MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 100–114 doi: 10.1111/j.1365-294X.2009.04474.x

Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines

J. B. WHITTALL,* J. SYRING,† M. PARKS,‡ J. BUENROSTRO,* C. DICK,* A. LISTON‡ and R. CRONN§

**Department of Biology, Santa Clara University, Santa Clara, CA 95053, USA, †Department of Biology, Linfield College, McMinnville, OR 97128, USA, ‡Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97330, USA, §Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, OR 97331, USA*

Abstract

Critical to conservation efforts and other investigations at low taxonomic levels, DNA sequence data offer important insights into the distinctiveness, biogeographic partitioning and evolutionary histories of species. The resolving power of DNA sequences is often limited by insufficient variability at the intraspecific level. This is particularly true of studies involving plant organelles, as the conservative mutation rate of chloroplasts and mitochondria makes it difficult to detect polymorphisms necessary to track genealogical relationships among individuals, populations and closely related taxa, through space and time. Massively parallel sequencing (MPS) makes it possible to acquire entire organelle genome sequences to identify cryptic variation that would be difficult to detect otherwise. We are using MPS to evaluate intraspecific chloroplast-level divergence across biogeographic boundaries in narrowly endemic and widespread species of *Pinus*. We focus on one of the world's rarest pines – Torrey pine (*Pinus torreyana*) – due to its conservation interest and because it provides a marked contrast to more widespread pine species. Detailed analysis of nearly 90% (~105 000 bp each) of these chloroplast genomes shows that mainland and island populations of Torrey pine differ at five sites in their plastome, with the differences fixed between populations. This is an exceptionally low level of divergence (1 polymorphism/~21 kb), yet it is comparable to intraspecific divergence present in widespread pine species and species complexes. Population-level organelle genome sequencing offers new vistas into the timing and magnitude of divergence within species, and is certain to provide greater insight into pollen dispersal, migration patterns and evolutionary dynamics in plants.

Keywords: chloroplast genome, multiplex sequencing-by-synthesis, next-generation sequencing, *Pinus*

Received 18 June 2009; revision received 19 August 2009; accepted 25 August 2009

Introduction

Next-generation (Next-Gen) sequencing is revolutionizing all facets of molecular ecology (Hudson 2007; Rokas & Abbot 2009; this issue), as rapid access to orders of magnitude more data at substantially reduced costs promises a wealth of new insights. The ability to sequence nearly complete organellar genomes is an important milestone in this revolution. In addition to the important population and evolutionary insights provided by these independent genomic partitions, the compact size, conserved genic content and structural organization, and low (to absent) intraindividual variability of organelle genomes make them an experimentally tractable system for testing and refining modern sequencing strate-

Correspondence: Richard Cronn, Fax: (541) 750-7329; E-mail: rcronn@fs.fed.us

This article is a US Government work and is in the public domain in the USA.

gies (Moore *et al.* 2006; Meyer *et al.* 2007; Cronn *et al.* 2008, Parks *et al.* 2009), and for developing and testing new bioinformatics tools (Bryant *et al.* 2009).

In plants, the chloroplast genome has been an invaluable resource for investigating inter- and intraspecific evolutionary histories (Birky 1978, 2001; Chase *et al.* 1993; McCauley 1995; Newton *et al.* 1999; Provan *et al.* 2001; Petit *et al.* 2003). The predominantly uniparental inheritance of chloroplasts (for exceptions, see Birky 2001; Mogensen 1996) is analytically attractive since a single, independent genealogical history can be readily obtained for hypothesis testing and comparison with the nuclear genome. In plants showing maternal chloroplast inheritance, the magnitude and pattern of differentiation reveals the relative importance of seed vs. pollen dispersal and matrilineal evolutionary history (Ennos 1994; Hu & Ennos 1997; Petit *et al.* 2005). In a subset of land plants (conifers and a few flowering plant lineages), the chloroplast is paternally inherited and thus tracks the evolutionary history of pollen dispersal independent of the nuclear genome, and is frequently independent of the mitochondrial genome (Neale & Sederoff 1989). This allows genetic variation to be partitioned into parental contributions (pollen vs. seed), and for each genome to serve as an independent partition in tests for genetic differentiation of geographically isolated or disjunct populations (Hu & Ennos 1999; Mitton *et al.* 2000).

In most plants, the usefulness of chloroplast-derived information is often offset by its conservative mutation rate. For example, the estimated per-base mutation rate for chloroplast genome in pines is on the order of 0.2–0.4×10^{-9} synonymous substitutions per site per year (Willyard *et al.* 2007; Gernandt *et al.* 2008). This is ~ 1/100 the value for animal mitochondria (Moritz *et al.* 1987), so it requires proportionately more chloroplast DNA sequence to yield resolutions comparable to those estimated from animal mitochondrial genomes for similarly aged divergence events. An impact of this limitation is that chloroplast-based inferences often focus on the fastest evolving fraction of the chloroplast genome, primarily microsatellites or repeated motifs (Provan *et al.* 1999; Ebert & Peakall 2009). These markers show high mutation rates and can provide excellent haplotypic discrimination (Afzal-Rafii & Dodd 2007; Höhn *et al.* 2009; Moreno-Letelier & Piñero 2009). Conversely, chloroplast microsatellites are constrained in length, which increases the probability of molecular homoplasy (Estoup *et al.* 2002; Jakobsson *et al.* 2006) and makes them poorly suited for investigating genealogical, mutational, and coalescent histories (Brumfield *et al.* 2003). Collectively, these types of studies highlight the need for evaluating *all* genetic variation contained within the chloroplast genome.

The current generation of genome sequencers possesses an overwhelming excess of capacity for accessing sequences from entire organellar genomes. Land plant organellar genomes range in size from ~70–220 kb for the chloroplast, to over 700 kb in mitochondria (survey of NCBI GenBank; Release 172.0). When combined with multiplex or barcoding methods (e.g. Meyer *et al.* 2007; Craig *et al.* 2008; Cronn *et al.* 2008; Erlich *et al.* 2009), modern sequencers could potentially sequence hundreds of organelle genomes in a single analysis. Although the sequencing of genomes is increasingly easy, Next-Gen sequencers are not without limitations. For example, some platforms have been characterized as showing higher positional error rates than Sanger sequencing, particularly in regions of low complexity (e.g. single nucleotide repeats, short perfect repeats; Bentley *et al.* 2008). These repeats can be abundant in organellar genomes, so they might be 'hotspots' for methodological errors. Similarly, biases in genome-wide base composition have been reported to result in biases in sequencing error (Dohm *et al.* 2008; Dolan & Denver 2008). Plant organelle genomes are generally A/T-rich, with chloroplasts showing the greatest skew in base composition compared to mitochondria (~62% A/T vs. 58% respectively; NCBI GenBank; Release 172.0). These kinds of errors are not problematic for many genomics applications, but they are certain to inflate estimates of nucleotide diversity when surveying populations for rare polymorphisms.

In this report, we show how whole chloroplast genomes can be rapidly sequenced and screened to identify intraspecific variation, with examples from the conifer genus *Pinus*. Results from Next-Gen sequencing are directly compared to Sanger sequencing in order to evaluate the relationship between sequencing depth, the discovery of putative SNPs, and the false-positive and false-negative discovery rate. The primary focus of this study, Torrey pine (*Pinus torreyana*), is one of the rarest temperate trees in the world (Critchfield & Little 1966) and a species of conservation concern. Torrey pine is restricted to two populations in California, USA, separated by 280 km of Pacific Ocean (Fig. 1a). The mainland grove located north of San Diego, CA (*P. torreyana* ssp. *torreyana*), comprises ~3400 trees, while another ~2000 trees occur on Santa Rosa Island, CA (*P. torreyana* ssp. *insularis*). The populations have been suggested to be evolutionarily distinct based on subtle morphological differences (cone features, growth rates in common garden) and have been described as subspecies (Haller 1986). Torrey pine is exceptional among pine species due to its unusually low levels of allozyme variation (Ledig & Conkle 1983), and attempts at distinguishing island from mainland populations have been stymied by a lack of genetic variation, especially in cpDNA (Waters & Schaal

This article is a US Government work and is in the public domain in the USA.

