Article (refereed) - postprint

FalseDR. KEVIN DONNELLY (Orcid ID : 0000-0001-7633-9113)

DR. STEPHEN CAVERS (Orcid ID : 0000-0003-2139-9236)

Received Date : 13-Sep-2016

Revised Date : 25-Nov-2016

Accepted Date : 14-Dec-2016

Article type : Resource Article

Reconstructing the plant mitochondrial genome for marker discovery: a case study using Pinus

Kevin Donnelly[1,2*], Joan Cottrell[3], Richard A. Ennos[2], Giovanni Guiseppe Vendramin[4], Stuart A'Hara[3], Sarah King[1], Annika Perry[1], Witold Wachowiak[1,5] and Stephen Cavers[1]

[1]*NERC Centre for Ecology and Hydrology, Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, Scotland*

[2]*Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, Scotland*

[3]*Forest Research, Northern Research Station, Roslin, Midlothian EH25 9SY, Scotland*

[4]*Institute of Biosciences and BioResources, Division of Florence, National Research Council, Via Madonna del Piano, 10 - 50019 Sesto Fiorentino (Firenze), Italy*

[5]*Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland*

*Corresponding author Email:kevnne12@ceh.ac.uk

Keywords: marker discovery, high-throughput sequencing, mitochondrion, Pinus, phylogeography

Abstract

Whole-genome-shotgun (WGS) sequencing of total genomic DNA was used to recover ~1 Mbp of novel mitochondrial (mtDNA) sequence from *Pinus sylvestris* (L.) and three members of the closely-related *Pinus mugo* species complex. DNA was extracted from megagametophyte tissue from six mother trees from locations across Europe and 100 bp paired-end sequencing was performed on the Illumina HiSeq platform. Candidate mtDNA sequences were identified by their size and coverage characteristics, and by comparison with published plant mitochondrial genomes. Novel variants were identified, and primers targeting these loci were trialled on a set of 28 individuals from across Europe. In total, 31 SNP loci were successfully resequenced, characterising 15 unique haplotypes. This approach offers a cost effective means of developing marker resources for mitochondrial genomes in other plant species where reference sequences are unavailable.

Introduction

Plants possess three genomes: nuclear, mitochondrial, and chloroplast, which are subject to differing modes of inheritance. Unlike the nuclear genome, the organellar genomes of the mitochondrion and chloroplast are typically inherited uniparentally, from either the female or male parent exclusively (Clegg, 1990; Birky, 1995). For this reason, organellar genomes can be particularly useful for recovering some types of population genetic structure as their lower effective population size makes them more susceptible to differentiation by drift and hence signals of historic events are retained, such as those associated with post-glacial dispersal (Ennos, 1994; Sinclair et al., 1999). In the majority of angiosperms, both mitochondrial and chloroplast genomes are maternally inherited and dispersed via seed; whereas, in gymnosperms, chloroplast genomes are normally paternally inherited and transmitted via pollen (Neale and Sederoff, 1988; Mogensen, 1996). The chloroplast has been the predominant source of molecular markers, as chloroplast genomes are short, highly

conserved, non-recombining and exhibit a comparatively high base mutation rate relative to mitochondrial genomes (Provan et al., 1998; Byrne et al., 2003; Liu et al., 2012). However, their paternal mode of inheritance limits their utility in analyses of demographic history in gymnosperms, as population structure is rapidly broken down due to the long range dispersal of pollen. Mitochondrial markers would accelerate the study of seed-mediated gene flow in these species, however their development presents numerous challenges: plant mitochondrial genomes are relatively large, exhibit a high level of internal rearrangement, a very low base mutation rate (Wolfe et al., 1987), and very few completed sequences have been published to date. New approaches to marker discovery in the gymnosperm mitochondrial genome are needed, that take advantage of new sequencing methods to enable more comprehensive surveys of variation.

Pinus sylvestris is the most widely distributed of all conifers, with a range encompassing much of the northern hemisphere in Eurasia. To date, a very limited assortment of mitochondrial markers have been available for *P. sylvestris*, comprising structural variation at *cox*3 detected by RFLP analysis (Sinclair et al., 1998, 1999), and the indels at *nad*1 and *nad*7 (Soranzo et al., 2000; Naydenov et al., 2007; Pyhäjärvi et al., 2008). Although useful, the small number and low variability of these markers have limited their use. Here, using *P. sylvestris* and three of its close relatives as a case study, we demonstrate an approach based on a common next-generation sequencing platform to fast reconstruction of mitochondrial genome sequence and its application for marker discovery.

Presently, there are no completed references available for either the nuclear or mitochondrial genomes of *P. sylvestris* (although a scaffolded draft assembly has recently been made available for *P. taeda* (Neale et al., 2014)). Taxonomically, the nearest complete mitochondrial genome available is that of the gymnosperm *Cycas taitungensis* (Chaw et al., 2008), which is approximately 415 Kbp in length. Like those of other seed plants previously sequenced, the *C. taitungensis* mitochondrial genome is non-compact: some 89.9 % is comprised of non-coding sequence (introns, spacers, and pseudogenes), and relatively short (typically < 2 Kbp) repeated sequences account for 15.1 % of its

length. The small number of sequences available for *P. sylvestris* are for regions which encode genes (including introns), which are likely to be subject to a high degree of conservation between both species and individuals. Therefore, there are large quantities of intergenic sequence which have yet to be investigated, and which are likely to yield informative variants. Pinus nuclear genomes are exceptionally long: the *P. taeda* draft assembly is ~22 Gbp in length (Neale et al., 2014); estimates of the *P. sylvestris* genome size from flow cytometry range from ~21 to 27 Gbp (Bogunic et al., 2003 and Valkonen et al., 1994, respectively). Being both radically shorter and present in many more copies, the organellar genomes should receive substantially greater coverage during whole-genome-shotgun (WGS) sequencing and therefore be more amenable to *de novo* assembly.

