Published online 27 August 2008

Nucleic Acids Research, 2008, Vol. 36, No. 19 ***e122***
doi:10.1093/nar/gkn502

Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology

Richard Cronn[1,*], Aaron Liston[2,3], Matthew Parks[2], David S. Gernandt[4], Rongkun Shen[2,3] and Todd Mockler[2,3]

[1]Pacific Northwest Research Station, USDA Forest Service, Corvallis, OR 97331, [2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, [3]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA and [4]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, México DF 04510, Mexico

Received April 9, 2008; Revised June 26, 2008; Accepted July 21, 2008

ABSTRACT

Organellar DNA sequences are widely used in evolutionary and population genetic studies, however, the conservative nature of chloroplast gene and genome evolution often limits phylogenetic resolution and statistical power. To gain maximal access to the historical record contained within chloroplast genomes, we have adapted multiplex sequencing-by-synthesis (MSBS) to simultaneously sequence multiple genomes using the Illumina Genome Analyzer. We PCR-amplified ~120 kb plastomes from eight species (seven *Pinus*, one *Picea*) in 35 reactions. Pooled products were ligated to modified adapters that included 3 bp indexing tags and samples were multiplexed at four genomes per lane. Tagged microreads were assembled by *de novo* and reference-guided assembly methods, using previously published *Pinus* plastomes as surrogate references. Assemblies for these eight genomes are estimated at 88–94% complete, with an average sequence depth of 55× to 186×. Mononucleotide repeats interrupt contig assembly with increasing repeat length, and we estimate that the limit for their assembly is 16 bp. Comparisons to 37 kb of Sanger sequence show a validated error rate of 0.056%, and conspicuous errors are evident from the assembly process. This efficient sequencing approach yields high-quality draft genomes and should have immediate applicability to genomes with comparable complexity.

INTRODUCTION

The sequencing of the first chloroplast genome, that of *Nicotiana tabacum* (1), was an early demonstration of the power of genomic approaches, and it heralded a revolution in studies of the photobiology, biochemistry, physiology and evolutionary history of this important organelle that had its origins as a free living cyanobacterium. Documentation of chloroplast sequence variation has also been an essential tool in plant population and evolutionary studies for over two decades. At an average size of 120–160 kb, and containing ~130 genes, chloroplast genomes are sufficiently large and complex to include structural and point mutations that mark population-level processes and deeper evolutionary divergence. Further, the haploid state and uniparental transmission give chloroplast genes and genomes an effective population size approximately one-fourth of a nuclear locus (2,3). This has the effect of making chloroplast genes more responsive than nuclear genes to stochastic processes like drift and founder events, a property that has been exploited for testing hypotheses of seed (and less commonly pollen) dispersal, migration/colonization routes, intraspecific differentiation and interspecific introgression (4–6).

An important feature of chloroplast genomes is their high degree of sequence conservation. Strong purifying selection, acting on photosynthetic machinery, imposes clear constraints on nucleotide and structural mutation rates (7,8). Because of these constraints, structural changes in noncoding regions are often used to study population differentiation (9–12), while noncoding and coding sequences have been most successfully used to resolve more distant genealogical relationships [genus and above; (13–15)].

*To whom correspondence should be addressed. Tel: +541 750 7291; Fax: +541 750 7329; Email: rcronn@fs.fed.us
Present address:
Rongkun Shen, Vollum Institute, Oregon Health & Science University, Portland, OR 97239, USA

Due to the severe limits imposed on chloroplast divergence, large amounts of chloroplast DNA sequence are often required to detect statistically robust population differentiation or genealogical resolution (16). Accessing large amounts of the genome is feasible, but it has necessitated the use of novel cloning and high-throughput Sanger sequencing (17), or laborious primer walking approaches (18).

The recent, dramatic improvements in second generation sequencing make it possible to acquire entire genomes—and possibly many simple genomes—at a fraction of the time and cost of traditional approaches (19–23). These sequencers maximize capacity by generating short pyrosequencing (100–250 bp; 454 Life Sciences, Branford, CT) to micro 'sequencing by synthesis' reads (35–50 bp; Illumina 1G/Solexa, Illumina Inc., San Diego, CA). While individual sequences are far shorter than traditional Sanger sequencing, reads are sufficiently numerous that 20–100 Mbp (short read platforms) to 2–4 Gbp (microread platforms) can be sequenced on a single run. Moore *et al.* (21) recently demonstrated the merit of this approach by sequencing the chloroplast genomes of *Nandina domestica* and *Platanus occidentalis* using one run per species on the 454 Life Sciences GS 20 (25 Mbp capacity per run). They reported that the genome assemblies accounted for ~99.7% of the predicted genome length, and that coverage depth averaged from 24.6× (*Nandina*) to 17.3× (*Platanus*).

While the size and complexity of the chloroplast genome is well suited to short read pyrosequencing platforms, microread technologies produce an overwhelming excess of data that are less well suited for small genomes; for example, a single 2 Gbp run could conceivably sequence an average-sized chloroplast genome to a depth of 12 500×. Multiplex tagging methods have the potential to spread the capacity of high-capacity sequencers across many genomes, and strike a better balance between coverage, throughput and cost. At present, multiplex sequencing methods have been developed for pyrosequencing applications using indexed adapters (24–26) or indexed amplification primers (27,28). In most cases, however, the complexity of the multiplex input pool was far less than that of a single chloroplast genome. For example, six human mitochondria total to ~99 kb of sequence complexity (25), while 64 mitochondrial rDNA fragment amplicons from 13 species only contained ~6 kb of sequence complexity (27). Despite the many advantages of microread sequencing platforms, index adapters have yet to be reported for the Illumina 1G Genome Analyzer. For relatively gene dense chloroplast genomes, a challenge could yet lie in the presence of many mononucleotide repeats that are comparable in size to the read length. As such, it remains unclear how efficiently chloroplast genomes could be assembled with microread lengths and empirical error rates.

In this report, we describe two experiments where index adapters developed for the Illumina 1G Genome Analyzer (29) were used to sequence eight tagged chloroplast genomes in two four-plex sequencing reactions, representing a combined total of ~960 kb of unique sequence. In this study, we address four key questions: (i) Can multiplexed microreads be assembled into contigs representing most of the chloroplast genome using *de novo* and reference guided methods? (ii) How efficiently do microreads assemble in regions of low complexity? (iii) What is the error rate in assemblies constructed by these methods? and (iv) What are reasonable upper limits for multiplex sequencing of chloroplast genomes?

MATERIALS AND METHODS

DNA source and template amplification

DNA from *Picea sitchensis* and seven pine species was used in this study, six from Subgenus *Strobus* (*P. gerardiana, P. krempfii, P. lambertiana*, *P. longaeva, P. monophylla, P. nelsonii*) and one from Subgenus *Pinus* (*Pinus contorta*) (Table 1). Total genomic DNA was isolated using the methods described in Willyard *et al.* (30). For this test, we amplified the entire chloroplast genome using PCR and 35 primer pairs derived from the consensus chloroplast genome sequences of *P. thunbergii* and *P. koraiensis* (see Supplementary Table 1 for primer sequences). Amplicons for these regions averaged ~3.6 kb, and best results were obtained with Phusion polymerase (New England Biolabs, Ipswich, MA, USA). Once chloroplast genomes were amplified in their entirety, amplicon DNA concentrations were determined by either A_{260} (Nanodrop 1000; ThermoFisher Scientific, Wilmington, DE, USA) or visual approximation using gel electrophoresis, and amplicons were pooled in roughly equal mass mixtures (~10–30 ng per amplicon, 500–1000 ng total).

