Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press



k-mer-based Upstream Preprocessing of long reads for Isoform Discovery

Molly Borowiak and Yun William Yu

Genome Res. published online September 14, 2026
Access the most recent version at doi:10.1101/gr.282250.126

P<P Published online September 14, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

k-mer-based Upstream Preprocessing of long reads for Isoform Discovery

Molly Borowiak[1 [0009−0009−1195−7709]] and Yun William Yu[1 [0000−0002−8275−9576]]

Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract. Eukaryotic genes can encode multiple protein isoforms based on alternative splicing of their transcribed regions. Most modern novel isoform discovery methods function by identifying and assembling exon splice junctions from an RNA-seq sample. However, splice junctions can only be accurately annotated with time-intensive dynamic programming alignment. This manuscript introduces KuPID, a method for preprocessing long RNA-seq reads with the goal of better identifying novel isoform transcripts. KuPID utilizes *k*-mer sketching as a pre-filter to quickly pseudo-align reads to known reference isoforms. Full alignment need only then be applied to reads that are most relevant to isoform discovery. Not only does KuPID speed up the discovery pipeline, it also increases downstream accuracy by filtering out extraneous reads. KuPID preprocessing simultaneously increases the f1 accuracy of isoform discovery pipelines by up to 11.6 points while decreasing the runtime by a factor of 2-3×. An optional mode permits a KuPID sample to be paired with both isoform discovery and transcript quantification.

Keywords: novel isoform discovery · sketching · RNA-seq

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

Transcription is the process by which DNA segments are copied and synthesized into proteins. Complex regulatory mechanisms enrich the size and diversity of the human transcriptome. The DNA sequence for a eukaryotic gene is composed of both protein-coding (exon) and non-coding (intron) regions. When transcribing a gene, pre-messenger RNAs are spliced to remove the gene's introns and ligate its exons into the final transcript. Alternative splicing (AS) enables that same gene to encode multiple alternative protein isoforms. AS events can involve intron retention, exon skipping, or alternative splice sites within the exon (Wang et al. 2015). Changes in transcriptional promoter usage and polyadenylation further increases the amount of isoform diversity (Reyes and Huber 2018).

As of now, more than 95% of human genes have been found to undergo alternative splicing (Nilsen and Graveley 2010). Novel isoform discovery methods aim to detect and annotate alternative transcripts to complete the expanding human transcriptome. By doing so, they can also grant insight into the mechanisms of various biological functions. Alternative splicing events have been found to have well-established roles in processes ranging from cell differentiation and survival, to responses to stress (Arzalluz-Luque and Conesa 2018). In addition, aberrant splicing has been associated with multiple disease-causing single nucleotide polymorphisms (Singh and Cooper 2012). Isoform discovery (ID) is a common feature of modern transcript quantifiers, which measure changes in isoform abundances through high-throughput RNA sequencing (RNA-seq). (Chen et al. 2023; Prjibelski et al. 2023; Wyman et al. 2020) This pairing enables analysis of how the expression of alternative isoforms can differ in response to conditions such as cell state and disease, granting further insight into their function (Finotello and Di Camillo 2015).

Despite their importance, it can be difficult to detect many functionally relevant novel isoforms. Unlike constitutively spliced transcripts, alternative transcripts often exhibit unique temporal or cell-type specific expression patterns (Leparc and Mitra 2007). Without prior knowledge of their required underlying conditions, such isoforms might rarely appear in an RNA-seq sample. Furthermore, many discovery methods employ read support thresholds when evaluating potential novel isoforms (Prjibelski et al. 2023; Tang et al. 2020). If the rare isoforms are expressed at low levels, ID methods could still fail to detect them in the samples where they are present. Accurate isoform discovery methods are vital to ensure that context-specific isoforms can be studied.

The recent rise of the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms sequencing have been vital for isoform discovery. The short reads produced by high-throughput sequencing are highly accurate, but can omit the ordering of exons and introns in the original transcript (Wyman et al. 2020). The PacBio and ONT platforms can instead sequence full-length transcripts, making it easier to

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

accurately assemble the final isoform (Amarasinghe et al. 2020; Branton et al. 2008; Rhoads and Au 2015) . Key inefficiencies remain in modern discovery pipelines even with the use of long transcript reads. Current methods require all RNA-seq reads to be aligned to a reference genome. The splice junctions of the reads are then identified and assembled to detect novel isoforms (Kovaka et al. 2019; Prjibelski et al. 2023; Tang et al. 2020). Unfortunately, alignment is time-intensive. The main output of isoform discovery is the exon annotations of the newly found transcripts. For this aim, only the alignment of reads transcribed from novel isoforms should be necessary. However, we typically cannot tell which reads are novel without some form of reference alignment. Given the size of the average RNA-seq experiment, isoform discovery methods end up processing millions of query reads that are irrelevant for discovery.

In this manuscript, we introduce k-mer -based Upstream Preprocessing for Isoform Discovery, or KuPID. KuPID reduces runtime by filtering RNA-seq reads to those likely to represent novel transcripts so only they need to be mapped. Surprisingly, despite being a lossy filtering step, KuPID actually improves downstream accuracy for the ID pipelines it's paired with. Instead of trading off one for the other, KuPID provides gains in both speed and accuracy.

Results

KuPID-discovery Enables More Accurate and Efficient Isoform Discovery

We evaluated KuPID's performance when paired with modern isoform discovery methods to analyze simulated RNA-seq HiFi reads. The datasets contained varying percentages of reads originating from novel isoforms. We omitted a trial for one of the ID methods, flair, where 80 % of the reads from the YASIM-generated isoforms were novel. Both StringTie2 and IsoQuant successfully produced ID results for this trial; however, flair threw an unresolved exception error while working with the KuPID processed reads. For the remaining trials, the application of KuPID preprocessing increased the f1 accuracy of all three ID pipelines in comparison to using the original RNA-seq sample (Fig 1A). This improvement was consistent across the YASIM and reduction methods of representing the novel transcripts. For the majority of the pipelines, preprocessing with KuPID increased both the precision and recall of the ID results (Fig 1B). We posit that the increase in precision was directly related to KuPID reducing the percentage of annotated reads in the sample (Supplementary Table S1). The precision would worsen whenever the discovery pipelines assembled a transcript model that was not actually present in the sample. Any annotated reads could be included in an assembly and lead to the ID method's hallucinating a false positive. By removing extraneous annotated reads, KuPID curtails the number of false positive models.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

We further validated that KuPID could improve isoform discovery results using the SIRV-Set4 sequencing data published with IsoQuant (Prjibelski et al. 2023). For several of the ONT samples, the application of KuPID increased the average recall and precision (Fig 2). In comparison to the simulated HiFi reads, there was more variability in the ID accuracy for the real ONT reads. With that said, the accuracy of the SIRV results typically increased once the reads were preprocessed with KuPID, even if only marginally (Fig 2).