We used WGS sequencing to obtain sequence data from the megagametophyte tissue of three individuals of *P. sylvestris* from native European populations and one individual of each of three closely related species from the *P. mugo* complex: *P. mugo*, *P. uliginosa*, and *P. uncinata*. The high throughput rate afforded by modern WGS sequencing and the ability subsequently to distinguish that which originates from the mitochondrial genome now enables large quantities of mitochondrial sequence to be acquired in the absence of prior enrichment for mitochondria via differential centrifugation or other means. By taking advantage of the increased coverage of organellar relative to nuclear genomes we were able to identify putative mitochondrial contigs from our assembly, which were then corroborated by comparison with published mtDNA sequences. In doing so, previously unexplored regions of the mitochondrial genome were recovered, and novel variants were captured at the single nucleotide level. Primers were developed targeting these regions, and trialled on a limited set of 28 individuals from 16 Pinus populations.

Methods

Sample Material

Seed collections were made from 18 geographically disparate native woodlands across Europe. For marker discovery, seeds were collected from three *Pinus sylvestris* mother trees and one of each of *P. mugo*, *P. uliginosa*, and *P. uncinata* (Table 1). For marker validation, we used 28 trees collected from sites across the European range (Table 1, Fig 1).

Preparation and DNA Extraction

In preparation for WGS sequencing, DNA extractions were performed using megagametophytes isolated from seed. The relative abundance of mtDNA to nuclear in megagametophyte tissue is unknown, however, (1C) megagametophyte tissue was preferred with a view to reducing the shotgun sequencing resources expended upon nuclear DNA content. Five to six seeds per sample (obtained from the same cone, and sharing a common maternal lineage) were placed on damp tissue in Petri dishes and allowed to germinate until the seed coats had visibly split, which typically occurred within one week. The seed was placed on the stage of a dissecting microscope, its coat was removed and the haploid megagametophyte tissue was then isolated from the embryo of each seed with the aid of forceps. To improve DNA yield all samples from one cone were then bulked in a single DNA extraction, performed using a *Qiagen DNeasy Plant Kit* as per the manufacturer's instructions (Qiagen, Venlo, Netherlands). For resequencing, DNA was extracted from both megagametophyte and needle tissue.

Illumina Sequencing

Sequencing of DNA extracts was performed at the *Istituto di Genomica Applicata* (IGA) in Udine, Italy. DNA was randomly fragmented by sonication using a Bioruptor (*Diagenode*); libraries based on DNA of endosperms of seeds obtained from single mother trees were enriched by 12-cycle PCR reaction, and following electrophoresis of the PCR products on a 2 % agarose gel, fragments in the

600 bp size range were selected and adapters were attached to the ends. A Genome Analyzer flowcell was prepared on the cluster station supplied and six 100 bp paired-end libraries were sequenced in multiplex on one lane of the *Illumina HiSeq2000* platform according to the manufacturer's instructions. Images from the instrument were processed using the manufacturer's pipeline software to generate FASTQ sequence files.

Assembly

FASTQ files were first assessed using 'FastQC' (Babraham Bioinformatics, Cambridge, UK) to examine per-base quality, and identify any anomalous patterns in base calling across read lengths. Using 'trimmomatic' (Bolger et al., 2014), adapter sequences were removed, and non-random G/C content was cropped from the ends of reads; low quality base calls (Phred score < 10) were then removed from leading and trailing ends of reads, and a 4 bp sliding window was applied, removing any reads where mean quality within the window was < 15. Following trimming, reads of length < 36 bp were excluded. Data from the sample that yielded the largest number of reads (Punkaharju) were used for *de novo* assembly using *fermi* (Li, 2012) with a default minimum overlap of 50 bp. Summary statistics for assembled contigs were produced via 'Quast' (Gurevich et al., 2013).

Identification of Candidate mtDNA Contigs

The Punkaharju assembly was used as a basis to develop a reference set of candidate mitochondrial sequences (Fig 2a). To estimate coverage depth, the reads used to produce the assembly were mapped to it using 'BWA' (Li and Durbin, 2009); filtering and sorting of mapped reads was performed using 'SAMtools' (Li et al., 2009), duplicate reads were removed using 'Picard' (https://broadinstitute.github.io/picard/), and coverage information was retrieved using 'Bedtools' (Quinlan and Hall, 2010).

Coverage of the nuclear genome was expected to be very low on account of its very large size: if we were to assume a genome size of 24 Gbp, then with 10 million 100 bp reads (similar to our data),

and in the *absence* of any organellar genomes, we would anticipate a mean coverage depth of around 4.2×10^{-4}. To further preclude short nuclear contigs from downstream analysis, the assembly was restricted to contigs ≥ 1 Kbp.

NCBI's BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was then used to perform nucleotide searches (blastn) of the subset of contigs against the mitochondrial genome of *C. taitungensis* (Genbank: AP009381.1) and the *P. taeda* draft mitochondrial genome (Neale et al., 2014) (available at http://pinegenome.org/pinerefseq/). Only hits with an expect value (e-value) < 0.0001 were recorded, and contigs sharing at least 500 bp identity with the published sequences were retained in the final candidate set.

Variant Detection

In order to reduce potential mismapping of genomic or chloroplast reads to the candidate mitochondrial sequences identified, a composite reference was produced which also incorporated the draft *P. taeda* nuclear genome (v1.01) (Neale et al., 2014; available at http://pinegenome.org/pinerefseq/), and the *P. sylvestris* chloroplast genome (GenBank: JN854158.1) (Fig 2b). The *P. taeda* nuclear genome was converted from a scaffold to individual contigs (dramatically reducing file size due to the large gaps present), and contigs were then filtered to include only those > 50 bp.

