Assembly Arena: Benchmarking RNA isoform reconstruction algorithms for nanopore sequencing

Mélanie Sagniez[1,2*], Anshul Budhraja[1,2*], Bastien Paré[1,2], Shawn M. Simpson[1*], Clément Vinet-Ouellette[2], Marieke Rozendaal[1], Martin A. Smith[1-3#]

[1] CHU Sainte-Justine Research Center, 3175 Chem. de la Côte-Sainte-Catherine, Montréal, H3S 2G4, Québec, Canada.

[2] Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, 900 Édouard Montpetit Blvd, Montreal, H3T 1J4, Québec, Canada.

[3] School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW Sydney, High Street, Kensington, New South Wales 2052, Australia

[*] Contributed equally

[#] Corresponding author: martin.smith+pub@unsw.edu.au

Abstract

Resolving the transcriptomes of higher eukaryotes is more tangible with the advent of long read sequencing, which greatly facilitates the identification of new transcripts and their splicing isoforms. However, the computational analysis of long read RNA sequencing data remains challenging as it is difficult to disentangle technical artifacts from *bona fide* biological information. To address this, we evaluated the performance of multiple leading transcriptome assembly algorithms on their ability to accurately reconstruct RNA transcript isoforms. We specifically focused on deep nanopore sequencing of synthetic RNA spike-in controls (Sequins™ and SIRVs) across different chemistries, including cDNA and direct RNA protocols. Our systematic comparative benchmarking exposes the strengths and limitations of the different surveyed strategies. We also highlight conceptual and technical challenges with the annotation of transcriptomes and the formalization of assembly quality metrics. Our results complement similar recent endeavors, helping forge a path towards a gold standard analytical pipeline for long read transcriptome assembly.

Introduction

Long read sequencing greatly improves transcriptome profiling via its ability to qualify full-length RNAs, helping resolve complete exon chains and unannotated biological features, such as long non-coding RNAs, which often contain repetitive sequences (Byrne et al. 2019; Lagarde et al. 2017). These technologies present a distinct advantage over short-read RNA sequencing (RNA-seq) that require a well-annotated reference and struggle to resolve long-range splicing dependencies. The utilization of long reads holds promise for expanding the repertoire of known RNA transcripts, improving the quantification genes and splicing at the isoform level while contributing to a more detailed portrait of the transcriptome.

Recent studies have compared software for estimating transcript abundances in long read sequencing data (Dong et al. 2023; Pardo-Palacios et al. 2023b) but few reports have detailed the effectiveness of tools for RNA isoform discovery and qualification using long reads. Transcriptome

assembly is an economical method for characterizing gene expression of model (Yang et al. 2021; Kuo et al. 2020) or non-model organisms (Kang et al. 2021; Zagorščak and Petek 2021; Matra et al. 2023) and for identifying novel RNAs in diseases such as cancer (Fang et al. 2021; Verma et al. 2015).

While the definition of de novo assembly encompasses many meanings, we categorize the existing assemblers into 3 distinct overall strategies: (i) guided strategies, which combine sequencing data aligned to a reference genome and an existing reference gene annotation; (ii) *de novo* strategies thah rely solely on reads aligned to a reference genome to reconstruct a transcriptome; and (iii) *ab initio* strategies, which require only sequencing reads (often used for non-model species without a genome assembly). We compared 12 tools that assemble transcripts from long reads according to one (or more) of the above mentioned categories. Some tools (i.e. Bambu, Flair, isoQuant, and Stringtie2) fit into two strategies–namely 'guided' and '*de novo*'–as they allow the optimal input of a reference transcriptome annotation, such as RefSeq (O'Leary et al. 2016) or Gencode (Frankish et al. 2019), typically in gene transfer file (.gtf) file format.

Bambu (Chen et al. 2023) is a "context aware quantification tool" for long reads. It is an R package that works by using genome-aligned reads (in .bam format) to quantify known and putative novel genes. Bambu is intended to be used as a transcript quantification tool and not a transcriptome assembler. However, it has been included herein as it has been reported in recent studies considering novel transcript identification (Dong et al. 2023; Pardo-Palacios et al. 2023b). The Bambu pipeline has a feature that allows for *de novo* transcript discovery by estimating a novel discovery rate (NDR) which optimizes the false discovery rate in a reproducible manner across samples and analyses. The algorithm annotates reads using assigned read-classes to label splice junctions as full or partial overlaps, thereby producing an intermediate annotation output for the transcriptome. When provided with a transcriptomic annotation (the above-mentioned 'guided mode'), Bambu outputs all isoforms of that given annotation while adding discovered transcripts below the NDR limit.

Full-Length Alternative Isoform analysis of RNA (or FLAIR) (Tang et al. 2020) is a python workflow created for differential isoform expression analysis that also includes a transcriptome assembly module. It can integrate matched short-reads alongside long-reads to correct errors in the latter and improve accuracy. FLAIR organizes transcripts into groups based on splicing patterns, then further groups these reads by analyzing transcription start sites (TSSs) and end sites (TESs) separately. This is followed by collapsing the grouped reads into the most common isoform within each window for TSS and TES, respectively. FLAIR then aligns these reads to the initial isoform groups. Each putative RNA isoform is then considered to be a valid and observed transcript depending on its observed coverage. FLAIR is composed of several modules that can be run consecutively or independently; thus enabling both *de novo* and *guided* assembly strategies.

Full-Length Analysis of Mutations and Splicing (FLAMES) is a long-read transcriptome analysis tool developed for single-cell analyses that also includes a pipeline for bulk RNA-seq analysis (Tian et al. 2021). FLAMES initially groups reads in a similar manner to FLAIR. However, it introduces a unique approach to modeling the probability of read truncation by incorporating a linear model of isoform length, which leverages the concept that longer isoforms are more likely to have truncated reads with incomplete 5′/3′ ends (especially in the case of single cell RNA-seq droplet-capture protocols where low input material and multiple rounds of PCR are employed). Following this, the sequence of each polished transcript serves as an updated reference for direct realignment of input reads using Minimap2 (Li 2021). During this realignment, transcripts with insufficient coverage are discarded. Reads are then assigned to transcripts based on alignment scores, fractions of reads aligned, and transcript coverage, resulting in an output of isoform-level counts.

IsoQuant (Prjibelski et al. 2023) is a long read transcript annotation tool, written in python, that uses intron graphs to reconstruct transcripts with and without a reference annotation. In the graph, two vertices of splice junctions are connected by a directed edge—if two splice sites are found one after the other in a read, they will be connected in the graph. In the case where a reference annotation is provided, reads are assigned to known isoforms by approximate intron-chain matching to account for error-prone reads. Like Bambu, when provided with a reference annotation, IsoQuant outputs all

transcripts from the reference in addition to new transcripts, as well as another assembly exclusively containing isoforms expressed in the input data.

Mandalorian (Volden et al. 2023) is a long-read transcript identification tool also tested in the LRGASP benchmarking project (Pardo-Palacios et al. 2023b). It takes as input accurate full-length transcriptome sequencing data (adaptor and poly-A tail trimmed as well as reoriented reads) and is composed of 5 modules, by default, sequentially executed: (i) alignment to genome; (ii) alignment clean-up; (iii) locus by locus grouping and consensus assembly; (iv) alignment to generate isoform models and group by gene; (v) quantification. Mandalorion is reported to be among the strongest methods for transcript discovery but was only tested on PacBio HiFi and Oxford Nanopore R2C2 reads (both of which employ consensus base calling) for increased sequence accuracy (Pardo-Palacios et al. 2023b). Consequently, the authors of Mandalorion have only tested their tool on such data and recommend using Flair for regular nanopore reads.

StringTie (Pertea et al. 2015) is a *de novo* transcriptome assembly and quantification tool written in C/C++ that operates on genome-aligned files from short- or long-read RNA-seq data. The tool implements a network flow algorithm on the assembled contiguous stretches of transcripts called ‘super-reads’. This algorithm determines the path with the maximum coverage, which allows for adjustments to abundance estimations and produces a more fitting transcript abundance result. The most recent update to the tool, StringTie2, has reduced memory requirements and removed a great percentage of false-positive spliced alignments (Kovaka et al. 2019).

Technology Agnostic LOng-read aNalysis (or TALON) (Wyman et al. 2020) is an annotation dependent tool that analyzes long-read transcriptomes across datasets. The recommended first step in the pipeline is to correct the non-canonical junctions using the python tool “TranscriptClean” (Wyman and Mortazavi 2019). This is followed by the python and R based TALON framework, which involves labeling potential internal priming events, annotating transcript models, recording the transcript abundance, filtering transcript models using biological replicates and ultimately generating a custom transcriptome, used to produce the final transcript and gene count tables. An isoform is considered

only when near-alike reads are found above a certain threshold count, in multiple replicates. TALON does not use read clustering or a consensus sequence in an attempt to prevent introduction of incorrect input for its transcript model. While capable of processing a single sample, TALON is meant to process technical/biological replicates (Amarasinghe et al. 2020)–a common occurrence in RNA-seq analyses. TALON has been known to produce a few fallacious transcript predictions in simulated long-read data (Kuo et al. 2020).

