H.D. Marshall · C. Newton · K. Ritland

Chloroplast phylogeography and evolution of highly polymorphic microsatellites in lodgepole pine (*Pinus contorta*)

Received: 7 December 2000 / Accepted: 17 May 2001

Abstract We employed a novel set of six highly polymophic chloroplastic simple sequence repeat (cpSSR) loci to investigate the phylogeography of lodgepole pine (*Pinus contorta* Dougl. Ex. Loud.), and to examine aspects of the evolutionary process operating on these repetitive DNA sequences. Chloroplast haplotypes of 500 trees, sampled throughout the range of lodgepole pine, were determined. We found a marked association of genetic distance with physical distance within the scale of 0 to 1,000 km, but no association beyond that range. Likewise, geographic clustering was observed only among recent clades in a dendrogram. These phylogeographic patterns are consistant with a rapid rangewide expansion ("big-bang") followed by recent, local population differentiation ("galaxy formation"). In support of this expansion, coalescent simulations of the genealogical process gave a long-term effective population size in the low thousands, and a time to common ancestry of about 1,500 generations (12,000 years), consistent with a post-Pleistocene population expansion as documented by previous pollen-sediment analyses. Two lines of evidence (mapping mutational events onto a phylogeny, and evaluation of observed versus expected gene diversity) suggest that five of the cpSSR loci evolve primarily by a stepwise model of evolution of single repeat changes (but with a small proportion of changes involving two or more repeats), and the coalescent simulations point to a mutation rate of about 10^{-3}.

Communicated by H.C. Becker

H. Dawn Marshall (✉) · K. Ritland
Department of Forest Sciences, University of British Columbia,
2424 Main Mall, Vancouver, B.C. V6T1Z4, Canada
e-mail: dawnm@mun.ca
Tel.: +1-709-737-4713, Fax: +1-709-737-3018

C. Newton
BC Research Incorporated, 3650 Wesbrook Mall, Vancouver,
B.C. V6S2L2, Canada

Present address:
H. Dawn Marshall, Department of Biology,
Memorial University of Newfoundland, St. John's A1B 3X9, Canada

Keywords Chloroplast · Microsatellites · Lodgepole pine · Phylogeography · Mutational model

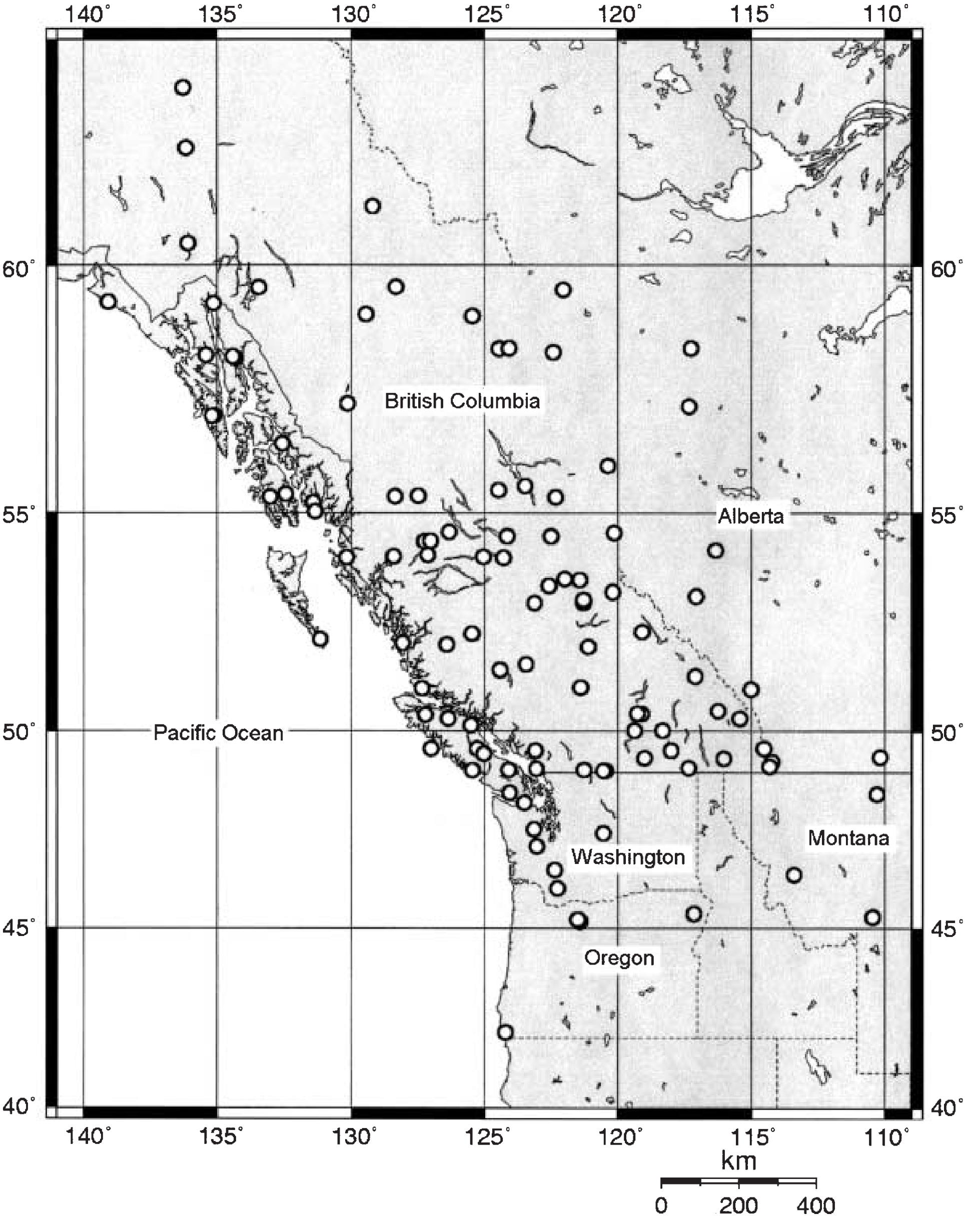
Introduction

Simple-sequence-repeat (SSR) loci provide superior information about the geographical distribution of genetic variation, and about population size and history (Wilson and Balding 1998). Having such markers on non-recombining chromosomes can make them even more informative, as ancestral haplotypes can be reconstructed with parsimony methods. In humans, SSR variation of the non-recombining human Y chromosome has been used to construct phylogenetic trees that indicate human origins (Seielstad et al. 1999) and the peopling of the New World (Torroni et al. 1994; Ruiz-Linares et al. 1999).

In plants, the presence of variable mononucleotide repeats in chloroplast DNA was first demonstrated by Powell et al. (1995a) in pine and by Powell et al. (1995b) in soybean. Vendramin et al. (1996) designed primers for species in the pine family (Pinaceae) that amplified 20 such chloroplast microsatellites; an additional four primer pairs were identified by Hamilton (1999). Primer pairs were also recently designed for dicotyledonous angiosperms (Weising and Gardner 1999). Chloroplast SSRs have provided inferences about the recent evolutionary history in some tree species, usually about postglacial migration routes and the locations of refugia (Newton et al. 1999).