Each of the six WGS sequenced samples were then mapped to this composite reference using an identical procedure as described above. Variant detection was performed with GATK (DePristo et al., 2011); variants with a read-depth < 10 or > 200 were not recorded, in order to prevent inclusion of SNPs with poor support, or those with a high likelihood of being attributable to paralogues, respectively (large spikes in read depth are likely caused by erroneous mapping of highly similar reads from more than one location to the same reference locus; this may result in spurious SNP calls). Alignment data were visualised using 'Tablet' (Milne et al., 2013).

Sanger Resequencing of Polymorphic Loci

Each sample submitted for WGS sequencing consisted of bulked DNA from 5 – 6 endosperms of seeds from the same mother tree, each of which would be expected to share the same mitochondrial haplotype. Despite this, some within-mother variation was observed after reads were aligned to the candidate mitochondrial reference. Primers were designed to flank SNPs where this source of variation was minimal i.e. where aligned sequences from a maternal endosperm sample were consistently identical or non-identical to the reference.

An initial 48 loci were targeted among 28 individuals from across Europe (Table 1). DNA (~20 ng) was amplified in a total reaction volume of 20 µl using the following mixture: 1 µl DNA, 1X PCR buffer (160 mM $(NH_4)SO_4$, 670 mM Tris-HCl, 25 mM $MgCl_2$, 0.1 % Tween-20 at pH 8.8 (Bioron, Germany), 5 µM of each primer (Eurofins-MWG), 0.2 mM of each dNTP (VWR International), and 0.25 U Superhot Taq DNA polymerase (Bioron). PCR was performed using an initial denaturing phase at 95°C for 5 mins, followed by 30 cycles of 30 s at 94°C for denaturation, 57°C for 60 s annealing, and 72°C for 60 s extension. Final elongation was performed at 72°C for 10 mins, after which samples were held at 10°C. PCR products were electrophoresed on a 1.4% agarose gel and if a single product was obtained the remainder of the product was cleaned up with EXO-SAP IT (Affymetrix, UK) prior to sequencing. Sequences for each locus from each sample that were successfully amplified (31 in total; Table 2) were concatenated to produce a multi-locus haplotype for each individual (Supplementary Table 1). A maximum likelihood (ML) tree based on the General Time Reversible (GTR) model was constructed via MEGA6 (Tamura et al., 2013) using 10,000 bootstrap replicates. For the purposes of tree construction, missing data were imputed from the most closely-related complete haplotype.

In addition to newly discovered polymorphic loci, samples were also genotyped for the indels previously identified at *nad*1 and *nad*7 gene regions (Soranzo et al., 2000; Naydenov et al., 2007; Pyhäjärvi et al., 2008); alternative primers for *nad*7 were obtained from Danusevičius et al., 2013. The *nad*1 haplotypes were determined via Sanger sequencing, and *nad*7 on the basis of size

differences identified using a Licor 4300 DNA sequencer (LI-COR Biosciences, Nebraska, USA). PCR reaction mix was as previously described, but the forward primer was labelled with 700 nm dye (Eurofins-MWG) for subsequent fluorescent detection. PCR conditions were as follows: initial denaturing phase at 95°C for 5 mins, followed by 32 cycles of 40 s at 94°C for denaturation, 65°C for 75 s annealing, and 72°C 60 s for extension. Final elongation was performed at 72°C for 10 mins, after which samples were held at 10°C.

Results

Assembly and Contig Identification

The total number of paired-end reads obtained from each sample (total genomic DNA) varied from 3.2×10^7 to 6.2×10^7; the greatest number were obtained for the sample from Punkaharju, Finland. Prior to enforcing a minimum contig size of 1 Kbp (which dramatically reduced overall assembly length), the Finnish sample assembly was ~109 Mbp in length, and relatively AT-rich (Table 3). Of the contigs ≥ 1 kbp, two main groups were discernible on the basis of read depth (Fig 3). A BLAST search revealed that the high-coverage group consisted of chloroplast contigs, with the exception of one which was identified as bacterial in origin. All contigs identified as mtDNA according to their similarity to *P. taeda* and *C. taitungensis* mitochondrial genomes lay within the lower coverage group, the origin of the remainder of contigs in this group was not determined. Taken together, this lower coverage group consisted of 1274 contigs, with an overall length of 3.2 Mbp and a mean GC content of 42 %. The candidate mtDNA sequences had a combined length of ~1 Mbp [GenBank accession nos pending], and an elevated GC content (46 %) relative to the original unmodified assembly (40 %), and the chloroplast genome (39 %).

Sequencing Depth

Sequencing depth is reported in terms of the original *de novo* assembly (before subsetting), and separately for the candidate mitochondrial sequences and chloroplast genome after mapping reads to the composite reference (Fig 4). Median coverage was by far greatest for the chloroplast at 698x. The unmodified *de novo* assembly and the candidate mitochondrial contigs (a subset of the former) received substantially lower median coverage at 15x and 29x, respectively. These coverage distributions were, however, markedly different: mode coverage (i.e. the most common depth) occurred at 3x for the complete *de novo* assembly, and 31x for the mitochondrial subset.

Candidate Mitochondrial Sequences

The candidate mitochondrial subset consisted of 224 contigs, and ranged from 1 – 21 Kbp in length (Supplementary Fig 1). Although contigs were allocated to the subset on the basis of their similarity to the *C. taitungensis* and *P. taeda* mitochondrial genomes, a number of them matched the mitochondrial genomes of other closely-related species: including pines, *P. sylvestris*, *P. strobus*, and *P. monophylla*; other conifers *Abies sachalinensis, Larix mastersiana*, and *Picea smithiana*; angiosperms *Phoenix dactylifera*, *Ricinus communis*, and *Tripsacum dactyloides*. Candidate sequences were predominantly intergenic, however, coding sequences were also present for a number of mitochondrial genes.

Of the contigs which originally aligned to the *P. taeda* mitochondrial genome, but not to that of *C. taitungensis*, very little coverage was found on GenBank, and for many, even the best matches (typically plant mitochondrial or Pinus genomic sequence) shared < 5% identity.

