bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2

Ka Ming Nip[1,2], Saber Hafezqorani[1,2], Kristina K. Gagalova[1,2], Readman Chiu[1], Chen Yang[1,2], René L. Warren[1], Inanc Birol[1,3]

[1] Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada V5Z 4S6
[2] Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada V5Z 4S6
[3] Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

Corresponding authors:
Ka Ming Nip (kmnip@bcgsc.ca) and Inanc Birol (ibirol@bcgsc.ca)

Keywords:
RNA sequencing
Long reads
Transcriptome assembly
Transcript isoforms
Nanopore sequencing
Algorithm

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

ABSTRACT

Long-read sequencing technologies have improved significantly since their emergence. Their read lengths, potentially spanning entire transcripts, is advantageous for reconstructing transcriptomes. Existing long-read transcriptome assembly methods are primarily reference-based and to date, there is little focus on reference-free transcriptome assembly. We introduce RNA-Bloom2, a reference-free assembly method for long-read transcriptome sequencing data. Using simulated datasets and spike-in control data, we show that the transcriptome assembly quality of RNA-Bloom2 is competitive to those of reference-based methods. Furthermore, RNA-Bloom2 requires 27.0 to 80.6% of the peak memory and 3.6 to 10.8% of the total wall-clock runtime of a competing reference-free method. Finally, we showcase RNA-Bloom2 in assembling a transcriptome sample of *Picea sitchensis* (Sitka spruce). Since our method does not rely on a reference, it sets up the groundwork for large-scale comparative transcriptomics where high-quality draft genome assemblies are not readily available.

INTRODUCTION

RNA sequencing (RNA-seq) has become the standard method for gene/transcript discovery, transcriptome profiling, and isoform expression quantification. Since the dawn of high-throughput short-read sequencing technologies, transcriptome assemblies have enabled discovery of novel isoforms[1], identification of foreign RNAs[2], intra-species gene-fusion transcripts[3] and inter-species chimeric transcripts[4,5], and guided scaffolding[6] and annotation of draft genome assemblies. Such applications have been key in enhancing our understanding of genome biology and etiology and progression of various diseases.

Pacific Biosciences of California, Inc. (PacBio, Menlo Park, CA) and Oxford Nanopore Technologies PLC (ONT, Oxford, UK) have been offering long-read sequencing technologies commercially since 2011 and 2014, respectively. Both sequencing technologies have improved significantly since their emergence to yield increased read length, base accuracy, and throughput[8]. In particular, ONT's MinION devices are small and portable, thus having the potential to allow rapid sequencing and downstream analyses[9]. Moreover, nanopore sequencing enables direct RNA (dRNA) sequencing without the need to generate complementary DNA (cDNA) libraries[10]. On the other hand, PacBio's single-molecule-real-time (SMRT) sequencing provides circular consensus sequencing (CCS) to produce reads that have a lower base error rate than that of ONT reads[11]. As a result, the number of computational methods designed for processing and analyzing long-read sequencing data is growing rapidly[8].

Compared to Illumina short-read sequencing technologies, long reads are noisier but are several orders of magnitude longer, making them able to span through multiple exons and even capture full-length transcripts in some instances, thus simplifying the transcriptome assembly problem. However, existing transcriptome assemblers, such as StringTie2[12], are predominantly reference-based where transcripts are derived from spliced-alignment of reads against the reference genome. Genome annotations contain rich information about gene structures that may be utilized for guiding reference-based transcriptome assembly; some examples include: refinement of spliced-alignment based on known splice junctions, inference of transcript strand based on

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

annotated gene orientation, and resolution for antisense transcripts based on known transcription start and end sites. Consequently, a subclass of reference-based assemblers, such as FLAIR[13], require a genome annotation in addition to the reference genome for accurate isoform reconstruction.

Reference-free assembly of transcriptomes is especially valuable when there is no available reference genome or the reference genome is still at the draft stage, which may not fully support reference-based assembly of all transcripts in a given transcriptome sequencing sample. In general, long-read reference-free genome assembly algorithms such as wtdbg2[14] (also known as Redbean) are not suitable for transcriptome data because they cannot reconstruct alternative isoforms and they typically assume a uniform sequencing depth, which is practically nonexistent in transcriptomic data due to varying transcript expression levels. Reference-free assembly methods typically rely on read-to-read mapping, whereas reference-based methods rely on read-to-reference alignments. Since read-to-read mapping is much more resource-intensive than read-to-reference alignment, reference-free methods tend to have a much higher computational cost than reference-based methods. To find wider applications, reference-free assembly algorithms need to overcome the challenges in managing computational resources.

Sequence clustering-based assembly follows the divide-and-conquer paradigm and thus it requires less resources than methods that align all reads against each other. RATTLE[15] is an example of such a method, and to the best of our knowledge, it is the only reference-free transcriptome assembler that can assemble transcripts solely from long-read sequencing data. RATTLE clusters input reads into isoform-based (or gene-based) groupings and derives consensus sequences from each read cluster to reconstruct full-length transcripts. Nevertheless, clustering accuracy is an important factor in the assembly quality and computational performance. A lenient clustering criterion would create few but large read clusters, resulting in slow runtime, high peak memory, and aggregation of reads from too many genes. A stringent clustering criterion, on the other hand, would create many small clusters, potentially resulting in insufficient aggregation of reads and incomplete transcript reconstruction.

Digital normalization[16], also known as *in silico* read normalization, is a simple but effective method to improve the computational performance of reference-free assemblers by reducing the number of overrepresented reads, such as those of high-expressed transcripts, based on the saturation of *k*-mers in the reads. In contrast to naive subsampling, digital normalization is better at preserving low-expression transcripts. However, it has been primarily utilized for the assembly of short-read RNA-seq data[17,18]. With the introduction of strobemers[19] as a mismatch and indel tolerant alternative to *k*-mers, digital normalization with strobemers should be highly applicable to transcriptome assembly of noisy long reads.

Here we present RNA-Bloom2, the successor to our short-read transcriptome assembly tool, RNA-Bloom[20], that extends support for reference-free transcriptome assembly of bulk RNA long sequencing reads. RNA-Bloom2 offers both memory- and time-efficient assembly by utilizing digital normalization of long reads with strobemers. Our benchmarking shows that RNA-Bloom2 requires 27.0 to 80.6% of the peak memory and 3.6 to 10.8% of the total wall-clock runtime of RATTLE. In simulated datasets, RNA-Bloom2 has 0.1 to 5.8% higher recall and 0.3 to 1.0% lower misassembly rates than RATTLE and it has the lowest false-discovery rates in five out of

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

six samples. In experimental datasets, RNA-Bloom2 has the highest recall, up to 9.0% higher than the next best method, and the lowest misassembly rates, tying with FLAIR, in two out of three sequencing platforms. Finally, we showcase RNA-Bloom2 in assembling a transcriptome sample of *Picea sitchensis* (Sitka spruce), without using a genomic reference.

RESULTS

Reference-free transcriptome assembly with RNA-Bloom2

RNA-Bloom2's six-stage workflow for the reference-free transcriptome assembly of long reads is summarized in **Fig. 1**. In stage one, long reads are corrected for errors in an alignment-free approach based on a Bloom filter de Bruijn graph of *k*-mers derived from input reads. Short reads can be optionally provided to aid in the error correction of long reads. In stage two, the set of corrected reads is digitally normalized with strobemers, such that overrepresented reads are removed to yield a target read depth. Stages one and two are highly integrated to reduce input-output operations. Since only a portion of corrected reads would be retained by digital normalization, stage one is not meant to exhaustively correct all errors in the reads and is instead intended to be fast and memory-efficient. In stage three, reads in the normalized set are overlapped against each other to identify low-depth regions in the reads to be trimmed or split. In stage four, trimmed reads are overlapped against each other to generate an overlap graph where reads on each unambiguous path are assembled into a "unitig". In stage five, the unitigs, which may still contain errors, are polished using the alignments of corrected reads from stage one. In stage six, the polished unitigs are aligned against each other to generate an overlap graph where transcripts are derived based on the length-normalized read depth of the unitigs. If the reads are produced by the cDNA sequencing protocol, sequences containing potential poly(A) tails are identified in order to prune the overlap graphs in stages four and six. A more detailed description of each stage is provided in the **Methods** section.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

