MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 2991–3005 doi: 10.1111/j.1365-294X.2012.05513.x

Genome-wide association genetics of an adaptive trait in lodgepole pine

THOMAS L. PARCHMAN,* ZACHARIAH GOMPERT,* JOANN MUDGE,† FAYE D. SCHILKEY,† CRAIG W. BENKMAN‡ and C. ALEX BUERKLE*

**Department of Botany, University of Wyoming, Laramie, WY 82071, USA, †National Center for Genome Resources, Santa Fe, NM, USA, ‡Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA*

Abstract

Pine cones that remain closed and retain seeds until fire causes the cones to open (cone serotiny) represent a key adaptive trait in a variety of pine species. In lodgepole pine, there is substantial geographical variation in serotiny across the Rocky Mountain region. This variation in serotiny has evolved as a result of geographically divergent selection, with consequences that extend to forest communities and ecosystems. An understanding of the genetic architecture of this trait is of interest owing to the wide-reaching ecological consequences of serotiny and also because of the repeated evolution of the trait across the genus. Here, we present and utilize an inexpensive and time-effective method for generating population genomic data. The method uses restriction enzymes and PCR amplification to generate a library of fragments that can be sequenced with a high level of multiplexing. We obtained data for more than 95 000 single nucleotide polymorphisms across 98 serotinous and nonserotinous lodgepole pines from three populations. We used a Bayesian generalized linear model (GLM) to test for an association between genotypic variation at these loci and serotiny. The probability of serotiny varied by genotype at 11 loci, and the association between genotype and serotiny at these loci was consistent in each of the three populations of pines. Genetic variation across these 11 loci explained 50% of the phenotypic variation in serotiny. Our results provide a first genome-wide association map of serotiny in pines and demonstrate an inexpensive and efficient method for generating population genomic data.

Keywords: association mapping, next-generation sequencing, *Pinus contorta*, population genomics, serotiny

Received 29 August 2011; revision received 15 January 2012; accepted 23 January 2012

Introduction

Analysis of the relationship between genetic and phenotypic variations has been an area of extensive research for decades because of the broad significance of understanding the heritability and evolution of traits. Recently, technological advances have begun to enable such studies in a broader range of taxa, including systems for which there is extensive knowledge of the ecological and evolutionary context of trait variation, but for which few genomic resources existed (Hudson 2008; Hohenlohe *et al.* 2010; Baxter *et al.* 2011). Various analytical approaches exist for finding associations between phenotypes and underlying genetic variation. In association mapping, researchers use naturally occurring recombination and linkage disequilibrium (LD) to map genetic regions affecting phenotype at a potentially fine genomic scale. Because association mapping does not require pedigrees or artificial crosses, it is feasible to map the genetic basis of traits in natural populations (Neale & Savolainen 2004; Gupta *et al.* 2005; Hirschhorn & Daly 2005). However, because genome-wide association mapping requires large sets of densely spaced genetic markers, its application has been limited by a paucity of genomic resources available for many organisms (Stinchcombe & Hoekstra 2007). Recently, various enrichment strategies have been developed that target a subset of the genome

Correspondence: Thomas L. Parchman, Fax: 307-766-2851; E-mail: tparchma@uwyo.edu

for DNA sequencing, even without previous genomic resources for the taxon. These enrichment methods, coupled with high levels of multiplexing of individuals, lead to DNA libraries that can be sequenced with sufficient coverage to generate population genomic data at a fraction of the previously required time and cost (Craig *et al.* 2008; Gompert *et al.* 2010; Hohenlohe *et al.* 2010; Andolfatto *et al.* 2011; Cosart *et al.* 2011; Elshire *et al.* 2011).

Conifers are well suited for association mapping because of their large, relatively unstructured populations, high levels of outcrossing and nucleotide diversity, and rapidly decaying LD (Neale & Savolainen 2004; González-Martínez *et al.* 2006; Neale 2007). The large physical size of conifer genomes and high costs of genotyping large numbers of individuals have limited association mapping to taxa with substantially developed genomic resources. To date, association studies in conifers have used gene-based analyses where SNPs were assayed across genes that had been characterized in expressed sequence tag (EST) sequencing projects or candidate-gene-based analyses where candidate genes were available for phenotypes of interest (Neale 2007). Such studies of complex traits in loblolly pine (*Pinus taeda*; González-Martínez *et al.* 2007; Eckert et al., 2010; Quesada *et al.* 2010), Douglas-fir (*Pseudotsuga menziesii* ; Eckert *et al.* 2009a, c; Cumbie *et al.* 2011) and Sitka spruce (*Picea sitchensis*; Holliday *et al.* 2010) have shown the promise of association approaches for detecting the genetic architecture of phenotypes of commercial or ecological interest in conifers. These studies have relied on relatively expensive and time-consuming SNP assay development and scoring, and substantial previous sequencing of expressed genes (González-Martínez *et al.* 2007, 2008; Eckert *et al.* 2009b). While such approaches are useful, genome-wide association mapping has the potential to identify a more complete genetic architecture and to enable mapping in cases where candidate genes for the phenotype of interest are unavailable. The ability to rapidly generate genome-wide sequence data for many individuals from any taxon should soon allow association genetic approaches to be applied more widely in conifers. Such studies will be important for tree breeding programmes and understanding many aspects of the ecological genetics of conifers.

Lodgepole pine (*Pinus contorta*) is one of the most commercially and ecologically important plants in the Rocky Mountain region, where it occurs in vast uniform stands and constitutes the structural basis (a foundation species) of montane forest ecosystems. Substantial phenotypic variation has evolved among populations in response to diverse selection pressures, including variation in seed predator communities (Benkman *et al.* 2001, 2003; Benkman & Siepielski 2004) and fire regime (Lotan 1975; Arno 1980). Many lodgepole pines hold seeds for years in serotinous cones, releasing millions of seeds only after fire, resulting in rapid and dense recolonization of burned regions (Turner *et al.* 1994). There is substantial geographical variation in the prevalence of serotiny with the percentage of serotinous trees in stands ranging from zero to nearly 100%, and much of this variation is related to geographically variable natural selection arising from fire frequency and seed predation (Lotan 1975; Benkman & Siepielski 2004; Benkman *et al.* 2008). In addition, prefire serotiny levels affect stand density in regenerating forests, the cover and density of understory plants, and species richness in these communities (Turner *et al.* 1997, 2003). Thus, through its effects on stand regeneration and community structure, serotiny is a trait with extended community and ecosystem level consequences (Wymore *et al.* 2011).

Serotiny is also a trait that is expressed in at least 22 different species across the genus *Pinus* and appears to have evolved independently multiple times (Grotkopp *et al.* 2004). Studies of the genetic control of serotiny in lodgepole pine are limited to a single progeny test (Rudolph *et al.* 1959) and to observations of cone type frequencies in natural populations (Teich 1970), where limited evidence was consistent with one or a small number of loci controlling the trait. Whereas serotiny is often considered a binary phenotype under strong genetic control (most trees tend to have mostly serotinous or nonserotinous cones), individuals of intermediate cone type do occur at low frequency. Consequently, the architecture of this trait could be more complex than previously thought. Genome-wide association mapping has the potential to more precisely describe the number of genetic regions associated with the trait, as well as eventually facilitating the identification of causal mutations within mapped regions. Knowledge of the genetic architecture of serotiny would aid our understanding of how genetic variation and natural selection have shaped adaptive phenotypic variation within a single species and our understanding of the manner in which genetic variation in this trait underlies community and ecosystem level phenomena.

Here, we use population genomic data to analyse genetic associations with serotiny across three populations of lodgepole pine. We first quantify genetic variation in the sampled pines to assess the suitability of the sequence data for detecting and characterizing population structure. We then test for associations between genetic variation and serotiny for tens of thousands of genetic regions using a Bayesian association mapping model. Importantly, our models treat genotypes as unknown parameters to be estimated and incorporate stochastic variation in sequence coverage that is common in next-generation sequencing data. We describe a laboratory method for genomic enrichment and high-throughput, multiplexed DNA sequencing (similar to CRoPS, GBS and other recently published methods; van Orsouw *et al.* 2007; Gompert *et al.* 2010; Andolfatto *et al.* 2011; Elshire *et al.* 2011). We generated individually bar-coded sequences for 98 serotinous and nonserotinous lodgepole pines from three populations in the Rocky Mountains of Wyoming. We focus on three populations that have nearly identical, intermediate frequencies of serotinous trees so that allele frequency differences between populations do not exhibit spurious associations with phenotypic variation. Our results reveal a number of genetic regions with polymorphisms associated with serotiny and cast doubt on the previous conception of a very simple genetic architecture for the trait. These results also indicate the promise of our laboratory method for rapid, cost-effective and highly multiplexed sequencing of thousands of unique genetic regions and further establish the feasibility of population genomics in taxa with limited genomic resources.