However, while such SSR primers targeted to mononucleotide repeats may serve as general tools to study chloroplast variation, their levels of variability are usually well below that found for nuclear SSRs. Recently, however, there are several reports of SSR multinucleotide-repeat loci in chloroplasts that show greater variability. Stoehr et al. (1998) found a single chloroplast locus in Douglas-fir that yielded 13 alleles in a sample of 20 individuals, with band sizes ranging from 859 to 1,110 base pairs. Likewise, higher levels of variability were discovered from chloroplast microsatellites characterized

Fig. 1 Locations of provenances sampled (indicated by circles)



in the genus *Abies* (Vendramin et al. 1999). In lodgepole pine (*Pinus contorta* Dougl.) a set of six hypervariable, multinucleotide-repeat cpSSR markers was recently characterized (Stoehr and Newton 2001).

Lodgepole pine is a dominant tree species on over 2.6×10^7 ha of forest land in western Canada and the United States, with a climactically and edaphically diverse natural range encompassing 33° of latitude, 35° of longitude and 3,400 m of elevation (Fig. 1). Taxonomically a complex of four subspecies (Critchfield 1980), lodgepole pine is commonly considered to consist of an inland variety (ssp. *latifolia*) and a coastal variety (ssp. *contorta*) sometimes termed "shore pine". It is a wind-pollinated outcrosser, with abundant seed production and dispersal, rapid juvenile growth, and serotinous cones. It is an aggressive pioneer and its present range is thought to have been colonized following the last glacial stage in a northward migration (MacDonald and Cwynar 1991).

In this study, we use the highly polymorphic cpSSR markers of Stoehr and Newton (2001) to examine the phylogeographic structure of lodgepole pine (spp. *latifolia*), and to make inferences about its evolutionary history. The genealogical content of choroplast haplotypes, as compared to more traditional markers, allows us to address the hypothesis that the present-day distribution of genetic variation in lodgepole pine represents a historical range-wide expansion superimposed upon recent, local population differentiation. We also examine the evolution of repetitive sequences by evaluating whether levels of variation correspond to expectations under a stepwise mutation model, and by mapping mutational events onto a phylogeny. We are therefore able to assess the phylogeographic utility of chloroplastic simple sequence repeats.

Material and methods

Bud or needle tissue was sampled from between one and four trees from each of 119 provenances, originally sampled from throughout the species range (Fig. 1) and grown at the Prince George Tree Improvement Station (PGTIS) just south of Prince George, British Columbia. In addition, three provenances were intensively sampled (sample sizes of 68, 71 and 82). Tissue was frozen at –80°C for transport, then DNA was extracted from the bud tissue using standard CTAB (cetyl trimethylammonium bromide) procedures (Doyle and Doyle 1987). All purified DNAs were stored at –20°C and the yield ranged from 0.1 to 1.0 µg of DNA. DNA was diluted to 10–20 ng/µl into sets of 96-well microtitre plates (available to others upon request to C.N.). The six primer pairs given in Stoehr and Newton (2001) were used to amplify these DNAs. For multiplex cpDNA amplification, 1 µl of total DNA (10–20 ng/µl) was used in 6.25 µl reactions containing 1×Ultratherm buffer (Eclipse Biochemicals), 1.5 mM of $MgCl_2$, 200 µM each of dGTP, dTTP, dCTP, 20 µM of dATP, 0.1 µCi of α^{32}P-dATP (3,000 Ci/mmole, Amersham), 0.1 U of Ultratherm DNA polymerase and 0.025 U of Vent DNA polymerase (New England Biolabs), and 0.5 µM of total amplification primers. The primers were a mixture of six primer pairs specific for polymorphic regions in the lodgepole chloroplast genome: ($G_{2.1}/R_1$) F 5' AGATCGGGACAATGT-ATGCC, R 5' TGTCCTATCCATTAGACGAT; (9F/87R) F 5' ACTGCAAGGAACAGTAGAAC, R 5' CGGAACGTTTTCT-GATGCAC; (10F/RR) F 5' CAGAAGCCCAAGCTTATGGC, R 5' CGGATTGATCCTAACCATAC; (69F/R) F 5’ TTTCGGGC-TCCACTGTTATC R 5' CGTACTCAATTTGTTACTAC; (L_2/T_1) F 5' ACCAATTCCGCCATATCCCC, R 5' CTAGGGGAGGAT-AATAACATTGC; (I_1/A_2) F 5' TTCAAGTCCAGGATAGCCCA, R 5' CTACCAACTGAGCTATATCC. Reactions were started at 95°C for 3 min followed by 25 cycles of 94°C (30 s), 55°C (30 s) and 72°C (60 s) in a 96-well format MJ Research PT-100 thermocycler. Reactions were then diluted with 2 vol of 98% formamide, 1 mM EDTA, 0.1% each of bromophenol blue and xylene cyanol, denatured at 95°C, and loaded directly onto 4.5% polyacrylamide gels containing 8 M urea. Following electrophoresis, gels were dried and exposed to autoradiography using Biomax MS film (Kodak) for 6–48 h. Size variants at each of the six polymorphic loci were scored manually by co-migration with known standards placed every 5–10 lanes per sequencing gel. Haplotypes for each cpDNA source were assigned by arranging the observed cpDNA variant types (A, B, C, etc.) at each locus in their linear order predicted from the complete black pine cpDNA sequence (Wakasugi et al. 1994).

To characterize phylogeographic structure, we first computed pairwise genetic distances among haplotypes, using the distance measure of Goldstein et al. (1995), which assumes a stepwise mutation model. Because the repeat length differs among our SSR loci, distances for each locus were normalized by the variance of allele length at that locus. Two analyses were then conducted. First, genetic distances were plotted as a function of physical distance. Second, a dendrogram of haplotype relationships was constructed via the unweighted pair group method. Then, mean physical distance within haplotype groups was computed at each level of the phylogeny. These mean physical distances within groups were all standardized by the mean physical distance among all haplotypes, irrespective of group. Values less than one indicate significant grouping with respect to geography. The significance of such an association was determined by randomizing haplotypes with respect to geography, and computing mean pairwise distance using the same grouping; if the mean physical distance in such randomized data was greater than the original distance in at least 95 of 100 randomizations, the association was deemed significant.

To characterize the evolution of haplotypes, we reconstructed the phylogenetic relationships among haplotypes using the parsimony approach in PAUP* (Swofford 1998). Each locus with the exception of $G_{2.1}/R_1$ was treated as an ordered character with each repeat representing one step (alleles differing by one repeat are one step apart, those by two repeats are two steps apart, etc.). Weights among states of $G_{2.1}/R_1$ were encoded using a step matrix whereby transitions between states are weighted as the sum of the number of 10-base repeats between them and the number of 1-base repeats between them. The purpose of ordering or weighting the characters was to attempt to recreate the constraints that would be imposed during evolution by a stepwise model of mutation, as may be expected to hold for sequence-repeat data (Shriver et al. 1993). The heuristic-search algorithm was employed, and a 50% majority rule unrooted consensus of the first 1,000 equally most-parsimonious trees was chosen to represent the phylogeny.

To draw inferences about effective population size, we employed a Bayesian version of the Markov Chain Monte Carlo (MCMC) algorithm, as proposed by Wilson and Balding (1998), which finds estimates using a coalescent model of genealogy, a stepwise model of mutation, and a prior distribution of mutation rates and effective population sizes. With these priors, one can simultaneously estimate the effective population size (N_e), mutation rate (μ),$\theta = N_e\mu$, and the time since the most-recent common ancestor (TMRCA). We used their program MICSAT (Wilson and Balding 1998), available at http://www.maths.abdn.ac.uk/~ijw/. The procedure uses the entire haplotype complement of loci; however, we omitted the complex locus $G_{2.1}/R_1$ because simple stepwise mutation is unlikely at this locus (see below). We used uniform priors of mutation rate and effective population size, and 10,000 samples were taken during the MCMC process. The output gives probability distributions for these parameters.