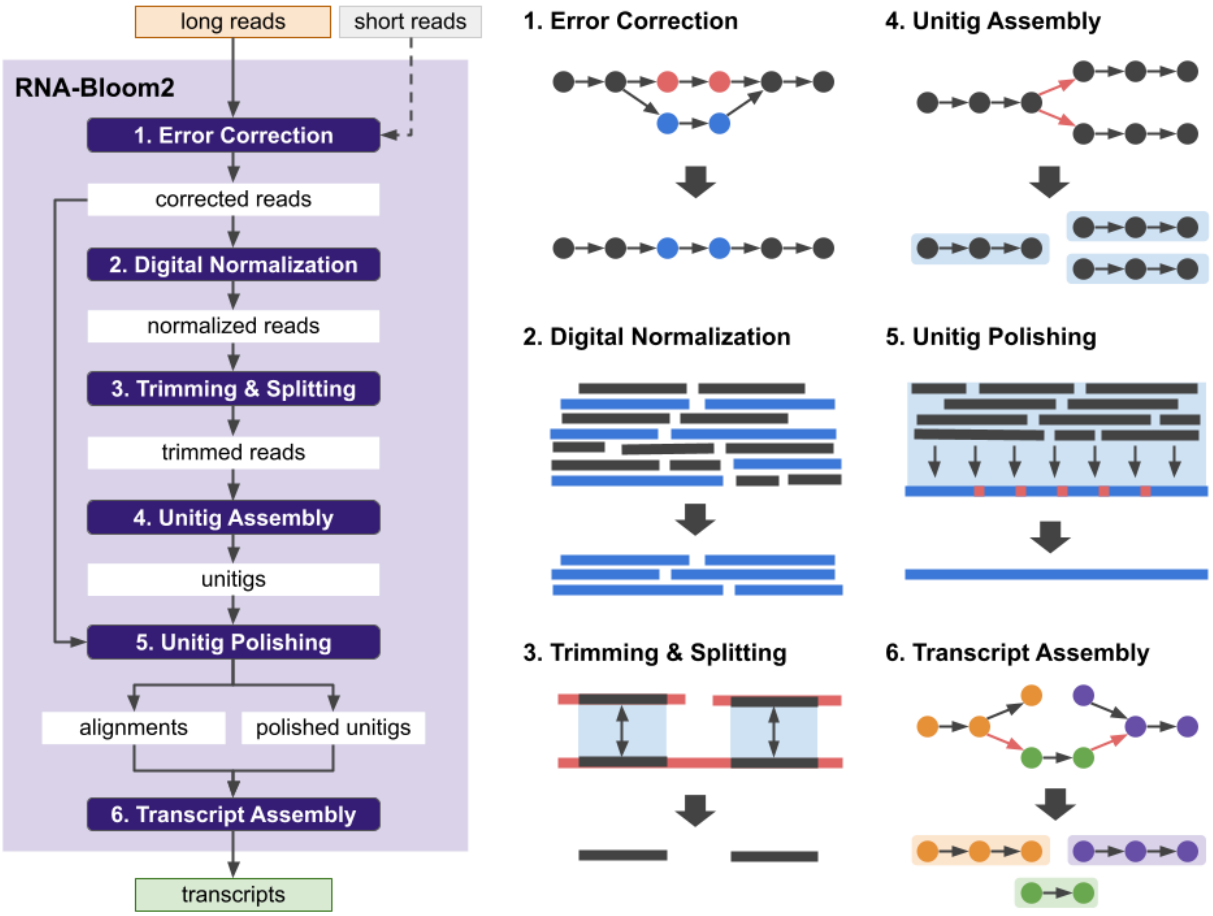

Fig. 1. RNA-Bloom2 assembly workflow overview.
The long-read assembly workflow of RNA-Bloom2 consists of six stages: (1) error correction with long reads (orange rectangle) and optionally with short reads (grey rectangle), (2) digital normalization, (3) trimming and splitting, (4) unitig assembly, (5) unitig polishing, and (6) transcript assembly.

Alignment-free error correction and digital normalization demonstrates utility in multiple data types

We evaluated the effectiveness of the error correction and digital normalization stages of RNA-Bloom2 using experimental data. We selected one mouse dataset from the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium[21] containing the matching sequencing data for ONT cDNA, ONT dRNA, PacBio CCS, and Illumina reads of the same biological sample (**Supplementary Table 1**). ONT dRNA and PacBio CCS reads do not contain adapters, but ONT cDNA reads and Illumina reads are trimmed for adapters with Pychopper[22] and Trimmomatic[23], respectively (**Supplementary Method 1**). Out of the three long-read samples, the ONT cDNA sample has the largest number of sequencing reads (13,127,667 reads) but the lowest read alignment rate (78.66%) against the combined reference genome for mouse

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

GRCm39 and Lexogen's Spike-In RNA Variant (SIRV) transcripts[24]. Compared to the ONT cDNA sample, the ONT dRNA and PacBio CCS samples have only one-sixth of the reads (2,153,439 reads and 2,144,172 reads, respectively) but higher read alignment rates (95.58% and 95.49%, respectively) against the reference genome. The Illumina sample has 40,225,298 read pairs (2 × 100 nucleotides (nt)) and is only used for the hybrid error correction of the long reads in RNA-Bloom2.

We first assess both methods of alignment-free error correction in RNA-Bloom2: (i) using only long reads, and (ii) using a hybrid of long and short reads. We investigated the nucleotide base error rates of the reads before and after error correction (**Supplementary Table 2**); error rates are measured by Trans-NanoSim[25] (**Supplementary Method 2**). The error rates of the reads before error correction in the ONT dRNA, ONT cDNA, and PacBio CCS samples are 12.17%, 7.18%, and 1.96%, respectively. Long-read-only error correction has reduced the error rates to 10.28%, 4.03%, and 1.35% for the ONT dRNA, ONT cDNA, and PacBio CCS samples, respectively. As expected, hybrid error correction has resulted in even lower error rates of 6.55%, 3.51%, and 1.34% for the ONT dRNA, ONT cDNA, and PacBio CCS samples, respectively. Long-read-only error correction has the largest reduction (-3.15%) in the error rate in the ONT cDNA sample, whereas hybrid error correction has the largest reduction (-5.62%) in the error rate in the ONT dRNA sample.

We next investigated the percentage of reads remaining after digital normalization and the percentage of input reads aligned to the final assembly (**Supplementary Table 3, Supplementary Method 3**). For assemblies with long-read-only error correction, 48.15%, 3.76%, and 11.66% of reads remained after digital normalization in the ONT dRNA, ONT cDNA, and PacBio CCS samples, respectively. For assemblies with hybrid error correction, 38.80%, 3.53%, and 11.63% of reads remained after digital normalization in the ONT dRNA, ONT cDNA, and PacBio CCS samples, respectively. The ONT dRNA assemblies have the highest percentages of reads remaining after digital normalization. This is likely due to the much higher error rate in the reads, which limits the number of matching strobemers among the reads. Despite the fact that a substantial proportion of reads are removed by digital normalization, 97.44 to 97.54% and 95.04 to 95.16% of input reads are still able to align to the final assemblies for the ONT dRNA and PacBio CCS samples, respectively. Although 73.52 to 74.02% of input reads aligned the final assembly for the ONT cDNA sample, it is important to note that only 78.66% of reads in the sample were aligned against the reference genome. These results confirm that digital normalization in RNA-Bloom2 is effective in removing overrepresented reads from long-read transcriptome sequencing data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Assembly benchmarking with simulated datasets

We benchmarked the assembly quality and the computational performance of RNA-Bloom2 on simulated data. We prepared two mouse simulated datasets with Trans-NanoSim[25] for the cDNA and dRNA sequencing protocols model on experimental ONT data (See **Methods**). To investigate the effect of sequencing depth, we subsampled each dataset to 2, 10, and 18 million reads, resulting in a total of six sets of reads for our benchmarking experiments. The features of the simulated datasets are presented in **Supplementary Table 4**. Compared to the cDNA dataset, the dRNA dataset has a higher error rate, longer N50 read length, and fewer simulated transcripts.

We compared RNA-Bloom2 against three other transcriptome assembly tools designed for long reads: RATTLE, StringTie2, and FLAIR. RATTLE is the only other reference-free method, whereas StringTie2 and FLAIR are entirely reference-based. In addition, all FLAIR assemblies were guided by the reference transcriptome annotation in conjunction with the associated reference genome. Since reference-based methods are expected to perform better than reference-free methods, StringTie2 and FLAIR serve as the gold-standard for evaluating the performance of RNA-Bloom2 and RATTLE. All assembly methods were run with 48 threads using the same compute nodes with the exception of FLAIR and RATTLE for the assemblies of the 18 million-read sets, which were reprocessed on a high-memory machine after failing the initial runs. Commands for all methods and computing hardware are documented in **Supplementary Method 6**.

