Meeting the Challenges of Non-Referenced Genome Assembly from Short-Read Sequence Data

M. Parks and A. Liston
Department of Botany and Plant Pathology
Oregon State University
Corvallis, Oregon 97331
USA

R. Cronn
Pacific Northwest Research Station
USDA Forest Service
Corvallis, Oregon 97331
USA

Keywords: next-generation sequencing, massively parallel sequencing, *Pinus,* Illumina

Abstract

Massively parallel sequencing technologies (MPST) offer unprecedented opportunities for novel sequencing projects. MPST, while offering tremendous sequencing capacity, are typically most effective in resequencing projects (as opposed to the sequencing of novel genomes) due to the fact that sequence is returned in relatively short reads. Nonetheless, there is great interest in applying MPST to genome sequencing in non-model organisms. We have developed a bioinformatics pipeline to assemble short read sequence data into nearly complete chloroplast genomes using a combination of de novo and reference-guided assembly, while decreasing reliance on a reference genome. Initially, short read sequences are assembled into larger contigs using de novo assembly. De novo contigs are then aligned to the corresponding reference genome of the most closely related taxon available and merged to form a consensus sequence. The consensus sequence and reference are in turn 'merged' such that aligned de novo sequence remains unaffected while missing sequence is filled in using the reference sequence. This chimeric reference is then utilized in reference-guided assembly to align the original short read data, resulting in a draft plastome. Using two established *Pinus* reference plastomes, our method has been effective in the assembly of 33 chloroplast genomes within the genus *Pinus,* and results with four species representing other genera of Pinaceae suggest the method will be of general use in land plants, particularly once limitations of PCR-based chloroplast enrichment are overcome.

INTRODUCTION

High throughput sequencing technologies have increased DNA and RNA sequencing capacity by orders of magnitude within the last decade. For example, the first human genomic sequence, finished in the early part of this decade, took over a decade to produce at an estimated total cost of between 0.3-3 billion US dollars (Lander et al., 2001; Venter et al., 2001; Collins et al., 2004; Bentley et al., 2008). In contrast, resequencing of human genomes today can be completed in a matter of weeks, with the cost measured in tens to hundreds of thousands of dollars (Bentley et al., 2008; Wang et al., 2008; Ahn et al., 2009; Mardis et al., 2009), and trends suggest the rate of progress in sequencing capacity is still increasing (Gupta, 2009). Currently at the forefront of sequencing efforts are massively parallel sequencing technologies (MPST). MPST platforms generate millions of short reads (currently 30-400 bp depending on the platform used (Simon et al., 2009)) in parallel during sequencing, which are then typically mapped back onto a previously sequenced reference genome to determine genomic sequence of sampled organismal or cellular lineages (Holt and Jones, 2008; Ley et al., 2008; Wang et al., 2008; Mardis et al., 2009). Even considering the tremendous sequencing capacity of these technologies, challenges remain in their application to a broad range of sequencing projects. For example, sequence capacity measured in Gbp is clearly excessive for the sequencing of small genomes (such as bacterial, organellar or viral genomes). Thus far, this challenge has been approached through the development of multiplex strategies in which multiple accessions indexed by short barcode tags are sequenced simultaneously (Porreca et al., 2007; Craig et al., 2008; Cronn et al., 2008). Another, and perhaps more

Proc. IS on Molecular Markers in Horticulture
Eds.: N.V. Bassil and R. Martin
Acta Hort. 859, ISHS 2010

daunting challenge, lies in utilizing MPST to sequence the genomes of organisms lacking a closely related reference genome. Clearly, with more distantly related species, the likelihood for sequence divergence and genomic structural rearrangements increases. Utilizing short read sequence data in such cases can quickly become problematic, as divergence and rearrangement make it difficult or impossible to map short reads onto the genomes of distantly related references (Whiteford et al., 2005; Pop and Salzberg, 2008). To counter this second problem, we have developed a short read assembly pipeline which transforms raw sequence data into genomic sequence in four basic steps: 1) de novo assembly of short read data into larger contigs, 2) alignment of de novo contigs to the most-closely related reference genome available, 3) formation of a chimeric reference using aligned de novo contigs, with gaps filled in by the reference genome, and 4) alignment of short read data to the chimeric assembly to form final genomic contigs. While several commercial "ail-in-one" software packages are currently available to serve a similar purpose, these tend to be fairly expensive (typically several thousands of US dollars). In contrast, our assembly pipeline can function entirely with open source software.

