Towards accurate, reference-free differential expression: 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], Nadia M. Davidson[1,2,*]

[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
*to whom correspondence should be addressed: Feng Yan yan.a@wehi.edu.au and Nadia M. Davidson davidson.n@wehi.edu.au

Abstract:

Long-read RNA sequencing has significantly advanced transcriptomics by enabling the full length of transcripts to be assessed. However, current analysis methods often depend on a high-quality reference genome and gene annotation. Recently, *de novo* 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.

Here, we comprehensively evaluate the current state-of-the-art 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, including assembly quality and computational efficiency, across a range of datasets, which include simulated data and spike-in sequin transcripts, where ground truth is known, and real data from human cell lines and pea (*Pisum sativum*) samples, using the reference-guided assembler Bambu to define truth. To represent contemporary analysis scenarios, the datasets cover depths from 6 million to 60 million reads, Oxford Nanopore Technologies (ONT) cDNA and direct RNA sequencing, and critically, we assessed the downstream impact of assembly choice on the detection of differential gene and transcript expression.

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 redundancy. Of the *de novo* pipelines, RNA-Bloom2, coupled with Corset for transcript clustering, was the best performing in 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

Introduction

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 (Pertea et al., 2015). This method is highly effective, often enabling the identification of novel transcripts even in well-characterised species like humans and mice (Zhao et al., 2023). 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 (Bushmanova et al., 2019; Grabherr et al., 2011; Nip et al., 2020; Schulz et al., 2012; Xie et al., 2014). Collectively they have facilitated the analyses of thousands of transcriptomes (Hölzer & Marz, 2019; Raghavan et al., 2022), expanding our knowledge of the diversity and complexity of life beyond model species. Although generally applied to non-model organisms (Marlétaz et al., 2018; G.-Q. Zhang et al., 2017; Y. Zhang et al., 2024), *de novo* assembly has also proven beneficial to the analysis of cancer transcriptomes, in

particular in the detection of gene rearrangements (Cmero et al., 2021; Davidson et al., 2015; Gonzalez-Bosquet et al., 2022; Wang et al., 2022).

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. Both technologies offer the promise of higher-quality assembled transcriptomes.

To leverage these advances, new reference-guided assembly methods have been built for long reads including Bambu, StringTie2, IsoQuant and FLAIR (Chen et al., 2023; Kovaka et al., 2019; Prjibelski et al., 2023; Tang et al., 2020). In contrast, few reference-free, *de novo*, assembly methods have been developed. The first effort of such kind was CARNAC-LR and isONclust which cluster reads into genes or gene families (Marchet et al., 2019; Sahlin & Medvedev, 2020). 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 (de la Rubia et al., 2022). 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 (Nip et al., 2023). 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 (Petri & Sahlin, 2023). 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.

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 (Hölzer & Marz, 2019), and informed data analysis workflows (Raghavan et al., 2022). 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 (Dong et al., 2023; Hardwick et al., 2016). 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 StingTie2 overall performed best in isoform detection (Su et al., 2024). 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 (Pardo-Palacios, Wang, et al., 2024). 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 (Sagniez et al., 2024). 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) read, >20k 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 ONT 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, 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 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. 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 (Grabherr et al., 2011; Hölzer & Marz, 2019). In addition, all datasets were assembled with the reference-guided tool Bambu (Chen et al., 2023). 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 (Dong et al., 2023; Su et al., 2024). For datasets without available ground truth, i.e., all except for the simulated and spike-in datasets, Bambu was also used to define the truth set.

For downstream differential expression analysis of *de novo* assembled transcriptomes, we employed Corset (Davidson & Oshlack, 2014) to group transcripts into gene clusters, Minimap2 (Li, 2018) for aligning reads back to the assembled transcriptome, and Salmon (Patro et al., 2017) for transcript abundance quantification with bootstrapping (Figure 1). Transcript and gene-level differential analysis was then conducted using the Bioconductor package limma (Law et al., 2014; Ritchie et al., 2015).

We evaluated all methods across four independent datasets, representing various sequencing technologies, read depths, and organisms (Table 1). 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, and PCR-cDNA Nanopore sequencing from pea (*Pisum sativum)* cultivars, a non-model species (Chen et al., 2021; Dong et al., 2023; Mestre-Tomás et al., 2023). To investigate the impact of coverage and library size, we downsampled the lung cancer cell line dataset at three different read depths. For each dataset we also had

matched short-read data available which we downsampled to the same number of nucleotides as the long-read data, and processed with Trinity (Table S1).

Aside from the pea cultivars, 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 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.

