duplex-specific nuclease (Shagin et al. 2002) and amplification of normalized fraction by PCR. Next-generation sequence data was generated using a Roche FS-FLX Titanium with 20X coverage at the McGill University and Genome Quebec Innovation Centre (Montreal, QC).

SNP Discovery and Validation

Sequences were assembled using Newbler (Roche, Mississauga, ON). Three assemblies were generated, one for each of the species, and one for the species combined. We completed our SNP search on the combined assembly using CLC Genomics Workbench 4 (CLC Bio, Cambridge, MA) with the following parameters: maximum coverage = 25, maximum gap and mismatch count = 0, minimum average quality = 15, minimum central quality = 20, minimum coverage = 10, minimum variant frequency (MVF) = 10% and a window length = 61. Next, SNPs were selected that were invariant within species and variant between species with equal representation from both species.

Because transcriptome data were used for *in silico* SNP discovery, one would expect false positives to be generated by false alignment of paralogous sequences within a single contig. Given that species in the *Pinus* family have a high proportion of paralogs (Cui et al. 2006), we used re-sequencing for validation of SNPs identified *in silico*. From the list of potential species discriminating SNPs, 75 primer pairs were designed for amplification and re-sequencing using PrimerBLAST (Rozen and Skaletsky 2000) with the default parameters, except the minimum temperature (T_m) was set to 45^{o}C and the *Pinus* non-redundant database was used. The primers were screened using an initial panel of four lodgepole and four jack pine individuals.

Genomic DNA was extracted from megagametophytes, young seedlings and needles of mature trees as previously described in Cullingham et al. (2011, 2012). DNA concentrations were assessed using the Infinite 200 NanoQuant (Tecan, Mannedorf, Switzerland) and approximately 200ng of template DNA was used for PCR amplification. Reactions were performed in 20µl final volume and consisted of 1X thermopol buffer (New England BioLabs, Pickering, ON), 2.5mM $MgCl_2$, 200µM each dNTP, 0.25 µM each primer, and 1U Taq DNA polymerase (New England BioLabs, Pickering, ON). PCR amplification was completed using the following cycling parameters 94^{o}C for 2 min, followed by 35 cycles of 94^{o}C for 30 s, T_a (50 – 60 ^{o}C) for 45 s, and 72^{o}C for 60 s, and a final extension at 72^{o}C for 30 min. In some instances, PCR reactions were optimized by modifying the annealing temperature. PCR products were prepared for sequencing using QIAquick PCR purification kit (Qiagen, Mississauga, ON) and Big Dye terminator reaction followed by ethanol/EDTA/sodium acetate precipitation (v3.1, Applied Biosystems, Foster City, CA). Sequencing was carried out on an Applied Biosystems 3730 DNA analyzer and sequence analysis conducted using CLC Genomic Workbench.

SNP Genotyping

We selected 921 pine samples for SNP genotyping as a subset of ca. 1900 pine samples from British Columbia, Alberta, Saskatchewan and Ontario that were previously genotyped at 10 microsatellite loci (Cullingham et al. 2011, 2012). SNP genotyping was carried out at McGill University and Genome Quebec Innovation Centre using SEQUENOM® iPLEX® Gold technology (Elrich et al. 2005).

Hybrid Identification

We estimated the proportion of ancestry among samples using both NEWHYBRIDS 1.1 beta (Anderson and Thompson 2002) and STRUCTURE 2.3.1 (Pritchard et al. 2000; Falush et al. 2003, 2007) using a similar approach used previously (Cullingham et al. 2011, 2012). For both programs we used a burn-in of 50 000 and 500 000 Markov chain Monte Carlo (MCMC) sweeps for data collection. For STRUCTURE *K* was set equal to 2. We ran five iterations for both programs to ensure consistent results.

Marker Resolution

To determine whether there was a difference in the discriminating abilities of the two marker sets (SNPs and microsatellites) we generated simulated data-sets using HYBRIDLAB ver. 1.0 (Nielsen et al. 2006). We generated five separate data sets of 450 lodgepole pine, 150 jack pine and 20 of each of the following hybrid classes: F1, F2, jack pine backcross, lodgepole pine backcross, jack pine F2 cross, lodgepole pine F2 cross and double backcrosses. These were chosen to reflect the composition of the actual data (see Results). To determine the effects of marker type (microsatellite, SNP) on the accuracy of species class assignment, we used the accuracy values from the simulated datasets to conduct a linear mixed effect model analysis using the package lme4 (http://lme4.r-forge.r-project.org/) in R 2.14.2 (R Core Development Team 2011), where the fixed effects were marker (microsatellite, SNP), program (NEWHYBRID, STRUCTURE) and species (lodgepole, jack, and the eight hybrid crosses) and the random effect was the simulation set. To assess the significance of the effects we used the 'pval.fnc' function to conduct a Markov Chain Monte Carlo simulation in the languageR package (http://cran.r-project.org/web/packages/languageR/index.html).