Preparation of fragment libraries and Illumina sequencing

Pooled amplified chloroplast DNAs were sheared, polished and prepared using minor modifications of the original Illumina Sample Preparation kit (29). Briefly, 0.5–1 μg of pooled chloroplast DNA was sheared by nebulization using 32 psi N_2 for 8 min, and the sheared fragments were purified and concentrated using a QIAquick PCR purification spin column (QIAgen Inc., Valencia, CA, USA). Sheared fragments were treated with T4 DNA polymerase, T4 phosphonucleotide kinase and the Klenow fragment of *Escherichia coli* DNA polymerase to fill 5′ overhangs and remove 3′ overhangs. Terminal (3′) A-residues were added following a brief incubation with dATP and Klenow 3′–5′ exo−. Next, custom-made adapters containing unique 3 bp tags were ligated to the fragments in place of the standard Illumina adapters. The 3 bp tags include 'CCT', 'GGT', 'AAT', 'CGT' and 'ATT' (Supplementary Table 2), and are compatible with the Illumina 1G Genome Analyzer flow cell. Adapter-ligated DNAs in the range of 170–250 bp were size selected using agarose electrophoresis, and products were isolated using the QIAgen MiniElute gel extraction spin columns. These small insert libraries were then amplified independently using 18-cycle PCR amplification and standard Illumina primers. Amplified libraries were again size selected using agarose electrophoresis to remove amplicons below 170 bp; these apparently represent adapter dimers and other artifacts. After spin column extraction and quantitation, libraries were mixed (multiplexed) at equimolar ratios to yield a total oligonucleotide mix of 10 nM.

Table 1. *Pinus* samples used in this study. Voucher specimens are deposited in the Oregon State University Herbarium

Species	Taxonomy	Source	Reference	Genbank
P. contorta	Subgenus *Pinus*, Section *Trifoliae*	Newport, Lincoln Co., OR, USA. Accession CONT40 (A. Liston 1315)	*P. ponderosa* (Liston/Cronn, draft unpublished)	EU998740
P. gerardiana	Subgenus *Strobus*, Section *Quinquefoliae*	Nanga Parbat Region, Gilgit, Pakistan. 35.400°N, 74.591°E. Accession GERA04 (R. Businský 41123)	*P. koraiensis* NC_004677	EU998741
P. lambertiana	Subgenus *Strobus*, Section *Quinquefoliae*	NE Montague, Siskiyou Co., CA, USA. 41.850°N, 122.313°W. Accession LAMB08 (USFS Region 5 Seed Orchard)	*P. koraiensis* NC_004677	EU998743
P. longaeva	Subgenus *Strobus*, Section *Parrya*	White Mountains, Inyo County, CA, USA. 37.612°N, 118.241°W. Accession LONG01 (Kazmierski s.n.)	*P. koraiensis* NC_004677	EU998744
P. monophylla	Subgenus *Strobus*, Section *Parrya*	Near Eureka, UT, USA. 39.941°N, 112.146°W. Accession MONO11 (D. Gernandt 479)	*P. koraiensis* NC_004677	EU998745
P. nelsonii	Subgenus *Strobus*, Section *Parrya*	Near San Antonio Peña Nevada, Nuevo León, Mexico. 23.767°N, 99.900°W. Accession NELS03 (D. Gernandt 10198-15098)	*P. koraiensis* NC_004677	EU998746
P. krempfii	Subgenus *Strobus*, Section *Quinquefoliae*	Bi Doup Mountain, Lam Dong, Vietnam. 12.0°N, 108.68°E. Accession KREM03 (Royal Botanic Garden First Darwin Expedition 242)	*P. koraiensis* NC_004677	EU998742
P. sitchensis		Newport, Lincoln Co., OR, USA. Accession PICSIT04 (Liston 1314)	*P. thunbergii* NC_001631	EU998739

Aliquots of multiplex libraries (5 pmol) were denatured and then processed with the Illumina Cluster Generation Station at the Center for Genome Research and Biocomputing at Oregon State University (Corvallis, OR, USA), following manufacturer's recommendations. The Illumina 1G Genome Analyzer was programmed to run for 36 cycles, which produces a theoretical fixed read length of 36 bp. Images were collected over 300 tiles, each of which contained an average of 16 994 clusters in multiplex S1 and 22 124 clusters in multiplex S6.

Data filtering and analysis pipeline

After the run, image analysis, base calling and error estimation were performed using Illumina/Solexa Pipeline (version 0.2.2.6). Perl scripts were used to sort and bin all sequences using the three (5′) nucleotide tags; these tags were removed prior to evaluation with Reference Guided Assembler (*RGA*; R. Shen and T. Mockler, in preparation) or *de novo* assembly. Examination of Illumina *Q*-values revealed a decrease after cycle 33 (data not shown), thus the three 3′ bases were trimmed, and 30-mers were used in all subsequent analyses (31). Binned 30-mers were evaluated relative to the appropriate *Pinus* reference (*P. thunbergii*, NC_001631; *P. koraiensis*, NC_004677) using the program *RGA* in order to estimate the genome coverage.

To assemble chloroplast genomes using Illumina/Solexa microreads, we used a three-step process. First, *de novo* assemblies were attempted using Velvet Assembler 0.4 (32) using a hash length of 19, minimum average coverage of 5×, and minimum contig length of 100 bp. Second, contigs were aligned to a reference genome sequence using CodonCode version 2.0.4 (CodonCode Corporation, Dedham, MA, USA; http://www.codoncode.com/) and standard settings for global alignments. *Picea sitchensis* was aligned to the previously published chloroplast genome of *P. thunbergii* (NC001631) and the species of *Pinus* subgenus *Strobus* were aligned to *P. koraiensis* (NC004677). The assembly of *P. contorta* used a draft plastome of *P. ponderosa* as its reference (A. Liston and R. Cronn, unpublished results). Prior to alignment, an 'N' was added to the ends of each contig, in order to differentiate assembly gaps (dashes flanked by the added 'N's) from deletions (dashes) relative to the reference. Contigs that failed to align to the reference genome were scanned for chloroplast sequence homology using BLASTN (http://www.ncbi.nlm.nih.gov/). Successful matches typically contained >100 bp insertions relative to the reference genome; these contigs were manually inserted into the alignment. Between 67% and 98% of the contigs aligned to the reference genome. Unaligned contigs apparently represent nontarget PCR amplicons (data not shown). The final *de novo* assemblies covered 78.1–94.6% of the reference genome (excluding deletions and including insertions relative to the reference). Third, gaps between the *de novo* contigs were replaced with the reference sequence, and this chimeric assembly was used as a 'pseudo-reference' for reference-guided assembly with the program *RGA*. *RGA* aligns microreads to their best match in a reference sequence, and then creates a guided consensus sequence from the aligned overlapping reads. *RGA* outputs the resulting contigs, singletons, the real coverage of each base in the assembly, and identifies SNPs based on microread density in the assembled sequence compared

to the reference and *Q*-values at specific position on each microread. *RGA* settings used were ≤2 mismatches per microread, *Q*-values ≥20, read depth ≥3 and SNP acceptance requiring ≥70% of reads in agreement. The pseudo-reference created from *de novo* assemblies and the reference sequences were corrected using *RGA*.