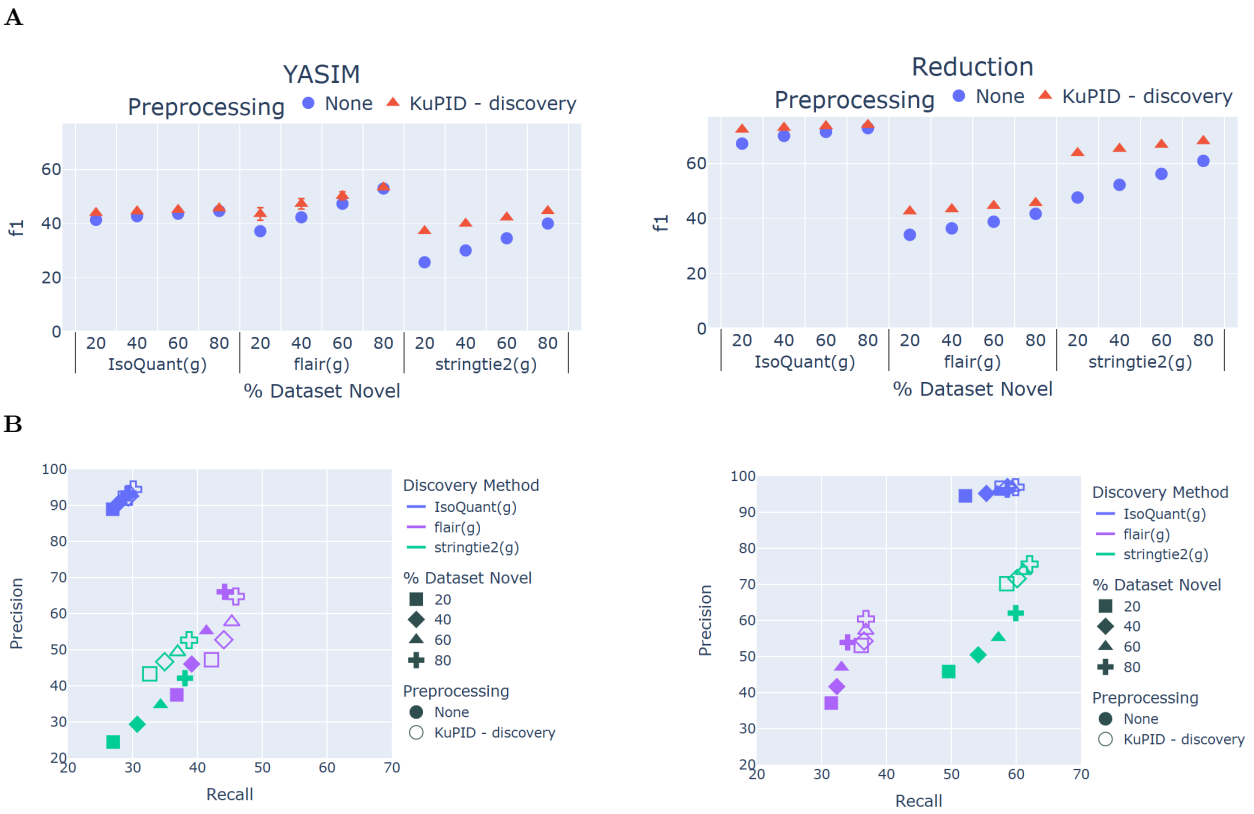


Fig. 1: **Evaluation of Isoform Discovery Results using simulated HiFi KuPID-discovery reads. A**, f1 accuracy of the novel transcripts found by the isoform discovery methods. **B**, Precision and recall of the true novel transcripts found by isoform discovery (ID) methods. Reads were simulated HiFi sequences. The ID methods were applied to either non-processed reads or reads processed with KuPID in 'discovery' mode. The percentage of reads from novel transcripts in the original sample ranged from 20-80%. The novel transcripts were generated using either the YASIM (left) or reduction method (right).

KuPID-discovery also minimized alignment bottlenecks by reducing the size of the sample file. For the majority of the simulated HiFi datasets, it was faster to run KuPID-discovery and then apply the minimap2 aligner than it was to align the non-processed reads (Fig 3). The exception to this was the extreme case of the dataset containing 80% or more novel reads. The runtime speed-up was most prominent for the datasets that initially contained few novel reads, as they saw the greatest reduction in sample size following the application of KuPID. In the best case scenario where as few as 20% of the initial sample reads were novel, it was at least

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

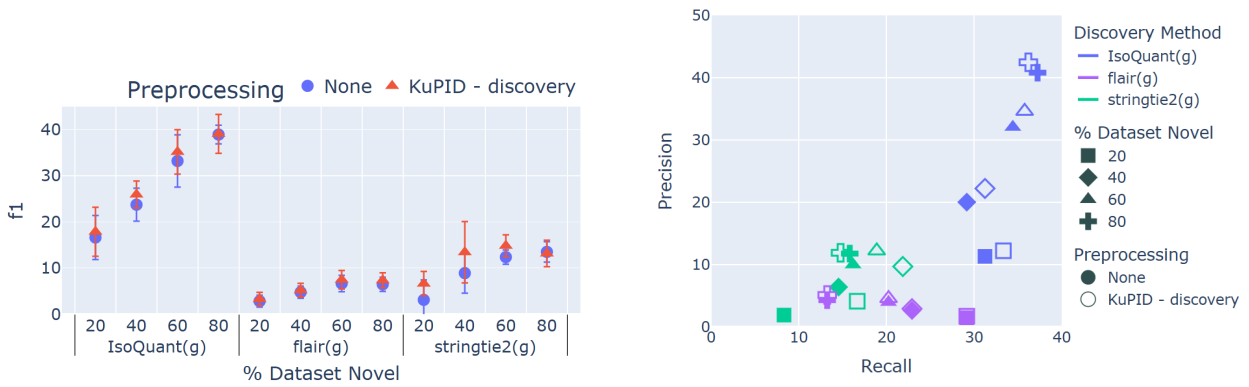


Fig. 2: **Evaluation of Isoform Discovery Results using real ONT KuPID-discovery reads.** f1 accuracy of the novel transcripts found by the isoform discovery methods (left). Precision and recall of the true novel transcripts found by isoform discovery (ID) methods (right). Reads were sequenced from SIRV-Set 4 with the R10.4 platform. The ID methods were applied to either non-processed reads or reads processed with KuPID in 'discovery' mode. Novel annotations were created via the reduction method. The percentage of the annotations that were novel ranged from 20-80%.

2-3x as fast to run the sample through KuPID than to apply full alignment. KuPID orchestrated a similar reduction in runtime when it was paired with the real ONT reads (Supplementary Figure S1). The ONT dataset was relatively small even without preprocessing, containing about 250,000 reads. Even so, applying KuPID to datasets where as few as 20% of annotations were novel led to a 2× speedup in alignment.

