Detection of alternative isoforms of gene fusions from long-read RNA-seq with FLAIR-fusion

Colette Felton[1], Alison D Tang[1], Binyamin A Knisbacher[2,3], Catherine J Wu[2,4,5], Angela N Brooks[1]*

[1]University of California Santa Cruz, Department of Biomolecular Engineering, Santa Cruz, CA, USA [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA [3]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, [4]Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, USA, [5]Brigham and Women's Hospital, Harvard Medical School, Department of Medicine, Boston, MA, USA
*correspondence: anbrooks@ucsc.edu

Abstract

Gene fusions are important cancer drivers and drug targets, but are difficult to reliably identify with short-read RNA-sequencing. Long-read RNA sequencing data are more likely to span a fusion breakpoint and provide more sequence context around the breakpoint. This allows for more reliable identification of gene fusions and for detecting alternative splicing in gene fusions. Notably, alternative splicing of fusions has been shown to be a mechanism for drug resistance and altered levels of oncogenicity. Here, we present FLAIR-fusion, a computational tool to identify gene fusions and their isoforms from long-read RNA-sequencing data. FLAIR-fusion can detect fusions and their isoforms with high precision and recall, even with error-prone reads. We also investigated different library preparation methods and found that direct-cDNA has a higher incidence of artifactual chimeras than direct-RNA and PCR-cDNA methods. FLAIR-fusion is able to filter these technical artifacts from all of these library prep methods and consistently identify known fusions and their isoforms across cell lines. We ran FLAIR-fusion on amplicon sequencing from multiple tumor samples and cell lines and detected alternative splicing in the previously validated fusion *GUCYA2-PIWIL4,* which shows that long-read sequencing can detect novel splicing events from cancer gene panels. We also detect fusion isoforms from long-read sequencing in chronic lymphocytic leukemias with the splicing factor mutation *SF3B1 K700E*, and find that up to 10% of gene fusions had more than one unique isoform. We also compared long-read fusion detection tools with short-read fusion detection tools on the same samples and found greater consensus in the long-read tools. Our results demonstrate that gene fusion isoforms can be effectively detected from long-read RNA-sequencing and are important in the characterization of the full complexity of cancer transcriptomes.

INTRODUCTION

Gene fusions are major somatic alterations with many established functions in multiple cancer types[1,2].Generally, gene fusions result from major translocations or deletions where two previously separate genes are fused together and expressed under a single promoter[1,3]. Previous work has shown that gene fusions are major drivers of about 16% of cancers and function as the sole driver in more than 1% of all cancers[4]. Many cancer-driving fusions contain known oncogenes which make them more virulent, while others contain exclusively genes oncogenic only within the fusion. An example of an oncogene-containing fusion is the *BCR-ABL1* fusion which is primarily found in chronic myeloid leukemia (CML) and contains the *ABL1* gene, which encodes a kinase involved in cell growth and proliferation. Usually this gene is turned off, but when fused to the BCR gene, it is always turned on under the *BCR* promoter. This results in a strong proliferative cancer-promoting phenotype[5].

Gene fusions can also be important diagnostic and prognostic markers in cancer and can be very good targets for treatment. Fusions are generally absent in healthy tissues making them a reliable marker for early detection of cancer[6]. Specific fusions are also recurrent in certain cancer types and can act as markers for the cancer type and severity[7]. Since they produce very unique chimeric proteins, they are also promising antigens for targeted therapies. For instance, there is a highly effective treatment targeted to the *BCR-ABL1* fusion[8,9].

Gene fusions can be detected either at the DNA or RNA level through whole genome or whole transcriptome sequencing[10,11,12,13]. Specific fusions can also be detected through more rapid and targeted PCR methods[14]. While most high-throughput sequencing has been done with short-read Illumina sequencing, it remains difficult to separate gene fusions from mapping artifacts (reads mapping to multiple loci due to similar sequence) as the short length of the reads means that the length of mappings in chimeric reads is especially short and a very small fraction of reads at a locus will span across the fusion breakpoint[15,16]. There are also short-read DNA sequencing tools that use *de novo* assembly techniques and even diploid assembly[17,18]. These tools have high false negative rates and unknown false positive rates[19,20]. They also only report structural variation and do not help researchers understand the functional impact of the mutation of the tumor. Consequently, running multiple short-read fusion detection tools on a single sequencing dataset will generally yield completely different sets of predicted fusions. Some of these tools are used in clinical research today, but the standard is currently to run multiple tools and look for fusions shared between their outputs, as well as doing extensive human curation and lab testing[21].

Long-read sequencing techniques have been used to improve gene fusion detection[22,23]. Long-read sequencing reads average at about 1,000 base pairs with

either the Pacific Biosciences (PacBio) or the Oxford Nanopore Technologies (ONT) platform and ONT can yield reads as long as 2 megabases[24]. They are better than short reads for this application because their length allows for a larger fraction of reads that span the fusion breakpoint, as well as being able to span repetitive and otherwise problematic sequences. There are a couple of tools that focus on fusion detection from long-read DNA sequencing, such as Sniffles and NanoSV[25,26]. They have already been shown to have higher accuracy than short-read techniques. However, they are also more expensive, as they use de novo assembly techniques and require very high read coverage. There are also multiple tools for identifying fusions from long-read RNA-seq including LongGF[27], JAFFAL[28], AERON[,29], FusionSeeker[44], and Genion[47].

Another advantage of long-read sequencing is that it provides more context around the fusion breakpoint. This is particularly important in transcriptome sequencing as it allows researchers to detect additional information about the gene fusion such as alternative splicing. Abnormal alternative splicing has been shown to contribute to tumor progression and can be especially impacted by splicing factor mutations[30]. Multiple tools exist to detect alternative splicing and full-length isoforms from long-read sequencing, something that cannot be done with short reads. Tools that do this include Bambu[53], FLAIR[31], ESPRESSO[45], IsoSeq[46], IsoQuant[48], Mandalorion[49], and TALON[50]. Detecting the full-length expressed isoform of a gene fusion allows for better protein product prediction and understanding of the functional output of the fusion. This additional information could allow researchers to better understand drug resistance in gene fusions. For instance, alternatively spliced isoforms of the *BCR-ABL1* fusion have been shown to create a protein product, which has been identified as a new target for therapies[33,34]. Splicing factor mutations are also common in certain cancer types, which makes it important to be able to understand differential or aberrant splicing of gene fusions and their effect on a tumor. When the splicing factor U2AF1 is mutated at the S34F site, it has been shown to cause alternative splicing in the *SLC34A2-ROS1* fusion, leading to upregulation of a more oncogenic isoform[35].

Here, we present FLAIR-fusion, a tool for a more comprehensive identification of gene fusions from long-read RNA-seq, which also identifies gene fusion isoforms. We compare FLAIR-fusion with existing long-read fusion tools on simulated and cancer cell line sequencing data and find the FLAIR-fusion gives better or comparable performance for gene fusion detection, while FLAIR-fusion is the only tool to report gene fusion isoforms. Surprisingly, we find many chimeric reads, representing true gene fusions and artifacts, even in PCR-free library preparation methods (direct CDNA and direct RNA) which challenge assumptions on causes of chimeric artifacts. Applying FLAIR-fusion to primary cancer samples, we find evidence of up to 10% of gene fusions having more than one isoform in a primary chronic lymphocytic leukemia (CLL) sample with a splicing factor mutation. FLAIR-fusion allows for high-confident and systematic gene

fusion isoform detection from long-read sequencing, thus allowing cancer researchers to better characterize gene fusion products and their potential clinical implications.

RESULTS

FLAIR-fusion was built to utilize the isoform detection capabilities of FLAIR[31]; however, FLAIR-fusion is amenable to the usage of other transcript-isoform detection methods such as StringTie2[32]. First, fastq or fasta reads are run through FLAIR-align and FLAIR-correct, which align using minimap2[54] and correct small deviations from splice sites in the alignment. Next, FLAIR-fusion is run, which first adds transcriptome annotation to the genomic alignment, then extracts all multi-mapping reads from the alignment. A subset of multi-mapping reads will be chimeric reads where sub-sequences will map to multiple genomic positions (**Figure 1**, **top panel**). For chimeric reads mapping to non-genic regions, close mappings are condensed into a single locus and filtered to ensure that they are farther apart than a user-defined distance parameter. The default read support that FLAIR-fusion requires is 3, although most analyses in this paper were run with a read support of 2 to match other tools. Potential chimeras are then filtered by read support and mapping score. Since the mitochondrial genome exists outside of the nucleus and therefore is unlikely to biologically fuse to the nuclear genome, any chimeras including mitochondrial genes are excluded. Chimeras containing known paralogous genes are excluded since those are likely due to mismapping to other paralogs.

Next, fusion breakpoints are determined by first referring to the position of the end of the genomic mappings on the fastq read, ensuring that the inferred breakpoint is where the read switches mapping to a separate genomic locus. Second, possible breakpoints are clustered by read support, with the breakpoint with the highest read support being selected for each locus. The mappings are checked for proximity to the start site of the gene, which is evidence for the fusion being expressed and is later used to score the fusion. Next, chimeric reads with mappings closer in genomic distance than the user-defined distance are filtered out. Chimeric reads are further filtered by the arrangement of mappings on the fastq read, where the portions of the read that are mapped to different regions of the genome should be arranged directly adjacent to each other on the read. This filter removes regions that multimap due to sequence similarity and also library preparation artifacts causing central adapters. The chimeric reads are then filtered by proximity to splice sites, since gene fusion breakpoints are more likely to occur in introns and are represented in RNA as being at splice sites (**Figure 1, middle panel**).

Finally, FLAIR-fusion filters by both gene coverage and fastq read coverage. For the gene coverage filter, we check that the mapping doesn't cover the entirety of the gene as these chimeric reads were likely due to cDNA or library preparation artifacts by visual inspection. For the read coverage filter we check that the combined chimeric

mappings cover the majority of the read. Chimeric reads that pass all filters are categorized as gene fusions.

Next, FLAIR-fusion detects the fusion isoforms by first splitting the fusion mappings into two files, with those mapping to one locus of a fusion in one file and those mapping to the other locus in a second file. FLAIR-collapse is then run on each file independently, with a file of reads attached to the isoforms they support being generated. FLAIR-fusion then matches the reads supporting each isoform of the fusion at each locus to each other. For instance, if a group of reads support both isoform A and B at locus 1 as well as isoform C at locus 2, FLAIR-fusion would report the full-length isoforms A-C and B-C. FLAIR-fusion generates five primary outputs - two .tsv files with the predicted fusions and predicted isoforms, two .bed files with the fusion reads and collapsed isoforms, and a .txt file with the chimeras that were filtered out.

Performance on simulated data

To test the ability of FLAIR-fusion to correctly identify gene fusions, we ran it on data simulated to closely match biological gene fusions. This dataset included simulating multiple isoforms of both loci involved in the gene fusion. For each fusion, two protein-coding genes were randomly selected from the GENCODE v37 annotations[38]. For each of these genes a breakpoint was randomly selected within the gene, then each isoform of one gene was fused to a unique isoform of its fusion partner. This mode of simulation allows us to explore both gene fusions and their isoforms. For each simulated sample, we simulated up to 50 gene fusions with background gene expression of 6000 randomly selected protein-coding genes.

We then used Badread[39] to simulate long-reads reads at different coverage and quality levels comparable to published Nanopore sequencing data (**Supplemental Table 1)**, then ran FLAIR-fusion on the resulting fastqs. Badread simulates read identities from a beta distribution, but using mean, max, and standard deviation parameters. The mean, max, and standard deviation parameters for high, average, and low quality reads were (95,100,4), (87.5,97.5,5), and (75,90,8), respectively. We analyzed 4 replicates each of 3 different quality levels and 4 levels of coverage. A fusion was considered detected if the genes in the reported fusion were identical to the genes selected in the simulation process. An isoform was considered correctly identified if the intron chain present in the fusion was identical to that portion of the intron chain present in the reference.

FLAIR-fusion was found to have >80% recall and >70% precision on high-quality and medium quality reads at all coverage levels (**Figure 2A**). While FLAIR-fusion maintained high precision in the low-quality read samples, it has low recall, primarily due to low mapping quality and fidelity. All of the fusions that were missed in the higher-quality samples were also due to low mapping quality. Some fusions, especially those with only a single exon at one locus, didn't reliably have chimeric alignments even in

high coverage samples, suggesting that improvements in alignment software could recover novel gene fusions.