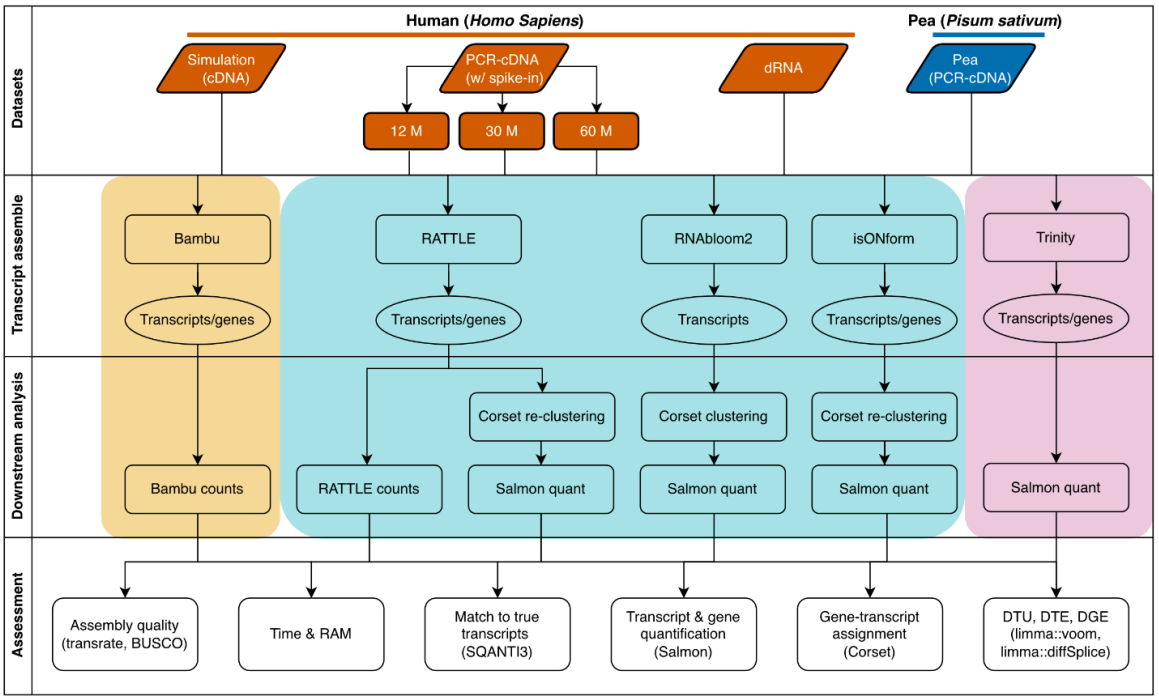

Figure 1: Schematic of the benchmarking study design. *Four datasets were utilized, including simulated unstranded cDNA (simulation), human stranded PCR-cDNA (PCR-cDNA) with sequin spike-ins, human direct RNA (dRNA), and stranded PCR-cDNA from Pisum sativum (Pea). The human PCR-cDNA was down-sampled to three sequencing depths (12, 30 and 60 million reads), and included sequin spike-ins (2% of total reads). Each ONT dataset was assembled using four different methods: the reference-free assemblers RATTLE, RNA-Bloom2, and isONform (teal), and the reference-guided assembler Bambu (orange). Trinity (pink) was applied to matched short-read data. 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.*

Dataset	ONT 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	cDNA (R9.4)	6	7.65	1085 (62-44010)	Simulation	2000, 6933, 2000	New data generated with SQANTI-SIM
PCR-cDNA:Human lung cancer cell lines: HCC827 vs NCI-H1975 + spike-ins	PCR-cDNA (SQK-PCS109, R9.4)	12	6.81	685 (50-71138)	Bambu	5523, 5405, 649	(Dong et al., 2023)
		30			Bambu	7793. 8550, 1203	
		60			Bambu	9481, 11148, 1743	
		2% of total reads			Sequin spike-in	44, 95, 28	
dRNA: SG-NEx human cancer cell lines: MCF7 vs A549	direct RNA (SQK-RNA002, R9.4)	11.6	15.9	1060 (50-16218)	Bambu	6329, 7231, 1639	(Chen et al., 2021)
Pea: (*Pisum sativum*) Thomas Laxton vs Daisy cultivars	PCR-cDNA (SQK-PCB109, R9.4)	16.5	7.75	649 (50-11561)	Bambu	254, 403, 87	In-house new data

Table 1: 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.*

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 1TB of RAM. RATTLE and isONform demonstrated significantly longer assembly times and higher memory usage compared to RNA-Bloom2 (Fig. 2, Table S2). Specifically, RATTLE required between 18 hours 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 hours to 7 days to complete assembly and failed to finish within two weeks on the 60 million reads PCR-cDNA dataset, thus it was not evaluated on that dataset.

In contrast, RNA-Bloom2 was notably faster and more memory-efficient, completing assemblies in 1 to 27 hours 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. As expected, the reference-guided assembler Bambu was the most resource-efficient, requiring the least computational resources across all datasets.

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.

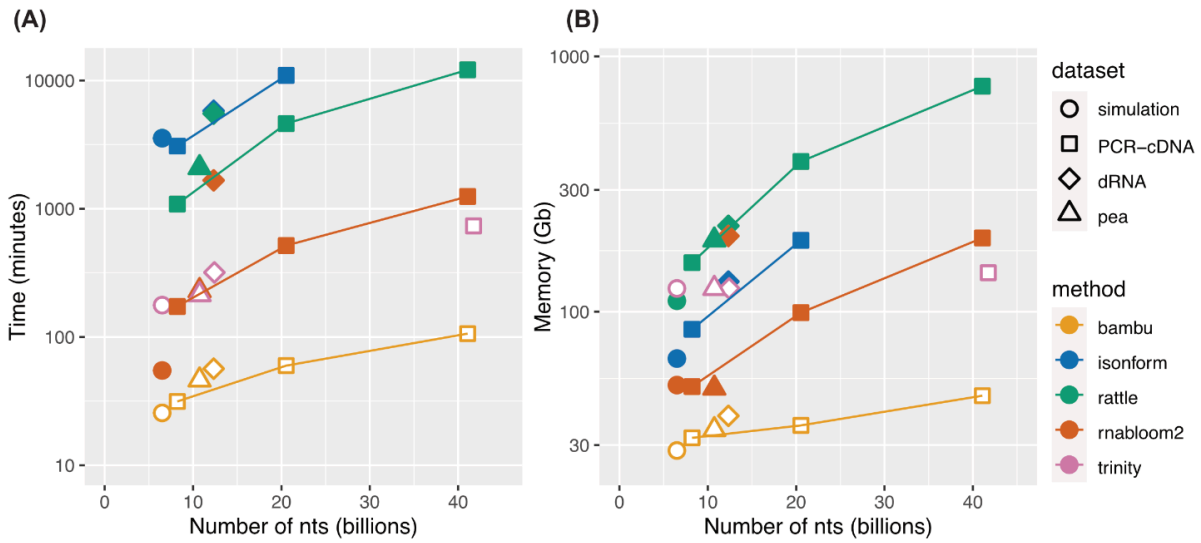

Figure 2: Computational resources required for long read de novo assembly. *(A) Time and (B) memory usage of each assembler as a function of the number of total nucleotides for each dataset. The change in resource usage between the three downsampled depths from the PCR-cDNA dataset are indicated with lines.*

Assembly quality

De novo assemblies from RATTLE, RNA-Bloom2 and isONform, together with Trinity and Bambu were next assessed for their quality, by comparing characteristics of the assembled transcriptome, such as the number, length and recall of reference transcripts (Figure 3). SQANTI3 was used to link assembled transcripts to reference transcripts using splice junctions (Pardo-Palacios, Arzalluz-Luque, et al., 2024). 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 (Aubry et al., 2014; Smith-Unna et al., 2016). As pea is likely to have incomplete gene annotation, we also used BUSCO genes to assess assembly completeness.