The computational performance of all four assembly methods is summarized in **Fig. 2** and **Supplementary Tables 5, 6**. StringTie2 has the fastest runtimes and consistently low peak-memory usage for all datasets. FLAIR has the worst peak-memory usages and RATTLE has the worst total runtimes. As expected, reference-based assemblers are faster than reference-free assemblers. RNA-Bloom2 has the lowest memory usage for the 2 million-read cDNA dataset. The peak memory usage and total runtimes of RATTLE are 1.24 to 3.70 and 9.22 to 28.12 times of those of RNA-Bloom2, respectively. Both RNA-Bloom2 and RATTLE require a higher peak-memory usage in assembling the dRNA datasets than the cDNA datasets, possibly due to the higher error rate and higher N50 read length of the dRNA datasets. However, the peak memory of RNA-Bloom2 for the dRNA datasets did not increase exponentially with respect to the number of input reads. This suggests that the digital normalization stage in RNA-Bloom2 is effective in reducing the number of reads because the number of transcripts in the 10 million-read set and the 18 million-read set only differs by 210 (**Supplementary Table 4**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

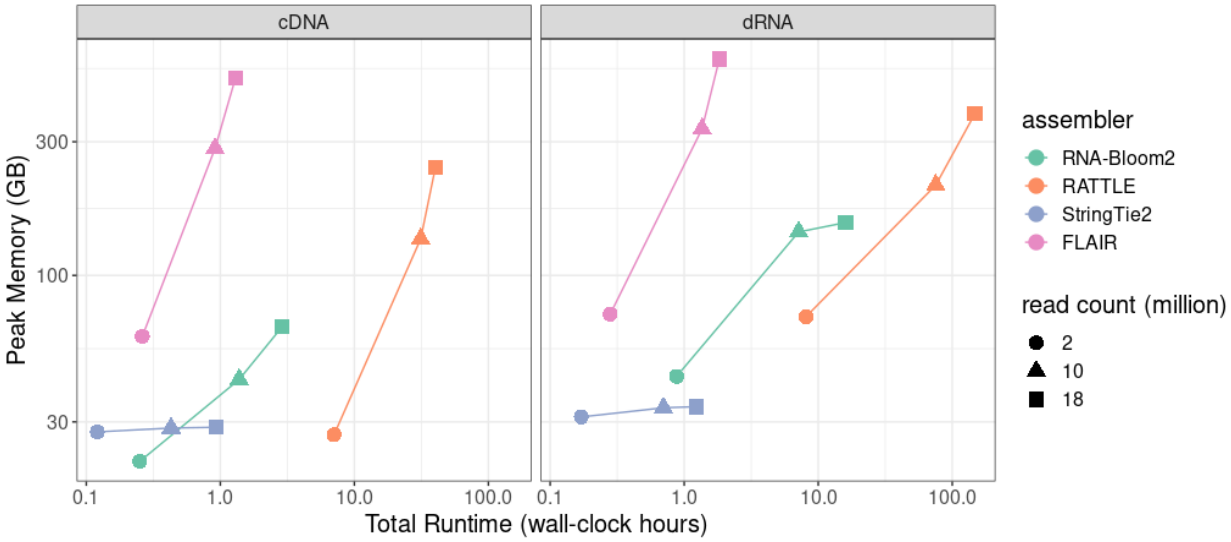

Fig. 2. Computational performance on simulated datasets.
All assemblers were run with 48 threads and assemblies were generated for 2, 10, and 18 million simulated reads of cDNA and dRNA samples. Peak memory usage was measured in GB, and runtime was measured in wall-clock hours. Both axes are in logarithmic scale. For StringTie2 and FLAIR, performance figures include read alignment and generation of indexed BAM files.

We evaluated the assembly quality in the simulated datasets based on the metrics described in **Table 1.** The assembly evaluation procedure is described in the **Methods** section. The benchmarking results are presented in **Fig 3**.

Metric	**Definition**
Complete reconstruction	Truth set transcript reconstructed at least 95% in length
Partial reconstruction	Truth set transcript reconstructed between 0 and 95% in length
Missing reconstruction	Truth set transcript with no detectable reconstruction
True positive	Truth set transcript with complete or partial reconstruction
False positive	Reference transcript not in the truth set
Misassembly	Incorrectly assembled sequence with segments from one or more reference transcripts
Intragenic misassembly	Incorrectly assembled sequence with segments from reference transcripts of the same gene
Intergenic misassembly	Incorrectly assembled sequence with segments from reference transcripts of different genes
Recall	Percentage of truth set transcripts reconstructed.
False discovery rate	= False positives / (False positives + True positives)
Misassembly rate	= Misassemblies / (Misassemblies + True positives)

Table 1. Transcriptome assembly quality assessment metrics.
These metrics are intended for sequencing data with a known ground truth where true-positives and false-positives can be easily discerned. The truth set transcripts are either the set of simulated transcripts or the set of spike-in transcripts in real data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

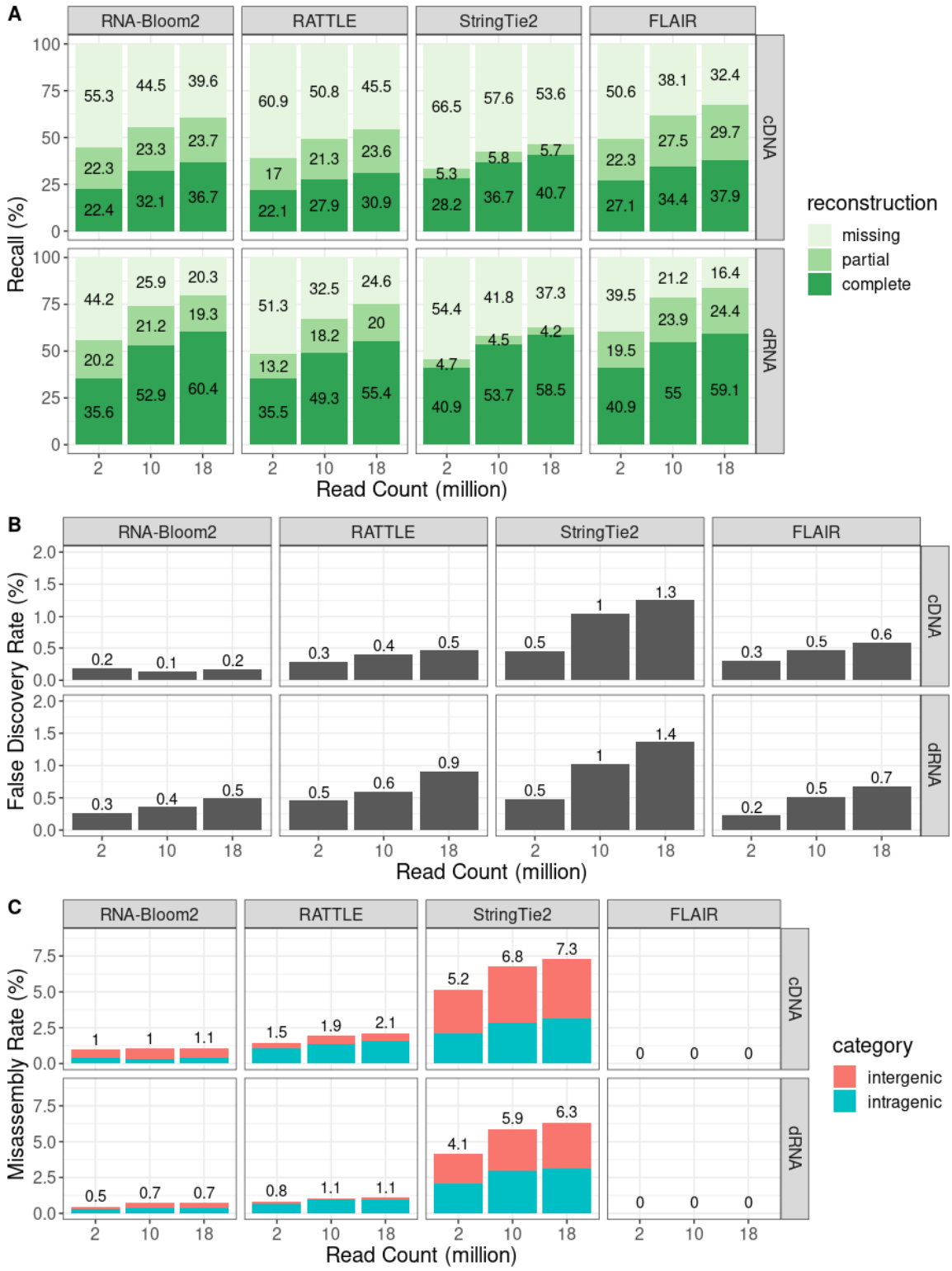

Fig. 3. Assembly quality evaluation of simulated datasets.