Materials and methods

Genetic material and phenotype

We obtained DNA from 98 lodgepole pines sampled from three mountain ranges in Wyoming: Wind River Range ($n = 20$), Absaroka Range, ($n = 36$) and Laramie Range (Vedauwoo; $n = 42$). To avoid uncertainty in phenotype ascertainment, we avoided sampling trees with both types of cones and were able to categorize serotiny as a binary phenotype. In each of the three populations, 44–49% of the sampled trees had serotinous cones. Across populations, we sampled needles and isolated DNA from 48 trees that had serotinous cones and 50 trees with nonserotinous cones. Importantly, the proportion of sampled trees in each population with serotinous cones was nearly identical, meaning that population structure is not a confounding issue for association analyses. All sampled trees were more than 50 years old (determined by known size–age relationship), so that absence of serotinous cones on trees was not because of young age (Lotan 1975; Critchfield 1980). DNA was extracted from 50 mg of dessicated needles using a CTAB-based method (Doyle 1991). DNA quality and concentration was assessed with agarose electrophoresis and with a NanoDrop spectrophotometer (Thermo Fisher, Inc.).

Illumina sequencing of restriction fragment libraries

We used a simple and cost-effective laboratory method for genomic enrichment prior to high-throughput sequencing, in this case with an Illumina GAIIx sequencer. Previously, we used a similar protocol with a 454 instrument (Gompert *et al.* 2010). Especially for large genomes, template reduction is necessary to ensure sufficient coverage, and we accomplish this with restriction enzyme digestion and size selection on agarose gels. Because this protocol also includes ligating bar codes to the DNA from each individual, it is suitable for studies of large numbers of individuals, as is typical in studies of molecular ecology, association mapping, and population genomics. The protocol involves similar steps as in AFLP protocols (Vos *et al.* 1995) and can be easily modified to alter the number of template regions produced for sequencing. It begins with digestion of template DNA with two restriction enzymes and is followed by PCR amplification and size selection on agarose gels to produce a pool of fragments for sequencing. By incorporating Illumina sequencing adaptors and individual 10-base bar codes into each fragment, the products are suitable for pooling and highly multiplexed sequencing. In addition, placing the bar code inside the sequenced fragments removes the need for using the second Illumina ‘indexing’ primer.

Restriction digestion and adaptor-ligation were carried out simultaneously on 0.5 μg of genomic DNA using the restriction endonucleases *Eco*RI and *Mse*I (NEB, Inc.). *Eco*RI is a methylation-sensitive enzyme, so that methylated sites in the genome (including noncoding DNA and repetitive elements) may be preferentially excluded from sequencing libraries. The adaptor sequences consist of the Illumina adaptor, a 10-bp individual bar code on one side of the fragment, and additional bases to match the restriction enzyme cut sites (*Eco*RI side: 5′-CTCTTTC CCTACACGACGCTCTTCCGATCT-3′ + 10-bp bar code + C; *Mse*I side: 5′-GCAGAAGACGGCATACGAGCTCTT CCGATCT-3′ + G; Fig. 1; the full protocol with all oligonucleotide sequences is available from Dryad-10.5061/dryad.m2271pf1). The 10-base bar codes used in this study came from a library of 151 unique sequences (454 Life Sciences Corp 2009), each of which differs by four bases from any other sequence in the library. This allows for recognition of sequencing errors in the bar codes and the correction of such errors during the parsing of bar codes. Adaptor sequences and their reverse complements were annealed by incubating at 95 °C for 5 min and slow cooling to room temperature. The adaptors were ligated to digested fragments using T4 DNA ligase (NEB, Inc.). Restriction and ligation were accomplished simultaneously in 11 μL volumes that were incubated for 18 h at 38 °C. After incubation, these reactions were diluted with 150 μL 0.1× TE buffer. We then PCR amplified these fragments using the Illumina PCR primers (1, 5′-AATGATACGGCGACCACCGAGATCTACACTCTTT CCCTACACGACGCTCTTCCGATCT-3′; 2, 5′-CAAGCA

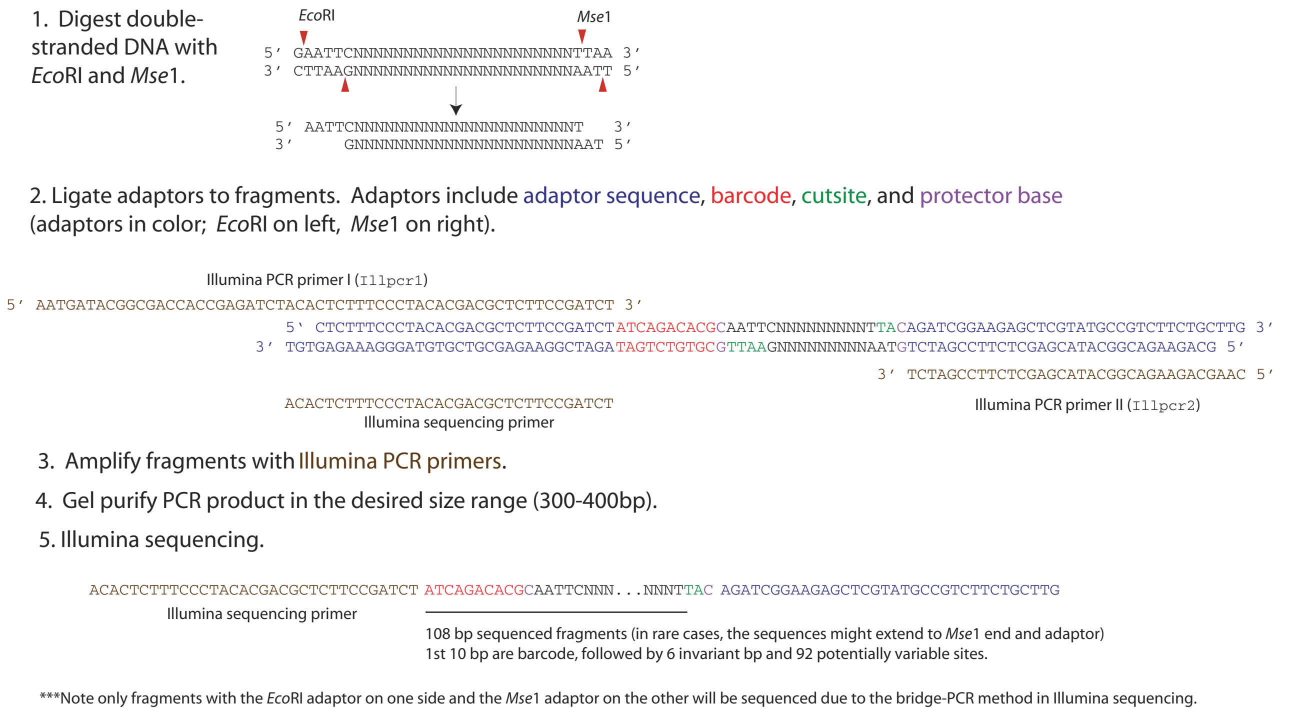
1. Digest double-stranded DNA with EcoRI and Mse1.
EcoRI
Mse1
5′ GAATTCNNNNNNNNNNNNNNNNNNNNNNNNNTTAA 3′
3′ CTTAAGNNNNNNNNNNNNNNNNNNNNNNNNNAATT 5′
5′ AATTCNNNNNNNNNNNNNNNNNNNNNNNNT 3′
3′ GNNNNNNNNNNNNNNNNNNNNNNNNNAAT 5′
2. Ligate adaptors to fragments. Adaptors include adaptor sequence, barcode, cutsite, and protector base (adaptors in color; EcoRI on left, Mse1 on right).
Illumina PCR primer I (Illpcr1)
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′
5′ CTCTTTCCCTACACGACGCTCTTCCGATCTATCAGACACGCAATTCNNNNNNNNNNTACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGATAGTCTGTGCGTTAAGNNNNNNNNNNAATGTCTAGCCTTCTCGAGCATACGGCAGAAGACG 5′
3′ TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC 5′
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Illumina sequencing primer
Illumina PCR primer II (Illpcr2)
3. Amplify fragments with Illumina PCR primers.
4. Gel purify PCR product in the desired size range (300-400bp).
5. Illumina sequencing.
ACACTCTTTCCCTACACGACGCTCTTCCGATCT ATCAGACACGCAATTCNNN...NNNTTAC AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
Illumina sequencing primer
108 bp sequenced fragments (in rare cases, the sequences might extend to Mse1 end and adaptor)
1st 10 bp are barcode, followed by 6 invariant bp and 92 potentially variable sites.
***Note only fragments with the EcoRI adaptor on one side and the Mse1 adaptor on the other will be sequenced due to the bridge-PCR method in Illumina sequencing.

Fig. 1 A schematic illustration of the laboratory protocol used here to prepare highly multiplexed libraries for Illumina sequencing.

