A Hitchhiker's Guide to Long-read RNA Isoform Sequencing

Christophe Georgescu[1], Brian Haas[1†], Akanksha Khorgade[1], Houlin Yu[1], James Webber[1], Ghamdan Al-Eryani[1], Daniel Bartlett[1], Emily White[1], Asa Shin[1], Niall Lennon[1], Victoria Popic[1†], Aziz Al'Khafaji[1†]
[1]Broad Clinical Labs, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA | †Corresponding Authors

Methods Development Lab

SUMMARY

Recent advances in long-read RNA isoform sequencing have enabled researchers to resolve the rich complexity of alternative splicing in the transcriptome at scale. This new capability naturally comes with a set of challenges; new tools must be developed for data processing, QC, analysis, and visualization. Further, appropriate computational infrastructure must be developed to reproducibly benchmark latest methods and derive principled best practices. Finally, robust and easy to use pipelines are needed to enable non-expert users to leverage these powerful methods. To this end, we have developed a suite of tools, best practices, and vignettes hosted on a publicly available website to guide researchers on their journey from RNA material to fully analyzed data.

cDNA synthesis artifacts and quantification

A myriad of artifacts can arise during cDNA synthesis (Fig. 1a), which can reduce sequencing throughput and complicate isoform analysis. We developed *MARTI* to efficiently identify artifacts, enabling optimization of cDNA synthesis chemistries and filtering data for higher performance analysis (Fig. 1b). For example, use of a modified TSO resulted in almost complete elimination of TSO-TSO artifact (Fig. 1c).