To date, we have used our pipeline to assemble 33 chloroplast genomes within the genus *Pinus* (32 of which lacked a same-species reference), as well as four chloroplast genomes of non-pine members of Pinaceae. All genomic sequencing was performed in multiplex (typically 4-6x) on the Illumina 1G genome analyzer, and resulting genomes were estimated to average 92% complete.

MATERIALS AND METHODS

Sequence Preparation

Amplification and sequence preparation followed Cronn et a1. (2008).

Processing of Raw Sequence Data

Microread sequence and quality files were converted from raw sequence output, sorted, binned and had their tags removed using custom perl scripts available at http://www.science.oregonstate.edU/~knausb/genomics_scripts/knaus_scripts.htm1.

Chloroplast Genome Assembly

Assembly from microreads to chloroplast genomes is described in detail elsewhere (Whittall et al., 2009). In brief, microreads from an accession were assembled into larger contigs using de novo assemblers, and aligned to the most closely related reference chloroplast genome available (Fig. lA). A chimeric reference sequence was then created by merging aligned contigs with the reference genome, such that aligned de novo sequence persisted and reference sequence was used in areas missing de novo coverage (Fig. lA). The accession's microreads were then aligned against this chimeric reference using a reference-guided assembler to form the contigs of the draft genome (Fig. 1B). These contigs were then checked for quality and manually edited also as previously described (Whittall et al., 2009).

RESULTS

Assemblies overall (including outgroups) averaged 92% complete (Fig. 2). Assemblies in subgenus *Strobus* averaged 117 kb, with an estimated 8.8% missing data (compared to *P. koraiensis* reference). Subgenus *Pinus* assemblies averaged just less than 120 kb (6% estimated missing data, compared to *P. thunbergii* reference). Outgroup assemblies averaged just over 119 kb (10.4% average estimated missing data compared to *P. thunbergii* reference). De novo assemblies ranged from 64% to over 97% of estimated plastome lengths (avg = 89.2±7.0% standard deviation), while finished assemblies were slightly higher (92.2±5.9% sd) as noted above (See Table 1 for details). Our alignment of all assemblies was 132,715 bp in length, with slightly less than half (62,298 bp) from exons encoding 71 conserved protein coding genes (20,638 amino acids), 36 tRNAs and 4

rRNAs. A high degree of co-linearity is inferred for these genomes due to: 1) the absence of major rearrangements within de novo contigs, and 2) the overall success of the PCR-based sequence isolation strategy (indicating conservation of the order of anchor genes containing primer sites). Nonetheless, several known structural rearrangements, including a tandem duplication of *psbA* in *P. contorta* (Lidholm and Gustafsson, 1991) and the apparent loss of duplicate copies of *psaM* and *rps4* in *P. koraiensis,* could not be confirmed. Two loci, *ycf1* and *ycf2,* stood out as highly variable regions among exons, accounting for 22% of all exon sequence but nearly 52% of exon variable sites. From our assemblies, these two loci also exhibit numerous indels in *Pinus,* although these are difficult to validate completely based on short-read assembly. Because of their variability, assembly success for protein-coding exons was determined both with and without these loci (see below).

Uncorrected p-distances between finished assemblies and their closest established references ranged greater than two orders of magnitude, from 0.000645 (76 total differences to reference in *P. thunbergii)* to 0.079221 (7615 total differences to reference in *Abies firma)* (Table 1). Assembly success generally was correlated weakly with divergence from reference and sequencing effort (here defined as microread count), although significant correlations were found in some cases (Table 2, Fig. 3). Assembly success and divergence from reference were correlated negatively in subgenus *Pinus* and outgroup accessions; this correlation was positive in subgenus *Strobus* and when all pines were considered together (Fig. 3A, Table 2). Assembly success was correlated positively with sequencing effort (here defined as microread count) in both subgenus *Pinus* and *Strobus,* but negatively correlated in outgroup accessions (Fig. 3B, Table 2). Significant correlation (i.e., 95% confidence interval for slope does not include zero) was found between assembly success and sequencing effort in subgenus *Pinus* and when all pines were considered together (positive correlation), and between assembly success and divergence in subgenus *Strobus* (Table 2).