Final sequences were annotated using standard settings in the program DOGMA [(33), http://dogma.ccbb.utexas.edu/]. Multiple alignments were made using MAFFT v. 5 (34), and full alignments with annotations were visualized using the VISTA viewer (34,35). See Supplementary Figure 1 for full annotation summaries. In addition, nucleotide positions corresponding to primer locations were changed to 'N', as the use of complementary forward and reverse primers at a single site precluded us from obtaining genomic sequence for these positions.

Sequence and gap validation

Given the absence of identical reference chloroplast genomes for these accessions, we have no direct way to determine exactly how complete our assemblies are, and whether repeated motifs inhibit microread assemblies. To provide an estimate of the completeness of these drafts, we calculated the proportion of exon nucleotides sequenced relative to the annotated *P. thunbergii* reference genome. To examine completeness through repeated regions, we determined the frequency spectrum of simple sequence repeats in reference genomes and contigs from the sequenced species. Searches for perfect repeats ≥6 units length were conducted using the Java program *Phobos* (36). Mononucleotide repeats were the most abundant class of repeat in the reference genomes, and they exceeded di- through tetranucleotide repeats both in repeat number and overall length (data not shown). For this reason, we restricted our screening to mononucleotide repeats. Frequencies for repeat classes ranging from 6 bp to 23 bp were determined for the original contigs (f_{total}), and for 'terminal repeat depleted contigs', where 5′ and 3′ terminal repeats were removed so that only interstitial repeats were present ($f_{interstitial}$). We calculated the proportion of repeats that interrupt contigs for all repeat length classes using the equation $(1-(f_{interstitial}/f_{total}))$.

Sequence accuracy relative to conventional Sanger sequencing was estimated by amplifying and sequencing six chloroplast regions (*cemA, psbC-psbD, rpoA, rpoB, rpoC1, rps4*) from the eight species included in this study. Amplification primers and conditions used to amplify these regions are included in Supplementary Table 1. Following PCR, products were sequenced by the High-Throughput Genomics Unit at the University of Washington (http://www.htseq.org/index.html). Sanger sequences and microread assemblies were aligned and differences were summarized using MEGA 4.0 (37).

We explored the potential limits of chloroplast genome multiplex sequencing by drawing random microread samples in triplicate from our *Pinus gerardiana* data set that correspond to a range of multiplex levels; 6× (905 534 reads), 8× (668 193 reads), 12× (445 460 reads) and 16× (334 096 reads). For this exercise, samples were only assembled *de novo* using Velvet with parameters described ealier. Assembled contigs were subsequently mapped onto our *P. koraiensis* reference to determine the number of aligned contigs, the summed length of all contigs, and the number of remaining gaps in the final assembly.

RESULTS

Analysis of multiplex tag frequencies

Sequencing of two multiplexed pools produced 11 445 101 36-bp reads (Table 2). Within each multiplex, correctly identified multiplex tags accounted for 90.1% (multiplex S1) and 90.5% (multiplex S6) of the total reads, with 634 951 and 482 082 of the reads being unassignable due to the absence of an expected tag sequence. Exclusion of unassigned tagged sequences and trimming of the 3-bp tags left 10 328 068 reads 33 bp in length, for a total sequence pool of 341 Mb.

Examination of tags from the ~1.12 million unassigned tagged sequences revealed three important findings. First, the pools of tagged microreads included a small number of

Table 2. Summaries of total tagged and aligned reads from two multiplex experiments on the Illumina/Solexa 1G Genome Sequencer

Experiment	MPLX S1				MPLX S6			
Total reads	6 391 206				5 053 895			
Adapters	167 038				147 632			
Net	6 224 168				4 906 263			
Tag	CCT	GGT	AAT	ATT	CCT	GGT	AAT	CGT
Taxon	CONT	GERA	KREM	LAMB	LONG	MONO	NELS	PICSIT
Total reads	1 423 449	1 336 385	1 552 811	1 420 032	930 019	1 232 647	1 111 158	1 263 800
Aligned reads	1 082 697	1 023 041	1 204 585	1 090 216	756 726	995 128	852 081	1 001 719
Aligned (%)	76.1	76.6	77.6	76.8	81.4	80.7	76.7	79.3
Mean coverage	59	138	149	72	117	186	75	55
Number of contigs (RGA)	57	9	24	68	25	39	104	183
Mean length	2066	13 017	4852	1697	4665	2959	1098	626
SD	4196	12 213	11 768	5709	5604	4994	3908	1615
Median contig length	136	9454	349	86	2586	409	76	82
N50	8012	26 178	10 580	10 437	9460	10 401	7135	4092
Sum contig lengths	117 784	117 153	116 448	115 444	117 189	116 456	114 246	114 679
Exon gaps	6467	4272	4861	6621	4918	7924	6200	8346
Exon complete (%)	91.0	94.0	93.2	90.7	93.1	88.9	91.3	88.3

sequences that correspond to adapters. These were found in 23 578 (S1) and 34 189 (S6) of the reads, accounting for 0.6% of the total reads from both experiments. Second, failure to correctly assign sequences due to an 'N' at one of three tag positions accounted for 92 505 (S1) and 88 582 (S6) reads, or 1.6% of the total reads from both experiments. This comparatively small value suggests that sequence quality plays a small but measurable role in the misassignment of reads. The largest pool of unassigned reads by far appears to arise from nucleotide misincorporation or errors in the sequencing process; these accounted for 619 695 reads in multiplex experiment S1, and 469 698 reads in multiplex experiment S6, or 9.8% of the total reads from both experiments. This class of errors appear nonrandom in nature, as the number of erroneously tagged reads in each of 60 potential mismatch sequence tags (potential mismatched tags = $y^n - z$, where y = number of potential nucleotide states [G, A, T, C], n is the number of tagged nucleotides and z is the number of bins corresponding to correctly tagged sequences) were highly correlated between multiplex runs S1 and S6 ($\log_{10}$[S1] = 0.05242 + 1.0138 * $\log_{10}$[S6]; r^2 = 0.980).

