


METHODOLOGY ARTICLE

Open Access



Discovery of clinically relevant fusions in pediatric cancer

Stephanie LaHaye¹, James R. Fitch¹, Kyle J. Voytovich¹, Adam C. Herman¹, Benjamin J. Kelly¹, Grant E. Lammi¹, Jeremy A. Arbesfeld¹, Saranga Wijeratne¹, Samuel J. Franklin¹, Kathleen M. Schieffer¹, Natalie Bir¹, Sean D. McGrath¹, Anthony R. Miller¹, Amy Wetzel¹, Katherine E. Miller¹, Tracy A. Bedrosian¹, Kristen Leraas¹, Elizabeth A. Varga¹, Kristy Lee¹, Ajay Gupta², Bhuvana Setty^{2,3}, Daniel R. Boué^{4,5}, Jeffrey R. Leonard^{3,6}, Jonathan L. Finlay^{2,3}, Mohamed S. Abdelbaki^{2,3}, Diana S. Osorio^{2,3}, Selene C. Koo^{4,5}, Daniel C. Koboldt¹, Alex H. Wagner^{1,3,7}, Ann-Kathrin Einfeld^{8,9,10}, Krzysztof Mrózek^{9,10}, Vincent Magrini^{1,3}, Catherine E. Cottrell^{1,3,4}, Elaine R. Mardis^{1,3}, Richard K. Wilson^{1,3} and Peter White^{1,3*} 

Abstract

Background: Pediatric cancers typically have a distinct genomic landscape when compared to adult cancers and frequently carry somatic gene fusion events that alter gene expression and drive tumorigenesis. Sensitive and specific detection of gene fusions through the analysis of next-generation-based RNA sequencing (RNA-Seq) data is computationally challenging and may be confounded by low tumor cellularity or underlying genomic complexity. Furthermore, numerous computational tools are available to identify fusions from supporting RNA-Seq reads, yet each algorithm demonstrates unique variability in sensitivity and precision, and no clearly superior approach currently exists. To overcome these challenges, we have developed an ensemble fusion calling approach to increase the accuracy of identifying fusions.

Results: Our Ensemble Fusion (EnFusion) approach utilizes seven fusion calling algorithms: Arriba, CICERO, FusionMap, FusionCatcher, JAFFA, MapSplice, and STAR-Fusion, which are packaged as a fully automated pipeline using Docker and Amazon Web Services (AWS) serverless technology. This method uses paired end RNA-Seq sequence reads as input, and the output from each algorithm is examined to identify fusions detected by a consensus of at least three algorithms. These consensus fusion results are filtered by comparison to an internal database to remove likely artifactual fusions occurring at high frequencies in our internal cohort, while a “known fusion list” prevents failure to report known pathogenic events. We have employed the EnFusion pipeline on RNA-Seq data from 229 patients with pediatric cancer or blood disorders studied under an IRB-approved protocol. The samples consist of 138 central nervous system tumors, 73 solid tumors, and 18 hematologic malignancies or disorders. The combination of an ensemble fusion-calling pipeline and a knowledge-based filtering strategy identified 67 clinically relevant fusions among our cohort (diagnostic yield of 29.3%), including *RBPM5-MET*, *BCAN-NTRK1*, and *TRIM22-BRAF* fusions. Following clinical confirmation and reporting in the patient’s medical record, both known and novel fusions provided medically meaningful information.

* Correspondence: peter.white@nationwidechildrens.org

¹The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children’s Hospital, Columbus, OH, USA

³Department of Pediatrics, The Ohio State University, Columbus, OH, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Conclusions: The EnFusion pipeline offers a streamlined approach to discover fusions in cancer, at higher levels of sensitivity and accuracy than single algorithm methods. Furthermore, this method accurately identifies driver fusions in pediatric cancer, providing clinical impact by contributing evidence to diagnosis and, when appropriate, indicating targeted therapies.

Keywords: Transcriptomics, Genomics, Pediatric neoplasms, Gene fusions, Cancer, RNA-Seq

Background

Globally, there are approximately 300,000 pediatric and adolescent cases of cancer diagnosed each year [1, 2]. While advances in medicine have led to a drastic improvement in 5-year overall survival rates (up to 84% in children under 15), pediatric cancer remains the most common cause of death by disease in developed countries [3, 4]. Pediatric cancers are defined by a distinct genomic landscape when compared to adult cancers, which includes an overall low number of somatic single nucleotide variants, common driver fusions and epigenetic changes that drive a specific transcriptional program. Pediatric cancers are often considered embryonic in origin and demonstrate a significant germline predisposition component approaching 10% [5–7].

Many pediatric tumors contain gene fusions resulting from the juxtaposition of two genes (Additional File 1: Fig. S1) [6]. Fusions typically occur through chromosomal rearrangements, and often lead to dysregulated gene expression of one or both gene partners [8–11]. Fusions can also generate chimeric oncoproteins, wherein functional domains from both genes are retained, often leading to aberrant and strong activation of nonspecific downstream targets [12]. The alterations in gene expression and activation of downstream targets induced by fusions are considered to be oncogenic events in pediatric cancer and increasingly may indicate response to specific targeted therapies.

The identification of an oncogenic fusion can provide medically meaningful information in the context of diagnosis, prognosis, and treatment regimens in pediatric cancers. Fusions may provide diagnostic evidence for a specific histological subgroup. For example, *EWSR1-FLII* fusions are highly associated with Ewing sarcoma, while the presence of a *ZFTA-RELA* fusion aids in subgrouping supratentorial ependymomas [12]. The detection of certain fusions, such as *BCR-ABL* in acute lymphocytic leukemia, can be used as a surrogate for residual tumor load and treatment response [13]. Fusions may also provide prognostic indication, such as *KIAA1549-BRAF* in low grade astrocytomas, which have a more favorable outcome compared to non-*BRAF* fused tumors [14, 15]. In addition, fusions that involve kinases can present therapeutic targets, including *FGFR1-TACC1*, *FGFR3-TACC3*, *NPM1-ALK*, and *NTRK* fusions [2, 12, 16–19].

However, regardless of the clear clinical benefits of characterizing fusion events in a given patient's tumor, accurate identification of fusions from next generation sequencing DNA data alone is not straightforward and they often go undiscovered. In particular, many fusions are not detectable by exome sequencing (ES) due to breakpoint locations that frequently occur in non-coding or intronic regions which may not have corresponding capture probes. Even whole genome sequencing (WGS) NGS data has proved difficult to evaluate complex rearrangements resulting in gene fusions due to a high false positive rate and due to the limitations of short read lengths [20, 21]. By contrast, next-generation RNA sequencing data, or RNA-Sequencing (RNA-Seq), offers an unbiased data type suitable for fusion detection, while also providing information about the expression of fusion transcripts, including multiple isoforms, and fusions that occur due to aberrant splicing events [22, 23].

While RNA-Seq is a powerful tool for fusion detection, it is not without its limitations. Notably, there is currently a major deficit in our ability to accurately identify fusions in spite of having many computational approaches available. Here, consistently identifying gene fusion events with high sensitivity and precision using one algorithm is unlikely and this is of critical importance in a clinical diagnostic setting [12]. Computational approaches that have been tuned for high sensitivity are limited by also calling numerous false positives, requiring extensive manual review of data, while those with a low false discovery rate (FDR) often miss true positives due to over-filtering [12]. To overcome these complications of sensitivity and specificity, we have employed an ensemble pipeline, which merges results from seven algorithmic approaches to identify, filter and output prioritized fusion predictions.

Another common issue encountered in fusion prediction is the identification of likely non-pathogenic fusions, due both to read-through events and fusions occurring in non-disease involved (normal) genomes [12, 24, 25]. We addressed these sources of false positivity through the implementation of a filtering strategy that removes known normal fusions and RNA transcription read-through events, based on internal frequency of detection and location of chromosomal breakpoints. Lastly, to prevent over-filtering and inadvertent removal of previously described known pathogenic fusion events, we have

developed and continually update a list containing known pathogenic fusion partners, that will return any data-supported fusions to the output list of prioritized fusion results for further evaluation.

The ensemble fusion detection pipeline outperformed all single algorithm methods we evaluated, achieving high levels of sensitivity, while simultaneously minimizing false positive calls and non-clinically relevant fusion predictions. Here, we describe our ensemble fusion detection approach, which we have named EnFusion (EnFusion), its performance on commercial control reference standards with known fusions, and its implementation on a pediatric cohort consisting of rare, treatment refractory, or relapsed cancers and hematologic diseases, as well as a secondary clinical acute myeloid leukemia (AML) cohort. Utilization of EnFusion resulted in a diagnostic yield of approximately 30% in our cohort, identified novel fusion partners, and has provided diagnostic information and/or targeted treatment options for this patient population.

Results

Development and optimization of ensemble pipeline on a control reference standard

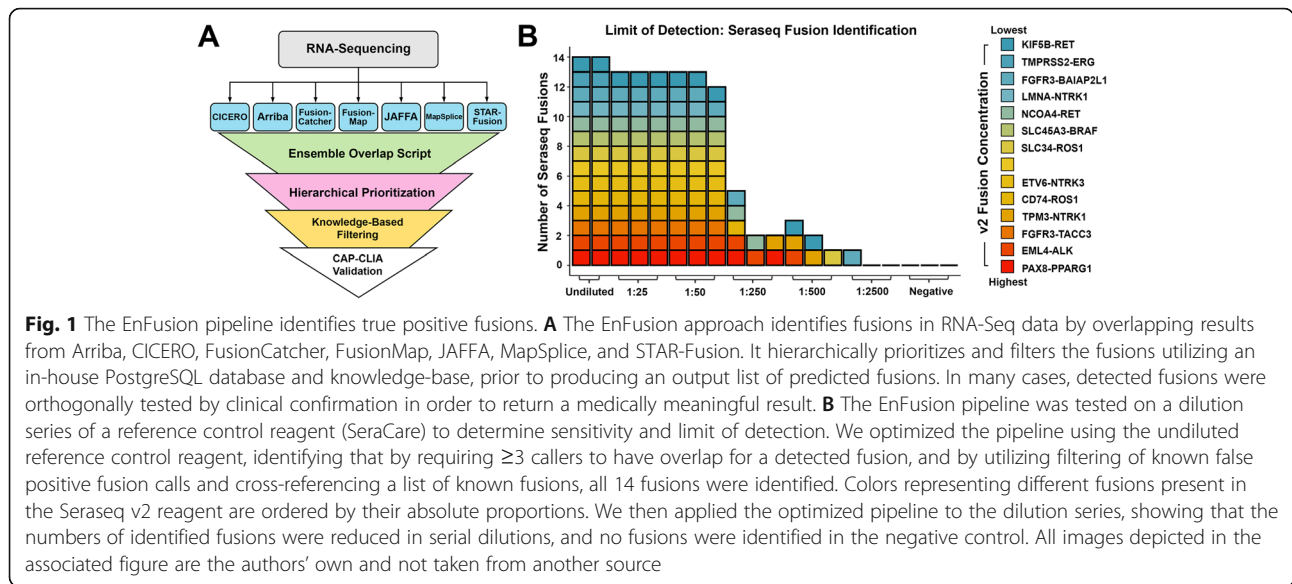
Identification of gene fusions through the use of a single algorithm is often associated with low specificity and poor precision [12]. Given prior literature supporting multi-algorithmic approaches to improve upon these deficits, we studied the intricacies of several fusion detection algorithms, and applied a defined set of algorithms with desired properties, aimed at detecting true positive fusions while minimizing false positive fusions [25–28]. After evaluating each algorithm's output, we developed the EnFusion pipeline that combines output consensus calls from seven different computational

approaches (Table 1, Fig. 1A), calculates the concordant fusion partners and breakpoints, and filters this output list based on internal frequency, reads of evidence, and breakpoint location. A list of known pathogenic fusions rescues any known pathogenic fusion gene partners with suitable algorithmic and read support for further evaluation (Additional File 1: Table S1).

