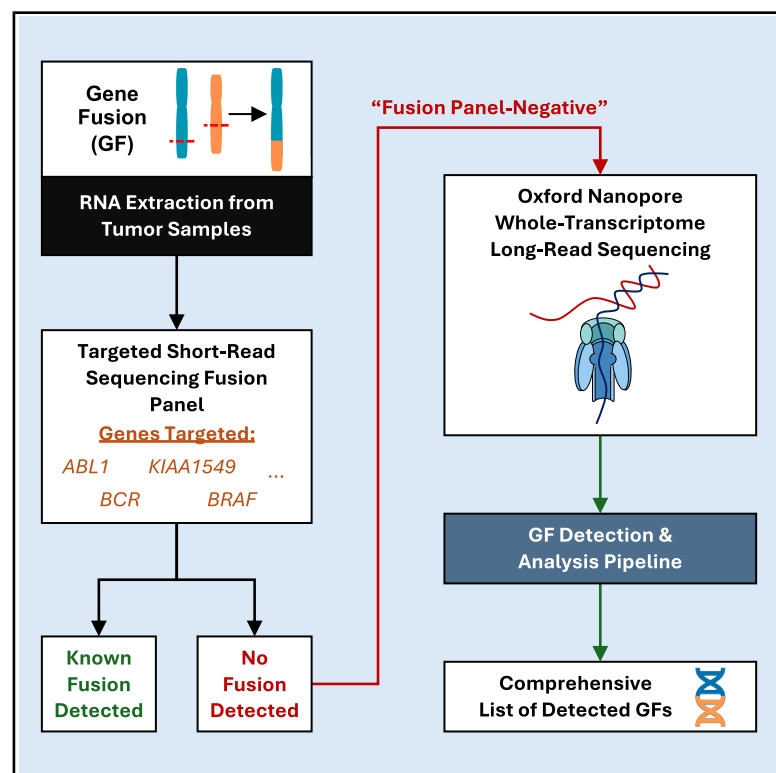


Combining panel-based and whole-transcriptome-based gene fusion detection by long-read sequencing

Graphical abstract



Authors

Karleena Rybacki, Feng Xu, Hannah M. Deutsch, ..., Yuanquan Song, Marilyn Li, Kai Wang

Correspondence

wangk@chop.edu

In brief

Rybacki et al. present a combined targeted and whole-transcriptome long-read sequencing strategy, with a computational pipeline to detect gene fusions in clinical cancer samples. By adapting a short-read panel and analyzing panel-negative gliomas, they reveal both known and candidate novel fusions, highlighting the strengths of long-read technology for fusion discovery.

Highlights

- Short-read CHOP Cancer Fusion Panel is adapted for ONT long-read sequencing
- Long-read analysis of panel-negative samples finds candidate gene fusions
- Long-read sequencing enables rapid, comprehensive fusion detection in cancer



Article

Combining panel-based and whole-transcriptome-based gene fusion detection by long-read sequencing

Karleena Rybacki,^{1,2} Feng Xu,³ Hannah M. Deutsch,⁴ Mian Umair Ahsan,² Joe Chan,² Zizhuo Liang,² Yuanquan Song,^{2,4,5} Marilyn Li,^{3,5} and Kai Wang^{1,2,5,6,*}

¹Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA

²Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

³Division of Genomic Diagnostics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

⁴Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA

⁵Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

⁶Lead contact

*Correspondence: wangk@chop.edu

<https://doi.org/10.1016/j.crmeth.2025.101111>

MOTIVATION Gene fusions (GFs) arise from chromosomal rearrangements and form hybrid genes that serve as diagnostic biomarkers and potential therapeutic targets in cancer. Conventional short-read fusion panels are limited by sequence alignment constraints and prolonged turnaround times from sample pooling. These factors can hinder timely diagnosis, treatment, and prognosis and may overlook novel, rare, or complex GFs. Long-read sequencing offers a way to address these challenges by enabling more comprehensive GF detection in a shorter time frame. Developing robust long-read sequencing GF detection and analysis pipelines is essential to comprehensively explore the landscape of GFs and advance our understanding of their roles in cancer development and progression.

SUMMARY

We present a comprehensive gene fusion (GF) detection and analysis workflow that combines targeted panel-based and whole-transcriptome long-read sequencing. We first adapted libraries from the short-read CHOP Cancer Fusion Panel, which targets 119 oncogenes commonly implicated in cancer fusions, for use on Oxford Nanopore Technologies' long-read sequencing platform. Long-read sequencing successfully detected known GFs in panel-positive samples, confirming compatibility, and enabled reduced turnaround times. To expand GF discovery in clinically challenging cases, we analyzed 24 glioma samples with negative short-read fusion panel results using whole-transcriptome long-read sequencing. This identified 20 candidate GFs in panel-negative samples that were absent from current fusion databases, all of which were experimentally validated. In summary, we introduce a computational workflow that combines panel-based and whole-transcriptome long-read sequencing with tailored analysis pipelines to enable fast and comprehensive GF detection in cancer.

INTRODUCTION

Gene fusions (GFs) are widely recognized as pivotal events in the development and progression of cancer, playing important roles in tumor characterization and diagnosis.^{1,2} GFs occur when two or more distinct genes fuse, forming a new hybrid gene, typically explained by genomic rearrangements such as chromosomal insertions, deletions, duplications, inversions, and translocations. Consequently, this process can lead to the formation of aberrant protein-coding genes, oncogene activation, or tumor suppressor gene suppression, all of which contribute to cancer develop-

ment.^{1–4} From the discovery of the Philadelphia chromosome in over 95% of chronic myeloid leukemia (CML) patients^{5,6} to the identification of GFs in breast,^{7–9} prostate,^{10,11} and non-small cell lung cancer,¹² these GFs play critical roles in cancer tumorigenesis, progression, and remission. GFs also serve as promising therapeutic targets, as demonstrated by imatinib, the first US Food and Drug Administration (FDA)-approved drug for CML, which targets the overactive *BCR::ABL* tyrosine kinase. This breakthrough therapy has led to improved survival rates and, in some cases, cured the disease.^{13,14} Similarly, entrectinib, approved by the FDA in August 2019, exemplifies the



potential of targeted therapies in treating tumors harboring neurotrophic tyrosine receptor kinase (NTRK) fusions.¹⁵ The prevalence of GFs in cancer and their therapeutic significance underscore the need for faster, more comprehensive methods for GF detection and characterization to advance personalized cancer care.^{16,17}

Traditional methods for GF detection, such as karyotyping, fluorescence *in-situ* hybridization (FISH), PCR-based assays (e.g., quantitative PCR), and short-read sequencing (typically ~150 bp), have long been utilized as reliable and robust diagnostic tools. However, these approaches may have inherent limitations. The low resolution of karyotyping limits its ability to identify GFs and is only suitable for dividing cells arrested in metaphase. FISH and PCR-based assays, while useful, are not scalable and are constrained to known GFs due to the requirement of fusion sequences for specific primer designs, making them unsuitable for detecting novel GFs. While short-read sequencing can identify GF breakpoints, it struggles to resolve full-length fused transcripts, especially in repetitive or low-complexity regions. This challenge reduces the accuracy of split-read alignment and requires additional computational reconstruction prior to analysis.^{18–21} Moreover, while targeted short-read sequencing panels are clinically validated, they are restricted to a predefined set of GFs and lack the flexibility to detect rare or novel GFs with clinical relevance. Additionally, the Illumina short-read sequencing platform often requires the indexing and pooling of multiple samples, resulting in longer turnaround times. These delays pose significant challenges for small-scale diagnostic labs, where quick and accurate results are critical for informed cancer diagnosis, prognosis, and treatment strategies.

In contrast, long-read sequencing technologies offer several advantages that can overcome these limitations and enhance diagnostic capabilities. Widely used long-read sequencing technologies include Pacific Biosciences and Oxford Nanopore Technologies (ONT), both of which have made significant strides in recent years. Advancements in long-read sequencing have led to improved read lengths (typically >10 kb for DNA), affordability, speed, and base-pair resolution, enhancing our ability to capture the wide spectrum of human genetic variation.^{22–24} These improvements also facilitate the detection of complex GFs, recovery of complete transcript structures, accurate breakpoint identification, and increased sensitivity for lowly expressed or novel GFs. As a result, genomics methods based on these cutting-edge sequencing platforms has been increasingly developed and applied in biomedical research.²⁵ ONT stands out for its flexible sequencing options (MinION, GridION, and PromethION) with varying capacities, reduced library preparation requirements, and real-time data generation, all enabling fast turnaround times.^{26–29} Although ONT historically had higher error rates, recent improvements have reduced it to approximately 1% with v14 chemistry and R10 flow cells.³⁰ As computational tools continue to advance, ONT's technology remains a highly effective platform for structural analysis, particularly for small-scale diagnostic laboratories where quick and accurate results are crucial.^{31,32} These advancements continue to pave the way for innovative diagnostic applications, deepening our understanding of rare disorders and uncovering new disease mechanisms.^{33,34}

The CHOP Cancer Fusion Panel, developed at the Children's Hospital of Philadelphia (CHOP), is a clinically validated, targeted short-read sequencing approach designed to detect GFs involving 117 cancer genes (updated to 119 genes) with known diagnostic and therapeutic significance.³⁵ The panel employs one-sided PCR, a strategy optimized for capturing GFs involving a set of predefined genes while maintaining high sensitivity and specificity.³⁵ Despite its utility, the CHOP Cancer Fusion Panel, like other short-read-based methods, is limited in its ability to detect novel or complex GFs due to read-length constraints and targeted enrichment. In addition, the turnaround time for the fusion panel is ~14–21 days for sample pooling, library preparation, sequencing, and fusion analysis,³⁶ due to the need to pool samples to optimize sequencing resources. To address these challenges, our study explores the translational potential of adapting the CHOP Cancer Fusion Panel through ONT's long-read sequencing platform, and if negative, followed by whole-transcriptome sequencing on the ONT platform. To test this workflow, we analyzed both panel-positive and panel-negative clinical samples, split into cohorts, to assess the effectiveness of long-read sequencing for GF detection. Cohort 1 included mostly panel-positive samples, initially prepared with Illumina library preparation prior to ONT sequencing. This allowed us to directly assess the feasibility and effectiveness of ONT's performance with the clinically validated short-read workflow using samples with a known disease-causal GF. Although the use of Illumina-prepared libraries limited some of the inherent advantages of long-read sequencing, such as longer read lengths, this strategy provided a foundation to build upon established clinical workflows. Cohort 2, composed entirely of panel-negative samples, was subjected to ONT whole-transcriptome sequencing. This allowed us to assess the ability of ONT to detect additional or novel GFs that may have been overlooked by conventional targeted short-read approaches. By expanding the analysis to include untargeted regions by long-read sequencing, we highlight the potential of ONT to uncover previously undetected cancer-related GFs, thereby demonstrating its utility for both clinical diagnostic applications and translational research.