GAAGACGGCATACGAGCTCTTCCGATCT-3′) (Illumina, Inc.) that match the sequences of the ligated adaptors. Amplification reactions contained 6 μL of the diluted restriction–ligation products, 21.7 μL 1× PCR buffer, 0.3 μL Iproof high-fidelity polymerase at 4 Units/μL (Bio-Rad, Inc.) and 2 μL of a 5 μM mix of forward and reverse Illumina PCR primers. PCR conditions included 20 PCR cycles (94 °C for 30 s, 60 °C for 30 s, 72 °C for 2 min) and a final extension at 60 °C for 30 min.

The products of these PCRs were combined into a single, homogenized pool and subjected to electrophoresis on 2.5% agarose gels at 85 V for 150 min. A volume of 80 μL of the pooled library was placed in each of 24 lanes, and DNA in the region of 300–400 bp (based on comparison to a 1-Kbp DNA ladder) was excised from the gel and purified with QiaQuick gel extraction kits (Qiagen, Inc.). Altering the size and location of the region of DNA excised from the gel during this step can alter the number of genetic regions in the sequencing library, which would affect the coverage depth obtained from sequencing. The quality and concentration of libraries was assessed with a NanoDrop spectrophotometer and quantitative electrophoresis in a Bioanalyzer (Agilent, Inc.). Suitability of libraries for sequencing was verified with real-time, quantitative PCR at NCGR (National Center for Genome Resources, Santa Fe, NM, USA). Sequencing was accomplished on a single lane run on an Illumina GAIIx device at NCGR.

Sequence assembly and data analysis

All reads from the Illumina GAIIx were 108 nucleotides in length and began with the 10-bp bar code at the *Eco*RI end of our amplified fragments and the six bases corresponding to the *Eco*RI cut site, followed by 92 bases of informative genomic sequence. Prior to further processing, we trimmed the bar codes and the six following nucleotides from each fragment. In this initial processing, we corrected bar codes in the Illumina reads that differed by a single base from the reference bar code. As all of the bar codes differ by a minimum of four bases, this unambiguously and conservatively corrected a small number of reads containing sequencing errors in the bar code. During this parsing of bar codes, we added the correct individual identification for each tree to the identification line associated with each sequence in the fastq format files.

We first executed a *de novo* assembly based on a subset of 20 million reads using SEQMAN NGEN 2.0 (DNAstar, Inc.). We then used the consensus sequences of the highest quality contigs that had a minimum coverage depth of 7×, a minimum length of 88 bases and a maximum length of 96 bases as reference sequences on to which we assembled the entire set of sequences. Repeat-rich regions often assemble into long contigs, and the removal of such contigs from the reference improved the quality of the assembled contigs considerably. A template-guided assembly based on this reference was then executed in SEQMAN NGEN 3.0 (DNAstar, Inc.). For both the *de novo* and reference-based assemblies, we used a gap penalty of 50, minimum match percentage of 90%, match size of 50 bp, mismatch penalty of 15 and used the repeat handling option. For the reference-based assembly, reads that aligned to multiple reference contigs were discarded from the assembly to limit the representation of repetitive DNA in the final assembly. The full parameters used in assemblies are available from the authors by request.

We used custom Perl scripts along with bcftools and samtools (Li *et al.* 2009) to call variant sites in the assembled contigs. samtools processes input BAM files (a compressed file format for storing assembly data), computes the probability of the data given each possible genotype and stores the probabilities in the BCF format. bcftools then executes the calling of variant sites based on a Bayesian model that accounts for uncertainty in the data. We considered only SNPs and disregarded insertions and deletions. We used the full prior in bcftools, only considered SNPs where reads were present for at least 30% of the individuals, and required the probability of the data to be <0.05 under the assumption that all samples were homozygous for the reference allele. The data from the called variant sites were parsed and placed in files containing information on the haplotype counts for each individual at each genetic region, the sequence information on each SNP by haplotype for each individual at each genetic region, and a file containing the number of reads for each SNP in each individual. We then removed all genetic regions where individuals appeared to have more than two haplotypes. To further ensure that all loci behaved as Mendelian units, we discarded any variable site where the observed allele counts from apparent heterozygous individuals were very unlikely given a binomial distribution with $p = 0.5$. Specifically, we discarded a locus if $P(Y \leq y \mid p = 0.5, n) \leq 0.05$, where y is the count of the less frequently observed allele.

We aligned the consensus sequences containing SNPs to 322 000 contig and singleton sequences from a *Pinus contorta* whole transcriptome sequencing project (Parchman *et al.* 2010) to assess the proportion of these sequences that might represent transcribed regions. We executed this alignment using a template-guided assembly in SEQMAN NGEN 3.0 with the same parameter settings as used for the assemblies described above.

We used a Bayesian model to estimate population allele frequencies and genotypic states for each called SNP based on the observed sequence data (Gompert *et al.* 2012). This model treats genotypes at each locus and the allele frequencies as unknown parameters and assumes that sequences are sampled stochastically with variable coverage per nucleotide. Thus, the model accounts for uncertainty in coverage across individuals, loci and homologous gene copies inherent to next-generation sequencing projects and allows the use of variable or stochastic sequencing coverage at all levels. A detailed description of this model is given by Gompert *et al.* (2012). We obtained posterior probability distributions for allele frequencies and genotypic state using Markov chain Monte Carlo (MCMC). We ran 20 000 MCMC steps where parameter values at every fourth step were retained. We used principal component analysis (PCA) to summarize genetic variation among the three populations. We treated the probability of two of three genotypic states (the heterozygote and one homozygote) at each locus as variables for PCA (the third genotypic probability is redundant, as these probabilities must sum to one). We conducted PCA in R using the prcomp function on the centred variables (R Development Core Team 2011).

We used a hierarchical Bayesian specification of the F model to estimate genetic differentiation among populations (details of this model are provided in Gompert *et al.* 2012). The F model is commonly used to quantify population structure and yields a parameter equivalent to F_{ST} under several neutral population genetic models, including the equilibrium infinite-island model and a model of divergence from a common ancestral population without gene flow (Balding & Nichols 1995; Rannala & Hartigan 1996; Nicholson *et al.* 2002; Balding 2003; Falush *et al.* 2003). The F model allows for uncertainty in F_{ST} because of evolutionary and statistical sampling and treats F_{ST} as an evolutionary parameter rather than a statistical summary of allele frequencies. Our implementation of the F model assumes Hardy–Weinberg and linkage equilibrium within populations and that sequences do not contain errors, but allows for uncertainty in genotypic state because of variable and limited sequence coverage (Gompert *et al.* 2012). We used MCMC to obtain genome-level estimates of F_{ST} for each pair of populations. Chains were iterated for 25 000 steps, and parameter values were recorded every fifth step.

We tested for an association between each biallelic marker locus and serotiny using a simple Bayesian linear model. We allowed the effect of each marker genotype to vary independently (i.e. we did not assume a simple additive model for allelic effects). Specifically, we assumed the observed phenotypic value (0 = nonserotinous, 1 = serotinous) for individual i (y_i) follows a Bernoulli distribution with $\pi = p_i$ and that the probability of serotiny for individual i (p_i) can be described by the linear model,

$$\mathrm{p}_i = \sum_{k=0}^{2} \beta_{kj} \delta_{x_{ij}}^{k}, \tag{1}$$

where $x_{ij} \in \{0,1,2\}$ is the genotype of individual i for locus j, and $\delta_n^{n'}$ is Kronecker's delta (i.e. $\delta_n^{n'} = 1$ if $n =$

n' and 0 otherwise). $x_{ij} = 1$ denotes the heterozygous genotype for locus j, whereas $x_{ij} = 0$ and $x_{ij} = 2$ denote alternative homozygous genotypes. We placed independent Beta priors on each of the regression coefficients,

$$\beta_{kj} \sim \text{Beta}(a, b) \tag{2}$$

where $a = 1.92$ and $b = 2.08$ were specified to provide a mildly informative and conservative prior with an expected value $(a/(a + b) = 0.48)$ equal to the fraction of serotinous pines in the sample.

Because of generally low sequence coverage, we modelled genotypes as unknown parameters, which we estimated coincident with the Bayesian association mapping analysis. Genotypic states were estimated using the allele frequency model described earlier. We used MCMC to estimate marginal posterior probability distributions for regression coefficients (β) and genotype probabilities. Source code in C for this analysis is available from Dryad (10.5061/dryad.m2271pf1). We conducted single-locus tests of association for all loci in a pooled analysis of all 98 trees and also separately within each of the three populations. We obtained 10 000 MCMC samples from the posterior distribution for each parameter. We classified loci as associated with serotiny, and thus as marker tags for candidate genetic regions containing functional variants that control serotiny, if the lower 0.05 quantile of the posterior distribution for any of the three regression coefficients was >0.48 (the frequency of serotiny in the sample), or if the upper 0.95 quantile was <0.48. The odds ratios for a locus were estimated as the ratio of the probability of serotiny among alternative genotypes. We executed BLAST searches against NCBI's nr repository with the sequences containing loci that had significant associations with serotiny.

