Genome Biology

SOFTWARE

Open Access

# IFDlong: a model-based isoform and fusion detector for accurate annotation and quantification of long-read RNA-seq data

Wenjia Wang[1†], Jia-Jun Liu[2,3,4†], Yuzhen Li[5], Sungjin Ko[4,6], Ning Feng[7], Manling Zhang[7], Qingqi Lin[2,3,4], Mengying Xia[2,3,4], Yan P. Yu[4,6], Jian-Hua Luo[4,6], Pedro L. Baldoni[1], George C. Tseng[1,8] and Silvia Liu[2,3,4,8,9*]

† Wenjia Wang and Jia-Jun Liu contributed equally to the work as the first authors.

*Correspondence: shl96@pitt.edu

[1] Department of Biostatistics and Health Data Science, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
[2] Organ Pathobiology and Therapeutics Institute, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[3] Department of Pharmacology and Chemical Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[4] Pittsburgh Liver Research Center, University of Pittsburgh, Pittsburgh, PA, USA
[5] Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[6] Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[7] Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[8] Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[9] Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

## Abstract

Long-read transcriptome sequencing (long-RNA-seq) revolutionizes transcriptome research by enabling full-length transcript analysis for comprehensive exploration of isoform diversity. We developed IFDlong, a probabilistic framework and software suite for detecting isoform and fusion transcripts from bulk or single-cell long-RNA-seq data. IFDlong annotates each long read, identifies novel isoforms, quantifies expression via an expectation-maximization algorithm, and profiles fusion transcripts. In large-scale simulation and real data analyses, IFDlong outperforms existing tools and demonstrated high accuracy and robustness across multiple in-house and public datasets, including healthy tissues, cell lines, and different diseases.

**Keywords:** Long-read RNA-seq, Isoform, Alternative splicing, Fusion transcript, Single-cell sequencing

## Background

Long-read technology has brought high-throughput genomic sequencing into the third generation to study genomes, transcriptomes, and metagenomes. Long-read sequencing is able to generate reads up to several million base pairs, while short-read sequencing typically yields reads with only 50 to 300 base pairs. For transcriptome sequencing, long-read RNA-seq (long-RNA-seq) can successfully reveal the full-length transcripts at an unprecedented resolution. Compared with short-read RNA-seq, long-RNA-seq presents additional advantages. First, long-RNA-seq can reduce the noise caused by artificial amplification and can dramatically increase the alignment certainty. Second, by covering the full-length (or almost full-length) of the transcript sequence, long reads not only enable the precise identification and quantification of known isoforms, but also show the power to discover novel isoforms. Next, long reads present the advantages of detecting transcriptomic structural variants by precisely locating the exon donor-acceptor

merging points for alternative splicing analyses or covering the fusion junction points for fusion transcript identification. Furthermore, most protocols for single-cell RNA-seq only sequence the 3' or 5' end of the transcript, while single-cell long-read RNA-seq allows isoform-level analyses at single-cell resolution for the first time. Several platforms have facilitated this cutting-edge technique. Two platforms have been widely applied for long-read sequencing: the Pacific Biosciences (PacBio) Iso-Seq [1] and the Oxford Nanopore Technology (ONT) whole-transcriptome sequencing [2, 3]. Other than these, several synthetic long-read strategies have been developed as well, including 10X Genomics Linked-reads [4] and Element Biosciences LoopSeq [5, 6].

One of the major applications for long-RNA-seq is to analyze isoforms and transcriptomic structural variants (TSVs). TSV refers to the transcriptome sequence alterations that are caused by genomic structural variants (SVs) or transcriptomic splicing. It usually describes large variants with more than thousands of base pairs. Other than genome-level analysis, transcriptome-level variant study can add flexibility to describe alternative splicing, differences in expression levels, and potential functional alteration in protein expression. These variant events will potentially role as biomarkers for disease diagnoses and serve as therapeutic targets. Two major TSVs will be explored in this study: alternative splicing and fusion transcripts. Alternative splicing (or pre-mRNA splicing) refers to the exon splicing events on pre-mRNA, where exons can be alternatively included or removed when forming mature mRNA [7, 8]. Given the fact that over- or under-expressed isoforms (or proteins) and novel isoforms will influence the regulation system, the deregulation of alternative splicing events is a hallmark of cancer. Thus, disease-associated isoform-specific events can serve as biomarkers for disease diagnosis and treatment [9–11]. For isoform-level analyses, short-read sequencing is limited by read length and is therefore insufficient to capture full-length transcripts. As a result, isoform abundance must be inferred statistically, and the discovery of novel isoforms is challenging. In contrast, long-read sequencing offers unique advantages in accurately identifying transcript isoforms and capturing splice junctions through full-length transcript coverage. However, in addition to complete long reads, truncated reads are frequently generated, which can introduce uncertainty in isoform assignment (Fig. 1A). As such, statistical estimation for isoform expression by long-RNA-seq is still urgently needed. Several bioinformatics tools have been developed to investigate isoforms from long-RNA-seq data. LIQA [12], Bambu [13] and miniQuant [14] employ model-based algorithms for isoform quantification, and IsoQuant [15] uses a graph-based method. Mandalorion [16], IsoQuant, and miniQuant can quantify isoform expression, but are not able to annotate individual long reads with their isoform of origin, thereby lacking information on isoform-supporting long reads for downstream analysis. In addition, TALON [17], Bambu, and IsoQuant can discover novel isoforms, while LIQA, Mandalorion, and miniQuant only quantify the known ones. FLAIR [18, 19] reports the annotation per aligned piece, but not at the individual long-read level. Major feature comparisons were summarized in Additional file 1: Table S1.

Fusion transcripts (Fig. 1B) are caused by fusion genes or trans-splicing events resulting from chromosomal rearrangements or ligation of two different primary RNA transcripts. These variants play important roles in tumorigenesis by potentially shifting the tail protein's open reading frame (ORF), forming novel chimera proteins,

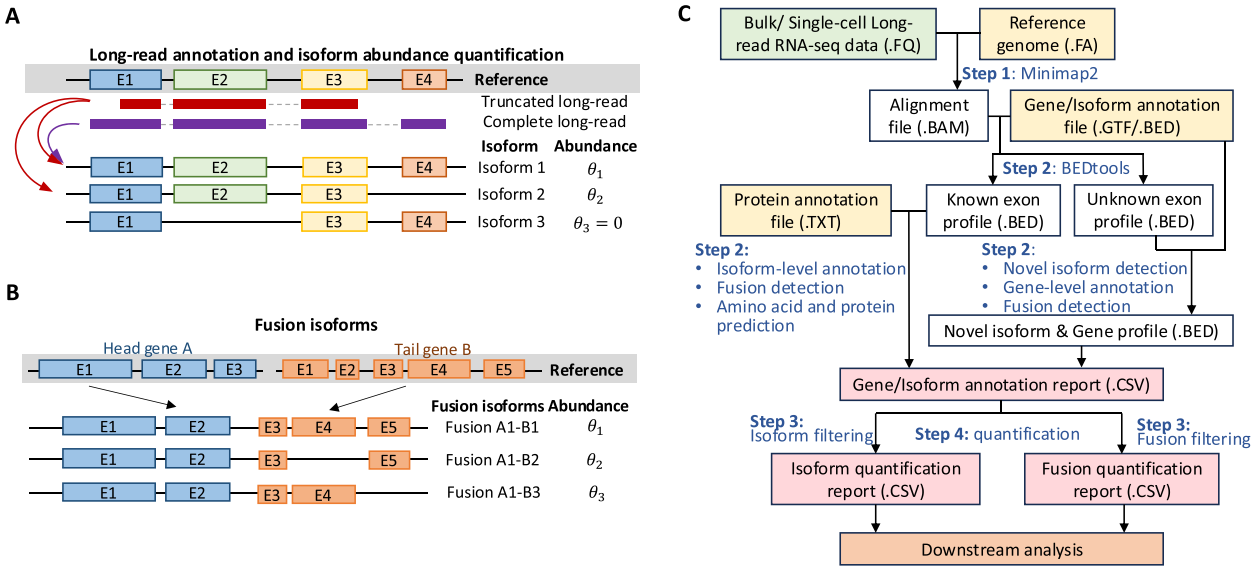



**Fig. 1** Illustration of alternative splicing, fusion isoform, and the IFDlong pipeline. **A** Long-read annotation and isoform abundance quantification. The red and purple reads represent the truncated and completed long reads, serving as the supporting reads for different isoforms. **B** Illustration of fusion isoforms. Head gene A and tail gene B can form multiple fusion isoforms by different fusion breakpoints. **C** The IFDlong pipeline. The proposed pipeline will take in the raw long-read sequences and annotation files, and ultimately output the gene/isoform annotation reports, isoform quantification reports, and fusion quantification reports for downstream analysis

or altering gene expression [20–23]. Previous studies have proved that fusions are highly associated with the occurrence and recurrence of cancers [24–26]. Suicide-gene insertion on the fusion gene can serve as a genotype-specific cancer therapy [27, 28]. In addition to the DNA-level of fusion gene study, fusion transcript analysis can reveal multiple alternative splicing events [29]. As illustrated in Fig. 1B, head gene A can fuse with three different isoforms of tail gene B. Short-read RNA-seq has been largely applied to discover fusions [30]. However, it has limitations in either performing isoform resolution analysis or requiring extremely deep sequencing (e.g., 1300x sequencing depth to detect rare fusion events in our previous study [31]). To overcome this, long reads possess an advantage over short reads in capturing the fusion isoform and precisely locating the fusion junction point [32]. Currently, bioinformatics tools, such as JAFFAL [33], Genion [34] and FLAIR-fusion [35], have been developed to call fusion transcripts from long-RNA-seq. These tools, however, can only detect fusion events at the gene level, not the isoform resolution. Additionally, these tools do not explicitly perform fusion isoform quantification to estimate the expression level of both normal and fusion isoforms (refer to Additional file 1: Table S1 for main feature comparisons).

In this article, we proposed the new bioinformatics pipeline *IFDlong*, an Isoform and Fusion Detector that was tailored for long-RNA-seq data for the annotation and quantification of isoform and fusion transcripts. As shown in Fig. 1C, the pipeline can process either bulk or single-cell long-RNA-seq data. Based on the reference genome and gene/isoform annotation profile, the long reads were aligned to the reference genome and were annotated to genes and isoforms. When a long read cannot be assigned to any known gene or isoform, it is classified as a novel isoform. The pipeline then applies model-based methods to generate isoform- and fusion-level quantification profiles as the final outputs. As shown in Additional file 1: Table S1, compared with the existing tools, IFDlong presents four major advantages: (1) IFDlong is compatible with both bulk and single-cell long-RNA-seq data; (2) IFDlong is able to perform read-level gene/isoform annotation and detect novel isoforms; (3) IFDlong performs model-based statistical estimation on the truncated long-reads with multiple alignment for highly accurate isoform and fusion quantification; (4) To our knowledge, IFDlong is currently the only tool that can discover fusion transcripts at isoform resolution. To test the performance of the proposed pipeline, we first compared IFDlong with several existing tools on *in-silico* data generated by a long-read simulator with different sequencing parameters. Next, multiple published long-read datasets were used to demonstrate that our tool is compatible with different long-read platforms and to show its ability to accurately profile alternative splicing and fusion events. In addition, different benchmarking metrics based on the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) [36] were adopted to enable comprehensive evaluation and comparison of our and several published tools. Our extensive simulation and real-data applications further offer a systematic assessment of existing tools in isoform and fusion transcript detection. In summary, IFDlong demonstrated improved performance in isoform quantification, novel isoform detection, and fusion isoform quantification. These results suggest

that IFDlong provides a user-friendly and integrated framework for long-RNA-seq data analysis.

## Results

### IFDlong development

We developed the software IFDlong, an isoform and fusion detector tailored for long-RNA-seq data. The software consists of four major steps, as shown in Fig. 1C.

#### *(Step 1) Long read alignment and filtering*

IFDlong first takes in long-RNA-seq data and aligns long reads to a reference genome (e.g., GRCh38 for human and GRCm38 for mouse) using the noise-tolerant long-read aligner minimap2 [37]. Unmapped long reads and reads with multiple alignments are removed. Long reads arising from paralogous sequences with secondary alignment to their pseudogenes or gene families are also filtered out. This pre-processing step is consistent with the existing reference-based tools.

#### *(Step 2) Gene, isoform, and amino acid (AA) annotation*

Long reads are intersected with gene/isoform annotation profile (described in the "Methods" section) using BEDTools [38]. Reads that consecutively overlap with known exons will be further annotated to known isoforms. While reads that cannot be fully covered by known exons will be defined as novel isoforms. These reads will be further assigned to known or novel genes through gene-level annotation. Ultimately, all the long reads will be annotated into three categories: known genes with known isoforms, known genes with novel isoforms, or novel genes. Note that, a long read will be either uniquely assigned to one isoform or uncertainly annotated to multiple isoforms (Fig. 1A). As a special case, a long read aligned to more than one gene (flagged by primary and supplementary alignment) is treated as supporting evidence for a fusion transcript. Based on the transcript level annotation, IFDlong predicts the corresponding amino acid sequence using the AA annotation profile (described in the "Methods" section). This precise annotation step enables subsequent isoform and fusion quantification.

#### *(Step 3) Fusion transcript and novel isoform filtering*

False positives in fusion detection may be introduced due to the nature of long-read sequencing (such as sequencing errors or artifacts) and transcriptome complexity (such as gene similarity and complex regulation mechanisms). To control the false positives, IFDlong innovatively applied the following filtering criteria. (1) Anchor length is defined as the number of base pairs that a supporting read is aligned to each fusion gene [30]. Short anchor length may result from misalignment, sequencing error, or genome similarity. A long read is defined as a fusion supporting read by IFDlong only if it has a minimum of 10 bp (by default) of anchor length to each fusion gene. (2) Fusion transcripts involving pseudo-genes are likely to be false positives. Long reads will be annotated based on the primary alignment rather than the fusion of a gene with its pseudo-genes. (3) Fusion candidates involving the same family

genes will be filtered out, because they often arise from multiple alignments driven by the transcriptome similarity. To distinguish novel isoforms from the known ones, a 'buffer length' is defined as the number of base pairs that differ from the known ones. Novel isoforms with buffer lengths below a specific threshold will be regarded as likely false positives and thus be filtered out. Finally, a read- level report file is generated to summarize the gene, isoform, amino acid, and fusion annotation. Users may optionally apply additional filters to remove read-through gene pairs or fusions between genes separated by short genomic distances. This filtering step substantially reduces false positives in IFDlong.

#### *(Step 4) Estimation of isoform and fusion transcript expression*

Based on the gene/isoform annotation report, IFDlong performs model-based estimation of the isoform and fusion transcript expression. Per gene or fusion gene group, IFDlong estimates the relative abundances of isoforms and fusion transcripts using an Expectation-Maximization (EM) algorithm to explicitly resolve ambiguously assigned long reads, such as truncated long reads in Fig 1A**,** or fusion genes involving multiple isoforms in Fig 1B. In the expectation step (E-step), read-level annotation information is used to calculate the likelihood of isoform/fusion expression. In the maximization step (M-step), relative abundances are updated to maximize the likelihood function. These two steps are performed iteratively until convergence. Full details are provided in the "Methods". Finally, IFDlong outputs reports containing isoform- and fusion-level quantity estimates. This model-based framework incorporates both complete and truncated long reads, thereby improving the accuracy of the transcript quantification.

### Isoform annotation and quantification in simulation data

Multiple simulation datasets were generated to evaluate the pipeline performance in terms of isoform annotation and quantification. To mimic the real sequencing data, expression intensities of 28157 isoforms were first quantified from the Universal Human Reference (UHR) sample [12] (Additional file 2: Fig. S1A). Then, based on this distribution, the Type I-1 simulation datasets were generated by tool PBSIM2 [39] with different accuracy settings (high-accuracy, PacBio and ONT), gradient number of long reads (around 0.5, 1, 2, and 4-million reads, see Additional file 2: Fig. S1B), and gradient mean read length (200 bp, 500 bp, 800 bp, 1000 bp, 1200 bp, 1500 bp, 1800 bp, and 2000 bp, see Additional file 2: Fig. S2). Simulation details are described in the "Methods" section.