As expected, the reference-guided method, Bambu, produced a transcriptome that most resembled the reference gene annotation across all datasets, with few unannotated genes and transcripts (Fig. 3A, S1A, Table S3), a similar transcript length distribution (Fig. 3B), more complete base coverage of the reference transcripts that were assembled (Figure 3C), fewer missing BUSCO genes than other assemblers (Figure 3D), and excellent recall of simulated transcripts (Figure 3E). However, we note that Bambu reported the sequence of the reference transcripts it was provided, and even after filtering of non-expressed transcripts, as was done here, the majority of Bambu's transcripts were identical to the reference. Therefore Bambu could not be directly compared to *de novo* approaches using the reference

gene annotation as truth. However, its assembly quality was informative to set an upper-limit on the best performance expected.

Amongst the *de novo* assembly methods, the number of transcripts assembled varied widely by assembler and dataset (Figure 3A). In the simulation, RNA-Bloom2, RATTLE and Bambu reported a similar number of transcripts (~40000) and genes (~17000), which were close to the true number simulated (40509 transcripts and 18145 genes). However, isONform assembled only 4849 transcripts and 444 genes. Combined with similar poor performance on the dRNA dataset, this 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 kmer and window size. Trinity, applied on the matched short-read simulation, assembled a similar number of genes as other tools, but over 60000 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 (Figure 3B), and fewest complete transcripts (Figure 3C, 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. 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 million 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 (Figure 3A-B).

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 (Figure 3A, S1A). When focusing on assembled transcripts on chrIS, we observed that for 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 (Figure 3A). In addition to false transcripts, we also observed insertion and deletion type errors being incorporated into the sequence of assembled transcripts from long-read data.

This impacted the predicted 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 (Figure S1C, Table S4). These errors were rescued if a reference genome was provided for polishing the long-read transcriptome assemblies (Figure 3D).

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) (Figure 3A, E), consistent with prior studies (Nip et al., 2023; Pardo-Palacios, Wang, et al., 2024). 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 (Figure 3E, S1D-E). Despite these differences, RNA-Bloom2 and RATTLE, had highly similar recall at gene-level, as indicated by the number of recovered reference genes (Figure 3A), transcript base-coverage (Figure 3C), and BUSCO genes (Figure 3D).