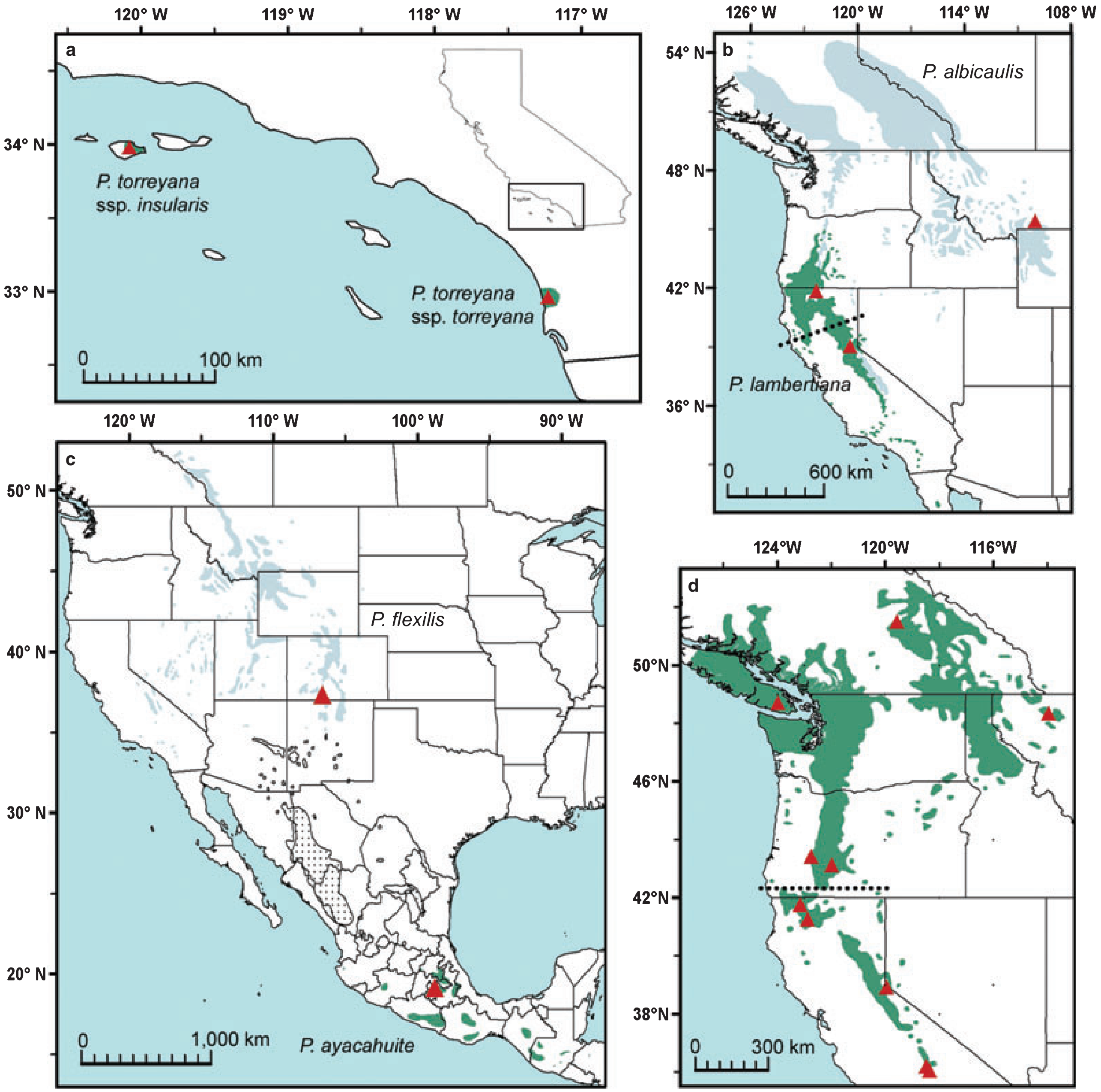

Fig. 1 Geographic distributions of the species examined and locations sampled (triangles). (a) *Pinus torreyana* is restricted to Santa Rosa Island (ssp. *insularis*) and the mainland (ssp. *torreyana*) near San Diego, California. (b) *Pinus albicaulis* (light shading) and *Pinus lambertiana* (dark shading); dotted line shows the division between north and south germplasm (see text for description). (c) *Pinus flexilis* and *Pinus ayacahuite*; *Pinus strobiformis* (stippled) is displayed to highlight the continuous distribution of this species complex (see text for description). (d) *Pinus monticola*; dotted line shows the division between north and south germplasm (see text for description).

1991). Despite three separate cpDNA studies and a combined total of 17 cpSSR loci (Provan *et al.* 1999), 150 cp restriction sites (Waters & Schaal 1991) and 3.5 kbp of cpDNA sequence (Gernandt *et al.* 2009), intraspecific variation has not been detected in this species. This, in turn, severely constrains our ability to understand the evolutionary history of this species.

Using Next-Gen sequencing, we can sequence and analyse whole chloroplast genomes from species of conservation concern such as Torrey pine, and begin to provide answers to important questions that bear upon their management: (i) can genetic variation be detected in the chloroplast genome of Torrey pine?; (ii) do extant populations represent an undifferentiated segregating

This article is a US Government work and is in the public domain in the USA.

metapopulation, or are they evolutionarily distinct in their chloroplast genomes?; (iii) is it possible to date the approximate divergence of chloroplast types detected?; and (iv) is Torrey pine unique among pines in its magnitude and partitioning of chloroplast divergence? To address these questions, we compare the results from Torrey pine to five estimates of intraspecific divergence that use partial or complete pine chloroplast genomes. Two of these comparisons compare divergent haplotypes within Sugar pine (*P. lambertiana*) and within Western White pine (*P. monticola*) that were sampled from previously identified, genetically divergent populations (Liston *et al.* 2007; Steinhoff *et al.* 1983; J. Syring, unpublished). The remaining comparisons are effectively intraspecific, as they focus on chloroplast genomes from taxa that have either been considered conspecific (*P.–P. sibirica*; Meusel *et al.* 1965; Shaw 1914), part of a *cembra* species complex (*P. flexilis–P. ayacahuite*; Moreno-Letelier & Piñero 2009; Syring *et al.* 2007), or are related through introgressive/chloroplast capture events (*P. lambertiana – P. albicaulis*; Liston *et al.* 2007). In total, these data offer an unprecedented view into the magnitude of intraspecific, cryptic chloroplast genome variation. They also highlight possible discrepancies between estimates of diversity/divergence from different classes of markers (microsatellites, single genes and whole genomes) that need to be reconciled in future comparisons.

Materials and methods

Haplotype sampling

Intraspecific samples were taken across previously identified biogeographic barriers and/or chosen to represent known haplotype variants for *Pinus torreyana*, *P. lambertiana* and *P. monticola* (Table 1, Fig. 1). *Pinus torreyana* plastomes were sequenced in one island (ssp. *insularis*) and one mainland (ssp. *torreyana*) individual grown at the Santa Barbara Botanical Garden. For haplotype screening, an additional 81 individuals were collected from both segments of the population within Torrey Pines State Natural Reserve, San Diego, CA, and 86 individuals were collected from the Santa Rosa Island population. *Pinus lambertiana* samples from two individuals represent previously identified and highly divergent haplotypes (Liston *et al.* 2007; Fig. 1b). Ten samples from *P. monticola* were chosen to evenly represent northern and southern populations of this species that have been previously determined to be phylogeographically distinct through isozyme studies (Steinhoff *et al.* 1983) and preliminary analyses of four low-copy nuclear loci (J. Syring, unpublished) (Fig. 1d).

Interspecific comparisons were also made, although these taxa are arguably conspecific (*P. cembra* and *P. sibirica*) or represent members of a species grade (*P. flexilis* and *P. ayacahuite*). Prior studies of chloroplast DNA show that divergence among these pairs of species is equivalent to conspecific comparisons in pines (Gernandt *et al.* 2005; Eckert & Hall 2006; Liston *et al.* 2007) and other gymnosperms (Little & Stevenson 2007). For example, *P. cembra* and *P. sibirica* show little morphological differentiation and have been considered conspecific (Shaw 1914; Meusel *et al.* 1965). Analysis of chloroplast microsatellites (Gugerli *et al.* 2001), chloroplast sequences (Liston *et al.* 2007), and nuclear gene sequences (Syring *et al.* 2007) reveal identical haplotypes in these species. Our samples (one per species) were collected from sites ~4800 km distant. *Pinus flexilis* and *P. ayacahuite* represent geographic extremes of a species complex that differ primarily in cone dimensions and seed wing development. This species complex spans 35° of latitude, from Mexico (*P. ayacahuite*), across the southwestern USA (*P. strobiformis*) and northward into Canada (*P. flexilis*) (Fig. 1c). Our samples of *P. flexilis* and *P. ayacahuite* (one each) were collected at sites ~2200 km distant. Finally, *P. albicaulis* (Fig. 1b) and northern populations of *P. lambertiana* are genetically and morphologically distinct, but they share nearly identical chloroplast haplotypes, possibly as a consequence of introgressive hybridization (Liston *et al.* 2007). Distribution maps of species (generated in ArcMap v9.3; ESRI) used digitized range maps of individual species (Critchfield & Little 1966; http://esp.cr.usgs.gov/data/atlas/little/).