To evaluate the base frequencies produced by this apparently nonrandom process, we tallied the erroneously tagged reads that differed by one mutational step from CCT and GGT tagged reads (e.g. CCT → ACT, CAT, CCA, etc.) and computed their observed and expected frequencies, assuming an equal probability of base misincorporation (excluding 'N's). As shown in Figure 1, one-step mutations from CCT and GGT tags in these experiments showed a remarkably consistent trend where the distribution of errors was nonuniform across bases and positions. Error was significantly higher in

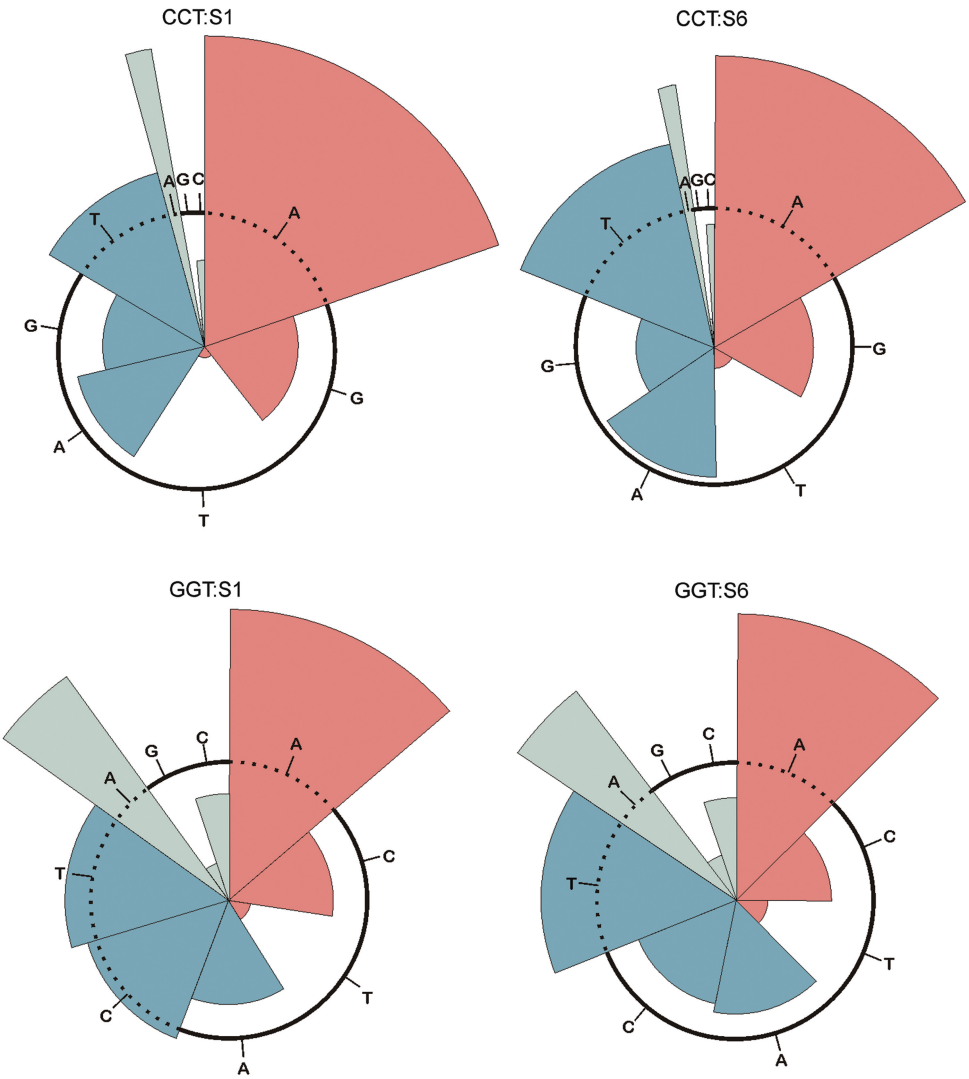

Figure 1. Relative frequencies of barcode error by barcode tag (CCT, GGT), experiment (S1, S6) and nucleotide position (1, 2, 3). Observed frequencies of erroneous, nontag nucleotides are indicated by position 1 (salmon), 2 (blue) and 3 (green); first and second position errors were far more common than third position errors. Slices within a position are scaled proportionately to the number of base calls for that nucleotide; if errors were present at equal frequencies within a base position, each slice would be of equal size and would not extend beyond the perimeter of the circle. In all experiments, errors involving substitutions to 'A' were more frequent than expected for position 1 and 3, where errors involving substitutions to 'T' were more frequent than expected for position 2.

positions 1 and 2 than position 3 ($G_{adj} = 53\,002$, $G_{crit} = 5.99$, $P = 0.000$; results not shown). In addition, errors for these index tag positions showed an excess of 'A' substitutions at positions 1 and 3, and an excess of 'T' substitutions at position 2. A mutational bias towards incorporating 'A' has been previously described for the Illumina/Solexa system (31), and it may be an important source of error in multiplex tagging of sequences for this platform.

Contig assembly of multiplexed sequences using *de novo* and reference guided approaches

Analysis of coverage using *RGA* software showed that the total number of reads differing by fewer than 3 bp from a reference ranged from 852 081 (in *P. nelsonii*) to 1 204 585 (*P. krempfii*) (Table 2). There was a significant association between the total number of reads per tagged pool and the number of *RGA* aligned reads ($READ_{RGA} = READ_{TOTAL} * 0.70767 + 92273$; $r^2 = 0.977$). A significant fit, however, was not observed for the number of total reads or *RGA* aligned reads to either the number of contigs or the sequencing depth ($r^2 \leq 0.02$ in all cases; $Prob[>F] \geq 0.71$).

The number of contigs produced by *RGA* for these chloroplast genomes ranged from 9 (*Pinus gerardiana*) to 183 (*P. sitchensis*), and they showed mean lengths ranging from 13 017 bp to 626 bp, respectively (Table 2). Sequencing depth varied substantially, not only between genomes but also within genomes. The variation in sequencing depth appears to be attributable to inaccuracies in PCR amplicon pooling, as the variation tracks the individual amplified regions (Figure 2). Despite these inaccuracies, the number of contigs produced showed a negative correlation with sequencing depth ($r = -0.678$; $P = 0.065$), and the summed length of contigs for all genomes showed a narrow range of lengths from 114.3 kb to 117.8 kb. These values are very similar to the known genome sizes for *P. thunbergii* (119 797 bp) and *P. koraiensis* (117 190 bp). Multiple sequence alignment of these eight species to *P. thunbergii* produced draft genomes that appear nearly complete with regard to gene content, as 129 of 132 genes could be accounted for in all species (see Supplementary Figure 1 for a detailed image of assemblies and gene locations).

Our assemblies showed 16 amplified regions with very low coverage (median of ≤ 5 microreads per position across the amplicon). Thirteen regions represent artifactual 'gaps', where we added insufficient PCR product to the sequence pool: these include regions from *P. contorta* (primers 1F/R and 13F/R), *P. krempfii* (primers 32F/R), *P. lambertiana* (primers 23F/R and 27F/R), *P. longaeva* (primers 13F/R), *P. nelsonii* (primers 4F/R, 28F/R and 32F/R) and *P. sitchensis* (primers 2F/R, 4F/R, 9F/R and 35F/R). Three gapped regions, however, represent lineage-specific gene deletions for *ycf12b* (78 bp at position 51 051), *psaM2* (93 bp at position 51 442) and *ndhI* (371 bp at position 101 988). These are shared deletions in the genomes of subgenus *Strobus* species (i.e. *P. gerardiana, P. krempfii, P, lambertiana, P. longaeva, P. monophylla, P. nelsonii* and the *P. koraiensis* reference), and the