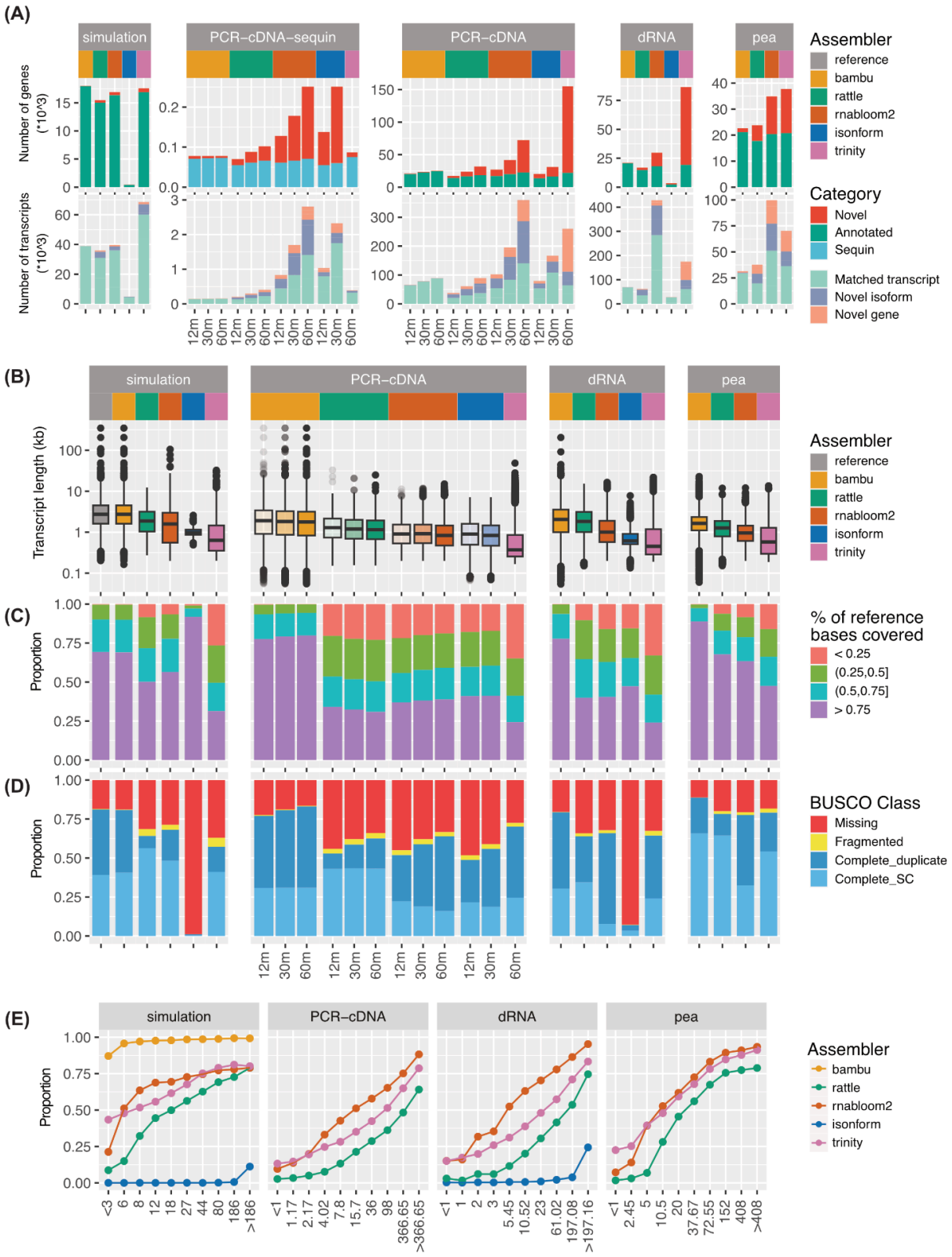

Figure 3: Assessment of assembled transcriptome quality *(A) Number of novel and annotated genes and transcripts in each assembled transcriptome. Transcripts were further divided into matched transcript (complete splice match and incomplete splice match), novel isoform (novel in catalogue and novel not in catalogue) and novel gene (intergenic, genic, intronic, antisense, and fusion).*

(B) Length distribution of assembled and reference transcriptomes.
(C) The proportion of bases recovered for each reference transcript using the Conditional Reciprocal Best BLAST (CRBB) approach.
(D) Proportion of BUSCO genes assembled. Assembled transcripts were corrected using the reference genome to remove errors and enable accurate open-reading-frame 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.

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 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 Bambu counts (Figure 4A).

Amongst the *de novo* methods, RNA-Bloom2 transcript counts were consistently one of the most correlated with truth (Figure 4A, Table S5). IsONform performed the best on the 12 million and 30 million PCR-cDNA 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 (Figure 4B, S2A-B).

RNA-Bloom2 and Trinity tended to have the highest number of clusters, and a large fraction which were redundant, meaning there was more than one cluster matching the same reference gene (Figure 4B, S2B, Table S7). 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 (Figure S2C). 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 (Figure 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 (Figure 4C, Table S5). For simulation, RATTLE was slightly better than RNA-Bloom2 and Trinity, 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 and 30 million PCR-cDNA 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 (Figure 4, S2A-B, Table S6). Based on these results, Corset was used in combination with RNA-Bloom2 to assess downstream differential expression.

Figure 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'. Clusters with transcripts from multiple reference genes were classed as 'mixed'. Clusters with only novel transcripts were classed as 'no match'.
(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.

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) (Baldoni et al., 2024). True changes were defined using either known simulated or sequin spike-in transcripts and genes when available. For other

datasets, we define truth using reference-guided results from 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 (Dong et al., 2023) we did not compare testing methods, but used methods amongst the top performing from prior evaluation, limma::voom and diffSplice (Ritchie et al., 2015) 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 (Figure 5A). All methods were close to the theoretical limit defined by the DGE results from applying limma to the simulated count table. Trinity performed marginally worse than the long-read pipelines. As expected, Bambu had superior performance almost identical to the theoretical limit, justifying its use to define truth on the real datasets. When DGE was evaluated using sequin spike-in genes, Trinity and RNA-Bloom2 marginally outperformed others (Figure 5B). On the PCR-cDNA data, all methods had similar performance (Figure 5C), while in the dRNA and pea data, RNA-Bloom2 performed best (Figure 5D).

For DTE, we observed that only Bambu was close to the theoretical limit. Of the long-read *de novo* assembly methods, RNA-Bloom2 showed the best performance across datasets. Interestingly, the short-read assembler, Trinity 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 4 out of 5 datasets. We also noticed for all methods, increasing read depth in PCR-cDNA data increased the number of true positives among top ranked results for DGE, DTE and DTU (Figure S3). But this effect was larger for DGE than DTE, and to a lesser extent for DTU. In DTE and DTU, the choice of assembler had a larger impact on performance (Figure 5).

Interestingly, in real data, all *de novo* methods missed a significant proportion of the differential transcripts detected by Bambu (black dashed line Figure 5C). 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.