Noncoding regions (aligned positions excluding exons, introns and RNA loci) contained the highest proportions of variable sites when all accessions were considered together, or when subgenus *Pinus,* subgenus *Strobus* and non-pine assemblies were considered separately (Table 3). Exonic (protein coding) sequence, while approximately the same overall length as total noncoding sequence, contained ca. 60-70% as many of the variable sites by comparison; this proportion decreased further with the exclusion of *ycf1* and *ycf2*. Both noncoding and exonic regions appear to have assembled with similar success (Table 3). In contrast, RNA loci had the lowest proportion of variable sites and also the lowest estimated assembly success (Table 3). Similarly, intron sequence was less variable than exonic sequence and noncoding sequence (but not exonic sequence without *ycf*1 and yc*f*2), and had a lower assembly success.

DISCUSSION

We have presented an effective and efficient method for assembling small, non-referenced genomes from short-read sequence data. Relying primarily on open source software, we assembled a total of 37 chloroplast genomes (36 non-referenced) of approximately 118 kb length to an average of 92% completion. The process described herein is an iterative process. Initially, preliminary genomic contigs are created through de novo assembly. These contigs are then refined through alignment to a closest reference, formation of a chimeric reference, and subsequent re-alignment of short read sequences (reference-guided assembly) to the chimeric reference. Key to this process is the formation of the chimeric reference prior to reference-guided assembly, consisting of aligned de novo contigs with reference sequence utilized in place of missing data. **In** theory, this allows for more accurate final assemblies for several reasons, including: 1) clear identification of indels within aligned de novo contigs, 2) potential identification of structural rearrangements through de novo assemblies, and 3) improved reference-guided assembly due to higher sequence identity between the chimeric reference and short read genomic sequences as opposed to simply relying on the closest reference without

manipulation. It is worth noting that we utilized what we considered to be the most up to date and applicable software at the time of our assemblies. However, since that time a considerable amount of effort has been put into developing and refining short read assembly software, such that for each step in our assembly process there are now several options. For example, the open source aligner Mummer (http://mummer.sourceforge.net/) could be used in place of the commercially available aligner CodonCode. Alternative reference-guided alignment programs, such as Maq (http://maq.sourceforge.net/maq-man.shtml), Bowtie (http://bowtie-bio.sourceforge.netlindex.shtml, Langmead et al., 2009) and Yasra (Miller and Ratan, unpublished) could be used in place of RGA.

While the bulk of our assemblies had fairly closely related references (within the same subgenus), it is noteworthy that 13 of our *Pinus* assemblies reside in different sections than the complete reference used in their assembly. This represents an estimated divergence period of 19-47 million years in the case of subgenus *Strobus* accessions and 14-31 million years in the case of subgenus *Pinus* accessions (Willyard et al., 2007; Gernandt et al., 2008). Further, the divergence period between the pines and our non-pine representatives is estimated at 87 to 193 million years (Willyard et al., 2007; Gernandt et al., 2008), thereby representing a several-fold greater divergence period yet. It is not surprising, then, that assembly success was somewhat negatively impacted by increasing phylogenetic distance from the reference (although the opposite trend was seen in subgenus *Strobus)*. Nonetheless, these trends were not particularly strong, and our worst assembly was estimated at 75% complete *(P. cembra)*. This provides validation for the effectiveness of our strategy in assembling genomes fairly divergent from the nearest available reference. Further support is found in the similar success rates of assembling coding and non-coding regions, in that one would expect to see decreased assembly success in more poorly conserved noncoding regions if our assembly strategy was lacking.