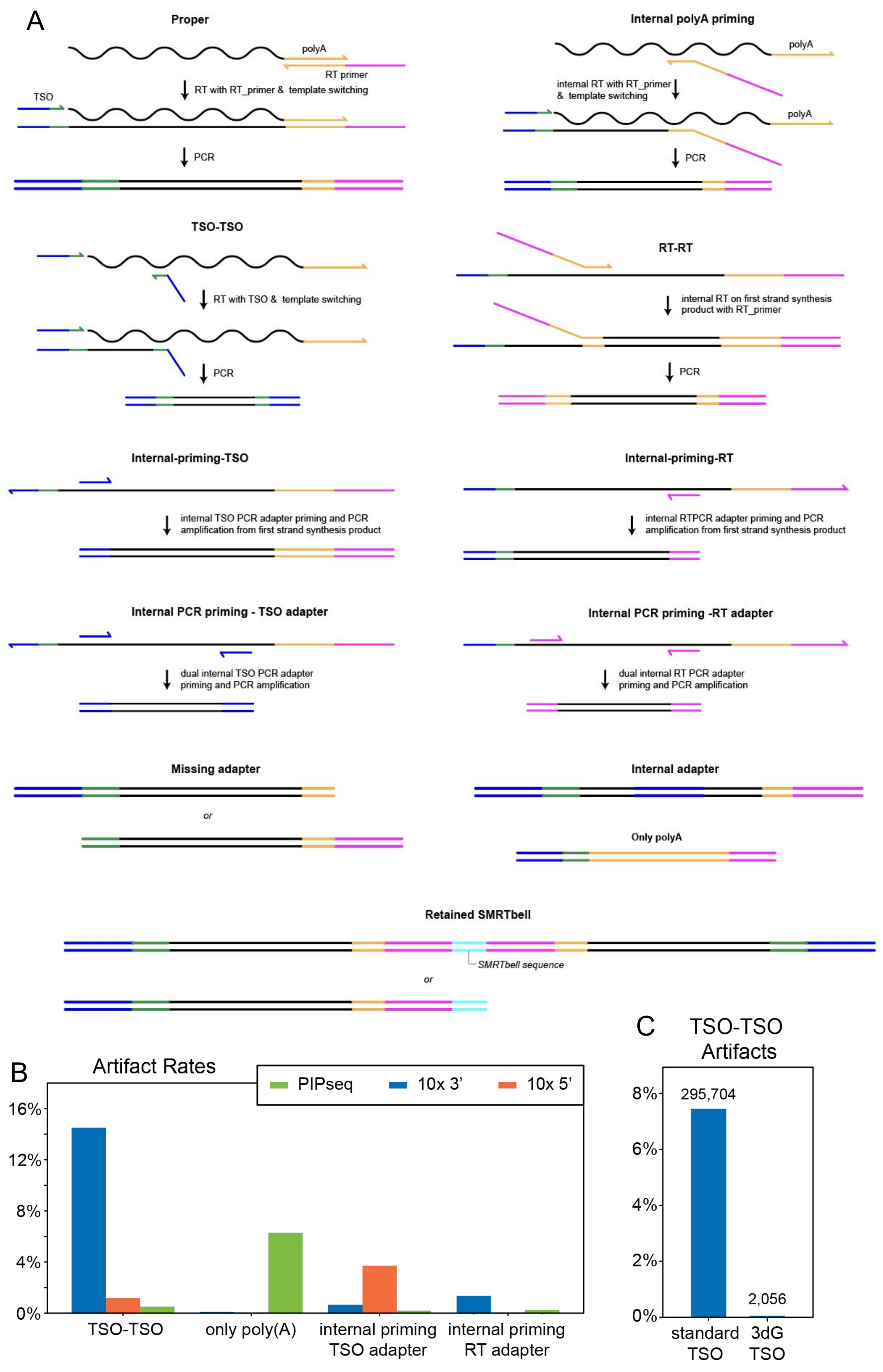


Figure 1. A) cDNA synthesis artifacts and their formation. B) MARTI analysis of single-cell platform artifact rates. C) Rate of TSO-TSO formation in 10x 3' single-cell with modified TSO.

Internal reverse transcriptase priming

Internal priming of adenine-rich tracks in mRNA results in truncated transcript capture and is a key obstacle in accurate isoform analysis. Strikingly, internal priming rates are ~20% across single-cell platforms (Fig. 2a) and drive the erroneous discovery of novel transcripts (Fig. 2b). Be sure to check the priming status of novel isoforms identified in your studies.

