BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 12 2013, pages 1492–1497
doi:10.1093/bioinformatics/btt178

Genome analysis

Advance Access publication May 22, 2013

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D. Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A. Taylor[1], Macaire Man Saint Yuen[4], Christopher I. Keeling[4], Dana Brand[1], Benjamin P. Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A. Moore[1], Yongjun Zhao[1], Andrew J. Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6] and Steven J.M. Jones[1,2,9]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada, [2]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada, [3]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [5]British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2, Canada, [6]Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [7]Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1K 7P4, Canada, [8]Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada and [9]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Michael Brudno

ABSTRACT

White spruce (*Picea glauca*) is a dominant conifer of the boreal forests of North America, and providing genomics resources for this commercially valuable tree will help improve forest management and conservation efforts. Sequencing and assembling the large and highly repetitive spruce genome though pushes the boundaries of the current technology. Here, we describe a whole-genome shotgun sequencing strategy using two Illumina sequencing platforms and an assembly approach using the ABySS software. We report a 20.8 giga base pairs draft genome in 4.9 million scaffolds, with a scaffold N50 of 20 356 bp. We demonstrate how recent improvements in the sequencing technology, especially increasing read lengths and paired end reads from longer fragments have a major impact on the assembly contiguity. We also note that scalable bioinformatics tools are instrumental in providing rapid draft assemblies.

Availability: The *Picea glauca* genome sequencing and assembly data are available through NCBI (Accession#: ALWZ0100000000 PID: PRJNA83435). http://www.ncbi.nlm.nih.gov/bioproject/83435.

Contact: ibirol@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 10, 2013; accepted on April 11, 2013

1 INTRODUCTION

The assembly of short reads to develop genomic resources for non-model species remains an active area of development (Schatz *et al*., 2012). The feasibility of the approach and its scalability to large genomes was demonstrated by the ABySS publication (Simpson *et al*., 2009) using human genome sequencing data and was later used to assemble the panda genome with the SOAPdenovo tool (Li *et al*., 2010). The technology provides high quality results, as demonstrated for bacteria (Bankevich *et al*., 2012; Ladner *et al*., 2013; Ribeiro *et al*., 2012), and has been successfully applied numerous times on more complex genomes (Chan *et al*., 2011; Chu *et al*., 2011; Diguistini *et al*., 2009, 2011; Godel *et al*., 2012; Swart *et al*., 2012).

Estimated at 20 giga base pairs (Gb) (Murray, 1998), sequencing and assembly of the genome of this gymnosperm species of the pine (*Pinaceae*) family present unique challenges. On the data generation end, those challenges include representation biases in whole-genome shotgun sequencing data, and difficulties in building reduced representation resources to scale down the magnitude of the problem. On the bioinformatics end, assembling massive sequencing datasets is extremely demanding on computing cycles, memory usage, storage requirements, and for parallel programming implementations on communication traffic.

We addressed the data representation challenges by preparing and sequencing multiple whole-genome shotgun libraries on the HiSeq 2000 and MiSeq sequencers from Illumina (San Diego, CA, USA). Compared with localized sequencing protocols, such as building and sequencing fosmid libraries, or the recent approach of isolating ~10 kb DNA strands to generate indexed sequencing fragments in high throughput (Moleculo, San Diego, CA, USA), a shotgun only sequencing approach rapidly provides sequence data effectively covering the target genome at a cost that can be an order of magnitude less. The difference in cost is especially substantial when sequencing a large genome.

In this work, we demonstrate that shotgun sequence assembly at this scale remains viable and produces valuable results. To

*To whom correspondence should be addressed.

assemble the spruce genome, we used the ABySS algorithm (Simpson *et al.*, 2009), which captures a representation of read-to-read overlaps by a distributed de Bruijn graph and uses parallel computations to build the target genome. The modular nature of the tool allowed us to execute a large number of tests to tune the message passing interface for a successful execution, train the assembly parameters for an optimal assembly and quantify the utility of long reads for large genome assemblies. To the best of our knowledge, the ABySS algorithm is unique in its ability to enable genome assemblies of this scale using whole-genome shotgun sequencing data.

