Genome Biology

RESEARCH

Open Access

A comprehensive evaluation of long-read de novo transcriptome assembly

Feng Yan[1,2*], Pedro L. Baldoni[1,2], James Lancaster[1,2,3], Matthew E. Ritchie[1,2], Mathew G. Lewsey[3,4,5], Quentin Gouil[1,2,3,6,7] and Nadia M. Davidson[1,2*]

*Correspondence:
yan.a@wehi.edu.au;
davidson.n@wehi.edu.au

[1] The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
[2] Department of Medical Biology, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Parkville, VIC 3010, Australia
[3] La Trobe Institute for Sustainable Agriculture and Food, Department of Ecological, Plant and Animal Sciences, La Trobe University, Bundoora, VIC 3086, Australia
[4] La Trobe Institute for Molecular Sciences, La Trobe University, Bundoora, VIC 3086, Australia
[5] Australian Research Council Centre of Excellence in Plants for Space, La Trobe University, Bundoora, VIC 3086, Australia
[6] Olivia Newton-John Cancer Research Institute, Heidelberg, VIC 3084, Australia
[7] School of Cancer Medicine, La Trobe University, Heidelberg, VIC 3084, Australia

Abstract

Introduction: Recently, de novo transcriptome assembly methods have been developed to utilise long-read data in cases where a reference genome is unavailable, such as in non-model organisms. Despite the potential of these tools, there remains a lack of benchmarking and established protocols for optimal reference-free, long-read transcriptome assembly and differential expression analysis.

Results: Here, we evaluate the long-read de novo transcriptome assembly tools, RATTLE, RNA-Bloom2 and isONform, and compare their performance to one of the leading short-read assemblers, Trinity. We assess various metrics across a range of datasets, which include simulated data and spike-in sequin transcripts, where ground truth is known, and real data from human and pea (*Pisum sativum*) samples, using a reference-based approach to define truth. To represent contemporary analysis scenarios, the datasets cover depths from 6 to 60 million reads, Oxford Nanopore Technologies (ONT) cDNA, ONT direct RNA and Pacific Biosciences (PacBio) 10× single-cell sequencing. Critically, we assess the downstream impact of assembly choice on the detection of differential gene and transcript expression.

Conclusions: Our results confirm that long reads generate longer assembled transcripts than short-reads for reference-free analysis, though limitations remain compared to reference-guided approaches, and suggest scope for improved accuracy and reduced redundancy. Of the de novo pipelines, RNA-Bloom2, coupled with Corset for transcript clustering, was the best performing in terms of both accuracy and computational efficiency. Our findings offer guidance when selecting the most effective strategy for long-read differential expression analysis, when a high-quality reference genome is unavailable.

Keywords: Long reads, Transcriptome assembly, Reference-free, Differential expression, Non-model organisms

Background

Understanding which genes and transcripts are present in an organism is essential for analysing RNA sequencing (RNA-seq) data and enables a variety of downstream analyses, such as differential expression, alternative splicing, and the identification of RNA variants. Transcriptome assembly is a key method to achieve this, particularly when reference genomes and gene annotations are unavailable. Transcriptome assembly reconstructs the sequences of expressed transcripts directly from RNA-seq data and is performed using two primary approaches: reference-guided and de novo. Reference-guided assembly relies on a reference genome to infer genes and their exon structures based on the positions of mapped reads [1]. This method is highly effective, often enabling the identification of novel transcripts even in well-characterised species like humans and mice [2]. However, its success relies on the availability and quality of a reference genome. In contrast, de novo assembly offers a robust solution for non-model organisms that lack reference genomes. By reconstructing transcriptomes directly from RNA-seq reads, this approach bypasses the need for prior genomic information. It allows gene expression and transcript diversity to be explored in a broad array of species, significantly extending the applicability of RNA-seq.

To achieve de novo assembly of massively parallel short-read sequencing, numerous tools and pipelines have been developed such as Trinity, RNA-Bloom, SOAPdenovo-Trans, rnaSPAdes and Oases [3–7]. Collectively they have facilitated the analyses of thousands of transcriptomes [8, 9], expanding our knowledge of the diversity and complexity of life beyond model species. Although generally applied to non-model organisms [10–12], de novo assembly has also proven beneficial to the analysis of cancer transcriptomes, in particular in the detection of gene rearrangements [13–16].

However, despite the success of short-read de novo transcriptome assembly, a fundamental limitation remains in that long-range relationships across exons and exon junctions can not be directly measured and must be inferred. Due to fragmentation, splicing events at one end of a gene are decoupled from those at the other end, and the true exon structure of the gene is obscured. This limitation is solved by long-read (or third-generation) sequencing. Both Pacific BioSciences (PacBio) and Oxford Nanopore Technologies (ONT) offer protocols to sequence full-length transcripts without fragmentation and ONT is even capable of sequencing native RNA with RNA modifications [17]. Both technologies offer the promise of higher-quality assembled transcriptomes. When combined with single-cell technologies, they enable isoform-resolved single-cell transcriptome analysis [18].

To leverage these advances, new reference-guided assembly methods have been built for long reads including Bambu, StringTie2, IsoQuant and FLAIR [19–22]. In contrast, few assembly methods which work without a reference genome or annotation have been developed. The first efforts of such kind were CARNAC-LR and isONclust which cluster reads into genes or gene families [23, 24]. However, these two methods do not generate a set of non-redundant transcript sequences, restricting their application. RATTLE was the first developed to address the transcript assembly problem, which it solved using a three-step pipeline. First, reads are clustered into gene and then transcript groups using a greedy *k*-mer based approach for efficiency. Next, errors are corrected by consensus calling the reads within each transcript cluster. A final polishing step refines the clusters

and reports the number of reads in each to give transcript quantification [25]. RATTLE was shown to improve clustering accuracy and provided comparable transcript abundance estimates to reference-guided tools when tested on SIRV spike-in transcripts. Following RATTLE, RNA-Bloom2 was developed. It builds upon RNA-Bloom, a successful short-read transcriptome assembler, and employs error-tolerant digital normalization to reduce the number of comparisons required for the subsequent steps of overlap assembly and polishing [26]. RNA-Bloom2 was reported to improve computational efficiency over RATTLE, and provide higher recall when tested on simulation and spike-in data. More recently, isONform was developed. It forms the final step in the isON-pipeline of isONclust, isONcorrect, and isONform, by using minimizer pairs from isONclust gene clusters for graph construction. Bubbles in the graph are iteratively popped to remove errors and the final output is a set of transcript sequences [27]. IsONform was reported to have higher recall than RATTLE on simulated and SIRV spike-in data, but was significantly slower to run and its performance was not compared against RNA-Bloom2.

Aside from de novo assemblers specifically developed for long-reads, there are also methods that support the hybrid assembly of both long and short read data. These tools generally fall into two main categories: (1) those that use short reads to error-correct long reads before assembling, such as RNA-Bloom2 hybrid mode, and (2) those that use short reads to build an assembly graph and then incorporate long reads for scaffolding gaps and resolving complex structures, such as rnaSPAdes and IDP-denovo [28, 29]. Table 1 provides a summary of the de novo assembly tools discussed (Table 1).

Table 1 Popular long read de novo assemblers. The key differences in algorithmic approach, input and output data between evaluated tools

Category	Tool	Core algorithmic approach	Input data supported	Primary output
Long-read de novo assembly	RNA-Bloom2	De novo assembly using Bloom filters and overlap graphs	Long-read RNA-seq (ONT, PacBio), FASTQ	Transcript sequences FASTA
	RATTLE	Reference-free clustering and consensus-based transcript reconstruction	Long-read RNA-seq (ONT, PacBio), FASTQ	Transcript sequences FASTQ, transcript clustering & expression estimates
	isONform (isONclust + isONcorrect)	Read clustering followed by error correction and consensus calling	Long-read RNA-seq (ONT, PacBio), FASTQ	Transcript sequences FASTA, transcript clustering
Short-read de novo assembly	Trinity	De Bruijn graph–based de novo assembly	Short-read RNA-seq, FASTQ	Transcript contigs FASTA, transcript clustering
Hybrid de novo assembly	rnaSPAdes (hybrid)	De Bruijn graph assembly with RNA-seq–specific heuristics, long reads are used to resolve complex structures	Long-read (ONT, PacBio) and short-read RNA-seq, FASTQ	Transcript contigs FASTA, transcript clustering
	RNA-Bloom2 (hybrid)	Short reads used to aid long read error correction prior to long-read assembly	Long-read (ONT, PacBio) and short-read RNA-seq, FASTQ	Transcript sequences FASTA

Independent benchmarking efforts are critical to the advancement of transcriptome assembly-based protocols. For short-read data, they have exposed the strengths and weaknesses of different short-read approaches [8], and informed data analysis workflows [9]. Similar efforts are now underway for long-read assembly tools, but these have largely focused on reference-guided approaches. Dong et al. benchmarked six tools using ONT in-silico mixtures of two cells lines with synthetic RNA spike-ins (sequins), and found Bambu had the highest precision and recall of known sequin transcripts and performed well on low sequencing depths [30, 31]. Similar results were observed by Su et al. who benchmarked 9 tools using simulated PacBio and ONT data and sequins from ONT and found that Bambu, IsoQuant and StringTie2 overall performed best in isoform detection [32]. The Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) presented a comprehensive assessment of 14 tools across three species with various library and sequencing protocols (both long-read and short-read) on three key tasks: isoform detection with high-quality reference genome and gene annotation, isoform quantification, and isoform detection given a genome but no gene annotation [33]. However, only one long-read de novo assembler (RNA-Bloom2) was included. It demonstrated high recall but low precision at recovering SIRV spike-in transcripts. Only one independent study has evaluated de novo methods thoroughly [34]. Sagniez et al. compared RATTLE, RNA-Bloom2 and isONclust assemblers using deeply sequenced synthetic RNA (SIRV and sequins) with ONT R9 and R10 chemistries. As expected, they found that transcript assembly was most accurate when both a reference genome and gene annotation were used, but de novo methods did outperform some reference-guided approaches. However, this study only evaluated the assembly of synthetic transcripts, and used datasets which are a fraction of the depth (O(1 million reads)) and complexity (69 SIRV and 160 sequin transcripts) seen for real transcriptomic projects (O(10 million reads), > 20 k transcripts). Moreover, the impact on downstream differential analysis, being one of the most common applications of RNA-Seq, was not examined.

To overcome the lack of independent benchmarking, we have comprehensively evaluated the performance of current de novo assembly methods for long-read sequencing data. As the performance of de novo methods varies with transcriptome complexity, sequencing error rate and depth, we tested each method on several datasets: simulated ONT cDNA data, ONT PCR-cDNA and ONT direct RNA data from multiple cancer cell lines with additional sequin spike-ins, PacBio Kinnex single-cell RNA-seq of human peripheral blood mononuclear cells (PBMCs), and in a plant species (PCR cDNA data from *Pisum sativum*). For each, we examined the assembly accuracy and completeness, as well as the software's computational time and memory. As the end goal of many de novo analyses is the identification of differentially expressed genes and transcripts, we assessed the impact that the choice of assembly tool, transcript to gene clustering and quantification makes on the detection of differential expression. We demonstrated that de novo methods can recover the majority of differentially expressed genes found with reference-guided approaches and some at the transcript-level. Based on our simulation results, long-read approaches are currently on par with short reads for quantification and differential expression, yet provide more complete transcript sequences. We also assessed hybrid short and

long-read assembly pipelines and found that they did not offer an advantage over using a single sequencing protocol. Critically, these results provide the foundational guidelines for the optimal workflow to support downstream long-read differential analysis in species lacking a completed reference genome and gene annotation.

Results

Study design

To evaluate the performance of long-read de novo assembly tools, we selected three recently developed methods: isONform, RATTLE and RNA-Bloom2. As a gold standard for short-read assembly, we included Trinity, which was chosen based on its strong performance in several independent evaluations [4, 8]. In addition, all datasets were assembled and expression quantified with the reference-based tool Bambu [19]. For datasets without available ground truth, i.e., all except for the simulated and spike-in datasets, Bambu was used to define the truth set. Bambu was chosen due to its high accuracy in independent benchmarking studies and thus represents the best-case scenario for assessing the potential of reference-free methods [30, 32].