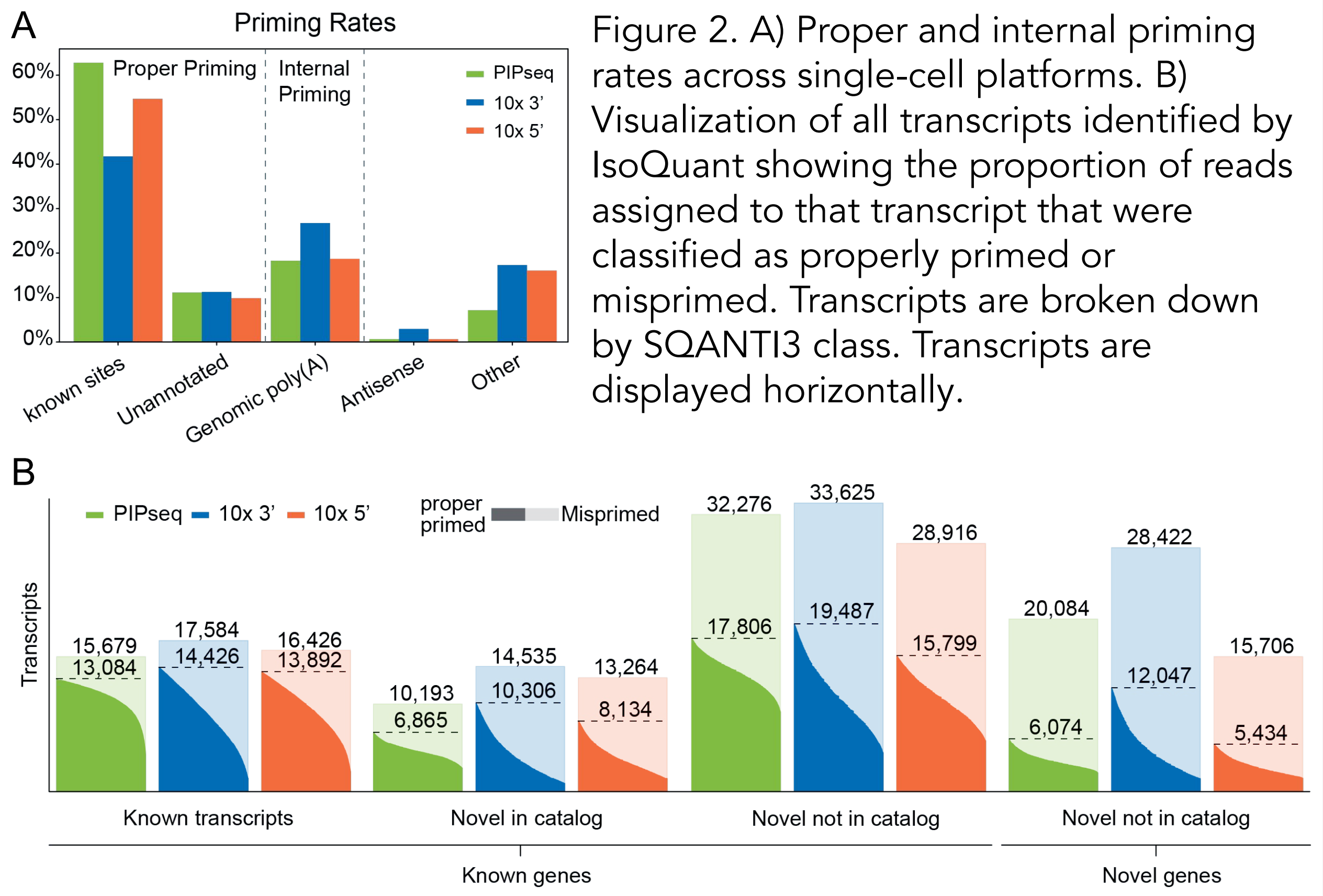


Figure 2. A) Proper and internal priming rates across single-cell platforms. B) Visualization of all transcripts identified by IsoQuant showing the proportion of reads assigned to that transcript that were classified as properly primed or misprimed. Transcripts are broken down by SQANTI3 class. Transcripts are displayed horizontally.

Long-read RNA-seq QC

To assess long-read RNA-seq data quality, we developed LongRNAqc+, a read level QC tool that reports key transcript coverage metrics. Taking as input an aligned BAM per sample, LongRNAqc+ runs a long-read adapted version of RNAqc, an adapted version of SQANTI3, and optionally LRAA (poster #206) and IsoQuant. From their outputs, it generates comparisons of important metrics across multiple samples such as the proportion of reads spanning all exons of an isoform (Full Splice match) and their length distributions, normalized against the total number of reads. These results can help identify issues such as degraded RNA or too stringent cleaning of smaller molecules, making it useful for both production runs and testing experimental conditions when trying to optimize protocols or chemistry.

In contrast to other approaches that perform similar analyses on collapsed reads, read level analyses naturally reflect the underlying data. While algorithms for merging reads are useful for isoform identification, performing QC after this step clouds our understanding of the primary data quality going into these processes. For example, samples that have degraded RNA will result in shorter sequenced cDNAs, though those incomplete transcript captures will be able to merged with non-degraded sequenced transcripts - hence masking the true quality of our library.

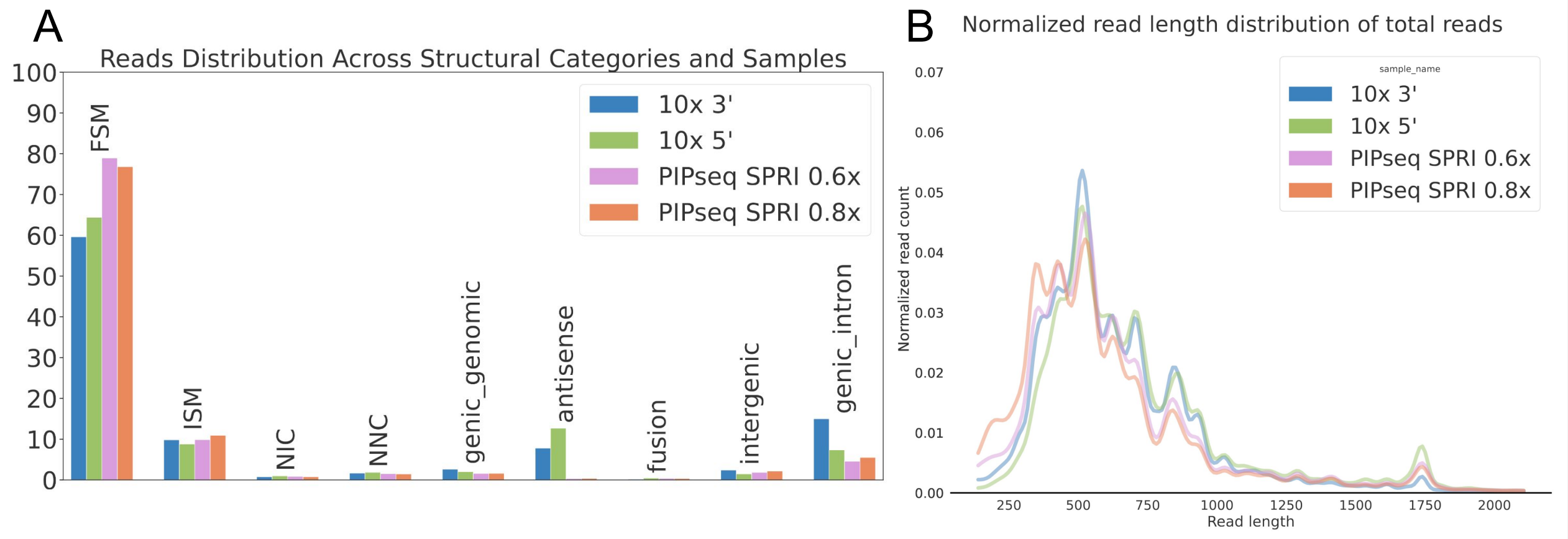


Figure 3. A) Squanti based classification of reads from single-cell platforms. FSM: Full Splice Match, ISM: Incomplete Splice Match, NIC: Novel In Catalog, NNC: Novel Not in Catalog B) Aligned read length distribution from single-cell platforms.