FLAIR-fusion's isoform-level recall and precision followed similar trends, with 60% recall and precision in high-quality and medium-quality reads but lower recall in low-quality reads (**Figure 2B**). However, in medium to high-quality reads, FLAIR-fusion was able to correctly identify up to 6 unique isoforms in a single fusion, following the isoform across the fusion breakpoint. The isoforms that were missed were ones that have been previously shown to be difficult to detect, such as shorter isoforms that contain a subset of exons of a longer fusion[31].

We also compared FLAIR-fusion to two other tools for long-read fusion detection, JAFFAL[28] and LongGF[27]. Since neither of these tools is able to identify fusion isoforms, we only compared based on fusion detection. We ran all tools requiring a minimum fusion support of two reads. No other parameters were standardized due to the lack of user-defined parameters in JAFFAL and LongGF. All of the tools showed an effect of read coverage on fusion recall in low-quality reads (**Figure 2C-D, Supplemental Table 2**). On the simulated dataset, FLAIR-fusion outperformed both JAFFAL ($p<0.001$) and LongGF ($p<0.001$) at multiple coverage and read quality levels (**Figure 2C-D, Supplemental Table 3**).

Detecting clinically-relevant fusion isoforms on amplicon sequencing data

To illustrate the importance of detecting isoforms of gene fusions, we ran FLAIR-fusion on a Nanopore cDNA amplicon sequencing dataset targeting fusions from multiple tumor types and cell lines[40]. This is an ideal dataset to detect alternative isoforms given the increased read coverage. These fusions have been shown to be important cancer drivers or markers of pathogenicity. All of these fusions were experimentally validated by PCR and their breakpoints validated by Sanger sequencing[40]. FLAIR-fusion is able to correctly identify all of these important cancer fusions and identify their breakpoints, but in addition, was able to identify alternatively spliced isoforms in *GUCY1A2-PIWIL4, EFHD1-UBR3* and *ERGIC2-CHRNA6*. However, some of this splicing was at low abundance (< 5% of total fusion reads), so we focused on the alternative splicing in *GUCY1A2-PIWIL4*. In this fusion there is an alternative isoform that makes up 13% of reads with a skipped exon in *PIWIL4*. PIWIL4 has been previously identified as a pro-migratory and anti-apoptotic factor in breast cancer that would make a good therapeutic target[36]. Understanding the expressed isoforms of this gene in tumors will allow for better designed therapies that cannot be evaded by alternative splicing.

Effect of library preparation method on chimeras and fusion detection

To investigate the ability of FLAIR-fusion to detect known fusions in whole-transcriptome data, we used Nanopore sequencing data generated from 5 cancer cell

lines of different tissue types with well-characterized gene fusions[41]. This dataset also allowed us to investigate the impact of library preparation method on fusion detection as they used cDNA, direct-cDNA, and direct-RNA (dRNA) approaches. This is important because the more common cDNA library preparation method uses PCR amplification, which can introduce artifactual chimeras through template switching[42]. Both direct-cDNA and dRNA do not use PCR amplification, although they still use reverse transcription. Understanding which library preparation method is best for detecting fusions with long-read sequencing is important for the wider adoption of this technique in the future.

We hypothesized that the cDNA samples would have the highest levels of chimeras due to PCR. We found instead that direct-cDNA had approximately 10 times higher levels of chimeras compared to cDNA and directRNA ($p<0.0001$, **Figure 3A**). Direct-RNA and cDNA had similar levels of chimeras contrary to our hypothesis. This suggests that PCR amplification artifacts are not the only or most significant source of chimeras, as previously thought. We then investigated the population composition of the chimeras by analyzing which FLAIR-fusion filters the chimeras failed (**Figure 3B**). We found that the direct-RNA and cDNA chimera populations were generally similar, while the direct-cDNA had significantly less chimeras with only one supporting read ($p<0.0001$), more chimeras with bad mapping quality ($p<0.0001$), less chimeras with unaligned portions of the read or overlapping alignments ($p<0.0001$), and more chimeras with short genomic distance between the loci ($p<0.0001$). Having more reads per chimera is consistent with the direct-cDNA having more chimeric reads overall, while the differences in the other filters indicate that these chimeras are likely due to library preparation errors and not alignment errors.

Direct-cDNA also has longer reads than direct-RNA and cDNA with a lower median percentage of the read aligned (direct-cDNA 64%, cDNA 79%, direct-RNA 94%; **Supplementary Figure 1**). This both reflects the higher incidence of chimeric alignments in direct-cDNA and may suggest that the additional length of reads in direct-cDNA may be the direct result of errors in the library preparation protocol that create chimeras. We next investigated whether the increase in chimeras in direct-cDNA was due to annealing of polyA tails by analyzing the frequency of polyA and polyT stretches in the center of chimeric reads. We found that direct-RNA had very low (0.1%) incidence of central polyA and polyT stretches in chimeric reads and that both direct-cDNA and cDNA had significantly ($p<0.0001$) higher levels of these events (~15% of chimeric reads, **Supplementary Figure 2**). The low overall levels of these events indicates that they do not explain the majority of chimeras.

We found one event type in direct-cDNA chimeras that was not present in direct-RNA or cDNA, which was chimeras with both the positive and negative strand sequence for two genes (**Supplementary Figure 3**). These events also tend to have central polyA or polyT sequence, indicating that a combination of different primer design in the cDNA

library prep that allows for internal priming and lack of successful splitting of $1D^2$ reads may be causing the increase in chimeric transcripts in the direct-cDNA.

We also found that the number of chimeric reads aligned to a gene strongly correlates with total gene expression consistently across all library prep methods, indicating that overall the formation of chimeras is a stochastic process not primarily driven by misalignment due to sequence similarity ($r > 0.9$, $p < 0.001$, **Supplementary Figure 4**).

Using a dataset of fusions in these cell lines that had been previously validated by at least two publications[28], we compared the recall of known fusions from sequencing with different library prep methods and found no differences in the recall, indicating that the higher levels of artifactual chimeras in direct-cDNA do not inhibit fusion detection using our method (**Figure 3C)**. We chose not to analyze precision on this set as there is a lack of a true negative set due to the possibility of real, but not previously validated, fusions.

Given that we did not detect all known fusions in these samples, we investigated what depth of sequencing it would take to reach saturation in true fusion detection by generating a saturation curve based on the MCF7 cDNA sequencing data (**Supplementary Figure 5**, **Methods**). We found that based on this data, we estimate that it would take >1 billion reads to approach saturation of detected true fusions. We then compared our results for Nanopore cDNA sequencing to a single sample of PacBio sequencing of MCF7[51]. 20 true fusions were detected in this sample, 4 times greater than the predicted fusion detection in Nanopore data with this sequencing depth. This indicates that with greater base-level accuracy, more fusions can be detected from less data.

We also compared the recall of FLAIR-fusion to JAFFAL and LongGF on this dataset, running the tools as described above with standardized fusion support of two reads. We found no significant difference between the performance of these tools on any of the cell lines sequenced, which have a range of 1-53 experimentally validated fusions (**Supplementary Figure 6**). While the tools have comparable recall of the experimentally validated fusions, the other fusions they detect in each sample vary. To better understand this variability between tools, we first identified what fraction of the set of putative fusions that JAFFAL and LongGF report is also detected but filtered out by FLAIR-fusion (**Figure 3D**). We found that FLAIR-fusion detects almost all of the fusions called by LongGF but only 80% of those detected by JAFFAL. FLAIR-fusion and LongGF both use genomic alignment to identify chimeras while JAFFAL uses transcriptomic alignment. This suggests that there are additional chimeras that can only be identified through transcriptomic alignment. We then took the set of putative fusions that JAFFAL and LongGF report, but FLAIR-fusion filters out, and identified at what filtering steps FLAIR-fusion removed them (**Figure 3E**). We found that most chimeras are filtered out by the fastq distance filter (~75%), indicating how important it is to

ensure that chimeric alignments are not composed of similar sequence and don't contain central unaligned sequence.

FLAIR-fusion can also be used to detect full-length isoforms of gene fusions from this cell line data. It reports multiple isoforms in the well-characterized fusion BCR-ABL1 in the K562 cell line (**Supplementary Figure 7**). These are consistent with the isoforms that have previously been reported through manual annotation[41]. These results demonstrate that FLAIR-fusion can be a useful tool for researchers interested in profiling cancer-associated fusions.

Detection of gene fusion isoforms in whole-transcriptome primary cancer samples

To assess the ability of FLAIR-fusion to detect alternative splicing in gene fusions in primary cancer samples, we applied it to 6 CLL samples that were sequenced with Nanopore using a cDNA library preparation[31]. Three normal B-cell samples were also sequenced as a control. Three of the CLL samples had the *SF3B1 K700E* hotspot mutation which has been shown to cause transcriptome-wide changes in splicing[31]. We hypothesized that we would detect alternative splicing in gene fusions detected in *K700E* samples. Two CLL samples of each genotype were sequenced with Nanopore MinION, while one sample was sequenced using the higher throughput Nanopore PromethION. The samples sequenced with the MinION had an average of 0.5M reads, which was not enough coverage to detect fusions. Therefore, all further analysis was performed on the samples sequenced on a PromethION, which had an average of 50M reads (CW95 and CLL043/CW109). These samples were also sequenced with short-read Illumina sequencing and two of the leading short-read fusion detection tools, STAR-fusion[55] and Arriba[56], were run on these samples. First, we compared the differences in fusion detection between all short-read and long-read tools on the CLL *SF3B1* WT sample. We found that all tools except JAFFAL detect a similar total number of fusions, with JAFFAL detecting an order of magnitude more fusions than other tools (**Figure 4A**). However, the overlap between fusions detected in short-read tools is two orders of magnitude smaller than the total fusions detected by those tools, while the majority of fusions detected by the long-read tools are in common between all three tools (**Figure 4A, red box; 4B**). All tools detect one fusion, *BIRC3–REXO2*.

Of the fusions detected by FLAIR-fusion in the *SF3B1 K700E* tumor, 10% had >1 unique isoform, compared to 5% in the *SF3B1 WT* tumor (**Figure 4C-D**). In addition, none of the *SF3B1 WT* fusions had alternative splicing in both of the genes involved in the fusion, while 2.5% of fusions in the *SF3B1 K700E* tumor had alternative splicing in both loci. One of these fusions, *MAX–CHURC1*, has one isoform with the retention of exon 2 in *MAX* and the skipping of exon 3 in *CHURC1*, while the other isoform has the skipping of exon 1 in *MAX* and the retention of exon 3 in *CHURC1* (**Figure 4E**). These could not be deconvoluted by short reads because of the lack of reads spanning the full transcript. Another fusion with alternative splicing in the *SF3B1 K700E* sample is the

SPIDR–UBE2V2 fusion. In the *SPIDR* gene, one isoform appears to have the chimeric junction after exon 4, while the other isoform has the chimeric junction after exon 6. There is an annotated isoform that skips exons 5 and 6, which we also find support for non-fusion transcripts of *SPIDR*. The isoform structure of the *SPIDR-UBE2V2* fusion can be representative of an underlying genomic breakpoint in the intron after exon 6 with alternative splicing causing different chimeric junctions in the RNA. (**Figure 4F**). Again, this is a structure that can be uniquely identified and understood with long-read sequencing and FLAIR-fusion.

Discussion

Long-read sequencing provides much longer reads and therefore more context around fusion breakpoints. While some tools have used long-reads for fusion detection, none have fully taken advantage of the ability to detect both gene fusions and their full-length isoforms at the same time, allowing for a more complete functional interpretation of the fusion. We developed FLAIR-fusion, a tool for the detection of gene fusions and their isoforms from long-read RNA-sequencing data. This tool is able to do splice site correction of all reads, gather chimeric reads, and then apply a number of specific filters to reduce fusion calls caused by artifacts. It then identifies the isoforms at each locus involved in a fusion, then combines those to identify full-length fusion isoforms matched across the fusion breakpoint. We benchmarked FLAIR-fusion using real and simulated Nanopore sequencing data with lower sequence accuracy. Other long-read transcriptome data with higher sequencing accuracy such as R2C2-cDNA/Nanopore or cDNA/PacBio would be expected to have increased accuracy (**Figure 2, Supplementary Figure 5**)[37].

Using simulated reads, we were able to show that FLAIR-fusion is able to detect both gene fusions and their full-length isoforms with high sensitivity and precision. On the simulated dataset, FLAIR-fusion also outperformed two other long-read fusion detection methods, JAFFAL and LongGF. We also used FLAIR-fusion to analyze amplicon sequencing of multiple previously identified fusions. FLAIR-fusion detected all expected fusions and detected alternative splicing in the *GUCY1A2-PIWIL4* fusion. Understanding alternative splicing of known fusions in cancer is important to more sensitively diagnose and profile these cancers.