2 METHODS

2.1 Sample collection

Apical shoot tissues were collected in April 2006 from a single white spruce (*Picea glauca*, genotype PG29) tree at the Kalamalka Research Station of the British Columbia Ministry of Forests and Ranges, Vernon, British Columbia, Canada. Genomic DNA was extracted from 60 gm tissue by BioS&T (http://www.biost.com/, Montreal, QC, Canada) using an organelle exclusion method yielding 300 μg of high quality purified nuclear DNA.

2.2 Library preparation and sequencing

DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 s using an E210 sonicator (Covaris) and then analysed on 8% PAGE gels. The 200–300 bp (for libraries with 250 bp insert size) or 450–550 bp (for libraries with 500 bp insert size) DNA size fractions were excised and eluted from the gel slices overnight at 4°C in 300 μl of elution buffer {5:1 [vol/vol] LoTE buffer [3 mM Tris–HCl (pH 7.5), 0.2 mM EDTA]/7.5 M ammonium acetate} and was purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adapters (with 5′ overhangs). Adapter-ligated products were purified on QIAquick spin columns (Qiagen) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen).

The mate pair (MPET, a.k.a. jumping) libraries were constructed using 4 μg of genomic DNA with the Illumina Nextera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin containing mate pair junction adapter, which left a short sequence gap in the tagmented DNA. The gap was filled by a strand displacement reaction using a polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9 kb and 9–13 kb fractions, which were purified using a Zymoclean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by a digestion to remove any linear molecules and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by Agilent DNA 1000 series II assay and KAPA Library Quantification protocol. The two fractions were pooled for sequencing paired end 100 bp using Illumina HiSeq2000.

The construction of the 12 kb mate pair libraries was achieved by a hybrid 454/Illumina procedure. Briefly, 50 μg of genomic DNA was fragmented for 20 cycles at speed code 12 using a Hydroshear (Digilab, Marlborough, MA) equipped with a large assembly module. The fragmented DNA was loaded on a 1% agarose gel, and fragments from 12 to 18 kb were extracted. Biotinylated circularization adapters from the GS Titanium Paired-end Adaptor set (454 Life Sciences/Roche, Branford, CT) were added to ends of the gel-extracted fragments. Homologous recombination of the ends was performed with Cre recombinase (New England Biolabs, Ipswich, MA), and linear molecules remaining in solution were removed with Plasmid Safe (Epicentre, Madison, WI). Circular molecules were fragmented using GS Rapid Library Nebulizers (454 Life Sciences/Roche, Branford, CT), and fragment end-repair followed by A-tailing was performed with the GS Rapid Libray preparation kit (454 Life Sciences/Roche, Branford, CT). TruSeq Adaptors (Illumina, San Diego, CA) were ligated to the repaired/A-tailed ends. Biotinylated fragments were enriched using Streptavidin-coupled Dynabeads (Life Technologies, Grand Island, NY) and amplified by PCR using Illumina specific primers.

Random bacterial artificial chromosome (BAC) sequencing was performed using DNA from the same genotype on 454 GS-FLX Titanium with 6 kb paired-end libraries at the PlateForme d'Analyses Génomiques of the Institute for Systems and Integrative Biology (Université Laval, Quebec City, QC). A single paired-end library was prepared on a pool of 15 BACs (equimolar concentrations) as described earlier in the text with the following modifications: 15 μg of DNA was fragmented using a Hydroshear with a standard assembly for 20 cycles at speed code 18, 6–10 kb fragments were extracted from the gel and GS-FLX library adaptors were ligated to the repaired/A-tailed fragments. GS-FLX sequencing using the titanium chemistry was performed according to manufacturer's instructions (454 Life Sciences/Roche, Branford, CT). Sanger sequencing method was used to obtain targeted BAC sequencing data as previously described (Hamberger *et al.*, 2009; Keeling *et al.*, 2010).