To optimize the SNP panel, we used INTROGRESS (Gompert and Buerkle 2010) to determine the informativeness of each locus. This program assigns alleles to each parental class and estimates a hybrid index for each individual. For this analysis we removed all samples with missing data and used only lodgepole and jack pine that assigned to their class with ≥ 0.95 probability resulting in 88 jack pine, 376 lodgepole pine and 157 hybrids analyzed. Using this information we conducted the classification in STRUCTURE iteratively by systematically removing each locus starting with the least informative/most introgressed to determine the minimum number of loci required to resolve species classes.

Results

FS-FLX Titanium sequencing yielded 1 598 694 reads for lodgepole pine and 1 559 458 reads for jack pine, with average read lengths of 358 and 346, respectively. Following trimming and removal of poor-quality reads according to Huse et al. (2007), 1 597 295 and 1 558 772 reads were retained for lodgepole pine and jack pine, respectively. These data are archived in the NCBI Short Read Archive (SRP004517). We assembled 93 364 contigs using the combined reads. Our SNP search parameters discovered 96 059 SNPs and of these 16 959 showed fixed differences between the species. Using only SNPs with the highest coverage (20X) and equal contribution from both species (45-50% MVF) resulted in 75 contigs for primer design.

Of the 75 primer pairs that were tested, 23 primer pairs, flanking 26 SNPs, resulted in PCR products with potential species specific polymorphisms. These were sequenced for an additional 32 individuals of each species using a DNA pooling approach (Pelgas et al. 2004) where eight pools, each with four individuals, were created using an equal amount of DNA from each individual. Amplicons generated from megagametophytes for each primer pair were also sequenced to verify that the SNP variant did not result from paralogs. Based on the additional sequencing, nine SNPs were species specific, 16 SNPs were shared polymorphisms with large frequency differences between the two species, and the last amplified region was a paralog. These 25 SNPs (Table 1) were used to genotype 921 individuals using Sequenom technology (contig sequence accessions: KC411636-KC411658).

Of the 25 SNPs that proceeded to genotyping, three had high failure rates and were removed from further analyses. Of the remaining 22 SNPs, two were tightly linked and six were of chloroplast origin, resulting in 14 SNPs for the hybrid discrimination analysis. We obtained complete or near complete genotypes for 822 of the 921 individuals as a result of sample quality/quantity. Using this SNP marker set together with the microsatellite marker set and assignment criteria developed in Cullingham et al. (2011) for this system based on the outputs from NEWHYBRID and STRUCTURE, 463 individuals assigned to lodgepole pine, 154 to jack pine and the remainder to hybrids (205). The six chloroplast loci resulted in one haplotype for lodgepole pine and one for jack pine. Of the 205 hybrids, 175 and 30 exhibited the lodgepole and jack pine haplotype, respectively. The majority of assignments were supported by the SNP markers (97%), while only 87% were supported using the microsatellite markers alone.

Marker resolution

The assignment accuracies among 10 different classes of individuals for the three marker sets using two different programs are summarized in Figure 1. SNP data for all of the classes has at least as high species discriminating power as the microsatellite data. The mixed effect model results (Table 2) indicate reduced accuracy of assignment with the microsatellites. There is also an effect of the program: use of STRUCTURE results in reduced accuracy of assignment compared to NEWHYBRID. Species class has the largest effect on assignment accuracy, where accuracy is decreased for the third generation backcrosses (JpJpBC, LpLpBC, JpF2 and LpF2).

In identifying the optimal SNP panel, we found a large difference in the degree of introgression among the SNPs (Figure 2). Considerably more SNPs had greater lodgepole introgression into jack pine than we observed with the microsatellites. The order of locus removal is indicated in Table 3 where the effect on the assignment accuracy of removing one locus systematically from the SNP panel is presented. Here we found consistent levels of assignment accuracies with 7-14 SNPs included in the panel with minor decreases in the second and third generation back-crosses. However, we saw decreasing accuracies among all classes when less than seven SNPs were used.

Discussion

Through the careful selection of SNP markers, we have been able to identify a panel that has greater discriminating power than microsatellites in resolving the ancestry among lodgepole pine, jack pine and their interspecific hybrids. Using an alignment of lodgepole and jack pine 454 transcriptome data we were able to identify SNPs that were either fully discriminating or had large frequency differences between the species. These markers have several practical applications, including identification of appropriate seed stock for future forest generations using a marker panel that is highly accurate, easy to interpret and transferable to other analytical platforms.