(**A**) Recall is measured with respect to transcription reconstruction levels. (**B**) False discovery rate is measured based on false positive transcripts detected. (**C**) Misassembly rate is determined based on the number of intergenic and intragenic misassemblies detected.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

The trends for recall are similar for both simulated cDNA and dRNA datasets (**Fig. 3A**). RNA-Bloom2 has higher percentages (+0.1 to +5.8%) of complete reconstruction than RATTLE in all simulated samples. The largest difference is observed in the 18 million-read cDNA sample, whereas the smallest difference is observed in the 2 million-read dRNA sample. RNA-Bloom2 also has lower percentages (–4.3 to –7.1%) of missing transcripts than RATTLE in all samples. Second to FLAIR, RNA-Bloom2 has the second smallest percentages of missing transcripts (29.6 to 55.3% for cDNA sets and 20.3 to 44.2% for dRNA sets). For all cDNA samples, StringTie2 has the highest percentages of complete reconstruction (28.2 to 40.7%) but the smallest percentages of partial reconstruction (5.3 to 5.8%) for all three dataset sizes. StringTie2 has the highest percentage of missing transcripts in all cDNA (53.6 to 66.5%) and dRNA (37.3 to 54.4%) samples. StringTie2 and FLAIR are tied for having the highest percentage of complete reconstruction (40.9%) for the 2 million-read dRNA sample. FLAIR has the highest percentage of complete reconstruction (55.0%) for the 10 million-read dRNA sample. RNA-Bloom2 has the highest percentage of complete reconstruction (60.4%) for the 18 million-read dRNA sample.

We further investigated assembly recall with respect to transcript expression levels. We assigned simulated transcripts to expression quartiles: low, medium-low, medium-high, and high. The expression-stratified assembly recall results for simulated cDNA and dRNA datasets are presented in **Supplementary Fig. 1 and 2**, respectively. StringTie2 has the most complete reconstruction in the medium-high (32.0 to 55.3%) and high expression (74.8 to 80.3%) cDNA quartiles. FLAIR has the most complete reconstruction in the medium-low (14.6 to 33.2%) and low (3.8 to 16.0%) expression cDNA quartiles. RNA-Bloom2 was second, behind StringTie2, in the high expression cDNA quartile (63.3 to 75.5%). In the high expression dRNA quartile, RNA-Bloom2 has the most complete reconstruction in the 18 million-read sample (88.2%), while StringTie2 has the most complete reconstruction in the 2 and 10 million-read samples (86.5% and 87.5%, respectively). In the medium-high expression dRNA quartile, RNA-Bloom2 has the most complete reconstruction in the 10 and 18 million-read samples (69.6% and 73.8%, respectively), while StringTie2 has the most complete reconstruction in the 2 million-read sample (57.5%). FLAIR has the most complete reconstruction in all sample sizes in the medium-low (30.6 to 55.1%) and low expression (9.6 to 43.9%) dRNA quartiles.

We also evaluated the false-discovery rates (FDR) and misassembly rates (**Fig. 3B**, **C**) for the four assemblers. In all simulated samples, RNA-Bloom2 has lower FDR (–0.1 to –0.3% for cDNA sets, –0.2 to –0.4% for dRNA sets) than RATTLE, and StringTie2 has the highest FDR (0.5 to 1.3% for cDNA sets, 0.5 to 1.4% for dRNA sets) except it was tied with RATTLE in the 2-million read dRNA sample (0.5%). RNA-Bloom2 has the lowest FDR in all simulated cDNA samples (0.3 to 0.5%). In the simulated dRNA dataset, FLAIR has the lowest FDR (0.2%) in the 2 million-read sample while RNA-Bloom2 has the lowest FDR in the 10 and 18 million-read samples (0.4% and 0.5%, respectively). RNA-Bloom2 has higher intergenic misassembly rates

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

and lower intragenic misassembly rates than RATTLE, but the combined misassembly rates are lower (–0.5 to –1.0% in cDNA sets, –0.3 to –0.4% in dRNA sets) in RNA-Bloom2 assemblies. FLAIR has no detectable misassemblies in all samples, while StringTie2 has the highest misassembly rates for intergenic misassembly (3.0 to 4.2% for cDNA sets, 2.0 to 3.2% for dRNA sets), intragenic misassembly (2.1 to 3.1% for both cDNA and dRNA sets), and combined (5.2 to 7.3% for cDNA sets, 4.1 to 6.3% for dRNA sets).

Assembly benchmarking with spike-in control data

In addition to simulated data, we also benchmarked the four assembly methods on experimental sequencing data of known sequences. We selected one mouse dataset from the LRGASP Consortium containing the matching sequencing data for ONT cDNA, ONT dRNA, and PacBio CCS of the same biological sample. The sequencing samples for this dataset were spiked with Lexogen's Spike-In RNA Variant (SIRV) transcripts[24] containing 92 External RNA Control Consortium (ERCC) spike-ins, 69 SIRV isoforms, and 15 long SIRVs. We extracted the reads corresponding to the spike-ins (See **Methods**) for assembly benchmarking and the features of the spike-in datasets are summarized in **Supplementary Table 7**. The PacBio CCS sample has the longest N50 read length (2,460 nt) and the lowest error rate (2.03%). The ONT cDNA sample has the shortest N50 read length (712 nt) but the highest number of reads (n=404,783). The ONT dRNA sample has the fewest reads (n=26,814) and the highest error rate (11.01%).

We evaluated the assembly quality of the spike-in samples based on the metrics described in **Table 1**, and the benchmarking results are presented in **Fig. 4**. For all three samples, RNA-Bloom2 has the smallest percentage of missing reconstruction. RNA-Bloom2 also has the highest percentages of complete reconstruction in both ONT samples (63.6% for cDNA sample, 46.6% for dRNA sample), and it is tied with FLAIR in the PacBio CCS sample (63.1%) (**Fig. 4A**). Unlike what we observed with the simulated datasets, there are no results for false discovery rate in the spike-in data because all reference spike-in transcripts are true positives. Misassembly rates are measured based on only intragenic misassemblies because intergenic misassemblies are not found (**Fig. 4B**). FLAIR has the lowest misassembly rates in all samples and it is tied with RNA-Bloom2 for having zero misassembly rates in the ONT dRNA and PacBio CCS samples. StringTie2, RNA-Bloom2, and RATTLE have the highest misassembly rates in the ONT dRNA (10.1%), ONT cDNA (9.0%), and PacBio CCS (12.0%) samples, respectively. The high misassembly rates in RNA-Bloom2, RATTLE, and StringTie2 for the ONT cDNA sample is likely a result of the low N50 read length.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