Microread sequencing and genome assembly

DNA was extracted from fresh needles or seed megagametophyte tissue using the FastDNA Kit (Q-BIO Gene) or the DNeasy Plant Mini Kit (QIAGEN). For all samples but *P. monticola*, chloroplast genomes were amplified in 35 separate PCR reactions as previously reported (Cronn *et al.* 2008). In *P. monticola* one-third of the chloroplast genome was amplified in 12 PCR reactions with primers 1F through 12R (Cronn *et al.* 2008). For each species, the PCR reactions were quantified, pooled into equal-molar mixtures and converted into barcoded Illumina sequencing libraries (Cronn *et al.* 2008). Individual libraries were pooled into multiplex sequencing libraries ranging from 4× (for full chloroplast genomes) to 16× (partial *P. monticola* chloroplast genomes).

Cluster generation of adapter-barcoded libraries used 5 pmol, and produced 870 000–2 870 000 microreads per sample for complete genomes, and 188 000–787 000 microreads for partial genomes (*P. monticola*). After the removal of barcodes, microreads (33–37 bp) from all

This article is a US Government work and is in the public domain in the USA.

Table 1 Species, sample locations and properties of microread assemblies used to construct full or partial chloroplast genomes

Species	Collection ID	Locality information	Latitude, longitude	Voucher	Microreads	RGA contigs (average length bp)	Average depth* (median)	Assembly length† (bp)	GenBank accession no.
P. albicaulis	ALBI05	USA: Montana, Stillwater Co.	45.44°N, 110.01°W	OSC 213500	869 509	56 (2079.0)	100.9 (56)	107 159	FJ899566
P. ayacahuite	AYAC01	Mexico: Tialmanatco, Mexico	19.17°N, 98.80°W	OSC 213762	1 173 420	54 (2038.9)	133.7 (97)	104 983	FJ899570
P. cembra	CEMB03	Austria: Styria, Resorts Predlitz-Turrach and Reichenau, Turracher Hohe	47.98°N, 13.89°E	OSC 213511	1 166 707	110 (905.8)	175.4 (44)	86 921	FJ899574
P. flexilis	FLEX13	USA: Arizona, Graham Co.	37.38°N, 106.58°W	OSC 21372	1 545 509	39 (2979.5)	186.6 (137)	110 415	FJ899576
P. lambertiana N	LAMB08	USA: California, Siskiyou Co.	41.85°N, 122.31°W	OSC 213878	2 870 153	19 (6122.7)	300.2 (102)	114 386	EU998743
P. lambertiana S	LAMB01	USA: California, Placer Co.	39.05°N, 120.38°W	OSC2 13878	1 180 289	55 (2011.1)	172.2 (114)	105 202	FJ899577
P. monticola	MONT06	Canada: mainland British Columbia	51.50°N, 119.55°W	Unvouchered	324 564	64 (587.8)	145.1 (52)	36 133	GQ478176
P. monticola	MONT07	Canada: Vancouver Island, British Columbia	48.72°N, 123.97°W	Unvouchered	306 676	55 (660.9)	140.7 (90)	34 231	GQ478177
P. monticola	MONT08	USA: California, Siskiyou Co.	41.75°N, 123.13°W	Unvouchered	512 154	35 (1111.3	231.0 (116)	37 696	GQ478178
P. monticola	MONT12	USA: California, Kern Co.	36.05°N, 118.35°W	Unvouchered	444 888	36 (1085.1)	212.3 (159)	38 578	GQ478179
P. monticola	MONT14	USA: California, Tulare Co.	36.20°N, 118.48°W	Unvouchered	448 833	41 (946.2)	219.4 (150)	38 202	GQ478180
P. monticola	MONT26	USA: Oregon, Douglas Co.	43.42°N, 122.73°W	Unvouchered	233 684	41 (940.3)	109.0 (82)	36 952	GQ478181
P. monticola	MONT30	USA: Oregon, Douglas Co.	43.13°N, 121.97°W	Unvouchered	358 422	37 (1044.5)	187.5 (121)	37 981	GQ478182
P. monticola	MONT36	USA: California, Siskiyou Co.	41.25°N, 122.87°W	Unvouchered	787 410	65 (586.2)	130.0 (50)	36 368	GQ478183
P. monticola	MONT38	USA: California, El Dorado Co.	38.92°N, 119.94°W	Unvouchered	193 449	52 (739.0)	92.9 (53)	36 733	GQ478184
P. monticola	MONT49	USA: Montana, Flathead Co.	48.34°N, 113.93°W	Unvouchered	188 214	50 (764.2)	91.8 (59)	35 893	GQ478185
P. sibirica	SIBI03	Russia: Kemorovo District	55.40°N, 86.10°E	OSC 213880	947 216	108 (995.5)	84.0 (62)	97 547	FJ899558

This article is a US Government work and is in the public domain in the USA.

Table 1 *Continued*

Species	Collection ID	Locality information	Latitude, longitude	Voucher	Microreads	RGA contigs (average length bp)	Average depth* (median)	Assembly length† (bp)	GenBank accession no.
P. torreyana ssp. *insularis*	SBBG 65–187	Santa Barbara Botanical Garden, CA, USA (grown from seed collected by Bob Haller from Santa Rosa Island)	34.27°N, 119.42°W	Whittall. 2008.245	1 157 851	60 (1920.5)	158.1 (89)	107 977 (109 041)	FJ899564
P. torreyana ssp. *torreyana*	SBBG *s.n.* ('pre-1937')	Santa Barbara Botanical Garden, CA, USA (grown from seed collected by Bob Haller from La Jolla, CA)	34.27°N, 119.42°W	Whittall. 2008.244	1 114 111	67 (1762.4)	104.4 (77)	104 432 (105 892)	FJ899563

*Average depth and median values reported only for those sites with ≥5× depth.
†Values in parentheses for *P. torreyana* include additional Sanger sequence.

accessions except *P. monticola* were assembled with *de novo* assemblers VELVET v. 0.6 (Zerbino & Birney 2008) and EDENA v. 2.1.1 (Hernandez *et al.* 2008), using minimum depth filters of 5×, minimum contig lengths of 100 bp and hash lengths of 25 bp. Generally, assembled contigs ranged from several hundred to several thousand bp in length; between 100 and 300 contigs were produced per complete genome, and 35 and 65 contigs were produced per partial genome (Table 1).

Genome assembly from *de novo* contigs followed a two-step process. *De novo* contigs were aligned to a reference chloroplast using CODONCODE v. 2.0.6 (Codoncode Corp., http://www.codoncode.com). The following reference sequences were used: *P. ponderosa* (GenBank FJ899555) for *P. torreyana* accessions; and *P. koraiensis* (GenBank AY228468) for *P. albicaulis*, *P. ayacahuite*, *P. cembra*, *P. flexilis*, *P. lambertiana* and *P. sibirica* accessions. Orphan contigs that failed to align to references were checked for chloroplast homology using BLASTN (http://www.ncbi.nlm.nih.gov/); where sequence coverage was lacking or where contig alignment failed due to indels, orphan contigs were manually inserted into the alignment. *De novo* assemblies from these two programs (VELVET, EDENA) were nearly identical, but a slight increase in aligned *de novo* assembly length was gained through the use of both assemblers. A consensus sequence of aligned VELVET and EDENA *de novo* contigs was made using BioEdit version 7.0.5.2 (Hall 1999). The terminal 30 bp of contig ends were also edited to match the reference sequence completely, as these regions often contained assembly error due to reduced sequencing depth at contig ends. The consensus sequence of aligned contigs was merged with the reference to form a 'chimeric pseudoreference', composed primarily of *de novo* sequence (typically >90%), and including a small proportion (<10%) of reference sequence where *de novo* sequence was missing. Original microreads from each accession were then re-mapped onto a pseudoreference using the reference-guided assembler RGA (Shen and Mockler, http://rga.cgrb.oregonstate.edu/), a minimum depth of 2×, maximum allowable error/mismatch of 0.033 and 70% majority minimum for SNP acceptance. *Pinus monticola* sequences were assembled against an unpublished *P. monticola* chloroplast genome sequence (R. Cronn, unpublished) using RGA with these same parameters.