The SNPs were better able to resolve the species and their hybrids, based on the analysis of the accuracy of the simulated data (Figure 1) and a comparison to the combined (microsatellites and SNPs) empirical data. For the empirical data, there were only a few discrepancies between the microsatellite/SNP set and the SNPs alone (3%) and over four times more between the microsatellites and the microsatellite/SNP set (13%). The improvement afforded by the SNPs relative to the microsatellites can be attributed to differences in the number of shared alleles exhibited by the two marker sets. The microsatellites presented a considerable number of shared alleles between jack and lodgepole pine (Figure 2). In contrast, many SNPs were fully discriminating between species, and SNPs that were not fully discriminating exhibited large frequency differences contributing to the resolving power of the SNP dataset (Figure 2). Greater informativeness of SNPs over microsatellites has also been observed for other systems despite the greater allelic diversity of microsatellites (Liu et al. 2005; Smith and Seeb 2008)

Further evidence of the SNP performance is derived from the comparative analysis of admixture results using simulated data across the different marker sets and programs. These data revealed that all three elements (program, marker set and admixture level) affect the accuracy of assignment. For the marker sets, SNPs outperformed microsatellites for the most part, but there was an interaction with the species class (Table 2) where the two markers performed similarly for the jack pine hybrid crosses (Figure 1). This is because the degree of introgression of lodgepole pine into jack pine for the SNPs is greater than for the microsatellites (Figure 2), which will result in hybrids that are genetically very similar to jack pine, thereby affecting resolution (Lexer et al.. 2007). We also found an effect of the program used for analysis, and an interaction with the species class (Table 2), in which NEWHYBRIDS outperformed STRUCTURE in the hybrid classes, but STRUCTURE outperformed for the parental classes. Differences in these programs have been documented previously for other hybrid systems (Vähä and Primmer 2006; Burgarella et al. 2009; Quintela et al. 2010).

To develop an optimal panel that balances accuracy with cost effectiveness, we used an analysis of introgression to identify the extent of shared polymorphism for each locus, then removed loci from the classification analysis starting with the most introgressed and therefore least informative (Table 3, Figure 2). This analysis highlights the importance of using near-diagnostic loci as only four are fully discriminating and based on the assignment accuracy anywhere from 7-13 loci would be needed to ensure a discriminatory level similar to the full SNP panel for most categories. For identification of the double back-cross class, no fewer than 10 loci should be maintained. Ten loci can easily be accommodated in a single multiplex reaction, for instance by using the Snapshot® Multiplex Kit (Applied Biosystems) resulting in rapid analysis of unknown seed stock for ancestry identification using readily accessible instrumentation.

There are several practical applications for these markers. Accurate identification of ancestry for seed stock using conventional approaches is challenging for the lodgepole × jack pine hybrid system given the close proximity of pure and hybrid individuals. The development of this SNP marker panel will allow for easy, reliable and rapid identification of ancestry which is a relevant concern as forest managers are faced with difficult challenges that include mitigating the impact of climate change on forest distributions (Gray and Hamann 2011). Reliable distribution maps for these species are also necessary to develop accurate models (Coops et al. 2012) that can investigate the invasion of a new pest species in jack pine (mountain pine beetle; Cullingham et al. 2011).

Acknowledgements

The authors would like to thank Leonard Barnhardt and Deogratias Rweyongeza (Environment and Sustainable Resource Development, Government of Alberta) for their expertise and access to samples;

Stephanie Boychuck, Joël Fillon, Dominik Royko (University of Alberta) and staff members at the McGill University and Genome Quebec Innovation Centre for laboratory and bioinformatic support; and Matt Bryman (University of Alberta) for logistics. We would also like to thank comments from Sally Aitken and two anonymous reviewers for improving the manuscript. We acknowledge funding for this research from the Government of Alberta (AAET/AFRI-859-G07), as well as grants from Genome Canada, the Government of Alberta through Genome Alberta, and Genome British Columbia in support of the Tria I and Tria II projects (http://www.thetriaproject.ca) of which J.E.K.C, D.W.C. are principle investigators.

References

Anderson E, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160:1217-1229.

Baric S, Monschein S, Hofer M, Grill D, Dalla Via J (2008) Comparability of genotyping data obtained by different procedures an inter-laboratory survey. J Hortic Sci Biotech 83:183-190

Bennuah SY, Wang T, Aitken SN (2004) Genetic analysis of the *Picea sitchensis* × *glauca* introgression zone in British Columbia. Forest Ecol Manag 197:65-77

Bleiker KP, Carroll AL (2011) Rating introgression between lodgepole and jack pine at the individual tree level using morphological traits. North J Appl For 28:138-145

Boland DJ, Brooker GM, Chippendale N et al. (2006) *Forest trees of Australia*, 5th Ed. CSIRO Publishing, Collingwood, VIC

Bucci G, Vendramin GG (2000) Delineation of genetics zones in the European Norway spruce natural range: preliminary evidence. Mol Ecol 9:923-934

Burgarella C, Lorenzo Z, Jabbour-Zahad R et al. (2009) Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). Heredity 102:442-452.