As a representative of the Type I-1 simulation, Fig. 2A shows the performance comparison in the simulation data with high-accuracy sequencing, 0.5 million reads, and 1500 bp mean read length. The proposed IFDlong pipeline and seven existing isoform quantification tools (LIQA [12], TALON [17], Mandalorion [16], Bambu [13], FLAIR [18, 19], miniQuant [14] and IsoQuant [15]) were applied to the simulation data and compared with the true isoform distribution. The distribution histograms of isoform intensities are shown in the diagonal blocks of Fig. 2A, and all their pairwise distribution scatter plots and Spearman's correlations are presented in the bottom and upper blocks, respectively. Among all the tools, the IFDlong achieves the highest correlation with the truth (Spearman's correlation = 0.82), followed by miniQuant (0.79), Bambu (0.75) and IsoQuant (0.7), while the other tools in

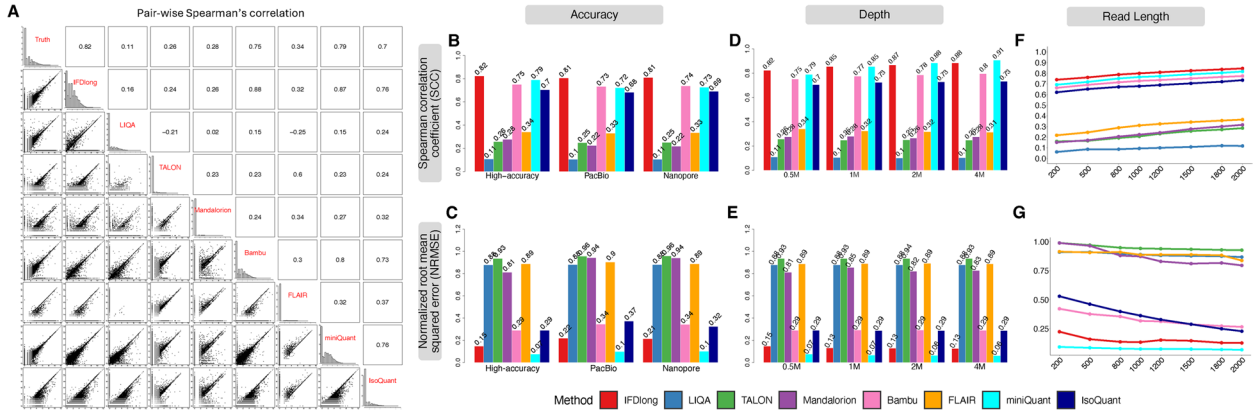



**Fig. 2** Isoform annotation and quantification in the Type I-1 simulation analysis. **A** Pairwise comparison of the isoform quantification tools with the setting of accurate sequencing, 0.5M reads, and 1500 bp mean length. The bottom left cells present the pairwise scatter plot of isoform expression. The upper right cells indicate Spearman's correlation coefficient (SCC). **B**, **C** SCC and normalized root mean squared error (NRMSE) of isoform expression between the truth and each tool under different sequencing accuracies with 0.5M reads and 1500 bp mean length. **D**, **E** SCC and NRMSE of isoform expression between the truth and each tool under different sequencing depths with high-accuracy sequencing and 1500 bp mean length. **F**, **G** SCC and NRMSE of isoform expression between the truth and each tool under gradient mean read lengths with 0.5M reads and high-accuracy sequencing. The bars are colored by different isoform quantification methods: IFDlong, LIQA, TALON, Mandalorion, Bambu, FLAIR, miniQuant, and IsoQuant

general under-estimate the isoform expression (scatter plot under the diagonal when compared with the truth) or are unable to capture a subset of the known isoforms (multiple zero-expressed isoforms in the scatter plot). The same conclusion holds for the simulation datasets with gradient parameter settings. Fig. 2B, C, and Additional file 2: Fig. S3 illustrate Spearman correlation coefficient (SCC), normalized root mean squared error (NRMSE), Pearson correlation coefficient (PCC), and median of the relative difference (MRD) in the three sequencing accuracy settings. The IFDlong and miniQuant overall reach the highest performance, followed by Bambu and IsoQuant. In general, all the tools are robust across different platforms. The IFDlong shows slightly better performance in the high-accuracy setting compared with the ONT and PacBio settings. When checking the pipeline performance in terms of sequencing depth, Fig. 2D, E, and Additional file 2: Fig. S3 indicate that all the tools are robust to different sequencing coverage and can quantify isoform expression at relatively lower depths (such as 0.5M). The IFDlong presents slightly improved performance with the increasing number of reads. In addition, tool performances were evaluated in the simulations with a gradient of read lengths. As shown in Fig. 2F, G, and Additional file 2: Fig. S3, all the tools yield dramatically increasing accuracy as the sequencing reads extend longer, which strongly proves the advantage of long-read techniques compared with short-read sequencing. Specifically, the IFDlong achieves the highest SCC (0.74) with the truth in the simulation data with 200 bp mean length, and its performance further increases to 0.85 in 2000 bp length. Among all the settings, IFDlong, miniQuant, Bambu and IsoQuant performed the best and detected the largest true positives when compared with the truth. Take the high-accurate setting in Fig. 2A as an example, among the 28,157 true isoforms, IFDlong is able to detect 28,004 of them, followed by miniQuant, Bambu, IsoQuant and LIQA with 23,154, 21,028, 16,920 and 15,595 true positives, while TALON, FLAIR and Mandalorion only report 3,666, 1,707 and 1,761 true isoforms. If only evaluating the common isoforms between the truth and the detected ones, IFDlong, miniQuant, IsoQuant, LIQA, Bambu and FLAIR achieved the highest similarity with the truth for different simulation settings (Additional file 2: Fig. S4). Besides isoform-resolution analysis, the SCCs and pairwise scatter plots at the gene level are shown in Additional file 2: Fig. S5. They maintain a consistent trend as the isoform resolution analysis (Fig. 2A) that IFDlong achieves among the best.

Besides the accuracy, computational costs were benchmarked among multiple tools. We applied 1 core for all the tools on the Type I-1 simulation dataset. The computational cost for IFDlong includes all the steps: alignment, individual long-read annotation, fusion filtering, and isoform and fusion quantification. While the other reference-based isoform tools only contain alignment and isoform quantification steps. In general, all the tools consume longer computing time for larger sequencing coverage and longer read length, with some exceptions (Additional file 2: Fig. S6A and B). While the largest memory used is not a monotone function of the read length (Additional file 2: Fig. S6C), but all within an acceptable cost. The IFDlong algorithm requires a reasonable computing time and memory footprint compared with existing tools. For example, as shown in Additional file 2: Fig. S6B, IFDlong completes the analysis of a sample with a 2,000 bp read length

in 106 minutes while using 83.4 MB of peak memory, whereas other tools require 65–352 minutes and 18–232 MB. These results demonstrate that IFDlong achieves an effective balance between computational accuracy and efficiency.

### Distinguishing multiple alternative splicing events arising from the same gene in simulation data

In addition to gene-level quantification, one of the major advantages of the long-read technique is to identify multiple alternative splicing events and quantify their expression. As illustrated in Fig. 1A, a long read will be either uniquely aligned to one isoform or uncertainly assigned to multiple isoforms. To address this issue, the IFDlong pipeline developed a model-based algorithm to estimate isoform relative abundance (see the "Methods" for algorithm details). In the Type I-2 simulation, we generated long reads originating from 200 genes, where each gene expresses at least 8 different isoforms. The isoform tools (IFDlong, LIQA, TALON, Mandalorion, Bambu, FLAIR, miniQuant and IsoQuant) were applied to this simulation data to quantify the abundance of all the alternative splicing events. Per gene, the cosine similarity between the true isoform proportion and the tool-predicted proportion was calculated. As a result, Fig. 3A shows the distribution of the cosine similarity

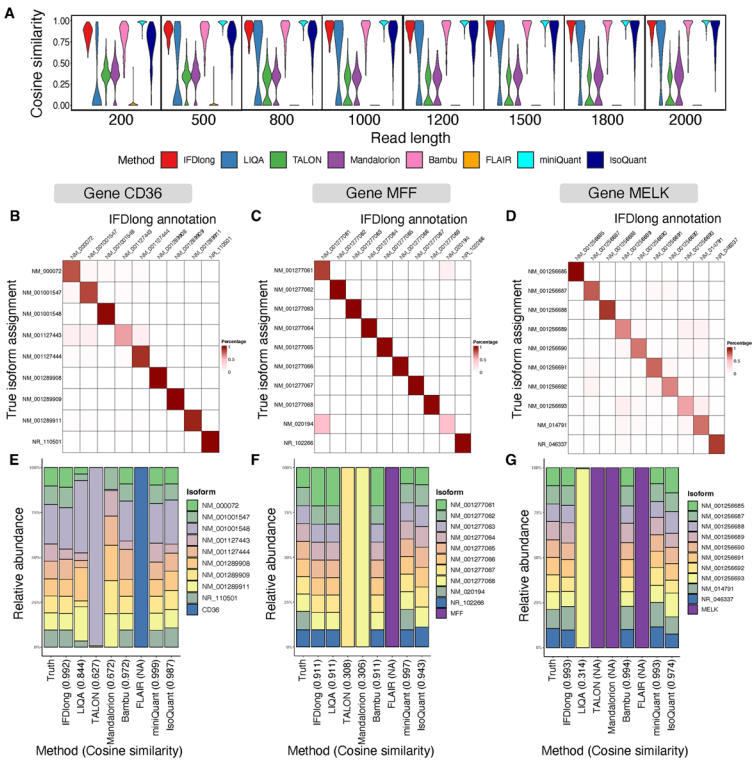



**Fig. 3** Distinguishing multiple alternative splicing events in the Type I-2 simulation analysis. **A** Cosine similarity of the isoform expression for 200 genes when compared with the truth across different sequencing lengths under high-accuracy sequencing. **B**-**D** The percentage contingency table between the true isoform assignments and the annotation by IFDlong for gene CD36, MFF and MELK. **E**-**G** Comparison of isoform relative abundance among the true simulations and the expressions estimated by different tools for gene CD36, MFF and MELK. The color bars indicate multiple isoforms arising from the same gene or gene-level quantification

of these 200 genes per tool along gradient read lengths. Overall, our IFDlong and miniQuant achieved the best similarity with the truth, followed by Bambu, IsoQuant and LIQA. As representatives, ten genes (CD36, MFF, MELK, ASPH, BLNK, CTCFL, ERG, THPO, TSACC, and YAP1) were simulated individually to evaluate the tool performance. Figure 3B-D and Additional file 2: Fig. S7 illustrate the contingency table (in percentage) between the truth and the IFDlong annotated isoform assignments for these genes, where larger percentages in the diagonal block of the heatmap indicate higher consistency. In terms of abundance evaluation, we compared the relative abundance (summing up to be 100%) among the truth and the estimates by the eight tools in Fig. 3E-G. The IFDlong demonstrated high consistency with the truth (cosine similarity to be 0.992, 0.911 and 0.993 for gene CD36, MFF and MELK, respectively), whereas miniQuant, Bambu, and IsoQuant presented equivalent similarities with IFDlong. LIQA and Mandalorion overall performed well but failed for some genes (Fig. 3F and Additional file 2: S7). TALON assigned all reads to one isoform, and FLAIR only detected genes but failed to call the exact isoforms. Similar to the Type I-1 simulation setting, we generated gradient read lengths in this Type

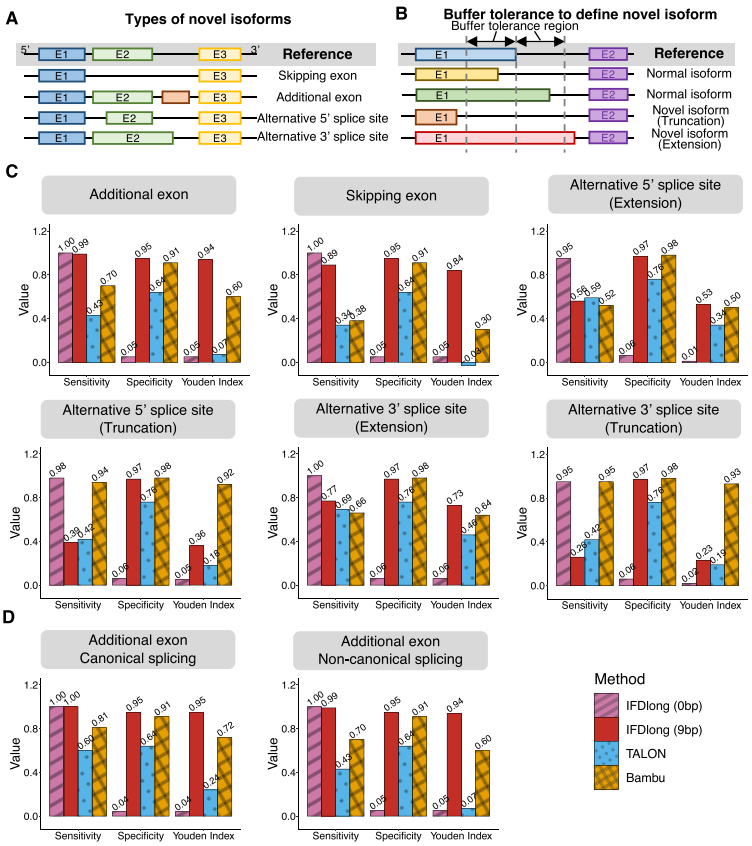



**Fig. 4** Illustration and detection of novel isoforms in the Type I-3 simulation analysis. **A** Major types of novel isoforms. **B** Buffer length tolerance to define a novel isoform. Compared with the reference exon (blue), if the difference is within the buffer tolerance region, the isoform will be defined as a normal one (yellow and green isoforms); otherwise, it will be described as a novel isoform with either truncation (orange) or extension (red). **C** Bar-graph for sensitivity, specificity and Youden index of IFDlong, TALON and Bambu when detecting novel isoforms from different types. The bars are colored by different settings for the buffer length (0 bp or 9 bp) of IFDlong and the other tools

I-2 simulation dataset. Take gene CD36, MFF and MELK as examples, Additional file 2: Figs. S8-10 illustrate the contingency heatmap of the IFDlong pipeline and the relative abundance bar-graph across different read lengths, where the results present increasing similarity when the read length extends longer. When calculating their cosine similarity with the truth, IFDlong presented robustly high performance for all the genes across different read lengths, while all the other tools failed for some settings (Additional file 1: Table S2).

#### Novel isoform detection in simulation data

In addition to quantifying known isoforms, IFDlong can call novel isoforms from the long-RNA-seq data. As typical examples shown in Fig. 4A, four types of novel isoforms were defined: skipping exon, additional exon, alternative 5' splice site (with truncation or extension), and alternative 3' splice site (with truncation or extension). Multiple factors will cause the alterations of the reads compared with the reference sequences, such as insertion, deletion, or mismatches caused by SNP, mutation and sequencing errors, or misalignment caused by multiple alignment and aligner performance. To avoid these false positives when calling novel isoforms, we proposed a 'buffer tolerance region' to describe the distance between the reference exon boundary and the detected one, as shown in Fig. 4B. The reference boundary was defined using the annotation GTF file, while the detected boundary was determined from the CIGAR information provided by minimap2 alignment. If the distance falls within the buffer length (yellow and green examples in Fig. 4B), IFDlong classifies it as a normal isoform. However, if the distance exceeds the buffer length, as illustrated in the orange and red examples in Fig. 4B, IFDlong annotates it as a novel isoform. The alternative 5' or 3' splice sites (Fig. 4A) with extension or truncation were defined using this buffer length. Choosing an optimal buffer length is therefore critical to balancing sensitivity and specificity in novel isoform discovery. Therefore, the Type I-3 datasets generated six libraries for each category of novel isoform. Each library contains half of the reads derived from normal isoforms and the other half generated from the novel isoforms in each category, as described in the "Methods". Note that, only IFDlong, TALON, and Bambu can annotate isoforms at the individual long-read level. Figure 4C shows the sensitivity, specificity, and Youden index for these three tools when detecting long reads supporting novel isoforms. Two buffer tolerance lengths were applied to IFDlong: 0 bp (exact edge-matching between the reference and detected exon) and 9 bp, while the buffer length for TALON and Bambu is not allowed to be customized. As shown in Fig. 4C, compared with zero buffer setting, IFDlong with 9 bps buffer region can significantly increase the specificity with reasonable sacrifice of sensitivity, thus achieving the highest Youden index when compared with TALON and Bambu. This specifically influences the performance of alternative 5' and 3' splice site novel types, because they are more sensitive when defining the cutoff for normal and novel. In this Type I-3 simulation, we generated datasets with mean read lengths of 1000 bp (Additional file 2: Fig. S11) and 2000 bp (Fig. 4C), where IFDlong achieved higher and more robust performance in the longer read-length datasets. For example, when detecting the novel type of additional exon, IFDlong with a 9 bp buffer region can reach 99% sensitivity and 95% specificity (Fig. 4C), while the 1000 bp can already reach a satisfying performance (e.g., with a 9 bp buffer, 58% sensitivity and 97%

specificity, Additional file 2: Fig. S11). In addition to the high-accuracy sequencing setting, simulation data with ONT and PacBio accuracy were generated, where a similar conclusion holds in these two platforms (Additional file 2: Fig. S11). Moreover, compared with the high-accuracy sequencing reads shown in Fig. 4C, a 9 bp buffer length provides greater benefit to IFDlong under high sequencing error rates, yielding a significant increase in specificity with negligible loss of sensitivity compared with a 0 bp buffer length (Additional file 2: Fig. S11). More systematic analyses were performed with gradient buffer length (0 to 10 bp) on all the novel isoform types and accuracy settings, where a similar conclusion is drawn (Additional file 2: Fig. S11). For the Type I-3 simulation data, a buffer length of 1 bp is the elbow point to significantly increase the Youden Index compared with 0 bp (Additional file 2: Fig. S11). The optimized selection of buffer length depends on the real application, where 3 to 10 bp is generally suggested to control the false positives without much loss of true positives.