We determined that most chimeras are likely not formed via PCR artifacts, as dRNA and direct-cDNA samples that were prepared without PCR showed similar or greater numbers of chimeras compared to cDNA samples prepared with PCR. We also found that the expression of a gene is highly correlated to the number of chimeric reads, which suggests that the process creating these artifacts is not due to specific repetitive sequence in the reads. We also found substantive differences in both the number and type of chimeras seen in direct-cDNA, indicating that this library prep method may not be optimal for those looking to identify gene fusions from RNA. Through comparison of

Nanopore cDNA sequencing and PacBio cDNA sequencing, we also found that the reduced error rate of PacBio sequencing allows for greater fusion detection with fewer reads, as well as a lack of saturation of fusion detection with less-accurate sequences at sequencing depth less than 15 million (**Supplementary Table 3**). This indicates that as long-read sequencing technology continues to improve, we will be able to continue to detect more gene fusions more accurately.

We also showed that tools for fusion detection from long-read data have much more consensus than tools using short-read data, showing the robustness of fusion identification from long-reads. We detected interesting differences in the alternative splicing of fusions between an *SF3B1 WT* sample and an *SF3B1 K700E* sample, suggesting that *SF3B1 K700E* may cause more alternative splicing of fusions. Finally, we identified complex alternative isoforms of gene fusions in a *SF3B1 K700E* leukemia sample, indicating that the *SF3B1 K700E* mutation impacts the splicing of gene fusions. These structures could only be resolved by long-read sequencing and FLAIR-fusion, which is more evidence that the combination of FLAIR-fusion and long-read sequencing is uniquely useful to better characterize fusions in primary tumor samples. More sequencing and research needs to be done to fully understand how splicing mutations impact gene fusion mutations and how this could impact disease outcomes.

METHODS

<u>FLAIR-fusion pipeline:</u> FLAIR-fusion is a python tool for fusion detection from long reads. There is the option of starting from a FASTQ file or a BAM file, although if starting from a mapped BAM file, the alignment must be run with –secondary=yes to allow for chimeric mappings. If starting from a FASTQ file, the pipeline first runs the FLAIR-align module, which runs minimap2[54] with the desired options. It then converts the minimap2 BAM file to a BED file, as the default BED file that minimap2 produces doesn't include chimeric reads. The FLAIR-correct module is then run on the BED file, which corrects splice sites with an error of a few base pairs. The corrected reads are then compared to the transcriptome and each read assigned to the correct gene. Next, all reads that map multiple times are extracted from the BED file. Paralogous mappings are identified and moved to the metadata file, then fusions involving non-genic regions are grouped and collapsed. Next, we read through the SAM file and identify how the mappings are located on the FASTQ read. If the mappings overlap or are too far apart, the chimera is moved to the metadata file. Other filters include:

- chimeras involving mitochondrial genes
- fraction of fastq read covered by alignments (default: >80%)
- genomic distance between alignments (default: 5kbp)
- shortest distance from alignment to start of known isoform (default: 300bp)

    - Alignments cannot be the second half of two genes - there needs to be proximity to a promoter for the chimera to be reasonably expected to be expressed
- distance to splice site of breakpoint (default: <5bp)
    - This filter compares against splice sites from Intropolis v1[52].
- fraction alignment covers gene (default <.95%)

Chimeras that pass all filters are classified as gene fusions. Their reads are written to the prefixReads.bed file, and the fusions are written to the prefixFusions.bed file. If the user does not specify -ij, FLAIR-fusion next identifies gene isoforms in the gene fusions by first separating the prefixReads.bed into two files, separating the different mapping loci for each fusion. Each file is then individually run through FLAIR-collapse using the –generate-map option, which collapses the reads at each locus into isoforms and produces a file that associates reads with isoforms. Next, using that file, FLAIR-fusion matches the isoforms detected at each locus involved in a fusion to the other locus, generating full length isoforms across the fusion breakpoint. FLAIR-fusion also has many options for running and adding/removing filters that can be found at https://github.com/cafelton/FLAIR-fusion. For this paper, we used FLAIR-v1.5, although any version can be used with FLAIR-fusion.

Fusion simulations: First, 50 protein coding genes are randomly selected from the GENCODE 37 annotations and all isoforms of those genes are retrieved. Next, a random breakpoint is generated in each gene and the gene is matched with a random partner from the set. The minimum total isoforms of the two genes in the pair is identified, then that minimum number of fusion isoforms is generated from the gene pair. Once all simulated fusions are generated, 6000 other random genes are selected from the annotation and added to the simulated reference transcriptome. This code can also be found in the FLAIR-fusion Github at makeHumanFusions.py That simulated reference transcriptome is then run via Badreads with different coverage levels and qualities as follows:

High-quality nanopore reads:
```
badread simulate --reference ref.fasta --quantity 50x --error_model random \
  --qscore_model ideal --glitches 0,0,0 --junk_reads 0 --random_reads 0 \
  --chimeras 0 --identity 95,100,4 --start_adapter_seq "" --end_adapter_seq "" \
  | gzip > reads.fastq.gz
```
Medium-quality nanopore reads:
```
badread simulate --reference ref.fasta --quantity 50x \
  | gzip > reads.fastq.gz
```
Bad-quality nanopore reads:
```
badread simulate --reference ref.fasta --quantity 50x --glitches 1000,100,100 \
  --junk_reads 5 --random_reads 5 --chimeras 10 --identity 75,90,8 \
  | gzip > reads.fastq.gz
```

These parameters were previously defined by the Badreads team and can be found on their GitHub at https://github.com/rrwick/Badread.

MCF7 fusion detection saturation analysis: We combined 3 PCR-cDNA runs for the MCF7 cell line for a total of ~16M reads, then took 3 random subsamples for each desired sequencing depth (2.5M, 5M, 10M, 15M). We then independently aligned each subsample to GRCh38 and identified all chimeric reads from the sam file with no filtering. We then compared these reads to a list of fusions that had been previously validated by at least two publications[28] and identified which fusions were found in our data using two different thresholds (3 & 5 reads). We next plotted the data and generated a logarithmic trendline.

Tool comparisons: For tool comparison on both the simulated and cell line data, we ran FLAIR-fusion with default settings except for k=2, which sets the minimum reads covering a fusion to two and sets it on par with the other tools. For LongGF, minimap2[54] was run with the same parameters as used in FLAIR-align, then LongGF was run with the command:
LongGF sorted.bam gencode.v37.annotation.gtf 100 50 200
JAFFA version 2.1 was run on all files with the command:
<path to JAFFA>/tools/bin/bpipe run <path to JAFFA>/JAFFAL.groovy fastq.gz
For the JAFFAL output, fusions reported as PotentialTransSplicing are excluded from the analysis to maintain a minimum read support level of 2 across all tools. Fusions were classed as true positives if both gene loci detected were correct. We also allowed fusions mapping to paralogous loci to be classed as true positives. We reported breakpoints on the fusion gene level, so any alternative fusion breakpoints (specifically from JAFFAL) were not counted separately.

Analyzing chimeric reads: To understand the makeup of chimeras (Figure 3 B&E), we ran all FLAIR-fusion filters concurrently on all chimeras and chimeric reads. This approach is distinct from the normal FLAIR-fusion approach as FLAIR-fusion increases efficiency by filtering out chimeras by stages.

Sample sequence access: SGNex ONT cell line sequencing is available at https://github.com/GoekeLab/sg-nex-data. CLL patient data was sequenced as described in Tang et al [31].

Short-read RNA-seq analysis: Short-read RNA-seq was processed in Terra (https://app.terra.bio/) using the GTEx V7 pipeline (https://github.com/broadinstitute/gtex-pipeline), which uses STAR (v2.6.1d) to perform alignment to hg19 (b37) using the GENCODE v19 annotation. For Gene fusion

detection, STAR-fusion v1.8.0[55] and Arriba v1.0.1[56] were run with default parameters using GENCODE v19 annotation and the Arriba blacklist dated 2018-11-04. All Arriba fusion calls were kept for further analysis, including low confidence calls.

AUTHOR CONTRIBUTIONS

C.F. and A.N.B. designed the study. B.A.K. performed short-read analysis of fusions. A.D.T. updated FLAIR to work with fusion detection and worked with C.F. on the CLL data. C.F. wrote code for FLAIR-fusion and did all other analyses. C.F., B.A.K., C.J.W., and A.N.B. interpreted the data. C.F. and A.N.B. wrote the manuscript with input from all other co-authors.

ACKNOWLEDGEMENTS

This work was funded by NIH R01HG010053 and R35GM138122 (to A.N.B.). This work was also funded by P01CA206978 and R01CA155010 (to C.J.W.).

CODE AVAILABILTY

Code is available at https://github.com/cafelton/FLAIR-fusion.

COMPETING INTERESTS

A.N.B. is a consultant for Remix Therapeutics, Inc.

REFERENCES

1) Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res*. 2016;44(10):4487-4503. doi:10.1093/nar/gkw282
2) Taniue, K., & Akimitsu, N. (2021). Fusion Genes and RNAs in Cancer Development. *Non-Coding RNA*, *7*(1). https://doi.org/10.3390/ncrna7010010
3) Varley KE, Gertz J, Roberts BS, et al. Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Res Treat*. 2014;146(2):287-297. doi:10.1007/s10549-014-3019-2
4) Gao Q, Liang WW, Foltz SM, et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep*. 2018;23(1):227-238.e3. doi:10.1016/j.celrep.2018.03.050
5) CA Westbrook, AL Hooberman, C Spino, RK Dodge, RA Larson, F Davey, DH Wurster- Hill, RE Sobol, C Schiffer, CD Bloomfield; Clinical significance of the BCR-ABL fusion gene in adult acute lymphoblastic leukemia: a Cancer and Leukemia Group B Study (8762). *Blood* 1992; 80 (12): 2983–2990. doi: https://doi.org/10.1182/blood.V80.12.2983.2983

6) Heyer, E.E., Deveson, I.W., Wooi, D. *et al.* Diagnosis of fusion genes using targeted RNA sequencing. *Nat Commun* 10, 1388 (2019). https://doi.org/10.1038/s41467-019-09374-9
7) Vellichirammal,NN et al. Pan-Cancer Analysis Reveals the Diverse Landscape of Novel Sense and Antisense Fusion Transcripts. *Molecular Therapy - Nucleic Acids* 19, 1379-1398 (2020). https://doi.org/10.1016/j.omtn.2020.01.023.
8) Druker, B., Tamura, S., Buchdunger, E. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells. *Nat Med* 2, 561–566 (1996). https://doi.org/10.1038/nm0596-561
9) Wong, S and Witte, ON. The BCR-ABL Story: Bench to Bedside and Back. *Annual Review of Immunology* 22:1, 247-306 (2004). https://doi.org/10.1146/annurev.immunol.22.012703.104753
10) Li, Y., Roberts, N.D., Wala, J.A. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 112–121 (2020). https://doi.org/10.1038/s41586-019-1913-9
11) Zhang, Jin et al. “INTEGRATE: gene fusion discovery using whole genome and transcriptome data.” *Genome research* vol. 26,1 (2016): 108-18. doi:10.1101/gr.186114.114
12) Luo, Jian-Hua et al. “Discovery and Classification of Fusion Transcripts in Prostate Cancer and Normal Prostate Tissue.” *The American journal of pathology* vol. 185,7 (2015): 1834-45. doi:10.1016/j.ajpath.2015.03.008
13) Kim CY, Na K, Park S, Jeong SK, Cho JY, Shin H, Lee MJ, Han G, Paik YK. FusionPro, a Versatile Proteogenomic Tool for Identification of Novel Fusion Transcripts and Their Potential Translation Products in Cancer Cells. Mol Cell Proteomics. 2019 Aug;18(8):1651-1668. doi: 10.1074/mcp.RA119.001456.
14) Tong, Y.-Q., Zhao, Z.-J., Liu, B., Bao, A.-Y., Zheng, H.-Y., Gu, J., Xia, Y., McGrath, M., Dovat, S., Song, C.-H., & Li, Y. (2018). New rapid method to detect BCR-ABL fusion genes with multiplex RT-qPCR in one-tube at a time. *Leukemia Research*, *69*, 47–53.
15) Tian, L., Li, Y., Edmonson, M.N. *et al.* CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol* 21, 126 (2020). https://doi.org/10.1186/s13059-020-02043-x
16) Kumar, Shailesh et al. “Identifying fusion transcripts using next generation sequencing.” *Wiley interdisciplinary reviews. RNA* vol. 7,6 (2016): 811-823. doi:10.1002/wrna.1382
17) Haas, B.J., Dobin, A., Li, B. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* 20, 213 (2019). https://doi.org/10.1186/s13059-019-1842-9

18) Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018 Oct 22:10.1038/nbt.4277. doi: 10.1038/nbt.4277.
19) Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240.
20) Khayat, M. M., Sahraeian, S. M. E., Zarate, S., Carroll, A., Hong, H., Pan, B., Shi, L., Gibbs, R. A., Mohiyuddin, M., Zheng, Y., & Sedlazeck, F. J. (2021). Hidden biases in germline structural variant detection. *Genome Biology*, *22*(1), 347.
21) Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U. H., Neumann, O., Stenzinger, A., Scholl, C., Fröhling, S., & Brors, B. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Research*, *31*(3), 448–460.
22) Davidson, N.M., Majewski, I.J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med* 7, 43 (2015). https://doi.org/10.1186/s13073-015-0167-x
23) Jason L. Weirather, Pegah Tootoonchi Afshar, Tyson A. Clark, Elizabeth Tseng, Linda S. Powers, Jason G. Underwood, Joseph Zabner, Jonas Korlach, Wing Hung Wong, Kin Fai Au, Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing, *Nucleic Acids Research*, Volume 43, Issue 18, 15 October 2015, Page e116, https://doi.org/10.1093/nar/gkv562
24) Soneson, C., Yao, Y., Bratus-Neuenschwander, A. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* 10, 3359 (2019).
25) Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018 Jun;15(6):461-468. doi: 10.1038/s41592-018-0001-7.
26) Cretu Stancu, M., van Roosmalen, M.J., Renkens, I. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* 8, 1326 (2017). https://doi.org/10.1038/s41467-017-01343-4
27) Liu, Q., Hu, Y., Stucky, A., Fang, L., Zhong, J. F., & Wang, K. (2020). LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics*, *21*(Suppl 11), 793.
28) Davidson, N., et al. JAFFAL: Detecting fusion genes with long read transcriptome sequencing. *bioRxiv* (2021). https://doi.org/10.1101/2021.04.26.441398

29) Rautiainen, M., et al. AERON: Transcript quantification and gene-fusion detection using long reads. *bioRxiv* (2020). https://doi.org/10.1101/2020.01.27.921338
30) Zhang, Y., Qian, J., Gu, C. *et al.* Alternative splicing and cancer: a systematic review. *Sig Transduct Target Ther* 6, 78 (2021). https://doi.org/10.1038/s41392-021-00486-7
31) Tang, A.D., Soulette, C.M., van Baren, M.J. *et al.* Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 11, 1438 (2020). https://doi.org/10.1038/s41467-020-15171-6
32) Kovaka, S., Zimin, A.V., Pertea, G.M. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* 20, 278 (2019). https://doi.org/10.1186/s13059-019-1910-1
33) Volpe G, Cignetti A, Panuzzo C, Kuka M, Vitaggio K, Brancaccio M, Perrone G, Rinaldi M, Prato G, Fava M, Geuna M, Pautasso M, Casnici C, Signori E, Tonon G, Tarone G, Marelli O, Fazio VM, Saglio G. Alternative BCR/ABL splice variants in Philadelphia chromosome-positive leukemias result in novel tumor-specific fusion proteins that may represent potential targets for immunotherapy approaches. Cancer Res. 2007 Jun 1;67(11):5300-7. doi: 10.1158/0008-5472.CAN-06-3737.
34) Neckles, C, Sundara Rajan, S, Caplen, NJ. Fusion transcripts: Unexploited vulnerabilities in cancer? *WIREs RNA*. 2020; 11:e1562. https://doi.org/10.1002/wrna.1562
35) Esfahani, M.S., Lee, L.J., Jeon, YJ. *et al.* Functional significance of U2AF1 S34F mutations in lung adenocarcinomas. *Nat Commun* 10, 5712 (2019). https://doi.org/10.1038/s41467-019-13392-y
36) Wang, Z., Liu, N., Shi, S., Liu, S., & Lin, H. (2016). The Role of PIWIL4, an Argonaute Family Protein, in Breast Cancer. *The Journal of Biological Chemistry*, *291*(20), 10646–10658.
37) Francisco Pardo-Palacios, Fairlie Reese, Silvia Carbonell-Sala et al (2021). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification, PREPRINT available at Research Square https://doi.org/10.21203/rs.3.rs-777702/v1
38) Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923.
39) Wick RR. Badread: simulation of error-prone long reads. Journal of Open Source Software. 2019;4(36):1316. doi:10.21105/joss.01316.

40) Suzuki, A., Suzuki, M., Mizushima-Sugano, J., Frith, M. C., Makalowski, W., Kohno, T., Sugano, S., Tsuchihara, K., & Suzuki, Y. (2017). Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *24*(6), 585–596.
41) Chen, Ying, et al. "A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines." *bioRxiv* (2021). doi: https://doi.org/10.1101/2021.04.21.440736
42) Wang, G. C. Y., & Wang, Y. (1996). The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*, *142 (Pt 5)*, 1107–1114.
43) White, R., Pellefigues, C., Ronchese, F., Lamiable, O., & Eccles, D. (2017). Investigation of chimeric reads using the MinION. *F1000Research*, *6*, 631.
44) Chen, Y., Wang, Y., Chen, W., Tan, Z., Song, Y., Human Genome Structural Variation Consortium, Chen, H., & Chong, Z. (2023). Gene Fusion Detection and Characterization in Long-Read Cancer Transcriptome Sequencing Data with FusionSeeker. *Cancer Research*, *83*(1), 28–33.
45) Gao, Y., Wang, F., Wang, R., Kutschera, E., Xu, Y., Xie, S., Wang, Y., Kadash-Edmondson, K. E., Lin, L., & Xing, Y. (2023). ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Science Advances*, *9*(3), eabq5072.
46) *IsoSeq: IsoSeq3 - Scalable De Novo Isoform Discovery from Single-Molecule PacBio Reads*. (n.d.). Github. Retrieved May 31, 2023, from https://github.com/PacificBiosciences/IsoSeq
47) Karaoglanoglu, F., Chauve, C., & Hach, F. (2022). Genion, an accurate tool to detect gene fusion from long transcriptomics reads. *BMC Genomics*, *23*(1), 129.
48) Prjibelski, A. D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A. L., & Tilgner, H. U. (2023). Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01565-y
49) Volden, R., Schimke, K., Byrne, A., Dubocanin, D., Adams, M., & Vollmers, C. (2022). Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. In *bioRxiv* (p. 2022.06.29.498139). https://doi.org/10.1101/2022.06.29.498139
50) Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., England, W., Chu, S.-H., Spitale, R. C., Tenner, A. J., Wold, B. J., & Mortazavi, A. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. In *bioRxiv* (p. 672931). https://doi.org/10.1101/672931
51) *IsoSeq Human MCF7 Transcriptome*. (n.d.). Github. Retrieved May 31, 2023, from https://github.com/PacificBiosciences/DevNet/wiki/IsoSeq-Human-MCF7-Transcriptome
52) Nellore, A., Jaffe, A. E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips, R. A., III, Karbhari, N., Hansen, K. D., Langmead, B., & Leek, J. T. (2016).

Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology*, *17*(1), 266.

53) Chen, Y., Sim, A., Wan, Y. K., Yeo, K., Lee, J. J. X., Ling, M. H., Love, M. I., & Göke, J. (2022). Context-Aware Transcript Quantification from Long Read RNA-Seq data with Bambu. In *bioRxiv* (p. 2022.11.14.516358). https://doi.org/10.1101/2022.11.14.516358
54) Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics , 34(18), 3094–3100.
55) Haas, B. J., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T. G., Pochet, N., Sun, J., Wu, C. J., Gingeras, T. R., & Regev, A. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. In *bioRxiv* (p. 120295). https://doi.org/10.1101/120295
56) Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U. H., Neumann, O., Stenzinger, A., Scholl, C., Fröhling, S., & Brors, B. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Research*, *31*(3), 448–460.

SUPPLEMENTARY FILES

Supplementary File 1: Fusion simulation files. Zipped folder containing four simulated reads.fasta files and the matching truth.tsv files with the true fusions used for simulation. These reads files are simulated transcriptomes that were then passed through Badreads as described in Methods to simulate nanopore error models.

Supplementary File 2: FLAIR-fusion results on the amplicon sequencing data from Suzuki et al. The first tab describes the fusions identified and the second tab has the isoforms identified in .bed format. The isoform names in the bed file are formatted as follows:
FusionGene1--FusionGene2-.-readsSupportingIsoform|combIso[IsoformID]|[double:aligns to both genes/single:aligns to only one side of fusion]-.-FusionGene1/geneDir/geneStart/geneEnd
There are two bed lines for each isoform that aligns to both sides of a fusion, these two lines will have the same Isoform ID number.

Supplementary File 3: FLAIR-fusion results from cancer cell lines. Results are on all replicates of all SGNex cell lines (from https://github.com/GoekeLab/sg-nex-data). Formatting is as follows:

- Column 1: [cell line]-[library prep method]-replicate[n]run[n]
- Column 2:[fusion gene 1]--[fusion gene 2]
- Column 3: number of reads supporting fusion
- Column 4: average mapping score as a decimal (0-1, higher is better)
- Column 5: sequence agreement at fusion breakpoint (0-1, higher is better)
- Column 6: fraction of reads aligning to fusion genes that align to the fusion, averaged across both genes
- Column 7: 3'-[fusion gene 1]-chr[x]-[breakpoint position]-[fraction of reads at this locus aligning to fusion]
- Column 7: 5'-[fusion gene 2]-chr[x]-[breakpoint position]-[fraction of reads at this locus aligning to fusion]

Supplementary File 4: FLAIR-fusion results on the CLL sequencing data. The first tab is the B-cell control data, the second tab is SF3B1 WT, and the third tab is SF3B1 K700E CLL. Columns are as described, although column 2 is now: [# isoforms aligning to both fusion loci]/[total # isoforms supported by fusion reads]/[# reads supporting isoforms aligning to fusion]/[total # reads supporting fusion]

Supplementary File 5: Short-read sequencing fusion calling results on the SF3B1 WT CLL. The first tab is the STAR-fusion results and the second is the Arriba results. The columns are as described in the file header.

Figure 1: FLAIR-Fusion pipeline for fusion isoform detection

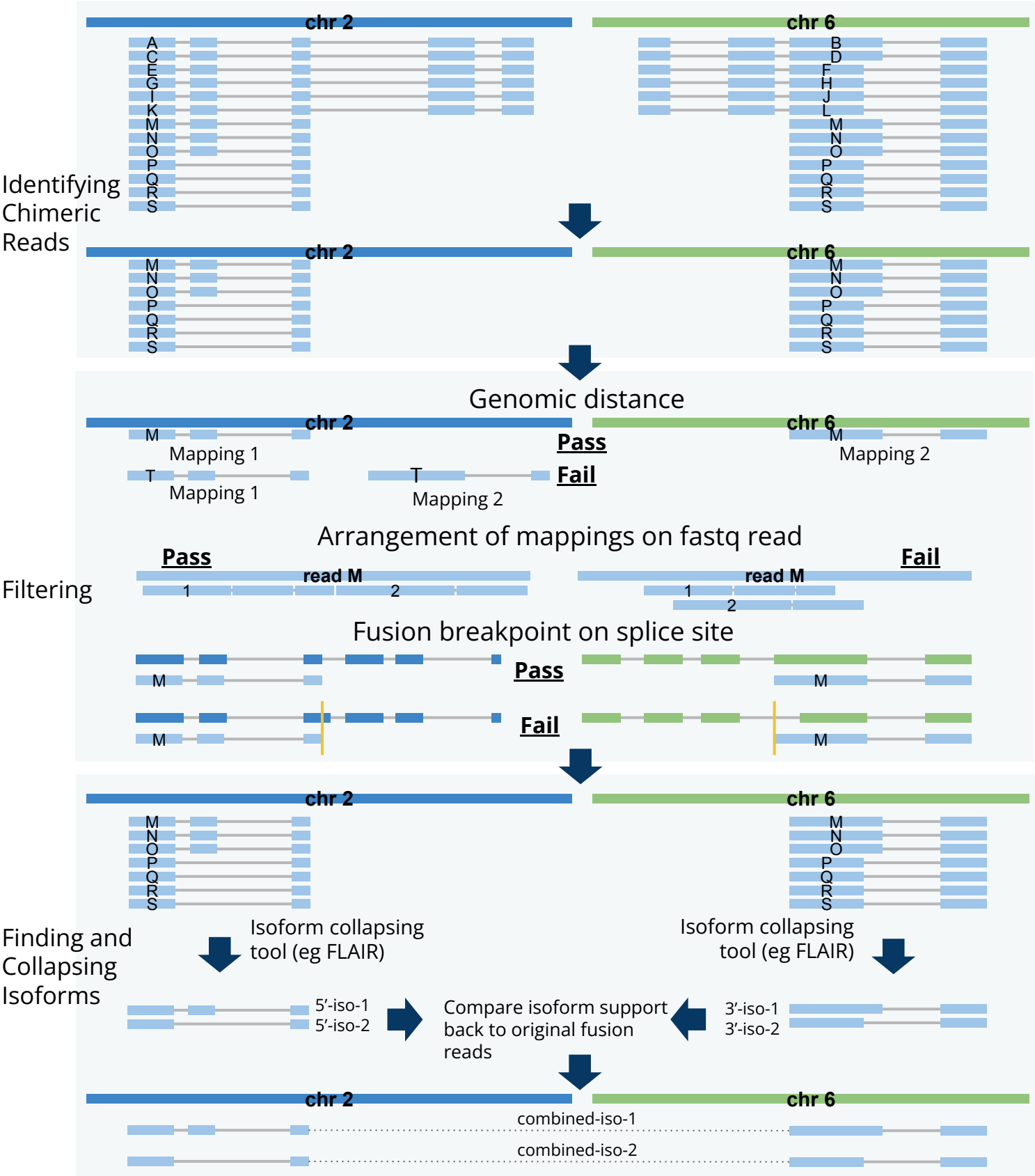


Figure 1: After reads are mapped and splice sites are corrected using FLAIR-align and FLAIR-correct, reads that map to multiple loci are identified (top panel). Next, multiple filters are applied to separate mapping or library preparation errors from true fusions. A subset of key filters are shown: ensuring genomic distance between mappings, checking that the mappings don't include overlapping sequence, and checking that the breakpoint between the mappings is at a splice site (middle panel). Finally, isoforms are identified separately for each locus in a fusion and then combined to create full-length gene fusion isoforms (bottom panel).