We developed a comprehensive computational pipeline for long-read GF detection that integrated GF detection tools such as LongGF,³⁷ JAFFAL,³⁸ and FusionSeeker,³⁹ coupled with a robust set of filtering criteria and a computational validation. This pipeline successfully identified the expected fusions in most samples, demonstrating its reliability and accuracy. Additionally, it enabled the validation of potentially novel GFs in panel-negative samples that were likely overlooked by the short-read panel, further highlighting the advantages of long-read sequencing. By adapting an existing short-read fusion panel for long-read sequencing, validating its performance on clinically relevant samples, and reanalyzing panel-negative samples, our study demonstrates the translational potential of long-read sequencing in diagnostic and research applications. More importantly, the establishment of the experimental and computational workflow described in this study offers several advantages, including enhanced detection sensitivity, reduced turnaround times, and mitigation of the need for and delays associated with sample pooling. This method enables more confident identification of disease-causal, clinically relevant,

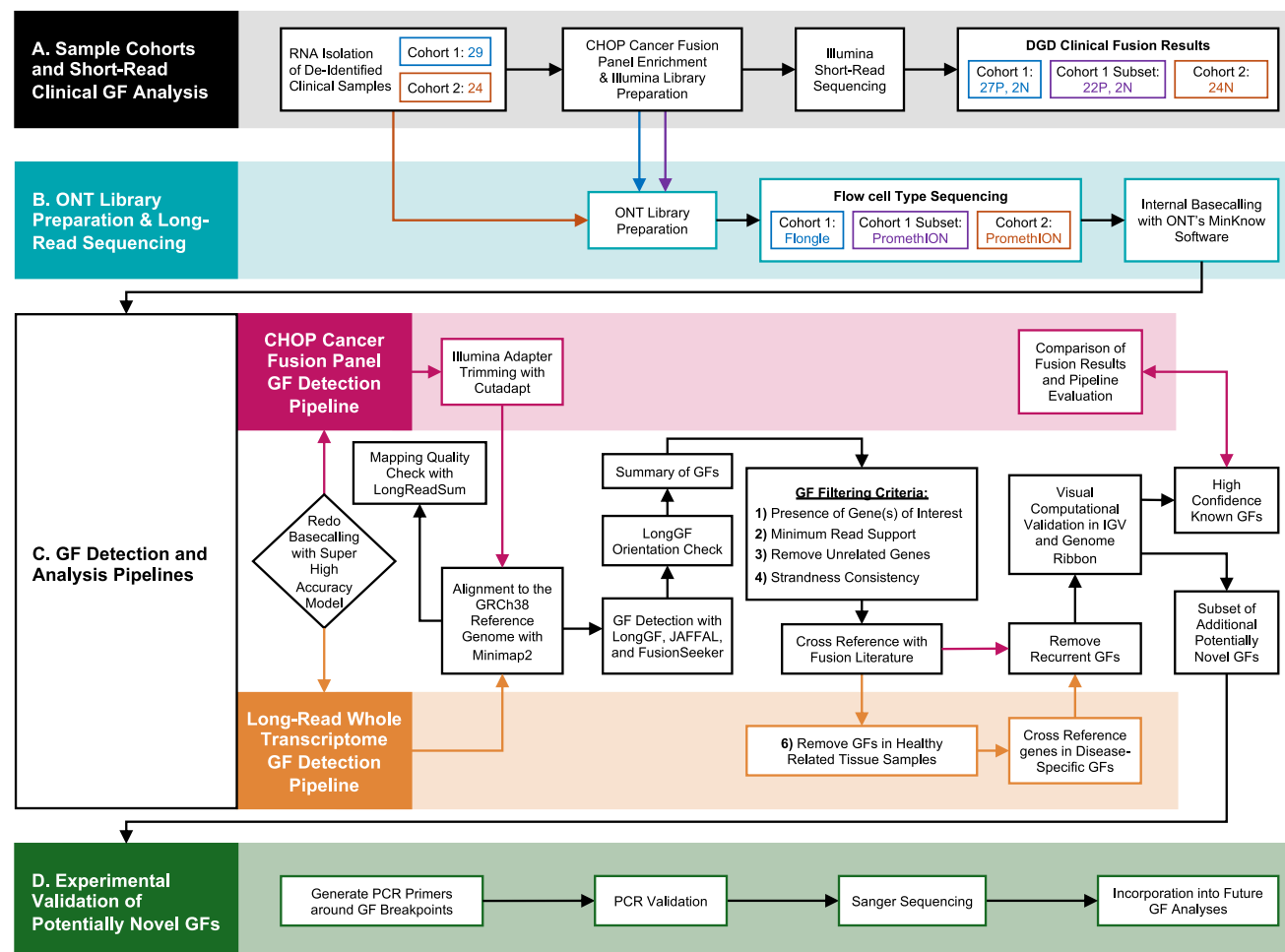


Figure 1. Comprehensive gene fusion detection workflow from sample collection to validation

(A) Sample cohorts and short-read clinical gene fusion (GF) analysis. Clinical samples from cohort 1 (29 samples, blue arrows) and cohort 2 (24 samples, orange arrows) were initially processed using the CHOP Cancer Fusion Panel, followed by Illumina library preparation, short-read sequencing, and GF analysis (black arrows, indicating steps applied to all samples). Cohort 1 included 27 panel-positive and 2 panel-negative samples, while cohort 2 included only panel-negative samples. A subset of cohort 1 (22 panel-positive and 2 panel-negative samples, purple arrows) underwent further analysis.

(B) ONT library preparation and long-read sequencing. Cohort 1 samples were prepared and sequenced on Flongle flow cells, while the cohort 1 subset samples were pooled, prepared, and sequenced on PromethION flow cells. Internal basecalling for all samples was completed using ONT's MinKNOW software.

(C) GF detection and analysis pipelines. Both pipelines begin with re-basecalling the long-read sequencing data using a super-high-accuracy model (diamond). Following this, two distinct pipelines were employed: the CHOP Cancer Fusion Panel GF detection pipeline (top pink section and arrows) and the long-read whole-transcriptome GF detection pipeline (bottom orange section and arrows). Shared core components of the GF detection and analysis approach (center, uncolored) are applied in both pipelines, with black arrows indicating steps common to all samples, regardless of the pipeline used. Cohort 1 and its subset followed the CHOP Cancer Fusion Panel pipeline, while cohort 2 followed the long-read whole-transcriptome pipeline. Detailed methods of these pipelines are outlined in the [STAR Methods](#).

(D) Experimental validation of potentially novel GFs. High-confidence, potentially novel GFs were experimentally validated using PCR and Sanger sequencing, with validated GFs considered for inclusion in future GF analyses.

and even potentially novel GFs, which are critical for improving diagnostic accuracy, guiding treatment decisions, and optimizing patient management.

RESULTS

In this study, we developed a comprehensive workflow for GF detection using ONT long-read sequencing. We evaluated its performance across two cohorts of clinical cancer patient sam-

ples, both of which had been previously assayed using the short-read CHOP Cancer Fusion Panel³⁵ (Figure 1). Cohort 1 included mostly panel-positive samples, allowing us to assess the identification of the expected disease-causal GFs. Cohort 2, composed entirely of panel-negative glioma samples, was subjected to long-read whole-transcriptome sequencing to uncover GFs not detectable by the fusion panel, some of which may be disease relevant. Table S1 details sample origin, cancer type, and clinical indications of all samples.

To identify high-confidence GFs, we implemented a systematic filtering and prioritization approach. This accounted for differences between long-read sequencing of short-read targeted enrichment libraries (cohort 1 and its subset) and long-read whole-transcriptome sequencing (cohort 2). In both cohorts, we excluded GFs involving non-relevant genes, such as mitochondrial, ribosomal, human leukocyte antigen (HLA), or pseudogenes, to reduce likely false positives. Read support thresholds were adjusted based on flow cell type and sample characteristics to maintain sensitivity, especially for rare or complex events. Strand consistency filters were applied to remove potential technical artifacts, and recurrent fusions detected in over 15% of all samples were flagged for manual review prior to removal. In cohort 2, we further excluded fusions observed in normal brain tissue from the Genotype-Tissue Expression (GTEx) long-read project (version 9)⁴⁰ and prioritized those involving a gene relevant to GFs implicated in central nervous system (CNS) diseases reported in the Mitelman fusion database. Detected GFs were considered potentially novel if they were not identified by the CHOP Cancer Fusion Panel and not previously reported in relevant fusion databases, specifically Mitelman,⁴¹ COSMIC Fusion,⁴² and ChimerDB 4.0.⁴³ These results demonstrate the effectiveness of ONT long-read sequencing in detecting a broader spectrum of GFs, including those overlooked by traditional targeted approaches.

Cohort 1: Long-read GF detection by Flongle flow cell

Cohort 1 consisted of short-read libraries from 29 clinical samples that were previously assayed by the CHOP Cancer Fusion Panel. These samples were evaluated through the ONT long-read sequencing platform on individual Flongle flow cells, with sequencing summary details provided in Table S2A. The Cutadapt metrics for cohort 1 samples, summarizing the removal of short-read library preparation adapters, are presented in Table S3A. Following alignment to the GRCh38 reference genome (no alternate contigs) and quality check analysis using LongReadSum,⁴⁴ the mapped N50 read lengths ranged from 139 to 300 bp, with an average of approximately 220 bp (Table S4A). This was consistent with the initial short-read library preparation of the samples from previous analysis. All samples had fewer than 1 million mapped reads for subsequent long-read analyses, as expected due to prior targeted enrichment from the fusion panel, yet the sequencing depth remained sufficient for GF detection.

GF detection for this cohort was conducted using the long-read CHOP Cancer Fusion Panel GF detection pipeline, illustrated by the pink and black arrows in Figure 1C. Given that the sequencing libraries originated from targeted enrichment of the genes involved in the CHOP Cancer Fusion Panel, the pipeline was inherently focused on these genes to identify the expected disease-causal GFs. A broad range of unique GFs across all samples were detected and after applying the filtering criteria (STAR Methods), the number of remaining unique GFs (based on genes involved in the fusion) in the panel-positive samples ranged from 1 to 34 (Table S5A). In the two panel-negative samples (samples 21 and 25), no fusions remained, which confirmed the expected fusion panel-negative status. In the panel-positive samples, at least one known GF previously re-

ported in literature was present in 21 samples. Visual inspection of the Integrated Genomics Viewer (IGV) and Genome Ribbon plots for these known GFs confirmed the presence and breakpoints for these fusions. For example, a few of the known GFs identified included (1) *EBF1::PDGFRB* (sample 1), (2) *GNAI2::BRAF* (sample 10), and (3) *ETV6::NTRK3* (sample 22), as shown in Figure 2.

In total, 20 samples contained exactly 1 curated GF each, 7 samples had 2 or more putative GFs, and 2 samples were deemed panel negative. When we evaluated the performance of our pipeline and compared the results with the original analysis on short-read data from the diagnostic lab, the expected disease-causal GF was identified among the putative GFs in 21 of 27 panel-positive samples and the 2 panel-negative samples were correctly classified. However, in the remaining 6 panel-positive samples, the true positive GF was not detected. To explore why these GFs were not detected by Flongle flow cell, we conducted a more in-depth analysis on a subset of samples in cohort 1 using PromethION flow cell ONT long-read sequencing, detailed in a subsection below.

Cohort 1 subset: Long-read GF detection by PromethION flow cell

A subset of cohort 1, referred to as cohort 1 subset, included 24 of the 29 clinical samples, excluding samples 1 and 3–6 due to insufficient starting material for sequencing. These 24 samples consisted of 22 panel-positive and 2 panel-negative samples, which were pooled for multiplex sequencing on a PromethION flow cell. Although the transition from Flongle to PromethION flow cell for the same samples exhibited minimal changes in N50 read lengths, the total number of reads initially generated increased severalfold, with sequencing summary details provided in Table S2B. Table S3B provides an overview of Cutadapt processing for cohort 1 subset samples, detailing the removal of short-read library preparation adapters. After aligning the reads to the GRCh38 reference genome and performing quality analysis with LongReadSum, the mapped read counts for subsequent analyses ranged between 1.62 and 10.2 million (Table S4B).