To optimize the approach, we utilized a reference standard from a commercial provider (SeraSeq Fusion RNA, SeraCare, Milford, MA), containing synthetic RNAs representing 14 cancer-associated fusions in varying proportions (Additional File 1: Tables S2 and S3). Data generated from these RNA-Seq libraries, performed as replicates for a range of dilutions, were analyzed using the ensemble pipeline. We compared the output derived from a consensus of two or more callers to that from a consensus of three or more callers by calculating sensitivity ($\#$ of SeraSeq fusions identified)/(14 possible SeraSeq fusions), and precision ($\#$ of SeraSeq fusions identified)/($\#$ of total fusions identified) prior to filtering or known fusion list comparison. The undiluted reference standard with consensus of at least two callers, had a sensitivity of 100% and precision of 35%. Inclusion of the knowledgebase filtering step reduced the sensitivity to 85.7% while increasing the precision to 77.4%, and the known fusion list rescue step increased sensitivity to 100% and precision to 80% (Additional File 1: Table S4, Fig. S2A). By increasing the consensus requirement to three callers, rather than just two, the prefiltered sensitivity was 100% and precision was 90.3%. Inclusion of the filtering step reduced the sensitivity to 85.7% while increasing the precision to 100%, and known fusion list rescue increased sensitivity to 100% and precision to 100% (Table 2; Additional File 1: Fig. S2A). The inclusion of the known fusion list prevented the removal of

Table 1 Performance comparison of individual fusion calling algorithms. Fusion calling algorithms utilized by EnFusion and their contributions to fusion calling in the NCH pediatric cancer and hematologic disease cohort

Tool	Version	Aligner	Reference	Average fusions called per case	Sensitivity (clinically relevant fusions called out of 67)
Arriba	v1.2.0	STAR aligner	Uhrig et al., 2021 [29] Genome Res	54	88.1% (59)
CICERO	v0.3.0	candidate SV (structural variant) breakpoints and splice junction	Tian et al., 2020 [30] Genome Biol	1909	92.5% (62)
FusionMap	v mono- 2.10.9	GSNAP (Genomic Short-read Nucleotide Alignment Program) - 12mer based	Ge et al., 2011 Bioinformatics [31]	34	86.6% (58)
FusionCatcher	v0.99.7c	4 aligners to identify junctions (Bowtie, BLAT, STAR, and Bowtie2)	Nicorici et al., 2014 [32] bioRxiv	1554	89.6% (60)
JAFFA	direct v1.09	BLAT, uses kmers to select reads that do not map to known transcripts	Davidson et al., 2015 [33] Genome Med	1134	97.0% (65)
MapSplice	v2.2.1	approximate sequence alignment combined with a local search	Wang et al., 2010 [34] Nucleic Acids Res	37	85.1% (57)
STAR-Fusion	v1.6.0	STAR aligner	Haas et al., 2019 [25] Genome Biol	71	94.0% (63)



known Seraseq fusions, due to too few reads of evidence or number of callers providing support, as well as a single Seraseq fusion, *EML4-ALK*, which was present at an artificially high frequency in our database (31.8%) due to false positive calls by FusionCatcher. Implementation of the known fusion list led to sensitivity scores of 100% for both levels of caller consensus. The individual fusion detection algorithms ranged in sensitivity and precision, and while certain algorithms maintain high levels of sensitivity in addition to moderate levels of precision, such

as STAR-Fusion (sensitivity = 100%, precision = 43.8%), others such as FusionCatcher (sensitivity = 92.9%, precision = 4.3%) and CICERO (sensitivity = 100%, precision 1.1%) had high levels of sensitivity with very low precision levels (Table 2; Additional File 2: Table S4). When considering the overall results from undiluted and serial dilutions of the reference standard, the required overlap of at least three callers, with filtering and utilization of the known fusion list, led to significantly fewer total fusions identified compared to two consensus callers ($p =$

Table 2 Improved precision in fusion detection, utilizing Seraseq controls, achieved by EnFusion. Data shown is from undiluted Seraseq v3 RNA-Seq, experiments performed in duplicate, averages are shown. Individual algorithms are listed by precision, in descending order. Seraseq fusions identified (true positive) are out of a possible 14 fusions

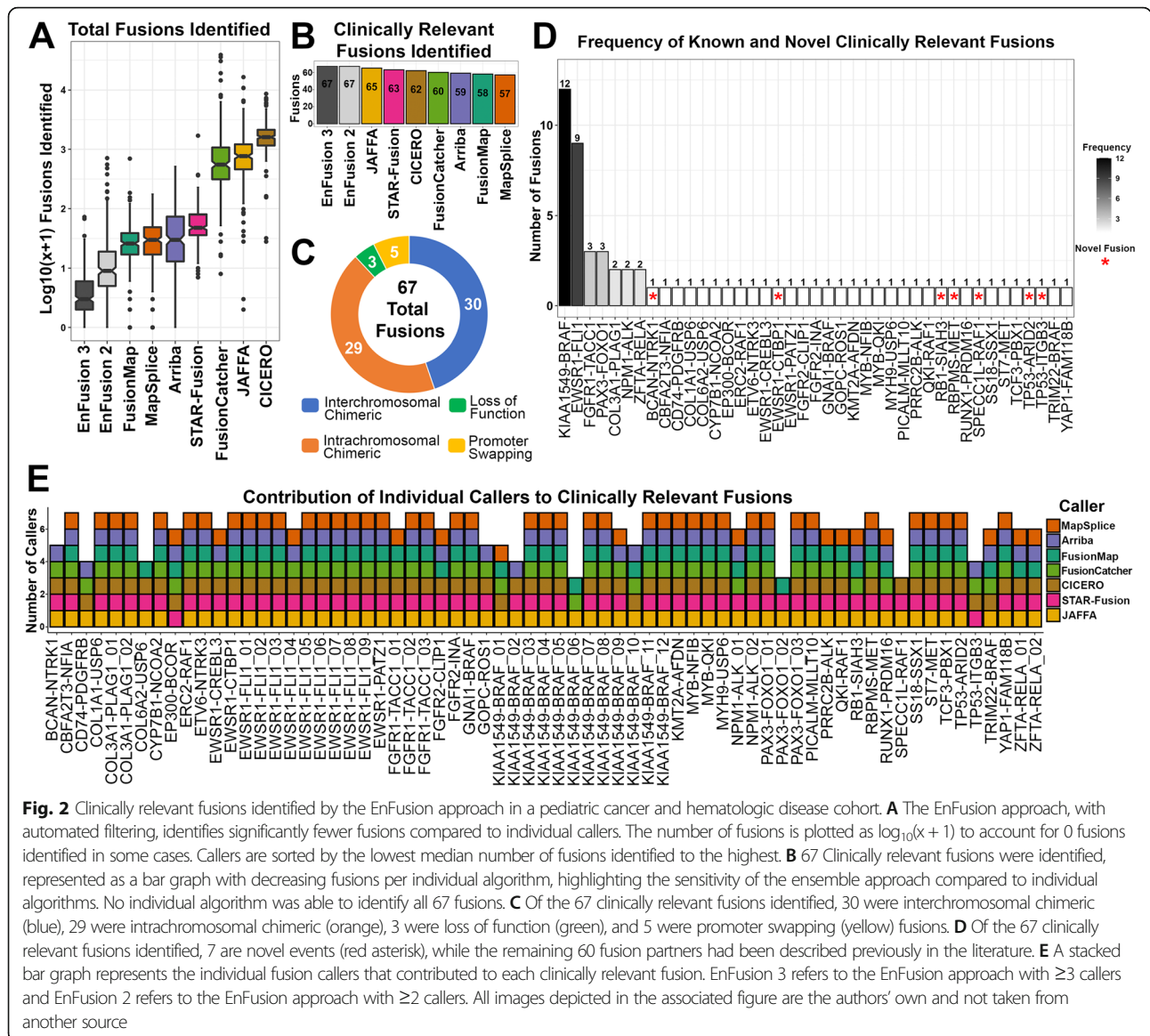
Algorithm	Total fusions identified	Seraseq fusions identified	Sensitivity	Precision
Arriba	23.5	13	92.9%	55.3%
MapSplice	22	12	85.7%	54.6%
STAR-Fusion	32	14	100.0%	43.6%
FusionMap	30	12.5	89.3%	41.7%
FusionCatcher	299.5	13	92.9%	4.3%
JAFFA	470.5	12.5	89.3%	2.7%
CICERO	1323	14	100.0%	1.1%
EnFusion 2 callers	40	14	100.0%	35.0%
EnFusion 2 callers + filter	15.5	12	85.7%	77.4%
EnFusion 2 callers + filter + known fusion list	17.5	14	100.0%	80.0%
EnFusion 3 callers	15.5	14	100.0%	90.3%
EnFusion 3 callers + filter	12	12	85.7%	100.0%
EnFusion 3 callers + filter + known fusion list	14	14	100.0%	100.0%

2.34E-07) (Table 2; Additional File 1: Fig. S2B, Table S5). The EnFusion results obtained from various reference standard dilutions, with a minimum of three callers in consensus, using filtering and known fusion list rescue are shown (Fig. 1B; Additional File 2: Table S4). The optimized EnFusion pipeline, consisting of a consensus of three callers, filtering, and the known fusion list, maintained high levels of sensitivity, (at least 90.5%), while maintaining 100% precision as low as the 1:500 dilution of the reference standard (Additional File 2: Table S4). In addition to the high levels of sensitivity and precision, the total number of fusions identified by this optimized EnFusion pipeline in undiluted and diluted samples was significantly fewer than the number identified by individual fusion detection algorithms, including STAR-Fusion ($p = 1.77E-12$), CICERO ($p = 3.39E-14$) and Fusion-Catcher ($p = 1.00E-08$) (Additional File 1: Table S5). These results highlight the removal of false positive fusions, which includes artifactual and benign fusion events, and subsequent reduction in manual evaluation requirements (Additional File 1: Fig. S2C, S2D). Notably, we only considered the 14 Seraseq synthetic fusions as true positives, therefore our precision statistics consider all other fusions to be false positives. While additional fusions may exist within the GM24385 cell line, these events are filtered out in the optimized EnFusion approach due either to high frequency across our cohort or supporting read evidence below our minimum threshold, suggesting that these fusions are likely artifactual or commonly occurring, and thus not clinically relevant.