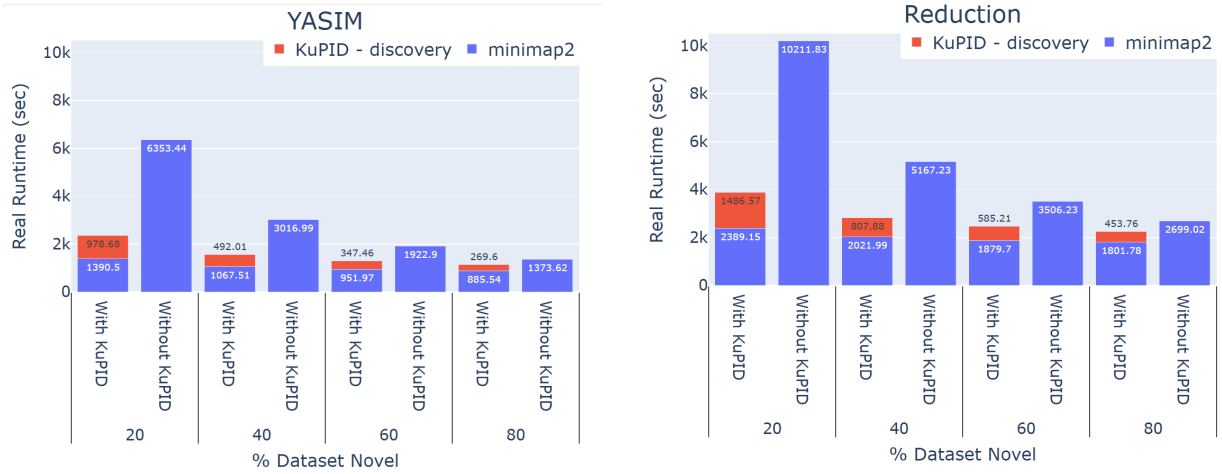


Fig. 3: **Applying KuPID to RNA-seq samples reduces alignment runtime.** Average time in seconds needed to align simulated HiFi RNA-seq reads with the minimap2 software. We compare the runtime of aligning the non-processed reads to the reads selected by KuPID-discovery. The runtime for the KuPID reads encompasses both the time of aligning the reads and the time of applying KuPID to the original sample. The novel transcripts were generated using either the YASIM (left) or reduction method (right).

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

KuPID-discovery can Detect Difficult Novel Transcripts

Surprisingly, the addition of KuPID to the ID pipelines often improved the recall performance (Fig 1B). With KuPID reads, the discovery methods could detect a number of novel transcripts that would otherwise have been hidden from them in the original RNA-seq sample. The blame for the worse recall seems to lie with the annotated reads. With the simulated HiFi datasets where KuPID was not applied, the ID methods consistently recalled fewer novel transcripts from samples with higher proportions of annotated reads (Fig 1B). These datasets contained identical novel reads, and only differed in their number of annotated transcripts. Thus, the inclusion of annotated reads in the sample appears to mask novel transcripts from the ID pipelines.

The impact of annotated reads was felt not just across the sample, but particularly at the gene-level. We found that the ID method's ability to detect novel transcripts was also impacted by the composition of isoforms expressed by a given gene. In our simulated data, any gene that expressed a novel transcript could simultaneously express annotated transcripts as well. Based on the original RNA-seq samples, we compared the ability of the ID methods to detect novel transcripts that were expressed alongside annotated reads versus novel transcripts that were not. ID pipelines often had less success in detecting novel isoforms transcribed from genes with mixed expression (Fig 4A). The addition of KuPID to the pipelines mitigated the effects of masking and enabled the discovery of these difficult transcripts. Preprocessing the RNA-seq samples increased recall of alternative transcripts at both the gene-(Fig 4B) and sample-wide levels (Fig 1B). The magnitude of KuPID's impact on recall was not uniform. For IsoQuant, the improvement of recall at the gene-level was somewhat marginal. Of all the pipelines, StringTie2 had the largest increase in the number of novel transcripts detected when paired with KuPID.

KuPID-quantify Enables More Efficient Transcript Quantification

KuPID's main mode (KuPID-discovery) prioritizes the accuracy of isoform discovery methods. KuPID's optional mode (KuPID-quantify) can accommodate both discovery and quantifcation analysis. To illustrate this, we employed KuPID-quantify to create a candidate set for each of the simulated HiFi samples. We then applied the discovery and quantification modules of IsoQuant and StringTie2. We omitted the results from flair due to unforeseen exception errors when running the software. We found that the discovery results from the KuPID reads closely mimicked the results from the non-processed reads (Fig 5A). The KuPID reads also performed reasonably well in quantifying the abundance of annotated transcripts. We calculated the Spearman's rank correlation between the true abundance of each reference isoform (reported in Transcripts per Million) and the predicted abundance. Across the data samples, the Spearman's correlation for the