The mutation model for SSR evolution was tested in two ways. First, we compared the distribution of allele sizes for each locus with a haplotype phylogeny obtained using all six loci. For the sake of simplicity, a subset of 44 haplotypes (those that occur more than twice) was chosen for this task, and each locus was left unordered. As before, the first 1,000 shortest trees were retained, and a 50% majority rule unrooted consensus was calculated. Second, the computer program BOTTLENECK (Cornuet and Luikart 1996) was used to generate expected heterozygosity (in the sense of gene diversity; Nei 1986) distributions, given the sample size and the number of alleles observed, for each locus under each of three models of evolution: the infinite alleles model (IAM), the stepwise mutation model (SMM) and the two phase model (TPM). Observed heterozygosity was then compared with expected for each model to determine the most appropriate model for each locus.

Results

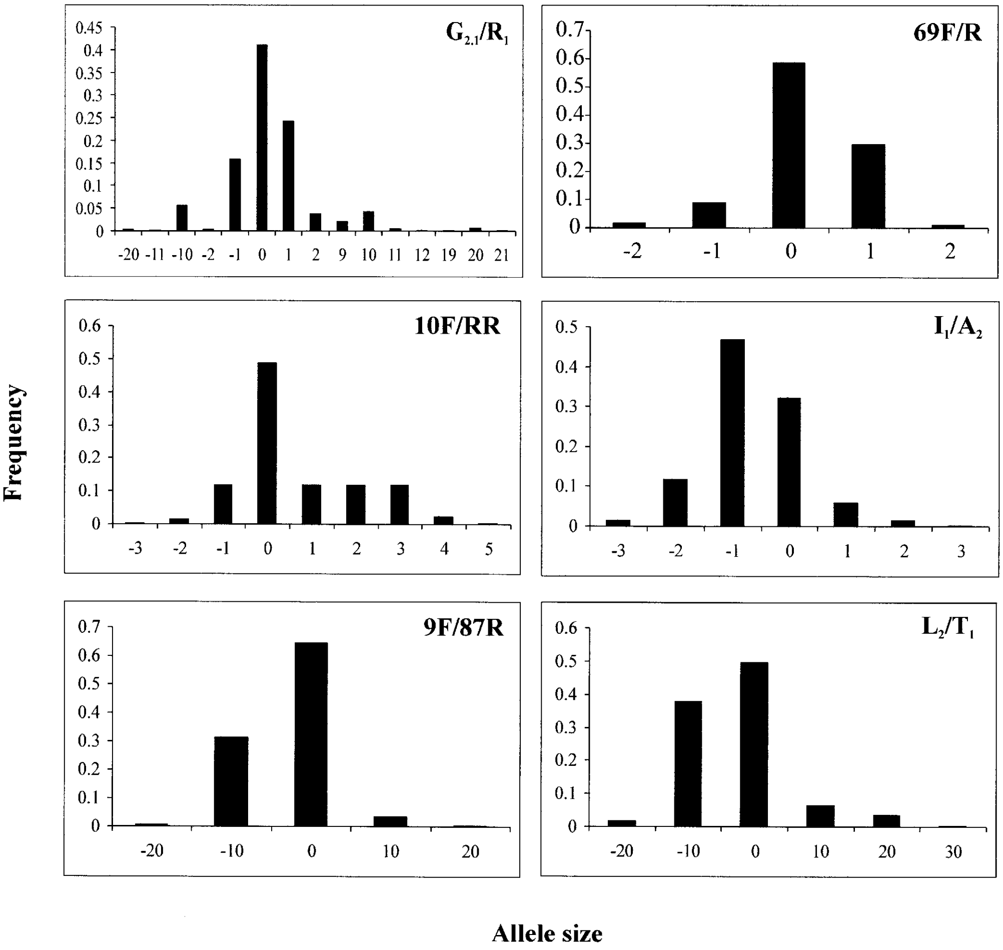
Sequence-repeat profiles

Between five and 17 alleles (size variants) were identified at each locus, and the distribution of allele sizes was largely unimodel with the mode near the middle (Fig. 2). Per-locus gene diversity ranged from 0.487 (9F/87R) to 0.739 ($G_{2.1}/R_1$). A total of 49 alleles at the six loci defined 204 haplotypes, of which 132 occurred in single individuals. The ten most-frequent haplotypes occurred in 10 to 29 individuals each and are not specific to provenance. Furthermore, many of the common haplotypes occur in each of the three intensely sampled provenances.

Geographic distribution of genetic variation in lodgepole pine

There was a marked association of genetic distance with physical distance within the scale of 0 to 1,000 km (Fig. 3). This association was also quite linear (Fig. 3). Genetic distance increased from 0.10 to 0.15 over this span, then stayed constant at about 0.15–0.17 over longer distances.

Fig. 2 Frequency distribution of polymorphic alleles at six lodgepole pine cpDNA loci. Columns in each frame indicate the frequency of each allele. One allele was arbitrarily set to zero, and + or – indicates the number of base pairs size difference greater or less than this allele



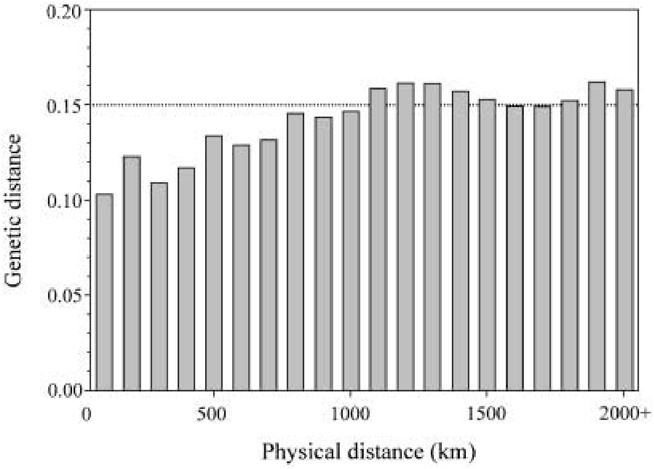


Fig. 3 Genetic distance as a function of physical distance, plotted in intervals of 100 km

The randomization test at any genetic distance less than about 0.15 indicated a significant association at that distance interval (dotted horizontal line in Fig. 3; the relationship between physical distance and genetic distance decreased by 0.00005 genetic-distance units every 1 km, up to 1,000 km with $P<0.00001$).

The dendrogram of relatedness among haplotypes (Fig. 4) showed a fairly continuous clustering, with two large clusters plus an outlier at the deepest levels of the dendrogram. No association of the two large clusters with a region of the distribution was observed. Figure 5 combines the isolation by distance analysis (Fig. 3) with the dendrogram analysis (Fig. 4). This figure shows the strength of the associations of groupings with geography, at each level of clustering or “depth of dendrogram”. Smaller values on the Y-axis indicate a stronger association of clades with geography. At the most-recent level of the dendrogram, associations are moderate but variable (0.95–1.01) and not statistically significant because of the small number of groups present at this level. At the level where clustering occurs with a genetic distance of about 0.2, the association is the strongest (0.95) and statistically significant. At the 0.4 clustering level, associations become dramatically weaker (0.98–0.99) but are still significant because of the larger number of groups at this level of the dendrogram. At the 0.6 level and above, there were few significant associations.