For downstream differential expression analysis of de novo assembled transcriptomes, we employed Corset [35] to group transcripts into gene clusters, minimap2 [36] for aligning reads back to the assembled transcriptome. Salmon [37] was used for transcript

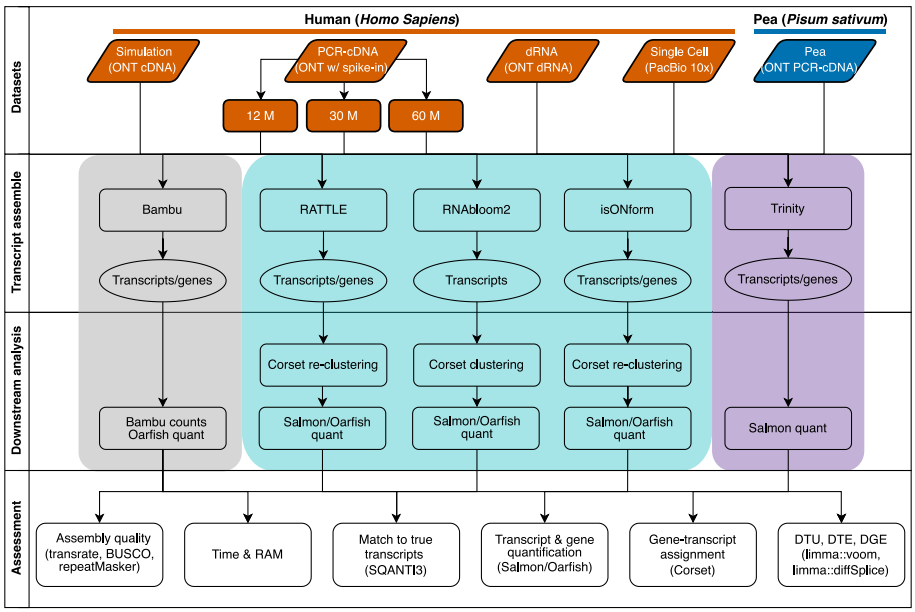

Fig. 1 Schematic of the benchmarking study design. Five datasets were utilized, including simulated unstranded cDNA ONT data (simulation), human stranded PCR-cDNA ONT data (PCR-cDNA) with sequin spike-ins, human direct RNA ONT data (dRNA), human single-cell RNA-seq generated with PacBio Kinnex (single-cell), and stranded PCR-cDNA ONT data from *Pisum sativum* (Pea). The human PCR-cDNA was down-sampled to three sequencing depths (12, 30 and 60 million reads), and included sequin spike-ins (2% of total reads). Each dataset was assembled using three long-read reference free methods: RATTLE, RNA-Bloom2, isONform (teal). Trinity (purple) was applied to matched short-read data when available. A reference-guided approach, Bambu was used to define ground truth (grey). The evaluation focused on both assembly quality and the results of differential gene and transcript expression analysis. Tools used throughout the assembly pipeline and evaluation are also indicated

Table 2 The datasets used for benchmarking. The DE analysis was performed comparing the 2 conditions listed, and the number of true DE genes, DE transcripts and genes undergoing DTU

Dataset	Protocol	Total reads (million)	Error rate (%)	Mean read length (bp) (range)	How truth was defined	True #DGE #DTE #DTU-genes	Source
Simulation: control vs perturbed	ONT cDNA (R9.4)	6	7.65	1085 (62–44,010)	Simulation	2000, 6933, 2000	New data generated with SQANTI-SIM
PCR-cDNA:Human lung cancer cell lines HCC827 vs NCI-H1975 + spike-ins	ONT PCR-cDNA (SQK-PCS109, R9.4)	12	6.81	685 (50–71138)	Bambu	5328, 5362, 635	GEO
		30			Bambu	7506, 8502, 1182	
		60			Bambu	9094, 11,098, 1726	
		2% of total reads			Sequin spike-in	44, 95, 28	
dRNA: SG-NEx human cancer cell lines MCF7 vs A549	ONT direct RNA (SQK-RNA002, R9.4)	11.6	15.9	1060 (50–16218)	Bambu	6280, 7231, 1637	SG-NEx AWS
Single-cell: Human PBMC myeloid vs lymphoid cells	PacBio Kinnex 10x 3′ kit	15 (assembly) 555 (quantification)	0.44	835 (50–7173)	Bambu	7352, 23,711, 7363	https://www.pacb.com/connect/datasets/#RNA-datasets
Pea: (*Pisum sativum*) Thomas Laxton vs Daisy cultivars	ONT PCR-cDNA (SQK-PCB109, R9.4)	16.5	7.75	649 (50–11561)	Bambu	254, 403, 87	In-house new data

abundance quantification with bootstrapping (Fig. 1) for bulk samples, and Oarfish [38] for the single-cell data. Transcript and gene-level differential analysis was then conducted using the Bioconductor package limma [39–42].

We evaluated all methods across five independent datasets, representing various sequencing technologies, read depths, and organisms (Table 2). The datasets included simulated unstranded cDNA ONT data, real PCR-cDNA ONT sequencing data from human lung cancer cell lines with sequin spike-ins (bioinformatically restranded), direct RNA ONT sequencing data from the SG-NEx project, 10x Genomics single-cell RNA-seq generated from the PacBio Kinnex protocol, and PCR-cDNA ONT sequencing data from pea (*Pisum sativum)* cultivars, a non-model species [30, 43, 44]. To investigate the impact of coverage and library size, we downsampled the lung cancer cell line dataset at three different read depths. For all datasets except the single-cell data, we also had matched short-reads available which we downsampled to the same number of nucleotides as the long-read data, and processed with Trinity (Additional file 1: Table S1).

Aside from the pea cultivars and single-cell data, each dataset consisted of two conditions, each in triplicate, with reads pooled for assembly. For the pea, samples from four conditions were pooled for assembly, but only two conditions contrasted for downstream differential analysis. For the single-cell dataset, we first performed gene expression clustering and compared the two major clusters, which corresponded to myeloid and lymphoid cells. For each dataset, we assessed: (1) computational efficiency, (2) the quality of assembled transcripts, based on transcript length, recall of reference transcripts and BUSCO gene recovery, (3) transcript and gene quantification accuracy, particularly focusing on clustering transcripts into gene clusters, and (4) the precision and recall in detecting differentially expressed genes (DGE), differentially expressed transcripts (DTE), and differential transcript usage (DTU), between the two conditions.

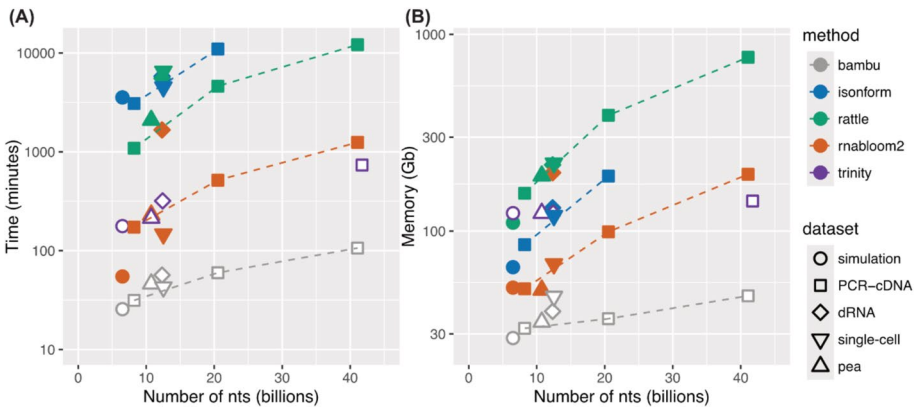

Fig. 2 Computational resources required for long read de novo assembly. **A** Time and (**B**) memory usage of each assembler as a function of the number of total nucleotides for each dataset. The change in resource usage between the three downsampled depths from the PCR-cDNA dataset are indicated with dashed lines. The reference-guided assembler, bambu, is included for comparison and highlights the significant computational resources required for reference-free assembly

Computational efficiency

Given the substantial time and memory requirements for de novo assembly, we first assessed the computational performance of each tool. All assemblers were run on a high-performance computing cluster where we requested 48 cores and a maximum of 1 TB of RAM. RATTLE and isONform demonstrated significantly longer assembly times and higher memory usage compared to RNA-Bloom2 (Fig. 2, Additional file 1: Table S2). Specifically, RATTLE required between 18 h and 8 days to finish, utilising 110 to 764 GB of RAM, depending on the dataset. While isONform used less memory (65–190 GB), it was slower, taking 51 h to 7 days to complete assembly and failed to finish within two weeks on the 60 million reads PCR-cDNA and the pea datasets, thus it was not evaluated on these datasets.

In contrast, RNA-Bloom2 was notably faster and more memory-efficient, completing assemblies in 1 to 27 h while consuming 50 to 198 GB of RAM. However, its performance on the dRNA dataset was an exception, where it ran 9 times slower and required 4 times more memory compared to cDNA at a similar read depth. We hypothesise that the increased error rate in dRNA sequencing reduced the effectiveness of digital normalisation, which is a key step in RNA-Bloom2 to reduce computational burden (60% of dRNA reads remaining compared to only 12.3% of PCR-cDNA reads after digital normalisation). Trinity, the short-read assembler, which also employs digital normalisation, showed comparable runtime to RNA-Bloom2 across the datasets and scaled consistently with increasing read depth. Notably, Trinity maintained relatively stable memory usage across different read depths.

Overall, we found RNA-Bloom2 to be the most computationally efficient long-read de novo assembler, although its performance may degrade on datasets with higher basecalling error rates, such as direct RNA. All de novo methods demanded substantial computational resources, highlighting a current limitation in assembling large datasets.

Assembly quality

De novo assemblies from RATTLE, RNA-Bloom2 and isONform, together with Trinity were next assessed for their quality, by comparing characteristics of the assembled transcriptome, such as the number, length and recall of reference transcripts (Fig. 3). SQANTI3 was used to link assembled transcripts to reference transcripts using splice junctions [45]. To assess whether complete transcripts were being assembled, we used the Conditional Reciprocal Best BLAST (CRBB) approach, which gives the percentage of bases recovered for each reference transcript [46, 47]. As pea is likely to have incomplete gene annotation, we also used BUSCO genes to assess assembly completeness. To set an upper-limit on the best performance expected, we also calculated metrics for the reference genome-guided approach, Bambu, which was filtered for only expressed transcripts (Methods).