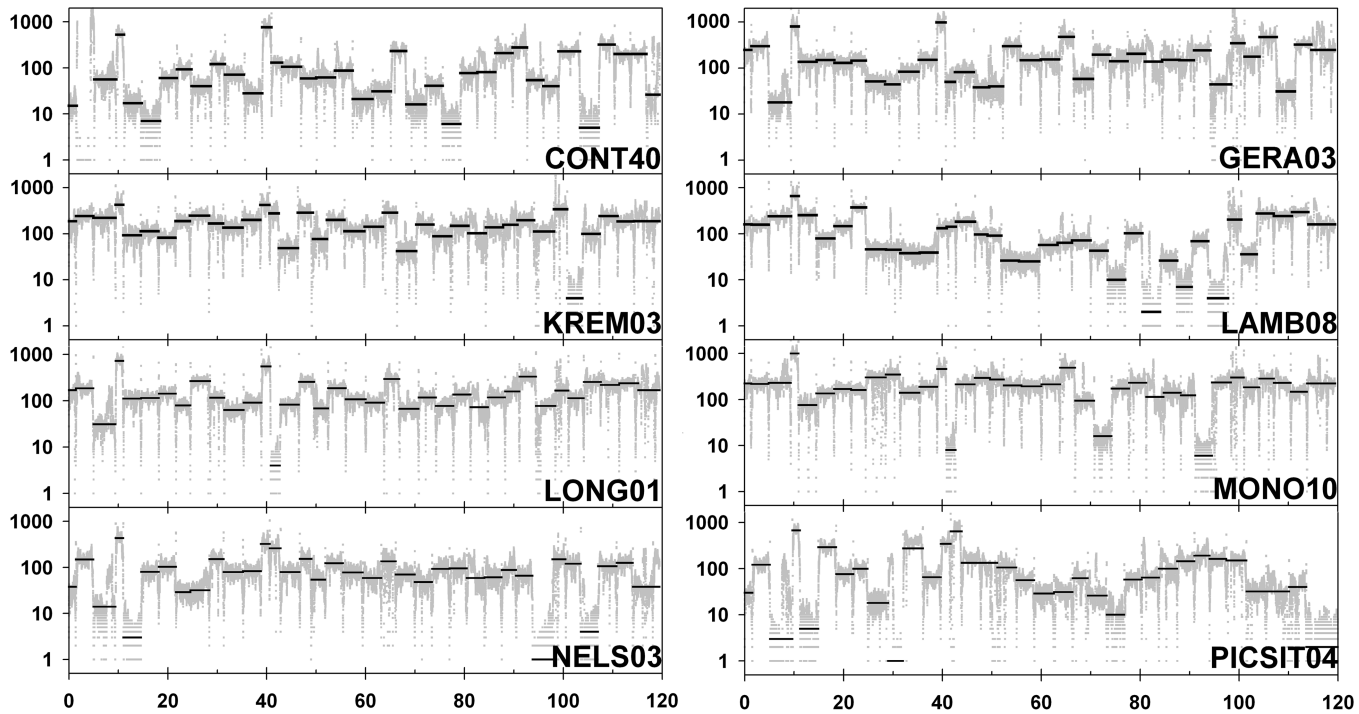

Figure 2. Plots showing sequencing depth by position for eight chloroplast genomes sequenced by multiplex sequencing-by-synthesis. Microreads per position (*y*-axis) are plotted in gray relative to the position in the assembly (*x*-axis, in kb). The median number of reads across each PCR amplicon is indicated by black lines.

corresponding genes are present in the subgenus *Pinus* reference (*P. thunbergii*), *P. contorta*, and *P. sitchensis*.

Degree of sequence completion

Since we lack exact reference chloroplast genomes for our accessions, we estimated the degree of completion by comparing the number of putative exon encoding nucleotides from each draft genome to the predicted gene exons based on the reference genome (*P. thunbergii*) annotation, which includes 4 rRNAs (4518 bp), 41 tRNAs (2653 bp) and 87 protein encoding genes (60 771 bp). Based on exon predictions, our assemblies averaged 90.1% complete, ranging from a low of 88.3% for *P. sitchensis* (8346 missing exon nucleotides), to a high of 94.0% for *Pinus gerardiana* (4272 missing exon nucleotides) (Table 2). To the extent that gaps in exons arise primarily from heterogeneity in read density, these values should be accurate predictors of the proportion of the genome that remains undersampled for assembly.

In contrast to exons, regions of low complexity cannot be easily assembled using microreads, particularly if repeated nucleotide motifs approach the microread length. To determine whether mononucleotide repeats limit the comprehensiveness of our assemblies, we screened individual contigs for mononucleotide repeats ≥6 bp in length. In our contigs from all eight genomes, the largest observed repeats were 18 bp (found in *P. contorta, P. krempfii, P. lambertiana, P. longaeva, P. nelsonii*, and *P. sitchensis*), which is comparable to the largest mononucleotide repeats from *P. thunbergii* (17 bp), but smaller than the largest repeat from *P. koraiensis* (23 bp). We computed the 95% confidence interval for counts of assembled repeats for these eight assembled sequences; with the exception of the 6 bp repeat class, these confidence intervals include one or both of the counts of repeats for the two reference genomes (Figure 3).

Close examination revealed that the longest repeats were located on the 5′ and 3′ termini of contigs, suggesting that contig assembly was frequently interrupted by mononucleotide repeats. To determine if repeat length influences the success of contig assembly, we estimated two parameters: (i) the frequency spectrum of repeats ≥6 bp from all contigs, irrespective of location on the contig; and (ii) the frequency spectrum of repeats ≥6 bp, which were fully assembled within the contig (= *interstitial* repeats). With these values, we calculated the proportion of repeats that interrupt contigs across all length classes (described in Materials and methods section). As shown in Figure 3, there is a significant linear relationship between the length of a mononucleotide repeat and the proportion of contigs interrupted by a repeat (proportion terminated = $-0.5980 + 0.09769$ * repeat length; $r^2 = 0.714$; Prob[>F] = 0.0001). Only 2% of repeats 6 bp in length interrupt a contig, while the remainder are assembled within contigs. However, each base pair increase in repeat length results in a ~10% increase in the frequency of repeats that interrupt a contig. Based on our data, mononucleotide repeats >12 bp in length are more likely to terminate an assembly than to be fully assembled, and by 16 bp all repeats are predicted to interrupt contigs and be assembled on 5′ or 3′ termini. In light of the low abundance of simple repeats in pine chloroplast genomes (mononucleotide repeats ≥6 bp total 2120 bp, or 1.8% of the *P. thunbergii* genome), these seem unlikely to present a major obstacle for obtaining nearly complete genome sequences based entirely upon microreads, particularly if microreads were extended to 50 bp or if paired reads were added (both are available for the Illumina 1G but were unavailable at the time of these experiments).

In addition to assembly problems associated with simple repeats, we observed two other methodological artifacts that presented obstacles for complete assembly. First, *de novo* assembly consistently produced a small proportion of contigs with high divergence in the terminal 10 bp of one or both contig ends. To correct this assembly error, the terminal 10 bp of all *de novo* contigs was adjusted to match the reference sequence prior to final assembly in *RGA*. Second, we noted a drop in read density within the 30 bp flanking the primer locations (Figure 2). Associated with this low coverage, ~4% of the regions within 30 bp of a primer showed an elevated divergence relative to the reference, as compared to the rest of the contig assembly. For flanking regions showing >10% divergence, we changed all mismatch nucleotides to 'N' to reflect the likelihood that these mismatches probably arise from methodology rather than true sequence divergence. In combination with masking of primer sites, these adjustments resulted in a <1% reduction in total assembly length.