2.3 MiSeq modification

In sequencing the spruce genome, we generated longer read lengths by modifying the MiSeq platform. The MiSeq uses a clamshell style cartridge (Supplementary Fig. S1A) to hold reagent tubes in an array that is accessed by the MiSeq's sippers. Most of the reagents are used for read length independent steps such as denaturation and cluster generation, but three reagents, the Scan, Cleavage and Incorporation mixes, are consumed at each cycle. Although the MiSeq allows any read length to be specified in the control software, the reagent cartridge cannot be replaced during the run without stopping it. Increasing the read length therefore requires increasing the quantity of the length-dependent mixes in the cartridge. This led to the solution of combining the length-dependent reagents of two kits into one.

A tool was designed that opens the snap-hook latches holding the cartridge together (Supplementary Figs S1B and S2), giving access to the reagent tubes, yet allowing the cartridge to be put back together without damage to its components (Supplementary Fig. S1C). At 40 ml, the stock length-dependent reagent containers allow for a maximum of ~650 cycles in total. To maximize the potential of the combined kit approach, a new reagent tray with 70 ml wells was designed and placed in a modified clamshell base.

2.4 Read merging

Reads from the HiSeq 2000 PET 250 bp libraries and the MiSeq PET 500 bp libraries were merged using abyss mergepairs (Supplementary Fig. S3). This utility performed a pair-wise Smith Waterman overlapped alignment (Smith and Waterman, 1981) between reads pairs, and selected the best quality base where alignments returned mismatching bases. An arbitrary base was selected when qualities were identical. In cases of read-to-read alignment ambiguity, read pairs were not merged.

Many of the MiSeq reads had significantly reduced qualities near the ends, and the second reads typically had more bases with reduced quality (Supplementary Fig. S4). Therefore, the first and second reads were initially trimmed to appropriate lengths, and further quality trimmed based on a quality score threshold of 15. In all datasets, the reads typically merged into long and high quality reads. Supplementary Box S1 shows the command lines and options used for read merging.

2.5 Assembly process

There were four sets of reads used in the assembly: (i) Merged reads, including MiSeq PET 500 bp libraries and HiSeq PET 250 bp, and (ii) HiSeq PET 500 reads were used in the initial sequence assembly. The HiSeq PET 500 and (iii) unmerged HiSeq PET 250 reads were used for paired linking information to generate contigs. Finally, (iv) the long fragment MPET libraries were used to bridge over segments of repetitive regions to form scaffolds (Supplementary Fig. S5). All stages of the assembly were run using the ABySS wrapper, abyss-pe, with the exception of generating the FM-index (see Supplementary Material for details), and the parameters used are outlined in Supplementary Box S2. Assembly execution times are given in Supplementary Table S1.

2.6 Read alignments

To generate contigs and scaffolds, we first needed to align the reads to the previous unitig and contig stages of the assembly, respectively. Owing to the size of the spruce genome, the fragmented nature of the assembly stages, and the size of the resulting fasta files (~40 GB), we note that general purpose read alignment tools were not suitable for the task.

For the same reasons, our standard method of generating an FM-index (Ferragina and Manzini, 2000) within the ABySS-map tool was too memory intensive, requiring >500 GB of RAM. We solved this problem by using bwte (Ferragina *et al.*, 2012), a tool for generating a Burrows Wheeler Transformation (Burrows and Wheeler, 1994), and converted its results to an FM-index with abyss-index (Supplementary Fig. S6). This method allowed us to index the unitigs and contigs using a maximum of ~60 GB of RAM on a single machine.

For the 500 bp PET libraries sequenced on the HiSeq 2000, we aligned each lane of data in parallel, merged alignments into a single file and inferred fragment size distributions using tools within the ABySS package. The read alignments to different unitigs were converted to distance estimates for the contig assembly stage. The alignments and distance estimates for each MPET Library were done using the standard wrapper for ABySS.