Cabezas JA, Ibáñez J, Kijavetski D et al. (2011) A 48 SNP set for grapevine cultivar identification. BMC Plant Biol 11:153

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep 11: 113-116

Coops N, Wulder MA, Waring RH (2012) Modeling lodgepole and jack pine vulnerability to mountain pine beetle expansion into the western Canadian boreal forest. Forest Ecol Manag 274:161-171

Cui L, Wall K, Leebens-Mack JH, Lindsay BG et al. (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738-749

Cullingham CI, James PMA, Cooke JECK, Coltman DW (2012) Characterizing the physical and genetic structure of the lodgepole pine × jack pine hybrid zone: mosaic structure and differential introgression. Evol Appl doi:10.1111/j.1752-4571.2012.00266.x

Cullingham CI, Cooke JEK, Dang S, Davis CS, Cooke BJ, Coltman DW (2011) Mountain pine beetle host-range expansion threatens the boreal forest. Mol Ecol 20:2157-2171

Ehrich M, Bocker S, van den Boom D (2005) Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. Nucleic Acids Res 33:e38

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574-578

Gompert Z, Buerkle CA (2010) INTROGRESS: a software package for mapping components of isolation in hybrids. Molecular Ecology Resour 10:378-384

Gray LK, Hamann A (2011) Strategies for reforestation under uncertain future climates: guidelines for Alberta, Canada. PLoS ONE 6:e22977

Guichoux E, Lagache L, Wagner S et al. (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591-611.

Hamann A, Gylander T, Chen P-Y (2011) Developing seed zones and transfer guidelines with multivariate regression trees. Tree Genet Genom 7:399-408.

Hamzeh M, Sawchyn C, Périnet P, Dayanandan S (2007) Asymmetrical natural hybridization between *Populus deltoids* and *P. balsamifera* (*Salicaceae*). Can J Bot 85:1227-1232

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Mol Ecol 14:599-612

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:RI 43

Kalinowski ST (2004) Genetic polymorphism and mixed-stock fisheries analysis. Can J Fish Aquat Sci 61:1075-1082

Lao O, van Duijin K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78:680-690

Lexer C, Joseph JA, van Loo M et al. (2010) Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. Genetics 186:699-712

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF (2007) Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. Heredity 98:74-84

Liu N, Chen L, Wang S, Oh C, Z H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet 6:S26

Martinsen GD, Whitham TG, Turek RJ, Kleim P (2001) Hybrid populations selectively filter gene introgression between species. Evolution 55:1325-1335.

Meirmans PG, Lamothe M, Gros-Louis M-C, Khasa D, Périnet P, Bousquet J, Isabel N (2010) Complex patterns of hybridization between exotic and native North American poplar species. Am J Bot 97:1688-1697

Morin PA, Luikart G, Wayne RK et al. (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208-216

Narum SR, Banks M, TD Beacham et al. (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. Mol Ecol 17:3464-3477

Nielsen EEG, Arvebach L, Kotlicki P (2006) HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. Mol Ecol Notes 6:971-973

Orgego J, Bonal R (2010) Natural hybridisation between kermes (*Quercus coccifera* L.) and holm oaks (*Q. ilex* L.) revealed by microsatellite markers. Plant Biol 12:234-238

Pelgas B, Isabel N, Bousquet J (2004) Efficient screening for expressed sequence tag polymorphisms (ESTPs) by DNA pool sequencing and denaturing gradient gel electrophoresis (DGGE) in spruces. Mol Breeding 13:263-279

Pollack JC, Dancik BP (1985) Monoterpene and morphological variation and hybridization of *Pinus contorta* and *P. banksiana* in Alberta. Canadian Journal of Botany 63:201-210

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Quintela M, Thulin C-G, Höglund J (2010) Detecting hybridization between willow grouse (*Lagopus lagopus*) and rock ptarmigan (*L. muta*) in central Sweden through Bayesian admixture analyses and mtDNA screening. Conserv Genet 11:557-569

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422.

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386

Rweyongeza DM, Dhir NK, Barnhardt LK, Hansen C, Yang R-C (2007) Population differentiation of lodgepole pine (*Pinus contorta*) and jack pine (*Pinus banksiana*) complex in Alberta: growth, survival, and responses to climate. Can J Bot 85:545-556

Sarmiento C, Detienne P, Heinz C, Molino JF, Brard P, Bonnet P (2011) Pl@ntwood: a computer-assisted identification tool for 110 species of amazon trees based on wood anatomical features. IAWA Journal 32:221-232

Shagin DA, Rebrikov DV, Kozhemyako VB (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Res 12:1935-1942

Smith CT, Seeb LW (2008) Number of alleles as a predictor of the relative assignment accuracy of short tandem repeat (STR) and single-nucleotide-polymorphism (SNP) baselines for chum salmon. T Am Fish Soc 137:751-762

Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Mol Ecol 15:63-72

Wheeler NC, Guries RP (1987) A quantitative measure of introgression between lodgepole and jack pines. Can J Bot 65:1876-1885

Winans GA, Paquin MZ, Van Doomik DM et al. (2004) Genetic stock identification of steelhead in the Columbia River basin: an evaluation of different molecular markers. N Am J Fish Manage 24:672-685

Yang R-C, Ye Z, Hiratsuka Y (1999) Susceptibility of *Pinus contorta – Pinus banksiana* compled to *Endocronartium harknessii*: host-pathogen interactions. Can J Bot 77:1035-1043.