Ab initio tools are not as commonly used as *de novo* or *guided* ones and, therefore, have not been subject to thorough independent benchmarking. RATTLE is a 'reference-free reconstructor and quantifier of transcriptomes' as described by the developers of the C++ and python based tool (de la Rubia et al. 2022). It works by clustering raw reads using a two-step greedy approach, which produces read clusters that should belong to the same gene. These 'gene clusters' are further classified into 'transcript clusters' by identifying internal gaps in the read sequences. Error correction and polishing of these clusters produces consensus sequences and matched abundance estimations. Using simulated and spike-in synthetic controls, RATTLE's reference-free assembly algorithm seems to perform better with dRNA data than cDNA(de la Rubia et al. 2022). Although RATTLE seems to perform on par with reference-based assemblers, it is unable to distinguish different isoforms when exon sizes are below a certain limit (de la Rubia et al. 2022). The authors have favorably compared RATTLE to similar clustering algorithms, namely CARNAC-LR and isONclust.

IsONclust (Sahlin and Medvedev 2020) is a greedy search algorithm, implemented in python, that uses minimizers in order to determine similarity and cluster reads together. This approach follows three simple steps for each read: (i) assigning certain reads as 'representative reads' of potential clusters and maintaining a hash table of these reads, queried using minimizers; (ii) assigning reads to the cluster with maximal shared minimizers; (iii) conditionally assigning other reads to existing clusters if the similarity is above a certain threshold, otherwise creating a new cluster with the unclassified read as a new representative. Oxford Nanopore has optimized this program and produced the C++ implementation, isONclust2 (https://github.com/nanoporetech/isONclust2) . IsONclust2 was made in collaboration with the authors of isONclust (Kristoffer Sahlin and Paul Medvedev), which

resulted in a tool with improved efficiency that takes strandedness into consideration, as well as additional features.

RNA-Bloom (Nip et al. 2020) is a Java based assembly algorithm originally designed for single-cell short-read transcriptome assembly combining bloom filters with De Bruijn graph assembly. This approach to *ab initio* transcriptome assembly was improved upon and applied to long-read bulk sequencing data with RNA-Bloom2 (Nip et al. 2023). The latter uses digital normalization with strobemers–a strategy reported to be less sensitive to mutations than k-mers–before assembling and polishing unitigs (Sahlin 2021). The authors have compared RNA-Bloom2 to RATTLE and found their tool to be faster, less memory intensive while reaching a higher recall and a lower false discovery rate (Nip et al. 2023).

The diversity of recently developed transcriptome assembly tools and the increasing adoption of long read sequencing technologies reinforce the need to distinguish between analytical artifacts (e.g., erroneous isoform detection) and experimental ones (e.g., fragmented RNA, reverse transcription drop off, internal priming, etc). Here, we describe a comprehensive and independent comparative performance benchmark of the aforementioned transcriptome assembly tools. We assessed algorithmic performance with two tools for comparative isoform qualification, SQANTI3 (Tardaguila et al. 2018) and GFFcompare (Pertea and Pertea 2020). SQANTI3 has recently been employed to evaluate long-read transcriptome tools in the Long-read RNA-Seq Genome Annotation Assessment Project (LR-GASP) study (Pardo-Palacios et al. 2023b). GFFcompare derives from Cuffcompare, a utility program from the well-known RNA-seq analysis suite Cufflinks (Pertea and Pertea 2020). We present a comprehensive overview of the strengths and weaknesses of *de novo* transcriptome assemblers for long reads, with the ultimate objective of helping the scientific community make informed decisions when selecting the most appropriate strategy for their isoform-level transcriptome analyses.

Results

Transcriptome assembly is a crucial step in deciphering the functional landscape of a cell or organism. We evaluated most available long read transcriptome assembly algorithms encompassing 3 assembly paradigms (*guided*, *de novo* & *ab initio*) using two distinct *in vitro* transcribed RNA spike-in controls, namely Sequins™ (Hardwick et al. 2016) and Spike-in RNA variants (or SIRVs) (Lexogen SIRV Set 4), and four distinct nanopore sequencing chemistries, including the latest direct RNA sequencing protocol (**Table 1**, **Figure 1**).

Table 1. Sequencing data

Dataset	Input	Library preparation protocol	Flowcell & ID	Library molarity	Sequencer	Runtime	Basecalling*	Reads (Pass/Fail)	Input number of reads
LSK109_ sequins	15ng RNA Sequins™ mixA	PCS109 RT protocol + LSK109	R9.4.1 FAN54376	50 fmol	GridION	72h	Guppy6.0.1_sup	6,455,950 P+F	5,2912,936
LSK114_ sequins	15ng RNA Sequins™ mixA	PCS109 RT protocol + LSK114	R10.4.1 FAV99142	20 fmol	PromethION	12h	Guppy6.4.1_sup	5,112,902 P+F	4,262,005
LSK114_ SIRVs	0.0375ng SIRV set-04 + 125ng human RNA	PCS109 RT protocol + LSK114	R10.4.1 PAK95982	18 fmol	PromethION	72 hours, wash + reload (18 fmol) after 29h	Guppy6.5.7_sup	42,764** P+F	42,764
RNA002 _sequins	300ng RNA Sequins™ mixA	RNA002	R9.4.1 PAM58324	20 fmol	PromethION	89h30m	Guppy6.0.7_hac	877,710 P+F	877,710
RNA004 _sequins	4 samples: 15ng RNA Sequins™ mixA 1.5µg human RNA	RNA004_beta-v2	RP4_beta-testing PAO83093 PAO83456 PAO84072 PAO96683	20 fmol	PromethION	68h, wash + reload (20 fmol) after 44h15m	Guppy6.4.6_sup	162,644 P	162,644

*sup = super high accuracy ; hac = high accuracy

**Sequins™ reads retrieved (full-length + rescued), after Pychopper was executed on human + Sequins™/SIRV reads

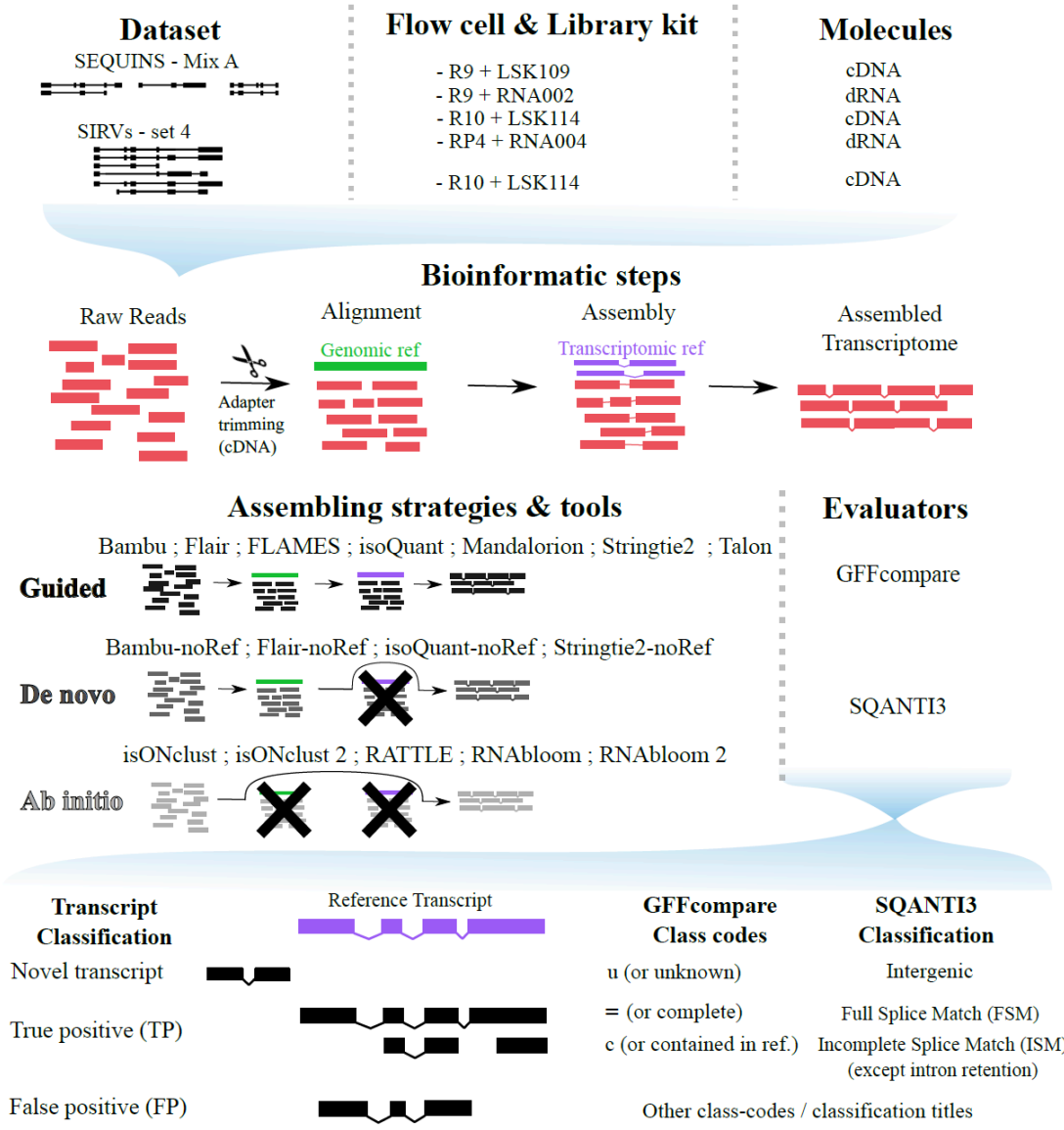

Figure 1. Overview of the study design for cDNA and dRNA assembly generation and quality assessment. LSK: Ligation sequencing kit; cDNA: complementary DNA; dRNA: direct RNA.