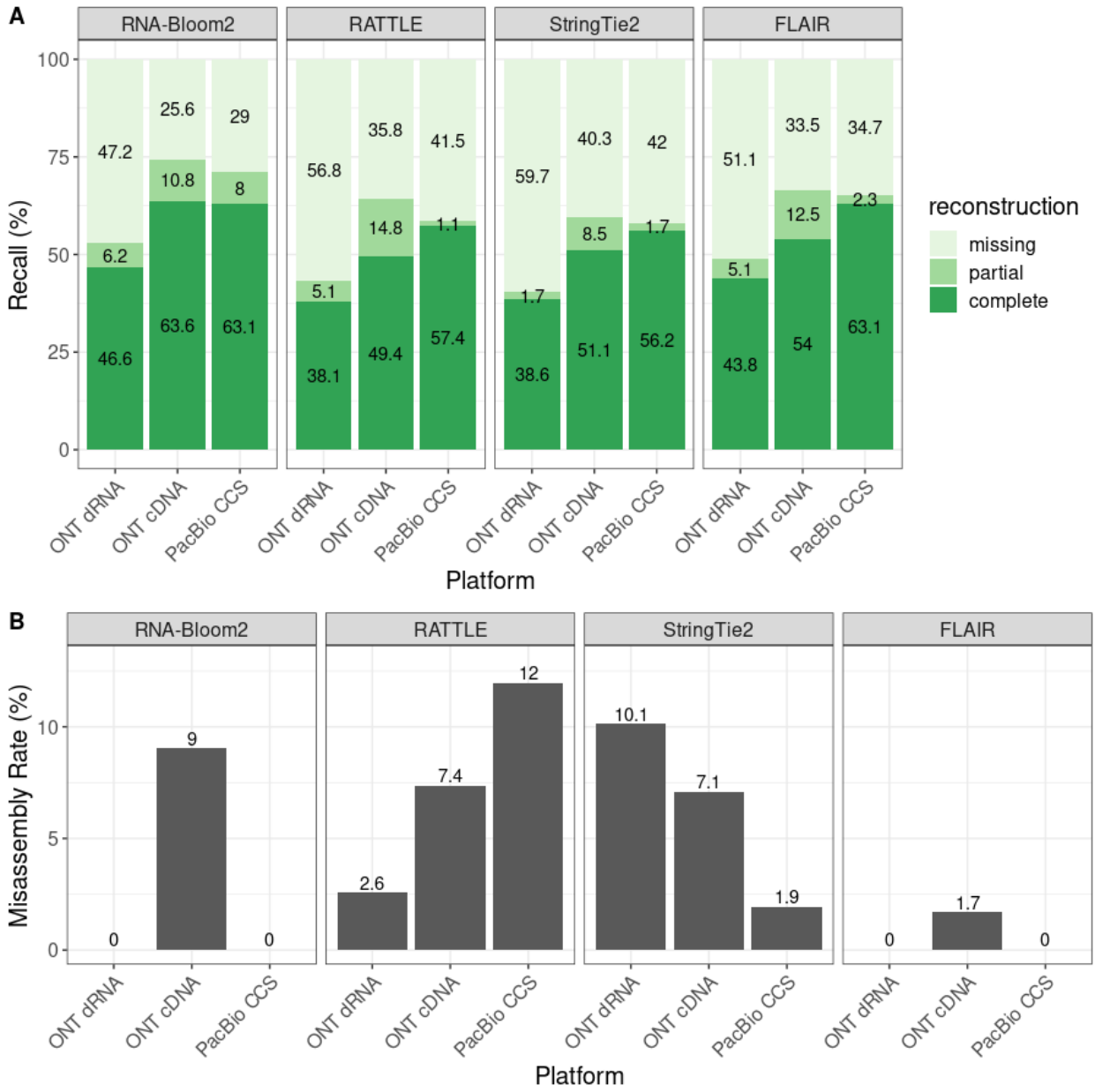

Fig. 4. Assembly quality evaluation of spike-in control datasets.
(**A**) Recall and (**B**) misassembly rates were evaluated for each assembly method on spike-in control data generated from three sequencing technologies: ONT direct RNA, ONT cDNA, and PacBio CCS. The spike-in control data were extracted from a mouse dataset from the LRGASP Consortium.

Reference-free assembly of a Sitka spruce transcriptome

The Sitka spruce (*Picea sitchensis*) is a large, evergreen, and long-living conifer species native to the Pacific Northwest in North America. Although a 20 Gbp draft genome assembly is publicly available[26], its scaffold N50 length is 56.8 kbp, which reflects the draft stage of this short-read

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

genome assembly. In particular, the Sitka spruce's alternative splicing pattern has not been fully investigated. Since conifers are known for their long introns[27,28], the fragmented draft genome would highly limit the effectiveness of reference-based transcriptome assembly methods. Thus, it is a valuable use case to illustrate the utility of reference-free transcriptome assembly methods. Using RNA-Bloom2 and RATTLE, we assembled RNA-seq data from mixed tissue (young needle, bark, xylem, and mature needle) cDNA sampled from a Sitka spruce Q903 spruce weevil-susceptible individual, originated from Haida Gwaii, British Columbia, Canada (53.917, -132.083). The cDNA sample was sequenced on a ONT MinION device (R9.4 flow cell) and the reads are basecalled with Guppy (See **Methods** and **Supplementary Method 7**). A total of 1,323,043 ONT reads with N50 read length of 1,543 nt remained after adapter-trimming with Porechop[30]. We also performed an additional RNA-Bloom2 assembly with hybrid error correction using Illumina paired-end RNA-seq data from a previous study[29].

First, we measured the completeness of single-copy orthologs with BUSCO[31] for the adapter-trimmed ONT reads, the two RNA-Bloom2 assemblies, and the RATTLE assembly (**Supplementary Method 7**). BUSCO provides a quantitative assessment of expected gene content for each set of transcript sequences, and the results are summarized in **Supplementary Table 8**. The RNA-Bloom2 assembly with hybrid error correction has the highest percentage of complete BUSCO and the lowest percentages of fragmented and missing BUSCO. Specifically, the complete BUSCO has improved from 73.4% in the adapter-trimmed reads to 87.6% in the RNA-Bloom2 assembly, whereas the percentages of fragmented and missing BUSCO in the reads (7.7% and 18.9%) have reduced by half after assembly with RNA-Bloom2 (3.4% and 9.0%). On the other hand, the RNA-Bloom2 assembly with long-read-only error correction has a higher percentage of complete BUSCO and lower percentages of fragmented and missing BUSCO than the input reads and the RATTLE assembly. Compared to the reads, the RATTLE assembly has a lower percentage of complete BUSCO and higher percentages fragmented and missing BUSCO.

We have selected the RNA-Bloom2 assembly with hybrid error correction for further analyses. This transcriptome assembly has a total of 68,514 transcripts, where 98.95% of adapter-trimmed reads were aligned to the transcriptome assembly with minimap2[32] (**Supplementary Method 3**). We also aligned the assembled transcripts against the draft genome with minimap2 (**Supplementary Table 9**). A total of 66,866 (97.59%) assembled transcripts were aligned to the draft genome. Of these aligned transcripts, 21,423 (32.04%) transcripts have at least one split-alignments. Since split-alignments on a high-quality genomic reference typically indicate incorrectly assembled transcripts, we compared these split-alignments of assembled transcripts to STAR[33] alignments of the Illumina paired-end RNA-seq data against the draft genome. We found that 13,376 (62.44%) transcripts with split-alignments contain at least one split supported by at least one STAR alignment (**Supplementary Method 7**). This suggests that these transcripts

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

were correctly assembled and the majority of split-alignments is likely a result of fragmented genic regions in the draft genome.

To understand the gene structure of transcripts contained in the genomic scaffolds, we supplied the RNA-Bloom2 assembly with hybrid error correction as full-length RNA sequences to PASA[34] to create a transcript structure annotation based on the draft genome. It is important to note that this annotation produced by PASA is only a partial representation of the Sitka spruce transcriptome due to fragmented genic regions. PASA generated an annotation consisting of 15,222 genes, 18,991 transcripts, 58,049 unique exons, 37,090 unique introns, and 19,079 poly(A) tails (**Fig. 5A**). There are more poly(A) tails than transcripts because PASA collapses transcripts with alternative polyadenylation. Overall, 95.7% of splice junctions from the PASA annotation overlaps with splice junctions in the Illumina paired-end RNA-seq data reported by STAR. We also tallied the frequencies of unique exons, introns, and transcripts per gene (**Fig. 5B**). On average, each gene has 3.8 exons, 2.4 introns, 1.2 transcripts. 59.12% genes contain 2 or more exons, and 16.1% genes contain at least 2 expressed transcripts. A maximum of 55 exons, 53 introns, 13 transcripts are observed per gene.

We calculated the length distributions of exons, introns, transcripts, genes, and poly(A) tails based on the output files from PASA (**Fig. 5C**). Exon lengths range from 10 to 18,115 nt with a primary peak at 116 nt and a slightly shorter secondary peak at 518 nt. Intron lengths range from 21 to 206,268 nt with a primary peak at 113 nt and a much shorter secondary peak at 25,474 nt. 10.3% of introns are longer than 10,000 bp, which is in congruence with the long intron characteristic of conifers. Likely as a result of long introns, gene lengths range from 115 to 364,125 nt with a primary peak at 1,865 nt and a secondary peak at 45,120 nt. Transcript lengths range from 115 to 18,115 nt with a peak at 1,863 nt, which is nearly identical to the peak gene length. Poly(A) tail length ranges between 10 to 104 nt long with a primary peak at 21 nt and a shorter secondary peak at 50 nt. The bimodal poly(A) tail length distribution is also observed in *Arabidopsis* seeding transcriptomes[35].