3 RESULTS

Prior experience indicates that sampling a large genome with multiple libraries and fragment lengths can mitigate potential sampling biases and capture a more even representation of the underlying genome (DiGuistini *et al.*, 2011; Earl *et al.*, 2011; Keeling *et al.*, 2013; Li *et al.*, 2010). One novel feature in our sequencing approach was to complement the high coverage data from the HiSeq 2000 sequencers with longer reads from the MiSeq, at low coverage, to support the assembly process. Using an early access 2×150 bp kit on the HiSeq 2000, we generated PET reads from two libraries with 250 bp nominal fragment lengths and 18 libraries with 500 bp nominal fragment lengths to a total of 64-fold raw coverage. Using the MiSeq, we generated 2×300 bp PET reads from four libraries with 500 bp nominal fragment lengths and 2×500 bp PET reads from one library with 500 bp nominal fragment length, contributing a further 4-fold raw coverage (Table 1). We also generated large fragment libraries of 6, 8 and 12 kb nominal fragment lengths, to provide linkage information across repeat structures. The first two of these libraries were prepared using the MPET protocol of Illumina, and the third was a pool of seven libraries prepared using a modified 454 paired end sequencing protocol (454 Life Sciences/Roche, Branford, CT). All long fragment libraries were sequenced at 2×100 bp, and the resulting sequences were used only for their linkage information during the scaffold stage of the assembly.

The first release of the sequencing chemistry on the MiSeq allowed for 150 read cycles, which increased to 250 cycles in a subsequent release. To obtain longer read lengths, we modified the sequencing instrument as detailed in the Supplementary Material. Longer reads obtained from this platform were instrumental in improving the contiguity of the genome assembly as described later in the text.

To build the white spruce genome, we used the ABySS *de novo* assembly tool (Simpson *et al.*, 2009), which has been rigorously evaluated (Earl *et al.*, 2011; Li, 2012; Rahman and Pachter, 2013). The process broadly has three stages: unitig, contig and scaffold building. Performed on a computer cluster of dual Intel Xeon® 6-core processors, each addressing 48 GB of memory, the three stages of the spruce genome assembly took approximately two, four and four days using 1560, 288 and 36 CPU cores, respectively. The first assembly stage constructs a distributed and scalable de Bruijn graph to represent read-to-read overlaps for unitig assembly. The second stage involves read-to-unitig alignments and path traversals on the unitig adjacency graph from the previous stage for contig construction, resolving short repeats along those paths, when possible. Similarly, the third stage involves read-to-contig alignments and path traversal on the contig-adjacency graph for scaffold construction, denoting unresolved sequence content with Ns.

The first assembly stage can further be conceptualized in two parts, with roughly equal run times. The first part provides preliminary unitigs and a graph representing their adjacency, which is defined as (k-1) bp overlaps, where k is the primary assembly parameter denoting the minimum read-to-read overlap length that is considered specific enough for assembly. The second part eliminates short redundant unitigs, when their neighbors on both sides share more than a certain length of sequence (default: 10 bp), and enables further graph simplification. We based the preliminary assessment of assembly quality on the first part of the first stage, and optimized our assembly parameters, using the pre-unitig contiguity statistics (Fig. 1).

The relatively short run time of this part of the assembly process allowed us to execute a large number of tests. Here, we also take this opportunity to demonstrate the utility of longer reads, as they pertain to the optimal *k*-mer length. As a larger k would resolve longer sequence ambiguities, it is desirable to choose it as

Table 1. Sequencing data

Library protocol	Read length (bp)	Sequencing platform	Nominal fragment length (bp)	# Libraries	# Reads (M)	Fold coverage
PET	150	HiSeq 2000	250	2	1520	11.4
PET	150	HiSeq 2000	500	18	7000	52.5
PET	300	MiSeq	500	4	170	2.6
PET	500	MiSeq	500	1	46	1.2
MPET	100	HiSeq 2000	6000	1	816	N/A
MPET	100	HiSeq 2000	8000	1	769	N/A
454	100	HiSeq 2000	12 000	7	34	N/A

Note: Short fragment libraries prepared by the Illumina PET protocol were used for their sequence content. Long fragment libraries prepared by the Illumina MPET, and a modified 454 protocols were used for linkage information during scaffolding. As long fragment libraries do not contribute to the sequence content of the assembly, their contribution to genome coverage is marked as N/A (not applicable).