Assembly accuracy

Conventional Sanger sequencing at six different chloroplast regions (*cemA, psbC-psbD, rpoA, rpoB, rpoC1, rps4*)

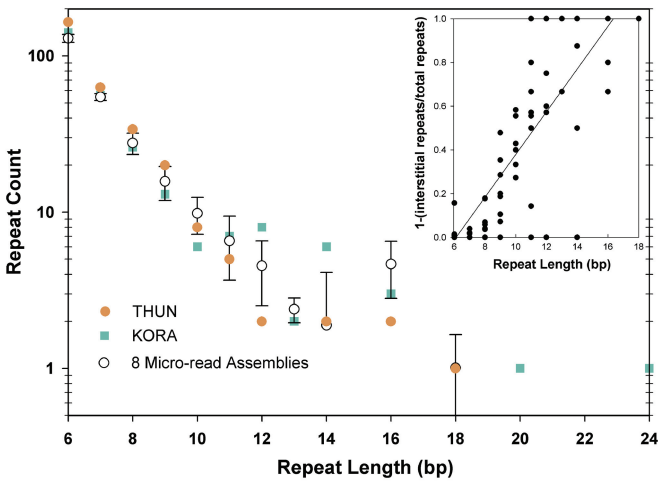

Figure 3. Frequency spectrum of mononucleotide repeats observed in reference and microread assemblies of *Pinus* chloroplast genomes. The number of repeats per length class (6–24 bp) is plotted for *P. thunbergii* (THUN; salmon) and *P. koraiensis* (KORA; blue). The average and 95% confidence interval for eight microread assemblies (seven *Pinus*, one *Picea*, white circles) are also shown. Inset: relationship between the proportions of repeats terminating contigs and the length of each repeat class for the eight microread assemblies. The least squares regression line is indicated.

in the eight species yielded 37 290 nt that could be directly compared to microread assemblies for accuracy (Table 3). In this comparison, we observed 21 sites that showed evidence of sequencing or assembly error, yielding a cumulative error of 0.056%. Fisher's exact test showed that sequencing error was nonuniform across taxa ($\chi^2 = 6.81$; 7 df; $P = 0.449$), but nearly uniform across genes ($\chi^2 = 10.37$; 5 df; $P = 0.065$). Errors in *rps4* and *cemA* accounted for 15 of the 21 observed errors. At present, we do not know the exact source of these errors but possible sources include sequencing error, assembly error and amplification of duplicated paralogous loci; we are examining this in greater detail. This error rate is nearly identical to the rate reported by Moore *et al.* (21) for pyrosequencing of chloroplast genomes (0.037%).

Limits to chloroplast multiplex sequencing

In designing these experiments, we chose four-genome multiplexing primarily for theoretically high sequencing depth and the ease of library assembly; nevertheless, the question remains as to whether higher level multiplexing can be attained without sacrificing assembly length. To address this question, we drew random samples of *Pinus gerardiana* microreads in a manner that simulated higher multiplex levels of 6× (895 606 reads), 8× (668 363 reads), 12× (441 119 reads) and 16× (334 181 reads). As expected, sequencing depth decreased from 88 in the original 4-plex sample to 20 in the simulated 16-plex samples (Figure 4). While this reduction in coverage had no significant impact on the number of contigs recovered from these treatments (Figure 4; $r^2 = 0.1567$, Prob[>F] = 0.181), contigs were substantially shorter, which resulted in significant decreases in the summed length of aligned contigs (Figure 4; $r^2 = 0.7597$, Prob[>F] = 0.0001). Based on this randomization, reductions in summed contig lengths in the 4-, 6- or 8-plex assemblies were modest, differing by <5%; in contrast, reductions in 12-plex (10.4%) and 16-plex (19.2%) assemblies were significant. In light of these results, we predict that multiplexed pools containing between four and eight chloroplast genomes are likely to return satisfactory draft assemblies.

DISCUSSION

This report provides the first demonstration of the feasibility for assembling microreads from multiplexed pools of chloroplast genomes into nearly complete draft sequences. For the eight genomes we sequenced, we predict that 88.3–94.0% of each genome has been sequenced, with assemblies ranging in number from 9 to 183 contigs, and with mean contig sizes ranging from 13 017 bp to 626 bp. These genomes were sequenced from pools containing four different templates, each of which contained ~500 000 bp of sequence complexity. In light of the challenges associated with established methods of plastome sequencing (17,21,38), microread sequencing of

Table 3. Error estimates for six genic regions across eight species

Locus	Length	CONT 40	GERA 03	KREM 03	LAMB 08	LONG 01	MONO 10	NELS 03	PICSIT 04	Row Total
cemA	641	0	0	0	0	1	0	7	0	8
psbCD	2182	0	0	0	2	0	0	0	N/A[a]	2
rpoA	454	0	0	0	0	0	4	0	0	4
rpoB	594	0	0	0	0	0	0	0	0	0
rpoC1	717	0	0	0	0	0	0	0	0	0
rps4	346	1	0	2	0	0	0	0	4	7
Total	4934	1	0	2	2	1	4	7	4	21
Error (%)		0.020	0	0.040	0.040	0.020	0.081	0.142	0.145	0.056

Total sequence differences between Illumina/Solexa-derived assemblies and traditional Sanger sequencing are shown.
[a]The *psbCD* region was not amplified from *Picea sitchensis*.

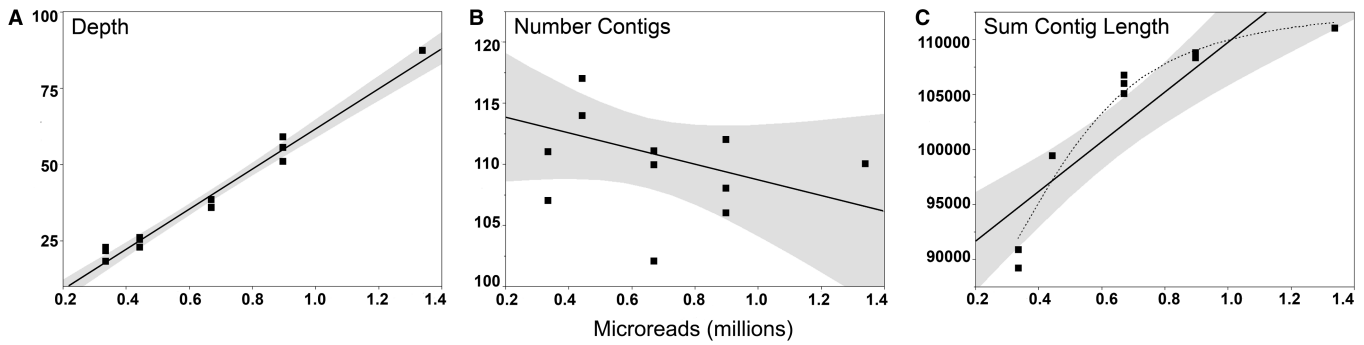

Figure 4. Simulations of higher level multiplex levels. Random subsets of microreads from the *P. gerardiana* data set were sampled to simulate multiplex levels ranging from 4× (1.37 million microreads) to 16× (0.34 million microreads). Triplicate random subsets were assembled with Velvet *de novo* assembly, and assemblies were evaluated for sequencing depth (**A**), the number of contigs (**B**) and the summed contig lengths (**C**). Solid lines show the best fit line from least squares regression and shaded regions show the 95% confidence interval of the best fit line. The curved line (C) shows the best fit with a smoothing spline ($\lambda = 5 \times 10^{15}$; $r^2 = 0.973$).

multiplexed genomes is an efficient and economical approach for creating draft genomes, and this approach will be useful at higher multiplex levels.