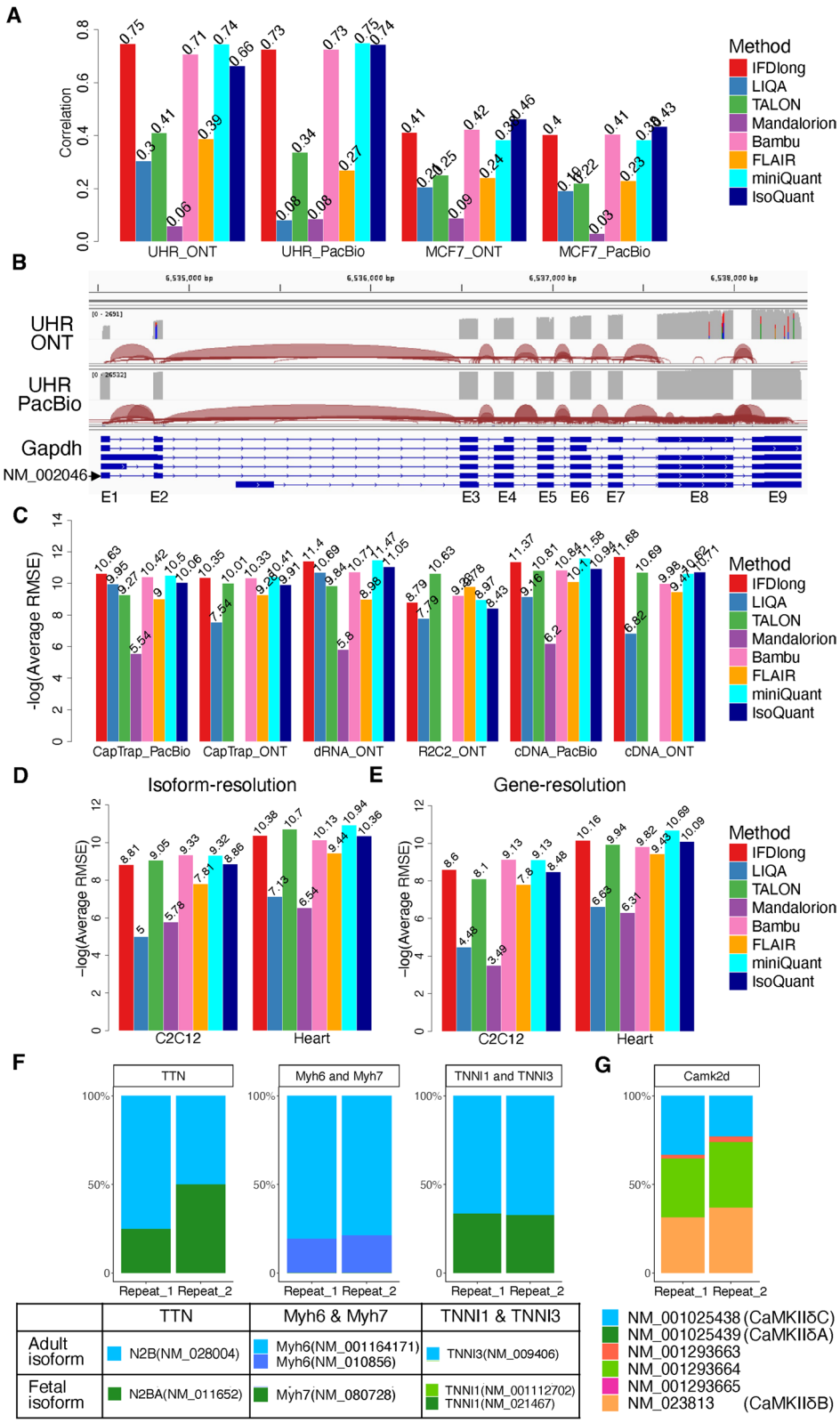
In addition to the above sequencing factors that influence the novel isoform detection, biological canonical splice sites are typically detected as conserved dinucleotides for alternative splicing events [40]. Given that both canonical and non-canonical splicing sites are important components [41, 42], our IFDlong can detect novel isoforms from both types. To check the tool performance on each type, Type I-3 data was split into two novel types: canonical sites defined by GT/GC and AG as donor and acceptor, and non-canonical sites otherwise. Taking additional exon as an example, IFDlong presents equivalent performance on both types (Fig. 4D), with a Youden index of 0.95 and 0.94 for canonical and non-canonical splicing types, respectively. Furthermore, IFDlong maintains equally robust high performances across different novel types, accuracy settings, and read lengths (Additional file 2: Fig. S12). Overall, compared with TALON and Bambu, our IFDlong pipeline can better balance sensitivity and specificity for most of the simulation scenarios (Figs. 4C, S11**,** and Additional file 2: Fig. S12).

#### Application in real human datasets for isoform detection and quantification

In addition to *in silico* data with complete underlying truths, we applied the proposed IFDlong pipeline and the seven existing isoform quantification tools to real sequencing data with partial experimentally measured truths. For human data application, RNA extractions from both Universal Human Reference (UHR) and MCF7 human breast cancer cell lines were sequenced by both the ONT and the PacBio platforms. Data were downloaded from public resources and fully described in the "Methods". As a result, Fig. 5A and Additional file 2: Fig. S13 illustrate the performance of the eight tools in

(See figure on next page.)

**Fig. 5** Application of isoform detection in real human and mouse studies. **A** Spearman's correlation of isoform quantification in human studies: universal human reference (UHR) and MCF7 breast cancer cell line sequenced by the ONT and PacBio platforms. The isoform quantification tools are represented in different colors. **B** Illustration of multiple isoforms of the gene Gapdh supported by long reads in the UHR dataset. **C** Spearman's correlation of isoform quantification on human H1-mix samples sequenced by different long-read platforms and library preparation strategies. **D**, **E** Consistency measurement (-log10(average RMSE)) of isoform and gene-level quantification in mouse studies: C2C12 cell line with four repeats and mouse heart tissues with two repeats. **F** Relative abundance of adult and fetal isoforms quantified by the IFDlong pipeline on mouse heart tissues. **G** Relative abundance of multiple alternative splicing events of Camk2d quantified by the IFDlong pipeline on mouse heart tissues




| | TTN | Myh6 & Myh7 | TNNI1 & TNNI3 |
|---|---|---|---|
| Adult isoform | N2B(NM_028004) | Myh6(NM_001164171)<br>Myh6(NM_010856) | TNNI3(NM_009406) |
| Fetal isoform | N2BA(NM_011652) | Myh7(NM_080728) | TNNI1(NM_001112702)<br>TNNI1(NM_021467) |

**Fig. 5** (See legend on previous page.)

terms of isoform-level SCC, MRD, NRMSE, and gene-level SCC. For UHR datasets, 964 isoforms quantified by TaqMan Real-Time PCR Assays serve as the truth [43]. At both isoform and gene resolution, IFDlong achieves the highest correlation with the RT-PCR results (isoform-level SCC to be 0.75 and 0.73 in the ONT and the PacBio datasets, Fig. 5A), indicating the great performance of our proposed tool. Among the 964 isoforms, the IFDlong successfully captures 677 (70%) and 812 (84%) of them in the ONT and the PacBio datasets, respectively. Beyond this PCR quantified isoform set, in total 16,578 and 28,078 isoforms are detected by the IFDlong pipeline in both datasets (Additional file 1: Table S3). These additional isoforms, however, cannot be evaluated given the limitation of the RT-PCR measurement set.

As an illustration, Fig. 5B takes the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (Gapdh) as an example to show how the long-reads can quantify the transcript expression at isoform resolution. Gapdh is a highly expressed gene in human samples. Among the multiple alternative splicing events, NM_002046 plays a dominant role in the UHR datasets as quantified by the IFDlong pipeline (Additional file 1: Table S3). As shown in the IGV splice junction and coverage plot, the long reads can successfully distinguish the differences in exons 1, 4, 6, and 7 among multiple isoforms, which were successfully captured by IFDlong (Additional file 1: Table S3).

A similar analysis was performed on MCF7 datasets, where transcript expressions quantified by short-read RNA-seq were used as the underlying truth. Admittedly, short-read data is not a golden standard for truth, our IFDlong pipeline still performed among the best (SCC of 0.41 and 0.40 for ONT and PacBio dataset in Fig. 5A and Additional file 2: Fig. S13). The full lists of detected isoforms by IFDlong are summarized in Additional file 1: Table S4.

Long-read measurement platforms and libraries are key factors for sequencing quality, and thus influence the performance of isoform quantification tools. We further evaluated IFDlong across multiple platforms. SGNex ONT sequencing was performed on MCF7 with multiple runs using cDNA, direct cDNA, direct RNA, and EV direct RNA (described in "Methods"). As shown in Additional file 2: Fig. S13D, IFDlong is compatible with all these platforms and IFDlong, Bambu, miniQuant and IsoQuant yields overall the best SCC with the truth. Moreover, LRGASP [36] has measured the H1-mix sample using both PacBio and ONT platform with different library preparation methods and repeats (described in "Methods"). Given no ground truth for this dataset, two consistency measures were employed: the minus log of average root mean squared error (RMSE) and the irreproducibility measure (IM), where high minus log(average RMSE) and low IM indicate high consistency among multiple repeats. As presented in Fig. 5C, Additional file 2: Fig. S14, and Additional file 1: Table S5, IFDlong can successfully achieve low variance among the three repeats of each library preparation strategy at both isoform and gene resolution. These results indicate that IFDlong is compatible with different long-read platforms and library preparation strategies and can in general present high performance in terms accuracy and consistency.

#### Application in real mouse datasets for isoform detection

In addition to human data, our proposed pipeline can be applied to other species if the reference genome and annotation files are provided as the input. In this application,

we collected long-read data on mouse cell cline C2C12 (with four repeats) and heart tissues (with two repeats) from the LRGASP/ENCODE project [44] to illustrate the generalizability of our proposed tool into broad applications. For both datasets, the ENCODE project employed the TALON pipeline to quantify the isoform expression. In addition, we applied IFDlong, LIQA, Mandalorion, Bambu, FLAIR, miniQuant and IsoQuant to these two mouse datasets for performance evaluation and application. Isoforms quantified by IFDlong were summarized in Additional file 1: Tables S6 and S7. Given no golden standard truth is available for these two datasets, we can only check the performance similarity among the tools. As shown in Additional file 2: Fig. S15, IFDlong presents high and moderate agreement with Bambu, miniQuant and IsoQuant in all the datasets, while both LIQA and Mandalorion generated few isoforms in general. In addition, the mouse datasets were applied to benchmark the consistency of isoform (or gene) quantification among multiple repeats. Per software, mean isoform (or gene) abundance was first calculated across all repeats, and then the average RMSE was derived based on this mean abundance. As shown in Fig. 5D and E, IFDlong, TALON (result provided by the ENCODE project [44]), Bambu, miniQuant, and IsoQuant achieved the largest minus log10(average RMSE) value at both isoform and gene resolution. Alternative benchmarking method IM is also applied and indicates the high robustness properties of IFDlong in both C2C12 and heart datasets (Additional file 2: Fig. S16).

In addition to robustness, isoform diversity and biological discoveries can evaluate the isoform quantification pipelines indirectly. From the two repeats of mouse heart tissues, IFDlong, TALON, Bambu, miniQuant and IsoQuant discovered 15972, 18794, 14092, 17129 and 12501 isoforms, while LIQA, Mandalorion and FLAIR only detected 293, 975 and 4777 isoforms. The sufficient estimation of the isoform diversity reveals biological insights. For example, the heart tissues were collected from mice at 14 days, which is the transition time of fetal isoforms to adult ones. The isoform quantification by IFDlong (Additional file 1: Table S7) has strongly supported this conclusion. For instance, during the perinatal period, there is a transition in titin (TTN) isoforms from the compliant fetal titin N2BA (NM_011652) to the stiffer N2B (NM_028004) adult isoform in the heart, adapting to postnatal cardiac load demands [45]. Altered splicing of TTN is implicated in heart failure, and lowering the N2BA to N2B ratio has been proposed as a therapeutic strategy for heart failure [46–48]. The expression levels of N2BA and N2B that have been quantified by IFDlong present this dynamic transition status (Fig. 5F). Besides TTN, the switching of several myofibril proteins during the postnatal period was observed. For example, the fetal isoform of myosin heavy chain 7 (Myh7, NM_080728) is gradually switching to adult isoform Myh6 (NM_001164171 and NM_010856) during the perinatal period [49]. The slow skeletal TNNI (TNNI1) is in the process of transferring to the cardiac-type troponin (TNNI3) [50]. As shown in Fig. 5F, the IFDlong successfully captured isoform diversity during this transition process, where more adult isoforms were observed than their corresponding fetal isoforms. However, the other isoform quantification tools cannot fully capture this process and don't show high agreement with each other (Additional file 2: Fig. S17).

In addition, IFDlong also reveals the distribution of multiple isoforms originating from the same gene in mouse heart data. For instance, calcium/calmodulin-dependent

protein kinase II delta (CaMKIIδ) plays a central role in a variety of cardiac diseases [51], which presents several alternative splicing forms as shown in Fig. 5G. CaMKIIδB (NM_023813) is a dominant isoform containing exon 14, which regulates gene transcription and may play a protective role in cardiac diseases [52]. CaMKIIδC (NM_001025438) is another dominant isoform lacking exons 14–16, which is the major contributor to the pathological process of cardiac diseases [53]. CaMKIIδA (NM_001025439) comprises exons 13 and 15–17 and regulates the L-type Ca2+ channel and contributes to calcium mishandling in heart failure [54]. IFDlong can successfully capture this isoform diversity (Fig. 5G), while LIQA, Mandalorion and FLAIR failed (Additional file 2: Fig. S18). Besides CaMKIIδ, as shown in Additional file 2: Fig. S19, the IFDlong analysis has revealed the isoform distribution of many other cardiac disease-related genes, such as Calcium Voltage-Gated Channel Subunit Alpha1 C (Cacna1c), Troponin T2 (Tnnt2), Calcium/Calmodulin Dependent Protein Kinase II Gamma (Camk2g) and Myocyte Enhancer Factor 2C (Mef2c). All these have indirectly proved the additional resolutions revealed by IFDlong for isoform-level analysis when compared with conventional gene-level quantification or the other existing long-read tools.

#### Detecting differentially expressed isoforms in single-cell long-read RNA-seq data

The IFDlong pipeline is not only compatible with the bulk long-RNA-seq, but also can be applied to the single-cell long-RNA-seq data to investigate the isoform expression at a single-cell resolution. A paired tumor-normal liver samples from a hepatocellular carcinoma (HCC) patient sequenced by the LoopSeq platform were used in this project (details described in the "Methods") [55]. By IFDlong pipeline, in total 32,124 isoforms originating from 14,279 genes across 162 cells from the benign libraries and 285 cells from the tumor libraries were detected. We innovatively performed single-cell clustering and differential analysis [56] based on the isoform profile (Fig. 6A) in addition to the traditional gene-based analysis (Fig. 6B). As shown in the uniform manifold approximation and projection (UMAP) dimension reduction plot, cells from tumor and normal libraries are well separable from each other. Due to the cell heterogeneity, some cells extracted from the tumor library present benign expression patterns. Cell clustering also reflects the heterogeneity of the tumor cells. Furthermore, downstream analysis can be performed on the isoform-based profile. For example, trajectory inference aims to estimate the pseudotime of cell development. In this HCC application, isoform-based trajectory inference (Fig. 6C) highlights the two tumorigenesis pathways from the benign toward tumor cells when compared with the gene-based analysis (Fig. 6D). More in-depth analysis will be investigated in the future. Next, differential expression analysis was performed comparing tumor and normal cells to define differentially expressed isoforms (Additional file 1: Table S8). Isoforms were further mapped to the genes for downstream gene set enrichment analysis on the KEGG database (Fig. 6E and Additional file 1: Table S9A) and GO database (Additional file 2: Fig. S20A and Additional file 1: Table S9B). Top differential isoforms were subject to pathway enrichment test by Ingenuity Pathway Analysis (IPA on up-regulated isoforms, Additional file 2: Fig. S20B and Additional file 1: Table S9C). Top pathways are primarily associated with the tumor immune microenvironment and immune evasion processes. These pathways play a crucial role in HCC tumorigenesis and the response to immune checkpoint inhibitor

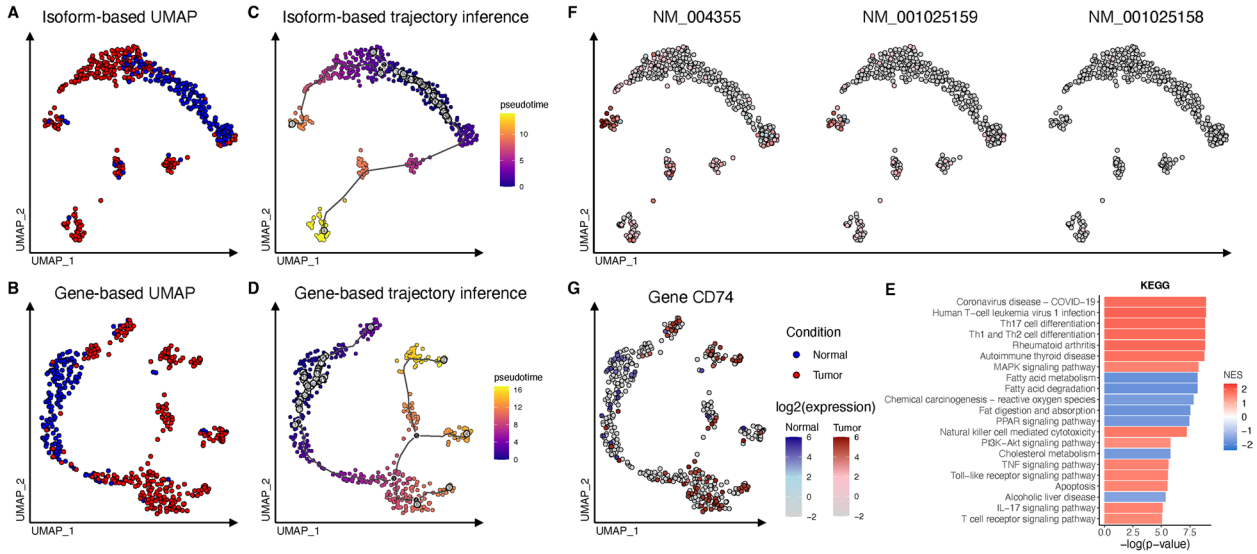



**Fig. 6** Application of IFDlong pipeline into single-cell long-read RNA-seq data on human HCC. **A**, **B** UMAP visualization on the single-cell data based on isoform or gene expression profile. Each dot indicates a single cell that is colored by blue (for normal) or red (for tumor). **C**, **D** Trajectory inference of tumorigenesis based on isoform or gene expression profile. Each dot indicates a single cell that is colored by the estimated pseudo-time. **E** Top KEGG pathways enriched by the differentially expressed isoforms comparing tumor and normal cells. **F**, **G** Multiple alternative splicing events of gene CD74 present differential expression patterns between normal and tumor cells (**F**) Expression levels of three isoforms from gene CD74: NM_004355, NM_001025159, and NM_001025158; (**G**) Expression levels of gene CD74

(ICI) therapies. Particularly, the top pathways encompass various aspects of immune processes, including heterogeneity within the tumor microenvironment, immune cell infiltration, immune evasion mechanisms employed by tumor cells, and potential targets for immunotherapy (Fig. 6E).