Parsimony reconstruction of haplotype evolution

The phylogenetic relationships among the 204 haplotypes are presented in Fig. 6 as the 50% majority rule unrooted

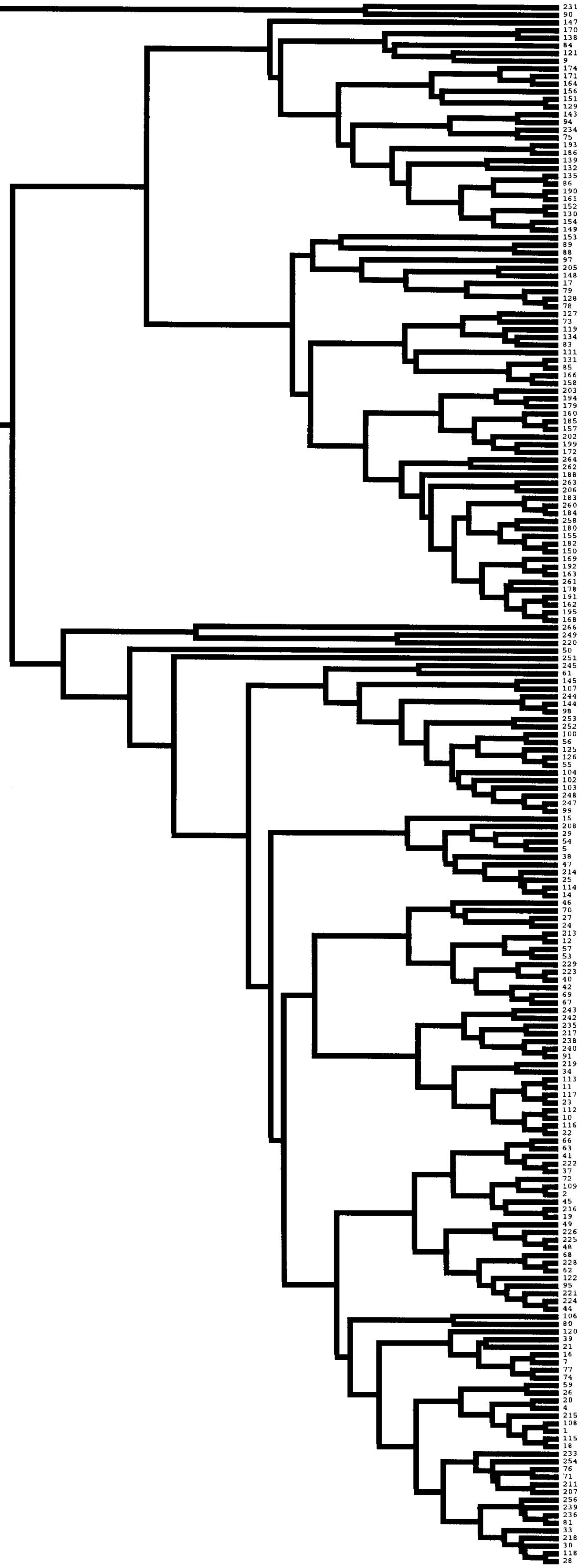


Fig. 4 Dendrogram of haplotypes, based by UWPM clustering of the matrix of Goldstein et al. (1996) distances

consensus of the first 1,000 shortest trees, each 251 steps in length with a consistency index (CI) of 0.345. The tree is poorly resolved, and mapping the phylogenetic locations of the three intensely sampled provenances reveals no obvious association of clade with geographical location. We also mapped the most northern and most southern of the unique haplotypes (those occurring only once in the data) onto the tree. The more-frequent haplotypes occur both basally and terminally in the tree, and most haplotypes differ from their closest relative by one mutational difference. Size homoplasy is evident from the large tree length (211 steps in excess of what would be required if each character state at each locus were attained only once on the tree), low CI, and poor resolution of our tree, and the very large number of equally most-parsimonious resolutions of the data (>10,000).

Genealogical inferences about population parameters

The probability density functions for the parameters μ, θ, N_e, and TMRCA generated using 10,000 simulated

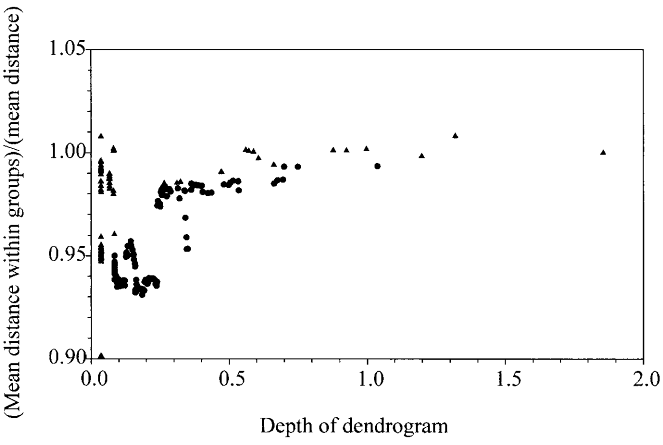


Fig. 5 Mean pairwise distance within groups, plotted as a function of dendrogram depth