Ying CC, Yanchuk AD (2006) The development of British Columbia’s tree seed transfer guidelines: Purpose, concept, methodology and implementation. Forest Ecol Manage 227:1-13

Zavarin E, Critchfield WB, Snajberk K (1969) Turpentine composition of *Pinus contorta* x *Pinus banksiana* hybrids and hybrid derivatives. Can J Bot 47:1443-1453.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 30:892-897.

Zhulidov PA, Bogdanova EA, Shcheglov AS (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32:e37

Data Archiving Statement

All 454 read data for lodgepole and jack pine have been archived on the NCBI Short Read Archive (SRP004517) and this is cross-referenced on the European Bioinformatics Institute, European Nucleotide

Archive (http://www.ebi.ac.uk/ena/data/view/SRP004517), and the DNA Databank of Japan (http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP004517). Contig sequences for species discriminating SNPs are archived on NCBI (accessions: KC411636-KC411658). Microsatellite genotyping data is available from Dryad, DOI: 10.5061/dryad.456q26k3.

Table 1. Set of SNPs used to profile 921 lodgepole, jack and hybrid pine for species discrimination including variant, DNA type (Region), missing data/locus (Null), primer sequences, annealing temperature used for resequencing and annotation for the genomic SNPs based on TAIR (http://www.arabidopsis.org/) and GCAT (https://gydleweb.gydle.com/arborea/gcat/).

Locus	Variant	Region	Null	Forward primer	Reverse primer	Ta	Annotation
*C17954-P346	T/G	Chloroplast	0	TGGTCGAAATGTACAATGAAGA	ACGATTGAAACGACGGAAGA	60	photosynthetic electron transfer A
*C38148-P707	G/T	Chloroplast	0	GGGCTCAATGTGATAATTGCG	TTAATGGAAGAGATTCCGCG	60	acetyl-CoA carboxylase carboxyl transferase subunit
*C52254-P578	C/G	Chloroplast	2	TGGTTCCAGGAAGTAGAATCCATGT	TGGCCAATGCGCCAATTGGT	50	photosystem I PsaA/PsaB protein
*C54855-P218	C/T	Chloroplast	3	AGAATTCAAGTAGAAATATCGATCT	TTGGGTACAACTAGTGCACT	60	PETG
*C55014-P315	A/C	Chloroplast	0	ACGAACGCATATAAAGAGCTCCTCG	AGACGTCCTGCAATTTGGATCTGA	60	RNA polymerase subunit
*C85071-P216	A/C	Chloroplast	2	GGAGCGGTCATATCTAGCCA	TCCTTTATGATAAGTGGTCTTTGC	65	Ycf1 protein
C26372-P562	G/C	Genomic	8	GAGCAGCCTCTGCTAGTGAA	ACAAAGAACTAGCTCACTTGTAC	60	calcium-dependent lipid binding family protein
C35213-P325	C/T	Genomic	1	GCCAAGGGACCACACGCTCT	CCTTGACTTGCTAATTGTGATGGCA	65	eukaryotic aspartyl protease family protein
C39371-P429	A/G	Genomic	41	CACTTGCTGTTGGGTGGCTGT	GCCCAGCAGGATTAATGAACTCA	65	protein of unknown function (DUF3353)
C54523-P103	A/T	Genomic	1	AGAACTTTTGTACACCTGACAAACT	GCGAGGCATCTATCCATAGCTCA	60	translation protein SH3-like family
C55350-P439	C/T	Genomic	6	AGAGCTAAAGGAGTACAATTGTGCA	TCAGAGGACTCACTTGGTTCA	60	chaperone protein dnaJ-related
C55378-P723	T/G	Genomic	2	GAACGTGGTGGCTGTGGCAA	GTGCAGCTGGACAGTACAAGAAA	65	transcription factor jumonji domain-containing protein
C55401-P415	T/G	Genomic	0	TGACACTAATATCAGCAATGTGGCA	TGGCGCACTTTTCTGACCCA	60	transcribed locus
C63961-P710	C/T	Genomic	1	CGCTCATCAGTGGCTCTTCTGGT	GTGGACGATTCTCCTGGCGCT	65	
C64907-P190	A/C	Genomic	0	AGGTACCGCTCCAATTATTGTGT	GTCGGATGATTGCACCTCTA	60	thioredoxin superfamily protein
C66807-P512	C/T	Genomic	1	TAAAACTTCTAGTCACGCTG	TAGCCATCTCTATCATGACA	60	beta-amylase/glycosyl hydrolase family 14
C84852-P331	A/T	Genomic	17	ACCTAATGCAATCCCTTCACCTCC	GGACTCTGAACATGACAGGTCCACA	65	CRAL/TRIO domain/Sep14p-like phosphatidylinositol transfer protein
C85320-P102	C/G	Genomic	11	TGAGCGAACAAACACTTAGGGT	CCATTGCCCTGTGACTCCGT	65	DEK domain-containing chromatin associated protein
C85407-P1002	C/G	Genomic	16	ACGCTTTCTAGATACAGCATG	TTTATTTTATATTCACTCACGTCTT	60	embryo defective 2737
Lp-C45579-P117	C/G	Genomic	1				myb-like HTH transcriptional regulatory family protein
†C55378-P723-2	T/C	Genomic	43				transcription factor jumonji domain-containing protein
†C63961-P710-2	G/C	Genomic	1				
‡C85506-P364	C/T		NA	GCGGCAGGACATGTTGCGAG	TGCCTGCCAAGGCTCATGCG	65	transcribed locus
‡C85506-P364-2	C/T		NA				
‡C85506-P364-3	A/G		NA				