To further benchmark our approach, we utilized an external cohort of adult acute myeloid leukemia (AML) samples ($n = 11$) containing known, clinically relevant fusions as true positives. Nine of the AML samples each contained a previously identified and clinically relevant fusion, identified by karyotyping and expert review by a clinical cytogeneticist. Utilizing the conditions optimized by the control reference standards, EnFusion successfully identified each of the nine clinically relevant fusions, resulting in 100% sensitivity. Two of the AML samples in this cohort did not have an identified fusion prior to EnFusion analysis, which identified a clinically relevant fusion. Upon reanalysis of the karyotyping results, a clinical cytogeneticist confirmed both EnFusion findings. (Additional File 1: Table S6). Additionally, we tested the specificity of our approach by utilizing negative control data generated by Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEERS) [35]. We generated two control datasets and utilized a previously published BEERS negative control dataset [28, 33]. EnFusion achieved high levels of specificity, identifying only one false positive across the three datasets, whereas the individual callers identified between 3 and 301 false negatives (Additional File 1: Table S7).

Implementation of the ensemble approach on an in-house pediatric cancer and hematologic disease cohort

Having demonstrated the efficacy of the optimized EnFusion pipeline using synthetic fusion samples, we further evaluated the utility of the pipeline on RNA-Seq data obtained from 229 patient samples, obtained from three prospective pediatric cancer and hematologic disease studies at Nationwide Children's Hospital (NCH) (Additional File 1: Fig. S3). Our approach identified significantly fewer total predicted fusions post-filtering, compared to all other single callers (Fig. 2A; Additional File 1: Table S8). Applying the known fusion list rescue altered the average number of fusions identified overall, as an average of 4.00 fusions per case were identified by 3 or more callers, while an average of 4.05 fusions were identified by 3 or more callers after applying the known fusion list; a total of 10 fusions were rescued by this approach, of which 1 (*KIAA1549-BRAF*; Additional File 3: Table S9) was clinically relevant. The retained *KIAA1549-BRAF* fusion was identified by three callers but was initially filtered out due to too few reads of evidence, possibly due to either low expression, low tumor cellularity or clonality. While the inclusion of the known fusion list increases false discovery, the benefit of increased sensitivity, through the rescue of filtered out clinically relevant fusions, greatly outweighs this slight decrease in specificity. In total, 67 clinically relevant fusions, identified in 67 different cases, (33 CNS, 7 heme, and 27 solid tumors; Additional File 1: Fig. S4A) were discovered using the optimized EnFusion pipeline with automated filtering, including the known fusion list feature, and a consensus of three callers (29.3% of tumors contained a clinically relevant fusion). Regardless of source material, there was roughly a 30% yield; with clinically relevant fusion identification in 44 of 148 frozen samples (30% yield), 19 of 68 FFPE samples (28% yield), and 4 of 13 other samples (31% yield), which included blood, cerebral spinal fluid, or bone marrow (Additional File 1: Fig. S4B). No single fusion detection algorithm was able to identify all 67 fusions. While JAFFA was the most sensitive algorithm, identifying the most clinically relevant fusions (65 out of 67), it also had one of the highest average numbers of fusions identified per sample, 1134 fusions, indicating a large number of likely false positives (Fig. 2B; Additional File 1: Table S8). Identified fusions were broken down into 4 types: Interchromosomal Chimeric ($n = 30$), Intrachromosomal Chimeric ($n = 29$), Loss of Function ($n = 3$), and Promoter Swapping ($n = 5$) (Fig. 2C). Of the 67 clinically relevant fusions, seven were considered novel events, defined as a gene fusion involving two partners not previously described in the literature at the time of identification (Fig. 2D). Of the 67 fusions detected, 43 (64.2%) were identified by all seven callers, 56 (83.6%) were identified by ≥ 6 callers, 60 (90%) were identified by ≥ 5 callers, 64 (96%) were identified by ≥ 4 callers, and 67 (100%) were identified by ≥ 3 callers. (Fig. 2E). One sample experienced an unresolvable failure of FusionMap, likely due to high sequencing read



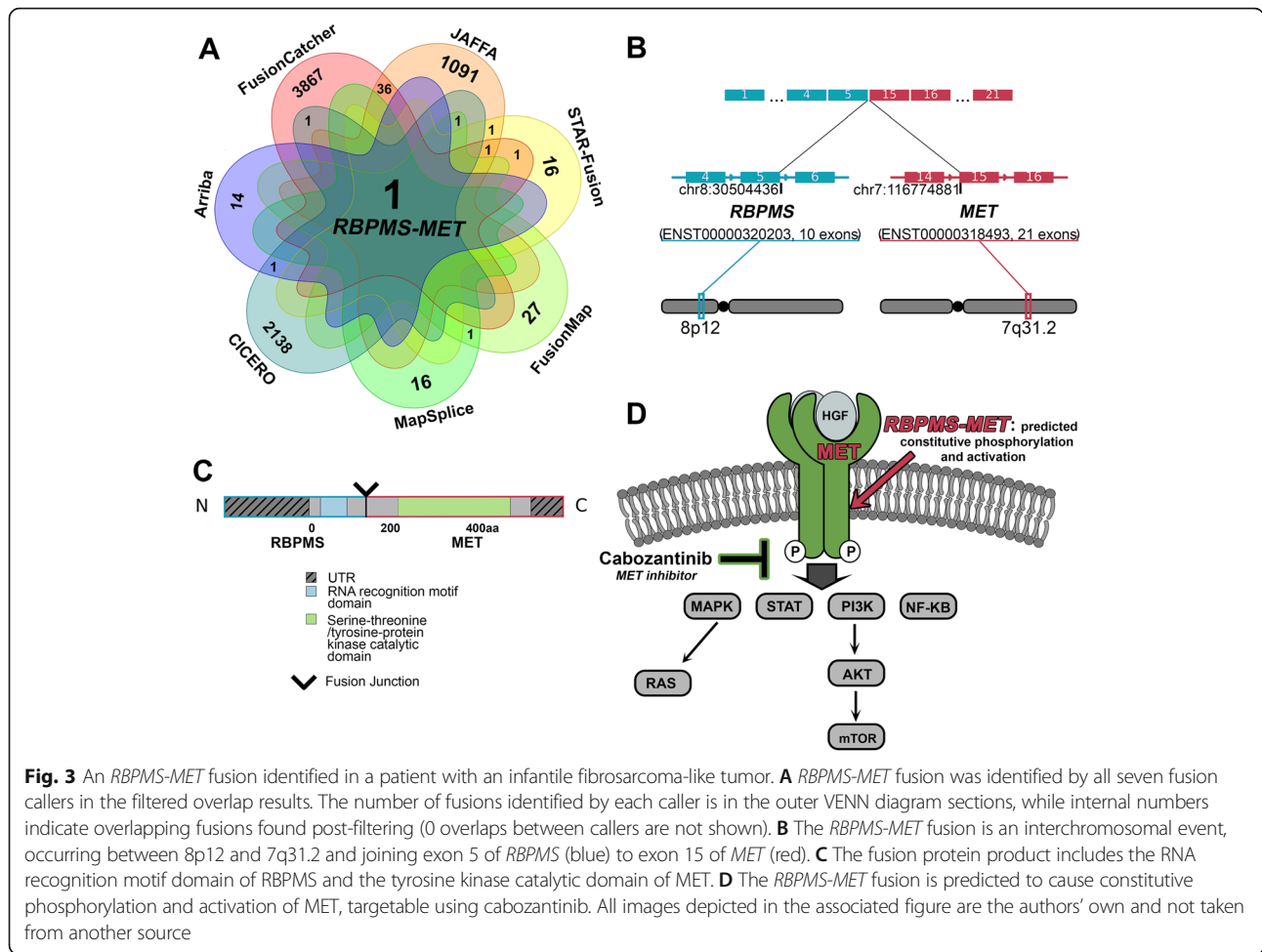
number. Results from the remaining callers, which successfully completed for this sample, were still included in our analysis. These results highlight the ability of the optimized EnFusion approach to identify gene fusions with a high level of confidence and a reduced number of false positive predictions, while preventing over-filtering by comparison to a list of known pathogenic fusions.