Like cohort 1, the CHOP Cancer Fusion Panel GF detection pipeline (illustrated by the pink and black arrows in Figure 1C) was applied to this subset of samples to identify the expected disease-causal GF. After applying the GF filtering criteria, the remaining unique GFs (based on gene names) ranged from 1 to 48 fusions, with an average of 10 per sample (Table S5B), while the previously deemed panel-negative samples contained only 2 unique fusions each. In these two panel-negative samples, sample 21 identified the reciprocal of a known fusion (*TACC3::FGFR3*) and an unknown GF (*RUNX1::PPARD*), both with low read support and a lack of validation across multiple programs. In sample 25, one GF (*LPP::PDGFRB*) had low read support and was only reported in one program, while the other fusion (*FUS::SAP30BP*) had higher read support, but the *FUS* gene breakpoints differed by 133 bp between the two fusion programs, further reducing confidence in true fusion events. These factors suggested that fusions in these samples were likely technical artifacts, confirming the panel-negative status. Of the samples, 16 contained 1 putative, likely disease-causal GF, 6 samples

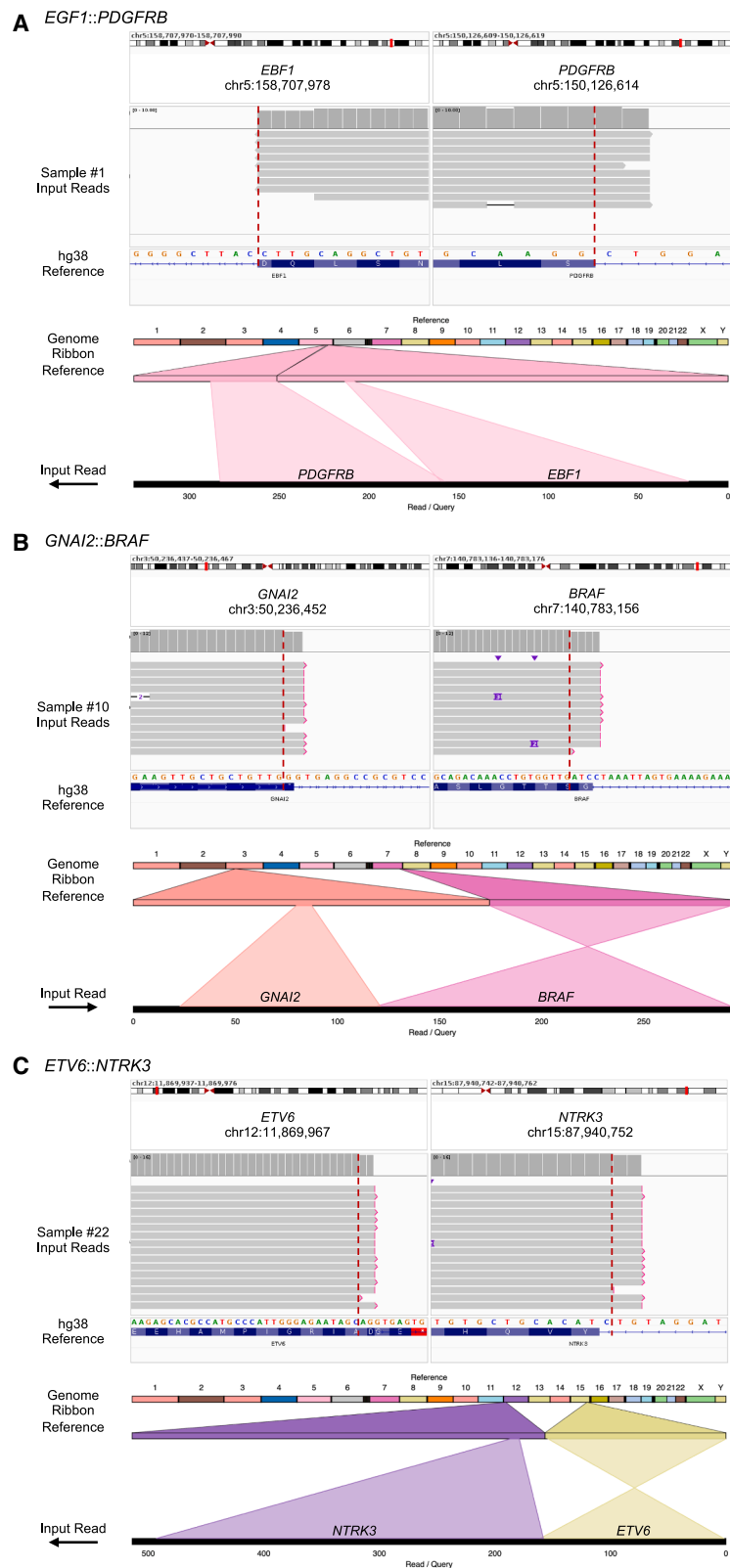


Figure 2. IGV and Genome Ribbon plots of expected fusions from the clinical cancer samples 1, 10, and 22

The IGV plots of the expected fusions in (A) *EBF1::PDGFRB* (sample 1), (B) *GNAI2::BRAF* (sample 10), and (C) *ETV6::RUNX1* (sample 24). In each plot, the top section features the IGV plot with input supporting reads annotated with the gene name and reported breakpoint indicated by the dotted lines. The lower section displays the Genome Ribbon single-read view, where gene regions are labeled by name, and the input read orientation is denoted by the arrow.

contained two or more GFs with supporting reads exceeding the adjusted threshold, and 2 samples were determined to be fusion-negative.

Following confirmation of the GFs previously confirmed by the CHOP Cancer Fusion Panel, 19 of the 22 panel-positive samples of cohort 1 subset correctly identified the expected disease-causal GF. Notably, three samples that had previously failed to report the disease-causal GF from Flongle (16, 18, 27) were later recovered in PromethION flow cell sequencing. This recovery suggests that failure to detect the expected GF in these samples could be attributed to the lower coverage from Flongle (average 62.98× coverage) compared to PromethION (average 864.57× coverage) sequencing. It is important to note that, despite the samples being previously analyzed on Flongle flow cells, the prior knowledge of the expected disease-causal GFs and panel-negative status did not influence the GF detection method or the application of the GF criteria. However, three samples (2, 9, and 26) still did not have their disease-causal GF detected by either Flongle (cohort 1) or PromethION (cohort 1 subset) flow cell sequencing.

Overall, of the GFs initially detected from Flongle or recovered from PromethION flow cell sequencing, 2 GFs were identified exclusively from 1 program, 7 by 2 programs, and 10 were detected by all 3 programs. This highlights the increased sensitivity achieved through an ensemble of GF detection programs. Table 1 presents the expected disease-causal GFs, breakpoints, and read support across the GF detection programs for all samples. Successful identification of cohort 1 samples by Flongle flow cell is denoted without an asterisk, while those undetected by Flongle but recovered by PromethION flow cell of the cohort 1 subset are marked with one asterisk (*), and samples undetected by both sequencing methods are marked with two asterisks (**).

Manual analysis of the undetected GFs in cohort 1 and cohort 1 subset

After multiplex sequencing of cohort 1 subset samples on the PromethION flow cell, we analyzed the possible causes of the three previously undetected GFs from samples 2, 9, and 26. This analysis focused on a subset of reads from each sample with a minimum mapping quality score of 55.

Sample 2 *IGH::CRLF2*

For sample 2, the disease-causal GF from the CHOP Cancer Fusion Panel was identified to be between the *CRLF2* gene and the *IGH* gene cluster, which consists of several genes. The GF detection pipeline identified 40 unique fusions involving the *CRLF2* gene but did not report any fusion to *IGH*. Manual inspection of the BAM file showed several split alignments between *IGH* (chr14:105,817,135) and *CRLF2* (chrX:1,206,569), as shown in Figure S1A. However, the breakpoint in *IGH* is considered intergenic since the coding region of this gene was not annotated as part of any gene in the GENCODE version 40⁴⁵ annotation. Since GF detection programs look for fusions between two annotated genes, this GF between the *CRLF2* gene and an unannotated intergenic locus in *IGH* was ignored by all GF detection methods. Typically, a candidate GF between a gene and an intergenic region would be attributed to technical or spurious alignment artifacts. This case suggests that a true

and likely oncogenic GF could exist between a gene and intergenic region (especially for immunoglobulin or T cell receptor), and GF detection methods should take this possibility into consideration when reporting results as opposed to discarding such GFs.

Sample 9 *EGFR::EGFR*

The disease-causal transcriptomic rearrangement for sample 9 was not a GF, but the skipping of exons 21 to 23 of the *EGFR* gene (Ensemble transcript ENST00000275493.7), whereas no known transcript for *EGFR* in GENCODE version 40⁴⁵ annotation exhibits this exon-skipping pattern. Minimap2⁴⁶ alignment for this gene shows an alternative splicing event between exons 20 and 24, with virtually no coverage for exons 21 to 23, as shown in Figure S1B. This rearrangement was not identified by GF detection tools such as LongGF, JAFFAL, and FusionSeeker because they only search for rearrangements between exons of distinct genes, as opposed to rearrangements within exons of the same gene. To detect such rearrangements in genes known to have oncogenic exon skipping events, an alternative strategy would be to employ novel isoform detection tools such as FLAIR,⁴⁷ TALON,⁴⁸ or LIQA,⁴⁹ that can detect both novel exon boundaries and novel combination of known exons.

Sample 26 *STIL::TAL1*

The GF detection pipeline incorporates splice-aware alignment to the reference genome, which failed to detect the *STIL::TAL1* GF in sample 26. These two genes are directly adjacent to each other, and the fusion breakpoints are only ~80 kbp apart. Minimap2 splice-aware alignment treats the two fused exons as part of a single transcript and reports an 80-kbp spliced region between the fused exons, instead of splitting the read alignment into two distinct alignment records. As a result, GF detection methods that require a split alignment between distinct genes were not able to detect this GF. However, when employing non-splice-aware mapping with Minimap2, the read alignments between the genes were split into two distinct segments. This enabled LongGF to successfully identify this GF with 37 supporting reads with PromethION flow cell (cohort 1 subset), confirming the fusion event, while it remained undetected in Flongle flow cell (cohort 1). This highlighted the potential limitations of the (1) alignment software used, such as Minimap2, which does not split alignments between distinct, nearby genes despite a provided reference splice junction file and (2) GF detection tools, which are not able to detect a GF from a single alignment record that spans multiple distinct and non-overlapping genes. In general, we recommend using a splice-aware alignment method over a non-splice-aware alignment method as the latter treats introns as deletions or unnecessarily splits alignments over introns, which can confound a true deletion from a transcriptomic rearrangement.

Cohort 2: Long-read whole-transcriptome GF detection by PromethION flow cell

Cohort 2 consisted of 24 clinical samples derived from high-grade gliomas (HGGs) and low-grade gliomas (LGGs) (Table S1). Like previous cohorts, these samples were previously processed through the CHOP Cancer Fusion Panel and deemed fusion panel negative. Note that we also included the K562 cell line, which harbors *BCR::ABL1* fusion, in each pool as an internal control.