For instance, considering the human leukocyte antigen (HLA) [57–59] and B2M [60, 61] genes, which encode proteins crucial for the antigen presentation process critical for activating T cells and initiating an immune response against tumor cells. Dysregulation or alterations in the expression of these HLA genes can impact the immune system's ability to recognize and eliminate cancer cells, directly influencing the success of ICIs [58, 60]. In addition, validated HCC immune players, CD74 [62, 63] (Fig. 6F and G), CCL5 [64, 65] (Additional file 2: Fig. S21A), and IL7R [66, 67] (Additional file 2: Fig. S21B), known for their involvement in HCC ICI response, have been identified as top isoforms in pathway analysis. Their isoforms exhibit a significant increase in tumor cells compared to non-tumor regions, supporting the validated role of these isoforms in HCC pathogenesis, which underscores the necessity for understanding the mechanisms of ICI resistance. Notably, CD74 contains three isoforms (Fig. 6F, NM_004355, NM_001025159, and NM_001025158). Compared with traditional short-read analysis that can only quantify gene-level expression (Fig. 6G), IFDlong reveals that isoforms NM_004355 and NM_001025159 present similar expression patterns with higher expression in tumor cells, while NM_001025158 shows overall low expression. CD74 plays multifaceted roles in HCC pathogenesis by directly influencing antigen presentation through regulating major histocompatibility complex (MHC) class II molecules, influencing immune modulation, and tumor cell behavior [62, 63]. Furthermore, several isoforms of CCL5 (Additional file 2: Fig. S21A, dominant by isoform NM_002985) and IL7R (Additional file 2: Fig. S21B, with both NR_120485 and NM_002185 highly expressed) have been reported, resulting from alternative splicing or post-translational modifications [68–70], may exert distinct biological functions in HCC, beyond their major roles as chemo-attractants and cognate receptors for various immune cell types, including T cells, monocytes, macrophages, and dendritic cells. Isoform diversity contributes to establishing an immunosuppressive microenvironment within the tumor, thereby facilitating tumor progression and immune evasion.

#### Fusion isoform detection and quantification in simulation data

IFDlong is designed to detect and quantify fusion transcript at isoform resolution. For performance evaluation, the Type F-1 dataset was generated by pooling reads simulated from 1000 normal isoforms and 1000 fusion transcripts. The proposed IFDlong pipeline and three existing fusion quantification tools (JAFFAL [33], Genion [34] and FLAIR-fusion [35]) were applied to the simulation data and compared with the true fusion distribution. As a representative, Fig. 7A shows the performance comparison in the data with approximately 0.5 million high-accuracy sequencing reads. Considering that the existing tools can only detect fusion transcripts at gene resolution, Fig. 7A illustrates the gene-level fusion expression in the diagonal blocks, as well as their pairwise expression scatter plots and SCC in the bottom and upper blocks. Among these tools, IFDlong achieved the highest correlation with the truth (SCC=0.83), followed by JAFFAL (SCC=0.66). The same conclusion holds for the simulation datasets with the ONT and

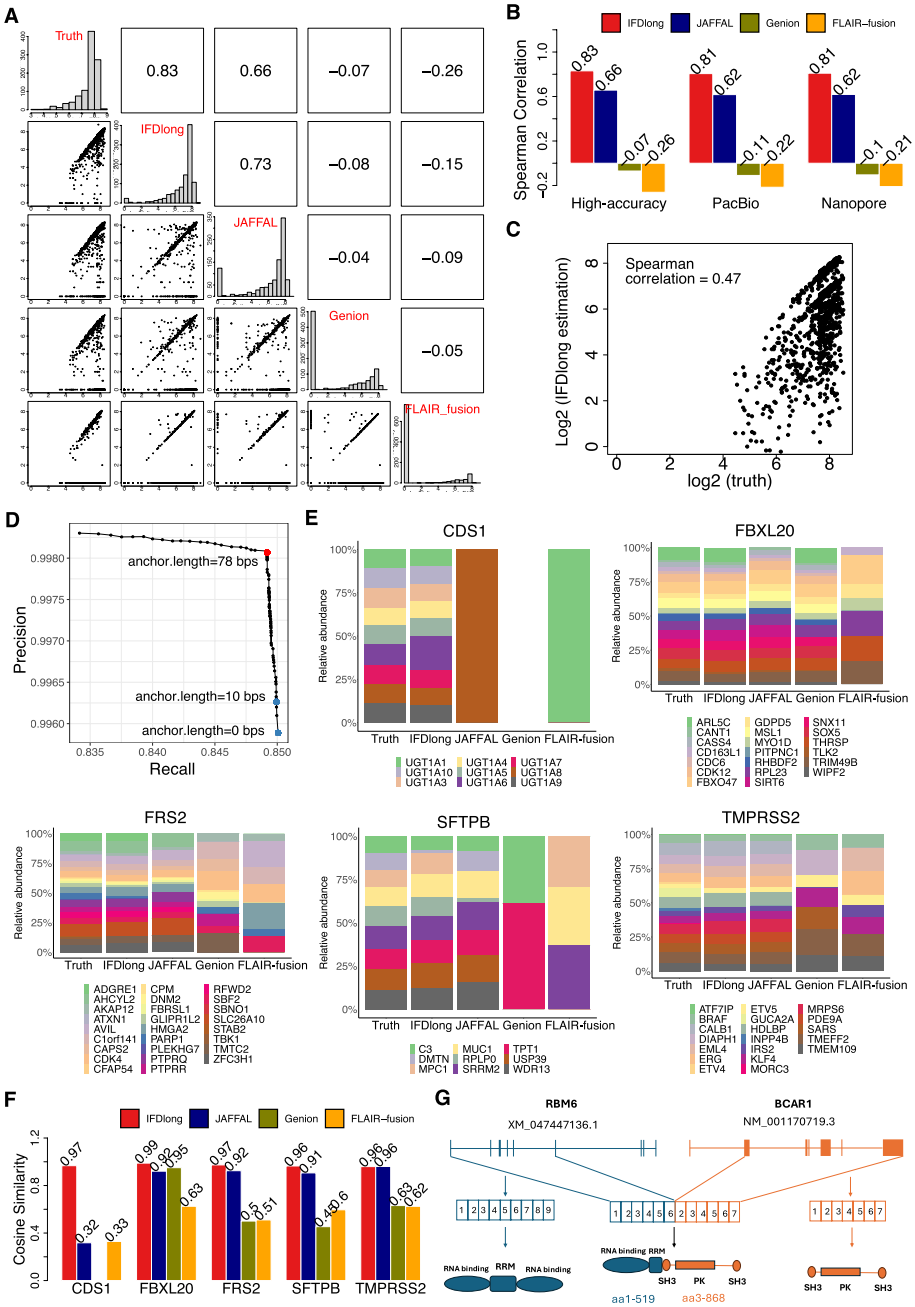



**Fig. 7** Detection and quantification of fusion transcripts in simulation studies. **A** Pairwise comparison of the fusion gene quantification tools in the Type F-1 simulation data with high-accuracy sequencing. The bottom left cells present the pairwise scatter plot of fusion expression. The upper right cells indicate Spearman's correlation coefficient. **B** SCC of fusion expression between the truth and each tool under different sequencing accuracies in the Type F-1 simulation data. The bars are colored by different fusion quantification methods: IFDlong, JAFFAL, Genion, and FLAIR-fusion. **C** Fusion isoform expression between the truth and the IFDlong estimation. **D** Precision and Recall curve for fusion transcript detection by IFDlong pipeline using different anchor lengths. **E** Relative abundance of the relative abundance of multiple paired genes in the Type F-2 simulation data, with high-accuracy setting and a 1500 bp mean read length. **F** Cosine similarity of multiple paired genes in the Type F-2 simulation data, with high-accuracy setting and a 1500 bp mean read length. **G** Functional modules of novel fusion transcript RBM6 – BCAR1 discovered by the IFDlong pipeline in the MCF7 cell line. Top: Genome structure of RBM6 (XM_047447136.1) and BCAR1 (NM_001170719.3). The karyotype genomic position of the gene is illustrated. Each bar and line indicates an exon or an intron, respectively. Middle: Wild-type mRNA and fusion RNA diagram from transcription. Bottom: Protein domains and structure diagrams for RBM6, RBM6-BCAR1, and BCAR1

the PacBio sequencing accuracy. As shown in Fig. 7B and Additional file 2: Fig. S22A-B, IFDlong always performed the best with different accuracy settings. Of note, tools such as Genion and FLAIR-fusion present comparatively low correlation with the truth, because the majority of the fusion genes were not detected (false negatives indicated by the dots sitting on the x-axis in the scatter plot of Fig. 7A and Additional file 2: Fig. S22A, B). For example, IFDlong was able to detect 977 out of 1000 fusion genes in the high-accuracy setting, while JAFFAL, Genion and FLAIR-fusion can only detect 876, 497 and 236 true positives. If we only count these detected fusions, all the tools present high correlation with the truth (Additional file 2: Fig. S22C-E).

Exploring fusion transcripts at isoform resolution is one of the major advantages of IFDlong. Compared with the three existing tools, IFDlong is the only tool that can annotate and quantify fusion isoforms (Fig. 1B). The scatter plots comparing the true fusion isoform intensities and the IFDlong estimated expressions are shown in Fig. 7C and Additional file 2: Fig. S23, where IFDlong can reach an SCC of 0.5 in the Type F-1 dataset. Given the increasing complexities of fusion quantification at isoform resolution compared with gene-level analysis, some fusion isoforms were underestimated by IFDlong (dots located underneath the diagonal). This is an expected outcome under the current estimation strategy, as reads originating from fusion transcripts are partially compatible with a single known transcript. Consequently, these reads are assigned to annotated transcripts, leading to an underestimation of the fusion transcripts' expression level.

One of the key factors for fusion detection is the setting of anchor length, which indicates the minimum base pairs of read alignment to each fusion gene [30]. A shorter anchor length will result in a higher recall rate, but at the cost of precision rate, because a larger number of false positives with short aligned supporting reads were introduced. In contrast, a longer anchor length can reach higher precision with the loss of recall, given true fusions with shorter alignment will be mis-detected (false negatives). To test the sensitivity of IFDlong to anchor length, Fig. 7D illustrates the precision and recall (PR) curve of fusion transcript detection with anchor length ranging from 0 bp to 100 bp. Fig. 7D shows that the highest F1-score for this high-accuracy simulation dataset occurs at an anchor length of 78 bp, and Additional file 2: Fig. S24 presents similar PR curves under both PacBio and ONT accuracy settings. IFDlong consistently maintains high precision (>0.99) and recall (>0.83) across the tested range. In practice, the optimal anchor length depends on sequencing quality and the desired balance between precision and recall. Lower sequencing accuracy reduces both precision and recall, requiring larger anchor lengths to achieve the best F1-score. In general, longer reads support the use of larger anchor lengths, which will largely reduce false positives while maintaining an acceptable level of true positives. Conversely, to maximize fusion candidate discovery in real practice, a smaller anchor length is preferable. By default, IFDlong uses 10 bp as suggested by many short-read RNA-seq pipelines [30], but users can adjust this parameter to best balance precision and recall for their datasets.

#### Distinguishing multiple fusion-mates paired with the same gene in simulation data

A gene frequently involved in fusion events may potentially pair with multiple mates in real practice. The Type F-2 datasets were simulated to estimate the software performance

in distinguishing different fusion-mates. Specifically, based on the TCGA fusion gene database, the top five frequently-fused genes (FBXL20, CDS1, TMPRSS2, FRS2, and SFTPB) with their corresponding mate gene panels (Additional file 1: Table S10) were selected as templates to generate the simulation long reads. IFDlong, Genion, JAFFAL, and FLAIR-fusion were then applied to this dataset to quantify the expression level of all fusion pairs. As a result, the relative abundances of all the fusion mates are compared across the truth and the software estimations in Fig. 7E, and their corresponding cosine similarities are summarized in Fig. 7F. For example, TMPRSS2 is a popular fusion gene in prostate cancer. Other than the most prevalent TMPRSS2-ERG fusion [71], in total 19 genes are reported to be fused with TMPRSS2 in the TCGA database (Additional file 1: Table S10). Compared with the truth, both IFDlong and JAFFAL achieved high agreement (cosine similarity = 0.96) when estimating the relevant abundance of these 19 fusion transcripts in the simulation data with high-accuracy sequencing and mean read length to be 1500bp (Fig. 7E and F). In all five simulation datasets, IFDlong presents the highest cosine similarity with the truth (higher than 0.96), while the three existing tools result in unstable performance. All the tools showed high consistency with the truth in the FBXL20-fusion simulation, but JAFFAL, Genion and FLAIR-fusion failed in the CDS1-fusion simulation. Similar results were observed in the simulation data with different accuracy settings (accurate, PacBio, and ONT) and mean read lengths (200 bp, 1000 bp, and 1500 bp) (Additional file 2: Figs. S25-S26). IFDlong achieved the highest cosine similarity (higher than 0.94) for all the simulation scenarios, followed by JAFFAL which performed well in all the settings other than the CDS1-fusion.

#### Two-segment and three-segment fusion detection in real data application

Besides the *in silico* simulation datasets, the fusion callers were applied to real sequencing data for performance evaluation and novel fusion transcript detection. The IFDlong fusion detection pipeline, as well as three existing tools (JAFFAL, Genion, and FLAIR-fusion), were evaluated in four real data sets: MCF7 breast cancer (bulk long-RNA-seq by both the PacBio and the ONT platforms), H838 Lung adenocarcinoma cell line (single-cell long-RNA-seq by the ONT platform), samples from colon cancer patients (bulk long-RNA-seq by the LoopSeq platform), and liver tissue samples from hepatocellular carcinoma (HCC) patients (single-cell long-RNA-seq by the LoopSeq platform). Among them, colon cancer and HCC cohorts are in-house samples, where novel fusions were detected and validated in the literature [6, 55]. As shown in Additional file 1: Table S11, the IFDlong pipeline successfully detects all these two-segment fusions, while Genion, JAFFAL and FLAIR-fusion can only identify 10, 5 and 6 fusions out of 11. All the fusion callers were applied to MCF7 and H383 datasets and the two-segment fusions identified by IFDlong are listed in Additional file 1: Table S12.

In addition to two-segment fusions that are frequently observed, three-segment fusions were discovered in the previous research [72], while most of the computationally predicted three-segment fusions may result from artificial chimeras or misalignment events. To test the ability of three-segment fusion discovery, Additional file 1: Table S11 lists 12 fusion transcripts detected from MCF7 and H838 by JAFFAL [33]. The IFDlong pipeline can successfully detect all of them, while Genion and FLAIR-fusion mis-filtered, reported partial or failed to detect them.

### Experimental validation of novel fusion transcript detected in real data

Computational analysis by the IFDlong generated a long list of fusion transcripts in real data applications. To further avoid the fusions that might be caused by artificial chimera events, we filtered these fusions by the criteria of sufficient supporting reads and no gap between the two fused sequences in the long reads. Further, real-time polymerase chain reaction (RT-PCR) was performed on selected fusions for experimental validation of the novel fusions discovered by IFDlong. The analysis reveals a fusion RNA species in the MCF7 cell line involving BCAR1. This fusion transcript is detected by both IFDlong (supported by 7 long reads) and JAFFAL from the ONT data, but mis-filtered by Genion and FLAIR-fusion pipeline (detected in the intermediate files, but not listed in the final reports). The RBM6-BCAR1 fusion results in the mRNA containing the first 6 exon sequences of RBM6 and exons 2–7 of BCAR1 (Fig. 7G and Additional file 2: Fig. S27). RBM6 is an RNA-binding protein involved in RNA splicing [73]. The fusion protein lost a part of the RNA recognition motif and the C-terminal hydrophilic domain which may involve RNA splicing activity. On the other hand, BCAR1 is an adaptor protein that contains SH3 and substrate motifs for several protein kinases. It plays a crucial role in cell motility [74], apoptosis [75], and cell cycle regulation [76]. Overexpression of this protein indicates a poor prognosis for several cancers [77, 78]. All the domains of BCAR1 were largely intact in the predicted fusion protein. Since BCAR1 is located in the cytoplasm [79] while RBM6 is in the nucleus [80], the fusion may alter the distribution of either of the proteins. Significant biological impact is expected for this fusion event that the IFDlong discovers.