The number of transcripts assembled by the de novo methods varied widely by assembler and dataset (Fig. 3A, Additional file 2: Fig. S1A, Additional file 1: Table S3). In the simulation, RNA-Bloom2 and RATTLE reported a similar number of transcripts (~40,000) and genes (~17,000), which were close to the true number simulated (40,509 transcripts and 18,145 genes). However, isONform assembled only 4849 transcripts and 444 genes. Combined with similar poor performance on the dRNA dataset, this

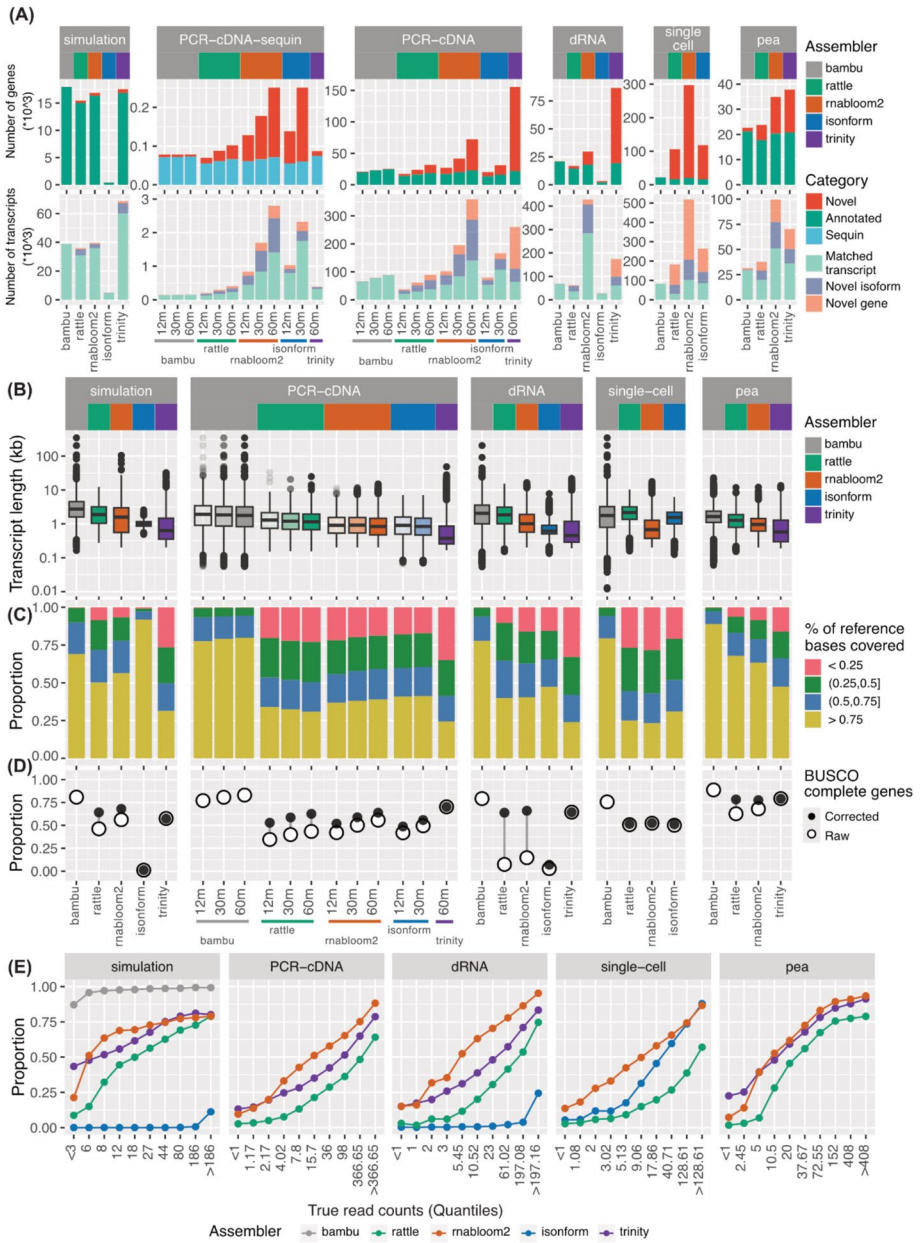

Fig. 3 Assessment of assembled transcriptome quality. **A** Number of novel and annotated genes and transcripts in each assembled transcriptome. Transcripts were further divided into matched transcript (complete splice match and incomplete splice match), novel isoform (novel in catalogue and novel not in catalogue) and novel gene (intergenic, genic, intronic, antisense, and fusion). **B** Length distribution of assembled transcriptomes. **C** The proportion of bases recovered for each reference transcript using the Conditional Reciprocal Best BLAST (CRBB) approach. **D** Proportion of BUSCO complete genes (both single copy and duplicated) assembled before and after sequence polishing. Assembled transcripts were corrected using the reference genome to remove errors and assess protein sequence predictions. **E** Proportion of reference transcripts that were assembled binned by true expression. True expression was taken from simulated or Bambu counts, and binned into 10% quantiles. X-axis labels show the corresponding read count range for each quantile. For PCR-cDNA data, only the 60 million data was shown

suggests that isONform with default parameters is not robust to variations in the data, such as error rate (dRNA), strandedness or read length (simulation), which may require adjusting isONform's parameters such as k-mer and window size. Trinity, applied on the matched short-read simulation, assembled a similar number of genes as other tools, but over 60,000 transcripts. Although this was significantly more than the number of simulated transcripts, the majority were found to have matching splice junctions with the reference gene annotation. Therefore the larger number of transcripts reported by Trinity is likely a result of a fragmented assembly (64% incomplete splice match), rather than errors. This is further supported by Trinity having the shortest transcript lengths (Fig. 3B), and fewest complete transcripts (Fig. 3C, Additional file 2: Fig. S1B) across all datasets, and highlights the benefit of long reads in reconstructing the full length of transcripts.

On the real datasets, we found that the reported number of transcripts was highest from RNA-Bloom2 and lowest in RATTLE, excluding isONform's failure on dRNA. This high number for RNA-Bloom2 can partially be explained by a high level of transcript redundancy, with multiple assembled contigs matching the same reference transcript (Additional file 2: Fig. S1C). As expected, we also observed that the number of transcripts increased with read depth in the PCR-cDNA dataset. Although new reference transcripts were assembled at higher depths, so too were novel unannotated sequences. The unannotated novel genes increased 3 and 4 folds in RATTLE and RNA-Bloom2 respectively from 12 to 60 million reads, and composed the majority of the Trinity assembly. Interestingly, these novel sequences also coincided with a drop in the mean transcript length (Fig. 3A, B). We also observed that the assemblies were enriched for simple repeats in the ONT cDNA datasets (PCR-cDNA and pea), and SINE and LINE repeats in the PacBio data (single-cell), hinting at artifacts in either assembly or sequencing protocol, such as representation of intronic regions resulted from internal priming, which are rich in repetitive sequence (Additional file 2: Fig. S1D, Additional file 1: Table S4) [48].

Although we could not definitively know if the novel transcripts were real or false positives, the rates of false positives from the simulation and sequin chromosome (chrIS) in the PCR-cDNA data provided evidence of a high rate of assembly error and redundancy. In the simulation, up to 14% of transcripts and 7% of genes were found to be false positives (Fig. 3A, Additional file 2: Fig. S1A). When focusing on assembled transcripts on chrIS, we observed that for the RNA-Bloom2 assembly on the deepest PCR-cDNA dataset, only 71 out of 251 assembled genes matched to sequin genes, and 1412 out of 2804 assembled transcripts matched to known sequin transcripts (Fig. 3A). In addition to false transcripts, we also observed insertion and deletion type errors being incorporated into the sequence of assembled transcripts from ONT data. This impacted the predicted protein translation and hence BUSCO completeness, and was particularly evident in the dRNA data, where up to 50% of conserved BUSCO protein-coding genes expressed were missed (Fig. 3D, Additional file 1: Table S5). These errors were rescued if a reference genome was provided for polishing the long-read transcriptome assemblies. This issue was not seen for the PacBio or short-read data and suggests that error rate is an important consideration when designing experiments requiring reference-free transcriptome analysis.

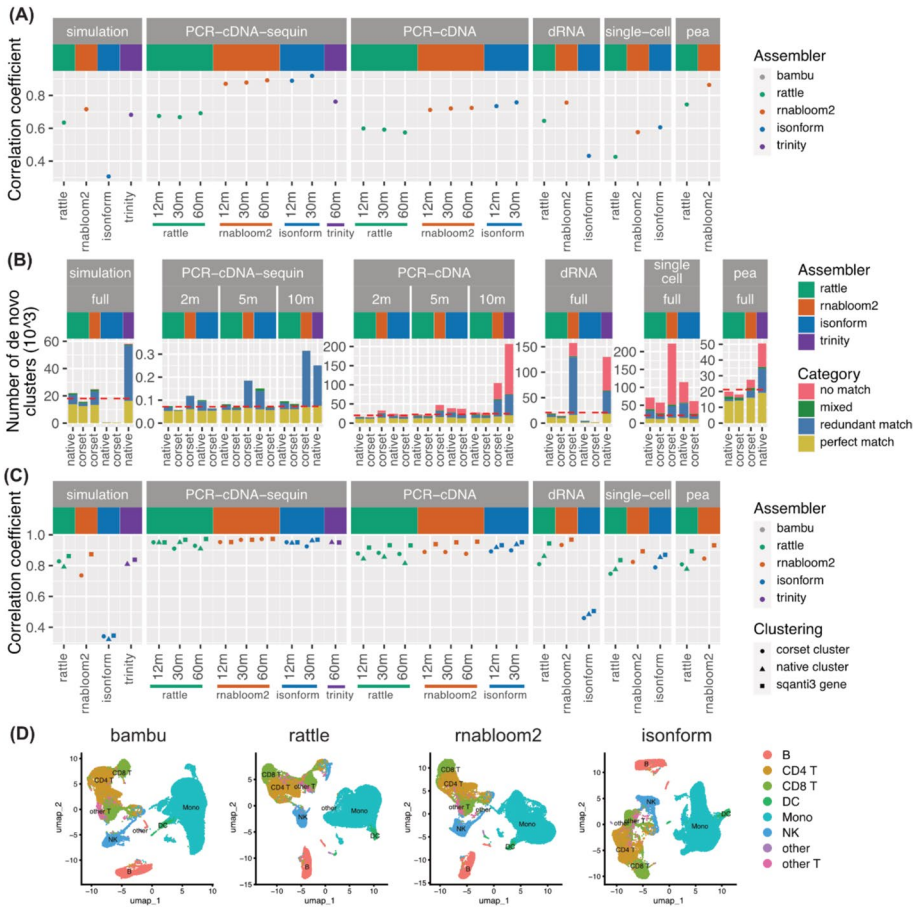

Fig. 4 Accuracy of transcript and gene abundance estimates. **A** Pearson correlation coefficients between estimated and true transcripts expression values (log2(count + 1)). **B** The number of gene-level clusters generated for each assembly. Match indicates when all transcripts within the cluster correspond to the same reference gene. Where one reference gene matched multiple clusters, the cluster with the most transcripts was classed as a 'perfect match' and the rest as 'redundant match'. Clusters with transcripts from multiple reference genes were classed as 'mixed'. Clusters with only novel transcripts were classed as 'no match'. The red dashed line denotes the number of reference gene-level clusters generated by Bambu. **C** Pearson correlation of true gene expression (log2(count + 1)) to Corset, native or SQANTI3 cluster expression. SQANTI3 clustering is an optimal scenario where transcripts are grouped based on their true reference genes, and it is included as an upper limit. **D** UMAP generated using gene-level counts from reference (bambu) and de novo assemblies and colored based on the cell type annotation obtained from a reference analysis

As a general trend, RNA-Bloom2 reported more false positives or unannotated transcripts than RATTLE, but also had higher recall of true positives (simulated, spike-in transcripts or reference) (Fig. 3A, E), consistent with prior studies [26, 33]. RATTLE's more conservative behaviour could be explained by requiring a minimum of 5 reads for consensus calling. However, we found that RATTLE missed transcripts even at higher coverage levels and increasing sequencing depth only improved recovery for lowly expressed genes and transcripts (Fig. 3E, Additional file 2: Fig. S1E-F). Despite these differences, RNA-Bloom2 and RATTLE had highly similar recall at gene-level, as indicated by the number of recovered reference genes (Fig. 3A), transcript base-coverage (Fig. 3C), and BUSCO genes (Fig. 3D).

Transcript and gene abundance estimates

Next, we examined how the differences between assemblies impacted transcript and gene-level expression estimates. To compare across different assemblies, we estimated abundances by mapping reads to assembled transcriptomes with minimap2 and applied Salmon or Oarfish for quantification (Methods). For each reference transcript, we identified the best match in the assembly, where the best match was defined as the assembled transcript with the highest expression estimate that was a SQANTI3 splice match. We then assessed the correlation between the estimated and 'true' expression, with truth derived from simulated counts, sequin concentration or reference-based quantification.

RNA-Bloom2 transcript counts were consistently one of the most correlated with truth (Fig. 4A, Additional file 2: Fig. S2A, Additional file 1: Table S6). IsONform performed the best on the 12 million and 30 million PCR-cDNA data and single-cell data, demonstrating that when it runs successfully, it can exceed the performance of other methods. Trinity performed equally well as RNA-Bloom2 on the simulation, but trailed it on the spike-in dataset. We did not assess Trinity on the real datasets, as the truth was defined from long reads, giving the long-read methods an advantage.