Identification and Resequencing of Variants

Following PCR and Sanger sequencing, 31 SNPs in 30 contigs were successfully genotyped in 28 samples from across Europe (primer pairs are listed in Table 2). With one exception, all of the SNPs were transversions (A ↔ C or G ↔ T). Following concatenation across loci, 15 multi-locus

haplotypes could be distinguished (Table 1). Eleven of the 31 SNP loci differentiated Finnish samples from the others included in the study.

Only three SNPs were species specific within the sample, distinguishing *P. uncinata* from the other pines. Nevertheless, unique haplotypes were found within each of the species, but not all samples; one *P. mugo* individual of Polish origin was found to share an identical haplotype with *P. sylvestris* from Sweden and Italy. Generally, there was high degree of geographic correspondence among the haplotypes, and neighbouring populations shared similar or identical haplotypes (Fig 5). However, there were exceptions, as Swedish populations shared haplotypes with samples from sites in Italy and Austria, and were strongly differentiated from those in neighbouring Finland.

Samples were also typed at previously identified polymorphic loci at *nad*1 intron B/C and *nad*7 intron 1 (Supplementary Table 1). Following the naming convention of Naydenov et al. (2007), all samples at *nad*1 were found to possess haplotype 'A' (lacking an insertion); at *nad*7 all samples were scored as type 'A' (lacking a deletion) with the exception of those from Finland, all of which possessed type 'B' (5 bp deletion). Integration of the *nad*1 and *nad*7 loci with the multilocus SNP haplotypes did not increase resolution.

Discussion

Assembly and Coverage

Coverage of the *P. sylvestris* nuclear genome was expected to be extremely low on account of its large size. The original *de novo* assembly (total genomic DNA) was predominantly comprised of short contigs, and half of the assembly length was accounted for by contigs of fewer than 128 bp. Equivalent distributions for the chloroplast genome and candidate mtDNA contigs exhibited distinctly higher coverage and lower variability, as would be expected given the short size and greater ratio of organelle to nuclear genomes present in a cell. The mean GC content of the

candidate mtDNA subset (46 %) was markedly differentiated from the background of the total genomic assembly (40 %), and from the chloroplast genome (39 %), but was almost equal to that of the *C. taitungensis* mitochondrial genome, and comparable to mitochondrial genomes of other plant species.

Of the contigs ≥ 1 kbp, many were not identified as mtDNA using the similarity criteria described, despite occupying a similar range of read depths. Were these contigs assumed to be mitochondrial, the mtDNA assembly would be substantially longer (~3.2 Mb), with a reduced GC content (42 %). On the basis of the *P. taeda* draft and *C. taitungensis* mitochondrial genomes, it was assumed that of *P. sylvestris* would be highly similar in GC content. In construction of the *P. taeda* draft, scaffolds were selected on the basis of a GC content ≥ 44 %, and therefore contigs selected on the basis of similarity to that assembly, and that of the *C. taitungensis* mitochondrial genome, are likely to exhibit comparable values. There is precedent for somewhat lower values: Sloan et al. (2012) describe the mitochondrial genomes of four Silene species, including the largest known to date (11.3 Mbp), which exhibit GC content values of ~41 – 43 %. We might anticipate lower than average GC content for Pinus mitochondrial genomes, should their relatively long lengths be in part attributable to the assimilation or repetition of large sections of nuclear DNA.

Plant mitochondrial genomes are prone to high recombinational activity, contain large intergenic regions, and are known to incorporate DNA from both chloroplast and nuclear genomes (Knoop, 2004; Wang et al., 2007), the latter having similarly assimilated mtDNA (Timmis et al., 2004). Many of the assembled contigs which showed a high degree of similarity to the mitochondrial genomes of *P. taeda* and *C. taitungensis* did so for only a short portion of their length, suggesting a low degree of conservation outside of coding regions. Inversions were common among the sequences that aligned to *C. taitungensis*; these are likely to be ancient and may offer a useful resource for phylogenetic inference, and indeed structural changes in mtDNA have previously been used to resolve plant phylogenies (Manhart and Palmer, 1990; Dombrovska and Qiu, 2004). Following development,

candidate mtDNA sequences were searched against those available on GenBank. Contigs which had been identified based on their similarity to *C. taitungensis* were often found to share greater identity with the mitochondrial genomes of other plant species, in particular Pinus when available. Sequences that were retained on the basis of their similarity with the *P. taeda* mitochondrial genome (but not *C. taitungensis*), generally did not align with other published sequences: these sizeable intergenic regions may have origins in the nuclear genome, and share little identity with the mitochondrial genomes published thus far.

Complete chloroplast genomes are available for numerous species, and are significantly shorter than mitochondrial genomes, typically ranging from 120 – 160 Kbp in length (Palmer, 1985). For the purposes of marker discovery, we included references for the chloroplast genome of *P. sylvestris* and the nuclear genome of *P. taeda* when aligning reads to the candidate mitochondrial reference, to reduce the potential for mismapping of similar sequences between genomes. The read depth attained for the chloroplast genome (approaching 679x) was substantive, and shows that WGS approaches are well suited to studies of chloroplast diversity, as large numbers of samples could be multiplexed whilst maintaining a significant level of coverage.

Marker Discovery

After deriving the mitochondrial subset from our original assembly, we aligned WGS reads from a further five samples taken from *P. sylvestris* and the *P. mugo* complex. On the initial pass, a large number of the potential SNPs detected between samples were also bi-allelic within samples. There are two possible reasons for this, which are not mutually exclusive: heteroplasmy, whereby an individual may possess more than one mitochondrial haplotype, or the occurrence of paralogous sequences elsewhere in the mitochondrial or nuclear genomes.