*****Chloroplast loci not included in the discriminating analyses

†Linked loci removed from analyses

‡Poor quality loci removed from all analyses

Table 2. Mixed effect model results to determine what factors (fixed effects) effect the accuracy of class assignment where the simulated dataset was the random effect. The third generation hybrid crosses (lodgepole and jack pine double backcrosses) were the least accurately assigned. As well, the program (ProgramST: STRUCTURE) and marker type (Markerµsat: microsatellite) resulted in reduced species assignment accuracy. Fixed effects were: species (classes of parentals: Jack, Lodge, and hybrids: F2, JpBC (jack pine backcross), JpF2 (jack pine F2 cross), JpJpBC (jack pine double backcross), LpBC (lodgepole pine backcross), LpF2 (lodgepole pine F2 cross), LpLpBC (lodgepole pine double backcross)), marker (microsatellite and SNPs) and program (NEWHYBRID and STRUCTURE)

Random effects				
	Groups	Name	Variance	Std.Dev.
	DataSet	(Intercept)	2.10E-13	4.58E-07
	Residual	6.62E-03	8.14E-02	

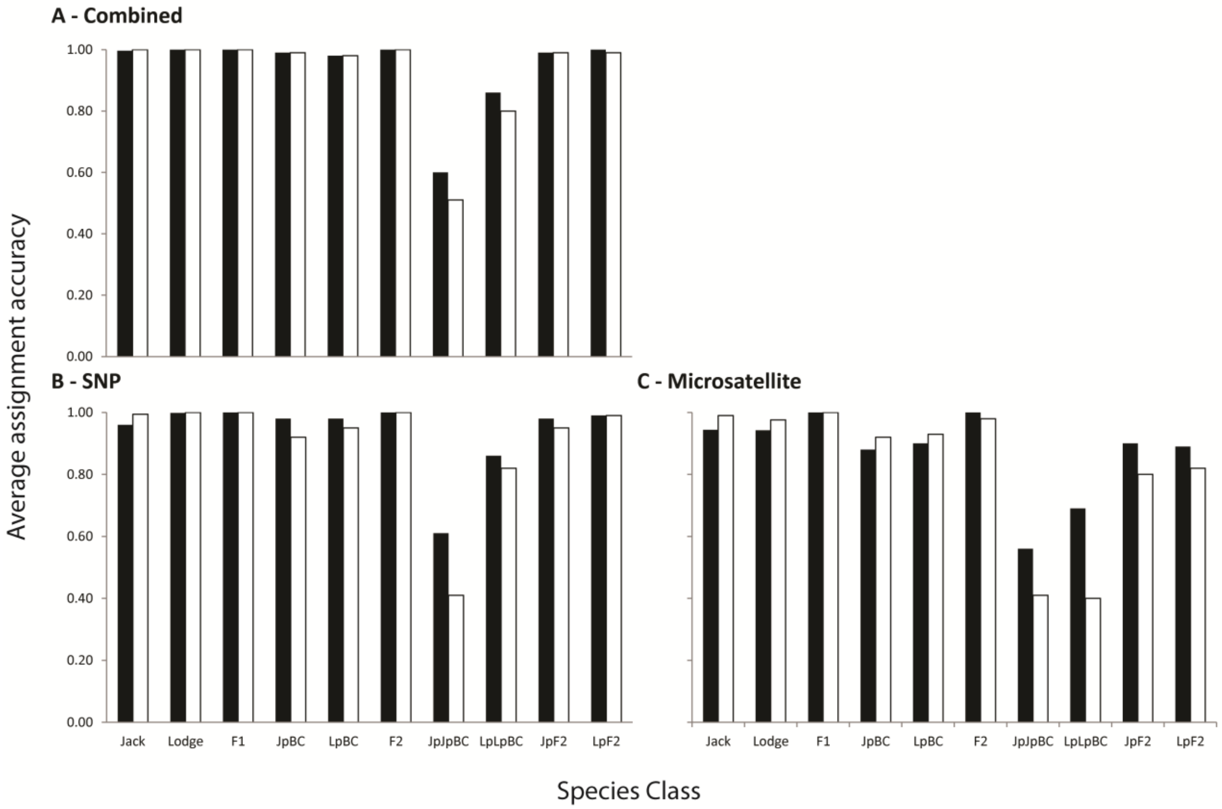
Fixed effects				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0451	0.01994	52.42	0.000
ProgramST	-0.0287	0.01151	-2.49	0.014
F2	-0.0050	0.02574	-0.19	0.846
Jack	-0.0280	0.02574	-1.09	0.278
JpBC	-0.0750	0.02574	-2.91	0.004
JpF2	-0.0925	0.02574	-3.59	0.000
JpJpBC	-0.5025	0.02574	-19.52	0.000
Lodge	-0.0210	0.02574	-0.82	0.416
LpBC	-0.0600	0.02574	-2.33	0.021
LpF2	-0.0775	0.02574	-3.01	0.003
LpLpBC	-0.2500	0.02574	-9.71	0.000
Markerµsat	-0.0615	0.01151	-5.34	0.000