Table 1. Overview of expected GFs in clinical cancer patient samples

Sample no.	Expected fusion	Reported gene 1 breakpoint (hg38)	Reported gene 2 breakpoint (hg38)	No. of supporting reads					
				Flongle flow cell			PromethION flow cell		
				L	J	FS	L	J	FS
1	<i>EBF1::PDGFRB</i>	chr5:158,707,978	chr5:150,126,613	6	4	0	NA	NA	NA
3	<i>KMT2A::AFF1</i>	chr11:118,484,972	chr4:87,084,118	25	21	24	NA	NA	NA
4	<i>P2RY8::CRLF2</i>	chrX:1,536,919	chrY:1,212,637	877	934	1,034	NA	NA	NA
5	<i>PML::RARA</i>	chr15:74,033,358	chr17:40,348,314	8	23	19	NA	NA	NA
6	<i>KMT2A::MLLT3</i>	chr11:118,482,493	chr9:20,365,746	20	37	0	NA	NA	NA
17	<i>FGFR3::TACC3</i>	chr4:1,806,934	chr4:1,739,701	473	465	887	4,057	4,060	0
23	<i>FUS::CREB3L2</i>	chr16:31,184,357	chr7:137,908,380	93	92	89	3,852	3,861	0
14	<i>KIAA1549::BRAF</i>	chr7:138,867,973	chr7:140,787,582	40	33	11	1,334	170	0
22	<i>ETV6::NTRK3</i>	chr12:11,869,967	chr15:87,940,754	16	16	16	834	826	809
29	<i>KIAA1549::BRAF</i>	chr7:138,879,536	chr7:140,783,155	5	4	4	404	401	358
15	<i>KIAA1549::BRAF</i>	chr7:138,861,137	chr7:140,787,582	28	5	6	339	315	37
20	<i>EWSR1::ERG</i>	chr22:29,287,133	chr21:38,383,921	10	8	7	302	242	0
19	<i>KIAA1549::BRAF</i>	chr7:138,867,973	chr7:140,787,583	21	19	4	281	265	0
12	<i>KIAA1549::BRAF</i>	chr7:138,861,137	chr7:140,787,583	12	2	0	243	219	30
24	<i>ETV6::RUNX1</i>	chr12:11,853,559	chr21:34,892,963	4	4	4	141	76	0
10	<i>GNAI2::BRAF</i>	chr3:50,236,452	chr7:140,783,156	11	10	8	120	115	92
8	<i>KIAA1549::BRAF</i>	chr7:138,867,973	chr7:140,787,583	11	9	0	115	15	12
13	<i>GTF2I::BRAF</i>	chr7:74,755,536	chr7:140,783,156	11	7	7	81	46	14
28	<i>EWSR1::FLI1</i>	chr22:29,287,132	chr11:128,805,365	2	0	0	71	60	54
7	<i>ETV6::RUNX1</i>	chr12:11,869,967	chr21:34,892,964	3	3	3	60	30	0
18*	<i>HIP1::MET</i>	chr7:75,554,112	chr7:116,774,880	0	0	0	24	24	21
16*	<i>WDCP::NTRK2</i>	chr2:24,031,024	chr9:84,867,237	0	0	0	20	0	0
27*	<i>NTRK2::BCR</i>	chr9:84,867,430	chr22:23,253,792	0	0	0	17	17	13
11	<i>GOPC::ROS1</i>	chr6:117,566,854	chr6:117,321,394	0	9	0	0	81	0
2**	<i>IGH::CRLF2</i>	chr14:105,817,223	chrY:1,212,636	0	0	0	0	0	0
9**	<i>EGFR::EGFR</i> (exon skipping)	exon 24	exon 20	0	0	0	0	0	0
26**	<i>STIL::TAL1</i>	chr1:47,314,034	chr1:47,225,889	0	0	0	0	0	0
21	Fusion negative	–	–	–	–	–	–	–	–
25	Fusion negative	–	–	–	–	–	–	–	–

An overview of the expected causal GF from each clinical cancer patient sample, detailing the de-identified sample number, GF name, and reported hg38 reference breakpoints from Nanopore long-read sequencing. The accompanying data includes the number of supporting reads from Flongle and PromethION flow cell sequencing as detected by LongGF (L), JAFFAL (J), and FusionSeeker (FS) programs. Successful identification by Flongle flow cell sequencing for detected GFs do not have an asterisk by the sample number and those undetected are marked with one (*) or two (**) asterisks. Similarly, successful identification by PromethION flow cell sequencing for detected GFs do not have an asterisk, samples that were previously undetected by Flongle but successfully recovered in PromethION flow cell sequencing are denoted with one asterisk (*), and those that remained undetected overall are denoted by two asterisks (**). For samples with NA, PromethION flow cell sequencing was not conducted, and the Fusion Negative samples do not possess an expected fusion, as indicated by a dash. The table is sorted from high to low in terms of the number of supporting reads from LongGF as sequenced from PromethION.

However, unlike previous cohorts, the availability of total RNA enabled ONT library preparation and full-length transcriptome sequencing. This provided a unique opportunity to directly leverage the inherent advantages of ONT long-read sequencing for detecting complex genetic rearrangements and identify likely additional GFs possibly overlooked by conventional short-read sequencing methods.

These samples were pooled across three sequencing pools and prepared for multiplex PromethION flow cell sequencing. The sequencing results demonstrated the long-read capability of ONT's PromethION flow cell sequencing, generating an

average of 8.2 million reads with an average N50 read length of 850 bp (excluding the K562 controls), demonstrating an improvement over the limitations of short-read sequencing (Table S2C). Sequencing depth varied across the pools of multiplexed samples, with pool 2 having the highest number of mapped reads at 9.2million, followed by pool 3 with 7.44 million, and lastly, pool 5 with 4.55 million (Table S2C). Following alignment to the GRCh38 reference genome and quality check analysis using LongReadSum, all samples in this cohort had an average of 4.5 million mapped reads for subsequent analyses (Table S4C). This read count reflected our sample pooling strategy, which

aimed to balance sample multiplexing with sufficient sequencing depth for long-read transcriptome analysis.

The long-read whole-transcriptome GF detection pipeline, represented by the orange and black arrows in [Figure 1C](#), was applied to identify known and potentially novel GFs that may have been overlooked in cohort 2. Initial GF detection tools LongGF,³⁷ JAFFAL,³⁸ and FusionSeeker³⁹ identified thousands of unique GFs, each defined by specific gene pairs and breakpoints. To refine these results, we applied the standard GF filtering criteria ([STAR Methods](#)) and an additional criterion to exclude GFs identified in parallel healthy brain tissue samples from the GTEx long-read project (version 9).⁴⁰ This approach helped narrow down the list of biologically relevant GFs for further analysis.

Excluding those detected in the K562 controls, 5 samples had no fusions remaining (49, 62, 63, 66, and 71), 2 samples had exactly 1 fusion (65 and 70), and the 17 remaining samples had more than 1 fusion. Following the application of a GF filtering criteria for cohort 2, this resulted in a drastic reduction of the initial set of detected GFs, leaving 180 unique fusions based on the gene name, with an average of 23 GFs per sample ([Table S5C](#)). The filtered set of GFs was then cross-referenced with fusion databases, revealing that four unique fusions (two with their reciprocal fusions) were detected within the samples. Among these were *NUP107::MDM2* (sample 64, also reported reciprocal), *RPL4::MAP3K13*, *ESR1::SYNE1* (reported in two samples, 45 and 52), and *MDM2::RAP1B* (sample 64, also reported reciprocal). Genes in the remaining GFs were also cross-referenced with a curated list of genes known to be involved in GFs found in CNS-related diseases, reported in the Mitelman database.⁴¹ Across all samples, this identified 255 unique GFs with only 1 gene and 71 GFs with both genes associated with fusions in CNS-related diseases, suggesting their potential clinical significance. Notably, sample 34 had the highest number of GFs with both genes associated (20 GFs), while sample 40 had 80 GFs with only 1 gene in the fusion associated with a known GF found in CNS-related diseases. This trend was expected, as samples 34 and 40 were sequenced from pool 2, which had the highest average number of reads sequenced and aligned, likely increasing the sensitivity of GF detection.

The distribution of GFs across the HGG and LGG samples revealed a higher prevalence of unique GFs associated with CNS disease-related genes in HGG. Specifically, HGG samples identified 57 unique GFs where both genes were associated with CNS diseases and 188 GFs where only 1 gene was linked. In contrast, LGG samples detected just 14 GFs with both genes and 67 with at least 1 associated gene. While both glioma cancer types exhibited fusions involving CNS-related disease genes, the HGG samples identified a higher number of detected fusions. This is consistent with increased genomic instability in HGG, often linked to structural variants or other chromosomal rearrangements driving tumorigenesis.^{50,51} These findings also suggest that with fewer detected fusions, alternative oncogenic mechanisms may drive tumor progression in LGGs. However, some LGG samples still contained fusions with potential clinical significance, highlighting the need for further investigation into the functional role of these GFs in lower-grade gliomas and their clinical relevance.

Identification of potentially novel GFs

Across all samples, 572 remaining unique GFs were reviewed, and 25 were classified as high confidence, largely based on their involvement in the CHOP Cancer Fusion Panel, relevance to genes in GFs associated with CNS diseases, breakpoint type, and breakpoint consistency. These GFs were then visualized in Genome Ribbon ([Figure S3A](#)) and selected for further experimental validation ([Figure 1D](#)) as potentially clinically relevant novel GFs. The distribution of detected GFs demonstrated the power of long-read sequencing for blind GF detection, especially in challenging glioma cases that were previously deemed fusion panel negative. However, with increased read lengths and fusion sensitivity, careful filtering criteria were applied to distinguish biologically relevant GFs from artifacts.

For the experimental validation, PCR primers are listed in [Table S6](#), and [Figure S3B](#) details PCR gel electrophoresis images. During PCR validation, a faint band corresponding to the *ATG4B::CNTRL* fusion was unexpectedly observed in the K562 negative control. Closer examination revealed that the downstream intronic region to the *ATG4B* gene breakpoint contains a SINE repeat element, MIRb, as indicated by RepeatMasker data on the UCSC Genome Browser.⁵² This suggested that the observed amplification in the negative control was likely due to primer binding to the MIRb repeat element, rather than a true fusion product. Additionally, other fusions involving *ATG4B* and *CNTRL* genes individually were initially detected in the K562 samples but had very low read support (fewer than two reads), further suggesting that the observed band could be a non-specific amplification artifact. Additionally, a faint PCR band corresponding to the *OLIG2::CHTOP* fusion (redesigned primers) was also observed in the K562 negative control. Unlike the *ATG4B::CNTRL* fusion, no repeat elements were identified near the gene breakpoints, and no fusions involving the individual genes were detected. This suggested possible low-level non-specific amplification. Both amplifications were specific to the K562 negative control, and no such amplification was detected in the clinical samples, as confirmed by the Sanger Traces for these validated fusions ([Figure S3C](#)), further supporting the validity of the fusion events.

Of the 25 selected GFs for experimental validation, 20 were successfully validated through PCR and Sanger sequencing. For the five GFs that were not validated, potential reasons include low read support, poor primer binding, or suboptimal PCR amplification. Two GFs (*OLIG2::CHTOP* and *PTPRK::NOX3*) were later successfully validated after primer redesign. These candidate GFs represent potentially novel chromosomal rearrangements, as they were not identified by the CHOP Cancer Fusion Panel and were absent from current fusion databases (e.g., Mitelman,⁴¹ COSMIC Fusion,⁴² ChimerDB 4.0⁴³). This demonstrates the ability of long-read sequencing to uncover complex chromosomal rearrangements in samples previously deemed fusion panel negative.