Heteroplasmy is believed to be common for plant mitochondria (Kmiec et al., 2006), and in species where mitochondria are maternally inherited, may arise from leakage of the mitochondrial genome

of the paternal parent. This can in turn facilitate the establishment of new recombinant haplotypes (Städler and Delph, 2002; Pearl et al., 2009), and may provide a means by which to offset 'Muller's Ratchet': the otherwise irreversible accumulation of deleterious mutations in asexually reproducing populations (Muller, 1964; Neiman and Taylor, 2009). Paternal leakage of the mitochondrial genome has previously been reported to occur in other Pinus species (Wagner et al., 1991), and it would be of interest to determine the rate at which it may occur in *P. sylvestris* and *P. mugo*. This would have implications for studies of population structure, as mtDNA markers have previously been assumed to disperse exclusively via seed. However, the strong spatial structure that has been observed suggests that pollen-mediated dispersal of mitochondria is rare. Another, more likely, explanation is the occurrence of paralogues which exhibit minor differences; these align to the same reference loci, superficially resembling SNPs. Duplicate sequences may be present throughout the length of the mitochondrial genome, but given the propensity for exchange of genetic material, it's conceivable that within sample variation also occurs as a result of paralogues between the nuclear and mitochondrial genomes. The nuclear genome is subject to a significantly higher base mutation rate, and paralogues that lie within the nuclear genome are therefore more likely to deviate than those within the mitochondrial genome. Primers were designed to target those SNP loci that exhibited little to no variation within samples to minimise the potential for ambiguous variant calls during resequencing.

In total, 31 SNPs were successfully resequenced, almost all of which were transversions. The rarity of transitions seems unusual, but is not without precedent: Wolfe et al. (1987) found that transitions comprised less than 50% of SNPs occurring in plant mitochondrial genes, and in a study of date palm cultivars, ~70% of the mitochondrial substitutions reported by Sabir et al. (2014) were transversions, in this case found overwhelmingly in intergenic regions. Although we performed a number of filtering steps to generate a set of mitochondrial contigs from an assembly of total genomic DNA, an analysis of the products of controlled crosses involving parents with differing haplotypes would be of benefit to strengthen the evidence that the new markers indeed lie within the mitochondrial

genome by exhibiting maternal inheritance and a consistently haploid profile. Each sample used for marker discovery comprised a pooled DNA extract of several megagametophytes from the same mother, and although SNPs were filtered on the basis of low within-sample variation, it is possible that nuclear polymorphisms detected between samples could be misconstrued as mitochondrial. This, however, seems unlikely given the extremely low nuclear coverage predicted; furthermore, heterozygotes were not apparent in the 28 samples resequenced via Sanger. Due to the potential homology between the two genomes, primers developed for the mitochondrial genome might simultaneously target nuclear sequence; nevertheless, due to the overwhelming number of copies present, any signal observed following amplification should originate predominantly from the mitochondrial genome.

Distribution of Variants in Europe

A number of investigators have attempted to retrace post-glacial migration routes and locate the refugia occupied by *P. sylvestris* using mtDNA variation, and with some success (Naydenov et al., 2007; Pyhäjärvi et al., 2008; Sinclair et al., 1998, 1999; Soranzo et al., 2000). However, these studies have been constrained by a very limited pool of markers located around gene coding regions that are highly conserved among plant taxa. We tested our panel of novel SNP markers on a pilot sample of 28 individuals drawn from across the European range, consisting of *P. sylvestris* and three subspecies of the closely related *P. mugo* complex (*P. mugo*, *P. uliginosa*, and *P. uncinata*). In doing so, 15 multilocus haplotypes were identified; however, the discovery of further unique haplotypes is likely should larger samples be genotyped at the same loci.

Our results were broadly consistent with those of earlier studies, which indicated the presence of distinct refugia in southern and northern Europe. The Iberian Peninsula is known to have harboured Scots pine populations during the last glacial maximum, but is not believed to have contributed to

the recolonisation of northern Europe (Soranzo et al., 2000): we observed two unique haplotypes among Spanish individuals, distinct both from one another and from those in the remainder of the mainland distribution. Strong differentiation was observed between the haplotypes found in Swedish and Finnish populations, contributing to the evidence that two ancestral lineages meet in Fennoscandia (Naydenov et al., 2007; Pyhäjärvi et al., 2008). Interestingly, Scottish populations, which represent the western range edge, were more similar to those in Poland than any of the others sampled: although these regions have been shown to share haplotypes previously, here we also found them to be differentiated from populations in Italy or Sweden.

Also included in the study were samples from the closely related *P. mugo* species complex; *P. mugo*, *P. uliginosa*, and *P. uncinata*. Haplotypes from these species did not form a distinct monophyletic clade, one of the Polish *P. mugo* samples shared a haplotype present in Swedish and Italian *P. sylvestris*. *P. sylvestris* and species from the *P. mugo* complex have a recent common ancestry and are known to hybridise in the wild, and no fixed differences have been observed at nuclear loci (Wachowiak and Prus-Głowacki, 2008; Wachowiak et al., 2013). As mitochondrial markers are maternally inherited, they should provide a useful means to investigate the extent and direction of hybridisation in future studies.

Modern sequencing techniques enable large quantities of novel sequence data to be obtained at relatively low cost: here, by exploiting the distinctive coverage profile of the mitochondrial genome, we recovered ~1 Mbp of sequence and identified 31 new SNP markers from an initial panel of only six samples. We designed and tested primers targeting these loci and successfully verified their application in a multispecies sample set, recovering 15 haplotypes. A more extensive sample of European Pinus is now required to ascertain the extent to which new inferences can be drawn on the basis of the expanded marker set. Efficient discovery of further novel polymorphisms may be facilitated by targeted resequencing of known mtDNA sequence, effectively concentrating high-throughput sequencing effort on specific regions of interest across multiple individuals. Here, the

emphasis was upon *P. sylvestris* and *P. mugo*, however, the same techniques may be applied to develop novel markers and genomic resources for organelles of other plant species.