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

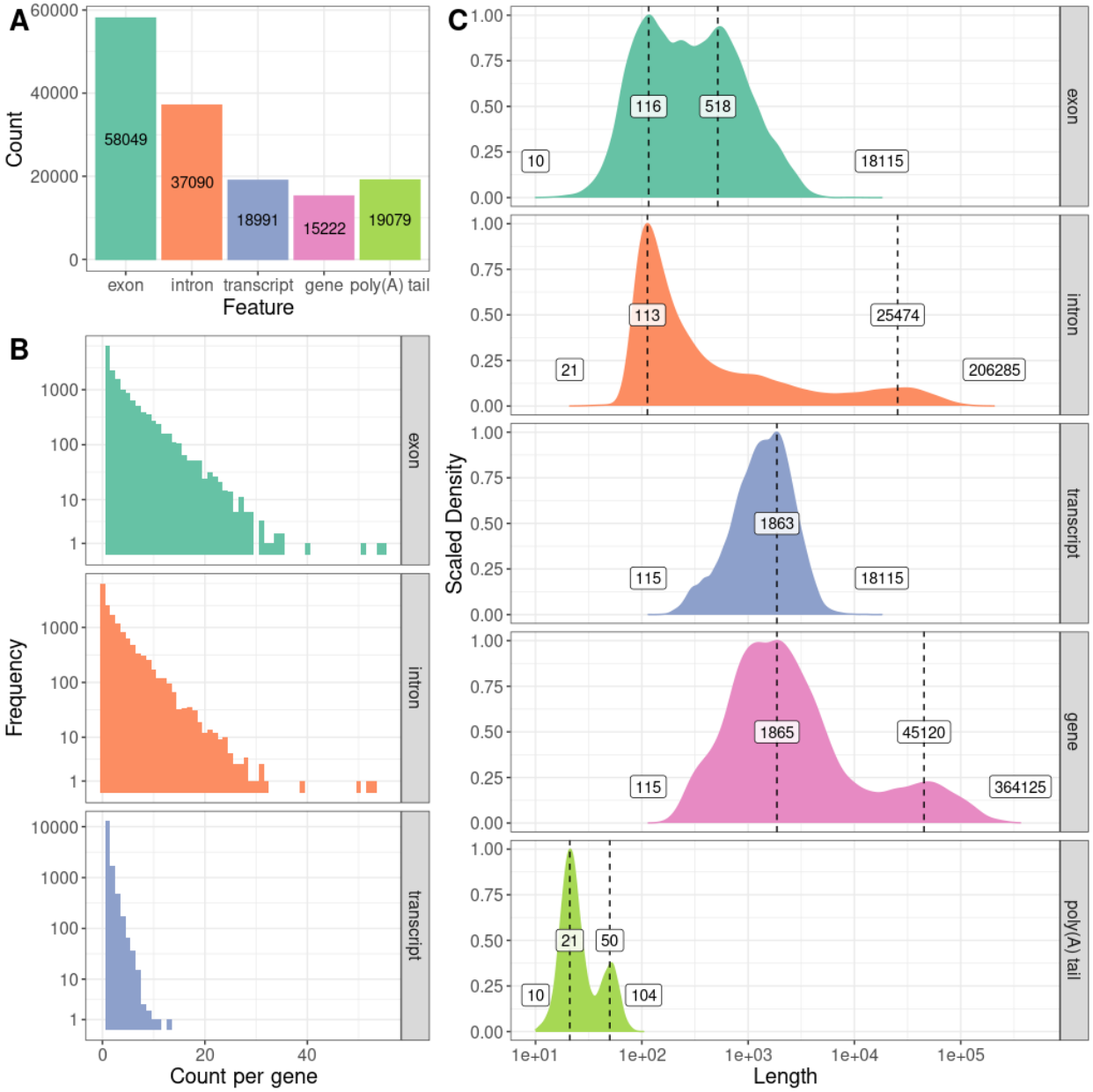

Fig. 5. Distributions of feature lengths and feature counts per gene for the Sitka spruce transcriptome.

(**A**) Total counts of exons, introns, transcripts, gene, and poly(A) tails. (**B**) The frequency of per-gene counts of exons, introns, and transcripts. The vertical "Frequency" axis is presented in logarithmic scale. **(C)** The length distributions of exons, introns, transcripts, genes, and poly(A) tails. The horizontal "Length" axis is presented in logarithmic scale. The vertical axis is scaled to the maximum value for each feature. The minimum and maximum values are indicated at both tails of the distributions. Peak values on the distributions are superimposed on the vertical dotted lines.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

We also investigated alternative splicing events in the assembled transcripts. PASA reports nine types of alternative splicing events (**Fig. 6**): spliced intron, retained intron, alternate acceptor, alternate donor, alternate exon, retain exon, skipped exon, starts in intron, and ends in intron. Spliced intron is the most common event (27.5%), followed by retained intron (24.3%). Retained intron and spliced intron are the most frequently co-occurring event types. Transcripts involving 3 or more event types are detected but are much rarer.

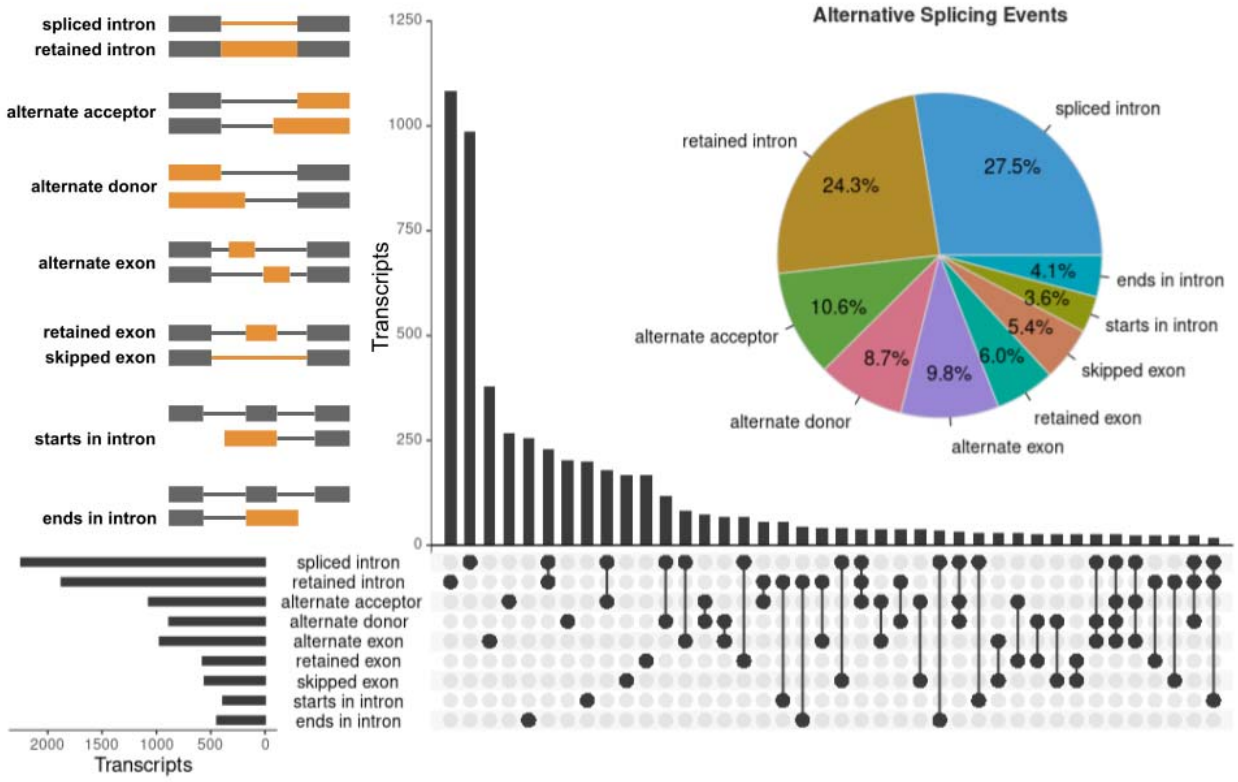

Fig. 6. Alternative splicing events in the Sitka spruce transcriptome.
Nine types of alternative splicing events are presented as they are defined in PASA and depicted by connected grey and orange rectangles in the top-left diagram. The exons, introns, or splice-junctions involved in each event type are highlighted in orange. The pie chart presents the relative proportions of all event types in the PASA annotation. The horizontal bar chart in the UpSet plot shows the total number of transcripts containing each event type. The vertical bar chart in the UpSet plot shows the number of transcripts containing single event types and co-occurring event types, which are indicated by single dots and connected dots in the matrix, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Finally, we applied the EnTAP pipeline[36] to produce protein sequence translation and functional annotation for the RNA-Bloom2 assembly (**Supplementary Method 7**). Using the functional annotation and similarity search to known spruce protein sequences, we have identified the following putative peptides: 15 terpene synthases (TPS), including 10 monoterpene synthases and seven diterpene synthases, 100 cytochrome P450 (CYP) peptides from 55 different subfamilies, and 17 NAM/ATAF/CUC (NAC) transcription factors from six different subfamilies. TPS, CYP, and NAC are gene families known for their contribution to constitutive and induced resistance to damage by the spruce weevils[29,37,38].

DISCUSSION

The rapid improvements to long-read sequencing technologies present a significant challenge to reference-free transcriptome assembly methods. As the throughput of long-read sequencers continues to increase, larger sequencing datasets are produced and thereby increasing the sequence assembly and analysis computational challenge. RNA-Bloom2 addresses this by digital normalization with strobemers. In our benchmarking with simulated data and spike-in control data, we showed that the computational performance and the assembly quality of RNA-Bloom2 significantly surpasses those of RATTLE, a previous reference-free transcriptome assembler that relies on clustering of long reads. In particular, RATTLE has total wall-clock runtimes over nine times that of RNA-Bloom2. Therefore, digital normalization with strobemers, within the RNA-Bloom2 assembly workflow, was a successful application of the concept in assembly of long-read sequencing data, and it was a superior alternative to clustering-based reference-free assembly.

We note that reference-based assembly methods tend to run much faster than reference-free methods, but their overall assembly quality varies depending on the metric used. In simulated data, StringTie2 has better recall than reference-free methods, but it has higher false discovery rates and misassembly rates. FLAIR, on the other hand, has the lowest misassembly rates, possibly due to guidance from the reference annotation. It is important to note that StringTie2 does not strictly require a reference annotation in addition to the reference genome, but FLAIR requires both reference annotation and reference genome. Therefore, the application of FLAIR is mainly limited to discovery of novel isoforms while StringTie2 only requires a good quality reference genome.