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

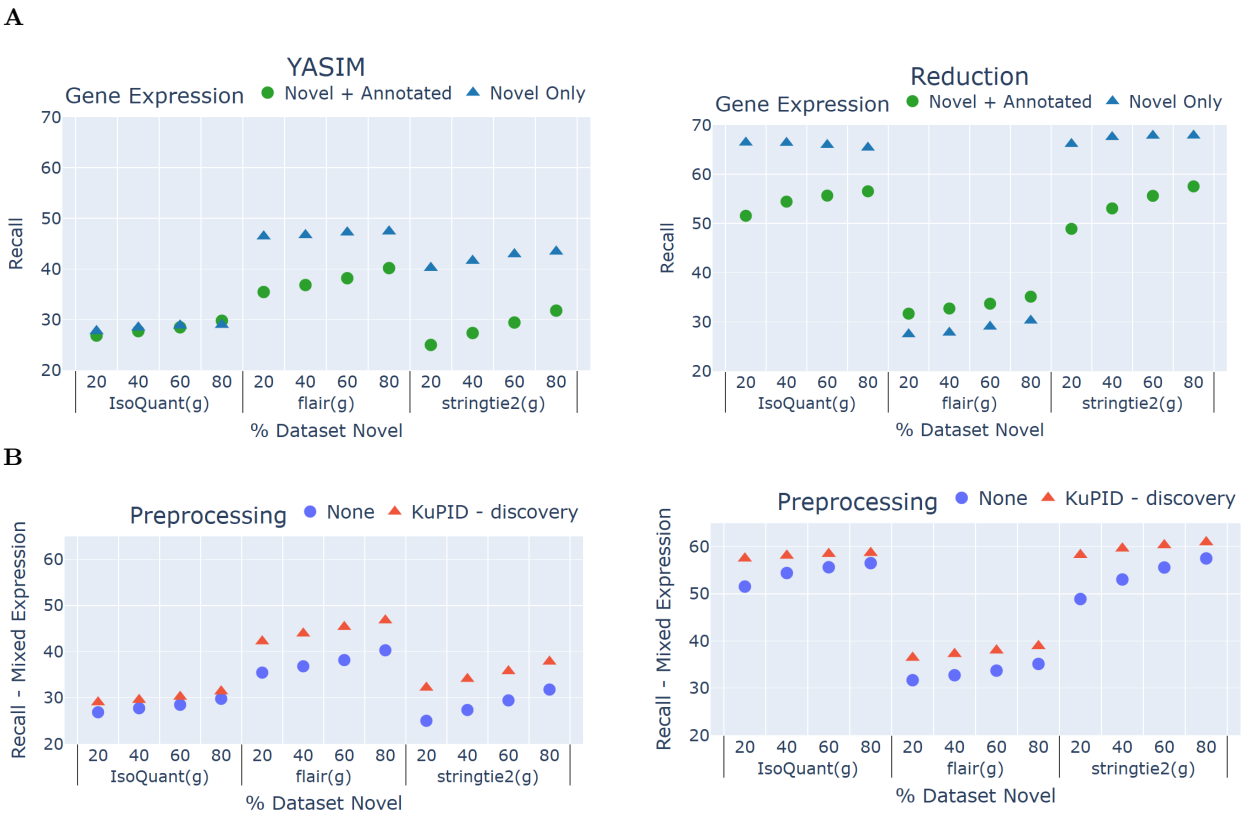


Fig. 4: **Isoform discovery methods have difficulty in detecting isoforms from genes with mixed expression. KuPID can detect novel transcripts that are otherwise masked by annotated reads.** **A**, Recall of the true novel transcripts found by isoform discovery (ID) methods when applied to non-processed RNA-seq samples. Recall is compared between genes that only expressed novel isoforms and genes that expressed both novel and annotated isoforms. **B**, Recall of the novel isoforms transcribed from genes with mixed expression. The isoform discovery methods were applied to either the non-processed RNA-seq sample or the reads selected by KuPID-discovery. RNA-seq samples were comprised of simulated PacBio HiFi reads. The novel transcripts were generated using either the YASIM (left) or reduction method (right).

KuPID processed reads was at least as strong as the correlation with the non-processed reads (Fig 5B). In terms of runtime, KuPID-quantify had similiar efficiency to KuPID-discovery. KuPID-quantify consistently reduced alignment time for all datasets with fewer than 80% novel reads (Fig 5C).

Discussion

Our method KuPID can preprocess long RNA-seq reads to create an optimal subsample for either novel isoform discovery or transcript quantification. By eliminating annotated reads from data samples, KuPID-quantify can closely approximate discovery and quantification results while decreasing the amount of time needed for downstream analysis. In comparison, KuPID-discovery can improve both the efficiency and the accuracy of the ID pipelines it's paired with. KuPID-discovery is most effective when it is applied to data

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

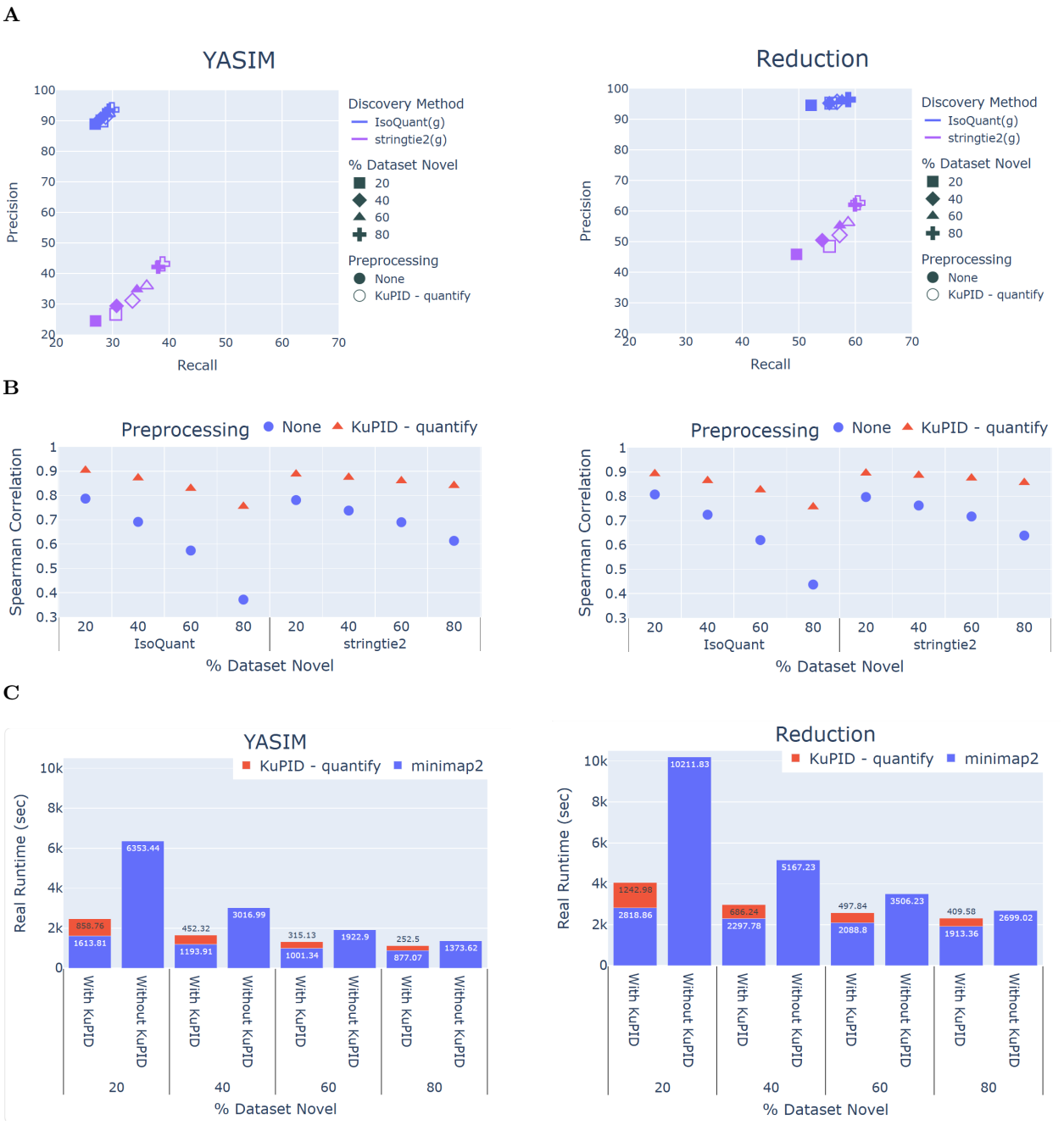


Fig. 5: **Evaluation of Transcript Discovery and Quantification Performance with KuPID-quantify reads. A**, Precision and recall of the true novel transcripts found by isoform discovery (ID) methods. The discovery pipelines were applied to either non-processed RNA-seq samples or to the reads selected by KuPID-quantify . **B**, Spearman's coefficient of the relationship between the true and predicted abundances of annotated isoforms in the sample. Abundances were reported in Transcripts Per Million. **C**, Average time in seconds needed to align RNA-seq reads with the minimap2 software. We compare the runtime of aligning the non-processed reads to the reads selected by KuPID-quantify. RNA-seq samples were comprised of simulated PacBio HiFi reads. The novel transcripts were generated using either the YASIM (left) or reduction method (right).