We estimated Burrow's composite measure of Hardy–Weinberg and linkage disequillibrium (Δ) between each of the 11 loci statistically associated with serotiny (see Results) and a haphazard set of 19 additional loci (Weir 1979). We estimated Δ to determine whether the candidate loci probably represent one or multiple genetic regions and to identify loci with severe deviation from Hardy–Weinberg equilibrium. The Δ metric does not require phased data and does not assume Hardy–Weinberg equilibrium at either locus, but instead provides a joint measure of intralocus and interlocus disequilibria (Weir 1979). Because of uncertainty in genotypic state for each individual, we used a Monte Carlo algorithm to sample genotypes for individuals and calculated the composite disequilibrium measure ($\Delta_{jj'}$) for each locus pair. We sampled genotypes for each individual and locus according to the estimated genotype probabilities obtained previously. We then obtained our estimate for $\Delta_{jj'}$ by taking the mean of the calculated δ values over 1000 repeated samples of individual genotypes. We used a simple permutation test to determine whether the extent of LD among the candidate loci exceeded LD among the noncandidate loci.

We fit GLMs for the association between genotypes at the 11 candidate loci and serotiny to obtain likelihood-based Monte Carlo estimates of the proportion of phenotypic variation explained by the genetic data. Our primary interest was to determine how much of the variation in serotiny could be explained by a full model that included all 11 candidate loci, but we first fit GLMs including individual loci for comparison. We assumed the phenotypic data were described by a binomial probability distribution and used a logit link function. As above, we iteratively sampled genotypes for each individual and locus according to the estimated genotype probabilities to account for genotype uncertainty. For each sampled set of genotypes, we fit a GLM that included only an intercept (null model), an intercept and genotype for one locus (single-locus model), or an intercept and genotypes for the 11 candidate loci (full model) using the iteratively reweighted least squares algorithm implemented in the R function glm. We obtained Monte Carlo estimates of the Akaike information criterion (AIC) and the coefficient of determination (r^2) for each sample by taking the mean over 1000 sampled sets of genotypes. The coefficient of determination can be interpreted as the proportion of phenotypic variation explained by the genetic data included in each model. We repeated this procedure after randomizing phenotypes (1000 repeated samples of genotypes for each of 10 randomizations of phenotypes) to obtain the expected distribution of AIC and r^2 under the null hypothesis of no association between genotype and phenotype.

To further examine the contribution of different numbers and subsets of loci, we conducted several additional analyses using the same Monte Carlo and glm approaches described above. We used the bestglm routine in R to analyse all possible combinations of loci included in general linear models predicting serotiny and to systematically choose the best subset of models. bestglm uses a simple exhaustive search algorithm across models including all possible numbers and combinations of parameters to find the models with the smallest sum of squares or deviances and evaluates models based on AIC. We ran 1000 iterations of bestglm using the Monte Carlo sampled genotypes and tracked the loci that were included in the best overall model for each iteration. Likewise, to evaluate the effect of adding additional loci to these GLMs, we calculated r^2 and AIC for a one locus model and then sequentially for each model with an additional locus added (up to the full model with 11 loci). For these analyses, we added loci in order of their r^2 values from the single-locus models described above.

Results

A total of 36 675 491 sequences were generated on an Illumina GAIIx platform. After trimming off bar codes and the preceding six bases associated with the *Eco*RI cut site, 30 719 069 reads averaging 92 bases in length were retained for analysis. An average of 313 459 sequences were generated per bar-coded individual. *De novo* assembly of a subset of 20 million reads placed 3 783 249 reads into 148 956 contigs containing a minimum of seven reads, for an average coverage depth across all individuals of 25×. We discarded contigs longer than 96 bases and shorter than 88 bases, as well as those with insufficient coverage depth, and concatenated the consensus sequences from these contigs into an artificial reference template. The template-guided assembly of all 30 719 069 reads then placed 9 395 072 reads onto the reference sequences and resulted in an average coverage depth of 66× per genetic region across all individuals (0.7× per individual). After processing the BAM files created from these assembly and calling variants using bcftools in samtools, 97 616 variant sites spanning 45 529 contigs were retained for use in the following analyses. A total of 3594 of the consensus sequences representing these contigs aligned to *Pinus contorta* transcriptomic sequences, indicating that at least 6% of the data probably represent transcribed regions.

The minor allele frequency (MAF) varied considerably among loci (Fig. 2). The sequence data included numerous low-frequency variants, but the majority of loci were quite variable. The first two principal components explained 8% (PC1) and 7% (PC2) of the variation in estimated genotypic state probabilities across all 97 616 loci. The first two principal components suggest a lack of differentiation between serotinous and nonserotinous trees within populations, but a geographical signal of isolation by distance (Fig. 3; Novembre & Stephens 2008). We note that because the PCA was based on genotype probabilities, and because of features of the model used to generate F_{ST} estimates among populations, the per cent variation explained in the PCA is not necessarily proportional to F_{ST} among populations (McVean 2009). Pairwise estimates of genome-level genetic differentiation (F_{ST}) between the sampled populations were very low but nonzero, and population differentiation exceeded the differentiation between serotinous and nonserotinous pines (Table 2).

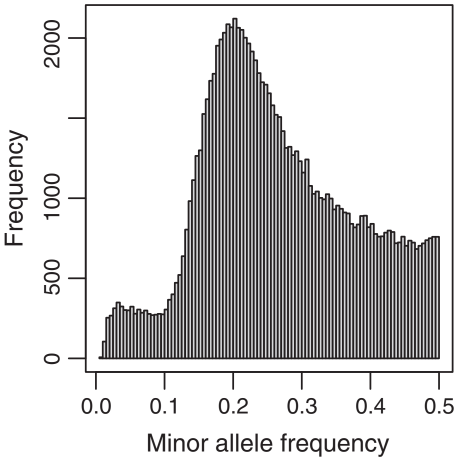

Fig. 2 The distribution of minor allele frequencies for the 97 616 loci used in this study.

We detected 11 loci statistically associated with variation in serotiny. Loci with low sequencing coverage contain less information and are thus unlikely to have significant parameter estimates in the Bayesian association analyses. The loci exhibiting significant associations had relatively high coverage depths (mean of five reads per individual per locus; Table 1), and there is comparatively little uncertainty in genotypes for these loci. The highest odds ratio (ratio of the probability of serotiny

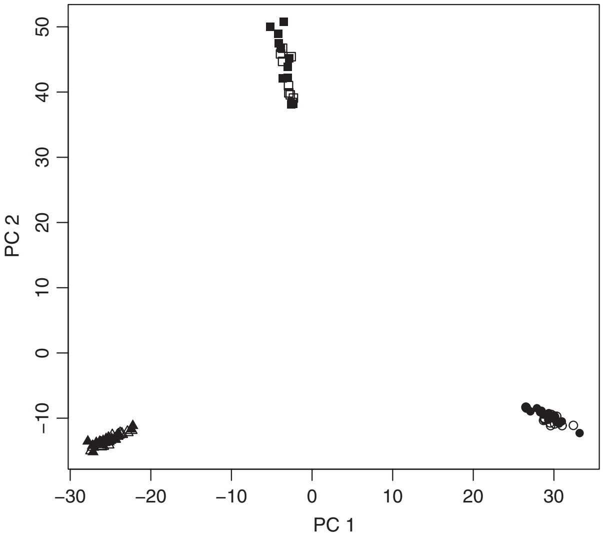

Fig. 3 A plot of the first two principal components of genetic variation among individuals from three lodgepole pine populations clearly distinguishes individuals from different sampling localities, but substantial overlap of serotinous and nonserotinous trees in each. The first two principal components account for 8% (PC 1) and 7% (PC 2) of the variation in genotype probabilities across the 97 616 loci. Populations are indicated by shape (triangles, Absaroka Range; squares, Wind River Range; and circles, Laramie Range–Vedauwoo), and filled symbols represent serotinous trees and open symbols represent nonserotinous trees.