Clinical impact of fusion prediction

An *RBPMS-MET* fusion in an infantile fibrosarcoma-like tumor

A female infant presented with a congenital tumor of the right face. Histologically, the tumor consisted of variably cellular fascicles of spindle cells with a nonspecific immunohistochemical staining profile, suspicious for infantile fibrosarcoma. However, the tumor was negative

for an *ETV6-NTRK3* fusion, one of the defining features of infantile fibrosarcoma [36]. RNA-Seq of the primary tumor and EnFusion analysis revealed an *RBPMS-MET* fusion as the only consensus call. By contrast, the individual callers identified numerous fusions as follows: Arriba: 16, CICERO: 2142, FusionMap: 29, FusionCatcher: 3907, JAFFA: 1130, MapSplice: 18, and STAR-Fusion: 20 (Fig. 3A, Additional File 3: Table S9). *RBPMS*, an RNA-binding protein, and *MET*, a proto-oncogene receptor tyrosine kinase, have been identified as fusion partners in a variety of cancers with other genes and as gene fusion partners in a patient with cholangiocarcinoma [37]. Although *MET* fusions are uncommon drivers of sarcoma [38], a *TFG-MET* fusion has been reported in a patient with an infantile spindle cell sarcoma with neural features [37, 39, 40]. The interchromosomal



in-frame fusion of *RBPMS* (ENST00000320203, exon 5) to *MET* (ENST00000318493, exon 15) juxtaposes the RNA recognition motif of *RBPMS* to the *MET* tyrosine kinase catalytic domain (Fig. 3B, C). Given the therapeutic implications of this driver fusion, the fusion was confirmed and reported in the patient's medical record. The identification of this fusion provided the molecular driver for this tumor, which enabled definitive classification as an infantile fibrosarcoma-like tumor with a *MET* fusion. The patient was initially treated with VAC (vincristine, actinomycin D, and cyclophosphamide) chemotherapy which reduced tumor burden. Surgical resection of the mass was performed with positive margins. Given the presence of a targetable gene fusion, the presence of residual tumor, and the morbidity associated with additional surgery or radiation, the patient was subsequently treated with the *MET* inhibitor cabozantinib and demonstrated a complete pathological response (Fig. 3D).

An *NTRK1* fusion in an infiltrating glioma/astrocytoma

A 6-month-old female was diagnosed with an infiltrating glioma/astrocytoma, with a mitotic index of 7 per single

high-power field (HPF) and a Ki-67 labeling index averaging nearly 20%, indicative of aggressive disease. RNA-Seq of the primary tumor revealed a *BCAN-NTRK1* fusion, identified by five callers as the only consensus fusion output from EnFusion (Fig. 4A). This fusion was clinically confirmed by RT-PCR as an in-frame event, resulting from an intrachromosomal deletion of 225 kb at 1q23.1, which juxtaposes *BCAN* (ENST00000329117, exon 6) to *NTRK1* (ENST00000368196, exon 8) (Fig. 4B, C). This fusion results in the loss of the ligand binding domain of *NTRK1*, while retaining the tyrosine kinase catalytic domain, leading to a predicted activation of downstream targets in a ligand-independent manner [41]. Comparison of the normalized read counts from RNA-Seq data revealed elevated *NTRK1* expression, over 7 standard deviations from the mean, relative to *NTRK1* expression for CNS tumors within the NCH cohort ($N = 138$) (Fig. 4D). This result indicates the use of first generation TRK inhibitor therapies, with recent regulatory approvals, that have exemplary response rates (75%) and are generally well tolerated by patients [41]. Although the patient has no evidence of disease following gross total resection and

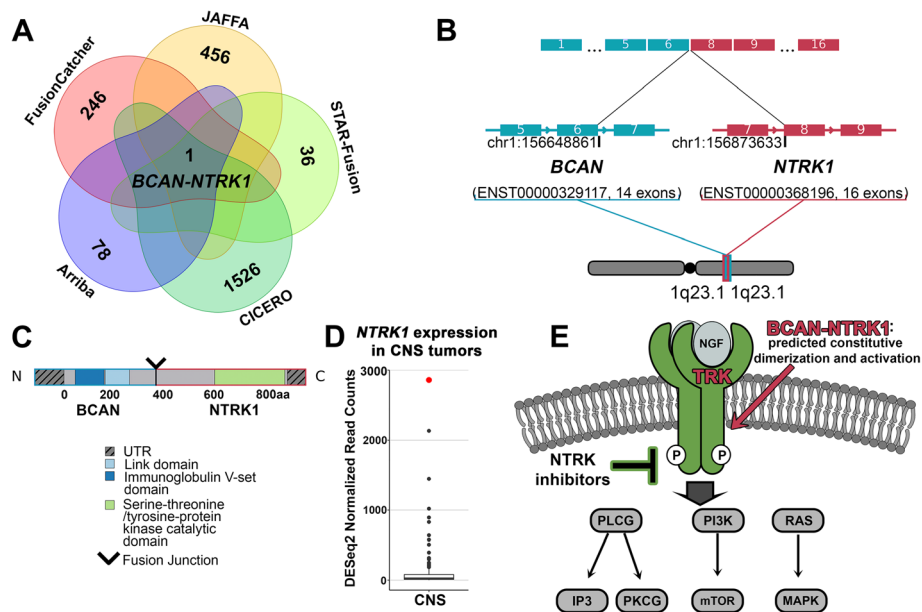


Fig. 4 Targetable *NTRK1* fusion identified in an infiltrating glioma. **A** The *BCAN-NTRK1* fusion was identified by 5 of 7 fusion callers, and was the only fusion returned by the filtered overlap results. Total fusions identified by each caller are shown, FusionMap and MapSplice identified no overlapping fusions that passed filtering (0 overlaps between callers are not shown). **B** The *BCAN-NTRK1* fusion is an intrachromosomal event occurring on 1q23.1, joining exon 6 of *BCAN* (blue) and exon 8 of *NTRK1* (red). **C** This fusion results in the juxtaposition of the tyrosine kinase catalytic domain of the *NTRK1* gene to the 5' end of the *BCAN* gene. **D** *NTRK1* is highly expressed in this patient (red) compared to CNS tumors (black) in the NCH cohort (CNS tumors: $n = 138$), with a normalized read count that is 7.70 standard deviations above the mean (131.2). **E** The *BCAN-NTRK1* fusion is predicted to increase expression and activation of the tyrosine kinase *NTRK1*, which may be inhibited by TRK inhibitor therapy (green). All images depicted in the associated figure are the authors' own and not taken from another source

treatment with conventional chemotherapy, TRK inhibitors may be clinically indicated in the setting of progressive disease given these findings (Fig. 4E).

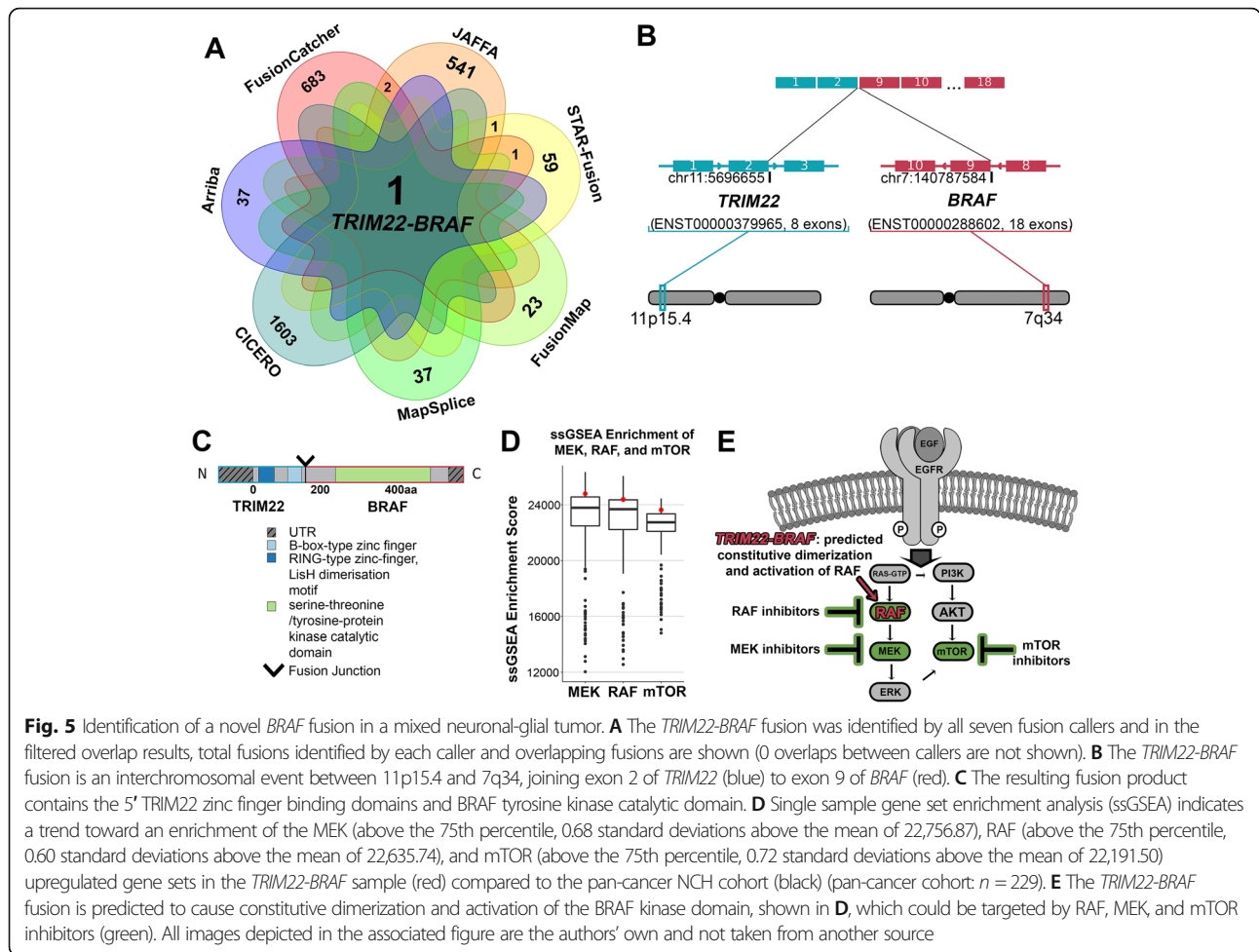
Novel *BRAF* fusion in a mixed neuronal-glioma tumor

A 14-year-old male with a lower brainstem tumor was diagnosed with a low-grade mixed neuronal-glioma tumor of unusual morphologic appearance. Tumor histology had features of both ganglioglioma and pilocytic astrocytoma. This tumor was negative for the somatic variant *BRAF* p.V600E, one of the most common somatic alterations associated with gangliogliomas and pilocytic astrocytomas [42]. Both the ganglioglioma and pilocytic astrocytoma-like portions of the primary tumor were studied separately by RNA-Seq. A novel *TRIM22-BRAF* fusion was identified in both histologies of the tumor, with EnFusion results from the ganglioglioma portion represented in Fig. 5A. *TRIM22-BRAF* was the only consensus fusion output by EnFusion and was clinically confirmed by RT-PCR. *TRIM22* and *BRAF* are novel fusion partners; however, *TRIM22* has been reported with other fusion partners in head/neck squamous cell carcinoma [43]. *BRAF* is a known oncogene that activates the RAS-MAPK signaling pathways, and has been described with numerous fusion partners, including the common *KIAA1549-BRAF* fusion in pediatric low-grade gliomas

[42]. This fusion is an interchromosomal translocation occurring between *TRIM22* (ENST00000379965, exon 2) at 11p15.4 and *BRAF* (ENST00000288602, exon 9) at 7q34. The resulting protein includes the *TRIM22* zinc finger domains and the *BRAF* tyrosine kinase domain (Fig. 5B, C). The *TRIM22-BRAF* fusion may lead to constitutive dimerization and activation of *BRAF* kinase domain, which is indicated by single sample Gene Set Enrichment Analysis (ssGSEA) and is theoretically targetable through RAF, MEK, or mTOR inhibitors (Fig. 5D, E).