Some of the outputs of the tools run by LongRNAqc+ can be further used for downstream inspection with ITV (next section).

Integrative Transcriptomics Viewer

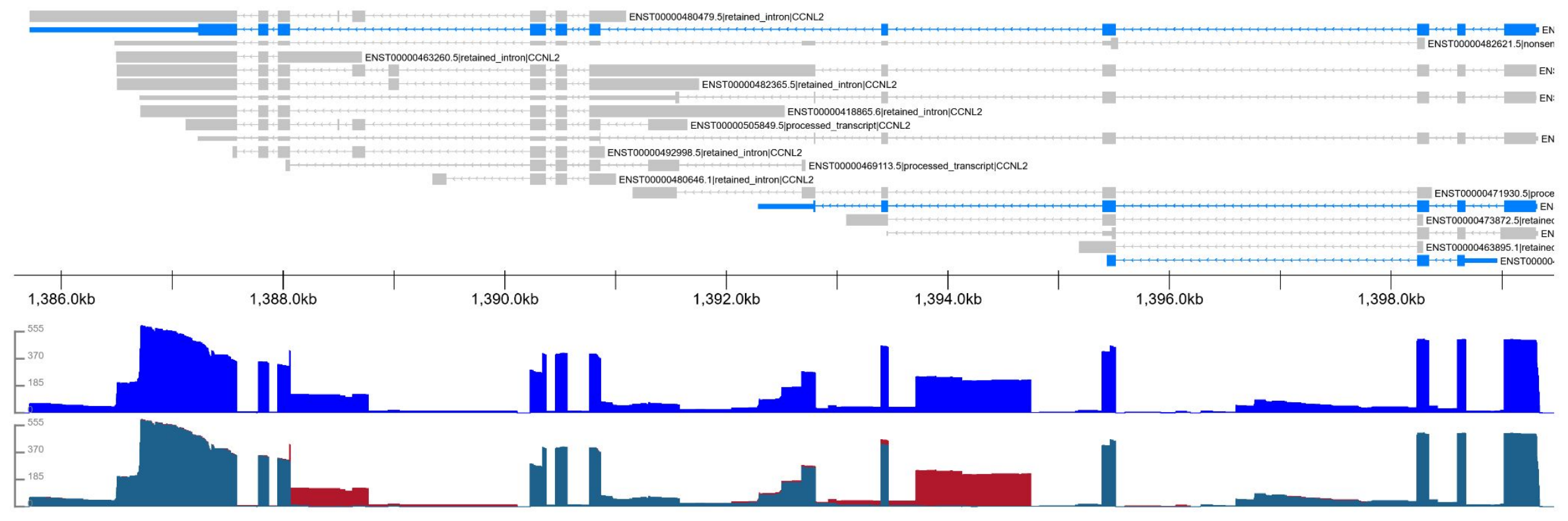
Integrative Transcriptomics Viewer (ITV) is a Python package built to better support RNA long reads and isoform level visualizations. Features include:

> plot a single read per line, making clear which reads are full length vs fragments.

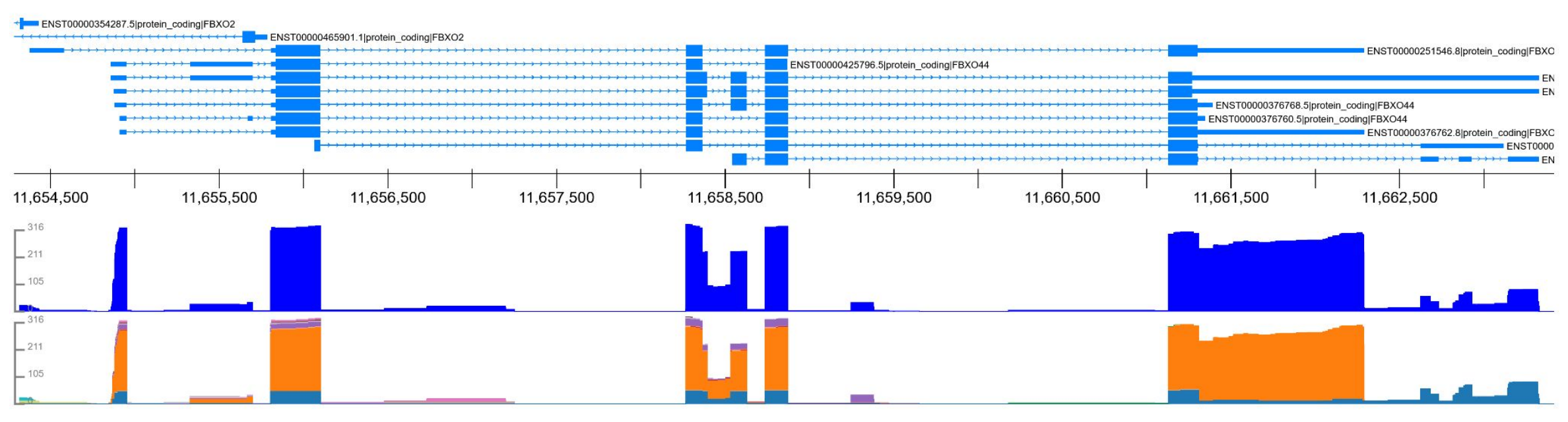
> plot only the exons of a gene, increasing readability by reclaiming the vast space otherwise used by introns.