## Discussion

In this paper, we introduce IFDlong, a novel bioinformatics tool tailored for the analysis of bulk and single-cell long-RNA-seq data. IFDlong offers several distinct advantages over existing tools, making it a comprehensive solution for accurate annotation and quantification of isoforms and fusion transcripts. First, IFDlong provides a suite of functions that encompass various aspects of long-RNA-seq analysis (Fig. 1), including long-read annotation at both gene and isoform resolution, prediction of amino acid sequences, known isoform quantification (Figs. 2, 3 and 5), novel isoform discovery (Fig. 4), and detection of fusion transcripts (Fig. 7). Second, IFDlong performs model-based estimation to infer the isoform and fusion abundance accurately. Specifically, the EM algorithm was developed to address the long reads with ambiguous annotations (reads with uncertain isoform assignments, Fig. 1A) and fusion transcripts consisting of multiple alternative splicing variants (Fig. 1B). IFDlong achieves higher accuracy in terms of isoform and fusion quantification when compared with the existing tools (Figs. 2, 3, 4, 5, 6, 7). Third, IFDlong employs multiple filtering criteria to control false positives when detecting novel isoforms and fusion transcripts. IFDlong enhances the accuracy of fusion detection by removing fusion candidates involving pseudogenes, genes from the same family, and readthrough events. Moreover, the user-adjustable parameters, such as buffer length for novel isoform detection (Fig. 4) and anchor length for fusion identification (Fig. 7D), provide flexibility in customizing the analysis pipeline to adapt to broad applications. Fourth, IFDlong innovatively detect and quantify fusion transcript at isoform resolution, where the current tools only report gene-level fusions. Fifth, IFDlong

presents advantages in its versatility and compatibility with diverse experimental setups and species. Specifically, IFDlong is applicable for both human and mouse data (Fig. 5) and can be easily generalized to other species given the corresponding reference and annotation profiles. Moreover, IFDlong is compatible with different experimental platforms (PacBio, ONT, and linked-short-read platforms) and multiple library preparation strategies (bulk and single-cell RNA-seq, as shown in Fig. 6; CapTrap, direct RNA, R2C2, cDNA libraries, as shown in Fig. 5C and Additional file 2: S13D). In addition, IFDlong can take in the alignment file by different long-read aligners, such as minimap2 [37] by default or STAR-long [81]. All these properties make IFDlong a generalizable tool that can be applied to broad applications and adjusted to best suit individual datasets.

In this project, comprehensive comparisons against existing tools were performed to demonstrate the superior performance of IFDlong through extensive simulations and real-data applications. For isoform analysis, our tools is compared with seven existing tools: LIQA [12], TALON [17], Mandalorion [16], Bambu [13], FLAIR [18, 19], mini-Quant [14] and IsoQuant [15]. And for fusion transcript detection, IFDlong is compared with JAFFAL [33], Genion [34] and FLAIR-fusion [35]. Five types of simulation datasets were generated to show the high performance of IFDlong in terms of isoform quantification (Type I-1 and I-2), novel isoform detection (Type I-3), and fusion quantification (Type F-1 and F-2). To better mimic real data for the co-existence of multiple TSV events, we further pooled the five simulation datasets and evaluated the IFDlong performance. IFDlong presents robust performance (Additional file 2: Fig. S28) when compared with the individual datasets (Figs. 2, 3, 4, and 7). Our tool can successfully deal with the co-existence of both novel isoforms and fusion transcripts, which could potentially impact the identification and quantification. In addition to simulation data, multiple real datasets were employed to test the generalizability of the IFDlong pipeline. We have applied different benchmarking methods from the LRGASP [36], including multiple accuracy metrics if knowing the underlying truth, and consistency measures without the truth. The robustness and reliability of IFDlong make it a prioritized tool for researchers seeking accurate and comprehensive analysis of long-RNA-seq data.

Existing comparative studies have highlighted the advantages and limitations of long-read approaches compared with short-read methods [14, 36, 82]. Quantification of transcript abundances has traditionally relied on short-read RNA-seq data [83], and tools such as kallisto [84], RSEM [85], and StringTie [86] are well established for isoform-level expression estimation. In contrast, long-read technology enables the direct observation of full-length isoforms, more accurate estimation of relative isoform abundance with fewer assumptions, and precise identification of splicing junctions, novel isoforms, and other TSVs. Nevertheless, long-read RNA-seq still faces key limitations—including lower throughput, higher error rates, and higher cost—which present challenges for transcript quantification. IFDlong is specifically designed for long-read RNA-seq and is compatible with both short- and long-read datasets (as demonstrated in Fig. 2F and G across varying read lengths), although it has not yet been fully optimized for short-read applications.

Our analysis pipeline also highlights the significant potential of the single-cell long-RNA-seq data analysis. The HCC study enables the delineation of multiple isoform components at the single-cell resolution. This approach facilitates the identification of

rare cell populations, evaluation of cell-to-cell variability, and the reconstruction of cellular trajectories, thereby offering a comprehensive view of the immune landscape in HCC. As a future direction of the single-cell analysis, clustering algorithms addressing the sparse isoform expression profile are urgently needed. Moreover, isoform-specific databases, such as molecular and signaling pathways and ligand-receptor interactions, are expected, so that the isoform-resolution single-cell analysis can be further extended. These approaches can be generalized to comprehend the molecular mechanisms underlying diverse diseases and biological processes. These insights guide the development of more targeted and effective therapeutic strategies tailored to individual patients. Given the great potential of the single-cell long-RNA-seq analysis, it will be an urgent need for power calculation of the lowly expressed transcripts and addressing the potential technical biases for single-cell data.

Fusion transcripts can be classified into different categories based on their formation mechanisms, junction points, and protein structures. By formation mechanism, fusion transcripts fall into two groups [32, 87]. Genomic rearrangement-dependent fusions arise from DNA-level changes and can be further subdivided into direct fusions, caused by a single structural rearrangement, and composite fusions, resulting from multiple rearrangements. In contrast, genomic rearrangement-independent fusions are RNA-level chimeric events, including trans-splicing, transcriptional readthrough, and cis-splicing between adjacent genes (cis-SAGe). IFDlong can detect both two-segment and three-segment fusions, except for intragenic splicing events involving sense and antisense strands of the same gene. Such cases are reported as novel isoforms since they involve only a single gene. Based on junction points, fusion transcripts are further classified as intact exon (IE) type, where breakpoints occur at exon boundaries, or broken exon (BE) type, where breakpoints occur within exons [30]. IFDlong identifies both fusion types and precisely locates their junction points. Finally, based on downstream protein structure, fusion transcripts can be categorized as chimera protein-forming, independent wild-type tail gene, or non-fusion-forming [31]. The IFDlong output file provides AA annotations for individual long reads, enabling detailed downstream protein structure analysis.

This study has limitations that should be addressed in the future. Compared with other existing tools, both the simulated data and the IFDlong pipeline used the same hg38 reference genome and gene-symbol–based annotation profiles downloaded from the UCSC Genome Browser. When comparing IFDlong with other tools, we used the same reference and annotation files whenever the tools allowed user-defined annotations. For tools that did not allow this, we relied on their embedded annotation files, which may differ from those used in IFDlong. Such discrepancies could lead to mismatches due to differences in the number of annotated genes or the use of alternative gene names across databases. We also recognize that some tools using GENCODE Ensembl-ID-based annotation files performed well in the LRGASP project, but when applied with our gene-symbol–based annotations, they may miss many true positive isoforms. On the other hand, the qPCR or short-read RNA-seq data served as the truth for the comparison on real datasets, avoiding the annotation bias issue. We note that short-read isoform quantification is not a gold standard truth for real data, and more indirect evaluation methods (such as biological interpretation of the results) were applied. For the performance

comparisons, all the tools applied their default parameter settings, including our IFDlong pipeline. For fair comparison, we didn't tune parameters for each method which may potentially further increase the performance.

Given the advantages and limitations of IFDlong, several directions for future improvements can be envisioned. (1) Computational efficiency: The current runtime of IFDlong was around 3 hours using 30 CPU cores to process the H1_mix samples with ~20 million long reads (≈19 billion base pairs; FASTQ file size ≈38 GB). While the current version of IFDlong is efficient and manageable, we aim to further reduce computational cost by incorporating more efficient C++ programming, optimizing the long-read annotation step, and leveraging k-mer-based techniques for faster convergence of the estimation algorithm. (2) Alignment flexibility: The current pipeline is based on alignment profile from a single aligner (minimap2 by default). To take advantage of multiple long-read aligners (such as STAR-long [81] and GMAP [88]), the pipeline can be improved by integrating alignment files generated from multiple tools. (3) Bias correction: IFDlong does not yet account for sequence-specific, fragment-GC, or positional biases. Inspired by short-read-based methods such as Sailfish [89] and Salmon [90], we will incorporate bias correction into IFDlong's estimation models. (4) Handling paralogous sequences: Long reads arising from pseudogenes or gene families are currently annotated by their primary alignment or filtered out to control false positives. More sophisticated approaches for handling paralogous sequences will be developed in future versions. (5) Tools, such as miniQuant [14], that can integrate both short and long reads from the same sample will be the future trend. Although IFDlong is compatible with both short- and long-read data, it has not yet been optimized for short-read or hybrid applications. Future work will focus on leveraging the complementary strengths of both read types to enhance accuracy and robustness. (6) Isoform-resolution downstream analyses: Most transcriptome downstream analyses remain gene-centric, such as pathway enrichment of differentially expressed genes and single-cell interaction analyses using gene-based ligand-receptor databases. Developing isoform-resolution databases and statistical approaches will be an important future direction.

## Conclusions

IFDlong presents significant advancements in long-RNA-seq analysis for annotating and quantifying isoforms and fusion transcripts. Its unique features, including integrating the EM algorithm, stringent false-positive control, compatibility, and superior performance, position IFDlong as a versatile tool for long-read transcriptome research. This novel bioinformatics software will contribute to the community through its broad biomedical applications.

## Methods

### The preparation of reference and annotation files

IFDlong requires the reference genome in FASTA format, the transcript annotation profile in GTF format, the gene annotation profile in GTF format, the amino acid (AA) annotation profile in TXT format, the pseudo gene database in RDS format, and the root gene symbols of all gene families in TXT format. All human and mouse reference and annotation files are well prepared (available for download on GitHub) and will be fully

described below. To apply to species other than human and mouse, IFDlong provides the "refDataSetup.sh" function to build up the reference files.

IFDlong can take in either a raw sequencing read file (in FASTQ format) or an aligned file (in BAM format). If working on the FASTQ file, a reference genome in FASTA format is required to call the minimap2 aligner [37] (Fig. 1C). For gene and isoform annotation, a gene annotation profile in BED format is required by IFDlong. For example, the annotation file in GTF format can be downloaded from the UCSC genome reference Consortium (e.g., GRCh38 for humans and GRCm38 for mice downloaded from https://hgdownload2.soe.ucsc.edu/downloads.html). Then the GTF file will be formatted by IFDlong (via the command IFDlong.sh) into BED format, where each row records the chromosome, start position, end position, name, score, and strand of each consisting exon. Ultimately, a known gene and a known transcript profile will be used by the IFDlong pipeline for gene and transcript annotation.

IFDlong will build the AA annotation profile for each known transcript. Based on the reference genome (in FASTA format, described above) and the transcript profile (in BED format, described above), IFDlong will first apply the *getfasta* function in *BEDtools* [38] to extract the DNA/RNA sequences per transcript, followed by the *translate* function in *Biostrings* R package to translate the DNA/RNA sequences into AA sequences.

IFDlong involves innovative pipelines to filter out false positive fusion candidates using the database for gene families and pseudogenes. The human pseudogene database was downloaded from Pseudogene.org (http://pseudogene.org/psicube/data/gencode.v10.pseudogene.txt), and the mouse pseudo gene database was collected from Mouse Genome Informatics (https://www.informatics.jax.org/downloads/reports/MGI_BioTypeConflict.rpt). To act as a complementary function, additional pseudo genes were identified based on the gene descriptions obtained from *Bioconductor* packages *org.Hs.eg.db* (for human) and *org.Mm.eg.db* (for mouse). The pseudogene names were saved in an RDS file. To collect gene family information, the human database was downloaded from the HGNC BioMart server (https://biomart.genenames.org), which contains the family names and common root gene symbols of each human gene family. For mouse gene family information, in light of the lack of a reliable open-source database, the mouse genes were first mapped to human homologous genes by MGI Data and Statistical Reports (https://www.informatics.jax.org/downloads/reports/index.html). Then the mapped mouse genes will employ the same gene family database as the human one.

#### Isoform-resolution quantity estimates

The IFDlong isoform quantification method is motivated by the LIQA [12], with an improved estimation algorithm utilizing the results from the upstream isoform annotation analysis. First, our long-read isoform annotation, as well as the subsequent filtering steps, contribute to a superior initialization of the isoform abundance estimation. Second, IFDlong comprehensively identifies known, novel and fusion isoforms, which makes the abundance of each category more accurate. IFDlong can achieve the simultaneous abundance estimation of known single-isoform, novel isoform and fusion isoform. Third, to make the best advantage of long-read sequencing, we estimate the fusion expression at isoform resolution, where the current existing tools only detect fusion transcript at gene resolution.

**Isoform quantity estimates**

Mathematically, given a gene of interest denoted by $g$, let $R^g$ represent the set of reads (denoted by $r$) that are aligned to gene $g$, and $I^g$ is the set of known isoforms (denoted by $i$) from gene $g$. For a specific isoform $i \in I^g$, let $\theta_i^g$ denote its relative abundance, with $0 \le \theta_i^g \le 1$ and $\sum_{i \in I^g} \theta_i^g = 1$. The probability that a read originates from isoform $i$ is defined as $\Pr(isoform = i) = \theta_i^g$. We further define the indicator matrix $Z_{R^g, I^g} \in \{0,1\}^{|R^g| \times |I^g|}$ that is unobserved with an entry $z_{r,i} = 1$ if read $r$ is generated from a molecule that originated from isoform $i$, and otherwise $z_{r,i} = 0$. For isoform quantification, the goal is to estimate the relative abundance $\Theta^g = \{\theta_i^g, i \in I^g\}$ based on RNA-seq long reads annotated to the gene $g$.

With the notation above, the complete data likelihood of the long-RNA-seq aligned to gene $g$ can be written as

$$L\left(R^g, Z_{R^g, I^g} | \Theta^g\right) = \prod_{r \in R^g} \prod_{i \in I^g} \left\{\Pr\left(read = r, isoform = i | \Theta^g\right)\right\}^{z_{r,i}}$$

$$= \prod\nolimits_{r \in R^g} \prod\nolimits_{i \in I^g} \left\{\Pr\left(read = r | isoform = i, \Theta^g\right) \theta_i^g\right\}^{z_{r,i}},$$

where $\Pr\left(read = r | isoform = i, \Theta^g\right) = \begin{cases} \frac{1}{|R^{gi}|}, i \in I^{gr} \\ 0, i \notin I^{gr} \end{cases}$, $R^{gi}$ denotes the set of reads that are annotated to isoform $i$ (i.e., a subset of $R^g$ with $R^{gi} \subset R^g$), and $I^{gr}$ represents the set of isoforms that read $r$ is annotated to (similarly, $I^{gr} \subset I^g$). Note that it is possible that IFDlong annotates a read to multiple known isoforms that are highly overlapped with each other, and in this case, $|I^{gr}| > 1$.

Next, with the complete data likelihood, the update procedure of the Expectation-Maximization (EM) algorithm to estimate the parameter $\Theta^g$ is as follows:

Initialization: Taking the reads aligned to multiple isoforms as a supporting read for each of the isoforms (e.g., if one read is aligned to two isoforms of gene $g$, it will serve as a supporting read for both isoforms, as the red long-read illustrated in Fig. 1A), $\Theta^g$ will be initialized as the proportion of the supporting reads aligned to each isoform for gene $g$.

Expectation-step (E-step): Calculate the expectation of the log-likelihood with the following function:

$$Q\left(\Theta^g | \Theta^{g(t)}\right) = \sum\nolimits_{r \in R^g} \sum\nolimits_{i \in I^g} E_{\left(Z_{R^g, I^g} | \Theta^{g(t)}\right)} \left\{z_{r,i}\right\} \log\left[Pr\left(read = r | isoform = i, \Theta^g\right) \theta_i^g\right],$$

where $E_{\left(Z_{R^g, I^g} | \Theta^{g(t)}\right)} \left\{z_{r,i}\right\} = \frac{\theta_i^{g(t)}}{\sum_{k \in I^{gr}} \theta_k^{g(t)}}$.