Genomes were aligned using MAFFT v. 6.240 (Katoh *et al.* 2005) with a gap opening penalty of 2–2.5 and a gap extension penalty of 0. Aligned sequences were annotated using DOGMA (Wyman *et al.* 2004) and the Chloroplast Genome Database (http://chloroplast.cbio.psu.edu/). Initial quality checks of exon translations (to identify errors and frameshift/nonsense mutations) and spatial patterning of SNPs showed some

This article is a US Government work and is in the public domain in the USA.

regions with unexpectedly high divergence, and these were inferred as misassemblies arising from one or more of the following sources: (i) rare misassembly error from RGA; (ii) errors arising due to low sequencing depth near primers; and (iii) amplification of paralogous pseudogenes. In these rare instances, preference was given to *de novo* sequence assemblies. If the problematic region was not represented in *de novo* assemblies, or if unexpectedly high divergence was found across an entire region (exon or amplicon) the region was coded as missing. We observed that highly divergent regions were commonly associated with nucleotides flanking primer locations (±100 bp of the primer), and this appears to be related to low sequencing depth near primers; these regions were changed to N's. Finally, due to the overlapping nature of our primers (Cronn *et al.* 2008), there was no way to unequivocally determine the sequence of primer regions, so primer sequences were changed to N's. The net impact of these corrections is that true 'hotspots' of divergence are only supported in our study if they are supported by *de novo and* reference guided assembly.

As noted below, chloroplast variation from *P. torreyana* was also evaluated by direct Sanger sequencing. Alignment of these sequences to Illumina-based assemblies identified 1064 bp (ssp. *insularis*) and 1460 bp (ssp. *torreyana*) of gaps that could be eliminated by merging these data. For the purpose of identifying false-positives and false-negatives, Sanger sequences were compared to assemblies derived *only* from Illumina microreads. Our final sequences to GenBank, however, include the Sanger additions.

Pairwise comparisons of pine plastomes

In order to assess the distinctiveness of the Torrey pine plastome results, we compared *P. torreyana* to nearly complete chloroplast genome divergence in four other cases [*P. lambertiana* northern (N) vs. southern (S) haplotypes, *P. lambertiana* N vs. *P. albicaulis*, *P. ayacahuite* vs. *P. flexilis*, *P. cembra* vs. *P. sibirica*], and from 10 partial plastomes of *P. monticola* (~39 kb). For these comparisons, all variable sites in initial assemblies were filtered for a minimum 25× coverage depth and 85% majority base call based on results of *P. torreyana* SNP validation (rationale for this minimum depth is provided below).

Uncorrected pairwise distances between haplotypes were calculated for the entirety of the aligned sequences, and partitioned into synonymous plus silent sites (d*S*) vs. non-synonymous sites (d*N*). All distance estimates were calculated using MEGA4 (Tamura *et al.* 2007), with P-distances for comparisons of overall nucleotides, Jukes–Cantor estimates of d*S* and d*N*, and pairwise deletion of unshared sites. Estimates of error were determined using 500 bootstrap replicates. AMOVA was conducted using GENALEX v. 6 (Peakall & Smouse 2006) to examine hierarchical structure of genetic variation in *P. monticola* between two regions. Input data was from pairwise distance matrices, and significance was assessed using 1000 permutations.

Sanger sequencing of variable sites in P. torreyana

For the two *P. torreyana* samples, variable sites were scrutinized based on the minimum number of microreads supporting the base call and the minimum base-call consistency to directly identify true SNPs and to estimate the rate of false-positive SNPs and false-negative SNPs. From this analysis, regions flanking 32 putative SNPs and 2 indels were examined by Sanger sequencing. Primers were developed to maximize the number of variable sites covered while limiting the amplification products to ~1 kb each (primer sequences available from authors by request). PCR reactions were done in 20 μL reaction volumes containing: $MgCl_2$ (2.5 mM), *Taq* PCR Buffer B (1×; Fisher Scientific), dNTPs (0.25 mM each), forward and reverse primers (1 μM each), *Taq* polymerase (2 units) and 50 ng of genomic DNA. Thermal cycling conditions were: 30 s denature at 92 °C, followed by 35 cycles of 8 s denature at 92 °C, 30 s annealing at 55–57 °C and 90 s extension at 72 °C. A final 10-min extension at 72 °C was followed by a 4 °C hold. PCR products were visualized on agarose gels and directly sequenced on an ABI 3730 (Applied Biosystems).

SNP genotyping in P. torreyana

For the SNPs confirmed with Sanger sequencing, we genotyped 167 trees (81 from mainland; 86 from island). All five variable sites overlapped with restriction enzyme recognition sites, yet in order to confidently determine genotypes, we developed a complementary dCAPs assay using a primer that introduced a restriction site into the allele that was not cut by the native restriction site (Neff *et al.* 1998). Using these genotyping primers, we amplified fragments from 121 to 203 bp following the aforementioned PCR protocol. Five to 10 μL of PCR product were digested with 5 units of restriction enzyme for 5 h and assayed on agarose gels. Cut vs. uncut fragments for each SNP differed by 21–32 bp.

Divergence dating

Although there is substantial error associated with coalescent approaches to estimating divergence times (Graur & Martin 2004; Morrison 2008), these analyses can be informative when comparing recently diverged

This article is a US Government work and is in the public domain in the USA.

taxa with similar mutation rates and generation times. We estimated approximate divergence dates for intraspecific haplotype pairs and average divergence of 10 haplotypes for *P. monticola*. For calibration, we used chloroplast-specific mutation rates estimated for *Pinus* (Willyard *et al.* 2007; Gernandt *et al.* 2008). These prior studies reported a range of mutation rates based on slightly different fossil calibrations. For simplicity, we used the most recent estimates for divergence of hard and soft pines (72–87 Ma; Gernandt *et al.* 2008; Willyard *et al.* 2007), and calibrated mutation rates at the midpoint of this estimate (79.5 Ma) with a 4-Myr standard deviation. Assuming a lognormal distribution (Morrison 2008), 95% confidence intervals include the estimated divergence dates of both recent studies (72.1 Ma, 87.9 Ma).

Under these assumptions, we calculate the mean silent divergence rate to be 0.24×10^{-10} silent substitutions per site per year (95% CI = 0.890–5.371×10^{-10}). To include error in this estimate, we assumed that error in divergence dates is lognormally distributed (Morrison 2008). Under assumptions of the neutral model, the absolute per-year mutation rate (μ) for a haploid organelle is represented as:

$$\mu = \frac{d}{2T_{\text{div}} + N_{\text{e}}}, \qquad (eqn1)$$

so

$$T_{\text{div}} = \frac{d - (N_{\text{e}}\mu)}{2\mu}, \qquad (eqn2)$$

where T_{div} is the time since species divergence (measured as absolute years), d is pairwise divergence between haplotypes, μ is the mutation rate and N_{e} is the ancestral effective population size (Kimura 1983). Divergence dates were estimated by Monte-Carlo simulation, using lognormally distributed mutation rates (0.24×10^{-10}; 95% CI = 0.890–5.371×10^{-10}), normally distributed silent (d*S*) genetic distances and errors, and values of N_{e} that span a reasonable range from 100 to 5000. Only results from $N_{\text{e}} = 1000$ are presented, as varying N_{e} over this range had minimal impact on estimated dates. Divergence dates are reported as means, and 95% confidence intervals are approximated from 2.5% to 97.5% percentiles of 10 000 simulations.

Results

Microread sequencing and genome assembly

When barcoded samples from these experiments were parsed, we retrieved an average of 1 336 085 microreads for each full genome and 379 829 microreads for partial *Pinus monticola* genomes (Table 1). By aligning *de novo* contigs onto reference genomes, we determined that *de novo* assemblies consistently were interrupted at priming sites (35 for whole genomes; 12 for partial genomes) and low complexity single nucleotide repeats; this phenomenon is evident in depth plots for genomes (Fig. 2), and is discussed in greater detail in Cronn *et al.* (2008). In addition, alignment of *de novo* contigs revealed no detectable structural rearrangements. RGA analysis resulted in an average of 63.1 contigs per full genome, with an average length of 2313 bp per contig. Assemblies for the partial *P. monticola* genome were proportionately less abundant (mean = 47.6 contigs) and shorter (mean = 846.5 bp). Full genome sequences produced by the pseudoreference-guided assembly process include an average of 104 336 bp (88.9%) for full genomes, and 36 887 bp (93.4%) for partial genomes.

Confirming SNPs and false-positives in P. torreyana

Initial pairwise comparisons of chloroplast genomes revealed a surprisingly large number of polymorphic sites, a finding seemingly inconsistent with expectations of a conservative chloroplast divergence rate. For example, analysis of 104 432 bp from paired samples of Torrey pine revealed 32 putative SNPs (Table 2; Fig. 2) that spanned a range of sequencing depth (Fig. 2). A plot of the majority base frequency vs. the sequencing depth for variable positions (Fig. 3) showed that sequencing depth was generally low among these sites (geometric mean = 18.9). At most putative SNP sites, only two of the possible four nucleotides were found (Table 2) and the minority nucleotide represents the ancestral state (*P. ponderosa*). We attribute this bias to either sequencing errors in the 3-bp barcode resulting in the incorrect assignment of microreads (Cronn *et al.* 2008) or to potential sample cross-over between adjacent lanes of the Illumina flow cell during the cluster generation process.