Acknowledgements

KD would like to thank Pär Ingvarsson for kindly agreeing to host him at Umeå University, Doug Scofield for his valuable bioinformatic guidance, and Procogen for funding this visit. KD was funded by a Natural Environment Research Council/Collaborative Awards in Science and Engineering PhD studentship in conjunction with Forest Research (Grant No: NE/H003959/1). WGS sequencing was undertaken with the support of EVOLTREE. WW acknowledges financial support from the Polish National Science Centre (Grant No: DEC-2012/05/E/NZ9/03476). We would also like to acknowledge valuable feedback from two anonymous reviewers in production of the manuscript.

References

Birky, C.W. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc. Natl. Acad. Sci. *92*, 11331–11338.

Bogunic, F., Muratovic, E., Brown, S.C., and Siljak-Yakovlev, S. (2003). Genome size and base composition of five Pinus species from the Balkan region. Plant Cell Rep. *22*, 59–63.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. *30*, 2114–2120.

Byrne, M., Macdonald, B., and Brand, J. (2003). Phylogeography and divergence in the chloroplast genome of Western Australian Sandalwood (*Santalum spicatum*). Heredity *91*, 389–395.

Chaw, S.-M., Shih, A.C.-C., Wang, D., Wu, Y.-W., Liu, S.-M., and Chou, T.-Y. (2008). The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol. Biol. Evol. *25*, 603–615.

Clegg, M.T. (1990). Molecular diversity in plant populations. 98–115.

Danusevičius, D., Buchovska, J., Stanys, V., Šikšnianienė, J.B., Marozas, V., and Bendokas, V. (2013). DNA marker based identification of spontaneous hybrids between *Pinus mugo* and *P. sylvestris* at the Lithuanian sea-side. Nord. J. Bot. *31*, 344–352.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dombrovska, O., and Qiu, Y.-L. (2004). Distribution of introns in the mitochondrial gene nad1 in land plants: phylogenetic and molecular evolutionary implications. Mol. Phylogenet. Evol. *32*, 246–263.

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

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinforma. Oxf. Engl. *29*, 1072–1075.

Kmiec, B., Woloszynska, M., and Janska, H. (2006). Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Curr. Genet. *50*, 149–159.

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. *46*, 123–139.

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly. Bioinformatics *28*, 1838–1844.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Liu, Y., Yang, S., Ji, P., and Gao, L. (2012). Phylogeography of *Camellia taliensis* (Theaceae) inferred from chloroplast and nuclear DNA: insights into evolutionary history and conservation. BMC Evol. Biol. *12*, 92.

Manhart, J.R., and Palmer, J.D. (1990). The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. Nature *345*, 268–270.

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., and Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. *14*, 193–202.

Mogensen, H.L. (1996). The Hows and Whys of Cytoplasmic Inheritance in Seed Plants. Am. J. Bot. *83*, 383–404.

Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. *1*, 2–9.

Naydenov, K., Senneville, S., Beaulieu, J., Tremblay, F., and Bousquet, J. (2007). Glacial vicariance in Eurasia: mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. BMC Evol. Biol. *7*, 233.

Neale, D.B., and Sederoff, R.R. (1988). Inheritance and Evolution of Conifer Organelle Genomes. In Genetic Manipulation of Woody Plants, J.W. Hanover, D.E. Keathley, C.M. Wilson, and G. Kuny, eds. (Springer US), pp. 251–264.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. *15*, R59.

Neiman, M., and Taylor, D.R. (2009). The causes of mutation accumulation in mitochondrial genomes. Proc. R. Soc. Lond. B Biol. Sci. *276*, 1201–1209.

Palmer, J.D. (1985). Comparative Organization of Chloroplast Genomes. Annu. Rev. Genet. *19*, 325–354.

Pearl, S.A., Welch, M.E., and McCauley, D.E. (2009). Mitochondrial Heteroplasmy and Paternal Leakage in Natural Populations of *Silene vulgaris*, a Gynodioecious Plant. Mol. Biol. Evol. *26*, 537–545.

Provan, J., Soranzo, N., Wilson, N.J., McNicol, J.W., Forrest, G.I., Cottrell, J., and Powell, W. (1998). Gene–pool variation in Caledonian and European Scots pine (*Pinus sylvestris* L.) revealed by chloroplast simple–sequence repeats. Proc. R. Soc. Lond. B Biol. Sci. *265*, 1697–1705.

Pyhäjärvi, T., Salmela, M.J., and Savolainen, O. (2008). Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. Tree Genet. Genomes *4*, 247–254.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Sabir, J.S.M., Arasappan, D., Bahieldin, A., Abo-Aba, S., Bafeel, S., Zari, T.A., Edris, S., Shokry, A.M., Gadalla, N.O., Ramadan, A.M., et al. (2014). Whole Mitochondrial and Plastid Genome SNP Analysis of Nine Date Palm Cultivars Reveals Plastid Heteroplasmy and Close Phylogenetic Relationships among Cultivars. PLoS ONE *9*, e94158.

Sinclair, W.T., Morman, J.D., and Ennos, R.A. (1998). Multiple origins for Scots pine (*Pinus sylvestris* L.) in Scotland: evidence from mitochondrial DNA variation. Heredity *80*, 233–240.

Sinclair, W.T., Morman, J.D., and Ennos, R.A. (1999). The postglacial history of Scots pine (*Pinus sylvestris* L.) in Western Europe: evidence from mitochondrial DNA variation. Mol. Ecol. *8*, 83–88.

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., and Taylor, D.R. (2012). Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLoS Biol *10*, e1001241.

Soranzo, N., Alia, R., Provan, J., and Powell, W. (2000). Patterns of variation at a mitochondrial sequence-tagged-site locus provides new insights into the postglacial history of European *Pinus sylvestris* populations. Mol. Ecol. *9*, 1205–1211.

Städler, T., and Delph, L.F. (2002). Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant. Proc. Natl. Acad. Sci. *99*, 11730–11735.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. *30*, 2725–2729.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. *5*, 123–135.