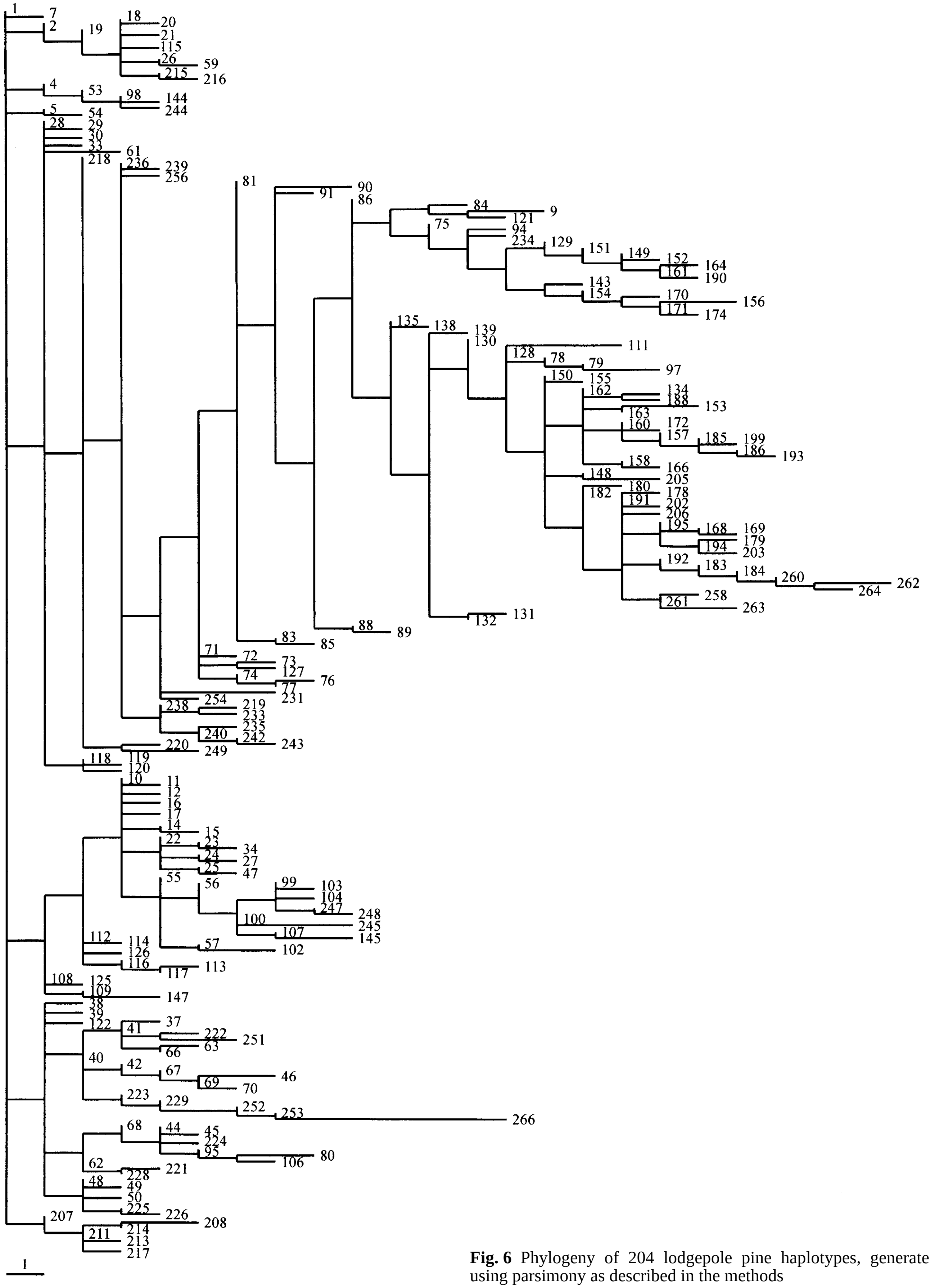


Fig. 6 Phylogeny of 204 lodgepole pine haplotypes, generated using parsimony as described in the methods

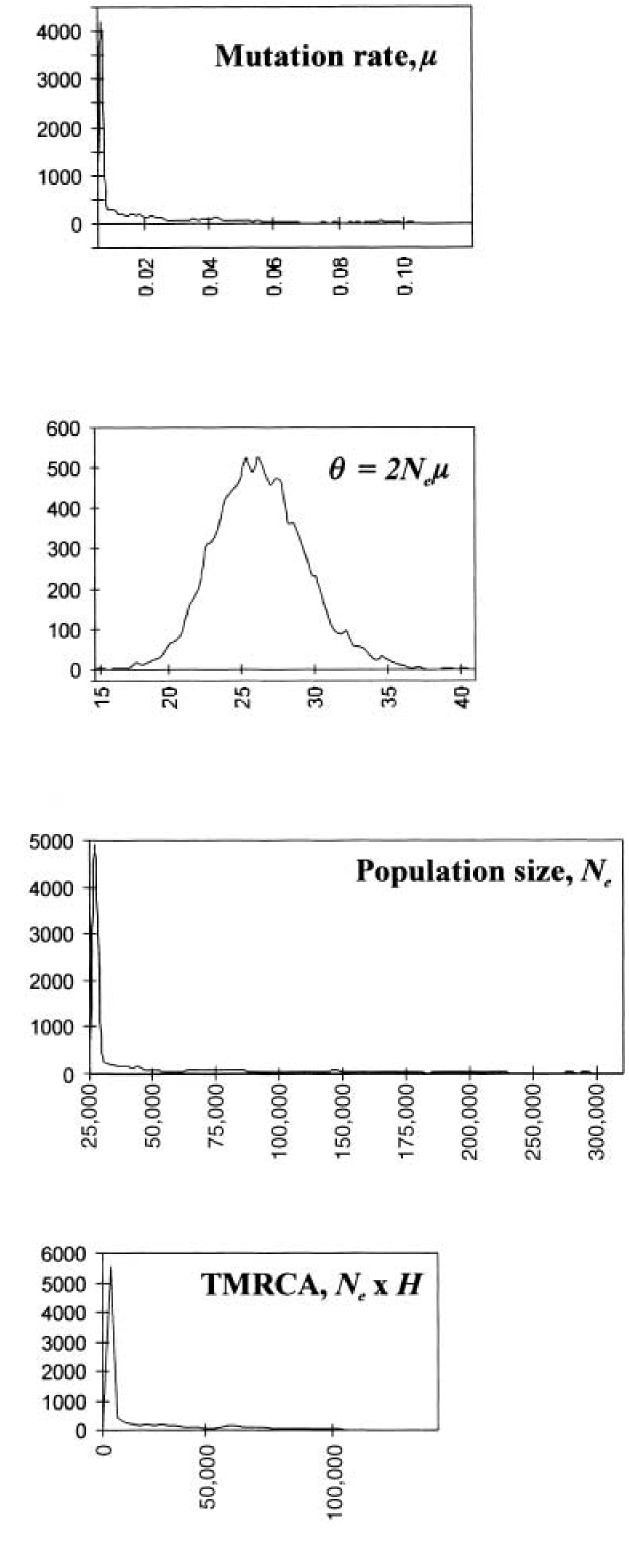


Fig. 7 Probability distributions of N_e, θ, μ and TMRCA generated for lodgepole pine chloroplast sequence repeats, using the Bayesian inference method of Wilson and Balding (1998)

samples of the genealogical process for the lodgepole five-locus profiles are presented in Fig. 7. The median and 95% probability interval values for each are 0.00371 (0.0000477, 0.0882), 24.6 (18.8, 31.7) 3,190 (146, 234,000) and 1,556 (36, 158,000), respectively. Although the numerical interval values suggest a lack of precision regarding these estimates (except in the case of θ), the density functions show fairly sharp modes for each. The discrepancy is in part due to limited support for a large range of high values of each parameter. More confidence in the θ estimate reflects the fact that under the standard coalescent model information about N_e and μ is available from allelic data only through their product (θ=2$N_e\mu$), making postdata inferences about θ more robust (Wilson and Balding 1998) than either N_e or μ individually. Nonetheless, it can be useful to distinguish the two for comparative purposes. The results are therefore consistent with a mutation rate of approximately 10^{-3} for lodgepole chloroplastic sequence repeats, a long-term effective population size in the low thousands, and a time to common ancestry of about 1,500 generations. The θ value corresponds to a mean heterozygosity of H_e=0.859 under a stepwise mutation model according to the formula H_e=1–1/(1+4N_e)$^{-1}$.