> in-memory splitting and filtering of BAMs or coverage track coloring based on features such as alignment orientation, barcodes whitelist or clustering, QC classification results, etc.

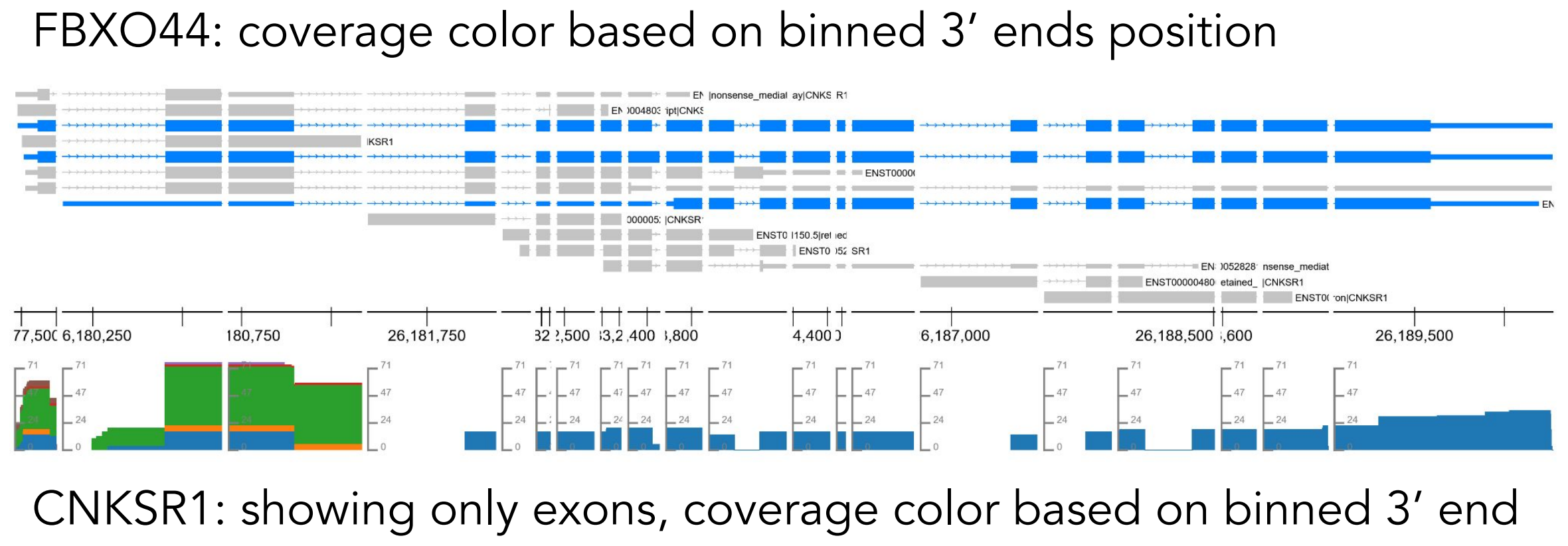
> creation of interactive, self contained HTML reports that can easily be shared, with for example splitting of reads based on isoform they are classified as and cell type they are found in.



CCNL2: coverage color based on read orientation, in blue sense reads, in red antisense reads.



FBXO44: coverage color based on binned 3' ends position



CNKSR1: showing only exons, coverage color based on binned 3' end position.

Isoform ID benchmarking

As transcriptome sequencing technologies and software tool co-evolve, it is essential that we have effective methods for routine evaluation of their accuracy. As such, we have generated new data sets and streamlined workflows for executing state-of-the-art software tools and benchmarking routines for quantification-only, reference-guided isoform identification, and for reference-annotation-free isoform identification. Examples of long read isoform based analysis tools and related benchmarking efforts are shown below.