In order to assess gene-level expression estimates, transcripts must be assigned into gene groups, a step known as clustering. While RATTLE, isONform and Trinity output gene-level clusters, RNA-Bloom2 does not. Therefore, we tested the performance of Corset, which is a clustering method built for short-read de novo assembled transcriptomes. Corset clustered the assembled transcripts into genes based on all-vs-all sequence alignments from minimap2. Corset and each assembler's native clustering were then assessed (Fig. 4B, Additional file 2: Figs. S2B-C, Additional file 1: Table S7).

RNA-Bloom2 and Trinity tended to have the highest number of clusters, and a large fraction of these were redundant, meaning there was more than one cluster matching the same reference gene (Fig. 4B, Additional file 2: Fig. S2C, Additional file 1: Table S8). For RNA-Bloom2, we found redundant clusters often contained shortened assembled transcripts, with a higher number of sequencing errors, which impacted the all-vs-all alignment (Additional file 2: Fig. S2D). Both RNA-Bloom2 and Trinity also had many clusters that did not match any reference gene. This is likely a consequence of more novel transcripts being assembled, as seen in the assembly quality assessment (Fig. 3A).

We next aggregated the transcript-level counts for each cluster, to obtain gene-level counts, and compared these values against the 'truth'. Each cluster was assigned a reference gene label based on the reference gene of the majority of its constituent transcripts. When more than one cluster matched a reference gene, we examined the expression of the cluster with the highest counts. When a reference gene had no matching cluster, its count was taken as zero. To assess the impact of clustering, we also looked at the expression correlation for perfect clustering (SQANTI3 gene), being the scenario that all assembled transcripts from the same gene are assigned to the same cluster based on SQANTI3 results.

Most de novo assemblies gave similar correlations at gene-level (Fig. 4C, Additional file 1: Table S6). For simulation, RATTLE was slightly better than RNA-Bloom2, while for other datasets, RNA-Bloom2 with Corset had a similar or better performance than RATTLE. Similar to the transcript-level comparison, isONform on the 12 million PCR-cDNA, 30 million PCR-cDNA and single-cell datasets, gave the highest correlation.

However, increasing read depth had very little impact on the correlation at both transcript and gene levels.

Overall, Corset was found to be an effective tool for clustering transcripts into genes, with a similar, and in some instances, better performance than native clustering. Based on these results, Corset was used in combination with RNA-Bloom2 to assess downstream analysis steps. The assembler's own clustering was used for RATTLE and isONform.

Finally, we also examined the impact of abundance estimation on single-cell gene-expression clustering (which is distinct from the transcript clustering examined previously), dimensionality reduction and visualisation. Cell types obtained from a reference analyses were annotated onto the reference-free UMAPs to indicate true groupings. We found that reference-free gene-expression (Fig. 4D) and transcript-expression (Additional file 2: Fig. S3A) were able to recapitulate the clusters seen from Bambu quantification for all assemblers. These results suggest that long-read single-cell analysis can be accurately performed in a reference-free manner, enabling transcriptome analysis at single-cell resolution for non-model species that lack a reference genome.

Differential expression

We next performed differential analysis using the transcript and gene counts as previously described, and examined each pipeline's ability to recover true differential gene expression (DGE), differential transcript expression (DTE), and differential transcript usage (DTU) [39, 40]. True changes were defined using either known simulated or sequin spike-in transcripts and genes when available. For other datasets, we defined truth using a reference-based approach (Bambu) with an FDR cut-off of 0.05 (see Table 1). Transcripts in the de novo assembly were assigned to reference transcripts using SQANTI3 and clusters were assigned to reference genes based on majority vote, as described earlier. As benchmarking of differential testing methods has been performed previously [30] we did not compare testing methods, but used methods amongst the top performing from prior evaluation, limma::voom and limma::diffSplice [42] for differential expression and transcript usage analysis respectively.

On the simulated data, all methods performed well in the task of identifying DGE with the true DE genes ranked higher than other clusters (Fig. 5A). All methods were close to the accuracy of the reference-guided tool Bambu. Trinity performed marginally worse than the long-read pipelines. When DGE was evaluated using sequin spike-in genes, Trinity and RNA-Bloom2 marginally outperformed others (Fig. 5B). On the real datasets, RNA-Bloom2 was consistently a top performer (Fig. 5C-F). However when isONform ran successfully, its performance was similar or exceeded that of RNA-Bloom2. For example, on the PCR-cDNA data, isONform with 30 million reads identified true positives with the same accuracy as RNA-Bloom2 with 60 million reads (Fig. 5C). As RNA-Bloom2 produces a highly redundant transcriptome, we also assessed true positives as a function of all top ranked transcripts including duplicated transcripts (Additional file 2: Figs. S3B-G). Aside from PCR-cDNA DGE, where RATTLE was marginally better, RNA-Bloom2 remained the best performing overall.

For DTE, amongst long-read de novo assembly methods, RNA-Bloom2 showed the best performance across datasets. Interestingly, the short-read assembler, Trinity

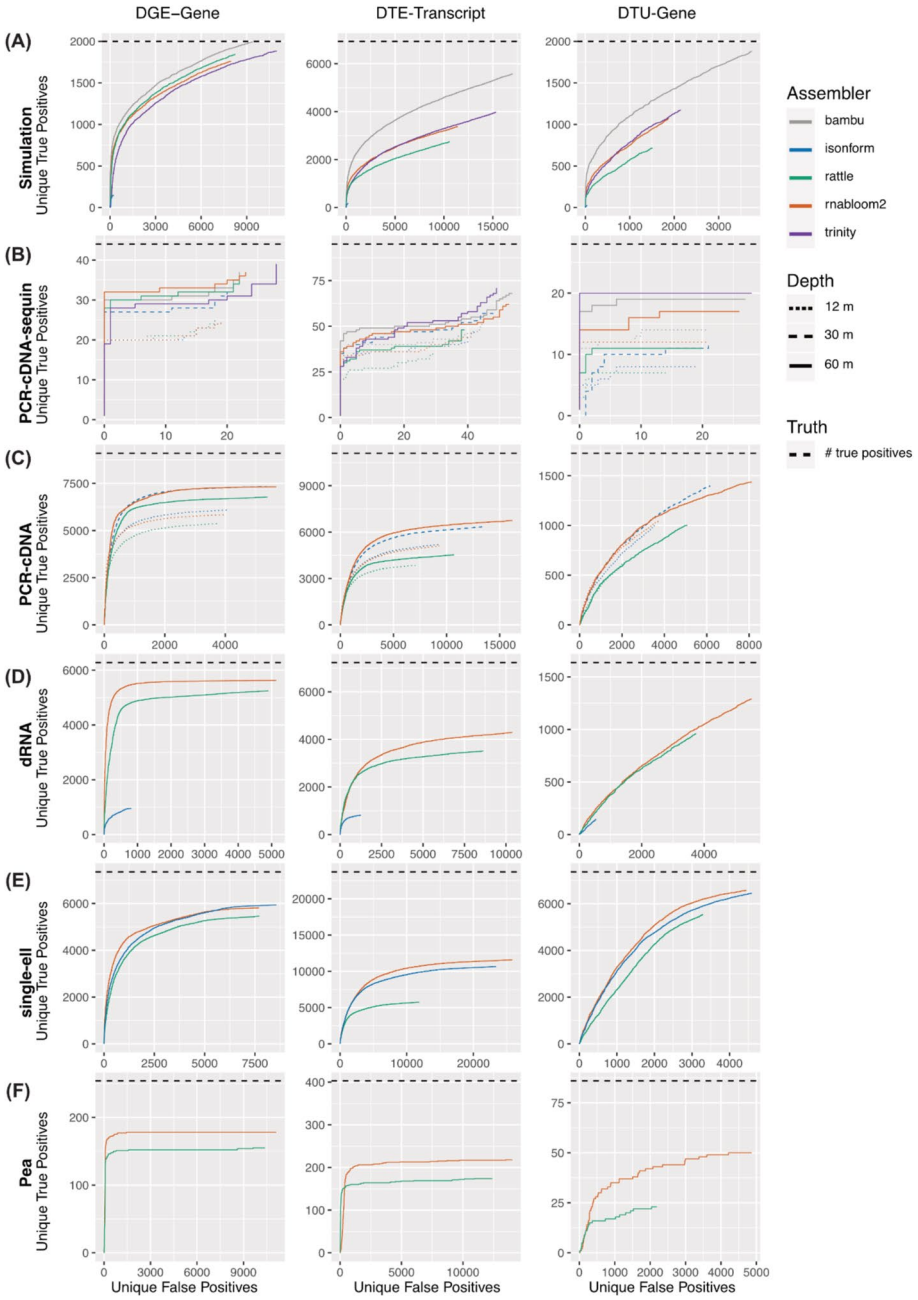

Fig. 5 ROC curves for differential analysis. DGE, DTE and DTU analysis was performed for (**A**) Simulation, (**B**) PCR-cDNA sequin, (**C**) PCR-cDNA, (**D**) dRNA, (**E**) single-cell, and (**F**) pea dataset. Unique false positives are shown against unique true positives. When multiple clusters match the same reference gene, the cluster with the lowest FDR is retained, hence 'unique' true positives and 'unique' false positives. Truth is defined from simulation, sequin or Bambu differential results, and the total number of true positives is indicated by the horizontal dashed line (black). For panels (**B**, **C**), we show the deepest and shallowest read depths of the PCR-cDNA dataset for which an assembler ran successfully, specifically 60 million reads for RNA-Bloom2 and RATTLE (solid line), 30 million reads for isONform (lines with longer dashes), and 12 million reads for all assemblers (lines with short dashes)

performed well on both simulation and sequin data. For differential transcript usage analysis (DTU), genes were ranked by the minimum Simes-adjusted p-values from their transcripts p-values. DTU results showed a similar trend to DGE and DTE, with RNA-Bloom2 being the best performing of the long-read de novo assembly methods in 5 out of 6 datasets. We also noticed that for all methods, increasing the read depth in the PCR-cDNA data increased the number of true positives among top ranked results for DGE, DTE and DTU as expected. However, surprisingly, for DTE and DTU, the choice of assembler had a larger impact on performance than quadrupling the sequencing depth (Fig. 5B, C).

Interestingly, in real data, all de novo methods missed a significant proportion of the differential transcripts reported by Bambu (black dashed line Fig. 5C-F). Assuming the Bambu results are true positives, this suggests a significant loss in performance when a reference genome is not used and leaves room for further optimisation of de novo methods. However, of the de novo assembly tools currently available, RNA-Bloom2 was found to have the best overall performance at both gene and transcript level.

De novo assembly captures novel transcripts

De novo transcriptome assembly enables the discovery of biologically significant transcripts, including those absent from the reference genome and gene annotation, making it an essential approach when the reference genome or gene annotation is incomplete or unavailable. To highlight this advantage, we investigated the differentially expressed de novo assembled transcripts with novel sequences. We considered any transcript to be 'novel' if (1) it did not align to reference genome, or (2) its primary alignment to the reference genome contained more than 400 consecutive unaligned bases, or (3) more than 100 or 200 accumulative mismatches, for cDNA and dRNA respectively. Since RNA-Bloom2 performed well on all datasets, we investigated the novel results in its transcriptome assemblies.