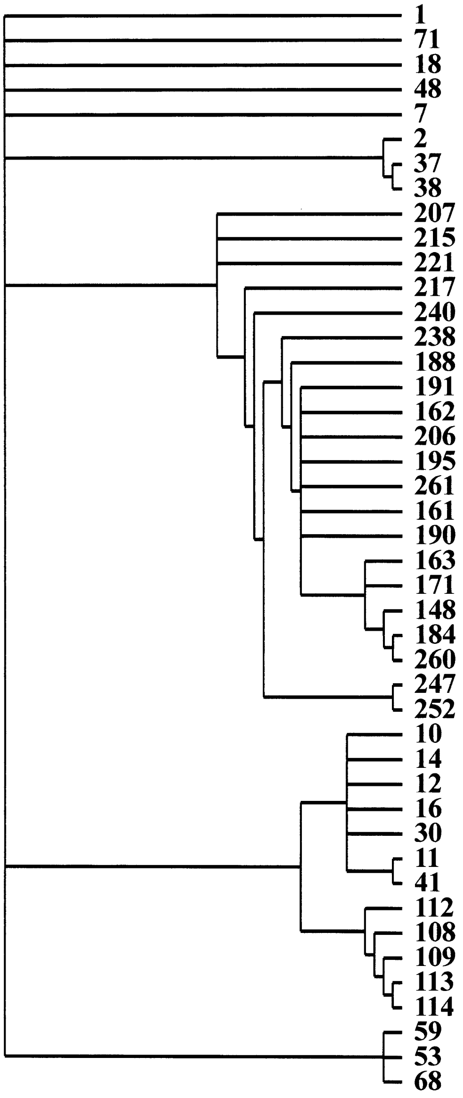
Evolution of chloroplast sequence repeats

The distribution of allele sizes on the phylogeny of 44 haplotypes (Fig. 8) suggests that most alleles were generated from alleles that differ by one size unit. For example, at locus 9F/87R one 20-to-10-bp change defines a large clade, and two separate 20-to-30-bp changes define individual taxa, comprising a homoplasy at this locus. At each of the other four simple loci, at least one two-or-more repeat difference needs to be invoked. At the complex locus $G_{2.1}/R_1$, seven of the 13 changes involve single repeats, while the other six require a combination of one and 10-bp changes. These observations are consistent with a stepwise model of evolution involving primarily single repeat changes but incorporating a proportion of changes involving two or more repeats (a TPM) for the five simple loci. The results of the BOTTLENECK simulations, presented in Table 1, corroborate this. The observed heterozygosity for each of these loci is most similar to that expected under the TPM. For locus $G_{2.1}/R_1$, the best fit appears to be to the IAM, a result that is also not inconsistent with the parsimony analysis.

Discussion

Lodgepole pine has been intensively studied genetically from three perspectives. First, it has been the subject of numerous provenance trials, which demonstrate marked elevational and latitudinal patterns of adaptive differentiation for morphometric traits (Wheeler and Guries 1982; Ying and Liang 1994) and for growth and survival traits (Ying 1991; Xie and Ying 1995). Second, a complement of population surveys documenting the genetic variation displayed by neutral markers such as isozymes (Yeh and Layton 1979; Dancik and Yeh 1982; Yeh et al. 1985; Epperson and Allard 1989) and mitochondrial DNA (Dong

Fig. 8 Phylogeny of 44 most-frequent lodgepole haplotypes, with the allelic state mapped for each locus



	$G_{2.1}/R_1$	9F/87R	L_2/T_1	10F/RR	69F/R	I_1/A_2
1	20	20	20	3	2	3
71	20	20	10	3	2	3
18	20	20	20	4	2	3
48	20	20	20	3	0	3
7	20	30	20	3	2	3
2	19	20	20	3	2	3
37	19	20	20	3	1	3
38	11	20	20	3	1	3
207	20	20	20	3	2	2
215	20	20	10	3	2	2
221	20	20	20	3	1	2
217	20	20	20	1	2	2
240	20	20	10	1	3	2
238	20	20	10	2	3	2
188	21	20	10	2	3	6
191	21	10	10	2	3	6
162	21	10	10	2	3	5
206	21	10	0	2	3	6
195	21	10	10	1	3	6
261	21	10	10	2	3	7
161	30	10	10	2	3	5
190	30	10	10	2	3	6
163	22	10	10	2	3	5
171	29	10	10	1	3	5
148	20	10	10	2	2	5
184	22	10	10	1	2	6
260	22	10	10	1	2	7
247	20	20	40	2	2	2
252	20	20	30	2	2	1
10	20	20	20	2	2	3
14	11	20	20	2	2	3
12	21	20	20	2	2	3
16	20	30	20	2	2	3
30	20	20	20	2	3	3
11	19	20	20	2	2	3
41	19	20	20	2	1	3
112	20	20	20	2	2	4
108	20	20	20	3	2	4
109	19	20	20	3	2	4
113	19	20	20	2	2	4
114	11	20	20	2	2	4
59	20	20	30	5	2	3
53	21	20	30	3	2	3
68	20	20	30	4	1	3

Table 1 Comparison of observed and expected per-locus gene diversity under three mutation models

Locus	Observed	IAM			TPM			SMM		
	H_e	H_e	SD	DH/SD[a]	H_e	SD	DH/SD	H_e	SD	DH/SD
$G_{2.1}/R_1$	0.739	0.731	0.104	0.077	0.820	0.049	–1.668	0.878	0.025	–5.408
9F/87R	0.487	0.421	0.186	0.352	0.531	0.146	–0.301	0.639	0.105	–1.186
10F/RR	0.659	0.524	0.173	0.780	0.642	0.116	0.142	0.747	0.069	–1.062
69F/R	0.561	0.41	0.194	0.777	0.524	0.148	0.246	0.641	0.102	–0.516
L_2T_1	0.604	0.469	0.186	0.722	0.591	0.124	0.105	0.704	0.083	–0.919
I_1A_2	0.707	0.604	0.149	0.686	0.714	0.089	–0.081	0.804	0.046	–1.873

[a] DH/SD is the observed less the expected gene diversity, divided by the standard deviation of the expected gene-diversity distribution

and Wagner 1993) have been conducted for lodgepole pine (discussed below). Third, lodgepole pine has served as a model organism to study the influence of genetic forces such as linkage and epistasis (Epperson and Allard 1987) and natural selection (Yang et al. 1996). Our current study adds a number of findings to these bodies of study.

Biogeography and evolutionary history of lodgepole pine

The genetic structure of populations represents the influences of both common ancestry and current gene exchange, and the concept of phylogeography attempts to simultaneously elucidate each by examining the geographic distribution of genealogical lineages (Schaal et al. 1998; Newton et al. 1999). Although the techniques and utility of phylogeography are well established in animals, plant studies have received limited attention from this approach, in part due to a paucity of appropriate genetic markers (Schaal et al. 1998). Typically, the ideal molecular marker for animal studies has been the mitochondrial DNA, which demonstrates a rapid rate of nucleotide substitution in combination with non-recombining maternal inheritance. Mitochondrial DNA is less appropriate for plant studies as its evolution involves slower nucleotide substitution rates and large structural rearrangements (Schaal et al. 1998), but the alternative non-recombining choroplast molecule is currently coming into focus for phylogeographic purposes. Recent examples include the use of chloroplast SSRs to infer the presence of refugia and re-colonization routes in maritime pine (Vendramin et al. 1998), black alder (King and Ferris 1998) and white oak (Dumolin-Lapegue et al. 1997).

Here, we demonstrate the use of novel hypervariable paternally inherited chloroplast repeat sequences to determine the phylogeographic structure of lodgepole pine throughout its range, and compare the outcome of this method with more traditional studies of the distribution of genetic variation in this species. Because lodgepole pine comprises large populations of wind-pollinated outcrossers with the potential for long-distance gene exchange via pollen and seed dispersal, one would not expect to observe population subdivision on the basis of ongoing genetic drift or reduced gene flow. Additionally, for allozymes the relatively undifferentiated status of conifers in this range has been attributed to a lack of significant barriers to gene flow and to the relatively low variability of allozymes (Yeh et al. 1985). Nonetheless, we found significant geographic structure within 1,000 km for the chloroplast SSRs.