Table 1 Identifiers for genetic regions containing nucleotide polymorphisms associated with serotiny, minor allele frequency for each locus (MAF), genotype-specific parameter estimates for the probability of serotiny conditional on genotype (the significant parameter for the locus is in bold; overall population frequency of serotiny was 47/98 = 0.48), and estimates for the ratio of probability in serotiny between pairs of genotypes (probability ratio)

Identifier	MAF	Depth	Probability of serotiny			Probability ratio			Accession
			AA	AA′	A′A′	AA/AA′	AA/A′A′	AA′/A′A′	
65253	0.18	5.6	0.542 (0.16–0.87)	0.429 (0.34–0.52)	**0.704** (0.51–0.86)	0.792 (0.46–2.59)	1.285 (0.72–4.21)	1.627 (1.11–2.26)	AC241283
112487	0.16	3.14	0.485 (0.13–0.85)	**0.760** (0.51–0.93)	0.456 (0.37–0.54)	0.650 (0.17–1.27)	1.067 (0.29–1.96)	1.664 (1.06–2.27)	AC241311
54398	0.22	5.24	0.488 (0.13–0.86)	**0.734** (0.50–0.90)	0.455 (0.37–0.54)	0.676 (0.18–1.34)	1.071 (0.28–1.97)	1.604 (1.02–2.18)	NA
1428	0.20	17.8	0.550 (0.21–00.86)	0.433 (0.35–0.52)	**0.781** (0.57–0.92)	1.258 (0.47–2.13)	0.710 (0.27–1.24)	0.559 (0.41–0.80)	NA
2539	0.31	3.54	0.444 (0.12–0.80)	0.385 (0.28–0.50)	**0.649** (0.51–0.77)	1.147 (0.30–2.31)	0.687 (0.18–1.30)	0.594 (0.41–0.86)	AC241288
103454	0.18	2.98	0.480 (0.13–0.86)	**0.274** (0.10–0.48)	0.539 (0.45–0.63)	1.762 (0.42–5.37)	0.897 (0.23–1.63)	0.510 (0.18–0.93)	AC241359
17466	0.19	3.7	0.494 (0.13–0.86)	**0.339** (0.20–0.49)	0.566 (0.46–0.67)	1.447 (0.38–3.08)	0.869 (0.24–1.55)	0.598 (0.35–0.91)	AC241351
1994	0.26	4.29	0.497 (0.14–0.86)	0.359 (0.24–0.48)	**0.605** (0.50–0.71)	1.381 (0.37–2.75)	0.823 (0.23–1.49)	0.595 (0.38–0.87)	AC241292
64526	0.25	1.57	0.483 (0.13–0.85)	**0.307** (0.15–0.47)	0.567 (0.46–0.66)	0.639 (0.25–2.55)	1.169 (0.64–4.51)	1.843 (1.08–4.03)	AC241284
1853	0.18	2.28	0.484 (0.39–0.57)	0.708 (0.48–0.88)	**0.188** (0.04–0.44)	0.687 (0.50–1.04)	2.555 (1.05–11.99)	3.697 (1.43–17.23)	EU998740
40518	0.20	5.36	0.571 (0.18–0.89)	0.601 (0.48–0.71)	**0.347** (0.23–0.48)	0.946 (0.29–1.55)	1.616 (0.49–3.07)	1.738 (1.17–2.68)	NA

The minor allele homozygous genotypes, typically characterized by large uncertainty in parameter estimates, are represented by the first genotype column. 95% credible is given below in parentheses for both the parameter estimates and each odds ratio. Coverage depth (Depth) represents the average number of sequences per individual obtained for each locus. NCBI accession numbers are given for sequences exhibiting BLASTN matches to loci where matches occurred.

for a pair of genotypes) for the probability of observing a serotinous tree among genotypes was 3.697, and the probability of serotiny conditional on genotype ranged from 0.19 to 0.78 (Table 1). One of the SNPs most clearly associated with serotiny was in locus 1853 (Table 1). Individuals with one of the homozygous genotypes only had an 18% chance of being serotinous, whereas heterozygous individuals had a 71% chance of being serotinous. Consistent with expectations from the Hardy–Weinberg law, individuals homozygous for the less frequent allele at each locus were rare in the sample. Thus, our estimates of the probability of serotiny for the rare genotype were associated with considerable uncertainty. The same was occasionally true for other genotypes, but for each of the 11 loci, the probability of serotiny for at least one genotype deviated significantly from the sample frequency of serotiny. Analyses conducted within individual populations implicated a smaller and mostly nonoverlapping set of loci associated with serotiny. That the same loci were not repeatedly detected in the within population association analyses is not surprising given the small number of individuals in these populations and the very low levels of LD characteristic of pine populations. While detecting similar loci exhibiting associations would be expected in populations with high levels of LD (such as mapping populations), this is less likely for populations of pines where LD is known to decay rapidly (Brown *et al.* 2004; Krutovsky & Neale 2005; Heuertz *et al.* 2006). However, some of the same loci were associated with serotiny within individual populations and in the whole sample, and parameter estimates for these 11 candidate loci were highly correlated across the three pine populations (Fig. 4A–C), indicating the loci have similar relationships with serotiny and providing some evidence for a shared genetic architecture.

Of the associated loci that had BLAST matches to NCBI's nr repository, nearly all were to large sequences identified as random *Pinus taeda* clones (Table 1). One sequence (locus 1853) had high sequence similarity to a *P. contorta* chloroplast accession (Table 1). However, many individuals were clearly heterozygous at this locus, indicating it is a nuclear copy of chloroplast DNA, which is common in plants (Ayliffe & Timmis 1992; Yuan *et al.* 2002; Huang *et al.* 2003).

Intra- and interlocus disequilibria for the 11 candidate loci were weak to nonexistent (Fig. 4D). The mean absolute value of $|\Delta|$ between candidate loci was 0.008 (minimum < 0.001, maximum = 0.026). Moreover, the magnitude of Δ between candidate loci was not significantly greater than between pairs of 19 arbitrarily chosen loci ($|\Delta_{\text{candidate}}| - |\Delta_{\text{noncandidate}}| = -0.0009$, $P = 0.7421$). Finally, none of the candidate loci mapped to the same sequence read. In addition, most of the 11 loci

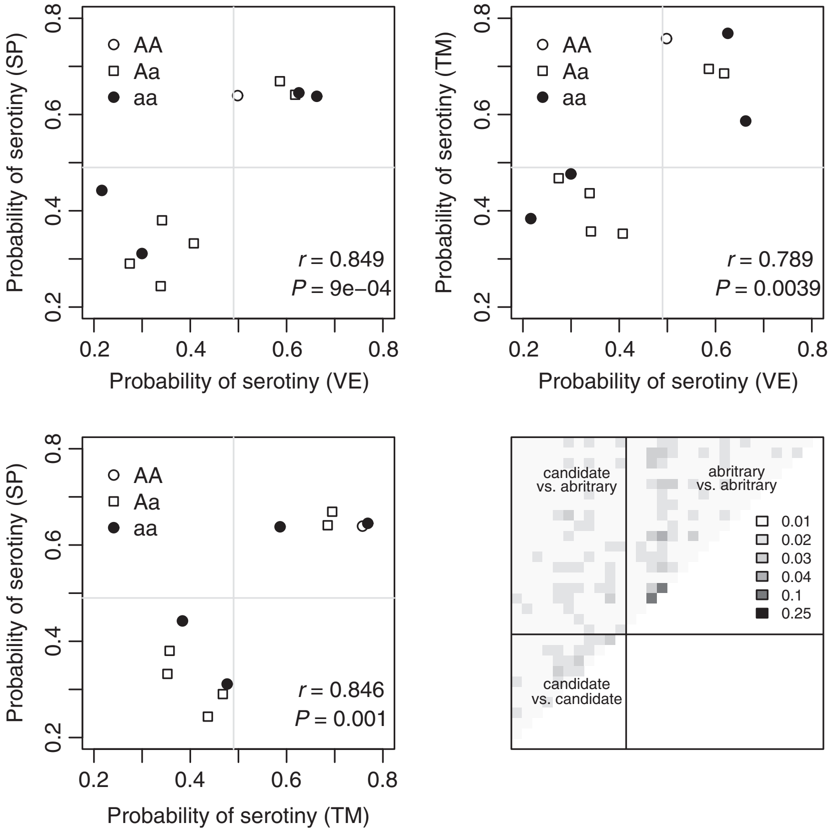

Fig. 4 Three panels contain plots of parameter estimates for the same 11 loci that were significantly associated with serotiny when all individuals across populations were used in the combined analysis. Each plot displays parameter estimates from Bayesian association mapping conducted within individual populations and presents the strength of association of parameter estimates from trees in different populations. The lower right panel represents the strength of correlations between allele frequencies of the 11 loci significantly associated with serotiny (bottom left quadrant), between the associated loci and 19 arbitrarily selected loci (upper left quadrant), and between pairs of the 19 arbitrarily selected loci (upper right quadrant). The axes on this plot correspond to loci used in pairwise linkage disequilibrium estimation in the above order.

had best BLAST matches to large arbitrary *P. taeda* clones of unknown function and location, and each locus matched a different and unique sequence.