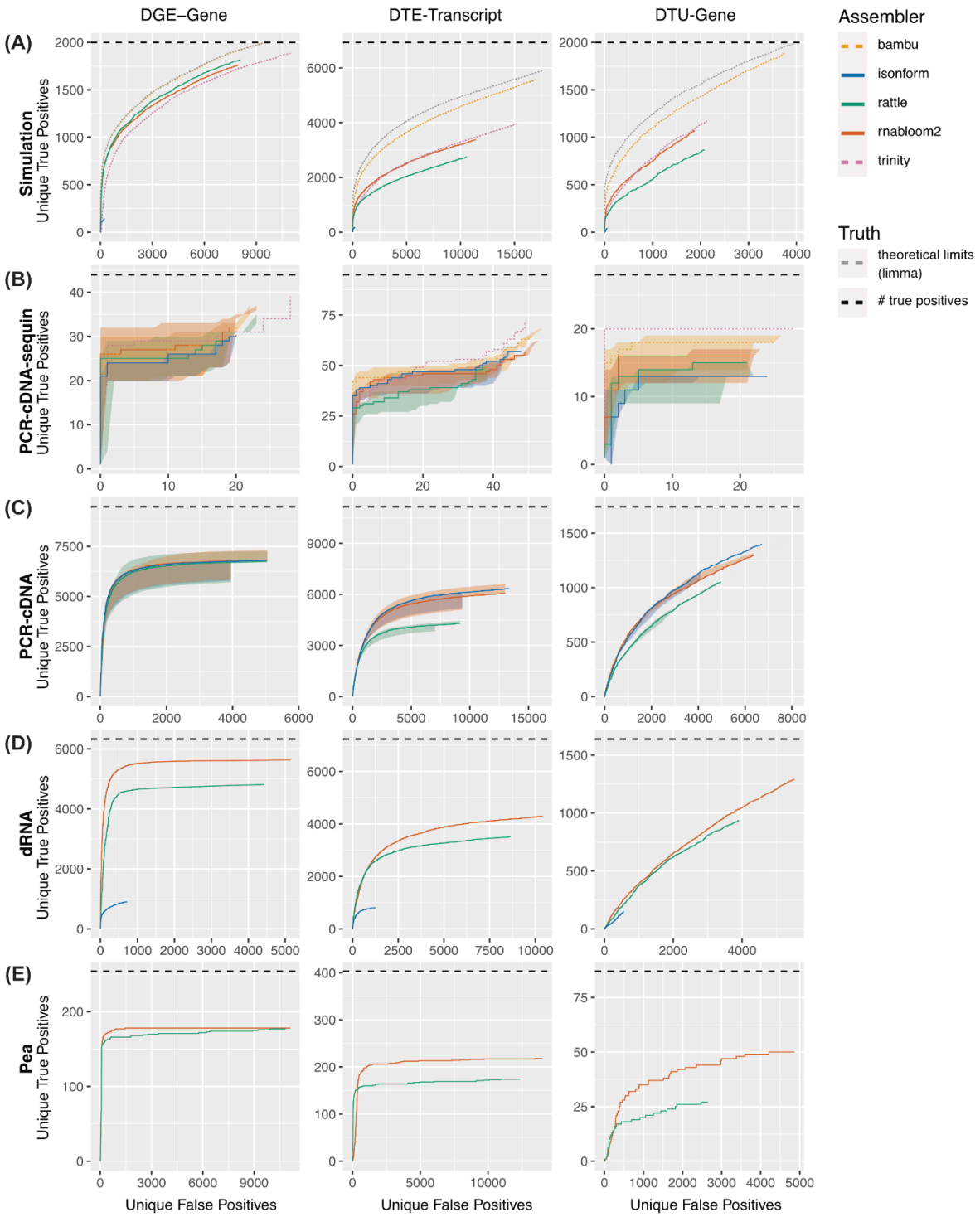

Figure 5: ROC curves for differential analysis. *DGE, DTE and DTU analysis was performed for (A) Simulation, (B) PCR-cDNA sequin, (C) PCR-cDNA, (D) dRNA, and (E) pea dataset. Unique false positives were shown against unique true positives. When multiple clusters match the same reference gene, the cluster with the lowest FDR is retained, hence 'unique' true positives and 'unique' false positives. Truth is defined from simulation and sequin or Bambu differential results for the rest, and the total number of true positives is indicated by the horizontal dashed line (black). For*

simulation (A), the grey dashed curve indicates the theoretical detection limit - defined by running limma on the simulated count matrix. For panels (B-C), the solid line shows the number of true positives and false positives from the 30 million PCR-cDNA dataset, and the shades indicate the ranges of true positives from 12 million, 30 million and 60 million PCR-cDNA data.

De novo assembly captures novel transcripts

De novo transcriptome assembly has enabled 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 genome (Figure 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, among 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) (Table S8-9) (Ghandi et al., 2019). One of these fuses the RPL7 gene to an intergenic region on chromosome 8 in the HCC827 lung cancer cell line (Figure 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 (Figure 6E), and the other is the well-reported fusion BCAS4-BCAS3 (Edgren et al., 2011). To assess whether other fusion genes were seen amongst the novel transcripts, we searched all assembled transcripts for fusions using JAFFAL (Davidson et al., 2022) 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, Table S10) (Ghandi et al., 2019). 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 cell line) (Table S11) (Ghandi et al., 2019).

For the pea dataset, we found 154 DE transcripts with novel sequences, including 7 matching BUSCO genes (Figure 6C, Table S12). 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). 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.