Discussion

Fusions play a significant role as common oncogenic drivers of pediatric cancers, and their identification may refine diagnosis, inform prognosis, or indicate potential response to molecularly targeted therapies. We have developed an optimized pipeline for fusion detection that harmonizes results from several fusion calling algorithms, filters the output to remove known false positive results, and evaluates the detected fusions compared to a list of known pathogenic fusions. Testing this pipeline on a reference standard indicated that it outperforms single fusion detection algorithms by reducing the number of false positive calls, producing a smaller number of fusions prioritized by the strength of supporting



evidence, and suitable for manual inspection. As such, our pipeline greatly simplifies the interpretation process, enabling our multidisciplinary oncology teams to focus on medically relevant findings.

We tested the optimized EnFusion pipeline in a prospective study of 229 pediatric cancer and hematologic disease cases and identified 67 fusions. Of these, the fusions from 50 patients were selected for clinical confirmation by an orthogonal method, in our CAP-accredited, CLIA-validated clinical laboratory. All 50 (100% true positive rate) were confirmed to be true fusion events, and were determined to be of clinical relevance by our multidisciplinary care team, providing a diagnostic yield of over 29% across the cohort. (Additional File 3: Table S9). Given the high number of putative fusions observed with any single caller, it can be difficult to manually identify a pathogenic fusion amongst a list of tens, if not hundreds, of output fusions. By taking into consideration the frequency in which each fusion occurs in an internal database, as well as the level of evidence based on the number of callers and number of supporting reads by each caller, one can more confidently remove false

positives and identify relevant fusions. While our approach does not remove the necessity of manual curation, which is required to determine true clinical relevance of a fusion, it is able to drastically reduce the number of fusions that must be manually assessed, down to ~4 fusions per case, and provides annotations, including a pathogenicity gene partner score, to ease manual interpretation efforts. Our fully automated pipeline aids in prioritization, filtering, and subsequent knowledge-based analysis, providing a more streamlined and less labor-intensive approach to identify fusions, compared to current fusion identification methodologies, drastically reducing the manual workload required to sort through unfiltered or unprioritized results.

The most frequent fusion identified within our pediatric cancer cohort was *KIAA1549-BRAF* ($n = 12$, frequency = 5.2%; Fig. 2B) [17]. This fusion is characteristically found in pilocytic astrocytomas, which comprise 8.7% of our pediatric cancer cohort (20 out of 229 cases) [44]. We identified five different sets of *KIAA1549-BRAF* breakpoints within our cohort (Additional File 1: Fig. S5A). The most common fusion patterns represented in

the literature are *KIAA1549* exon 16-*BRAF* exon 9 (16–9) or *KIAA1549* exon 15-*BRAF* exon 9 (15–9), and these two breakpoints represent 9 of the 12 *KIAA1519-BRAF* fusions we identified (Additional File 1: Fig. S5B) [45, 46]. Three additional previously described sets of breakpoints were also identified, *KIAA1549* exon 16-*BRAF* exon 11 (16–11; $n = 1$), *KIAA1549* exon 15-*BRAF* exon 11 (15–11; $n = 1$), and *KIAA1549* exon 13-*BRAF* exon 9 (13–9; $n = 1$; Additional File 1: Fig. S5). While the 16–11 and 15–11 breakpoints occur less frequently than 16–9 or 15–9, they have been well described in the literature [45]; whereas only a single case with 13–9 breakpoints was reported as part of a pilocytic astrocytoma cohort study [47]. *KIAA1549-BRAF* fusions often have low levels of expression, a phenomenon that has been described in the literature and is associated with difficulties in its identification through RNA-Seq based methodologies, which lack fusion product amplification [30]. The ability of EnFusion to identify *KIAA1549-BRAF* fusions, and others that have very low levels of expression, highlights the sensitivity of our approach. Additionally, a supplementary “singleton” file for fusions that are identified by individual algorithms and on the known fusion list is also output by our approach, allowing users the opportunity to manually interpret singleton results. In one case, a *KIAA1549-BRAF* fusion was missed in the overlap output but rescued using EnFusion’s singleton output (it was identified by a single read of evidence from JAFFA; data not shown). Inclusion of the Singleton output allowed for the rescue of this finding and the subsequent clinical confirmation of this fusion by Sanger sequencing. This approach ensures that fusions on the known fusion list are retained, even with minimal evidence by a single caller, and prevents users from having to sift through individual caller’s outputs for these types of low evidence fusions.

Our approach has also identified other fusions commonly associated with pediatric cancer, including *EWSR1-FLI1* ($n = 9$), *FGFR1-TACC1* ($n = 3$), *PAX3-FOXO1* ($n = 3$), *ZFTA-RELA* ($n = 2$), *COL3A1-PLAG1* ($n = 2$), and *NPM1-ALK* ($n = 2$) (Fig. 2B). In addition to common fusions, EnFusion also identified seven novel fusions (Fig. 2B). Five of the seven novel fusions were confirmed by an orthogonal assay in our clinical lab (Additional File 3: Table S9). Chimeric fusions, which include both interchromosomal ($n = 30$) and intrachromosomal ($n = 29$) events, were the most common type of fusion identified within the cohort, however, 5 promoter swapping and 3 loss of function fusions were also identified, highlighting the range of fusions this approach can detect (Fig. 2C).

Running seven different fusion callers is computationally complex, as each has its own set of dependencies and environmental requirements. To overcome this, we

utilize modern cloud computing technologies. Most notable, our entire pipeline has been built in an AWS serverless environment, removing the requirement for high performance computing (HPC) clusters, while producing highly reproducible results and enabling pipeline sharing (Additional File 1: Fig. S6). The use of a serverless environment provides flexibility to deploy and scale applications regardless of the application’s size, without needed concern for the underlying infrastructure. We are also leveraging containers to process the data within the serverless environment, as they can be easily utilized by outside institutions with little to no adjustment to their own environments. Another benefit to the current structure of our approach is the ability to assess output from the individual algorithms in real time, as the ensemble pipeline is automatically run after each individual caller completes, allowing for interpretation of at least 3 of the 7 callers within ~3.5 hours (h), which can be beneficial in situations that necessitate fast turnaround times (Additional File 1: Fig. S7). Overall, our novel use of serverless technology provides a robust computational solution that is fully automatable and easy to distribute.

There are numerous benefits to the utilization of this optimized pipeline, in that detected fusion events are agnostic to gene partner, allowing identification of common, rare, and novel fusions. In addition, the RNA-Seq data set can be utilized for other types of downstream and correlative analyses, including evaluation of gene expression for loci disrupted by the fusion (Fig. 4D). Utilization of cohort data to assess outlier gene expression can provide valuable insights into pathway disruptions that may occur due to the gene fusion (Fig. 5D) and may provide information about disease subtyping.

The EnFusion pipeline is customizable, allowing users to select how many and which callers to deploy. This may impact potential cost savings, time-to-result, or permit customization that eliminates specific callers that require excessive compute requirements or run times, as suitable in a clinical diagnostic or research setting. Users can also determine the number of consensus calls required to support fusion prediction, which can reduce the number of fusions to assess manually. Callers with a higher percentage of false positives, FusionCatcher and JAFFA, often overlap in their predictions, leading to an increased average number of fusions output by the ensemble pipeline with a consensus of only two callers; a problem diminished by requiring predictions from at least three callers to overlap. In our study, precision was found to be highest in the three-caller consensus version of the ensemble pipeline (Table 2; Additional File 2: Table S4). Another benefit to utilizing different algorithms is the ability to assess supplementary output data, in addition to traditional fusion calling. We have made use of this through the inclusion of the internal tandem

duplication (ITD) detection which is performed by CICERO. CICERO has identified 7 clinically relevant ITDs within our cohort, 4 of which we have confirmed using orthogonal assays (Additional File 1: Table S10).

Future developments to the pipeline could include a weighting system for each caller, based on the precision and sensitivity of the algorithm and on which callers have overlapping predictions, leading to a more sophisticated prioritization strategy. Additional fusion calling algorithms may also be considered and provided as options for users. As each algorithm has its own strengths and weaknesses, future iterations of this pipeline could include the ability to select algorithms based on the specific needs of the user and dataset at hand. For example, while we identified a handful of loss of function fusions in our study, only two of the included algorithms can identify intergenic breaks (Arriba and CICERO). As such, our approach may be under-representing loss of function rearrangements due to having only two callers that report this type of alteration. To ameliorate this, one might be able to selectively utilize additional fusion detection algorithms that specifically identify this type of break, such as InFusion, deFuse, or Mintie [48–50]. The known fusion list can also be modified and tailored to include specific gene pairs, or even single genes of interest, providing another layer of customization. Importantly, through the utilization of a proper internal database for frequency filtering purposes, considering age and/or cancer diagnosis, and with the deployment of the appropriate known fusion list, the ensemble approach could be readily implemented in adult cancer fusion detection. On a similar note, while we have employed the EnFusion approach to analyze data from hematologic malignancies, our cohort size was limited, and unable to identify complex enhancer hijacking events, known to occur in this patient population. This disparity is likely due to the specific pediatric cohort not containing these events, as none were identified by any of the single callers we used. Lastly, not all predictors performed equally, and there was a single unresolvable failure of FusionMap to complete. This failure was likely due to the sequencing depth of the sample, however further analysis is required to determine whether parameter modification would permit completion of FusionMap in this case (Additional File 3: Table S9). Importantly, our approach was able to circumvent this failure due to the multi-caller nature of the pipeline. Lastly, there are many modalities of RNA-seq analysis that may be harnessed in future developments of the ensemble fusion detection pipeline, which may include an integrative approach exploiting expression-based analysis and ranking. In summary, the ensemble pipeline provides

a highly customizable approach to fusion detection that can be applied to numerous settings, with opportunities for future improvements based on additional features and applications.