In the 60 million PCR-cDNA and dRNA datasets, we identified 2562 and 2307 transcripts respectively which were both differentially expressed and novel compared to the reference transcriptome (Fig. 6A, B). A number of these transcripts are likely to contain errors, in particular, some were found to have substantial mismatches likely due to sequencing errors being retained during assembly. However, amongst the top 10 ranked differentially expressed novel transcripts in PCR-cDNA and dRNA, there were three consistent with genomic rearrangement found in the Cancer Cell Line Encyclopedia (CCLE) (Additional file 1: Tables S9-10) [49]. One of these fuses the *RPL7* gene to an intergenic region on chromosome 8 in the HCC827 lung cancer cell line (Fig. 6D). The other two were found in the MCF7 breast cancer cell line. One is a fusion of *ATXN7* to an intergenic region on chromosome 1 (Fig. 6E), and the other is the well-reported fusion *BCAS4-BCAS3* [50]. To assess whether other fusion genes were seen amongst the novel transcripts, we searched all assembled transcripts for fusions using JAFFAL [51] and compared them to previously reported fusion genes in the CCLE. For PCR-cDNA, 8 novel transcripts were identified as DE fusion transcripts with 4 previously reported fusion genes in CCLE (2 from H1975 and 2 from HCC827, Additional file 1: Table S11) [49]. For the dRNA data, 26 novel transcripts were identified as DE fusion transcripts with 11 previously reported fusion genes in CCLE (10 from MCF7 and 1 from the A549

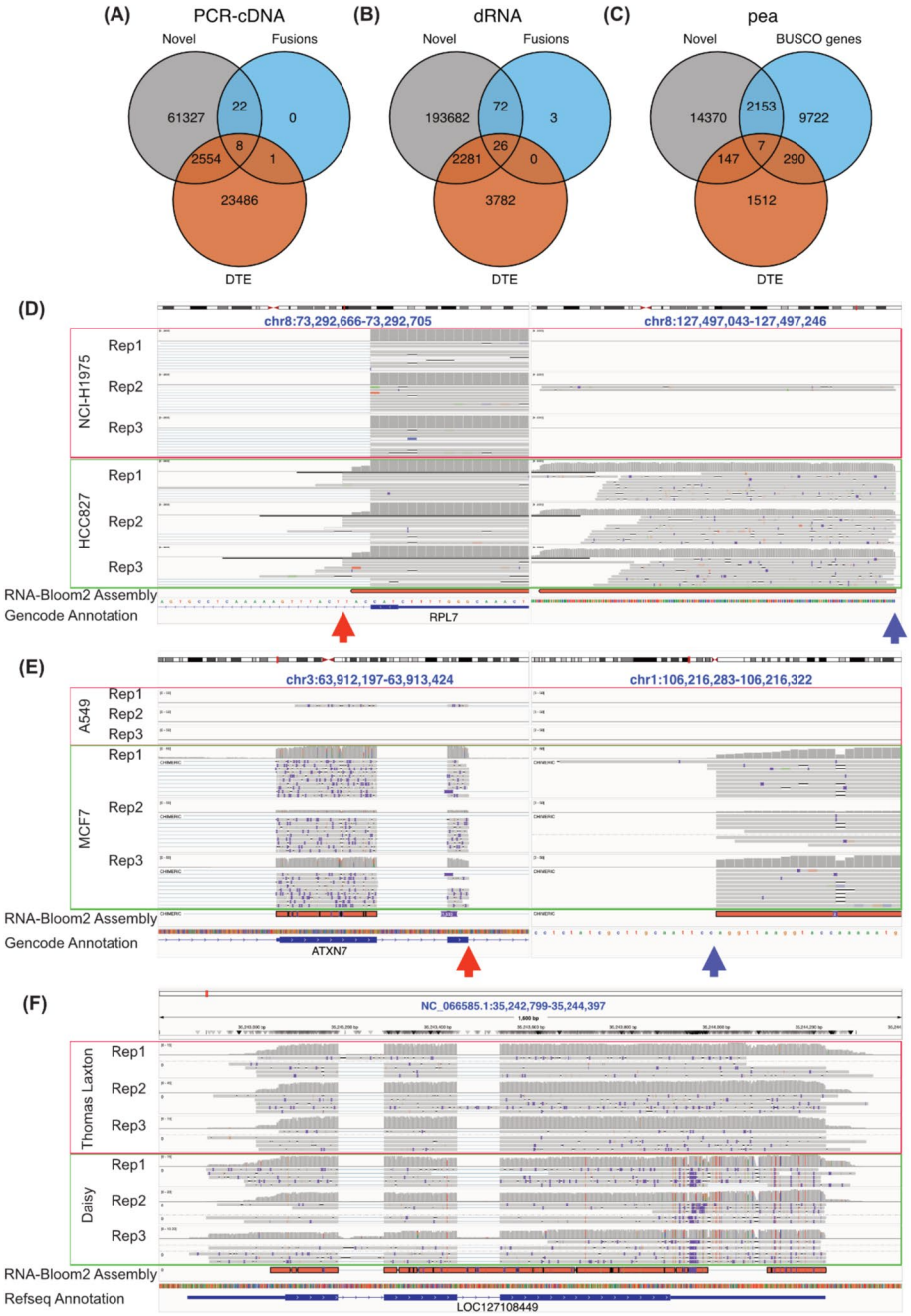

Fig. 6 Novel transcripts in de novo assemblies. **A-B** Venn diagram showing the number of transcripts from RNA-Bloom2 which were differentially expressed, novel, and/or known fusion genes in (**A**) PCR-cDNA 60 million and (**B**) dRNA. **C** Venn diagram showing the number of transcripts from RNA-Bloom2 which were differentially expressed, novel, and/or has a BUSCO gene match (pea). **D** IGV visualisation of potential *RPL7*-chr8 intergenic fusion from the PCR-cDNA dataset, split view to show the break point (orange and blue arrows showed left and right breakpoints respectively). **E** IGV visualisation of potential *ATXN7*-chr1 intergenic fusion from the dRNA dataset, split view to show the break point (orange and blue arrows showed left and right breakpoints respectively). **F** IGV visualisation of *LOC127108449* transcript from pea data showed cultivar specific SNPs, insertions and deletions

cell line) (Additional file 1: Table S12) [49]. When we performed a similar analysis using the novel transcripts from RATTLE, we found 4 DE fusion transcripts in PCR-cDNA, matching 3 CCLE fusion genes, all of which were also detected with RNA-Bloom2. In the dRNA dataset, 9 DE fusion transcripts from RATTLE matched 9 CCLE fusion genes, of which 7 were also detected in RNA-Bloom2 (Additional file 2: Figs. S4A-B, Additional file 1: Tables S11-12).

For the pea dataset, analysed with RNA-Bloom2, we found 154 DE transcripts with novel sequences, including 7 matching BUSCO genes (Fig. 6C, Additional file 1: Table S13). These 7 transcripts all showed large sections of divergent sequences when aligned to the reference genome. In addition, these transcripts harboured phased cultivar-specific SNPs. An example of this is in the uncharacterised gene *LOC127108449*, which has homology with DBH-like monooxygenase in eudicotyledons. We identified a transcript of *LOC127108449* which had a small insertion (~20 bp total), several small deletions and a novel splice variant in its 3' UTR which was significantly more highly expression in Daisy compared to the Thomas Laxton cultivar (logFC = 4.7) (Fig. 6F). Read coverage over this loci was consistent with the small novel insertion and deletions in Daisy, but limited reads supported the splice variant, highlighting the need for manual curation of de novo assembly results. Interestingly, RATTLE detected more novel DE transcripts matching to BUSCO genes (13 genes), with only 3 matching between the two assemblers (Additional file 2: Fig. S4C). These results suggest that assembly recall may be improved if multiple assemblies are combined.

These novel examples collectively demonstrate that de novo methods do provide insights beyond known genes and transcripts and enable reference-unbiased differential analysis.

Long-read only pipelines outperform hybrid approaches

Despite the theoretical advantages of long-read sequencing for reference-free differential transcript analysis, our benchmarking results indicate that short reads still perform well. Short-read sequencing currently benefits from lower error rates. It also produces a greater number of counts for a fixed total number of sequenced bases, increasing statistical power for differential testing. Therefore, as a final assessment in our study, we

(See figure on next page.)

Fig. 7 Performance of hybrid short and long-read data assembly. **A** Length distribution of assembled and reference transcriptomes. **B** Proportion of BUSCO complete genes (both single copy and duplicated) assembled before and after sequence polishing. Assembled transcripts were corrected using the reference genome to remove errors and assess protein sequence predictions. **C** Number of novel and annotated genes and transcripts in each assembled transcriptome. Transcripts were further divided into matched transcript (complete splice match and incomplete splice match), novel isoform (novel in catalogue and novel not in catalogue) and novel gene (intergenic, genic, intronic, antisense, and fusion). **D** Proportion of reference transcripts that were assembled binned by true expression. True expression was taken from simulated or Bambu counts, and binned into 10% quantiles. X-axis labels show the corresponding read count range for each quantile. **E** The number of gene-level clusters generated for each assembly. 'match' indicates when all transcripts within the cluster correspond to the same reference gene. Where one reference gene matched multiple clusters, the cluster with the most transcripts was classed as a 'perfect match' and the rest as a 'redundant match'. Clusters with transcripts from multiple reference genes were classed as 'mixed'. Clusters with only novel transcripts were classed as 'no match'. **F**, **G** and **H** Unique false positives were shown against unique true positives for DGE, DTE and DTU analysis for simulation (**F**), PCR-cDNA-sequin (**G**) and PCR-cDNA (**H**) data. Where unique true and positives are described as in Fig. 5. The total number of true positives is indicated by the horizontal dashed line (black)

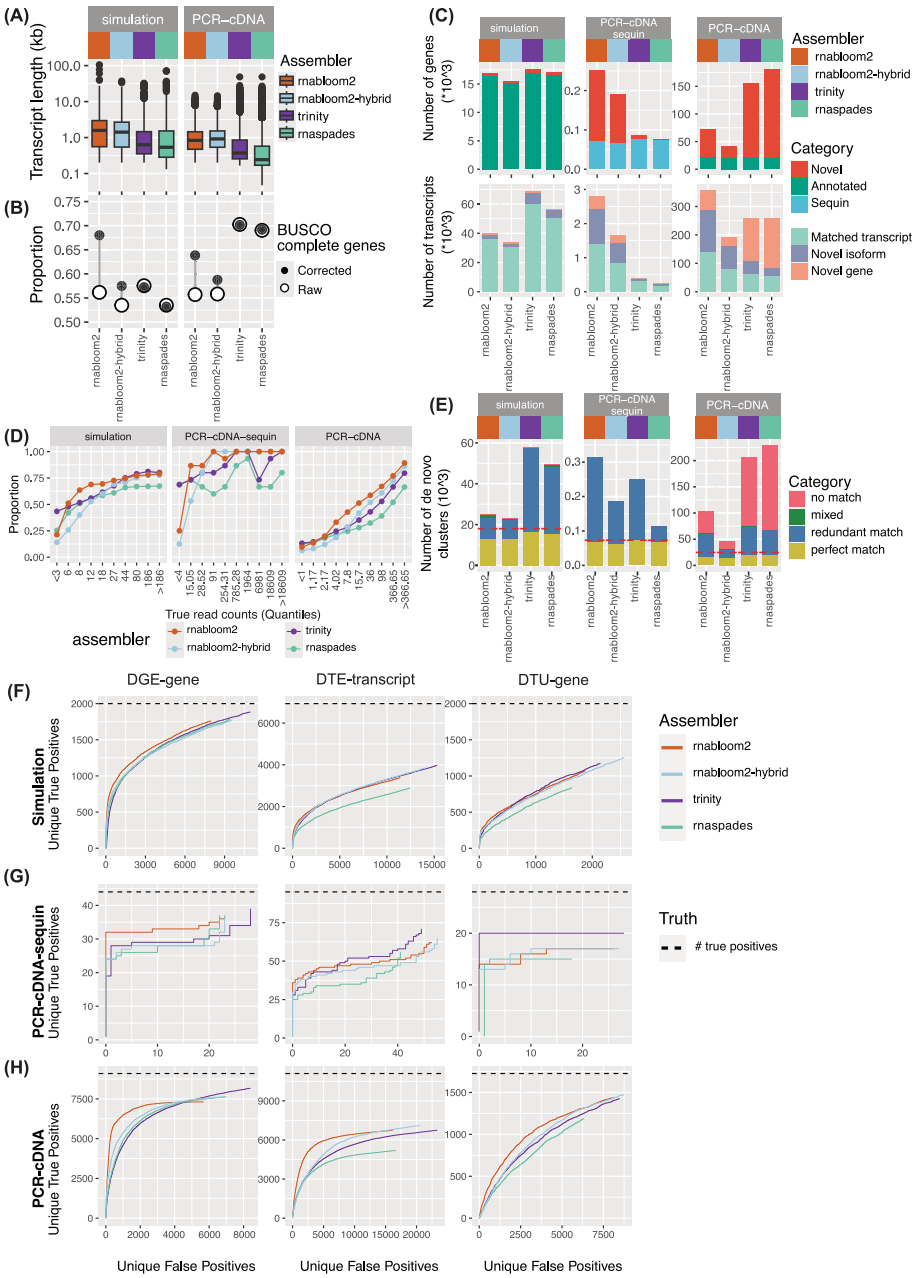

Fig. 7 (See legend on previous page.)

evaluated whether hybrid approaches could integrate the strengths of both long- and short-read sequencing to outperform either technology alone.