DISCUSSION

Our study presents a two-tiered long-read approach to achieve a more comprehensive profile of GFs, thereby improving our understanding of their role in cancer development and as potential therapeutic targets. Targeted short-read next generation sequencing

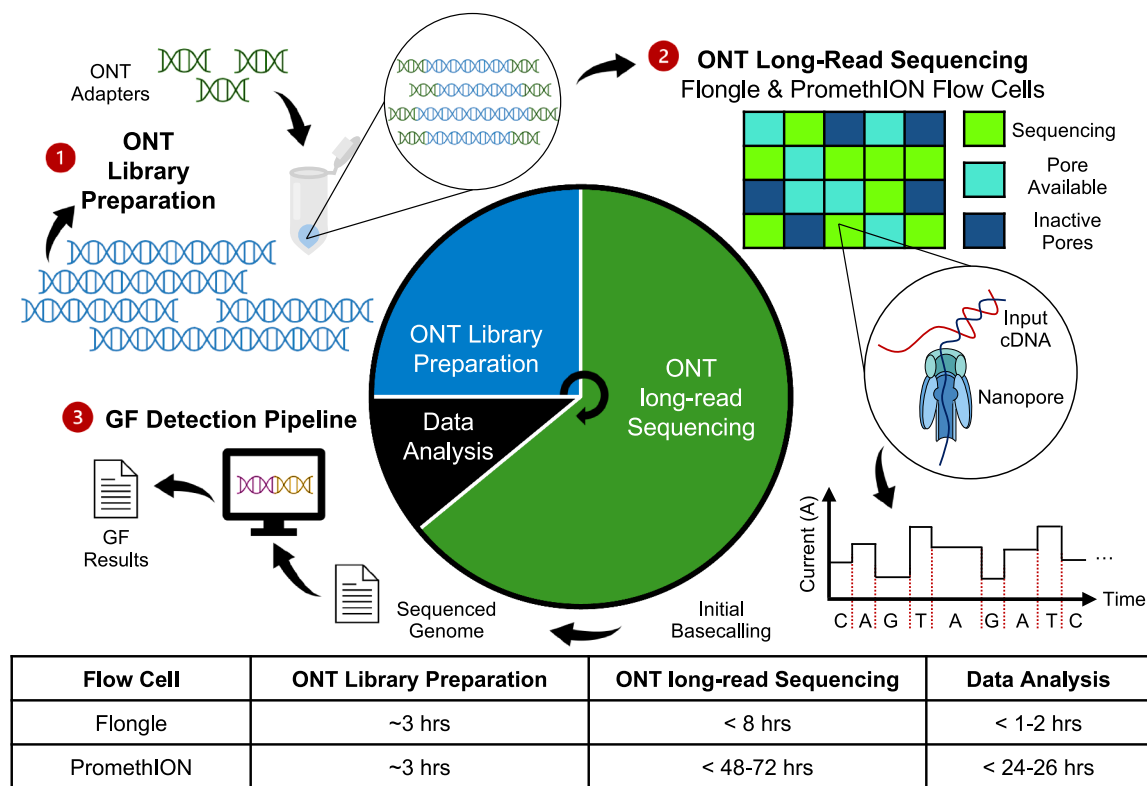


Figure 3. Overview of Flongle and PromethION flow cell library preparation, sequencing, and analysis

Clinical samples initially prepared for short-read sequencing underwent ONT library preparation, which typically requires ~3 h. These libraries were subjected to long-read sequencing using both Flongle and PromethION flow cells, with sequencing run times of up to 8 h and less than 48 h, respectively. The resulting reads were basecalled using ONT's integrated basecaller, MinKNOW, and processed through the GF detection pipeline. Downstream computational analysis typically requires ~1–2 h for Flongle and ~24–26 h for PromethION. Notably, ONT sequencing supports real-time analysis and can be stopped early upon detection of a target fusion, enabling reductions in both sequencing and analysis times, and improving overall turnaround time.

(NGS) panels have their limitations, particularly in identifying fusions involving non-conventional genes and capturing the full-length sequences due to their short sequence lengths. Moreover, batching multiple samples for gene panels often results in prolonged turnaround times, which can delay time-sensitive cancer diagnoses. In response to these challenges, our study explored the effectiveness of ONT long-read sequencing in detecting GFs in clinical cancer patient samples that were previously assayed by the CHOP Cancer Fusion Panel. Although cohort 1 samples were initially processed with the short-read library preparation protocol, we were able to adapt the short reads through the ONT sequencing platform with minimal additional processing. Cohort 2 provided an opportunity to leverage the advantages of long-read sequencing in terms of identifying additional GFs, possibly previously overlooked, in challenging clinical cases. This approach enabled us to detect potentially novel GFs with a relatively quick turnaround time (Figure 3). Below, we discuss the different approaches and limitations of the present study.

Advantages of ONT long-read sequencing in GF detection

While conventional short-read NGS fusion panels have become the standard for GF detection, ONT long-read sequencing offers

distinct advantages, particularly when working with challenging clinical samples. The Flongle flow cell provides quicker turnaround time, while PromethION offers higher sensitivity and a greater amount of data. In our study, we sequenced all samples for the full allotted time, 24 h for Flongle (cohort 1) and 72 h for PromethION (cohort 1 subset). Even with just 8 h for Flongle and 48 h for PromethION, sequencing proficiency for the short-read fusion library samples was sufficient for GF detection, as outlined in Tables S7A and S7B. These results highlight that Flongle, with less than 3 h of sample preparation and a minimum of 8 h of sequencing, followed by 1–2 h of data analysis, allows for comprehensive GF detection within a single day, as illustrated in Figure 3. In contrast, the PromethION flow cell, with its higher throughput capabilities (up to 290 Gb of data), provides a greater data yield at the cost of longer sequencing times. While the sample preparation time remains the same, the sequencing time can be completed within a minimum of 48 h, with an additional 24–26 h for downstream analysis. However, by multiplexing multiple samples prior to PromethION flow cell sequencing, we were able to effectively increase and optimize flow cell resources, while processing more samples simultaneously. In cohort 2, the sample sequencing time was increased to 96 h; however, from Table S7C, we can see that the sequencing proficiency for the long-read libraries

was sufficient for GF detection after 72 h. While this is a longer sequencing time, similarly, multiplexing multiple samples for true long-read sequencing warrants additional time to optimize flow cell resources.

The CHOP Cancer Fusion Panel requires 14 to 21 days for sample pooling, library preparation, sequencing, and fusion analysis.³⁶ This prolonged timeline is influenced by the need to pool samples to optimize sequencing resources and, in some cases, requires a more individualized approach for a quicker response time. ONT long-read sequencing mitigates these delays by offering faster and flexible approaches with Flongle flow cell sequencing. With ONT and proper downstream analysis workflows, diagnostic outcomes can be determined within just a couple of days, significantly reducing turnaround time and improving overall efficiency.

The robustness of PromethION flow cell sequencing was further demonstrated by its ability to accommodate variability in tissue sample collection type, as outlined in Table S1. It handled the challenges associated with formalin-fixed paraffin-embedded (FFPE) samples, which often result in degradation and compromised nucleic acid quality. In cohort 1, samples 21, 25, 26, and 28 were originally prepared using FFPE sample collection and previously analyzed with short-read targeted sequencing. Among these, we successfully confirmed the two fusion panel-negative samples (21 and 25) and identified the expected fusion gene (*ESW1::FLI1*) in the fusion panel-positive sample 28 using ONT sequencing. However, in sample 28, the fusion was detected only by LongGF with limited read support (Table 1). Overall, several GFs identified from Flongle sequencing exhibited low read support, often fewer than five reads, which increased the risk of overlooking these fusions due to random molecule selection. For fresh/frozen samples, PromethION sequencing of cohort 1 subset samples improved fusion detection. Of the 17 fresh/frozen samples (excluding sample 9, where the expected GF was undetected), three samples (16, 18, and 27) failed to identify their expected causal fusion using Flongle but were successfully detected with PromethION. The remaining sequenced fresh/frozen tissue samples identified the expected GF in at least two of the GF detection programs across both sequencing approaches. This was not the case for sample 11 (*GOPC::ROS1*), which was detected by LongGF and JAFFAL in Flongle but only LongGF in PromethION sequencing. In contrast, the blood/bone marrow tissue samples demonstrated more consistent results across both sequencing platforms. All eight samples in this category successfully identified the expected causal fusion using Flongle sequencing; however, due to limited sample availability, five samples were not included in cohort 1 subset or sequenced on PromethION.

Moreover, ONT long-read sequencing proved to be highly advantageous for samples with total RNA available (cohort 2). True long-read sequencing allowed us to capture full-length transcripts, offering a more comprehensive view of the potential structural rearrangements. In contrast to traditional NGS fusion panels, which may have overlooked GFs due to their reliance on experimental enrichment methods, our approach directly used ONT sequencing to ensure that no filtering occurred at the library preparation stage. This eliminated the risk of losing important sequencing information over areas that may not have

seemed to be of interest initially but later revealed additional, validated chromosomal rearrangements. Furthermore, this approach provided the flexibility to computationally focus the analysis on specific genes or regions of interest throughout the analysis. This eliminated the need for additional patient sampling while enabling the detection of rare and low-frequency GFs, which are essential for enhancing cancer diagnosis accuracy and guiding downstream treatment planning.

Evaluation of the computational detection pipeline for GF detection

In our workflow, we developed an ensemble-based GF detection pipeline with custom filtering procedure, which plays a crucial role in accurately identifying GFs within our samples by integrating three complementary detection tools. The ensemble approach combined the advantages to each program and contributed additional layers of information. LongGF reported the locations of the genes involved in the GF on the input strand, the strandness of the individual genes, and the IDs of the supporting reads, while JAFFAL also offered strandness information and cross-referenced fusions with the Mitelman database or a unique curated list. FusionSeeker provided GF detection and supporting reads but lacks strandness information, which was assumed from the reference strandness. By combining these tools, we ensured robust GF detection, checked gene orientation via strandness and input positions where available, and processed the GFs through the GF criteria to ensure complete analysis.

Implications of long-read sequencing in cancer diagnosis

The integration of long-read sequencing, such as ONT's platform, into clinical cancer diagnostics holds significant promise for advancing the detection of GFs, which are critical biomarkers in various cancer types. Our study demonstrates that long-read sequencing provides a more comprehensive view of GFs, identifying those likely missed by traditional short-read sequencing approaches. This is particularly valuable in oncology, where GFs are often involved in tumorigenesis and can guide treatment strategies. Long-read sequencing holds the potential to not only detect expected and novel GFs but also provide enhanced resolution of complex genomic rearrangements. Ultimately, this technology has the power to advance personalized treatment strategies, particularly for cancers with poorly understood molecular mechanisms.

Long-read sequencing had noteworthy implications for fusion panel-negative samples, as in cohort 2, where it allowed for the detection of low-abundance, known, and potentially novel GFs that might have otherwise been overlooked by targeted short-read panels. The *MDM2::NUP107* fusion was reported in both Mitelman and ChimerDB, with cancer associations spanning lung adenocarcinoma, squamous cell lung cancer, and breast cancer, although gene breakpoints varied from those reported in the reference. Similarly, *RPL4::MAP3K13* and *ESR1::SYNE1* were linked to multiple myeloma and adenocarcinoma in Mitelman, respectively. The *MDM2::RAP1B* fusion, reported in ChimerDB from a glioblastoma patient (TCGA-06-0686-01A), had variable breakpoints compared to those documented in

the database. Given that this fusion was also reported in the glioblastoma sample, the fusion was selected for experimental validation. However, the fusion was not validated by PCR. This highlighted the need for primer redesign to ensure confirmation of this fusion and/or its reciprocal form. Importantly, these fusions involved genes that were initially targeted in the CHOP Cancer Fusion Panel. The identification of these fusions through long-read sequencing reaffirms the utility of this platform in advancing GF detection toward clinical applications.

In terms of low-abundance GFs, two validated fusions from sequencing pool 5 (*CLDND1::WRN* and *BAZ1A::PI3*) had read support below the predefined filtering threshold. Despite this, they were included in the experimental validation due to the lower overall sequencing depth of pool 5. These fusions were selected based on their exonic breakpoints, which are likely to produce a functional fusion protein with potential oncogenic properties. Given the reduced sequencing depth of pool 5, most fusions did not meet the read support threshold and had samples with zero fusions remaining. This was evident in the K562 control sample showing only one supporting read for an expected abundant GF. Nevertheless, the *CLDND1::WRN* and *BAZ1A::PI3* GFs, each with three supporting reads, were experimentally validated. This highlights that even with lower sequencing depths, long-read sequencing successfully detected and validated low-abundance GFs, reinforcing its utility in unbiased GF discovery in samples previously deemed fusion panel negative.

However, while the clinical implementation of long-read sequencing has yet to be fully incorporated, our study marks a significant step forward in this direction. The advantages of long-read sequencing include lower costs, shorter turnaround times (Figure 3), and a growing availability of bioinformatics tools and analysis pipelines. As long-read sequencing technologies continue to progress and additional workflows are developed, their combined integration into routine clinical practice could significantly enhance the precision of cancer diagnostics, particularly in cases involving rare or complex fusions.