As qualitative assessment of assemblies is somewhat subjective, we opted to utilize and compare two different tools with distinct yet similar classification categories and scoring metrics, namely GFFcompare and SQANTI3. These were used to compare the assembled isoforms against 'truth set' annotations provided by the respective manufacturers of the spike-in controls. N.B. biological samples were purposefully omitted herein given their lack of an absolute truth and the difficulty in distinguishing biological variation from analytical artifacts. Because long-read

sequencing sometimes results in low confidence sequences at the 3' and 5' ends, we defined a True Positive as a perfect exon-chain match or contained in reference ("c" classcode for GFFcompare and "ISM" for SQANTI3). Precision and sensitivity were calculated for each tool and chemistry combination based on individual classifications of each assembled transcript (see Methods for details). In this context, precision reflects the integrity of the transcriptome, while sensitivity reflects its completeness.

As expected, *guided* strategies perform better than others in sensitivity and/or precision (**Figure 2 & Supplemental File 1**). It is important to note that the most highly performing tool, Bambu, systematically outputs all transcripts from the reference provided to it, in addition to the potential *de novo* transcripts that it assembles, hence surpassing every other tool in both precision and sensitivity in all datasets. Isoquant and Stringtie2 tend to perform comparably, with IsoQuant a step ahead in precision with the Sequins™ controls. Interestingly, the performance of reference-guided IsoQuant appears much inferior on the SIRV controls, which generally contrast less gene diversity but more alternative splicing isoforms than Sequins™. However, IsoQuant consistently performs well across both cDNA and dRNA. Unlike Bambu, which systematically outputs all transcripts irrespective of their expression status, IsoQuant provides the option to only produce transcripts demonstrably expressed within the input dataset (similar to Stringtie2), which was considered herein. Stringtie2 also performs decently in its *guided* mode, although its accuracy decreases notably in dRNA datasets (RNA002 and RNA004 in Figure 2B). For TALON, incorporating the authors' recommendations for parameters (providing biological/technical replicates; labeled as TALON_reco in Figure 2) significantly improved precision, which nonetheless falls short of most other tools despite decent sensitivity. Among other reference-guided tools, FLAMES displays poor sensitivity but the assemblies, albeit incomplete, remain accurate, presumably due to a stringent filtering step in the pipeline (see Methods for further details). FLAIR displays reasonable precision but ranks relatively poorly overall in *guided* mode.

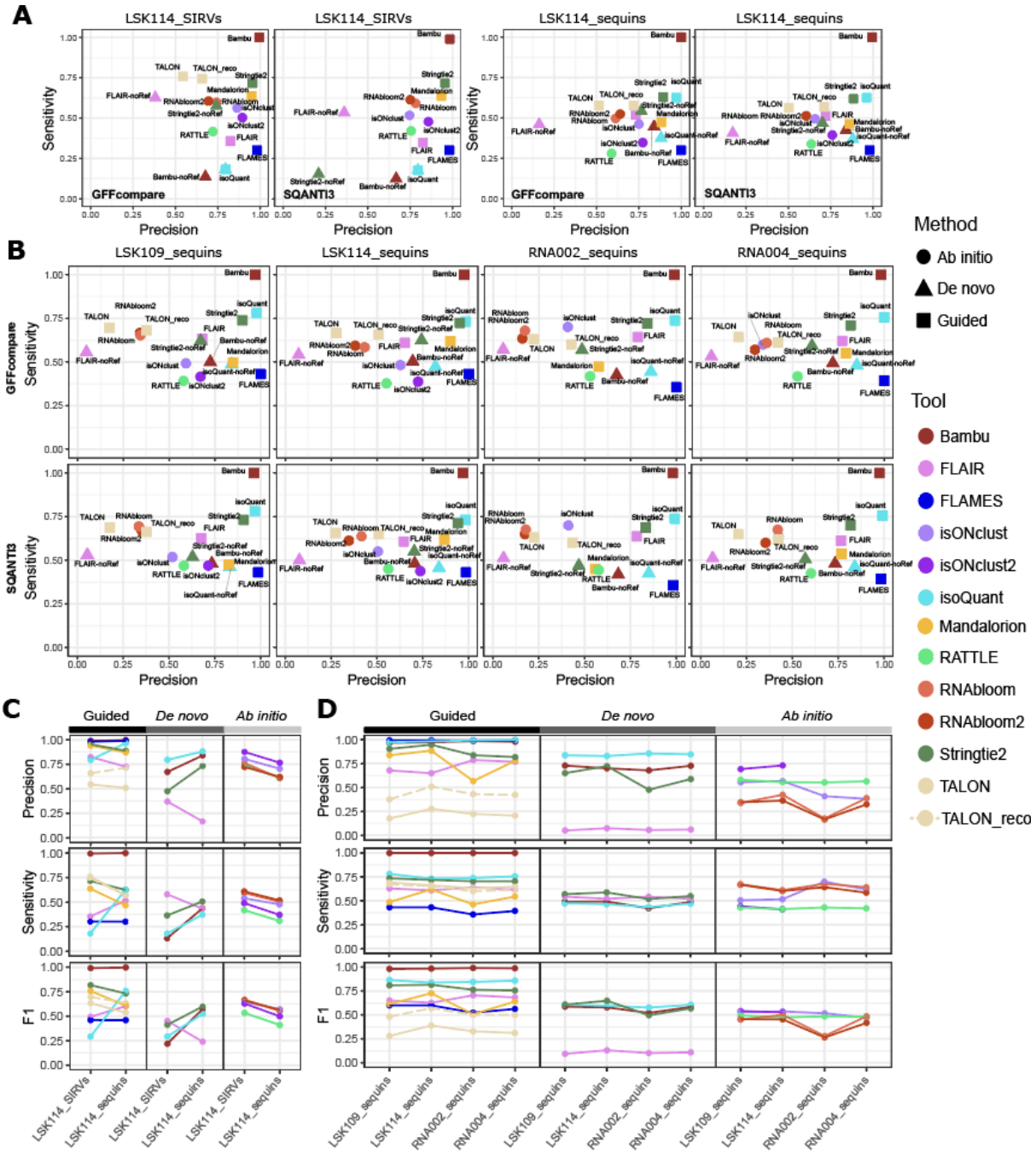

Figure 2. Precision, sensitivity and F1 for all sampled datasets and algorithms. (A) Precision and sensitivity values calculated as per GFFcompare and SQANTI3 classifications for assembly outputs from both 40,000 reads from SIRV (left panels) and Sequins™ (right panels) with cDNA+PCR ligation protocol LSK114. **(B)** Precision and sensitivity values calculated as per GFFcompare and SQANTI3 classifications for all Sequins™ dataset assembly outputs available generated from 150,000 input reads. **(C)** Mean precision, sensitivity and F1 values calculated from GFFcompare and SQANTI3 classifications for the same datasets as (A). **(D)** Mean precision, sensitivity and F1 values

calculated from GFFcompare and SQANTI3 classifications for the same datasets as (B). For individual values, in (C) & (D) see **Supplemental Figure 1**.

De novo methods generally have lower precision and sensitivity compared to *guided* methodologies, even when employing the same software. Here, IsoQuant's *de novo* approach (labeled as IsoQuant-noRef in **Figure 2**) displays better precision than Stringtie2 but lower sensitivity. Stringtie2 (Stringtie2-noRef) has better overall sensitivity than the other *de novo* tools, especially in cDNA (LSK109 & LSK114 in **Figure 2B**). Although Bambu's *de novo* mode has a drastic drop in precision and sensitivity when compared to *guided* mode, it performs slightly better than Stringtie2-noRef (especially in dRNA; **Figure 2B**). FLAIR in *de novo* mode regrettably does not instill confidence in any tested modality, highlighting an issue with a very high false positive rate. Interestingly, all *de novo* methods performed relatively poorly compared to *guided* and *ab initio* assembly modalities on the LSK114 SIRV dataset, suggesting that *guided* and *ab initio* strategies might be most suitable for resolving isoforms in genes presenting complex alternative splicing.