The assembly strategy we used—*de novo* assembly of microreads into contigs, mapping contigs on a reference to create a 'pseudoreference', and using reference-guided assembly to correct the pseudoreference—has advantages over individual methods alone. *De novo* assembly proceeds independent of divergence from a reference; for this reason, comparison of contigs assembled *de novo* relative to the reference can be used to identify putative insertions, deletions and rearrangements that would otherwise be incorrectly assembled in a reference guided assembly. In contrast, *RGA* aligns microreads to their best match in a reference sequence based on user-defined limits for read depth, quality and SNP acceptance. A final pass with *RGA* extends assemblies into regions of low coverage and corrects misassembly errors created during *de novo* assembly. Furthermore, *RGA* allows the incorporation of information from base quality scores, a feature that is currently lacking in *de novo* assembly programs.

While these methods assemble nearly all of the chloroplast genome, they do not assemble microreads into a single, full-length genome sequence. Defining the properties of contig-interrupting regions will be an important next step in constructing full genome assemblies from microreads. In our samples, long mononucleotide repeats are one important factor that contributes to contig interruption (Figure 3). In addition, our amplification strategy reduced the likelihood of a complete assembly, since the PCR products were amplified using complementary primer sites (e.g. sequence of a reverse primer was the complement of the next forward primer); ideally, these should be staggered to include 30 or more overlapping bases. Other sequence motifs may also play a role in contig interruption, but these cannot be identified conclusively without an exact reference. We are currently evaluating the properties of sequences associated with contig breaks by comparing microread assemblies of a chloroplast genome with a Sanger-derived assembly from the same genome. It is important to note that the use of longer single reads or paired-end reads may resolve many of the problems responsible for contig breaks.

Based on our experiments, three factors bear upon the successful application of multiplex microread sequencing for assembling complete chloroplast genomes; (i) the availability of a high quality reference genome; (ii) minimizing nontarget DNA in the library; and (iii) an efficient strategy for isolating chloroplast genomes.

Criterion 1: reference genomes

To assemble and evaluate the completeness and accuracy of our pine chloroplast genomes, we relied on closely related reference genomes to order contigs and to fill contig gaps. This is possible because the average nucleotide distances between our samples and their respective references are <0.8% for hard pines (relative to *P. thunbergii*) and 1.4% for soft pines (relative to *P. koraiensis*). Indeed, the higher divergence between *P. sitchensis* and these references (6.3% and 5.6%, respectively) is a major factor contributing to the higher number of contigs, shorter average contig size and shorter genome assembly (Table 2). Equally as important, the predicted gene content and gene order appears conserved between the references and our samples; evidence for this is provided by PCR success across 35 amplicons (rearrangements between primers would lead to PCR failure, which was not observed), and the absence of rearrangements in comparisons between contigs assembled using *de novo* assembly relative to the references. If extensive rearrangements had been present, this would have complicated our assemblies.

In the absence of a reference genome, we would have faced two options: rely entirely on *de novo* assembly, or assembly of coding sequences only. Initial experiments using *de novo* assembly methods alone retrieved contigs that were generally shorter and more numerous than those produced by a combination of *de novo* and reference guided approaches together (data not shown). These contigs would account for 78.1–94.6% of the assembled sequence length, but ordering contigs would prove problematic. The alternative, assembling the 'gene space' of the chloroplast genome, would be possible by referencing a composite sequence made up of the conserved genes (protein coding, mRNA, tRNA) from a more divergent species. This strategy would fall short for assembling the noncoding portion of the genome, and variation in genome organization would also go undetected. Nevertheless, this strategy would allow the 67 942 bp encoding 132 putative genes in the pine chloroplast genome to be compared with more distantly related gymnosperms.

For population and evolutionary genetic applications, microread assemblies covering ≥88% of the genomic coding sequence (>59 kb) would provide an exceptionally high level of haplotypic discrimination and genealogical resolution. Simple sequence repeats would need to be excluded from analyses, but chloroplast microsatellites are known to display unusual mutation patterns that limit the accuracy of estimates of differentiation (Fst), genealogies and coalescent inferences (39). Microread assemblies may contain a large number of gaps; these can either be considered as missing data in the analysis or else resolved using imputation approaches (40).

Criterion 2: minimizing nontarget DNA

Excluding nontarget DNAs during sample preparation is exceedingly important for higher level multiplexing. As shown in Table 2, many of the reads in our multiplex pools were not chloroplast in origin; these reads include 'unidentified' nonchloroplast DNA and adapters from library construction. Up to 24% of the microreads in our samples sample could not be mapped to the reference genome or the adapter sequences (Table 2), and are likely attributable to nonspecific amplification during PCR. Nonspecific amplification products can be minimized by modifying PCR conditions, but complete elimination may not be possible, particularly in the case of paralogous sequences that have integrated into the nuclear genome. In addition, adapter dimers isolated during the library construction stage added to the number of nonchloroplast

sequences retrieved. In these experiments, the contribution of adapters was very low (<1%); however, in a preliminary multiplex experiment, adapter dimers accounted for 31% of the total microreads (data not shown). We found that selecting slightly larger libraries (170–300 bp, as opposed to the 120–170 bp range recommended by Illumina) reliably eliminates adapter contamination and has no observable impact on the resulting sequences.

Criterion 3: an efficient enrichment strategy

Our approach to isolating chloroplast genomes was to use multiple long PCR reactions; this provides a highly enriched template, but it is inefficient for processing a large number of samples. Isolating chloroplast DNA from organelles is one route for template isolation, but these preparations take substantial effort and often contain a large amount of contaminating nontarget DNA (17). Recent advances in genome-wide selective hybridization and amplification (41,42) may offer the most efficient and high-throughput means to isolate whole chloroplast genomes, and they have the added benefit of being potentially useful across a broad spectrum of species.

Applications beyond chloroplast genomes

The vast improvements being made in DNA sequencing technologies offer new and seemingly limitless applications for studies in population and evolutionary genetics. Nevertheless, the acquisition of sub-genome scale data from many templates is a challenge with these technologies, since they are designed to obtain high sequence depth from a comparatively small number of samples. Multiplex sequencing approaches, such as those outlined here and elsewhere, make maximal use of the substantial capacity offered by microread sequencers and offer a means to produce draft assemblies of multiple chloroplast-sized genomes in a high-throughput, cost-effective manner. This method also has immediate application for multiplex sequencing of similar sized clones (BACs, fosmids) or nucleic acid pools of limited sequence complexity (siRNA, mRNA and ChIP sequence pools; polymerase fidelity studies; environmental samples of target genes).