To determine whether these sites were false-positives arising from low sequencing depth, we resequenced these 32 sites from both Torrey pine samples using standard Sanger sequencing. The resulting 23.1 kb of sequence (11 628 bp for mainland; 11 528 bp for island) validated 5 SNPs. These were located in the *trnV–trnH* spacer (127× depth minimum, 98% consistency), *trnS–psbB* spacer (127× min, 98%), *ycf1* coding region (a replacement substitution; 86× min, 99%), *rps4–ycf12* spacer (61× min, 97%) and 23S rRNA (10× min, 83%) (Fig. 3). With the exception of 23S rRNA, these positions showed the highest combined depths (all >60×) and consistency (>95%). All remaining variable positions that were confirmed as false-positives showed generally low average sequencing depth and read con-

This article is a US Government work and is in the public domain in the USA.

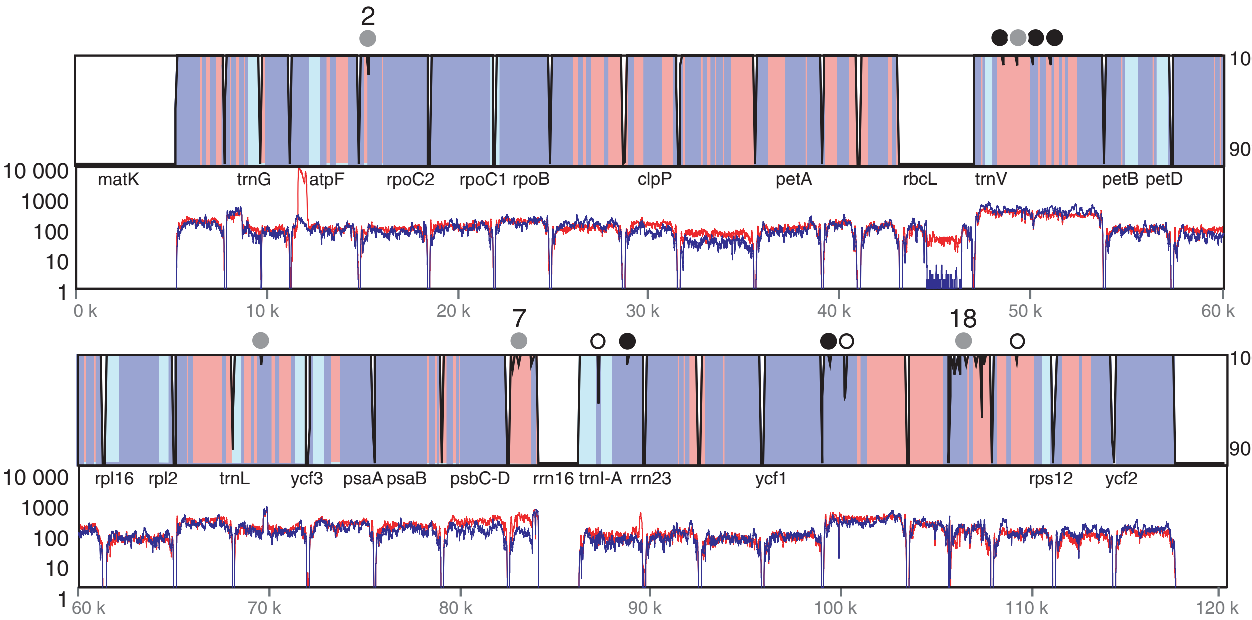

Fig. 2 Assembly, sequence depth and variable sites for aligned *P. torreyana* chloroplast genomes. Black circles are confirmed SNPs, grey circles are confirmed false-positives and indels, open circles are unconfirmed indels. Numerals above circles indicate multiple polymorphisms. In the sequence alignment, exons are blue, introns are light blue and intergenic regions are pink. The *matK* and *16S rRNA* regions were not obtained in either sample; the *rbcL* region was amplified in ssp. *insularis*, a putative non-plastid pseudogene was amplified in ssp. *torreyana*. In the sequence density plots, blue lines (ssp. *torreyana*, mainland) and red lines (ssp. *insularis*, island) indicate sequencing depth at each position.

sistency (with means of 21× and 76% respectively). The 23.1 kb of Sanger sequence used to validate SNPs is also useful in estimating the false-negative rate for regions that are readily accessible for sampling by short read and Sanger sequencing. This additional Sanger sequence differed from Illumina base calls at seven positions that were fixed in both subspecies. At present, we do not know the source for this systematic bias, but it is important to recognize that these differences are rare (seven sites out of 11 528 bp), consistent and do not result in novel SNPs. After confirmation of SNPs through Sanger sequencing, 99.995% of the *P. torreyana* genome was found to be identical between the two subspecies. Based on the results of this detailed screening, we used similar filtering criteria (depth ≥25×; consistency ≥85%) for all subsequent analyses where Sanger validation was unavailable.

Relative and absolute chloroplast genome divergence in pines

Using the filtering criteria identified above, ‘intraspecific’ pairwise differences between chloroplast genomes of widespread pine taxa ranged from zero differences across 75 195 bp in *P. cembra* vs. *P. sibirica*, to a high of 382 differences across 88 768 bp within *P. lambertiana* (Table 3; Fig. 4). In general, variable sites were unevenly dispersed across genome, with no mutational ‘hot-spots’ apparent across all comparisons (Fig. 4). Replacement substitutions were found in all comparisons except between *P. cembra* and *P. sibirica* and partial genomes of *P. monticola*. As expected for this conservative genome, silent substitutions outnumbered replacement substitutions 3.8:1 across all positions. Comparison of *P. torreyana* to other pairwise calculations shows that the average pairwise distance for *P. torreyana* (0.000047) is considerably higher than the comparison between the identical sequences from *P. cembra* vs. *P. sibirica*, approximately equal to the average pairwise divergence within *P. monticola* (0.000050), and substantially lower than the divergence for the *P. ayacahuite*–*P. flexilis* comparison (0.000165; Table 3).

Based on previously calibrated silent substitution rates for pine chloroplast genes, we estimated the divergence times between paired haplotypes in four comparisons and the average haplotype divergence date for 10 haplotypes within *P. monticola* (Table 3). Mainland and island Torrey pine plastomes diverged *c.* 160 000 years ago. In the absence of detectable divergence between *P. cembra* vs. *P. sibirica*, we estimated a maximum divergence date for these individuals by assuming that they differed maximally by one substitution across the range of sampled silent sites (45 949 bp); this places the mean estimated divergence date at <60 000 years ago. The

This article is a US Government work and is in the public domain in the USA.

Table 2 Read densities for all variable sites detected in *Pinus torreyana* and Sanger sequencing validation

		ssp. *torreyana* (mainland)					ssp. *insularis* (island)				
Position*	*Ancestral*	A	C	G	T	Consistency	A	C	G	T	Consistency
15316	T	**12**	0	0	2	0.857	0	0	0	**36**	1
15318	T	2	0	0	**10**	0.833	**31**	0	0	2	0.939
48669	C	0	**273**	0	2	0.993	0	2	0	**127**	0.984
50203†	T	0	0	1	**39**	0.975	0	0	0	**18**	1
51143	T	0	**169**	6	0	0.966	0	0	2	**61**	0.968
52253	C	**312**	4	0	5	0.972	2	**127**	0	0	0.984
69968	C	1	31	0	164	0.837	0	36	0	**63**	0.636
82965	G	0	**12**	1	0	0.923	1	2	**42**	0	0.933
82992	G	0	0	**9**	0	1	**8**	0	5	1	0.571
82997	C	**7**	2	0	0	0.778	1	**17**	0	0	0.944
83340	C	**58**	4	0	0	0.935	22	131	0	0	0.856
83907	A	5	**25**	0	0	0.833	68	14	0	1	0.819
84063	T	0	91	0	2	0.978	0	**27**	0	23	0.54
89077	A	1	1	0	**10**	0.833	**17**	0	0	0	1
99677	A	**86**	0	0	1	0.989	1	**98**	0	0	0.99
106136	C	0	**11**	0	0	1	0	3	0	**4**	0.571
106278	C	0	**25**	0	2	0.926	0	0	0	**4**	1
106303	T	1	2	0	**39**	0.929	0	**10**	0	0	1
106315	G	1	0	**35**	1	0.946	**8**	0	0	0	1
106324	G	2	0	**37**	0	0.949	**6**	1	0	0	0.857
106475	C	0	**53**	0	9	0.855	0	**11**	0	**11**	0.5
106489	A	**47**	0	1	0	0.979	0	0	**12**	0	1
106515	C	1	**60**	0	4	0.923	0	3	0	**11**	0.786
106537	C	0	**48**	1	2	0.941	0	12	0	**16**	0.571
106855	G	**18**	0	16	0	0.529	3	0	**21**	0	0.875
106958	G	1	1	**40**	5	0.851	0	0	3	**20**	0.87
107207	A	18	**25**	0	0	0.581	**24**	1	0	1	0.923
107383	C	0	**49**	1	13	0.778	0	6	0	**19**	0.76
107597	A	**105**	0	1	0	0.991	9	0	**12**	0	0.571
107613	G	8	0	**84**	0	0.913	**19**	0	1	0	0.95
107638	C	0	**66**	7	0	0.904	0	4	**14**	0	0.778
107817	A	**39**	0	0	12	0.765	8	0	0	**18**	0.692

The positions of all variable sites are shown, with the five validated SNP positions indicated in bold italic type; the remaining positions are false-positives. Positional base calls are shaded proportionally to read depth; majority base calls for a position are also indicated in bold. The ancestral nucleotide state is represented by the sequence of *Pinus ponderosa.*

*Position in alignment of *P. torreyana* and *P. ponderosa* assemblies.