Maximization-step (M-step): Maximize function $Q\left(\Theta^g | \Theta^{g(t)}\right)$ by $\Theta^g$ with the following formula,

$$\theta_i^{g(t+1)} = \sum\nolimits_{r \in R^g} E_{\left(Z_{R^g, I^g} | \Theta^{g(t)}\right)} \left\{z_{r,i}\right\} / \left|R^g\right|, \text{for } i \in I^g.$$

The EM algorithm iteratively updates the parameter by alternating between the E-step and M-step until convergence.

### Fusion quantity estimates

A similar idea is applied to fusion quantification. Given a gene fusion $f$ consisting of all genes $g_j^f$ in the gene set $G^f$, let $R^f$ denote the set of reads that are annotated to the gene fusion $f$, and $C^f$ is the set of fusion transcripts from the gene fusion (i.e., each isoform component $i_{cj}$ of a fusion transcript $c \in C^f$ is from one of the gene $g_j^f$ in $G^f$). Then, we denote the relative abundance of each fusion transcript to be $\theta_c^f$ with $0 \le \theta_c^f \le 1$ and $\sum_{c \in C^f} \theta_c^f = 1$, so the probability that a read originates from the fusion transcript $t$ is $\Pr(fusion = c) = \theta_c^f$. Similarly, we define the unobserved read-fusion transcript compatibility matrix $Z_{R^f,C^f} \in \{0,1\}^{|R^f| \times |C^f|}$ with entry $z_{r,c} = 1$ if the read $r$ is generated from a molecule that originated from fusion transcript $c^f$, and otherwise $z_{r,c} = 0$. For fusion transcript quantification, our goal is to estimate the relative abundance $\Theta^f = \{\theta_c^f, c \in C^f\}$ based on RNA-seq long reads annotated to the gene fusion $f$.

Similarly, the complete data likelihood is

$$L\left(R^f, Z_{R^f,C^f} | \Theta^f\right) = \prod_{r \in R^f} \prod_{c \in C^f} \left\{ \Pr\left(read = r, fusion = c | \Theta^f\right) \right\}^{z_{r,c}}$$
$$= \prod\nolimits_{r \in R^f} \prod\nolimits_{c \in C^f} \left\{ \Pr\left(read = r | fusion = c, \Theta^f\right) \theta_c^f \right\}^{z_{r,c}},$$

where $\Pr\left(read = r | fusion = c, \Theta^f\right) = \begin{cases} \frac{1}{|R^{fc}|}, c \in C^{fr} \\ 0, c \notin C^{fr} \end{cases}$, $R^{fc}$ denotes the set of reads that are annotated to fusion transcript $c$ (i.e., a subset of $R^f$, $R^{fc} \subset R^f$), and $C^{fr}$ represents the set of fusion transcripts that the read $r$ is annotated to (similarly, $C^{fr} \subset C^f$). Again, it is possible that IFDlong annotates a read to multiple fusion transcripts which are highly overlapped with each other, and in this case, $|C^{fr}| > 1$.

The EM algorithm iteratively updates the parameter $\Theta^f$ by alternating between the E-step and M-step as below until convergence.

E-step: calculate the expectation of the log-likelihood by

$$Q\left(\Theta^f | \Theta^{f(t)}\right) = \sum\nolimits_{r \in R^f} \sum\nolimits_{c \in C^f} E_{\left(Z_{R^f,C^f} | \Theta^{f(t)}\right)} \left\{ z_{r,c} \right\} \log \left[ Pr\left(read = r | fusion = c, \Theta^f\right) \theta_c^f \right],$$

where $E_{\left(Z_{R^f,C^f} | \Theta^{f(t)}\right)} \left\{ z_{r,c} \right\} = \frac{\theta_c^{f(t)}}{\sum_{k \in C^{fr}} \theta_k^{f(t)}}$.

M-step: maximize function $Q\left(\Theta^f | \Theta^{f(t)}\right)$ with respect to $\Theta^f$ and have

$$\theta_c^{f(t+1)} = \sum\nolimits_{r \in R^f} E_{\left(Z_{R^f,C^f} | \Theta^{f(t)}\right)} \left\{ z_{r,c} \right\} / \left| R^f \right|, \text{for } c \in C^f.$$

### Simulation data generation

*In silico* simulation data were generated to benchmark the software performance with the underlying truth. Type I and Type F simulation datasets were generated for isoform and fusion evaluation, respectively.

The Type I-1 simulation datasets were generated to mimic the distribution of whole transcriptome expression. First, to simulate the distribution of isoform expression from real data, Universal Human Reference (UHR) data (details described later) was employed, where 28157 isoforms were detected and quantified (Additional file 2: Fig. S1A). Second, sequence templates were constructed using these isoforms by synthesizing the exons of each isoform according to the transcript annotation profile. Next, multiple datasets of long-RNA-seq were generated by the long-read simulator PBSIM2 [39] with different accuracy settings and gradient mean read lengths. Three accuracy setting were generated by adjusting the PBSIM2 parameters: high-accuracy sequencing (*--difference-ratio=1:1:0* for substitution:insertion:deletion, and *--accuracy-mean=0.99*); Naopore-like accuracy (*--difference-ratio=23:31:46* and *--accuracy-mean=0.85)*; and PacBio-like accuracy (*--difference-ratio=6:50:54* and *--accuracy-mean=0.85)*. Under the setting of high-accuracy sequencing, simulation datasets with gradient mean read lengths were generated by setting *--length-mean* to be 200 bp, 500 bp, 800 bp, 1000 bp, 1200 bp, 1500 bp, 1800 bp, and 2000 bp (Additional file 2: Fig. S2). Finally, each pool of simulated reads with a specific accuracy and mean read length setting was subsampled to gradient sequencing depth with 0.5, 1, 2, and 4 million reads, respectively (Additional file 2: Fig. S1B). Note that, the expression intensities of the 28157 isoforms follow the distribution of isoform expressions in the UHR data. In the meanwhile, the transcript and gene origin of each simulated long read were recorded to serve as the underlying truth.

The Type I-2 simulation datasets were designed to generate long reads from multiple isoforms arising from the same gene. The sequencing template contained 200 non-pseudo genes, where each gene consisted of 8–12 different isoforms. Next, similar to the Type I-1 simulation, multiple Type I-2 datasets were generated with a sequencing depth of 400 under different mean read lengths (1000 bp and 2000 bp) and the three accuracy settings (accurate, PacBio, and ONT).

The Type I-3 simulation datasets aim to benchmark the novel isoform detection. First, the sequence templates were constructed by two isoform sets: 1000 known isoforms and novel isoforms. Novel isoforms were derived from the 1000 known transcripts with six different types: 859 templates with skipping exon, 1406 templates with additional exon, 873 templates with truncated alternative 5' splice site, 1329 templates with extended alternative 5' splice site, 873 templates with truncated alternative 3' splice site, and 1325 templates with extended alternative 3' splice site. Each type of novel isoform was pooled with the 1000 known transcripts to generate six sequence templates in total. Per novelty type, simulator PBSIM2 was employed to generate long-RNA-seq datasets with a sequencing depth of 40 under the mean read length of 1000 bp and 2000 bp, and the three different accuracy settings. The origin source and the novel type of each simulated long read were recorded to serve as the truth for performance evaluation.

The Type F-1 datasets were simulated for fusion transcript quantification. The sequence templates consist of two sets of isoforms: 1000 selected known isoforms arising from different genes, and 1000 mosaic fusion transcripts combining two known isoforms from the 1000 known list with random breakpoints. Based on these template sequences, PBSIM2 was employed to generate simulation datasets under three accuracy settings with a mean read length of 1500 bp and a sequencing depth of 400. According

to the simulation log files, the read fusion status (fusion transcripts or normal isoform) and the isoform/gene origin information of each long-read sequence were collected as underlying truth.

The Type F-2 simulation datasets were intentionally designed to mimic the genes fused with multiple mates. To simulate the gene pairs that have high occurrence in real data, the TCGA fusion database (https://www.kobic.re.kr/chimerdb/download) was used as the reference. Among the top 10 cancer types with the highest number of fusion transcripts, top 5 genes (referred as popular genes) with the highest number of paired genes were selected: FBXL20 (Breast Invasive Carcinoma, BRCA), CDS1 (Bladder Urothelial Carcinoma, BLCA), TMPRSS2 (Prostate Adenocarcinoma, PRAD), FRS2 (Sarcoma, SARC) and SFTPB (Lung Adenocarcinoma, LUAD). For each popular gene, the fusion templates were constructed by fusing the popular gene and its mates according to the fusion breakpoints in the TCGA fusion database. Then based on these fusion templates, long-RNA-seq datasets were simulated under different mean read lengths (200 bp, 1000 bp, and 1500 bp) and three accuracy settings (high-accuracy, PacBio, and ONT) with a sequencing depth of 40 by PBSIM2. All the fusion origins and breakpoints of each simulated long-read sequence were collected as the truth.

A merged simulation dataset combining all the above five types was generated to mimic the real data, where both novel isoforms and fusion transcripts exist and may potentially impact the detection and quantification. Per accuracy setting (high-accuracy, PacBio, and ONT), the 1500 bp reads of each five types were concatenated as the merged simulation data. Correspondingly, the true isoform and fusion expressions were summed up across the five datasets.

### Real data collection and pre-processing

#### *Universal human reference (UHR)*

Long-RNA-seq on the UHR sample was used to evaluate isoform quantification. The direct mRNA data sequenced by the ONT platform was downloaded from Gene Expression Omnibus (GEO) PRJNA639366 [12] with a total of 476,000 reads and 896 base pairs in median, and 441,138 reads were aligned to human reference by minimap2 aligner [37]. The UHR RNA (Agilent) + SIRV Isoform Mix E0 (Lexogen) sample measured by the PacBio platform was downloaded from the PacBio website (https://downloads.pacbcloud.com/public/dataset/UHR_IsoSeq/). In total 6,775,127 reads with 1835 base pairs in median length were collected, and 6,385,883 reads were aligned to human reference by minimap2 aligner. To serve as the underlying truth, measurements by the TaqMan Real-Time PCR Assays on the Stratagene UHR RNA samples were collected from the MicroArray Quality Control (MAQC) project via GSE5350 [43]. The geometric mean of the four repeating measurements was calculated as the true expression. In total 1044 isoforms were quantified and 964 of them have passed the detection flag that will be used for this project.

#### *Mouse data*

For isoform quantification, long-RNA-seq data on mouse C2C12 cell line and heart tissue were downloaded from the ENCODE project [44]. Four repeats of C2C12 were downloaded from the GEO database with accession ID GSE219813 and GSE219595, and

two repeats of heart data were downloaded from GSE219825. The isoform expressions of both mouse datasets were quantified by the TALON pipeline [17] in the ENCODE project [44].

#### *Single-cell long-RNA-seq on hepatocellular carcinoma (HCC) patient*

The dataset was downloaded from GSE223743, including one paired liver benign – tumor sample from an HCC patient [55]. This dataset was applied for isoform quantification, differential expression analysis, and fusion detection analysis. Single cells with at least 1000 long-reads were defined as valid cells, and in total 162 and 285 cells were analyzed from benign and tumor libraries. For fusion transcript detection, three fusion transcripts (EML4 - ACTR2, CCDC127 - PDCD6, and FGG - PLG) were validated by Sanger sequencing in the previous study [55] that will serve as the truth.

#### *Colon cancer samples*

Tissue samples were collected from 3 colon cancer patients, including benign colon tissues (N), primary colon cancer samples (T), and lymph node metastasis samples (M). The samples were sequenced by the Element Biosciences LoopSeq platform and the long-read data was deposited to the GEO database with accession GSE155921 [6]. Among the 8 samples, in total 4 two-way fusions (STAMBPL1 - FAS, SMYD3 - ZNF124, VAPB - GNAS, and ECHDC1 - PTPRK) were detected and validated using Taqman qRT-PCR and Sanger sequencing [6], which will serve as the underlying truth for this study.

#### *MCF7 human breast cancer cell line*

MCF7 RNA extractions sequenced by both the PacBio and the ONT platforms were downloaded for isoform and fusion analysis. For PacBio SMRT sequencing, long-read FASTQ files were downloaded from the NCBI SRA database with accession ID SRP055913 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP055913). The 113 runs were concatenated into one file for analysis. For SGNex ONT sequencing, data was downloaded from https://github.com/GoekeLab/sg-nex-data [91]. MCF7-EV_directRNA, MCF7_cDNA, MCF7_directRNA, and MCF7_directcDNA with multiple replicates were used in this study. As the true isoform expression, the RNA-seq on MCF7 with RSEM quantification of transcripts [85] downloaded from Cancer Cell Line Encyclopedia (CCLE) [92] was employed. Both two-way and three-way fusions that have been summarized in the previous research were used as fusion truth [33].

#### *Human H1-mix*

This ENCODE biosample contains the embryonic stem cells and definitive endoderm derived from H1 human. The LRGASP [36] has measured this sample with multiple platforms and library preparation strategies: PacBio – CapTrap (with three repeats), PacBio – cDNA (with three repeats), ONT – CapTrap (with three repeats), ONT – R2C2 (with six repeats), and ONT – cDNA (with three repeats). No dataset was used as true isoform expression, but the consistency among the multiple repeats were evaluated in this project.

#### *H838 human lung adenocarcinoma cell line*

Single-cell long-RNA-seq data on H838 sequenced by the ONT platform was downloaded from the GEO database with accession ID GSE154869 [93]. Both two-way and three-way fusion transcripts were detected from this dataset for performance evaluation.

### Existing Isoform and fusion tools for comprehensive comparison

long-RNA-seq alignment was performed by the tool minimap2 [37], and the BAM file was filtered by SAMtools [94] if needed. For isoform analysis, IFDlong was compared with seven tools: LIQA [12], TALON [17], Mandalorion [16], Bambu [13], FLAIR [18, 19], miniQuant [14] and IsoQuant [15]. For fusion detection, IFDlong was compared with three tools: JAFFAL [33], Genion [34] and FLAIR-fusion [35]. Their version and key script lines were summarized below.

#### *LIQA (v1.2.0)*

Liqa -task quantify -refgene $refgeneFile -out $outPath/$sample.isoform_expression_estimates -max_distance 10 -f_weight 1 -bam $BAMfile

#### *TALON (v6.0.1)*

Talon_label_reads --f SAMfile --g $genomeFAfile --t 1 --ar 20 --deleteTmp --o $sample
talon --f config.csv --db hg38_talon.db --build hg38 --o $sample
talon_summarize --db hg38_talon.db --v --o $sample
talon_abundance --db hg38_talon.db -a hg38_annot --build hg38 --o $sample
talon_filter_transcripts --db hg38_talon.db --datasets $sample -a hg38_annot --maxFracA 0.5 --minCount 5 --minDatasets 2 --o filtered_transcripts.csv
talon_abundance --db hg38_talon.db --whitelist filtered_transcripts.csv --db hg38_talon.db -a hg38_annot --build hg38 --o $sample
talon_create_GTF --db hg38_talon.db --whitelist filtered_transcripts.csv -a hg38_annot --build hg38 --o $sample

#### *Mandalorion (v1.0; Python version 3.8)*

python Mando.py -p $outputDir -g $GTFfile -G $genomeFAfile -f $FASTQfile

#### *Bambu*

(v3.2.4-Bioconductor; gcc version 10.2.0; R version 4.2.0)
se <- bambu(reads = BAMfile, annotations = annotations, genome = genomeFAfile, trackReads=TRUE)
writeBambuOutput(se, path = outputDir)

#### *FLAIR (v2.0.0)*

flair align -g $genomeFAfile -r $FASTQfile
flair correct -q flair.aligned.bed -g $genomeFAfile -f $GTFfile
flair collapse --check_splice -g $genomeFAfile -f $GTFfile -q flair_all_corrected.bed -r $FASTQfile

flair quantify --sample_id_only --reads_manifest manifest.tsv -i flair.collapse.isoforms.fa

***miniQuant (v1.2.1)***

miniQuant quant -r pathto/isoquant/ref_DB/transcriptome.v2.fa -l $FASTQfile

***IsoQuant (v3.7.0; Python version 3.8)***

isoquant.py --reference $genomeFAfile --genedb $GTFfile --fastq $FASTQfile --data_type nanopore[pacbio] -o $outputDir --report_novel_unspliced true

***JAFFAL (JAFFA v2.3)***

bpipe run JAFFAL.groovy $FASTQGZfile
Rscript make_final_table.R $sample.fastq_genome.psl $sample.reads pathto/JAFFA/hg38_genCode22.tabpathto/JAFFA/known_fusions.txt10000NoSupport,PotentialReadThrough 50 $sample.summary

***Genion (v1.1.1)***

minimap2 -x splice $genomeFAfile $FASTQfile -c -o $sample.align.paf
genion-i$FASTQfile-dpathto/genomicSuperDups.txt--gtf$GTFfile-g$outDir/$sample.align.paf -s pathto/cdna.selfalign.tsv -o $sample.tsv

***FLAIR-fusion (preprint version 19-03−2021)***

python flair.py align -g $genomeFAfile -r $FASTQfile
bamToBed -bed12 -i flair.aligned.bam > flair.aligned.bed
samtools view -h -o $SAMfile flair.aligned.bam
bedtools intersect -wao -a flair.aligned.bed -b pathto/FLAIR-fusion/gencode.v37.annotation-short.gtf > flair-bedtools-genes.txt
python bedtoolsGeneHelper.py -i flair-bedtools-genes.txt
python 19-03−2021-fasta-to-fusions-pipe.py -c -u -i -j -r $FASTQfile -m flair.aligned.bed -f flair.py -g $genomeFAfile -t $GTFfile -a pathto/FLAIR-fusion/gencode.v37.annotation-short.gtf

**Performance evaluation**

*Spearman's rank correlation coefficient (SCC)* is a nonparametric measure of statistical dependence between the rankings of two vectors. The correlation ranges from −1 to 1, with 1 indicating a perfectly identical ranking between the two vectors, and −1 representing a fully opposed correlation. For both isoform or fusion transcript quantification, the SCC between the true number of supporting reads (denoted as vector $\overrightarrow{X}$) and the tool-estimated number of supporting reads (denoted as vector $\overrightarrow{Y}$) by each algorithm can be evaluated by the formula:

$$\rho = \frac{cov\left\{R\left(\overrightarrow{X}\right), R\left(\overrightarrow{Y}\right)\right\}}{\sigma_{R\left(\overrightarrow{X}\right)}\sigma_{R\left(\overrightarrow{Y}\right)}},$$

where $R(\overrightarrow{X})$ and $R(\overrightarrow{Y})$ are the ranks of $\overrightarrow{X}$ and $\overrightarrow{Y}$ respectively; $\sigma_{R\left(\overrightarrow{X}\right)}$ and $\sigma_{R\left(\overrightarrow{Y}\right)}$ are the standard deviations of the rank variables, and $cov\left\{R\left(\overrightarrow{X}\right), R\left(\overrightarrow{Y}\right)\right\}$ is the covariance of the rank variables.