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

samples where the minority of reads are novel. Even so, KuPID can improve the accuracy of discovery methods irrespective of the total amount of novel reads in the sample.

KuPID is robust to not only the amount of novel reads present, but to the types of novel isoforms as well. We generated novel isoforms with two methods: either creating new isoforms with the external software YASIM, or by randomly reducing the reference transcriptome and treating the removed isoforms as novel. KuPID-discovery successfully improved the accuracy of the ID results of the datasets created with both methods. Notably, the IsoQuant and StringTie2 methods had greater success in recalling novel transcripts from the reduced method than from the YASIM method (Fig 1B). This trend was consistent across both the original RNA-seq samples and the KuPID processed reads. One possible explanation could be that the reduction method produced a high number of transcripts with novel exons. With our simulated HiFi reads, each set of novel isoforms created through reduction had $\approx$ 20,000 novel exons, while the YASIM set did not contain any novel exons. This seems to indicate a possible trend of the presence of novel exons influencing whether or not a novel isoform will be detected.

Unexpectedly, ID pipelines that pre-processed the HiFi samples with KuPID were able to detect a greater amount of the novel transcripts that were present. This phenomenon likely stems from KuPID reducing bias from annotated reads from the prospective transcript models. Besides employing support thresholds, current isoform discovery methods tend to favor transcripts with the greatest amount of read support. The base stringtie algorithm employs a network flow algorithm to reconstruct transcripts (Pertea et al. 2015). A splice graph is constructed from the sample reads, with each splice path representing a transcript. The algorithm iteratively finds the transcript with the heaviest read coverage before assigning reads to it. IsoQuant uses a similar intron graph algorithm, and removes any transcript with less than 2% read support from the maximum graph coverage (Prjibelski et al. 2023). Selecting models with the greatest amount of support is intuitive, but read support can be inflated if a model is covered by both annotated and novel reads. To illustrate this, we found that isoform pipelines have more difficulty in detecting novel transcripts when they are expressed by genes that also express annotated transcripts (Fig 4B). By filtering out annotated reads at the gene-level, KuPID can likely remove bias and improve the recall of novel isoforms.

Ensuring that isoform discovery pipelines have sufficiently high recall is vital for detecting context-specific novel transcripts. Alternative transcripts can have unique expression patterns based on a given environment or cell-type (Leparc and Mitra 2007). Practically, rare or expensive conditions will have a limited amount of replicates to analyze for novel transcripts. ID pipelines that involve KuPID are able to remove extraneous reads that would otherwise have occluded context-specific transcripts from discovery.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Methods

KuPID Algorithm

KuPID processes long RNA read sequences exclusively. It requires a reference transcriptome formatted as the sequences of all known transcripts. In its main 'discovery' mode, KuPID will output a subset of the reads that are most likely to be sequenced from a novel isoform. The read subset can be submitted to the isoform discovery method of the user's choice. KuPID consists of three main stages (Fig 6):

1. ***k*-mer sketching**: We create simplified representations of the RNA-seq reads and reference transcriptome that will enable fast comparisons between the two.
2. **Pseudo-alignment to reference transcriptome**: Each RNA-seq read is pseudo-aligned with a sparse chaining procedure that finds the largest set of colinear k-mer matches between the read and the reference transcriptome.
3. **Read Selection**: We use the sparse chaining results to gauge if any of the reads lack a strong match with the set of known isoforms. If there are substantial gaps in the pseudo-alignment, KuPID concludes that it is likely transcribed from a novel isoform.

KuPID offers an optional mode for end-users that also want to quantify the transcripts in their RNA-seq sample. In its 'quantify' mode, KuPID will output a subsample of the reads that map to each annotated isoform in addition to the likely novel reads. This mode speeds up alignment, but does not prioritize improvements in isoform discovery.

***k*-mer Sketching via FracMinHash** k-mer sketching methods are a class of randomized algorithms that focus on constructing smaller and simplified representations of string-formatted data. When applied to large datasets with complex tasks, such as large-scale sequence comparison, sketching offers improved storage and runtime efficiency (Roberts et al. 2004). We convert every RNA-seq read and reference transcript into a subset of its representative k-mer s using the FracMinHash sketch method (Irber et al. 2022). We define $\Sigma = \{A, C, G, T\}$, and let Σ^k represent the set of nucleotide k-mer s, or k-length substrings. A hash function $h : \Sigma^k \rightarrow [1, M]$ maps every k-mer in the sequence to a numerical value. In our implementation, we apply the invertible integer hash function from minimap (Li 2016). The k-mer sketch for a sequence S with a set of k-mer s $W \subset \Sigma^k$ is given as $F(S) = \{w | w \in W, h(w) \leq \frac{M}{\gamma}\}$, where γ determines the fraction of k-mer s chosen.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