Using the simulation and 60 million PCR-cDNA datasets, we evaluated two representative hybrid assembly approaches: RNABloom2-hybrid and rnaSPAdes. These two tools represent current mainstream approaches for hybrid assembly. RNABloom2-hybrid extends RNABloom2 by using short-reads to polish the long-read prior to assembly. In contrast, rnaSPAdes uses the short reads to build an assembly graph, then long reads are aligned to the graph to resolve repeats and close gaps.

Each hybrid assembler was applied to a dataset comprising 50% ONT and 50% Illumina reads, with reads randomly selected and the total number of sequenced bases matched to the single-technology datasets. Following assembly, read alignment and quantification were performed separately for long and short reads as described previously, and the resulting count matrices were aggregated to generate the final gene- or transcript-level quantifications. Pipeline performance was then assessed using the same metrics as before, including assembly quality (Fig. 7A-C, Additional file 2: Figs. S5A-B), clustering (Fig. 7D-E, Additional file 2: Figs. S5C-E), and differential expression accuracy (Fig. 7F-H).

We found that assemblies generated from long reads (RNA-Bloom2) and short reads (Trinity) performed as well as, or better than, the hybrid approaches across most metrics and datasets. RNABloom2-hybrid exhibited similar characteristics to RNA-Bloom2, such as longer transcript lengths (Fig. 7A) and retained sequencing errors (Fig. 7B). However, its overall performance resembled that of RNA-Bloom2 applied to only 30 million reads (Additional file 1: Table S3, 4, 5, 6), suggesting that the short-read component of the input data may not have been fully utilised during assembly. Nevertheless, as short reads are used for polishing, RNABloom2-hybrid showed a lower sequence error rate than RNA-Bloom2, as indicated by the smaller improvement observed after error correction when evaluating BUSCO completeness (Fig. 7B). RnaSPAdes, by contrast, exhibited features typical of short-read assemblers, producing shorter transcript contigs (Fig. 7A) but with low sequence error rates (Fig. 7B). We therefore speculate that the long-read component of the input data was underutilised in this case. Collectively, these results suggest that further optimisation is required for hybrid transcriptome assembly approaches to fully exploit the complementary advantages of long- and short-read sequencing technologies.

Discussion

Our comprehensive evaluation of long-read de novo transcriptome assembly tools highlights both the promise and ongoing challenges of performing transcript-level analyses in the absence of a high-quality reference genome and/or transcriptome. We assessed RATTLE, RNA-Bloom2 and isONform, using the short-read assembler Trinity and reference-guided tool Bambu as a guide for expected performance. Using simulated data, sequin spike-ins and real experimental data spanning different library types, read depths, and biological applications, we tested a broad range of experimental scenarios and identified the most effective general strategy for reference-free long-read differential expression analysis.

An important finding of this study is that computational performance is a bottleneck for most tools. Current tools require major improvement to efficiently handle large datasets. This study was one of the first to benchmark these methods with a dataset of a realistic size (60 million reads). However, it is no longer unusual for a PromethION flow cell to generate over 100 million reads, while the recent Pacbio Kinnex SPRQ upgrade on the Revio sequencer now generates 60 million transcript reads per flow cell, and we can only expect the size of long-read transcriptome datasets to increase in the future. Notably, in this study we were unable to assemble the full PacBio single-cell dataset of 555 million reads because of the computational demand.

Our strategy of assembling a subset of the data, but leveraging all reads for quantification is one approach for managing the computational burden.

Another key finding is that, although long-read methods inherently produce longer assembled transcripts with more complete isoform structures than short-reads, they still do not reach the capability of reference-guided assemblies. This limitation was seen across multiple metrics, especially in recovering lowly expressed transcripts and performing transcript-level differential analyses.

Among the de novo long-read pipelines evaluated, RNA-Bloom2 consistently performed among the top, especially when paired with Corset for transcript clustering and downstream differential analysis. However RNA-Bloom2 had a tendency to produce very large assemblies, with a high amount of redundancy. RATTLE and isONform were both limited by their computational needs, and when applied to a smaller and less noisy dataset, isONform performed well, suggesting its default parameters were not robust on different datasets.

Although Trinity assembled shorter and incomplete transcripts, higher sequencing depth in short-read RNA-seq increased the power in differential testing. Another advantage of short-read RNA-seq is lower error rates in the assembled contigs. We found a surprisingly high rate of sequencing errors were retained in long-read ONT de novo assembly. This inadversely impacted multiple downstream analysis steps such as clustering and identification of novel transcripts, and we found it to have a profound impact on protein sequence predictions. However, ongoing improvements in basecalling, such as dorado duplex basecalling, and sequencing chemistries, such as the ONT RNA004 kit and R10 flowcell, are expected to reduce these errors. As error rates decrease, we anticipate a corresponding enhancement in transcriptome assembly quality, downstream applications such as DE analysis, as well as an improvement in computational efficiency.

Although comprehensive, our study did not examine ways to improve assembly quality, such as polishing reads or contigs, or combining multiple assemblies produced by different tools or k-mers. Benchmarking could also be expanded to examine alternative methods for post-assembly transcript clustering and quantification, which may influence the final DE results. With the recent development of numerous long-read specific methods, this could improve downstream analysis further [38, 52].

We also noted that wider applications of de novo transcriptome assembly beyond DE analysis were not benchmarked, but our results have implications for them too. With improved transcript reconstructions, it will become increasingly feasible to perform accurate variant calling and allele-specific expression analyses using long-read data, enabling detailed comparative transcriptomics in non-model organisms. Such analyses can reveal subtle differences across populations or environmental conditions. Although these types of analyses are already performed using short reads, long reads enable phasing of variants and splicing, which can now be done reference-free, and even at single-cell resolution. Additionally, long-read assemblies will allow the study of epitranscriptomic modifications in non-model species. These extensions highlight that assembled transcriptomes are not just a means to perform simple quantification, but can serve as a reference for multiple types of analyses.

Short-read sequencing remains an important approach due to its high throughput, high base quality and well established analytical workflows, allowing the detection of

subtle differential expression changes. However, short-read assemblies, as demonstrated, generally produce shorter contigs and are less capable of resolving isoform complexity without a reference. In contrast, long-read assemblies are better at capturing complete transcripts, which is crucial for a comprehensive understanding of gene structure. Thus, the decision to adopt a short-read or long-read strategy, or even a hybrid approach, depends on experimental goals, sample availability, and the quality of the reference genome. Further work is required to identify the optimal and most cost-effective combination of these technologies.

Conclusions

Our findings demonstrate that while contemporary long-read transcriptome assemblers perform well, weaknesses exist in terms of performance and usage, highlighting the need for ongoing development in this field. By demonstrating the strengths and weaknesses of different combinations of tools, such as RNA-Bloom2 and Corset, we provide a practical guide for researchers designing DE studies without a high-quality reference genome or transcriptome. With future improvements in error-correction methods, assembly algorithms and reference-free quantification strategies, we anticipate that long-read transcriptome assembly will enable more accurate, efficient and wide-ranging applications for long-read data.

Methods

Pea growth conditions and RNA extraction

Pisum sativum cultivars 'Thomas Laxton' and 'Daisy' (supplied by the Australian Grains Genebank and the John Innes Centre) were grown within a controlled glasshouse environment at an average ambient temperature of 22 °C and subjected to long-day light cycles (8 h dark, 16 h light). The soil medium macro/micro nutrients and other constituents consisted of 0.5 L Vermiculite, 0.5 L Perlite, 35 g Macracote Coloniser Plus 4-month slow release fertiliser (15N: 3P: 9 K), 30 g Nitrogen slow-release fertiliser (40N: 0P: 0 K), 25 g Water holding granules, 15 g Trace elements (6 Mg: 6.5FE:5.4S: 1.5Mn: 0.4Zn: 0.14B: 0.07Mo), and 5 g Garden lime per 30 L of soil. Whole stipules were harvested from the third leaflet node of individual seedlings 2 weeks post-emergence and snap-frozen in liquid nitrogen. Each sample was crushed into a fine powder by pestle and mortar. Samples were then weighed on a microgram scale to 1 mg of whole stipule tissue per extraction. Total RNA was then isolated using Invitrogen TRIzol reagent (ThermoFisher Scientific) using the manufacturer's protocol. Minor modifications to the protocol included an extended centrifugation for 15 min, as some RNA precipitate was still suspended in solution for some samples. RNA precipitate pellets were washed twice rather than once, as it was found this removed further impurities and improved the overall quality of the RNA. Centrifugation between washes was also extended to 10 min, to ensure the RNA pellet was anchored to the bottom of the Eppendorf tube during wash steps. Total RNA pellets for each sample were then re-suspended in RNase-free water, and subsequently quality-checked and quantified utilising the NanoDropTM spectrophotometer.

Simulating Nanopore cDNA data and matched short-read sequencing data

We first obtained a subset of transcripts that are widely expressed in the GTEx v9 long read dataset (92 samples) using Gencode comprehensive annotation (v44). We kept transcripts with more than 5 reads in at least 15 samples after Salmon quantification (18,145 genes, 40,509 transcripts) and stored their mean count per million (CPM) values as the control group's baseline expression. We then generated a perturbed set of CPM values where transcript expression was changed by: (1) randomly selecting 1000 genes and changing all transcripts belonging to that gene concordantly (500 genes two fold up and 500 genes two fold down), (2) selected another 1000 genes randomly, and then select 2 random transcripts from the gene and swap their expression, (3) selected another 1000 genes randomly and then selected 1 random transcript to change its expression (500 transcripts two fold up and 500 transcripts two fold down). The updated CPM were stored as the perturbed group baseline expression. We then generated a count matrix and CPM matrix for 3 control replicates and 3 perturbed replicates with gamma distribution, followed by a Poisson distribution [39]. Both long-read and short-read FASTQ files were simulated using SQANTI-SIM with default settings and ONT R9.4 cDNA error profile (v 0.2.1) [44]. The long read data contained 6 million reads in total, and an average read length of 1085 bp, and short read data was 100 bp paired-end. We then subsampled the short-read data to match the total number of base pairs in the long read data (6.5 billion bases). The simulated data was non-stranded, and contains 2000 DE genes, 2000 genes with DTU, 5927 transcripts with DTU and 6933 DE transcripts.

Nanopore PCR-cDNA, direct-RNA data and matched short-read RNA-seq

To investigate how the assemblers perform with different sizes and types of data, we downsampled the PCR-cDNA data for 2 cancer cell lines (HCC827 and NCI-H1975, 3 replicates each) from GSE172421 to 12 million, 30 million and 60 million reads in total (2 million, 5 million and 10 million for each replicate, with a mean read length 685 bp) and included short-read stranded RNA-seq with matched number of bases to the 60 million PCR-cDNA data (75 bp paired-end and 41 billion bases) [30]. The PCR-cDNA data includes sequin spike-in transcripts from 2 mixes of different concentrations. Our benchmarking study also used direct-RNA data from the SG-NEx project for the A549 and MCF7 cell lines (3 replicates each, 11.6 million total long reads with a mean read length of 1060 bp, covering 12.3 billion bases) and matched short-read RNA-seq [43]. Lastly, a PCR-cDNA ONT dataset with 12 samples (3 replicates across 4 conditions) from pea (*Pisum sativum*) species was prepared with the Oxford Nanopore PCR-cDNA Barcoding kit (SQK-PCB109) following the manufacturer's guidelines. Fifty-nanograms total RNA input was used for each of the 12 samples. Samples were reverse-transcribed into cDNA and PCR amplified and barcoded for 15 cycles in a thermocycler. The final pool was loaded onto a PromethION flow cell (FLO-PRO002, R9.4.1) and sequenced on the PromethION P24 (Oxford Nanopore Technologies Limited) over 72 h. Raw nanopore reads were basecalled in super-accurate mode and demultiplexed with guppy (v 6.3.7). This data contained 16.6 million reads covering 10.7 billion bases. The same RNA was used to prepare short-read libraries with the Illumina TruSeq stranded kit v2 according to manufacturer's instructions. Pooled libraries were sequenced on an

Illumina NextSeq 550 with 75 bp single end reads. All long-read PCR-cDNA data, dRNA data and their matched short read were stranded.