Conclusions

The EnFusion pipeline provides a highly automated and accurate approach to fusion detection, developed to identify high confidence, clinically relevant gene fusions from RNA-Seq data produced from pediatric cancer and hematologic disease samples, that could be readily implemented in adult cancer RNA-Seq data analysis. The clinical impact of accurately identifying gene fusions in a given patient's tumor sample is critical, not only in refining diagnosis and providing prognostic information, but also indicating potential therapeutic vulnerabilities that may shape treatment decisions. These important advantages have led us to incorporate EnFusion as a necessary component of our translational pediatric cancer analysis pipeline.

Methods

Description of an internal patient cohort

In total, 229 patients were consented and enrolled onto one of three Institutional Review Board (IRB) approved protocols (IRB17–00206, IRB16–00777, IRB18–00786) and studied at the Institute for Genomic Medicine (IGM) at Nationwide Children's Hospital (NCH) in Columbus, Ohio. Through the utilization of genomic and transcriptomic profiling, these protocols aim to refine diagnosis and prognosis, detect germline cancer predisposition, identify targeted therapeutic options, and/or to determine eligibility for clinical trials in patients with rare, treatment-refractory, relapsed, pediatric cancers or hematologic diseases, or with epilepsy arising in the setting of a low grade central nervous system (CNS) cancer. Our in-house NCH cohort as studied here, consisted of samples from CNS tumors ($n = 138$), hematologic diseases ($n = 18$), and non-CNS solid tumors ($n = 73$), as represented in Additional File 1: Fig. S3.

RNA-Seq of patient tissues and positive controls used for benchmarking

RNA was extracted from snap frozen tissue, formalin-fixed paraffin-embedded (FFPE) tissue, peripheral blood, bone marrow, and cerebral spinal fluid utilizing dual RNA and DNA co-extraction methods originally developed by our group for The Cancer Genome Atlas project [51]. White blood cells were isolated from peripheral blood or bone marrow using the lymphocyte separation medium Ficoll-histopaque. Frozen tissue, white blood cells, or pelleted cells from cerebrospinal fluid were homogenized in Buffer RLT, with beta-Mercaptoethanol to denature RNases, plus Reagent DX and separated on an AllPrep (Qiagen) DNA column to

remove DNA for subsequent RNA steps. The eluate was processed for RNA extraction using acid-phenol:chloroform (Sigma) and added to the mirVana miRNA (Applied Biosystems) column, washed, and RNA was eluted using DEPC-treated water (Ambion). DNase treatment (Zymo) was performed post RNA purification. FFPE tissues were deparaffinized using heptane/methanol (VWR) and lysed with Paraffin Tissue Lysis Buffer and Proteinase K from the HighPure miRNA kit (Roche). The sample was pelleted to remove the DNA, and the supernatant was processed for RNA extraction with the HighPure miRNA column, followed by DNase treatment (Qiagen). RNA quantification was performed with Qubit (Life Sciences).

RNA-Seq libraries were generated using 100 ng to 1 µg of DNase-treated RNA input, either by ribodepletion using the Ribo-Zero Globin kit (Illumina) followed by library construction using the TruSeq Stranded RNA-Seq protocol (Illumina), or by ribodepletion with NEBNext Human/Mouse/Rat rRNA Depletion kit followed by library construction using the NEBNext Ultra II Directional RNA-Seq protocol (New England BioLabs). Illumina 2 × 151 paired end reads were generated either on the HiSeq 4000 or NovaSeq 6000 sequencing platforms (Illumina). An average of 104 million read pairs were obtained per sample (range 37 M to 380 M read pairs).

Following data production and post-run processing, FASTQ files were aligned to the GRCh38 human reference (hg38) using STAR aligner (version 2.6.0c) [52]. Feature counts were calculated using HTSeq and normalized read counts were calculated for all samples using DESeq2 [53, 54]. Single sample Gene Set Enrichment Analysis (ssGSEA), v10.0.3, was performed on DESeq2 normalized read counts using Molecular Signatures Database (MSigDB) Oncogenic Signatures (c6.all.v7.2.symbols.gmt), which included MEK-upregulated genes (MEK_UP.V1_UP), RAF-upregulated genes (RAF_UP.V1_UP), and mTOR-upregulated genes (MTOR_UP.N4.V1_UP) [55].

RNA-Seq of SeraCare control reference standards

Seraseq Fusion RNA Mix (SeraCare Inc., Milford, MA) was utilized as a control reference standard reagent to test and optimize the ensemble fusion detection pipeline. This product contains 14 synthetic gene fusions in vitro transcribed, utilizing the GM24385 cell line RNA as a background. RNA-Seq libraries were prepared utilizing 500 ng input of neat (undiluted) Seraseq Fusion RNA v2, a non-commercially available concentrated product, as input (SeraCare). RNA-Seq libraries were also prepared using 500 ng input of diluted control reference standard (Seraseq Fusion RNA v3 (SeraCare)), which, as a neat reagent is roughly equivalent to a 1:25 dilution of the v2 product, and of total human RNA (GM24385, Coriell)

for the fusion-negative controls. Concentrations of individual fusions in the control reference standard were determined by the manufacturer using a custom fluorescent probe set (based on TaqMan probe design) for each fusion and evaluation by droplet digital PCR. Digital PCR-based concentration data (copies/ul) are available in Additional File 1: Table S2 for the undiluted sample and Additional File 1: Table S3 for the diluted sample [56].

Dilutions of the Seraseq Fusion RNA v3 reference standard were performed by mixing with control total human RNA (GM24385, Coriell) for final dilutions of 1: 25, 1:50, 1:250, 1:500, 1:2500. We also evaluated undiluted Seraseq Fusion RNA v2. For neat and diluted samples, 500 ng input RNA was treated using the NEBNext Human/Mouse/Rat rRNA Depletion kit and libraries were prepared following the NEBNext Ultra II Directional RNA-Seq protocol (New England BioLabs). Paired end 2 × 151 bp reads were produced using the HiSeq 4000 platform (Illumina). An average of 149 million read pairs were obtained per Seraseq sample (range of 86 M to 227 M read pairs).

Generation of negative control dataset

To assess the false discovery rate of the EnFusion pipeline, we utilized three synthetic control datasets generated by Benchmark for Evaluating the Effectiveness of RNA-Seq Software (BEERS) [35]. One dataset (BEERS1) was previously described in the literature [28, 33], and we generated two new datasets (BEERS2 and BEERS3) utilizing the BEERS default parameters for 151 paired end data with 50 million reads.

Optimized fusion detection pipeline

Fusions were detected from paired end RNA-Seq FASTQ files utilizing an automated ensemble fusion detection pipeline that employs seven fusion-calling algorithms described in Table 1: Arriba (v1.2.0), CICERO (v0.3.0), FusionMap (v mono-2.10.9), FusionCatcher (v0.99.7c), JAFFA (direct v1.09), MapSplice (v2.2.1), and STAR-Fusion (v1.6.0) [25, 29–34]. STAR-Fusion parameters were altered to reduce the stringency setting for the fusion fragments per million reads (FFPM) to 0.02 (`-min_FFPM 0.02`), while default parameters were retained for all other callers. After fusion calling with each independent algorithm, a custom algorithm written in the R programming language, was used to “overlap,” or align and compare, the unordered gene partners identified by individual fusion callers. The utilization of unordered gene partners allows for fusions to be compared, even if different breakpoints were identified by individual algorithms, and to include reciprocal fusions. To ameliorate naming issues that may be encountered due to different references utilized across

individual callers, we used an “alias matching algorithm” to harmonize gene names by conversion to HUGO Gene Nomenclature Committee (HGNC) gene symbols. To mitigate ambiguously matching aliases ($n = 1392$), gene symbol cytogenetic band locations are utilized in accordance with breakpoints, to identify the correct gene symbol. Fusion partners identified by at least three of the seven callers are retained and prioritized based on the number of contributing algorithms first and then by the number of sequence reads providing evidence for each fusion. The overlap output retains annotations from the individual callers, including breakpoints, distance between breakpoints, donor and acceptor genes, reads of evidence, nucleotide sequence at breakpoint (if available), frequency information from the database, and whether the identified fusion contains “known pathogenic fusion partners”. If discordant breakpoints are identified across callers for a set of fusion partners, the breakpoints with the most evidence, determined by number of supporting reads, are prioritized in the output.

The fusions are filtered by the following steps (Fig. 1A). Read-through events, which occur between neighboring genes and are typically identified in both healthy and disease states, are not expected to impact cellular functions [12, 24]. This type of fusion prediction is a source of false positive results, so we have implemented a filter that removes fusions detected between genes fewer than 200,000 bases apart, that occur on the same strand and chromosome. Recurrent fusions with uncertain biological significance have also been identified in normal tissues. To prevent the inclusion of commonly occurring, benign fusions in our output, a PostgreSQL database was used to filter commonly occurring artifactual fusions. This filter removes any expected fusion artifact with greater than a 10% frequency of detection based on our internal cohort. Lastly, to ensure a high level of confidence in the identified fusions, we utilize a minimum threshold for level of evidence, removing fusions that contain fewer than four reads of support from at least one contributing algorithm.

While filtering can remove false positive results and reduces the time needed to review predicted fusions, it also can remove true positive fusions in certain circumstances. To prevent the inadvertent filtering of known fusions, a known fusion list was developed containing 325 pairs of common fusion partners associated with cancer, as identified in COSMIC and TCGA (Additional File 1: Table S1) [27, 57]. To increase sensitivity in the identification of known pathogenic fusions, fusion partners that are on the known fusion list are retained if at least two callers have identified the fusion. The EnFusion pipeline also outputs a supplementary singleton fusion file, containing fusions identified by a single caller

that are on the known fusion list, allowing users to examine low evidence fusions that may be of interest.

To prioritize fusions that contain gene partners commonly found in the known fusion list, we developed the “Gene Partner Predicted Pathogenicity Score” based on the frequency of the individual partners in the known fusion list. Of the 325 fusions on the known fusion list, 38 genes are present as a fusion partner ≥ 3 times (Additional File 1: Table S11, Fig. S8). The most common partners are *BRAF* and *KMT2A*, which are present as fusion partners 28 times each. To aid prediction of novel, or not well described, pathogenic fusions, we developed a score based on known pathogenic gene partners. This score utilizes the frequency of partners present on the known fusion list. The pathogenic frequency score ranges from 10 (most frequent) to 1 (least frequent, but present at least 3 times):

$$\text{Pathogenic Frequency Score} = 10 / (f_{max} - f)$$

Where f is the gene frequency and f_{max} is the maximum observed frequency. The following annotations are included in the EnFusion results if an identified fusion contains one of the 38 common pathogenic gene partners: designation as a known pathogenic gene partner, inclusion of the frequency score (1–10), and gene type based on UniProt description [58].