†Site 50203 was polymorphic in the original sequence assemblies, but this is not supported in the read density analysis (nor Sanger sequencing).

remaining estimates ranged from *c.* 145 000 to 598 000 years ago, placing the divergence of these haplotype pairs to the mid- to upper-Pleistocene. At the far extreme, the divergent haplotypes residing within *P. lambertiana* date to a far more ancient divergence of *c.*14.8 Ma.

Spatial differentiation in pine plastomes

In this study, we are able to provide estimates of genome-wide geographic differentiation for two of the examined species, *P. torreyana* and *P. monticola*. Restriction enzyme genotyping of 167 mainland and island Torrey pine trees demonstrated that the 5 validated SNPs present in our two exemplars represented fixed differences between these populations. Based on these results, we predict that the mainland and island populations are distinct and fully differentiated in their plastomes. In contrast, our sample of chloroplast genomes from 10 *P. monticola* individuals resulted in 9 distinct haplotypes. Based on prior studies of nuclear genetic variation in this species (Steinhoff *et al.* 1983), we explicitly divided our sample into 'northern' and 'southern' geographic groups (Fig. 1d), and examined chloroplast variation using AMOVA. This analysis shows that haplotype variation does not follow the pattern of nuclear differentiation, as ϕ*PT* (the partitioning of variance among groups, relative to total variance) was

This article is a US Government work and is in the public domain in the USA.

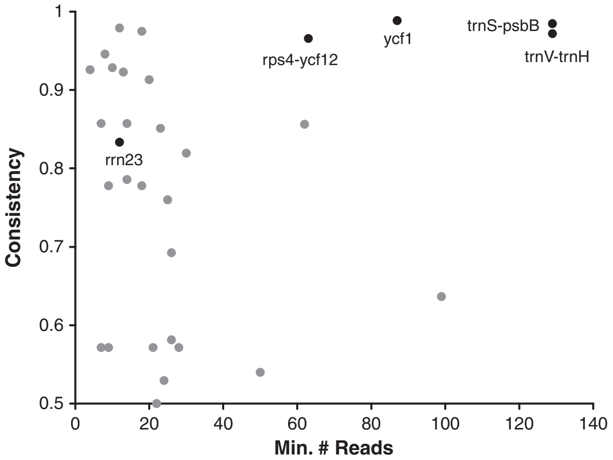

Fig. 3 A comparison of minimum read density and minimum base-call consistency was used to predict SNPs and false-positive SNPs from the variable sites identified in comparing two *Pinus torreyana* plastomes. Black circles are confirmed SNPs, and the identity of each region is noted; grey circles indicate confirmed false-positives.

insignificant for these chloroplast genomes (0%, $P = 0.861$).

Discussion

Recent dramatic improvements in DNA sequencing make it possible for simple genomes to be completely sequenced and compared in population and evolutionary genomics studies. In this analysis, we sequenced multiple barcoded chloroplast genomes simultaneously (four to six complete genomes, 16 partial genomes per lane), and have compared pairwise divergences of genomes reflective of intraspecific comparisons (*Pinus lambertiana, P. monticola, P. torreyan*a) or effectively intraspecific comparisons (*P. ayacahuite–P. flexilis, P. cembra–P. sibirica, P. albicaulis–P. lambertiana*). These intraspecific comparisons are based on 1.3 million aligned bases, and they add substantially to our understanding of the magnitude of intraspecific chloroplast genome variation in conifer trees.

One of the striking results to emerge from our analysis of full chloroplast genomes is that genome-wide sequence variation is very low within pine species. In all instances except one (*P. lambertiana*; discussed next), two selected chloroplast genomes from pine species showed fewer than 18 differences across the span of their full genome. This value is substantially lower than a comparison of two samples representing unique varieties of *Oryza sativa*, which showed 72 SNPs (Tang *et al.* 2004). As intraspecific chloroplast genome sequencing is in its infancy, we do not know if low divergence is an outcome specific to our sampling, a general condition for conifers (perhaps attributable to low absolute mutation rate, combined with a recent population expansion) or common throughout land plant chloroplast genomes. For the species we examined, it is clear that accurate estimates of nucleotide divergence and genealogical

Table 3 Divergence statistics for complete and partial chloroplast genomes in *Pinus*

	Chloroplast genome comparison					
	P. torreyana	*P. monticola* N–*P. monticola* S	*P. lambertiana* N–*P. lambertiana* S	*P. lambertiana* N–*P. albicaulis*	*P. ayacahuite–* *P. flexilis*	*P. cembra–* *P. sibirica*
Alignment length (bp)	120 362	39 150	114 000	117 504	117 546	117 228
Filtered SNPs	5	7	382	12	17	0
Pairwise distance	0.000047	0.000050	0.004303	0.000113	0.000165	0.0
(SE)Average bp compared	(0.000015)	(0.000017)	(0.000148)	(0.000031)	(0.000041)	(0.0)
	105 308	35 535	88 768	106 058	102 920	75 195
d*N*	0.000029	0.000000	0.002528	0.00008	0.000079	0.0
(SE)Average bp compared	(0.000026)	(0.000000)	(0.000247)	(0.000052)	(0.000037)	(0.0)
	34 547	12 727	32 432	37 636	37 779	29 277
d*S*	0.000057	0.000077	0.005344	0.000132	0.000215	0.0
(SE)Average bp compared	(0.000022)	(0.000025)	(0.000264)	(0.00003)	(0.000061)	(0.0)
	70 603	22 808	56 327	68 263	65 209	45 949
Estimated T_{div}	160	214	14 881	369	598	<60.6
(LCL, UCL)*	(41.5, 433)	(61.3, 547)	(5353, 33 172)	(120, 884)	(182, 1448)	(<3.4, 182)

Comparisons reflect intraspecific divergence in *P. torreyana*, *P. monticola* and *P. lambertiana*, and divergences that reflect near-conspecific comparisons (*P. ayacahuite–P. flexilis*; *P. cembra–P. sibirica*). The large values in the *P. lambertiana* N–S comparison result from introgression with a *P. albicaulis*-like chloroplast genome donor; for this reason, comparisons within *P. lambertiana* and between *P. lambertiana* N and *P. albicaulis* are shown. Standard errors (SEs) were determined using 500 bootstrap replicates.

*T_{div} is reported in thousands of years, with lower confidence (LCL) and upper confidence levels (UCL) noted. To calculate T_{div} for *P. cembra–P. sibirica*, we assumed an upper bound of one synonymous substitution for these genomes.

This article is a US Government work and is in the public domain in the USA.

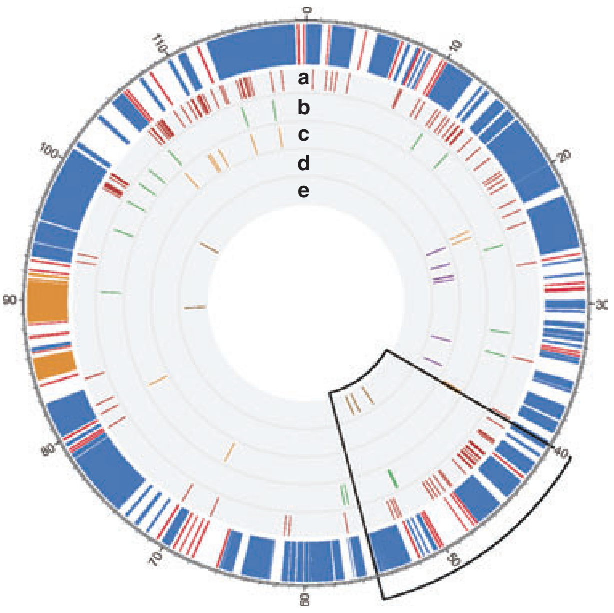

Fig. 4 Location of chloroplast genome SNPs in pairwise and population comparisons. Outer track shows the location of protein coding (blue), tRNA (red) and rRNA (orange) genes in the *Pinus* chloroplast genome; scale is in kbp. Inner tracks show the location of filtered SNPs for each comparison: (a) *Pinus lambertiana* N vs. S; (b) *Pinus ayacahuite* vs. *Pinus flexilis*; (c) *Pinus albicaulis* vs. *Pinus lambertiana* N; (d) *Pinus monticola* populations (partial genomes, positions 1–39 000); (e) *Pinus torreyana* island vs. mainland.

relationships will require full – not partial – genomes for robust resolution, and even this may be insufficient.