A

Tools	Ref. Guided ID & Quant	Ref. Free ID & Quant	SC Read Assignments
LRAA	✅	✅	✓
IsoQuant	✅	✅	✓
Mandalorion	✅	✅	
Bambu	✅	✅	
StringTie	✅	✅	
IsoSeq	✅	✅	
Isosceles	✅		✓
FLAIR	✅		
FLAMES	✅		
ESPRESSO	✅		
TALON	✅		
Oarfish	QUANT ONLY		

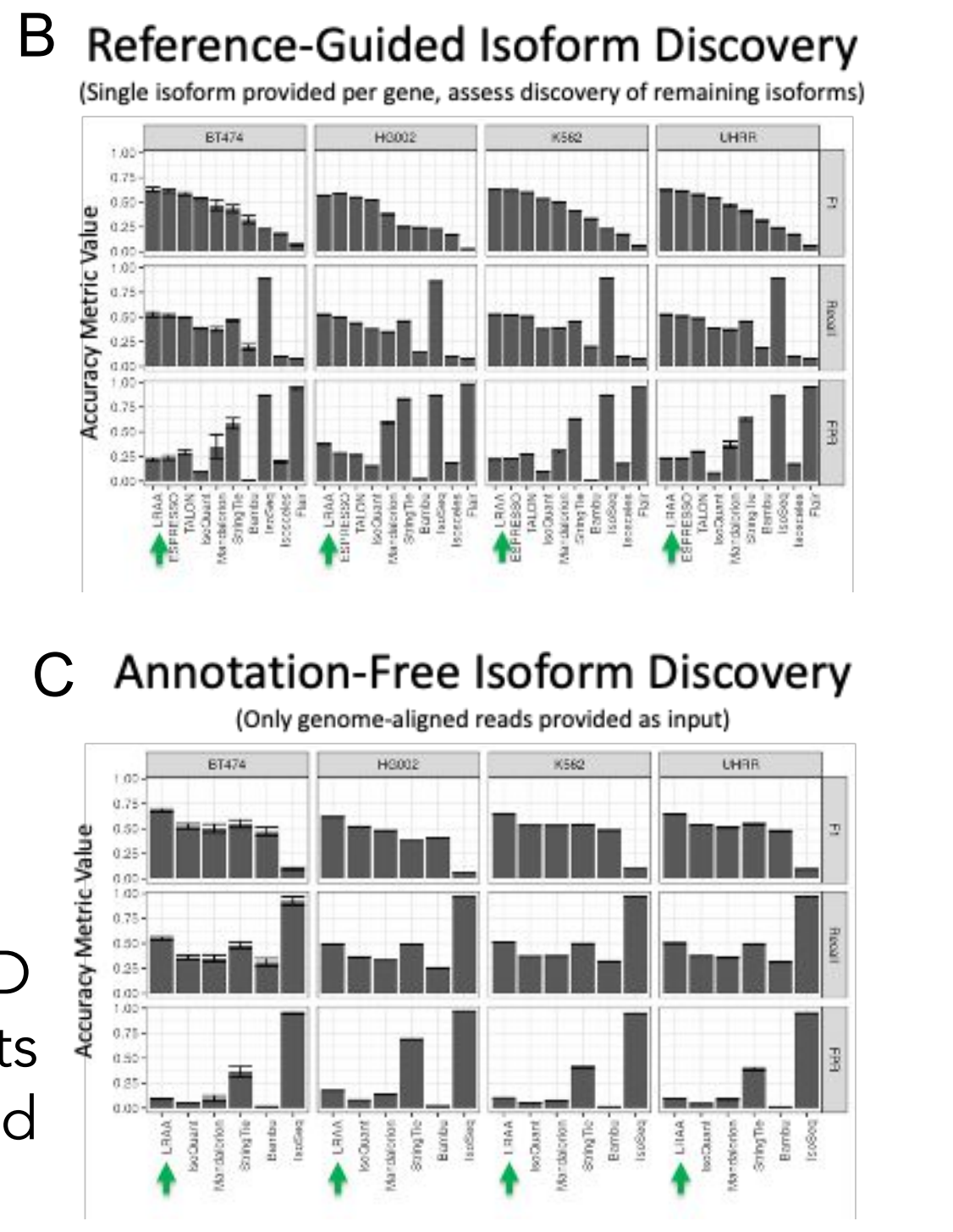


Figure 4. A) Table of benchmarked ID & quant tools. B,C) Benchmark results from BT474, K562, UHRR, and NA24385 Kinnex RNA-seq.