Limitations of the study

The clinical cancer patient samples from cohort 1 were initially prepared for short-read sequencing, with fragment sizes ranging from 150 to 300 bp, where ONT sequencing did not fully exploit the benefits of long-read sequencing technology. This limitation may lead to the identification of potentially novel fusions that could be artifacts of the shorter read lengths.

Moving forward, incorporating long-read, full-length transcript sequencing for fusion positive cases would further enhance the accuracy of split-read alignments and reduce spurious alignment artifacts, thus minimizing false positive GFs. In this study, our long-read sequencing efforts primarily focused on fusion-negative cases and successfully identified additional GFs that may have been overlooked by previous analysis with the short-read targeted NGS fusion panel. However, sequencing fusion-positive samples directly with long-read sequencing could provide a more comprehensive evaluation of GF detection sensitivity and specificity.

Although our clinical sample cohort is limited, we aim to expand the application of our pipeline to include additional clinical cancer types with implicated GFs. This expansion

will not only contribute to the evolving landscape of GF research but also enhance its translational impact on personalized medicine and therapeutic strategies, using long-read sequencing. However, there is a clear need for better filtering methods to process the GF results, where incorporating machine learning algorithms could further improve GF interpretation for refinement. This approach is particularly valuable for detecting and prioritizing complex GFs autonomously rather than prolonged manual clinical interpretation. Overall, our study represents a significant advancement in GF detection and characterization and highlights that single-end molecule sequencing remains an asset toward improved GF detection in clinical settings.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Kai Wang (wangk@chop.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- ONT long-read sequencing data for both sample cohorts have been deposited at the NCBI Sequence Read Archive (SRA) under BioProject: PRJNA1087427 (cohort 1) and BioProject: PRJNA1267462 (cohort 2), which are publicly available as of the date of publication.
- All original workflow code is available on GitHub at <https://github.com/WGLab/Gene-Fusion-Detection-Pipeline-LRS>. An archival DOI is included in the [key resources table](#).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

ACKNOWLEDGMENTS

The authors would like to thank our collaborators at the Division of Genomic Diagnostics (DGD) at CHOP and the University of Pennsylvania. In addition, we would like to thank current and former members of the Wang lab, Qian Liu, Li Fang and Hu Yu, for their technical assistance and valuable comments and feedback. We are grateful for the technical consultation from the IDRC Biostatistics and Data Science core (HD105354). This study is supported in part by NIH grant GM132713 and HG013359, an NIH/NHGRI T32 training grant on computational genomics (HG000046), and a CHOP Omics Initiative grant.

AUTHOR CONTRIBUTIONS

Conceptualization, K.W., M.L., Y.S., and K.R.; clinical sample collection and short-read analysis, F.X.; methodology, K.W., M.L., Y.S., K.R., M.U.A., and J.C.; GF detection pipeline development, K.R., M.U.A., and J.C.; GF analysis, K.R. and M.U.A.; sequencing and experimental validation, J.C., H.M.D., Z.L., and Y.S.; comparisons to DGD original fusion results, K.R., F.X., M.U.A., M.L., and K.W.; data curation, K.R., M.U.A., J.C., H.M.D., and Z.L.; writing – original draft, K.R.; writing – review & editing, K.R., H.M.D., M.U.A., J.C., F.X., Z.L., Y.S., M.L., and K.W.; figure creation, K.R.; all authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
 - Clinical cancer patient samples and CHOP Cancer Fusion Panel analysis
- **METHOD DETAILS**
 - CHOP Cancer Fusion Panel Illumina library preparation and short-read sequencing
 - ONT Library Preparation and Long-Read sequencing of clinical cancer cohorts
 - GF detection and analysis pipeline
 - GF experimental validation
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.crmeth.2025.101111>.

Received: September 26, 2024

Revised: April 28, 2025

Accepted: June 24, 2025

Published: July 21, 2025

REFERENCES

1. Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* 7, 233–245. <https://doi.org/10.1038/nrc2091>.
2. Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* 15, 371–381. <https://doi.org/10.1038/nrc3947>.
3. Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F.J., Reschneider, P., Garvin, T., Fang, H., Gurtowski, J., Hutton, E., et al. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* 28, 1126–1135. <https://doi.org/10.1101/gr.231100.117>.
4. Singh, A., Zahra, S., Das, D., and Kumar, S. (2019). AtFusionDB: a database of fusion transcripts in Arabidopsis thaliana. *Database* 2019, 135. <https://doi.org/10.1093/database/bay135>.
5. de Klein, A., van Kessel, A.G., Grosveld, G., Bartram, C.R., Hagemeijer, A., Bootsma, D., Spurr, N.K., Heisterkamp, N., Groffen, J., and Stephenson, J.R. (1982). A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* 300, 765–767. <https://doi.org/10.1038/300765a0>.
6. Rowley, J.D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* 243, 290–293. <https://doi.org/10.1038/243290a0>.
7. Robinson, D.R., Kalyana-Sundaram, S., Wu, Y.-M., Shankar, S., Cao, X., Ateeq, B., Asangani, I.A., Iyer, M., Maher, C.A., Grasso, C.S., et al. (2011). Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.* 17, 1646–1651. <https://doi.org/10.1038/nm.2580>.
8. Tognon, C., Knezevich, S.R., Huntsman, D., Roskelley, C.D., Melnyk, N., Mathers, J.A., Becker, L., Carneiro, F., MacPherson, N., Horsman, D., et al. (2002). Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell* 2, 367–376. [https://doi.org/10.1016/s1535-6108\(02\)00180-0](https://doi.org/10.1016/s1535-6108(02)00180-0).
9. Veeraraghavan, J., Ma, J., Hu, Y., and Wang, X.-S. (2016). Recurrent and pathological gene fusions in breast cancer: current advances in genomic discovery and clinical implications. *Breast Cancer Res. Treat.* 158, 219–232. <https://doi.org/10.1007/s10549-016-3876-y>.
10. Kumar-Sinha, C., Tomlins, S.A., and Chinnaiyan, A.M. (2008). Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* 8, 497–511. <https://doi.org/10.1038/nrc2402>.
11. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310, 644–648. <https://doi.org/10.1126/science.1117679>.
12. Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S.I., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448, 561–566. <https://doi.org/10.1038/nature05945>.
13. Druker, B.J. (2008). Translation of the Philadelphia chromosome into therapy for CML. *Blood* 112, 4808–4817. <https://doi.org/10.1182/blood-2008-07-077958>.
14. Braun, T.P., Eide, C.A., and Druker, B.J. (2020). Response and resistance to BCR-ABL1-targeted therapies. *Cancer Cell* 37, 530–542. <https://doi.org/10.1016/j.ccell.2020.03.006>.
15. Marcus, L., Donoghue, M., Aungst, S., Myers, C.E., Helms, W.S., Shen, G., Zhao, H., Stephens, O., Keegan, P., and Pazdur, R. (2021). FDA approval summary: entrectinib for the treatment of NTRK gene fusion solid tumors. *Clin. Cancer Res.* 27, 928–932. <https://doi.org/10.1158/1078-0432.CCR-20-2771>.
16. Gatalica, Z., Xiu, J., Swensen, J., and Vranic, S. (2019). Molecular characterization of cancers with NTRK gene fusions. *Mod. Pathol.* 32, 147–153. <https://doi.org/10.1038/s41379-018-0118-3>.
17. Ling, Q., Li, B., Wu, X., Wang, H., Shen, Y., Xiao, M., Yang, Z., Ma, R., Chen, D., Chen, H., et al. (2018). The landscape of NTRK fusions in Chinese patients with solid tumor. *Ann. Oncol.* 29, viii.22–viii.23. <https://doi.org/10.1093/annonc/mdy269.073>.
18. Duncavage, E.J., Schroeder, M.C., O’Laughlin, M., Wilson, R., MacMillan, S., Bohannon, A., Kruchowski, S., Garza, J., Du, F., Hughes, A.E.O., et al. (2021). Genome sequencing as an alternative to cytogenetic analysis in myeloid cancers. *N. Engl. J. Med.* 384, 924–935. <https://doi.org/10.1056/NEJMoa2024534>.
19. Heydt, C., Wölwer, C.B., Velazquez Camacho, O., Wagener-Rydzek, S., Pappesch, R., Siemanowski, J., Rehker, J., Haller, F., Agaimy, A., and Worm, K. (2021). Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation. *BMC Med. Genom.* 14, 1–14. <https://doi.org/10.1186/s12920-021-00909-y>.
20. Heyer, E.E., Deveson, I.W., Wooi, D., Selinger, C.I., Lyons, R.J., Hayes, V. M., O’Toole, S.A., Ballinger, M.L., Gill, D., Thomas, D.M., et al. (2019). Diagnosis of fusion genes using targeted RNA sequencing. *Nat. Commun.* 10, 1388. <https://doi.org/10.1038/s41467-019-09374-9>.
21. Musich, R., Cadle-Davidson, L., and Osier, M.V. (2021). Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Front. Plant Sci.* 12, 657240. <https://doi.org/10.3389/fpls.2021.657240>.
22. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21, 597–614. <https://doi.org/10.1038/s41576-020-0236-x>.
23. Latysheva, N.S., and Babu, M.M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.* 44, 4487–4503. <https://doi.org/10.1093/nar/gkw282>.
24. Warburton, P.E., and Sebra, R.P. (2023). Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. *Annu. Rev. Genomics Hum. Genet.* 24, 109–132. <https://doi.org/10.1146/annurev-genom-101722-103045>.
25. Marx, V. (2023). Method of the year: long-read sequencing. *Nat. Methods* 20, 6–11. <https://doi.org/10.1038/s41592-022-01730-w>.

26. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 30. <https://doi.org/10.1186/s13059-020-1935-5>.
27. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. <https://doi.org/10.1126/science.1162986>.
28. Payne, A., Holmes, N., Rakyen, V., and Loose, M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 35, 2193–2198. <https://doi.org/10.1093/bioinformatics/bty841>.
29. Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 239. <https://doi.org/10.1186/s13059-016-1103-0>.
30. Kolmogorov, M., Billingsley, K.J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., Asri, M., Alvarez Jerez, P., Malik, L., Dewan, R., et al. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* 20, 1483–1492. <https://doi.org/10.1038/s41592-023-01993-x>.
31. Sahlin, K., and Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* 12, 2. <https://doi.org/10.1038/s41467-020-20340-8>.
32. Zhang, H., Jain, C., and Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. *BMC Genom.* 21, 889. <https://doi.org/10.1186/s12864-020-07227-0>.
33. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Front. Genet.* 10, 426. <https://doi.org/10.3389/fgene.2019.00426>.
34. Oehler, J.B., Wright, H., Stark, Z., Mallett, A.J., and Schmitz, U. (2023). The application of long-read sequencing in clinical settings. *Hum. Genomics* 17, 73. <https://doi.org/10.1186/s40246-023-00522-3>.
35. Chang, F., Lin, F., Cao, K., Surrey, L.F., Aplenc, R., Bagatell, R., Resnick, A.C., Santi, M., Storm, P.B., Tasian, S.K., et al. (2019). Development and clinical validation of a large fusion gene panel for pediatric cancers. *J. Mol. Diagn.* 21, 873–883. <https://doi.org/10.1016/j.jmoldx.2019.05.006>.
36. CHOP (2023). Children's Hospital of Philadelphia Pathology & Laboratory Medicine: Comprehensive Hematologic Cancer Panel. <https://www.testmenu.com/chop/Tests/786447>.
37. Liu, Q., Hu, Y., Stucky, A., Fang, L., Zhong, J.F., and Wang, K. (2020). LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genom.* 21, 793. <https://doi.org/10.1186/s12864-020-07207-4>.
38. Davidson, N.M., Chen, Y., Sadras, T., Ryland, G.L., Blombery, P., Ekert, P. G., Göke, J., and Oshlack, A. (2022). JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol.* 23, 10–20. <https://doi.org/10.1186/s13059-021-02588-5>.
39. Chen, Y., Wang, Y., Chen, W., Tan, Z., Song, Y., Human Genome Structural Variation Consortium; Chen, H., and Chong, Z. (2023). Gene fusion detection and characterization in long-read cancer transcriptome sequencing data with FusionSeeker. *Cancer Res.* 83, 28–33. <https://doi.org/10.1158/0008-5472.CAN-22-1628>.
40. Glinos, D.A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K.L., Garimella, K., et al. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 608, 353–359. <https://doi.org/10.1038/s41586-022-05035-y>.
41. Mitelman, F., Johansson, B., and Mertens, F. (2023). Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. <https://mitelman-database.isb-cgc.org>.
42. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 47, D941–D947. <https://doi.org/10.1093/nar/gky1015>.
43. Jang, Y.E., Jang, I., Kim, S., Cho, S., Kim, D., Kim, K., Kim, J., Hwang, J., Kim, S., Kim, J., et al. (2020). ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res.* 48, D817–D824. <https://doi.org/10.1093/nar/gkz1013>.
44. Perdomo, J.E., Ahsan, M.U., Liu, Q., Fang, L., and Wang, K. (2025). Long-ReadSum: A fast and flexible quality control and signal summarization tool for long-read sequencing data. *Comput. Struct. Biotechnol. J.* 27, 556–563. <https://doi.org/10.1016/j.csbj.2025.01.019>.
45. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. *Nucleic Acids Res.* 49, D916–D923. <https://doi.org/10.1093/nar/gkaa1087>.
46. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
47. Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., and Brooks, A.N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* 11, 1438. <https://doi.org/10.1038/s41467-020-15171-6>.
48. Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., et al. (2019). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. Preprint at bioRxiv. <https://doi.org/10.1101/672931>.
49. Hu, Y., Fang, L., Chen, X., Zhong, J.F., Li, M., and Wang, K. (2021). LIQA: long-read isoform quantification and analysis. *Genome Biol.* 22, 182. <https://doi.org/10.1186/s13059-021-02399-8>.
50. Spoor, J.K.H., den Braber, M., Dirven, C.M.F., Pennycuik, A., Bartkova, J., Bartek, J., van Dis, V., van den Bosch, T.P.P., Leenstra, S., and Venkatesan, S. (2023). Investigating chromosomal instability in long-term survivors with glioblastoma and grade 4 astrocytoma. *Front. Oncol.* 13, 1218297. <https://doi.org/10.3389/fonc.2023.1218297>.
51. Richardson, T.E., Walker, J.M., Hambardzumyan, D., Brem, S., Hatanpaa, K.J., Viapiano, M.S., Pai, B., Umphlett, M., Becher, O.J., Snuderl, M., et al. (2024). Genetic and epigenetic instability as an underlying driver of progression and aggressive behavior in IDH-mutant astrocytoma. *Acta Neuropathol.* 148, 5. <https://doi.org/10.1007/s00401-024-02761-7>.
52. Perez, G., Barber, G.P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., et al. (2025). The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* 53, D1243–D1249. <https://doi.org/10.1093/nar/gkae974>.
53. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. j.* 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>.
54. Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664. <https://doi.org/10.1101/gr.229202>.
55. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115. <https://doi.org/10.1093/nar/gks596>.
56. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J.P. (2011). Integrative Genomics Viewer. *Nat. Biotechnol.* 29, 24–26. <https://doi.org/10.1038/nbt.1754>.
57. Nattestad, M., Aboukhalil, R., Chin, C.-S., and Schatz, M.C. (2021). Ribbon: intuitive visualization for complex genomic variation. *Bioinformatics* 37, 413–415. <https://doi.org/10.1093/bioinformatics/btaa680>.
58. Oxford Nanopore Technologies. Guppy. <https://community.nanoporetech.com>.
59. Oxford Nanopore Technologies. Dorado. <https://github.com/nanoporetech/dorado>.
60. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. <https://doi.org/10.1038/s41568-018-0060-1>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human tumor tissue RNA samples	Children's Hospital of Philadelphia	N/A
Deposited data		
Nanopore sequencing data	This paper	NCBI: PRJNA1087427 and NCBI: PRJNA1267462
Oligonucleotides		
Primers for candidate novel gene fusions, see Table S6	This paper	N/A
Software and algorithms		
Guppy	Oxford Nanopore Technologies	https://community.nanoporetech.com
Dorado	Oxford Nanopore Technologies	https://github.com/nanoporetech/dorado
Minimap2	Li et al. ⁴⁶	https://github.com/lh3/minimap2
LongReadSum	Perdomo et al. ⁴⁴	https://github.com/WGLab/LongReadSum
Cutadapt	Martin et al. ⁵³	https://github.com/marcelm/cutadapt?tab=readme-ov-file
LongGF	Liu et al. ³⁷	https://github.com/WGLab/LongGF
JAFFAL	Davidson et al. ³⁸	https://github.com/Oshlack/JAFFA/tree/master
FusionSeeker	Chen et al. ³⁹	https://github.com/Maggi-Chen/FusionSeeker
BLAST	Kent et al. ⁵⁴	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Primer3	Untergasser et al. ⁵⁵	https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome
Integrative Genome Viewer	Robinson et al. ⁵⁶	https://igv.org/
Genome Ribbon	Nattestad et al. ⁵⁷	https://v2.genomeribbon.com/
UCSC Genome Browser	Perez et al. ⁵²	https://genome.ucsc.edu/
Other		
Custom gene fusion detection panel-based and whole transcriptome-based pipelines	This paper	https://doi.org/10.5281/zenodo.15617148

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Clinical cancer patient samples and CHOP Cancer Fusion Panel analysis

This study analyzed two cohorts of clinical cancer patient samples obtained from the Division of Genomic Diagnostics (DGD) at the Children's Hospital of Philadelphia (CHOP). As outlined in Figure 1A, these samples were previously analyzed using the CHOP Cancer Fusion Panel with the Illumina short-read sequencing platform,³⁵ which resulted in both panel-positive and negative cases. Panel-positive samples identified a disease-causal GF, while panel-negative samples did not exhibit any detectable disease-causal GFs. This dynamic fusion panel targets over 700 exons of 117 cancer genes (updated to 119 genes) known to be involved in GFs. It employs anchored multiplex one-sided PCR to amplify sequences containing fusions involving these specific genes, as listed on CHOP's test menu website.³⁶ The CHOP Cancer Fusion Panel was previously validated in the Journal of Molecular Diagnostics,³⁵ where characterized samples with known fusion transcripts were evaluated. This study was performed under the protocol approved by CHOP.

Sample Cohort 1 consisted of 29 samples, with 27 panel-positive and 2 panel-negative samples. Cohort 1 represented a range of cancer types, including leukemia (#1–7, 24), CNS tumors (#8–19, 27, 29), and non-CNS tumors (#20–23, 25, 26, 28), sourced from bone marrow, peripheral blood, fresh/frozen tissues (FT), and formalin-fixed paraffin-embedded (FFPE) tissues. Notably, two non-CNS tumor samples (#21, 25) were panel-negative. Cohort 2 consisted of all panel-negative samples from patients with high- and low-grade gliomas, primarily from FT CNS tumors. Detailed information about cohort composition, sample origin, cancer type, and clinical indications are provided in Table S1. Prior to analysis, all samples were de-identified of any patient information.

METHOD DETAILS

This study employed a comprehensive workflow utilizing Oxford Nanopore Technologies (ONT) long-read sequencing platform, combined with bioinformatics tools to detect diagnostically relevant, disease-causal, gene fusions (GFs) and validate potentially novel ones. An overview of the entire workflow is presented in [Figure 1](#) and described below, detailing each step of the approach from the initial analysis using the Children's Hospital of Philadelphia (CHOP) Cancer Fusion Panel, through long-read sequencing, GF detection and filtering, to experimental validation. Together, these methods aim to enhance our understanding of GFs and provide a foundation for their incorporation into future diagnostic and research applications. The workflow code is available on GitHub (<https://github.com/WGLab/Gene-Fusion-Detection-Pipeline-LRS>).

CHOP Cancer Fusion Panel Illumina library preparation and short-read sequencing

All samples were simultaneously prepared for CHOP Cancer Fusion Panel processing and Illumina short-read sequencing on the Hi-Seq 2500 platform, performed by the DGD at CHOP. The panel includes a unique molecular barcode, two indexes for HiSeq deduplication, and P5/P7 Illumina adapters, ensuring compatibility with flow cells and multiplexing. The specific indexes were recognized by the Illumina HiSeq program for downstream demultiplexing and error correcting purposes, as outlined in the Archer protocol (Archer Universal RNA Reagent Kit v2 for Illumina-8). The Illumina short-read cDNA libraries, used directly as sample input for Cohort 1 and its subset, were reported to be on average 150-300bps in length.

ONT Library Preparation and Long-Read sequencing of clinical cancer cohorts

Cohort 1

The 29 samples in Cohort 1 were prepared for ONT long-read sequencing ([Figure 1B](#)) directly from the previous Illumina-prepared libraries, which restricted the resulting sequenced read lengths to those typical of short-read sequencing (~150–300 bp). 29 individual long-read libraries were then generated using the ONT ligation sequencing kit (SQK-LSK110) and sequenced on individual Flongle flow cells ([Figure 1B](#), **blue arrows**) using our in-house GridION sequencer for 24 h.

Of the initial 29 samples, 24 with sufficient remaining material and designated as Cohort 1 Subset, were prepared for ONT library preparation using the Native Barcoding Kit V14 (SQK-NBD114.24) for multiplexing on a PromethION flow cell ([Figure 1B](#), **purple arrows**). The barcoded samples were pooled and sequenced on a PromethION R10.4.1 flow cell using our in-house P2 Solo sequencer for 72 h. After both sequencing approaches, internal basecalling was performed using ONT's MinKnow software to generate the initial raw sequence data for further analysis. Sample input concentrations and sequencing data generated from both Flongle and PromethION runs are summarized in [Tables S2A](#) and [S2B](#), respectively. Detailed sequencing metrics, including data generation over various timepoints for both flow cell types, are provided in [Tables S7A](#) and [S7B](#).

Cohort 2

A total of 15 additional samples, comprising Cohort 2, were classified as panel-negative after CHOP Cancer Fusion Panel processing ([Figure 1A](#)). The 15 samples in Cohort 2 were prepared for ONT long-read sequencing directly from the isolated total RNA ([Figure 1B](#), **orange arrows**). The samples were processed using cDNA-PCR sequencing, which involved ONT's cDNA-PCR Sequencing V14 Barcoding Kit (SQK-PCB114.24) protocol. Modifications included 250 ng of input total RNA instead of the recommended 500 ng and adjusting the beads ratio in the final cleanup step from 0.7 × to 0.4 ×. All remaining steps were followed according to the protocol.

After ONT library preparation, the samples were pooled, and multiplex sequenced on an R10.4.1 PromethION flow cell using our in-house P2 Solo sequencer for 72 h. Similarly, after sequencing, internal basecalling was performed using ONT's software. [Table S2C](#) summarizes the sample input concentrations and sequencing data generated from Cohort 2, and [Table S7C](#) provides detailed sequencing generation metrics at various timepoints for the PromethION flow cell sequencing.

GF detection and analysis pipeline

The GF detection and analysis pipeline, illustrated in [Figure 1C](#), was used to analyze the long-read sequenced data from all sample cohorts. This pipeline integrates multiple computational tools to ensure the detection of all expected disease-causal GFs in panel-positive cases and identify additional potentially novel GFs in panel-negative cases. To evaluate the pipeline's effectiveness, all samples were de-identified with respect to their fusion status and expected disease-causal GFs prior to analysis, facilitating a blind GF detection process. As shown in [Figure 1C](#), the pipeline begins with re-basecalling with using a super high-accuracy model (represented by a diamond shape). From this point, users can select one of two analysis approaches: the CHOP Cancer Fusion Panel GF pipeline (depicted in pink) or the Long-Read Whole Transcriptome GF pipeline (depicted in orange). Shared components between both pipelines are displayed in black in the central lane. Cohort 1 and Cohort 1 Subset were analyzed using the CHOP Cancer Fusion Panel GF pipeline, while Cohort 2 was analyzed using the long-read whole transcriptome GF pipeline.