GLMs that included single candidate locus effects had higher probabilities relative to the null model, consistent with the single-locus Bayesian association mapping results ($\mathrm{AIC_{null}} - \mathrm{AIC_{single}}$: mean across 11 loci = 4.5, range = 3.1–7.0). The proportion of phenotypic variation explained by genetic variation at each candidate locus ranged from 0.04 to 0.08 (mean = 0.06). Conversely, when we randomized phenotypes, little variation in serotiny was explained by genetic variation at individual loci and the single-locus models were no more probable than the null model ($\mathrm{AIC_{null}} - \mathrm{AIC_{single}}$: mean across 11 loci = −1.3, range = −2.0 – −0.5). The full model that included genotype effects for all 11 candidate loci was preferred over both the null model and single-locus models ($\mathrm{AIC_{null}} - \mathrm{AIC_{full}} = 32.2$), and this model explained 50% of the variation in serotiny. A moderate proportion of variance in serotiny could also be explained by the full model when phenotypes were randomized ($r^2 = 0.17$). This is perhaps not surprising as the full model contains a substantial number of parameters relative to the number of individuals. However, after considering the number of parameters, the full model with randomized phenotypes was clearly inferior to the null model ($\mathrm{AIC_{null}} - \mathrm{AIC_{full}} = -13.1$). Additionally, the proportion of variation in serotiny explained by the 11 locus model exceeds that explained by the 11 locus model with permuted phenotypes by 0.33 (i.e. most of the variation explained by the full model is not simply a product of the number of parameters included in the model).

Selection among GLMs with all possible combinations of 11 loci indicated that the best models contained more than seven loci. Models with fewer than five loci were consistently ranked lowest. Each locus was included at least 43% of the time in the best model across 1000 iterations of the model selection procedure, while one locus (10) was included in the best model nearly every time.

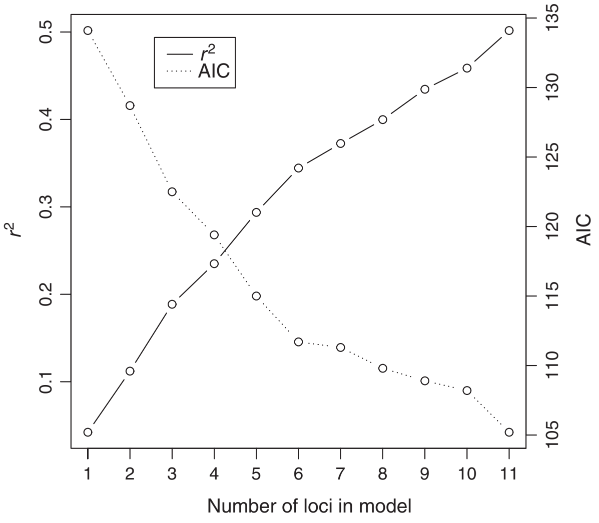

Fig. 5 r^2 and Akaike information criterion for generalized linear models (GLMs) including different numbers of the loci that exhibited significant associations with serotiny. Loci were added to the models in order of r^2 values estimated from single-locus GLMs.

Consequently, there was little basis for excluding loci from the full model, and all loci appeared to contain some information. In the sequential addition of loci to GLMs, r^2 continued to increase, and AIC continued to decrease, indicating improvement in the quality and explanatory power of these models as more loci were included (Fig. 5). These results provide further support that the models with more loci explain more variation in serotiny, and that multiple genetic regions are associated with the trait.

Discussion

This genome-wide association study provides an initial characterization of the genetic architecture of serotiny and suggests that the inheritance of serotiny is more complex than the one gene, two-allele scenario suggested previously (Rudolph *et al.* 1959; Teich 1970) and widely repeated in the ecological literature (Wymore *et al.* 2011). Remarkably, with only 98 individuals, strong statistical associations were found between serotiny and genotypes at 11 loci, and together these 11 loci can account for nearly 50% of the phenotypic variation in the sample. Linear model analyses indicated that each of the loci contribute to serotiny, and models containing all loci explained substantially more phenotypic variation than those containing subsets of loci (even when penalized for the increased number of parameters in the model, as indicated by AIC) (Fig. 5). The percentage of variation explained by associated loci, individually and combined, is in the range of that commonly reported in gene-based association mapping of quantitative traits in other conifers (González-Martínez *et al.* 2007; Eckert *et al.* 2009a; Holliday *et al.* 2010; Quesada *et al.* 2010). Furthermore, the genotype parameter estimates for these 11 loci were highly correlated in the three different populations (Fig. 4), indicating a similar genetic architecture of serotiny in different populations and strengthening our confidence that these loci tag genetic regions linked with causal variants. A shared genetic architecture is perhaps not surprising given the low levels of genetic differentiation among the sampled populations. LD among the 11 candidate loci was low to nonexistent and quite similar to LD among other arbitrarily chosen loci. These results suggest that the 11 loci tag genetic regions in LD with multiple, independent functional variants. A polygenic basis for serotiny is also supported by the existence of trees with intermediate levels of serotiny in natural populations.

These initial findings are based on relatively low sequence coverage for unmapped nucleotide polymorphisms in a sample of only 98 trees. Additional sequencing of these individuals would increase the precision of parameter estimates and might implicate additional (or fewer) loci as associated with serotiny or lead to the identification of multiple SNPs as nonindependent (i.e. in LD). Likewise, sampling a larger number of individuals would lead to more precise estimates of phenotypic effects of genotypic variation and a larger difference between the full and null models. At present, a null model that includes genotypes for all 11 loci explains a substantial fraction of the variation in randomized phenotypes ($r^2 = 0.17$); with a larger sample of trees, this fraction of variation explained in the null model would drop. Finally, the current description of the genetic architecture of serotiny is based on contrasts between extreme individuals that expressed serotiny or did not. However, serotiny in cones could also be scored as a continuous trait, with direct measurement of the temperature that is required for cones to open (Perry & Lotan 1979). A threshold response to temperature among different trees could give the appearance of a binary phenotypic trait and lead to simple hypotheses for the genetic architecture of serotiny. Future work measuring serotiny as a continuous trait (temperature that is required for cones to open) could lead to additional insights about the genetic architecture. Clonally replicated common garden designs have often bolstered association genetic studies in conifers (Neale & Kremer 2011), but are unavailable for serotinous pines. However, accounting for environmental variation across individuals and populations in a landscape genomics approach could also help to evaluate the extent to which environmentally induced plasticity influences serotiny. In on-going research, we are sequencing DNA from parent trees and haploid megagametophyte tissues to generate a high-density linkage map for these loci in lodgepole pine. A linkage map for the sequenced loci will allow better discrimination between linked and unlinked markers and better estimates of the number of independent loci that are associated with serotiny. Meanwhile, the current analyses implicate 11 loci as significantly associated with variation in serotiny and provide a basis for future research.

Genome-wide estimates of F_{ST} indicate low levels of genetic differentiation among the populations, in agreement with previous analyses across these and other lodgepole pine populations in this region (Wheeler & Guries 1982; Dancik & Yeh 1983; Yeh *et al.* 1985; Epperson & Allard 1989; Dong & Wagner 1994; Godbout *et al.* 2008; Parchman *et al.* 2011). High levels of diversity within and low levels of differentiation between populations of forest trees are common (González-Martínez *et al.* 2006; Neale 2007; Krutovsky *et al.* 2009). However, in contrast to previous studies, the large number of loci analysed here revealed clear geographical differences in allele frequencies (Fig. 3), highlighting the potential of genome-scale SNP data to recover geographical signal that may have been previously subtle or undetectable (Li *et al.* 2008; Novembre *et al.* 2008). Genome-wide differentiation between serotinous and nonserotinous trees within populations was absent (Table 2), and there was no clustering of individuals by phenotype in genotypic principal component space (Fig. 3). Population structure is often the principal confounding factor in association genetic studies and is often modelled as a covariate in such analyses (Yu *et al.* 2006). Because each of the populations analysed had nearly identical, intermediate frequencies of serotinous trees, any population structure would not have been a confounding factor for mapping serotiny. Nevertheless, the ability of this sequencing approach to detect fine-scale population structure will be useful for future studies surveying variation across a wider geographical range of populations and frequencies of serotiny.

Researchers have recently used candidate-gene-based as well as EST-based association studies to dissect the genetic architecture of complex traits in various conifers (González-Martínez *et al.* 2007; Eckert *et al.* 2009a, c, 2010; Holliday *et al.* 2010; Quesada *et al.* 2010; Cumbie *et al.* 2011). In contrast, the current study is a first application of genome-wide association mapping in conifers. Gene-based and genome-wide association mapping methods each have strengths and limitations, and both will continue to contribute to advances in genetics. Tests of association involving candidate genes have the advantage of a relatively small, focal set of loci representing functional variation and will be most useful for fine-scale mapping to individual polymorphisms within previously mapped regions (Neale & Savolainen 2004; Neale 2007; Eckert *et al.* 2010). A drawback of the candidate-gene-based approach is that it can only test for associations at loci that are defined a priori and may therefore miss substantial components of the genetic architecture or may not be feasible if no reasonable genic resources exist. By contrast, genome-wide association methods are not constrained by a priori hypotheses and can test for associations in both genic and intergenic regions of the genome (Gupta *et al.* 2005; Hirschhorn & Daly 2005). Given that functional variants need not be restricted to protein coding regions (Lynch 2007), progress in the genetics of previously unstudied traits might proceed through initial genome-wide association mapping, followed by more fine-scale mapping within candidate regions. Finally, it is noteworthy that the manner in which we identified polymorphisms means that our mapping was unaffected by ascertainment bias. This is in contrast to many mapping studies that do suffer from bias in ascertainment of marker loci, because they utilize nucleotide polymorphisms that are segregating within particular reference populations (i.e. have a minimal MAF in the reference population that exceeds some threshold; Kuhner *et al.* 2000; Clark *et al.* 2005; Albrechtsen *et al.* 2010).