The *ab initio* tools RNAbloom and RNAbloom2 present low precision but relatively high sensitivity, akin to the *guided* version of TALON. A sharp eye will notice performance values differ between **Figure 2A** & **2B**, indicating that depth of sequencing can affect assembly precision (see below). The poorest performer among *ab initio* methods appears to be RATTLE, with low sensitivity and precision slightly above 50%. Despite precision values of ~50-60%, RATTLE outperforms other *ab initio* methods for precision on dRNA, while isONclust2 is the most precise for cDNA (N.B. it does not process dRNA)(**Figure 2B**). Despite the high precision of isONclust2, it has a low sensitivity of ~50%, while the previous version (isONclust) has a higher sensitivity ranging between 50-75% but lower precision.

Noting the differences between the latest (LSK114) and earlier (LSK109) ONT cDNA library technologies (**Figure 2D**) one can observe the improvement in average precision of the TALON_reco, Stringtie2 (both, the guided and *de novo* versions), RNAbloom and RNAbloom2 with the latest technology, which presents less sequencing errors. The precision and sensitivity shown in **Figure 2C**

& **2D** are the mean of values generated from GFFcompare and SQANTI3's TP, FP and FN classifications, which are shown distinctly in **Supplemental Figure 1**. Similar precision improvements are observed for RNAbloom and RNAbloom2 with the latest version of the dRNA chemistry (RNA004) versus the older version (RNA002). Certain tools, such as Mandalorian and Stringtie2-noRef, have better average sensitivity and precision, respectively, with RNA004 compared to RNA002 (**Figure 2D**). FLAMES is also affected in average sensitivity measures for the RNA002 dataset, although to a lesser extent than the gap between Mandalorian's performance between the chemistries.

Besides the performance of individual tools, an interesting trend is observed concerning the choice of reference material: Methods applied to SIRV data present a broader range of sensitivity versus a broader range of specificity for the Sequins™ data (**Figure 1A**). This may reflect the different composition of the mixtures, with SIRVs containing relatively less genes (and more isoforms per gene) than Sequins™. Control dependent differences are also observed for individual tools (**Figure 2A,C**). There is a clear decrease in sensitivity and overall F1 score for isoQuant (*guided*) when applied to the SIRV data, with Flair (*guided*) following a similar pattern. Mandalorion, Stringtie2, TALON and TALON_reco seem to have improved sensitivity and F1 score with SIRV versus Sequins™. Among the *de novo* strategies, we can observe a dramatic decrease in the precision and sensitivity of Bambu, isoQuant and Stringtie2 for the SIRV dataset, while Flair follows the opposite trend (corroborated in **Figure 2A,C**). There is an improvement in performance of all *ab initio* tools when used on the SIRV dataset (**Figure 2C;** absolute values in **Supplemental File 1**), suggesting that different transcriptomic features will globally influence accuracy. This is exemplified with RNAbloom and RNAbloom2, whose accuracies are highly influenced by input parameters. N.B. TALON and TALON_reco pipelines produced no results when evaluated on SIRV with SQANTI3 (**Figure 1A**) due to issues with SQANTI3 handling of the strandedness in the assembly.

Overall, the F1 score of reference-guided tools remains fairly consistent across datasets except IsoQuant, which performs a lot better on Sequins™ than SIRVs (75% and 25%, respectively; **Figure 2C**). We observed that *de novo* and *ab initio* methods have a very similar predictive

performance overall (average F1 score, **Figure 2C,D**), with RNA002 and FLAIR as notable exceptions.

Significant differences in performance were observed between many assembly tools that can be attributed to the method used for qualitative evaluation. Stringtie2-noRef performs at 55% and 75% in sensitivity and precision, respectively, when evaluated with GFFcompare but both metrics fall to 15% with SQANTI3 for the same assembly (**Supplemental Figure 1A,B**). This suggests that the choice of quality metrics used to benchmark assembly and isoform detection performance can have a drastic impact on the resulting assembly strategy, while highlighting that isoform class definitions is a somewhat subjective evaluation.

These comparisons were performed with static read depths to directly compare assembly accuracies in a normalized manner. Since depth of sequencing can impact assembly accuracy by increased sampling of lower abundance transcripts and experimental artifacts, we evaluate the performance of representative tools on different sampling depths of the LSK114_sequins dataset, from 1,000 reads to 4.3 million reads (**Figure 3**). Overall, higher depth of sequencing is associated with increased sensitivity and lower specificity, with the notable exception of *guided* assembly modalities. Bambu, in particular, outputs all transcripts from the reference, yielding a sensitivity above the maximal theoretical value. Both IsoQuant and Stringtie2 approach 100% sensitivity and 75% precision at maximal depth, irregardless of the scoring metric. However, the *de novo* versions of Bambu and IsoQuant stagnate after 100k and 500k reads, respectively, evolving within the 48-55% sensitivity and 60-75% precision ranges (**Figure 3**, second column). In contrast, *de novo* Stringtie2's sensitivity scales with sampling depth. Of the two *ab initio* methods we tested, isONclust2 presents relatively stable precision in function of depth, with comparable if not superior performance to the 3 *de novo* methods in the middle panel. A drastic increase in precision for the isONclust result at 500k reads, evaluated using SQANTI3, presents an unusual outlier that we could not dismiss. Overall, the surveyed *ab initio* and *de novo assembly* modalities display a significantly lower sensitivity than theoretically expected, suggesting that the use of a minimum coverage threshold might restrict isoform discovery. Interestingly, *ab initio* methods appear to perform slightly better than *de novo*

strategies when evaluating with SQANTI3 (at least, for sensitivity), while the converse is observed with GFFcompare.

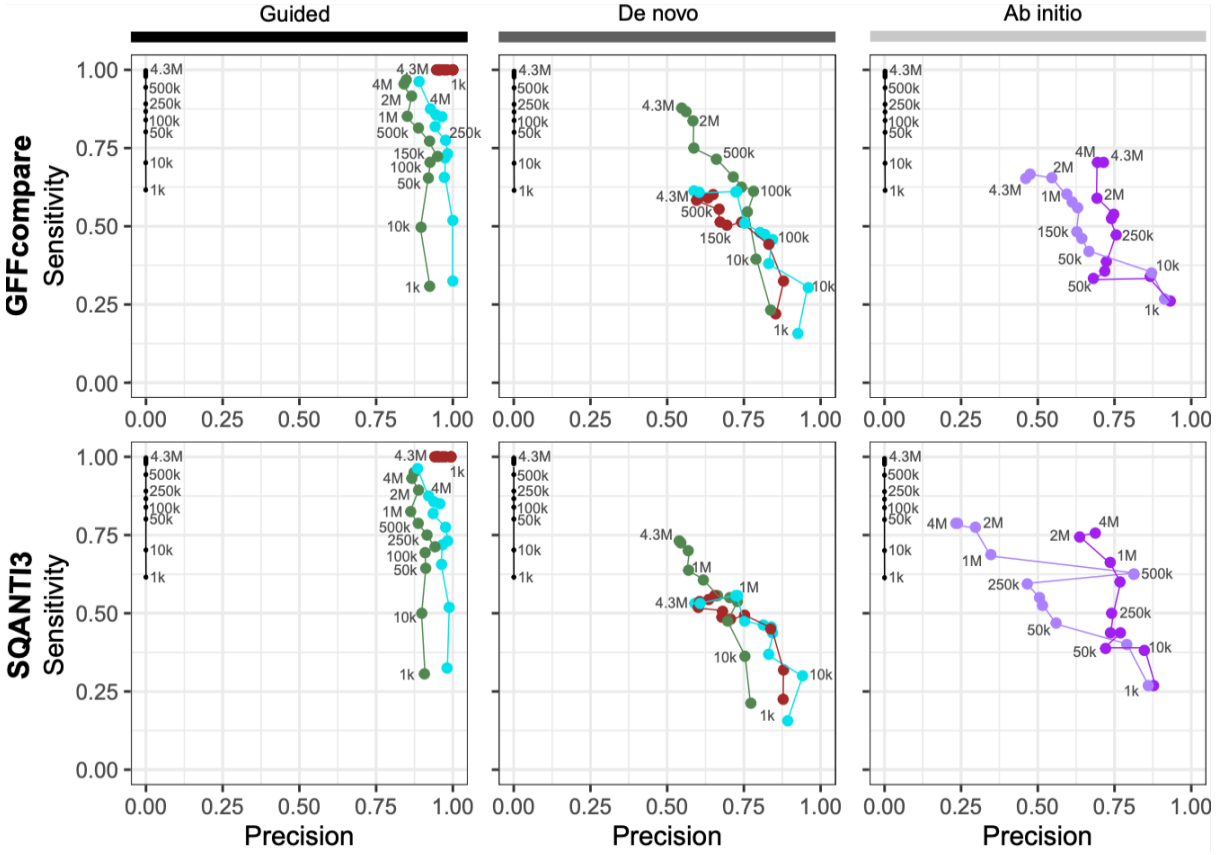

Figure 3. Sequencing depth impact on assembly. Precision and sensitivity values calculated as per GFFcompare (top panels) and SQANTI3 (bottom panels) classifications for all subsets available (ranging from 1,000=1k to 4,300,000=4.3M reads) of the LSK114_sequins dataset. Selected pipelines are Bambu, isoQuant and Stringtie2 for guided strategies; Bambu-noRef, isoQuant-noRef and Stringtie2-noRef for *de novo* strategies; and isONclust and isONclust2 for *ab initio* assembly strategies. The "max value per relative abundance" is calculated, from expected abundance as disclosed by Sequins™, as the maximum sensitivity for the number of transcripts detectable with at least 1 transcript per million (TPM) based on the number of input reads.