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

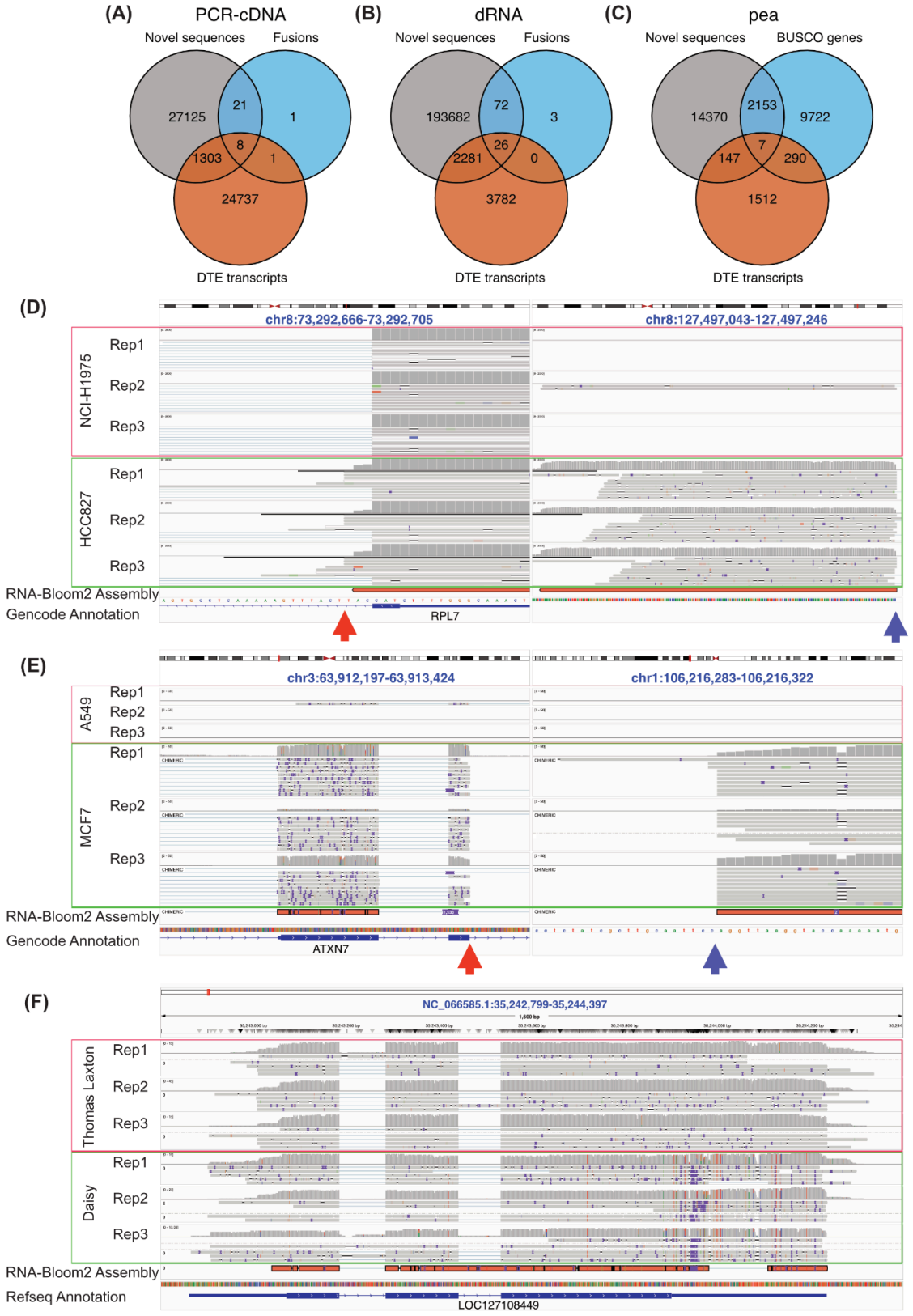

Figure 6: Novel transcripts in de novo assemblies.

(A-B) Venn diagram showing the overlap between differentially expressed transcripts, novel transcripts, and fusion transcripts from CCLE in (A) PCR-cDNA and (B) dRNA.
(C) Venn diagram showing the overlap between differentially expressed transcripts, novel transcripts, and transcripts with 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.

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, alongside the widely used short-read assembler Trinity, and reference-guided tool Bambu. 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.

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. 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 surprising high rate of sequencing errors were retained in long-read *de novo* assembly. This adversely 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 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.

Assembly tools that are highly optimised for PacBio data were not included in our benchmark, nor did we include PacBio data. However, we hope the finding here would also be applicable for PacBio data, which generally has fewer reads and higher base quality, making ONT assembly the more challenging task. 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 (Jousheghani & Patro, 2024; Loving et al., 2024).

We also noted the wider application of *de novo* transcriptome assembly beyond DE analysis. 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. 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.

Conclusion

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 : 9K), 30 g Nitrogen slow-release fertiliser (40N : 0P : 0K), 25 g Water holding granules, 15 g Trace elements (6Mg : 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 minutes, 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 minutes, 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 (18145 genes, 40509 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 2 fold up and 500 genes 2 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 select 1 random transcript to change its expression (500 transcripts 2 fold up and 500 transcripts 2 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 (Baldoni et al., 2024). 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) (Mestre-Tomás et al., 2023). 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. It is available for download at https://doi.org/10.5281/zenodo.14263456.

Nanopore PCR-cDNA, direct-RNA data and short-read RNAseq

To investigate how these 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 replicates, with a mean read length 685 bp), and included a short-read stranded RNAseq with matched number of bases to the 60 million PCR-cDNA data (75bp paired-end, and 41 billion bases) (Dong et al., 2023). 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 RNAseq (Chen et al., 2021). 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 hours. 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 library was sequenced on an Illumina NextSeq 2000 with 75 bp single end reads. All long-read PCR-cDNA data, dRNA data and their matched short read were stranded.

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. Simulated cDNA data were assembled using an unstranded setting, while the PCR-cDNA data from cancer cell lines and pea were preprocessed 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 `--ont` option), isONcorrect (0.1.3.5, default option) 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`). Details can be found in GitHub repository https://github.com/DavidsonGroup/longread_denovo_benchmark .

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 and cell lines, 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 (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 and 1Tb max RAM requested. Assemblies that did not finish in 2 weeks were terminated and excluded.

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 (Smith-Unna et al., 2016). We also run BUSCO (v 5.4.7, database `primates_odb10` and `fabales_odb10`) to obtain the proportion of conserved genes that were assembled (Manni et al., 2021).

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 (Pardo-Palacios, Arzalluz-Luque, et al., 2024). 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. For simulated non-stranded data, all transcripts classified as ‘antisense’ will undergo 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 to align the FASTQ from individual samples to the *de novo* transcriptome (options: `-p 1.0 -N 100`) and then ran salmon (v 1.10.2) on the aligned BAM with options `--ont --numBootstraps 100`. To measure correlation with truth transcript expression, *de novo* transcripts were assigned to reference transcripts based on SQANTI3 results, and 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, Trinity and RATTLE perform clustering of transcripts to genes using their in-built algorithms, we included their transcript to gene cluster information. For all 4 *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`) (Davidson & Oshlack, 2014). 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 cluster from native clustering if applicable, (3) gene cluster defined by SQANTI3 using reference gene annotation. Gene cluster expressions were 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 perfect match, and the rest defined as redundant match.

Assessing differential analysis using *de novo* assembly

We used `edgeR::catchSalmon` 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). We filtered the data to retain transcripts with at least 2 reads in 3 samples and a total of 5 reads, and genes with a minimum of 5 reads in 3 samples. All analyses were performed using the `limma::voom` pipeline after filtering (Baldoni et al., 2024). 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 reference. We also ran limma pipeline directly on the simulated count matrix to indicate the theoretical detection limits. 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 GitHub repository https://github.com/DavidsonGroup/longread_denovo_benchmark .