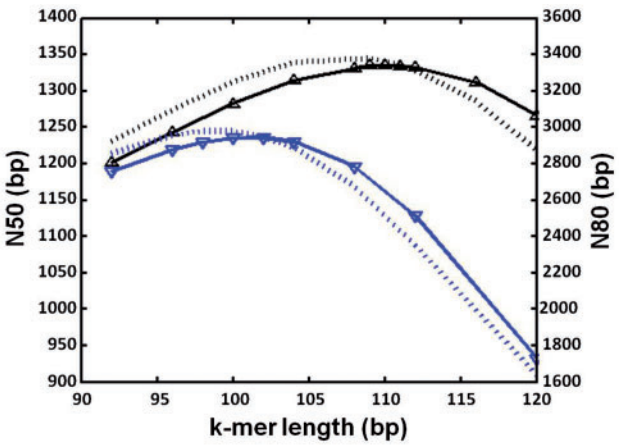

Fig. 1. Assembly optimization. The de Bruijn graph stage (pre-unitig) of the assembly was used to optimize the overlap parameter, *k*-mer length, and the effect of inclusion of longer reads was assessed. The contiguity metrics N50 (solid curves, left *y*-axis) and N20 (dotted curves, right *y*-axis) are shown for assemblies that use the short reads only (blue) and short and long reads (black). The contiguity of the two datasets peaked for different *k*-mer lengths, with dataset of short and long reads having a maximum N50 and a maximum N20 for the same k = 109 bp. For short reads only, optimization with respect to N20 resulted in a slightly lower *k*-mer length (98 bp) compared with optimization with respect to N50 (101 bp), both of which are lower than the optimum *k*-mer length for the full dataset. Longer *k*-mers were desirable, as they help disambiguate longer repeat motifs

large as possible, yet not too large; otherwise, we lose the sensitivity to detect valid read-to-read overlaps.

The choice of this parameter is determined by several factors, including read lengths; fold coverage; genome complexity; genome size; and experimental noise. Among these, genome complexity and size are determined by the choice of the species to study, and the experimental noise is determined by the sample collection methods and the choice of the sequencing platform. Longer reads and higher fold coverage would enable one to use a longer *k*-mer length.

Figure 1 shows results of our parameter search for an optimum *k*-mer length using the short HiSeq 2000 reads only and combined short HiSeq 2000 and long MiSeq reads. The contiguity statistics NX (describing X% reconstruction in assembled sequence lengths NX or longer) are typically locally concave-down functions of the *k*-mer length in a neighborhood where the total sequence reconstruction is close to the genome length. Controlled for the misassemblies, contig or scaffold N50 lengths are widely used quality metrics for genome assemblies. We demonstrate that the optimum pre-unitig N50 occurred at k = 109 bp when using both short and long reads, and at k = 101 bp when using just the short reads. Thus, we observe that incorporating the modest low coverage long read data from the MiSeq allowed us to use a more stringent overlap parameter for the assembly process, and resulted in improved assembly statistics. Optimum k using short and long read data yielded N50 = 1335 bp, whereas the optimum k using short read data only yielded N50 = 1236 bp. The difference between assembly contiguity numbers (7.4%) became more pronounced when the assembly process proceeded to use 158 million and 186 million pre-unitigs in these two cases, respectively, to construct unitigs (9.7%), contigs (9.8%) and scaffolds (17.8%). We also note that the optimum k values in both datasets were longer than the sequencing length of the MPET libraries.