One of the advantages of generating bespoke transcriptome annotations is to improve isoform-level quantification, thereby providing a robust foundation upon which subsequent post-processing procedures can be reliably executed. We compared the expected vs observed quantifications obtained based on the assemblies generated by Bambu, IsoQuant, Stringtie2-noRef

and Flair-noRef. These tools have been selected to highlight biases in assembly precision or sensitivity and their potential effect on resulting quantification. Bambu and IsoQuant have a similar precision (99% and 98% respectively) but a gap of 26% in sensitivity, whilst Stringtie2-noRef and Flair-noRef have a similar sensitivities (62% and 55% respectively) but a gap of 65% in precision (**Figure 2B-D**). We only assessed correlations for assembled transcripts that were classified as true positives while using the complete assembly as a reference for quantification (see Methods). Results indicate that *guided* strategies (i.e. Bambu and IsoQuant) are very close to the expected abundances ($R^2 > 0.8$) and stable across evaluators (**Figure 4**). However, although Stringtie2-noRef and FLAIR-noRef yield similar correlation coefficients (between 0.69-0.76 and 0.6-0.76 respectively), the FLAIR-noRef quantification generates a regression curve substantially lower than the expected x=y line, indicating that erroneous transcripts might be 'absorbing' read assignments from true positive isoforms. As expected given the lack of reverse transcription and PCR, direct RNA reads (RNA002 & RNA004) generate quantifications that correlate better with expected values than cDNA for all assemblies.

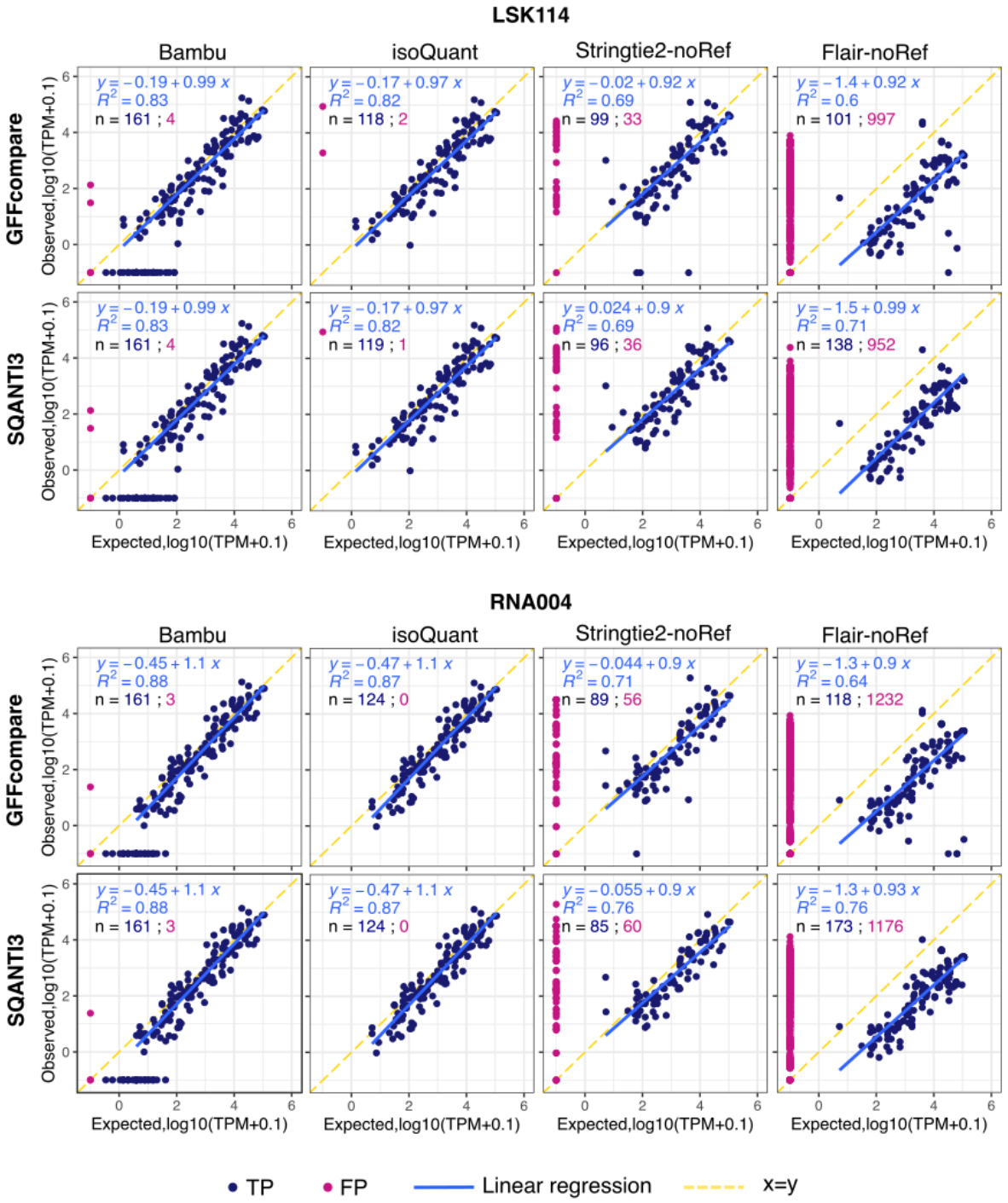

Figure 4. Assembly impact on quantification. Transcript-level quantification results as generated by Salmon from read alignments to transcriptome assemblies generated by Bambu, isoQuant, Stringtie2-noRef and Flair-noRef pipelines outputs on the LSK114_sequins_sub150k and the RNA004_sequins_sub150k datasets. Only True Positive (TP, see Methods for definition) transcripts as defined for GFFcompare (top panels) and SQANTI3 (bottom panels) are assigned to an expected

abundance, as per abundance information provided by Sequins™. False Positive (FP) transcripts were assigned a null expected abundance and removed from the linear regression equation and correlation coefficient (R^2) calculation. Number of transcripts for each category (TP, FP) is available on each panel as n = [number of TP transcripts in assembly] - [nb of FP transcripts in assembly], as well as the correlation coefficient.

With respect to ease of use and the execution time, most (but not all) tools leverage CPU multithreading (20 threads applied when possible). Depending on the input file, some *guided* methods (FLAIR, FLAMES, IsoQuant, Mandalorion) integrate the alignment step, while others do not (TALON, Stringtie2, Bambu). As shown in **Figure 5**, TALON_reco displays a significantly shorter alignment step and a significantly longer pre-processing step due to the alignment options and the use of the additional TranscriptClean script, which consumes 98% of the pre-processing time. Otherwise, all assemblers require similar computation time for 150 k reads except FLAIR, which, like TALON, includes a correction step. Notably, FLAMES performs similarly to other *guided* methods despite using a single thread. For *de novo* and *ab initio* methods, IsoQuant-noRef (13min15sec), isONclust (7min) and isONclust2 (1min30sec) stand out with significantly shorter execution times than all the other tools.