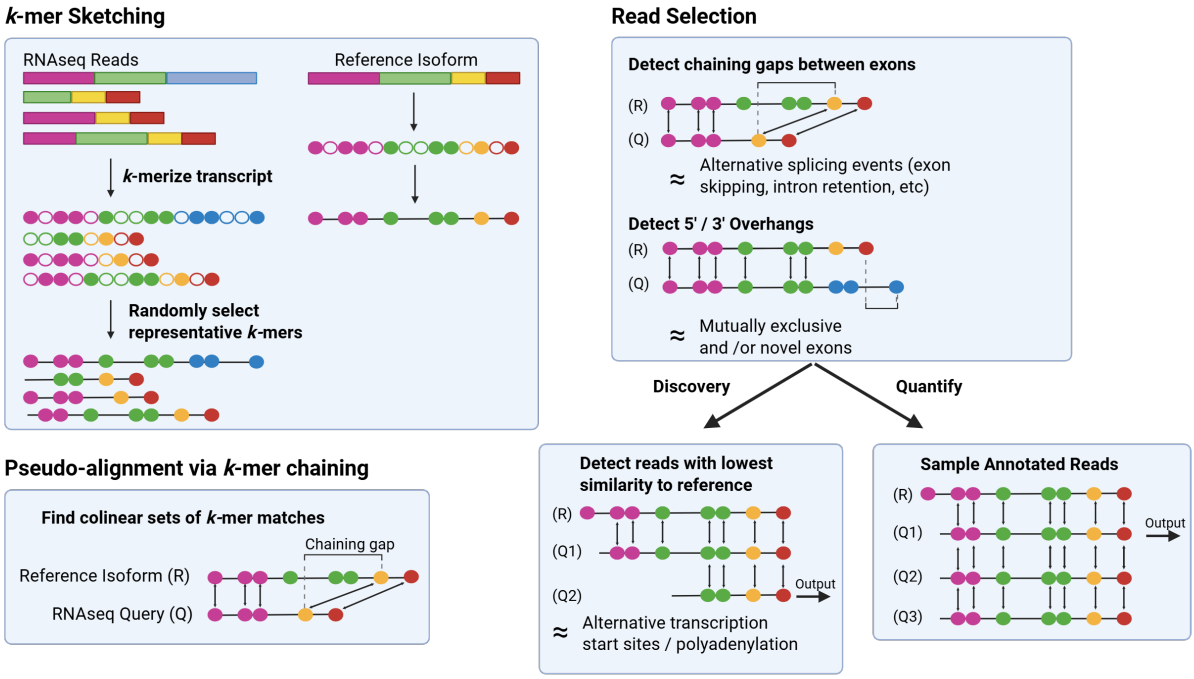


Fig. 6: **KuPID pipeline for detecting novel isoform reads.** Given an RNA-seq sample, KuPID returns a FASTA of the reads that were likely transcribed from novel isoforms. First, it converts the sample and a reference transcriptome into compact k-mer sketches. Then, it finds k-mer matches between the sketches to pseudo-align the RNA-seq queries to the reference. KuPID determines which reads are novel based on if they have substantial unmatched regions. Given the user's choice of mode, KuPID can also output a subsample of the reads mapped to each reference isoform for quantification (Borowiak 2025)

Pseudo-alignment to Reference Transcriptome We apply pseudoalignment to quickly determine the most similar reference isoform for a given query read. We first determine a limited reference space $S(q)$ by the set of isoforms that share at least 1 sketched k-mer with the mapping query q.

We define R as the set of all reference isoforms, Q as the set of all RNA-seq queries, and Σ^k as the set of all potential k-mer s. To quickly find $S(q)$, we use a separate data structure $\kappa : \Sigma^k \rightarrow \mathcal{P}(R)$ that maps each sketched k-mer to the isoforms it's found in. For a given k-mer x, $\kappa(x) = \{r | x \in F(r)\}$. A size limit of $|\kappa(x)| \leq b$ is placed to remove common k-mer s and to curtail the size of the search space. We can then formally define the search space $S(q)$ by $S(q) = \cup_{x \in F(q)} \kappa(x)$.

We narrow the search space further by selecting all isoforms from $S(q)$ that have the maximum number of unique k-mer matches with the query. We define the set of selected isoforms as $M(q) = \text{argmax}_{r \in S(q)} |F(r) \cap F(q)|$. We then create an anchor table A_{qr} of exact k-mer matches between the query q and each of the selected isoforms $r \in M(q)$. Each anchor $x_i = (x_{1i}, x_{2i})$ represents a match between the k-mer at position x_{1i} on q and position x_{2i} on r. We apply a dynamic programming (DP) algorithm to determine the optimal

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

chain, or subsequence of colinear anchors. Every anchor in the optimal chain $A^*_{qr} \subseteq A_{qr}$ is ordered and strictly increasing so that $\forall x \in A^*_{qr}, x_i \prec x_j \iff x_{1i} < x_{1j}$ and $x_{2i} < x_{2j}$ (Jain et al. 2022).

We define $f[i]$ as our optimal chaining score up to the ith anchor in A_{qr}. $f[i]$ is equivalent to the total number of anchors in the ordering, and is designed to maximize the amount of viable anchor matches between the query and the reference. We define $A^*[i] \subseteq A_{qr}$ as the optimal chain up to the ith anchor in A_{qr}. Throughout the chaining process we track $FP[i]$ and $LP[i]$, the first and last anchor in the optimal chain's ordering, along with $g[i]$, the largest gap up to the ith anchor. We define the gap between two adjacent anchors x_i and x_j in the chain as $g = |(x_{1j} - x_{1i}) - (x_{2j} - x_{2i})|$. The DP table for the optimal chaining score is filled out as follows:

$$f[i] = max \left\{ \begin{array}{l} 1, \\ max_{i>j\geq i-m} f(j) + 1, \text{ if } LP[j] \prec x_i \end{array} \right\} \quad (1)$$

Like similar anchor-based chaining methods, we employ a banded heuristic to limit the runtime of the DP algorithm (Li 2018; Shaw and Yu 2023). With our simplified chaining formula, many of the anchors will chain to the anchor that directly precedes it. The rare exception to this rule will be if an anchor match involves an erroneous k-mer , allowing the use of a small bandwidth m. During the chaining process, we omit repetitive k-mer s that occur at multiple positions along the query or reference sequences. Accurately incorporating repetitive regions into the chains often necessitates a larger bandwidth. We found that the presence of gaps and the score of the optimal chain are accurately estimated even when repetitive k-mer s are ignored.

Following pseudoalignment, we calculate a similarity score for each RNA-seq query to gauge how well it matches to its most probable reference sequence. The score is based on the Jaccard index, a known metric in gauging overlap between k-mer sketches (Ondov et al. 2016). The Jaccard index corresponds to $J(A, B) = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$, where A and B are k-mer sets from different sequences. We found the Jaccard index was an excellent base metric, but wanted to incorporate the multiplicity and ordering of the k-mer s in the chain. We define $\{F[S]\}$ as the multiset representation of the sketch $F[S]$. The multiplicity of a k-mer $x \in \{F[S]\}$ reflects the number of times the k-mer occurs in the sequence S. The final similarity score for query read q is:

$$I(q) = max_{r \in M(q)} \left\{ \frac{f(|A_{qr}|)}{|\{F(r)\}| + |\{F(q)\}| - f(|A_{qr}|)\}} \right\}$$

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Our chaining procedure is noticeably simpler than other anchor-based methods. Anchor-based chaining methods are often used as an intermediate step in seed-chain-extend alignment. After finding the optimal set of colinear anchors for two sequences, dynamic programming alignment can be extended to the unmatched regions between chains. Existing chaining methods often penalize large gaps between anchors and / or incorporate the amount of anchor overlap into the chain score (Jain et al. 2022). These scores are more appropriate for producing an accurate final alignment. However, we found that our pseudo-alignment without extension is sufficient to judge if a query is similar to any reference isoform. Our DP algorithm also allows large gaps in the optimal chain instead of penalizing them. Tracking the presence of gaps allows us to detect discrepancies in the distances between k-mer matches that could indicate alternative splicing.

Read Selection Novel candidate reads are chosen based on the following criteria:

1. Evidence of alternative splicing (AS) events
2. Evidence of mutually exclusive and / or novel exons
3. Evidence of alternative transcription start / stop sites (ATSS)

For a query read q, the chaining procedure returns the optimal chaining score and the largest gap $g^*(q)$ in the chain. We define n as the expected minimum length of an exon in the reference transcriptome. If $g^*(q) > n$, we infer the read has an AS event and output it as a novel candidate.