Valkonen, J.P.T., Nygren, M., Ylönen, A., and Mannonen, L. (1994). Nuclear DNA content of *Pinus sylvestris* (L.) as determined by laser flow cytometry. Genetica *92*, 203–207.

Wachowiak, W., and Prus-Głowacki, W. (2008). Hybridisation processes in sympatric populations of pines *Pinus sylvestris* L., *P. mugo* Turra and *P. uliginosa* Neumann. Plant Syst. Evol. *271*, 29–40.

Wachowiak, W., Boratyńska, K., and Cavers, S. (2013). Geographical patterns of nucleotide diversity and population differentiation in three closely related European pine species in the *Pinus mugo* complex. Bot. J. Linn. Soc. *172*, 225–238.

Wagner, D.B., Dong, J., Carlson, M.R., and Yanchuk, A.D. (1991). Paternal leakage of mitochondrial DNA in Pinus. Theor. Appl. Genet. *82*, 510–514.

Wang, D., Wu, Y.-W., Shih, A.C.-C., Wu, C.-S., Wang, Y.-N., and Chaw, S.-M. (2007). Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA. Mol. Biol. Evol. *24*, 2040–2048.

Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. *84*, 9054–9058.

Data Accessibility

Sequence data will be made available via GenBank prior to publication, and accession numbers will be provided in the text.

Author Contributions

The study was conceived by KD, JC, RAE, GGV, and SC; lab work was performed by KD, SAH, SK, AP, and WW; bioinformatic analysis was performed by KD; manuscript was written by KD, with input and revisions provided by JC, RAE, GGV, WW, and SC.

Tables and Figures

Figure 1 Map displaying species and location of origin for samples analysed.

Figure 2 Flowchart describing development of candidate mtDNA contigs and marker discover. A) *De novo* assembly built from sample with largest data set, and reads mapped to determine coverage; contigs then filtered by size, before comparison with published mtDNA sequences to produce final set. B) Composite reference constructed from the candidate mtDNA sequences as well as the draft nuclear genome of *P. taeda*, and the *P. sylvestris* chloroplast genome to reduce potential for erroneous variant calls due to read mismapping. WGS sequenced reads from all samples mapped to this reference and screened for variants.

Figure 3 Assembled contigs ≥ 1 Kbp plotted with respect to GC content and mean read depth; identity was assigned via BLAST search (see methods). All contigs with exceptionally high read depth were attributed to the chloroplast, with the exception of one which was found to be bacterial in origin. Contigs identified as mitochondrial exhibited markedly lower coverage, and relatively high GC content; coverage was comparable for contigs of undetermined origin.

Figure 4 Distribution of read depth for a) the unmodified *de novo* assembly, and b) the candidate mitochondrial contigs and chloroplast reference genome. Read depth was log transformed prior to plotting ($\log_{10}(\text{Depth} + 1)$); values are presented on the original scale. All data shown are based upon read-mapping of the *P. sylvestris* sample from Punkaharju, Finland.

Figure 5 Unrooted cladogram constructed via maximum likelihood using the general time reversible model and 10,000 bootstraps. Node values represent bootstrap support. Due to the occurrence of missing data, 25 of the possible 31 loci were used in tree construction. Sample abbreviations correspond to those listed in Table 1; those used in WGS marker discovery are marked with a '*'.

Table 1 Origin of samples sequenced by WGS for marker discovery, and samples sequenced by Sanger for marker validation. The mitochondrial haplotypes (A - O) are listed for each of the populations in which they were observed.

Table 2 Forward and reverse primers targeting each of the 31 SNP loci located on 30 contigs. PCR conditions were held constant throughout (see methods).

Table 3 Descriptive statistics for the *de novo* assembly produced using reads from the Finnish (Punkaharju) sample of *P. sylvestris*. When ordered by increasing length, the N50 and L50 describe the length of the contig which occurs at 50 % of the total assembly length, and the minimal number of contigs required to cover that length, respectively.

Table 1 Origin of samples sequenced by WGS for marker discovery, and samples sequenced by Sanger for marker validation. The mitochondrial haplotypes (A - O) are listed for each of the populations in which they were observed.

	Species	**Country**	**Site**	**Abbreviation**	**Lat**	**Long**	**No Samples**	**Haplotype(s)**
Used for marker discovery	*P. sylvestris*	Finland	Punkaharju	Fin_PK	61.76	29.29	1	E
		Scotland	Glen Loy	Scot_GL	56.91	−5.13	1	A
		Spain	Valsain	Spa_VS	40.87	-4.04	1	K
	P. mugo	Poland	Slaskie Kamienie	Pol_SK	50.77	15.60	1	N
	P. uliginosa	Poland	Węgliniec Reserve	Pol_WR	51.28	15.24	1	M
	P. uncinata	Andorra	Vall de Ransol	And_VD	42.55	1.61	1	I
Used for marker validation	*P. sylvestris*	Austria	Pernitz	Aus_PZ	47.91	16.00	1	C
		Finland	Kolari	Fin_KL	67.18	24.05	2	E
		Finland	Punkaharju	Fin_PK	61.76	29.29	2	D/E
		Italy	Cella di Palmia	Ita_CP	43.71	11.15	3	O
		Poland	Jarocin	Pol_JR	51.97	17.48	2	B
		Scotland	Glen Tanar	Scot_GT	57.05	-2.86	1	A
		Scotland	Rothimurchus	Scot_RM	57.15	-3.77	2	A
		Scotlan	Shieldaig	Scot_SD	57.5	-	3	A

		d			1	5.64		
		Spain	Trevenque	Spa_TV	37.08	-3.55	2	F
		Spain	Valsain	Spa_VS	40.87	-4.04	1	J
		Sweden	Krp. Tjärnbergshed en	Swed_KT	64.62	20.80	3	C/O
		Sweden	Väster Mjöingen	Swed_VM	62.75	13.57	2	G/O
	P. mugo	Poland	Slaskie Kamienie	Pol_SK	50.77	15.60	1	O
		Romania	Eastern Carpathians	Rom_EC	47.57	24.80	1	L
	P. uliginosa	Germany	Mitelwalde	Ger_MW	47.48	11.27	1	M
	P. uncinata	Andorra	Vall de Ransol	And_VD	42.58	1.64	1	H

Table 2 Forward and reverse primers targeting each of the 31 SNP loci located on 30 contigs. PCR conditions were held constant throughout (see methods).