Figure 2: FLAIR-Fusion outperforms other methods on simulated data

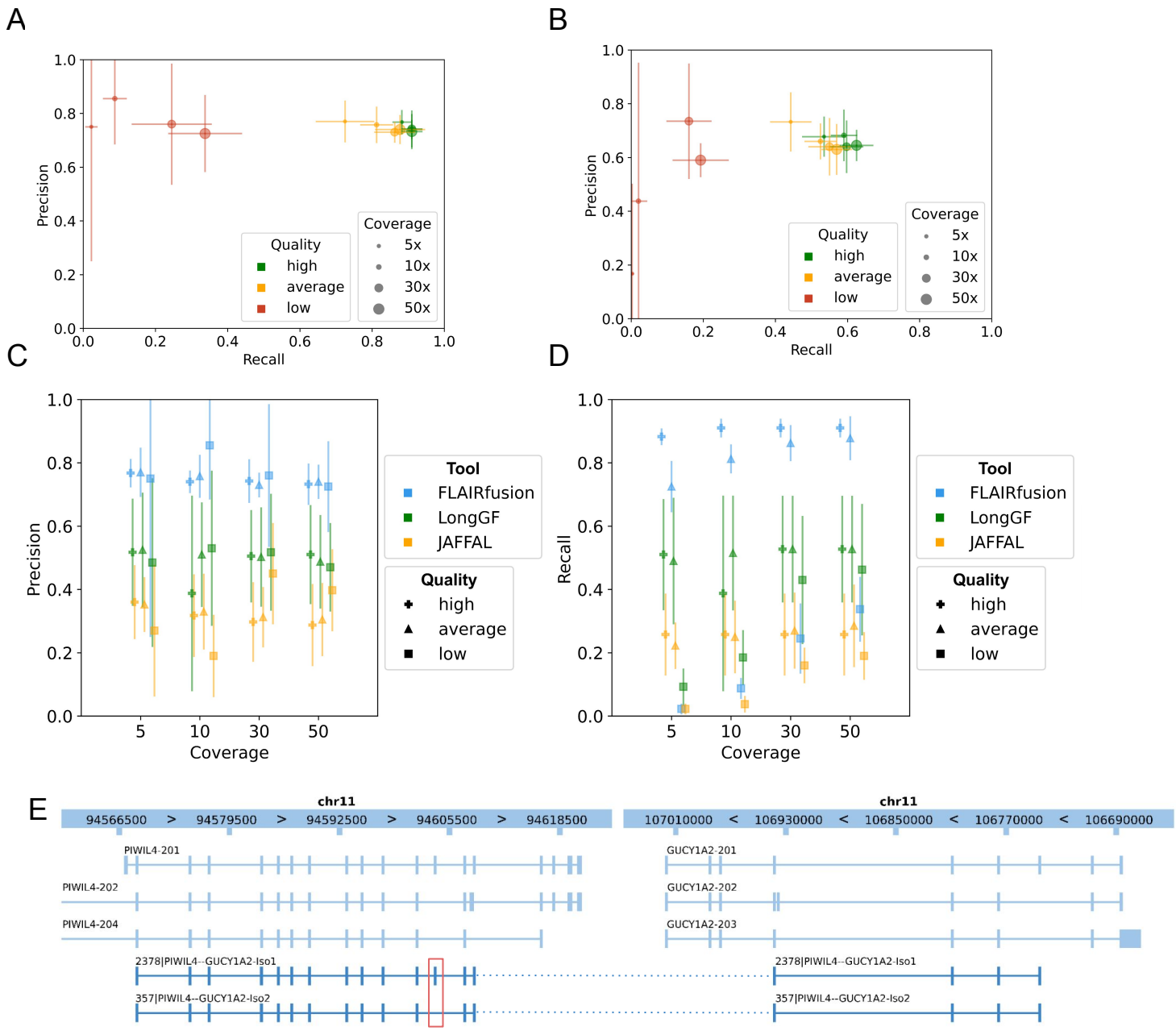


Figure 2: FLAIR-Fusion outperforms other methods on simulated data A+B Precision (true positives found(TP)/total fusions found) and recall (TP/total set of true fusions) of FLAIR-fusion for multiple coverage levels and simulated read qualities (see Methods), n=4 unique simulated transcriptomes. **A** fusion-level, **B** isoform level, with up to 10 simulated isoforms per fusion. **C** Precision and **D** recall of FLAIR-fusion, JAFFAL, and LongGF on the same simulated dataset as A+B. Almost all comparisons between tools are significant, for values see Supplementary Table 1. **E** Alignment of the fusion isoforms of the amplicon-sequenced PIWIL4-GUCY1A2 fusion. The first number in the fusion isoform label is the number of supporting reads for that isoform. A selection of the annotated isoforms of these genes is also shown with HUGO isoform IDs from gencode 38. Note that there is an inversion between these loci.

Figure 3: Library preparation method has no effect on chimeras found

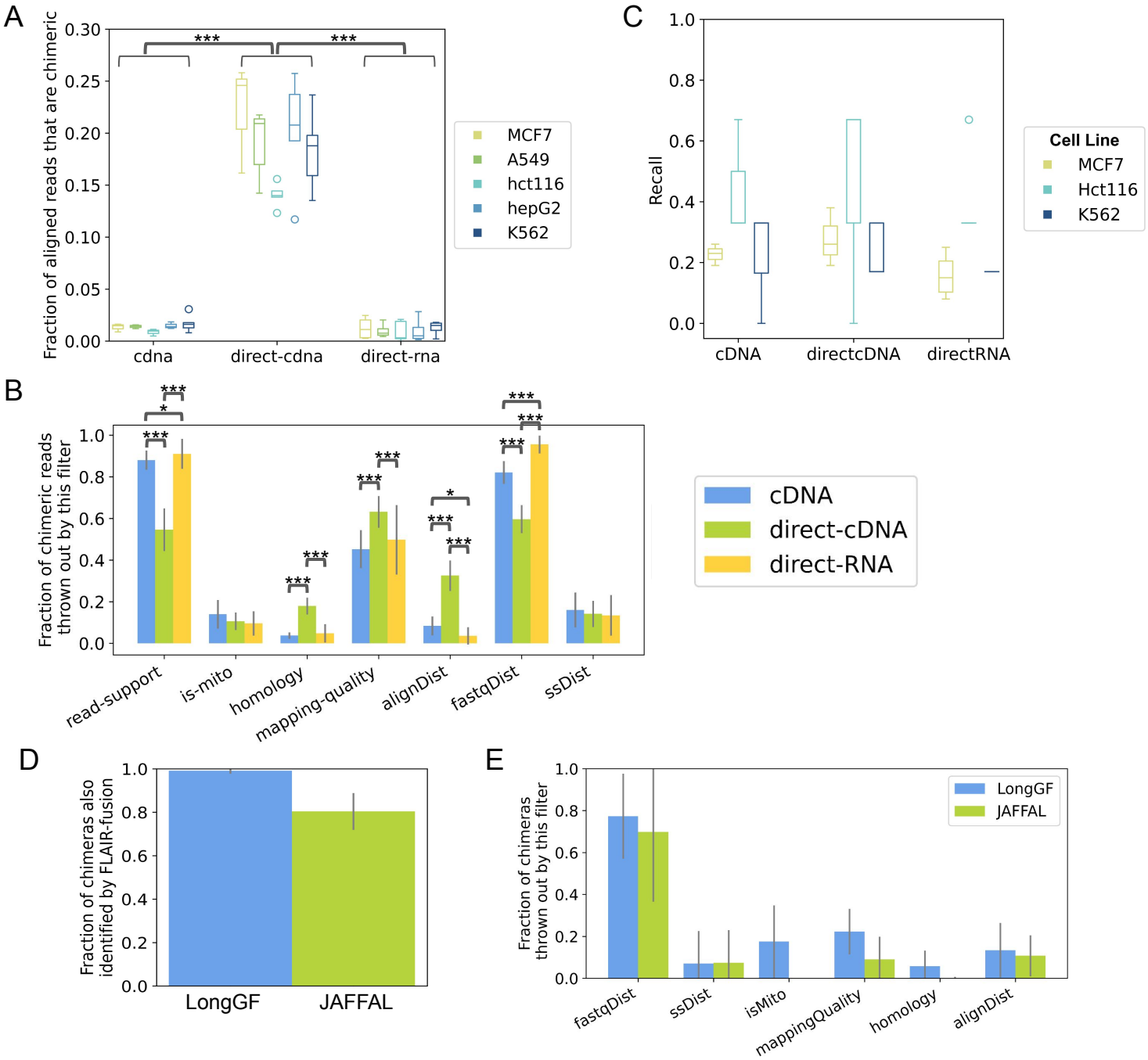


Figure 3: A Fraction of aligned reads that are chimeric (align to more than one distinct genic locus) in ONT sequencing of MCF7, A549, Hct116, HepG2, and K562 cell lines with cDNA, direct-cDNA, and direct-RNA library preparation methods. **B** Comparison of different FLAIR-fusion filters for removing chimeras to identify differences in sources of chimeras between library preparation methods. This is only for the MCF-7 cell line. A single chimera can be thrown out by multiple filters. **C** Recall of FLAIR-Fusion on MCF7(n known fusions = 53), K562 (n=6), and Hct116 (n=2) cell lines. HCT-115 and HEP-G2 were excluded due to lack of known fusions. No significant difference in fusion detection based on library prep method for each cell line was found. **D** Fraction of LongGF and JAFFAL chimeras in MCF7 also found by FLAIR-fusion. This compares the different base chimera detection methods. **E** For the fusions called in MCF7 by LongGF and JAFFAL that are also found as chimeras by FLAIR-fusion, which fraction of these are thrown out by which FLAIR-fusion filter. A single chimera can be thrown out by multiple filters. *-p<0.05, ***-p<0.00001

Figure 4: FLAIR-fusion detects alternative splicing in gene fusions in CLL SF3B1 K700E tumor samples