Novel transcripts in the *de novo* assembly

We ran JAFFAL (v 2.3) on the RNABloom2 assembled transcriptome in FASTA format from the cancer cell lines (PCR-cDNA and dRNA) (Davidson et al., 2022). Known structural variants and fusions from corresponding cell lines were downloaded from the DepMap portal (Public 24Q2 and CCLE2019 dataset) (Ghandi et al., 2019; Tsherniak et al., 2017). Alignment and tracks were visualised in IGV (Robinson et al., 2011).

Data access

The code to reproduce the analysis can be found in GitHub: https://github.com/DavidsonGroup/longread_denovo_benchmark. The simulated data can be downloaded from https://doi.org/10.5281/zenodo.14263456. The PCR-cDNA data from cancer cell lines were obtained from GEO (GSE172421), and dRNA data were obtained from SG-NEx AWS Open Data Registry: https://registry.opendata.aws/sgnex/. The PCR-cDNA data from pea are available upon request.

Conflict of interest

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

Funding

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

Acknowledgement

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.

Author contribution

FY and NMD developed the conceptual design of the benchmarking project. FY performed all analyses, generated figures and wrote the manuscript. NMD contributed to original draft writing and revision of the manuscript. PLB designed the simulation dataset with differential expression. JL and QG collected *Pisum sativum* samples, prepared library, performed sequencing, provided valuable suggestions for the benchmark and manuscript. NMD, MGL and MER supervised analyses. All authors provided feedback and approved the final manuscript.

Reference

Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D., & Hibberd, J. M. (2014). Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLOS Genetics*, *10*(6), e1004365. https://doi.org/10.1371/journal.pgen.1004365

Baldoni, P. L., Chen, Y., Hediyeh-zadeh, S., Liao, Y., Dong, X., Ritchie, M. E., Shi, W., & Smyth, G. K. (2024). Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. *Nucleic Acids Research*, gkad1167. https://doi.org/10.1093/nar/gkad1167

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), giz100. https://doi.org/10.1093/gigascience/giz100

Chen, Y., Davidson, N. M., Wan, Y. K., Patel, H., Yao, F., Low, H. M., Hendra, C., Watten, L., Sim, A., Sawyer, C., Iakovleva, V., Lee, P. L., Xin, L., Ng, H. E. V., Loo, J. M., Ong, X., Ng, H. Q. A., Wang, J., Koh, W. Q. C., … SG-NEx consortium. (2021). *A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines* [Preprint]. Genomics. https://doi.org/10.1101/2021.04.21.440736

Chen, Y., Sim, A., Wan, Y. K., Yeo, K., Lee, J. J. X., Ling, M. H., Love, M. I., & Göke, J. (2023). Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods*, 1–9. https://doi.org/10.1038/s41592-023-01908-w

Cmero, M., Schmidt, B., Majewski, I. J., Ekert, P. G., Oshlack, A., & Davidson, N. M. (2021). MINTIE: Identifying novel structural and splice variants in

transcriptomes using RNA-seq data. *Genome Biology*, *22*(1), 296. https://doi.org/10.1186/s13059-021-02507-8

Davidson, N. M., Chen, Y., Sadras, T., Ryland, G. L., Blombery, P., Ekert, P. G., Göke, J., & Oshlack, A. (2022). JAFFAL: Detecting fusion genes with long-read transcriptome sequencing. *Genome Biology*, *23*(1), 10. https://doi.org/10.1186/s13059-021-02588-5

Davidson, N. M., Majewski, I. J., & Oshlack, A. (2015). JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Medicine*, *7*(1), 43. https://doi.org/10.1186/s13073-015-0167-x

Davidson, N. M., & Oshlack, A. (2014). Corset: Enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biology*, *15*(7), 410. https://doi.org/10.1186/s13059-014-0410-6

de la Rubia, I., Srivastava, A., Xue, W., Indi, J. A., Carbonell-Sala, S., Lagarde, J., Albà, M. M., & Eyras, E. (2022). RATTLE: Reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *Genome Biology*, *23*(1), 153. https://doi.org/10.1186/s13059-022-02715-w

Dong, X., Du, M. R. M., Gouil, Q., Tian, L., Jabbari, J. S., Bowden, R., Baldoni, P. L., Chen, Y., Smyth, G. K., Amarasinghe, S. L., Law, C. W., & Ritchie, M. E. (2023). Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nature Methods*, *20*(11), Article 11. https://doi.org/10.1038/s41592-023-02026-3

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I. H., Nyberg, S., Wolf, M., Borresen-Dale, A.-L., & Kallioniemi, O. (2011). Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology*, *12*(1), R6.

https://doi.org/10.1186/gb-2011-12-1-r6

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3

Gonzalez-Bosquet, J., Cardillo, N. D., Reyes, H. D., Smith, B. J., Leslie, K. K., Bender, D. P., Goodheart, M. J., & Devor, E. J. (2022). Using Genomic Variation to Distinguish Ovarian High-Grade Serous Carcinoma from Benign Fallopian Tubes. *International Journal of Molecular Sciences*, *23*(23), Article 23. https://doi.org/10.3390/ijms232314814

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Hardwick, S. A., Chen, W. Y., Wong, T., Deveson, I. W., Blackburn, J., Andersen, S. B., Nielsen, L. K., Mattick, J. S., & Mercer, T. R. (2016). Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nature Methods*, *13*(9), 792–798. https://doi.org/10.1038/nmeth.3958

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5), giz039. https://doi.org/10.1093/gigascience/giz039

Jousheghani, Z. Z., & Patro, R. (2024). *Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification* (p. 2024.02.28.582591). bioRxiv. https://doi.org/10.1101/2024.02.28.582591