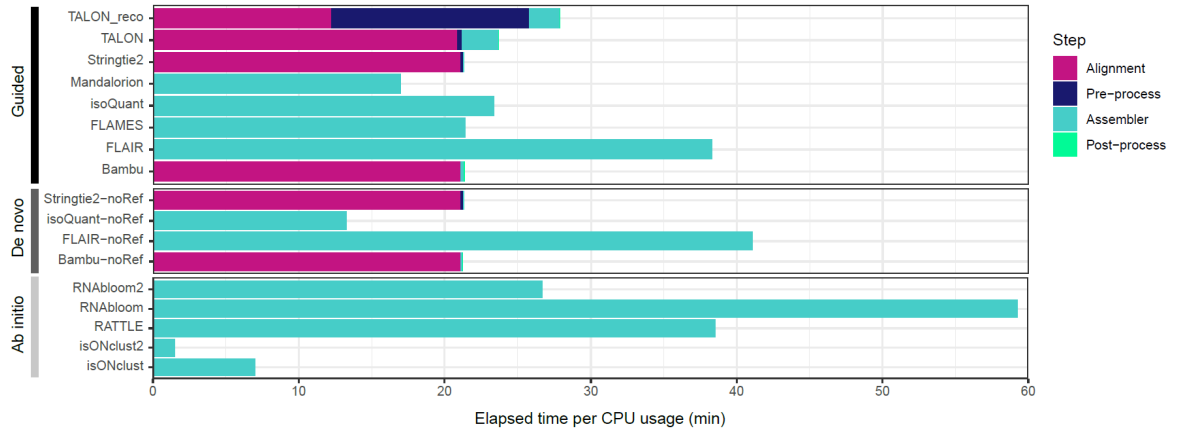

Figure 5. Execution time. Elapsed time as outputted by the /user/bin/time system function multiplied by the CPU usage percentage for all steps of all pipelines applied to the LSK114_sequins_sub150k dataset. When allowed, all tools were executed with 20 threads on a workstation with 2x16 core 2.30 GHz Intel Xeon Gold 5218 workstation with M.2 NVMe solid state drives.

Discussion

This benchmarking study on assembly algorithms using nanopore RNA-seq data was originally intended to expose the best methodology for the resolution of RNA isoforms. This was prompted by improved nanopore sequencing accessibility and accuracy that can easily identify numerous novel RNA isoforms, even in extensively sequenced model organisms. In our opinion, the common practice of relying on a reference transcriptome annotation (i.e., Refseq or Gencode) can lead to significant biases by including genes and isoforms that are not expressed in a given experimental condition and, more importantly, by ignoring or omitting unannotated transcripts that might be highly expressed or significantly regulated between conditions. This is notably the case for extensively spliced long non-coding RNA genes, which now outnumber protein-coding genes in recent Gencode annotations. The results we have obtained from RNA standards indicate that there is substantial variability in the accuracy of isoform reconstruction tools; a greater diversity than we expected given the single molecule nature of nanopore sequencing data. Hence, there appears to be no clear "one size fits all" solution for transcriptome assembly at the moment.

There are nonetheless several notable observations that can help guide future nanopore RNA-seq analyses. Providing a reference annotation and a reference genome unsurprisingly yields the most accurate (although not necessarily perfect) results. This modality might not detect novel isoforms beyond those in the reference–a possibility we briefly explored by removing a fraction of the isoforms in the reference annotation and assessing their discovery thereafter (not shown). In practice, this equates to the *de novo* modality, where reads are aligned to a genome prior to assembly and without a reference. Although this strategy greatly facilitates isoform discovery, the assembly must then be annotated by comparison to a gene reference with tools like GFFcompare and SQANTI3, which is a somewhat subjective undertaking that can complexify downstream analyses (see below). For *de novo* assembly, IsoQuant consistently demonstrates the least amount of false positive transcripts across all surveyed datasets, while StringTie2 might provide the greatest overall accuracy (F1 score), closely followed by reference-free Bambu and IsoQuant (**Figure 2**). Surprisingly to us, the

performance of *ab inito* methods compares favorably to *de novo* techniques in accuracy and execution time, especially isONclust & isONclust2. However, isONclust2 requires some enhancements to accommodate dRNA reads and to facilitate its integration into a streamlined analytical workflow. These *ab initio* techniques are often overlooked in model organisms with high-quality genomes, as well as many benchmarking studies. Their assembled transcriptome output is typically aligned to a genome prior to comparative reference gene annotation, thus also requiring tools such as GFFcompare and SQANTI3. Our analysis concurs with the conclusions of recent studies on the accuracy, effectiveness and versatility of Bambu, IsoQuant and Stringtie2 (Dong et al. 2023; Pardo-Palacios et al. 2023b). However, our assessment contradicts conclusions by the LRGASP consortium on the utility of FLAIR, found to be one of the most robust strategies for both well annotated organisms and orthogonal data integration. Our results suggest that the accuracy of the versions we tested lies far behind other tools, particularly when executed without a reference annotation (**Figure 2**).

Given the unsurprising tradeoff between sensitivity and specificity, some methods should be privileged over others depending on the objective of the sequencing experiment. For example, when seeking to identify new isoforms of known genes, certain tools might not be suitable. This is the case for IsoQuant and Bambu in *de novo* mode on SIRVs (sensitivity and specificity of ~15%, **Figure 2A**), a spike-in control set with fewer genes and more alternative splicing isoforms than Sequins™, where these tools performed much better (45-55% sensitivity). It should be mentioned that the depth of sequencing may also play a role in assessing each method's perceived accuracy, as we demonstrated in **Figure 3**. Unfortunately, the limited SIRV and dRNA sequencing depth and the large number of surveyed tools restricted a more comprehensive evaluation of the impact of depth on assembly accuracy, which undoubtedly has an impact on the assessment of true and false positive isoforms.

Certain factors that can affect transcriptome assembly were not directly evaluated in this study, either by design or practicality. Firstly, base calling algorithms and their associated error rates could potentially impact alignment and assembly accuracy, notably at exon-intron boundaries. We applied the latest base calling algorithms when preparing this manuscript; older versions (with high error rates) might produce different results. Secondly, we employed a single pairwise alignment tool (minimap2) when alignment was required. Alternative aligners could significantly impact assemblies,

particularly in *guided* and *de novo* modalities. The impact of alignment can be major—when confronted with fragmented reads, alignment algorithms can force terminal sequences spanning an exon junction to align within the intron as the affine gap penalty may be substantially larger than the associated mismatch penalty, leading to an apparently longer 'read-through' exon, a common observation in de novo transcriptome assemblies. Thirdly, the choice of reverse transcription enzymes, elongation times and PCR cycles can also have a significant impact on the results; the manufacturer's recommended protocols at the time of data generation were used. An additional experimental variable could be freeze-thaw cycles of the spike-in controls. To the best of our knowledge, only fresh controls were used. Fourthly, we used the default or recommended execution parameters for the assembly algorithms. It is to be expected that tweaking these would significantly impact results. In particular, we did not attempt to normalize the minimum coverage threshold when this parameter was available, defaulting to the developers' recommended settings. It is to be expected that a lower threshold would increase sensitivity and, potentially, reduce specificity for a given tool. In general, the developers bear the responsibility of optimal default parameter settings. Lastly, we did not evaluate the relative impact of transcript and sequence features on assembly accuracy. GC content, transcript length, exon length, number of exons, isoform diversity per gene are features that might help pinpoint the strengths and limitations of certain tools, while guiding algorithmic improvements to distinguish between biological signal and noise.

This study substantiates that assembling a long read transcriptome–even from relatively simple, well-defined spike-in controls–remains a complex and error-prone endeavor. Even the methods to compare and evaluate transcriptome annotations are subject to variation, exposing the subjective nature of qualitative transcript classifications. Most studies rely on either GFFcompare (Stuart et al. 2024; Saha et al. 2024) or SQANTI3 (Pardo-Palacios et al. 2023b; Brooks et al. 2024) to this avail. When comparing both methods on the same data, we observed surprisingly different assembly accuracies (c.f. the performance of Stringtie2-noRef with SIRV data in **Figure 2**, **Supplemental Figure 1**). While GFFcompare assesses a *de novo* transcriptome assembly with 75% precision and 55% sensitivity; SQANTI3 assesses 15% for both. This can partially be explained by

what transcript classifications we esteemed to define true or false positives (we considered incomplete transcripts, 'c' for GFFcompare and 'ISM' for SQANTI3, as true positives) but also by the way both tools classify transcripts, such as nuances in the assessment of intron chains or splice junction mappings. Another likely source of divergence is how GFFcompare systematically selects the most similar reference transcript to qualify an isoform, while SQANTI3 remains impartial and simply classifies any isoform that doesn't perfectly match the reference features as "new". Nonetheless, the isoform definitions we employed herein form an arguably more stringent approach compared to other studies, where all transcripts classifications that overlap the reference genes are considered true positives (Song et al. 2019). An interesting solution lies in the sequential use of GFFcompare and SQANTI3 annotation evaluators described by Wijeratne et al. (Wijeratne et al. 2024). The authors use GFFcompare to compare transcriptome features across samples and generate a combined list of non-redundant isoforms, while SQANTI3 is employed for comprehensive characterization of long-read transcript sequences, including quality control, identification, and quantification of full-length transcripts.