We propagated the optimum k = 109 bp assembly with short and long reads through the assembly pipeline. The full assembly of our data yielded 191 347 (171 971) scaffolds over 20 356 bp (22 967 bp) in length, representing >50% of the 20.8 Gb reconstructed (20 Gb estimated) spruce genome. The assembly statistics of the white spruce genome are presented in Table 2 in comparison with the whole-genome shotgun sequence assemblies of three barley cultivars (Mayer *et al.*, 2012). The recent barley genome assemblies represent results from a whole-genome shotgun sequencing and assembly project using similar data and offer a context for plant genomics.

As a means to assess the quality of the assembled spruce genome, the sequences of several BACs from the same genotype were aligned against the assembly using BLAST (Altschul *et al.*, 1990). Six previously sequenced targeted BACs containing known terpene synthase and cytochrome P450 genes (see Supplementary Material) were then compared with MUMmer dot plots (Kurtz *et al.*, 2004). Twenty-six scaffolds longer than 1000 bp aligned with >95% similarity to reconstruct 62% of the sequence of these six BACs.

Table 2. Assembly Statistics

Species	White spruce			Barley		
	Unitig	Contig	Scaffold	Morex Contig	Bowman Contig	Barke Contig
Number ≥ 500 bp	12.0 M	6.7 M	4.9 M	715 k	729 k	823 k
Number ≥ N50	2.2 M	1.2 M	191 k	121 k	124 k	170 k
Number ≥ NG50	3.0 M	1.0 M	172 k	N/A	N/A	N/A
N80 (bp)	824	2041	2041	1054	1131	998
N50 (bp)	1928	4996	20 356	2793	2994	2330
NG50 (bp)	1548	5351	22 967	N/A	N/A	N/A
N20 (bp)	4070	10 791	80 133	6537	6742	5077
Max (bp)	61 182	99 924	1 047 232	36 062	37 442	38 285
Reconstruction (Gb)	17.4	21.2	20.8	1.3	1.4	1.4

Note: Number, contiguity and reconstruction statistics at the three assembly stages of the white spruce genome in comparison with the contig statistics of whole-genome shotgun assemblies of three barley cultivars (Mayer *et al.*, 2012). The NG50 calculations assume a predicted genome size of 20 Gb for white spruce. As barley assemblies reconstruct less than half the estimated genome size of 5.1 Gb, the NG50 calculations are not applicable and are marked as N/A.

Even though we did not use any protein, expressed sequence tag or RNA-seq data to improve the scaffolds, the whole-genome shotgun approach produced a good quality assembly as measured by its representation of genic regions. First, we searched our assembly for a list of 248 ultra-conserved core eukaryotic genes (CEGs) as selected by Parra, Bradnam and Korf (Parra *et al.*, 2007). Using their Core Eukaryotic Genes Mapping Approach, we identified 95 (38%) of the CEGs as complete sequences and found 184 (74%) of them to have at least a partial representation in our assembly. Similarly, and on a broader scale, we used a set of 13 036 full-length cDNA clones from a closely related species, Sitka spruce (*Picea sitchensis*) (Ralph *et al.*, 2008), and measured their representation in the white spruce assembly. We identified 11 108 (83%) of them in our scaffolds, with 5895 (45%) presented in a single scaffold over 90% of their length. We verified that the exon orders of the Sitka spruce transcripts we investigated were conserved in the white spruce assembly.

From those 5895 genes found in the white spruce, we estimated the mean and the median gene lengths to be 5151 bp and 804 bp, respectively. These lengths represent genomic distances much less than our typical scaffold lengths.

4 CONCLUSION

The choice between a whole-genome shotgun sequencing approach and sequencing reduced representation libraries was extensively discussed during the Human Genome Project (Lander *et al.*, 2001; Venter *et al.*, 2001), and the former became the dominant technology as the sequencing throughput rapidly increased, rendering library techniques to prepare data for the latter approach relatively expensive. A decade later, researchers studying conifer genomes are trying to answer the same question. In our study, we demonstrate that modern whole-genome shotgun sequencing and assembly methods can provide competitive draft genome assemblies at the multi-Gb scale for downstream biological studies in a cost-effective way, even if it is far from producing chromosome level contiguous sequence.