Primer Pair	Contig	Forward (5' - 3')	Reverse (5' - 3')	Amplicon Size	SNP Position	SNP
1	*GenBank_Acc_1*	CTACAAGCGACACAGGAGCA	AGTTTCAATTTACTTATTGGCCCCC	339	3630	G/T
2	*GenBank_Acc_2*	ACGAAGTCAACACCGGGAAA	GTGAGAGAGAAAGAGCTCAGGT	349	3792	A/C
3	*GenBank_Acc_3*	ATTCCTGTGCTTGGTTGGGA	GGCGCTTACCCACACACTTA	570	14419	G/T
4	*GenBank_Acc_4*	TTTGATGGGGTACGGCACTT	ACCCAGAAGGTACGTGTGGT	597	7782	A/C
5	*GenBank_Acc_5*	TGAGTTCGTTGACCGCGTAA	TCAGGCGAGCTTGTGCTTTA	514	6408	A/C
6	*GenBank_Acc_5*	AGAAAGCAGTGATCCCGAGC	TTGAAGCGGACCTCATCGAC	444	1688	A/C
7	*GenBank_Acc_6*	TTCCATTCTTCGCCACGGAA	ATCTGCCGAACAAGGACCAG	574	1375	G/T
8	*GenBank_Acc_7*	ACACAGCAATGATGCAACGG	TGGTGAAACTGATGCCCCTT	417	2662	G/T
9	*GenBank_Acc_8*	CGTTGAACGGACCTTTGCAG	TAAAATACGGGTCCACCGGC	311	422	A/C
10	*GenBank_Acc_9*	CACTCAGAACCGGCTTGACA	TTTAGGCTTCTGGCCCTTGG	564	1301	A/C
11	*GenBank_Ac*	GATCGGGTCGGAGGC	AGTTGAAGCAAGCCAGC	369	3250	A/C

	c_10	ATAAT	AAG			
12	*GenBank_Acc_11*	TTTACGAAGCCCTTGGCGAT	CTGAACCGGGTGTAGCCTTT	548	1866	G/T
13	*GenBank_Acc_12*	CATCCTCTCCTCTCGATGGC	GCTTTTGGCTTGGTGCGAAT	358	208	G/T
14	*GenBank_Acc_13*	AGCTGGTCATAGCCAATAGCC	CCAAGTTTCATGCGCTCACA	600	12002	G/T
15	*GenBank_Acc_14*	TTCGAGGGTCGAGAACATGG	GTCAGTGCTTGTCAATGCCG	502	5460	A/C
16	*GenBank_Acc_15*	GGCCTTACGGCTCCTGAAAT	GTACCCTGGACGGAACACAG	486	286	A/C
17	*GenBank_Acc_16*	CGGAGCGAGGTGAAGAAACT	GCGAGAAGCAGTAGTGGGTT	593	2178	G/T
18	*GenBank_Acc_17*	TCCGATGATGAGGTGGAGGT	AGTTGAAGGCAGGAAGGTCG	522	4108	G/T
19	*GenBank_Acc_18*	TGCATTCTGGCTGGCTTTCT	GGCGTCGATAGACTCGGTTT	434	1528	G/T
20	*GenBank_Acc_19*	GGCATGTCCGCTATGGAAGT	AGGCTCCGGAAGTACCTGT	398	1517	G/T
21	*GenBank_Acc_20*	ATCGGCTCGACTGTTAAGGC	ACTGGTGCTCAAACCACACT	419	558	G/T
22	*GenBank_Acc_21*	GGTTGGTTGATCCATCCGGT	CCGGCTTGGGTACGTCTTTT	558	4134	G/T
23	*GenBank_Acc_22*	TGCGACCTGTGAATGGATGT	CGGCGGTTCTAGCCTTGATT	558	4372	A/C
24	*GenBank_Acc_23*	TTTGCTCCTGCTGGTGAGAC	GGCGAAATCCTCTCTCGACG	331	3726	C/T
25	*GenBank_Acc_24*	GGGCACAAGGGGATCTATGG	TTGGCTTAACGCTTCGGAAC	484	3457	G/T
26	*GenBank_Acc_25*	ATTGCCTGCCACTACAGACC	GGGAAAAGGTCTCCCCAGTG	541	1559	G/T
27	*GenBank_Acc_26*	GGGATTAGACCGGCAAACCT	CCCTGCAACTTGCCTCTGAA	410	3864	G/T
28	*GenBank_Acc_27*	GAGGGAGAGAACGCGAAATC	CCCATTTCCTCCCTACACGA	574	3963	G/T
29	*GenBank_Acc_28*	GGCGGCGGTATAGGGAAATA	GACCGGGTCAATCCTCTGTT	600	2147	A/C
30	*GenBank_Acc_29*	GGTGTACGTGTGGTAGGTGG	CACCACCGAAAGACGAGGAA	585	6557	G/T
31	*GenBank_Acc_30*	GCCAATCCACCTGCATGTTC	GAGGTGCGGGAAGTAATGCT	592	679	G/T

Table 3 Descriptive statistics for the *de novo* assembly produced using reads from the Finnish (Punkaharju) sample of *P. sylvestris*. When ordered by increasing length, the N50 and L50 describe the length of the contig which occurs at 50 % of the total assembly length, and the minimal number of contigs required to cover that length, respectively.

Reference	Length	GC%	N50	L50
De novo Assembly	108,996,556	39.8	128	286,153
Chloroplast	119,793	38.5	Single Contig	Single Contig
Mitochondrial	985,624	46.3	6,425	52