PacBio Kinnex PBMC single cell RNA-seq data

PacBio Kinnex PMBC single cell RNA-seq data was downloaded from PacBio website [53]. The deconcatenated and UMI-deduplicated FASTQ files for three 10x 3' PBMC samples were downloaded (10 k, 20k_rep1, 20k_rep2). The full data were subsampled to 5 million reads (mean read length 835 bp) per sample for assembling. The full dataset containing 555 millions reads was used for quantification and differential analysis.

Transcriptome assembly

RATTLE (v 1.0), RNA-Bloom2 (v 2.0.1), isONform (v 0.3.4) and Trinity (v 2.15.1) were run using FASTQ files according to the default workflow. RATTLE was compiled from source, RNA-Bloom2 and isONform were installed from conda, and Trinity was run within a singularity container image. Simulated cDNA data were assembled using an unstranded setting, while the PCR-cDNA data from cancer cell lines and pea were pre-processed with pychopper (v 2.7.9) to reorientate reads retaining only full length and rescued reads with properly paired primers at both ends, and assembled using a stranded setting. Direct RNA was assembled with a stranded setting, and U was converted to T in the final assembly. The isONform pipeline was run with isONclust (0.0.6.1, with the –ont option), isONcorrect (0.1.3.5, default options) and isONform (–exact_instance_limit 50 –k 20 –w 31 –xmin 14 –xmax 80 –max_seqs_to_spoa 200 –delta_len 10 –iso_abundance 5 –delta_iso_len_3 30 –delta_iso_len_5 50). For the single-cell PacBio data, we ran RATTLE in stranded mode, RNA-Bloom2 in stranded mode with –lrpb flag, and isONform pipeline without the correction step suggested by the authors. Details can be found in the GitHub repository [54].

For reference-guided assembly, minimap2 (v 2.26-r1175) was used to generate BAM files (-ax splice for simulated cDNA, -ax splice -uf for PCR-cDNA data from cancer cell lines and pea and -ax splice -uf -k14 for dRNA). For simulation, cell lines and single-cell data the GRCh38 genome and Gencode v44 comprehensive GTF annotation were used (with sequin annotation included for PCR-cDNA data). For the pea data, the RefSeq ZW6 reference genome and corresponding gene annotation were used (Additional file 1: Table S1).

Bambu (v 3.3.4) was then run with the corresponding strandness of the data and other parameters default. Only expressed transcripts and genes (based on a Bambu transcript count of $>=1$ in at least one sample) were retained for downstream comparisons and differential analysis. The Bambu GTFs were converted to FASTA using gffread (v 0.12.8).

All assemblies were executed on a high-performance cluster at the Walter and Eliza Hall Institute, Australia, with 48 cores (Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60 GHz (Broadwell), Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz (Icelake), Intel(R) Xeon(R) Gold 6130 CPU @ 2.10 GHz (Skylake)) and 1 Tb max RAM (DDR4 2400 MT/s, DDR4 3200 MT/s, DDR4 2666 MT/s) requested. Assemblies that did not finish in 2 weeks were terminated and excluded.

Hybrid assembly

To perform hybrid assembly on the simulation data and 60 million PCR-cDNA data, we subset 50% of Nanopore reads and 50% of Illumina reads. RNA-Bloom2 (hybrid mode) and rnaSPAdes (4.2.0, binary downloaded from github) were used for hybrid assembly with the same strandness for each dataset as described earlier.

Assessing basic assembly quality metrics

We assessed all assemblies with transrate (v 1.0.3) for assembly length, GC content and reference recovery using Conditional Reciprocal Best BLAST [47]. We also run BUSCO (v 5.4.7, database primates_odb10 and fabales_odb10) to obtain the proportion of conserved genes that were assembled [55]. We used RepeatMasker (4.2.2) to search repetitive sequences in human and *Pisum sativum* databases (Dfam 3.9) [56]. We used PrimeSpotter to detect internally primed sequences from the assembled transcriptome [48].

Assessing assembly based on splice junction using SQANTI3

We ran SQANTI3 (v 5.1.2) to assign de novo assembled transcripts and Bambu novel transcripts to reference gene and transcript annotation [45]. We provided a junction file to correctly align assembled transcripts to the reference genome. The aligned BAM files were converted to GTF and then FASTA files based on reference genome sequences using gffread to remove small insertion and deletions. These genome-corrected FASTA files were also assessed for their BUSCO score. The original uncorrected FASTA files were used for all downstream analysis. For simulated non-stranded data, all transcripts classified as 'antisense' were subjected to a second round of SQANTI3 annotation using their reverse complement sequences.

Assessing the recovery of reference transcripts and genes

To measure the recovery of reference transcripts and genes, we binned the reference transcripts and genes into 10 quantiles based on quantification from simulation, sequin or Bambu with reference. The proportion of assembled reference transcripts and genes was extracted from SQANTI3 results.

Assessing transcript quantification correlation

To quantify the expression of assembled transcripts, we ran minimap2 (map-ont for ONT data and map-hifi for PacBio data) to align the FASTQ files from individual samples to the de novo transcriptome (options: -p 1.0 -N 100). For the ONT and short-read data, we then ran salmon (v 1.10.2) on the aligned BAMs with 100 bootstraps, for the ONT data we used the option –ont. For the hybrid assembly data, the counts from short read (salmon quant) and long reads (minimap2 + salmon) were summed to obtain the final counts for each assembled transcript in each replicate. For the single-cell data, oarfish were used and single-cell counts were aggregated to sample level before assessing quantification. To measure correlation with truth transcript expression, de novo transcripts were assigned to reference transcripts based on SQANTI3 results. When multiple de novo transcripts were assigned to the same reference transcript, only the one

with the highest expression was used to calculate the correlation. De novo transcript expression was calculated as log2(count + 1). The expression values for each sample were then correlated to known expressions for simulation (log2(count + 1)), sequin data (log2(concentration)), or Bambu quantification (log2(count + 1)).

Assessing transcript to gene cluster assignment

Because Bambu, isONform, RATTLE, Trinity and rnaSPAdes perform clustering of transcripts to genes using their in-built algorithms, we included their transcript to gene cluster information. For all de novo methods, we further clustered assembled transcripts into genes using Corset based on the all-to-all overlap of transcripts using minimap2 (with options: -a -k15 -w5 -e0 -m100 -r2k -P –dual=yes –no-long-join) [35]. Furthermore, the performance of clustering was evaluated using several metrics including precision (de novo transcripts associated with transcripts from the same gene are clustered into one gene cluster) and recall (de novo transcripts associated with transcripts from the different genes are separated into different gene clusters) using SQANTI3 assigned transcripts and genes as truth.

Assessing gene cluster quantification

We use 3 different methods to obtain gene clusters for quantification: (1) de novo gene clustered from Corset, (2) de novo gene clusters from native clustering if applicable, (3) gene clusters defined by SQANTI3 using the reference gene annotation. Gene cluster expression was calculated as aggregated counts from all transcripts in the cluster and converted to log2(count + 1). Using SQANTI3 results, each gene cluster was assigned to a known gene if the majority of its transcripts were assigned to that gene. The correlation was then evaluated in a manner similar to that used for transcripts.

Assessing the redundancy of gene clusters

For de novo gene clusters from Corset or native clustering, we further classified them into (1) no match: if all transcripts from the cluster have no match to reference; (2) mixed: if transcripts from the cluster match to multiple reference genes, (3) match: all transcripts from the cluster match to one reference gene, and if multiple clusters match to the same gene, the one with most transcripts are defined as a perfect match and the rest defined as a redundant match.

Assessing differential analysis using de novo assembly

For ONT and short read assemblies, we used edgeR::catchSalmon (4.2.0) to import counts and read to transcript ambiguity. The scaled transcript counts (counts divided by read-to-transcript ambiguity) were used for differential transcript analysis (DTE and DTU), and the sum of transcript counts from gene clusters was used for differential gene expression analysis (DGE). For hybrid assemblies, the scaled transcript counts from short and long read were aggregated for DTE/DTU analysis and the sum of short and long read transcript counts from gene clusters was used for differential gene expression analysis (DGE). For single-cell data, due to the sparsity of the data, we performed pseudo-bulk differential analysis at cluster level. First, we obtained true cell type labels. We downloaded the Seurat count matrix for the 3 samples from the PacBio website and

annotated with Azimuth (0.5.0) using PBMC reference data to obtain true cell types for each cell [57] . Then for each of the de novo assemblies, a single cell UMI count matrix was generated using oarfish [38], and transcript UMI counts were aggregated into gene UMI counts based on gene clusters. Gene expression clustering was then performed to confirm the presence of two major clusters, corresponding to lymphoid and monocytic cells based on true cell type labels, and pseudo-bulking was performed for the two major clusters in each replicate. All analyses were performed using the limma::voom (3.52.4) pipeline after filtering [39]. For the single-cell data, we also ran limma::treat to select differential genes and transcript with fold-change > 1.5. For differential transcript usage (DTU) analysis, we used limma::diffSplice, and Simes correction to obtain gene-level p values. The differential lists were compared to the truth of either simulated differential gene/transcript, sequin with known differential expression or abundance, or Bambu differential results with the reference. If multiple de novo transcripts or genes matched to a reference transcript or gene, only the one with the smallest FDR was used when assessing unique true or false positives. The code for differential analysis can be found in the GitHub repository https://github.com/DavidsonGroup/longread_denovo_benchmark.

Novel transcripts in the de novo assembly

We ran JAFFAL (v 2.3) on the RNA-Bloom2 and RATTLE assembled transcriptome in FASTA format from the cancer cell lines (PCR-cDNA and dRNA) [51]. Known structural variants and fusions from corresponding cell lines were downloaded from the DepMap portal (Public 24Q2 and CCLE2019 dataset) [49, 58]. Alignment and tracks were visualised in IGV [59].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04001-5.

Additional file 1: Table S1. Summary of dataset characteristics, parameters and analyses done for each dataset. Table S2. Running time and memory usage for each assembler in each dataset. Table S3. The number of transcripts and genes by SQANTI3 category for each assembly. Table S4. Percentage of repetitive sequences and internally primed sequences. Table S5. Number of transcripts matching BUSCO genes for each assembly usingthe raw uncorrected transcriptome sequences andgenome-corrected transcriptome sequences. Table S6. Correlation coefficient between de novo assembly and reference transcripts and gene abundances. Table S7. Clustering evaluation metrics of native and Corset clustering using SQANTI3 to define true clusters. Table S8. The number of clusters classes as perfect, redundant, mixed and no match for each clustering method. Table S9. Top 10 novel DTEs ranked by FDR in 60 million PCR-cDNA data. Table S10. Top 10 novel DTEs ranked by FDR in dRNA data. Table S11. Novel fusion DTEs in 60 million PCR-cDNA data. Table S12. Novel fusion DTEs in dRNA data. Table S13. Novel DTE with BUSCO hits in pea data.

Additional file 2: Fig. S1. Assessment of assembled transcriptome quality. Fig. S2. Accuracy of transcript and gene abundance estimates. Fig. S3. ROC-style curves for differential analysis. Fig. S4. Novel transcripts in RATTLE. Fig. S5. Hybrid de novo assembly.