A knowledge-based interpretation strategy was applied to the filtered list of fusion partners output by the pipeline, including the use of FusionHub [59], to inform clinical relevance, such as diagnostic and/or prognostic information or a potential therapeutic target. Visual assessment of the fusion events was performed by examining RNA-Seq BAM files with Integrated Genome Viewer (IGV). Fusions were also assessed at the DNA level by IGV-based evaluation of gene-specific paired end read alignments from ES or WGS BAM files, for potential evidence of mapping discordance. Clinically relevant fusions were then assayed in a College of American Pathologists (CAP)-accredited clinical laboratory using one or more of the following methods: RT-PCR followed by Sanger sequencing of the resulting products, fluorescence in situ hybridization (FISH), chromosomal analysis, and/or by Archer FusionPlex Solid Tumor panel (ArcherDx) for clinical confirmation.

AWS implementation of the ensemble approach

The EnFusion pipeline is implemented utilizing an Amazon Web Services (AWS) serverless environment (Additional File 1: Fig. S6). The workflow is initiated via a call to Amazon API Gateway, which passes parameters, including the location of the input FASTQ files, to an AWS Lambda function. The Lambda

function initiates the AWS Batch job to load and executes a custom fusion detection Docker image, which launches Arriba, CICERO, FusionMap, FusionCatcher, JAFFA, MapSplice, and STAR-Fusion. We utilize the R5 family of instances for the fusion detection algorithms. Due to the efficiency by which different algorithms can multi-thread, each fusion detection tool is allocated 32 virtual CPUs (vCPUs), except for CICERO which is allocated 16 vCPUs and JAFFA which is allocated 8 vCPUs. Using the described allocations, Arriba completes the fastest (~ 37 min; minutes) for the runs completed year to date in 2020, followed by FusionMap (~ 1 h 12 min), STAR-fusion (~ 3 h 25 min), FusionCatcher (~ 10 h 35 min), CICERO (~ 11 h 49 min), MapSplice (~ 15 h 2 min), and JAFFA (~ 27 h 16 min), data is summarized in Additional File 1: Fig. S7. The results from the fusion callers are sent to an AWS S3 output bucket, which invokes AWS Batch to load and execute a Docker image with our overlap script upon completion. This allows for real-time examination of results as each caller finishes, as the overlapping output is updated upon completion of each individual caller, which is particularly advantageous given the long execution times for some of the fusion callers. It is possible to examine results upon completion of the three fastest algorithms within ~ 3.5 h, which is of great benefit for cases necessitating fast turnaround times, and complete results are made available by the next day. The overlap Docker image queries and writes to an Aurora PostgreSQL database and performs all necessary filtering. The final results, including annotated filtered and unfiltered fusion lists, are stored in an AWS S3 output bucket for subsequent interpretation. Dockerfile to build the Docker image used to run the overlap algorithm is available at our GitHub repository (<https://github.com/nch-igm/EnFusion>), DOI: <https://doi.org/10.5281/zenodo.5172341>.

Data analysis and statistics

Figures were plotted using R version 4.0.2. Statistical analysis was performed by GraphPad Prism 9 software. Graphical representation of fusion breakpoints and products were generated using a modified version of INTEGRATE-Vis [60].

Abbreviations

AWS: Amazon Web Services; CNS: Central nervous system; ES: Exome sequencing; FDR: False discovery rate; FFPE: Formalin fixed, paraffin embedded; FFP: Fusion fragments per million; GSNAP: Genomic Short-read Nucleotide Alignment Program; Heme: Hematologic diseases; HPF: High power field; HPC: High performance computing; IGM: Institute for Genomic Medicine; IGV: Integrated Genome Viewer; ITD: Internal tandem duplication; NCH: Nationwide Children's Hospital; QC: Quality control; RNA-Seq: RNA-sequencing; ssGSEA: Single Sample Gene Set Enrichment Analysis; vCPU: Virtual central processing unit; WGS: Whole genome sequencing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-08094-z>.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements

We thank the patients and their families for participating in our translational research protocol.

Authors' contributions

SL analyzed and interpreted fusion data, contributed to development of overlap algorithm and Docker images, contributed to AWS serverless workflow, and wrote the manuscript. JF and KV contributed development of overlap algorithm and Docker images, contributed to AWS serverless workflow, and contributed to manuscript writing. AHW and JA conceived and developed the alias functionality. AH, BJK, GL, and SW provided data analysis support, designed AWS serverless workflow, and contributed to manuscript writing. SF contributed to manuscript revisions and oversaw, organized, and performed data upload to SRA. KMS contributed to analysis and interpretation of RNA-Seq results and performed clinical data acquisition. KM, TAB, KL and DK provided NGS analysis and interpretation for cancer cohort data. NB, SDM, and ARM perform library preparations and developed laboratory procedures for RNA-Seq processing/QC. AW managed RNA-Seq processing and analysis. KL managed and coordinated all clinical samples. DRB, JRL, JLF, MA, DSO, AG, BS, EAV, and SCK contributed to the enrollment of patients onto the NCH cancer protocols and provided clinical expertise, DRB and SCK also contributed pathology materials (fixed or frozen tissues etc.) following QA and/or QC reviews of enrollee pathology. AKE and KM provided AML control samples and clinical confirmation information. VM oversaw technology development and contributed to the conceptual design of project. CEC, ERM, and RKW developed, led, and supervised work performed on cancer protocol, contributed to conceptual design of project, contributed to analysis and interpretation of RNA-Seq results, and contributed to manuscript writing and revision. PW conceived, designed, and supervised the project, oversaw and contributed to algorithm development, provided support for utilization of AWS and computational resources, and contributed significantly to manuscript writing and revision. All authors read and approved the final manuscript.

Funding

We thank the Nationwide Children's Foundation and The Abigail Wexner Research Institute at Nationwide Children's Hospital for generously supporting this body of work. These funding bodies had no role in the design of the study, no role in the collection, analysis, and interpretation of data and no role in writing the manuscript.

Availability of data and materials

DNA and RNA sequencing data for this study have been deposited to dbGAP, accession number phs001820.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001820.v1.p1). Seraseq fastq files, from the benchmarking studies, have been deposited to the NIH Sequence Read Archive (SRA), accession number PRJNA679580. Code for the overlap algorithm is available at our GitHub repository (<https://github.com/nch-igm/EnFusion>), DOI: <https://doi.org/10.5281/zenodo.5172341> (<https://doi.org/10.5281/zenodo.5172341>).

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Institutional Review Board (IRB) of The Research Institute at Nationwide Children's Hospital. Written informed consent was obtained from the patients and/or parents for molecular genetic analysis, which included RNA-sequencing. These protocols allowed for return of results from research sequencing studies after confirmation in a CLIA-certified laboratory.

Consent for publication

Not applicable.

Competing interests

No Competing interests: Stephanie LaHaye, James Fitch, Kyle Voytovich, Adam Herman, Benjamin Kelly, Grant Lammi, Jeremy Arbesfeld, Saranga Wijeratne, Kathleen Schieffer, Natalie Bir, Sean McGrath, Anthony Miller, Amy Wetzell, Katherine Miller, Tracy Bedrosian, Kristen Leraas, Elizabeth Varga, Ajay Gupta, Bhuvana Setty, Jeffrey Leonard, Jonathan Finlay, Mohamed Abdelbaki, Diana Osorio, Selene Koo, Daniel Koboldt, Alex Wagner, Ann-Kathrin Eisfeld, Krzysztof Mrozek, Vincent Magrini, Catherine Cottrell, Richard Wilson and Peter White.

Elaine Mardis: Qiagen N.V., supervisory board member, honorarium and stock-based compensation.

Daniel Boué: Illumina (ILMN) share holder.

Author details

¹The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. ²Division of Hematology, Oncology, Blood and Marrow Transplant, Nationwide Children's Hospital, Columbus, OH, USA. ³Department of Pediatrics, The Ohio State University, Columbus, OH, USA. ⁴Department of Pathology, The Ohio State University, Columbus, OH, USA. ⁵Department of Pathology, Nationwide Children's Hospital, Columbus, OH, USA. ⁶Section of Neurosurgery, Nationwide Children's Hospital, Columbus, OH, USA. ⁷Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA. ⁸Division of Hematology, The Ohio State University, Columbus, OH, USA. ⁹Clara D. Bloomfield Center for Leukemia Outcomes Research, The Ohio State University, Columbus, OH, USA. ¹⁰The Ohio State Comprehensive Cancer Center, Columbus, OH, USA.