Table 3. Species assignment accuracies (Jack: jack pine, Lodge: lodgepole pine, F1: jack × lodgepole, F2: F1 × F1, LpF2: lodgepole × F2, JpF2: jack × F2, LpBC: lodgepole × F1, JpBC: jack × F1, LpLpBC: lodgepole × LpBC, JpJpBC: jack × JpBC) estimated using simulated SNP data analyzed in STRUCTURE, following the systematic removal of the most informative loci based on the level of introgression (Figure 2). Highlighted rows indicate where significant decreases in assignment accuracy across the majority of classes occurred.

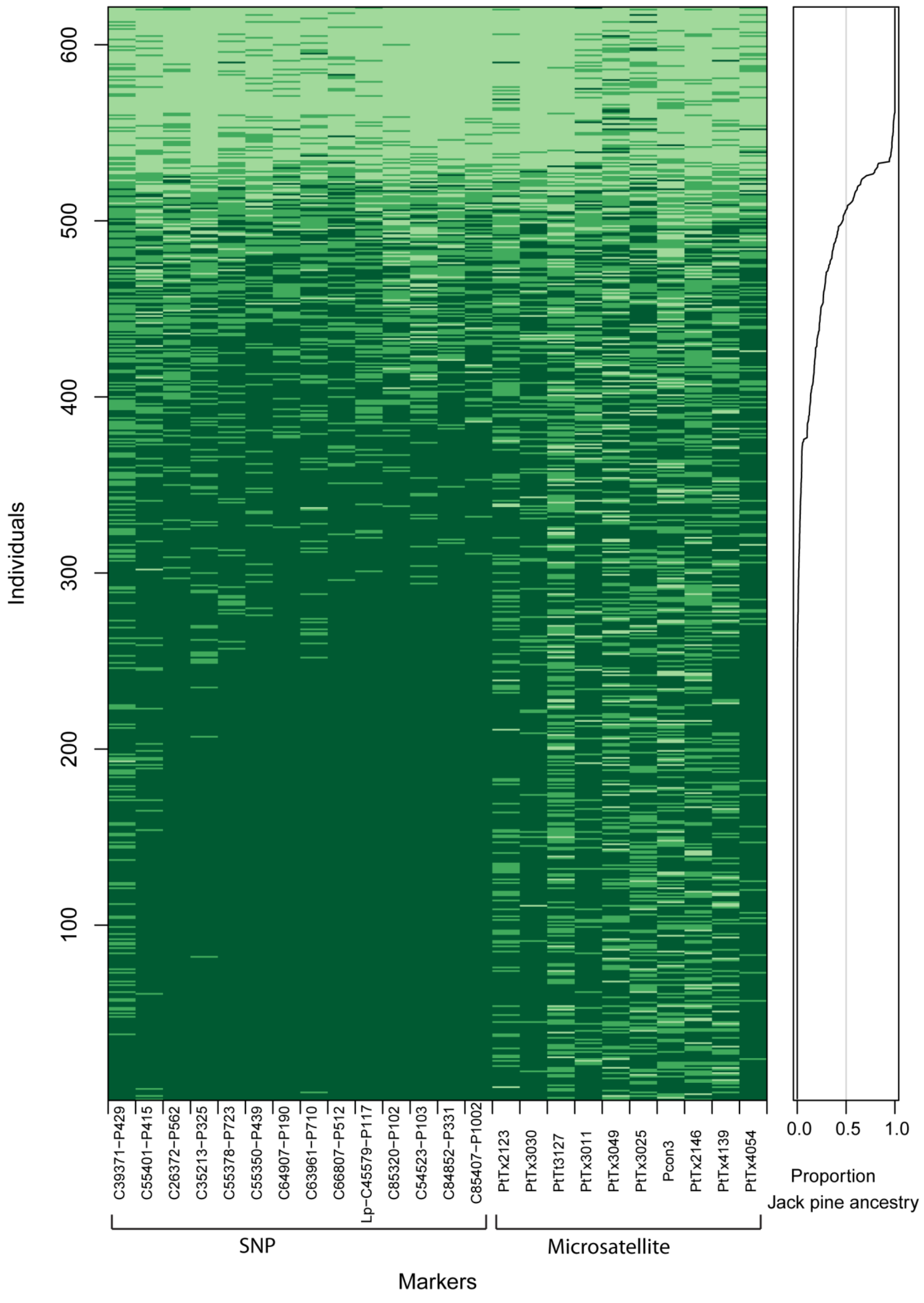
# of SNPs	Removed	Jack	Lodge	F1	F2	LpF2	JpF2	LpBC	JpBC	LpLpBC	JpJpBC
14	-	0.99	1.00	1.00	1.00	1.00	0.90	1.00	0.80	0.80	0.70
13	C39371-P429	0.99	1.00	1.00	1.00	1.00	0.90	1.00	0.80	0.70	0.50
12	C55401-P415	0.99	1.00	1.00	1.00	1.00	0.90	1.00	0.90	0.70	0.50
11	C26372-P562	0.99	1.00	1.00	1.00	1.00	0.90	1.00	0.90	0.60	0.40
10	C35213-P325	0.99	1.00	1.00	1.00	1.00	0.80	1.00	0.80	0.60	0.60
9	C55378-P723	0.99	1.00	1.00	1.00	1.00	0.90	1.00	0.70	0.70	0.50
8	C55350-P439	0.98	1.00	1.00	1.00	1.00	0.80	1.00	0.70	0.50	0.60
7	C64907-P190	0.99	1.00	1.00	1.00	1.00	0.80	1.00	0.70	0.50	0.50
6	C63961-P710	0.99	1.00	1.00	0.90	0.90	0.80	0.70	0.70	0.40	0.40
5	C66807-P512	0.97	0.99	1.00	1.00	0.90	0.70	1.00	0.60	0.70	0.40
4	LpC45579-P117	0.96	0.99	1.00	0.90	0.90	0.70	0.90	0.60	0.70	0.50
3	C85320-P102	0.98	0.99	1.00	0.70	0.90	0.30	0.90	0.40	0.30	0.30