We then select the reads that have discernible novel exons. While novel exons can create gaps in the pseudo-alignment, the chaining procedure doesn't check for reads that either begin or terminate with a novel exon. For each pairwise pseudo-alignment, we calculate the lengths of any unmatched 5'/3' overhangs along the reference and query sequences.

We define

$$O(q) = max \left\{ \begin{array}{l} 0 \\ FP^*[|A_{qr}|]_1 - FP^*[|A_{qr}|]_2 \\ (|q| - LP^*[|A_{rq}|]_1) - (|r| - LP^*[|A_{rq}|]_2) \end{array} \right\}$$

$$O(r) = max \left\{ \begin{array}{l} 0 \\ FP^*[|A_{qr}|]_2 - FP^*[|A_{qr}|]_1 \\ (|r| - LP^*[|A_{rq}|]_2) - (|q| - LP^*[|A_{rq}|]_1) \end{array} \right\}$$

When $O(q) > 0$ or $O(r) > 0$, there is a discrepancy in the lengths of the exons that lie at the 5' / 3' ends of the two sequences. If the discrepancy is sufficiently large, it indicates the presence of a mutually exclusive or novel exon in the query. We automatically select the query read if $O(q) > n$, as it indicates that the query has at least 1 exon that is not present in the reference transcript. If $O(r) > 0$, there are 2 possible

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

scenarios: 1) The query read possesses a shorter novel exon, or 2) the query read was prematurely truncated. To check for truncation, we compare the k-mer s in the unmatched 5' / 3' region of the query and reference. We specifically use every k-mer , not just those that were initially selected for the sketch. If the query read was truncated, then its unmatched region should only map to exons contained in the reference. Under this assumption, a query region of length q_l will be expected to have $(q_l - k + 1) * (1 - ke)$ matches with the reference. Any query read with fewer matches will be selected as having a novel exon. On the other hand, if the unmatched region is shorter than n bps, we will assume that it does not represent a full novel exon.

KuPID selects additional reads based on the mode selected by the end-user. KuPID offers two modes: 'discovery' and 'quantify'. While the former prioritizes novel isoform discovery alone, the latter can be paired with transcript quantification. In 'discovery' mode, KuPID will detect reads with ATSS based on their similarity scores. In theory, the lower the similarity score, the more likely the read is to represent a novel isoform. In practice, it can be difficult to distinguish individual novel and annotated reads from each other. Long sequence reads are frequently truncated due to nonsense mutations, fragmentation during library prep, or RNA degradation (Amarasinghe et al. 2020; Monzó et al. 2025). As such, we focus primarily on classifying whole groups of reads that map to a single isoform. The higher the average similarity score of a read group, the more likely it contains only annotated reads. We iteratively add all the reads from the group with the lowest squared average score until we reach the desired number of novel candidates. We limit the size of the candidate set to $c|AS| - |NE|$, where AS is the set of chosen reads with alternative splicing and NE is the set of reads with novel exons. c is a user-selected parameter that reflects the expected ratio of novel reads without AS to novel reads with AS within the sample. Additionally, we filter out truncation artifacts based on the assumption that true novel isoforms will express more than 1 read. While this may not be accurate for experiments with low sequencing depth, many discovery methods employ read support cut-offs when creating their transcript models (Prjibelski et al. 2023, Tang et al. 2020). Transcript models supported by a single read are often omitted by default. We assume any two reads encompassing the same set of exons will have comparable similarity scores within $\pm w = \frac{d}{R_l}$ of each other. d is defined as the expected median exon length in bps, and R_l as the length of the reference sequence. If any read does not have a similarity score within $\pm w$ of the other reads among its isoform group, it will be discarded.

In 'quantify' mode, KuPID will randomly select at most l of the reads that map to each known isoform in the RNA-seq sample. KuPID will also return a scale factor for each annotated isoform based on the total number of reads that mapped to it. After running the KuPID sample through the quantification method of their choice, a provided script will derive the final quantification results by multiplying the abundance estimates by the scale factors.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Data Preparation

KuPID was evaluated using a set of simulated PacBio HiFi reads from Chr1-22 of the human p14 genome. We created novel isforms via two methods:

1. **YASIM**: An external method that generates isoforms as novel combinations of existing splice junctions (Su et al. 2024)
2. **Reduction**: Selecting a random subset of isoforms to treat as novel

We generated 4 novel annotation sets with each method. Based on our chosen complexity index of 9, YASIM automatically generated approximately 43000 isoforms for each annotation set, with a mean of 4.46 isoforms per gene. For the reduction method, we randomly chose 20% of the isoforms in Chr1-22 of the human genome to be novel and removed those isoforms from the reference set. The isoforms from the reduction method were filtered so that each gene in the reference retained at least 1 transcript. For both methods, the novel annotation sets were filtered so that every transcript was at least 250 base pairs (bp) long and so there were no intron chain matches with the reference.

We used YASIM to generate the sequencing depth of transcripts within the novel and reference annotation sets. The novel and reference sets each had an average gene depth of 20, and their abundances were generated independently of each other. Multi-pass sequencing of the transcripts was simulated via pbsim3 with a total of 10 passes and an error rate of 10% per run (Ono et al. 2022). The multi-pass results were ran through the biopython ccs package to generate HiFi reads with a mean gap-compressed identity of 99.82 (Cock et al. 2009; De Coster and Rademakers 2023). We subsampled the HiFi reads to create datasets with varying percentages of novel reads. Each dataset contained all of the novel isoform reads, and a random subset of the annotated reads.

KuPID was also validated on a set of Oxford Nanopore MinION sequenced reads. We used Prjlbelski et al.'s sequencing data of Lexogen spike-in RNA variant (SIRV) synthetic molecules, provided in their benchmark of IsoQuant (Prjibelski et al. 2023). The dataset contained approximately 250,000 reads from SIRV-Set 4 (Iso Mix E0/ERCC/Long SIRVs), and were sequenced with the R10.4 platform. We used Lexogen's provided annotations of SIRV-Set4 for the reference, available at https://www.lexogen.com/sirvs/download/. Novel annotations were generated using the reduction method. Unlike with the simulated data, we did not subsample the SIRV reads to vary the percentage of reads that were novel. Instead, we varied the percentage of the reference transcriptome that was reduced, creating 4 replicates of each.

The SIRV reads were trimmed of their polyA tails prior to alignment and application of KuPID. A sample of 50 SIRV sequences were aligned to each other to identify common motifs in the 5' ends (The UniProt Consortium 2025). fastp was used to confirm that the motifs was present in the majority of the dataset's

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

reads, and to trim the likely adapter sequences (Chen et al. 2018). Additional clarification on our procedure can be found in Supplementary Note S1.