Considering that divergence plays a limited role in assembly success, it is then reasonable to ask what the most difficult obstacles are in assembling non-referenced genomic sequences. As reported by Cronn et al. (2008), assembly gaps are consistently found directly adjacent to primer sites used in PCR amplifications with our sequencing strategy. In addition, sequence repeats (such as microsatellites) may also be difficult or impossible to bridge with short read data (Cronn et al., 2008). In our assemblies, primer regions are likely a more significant problem, as they accounted for a substantial portion of missing sequence and precluded the assembly of any contig greater than the largest amplicon (just over 4 kb). In addition, a PCR-based method is more prone to failure in capturing genomic structural rearrangements, as amplification failures will occur when rearrangements span more than one amplicon. This could result in an amplicon being scored as missing or failed without any indication of a rearrangement. Regions with problematic amplification due to technical difficulties, primer divergence, or rearrangements can also eliminate substantial regions of the genomic assembly. For example, missing amplicons are part of the reason for the lower assembly success in rRNA regions, particularly in subgenus *Strobus* (likely due to technical problems with amplification and primer divergence, data not shown).

The limitations of PeR-based approaches noted above may be overcome through hybridization-based strategies (Gnirke et al., 2009; Herman et al., 2009), as these methods promise both more even and more thorough coverage of targeted regions. However, these methods have yet to be proven widely applicable. Alternatively, paired-end sequencing of whole genomic extractions may allow the simultaneous capture of large portions of chloroplast, nuclear, and mitochondrial genomes from a single organism (Meyers and Liston, 2010), particularly if nuclear genomes are relatively small (<1.5 Gbp).

Sequencing effort also clearly plays a role in the overall success and accuracy of assemblies, although there may be a point of diminishing returns and the phylogenetic distance to reference may still playa significant role. For example, with the exception of *P. ponderosa* (which was sequenced over several sequencing runs and prepared with variable methodology), our greatest sequencing effort was in *P. thunbergii* (4.54 million

reads, >1.8x sequencing effort of any other assembly). Nonetheless, this assembly was essentially identical in its completion to those of *P. taeda* and *P. pinaster,* which had 56% and 38% of the sequencing effort, respectively. On the other hand, missing sequence in these accessions is mostly associated with primer locations, which are impossible to recover with our strategy. Notably, other studies (Hillier et al., 2008; Whittall et al., 2009) have also demonstrated improved SNP discrimination with increasing coverage depth. For these reasons, it may be a good strategy in future projects to overestimate necessary sequencing effort rather than trying to maximize taxon density through low coverage levels, depending on the specific aims of the project. Alternatively, when assembling numerous non-referenced genomic sequences a reasonable strategy might be to initially dedicate a larger proportion of sequencing effort to the assembly of one or several representative reference genomes. This could in tum be followed by higher levels of multiplexing (relatively less sequencing effort) for subsequent accessions/taxa.

In the near future, it is reasonable to expect that increasing sequence capacity, as well as concurrent improvements in both targeted sequencing and short read assembly strategies will make de novo assembly of small genomes an increasingly simple process. Illumina predicts that their sequencing capacity will approach 100 Gbp per run by the end of 2009. Theoretically, this is sufficient capacity to sequence over 600 average-sized chloroplast genomes or 5500 average sized animal mitochondria to a depth of 100x in a single sequencing run. In order to efficiently utilize this capacity, however, it is clearly incumbent upon those involved in the sequencing of small genomes to maintain a similar pace of development in the areas of sample preparation and downstream assembly.

ACKNOWLEDGEMENTS

We would like to thank Mariah Parker-deFeniks and Sarah Sundholm for lab assistance, Uranbileg Daalkhaijav, Zachary Foster and Brian Knaus for computing assistance, Linda Raubeson for providing a chloroplast isolation of *Larix occidentalis,* David Gernandt for Sanger sequencing, and Christopher Campbell and Justen Whittall for DNA samples. We also would like to thank Mark Dasenko, Scott Givan, Chris Sullivan and Steve Drake of the OSU Center for Genome Research and Biocomputing. This work was supported by National Science Foundation grants (ATOL-0629508 and DEB-0317103 to A.L. and R.C.), the Oregon State University College of Science and the US Forest Service Pacific Northwest Research Station. New sequences have been deposited in GenBank with accessions numbers FJ899555-FJ899583.