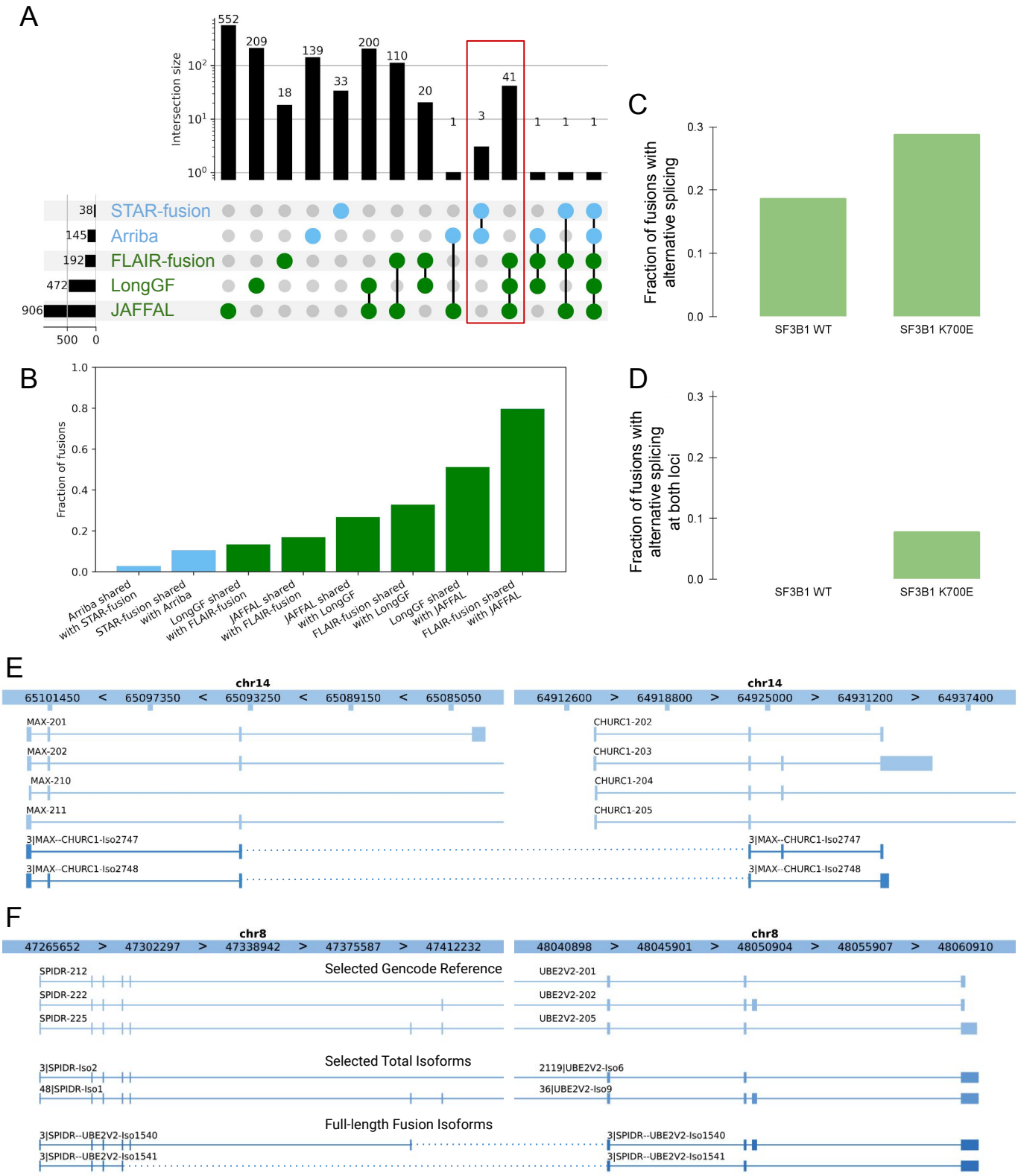
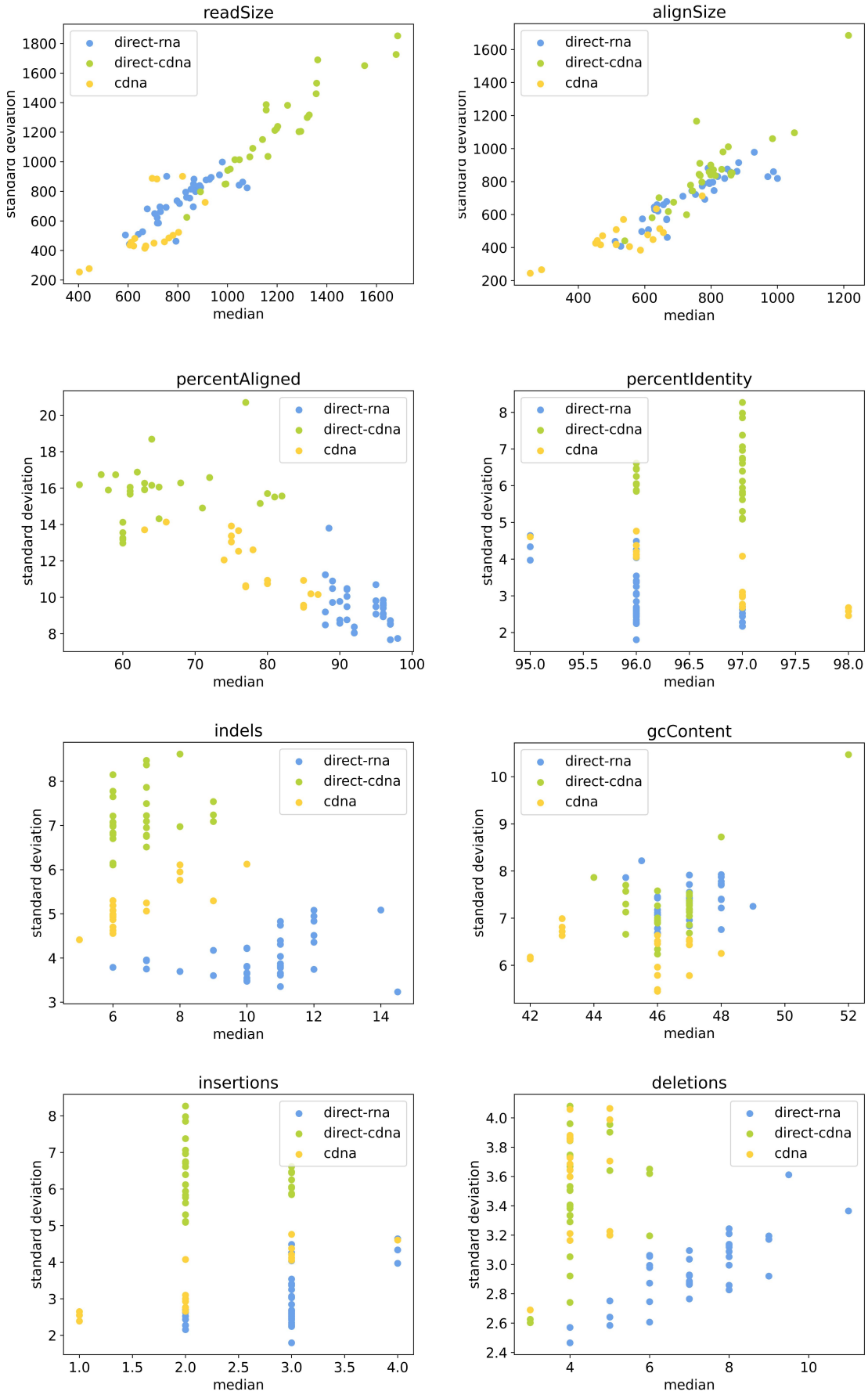
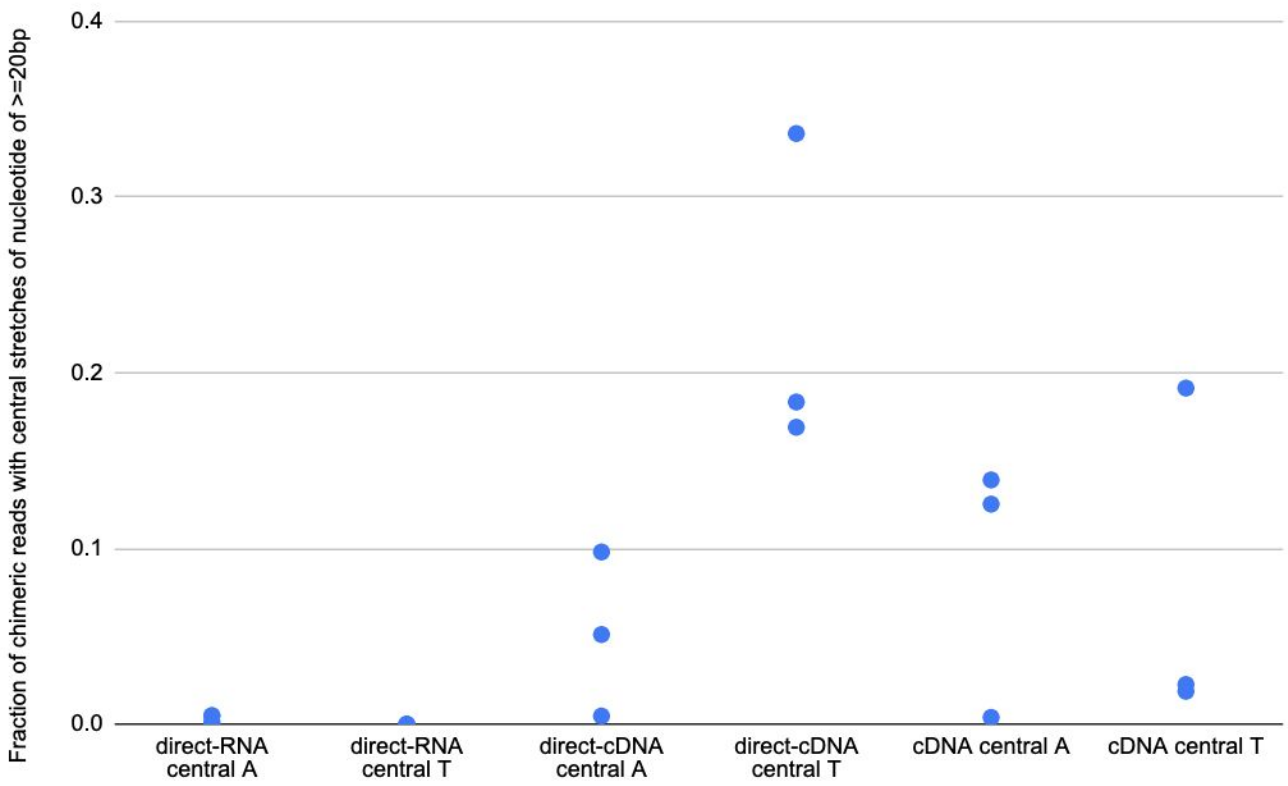


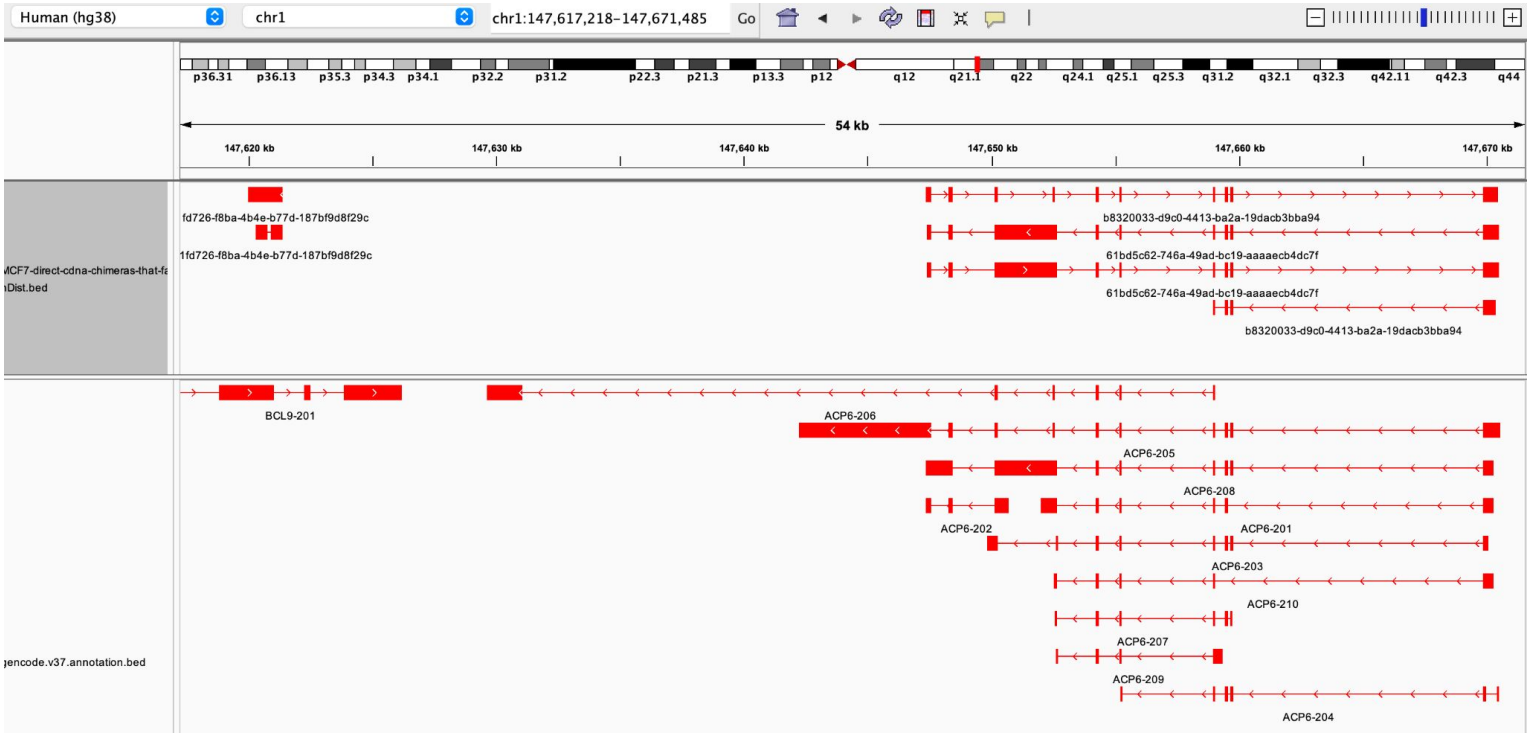
Figure 4: A Upset plot comparing performance of short-read tools STAR-fusion and Arriba and long-read tools FLAIR-fusion, LongGF, and JAFFAL on a CLL K700E WT sample. The red box highlights the greater overlap of the long-read tools compared to the short-read tools. **B** Shows fraction of fusions shared between each pair of short-read and long-read fusion detection tools. **C** Comparison of the fraction of gene fusions detected by FLAIR-fusion with alternative splicing at at least one of the fusion loci in SF3B1 WT to SF3B1 K700E. **D** Same as B, but there must be unique alternative splicing at both fusion loci. **E** Alignment of the fusion isoforms of the MAX-CHURC1 fusion in the SF3B1 K700E sample. A selection of the annotated isoforms of these genes is also shown. Note that there is an inversion between these loci. **F** Alignment of the SPIDR-UBE2V2 fusion with a selection of annotated isoforms.