As good quality reference annotations are not always readily available, transcriptome assembly methods must manage the lack of known transcription start and end sites, which are crucial in distinguishing the orientation of transcripts and discerning transcripts from antisense overlapping genes. Unlike direct RNA-sequencing, this is a major challenge for cDNA sequencing data, where the strand of reads cannot always be safely assumed. RNA-Bloom2 overcomes this

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

problem by identifying potential poly(A) tail-containing reads. In addition, RNA-Bloom2 filters edges in its overlap graph based on read counts of each edge and its incident nodes, thus removing false overlaps between sequences. The positive effects of these solutions in RNA-Bloom2 are supported by the relatively low false discovery rates and misassembly rates of RNA-Bloom2 in our benchmarking experiments.

In our analyses of the Sitka spruce transcriptome, we illustrated that RNA-Bloom2 assemblies have higher BUSCO completeness than input reads and a RATTLE assembly. We note that a portion of our assembled transcripts have split-alignments across genome scaffolds, but the majority of them are supported by paired-end short reads. We expect that a transcript-informed targeted gene reconstruction[39], using a long-read reference-free transcriptome assembly by RNA-Bloom2, may significantly improve discovery of new splice isoforms and the annotation of genes.

In summary, we illustrated the performance of RNA-Bloom2 with respect to state-of-the-art long-read transcriptome assembly methods, highlighting the strengths and weaknesses of each. We showed that RNA-Bloom2 is suitable for both ONT and PacBio sequencing technologies and it is competitive to reference-based methods. We expect RNA-Bloom2 to be scalable to increasing volumes of long-read data, and we anticipate RNA-Bloom2 will facilitate the gene annotation and transcriptome analyses of many species to be investigated.

METHODS

Alignment-free error correction

We have modified the error correction routine for short reads in RNA-Bloom to support error correction in long reads. First, *k*-mers and their multiplicities in the input reads are stored in a Bloom filter de Bruijn graph. Reads are then split into fixed-length tiles (default: 500 nt) that are evaluated independently of each other. To account for varying transcript expression levels, a multiplicity threshold is dynamically determined (**Supplementary Fig. 3**) within each tile to identify “weak” and “solid” *k*-mers, which have multiplicities lower and higher than or equal to the threshold, respectively. Weak *k*-mers represent potentially erroneous regions in the read while solid *k*-mers represent error-free regions in the read. To avoid introducing incorrect edits to the read, weak *k*-mers are replaced with an alternative path of solid *k*-mers in the de Bruijn graph only if this path shares a high sequence identity (default: 70%) as the target region spanned by the weak *k*-mers. The tiling nature of this routine ensures that more refined multiplicity thresholds are set for sub-regions in the read. The error correction process for each read may be repeated for additional iterations if at least one tile were modified in the previous iteration. In each successive iteration, the tiling positions are shifted by half a tile length to allow errors at tile

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

boundaries in the previous iteration to be corrected. After error correction is completed for the read, *k*-mers from all neighboring tiles are joined together and are assembled into an edited read sequence.

Digital normalization with strobemers

Digital normalization is intended to reduce the overall read depth to a much lower target depth (default: 3) by identifying the minimal longest reads set (MLRS) supporting the target depth. Digital normalization reduces the computational resource requirements of subsequent stages and it is most effective in reducing the number of reads for high-expressed transcripts, which are the main culprit for long runtimes and high memory usage in read-to-read alignments. The read depths represented by the MLRS are approximated by multiplicities of strobemers, which is an error-tolerant alternative to *k*-mers for sequence comparison. Three variants of strobemers have been introduced in previous work[19]: minstrobes, randstrobes, and hybridstrobes. In RNA-Bloom2, we used randstrobes of order three because it was shown to perform favorably on transcriptome data. Strobemer multiplicities are tracked by a counting Bloom filter, which is populated as reads are added to the MLRS.

Digital normalization begins by first sorting the input reads by their length in descending order. Only one read is evaluated at a time to maintain proper tracking of read depth in the MLRS. A read is designated as represented by the MLRS if nearly the entire read (default threshold of 50 nt from the read extremities) contains overlapping strobemers with multiplicities at or above the target depth (**Supplementary Fig. 4**). A read is designated as not represented by the MLRS if it has a region not containing any strobemers with multiplicities at or above the target depth. Each non-represented read is added to the MLRS and its strobemer multiplicities (that are lower than the target depth) would be incremented by one before the next read is evaluated. Represented reads are not included in the MLRS and their strobemer multiplicities are not incremented.

Read trimming and splitting

RNA-Bloom2 relies on minimap2 for overlapping reads against each other to identify sufficiently covered regions of each read. By default, the minimum required read depth for long-read assembly is set to three in RNA-Bloom2; a sufficiently covered region of a read must overlap with at least two other reads. Insufficiently covered head and tail regions of the reads are trimmed. Reads containing insufficiently covered middle region(s) are potentially chimera artifacts and thus are split into shorter sub-sequences (**Supplementary Fig. 5**). Completely contained reads are removed.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Unitig assembly

Trimmed reads are overlapped against each other with minimap2 to construct an overlap graph where the vertices and edges are reads and their overlaps, respectively. As was done in the read trimming stage, contained reads are removed. If the input data is strand-specific (i.e. ONT dRNA), only alignments on the same strand are retained and the overlap graph would only contain vertices for the forward strand. If the input data is not strand-specific (i.e. ONT cDNA), then vertices for both strands are created in the overlap graph and the overlap graph is pruned based on whether each read contains poly(A) tail or poly(T) head (**Supplementary Fig. 6**). The overlap graph is simplified by removing transitive edges. Unitigs are derived by assembling reads along unambiguous paths in the overlap graph.

Unitig polishing

Although alignment-free error correction has been performed on the reads that were used to generate unitigs, there are still residual base errors that can be polished using an alignment-based approach. Output reads from the error correction stage are aligned to the unitigs with minimap2. To avoid unintentional removal of short alternatively spliced exons during polishing, only alignments with large indels (default: > 50 nt) or low sequence identity (default: < 70%) are removed. The filtered alignments are passed to Racon[40] for polishing the unitigs.

Transcript assembly

An overlap graph of polished unitigs is constructed based on minimap2 overlaps between polished unitigs. Reusing the read alignments from the unitig polishing stage, the overlap graph is annotated with: (i) length-normalized read counts for the unitigs, and (ii) the number of reads spanning across the unitig overlaps.

If the input data is not strand-specific, then the overlap graph is pruned as it was done in unitig assembly and the read alignments are also examined for poly-A tail reads that are aligned to the unitigs. The unitigs are reoriented based on the poly-A tail read alignment orientations and the overlap graph is filtered accordingly (**Supplementary Fig. 6**). This procedure is crucial in discerning transcripts originating from overlapping genes on opposite strands of the chromosome. In addition, edges in the overlap graph are filtered by applying a binomial test on the number of reads supporting the edge with respect to the normalized read counts of the incident vertices.

After all filtering on the overlap graph has been performed, vertices are sorted by their read counts in descending order. Each vertex serves as the seed for a bidirectional greedy extension path with each extension choosing the neighbor vertex with the highest read count. Greedy

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

extension terminates upon reaching either a dead-end, a cycle, or a vertex with a read count of zero. The reads along this path are assembled into a transcript. All vertices along this path would be flagged from seeding new extension paths, and their read counts are decremented by the minimum read count in the path. Transcript assembly is complete when all vertices have been visited.

Benchmark dataset simulation

We used Trans-NanoSim v3.1.0 to simulate ONT cDNA and dRNA datasets based on the mouse ENSEMBL annotation for GRCm39. Mouse samples ENCFF232YSU and ENCFF349BIN from the LRGASP Consortium were selected for training Trans-NanoSim sequencing profiles for cDNA and dRNA data, respectively. Sequencing adapters were trimmed from raw reads using Pychopper v2.5.0[22] (**Supplementary Method 1**). Since no adapters were detected in the dRNA data, the raw reads were supplied to Trans-NanoSim for training the dRNA profile. On the contrary, adapters were found in the cDNA data; the adapter-trimmed "full-length" and "rescued" reads, as defined by Pychopper, were supplied to Trans-NanoSim for training the cDNA profile. We discarded all simulated reads defined as "unaligned" by Trans-NanoSim and we subsample the "aligned" simulated reads to 2, 10, 18 million reads using seqtk[41]. All software command parameters are documented in **Supplementary Method 4**.

Spike-in control reads extraction

Using minimap2 2.24-r1122, reads from three replicates for each platform were aligned against the hybrid reference genome of mouse and spike-ins provided by LRGASP. Only reads that are aligned uniquely to ERCC and SIRV sequences are kept (**Supplementary Method 5**).