Literature Cited

Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B-C., Kim, S.Y., Kim, W.-Y., Kim, C. et al., 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19:1622-1629.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, L, Brown, C.G., Hall, K.P., Evers, D.J, Barnes, C.L., Bignell, H.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53.

Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. and Conso, I. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G. and Stephan, D.A. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Meth 5:887-893.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R. and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36:e122.

Gernandt, D.S., Magallon, S., Geada Lopez, G., Zeron Flores, O., Willyard, A. and Liston, A. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci. 169:1086-1099.

Gnirke, A., Melnikov, A., Maguire, J, Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S. and Russ, C. 2009. Solution hybrid selection

with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotech. 27:182-189.
Gupta, P.K. 2009. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. 26:602-611.
Herman, D.S., Hovingh, G.K., Iartchouk, O., Rehm, H.L., Kucherlapati, R., Seidman, J.G. and Seidman, C.E. 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nat. Methods 6:507-510.
Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W. et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods 5: 183-188.
Holt, RA. and Jones, S.J.M. 2008. The new paradigm of flow cell sequencing. Genome Res. 18:839.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C, Baldwin, J., Devon, K., Dewar, K., Doyle, M. and FitzHugh, W. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.
Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S. and Hickenbotham, M. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66-72.
Lidholm, J. and Gustafsson, P. 1991. The chloroplast genome of the gymnosperm *Pinus contorta:* a physical map and a complete collection of overlapping clones. Curro Genet. 20:161-166.
Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D. et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. New Engl. J. Med.: 361 :1058-1066. NEJMoa0903840.
Meyers, S.C. and Liston, A. 2010. Characterizing the genome of wild relatives of *Limnanthes alba* (Meadowfoam) using massively parallel sequencing. Acta Hort. 859:309-314.
Pop, M. and Salzberg, S.L. 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24:142-149.
Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F. et al., 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4:931-936.
Simon, S.A., Zhai, J., Nandety, R.S., McCormick, K.P., Zeng, J., Mejia, D. and Meyers, B.C. 2009. Short-read sequencing technologies for transcriptional analyses. Annu. Rev. Plant Biol. 60:305-333.
Venter, J.C, Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A. and Holt, R.A. 200l. The sequence of the human genome. Science 291: 1304-13 5l.
Wang, J., Wang, W., Li, R., Li, Y, Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J. and Zhang, J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60-65.
Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A., Essex, J.W., Roach, P.L., Bradley, M. and Neylon, C. 2005. An analysis of the feasibility of short read sequencing. Nucleic Acids Res. 33:e 17l.
Whittall, J.B., Syring, J., Parks, M., Buenrostro, J., Dick, C, Liston, A. and Cronn, R. 2009. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol. Ecol. (in press)
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.

Tables

Table 1. Summary of genome assemblies and pairwise distances between assembled accessions and their original reference.