The ability to conduct genome-wide association studies has been hindered in most taxa by the large number of markers needed to reliably detect polymorphisms in LD with causal variants (Neale & Savolainen 2004). Ideally, association mapping requires a high enough

Table 2 Summary of genome-wide genetic differentiation estimates between populations and between serotinous and nonserotinous lodgepole pines using the hierarchical Bayesian F model of Gompert *et al.* (2012)

Type	Comparison	lower CI	median	upper CI
Geographical	Wind River Range vs. Absaroka Range	0.0111	**0.0133**	0.0145
	Wind River Range vs. Laramie Range	0.0092	**0.0113**	0.0124
	Absaroka Range vs. Laramie Range	0.0004	**0.0005**	0.0006
Phenotypic	Serotinous vs. nonserotinous	4×10^{-6}	$\mathbf{5{\times}10^{-6}}$	$6{\times}10^{-6}$

The median value of the posterior distribution (in bold face) and the upper and lower limits for the 95% credible intervals are presented for each pairwise comparison.

number of markers to expect that most of the genome would be in LD with genotyped SNPs, including causative polymorphisms. Indeed, successful association mapping studies in humans commonly utilize 1–2 million SNPs (Carlson *et al.* 2003; McCarthy *et al.* 2008). Pine genomes are very large (>10^{10} bases; Hall *et al.* 2000; Joyner *et al.* 2001), and LD has been found to decay relatively rapidly in conifers (Brown *et al.* 2004; Krutovsky & Neale 2005; Heuertz Myriam *et al.* 2006). While the rapid decay of LD should facilitate fine mapping in association studies, it also means that enormous numbers of SNPs (>2 million) would be preferable for genome-wide association studies in pines (Neale & Savolainen 2004; Neale 2007; Neale & Kremer 2011). Although the 97 716 SNPs analysed here have made a genome-wide association study feasible, the genetic regions in LD with the genotyped SNPs still represent a relatively small portion of the pine genome. Consequently, the loci for which we detected associations with serotiny are likely to tag genetic regions in LD with causal variants, rather than the variants themselves. With increases in the ability to query larger numbers of markers with high-throughput sequencing, it should soon become possible to conduct association studies that query a more substantial fraction of the genome and to produce meaningful and thorough genome-wide association studies in conifers and other nonmodel organisms.

Genomic reduction and highly multiplexed Illumina sequencing

The sequencing approach outlined here represents a time- and cost-effective route for producing population genomic data for nearly any taxon, without the need for any previous genome sequencing. The method is similar to other recently published methods (van Orsouw *et al.* 2007; Gompert *et al.* 2010; Hohenlohe *et al.* 2010; Andolfatto *et al.* 2011; Elshire *et al.* 2011), but also differs in several details. By altering the size range of gel-purified fragments, the degree of genomic reduction and coverage depth for fragments can be customized to individual sequencing projects. Sequencing of highly multiplexed samples is made possible by the enrichment procedure and by indexing DNA samples, which can be performed at the individual or population level (Gompert *et al.* 2010, 2012). Although we used 96 bar codes in the present study, many more bar codes can be generated for higher levels of multiplexing (Meyer & Kircher 2010). Compared to more traditional Illumina library preparation, this type of method has the advantages of reduced sample handling, only a single PCR step and gel purification, and inexpensive and flexible bar coding. In addition, the use of restriction enzymes may aid in avoiding repetitive portions of the genome (Elshire *et al.* 2011). Compared to traditional SNP assay techniques (e.g. Illumina's Golden Gate Assay) where SNPs are detected in a small set of individuals and then scored in broader panels, an important advantage of this type of multiplexed sequencing is the absence of ascertainment bias (Kuhner *et al.* 2000; Clark *et al.* 2005; Albrechtsen *et al.* 2010).

After calling variants in the reference-based assembly, we obtained haplotypic data for 45,529 genetic regions and SNP data for 97 616 sites. As expected, coverage varied across these regions and across individuals, but the average coverage depth per genetic region and individual was 0.7×. Even though there is no draft genome sequence for *Pinus*, we were able to construct a well-defined reference sequence based on the consensus sequences of each of the highest quality contigs from the *de novo* assembly. This reference sequence could then be used to assemble all other, well-represented reads in the libraries, which is an important aspect of this study for two reasons. First, *de novo* assemblies of large Illumina sequencing projects are computationally intensive, in both time and RAM requirements, whereas in some software (including DNASTAR's SEQMAN NGEN 3.0), reference-based assemblies are very rapid and can also utilize hard disk space. Second, using only the high-quality contigs in the reference sequence insured that genomic regions that would be problematic in the assembly, variant calling, and other downstream analyses (e.g. highly repetitive regions, paralogous regions) are not included in the alignments. Such an approach is likely particularly critical for organisms such as pines that have large genomes and large amounts of repetitive DNA (Morse *et al.* 2009). Finally, by using two different restriction enzymes and sequencing only from the end of each fragment cut by *Eco*RI, the protocol led to alignments that were typically rectangular contigs, with all reads sequenced in the same orientation and beginning and ending at the same position.

Conclusions

The presence or absence of serotiny in pines has substantial fitness consequences, yet polymorphism for serotiny is retained in several pine species (Givnish 1981; Tapias *et al.* 2004; Moya *et al.* 2008). Serotiny has probably had multiple independent origins in pines inhabiting fire prone environments (Grotkopp *et al.* 2004). Additionally, the striking variation in the occurrence of serotiny among populations of lodgepole pine has extended consequences for the forest communities and ecosystems in which lodgepole pine is often the foundation species (Turner *et al.* 2003; Benkman & Siepielski 2004; Wymore *et al.* 2011). Consequently, there is considerable interest in advancing our understanding of the genetic complexity of this adaptation. Greater knowledge of the underlying genetics will influence our thinking about the maintenance of phenotypic variation in the context of the geographically variable fitness consequences of the trait. A more complete description of the genetics will also enable comparative genomics to resolve questions of multiple origins of serotiny and common genetic architectures of the cone traits responsible for serotiny in different species of pine and even other conifers (*Cupressus* spp.). This study identifies a set of loci statistically associated with serotiny in lodgepole pine and demonstrates the feasibility of genome-wide association mapping in conifers. The sequencing approach described here can be readily used to generate large amounts of data for further study of this trait. Future work involving higher coverage sequencing and screening a larger number of trees and mapped loci will further increase our understanding of the genetic architecture of serotiny.

Acknowledgements

This manuscript was improved by the comments of two referees and Subject Editor R. Petit. This research was funded by the National Science Foundation to CAB (DBI-0701757) and the University of Wyoming Foundation to CWB.

References

454 Life Sciences Corp. (2009) Using multiplex identifier (MID) adaptors for the GS FLX Titanium chemistry-extended MID set. Tech. rep., Technical Bulletin: Genome Sequencer FLX System.

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in snp chips affect measures of population divergence. *Molecular Biology and Evolution*, **27**, 2534–2547.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Arno SF (1980) Forest fire history in the northern rockies. *Journal of Forestry*, **78**, 460–465.

Ayliffe MA, Timmis JN (1992) Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. *TAG Theoretical and Applied Genetics*, **85**, 229–238.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, **63**, 221–230.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

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

Benkman CW, Siepielski AM (2004) A keystone selective agent? Pine squirrels and the frequency of serotiny in lodgepole pine. *Ecology*, **85**, 2082–2087.

Benkman CW, Holimon WC, Smith JW (2001) The influence of a competitor on the geographic mosaic of coevolution between crossbills and lodgepole pine. *Evolution*, **55**, 282–294.

Benkman CW, Parchman TL, Favis A, Siepielski AM (2003) Reciprocal selection causes a coevolutionary arms race between crossbills and lodgepole pine. *American Naturalist*, **162**, 182–194.

Benkman CW, Siepielski AM, Parchman TL (2008) The local introduction of strongly interacting species and the loss of geographic variation in species and species interactions. *Molecular Ecology*, **17**, 395–404.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of National Academy of Sciences of the United States of America*, **101**, 15255–15260.

Carlson CS, Eberle MA, Kruglyak L *et al.* (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**, 518-521.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, **15**, 1496–1502.

Cosart T, Beja-Pereira A, Chen S *et al.* (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, **12**, 347.