We note that a rigorous assessment of the reported assembly is not a trivial undertaking and will need to be performed, as the assembly evolves from its draft stage toward a more established reference. For example, *de novo* assembly evaluation tools such as CGAL (Rahman and Pachter, 2013), FRCbam (Vezzi *et al.*, 2012) and ALE (Clark *et al.*, 2013) would either not scale to the size of the problem or require substantial time and computational resources. Still, compared with previous targeted gene and genome subsampling studies, the assembly introduced in this article already gives the community considerably greater power to identify and study gymnosperm genes, to assist forest management strategies and to understand the environmental biological interactions that involve spruce trees at a basic level.

ACKNOWLEDGEMENTS

The authors would like to thank the British Columbia Cancer Agency, Michael Smith Genome Sciences Centre for Illumina sequencing and targeted BAC Sanger sequencing and the Institute for Systems and Integrative Biology for random BAC Roche 454 sequencing used in this letter.

Funding: Funds were received by (J.M., J.Boh., I.B., A.Y., J.Bou., K.R., S.J.M.J.) through Genome Canada, Genome Quebec, Genome British Columbia and Genome Alberta for the SMarTForests Project (www.smartforests.ca).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Burrows,M. and Wheeler,D. (1994) A block sorting lossless data compression algorithm. *Technical Report 124*, Digital Equipment Corporation, Palo Alto, CA.

Chan,Q.W. *et al.* (2011) Updated genome assembly and annotation of Paenibacillus larvae, the agent of American foulbrood disease of honey bees. *BMC Genomics*, **12**, 450.

Chu,Y. *et al.* (2011) Genome sequence of Mycoplasma capricolum subsp. capripneumoniae strain M1601. *J. Bacteriol.*, **193**, 6098–6099.

Clark,S.C. *et al.* (2013) ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, **29**, 435–443.

Diguistini,S. *et al.* (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol.*, **10**, R94.

DiGuistini,S. *et al.* (2011) Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen. *Proc. Natl Acad. Sci. USA*, **108**, 2504–2509.

Earl,D. *et al.* (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.*, **21**, 2224–2241.

Ferragina,P. *et al.* (2012) Lightweight data indexing and compression in external memory. *Algorithmica*, **63**, 707–730.

Ferragina,P. and Manzini,G. (2000) Opportunistic data structures with applications. In: *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. Los Alamitos, CA.

Godel,C. *et al.* (2012) The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB J.*, **26**, 4650–4661.

Hamberger,B. *et al.* (2009) Targeted isolation, sequence assembly and characterization of two white spruce (Picea glauca) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. *BMC Plant Biol.*, **9**, 106.

Keeling,C.I. *et al.* (2010) Identification and functional characterization of monofunctional *ent*-copalyl diphosphate and *ent*-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. *Plant Physiol.*, **152**, 1197–1208.

Keeling,C.I. *et al.* (2013) Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. *Genome Biol.*, **14**, R27.

Kurtz,S. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Ladner,J.T. *et al.* (2013) Genome sequence of Weissella ceti NC36, an emerging pathogen of farmed rainbow trout in the United States. *Genome Announc.*, **1**, e00187–12.

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

Li,R. *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Mayer,K.F. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

Murray,B.G. (1998) Nuclear DNA amounts in gymnosperms. *Ann. Bot.*, **82**, 13.

Parra,G. *et al.* (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Rahman,A. and Pachter,L. (2013) CGAL: computing genome assembly likelihoods. *Genome Biol.*, **14**, R8.

Ralph,S.G. *et al.* (2008) A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). *BMC Genomics*, **9**, 484.

Ribeiro,F.J. *et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Res.*, **22**, 2270–2277.

Schatz,M.C. *et al.* (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.*, **13**, 243.

Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Swart,E.C. *et al.* (2012) The Oxytricha trifallax mitochondrial genome. *Genome Biol. Evol.*, **4**, 136–154.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Vezzi,F. *et al.* (2012) Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PloS One*, **7**, e52210.