A second important finding from this analysis is that mutational variability across the genome is sufficiently heterogeneous that divergence estimates from a small number of loci could be misleading. For example, based on complete genomes, we find that *P. torreyana* shows the lowest pairwise divergence among the comparisons examined. In contrast, if we had chosen a 15 000 bp contiguous region spanning nucleotide positions 40 000–55 000 for our analysis (e.g. Fig. 4), we would have reached a different conclusion, namely, that the two samples of *P. torreyana* have greater pairwise divergence than samples from *P. ayacahuite*–*P. flexis* and *P. albicaulis*–*P. lambertiana* N. The uneven distribution of variation across closely related chloroplast genomes argues strongly for a plastome-scale approach to intraspecific evolutionary studies, an approach now feasible with Next-Gen sequencing.

Organismal insights from pairwise chloroplast genome divergences

A key motivation for this analysis was to determine whether mainland and island populations of *P. torreyana* showed detectable chloroplast genome divergence, and to frame that divergence in the context of more widespread species and species complexes. As noted, traditional molecular approaches to distinguish the remaining populations of *P. torreyana* – a species distributed across two locations with a total range of <30 km^2 – have been largely inconclusive due to the absence of molecular variation in this species (Ledig & Conkle 1983; Provan *et al.* 1999). One study of 59 allozymes identified two variable loci in a survey of 157 trees representing the island and mainland populations (Ledig & Conkle 1983). These polymorphisms represented fixed differences between the island and mainland populations, a finding consistent with the complete partitioning of plastome variation reported herein. The unusual partitioning of plastome variation in *P. torreyana* is consistent with subspecific recognition of these two disjunct populations (Haller 1986).

In the absence of other comparisons, it would have been reasonable to conclude that the low divergence observed within *P. torreyana* was related to its restricted range or its low census and (presumably) low effective population size. From these initial intraspecific comparisons, however, we have learned that chloroplast genome divergence within many pine species and species complexes is low, even for geographically widespread species (Table 3). For example, *P. monticola* is known to consist of geographically differentiated populations (Fig. 1d) based on isozyme data from 12 isozyme loci (Steinhoff *et al.* 1983) and nuclear sequence data (J. Syring, unpublished). This species has a range of ~370 000 km^2 (Fig. 1d), spanning 17° of latitude and 13° of longitude, and occurring in ecologically disparate regions (e.g. northern Rocky Mountains of British Columbia, serpentine barrens of the Klamath-Siskiyous, the southern Sierra Nevada of California) from sea level to 3350 m in elevation (Mirov 1967). Despite this larger range and census counts for *P. monticola* (perhaps 2–3 orders of magnitude larger) than *P. torreyana*, pairwise chloroplast genome divergence values for these two species are nearly equal (0.000047 for *P. torreyana*, 0.000050 for *P. monticola*; Table 3). Perhaps more sobering, *P. cembra* and *P. sibirica* have a combined range that is greater than 5 million km^2, with our samples separated by 4800 km. Sequencing of 75 kbp turned up no detectable differences between these two haplotypes, providing us with a clear lower bound for expected pairwise divergence. The low intraspecific divergence uncovered in *P. torreyana* appears not to be solely attributable to its rarity, as this feature appears to be the norm for *Pinus* (Table 3).

Based on our sample, pairwise divergence of *P. ayacahuite*–*P. flexilis* (0.000165; Table 3) set a realistic expectation for the upper bound of intraspecific comparisons in *Pinus*. This species complex is distributed from southern Alberta, Canada south to Honduras, with our

This article is a US Government work and is in the public domain in the USA.

samples collected from sites 2200 km apart (Fig. 1c). Analysis revealed a total of 17 SNPs across a comparison of 103 kbp, or ~1 SNP per 6 kbp. Even at this upper bound of intraspecific divergence, this comparison highlights the daunting challenge of locating SNPs for use in population genetic analysis, and reinforces the importance of massively parallel sequencing efforts. Figure 4 indicates that there is not a single gene, intron or spacer region found in our analyses that would serve as a 'marker locus' for future studies in *Pinus*, as SNPs are spaced irregularly across the chloroplast genome.

Although pairwise genome divergences for our chosen species pairs are comparable, the partitioning of genetic variation is uniquely structured by species. Genotyping in *P. torreyana* indicates that the 5 validated SNPs are fixed across populations, yielding estimates of complete differentiation ($\phi PT = 1.0$) for these populations. In contrast, our sampling of haplotype diversity in 10 accessions of *P. monticola* appears to show no geographic partitioning, with a calculated ϕPT of zero. Geographic subdivision of *P. lambertiana* into northern and southern chloroplast haplogroups was recently documented by Liston *et al.* (2007). This research found two major haplotypes that shared 10 fixed differences across a narrow geographic zone 150 km in width (demarcated in Fig. 1b), relative to the 1600 km latitudinal range of the species. Based on Liston *et al.*'s (2007) data (~2300 bp of sequence from *mat*K and the *trn*G intron), the preponderance of the variation was found between geographic groups ($\phi PT = 0.98$; $P = 0.003$). Therefore, we have documented cases of narrowly endemic pines with high plastid differentiation (*P. torreyana*), widespread pines with high plastid differentiation (*P. lambertiana*; Liston *et al.* 2007) and widespread pines with essentially no plastid differentiation (*P. monticola*). These three examples demonstrate the impact that each unique history has had on these species and genomes.

Genome-scale data continues to show the uniqueness of *P. lambertiana*. The pairwise divergence between the northern and southern populations is 26-fold greater than the next highest comparison (*P. ayacahuite–P. flexilis*). Prior phylogenetic analyses confidently placed the northern haplotype in a clade that includes *P. albicaulis* (whitebark pine) and other East Asian white pines, and the southern haplotype in a clade with North American white pines (Liston *et al.* 2007; Parks *et al.*, 2009). Liston *et al.* (2007) interpreted this phylogeographic pattern as a case of chloroplast introgression from *P. albicaulis* into the northern population of *P. lambertiana.* In this case, the high pairwise divergence value is more indicative of an interspecific rather than intraspecific comparison and suggests a cautionary approach be taken if large haplotypic divergences are uncovered in *Pinus.* Our estimate for the time of this introgression event was *c.* 370 000 years bp (Table 3). Pairwise divergence between the northern *P. lambertiana* haplotype and *P. albicaulis* is 0.000113, a value within the range of our other intraspecific comparisons.

To summarize, low plastome variation in *Pinus* species appears to be commonplace. Even in *P. monticola*, where we uncovered 9 unique haplotypes in 10 individuals, inter-population level diversity averaged 1 SNP per 20 kbp for each pairwise comparison. Where deviations from the expectation of low plastome diversity occur, as in the case of *P. lambertiana*, further investigation as to the cause is warranted. Although there appear to be narrow limits on plastome diversity, the hierarchical structure of that genetic diversity should be anticipated to vary according to the unique history of each species. Contextually, this indicates that there is nothing unusual about the haplotypic diversity of *P. torreyana*. On the one hand, the identified fixed differences found between the mainland and Santa Rosa Island populations support the uniqueness of these populations, and are suggestive that both populations should be a part of any long-term conservation plan. On the other hand, the low intraspecific plastome diversity is a trait that is shared with much more common and geographically widespread species.

What is next in 'Next-Generation' organelle sequencing?

A significant question remaining to be addressed in intraspecific organellar genome sequencing is the congruence between estimates of diversity and differentiation from nucleotides and microsatellites. As noted, chloroplast microsatellites have been successfully used to address many population and landscape level questions (Provan *et al.* 2001; Petit *et al.* 2005; Ebert & Peakall 2009). This is particularly true for conifers, where microsatellite-based estimates of haplotype variation can be striking, and as many as 235 haplotypes have been recorded from 311 individuals (Afzal-Rafii & Dodd 2007). This extreme variability seems unusual in light of the apparent quiescence of the remainder of the genome, but these differences could be expected given the magnitude of difference in positional mutation rates of nucleotides ($0.890–5.371 \times 10^{-10}$) and microsatellites ($3.2–7.9 \times 10^{-5}$; Provan *et al.* 1999). The extreme variability in microsatellites, combined with length constraints, has led many to suspect that genealogical estimates may be obscured through mutational 'homoplasy' (Estoup *et al.* 2002). The methods we used in our analysis are poorly suited to directly comparing sequence and microsatellite variation, because long single nucleotide repeats are difficult to assemble with short microreads (Cronn *et al.* 2008). With the development of paired-end sequencing

This article is a US Government work and is in the public domain in the USA.

and longer sequence reads, direct comparison of sequence- and microsatellite-based population genetic and genealogical estimates should be a high priority to evaluate the consistency of these methods.