Although isozyme studies have generally demonstrated that lodgepole populations, like those of other conifers, are largely unstructured at neutral marker loci (Yeh and Layton 1979; Dancik and Yeh 1982; Epperson and Allard 1989), Yeh et al. (1985) reported an association of genetic variation with geography in which both latitudinal and altitudinal gradients were apparent. In particular, northern populations demonstrated a greater degree of inter-population differentiation than southern ones. Furthermore Dong and Wagner (1993) documented significant population differentiation at maternally inherited mitochondrial DNA marker loci. Yeh et al. (1985) attributed their patterns of clinal genetic variation in lodgepole pine to the contribution of different factors including genetic drift, divergent and balancing selection, and recent historical events such as migration from ancestrally divergent populations. While patterns of selection are generally restricted to certain loci, both genetic drift and historical influences should be detectable at a genome-wide level and therefore should be apparent in different genetic systems. In this context, a phylogeographic analysis of hypervariable chloroplast haplotypes may reveal the presence of structure due to historical events underlying a superficial pattern of ongoing high gene flow and currently large population sizes.

Like us, Vendramin et al. (1999) also found significant associations of genetic and physical distance for cpSSRs in silver fir (*Abies alba*). They suggested that the organization of levels of allelic richness across the range of silver fir was likely to have been shaped by the distribution of refugia during the last glaciation and the subsequent re-colonization processes. In the case of lodgepole pine, colonization of the current range is thought to have taken place following the last ice age in a northward migration from refugia located south of the continental glacial limits (MacDonald and Cwynar 1985). The expansion involved a time period of about 12,000 years and a range of 2,200 km, and probably proceeded by the founding of small populations via long-distance dispersal beyond a front such that current peripheral northern populations are the smallest and youngest (for example, the northernmost population at Gravel Lake, Yukon, is less than 100 years old; Cwynar and MacDonald 1987). The results of Yeh et al. (1985) may therefore be explained by reduced allelic diversity in these most recent, founding populations. Furthermore, high apparent gene flow would be expected in populations that are not in genetic equlibrium following a range extension, due to the sharing of ancestral haplotypes.

Phylogeographically under such a scenario, one might expect a relatively weak phylogenetic structure because founding haplotypes will occur in two or more populations. This situation was observed in the lodgepole pine phylogeny (Figs. 4 and 6). Additionally, some of the older or more-basal haplotypes should be restricted to the ancestral pre-expansion area, while the younger or more-derived haplotypes will be geographically widespread (Cann et al. 1987; Templeton et al. 1995). Although a pattern like this is difficult to demonstrate without a statistical procedure such as nested cladistic analysis (Templeton et al. 1995), there does seem to be a trend in this direction in the lodgepole pine phylogeny, in that many of the haplotypes found only at the most southern latitudes appear basally in the tree (Figs. 4 and 6).

However, the association of clades with geography was strongest for relatively young clades (Fig. 5), indicating that most phylogenetic structure has emerged relatively recently, within approximately 10–30% of the time since the common ancestor of all haplotypes. Assuming this time to be 12,000 years ago (see below), this indicates that most associations with geography have arisen in the past 1,000–3,000 years. This is also consistent with the isolation by distance pattern (Fig. 3) wherein populations only within 1,000 km show an association. This overall pattern accords with the scenario that lodgepole pine went through a “big bang” of population expansion after the end of glaciation, wherein populations were initially spread evenly across the range, with “galaxies” of geographic association developing relatively recently. Given this pattern one may expect stronger structure at the northernmost edges of the range where populations have undergone recent bottlenecks, or at the southernmost edge where populations have had longer periods to develop structure. The association between genetic and physical distances was not observed to be stronger at northern versus southern latitudes (data not shown) but the power of such comparisons is low.

Genealogical inferences about population size and history

Using both simulations and real data, Wilson and Balding (1998) point to the difficulties inherent in recovering the true tree from a set of linked microsatellite loci evolving in a stepwise manner. In humans, for example, a phylogeny of Y-chromosome microsatellite haplotypes from Nigeria, Sardinia and East Anglia failed to recover clades of more than six haplotypes from a single location. Reasonably robust estimates of population parameters such as θ and TMRCA on the other hand can be obtained from such a data set, and are equally as informative about certain aspects of evolutionary history as trees. For humans Wilson and Balding (1998) were able to detect a long-term effective population size in the low thousands, in agreement with previous genetic analyses on autosomal loci and mitochondrial DNA. Their TMRCA estimate of 30,000 years was low relative to previous estimates, but more precise than one based on 15 kb of Y-chromosome sequence, despite the limitations imposed by recurrent mutations.

Similar methodology, applied to the lodgepole pine chloroplast data, produced a long-term effective population size estimate also in the low thousands and a TMRCA of about 1,500 generations. Assuming a generation time for lodgepole pine of 80 years, this TMRCA corresponds to 12,000 years ago which fits perfectly with expectations based on colonization history. The low long-term effective population size, well below current lodgepole pine population sizes, is indicative of a historical population size expansion such as one would expect to accompany a range extension. Following each founding event during a northward migration, population sizes would be expected to increase as long as environmental conditions are favourable. Pollen analysis has documented this process in lodgepole pine (MacDonald and Cwynar 1991). Our data therefore provide further evidence of a post-glacial range extension accompanied by population growth as part of the biogeographical explanation for the current range and diversity of lodgepole pine populations in western North America.

Evolution of chloroplast sequence repeats

Some of the analyses performed in this study have relied upon the assumption that the chloroplast sequence repeats comprise a series of linked haplotypic loci, each of which evolves according to a stepwise model of mutation (with the exception of the complex locus $G_{2.1}/R_1$). The stepwise mutation model (or SMM) is perhaps the simplest plausible model usually invoked to explain mutational changes in repeat number at microsatellite loci (1–6 bp repeats). Under this model the repeat number increases or decreases by a single unit with equal probability and changes of more than one unit do not occur. In a study of different VNTR (variable number of tandem repeats) loci, Shriver et al. (1993) showed the SMM to be appropriate for repeat units of 3–5 bp and approximately appropriate for 1–2 bp repeats, but that 15–70 bp repeats (minisatellites) deviated in the direction of an infinite-alleles model (or IAM) which stipulates that each mutation gives rise to a new allele not already present in the population. Recently, a two-phased model of mutation (TPM) has been proposed. The TPM is intermediate to the SMM and IAM in that it operates like the SMM but allows for a proportion of changes involving more than one repeat unit. Most microsatellite data sets are now thought to better fit the TPM than the SMM or IAM (Di Rienzo et al. 1994).

Knowledge of the patterns and processes by which mutations accumulate at particular loci is important for different reasons. For example, it helps to evaluate the utility of different marker types for assaying population parameters, and enables the appropriate interpretation of statistics derived using those markers. Furthermore, it ultimately provides insight into the molecular mechanisms driving genome evolution. The results obtained here stand in tentative support of the TPM for the five simple chloroplast sequence-repeat loci, and the IAM for the more complex combination locus $G_{2.1}/R_1$. One potential problem in using the BOTTLENECK algorithms to assess the fit of the loci to these models is that, in order to do this, we must assume that deviation from mutation-drift equilibrium is not due to a recent reduction or expansion in population size. Obviously, this cannot reasonably be concluded for lodgepole pine, especially considering that the northernmost populations are likely very recent and perhaps only in their first generation. However, it has been pointed out that for a set of linked microsatellite loci departures from mutation-drift equilibrium due to variation in population size will not be apparent with only a few loci (Goldstein et al. 1996). The phylogenetic distribution of allele sizes also favours the TPM for the five simpler loci.