Supplemental Figure 1: Read quality and alignment statistics obtained from running https://github.com/BrooksLabUCSC/labtools/blob/master/samAlignStats.py. This is all independent replicates direct-RNA, direct-cDNA, and PCR-cDNA sequencing of the cell lines MCF7, A549, K562, HepG2, and Hct116.



Supplemental Figure 2: Fraction of chimeric reads with central stretches of nucleotide (A or T) of >=20bp in direct-RNA (n=4), direct-cDNA (n=3), and cDNA (n=3) sequencing of the MCF7 cell line.

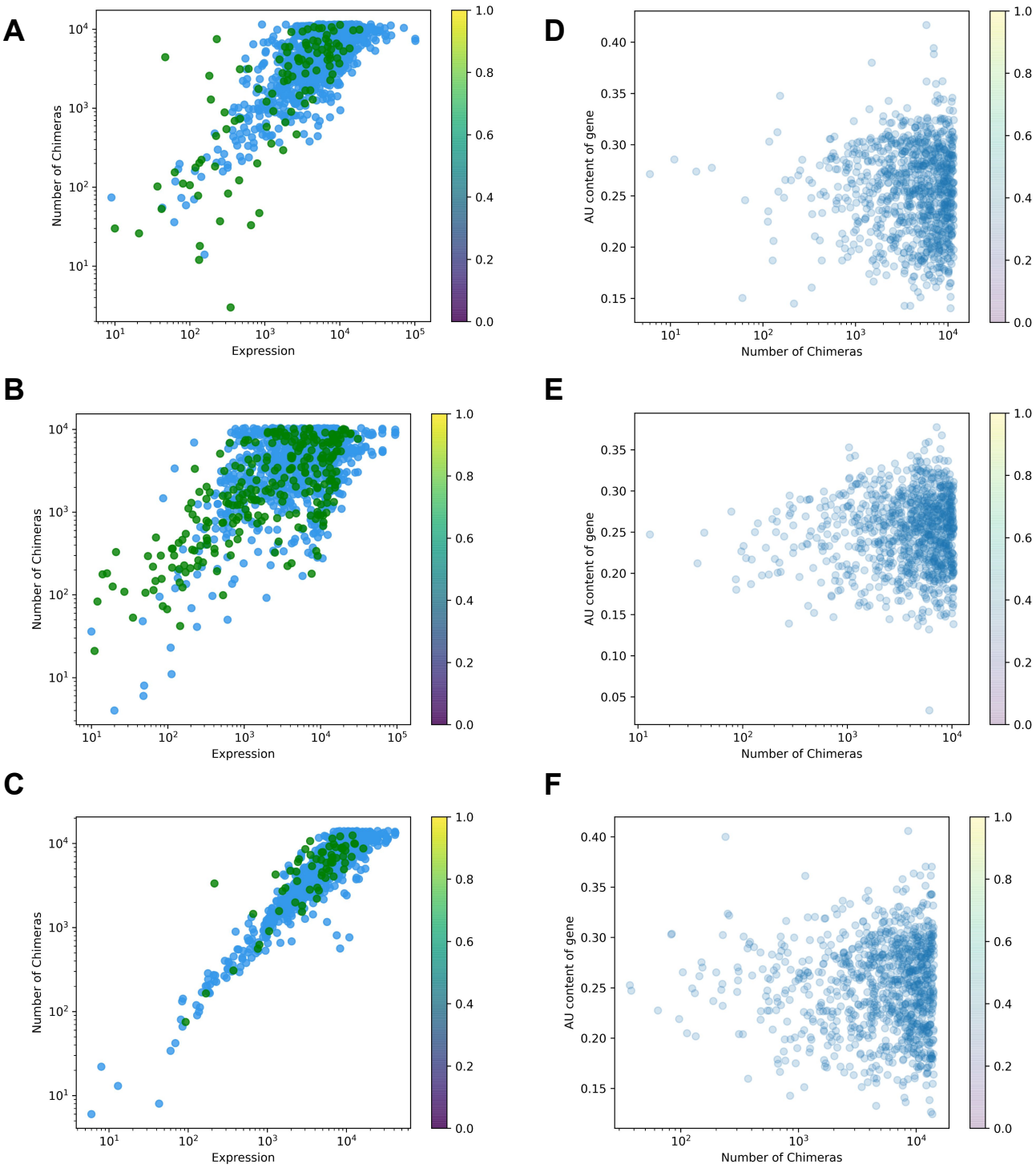


Purple - negative strand alignment to ACP6 (chr1)
Green - positive strand alignment to TUFM (chr 16)
Blue - negative strand alignment to TUFM (chr 16)
Orange - positive strand alignment to ACP6 (chr 1)

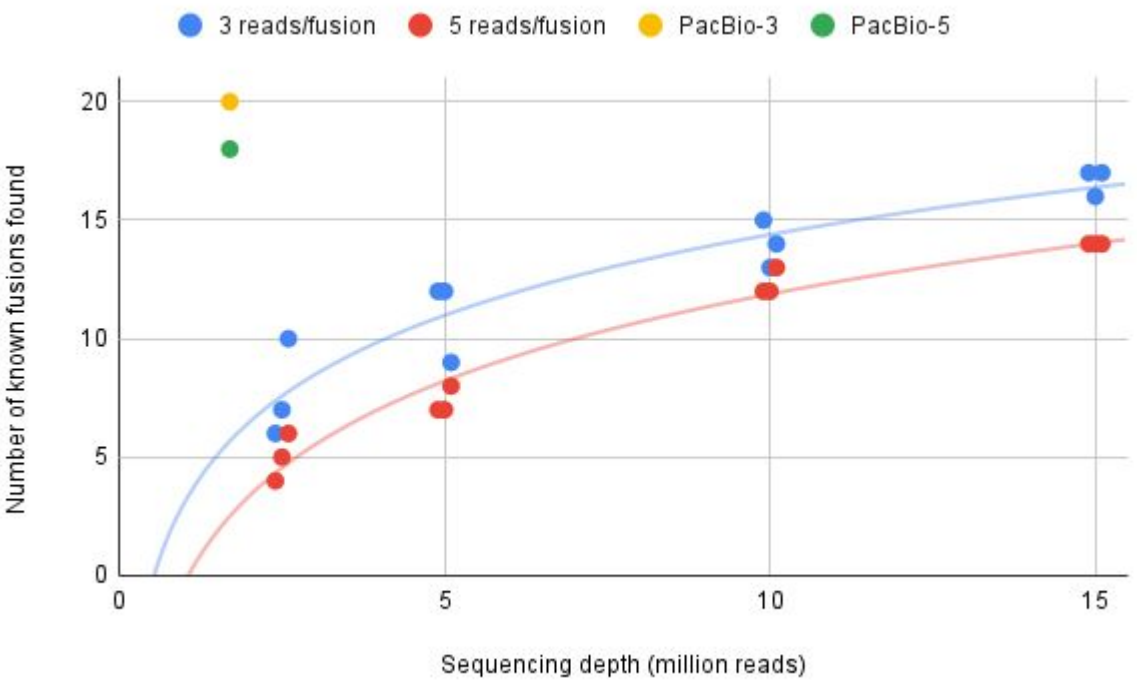
CTTAAATGTTGTTCACTTTTCACAGGGAAAGTCAGATGGCAGTAGGCCCATTCCTCCACATCCTAGAAGAGCAACCTGCT
GAAAGCCATGGACTCTACCACGCTGCCCCGACAAAGTCAGAAAGCTTCAACATCTCTACGGCTCATGATGTGACCTTCAT
ACACCGCTCTTAATGACCCTGGGGATTTTGACCACAAATGGCCACCGTTTGCTGTTGACCTGACCATGGAACTTTACCAG
CACCTGGAATCTAAGTGGTTTGTGCAGCTCTATTACCACGGGAAGGAGCAGGTGCCGAGAGGTTGCCCTGATGGGCTCT
GCCCGCTGGACATGTTCTTGAATGTATGTCGGTTATACCTTAAGCCCAGAAAAATACCACGCACTCTGCTCTCAAACTCAG
GTGATGGAAGTTGGAAGTAATGAAGGTAACTGATTTATAAAGCAGGATGTGTTGATTTAAAATAAAGTGCCTTTATACAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAGAAGAACTTGCCTGTCCGCTCTATCTTCTTTTTGTTGAGGTTGCAGCCTC
TGGCAGGTAAGCCAACCGAGCAGGAAATGTCCATCTAGCTGCCCTGCTGGGTTGCAGCCCTATGCCATGAGAGGGTACT
GGAAGCAGGAGGGAGCCCTGGCTAGGGCAGGCCCAAACGCAAGGGGAAGCTGAGCAGAGATCTGCACACTCAACCC
CATTTGATATTCTTCTCCTCCTCAGTCATGGCCAGCGTGTTGGTGACTAGACCGGTGCCAATAGTCGGTTGCCATCTCGC
AGGGTGAAACGCTGGCCTTTCTCTAAGATCATTGGCTGCCGCAAGATTAAGGTTGAACTTCAGGTCCTCCCGGGCATGG
CAAGCTCCTTCTCTGGGGGGCAGGATAATCCGACAGGCCATGTCCGGGAACCACAAGTTCAACCTAATCTTGCGACCAG
CCAATGATCTTAGAAAGGCCAGCGTTTCACCCTGCGAGATGGCAACCGGACTATTAACAATCTAGTCACCAACACGCTGG
CCATGACTGAGGAGGAGAAAGAATATCAAATGGGGTTGAGTGTGCAGATCTCTGCTCAGCTTCCCTTGCGTTTTAAGGC
CTGCCTAGCCAGGGCTCCCTCCTGCTTCAGTACCCTCTCATGGCATAGGCTGCAGCCCAGCAGAGGCAGCTAGATGGA
CATTCCCCTGCTCAGGAAGGGTTGGCCTGCCTGCTGGGAATCAGTAAACTTTGAATAGTAAAGCCAAAACAAAAAAAAAA
GAGATGGGAACAGACAGGTTCTTCTTTTTTCTAACTTATTTTAAAATCAACACATCCTGCTTTTATAAATCAGTTACTCTTCA
TTTCCAACTTCCATCACCTGAGTTTGAGAGTGCGTGGTATTTTCTGGGCTTAAGGTATAAACTGACATAGCATTCAAGAAC
ATGTCCAGCAGGCAGAGCCATGGGGCAACTTTCTCGGCACCTGCTCCTTCCCGTGGTAATAGAGCTGCACAAACCACTC
CTTAGATTCCAGGTGCTGGTAAAGTTCCATGGTCAGATCAACAGCAAACGGTGGCCATTTGTGGTCAAAATCCCCAGGGT
CATTAAGAGCGGTATGAAGGTCACATCATGAGCCGCATAGAGATACAACTTTCTGATCTTGTCGGGGCAGTGGCAGAGTC
ATAACCAACAGGTTACTCTCTCTAGGATGTGGAGAATGGGCCTACTGCCATCTGAAGACTTTCCCTGTGAAAAGTGAACA
ACATTTAAGCACCTAAGAGGGCACTCAGCATTCAGGGCAAGACTGATTCTGCCCCCAAGGACCTCATGAGTCCCAGACA
GAGCTGAGAAAAAAGGCAAGTAATGCATAAGGGCCAGCAAGCAGACCAGAAGGAGAAAATAGCTTCGAGTCCGGTCTTG
GGGTCCTCTGTGAATGAAGACTCCCAACAACGGGCACCAGCACACACAGACCTGTTGTGAACACAACTGAGATGGGAG