Reference	Accession	Estimated plastome length[1] (bp)	Aligned de novo length[2] (bp)	Determined genome length[3] (bp)	Uncorrected p-distance from reference	# of differences to reference
Pinus thunbergii	*Abies firma*	119207	97641	100921	0.079221	7615
	Cedrus deodara	118072	92231	95083	0.069649	6337
	Larix occidentalis	119680	114509	118797	0.065859	7411
	Picea sitchensis	120176	106899	109548	0.059725	6243
	Pinus attenuata	118229	102684	107865	0.013953	1494
	P. banksiana	120166	106027	110792	0.014866	1624
	P. canariensis	120158	114492	112900	0.008007	895
	P. chihuahuana	119202	110034	114793	0.013127	1373
	P. contorta	120011	76276	105833	0.014363	1628
	P. merkusii	119665	107634	113005	0.005494	616
	P. resinosa	120179	114145	117914	0.004079	460
	P. pinaster	119904	109302	119246	0.008528	995
	P. ponderosa	120289	108301	113659	0.012249	1444
	P. taeda	120422	116074	119057	0.012867	1512
	P. thunbergii	119717	114086	117936	0.000645	76
	P. torreyana ssp. *torreyana*	120401	108722	104432	0.012955	1335
	P. torreyana ssp. *insularis*	120412	108402	107978	0.013025	1388
Pinus koraiensis	*P. albicaulis*	117266	106403	107163	0.001573	168
	P. aristata	118226	111542	114628	0.014976	1683
	P. armandii	117141	99769	100399	0.002927	293
	P. ayacahuite	117424	103665	104985	0.005723	596
	P. cembra	115825	81630	86922	0.001317	114
	P. flexilis	117346	110273	110415	0.005543	607
	P. gerardiana	117615	114769	115466	0.007133	814
	P. krempfii	116598	110554	113730	0.007139	803
	P. lambertiana S	117515	104050	105203	0.005472	571
	P. lambertiana N	116449	107317	114390	0.001759	200
	P. longaeva	117726	111303	114798	0.014308	1611
	P. monophylla	116104	107095	112804	0.014291	1582
	P. monticola	116841	94630	93800	0.003391	316
	P. nelsonii	116616	106160	109434	0.015587	1671
	P. parviflora	115986	106600	109043	0.002189	238
	P. peuce	116697	106938	108157	0.004556	489
	P. rzedowskii	116802	106225	111128	0.015918	1727
	P. sibirica	116593	96630	97547	0.001572	153
	P. squamata	117848	109936	112199	0.007035	780
	P. strobus	116854	100714	103545	0.004956	511

[1]Determined based on full alignment.
[2]Total length of aligned de novo contigs.
[3]Total of all positions in length with determined base call.

Table 2. Statistical summaries of relationships shown in Figure 3.

Group	Regression	Slope of linear regression line	Confidence interval (95%) for slope	Correlation (R^2)
Subgenus *Pinus*	assembly success/ p-distance	-4.22	-9.38, 0.94	0.249
	assembly success/ microread count	2.55 ($x10^{-8}$)	0.45 ($x10^{-8}$), 4.66 ($x10^{-8}$)	0.422
Subgenus *Strobus*	assembly success/ p-distance	6.70	1.56, 11.83	0.294
	assembly success/ microread count	4.70 ($x10^{-8}$)	-0.88 ($x10^{-8}$), 10.27 ($x10^{-8}$)	0.148
All pines	assembly success/ p-distance	3.42	-0.41, 7.25	0.100
	assembly success/ microread count	3.34 ($x10^{-8}$)	0.85, 5.82	0.201
Outgroups	assembly success/ p-distance	-5.12	-31.33, 21.09	0.261
	assembly success/ microread count	-4.00 ($x10^{-8}$)[1]	-0.88 ($x10^{-8}$), 10.27 ($x10^{-8}$)	0.249

[1]Evidence of non-normality in data set.

Table 3. Summaries of alignment length, variable sites and assembly success for different partitions of aligned assemblies.

Type of region	Accessions included	Alignment length (bp)	No. variable sites in alignment	Percentage of variable sites in alignment	Ave. percent sequence completion
Non-coding	all (*n*=37)	58967	13209	22.4	93.5
	subgenus *Pinus* (*n*=13)	51342	2081	4.1	96.3
	subgenus *Strobus* (*n*=20)	49497	2526	5.1	92.9
	non-pines only (*n*=4)	55134	7495	13.6	87.1
Exons	all	57694	7924	13.7	93.4
	subgenus *Pinus*	55464	1726	3.1	93.9
	subgenus *Strobus*	56479	1797	3.2	93.5
	non-pines only	56508	3844	6.8	90.8
Exons, no *ycf*1 or *ycf*2	all	48919	4500	9.2	94.2
	subgenus *Pinus*	48769	815	1.7	94.2
	subgenus *Strobus*	48726	974	2.0	94.7
	non-pines only	48776	2377	4.9	91.3
Introns	all	13053	1780	13.6	87.7
	subgenus *Pinus*	12570	265	2.1	87.7
	subgenus *Strobus*	12627	387	3.1	87.2
	non-pines only	12784	1065	8.3	90.1
rRNA	all	4524	99	2.2	78.4
	subgenus *Pinus*	4518	18	0.4	79.0
	subgenus *Strobus*	4515	16	0.4	74.6
	non-pines only	4523	47	1.0	95.3
tRNA	all	1356	47	3.5	83.4
	subgenus *Pinus*	1356	5	0.4	90.6
	subgenus *Strobus*	1356	13	1.0	81.3
	non-pines only	1356	28	2.0	69.9