The main consequence of a stepwise mutation model, or any system that approximates it, is that identical alleles are expected to rise independently in different lineages, resulting in recurrent mutation or homoplasy. When considering haplotypes based on several loci the extent of recurrent haplotype occurrence is less than that of recurrent allele occurrence, but still problematical. Both types of homoplasy limit the ability of microsatellite-type data to detect phylogeographic relationships (Cooper et al. 1996; Doyle et al. 1998); recurrent-alleles homoplasy prevents high resolution of the phylogeny, and recurrent-haplotype homoplasy masks the presence of phylogeographic structure. Therefore, although the chloroplast sequence-repeat loci have proven useful for detecting certain aspects of lodgepole pine evolutionary history, more loci with combination alleles such as $G_{2.1}/R_1$ would be an invaluable addition for phylogeographic purposes. Finally, the assumption of linkage of chloroplastic loci may also be in error. Although reports of chloroplast recombination are rare if not absent in the literature, we have yet to rule this possibility out. The phylogenetic consequence of recombination is also homoplasy, a factor which would further compromise the utility of these markers for phylogeography assessment.

Acknowledgements The experiments carried out for this study comply with the current laws of Canada. We thank the Natural Sciences and Engineering Research Council of Canada for postdoctoral support to H.D.M. and an operating grant to K.R. Rita Wagner and Carol Fleetham were responsible for tree sampling at the Prince George Tree Improvement Station, and Forest Renewal British Columbia provided funding for this research.

References

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cooper G, Amos W, Hoffman, D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014

Critchfield WB (1980) Genetics of lodgepole pine. USDA Forest Service Research Paper, WO–37

Cwynar LC, MacDonald GM (1987) Geographical variation of lodgepole pine in relation to population history. Am Nat 129: 463–469

Dancik BP,Yeh FC (1982) Allozyme variability and evolution of lodgepole pine (*Pinus contorta* var. *latifolia*) and jack pine (*P. banksiana*) in Alberta. Can J Genet Cytol 25:57–64

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, et al. (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Dong J,Wagner DB (1993) Taxonomic and population differentiation of mitochondrial diversity in *Pinus banksiana* and *Pinus contorta*. Theor Appl Genet 86:573–578

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Doyle JJ, Morgante M, Tingey SV, Powell W (1998) Size homoplasy in chloroplast microsatellites of wild perennial relatives of soybean (*Glycine* subgenus Glycine). Mol Biol Evol 15:215–218

Dumolin-Lapegue S, Demesure B, Fineschi S, LeCorre V, Petit RJ (1997) Phylogeographic structure of white oaks throughout the European continent. Genetics 146:1475–1487

Epperson BK, Allard RW (1987) Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics 115:341–352

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

Goldstein DB, Ruíz-Linares A, Feldman M,Cavalli-Sforza LL (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavelli-Sforza LL, et al. (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13:1213–1218

Hamilton MB (1999) Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. Mol Ecol 8:521–523

King RA, Ferris C (1998) Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. Mol Ecol 7:1151–1161

MacDonald GM, Cwynar LC (1985) A fossil pollen based reconstruction of the late Quaternary history of lodgepole pine (*Pinus contorta* ssp. *latifolia*) in the western interior of Canada. Can J For Res 15:1039–1044

MacDonald GM, Cwynar LC (1991) Post-glacial population growth rates of *Pinus contorta* ssp. *latifolia* in western Canada. J Ecol 79:417–429

Nei M (1986) Molecular evolutionary genetics. Columbia University Press, New York

Newton AC, Allnutt TR, Gillies ACM, Lowe AJ, Ennos RA (1999) Molecular phylogeography, intraspecific variation and the conservation of tree species. Trends Ecol Evol 14:140–145

Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA (1995a) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc Natl Acad Sci USA 92:7759–7763

Powell W, Morgante M, Andre C, McNicol JW, Machray GC, Doyle JJ, Tingey SV, Rafalski JA (1995b) Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Curr Biol 5:1023–1029

Ruiz-Linares A, Ortíz-Barrientos D, Figueroa M, Mesa N, Múnera JG, Bedoya G, Vélez ID, García LF, Pérez-Lezaun A, Bertranpetit J, Feldman MW, Goldstein DB (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. Proc Natl Acad Sci USA 96: 6312–6317

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: problems and prospects. Mol Ecol 7:465–474

Seielstad M, Bekele E, Ibrahim M, Toure A, Traore M (1999) A view of modern human origins from Y chromosome microsatellite variation. Genome Res 9:558–67

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134: 983–993

Stoehr MU, Newton C (2001) Evaluation of mating dynamics and pollen contamination in a lodgepole pine seed orchard using chloroplast DNA markers. Can J For Res (in press)

Stoehr MU, Orvar BL, Vo TM, Gawley JR, Webber JE, Newton CH (1998) Application of a chloroplast DNA marker in seed orchard management evaluations of Douglas-fir. Can J For Res 28:187–195

Swofford DL (1998) Paup*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics 140:767–782

Torroni A, Chen JS, Simino O, Santachiarabeneceretti AS, Scott CR, Lott MT, Winter M, Wallace DC (1994) MtDNA and Y-chromosomal polymorphisms in four native American populations from southern Mexico. Am J Hum Genet 54:303–318

Vendramin GG, Lelli L, Rossi P, Morgante M (1996) A set of primers for the amplification of 20 chloroplast microsatellites in Pinaceae. Mol Ecol 5:595–598

Vendramin GG, Anzidei M, Madaghiele A, Bucci G (1998) Distribution of genetic diversity in *Pinus pinaster* Ait. as revealed by chloroplast microsatellites. Theor Appl Genet 97:456–463

Vendramin, GG, Degen B, Petit RJ, Anzidei M, Madaghiele A, Ziegenhagen B (1999) High level of variation at *Abies alba* chloroplast microsatellite loci in Europe. Mol Ecol 8:1117–1126

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA 91:9794–9798

Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9–19

Wheeler NC, Guries RP (1982) Population structure, genic diversity, and morphological variation in *Pinus contorta* Dougl. Can J For Res 12:595–606

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Xie C-Y, Ying CC (1995) Genetic architecture and adaptive landscape of interior lodgepole pine (*Pinus contorta* ssp. *latifolia*) in Canada. Can J For Res 25:2010–2021

Yang R-C, Yeh FC, Yanchuk AD (1996) A comparison of isozyme and quantitative genetic variation in *Pinus contorta* ssp. *latifolia* by F_{ST}. Genetics 142:1045–1052

Yeh FC, Layton C (1979) The organization of genetic variability in central and marginal populations of lodgepole pine *Pinus contorta* spp. *latifolia*. Can J Genet Cytol 21:487–503

Yeh FC, Cheliak WM, Dancik BP, Illingworth K, Trust DC, Pryhitka BA (1985) Population differentiation in lodgepole pine, *Pinus contorta* spp. *latifolia*: a discriminant analysis of allozyme variation. Can J Genet Cytol 27:210–218

Ying CC (1991) Performance of lodgepole pine provenances at sites in southwestern British Columbia. Silvae Genet 40: 215–223

Ying CC, Liang Q (1994) Geographic pattern of adaptive variation of lodgepole pine (*Pinus contorta* Dougl.) within the species' coastal range: field performance at age 20 years. For Ecol Man 67:281–298