MDL Long-read analysis resource

We developed an online resource for analyzing bulk and single-cell RNA sequencing data generated using Kinnex long read technology (QR code above). With the latest methodologies and best practices, we've curated a series of workflows and, R and Python based notebook vignettes to support data processing of raw HiFi reads coming off sequencers through extracting cleaned segmented reads, alignment, performing read QC to eventual downstream analyses such as isoform quantification / discovery and differential splicing analysis. The workflows are made available via Dockstore as WDLs to enable deployment and scaling. The R vignettes are currently static needing offline execution whereas the Py based vignettes such as ITV are interactive notebook with code-along enabled using plug-in options such as thebe, Binder and Google Collab. Sub-sampled test data sets for various sub-workflows are made available via public google storage buckets to encourage further exploration.

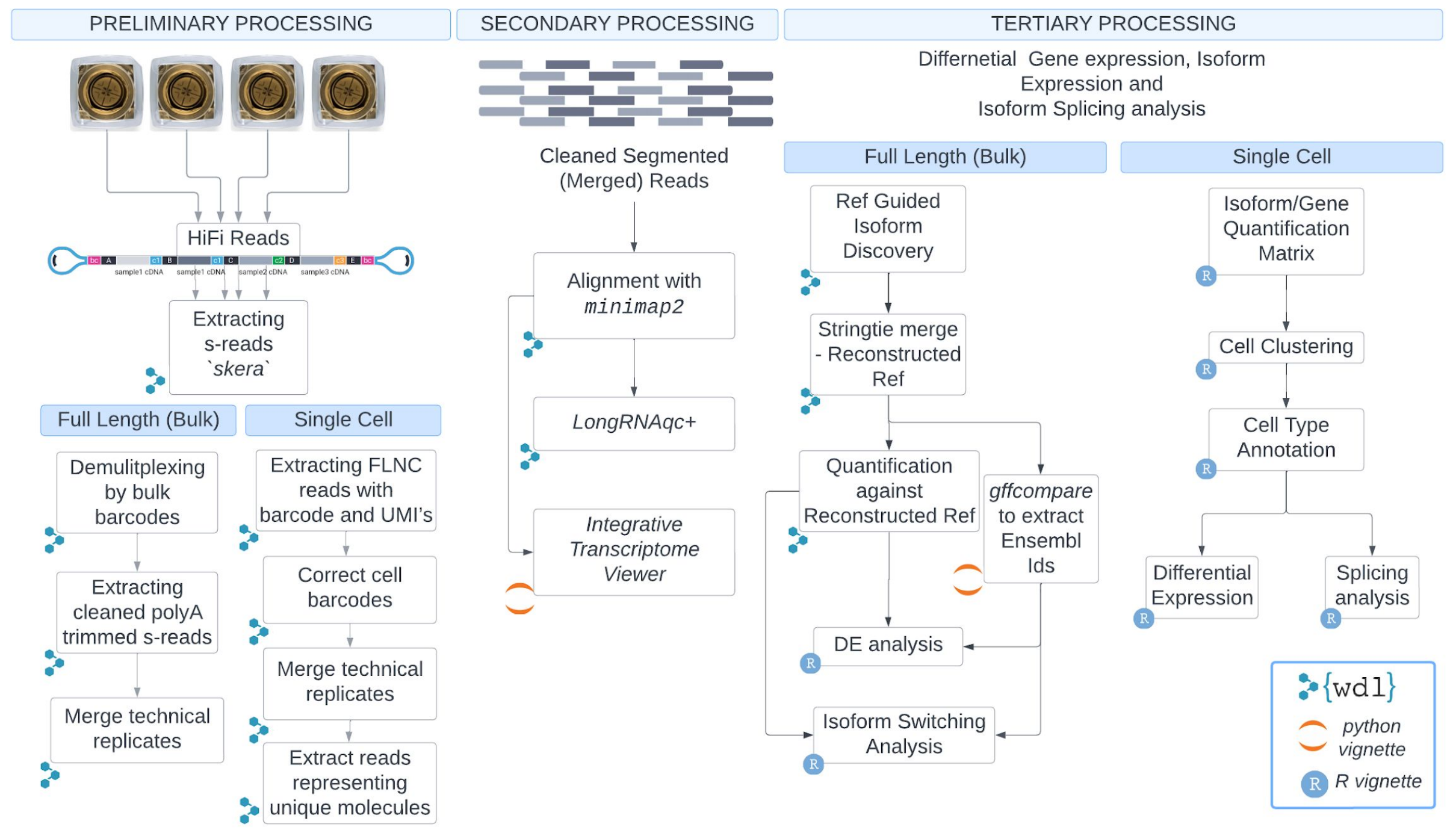

PRELIMINARY PROCESSING
SECONDARY PROCESSING
TERTIARY PROCESSING
HiFi Reads
Extracting s-reads `skera`
Full Length (Bulk)
Single Cell
Demulitplexing by bulk barcodes
Extracting cleaned polyA trimmed s-reads
Merge technical replicates