Figures

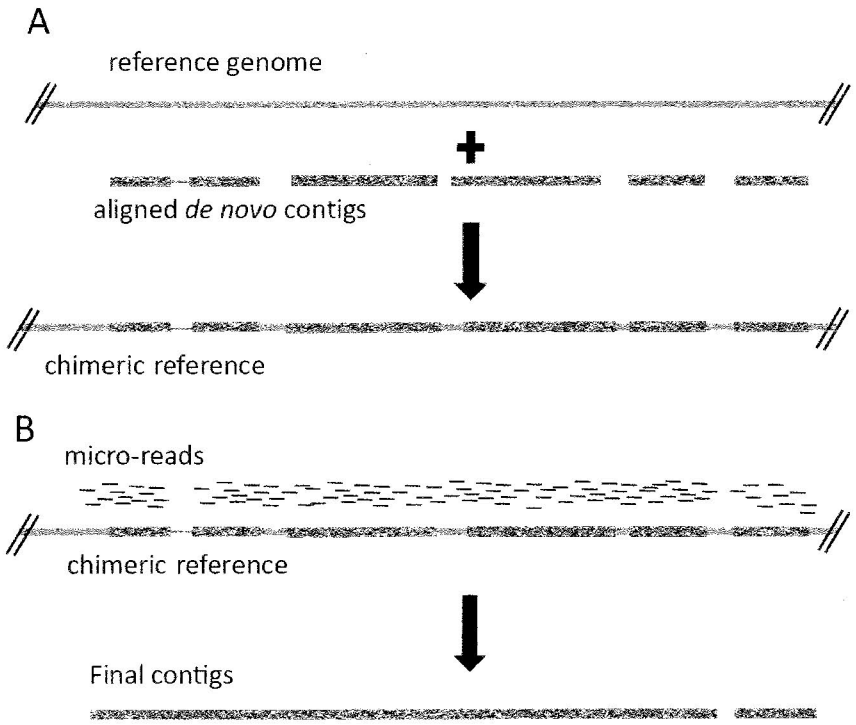

Fig. 1. Schematic diagram of assembly process from short read data. A) Alignment of de novo contigs to reference genome and merging to form chimeric reference. B) Alignment of microreads to chimeric reference to form genomic contigs.

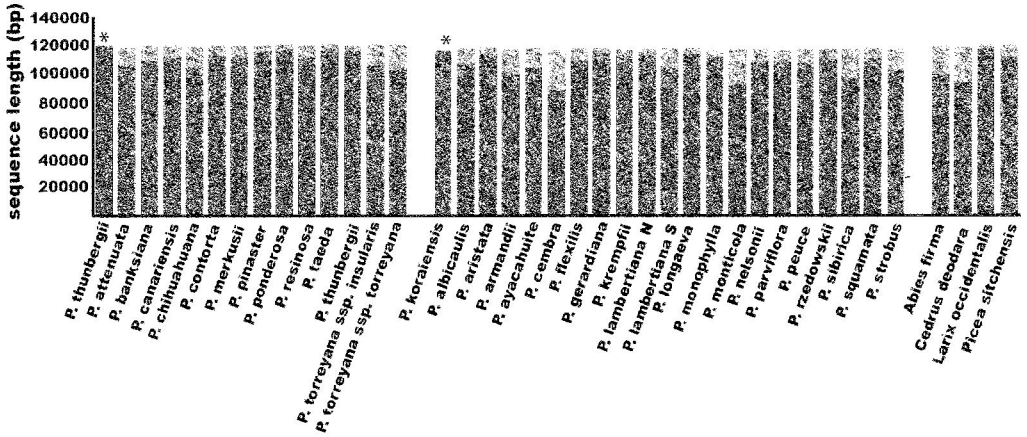

Fig. 2. Assembly success for *Pinus* and outgroup assemblies. Dark shading indicates amount of chloroplast genomic sequence successfully assembled; light shading indicates estimated unassembled sequence. *Indicate reference genomes.