*Pearson correlation coefficient (PCC)* measures the linear correlation between two vectors. The correlation ranges from −1 to 1, with 1 indicating a perfectly identical ranking between the two vectors, and −1 representing a fully opposed correlation. Following the above notation, the PCC is calculated as:

$$\rho = \frac{cov\left\{\overrightarrow{X}, \overrightarrow{Y}\right\}}{\sigma_{\overrightarrow{X}}\sigma_{\overrightarrow{Y}}},$$

where $\sigma_{\overrightarrow{X}}$ and $\sigma_{\overrightarrow{Y}}$ are the standard deviations, and $cov\left\{\overrightarrow{X}, \overrightarrow{Y}\right\}$ is the covariance.

*Root mean squared error (RMSE)* is employed to measure the differences between the true and estimated values. Smaller values indicate higher similarity. Following the above notation, the RMSE is applied to benchmark the isoform quantification by the formula:

$$RMSE = \sqrt{\frac{\|\overrightarrow{X} - \overrightarrow{Y}\|^2}{n}},$$

where $\|\cdot\|$ is the $l_2$-norm operator, and $n$ is the length of the vector $\overrightarrow{X}$ and $\overrightarrow{Y}$.

Normalized root mean squared error (NRMSE) is derived from RMSE, with the additional normalization term:

$$NRMSE = \frac{RMSE}{\sigma_{\overrightarrow{Y}}} = \frac{\sqrt{\frac{\|\overrightarrow{X} - \overrightarrow{Y}\|^2}{n}}}{\sigma_{\overrightarrow{Y}}},$$

where $\sigma_{\overrightarrow{Y}}$ is the standard deviation of the estimated isoform expression.

*Median of relative difference (MRD)* is used to measure the difference between true and estimated isoform expression. A small value indicates a higher similarity. MRD is calculated as:

$$MRD = median\left\{\frac{\left|\overrightarrow{X} - \overrightarrow{Y}\right|}{\overrightarrow{Y}}\right\}.$$

*–log10(averageRMSE)* is used to evaluate the agreement across multiple repeats for isoform quantification. Isoform proportions were first averaged across all the repeats. Then for each repeat, the root mean squared error (RMSE) of the estimated isoform proportions was calculated compared to this mean proportion. Ultimately, RMSE values of all the repeats were averaged and minus log10 transformed:

$$-\log_{10} average\ RMSE = -\log_{10}\frac{1}{R}\sum_{r}\sqrt{\frac{\|\overrightarrow{Y}_r - \overrightarrow{\overline{Y}}_.\|^2}{n}},$$

where $\|\cdot\|$ is the $l_2$-norm operator; $n$ is the length of the isoform proportion vector $\overrightarrow{Y}_r$; $\overrightarrow{\overline{Y}}_.$ is the mean of $\overrightarrow{Y}_r$ over $r = \{1,2,\ldots,R\}$, and $R$ is the number of repeats (e.g. four repeats for C2C12 mouse cell line or two repeats for mouse heart dataset).

*Irreproducibility measure (IM)* quantifies the consistency among multiple repeats. For isoform expression with different replicates, the IM is calculated as:

$$IM = \sqrt{\frac{\|CV(\log\left(\overrightarrow{Y}+1\right))\|^2}{n}},$$

where $CV$ is the coefficient of variation among $R$ repeats. Details were described by LRGASP [36].

*Sensitivity, Specificity, and Youden Index.* Sensitivity (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive. Specificity (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative. To benchmark novel isoform detection in the Type I-3 simulation data, the sensitivity and specificity are calculated as,

$$\text{Sensitivity} = \frac{\text{the number of isoforms that are truly novel and detected novel}}{\text{the number of isoforms that are trully novel}},$$

$$\text{Specificity} = \frac{\text{the number of isoforms that are truly normal and detected normal}}{\text{the number of isoforms that are trully normal}}.$$

To balance Sensitivity and Specificity, the Youden Index was defined as,

$$\text{Youden Index} = \text{Sensitiviety} + \text{Specificity} - 1.$$

*Precision, Recall, and F1 score.* Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. The precision and recall are defined as,

$$\text{Precision} = \frac{\text{the number of true fusions detected}}{\text{the number of fusions detected}},$$

$$\text{Recall} = \frac{\text{the number of true fusions detected}}{\text{the number of true fusions}}.$$

To balance the precision and recall, the F-measure (or F1 score) is defined as the harmonic mean of these two values,

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

*Cosine similarity* is a measure of similarity between two non-zero vectors defined in an inner product space. It is defined as the cosine of the angle between the two vectors, i.e., the dot product of the vectors divided by the product of their lengths. In the Type I-2 simulation, the cosine similarity for each software is the cosine of the angle (denoted as $\theta$) between the vector of true isoform proportion (denoted as $\overrightarrow{A}$) and the vector of estimated isoform proportion by the method (denoted as $\overrightarrow{B}$). Similarly, in the Type F-2. simulation, the cosine similarity for each method is the cosine of the angle between the vector of true fusion proportion and the vector of estimated fusion proportion by the method. The formula of cosine similarity is defined as

$$\cos(\theta) = \frac{\overrightarrow{A} \cdot \overrightarrow{B}}{\| \overrightarrow{A} \| \| \overrightarrow{B} \|},$$

where $\| \cdot \|$ is the $l_2$-norm operator.

#### Downstream analyses for real data application

Long-read RNA-seq alignment profiles and splicing junctions were visualized by the Integrative Genomics Viewer (IGV) [95]. For single-cell analysis, the isoform-level expression profiles across individual cells were imported into the R package Seurat [96] for data pre-processing, dimension reduction, visualization by the uniform manifold approximation and projection (UMAP) algorithm, and differential expression analysis. Trajectory inference at both gene and isoform resolutions were analyzed by Monocle3 [97]. Significant differentially expressed isoforms were further analyzed by the Ingenuity Pathways Analysis (IPA) and the Gene Set Enrichment Analysis (GSEA) using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases [98, 99].

#### RT-PCR for fusion validation

The procedure was described previously [100–118]. Total RNA was extracted using TRIzol to lyse the cancer tissues (Invitrogen, Inc, Carlsbad, CA). First-strand cDNA was synthesized using ~5 ng of total RNA from each sample, random hexamers, and Superscript IITM (Invitrogen, Inc) at 42 °C for 2 hours. One microliter each cDNA sample was used for TaqMan PCR reactions with 50 heat cycles at 94 °C for 30 seconds, 61 °C for 30 seconds, and 72 °C for 30 seconds using primers and probes specific for RBM6-BCAR1 (GAGTTCACTCTTGAAGATGCCA/GTCATAAAGCGCTTTGGCCAG; TaqMan probe, 5'/56-FAM/ATGCTAGAG/ZEN/GCCA ACCAGAAC/3IABkFQ/3'). A negative control and synthetic positive control were included in each batch of reactions. The threshold determined as positive for fusion transcript is Ct<45.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04023-z.

Additional file 1: Supplementary tables. Tables S1 to S12.

Additional file 2: Supplementary figures. Figs. S1 to S28.

**Acknowledgments**

We appreciate Songyang Zheng, who worked in Dr. Jianhua Luo's lab to assist with the fusion validation.

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

**Authors' contributions**
WW, JJL, YL, QL, MX, and SL developed the IFDlong pipeline, performed the simulation study and real data application, and generated the figures for data visualization. SK assisted in the biological interpretation of the HCC study. NF and MZ analyzed the isoform expression of mouse heart tissue samples. YPY and JHL performed experimental validation of the novel fusion transcripts and analyzed the biological function of the chimera protein. GCT and SL oversaw the progress of the project. WW, JJL, SK, NF, JHL, PB, GCT, and SL wrote the manuscript. All authors read and approved the final manuscript.

**Funding**
This research is supported in part by the National Institute of Health (NIH) NIGMS funding R35GM159862 (to SL), the Pittsburgh Liver Research Centre through NIH/NIDDK Digestive Disease Research Core Center grant P30DK120531 (Pilot and Feasibility Grants to SL, and Genomics and Systems Biology Core), the NIH funding UL1TR001857 (Pilot Grant to SL), the Innovation in Cancer Informatics Discovery Grant (to SL and SK), the Competitive Medical Research Fund (CMRF) of the UPMC Health System (to SL and SK), the NIH/NCI funding R01CA258449 (to SK), the NIH/NCI funding R01CA300059 (to SK), and the University of Pittsburgh Center for Research Computing through the NIH award S10OD028483.

**Data availability**
For Universal Human Reference (UHR) data, ONT data were downloaded from Gene Expression Omnibus (GEO) PRJNA639366 [12, 119], PacBio data were downloaded from https://downloads.pacbcloud.com/public/dataset/UHR_IsoSeq/ [120], and PCR data were collected from the MicroArray Quality Control (MAQC) project via GEO GSE5350 [43, 121]. Mouse C2C12 data were downloaded from the ENCODE project [44] via the GEO database with accession ID GSE219813 [122] and GSE219595 [123], and the mouse heart data were downloaded from GSE219825 [124]. Single-cell long-RNA-seq data on HCC patients were downloaded from the GEO GSE223743 [55, 125]. Human colon cancer long-read RNA-seq data were downloaded from GSE155921 [6, 126]. For MCF7, PacBio data were downloaded from the NCBI SRA database with accession ID SRP055913 [127], ONT data were downloaded from https://github.com/GoekeLab/sg-nex-data [91, 128], and the short-read RNA-seq expression profiles were downloaded from Cancer Cell Line Encyclopedia (CCLE) [92, 129]. The single-cell long-RNA-seq data on H838 sequenced by the ONT platform were downloaded from GSE154869 [93, 130]. The H1-mix datasets were downloaded from the LRGASP [36, 131].
The IFDlong software is available at GitHub (https://github.com/SilviaLiu12345/IFDlong2), which includes the source script, installation instructions, user manual, demo data, reference, and annotation files [132].
The source code used in the manuscript is available via Zenodo (https://doi.org/10.5281/zenodo.18776544) under an MIT license [133].

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 10 December 2024 Accepted: 2 March 2026
Published online: 18 March 2026

**References**

1. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13(5):278–89.
2. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16(12):1297–305.
3. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
4. Dreau A, Venu V, Avdievich E, Gaspar L, Jones FC. Genome-wide recombination map construction from single individuals using linked-read sequencing. Nat Commun. 2019;10(1):4309.
5. Hong LZ, Hong S, Wong HT, Aw PP, Cheng Y, Wilm A, et al. BAsE-seq: a method for obtaining long viral haplotypes from short sequence reads. Genome Biol. 2014;15(11):517.
6. Liu S, Wu I, Yu YP, Balamotis M, Ren B, Yehezkel T. Ben, et al. Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer. Commun Biol. 2021;4(1):506.
7. Zhang Y, Qian J, Gu C, Yang Y. Alternative splicing and cancer: a systematic review. Signal Transduct Target Ther. 2021;6(1):78.
8. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013;514(1):1–30.

9. Gonzalez E, McGraw TE. The Akt kinases: isoform specificity in metabolism and cancer. Cell Cycle. 2009;8(16):2502–8.
10. Qi W, Liu X, Qiao D, Martinez JD. Isoform-specific expression of 14-3-3 proteins in human lung cancer tissues. Int J Cancer. 2005;113(3):359–63.
11. Quinlan MP, Settleman J. Isoform-specific ras functions in development and cancer. Future Oncol. 2009;5(1):105–16.
12. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. Liqa: long-read isoform quantification and analysis. Genome Biol. 2021;22(1):182.
13. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods. 2023;20(8):1187–95.
14. Li H, Wang D, Gao Q, Tan P, Wang Y, Cai X, Li A, Zhao Y, Thurman AL, Malekpour SA, Zhang Y, Sala R, Cipriano A, Wei CL, Sebastiano V, Song C, Zhang KF. Improving gene isoform quantification with miniQuant. Nat Biotechnol. 2025;1–13.
15. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, et al. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol. 2023;41(7):915–8.
16. Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. Genome Biol. 2023;24(1):167.
17. Wyman D, Balderrama-Gutierrez G, Mortazavi A. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv. 2019;672931.
18. Tang AD, Felton C, Hrabeta-Robinson E, Volden R, Vollmers C, Brooks AN. Detecting haplotype-specific transcript variation in long reads with FLAIR2. Genome Biol. 2024;25(1):173.
19. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11(1):1438.
20. Hu X, Wang QH, Tang M, Barthel F, Amin S, Yoshihara K, et al. Tumorfusions: an integrative resource for cancer-associated transcript fusions. Nucleic Acids Res. 2018;46(D1):D1144–9.
21. Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. Nature. 2008;453(7191):110.
22. Quintas-Cardama A, Cortes J. Molecular biology of bcr-abl1-positive chronic myeloid leukemia. Blood. 2009;113(8):1619–30.
23. Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell. 2010;17(5):443–54.
24. Yu YP, Ding Y, Chen Z, Liu S, Michalopoulos A, Chen R, et al. Novel fusion transcripts associate with progressive prostate cancer. Am J Pathol. 2014;184(10):2840–9.
25. Yu YP, Liu S, Nelson J, Luo JH. Detection of fusion gene transcripts in the blood samples of prostate cancer patients. Sci Rep. 2021;11(1):16995.
26. Yu YP, Tsung A, Liu S, Nalesnick M, Geller D, Michalopoulos G, et al. Detection of fusion transcripts in the serum samples of patients with hepatocellular carcinoma. Oncotarget. 2019;10(36):3352–60.
27. Zarogoulidis P, Darwiche K, Sakkas A, Yarmus L, Huang H, Li Q, et al. Suicide gene therapy for cancer - current strategies. J Genet Syndr Gene Ther. 2013. https://doi.org/10.4172/2157-7412.1000139.
28. Chen ZH, Yu YP, Zuo ZH, Nelson JB, Michalopoulos GK, Monga S, et al. Targeting genomic rearrangements in tumor cells through Cas9-mediated insertion of a suicide gene. Nat Biotechnol. 2017;35(6):543–50.
29. Loo SK, Yates ME, Yang S, Oesterreich S, Lee AV, Wang XS. Fusion-associated carcinomas of the breast: diagnostic, prognostic, and therapeutic significance. Genes Chromosomes Cancer. 2022;61(5):261–73.
30. Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res. 2016;44(5):e47.
31. Luo JH, Liu S, Zuo ZH, Chen R, Tseng GC, Yu YP. Discovery and classification of fusion transcripts in prostate cancer and normal prostate tissue. Am J Pathol. 2015;185(7):1834–45.
32. Dorney R, Dhungel BP, Rasko JE, Hebbard L, Schmitz U. Recent advances in cancer fusion transcript detection. Brief Bioinform. 2023. https://doi.org/10.1093/bib/bbac519.
33. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, et al. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. Genome Biol. 2022;23(1):10.
34. Karaoglanoglu F, Chauve C, Hach F. Genion, an accurate tool to detect gene fusion from long transcriptomics reads. BMC Genomics. 2022;23(1):129.
35. Felton C, Tang AD, Knisbacher BA, Wu CJ, Brooks AN. Detection of alternative isoforms of gene fusions from long-read RNA-seq with FLAIR-fusion. bioRxiv. 2022: p. 2022.08.01.502364.
36. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nat Methods. 2024;21(7):1349–63.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
39. Ono Y, Asai K, Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. Bioinformatics. 2021;37(5):589–95.
40. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000;28(21):4364–75.
41. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. Nat Rev Genet. 2016;17(7):407–21.
42. Blakes AJ, Wai HA, Davies I, Moledina HE, Ruiz A, Thomas T, et al. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med. 2022;14(1):79.
43. Consortium M, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, et al. The MicroArray Quality

Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006;24(9):1151–61.
44. Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Celik MH, Rebboah E, et al. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. bioRxiv. 2023;2023.05.15.540865.
45. Opitz CA, Leake MC, Makarenko I, Benes V, Linke WA. Developmentally regulated switching of titin size alters myofibrillar stiffness in the perinatal heart. Circ Res. 2004;94(7):967–75.
46. Nagueh SF, Shah G, Wu Y, Torre-Amione G, King NM, Lahmers S, et al. Altered titin expression, myocardial stiffness, and left ventricular function in patients with dilated cardiomyopathy. Circulation. 2004;110(2):155–62.
47. Makarenko I, Opitz CA, Leake MC, Neagoe C, Kulke M, Gwathmey JK, et al. Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts. Circ Res. 2004;95(7):708–16.
48. Chauveau C, Rowell J, Ferreiro A. A rising titan: TTN review and mutation update. Hum Mutat. 2014;35(9):1046–59.
49. Mahdavi V, Lompre AM, Chambers AP, Nadal-Ginard B. Cardiac myosin heavy chain isozymic transitions during development and under pathological conditions are regulated at the level of mRNA availability. Eur Heart J. 1984;5(Suppl F):181–91.
50. Marston SB, Redwood CS. Modulation of thin filament activation by breakdown or isoform switching of thin filament proteins: physiological and pathological implications. Circ Res. 2003;93(12):1170–8.
51. Reyes Gaido OE, Nkashama LJ, Schole KL, Wang Q, Umapathi P, Mesubi OO, et al. CaMKII as a therapeutic target in cardiovascular disease. Annu Rev Pharmacol Toxicol. 2023;63:249–72.
52. Ramirez MT, Zhao XL, Schulman H, Brown JH. The nuclear deltaB isoform of Ca2+/calmodulin-dependent protein kinase II regulates atrial natriuretic factor gene expression in ventricular myocytes. J Biol Chem. 1997;272(49):31203–8.
53. Backs J, Backs T, Neef S, Kreusser MM, Lehmann LH, Patrick DM, et al. The delta isoform of CaM kinase II is required for pathological cardiac hypertrophy and remodeling after pressure overload. Proc Natl Acad Sci U S A. 2009;106(7):2342–7.
54. Xu X, Yang D, Ding JH, Wang W, Chu PH, Dalton ND, et al. ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. Cell. 2005;120(1):59–72.
55. Liu S, Yu YP, Ren BG, Ben-Yehezkel T, Obert C, Smith M, Wang W, Ostrowska A, Soto-Gutierrez A, Luo JH. Long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells. Elife. 2024;12:RP87607. https://elifesciences.org/articles/87607.
56. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42(2):293–304.
57. Liu L, Guo W, Zhang J. Association of HLA-DRB1 gene polymorphisms with hepatocellular carcinoma risk: a meta-analysis. Minerva Med. 2017;108(2):176–84.
58. De Re V, Tornesello ML, Racanelli V, Prete M, Steffan A. Non-classical HLA class 1b and hepatocellular carcinoma. Biomedicines. 2023. https://doi.org/10.3390/biomedicines11061672.
59. Catamo E, Zupin L, Crovella S, Celsi F, Segat L. Non-classical MHC-I human leukocyte antigen (HLA-G) in hepatotropic viral infections and in hepatocellular carcinoma. Hum Immunol. 2014;75(12):1225–31.
60. Lei WY, Hsiung SC, Wen SH, Hsieh CH, Chen CL, Wallace CG, et al. Total HLA class I antigen loss with the downregulation of antigen-processing machinery components in two newly established sarcomatoid hepatocellular carcinoma cell lines. J Immunol Res. 2018;2018:8363265.
61. Cicinnati VR, Shen Q, Sotiropoulos GC, Radtke A, Gerken G, Beckebaum S. Validation of putative reference genes for gene expression studies in human hepatocellular carcinoma using real-time quantitative RT-PCR. BMC Cancer. 2008;8:350.
62. Xiao N, Li K, Zhu X, Xu B, Liu X, Lei M, et al. CD74(+) macrophages are associated with favorable prognosis and immune contexture in hepatocellular carcinoma. Cancer Immunol Immunother. 2022;71(1):57–69.
63. Zhengdong A, Xiaoying X, Shuhui F, Rui L, Zehui T, Guanbin S, et al. Identification of fatty acids synthesis and metabolism-related gene signature and prediction of prognostic model in hepatocellular carcinoma. Cancer Cell Int. 2024;24(1):130.
64. Xu H, Zhao J, Li J, Zhu Z, Cui Z, Liu R. Cancer associated fibroblast-derived CCL5 promotes hepatocellular carcinoma metastasis through activating HIF1alpha/ZEB1 axis. Cell Death Dis. 2022;13(5):478.
65. Suzuki T, Matsuura K, Suzuki Y, Okumura F, Nagura Y, Sobue S, et al. Serum CXCL10 levels at the start of the second course of atezolizumab plus bevacizumab therapy predict therapeutic efficacy in patients with advanced BCLC stage C hepatocellular carcinoma: A multicenter analysis. Cancer Med. 2023. https://doi.org/10.1002/cam4.6876.
66. Kong F, Hu W, Zhou K, Wei X, Kou Y, You H, et al. Hepatitis B virus X protein promotes interleukin-7 receptor expression via NF-kappaB and Notch1 pathway to facilitate proliferation and migration of hepatitis B virus-related hepatoma cells. J Exp Clin Cancer Res. 2016;35(1):172.
67. Yin L, Zhou L, Xu R. Identification of tumor mutation burden and immune infiltrates in hepatocellular carcinoma based on multi-omics analysis. Front Mol Biosci. 2020;7:599142.
68. Li W, Li F, Zhang X, Lin HK, Xu C. Insights into the post-translational modification and its emerging role in shaping the tumor microenvironment. Signal Transduct Target Ther. 2021;6(1):422.
69. Xu H, Lin S, Zhou Z, Li D, Zhang X, Yu M, et al. New genetic and epigenetic insights into the chemokine system: the latest discoveries aiding progression toward precision medicine. Cell Mol Immunol. 2023;20(7):739–76.
70. De Zutter A, Van Damme J, Struyf S. The role of post-translational modifications of chemokines by CD26 in cancer. Cancers (Basel). 2021. https://doi.org/10.3390/cancers13174247.
71. Kron KJ, Murison A, Zhou S, Huang V, Yamaguchi TN, Shiah YJ, et al. TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nat Genet. 2017;49(9):1336–45.
72. Hu T, Li J, Long M, Wu J, Zhang Z, Xie F, et al. Detection of structural variations and fusion genes in breast cancer samples using third-generation sequencing. Frontiers in Cell and Developmental Biology. 2022;10:854640.
73. Machour FE, Abu-Zhayia ER, Awwad SW, Bidany-Mizrahi T, Meinke S, Bishara LA, et al. RBM6 splicing factor promotes homologous recombination repair of double-strand breaks and modulates sensitivity to chemotherapeutic drugs. Nucleic Acids Res. 2021;49(20):11708–27.

74. Li H, Li L, Qiu X, Zhang J, Hua Z. The interaction of CFLAR with p130Cas promotes cell migration. Biochim Biophys Acta Mol Cell Res. 2023;1870(2):119390.
75. Waters AM, Khatib TO, Papke B, Goodwin CM, Hobbs GA, Diehl JN, et al. Targeting p130Cas- and microtubule-dependent MYC regulation sensitizes pancreatic cancer to ERK MAPK inhibition. Cell Rep. 2021;35(13):109291.
76. Mao CG, Jiang SS, Shen C, Long T, Jin H, Tan QY, et al. BCAR1 promotes proliferation and cell growth in lung adenocarcinoma via upregulation of POLR2A. Thorac Cancer. 2020;11(11):3326–36.
77. Jiang S, Mao C, Jiang B, Tan Q, Deng B. High expression of BCAR1 by circulating tumor cells and tumor tissues is predictive of a poor prognosis of early-stage lung adenocarcinoma potentially due to regulation of epithelial-mesenchymal transition. Technol Cancer Res Treat. 2020;19:1533033820983086.
78. Heumann A, Heinemann N, Hube-Magg C, Lang DS, Grupp K, Kluth M, et al. High BCAR1 expression is associated with early PSA recurrence in ERG negative prostate cancer. BMC Cancer. 2018;18(1):37.
79. Sun G, Cheng SY, Chen M, Lim CJ, Pallen CJ. Protein tyrosine phosphatase alpha phosphotyrosyl-789 binds BCAR3 to position Cas for activation at integrin-mediated focal adhesions. Mol Cell Biol. 2012;32(18):3776–89.
80. Wang K, Ubriaco G, Sutherland LC. RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours. BMC Genomics. 2007;8:348.
81. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
82. Han SW, Jewell S, Thomas-Tikhonenko A, Barash Y. Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads. Genome Res. 2024;34(10):1624–35.
83. Sarantopoulou D, Brooks TG, Nayak S, Mrcela A, Lahens NF, Grant GR. Comparative evaluation of full-length isoform quantification from RNA-Seq. BMC Bioinformatics. 2021;22(1):266.
84. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
85. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011. https://doi.org/10.1186/1471-2105-12-323.
86. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. Stringtie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
87. Group, P.T.C., Calabrese C, Davidson NR, Demircioglu D, Fonseca NA, He Y, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578(7793):129–36.
88. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
89. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462-U174.
90. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417.
91. Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, et al. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. Nat Methods. 2025;22(4):801–12.
92. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8.
93. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22(1):310.
94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
95. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
96. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902 e21.
97. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82.
98. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141.
99. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
100. Zuo ZH, Yu YP, Martin A, Luo JH. Cellular stress response 1 down-regulates the expression of epidermal growth factor receptor and platelet-derived growth factor receptor through inactivation of splicing factor 3A3. Mol Carcinog. 2017;56(2):315–24.
101. He DM, Ren BG, Liu S, Tan LZ, Cieply K, Tseng G, et al. Oncogenic activity of amplified miniature chromosome maintenance 8 in human malignancies. Oncogene. 2017;36(25):3629–39.
102. Zuo ZH, Yu YP, Ding Y, Liu S, Martin A, Tseng G, et al. Oncogenic activity of miR-650 in prostate cancer is mediated by suppression of CSR1 expression. Am J Pathol. 2015;185(7):1991–9.
103. Yu YP, Ding Y, Chen R, Liao SG, Ren BG, Michalopoulos A, et al. Whole-genome methylation sequencing reveals distinct impact of differential methylations on gene transcription in prostate cancer. Am J Pathol. 2013. https://doi.org/10.1016/j.ajpath.2013.08.018.
104. Luo JH, Ding Y, Chen R, Michalopoulos G, Nelson J, Tseng G, et al. Genome-wide methylation analysis of prostate tissues reveals global methylation patterns of prostate cancer. Am J Pathol. 2013;182(6):2028–36.
105. Luo KL, Luo JH, Yu YP. (-)-Epigallocatechin-3-gallate induces Du145 prostate cancer cell death via downregulation of inhibitor of DNA binding 2, a dominant negative helix-loop-helix protein. Cancer Sci. 2010;101(3):707–12.
106. Zhu ZH, Yu YP, Shi YK, Nelson JB, Luo JH. CSR1 induces cell death through inactivation of CPSF3. Oncogene. 2009;28(1):41–51.

107. Shi YK, Yu YP, Zhu ZH, Han YC, Ren B, Nelson JB, et al. MCM7 interacts with androgen receptor. Am J Pathol. 2008;173(6):1758–67.
108. Yu YP, Yu G, Tseng G, Cieply K, Nelson J, Defrances M, et al. Glutathione peroxidase 3, deleted or methylated in prostate cancer, suppresses prostate cancer growth and metastasis. Cancer Res. 2007;67(17):8043–50.
109. Ren B, Yu YP, Tseng GC, Wu C, Chen K, Rao UN, et al. Analysis of integrin alpha7 mutations in prostate cancer, liver cancer, glioblastoma multiforme, and leiomyosarcoma. J Natl Cancer Inst. 2007;99(11):868–80.
110. Yu G, Tseng GC, Yu YP, Gavel T, Nelson J, Wells A, et al. CSR1 suppresses tumor growth and metastasis of prostate cancer. Am J Pathol. 2006;168(2):597–607.
111. Ren B, Yu G, Tseng GC, Cieply K, Gavel T, Nelson J, et al. MCM7 amplification and overexpression are associated with prostate cancer progression. Oncogene. 2006;25(7):1090–8.
112. Yu YP, Paranjpe S, Nelson J, Finkelstein S, Ren B, Kokkinakis D, et al. High throughput screening of methylation status of genes in prostate cancer using an oligonucleotide methylation array. Carcinogenesis. 2005;26(2):471–9.
113. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. J Clin Oncol. 2004;22(14):2790–9.
114. Jing L, Liu L, Yu YP, Dhir R, Acquafondada M, Landsittel D, et al. Expression of myopodin induces suppression of tumor growth and metastasis. Am J Pathol. 2004;164(5):1799–806.
115. Ren B, Yu YP, Jing L, Liu L, Michalopoulos GK, Luo JH, et al. Gene expression analysis of human soft tissue leiomyosarcomas. Hum Pathol. 2003;34(6):549–58.
116. Luo JH, Yu YP, Cieply K, Lin F, Deflavia P, Dhir R, et al. Gene expression analysis of prostate cancers. Mol Carcinog. 2002;33(1):25–35.
117. Yu YP, Lin F, Dhir R, Krill D, Becich MJ, Luo JH. Linear amplification of gene-specific cDNA ends to isolate full-length of a cDNA. Anal Biochem. 2001;292(2):297–301.
118. Lin F, Yu YP, Woods J, Cieply K, Gooding B, Finkelstein P, et al. Myopodin, a synaptopodin homologue, is frequently deleted in invasive prostate cancers. Am J Pathol. 2001;159(5):1603–12.
119. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. Oxford nanopore sequencing of acute myeloid leukemia samples. 2020. Datasets Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA640456.
120. PacBio. Universal Human Reference RNA (Agilent) + SIRV Isoform Mix E0 (Lexogen). 2019. Pacific Biosciences website. https://downloads.pacbcloud.com/public/dataset/UHR_IsoSeq/.
121. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Slikker WJ. MicroArray Quality Control (MAQC) Project. 2006. Datasets Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5350.
122. ENCODE. Long read RNA-seq from C2C12 (ENCSR418ZYU). 2022. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE219813.
123. ENCODE. Long read RNA-seq from C2C12 (ENCSR221XGR). 2022. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE219595.
124. ENCODE. Long read RNA-seq from heart (ENCSR842KTN). 2022. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE219825.
125. Liu S, Yu Y, Ren B, Yehezkel TB, Colbert C, Wang W, Ostrowska A, Soto-Gutierrez A, Luo J. Ultra-low error synthetic long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells. 2024. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE223743.
126. Liu S, Wu I, Yu Y, Balamotis M, Ren B, Yehezkel TB, Luo J. LoopSeq Synthetic Long Read Sequencing Uncovers Isoform Modulation in the progression of Colon Cancer. 2021. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155921.
127. University_of_Iowa. Full-length transcripts of the MCF-7 breast cancer cell line by PacBio SMRT sequencing. 2015. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/sra/?term=SRP055913.
128. Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, Sim A, Patel H, Low HM, Hendra C, Wratten L, Hakkaart C, Sawyer C, Iakovleva V, Lee PL, Xin L, Ng HEV, Loo JM, Ong X, Ng HQA, Wang J, Koh WQC, Poon SYP, Stanojevic D, Tran HD, Lim KHE, Toh SY, Ewels PA, Ng HH, Iyer NG, Thiery A, Chng WJ, Chen L, DasGupta R, Sikic M, Chan YS, Tan BOP, Wan Y, Tam WL, Yu Q, Khor CC, Wustefeld T, Lezhava A, Pratanwanich PN, Love MI, Goh WSS, Ng SB, Oshlack A, consortium SG-N, Goke J. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. 2025. Github. https://github.com/GoekeLab/sg-nex-data.
129. CCLE. CCLE RNAseq gene expression data for 1019 cell lines. Cancer Cell Line Encyclopedia (CCLE) 2019. https://sites.broadinstitute.org/ccle/.
130. Tian L, Jabbari JS, Seidi A, Ritchie ME. Long and short-read single cell RNA-seq profiling of human lung adenocarcinoma cell lines using 10X version 2 chemstry. 2020. Dataset Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154869.
131. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De Maria M, Adams MS, Balderrama-Gutierrez G, Behera AK, Gonzalez Martinez JM, Hunt T, Lagarde J, Liang CE, Li H, Meade MJ, Moraga Amador DA, Prjibelski AD, Birol I, Bostan H, Brooks AM, Celik MH, Chen Y, Du MRM, Felton C, Goke J, Hafezqorani S, Herwig R, Kawaji H, Lee J, Li JL, et al. Long-read RGASP. 2024. ENCODE. https://www.gencodegenes.org/pages/LRGASP/.
132. Wang W, Liu JJ, Li Y, Ko S, Feng N, Zhang M, Lin Q, Xia M, Yu YP, Luo JH, Baldoni P, Tseng GC, Liu S. IFDlong: a model-based isoform and fusion detector for accurate annotation and quantification of long-read RNA-seq data. 2026. Github. https://github.com/SilviaLiu12345/IFDlong2.
133. Wang W, Liu JJ, Li Y, Ko S, Feng N, Zhang M, et al. IFDlong: a model-based isoform and fusion detector for accurate annotation and quantification of long-read RNA-seq data. 2026. Zenodo. https://doi.org/10.5281/zenodo.18776544.


## Publisher's Note