Supplemental Figure 3: Top Sample view of chimeric reads from direct-cDNA sequencing of MCF7 with positive and negative strand alignments to the same locus **Bottom** View of how part of one (beginning and end of read aligning to ACP6 truncated for viewing) representative read aligns to the genome, highlighting central polyA sequence between alignments to different genes. Turquoise is sequence that aligns both positive and negative strands of TUFM

Supplemental Figure 4: Correlation between expression, AU-content, and chimeras

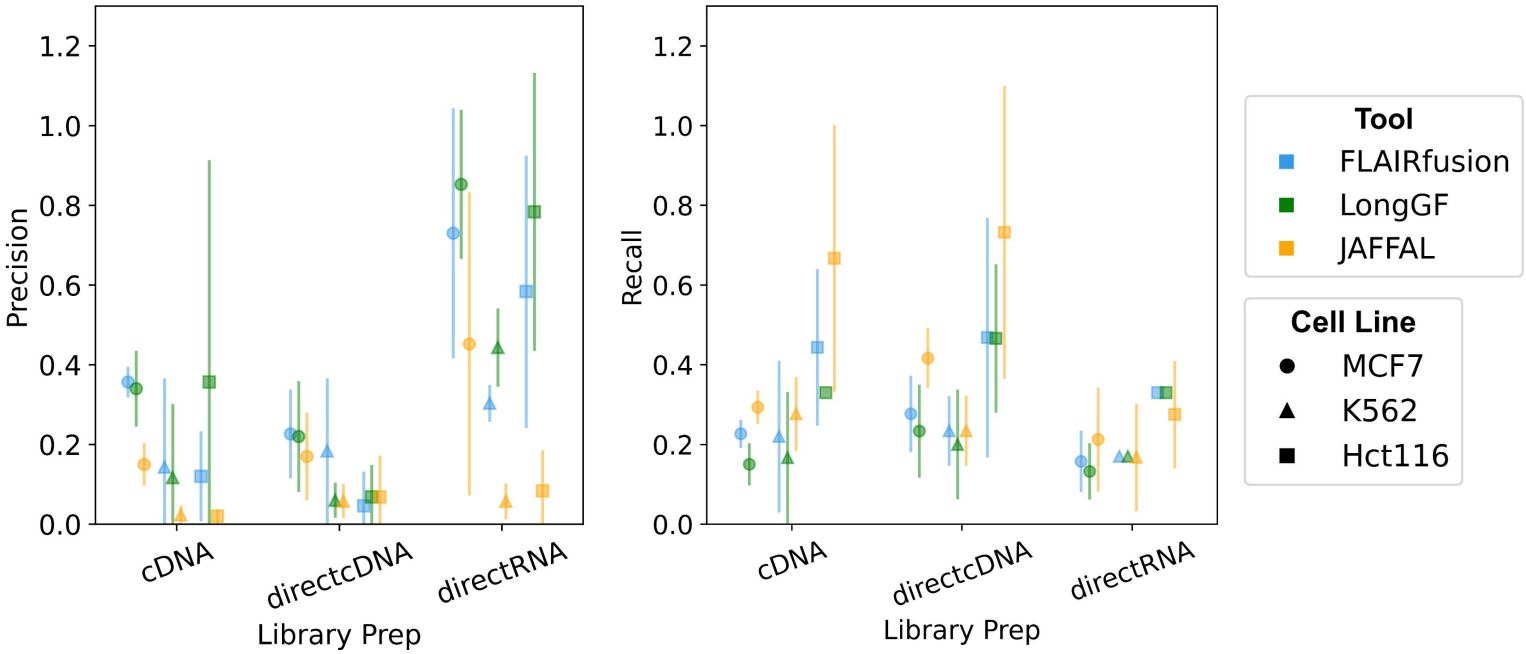


Supplemental Figure 4: A,B,C Correlation between number of chimeras and expression subsampled to 1000 genes for the MCF7 cell line - each point is an expressed gene, expression is number of reads of that gene. Green is the genes involved in previously identified gene fusions **A** cDNA, **B** direct-cDNA, **C** direct-RNA. **D,E,F** Correlation between number of chimeras and AU content of gene (potential source of artifacts due to binding with polyA tail - note: just U content is similar to these plots) subsampled to 1000 genes in the MCF7 cell line. **D** cDNA, **E** direct-cDNA, **F** direct-RNA

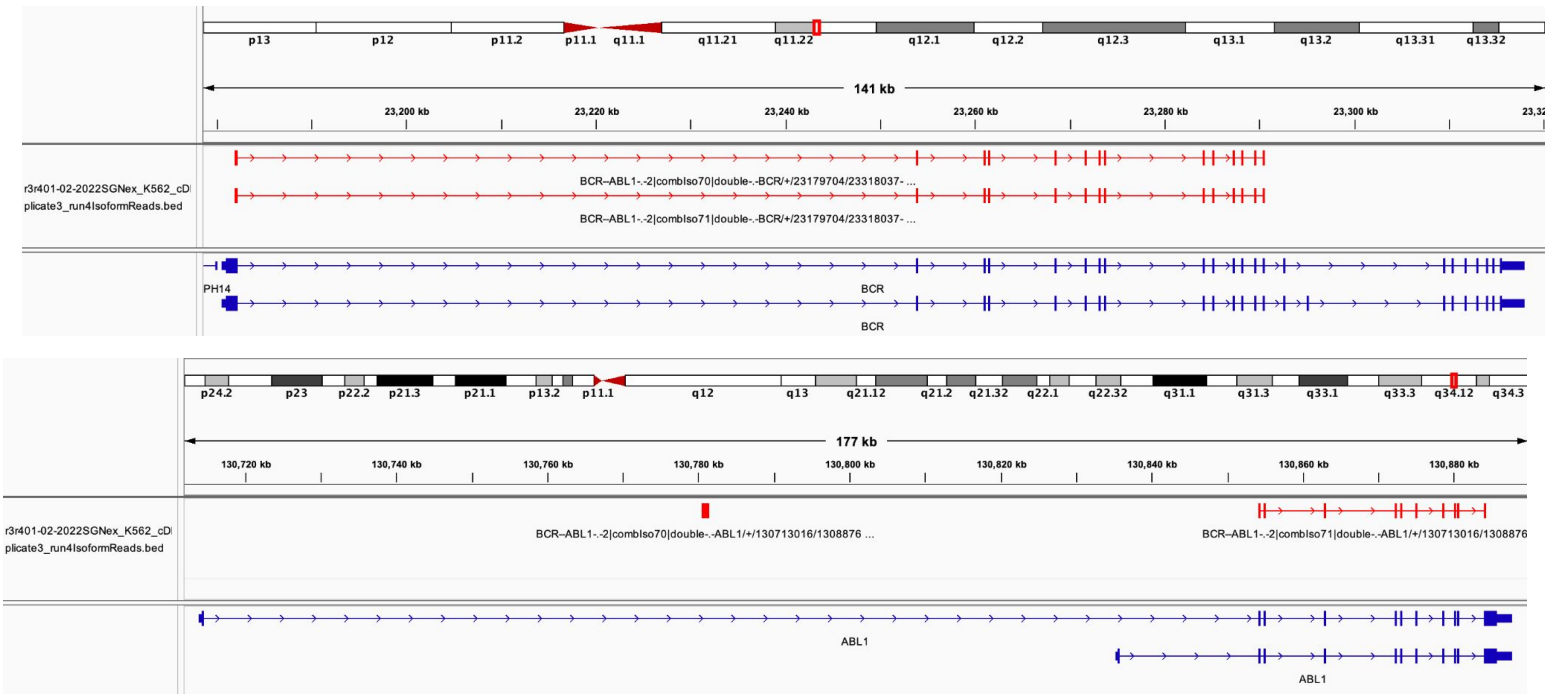


Supplemental Figure 5: Detection of known fusions in PCR-cDNA nanopore sequencing of the MCF7 cell line compared to PacBio sequencing of the same cell line. To obtain a saturation curve for the nanopore data, we combined reads from 3 replicates of cDNA sequencing for a total of 16.7M reads, then randomly subsampled these reads to obtain 3 replicates at each desired depth. Chimeras were then detected from the alignment of each replicate at two levels of stringency (3 reads vs 5 reads) and compared to known fusions. No other filters were applied for this analysis.

Supplemental Figure 6: Comparison of fusion-detection tools on cell line data



Supplemental Figure 6: Precision (left) and recall (right) of fusion calling tools on nanopore sequencing of cell lines. For significant differences between tools, see Supplemental Table 3. However, note that this dataset does not have a true negative, so evaluating true precision is difficult.



Supplemental Figure 7: Alternative splicing of BCR-ABL1 detected by FLAIR-fusion in one representative K562 cDNA sequencing sample.

Read percent sequence identity			
Sample	Mean	SD	Max
cell line cdna	87.41%	6.31%	99.26%
cell line direct cdna	88.76%	11.39%	100.00%
cell line direct rna	83.82%	4.44%	98.78%
cll control	90.78%	5.44%	100.00%
cll sf3b1 wt	90.89%	5.89%	100.00%
cll sf3b1 mut	90.86%	5.58%	100.00%

Supplemental Table 1: Read percent sequence identity for all datasets used in paper. Cell line data is from Chen et al. bioRxiv 2021. CLL data is from Tang et al. 2020.

Spearman correlation of coverage vs metric for all tools and read qualities			
Tool	Quality	Metric	P-value
JAFFAL	avg	precision	0.1957446889
JAFFAL	avg	recall	0.08264368227
JAFFAL	high	precision	0.1639816277
JAFFAL	high	recall	1
JAFFAL	low	precision	0.1355885946
JAFFAL	low	recall	9.92E-05
LongGF	avg	precision	0.7368847954
LongGF	avg	recall	0.6824890017
LongGF	high	precision	0.9821274886
LongGF	high	recall	0.6650551136
LongGF	low	precision	0.9111901055
LongGF	low	recall	5.06E-05
FLAIR-fusion	avg	precision	0.4398614018
FLAIR-fusion	avg	recall	0.005148848815
FLAIR-fusion	high	precision	0.1764442935
FLAIR-fusion	high	recall	0.2635807159
FLAIR-fusion	low	precision	0.1243894282
FLAIR-fusion	low	recall	1.05E-06

Supplemental Table 2: P-values for spearman of effect of read coverage on precision and recall for simulated data

Quality	Tool Comparison	Metric	T-test P-value
avg	FLAIRfusion-vs-JAFFAL	precision	1.27E-15
avg	FLAIRfusion-vs-JAFFAL	recall	6.99E-17
avg	FLAIRfusion-vs-LongGF	precision	8.44E-07
avg	FLAIRfusion-vs-LongGF	recall	2.16E-07
avg	LongGF-vs-JAFFAL	precision	0.0002607223323
avg	LongGF-vs-JAFFAL	recall	7.82E-06
high	FLAIRfusion-vs-JAFFAL	precision	2.60E-14
high	FLAIRfusion-vs-JAFFAL	recall	7.18E-20
high	FLAIRfusion-vs-LongGF	precision	9.07E-06
high	FLAIRfusion-vs-LongGF	recall	3.82E-09
high	LongGF-vs-JAFFAL	precision	0.006498592278
high	LongGF-vs-JAFFAL	recall	0.0004009184571
low	FLAIRfusion-vs-JAFFAL	precision	5.43E-06
low	FLAIRfusion-vs-JAFFAL	recall	0.1083478066
low	FLAIRfusion-vs-LongGF	precision	0.00266607232
low	FLAIRfusion-vs-LongGF	recall	0.0747826589
low	LongGF-vs-JAFFAL	precision	0.013015663
low	LongGF-vs-JAFFAL	recall	0.002531948064

Supplemental Table 3: P-values for t-test comparing each tool to another for different read qualities in simulated data.