Figure 1. Accuracy of assignment for simulated genotypes of lodgepole, jack pine and hybrid crosses (JpBC = jack pine backcross, LpBC = lodgepole pine backcross, JpJpBC = double jack pine backcross, LpLpBC = double lodgepole pine backcross, JpF2 = jack pine crossed with F2 and LpF2 = lodgepole pine crossed with F2) using SNP and microsatellite datasets combined (A), and separately (B and C, respectively) using STRUCTURE (ST) and NEWHYBRIDS (NH). Values reflect the average results across five simulated datasets each of which were based on five iterations.

Figure 2. Ancestry plot generated in INTROGRESS, dark green indicates homozygote lodgepole pine, light green indicates homozygote jack pine and medium green indicates heterozygote. The first 14 markers are SNPs and the remaining ten are microsatellites. Along the right panel is the proportion of jack pine ancestry for each individual (the inverse of which is lodgepole pine ancestry).


A - Combined
B - SNP
C - Microsatellite
Average assignment accuracy
1.00
0.80
0.60
0.40
0.20
0.00
Jack
Lodge
F1
JpBC
LpBC
F2
JpJpBC
LpLpBC
JpF2
LpF2
Species Class

425


Individuals
100
200
300
400
500
600
C39371−P429
C55401−P415
C26372−P562
C35213−P325
C55378−P723
C55350−P439
C64907−P190
C63961−P710
C66807−P512
Lp−C45579−P117
C85320−P102
C54523−P103
C84852−P331
C85407−P1002
PtTx2123
PtTx3030
PtTt3127
PtTx3011
PtTx3049
PtTx3025
Pcon3
PtTx2146
PtTx4139
PtTx4054
SNP
Microsatellite
Markers
0.0
0.5
1.0
Proportion
Jack pine ancestry