An important issue remaining to be resolved is the upper limit of multiplexing and how this relates to genome size and complexity. We explored the impact that reduced reads had on the assembly process, and in our samples the impact appears to be severe beyond 8-plex; for this reason, we recommend that chloroplast genomes be multiplexed at levels under 8-plex. As noted earlier, a key factor inherent to this recommendation is the amount of nontarget DNA in the template pool. Even with this caveat, multiplex sequencing-by-synthesis is an efficient approach for obtaining many small genomes in a cost-efficient manner. If we assume other organellar sequences can be isolated in comparable purity levels, multiplexing approaches like those described in this article could be used to sequence approximately 2.5 *Arabidopsis* mitochondrial genomes (~370 kb) or 60 animal mitochondrial genomes (~16 kb) in one channel of a flow cell. For templates that can be obtained in higher purity (e.g. fosmid or BAC clones; viral genomes), we expect that the level of multiplexing could be substantially higher than what we attempted here.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mariah Parker-deFeniks and Sarah Sundholm for laboratory assistance, John Reeves, Uranbileg Daalkhaijav and Daniel Zerbino for bioinformatics assistance, and Chris Campbell for unpublished data from *Picea*. We also thank Mark Dasenko, Scott Givan and Chris Sullivan of the OSU Center for Genome Research and Biocomputing. National Science Foundation grants (ATOL-0629508 and DEB-0317103 to A.L. and R.C.); Oregon State University College of Science; Oregon State University College of Agriculture; US Forest Service Pacific Northwest Research Station. Funding to pay the Open Access publication charges for this article was provided by the US Forest Service Pacific Northwest Research Station.

Conflict of interest statement. None declared.

REFERENCES

1. Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K. *et al.* (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.*, **5**, 2043–2049.
2. Birky,C.W. (1978) Transmission genetics of mitochondria and chloroplasts. *Annu. Rev. Genet.*, **12**, 471–512.
3. Moore,W.S. (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, **49**, 718–726.
4. Brunsfeld,S.J., Sullivan,J., Soltis,D.E. and Soltis,P.S. (2001) Comparative phylogeography of northwestern North America: a synthesis. In Silvertown,J. and Antonovics,J. (eds), *Integrating Ecological and Evolutionary Processes in a Spatial Context*. Blackwell Science, Oxford, pp. 319–339.
5. Petit,R.J., Aguinagalde,I., de Beaulieu,J.-L., Bittkau,C., Brewer,S., Cheddadi,R., Ennos,R., Fineschi,S., Grivet,D., Lascoux,M. *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.
6. Petit,R.J., Duminil,J., Fineschi,S., Hampe,A., Salvini,D. and Vendramin,G.G. (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol. Ecol.*, **14**, 689–701.
7. Kapralov,M.V. and Filatov,D.A. (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol. Biol.*, **7**, 73.
8. 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. USA*, **84**, 9054–9058.
9. Neale,D.B., Saghai-Maroof,M.A., Allard,R.W., Zhang,Q. and Jorgensen,R.A. (1988) Chloroplast DNA diversity in populations of wild and cultivated barley. *Genetics*, **120**, 1105–1110.
10. Soltis,D.E., Soltis,P.S. and Milligan,B.G. (1992) Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. In Soltis,P.S., Soltis,D.E. and Doyle,J.J. (eds), *Molecular Systematics of Plants*. Springer, London, pp. 117–150.
11. McCauley,D.E. (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends Ecol. Evol.*, **10**, 198–202.
12. Provan,J., Powell,W. and Hollingsworth,P.M. (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol. Evol.*, **16**, 142–147.

13. Graham,S.W. and Olmstead,R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.*, **87**, 1712–1730.
14. Shaw,J., Lickey,E.B., Beck,J.T., Farmer,S.B., Liu,W., Miller,J., Siripun,K.C., Winder,C.T., Schilling,E.E. and Small,R.L. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.*, **92**, 142–166.
15. Shaw,J., Lickey,E.B., Schilling,E.E. and Small,R.L. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.*, **94**, 275–288.
16. Small,R.L., Ryburn,J.A., Cronn,R.C., Seelanan,T. and Wendel,J.F. (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.*, **85**, 1301–1315.
17. Jansen,R., Raubeson,L., Boore,J., dePamphilis,C., Chumley,T., Haberle,R., Wyman,S., Alverson,A., Peery,R., Herman,S. *et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.*, **395**, 348–384.
18. Masooda,M.S., Nishikawaa,T., Fukuokaa,S.-i., Njengaa,P.K., Tsudzukib,T. and Kadowaki,K.-i. (2004) The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. *Gene*, **340**, 133–139.
19. Shaffer,C. (2007) Next-generation sequencing outpaces expectations. *Nat. Biotechnol.*, **25**, 149.
20. Hudson,M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.*, **8**, 3–17.
21. Moore,M., Dhingra,A., Soltis,P., Shaw,R., Farmerie,W., Folta,K. and Soltis,D. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.*, **6**, 17.
22. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
23. Fox,S., Filichkin,S. and Mockler,T. (2008) Applications of ultra-high throughput sequencing in plants. In: Belostotsky,D. (ed), *Plant Systems Biology*. Humana Press, New York, NY (in press).
24. Kasschau,K.D., Fahlgren,N., Chapman,E.J., Sullivan,C.M., Cumbie,J.S., Givan,S.A. and Carrington,J.C. (2007) Genome-Wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.*, **5**, e57.
25. Meyer,M., Stenzel,U., Myles,S., Prufer,K. and Hofreiter,M. (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.*, **35**, e97.
26. Jin,H., Vacic,V., Girke,T., Lonardi,S. and Zhu,J.-K. (2008) Small RNAs and the regulation of cis-natural antisense transcripts in *Arabidopsis*. *BMC Mol. Biol.*, **9**, 6.
27. Binladen,J., Gilbert,M.T.P., Bollback,J.P., Panitz,F., Bendixen,C., Nielsen,R. and Willerslev,E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.
28. Hamady,M., Walker,J., Harris,J., Gold,N. and Knight,R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5**, 237–237.
29. Illumina. (2006) Protocol for Whole Genome Sequencing using Solexa Technology. *BioTechniques Protocol Guide*. Biotechniques, New York, NY.
30. Willyard,A., Syring,J., Gernandt,D.S., Liston,A. and Cronn,R. (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Mol. Biol. Evol.*, **24**, 90–101.
31. Hillier,L., Marth,G., Quinlan,A., Dooling,D., Fewell,G., Barnett,D., Fox,P., Glasscock,J., Hickenbotham,M., Huang,W. *et al.* (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods*, **4**, 183–188.
32. Zerbino,D. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using De Bruijn graphs. *Genome Res.*, **18**, 821–829.
33. Wyman,S.K., Jansen,R.K. and Boore,J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.
34. Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046.
35. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
36. Mayer,C. (2008). *Phobos, a Tandem Repeat Search Tool for Complete Genomes*. Version 3.2.6. http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm (1 August 2008, date last accessed).
37. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
38. Dhingra,A. and Folta,K. (2005) ASAP: amplification, sequencing & annotation of plastomes. *BMC Genomics*, **6**, 176.
39. Jakobsson,M., Sall,T., Lind-Hallden,C. and Hallden,C. (2007) The evolutionary history of the common chloroplast genome of *Arabidopsis thaliana* and *A. suecica*. *J. Evol. Biol.*, **20**, 104–121.
40. Landwehr,N., Mielikäinen,T., Eronen,L., Toivonen,H. and Mannila,H. (2007) Constrained hidden Markov models for population-based haplotyping. *BMC Bioinformatics*, **8** (Suppl. 2), S9.
41. Okou,D.T., Steinberg,K.M., Middle,C., Cutler,D.J., Albert,T.J. and Zwick,M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.
42. Albert,T.J., Molla,M.N., Muzny,D.M., Nazareth,L., Wheeler,D., Song,X., Richmond,T.A., Middle,C.M., Rodesch,M.J., Packard,C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.