In summary, evaluating the accuracy of long-read transcriptome assembly strategies exposes the multi-faceted considerations that impact isoform-level analyses. In addition to the choice of assembly modality and algorithms, the nature of queried sequencing data and assembly evaluator characteristics impose nuanced trade-offs for isoform discovery and transcriptome annotation. The most important consideration, however, is the inclusion of well defined spike-in controls to help quantify artifacts of transcriptome assembly and establish a baseline of truthfulness in the output. By integrating these insights, researchers can make informed choices in selecting assembly methods tailored to their specific experimental objectives and resource constraints.

Methods

Samples sequencing and data acquisition

The following datasets, LSK109_sequins and LSK114_sequins were obtained by preparing 15ng of sequin standards RNA mix A (Hardwick et al. 2016) following the PCS109 library preparation protocol as recommended by manufacturer [Oxford Nanopore Technologies©, UK] up until the PCR step. Then, the protocol LSK109 (amplicons) and LSK114 (amplicons) was applied to the resulting cDNA. 50 fmol and 20 fmol of the respective final libraries were loaded onto a R9.4.1 flowcell (flowcell ID: FAN54376) and a R10.4.1 flowcell (flowcell ID: FAV99142) and ran for 72 hours on a GridION sequencer and 12 hours on a PromethION sequencer, respectively.

The LSK114_SIRVs dataset was obtained by preparing 0,0375ng of SIRV set-04 [Lexogen, Austria] spiked-in with 125ng of a human RNA sample, as recommended by the manufacturer. As for previous datasets, samples were processed following the PCS109 library preparation protocol up until the PCR step; then the protocol LSK114 (amplicons) was applied to the resulting cDNA as recommended by the manufacturer [Oxford Nanopore Technologies©, UK]. 18 fmol of the final library was loaded onto a R10.4.1 flowcell (flowcell ID: PAK95982) that ran for 72 hours on a PromethION sequencer. A reload was done with an additional 18 fmol of the library after 29 hours. Only reads mapping to the SIRV set-04 genome were retrieved using minimap2 and seqtk toolkit (https://github.com/lh3/seqtk).

For the RNA002_sequins dataset, 300ng of sequin standards RNA mix A (Hardwick et al. 2016) was prepared following the RNA002 library preparation protocol as recommended by the manufacturer [Oxford Nanopore Technologies©, UK]. 20ng of the final library was loaded onto a R9.4.1 flowcell (flowcell ID: PAM58324) and ran on a PromethION sequencer for 72 hours.

RNA004_sequins dataset was acquired by preparing 4 samples with 1.5µg of human spiked-in with 15ng of sequin standards RNA mix A (Hardwick et al. 2016). Samples were prepared following the RNA004 protocol (bêta testing phase: V2 April 2023). 20ng of the final libraries for each sample

was loaded onto individual RP4_bêta-testing flowcells (flowcell IDs: PAO83093 - PAO83456 - PAO84072 - PAO96683) and ran for 68 hours on a PromethION sequencer. A reload with the same amount of library was done after 44 hours and 15 minutes of sequencing. Only reads mapping to the Sequins™ reference genome were retrieved using minimap2 and seqtk toolkit (https://github.com/lh3/seqtk).

Datasets pre-processing

All datasets were basecalled with Guppy v6 [6.0.1 to 6.5.7] super high accuracy mode. For cDNA datasets (LSK109_sequins, LSK114_seuins, LSK114_SIRVs), pass and fail reads from basecalling were trimmed and reoriented using Pychopper v2.7.9 (https://github.com/epi2me-labs/pychopper). Full-length and rescued reads from Pychopper were then processed like raw basecalled pass/fail reads for dRNA datasets (RNA002_sequins, RNA004_sequins). Minimap2 v2.24 (Li 2021) was used to generate .bam, and .sam files (-ax splice --secondary=no). The option --MD and .sam output was used for TALON pipeline only, as required by the tool.

Subsampling and transcriptome assembly

The five datasets (PCS109_sequins, LSK114_sequins, LSK114_SIRVs, RNA002_sequins, RNA004_sequins) were randomly subsampled after pychopper pre-porcessing for cDNA datasets and after base calling for dRNA datasets to have the following amount of reads: 1k - 10k - 50k - 100k - 150k - 250k - 500k - 1M - 2M - 4M - Total. All subsampled datasets were processed through assembly as described below. Subsampled datasets are referred to as subX, where X is the number of reads in the dataset.

Each tool was run with default parameters using 20 threads when possible. Some tools were executed in two different ways (Stringtie2/Stringtie2-noRef, Bambu/Bambu-noRef, Flair/Flair-noRef, isoQuant/isoQuant-noRef) when assembly with and without a transcriptomic reference was allowed. For RNAbloom, the final assembly assessed doesn't contain the assembled short transcripts. isONclust2 was applied to cDNA datasets only (as it is not yet available for dRNA). FLAMES .gff3

output is filtered based on its transcripts evaluation provided : only transcripts classified as “true” in the ‘isoform_FSM_annotation.csv’ file are filtered and considered as final assembly. Further details of pipelines and their execution is available on github (https://github.com/msagniez/LRassBench).

Table 2. List of tools and corresponding input files

Strategy	Assembler	Version	Input*	Comment
Ab initio	RATTLE	1.0	.fastq	pre-filtering with reads longer than 150bp only
Ab initio	isONclust	0.0.6.1	.fastq	
Ab initio	isONclust2	2.4	.fastq	Only for cDNA datasets
Ab initio	RNAbloom	1.4.3	.fastq	
Ab initio	RNAbloom2	2.0.1	.fastq	
De novo	Bambu-noRef	3.3.3	.bam	
De novo	Flair-noRef	2.0.0	.bam	
De novo	isoQuant-noRef	3.3.1	.fastq	
De novo	Stringtie2-noRef	2.2.1	.bam	
Guided	Bambu	3.3.3	.bam	
Guided	Flair	2.0.0	.bam	
Guided	FLAMES	0.1	.fastq	
Guided	isoQuant	3.3.1	.fastq	
Guided	Mandalorion	4.3.0	.fastq	
Guided	Stringtie2	2.2.1	.bam	
Guided	Talon	5.0	MD.sam	
Guided	Talon_reco	5.0	.sam**	Alignment with options recommended by Talon authors

* .fastq = fastq file constituted of full-length and rescued reads outputted by Pychopper

.bam = sam file outputted by minimap2 v2.24 (options: -ax splice --secondary=no) and converted to bam file with samtools v1.19.2

.sam = sam file outputted by minimap2 v2.24 (options: -ax splice --secondary=no)

MD.sam = sam file outputted by minimap2 v2.24 (options: -ax splice --secondary=no --MD)

**.sam = sam file outputted by minimap2 v2.24 as recommended by Talon's authors (options: -ax splice -uf -k14)

Assembly quality assessment

Final assemblies were mapped to the original reference transcriptomes using two different tools to define the mappings: GFFcompare (Pertea and Pertea 2020) (.tmap file) and SQANTI3 (Pardo-Palacios et al. 2023a) (_classification.txt file). All Figures were plotted using the ggplot2 library in R.

Sensitivity, Precision and F1 calculation

As described in Figure 1, true positive transcripts (TP) are transcripts labeled "=" and "c" for GFFcompare as well as "Full Splice Match" and "Incomplete Splice Match" (except intronic transcripts) for SQANTI3. Potential new transcripts are the transcripts labeled "u" by GFFcompare and "Intergenic" by SQANTI3. These potential new transcripts are excluded from true positive, false positive and false negative sets. False positive transcripts (FP) are the ones that don't match the labels cited above. False negatives (FN) are the transcripts found in the reference but not assembled by the tools. The sensitivity is calculated as $TP/(TP + FN)$ and the precision is calculated as $TP/(TP + FP)$. The F1 score is calculated as $F1 = (2 * Precision * Sensitivity)/(Precision + Sensitivity)$.

Transcripts discovery accuracy assessment

We selected 9 transcripts to remove from the original Sequin™ .gtf annotation (R1_13_1 ; R1_21_2 ; R1_51_1 ; R2_116_1 ; R2_116_2 ; R2_47_2 ; R2_59_3 ; R2_6_2 ; R2_72_1) based on their theoretical abundance, length and number of isoforms. All guided tools were executed again using LSK114_sequins_sub150k dataset and the incomplete .gtf annotation. The GTFs from all the tools were parsed in R, and all the genomic coordinates of the assembled transcripts were plotted. The data was used to narrow down the assemblies that mapped to the region of the 9 artificially created 'novel'

transcripts. This subset was then filtered for exons that completely mapped to the known exons, as we wished to retain the ones that mapped to the nine novel isoforms only. This filtered dataset was used to visualize the assemblies, compared to the nine novel isoforms, using ggplot2 and ggtranscript (Supplemental figure 1 and 2).