Parameter Selection To evaluate the PacBio HiFi reads, we used 22-bp k-mer s and assumed an error rate of .002. For the more erroneous ONT reads, we used 14-bp k-mer s and an error rate of .05. Our sketches used $\frac{1}{\gamma} = .1$ of all k-mer s in the reference transcriptome and RNA-seq reads. By default, we set n = 30 as the expected minimum exon length, and d = 130 as the expected median exon length. In addition, we set B = .98 as the expected maximum score of a novel read and b = 16 as the maximum amount of unique isoforms we'd allow a k-mer to map to. For 'quantify' mode, we set $l = 5$ as the default amount of reads to sample from each annotated isoform. We found that $m = 2$ works well as the band width for our chaining procedure, as it was rare to have erroneous anchor matches. In general, the bandwidth should be greater than the expected number of contiguous erroneous k-mer s in a query sketch that match to the reference. The probability that a query k-mer is erroneous and still matches to the reference is equivalent to $ke \cdot \frac{r_{bp}}{4^k}$, where r_{bp} is the total number of base pairs in the reference transcriptome. Thus, a good rule of thumb is to use $m = 1 + \frac{1}{1-ke \cdot \frac{r_{bp}}{4^k}}$. We settled on using a scale factor of $c = 1.5$ when selecting potential novel reads. End-users can increase c to improve the method's sensitivity to isoforms with ATSS. However, overly high values of c can hinder the precision of discovery results by allowing additional annotated reads into the sample.

Pipeline Overview

We compared the performance of isoform discovery and quantification methods using the original RNA-seq sample or the KuPID-processed reads. All experiments were run on an Intel(R) Xeon(R) Gold 5416S processor with 3 threads. The simulated read sets were first aligned to release 114 of the Ensembl human genome with the minimap2 aligner (Harrison et al. 2024; Li 2018). The real ONT reads were aligned to Lexogen's SIRV-Set4 genome. The alignment output was submitted to three ID pipelines: IsoQuant, flair, and StringTie2 (Kovaka et al. 2019; Prjibelski et al. 2023; Tang et al. 2020). Each method employed a reference annotation in guided mode to aid their discovery process. We used release 48 of the GENCODE human database as the the reference transcriptome for the simulated HiFi reads (Frankish et al. 2023). For the real ONT reads, we took the reference annotation of SIRV-Set4 and converted it into the transcript sequences using the 'transcribe' module of YASIM (Su et al. 2024). The ID pipelines assembled the alignment results into prospective models of the transcripts they believed to be present within the RNA-seq sample. The pipelines then outputted an annotation file of the exon features for each transcript model.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Data Access

All simulated reads and novel annotation have been submitted to the Zenodo (https://doi.org/10.5281/zenodo.20713444).

Code Availability

The source code for KuPID and additional scripts are available at GitHub `https://github.com/mboro2000/KuPID.git` and as Supplemental Code.

Competing Interests Statement

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments

This research was funded in part by the US National Science Foundation BIO/DBI grant 2531433. We'd like to thank Aidan Zhang from the Yu Lab for his help in developing this manscript.

M.B. conceived the project, designed the algorithms, and implemented the software. Y.W.Y. supervised and contributed to the development of the methods. Both authors wrote and edited the manuscript.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

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. 10.1186/s13059-020-1935-5

Arzalluz-Luque Á, Conesa A. 2018. Single-cell RNA-seq for the study of isoforms—how is that possible? *Genome Biol* **19**: 110. 10.1186/s13059-018-1496-z

Borowiak, M. 2025. Figures created in BioRender. https://BioRender.com/4rjw8yr

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X et al. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**: 1146-53. 10.1038/nbt.1495

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890. 10.1093/bioinformatics/bty560

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. 10.1038/s41592-023-01908-w

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Freidberg I, Hamelryck T, Kauff F, Wilczynski B et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422-1423. 10.1093/bioinformatics/btp163

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data, *Bioinformatics* **39**. 10.1093/bioinformatics/btad311

Finotello F, Di Camillo B. 2015. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* **14**: 130-42. 10.1093/bfgp/elu035

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**: D942-D949. 10.1093/nar/gkac1071

Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Becker A, Bennett R, Berry A, Bhai J et al. 2024. Ensembl 2024. Nucleic Acids Res **52**: D891-D899. 10.1093/nar/gkad1049

Irber L, Brooks PT, Reiter T, Pierce-Ward NT, Hera MR, Koslicki D, Brown CT. 2022. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. bioRxiv doi: 10.1101/2022.01.11.475838

Jain C, Gibney D, Thankachan SV. 2022. Algorithms for colinear chaining with overlaps and gap costs. *J Comput Biol* **29**: 1237-1251. 10.1089/cmb.2022.0266

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. 10.1186/s13059-019-1910-1.

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Leparc GG, Mitra RD. 2007. A sensitive procedure to detect alternatively spliced mRNA in pooled-tissue samples. Nucleic Acids Res **35**: 146. 10.1093/nar/gkm989

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110. 10.1093/bioinformatics/btw152

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. 10.1093/bioinformatics/bty191

Monzó C, Frankish A, Conesa A. 2025. Notable challenges posed by long-read sequencing for the study of transcriptional diversity and genome annotation. *Genome Res* **35**: 583-592. 10.1101/gr.279865.124.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463. 10.1038/nature08909

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 132. 10.1186/s13059-016-0997-x

Ono Y, Hamada M, Asai K. 2022. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom Bioinform* **4**. 10.1093/nargab/lqac092

Pertea M, Pertea GM, Antonescu CM, Chang T, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295. 10.1038/nbt.3122

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. 10.1038/s41587-022-01565-y

Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Tilgner HU. 2023. Mouse SIRV and simulated data used in the IsoQuant publication (Version 3) [Data set]. Zenodo. 10.5281/zenodo.7611877

Reyes A, Huber W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46**: 582–592. 10.1093/nar/gkx1165

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genom Proteom Bioinform* **13**:278-89. 10.1016/j.gpb.2015.08.002.

Roberts M, Hayes W, Hunt BR, Mount SM, Yorke JA. 2004. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**: 3363–3369. 10.1093/bioinformatics/bth408

Shaw J, Yu YW. 2023. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat Methods* **20**: 1661–1665. 10.1038/s41592-023-02018-3

Singh RK, Cooper TA. 2012. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med*, **18**: 472–482. 10.1016/j.molmed.2012.06.006

Downloaded from genome.cshlp.org on November 2, 2026 . Published by Cold Spring Harbor Laboratory Press

Su Y, Yu Z, Jin S, Ai Z, Yuan R, Chen X, Xue Z, Guo Y, Chen D, Liang H et al. 2024. Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. *Nat Commun* **15**: 3972. 10.1038/s41467-024-48117-3

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. 10.1038/s41467-020-15171-6

The UniProt Consortium. 2025. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**: D609–D617. 10.1093/nar/gkae1010

Wang Y, Liu J, Huang B, Xu YM, Li J, Huang LF, Lin J, Zhang J, Min QH, Yang WM, et al. 2014. Mechanism of alternative splicing and its regulation. *Biomed Rep* **3**: 152-158. 10.3892/br.2014.407

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 doi: 10.1101/672931