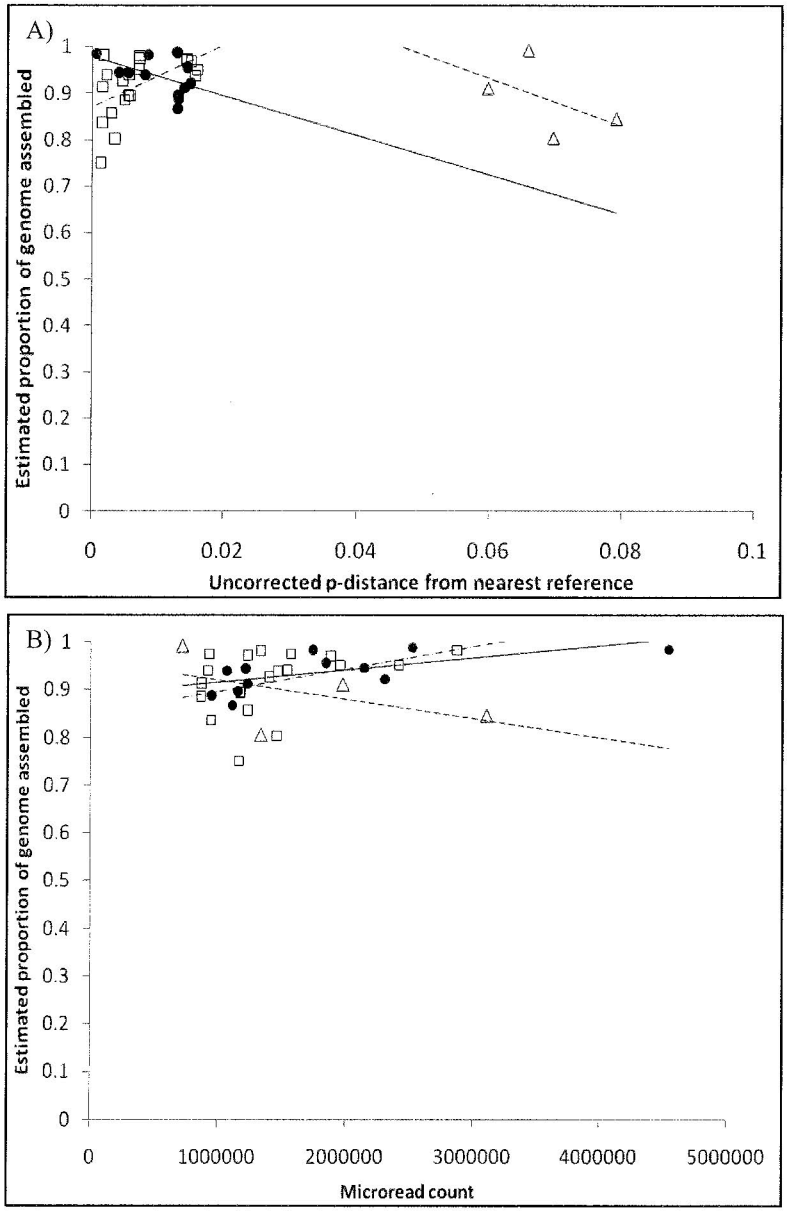

Fig. 3. Potential factors contributing to assembly success in ingroup accession assemblies. A) Relationship of assembly success and divergence from nearest complete reference used in assemblies. B) Relationship between assembly success and sequencing effort (i.e., number of microreads). In each chart, relationships are shown for subgenus *Pinus* (solid circles and lines), subgenus *Strobus* (squares and dash-dot lines) and outgroup (triangles and dashed lines) accessions. Results for *P. ponderosa* were not included in these estimations as the sequencing effort for this accession was substantially higher (10-20x the number of reads from several sequencing runs) than that for other accessions. Regression lines for analyses of all pines not shown.