Acknowledgements

We thank Bob Haller and Dieter Wilken for providing Torrey pine needles from the Santa Barbara Botanic Garden. Brian Knaus, Darren Smith, Stephanie Kim, Tim Butler, Mariah Parker-DeFeniks, Ugi Daalkhaijav, Sarah Sundholm, Angela Rodriguez, David Gernandt, Rongkun Shen, Todd Mockler, Mark Dasenko, Scott Givan, Chris Sullivan, Ismael Grachico, Jon Laurent and John Reeves provided critical assistance. Support provided by the Santa Barbara Botanic Garden and NSF IPY No. 0733078 to JBW, the OSU College of Science Venture Fund to AL, NSF ATOL No. 0629508 to AL and RC, and USFS PNW Research Station.

Conflicts of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Afzal-Rafii Z, Dodd RS (2007) Chloroplast DNA supports a hypothesis of glacial refugia over postglacial recolonization in disjunct populations of black pine (*Pinus nigra*) in western Europe. *Molecular Ecology*, **16**, 723–736.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Birky Jr CW (1978) Transmission genetics of mitochondria and chloroplasts. *Annual Review of Genetics*, **12**, 471–512.

Birky Jr CW (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms and models. *Annual Review of Genetics*, **35**, 125–148.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Bryant D, Wong W-K, Mockler T (2009) QSRA – a quality-value guided *de novo* short read assembler. *BMC Bioinformatics*, **10**, 69–75.

Chase MW, Soltis DE, Olmstead RG *et al.* (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden*, **80**, 528–580.

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, Little ELJ (1966) *Geographic Distribution of the Pines of the World*. US Department of Agriculture, Washington, DC, USA.

Cronn R, Liston A, Parks M *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**, e122.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**, e105.

Dolan P, Denver D (2008) TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics*, **9**, 250.

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resource*, **9**, 673–690.

Eckert A, Hall B (2006) Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. *Molecular Phylogenetics and Evolution*, **40**, 166–182.

Ennos RA (1994) Estimating the relative rates of pollen and seed migration among plant populations. *Heredity*, **72**, 250–259.

Erlich Y, Chang K, Gordon A *et al.* (2009) DNA Sudoku: harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research*, **19**, 1243–1253.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

Gernandt DS, Lopez G, Garcia SO, Liston A (2005) Phylogeny and classification of *Pinus*. *Taxon*, **54**, 29–42.

Gernandt DS, Magallon S, Geada Lopez G *et al.* (2008) Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. *International Journal of Plant Sciences*, **169**, 1086–1099.

Gernandt DS *et al.* (2009) Phylogenetic relationships of *Pinus* subsection *Ponderosae* inferred from rapidly evolving cpDNA regions. *Systematic Botany*; in press.

Gernandt DS, Hernández-León S, Salgado-Hernández E, Pérez de la Rosa JA (2009) Phylogenetic Relationships of *Pinus* Subsection *Ponderosae* Inferred from Rapidly Evolving cpDNA Regions. *Systematic Botany*, **34**, 481–491.

Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, **20**, 80–86.

Gugerli F, Senn J, Anzidei M *et al.* (2001) Chloroplast microsatellites and mitochondrial *nad1* intron 2 sequences indicate congruent phylogenetic relationships among Swiss stone pine (*Pinus cembra*), Siberian stone pine (*Pinus sibirica*), and Siberian dwarf pine (*Pinus pumila*). *Molecular Ecology*, **10**, 1489–1497.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Haller JR (1986) Taxonomy and relationships of the mainland and island populations of *Pinus torreyana* (Pinaceae). *Systematic Botany*, **11**, 39–50.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*, **18**, 802–809.

Höhn M, Gugerli F, Abran P *et al.* (2009) Variation in the chloroplast DNA of Swiss stone pine (*Pinus cembra* L.) reflects contrasting post-glacial history of populations from the Carpathians and the Alps. *Journal of Biogeography*, **36**, 1798–1806.

Hu X-S, Ennos RA (1997) On estimation of the ratio of pollen to seed flow among plant populations. *Heredity*, **79**, 541–552.

Hu X-S, Ennos RA (1999) Impacts of seed and pollen flow on population genetic structure for plant genomes With three contrasting modes of inheritance. *Genetics*, **152**, 441–450.

This article is a US Government work and is in the public domain in the USA.

Hudson ME (2007) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Jakobsson M, Hagenblad J, Tavare S *et al.* (2006) A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Molecular Biology and Evolution*, **23**, 1217–1231.

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, U.K.

Ledig FT, Conkle MT (1983) Gene diversity and genetic structure in a narrow endemic, Torrey pine (*Pinus torreyana* Parry ex Carr.). *Evolution*, **37**, 79–85.

Liston A, Parker-Defeniks M, Syring JV, Willyard A, Cronn R (2007) Interspecific phylogenetic analysis enhances intraspecific phylogeographical inference: a case study in *Pinus lambertiana*. *Molecular Ecology*, **16**, 3926–3937.

Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, **23**, 1–21.

McCauley DE (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology & Evolution*, **10**, 198–202.

Meusel H, Jäger E, Weinert E (1965) *Vergleichende Chorologie der Zentraleuropäischen Flora*. Gustav Fischer, Jena, Germany.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, e97; doi: 10.1093/nar/gkm566

Mirov NT (1967) The genus *Pinus*. *Ronald Press*, New York, NY, USA. 602 p.

Mitton JB, Kreiser BR, Latta RG (2000) Glacial refugia of limber pine (*Pinus flexilis* James) inferred from the population structure of mitochondrial DNA. *Molecular Ecology*, **9**, 91–97.

Mogensen HL (1996) The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany*, **83**, 383–404.

Moore M, Dhingra A, Soltis P *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Moreno-Letelier A, Piñero D (2009) Phylogeographic structure of *Pinus strobiformis* Engelm. across the Chihuahuan Desert filter-barrier. *Journal of Biogeography*, **36**, 121–131.

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics*, **18**, 269.

Morrison DA (2008) How to summarize estimates of ancestral divergence times. *Evolutionary Bioinformatics*, **4**, 75–95.

Neale DB, Sederoff RR (1989) Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. *Theoretical and Applied Genetics*, **77**, 212–216.

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. *Plant Journal*, **14**, 387–392.

Newton AC, Alnutt TR, Gillies ACM, Lowe AJ, Ennos RA (1999) Molecular phylogeography, intraspecific variation and the conservation of tree species. *Trends in Ecology & Evolution*, **4**, 140–145.

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at lowtaxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, **7**, 84.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Petit RJ, Aguinagalde I, de Beaulieu JL *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Petit RJ, Duminil J, Fineschi S *et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology*, **14**, 689–701.

Provan J, Soranzo N, Wilson NJ, Goldstein DB, Powell W (1999) A low mutation rate for chloroplast microsatellites. *Genetics*, **153**, 943–947.

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution*, **16**, 142–147.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution*, **24**, 192–200.

Shaw G (1914) *The Genus* Pinus. Harvard University, Cambridge, MA.

Steinhoff RJ, Joyce DG, Fins L (1983) Isozyme variation in *Pinus monticola*. *Canadian Journal of Forest Research*, **13**, 1122–1131.

Syring J, Farrell K, Businský R, Cronn R, Liston A (2007) Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Systematic Biology*, **56**, 163–181.

Tang J, Xia H, Cao M, Zhang X, Zeng W, Hu S, Tong W, Wang J, Wang J, Yu J, Yang H, Zhu L (2004) A comparison of rice chloroplast genomes. *Plant Physiology*, **135**, 412–420.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Waters ER, Schaal BA (1991) No variation is detected in the chloroplast genome of *Pinus torreyana*. *Canadian Journal of Forest Research*, **21**, 1832–1835.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*, **24**, 90–101.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Zerbino DR, Birney E (2008) VELVET: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Rich Cronn (US Forest Service, Pacific Northwest Research Station), Aaron Liston (Oregon State University) and PhD student Matt Parks are developing methods to facilitate the application of genomic data to population genetic and phylogenetic questions in non-model organisms. John Syring (Linfield College) studies the systematics, population genetics, and evolution of pines. Justen Whittall (Santa Clara University) studies plant evolution, applying genome-scale approaches to understanding the process of adaptation and speciation. Jason Buenrostro and Cindy Dick are members of the Whittall lab, and are responsible for validating Illumina SNPs.

This article is a US Government work and is in the public domain in the USA.