Acknowledgements

This research was undertaken with the assistance of Milton HPC/Virtual Machines, supported by WEHI Research Computing. We thank Ashley L. Weir for her valuable comments on the manuscript. We also acknowledge the use of ChatGPT for assistance with language and grammar editing.

Peer review information

Claudia Feng and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Author's information

Nadia M. Davidson, @nadia-davidson.bsky.social.
Mathew G Lewsey, @lewseylab.bsky.social.
Quentin Gouil, @qgouil.bsky.social.
Pedro L. Baldoni, @plbaldoni.bsky.social.
Feng Yan, @alexyfyf.bsky.social.

Authors' contributions
FY and NMD developed the conceptual design of the benchmarking project. FY performed all analyses and generated figures. FY and NMD wrote and revised the manuscript. PLB designed the simulation dataset with differential expression. JL and QG collected ***Pisum sativum*** samples, prepared libraries, performed sequencing, provided valuable suggestions for the benchmark and manuscript. NMD, MGL and MER supervised data generation and analyses. All authors provided feedback and approved the final manuscript.

Funding
NMD, QG and MER are supported by Australian National Health and Medical Research Council (NHMRC) Investigator Grants (GNT2016547, GNT2007996 and GNT2017257 respectively). NMD is also supported by the Estate of Judith Corrie Philpots. Work in MGL's laboratory was supported by the La Trobe University Research Focus Areas (Understanding Disease & Securing Food Water and the Environment).

Data availability
The code to reproduce the analysis can be found in GitHub under MIT licence [54]. The code for generating the simulation, the simulated FASTQ files, and all assemblies generated in this study can be obtained from Zenodo [60, 61]. The PCR-cDNA data from cancer cell lines were obtained from GEO (GSE172421) [62], and dRNA data were obtained from SG-NEx AWS Open Data Registry [63]. The long-read PCR-cDNA and short-read RNA-seq data from pea are available in ArrayExpress under accessions E-MTAB-14841 and E-MTAB-14845, respectively [64, 65]. The PacBio Kinnex single cell RNA-seq data were downloaded from PacBio's website [53].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
NMD and QG have had conference travel funded by Oxford Nanopore Technologies.

Received: 10 February 2025 Accepted: 5 February 2026
Published online: 18 February 2026

References

1. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5. https://doi.org/10.1038/nbt.3122.
2. Zhao Z, Chen Y, Cheng X, Huang L, Wen H, Xu Q, et al. The landscape of cryptic antisense transcription in human cancers reveals an oncogenic noncoding RNA in lung cancer. Science Advances. 2023;9:eadf3264. https://doi.org/10.1126/sciadv.adf3264.
3. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. RnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. 2019;8:giz100. https://doi.org/10.1093/gigascience/giz100.
4. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. https://doi.org/10.1038/nbt.1883.
5. Nip KM, Chiu R, Yang C, Chu J, Mohamadi H, Warren RL, et al. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. Genome Res. 2020;30:1191–200. https://doi.org/10.1101/gr.260174.119.
6. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92. https://doi.org/10.1093/bioinformatics/bts094.
7. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660–6. https://doi.org/10.1093/bioinformatics/btu077.
8. Hölzer M, Marz M. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8:giz039. https://doi.org/10.1093/gigascience/giz039.
9. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform. 2022;23:bbab563. https://doi.org/10.1093/bib/bbab563.
10. Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 2018;564:64–70. https://doi.org/10.1038/s41586-018-0734-6.
11. Zhang G-Q, Liu K-W, Li Z, Lohaus R, Hsiao Y-Y, Niu S-C, et al. The Apostasia genome and the evolution of orchids. Nature. 2017;549:379–83. https://doi.org/10.1038/nature23897.
12. Zhang Y, Chen H, Lian C, Cao L, Guo Y, Wang M, et al. Insights into phage-bacteria interaction in cold seep *Gigantidas platifrons* through metagenomics and transcriptome analyses. Sci Rep. 2024;14:10540. https://doi.org/10.1038/s41598-024-61272-3.
13. Cmero M, Schmidt B, Majewski IJ, Ekert PG, Oshlack A, Davidson NM. MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol. 2021;22:296. https://doi.org/10.1186/s13059-021-02507-8.
14. Davidson NM, Majewski IJ, Oshlack A. JAFFA: high sensitivity transcriptome-focused fusion gene detection. Genome Med. 2015;7:43. https://doi.org/10.1186/s13073-015-0167-x.

15. Gonzalez-Bosquet J, Cardillo ND, Reyes HD, Smith BJ, Leslie KK, Bender DP, et al. Using genomic variation to distinguish ovarian high-grade serous carcinoma from benign Fallopian tubes. Int J Mol Sci. 2022;23:14814. https://doi.org/10.3390/ijms232314814.
16. Wang Y, Xue H, Aglave M, Lainé A, Gallopin M, Gautheret D. The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma. NAR Cancer. 2022;4:zcac001. https://doi.org/10.1093/narcan/zcac001.
17. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15:201–6. https://doi.org/10.1038/nmeth.4577.
18. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22:1–24. https://doi.org/10.1186/S13059-021-02525-6.
19. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods. 2023. https://doi.org/10.1038/s41592-023-01908-w.
20. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278. https://doi.org/10.1186/s13059-019-1910-1.
21. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, et al. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol. 2023;41:915–8. https://doi.org/10.1038/s41587-022-01565-y.
22. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11:1–12. https://doi.org/10.1038/s41467-020-15171-6.
23. Marchet C, Lecompte L, Silva CD, Cruaud C, Aury J-M, Nicolas J, et al. De novo clustering of long reads by gene from transcriptomics data. Nucleic Acids Res. 2019;47:e2. https://doi.org/10.1093/nar/gky834.
24. Sahlin K, Medvedev P, Mary Ann Liebert, Inc., publishers. De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. J Comput Biol. 2020;27:472–84. https://doi.org/10.1089/cmb.2019.0299.
25. de la Rubia I, Srivastava A, Xue W, Indi JA, Carbonell-Sala S, Lagarde J, et al. RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. Genome Biol. 2022;23:153. https://doi.org/10.1186/s13059-022-02715-w.
26. Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, et al. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. Nat Commun. 2023;14:2940. https://doi.org/10.1038/s41467-023-38553-y.
27. Petri AJ, Sahlin K. IsONform: reference-free transcriptome reconstruction from Oxford Nanopore data. Bioinformatics. 2023;39:i222-31. https://doi.org/10.1093/bioinformatics/btad264.
28. Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J, et al. IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics. 2018;34:2168–76. https://doi.org/10.1093/bioinformatics/bty098.
29. Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, et al. Extending rnaSPAdes functionality for hybrid transcriptome assembly. BMC Bioinformatics. 2020;21:302. https://doi.org/10.1186/s12859-020-03614-2.
30. Dong X, Du MRM, Gouil Q, Tian L, Jabbari JS, Bowden R, et al. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. Nat Methods. 2023;20:1810–21. https://doi.org/10.1038/s41592-023-02026-3.
31. Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB, et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat Methods. 2016;13:792–8. https://doi.org/10.1038/nmeth.3958.
32. Su Y, Yu Z, Jin S, Ai Z, Yuan R, Chen X, et al. Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. Nat Commun. 2024;15:3972. https://doi.org/10.1038/s41467-024-48117-3.
33. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nat Methods. 2024;21:1349–63. https://doi.org/10.1038/s41592-024-02298-3.
34. Sagniez M, Budhraja A, Paré B, Simpson SM, Vinet-Ouellette C, Rozendaal M, et al. Assembly Arena: Benchmarking RNA isoform reconstruction algorithms for nanopore sequencing. bioRxiv; 2024. p. 2024.03.21.586080. https://doi.org/10.1101/2024.03.21.586080. Cited 2024 Aug 22.
35. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 2014;15:410. https://doi.org/10.1186/s13059-014-0410-6.
36. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
37. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. https://doi.org/10.1038/nmeth.4197.
38. Zare Jousheghani Z, Singh NP, Patro R. Oarfish: enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. Bioinformatics. 2025;41:i304. https://doi.org/10.1093/bioinformatics/btaf240.
39. Baldoni PL, Chen Y, Hediyeh-zadeh S, Liao Y, Dong X, Ritchie ME, et al. Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. Nucleic Acids Res. 2024;gkad1167. https://doi.org/10.1093/nar/gkad1167.
40. Baldoni PL, Chen L, Li M, Chen Y, Smyth GK. Dividing out quantification uncertainty enables assessment of differential transcript usage with limma and edgeR. Nucleic Acids Res. 2025. https://doi.org/10.1093/nar/gkaf1305.
41. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. https://doi.org/10.1093/nar/gkv007.
43. Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, et al. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. Nat Methods. 2025;22:801–12. https://doi.org/10.1038/s41592-025-02623-4.
44. Mestre-Tomás J, Liu T, Pardo-Palacios F, Conesa A. SQANTI-SIM: a simulator of controlled transcript novelty for lrRNA-seq benchmark. Genome Biol. 2023;24:286. https://doi.org/10.1186/s13059-023-03127-0.

45. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods. 2024;21:793–7. https://doi.org/10.1038/s41592-024-02229-2.
46. Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM, Public Library of Science. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. PLoS Genet. 2014;10:e1004365. https://doi.org/10.1371/journal.pgen.1004365.
47. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. Transrate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;26:1134–44. https://doi.org/10.1101/gr.196469.115.
48. You Y, Solano A, Lancaster J, David M, Wang C, Su S, et al. Benchmarking long-read RNA-sequencing technologies with LongBench: a cross-platform reference dataset profiling cancer cell lines with bulk and single-cell approaches. bioRxiv; 2025. p. 2025.09.11.675724. https://doi.org/10.1101/2025.09.11.675724. Cited 2025 Oct 7.
49. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8. https://doi.org/10.1038/s41586-019-1186-3.
50. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011;12:R6. https://doi.org/10.1186/gb-2011-12-1-r6.
51. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, et al. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. Genome Biol. 2022;23:10. https://doi.org/10.1186/s13059-021-02588-5.
52. Loving RK, Sullivan DK, Reese F, Rebboah E, Sakr J, Rezaie N, et al. Long-read sequencing transcriptome quantification with lr-kallisto. PLoS Comput Biol. 2025;21:e1013692. https://doi.org/10.1371/journal.pcbi.1013692
53. Pacific Biosciences. Kinnex single-cell RNA: Homo sapiens - PBMC. PacBio Datasets; 2025. https://downloads.pacbcloud.com/public/dataset/Kinnex-single-cell-RNA/
54. Yan F. A comprehensive evaluation of long-read de novo transcriptome assembly. 2024. https://doi.org/10.5281/zenodo.18448217. Cited 2026 Feb 1.
55. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199.
56. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12:2. https://doi.org/10.1186/s13100-020-00230-y.
57. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-87.e29. https://doi.org/10.1016/j.cell.2021.04.048.
58. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell. 2017;170:564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010.
59. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6. https://doi.org/10.1038/nbt.1754.
60. Yan F. Simulation data for benchmarking de novo long read transcriptome assembly software. Zenodo; 2024. https://doi.org/10.5281/zenodo.14263456. Cited 2026 Feb 1.
61. Yan F. De novo transcriptome assemblies for various RNAseq technologies. Zenodo; 2025. https://doi.org/10.5281/zenodo.17538009. Cited 2026 Feb 1.
62. Tian L, Gouil Q, Ritchie ME, Du MRM. Long and short-read transcriptome profiling of human lung cancer cell lines. Gene Expression Omnibus; 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172421. Accessed 1 Feb 2026. Cited 2026 Feb 1.
63. SG-NEx Consortium. Singapore Nanopore Expression Project (SG-NEx). Registry of Open Data on AWS; 2020. https://registry.opendata.aws/sgnex/
64. Gouil Q. short-read RNA-seq of pea stipules. Array Express; 2025. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-14845. Accessed 1 Feb 2026. Cited 2026 Feb 1.
65. Gouil Q. Long-read nanopore PCR-cDNA sequencing of pea stipules. Array Express; 2025. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-14841. Accessed 1 Feb 2026. Cited 2026 Feb 1.

Publisher's Note