Sequencing depth analysis

We show Precision and Sensitivity results for the 5 most performant tools (Bambu, isoQuant, Stringie2, isONclust, isONclust2) on all subsets of the LSK114_sequins dataset. For each of these subsets we also calculated the maximum sensitivity per relative abundance (MaxSensitivity) considering the theoretical abundance of each transcript in the mix as:

$Limit\ of\ Detection\ LoD\ =\ 1/number\ of\ reads\ in\ subset$ (in TPM)

$FN_{LoD} = \sum(sequin\ transcripts\ <\ LoD\ (in\ TPM)\ for\ mixA)$

$$MaxSensitivity\ = \frac{Total\ number\ of\ sequin\ transcripts}{Total\ number\ of\ sequin\ transcripts + FN_{LoD}}$$

Quantification

Quantification was obtained after mapping the original .fastq files of datasets LSK114_sequins_sub150k and RNA004_sequins_sub150k to the corresponding output assembly of Bambu, isoQuant, Stringtie2-noRef and Flair-noRef with minimap2 v2.24 [options: -a -N 100] and samtools v1.19.2 to convert to bam. Salmon v1.10.1 (Patro et al. 2017) was then executed. Experimental quantifications (TPM) of TP (as described in Sensitivity, Precision and F1 calculation section) were compared to theoretical abundances converted to TPMs to generate Pearson's correlation coefficient. TPs with null experimental quantifications were excluded from the coefficient calculation, as they are considered not found in assembly. FP transcripts were attributed a theoretical abundance of 0, as these are not real transcripts and should not be present.

Data access

Assembly annotations evaluated as well as GFFcompare (.tmap file) and SQANTI3 (_classification.txt file) evaluations are available along with the code used in our github page : https://github.com/msagniez/LRassBench. Input .fastq files as used in this study for SIRV and sequin data are available on SRA [accession numbers]

Competing interest statement

MS, AB & MAS have received financial support for travel to conferences from Oxford Nanopore Technologies. MAS has received free research consumables from Oxford Nanopore Technologies, who were not involved in the study design or the interpretation of results.

Acknowledgments

We would like to thank Eduardo Eyras and Richard Kuo for comments and feedback during the preparation of this work; David Barda and Tim Mercer from SequinsSequin™ who donated spike-in controls essential for this work; Libby Snell and Etienne Rainmondeau from Oxford Nanopore Technologies for support with RNA004 direct RNA beta testing; and the developers and bioinformatics community that contributed to the various open-source algorithms that were tested.

This work was supported by a Fonds de recherche du Québec–Santé Junior 1 fellowship [project 284217], a Canadian National Science and Engineering Research Consortium Discovery grant [DGECR-2022-00207], a Cole Foundation Transition Award, a Canadian New Frontiers in Research Fund Exploration Grant [NFRFE-2021-01005], by the Canadian Foundation for Innovation John R. Evans Leaders Fund [project 40767] to MAS, and by a Cole Foundation PhD fellowship to MS. Computing resources were partially provided by a research allocation from the Digital Research Alliance of Canada and Calcul Québec.

Author contributions

MAS conceived the study and secured funding. MS, BP, MR and MAS generated sequencing libraries and generated sequencing data. MS, AB, SMS & CVO performed bioinformatics analyses. MS & AB generated Figures. MS, AB & MAS wrote the manuscript with input from all authors.

References

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.

Brooks TG, Lahens NF, Mrčela A, Grant GR. 2024. Challenges and best practices in omics benchmarking. *Nat Rev Genet*. http://dx.doi.org/10.1038/s41576-023-00679-6.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**. http://dx.doi.org/10.1093/gigascience/giz100.

Byrne A, Cole C, Volden R, Vollmers C. 2019. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* **374**: 20190097.

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195.

Cuber P, Chooneea D, Geeves C, Salatino S, Creedy TJ, Griffin C, Sivess L, Barnes I, Price B, Misra R. 2023. Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore Technologies, and Pacific Biosciences, for DNA barcode sequencing applications. *Ecological Genetics and Genomics* **28**: 100181.

de la Rubia I, Srivastava A, Xue W, Indi JA, Carbonell-Sala S, Lagarde J, Albà MM, Eyras E. 2022. RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *Genome Biol* **23**: 153.

Dong X, Du MRM, Gouil Q, Tian L, Jabbari JS, Bowden R, Baldoni PL, Chen Y, Smyth GK, Amarasinghe SL, et al. 2023. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *bioRxiv* 2022.07.22.501076. https://www.biorxiv.org/content/10.1101/2022.07.22.501076v3 (Accessed August 15, 2023).

Fang Y, Chen G, Chen F, Hu E, Dong X, Li Z, He L, Sun Y, Qiu L, Xu H, et al. 2021. Accurate transcriptome assembly by Nanopore RNA sequencing reveals novel functional transcripts in hepatocellular carcinoma. *Cancer Sci* **112**: 3555–3568.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB, Nielsen LK, Mattick JS, Mercer TR. 2016. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods* **13**: 792–798.

Kang S-H, Lee W-H, Sim J-S, Thaku N, Chang S, Hong J-P, Oh T-J. 2021. De novo Transcriptome Assembly of Senna occidentalis Sheds Light on the Anthraquinone Biosynthesis Pathway. *Front Plant Sci* **12**: 773553.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**: 751.

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**: 1731–1740.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572–4574.

Marchet C, Lecompte L, Da Silva C. 2018. CARNAC-LR: Clustering coefficient-based Acquisition of RNA Communities in Long Reads. *JOBIM 2018*. https://hal.archives-ouvertes.fr/hal-01930211/.

Matra DD, Adrian M, Karmanah, Kusuma J, Duminil J, Sobir, Poerwanto R. 2023. Dataset from de novo transcriptome assembly of Myristica fatua leaves using MinION nanopore sequencer. *Data Brief* **46**: 108838.

Nip KM, Chiu R, Yang C, Chu J, Mohamadi H, Warren RL, Birol I. 2020. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res* **30**: 1191–1200.

Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, Birol I. 2023. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nat Commun* **14**: 2940.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–45.

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2023a. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *bioRxiv*. http://dx.doi.org/10.1101/2023.05.17.541248.

Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, et al. 2023b. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv*. http://dx.doi.org/10.1101/2023.07.25.550582.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**.

http://dx.doi.org/10.12688/f1000research.23297.2.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295.

Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**: 915–918.

Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, Vitale D, Lapidus A. 2020. Extending rnaSPAdes functionality for hybrid transcriptome assembly. *BMC Bioinformatics* **21**: 302.

Saha B, McNinch CM, Lu S, Ho MCW, De Carvalho SS, Barillas-Mury C. 2024. In-depth transcriptomic analysis of Anopheles gambiae hemocytes uncovers novel genes and the oenocytoid developmental lineage. *BMC Genomics* **25**: 80.

Sahlin K. 2021. Effective sequence similarity detection with strobemers. *Genome Res* **31**: 2080–2094.

Sahlin K, Medvedev P. 2020. De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. *J Comput Biol* **27**: 472–484.

Shao M, Kingsford C. 2017. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat Biotechnol* **35**: 1167–1169.

Song L, Sabunciyan S, Yang G, Florea L. 2019. A multi-sample approach increases the accuracy of transcript assembly. *Nat Commun* **10**: 5000.

Stuart KC, Johnson RN, Major RE, Atsawawaranunt K, Ewart KM, Rollins LA, Santure AW, Whibley A. 2024. The genome of a globally invasive passerine, the common myna, Acridotheres tristis. *DNA Res* **31**. http://dx.doi.org/10.1093/dnares/dsae005.

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438.

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. Corrigendum: SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**: 1096.

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 310.

Verma A, Jiang Y, Du W, Fairchild L, Melnick A, Elemento O. 2015. Transcriptome sequencing reveals thousands of novel long non-coding RNAs in B cell lymphoma. *Genome Med* **7**: 110.

Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. 2023. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. *Genome Biol* **24**: 167.

Wijeratne S, Gonzalez MEH, Roach K, Miller KE, Schieffer KM, Fitch JR, Leonard J, White P, Kelly BJ, Cottrell CE, et al. 2024. Full-length isoform concatenation sequencing to resolve cancer transcriptome complexity. *BMC Genomics* **25**: 122.

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al. 2020. A technology-agnostic long-read analysis pipeline for

transcriptome discovery and quantification. *bioRxiv* 672931. https://www.biorxiv.org/content/10.1101/672931v2 (Accessed August 16, 2023).

Wyman D, Mortazavi A. 2019. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics* **35**: 340–342.

Yang M, Shang X, Zhou Y, Wang C, Wei G, Tang J, Zhang M, Liu Y, Cao J, Zhang Q. 2021. Full-Length Transcriptome Analysis of Plasmodium falciparum by Single-Molecule Long-Read Sequencing. *Front Cell Infect Microbiol* **11**: 631545.

Zagorščak M, Petek M. 2021. A Comprehensive Guide to Potato Transcriptome Assembly. *Methods Mol Biol* **2354**: 155–192.

Zhang Q, Shi Q, Shao M. 2022. Accurate assembly of multi-end RNA-seq data with Scallop2. *Nat Comput Sci* **2**: 148–152.