Received: 11 March 2021 Accepted: 15 October 2021

Published online: 04 December 2021

References

- Steliarova-Foucher E, Colombet M, Ries LAG, Moreno F, Dolya A, Bray F, et al. IICC-3 contributors: international incidence of childhood cancer, 2001-10: a population-based registry study. *Lancet Oncol*. 2017;18(6):719–31. [https://doi.org/10.1016/S1470-2045\(17\)30186-9](https://doi.org/10.1016/S1470-2045(17)30186-9).
- Amatu A, Sartore-Bianchi A, Siena S. NTRK gene fusions as novel targets of cancer therapy across multiple tumour types. *ESMO Open*. 2016;1(2):e000023. <https://doi.org/10.1136/esmoopen-2015-000023>.
- Pui CH, Gajjar AJ, Kane JR, Qaddoumi IA, Pappo AS. Challenging issues in pediatric oncology. *Nat Rev Clin Oncol*. 2011;8(9):540–9. <https://doi.org/10.1038/nrclinonc.2011.95>.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin*. 2016; 66(1):7–30. <https://doi.org/10.3322/caac.21332>.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546–58. <https://doi.org/10.1126/science.1235122>.
- Grobner SN, Worst BC, Weischenfeldt J, Buchhalter J, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. *Nature*. 2018;555(7696):321–7. <https://doi.org/10.1038/nature25480>.
- Marshall GM, Carter DR, Cheung BB, Liu T, Mateos MK, Meyerowitz JG, et al. The prenatal origins of cancer. *Nat Rev Cancer*. 2014;14(4):277–89. <https://doi.org/10.1038/nrc3679>.
- Rowley JD. Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*. 1973;243(5405):290–3. <https://doi.org/10.1038/243290a0>.
- Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448(7153):561–6. <https://doi.org/10.1038/nature05945>.
- Jia Y, Xie Z, Li H. Intergenicly spliced chimeric RNAs in cancer. *Trends Cancer*. 2016;2(9):475–84. <https://doi.org/10.1016/j.trecan.2016.07.006>.
- Li Y, Li Y, Yang T, Wei S, Wang J, Wang M, et al. Clinical significance of EML4-ALK fusion gene and association with EGFR and KRAS gene mutations in 208 Chinese patients with non-small cell lung cancer. *PLoS One*. 2013; 8(1):e52093. <https://doi.org/10.1371/journal.pone.0052093>.
- Dupain C, Harttrampf AC, Urbinati G, Georger B, Massaad-Massade L. Relevance of fusion genes in pediatric cancers: toward precision medicine. *Mol Ther Nucleic Acids*. 2017;6:315–26. <https://doi.org/10.1016/j.omtn.2017.01.005>.
- Bernt KM, Hunger SP. Current concepts in pediatric Philadelphia chromosome-positive acute lymphoblastic leukemia. *Front Oncol*. 2014;4:54. <https://doi.org/10.3389/fonc.2014.00054>.
- Hawkins C, Walker E, Mohamed N, Zhang C, Jacob K, Shirinian M, et al. BRAF-KIAA1549 fusion predicts better clinical outcome in pediatric low-grade astrocytoma. *Clin Cancer Res*. 2011;17(14):4790–8. <https://doi.org/10.1158/1078-0432.CCR-11-0034>.
- Park SH, Won J, Kim SI, Lee Y, Park CK, Kim SK, et al. Molecular testing of brain tumor. *J Pathol Transl Med*. 2017;51(3):205–23. <https://doi.org/10.4132/jptm.2017.03.08>.
- Yuan L, Liu ZH, Lin ZR, Xu LH, Zhong Q, Zeng MS. Recurrent FGFR3-TACC3 fusion gene in nasopharyngeal carcinoma. *Cancer Biol Ther*. 2014;15(12): 1613–21. <https://doi.org/10.4161/15384047.2014.961874>.
- Jones DT, Kocalkowski S, Liu L, Pearson DM, Backlund LM, Ichimura K, et al. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res*. 2008;68(21):8673–7. <https://doi.org/10.1158/0008-5472.CAN-08-2097>.
- Morris SW, Kirstein MN, Valentine MB, Dittmer K, Shapiro DN, Look AT, et al. Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. *Science*. 1995;267(5196):316–7. <https://doi.org/10.1126/science.267.5196.316-b>.
- Mosse YP, Lim MS, Voss SD, Wilner K, Ruffner K, Laliberte J, et al. Safety and activity of crizotinib for paediatric patients with refractory solid tumours or anaplastic large-cell lymphoma: a Children's Oncology Group phase 1 consortium study. *Lancet Oncol*. 2013;14(6):472–80. [https://doi.org/10.1016/S1470-2045\(13\)70095-0](https://doi.org/10.1016/S1470-2045(13)70095-0).
- Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*. 2017;27(5):677–85. <https://doi.org/10.1101/gr.214007.116>.
- Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res*. 2018;28(8):1126–35. <https://doi.org/10.1101/gr.231100.117>.
- Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009; 458(7234):97–101. <https://doi.org/10.1038/nature07638>.
- Wang Q, Xia J, Jia P, Pao W, Zhao Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform*. 2013;14(4):506–19. <https://doi.org/10.1093/bib/bbs044>.
- He Y, Yuan C, Chen L, Lei M, Zellmer L, Huang H, et al. Transcriptional-readthrough RNAs reflect the phenomenon of "a gene contains gene(s)" or "gene(s) within a gene" in the human genome, and thus are not chimeric RNAs. *Genes (Basel)*. 2018;9(1). <https://doi.org/10.3390/genes910040>.
- Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Reggev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):213. <https://doi.org/10.1186/s13059-019-1842-9>.
- Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res*. 2016;44(5):e47. <https://doi.org/10.1093/nar/gkv1234>.
- Gao Q, Liang WW, Foltz SM, Mutharasu G, Jayasinghe RG, Cao S, et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep*. 2018;23:227–238.e3.
- Apostolides M, Jiang Y, Husic M, Siddaway R, Hawkins C, Turinsky AL, et al. MetaFusion: a high-confidence metacaller for filtering and prioritizing RNA-seq gene fusion candidates. *Bioinformatics*. 2021;37(19):3144–51. <https://doi.org/10.1093/bioinformatics/btab249>.
- Uhrig S, Ellermann J, Walther T, Burkhardt P, Frohlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res*. 2021;31(3):448–60. <https://doi.org/10.1101/gr.257246.119>.
- Tian L, Li Y, Edmonson MN, Zhou X, Newman S, McLeod C, et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol*. 2020;21(1):126. <https://doi.org/10.1186/s13059-020-02043-x>.
- Ge H, Liu K, Juan T, Fang F, Newman M, Hoek W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution.

- Bioinformatics. 2011;27(14):1922–8. <https://doi.org/10.1093/bioinformatics/btr310>.
32. Nicorici D, Şatalan M, Edgren H, Kangaspeska S, Murumägi A, Kallioniemi O, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014. <https://doi.org/10.1101/011650>.
 33. Davidson NM, Majewski IJ, Oshlack A. JAFFA: high sensitivity transcriptome-focused fusion gene detection. *Genome Med*. 2015;7(1):43. <https://doi.org/10.1186/s13073-015-0167-x>.
 34. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178. <https://doi.org/10.1093/nar/gkq622>.
 35. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011;27(18):2518–28. <https://doi.org/10.1093/bioinformatics/btr427>.
 36. Church AJ, Calicchio ML, Nardi V, Skalova A, Pinto A, Dillon DA, et al. Recurrent EML4-NTRK3 fusions in infantile fibrosarcoma and congenital mesoblastic nephroma suggest a revised testing strategy. *Mod Pathol*. 2018; 31(3):463–73. <https://doi.org/10.1038/modpathol.2017.127>.
 37. Flucke U, van Noesel MM, Wijnen M, Zhang L, Chen CL, Sung YS, et al. TFG-MET fusion in an infantile spindle cell sarcoma with neural features. *Genes Chromosom Cancer*. 2017;56(9):663–7. <https://doi.org/10.1002/gcc.22470>.
 38. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. *Nat Commun*. 2014;5(1):4846. <https://doi.org/10.1038/ncomms5846>.
 39. International Cancer Genome Consortium PedBrain Tumor Project. Recurrent MET fusion genes represent a drug target in pediatric glioblastoma. *Nat Med*. 2016;22(11):1314–20. <https://doi.org/10.1038/nm.4204>.
 40. Torre M, Jessop N, Hornick JL, Alexandrescu S. Expanding the spectrum of pediatric NTRK-rearranged fibroblastic tumors to the central nervous system: a case report with RBPMS-NTRK3 fusion. *Neuropathology*. 2018;38(6):624–30. <https://doi.org/10.1111/neup.12513>.
 41. Cocco E, Scaltriti M, Drilon A. NTRK fusion-positive cancers and TRK inhibitor therapy. *Nat Rev Clin Oncol*. 2018;15(12):731–47. <https://doi.org/10.1038/s41571-018-0113-0>.
 42. Pekmezci M, Villanueva-Meyer JE, Goode B, Van Ziffle J, Onodera C, Grenert JP, et al. The genetic landscape of ganglioglioma. *Acta Neuropathol Commun*. 2018;6(1):47. <https://doi.org/10.1186/s40478-018-0551-z>.
 43. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*. 2015;34(37):4845–54. <https://doi.org/10.1038/onc.2014.406>.
 44. Bar EE, Lin A, Tihan T, Burger PC, Eberhart CG. Frequent gains at chromosome 7q34 involving BRAF in pilocytic astrocytoma. *J Neuropathol Exp Neurol*. 2008;67(9):878–87. <https://doi.org/10.1097/NEN.0b013e3181845622>.
 45. Lin A, Rodriguez FJ, Karajannis MA, Williams SC, Legault G, Zagzag D, et al. BRAF alterations in primary glial and glioneuronal neoplasms of the central nervous system with identification of 2 novel KIAA1549:BRAF fusion variants. *J Neuropathol Exp Neurol*. 2012;71(1):66–72. <https://doi.org/10.1097/NEN.0b013e31823f2cb0>.
 46. Yamashita S, Takeshima H, Matsumoto F, Yamasaki K, Fukushima T, Sakoda H, et al. Detection of the KIAA1549-BRAF fusion gene in cells forming microvascular proliferations in pilocytic astrocytoma. *PLoS One*. 2019;14(7): e0220146. <https://doi.org/10.1371/journal.pone.0220146>.
 47. Jones DT, Hutter B, Jager N, Korshunov A, Kool M, Warnatz HJ, et al. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat Genet*. 2013;45(8):927–32. <https://doi.org/10.1038/ng.2682>.
 48. Okonechnikov K, Imai-Matsushima A, Paul L, Seitz A, Meyer TF, Garcia-Alcalde F. InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data. *PLoS One*. 2016;11(12): e0167417. <https://doi.org/10.1371/journal.pone.0167417>.
 49. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*. 2011;7(5):e1001138. <https://doi.org/10.1371/journal.pcbi.1001138>.
 50. Cmero M, Schmidt B, Majewski IJ, Ekert PG, Oshlack A, Davidson NM. MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. *bioRxiv*. 2021. <https://doi.org/10.1101/2020.06.03.131532>.
 51. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*. 2018;33:690–705.e9.
 52. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
 53. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166–9. <https://doi.org/10.1093/bioinformatics/btu638>.
 54. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>.
 55. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50. <https://doi.org/10.1073/pnas.0506580102>.
 56. Seraseq Tumor Fusion RNA Mix3. <https://www.seracare.com/globalassets/seracare-resources/pr-0710-0431-seraseq-tumor-fusion-rna-mix-v3-10330722.pdf>. Accessed 29 Sept 2021.
 57. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019; 47(D1):D941–7. <https://doi.org/10.1093/nar/gky1015>.
 58. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480–9. <https://doi.org/10.1093/nar/gkaa1100>.
 59. Panigrahi P, Jere A, Anamika K. FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. *PLoS One*. 2018;13(5):e0196588. <https://doi.org/10.1371/journal.pone.0196588>.
 60. Zhang J, Gao T, Maher CA. INTEGRATE-Vis: a tool for comprehensive gene fusion visualization. *Sci Rep*. 2017;7(1):17808. <https://doi.org/10.1038/s41598-017-18257-2>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