Craig DW, Pearson JV, Szelinger S *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, **5**, 887–893.

Critchfield WB (1980) Genetics of lodgepole pine. United States Forest Service Research Paper WO-37.

Cumbie WP, Eckert A, Wegrzyn J *et al.* (2011) Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of *Pinus taeda L. Heredity*, **107**, 105–114.

Dancik BP, Yeh FC (1983) Allozyme variability and evolution of lodgepole pine (*Pinus contorta* var *latifolia*) and Jack pine (*Pinus banksiana*) in Alberta. *Canadian Journal of Genetics and Cytology*, **25**, 57–64.

Dong JS, Wagner DB (1994) Paternally inherited chloroplast polymorphism in *Pinus* – estimation of diversity and population subdivision, and tests of disequilibrium with a maternally inherited mitochondrial polymorphism. *Genetics*, **136**, 1187–1194.

Doyle J (1991) DNA protocols for plants: a CTAB total DNA isolation. In: *Molecular Techniques in Taxonomy* (eds Hewitt GM, Johnston A), pp. 283–293. Springer, New York.

Eckert AJ, Bower AD, Wegrzyn JL *et al.* (2009a) Association genetics of coastal douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. cold-hardiness related traits. *Genetics*, **182**, 1289–1302.

Eckert AJ, Pande B, Ersoz ES *et al.* (2009b) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics & Genomes*, **5**, 225–234.

Eckert AJ, Wegrzyn JL, Pande B *et al.* (2009c) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*). *Genetics*, **183**, 289–98.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Epperson BK, Allard RW (1989) Spatial auto-correlation analysis of the distribution of genotypes within populations of lodgepole pine. *Genetics*, **121**, 369–377.

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.

Givnish TJ (1981) Serotiny, geography, and fire in the pine barrens of New Jersey. *Evolution*, **35**, 101–123.

Godbout J, Fazekas A, Newton CH, Yeh FC, Bousquet J (2008) Glacial vicariance in the Pacific Northwest: evidence from a lodgepole pine mitochondrial DNA minisatellite for multiple genetically distinct and widely separated refugia. *Molecular Ecology*, **17**, 2463–2475.

Gompert Z, Forister ML, Fordyce JA *et al.* (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, **19**, 2455–2473.

Gompert Z, Lucas LK, Nice CC *et al.* (2012) Reproductive isolation between two butterfly species evolved by divergent selection. *Evolution*, DOI:10.1111/j.1558-5646.2012.01587.x.

González-Martínez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. *Heredity*, **101**, 19–26.

González-Martínez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. *New Phytologist*, **170**, 227–238.

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics*, **175**, 399–409.

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL (2004) Evolution of genome size in pines (pinus) and its life-history correlates: supertree analyses. *Evolution*, **58**, 1705–1729.

Gupta P, Rustgi S, Kulwal P (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology*, **57**, 461–485.

Hall SE, Dvorak WS, Johnston JS, Price HJ, Williams CG (2000) Flow cytometric analysis of dna content for tropical and temperate new world pines. *Annals of Botany*, **86**, 1081–1086.

Heuertz M, De Paoli E *et al.* (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of norway spruce [*Picea abies* (l.) karst]. *Genetics*, **174**, 2095–2105.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, **6**, 95–108.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). *New Phytologist*, **188**, 501–514.

Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast dna into the nucleus. *Nature*, **422**, 72–76.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Joyner KL, Wang XR, Johnston JS, Price HJ, Williams CG (2001) DNA content for asian pines parallels new world relatives. *Canadian Journal of Botany*, **79**, 192–196.

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas-fir. *Genetics*, **171**, 2029–2041.

Krutovsky KV, Clair JBS, Saich R, Hipkins VD, Neale DB (2009) Estimation of population structure in coastal Douglas-fir [*Pseudotsuga menziesii* (mirb.) franco var. *menziesii*] using allozyme and microsatellite markers. *Tree Genetics & Genomes*, **5**, 641–658.

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

Li J, Li Q, Hou W *et al.* (2009) An algorithmic model for constructing a linkage and linkage disequilibrium map in outcrossing plant populations. *Genetics Research*, **91**, 9–21.

Lotan JE (1975) The role of cone serotiny in lodgepole pine forests. In: *Symposium Proceedings, Washington State University, Pullman, Washington, USA* (ed. Baumgartner DM), pp. 471–495.

Lynch M (2007) *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

McCarthy MI, Abecasis GR, Cardon LR *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics*, **9**, 356–369.

McVean G (2009) A genealogical interpretation of principal components analysis. *Plos Genetics*, **5**, e1000686.

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **2010**, DOI:10/1101/pdb.prot5448.

Morse AM, Peterson DG, Islam-Faridi MN *et al.* (2009) Evolution of genome size and complexity in *Pinus*. *PLoS One*, **4**, e4332.

Moya D, Saracino A, Salvatore R *et al.* (2008) Anatomic basis and insulation of serotinous cones in *Pinus halepensis*. *Trees – Structure and Function*, **22**, 511–519.

Neale DB (2007) Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development*, **17**, 539–544.

Neale D, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews. Genetics*, **12**, 111–122.

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

Nicholson G, Smith AV, Jonsson F *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **64**, 695–715.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Novembre J, Johnson T, Bryc K *et al.* (2008) Genes mirror geography within Europe. *Nature*, **456**, 98–101.

van Orsouw NJ, Hogers RCJ, Janssen A *et al.* (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

Parchman TL, Benkman CW, Jenkins B, Buerkle CA (2011) Low levels of population genetic structure in *Pinus contorta* (Pinaceae) across a geographic mosaic of co-evolution. *American Journal of Botany*, **98**, 669–679.

Perry DA, Lotan JE (1979) Opening temperatures of serotinous cones of lodgepole pine. Tech. rep., Forest Service, U.S. Department of Agriculture Research Note INT-228.

Quesada T, Vikneswaran G, Cumbie WP *et al.* (2010) Association mapping of quantitative disease resistance in a natural population of Loblolly Pine (*Pinus taeda* L.). *Genetics*, **186**, 677–686.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, **67**, 147–158.

Rudolph TC, Schoenike RE, Schantz-Hanzen T (1959) Results of one-parent progeny tests relating to the inheritance of open and closed cones in Jack pine. *Minnesota Forestry Notes*, **78**, 1–2.

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Tapias R, Climent J, Pardos J, Gil L (2004) Life histories of mediterranean pines. *Plant Ecology*, **171**, 53–68.

Teich AH (1970) Cone serotiny and inbreeding in natural populations of *Pinus banksiana* and *Pinus contorta*. *Canadian Journal of Botany*, **48**, 1805–1809.

Turner MG, Hargrove WW, Gardner RH, Romme WH (1994) Effects of fire on landscape heterogeneity in Yellowstone National Park, Wyoming. *Journal of Vegetation Science*, **5**, 731–742.

Turner MG, Romme WH, Gardner RH, Hargrove WW (1997) Effects of fire size and patterns on early succession in Yellowstone National Park. *Ecological Monographs*, **67**, 411–433.

Turner MG, Romme WH, Tinker DB (2003) Surprises and lessons from the 1988 yellowstone fires. *Frontiers in Ecology and the Environment*, **1**, 351–358.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP – a new technique for DNA-fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.

Wheeler NC, Guries RP (1982) Biogeography of lodgepole pine. *Canadian Journal of Botany*, **60**, 1805–1814.

Wymore AS, Keeley ATH, Yturralde KM *et al.* (2011) Genes to ecosystems: exploring the frontiers of ecology with one of the smallest biological units. *New Phytologist*, **191**, 19–36.

Yeh FC, Cheliak WM, Dancik BP *et al.* (1985) Population differentiation in lodgepole pine, *Pinus contorta* spp *latifolia* – a discriminant-analysis of allozyme variation. *Canadian Journal of Genetics and Cytology*, **27**, 210–218.

Yu J, Pressoir W, Briggs W *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Yuan Q, Hill J, Hsiao J *et al.* (2002) Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast dna insertion. *Molecular Genetics and Genomics*, **267**, 713–720.

T.P. is a postdoctoral scientist at the University of Wyoming whose research involves population genomics and evolutionary ecology of pines and a wide range of other organisms. C.B. is a professor at the University of Wyoming and has research interests in species interactions, coevolution, and the evolutionary ecology of crossbills (Aves) and conifers. Z.G. is a postdoctoral scientist at Texas State University with interests in the genetics of speciation and hybridization, and Bayesian methods in population genetics and genomics. F.S. and J.M. are research scientists at the National Center for Genomic Resources in Santa Fe, NM, with interests in applications of next generation sequencing technologies to genome biology. A.B. is an associate professor at the University of Wyoming with interests in evolutionary genetics, hybridization, and speciation.

Data accessibility

Library preparation protocol, C code for association mapping model, Reference sequence, and .bam Assembly: 10.5061/dryad.m2271pf1.