Extracting FLNC reads with barcode and UMI's
Correct cell barcodes
Merge technical replicates
Extract reads representing unique molecules
Cleaned Segmented (Merged) Reads
Alignment with minimap2
LongRNAqc+
Integrative Transcriptome Viewer
Differential Gene expression, Isoform Expression and Isoform Splicing analysis
Full Length (Bulk)
Single Cell
Ref Guided Isoform Discovery
Stringtie merge - Reconstructed Ref
Quantification against Reconstructed Ref
gffcompare to extract Ensembl Ids
DE analysis
Isoform Switching Analysis
Isoform/Gene Quantification Matrix
Cell Clustering
Cell Type Annotation
Differential Expression
Splicing analysis
{wdl}
python vignette
R vignette

Figure 5. Overview of workflows, tools, vignettes.

Integrated long-read RNA-seq analysis

Given the length, throughput, and accuracy of long-read transcriptomics, fusions and mutations can be derived from this rich datatype.

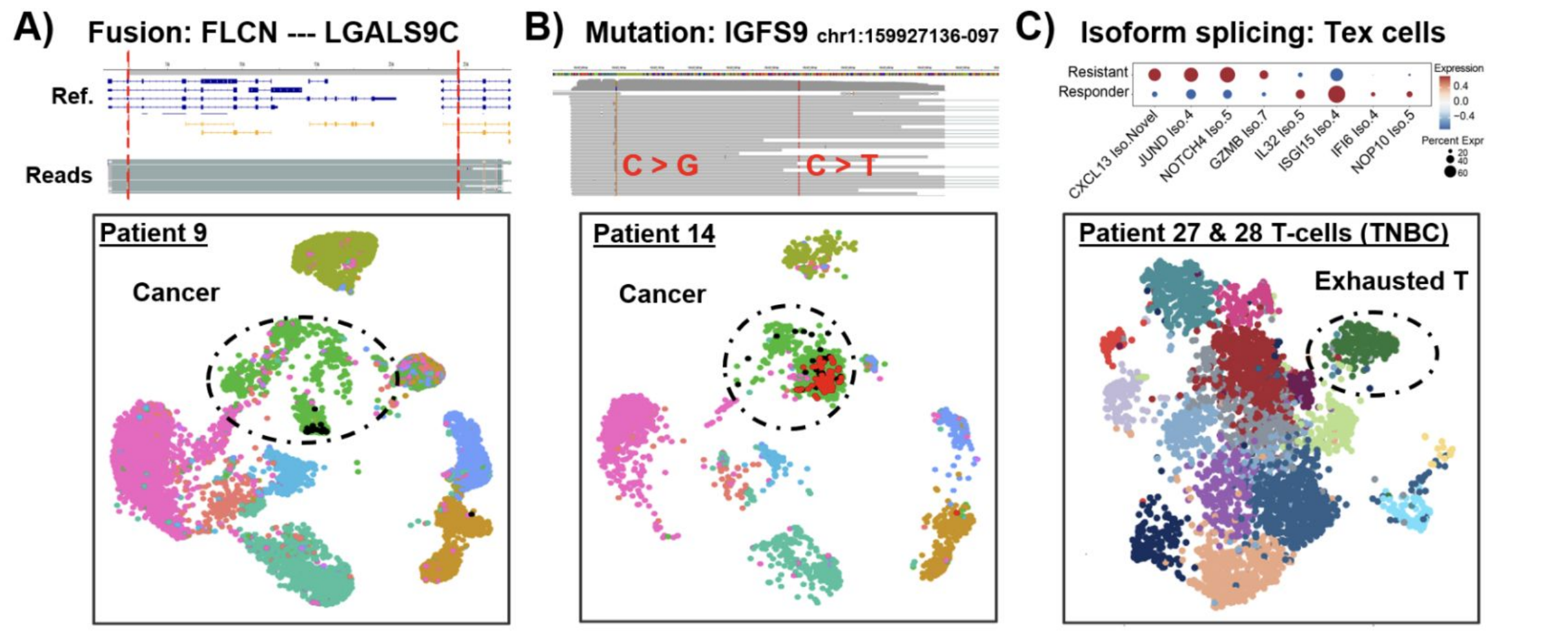


Figure 6. A) Identification of FLCN – LGALS9C fusion from patient 9. B) Pathogenic associated mutations (per COSMIC and gnomAD) found to be expressed in a subset of cancer cells. C) Top differentially expressed isoforms found in exhausted T-cells between TNBC patient 27 (responder) and 28 (non-responder).

PacBio and Broad Clinical Labs supported this work