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Loving, R., Sullivan, D. K., Reese, F., Rebboah, E., Sakr, J., Rezaie, N., Liang, H. Y., Filimban, G., Kawauchi, S., Oakes, C., Trout, D., Williams, B. A., MacGregor, G., Wold, B. J., Mortazavi, A., & Pachter, L. (2024). *Long-read sequencing transcriptome quantification with lr-kallisto* (p. 2024.07.19.604364). bioRxiv. https://doi.org/10.1101/2024.07.19.604364

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Marchet, C., Lecompte, L., Silva, C. D., Cruaud, C., Aury, J.-M., Nicolas, J., & Peterlongo, P. (2019). *De novo* clustering of long reads by gene from

transcriptomics data. *Nucleic Acids Research*, *47*(1), e2. https://doi.org/10.1093/nar/gky834

Marlétaz, F., Firbas, P. N., Maeso, I., Tena, J. J., Bogdanovic, O., Perry, M., Wyatt, C. D. R., de la Calle-Mustienes, E., Bertrand, S., Burguera, D., Acemel, R. D., van Heeringen, S. J., Naranjo, S., Herrera-Ubeda, C., Skvortsova, K., Jimenez-Gancedo, S., Aldea, D., Marquez, Y., Buono, L., … Irimia, M. (2018). Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, *564*(7734), 64–70. https://doi.org/10.1038/s41586-018-0734-6

Mestre-Tomás, J., Liu, T., Pardo-Palacios, F., & Conesa, A. (2023). SQANTI-SIM: A simulator of controlled transcript novelty for lrRNA-seq benchmark. *Genome Biology*, *24*(1), 286. https://doi.org/10.1186/s13059-023-03127-0

Nip, K. M., Chiu, R., Yang, C., Chu, J., Mohamadi, H., Warren, R. L., & Birol, I. (2020). RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research*, *30*(8), 1191–1200. https://doi.org/10.1101/gr.260174.119

Nip, K. M., Hafezqorani, S., Gagalova, K. K., Chiu, R., Yang, C., Warren, R. L., & Birol, I. (2023). Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-38553-y

Pardo-Palacios, F. J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., Tseng, E., & Conesa, A. (2024). SQANTI3: Curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods*, *21*(5), 793–797. https://doi.org/10.1038/s41592-024-02229-2

Pardo-Palacios, F. J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S.,

Williams, B., Loveland, J. E., De María, M., Adams, M. S., Balderrama-Gutierrez, G., Behera, A. K., Gonzalez Martinez, J. M., Hunt, T., Lagarde, J., Liang, C. E., Li, H., Meade, M. J., Moraga Amador, D. A., Prjibelski, A. D., … Brooks, A. N. (2024). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Methods*, *21*(7), 1349–1363. https://doi.org/10.1038/s41592-024-02298-3

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Petri, A. J., & Sahlin, K. (2023). isONform: Reference-free transcriptome reconstruction from Oxford Nanopore data. *Bioinformatics*, *39*(Supplement_1), i222–i231. https://doi.org/10.1093/bioinformatics/btad264

Prjibelski, A. D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A. L., & Tilgner, H. U. (2023). Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*, *41*(7), 915–918. https://doi.org/10.1038/s41587-022-01565-y

Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. *Briefings in Bioinformatics*, *23*(2), bbab563. https://doi.org/10.1093/bib/bbab563

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing

and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Sagniez, M., Budhraja, A., Paré, B., Simpson, S. M., Vinet-Ouellette, C., Rozendaal, M., & Smith, M. A. (2024). *Assembly Arena: Benchmarking RNA isoform reconstruction algorithms for nanopore sequencing* (p. 2024.03.21.586080). bioRxiv. https://doi.org/10.1101/2024.03.21.586080

Sahlin, K., & Medvedev, P. (2020). De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. *Journal of Computational Biology*, *27*(4), 472–484. https://doi.org/10.1089/cmb.2019.0299

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. https://doi.org/10.1101/gr.196469.115

Su, Y., Yu, Z., Jin, S., Ai, Z., Yuan, R., Chen, X., Xue, Z., Guo, Y., Chen, D., Liang, H., Liu, Z., & Liu, W. (2024). Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. *Nature Communications*, *15*(1), 3972. https://doi.org/10.1038/s41467-024-48117-3

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-15171-6

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, *170*(3), 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010

Wang, Y., Xue, H., Aglave, M., Lainé, A., Gallopin, M., & Gautheret, D. (2022). The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma. *NAR Cancer*, *4*(1), zcac001. https://doi.org/10.1093/narcan/zcac001

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S., & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., Wang, J.-Y., Lin, Y.-C., Xu, Q., Chen, L.-J., Yoshida, K., Fujiwara, S., Wang, Z.-W., Zhang, Y.-Q., Mitsuda, N., Wang, M., Liu, G.-H., Pecoraro, L., Huang, H.-X., … Liu, Z.-J. (2017). The Apostasia genome and the evolution of orchids. *Nature*, *549*(7672), 379–383. https://doi.org/10.1038/nature23897

Zhang, Y., Chen, H., Lian, C., Cao, L., Guo, Y., Wang, M., Zhong, Z., Li, M., Zhang,

H., & Li, C. (2024). Insights into phage-bacteria interaction in cold seep Gigantidas platifrons through metagenomics and transcriptome analyses. *Scientific Reports*, *14*(1), 10540. https://doi.org/10.1038/s41598-024-61272-3

Zhao, Z., Chen, Y., Cheng, X., Huang, L., Wen, H., Xu, Q., Zhou, X., Zhang, X., Chen, J., & Ni, T. (2023). The landscape of cryptic antisense transcription in human cancers reveals an oncogenic noncoding RNA in lung cancer. *Science Advances*, *9*(14), eadf3264. https://doi.org/10.1126/sciadv.adf3264