Re-basecalling with super high accuracy model

Cohort 1 and its subset were subjected to additional external basecalling using the Guppy⁵⁸ (v6.5.7) basecaller to generate high-confidence FASTQ files with a minimum base quality Q-score threshold of 7. At the time of this study, Guppy was the most readily available basecaller and Guppy⁵⁸ v6.5.7 (model dna_r9.4.1_450bps_sup.cfg) is equivalent to the more recent Dorado⁵⁹ basecalling model (dna_r9.4.1_e8_sup@v3.3). Cohort 2's external basecalling was subsequently performed with Dorado.⁵⁹

Illumina adapter removal

Illumina adapter trimming was performed using Cutadapt⁵³ to remove adapter sequences from the previous Illumina library preparation for all samples in Cohort 1 and its subset. Both 5' (Forward: ACACTCTTCCCTACACGACGCTCTCCGATCT, Reverse: TGACTGGAGTTCAGACGTGTGCTCTCCGATCT) and 3' (Forward: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA, Reverse: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) adapters were targeted for removal. To ensure complete adapter removal, Cutadapt was applied in two stages. First, we employed adapter trimming around the region of interest, as specified by the adapter sequences, and retained the sequence between adapters. The maximum allowed error rate during adapter matching was set to 30% to accommodate potential sequencing errors. Reads with fewer than 10 occurrences of a specified adapter were trimmed, generating both trimmed and untrimmed files. In the second stage, the untrimmed output was used as input for additional individual adapter trimming. Specific adapter sequences were provided, allowing a 30% error rate, and reads with less than 20 occurrences of the specified adapter were discarded. The resulting trimmed reads from each stage were combined. The success of the adapter trimming process was assessed using various metrics outlined in Tables S3A and S3B, for both Flongle (Cohort 1) and PromethION (Cohort 1 Subset) sequenced samples, respectively.

Alignment to the reference genome and quality assessment

The remaining reads were aligned to the GRCh38 reference genome (no alternate contigs) using Minimap2⁴⁶ (v2.24), with the supplied reference annotation from GENCODE⁴⁵ (v44). Alignment quality for each sample cohort was assessed using LongReadSum (v1.3.0) and is presented in Tables S4A–S4C for Cohort 1, Cohort 1 Subset, and Cohort 2, respectively. The totals and proportions of unclassified bases and reads were included, where unclassified refers to the reads that were not confidently assigned to specific samples during the demultiplexing process. To evaluate the sequencing depth across Flongle and PromethION sequenced samples and its impact on GF detection, coverage was calculated as the total number of bases sequenced divided by the sum of the lengths of the longest transcript of each gene targeted by the CHOP Cancer Fusion Panel.

GF detection and filtering criteria

GFs were detected using a combination of three long-read GF detection programs: LongGF,³⁷ JAFFAL³⁸ (v2.3), and FusionSeeker³⁹ (v1.0.1). For LongGF, multiple detection parameters were employed and included: minimum of two supporting reads, minimum of 10, 25, 50, 75, and 100 bp alignment record overlap with a gene, minimum 10 and 25 bin size, and a minimum alignment record against the reference genome of 10, 25, 50, 75, and 100 bps. The multiple parameters were implemented to enhance GF detection sensitivity, particularly in Cohort 1 and Cohort 1 Subset with shorter initial read lengths. The resulting GFs were a combination of the unique GFs reported across all parameters, prioritizing those with a higher number of reads if multiple instances of the same GF existed. Fusion orientation was verified after LongGF reporting by comparing the input strand information, reference strand orientation, and gene positions for each individual gene involved in the fusion. JAFFAL was set to its “Long” parameters for single-end ONT reads, and FusionSeeker was set to the Nanopore data type with GENCODE⁴⁵ (v44) reference annotation and the GRCh38 reference human genome supplied.

To refine and identify confident GFs detected in all sample cohorts, a set of shared filtering criteria were applied across both the CHOP Cancer Fusion Panel and the Whole Transcriptome analysis pipelines. These criteria were specifically used in the CHOP Cancer Fusion Panel analysis, while an additional filtering criterion and annotation was introduced in the Whole Transcriptome analysis pipeline. (1) Presence of Gene(s) of Interest: At least one gene in the fusion had to involve one of the 119 CHOP Cancer Fusion Panel genes³⁵ for Cohort 1 and its subset. For Cohort 2, this list expanded to include the Cancer Gene Census.⁶⁰ (2) Minimum Read Support Threshold: The threshold for supporting reads was set to a minimum of 2 reads for Cohort 1, 10 reads for Cohort 1 Subset, and 4 reads for Cohort 2 samples. These thresholds were determined based on the sample cohort and flow cell type to accommodate shorter initial read lengths in Cohort 1 and its subset, relative to the volume of data generated. For Cohort 2, which consisted of fusion-negative samples, a slightly higher threshold of 4 reads was set to ensure robust detection of possibly overlooked GFs from previous analyses. It is important to note that lower frequency fusions were not excluded from the analysis entirely, as they could be attributed to potentially novel GFs; this criterion prioritized the most probable GFs while still accounting for the possibility of novel events. (3) Removal of Unrelated Genes: Fusions involving mitochondrial, ribosomal, HLA, and pseudogenes were excluded from the analysis. These gene groups are often not relevant for disease-associated GFs, particularly in cancer. (4) Strandness Consistency: GFs were checked for consistency between reported and reference strandness, from the GENCODE (v46) annotation, ensuring either a direct or complementary match with the reference, Figure S2. This check was primarily applied to GFs detected by LongGF, as JAFFAL and FusionSeeker by default did not provide sufficient strand and gene input read location information to make these comparisons. However, direct and complementary matches to the reference strandness were noted when identified. Figures S2A–S2D depict direct strandness matches between the reported and reference strandness, while S2E–H show GFs involving complementary gene strandness matches. S2I–J, outlined in red, highlight instances where the strandness of one gene did not align with its reference, indicating a fusion between a gene and the complementary sequence of the other gene, and were therefore not retained. While these examples highlight some of the common strandness conditions and mismatches, they are not exhaustive. (5) Cross Reference with Fusion Literature: GFs that passed these filtering criteria in all sample cohorts were then cross-referenced with online fusion databases, including Mitelman,⁴¹ COSMIC Fusion,⁴² and ChimerDB 4.0,⁴³ to identify any known disease-causing GFs within the results.

The additional criteria specific to the long-read whole transcriptome samples in Cohort 2 was: (6) Exclusion of GFs Detected in Healthy Brain Tissue Samples: GFs detected in healthy brain tissue samples were filtered out to remove commonly observed fusions in non-cancerous brain tissue. Healthy brain samples were obtained from the Genotype-Tissue Expression (GTEx) project (v9),

comprising 88 tissue and K562 cell line samples generated using the ONT sequencing platform (dbGAP accession number phs000424.v9).⁴⁰ Among these, 22 brain tissue samples were analyzed using the same fusion detection programs applied to Cohort 2: LongGF, JAFFAL, and FusionSeeker. This filtering step ensured that the remaining fusions identified in Cohort 2 were relevant to glioma cancer pathology. By using the GTEx brain tissue dataset as a healthy control, this approach effectively minimized the inclusion of GFs commonly present in normal brain tissue. Altogether, the GF filtering criteria helped to remove fusions that may have been attributed to potential PCR artifacts during library preparation or technical artifacts from alignment, a feature lacking in most detection programs. (7) Cross Reference Genes in GFs in Disease-Specific Fusions: To further prioritize GFs relevant to CNS pathologies in Cohort 2 samples, the remaining GFs were cross-referenced with genes involved in known fusions implicated in brain-related diseases using topography annotations from the Mitelman database.⁴¹ This process involved selecting known GFs with brain and brainstem topographies (tissues used in clinical investigations) and identifying their associated tumor morphology (tumor histology). For each remaining GF in Cohort 2, we evaluated whether either one of the gene had been previously reported as part of a known GF associated with a CNS-related disease. If a gene in the fusion was also reported in the CNS-related disease fusion (different fusion partner), the associated disease morphology was annotated for that gene. This allowed us to prioritize novel GFs that included genes already implicated in CNS-related diseases, providing additional biological and clinical context for their potential relevance. (8) Removal of Recurrent GFs: GFs that appeared repeatedly in more than 15% of samples across all cohorts were excluded to eliminate potential artifacts. Fusions that were known within literature or a reciprocal of a known GF were excluded from this filtering. The fusion criteria helped to filter through the multitude of fusions reported, listed in [Tables S5A–S5C](#) for Cohort 1, Cohort 1 Subset, and Cohort 2, respectively. (9) Visual Computational Validation in IGV and Genome Ribbon: All GFs, regardless of fusion literature database or CNS-related disease morphology overlap (Cohort 2), were evaluated using the Integrative Genome Viewer⁵⁶ (IGV, v2.16) and Genome Ribbon.⁵⁷ These evaluations focused on read alignment patterns, depth, and presence of split reads at the reported breakpoint locations. We prioritized GFs with higher read support, as true oncogenic GFs were expected to be supported by a larger number of reads compared to passenger or artifact fusions. If a known fusion was identified it was recorded as the likely disease-causal GF for each sample in Cohort 1 and its subset and was compared to the DGD Illumina short-read fusion analysis results as part of the pipeline evaluation process as a validation and a quality check of the computational GF detection pipeline. High-confidence GFs not found in databases and visually validated were considered for experimental validation as potentially novel GFs ([Figure 1D](#)), as shown in [Figure S3A](#) from Cohort 2.

GF experimental validation

A subset of GFs that met the fusion filtering criteria were identified as potentially novel and selected for experimental validation, as outlined in [Figure 1D](#). PCR primers were designed with Primer3⁵⁵ software, targeting the 200bp regions upstream and downstream of the reported fusion breakpoints. The uniqueness of the primer sequences was confirmed using BLAT,⁵⁴ a sequence alignment tool, to avoid false-positive amplification. PCR was performed with Invitrogen's Platinum II Hot-Start PCR Master Mix 2× (14000012), which includes Invitrogen's Platinum II Taq Hot-Start DNA Polymerase premixed with Invitrogen's Platinum II PCR buffer and dNTPs. For Cohort 2, the targeted GFs for validation, including breakpoint positions (GRCh38), primer sequences, primer information, PCR product length, and supporting read IDs, are provided in [Table S6](#). PCR for potentially novel GFs identified in Cohort 2 was performed with Invitrogen's AccuPrime Taq DNA Polymerase, High Fidelity (12346086) which contains AccuPrime Taq DNA Polymerase, High Fidelity (5 U/μL) and the 10X AccuPrime PCR Buffer I. The PCR reaction mixture for each sample contains 5 μL 10X AccuPrime PCR Buffer I, 1 μL each of the 10 μM forward and reverse primers, 1 μL of the template DNA, 0.2 μL of AccuPrime Taq DNA Polymerase, and 41.8 μL of nuclease-free water. The thermal cycler conditions for the reactions are as follows: initial denaturation at 94°C for 2 min, 35 cycles of denaturation at 94°C for 30 s, annealing at 57°C–61°C (depending on the primers) for 30 s and extension at 64°C for 1 min, and a hold at 4°C. HG002 was included as a negative control for each primer pair to indicate no off-target interactions. Representative gel electrophoresis results and corresponding Sanger sequencing traces of successful GF amplifications are shown in [Figures S3B](#) and [S3C](#), respectively.

QUANTIFICATION AND STATISTICAL ANALYSIS

Detailed procedures and thresholds for gene fusion detection are described in the sections above. No additional statistical tests were performed.