Transcriptome assembly benchmarking

The command parameters for each assembler are documented in **Supplementary Method 6**. For the simulated datasets, transcriptome assemblies are aligned against the mouse ENSEMBL reference transcriptome with minimap2. The output alignment PAF files are processed with our in-house Python script `tns_eval.py`, which is available at https://github.com/bcgsc/rnaseq_utils. Only alignment segments of at least 150 nt in length, at least 90% sequence identity, and at most indels of 70-nt in length are considered. The ground truth transcript set is determined using the transcript identifiers in the simulated read names. Since not all known transcripts were simulated, the truth set is a subset of the ENSEMBL annotation. Any transcripts that are not in the truth set are designated as false-positives. If an assembled sequence aligns equally well to both a truth set transcript and a false-positive transcript, the assembled sequence would be assigned to the truth set instead of the false-positive. Any assembled sequences that have split-alignments to more than one transcript are designated as misassemblies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

For the spike-in datasets, transcriptome assemblies are aligned against the ERCC and SIRV sequences with minimap2. Since the ground truth transcript set is identical to the spike-in annotated transcripts, there are no false-positives. However, misassemblies are still detected as it was done for the simulated datasets.

Sitka spruce transcriptome analysis

All software command parameters are documented in **Supplementary Method 7**. The ONT cDNA reads were basecalled with Guppy v5.0.15. Since there are non-standard adapter and primer sequences, we used Porechop instead of Pychopper. We assembled the adapter-trimmed reads with RNA-Bloom2 v2.0.0 and short-read RNA-seq samples from previous work[29] were also included only for error correction of the ONT reads. The transcriptome completeness was benchmarked with BUSCO v5.3.2[31] and the embryophyte core gene set (odb10). The resulting RNA-Bloom2 assembly was supplied to PASA v2.5.2[34] for gene structure annotation, using minimap2 for transcriptome alignments against the draft genome. The figure for alternative splicing was generated with UpSetR[42]. The transcriptome assembly was annotated with EnTAP v0.10.8-beta[36] using TransDecoder v5.3.0[43] for protein sequence translation. Functional annotation was assigned based on Swiss-prot plant proteins[44], UniRef90 gene clusters[45], embryophyte orthologs from OrthoDB10[46] and high-quality proteins derived from NCBI RefSeq 99[47]. We performed the annotation of TPS, CYP and NAC through a BLASTP search against target spruce protein sequences reported previously[29,37,38], with minimum match of 95% identity and 90% query coverage.

SUPPLEMENTARY DATA

Simulated data used for benchmarking experiments is available at https://datadryad.org/stash/share/wtllL942eRij03GuEKvKTwEV6Kxl7Ia4Eus8lTu5TEM.

DATA AND SOFTWARE AVAILABILITY

The rebasecalled Nanopore sequencing data for the Sitka spruce cDNA sample has been deposited in the Sequence Read Archive (SRA) with run accession SRR19510936.

RNA-Bloom2 is implemented in Java and it is publicly available under GPLv3 license on GitHub at https://github.com/bcgsc/RNA-Bloom. The scripts we wrote to analyze our results are also publicly available on GitHub at https://github.com/bcgsc/rnaseq_utils.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Steven Jones and Dr. Marco Marra at Canada's Michael Smith Genome Sciences Centre (GSC) for graciously providing their nanopore sequencing data to test an earlier version of RNA-Bloom2, Dr. Kieran O'Neil and Vanessa Porter at GSC for evaluating the performance of an earlier version of RNA-Bloom2, Young Cheng at GSC for submitting the Sitka spruce nanopore sequencing data to the SRA, and the organizers of the LRGASP Consortium for making their sequencing data publicly available and sharing their evaluation results on an earlier version of RNA-Bloom2.

AUTHORS' CONTRIBUTIONS

KMN and IB conceived the study. KMN designed and developed RNA-Bloom2 under the supervision of IB. KMN, SH, and CY generated the simulated datasets and analyzed the LRGASP Consortium dataset. SH, RC, RLW, and IB advised on the benchmarking design. KMN, KKG, and RC analyzed the Sitka spruce dataset. All authors contributed to the interpretation of the results. KMN took the lead in writing the manuscript with input from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

FUNDING

This work was supported by Genome Canada and Genome British Columbia (**243FOR**); the National Institutes of Health (**2R01HG007182-04A1**); the Natural Sciences and Engineering Research Council of Canada (NSERC); and the Canadian Institutes of Health Research (CIHR). The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

REFERENCES

1. Moreno-Santillán, D. D., Machain-Williams, C., Hernández-Montes, G. & Ortega, J. De Novo Transcriptome Assembly and Functional Annotation in Five Species of Bats. *Sci. Rep.* **9**, 1–12 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

2. Jo, Y. *et al.* Integrated analyses using RNA-Seq data reveal viral genomes, single nucleotide variations, the phylogenetic relationship, and recombination for Apple stem grooving virus. *BMC Genomics* **17**, 1–12 (2016).
3. Mittal, V. K. & McDonald, J. F. De novo assembly and characterization of breast cancer transcriptomes identifies large numbers of novel fusion-gene transcripts of potential functional significance. *BMC Med. Genomics* **10**, 1–20 (2017).
4. Schelhorn, S.-E. *et al.* Sensitive Detection of Viral Transcripts in Human Tumor Transcriptomes. *PLoS Comput. Biol.* **9**, e1003228 (2013).
5. Lau, C.-C. *et al.* Viral-Human Chimeric Transcript Predisposes Risk to Liver Cancer Development and Progression. *Cancer Cell* **25**, 335–349 (2014).
6. Xue, W. *et al.* L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**, 1–14 (2013).
7. Raghavan, V., Kraft, L., Mesny, F. & Rigerte, L. A simple guide to de novo transcriptome assembly and annotation. *Brief. Bioinform.* **23**, (2022).
8. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 1–16 (2020).
9. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 1–11 (2016).
10. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
11. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100 (2017).
12. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 1–13 (2019).
13. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1–12 (2020).
14. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2019).
15. de la Rubia, I. *et al.* RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *Genome Biol.* **23**, 1–21 (2022).
16. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. (2012) doi:10.48550/arXiv.1203.4802.
17. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
18. Durai, D. A. & Schulz, M. H. Improving in-silico normalization using read weights. *Sci. Rep.* **9**, 1–10 (2019).
19. Sahlin, K. Effective sequence similarity detection with strobemers. *Genome Res.* **31**, 2080–2094 (2021).
20. Nip, K. M. *et al.* RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res.* **30**, 1191–1200 (2020).
21. Pardo-Palacios, F. *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. (2021) doi:10.21203/rs.3.rs-777702/v1.
22. GitHub - nanoporetech/pychopper: A tool to identify, orient, trim and rescue full length

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

cDNA reads. *GitHub* https://github.com/nanoporetech/pychopper.
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
24. Paul, L. *et al.* SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. *bioRxiv* 080747 (2016) doi:10.1101/080747.
25. Hafezqorani, S. *et al.* Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *Gigascience* **9**, giaa061 (2020).
26. Gagalova, K. K. *et al.* Spruce giga-genomes: structurally similar yet distinctive with differentially expanding gene families and rapidly evolving genes. *Plant J.* (2022) doi:10.1111/tpj.15889.
27. Stival Sena, J. *et al.* Evolution of gene structure in the conifer Picea glauca: a comparative analysis of the impact of intron size. *BMC Plant Biol.* **14**, 1–16 (2014).
28. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217.e14 (2022).
29. Whitehill, J. G. A., Yuen, M. M. S. & Bohlmann, J. Constitutive and insect□induced transcriptomes of weevil□resistant and susceptible Sitka spruce. *Plant-Environment Interactions* vol. 2 137–147 (2021).
30. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3**, e000132 (2017).
31. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
33. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).
34. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
35. Jia, J. *et al.* An atlas of plant full-length RNA reveals tissue-specific and evolutionarily-conserved regulation of poly(A) tail length. *bioRxiv* 2022.01.21.477033 (2022) doi:10.1101/2022.01.21.477033.
36. Hart, A. J. *et al.* EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
37. Warren, R. L. *et al.* Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* **83**, 189–212 (2015).
38. Whitehill, J. G. A. *et al.* Functions of stone cells and oleoresin terpenes in the conifer defense syndrome. *New Phytol.* **221**, 1503–1517 (2019).
39. Kucuk, E. *et al.* Kollector: transcript-informed, targeted de novo assembly of gene loci. *Bioinformatics* **33**, 1782–1788 (2017).
40. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
41. lh3/seqtk. *GitHub* https://github.com/lh3/seqtk.
42. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.07.503110; this version posted August 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

43. GitHub - TransDecoder/TransDecoder: TransDecoder source. *GitHub* https://github.com/TransDecoder/TransDecoder.
44. The UniProt Consortium *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).
45. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014).
46. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2018).
47. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2011).