annoFuse: an R Package to annotate and prioritize putative oncogenic RNA fusions

Krutika S. Gaonkar[1,2,4], Komal S. Rathi[1,2,4], Payal Jain[1,4], Yuankun Zhu[1,4], Nicholas A. Chimicles[1,2], Ammar S. Naqvi[1,2,4], Miguel A. Brown[1,4], Bo Zhang[1,4], Phillip B. Storm[4], John M. Maris[3], Adam C. Resnick[1,2,4], Pichai Raman[1,2,4], Jaclyn N. Taroni[5,*], Jo Lynne Rokita[1,2,4*]

[1]Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104

[2]Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104

[3]Division of Oncology, Children's Hospital of Philadelphia and Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, 19104

[4]Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104

[5]Alex's Lemonade Stand Foundation Childhood Cancer Data Lab, Philadelphia, PA, 19102

*Corresponding Authors: Jo Lynne Rokita, Ph.D. (rokita@email.chop.edu) and Jaclyn N. Taroni, Ph.D. (jaclyn.taroni@ccdatalab.org)

Abstract

Background

Gene fusion events are a significant source of somatic variation across adult and pediatric cancers and have provided some of the most effective clinically relevant therapeutic targets, yet computational algorithms for fusion detection from RNA sequencing data show low overlap of predictions across methods. In addition, events such as polymerase read-throughs, mis-mapping due to gene homology, and fusions occurring in healthy normal tissue require stringent filtering, making it difficult for researchers and clinicians to discern gene fusions that might be true underlying oncogenic drivers of a tumor and in some cases, appropriate targets for therapy.

Results

Here, we present *annoFuse*, an R package developed to annotate and identify biologically-relevant expressed gene fusions, along with highlighting recurrent novel fusions in a given cohort. We applied *annoFuse* to STAR-Fusion and Arriba results for 1,028 pediatric brain tumor samples provided as part of the Open Pediatric Brain Tumor Atlas (OpenPBTA) Project. First, we used FusionAnnotator to identify and filter “red flag” fusions found in healthy tissues or in gene homology databases. Using *annoFuse*, we filtered out fusions known to be artifactual and retained high-quality fusion calls. Second, we prioritized and captured known and putative oncogenic driver fusions previously reported in TCGA or fusions containing gene partners that are known oncogenes, tumor suppressor genes, COSMIC genes, or transcription factors. Using the PFAM database, we defined fusions containing kinases and annotated each for domain retention. Fusions can be plotted to visualize breakpoints by transcript with domain annotation. Finally, we determined recurrent fusions across the OpenPBTA cohort and recurrently-fused genes within each histology.

Conclusions

annoFuse provides a standardized filtering and annotating method for gene fusion calls from STAR-Fusion and Arriba by merging, filtering, and prioritizing putative oncogenic fusions across large cancer datasets, as demonstrated here with the OpenPBTA project. We are

expanding the package to be widely-applicable to other fusion algorithms and expect *annoFuse* to provide researchers a method for quickly evaluating and prioritizing fusions in patient tumors.

Keywords (3-10)

RNA-Seq, gene fusions, annotation tool, oncogenes, cancer

Background

Gene fusions arise in cancer as a result of aberrant chromosomal rearrangements or defective splicing, which bring together two unrelated genes that are then expressed as a novel fusion transcript (1). Detection of therapeutically-targetable fusion calls is of clinical importance and computational methods are constantly being developed to detect these events in real-time. Recent comparative studies show low concordance of fusion predictions across methods (2), suggesting that many predictions may not represent true events. Additionally, transcriptional read-throughs (3), in which the polymerase machinery skips a stop codon and reads through a neighbouring gene, as well as fusions that involve non-canonical transcripts or gene-homologs, are prevalent in disease datasets, yet the biological relevance of such events is still unclear. This makes it difficult for both researchers and clinicians to prioritize disease-relevant fusions and discern the underlying biological mechanisms and thus, appropriate fusion-directed therapy. Gene fusion events leading to gain-of-function or loss-of-function in kinases and putative tumor suppressor genes, respectively, have been shown to be oncogenic drivers with therapeutic potential, especially in pediatric tumors (4–6). For example, the recurrent fusion *KIAA1549-BRAF* is found across 66-80% of low grade gliomas and results in a fusion transcript that has constitutive BRAF kinase activity (7). *EWSR1-FLI1* is found in nearly 100% of Ewing's sarcoma and forms an oncogenic RNA complex, driving tumorigenesis (8). To capture highly recurrent and validated fusions such as these, the fusion databases ChimerDB (9) and TumorFusions (10) were developed from RNA fusions called in The Cancer Genome Atlas (TCGA) (11,12) samples. In such large-scale cancer studies, a single algorithm was routinely used to detect fusion calls because using multiple callers often adds complexity of annotation and integration. However, it is

now common practice to incorporate data from multiple algorithms to reliably define the fusion landscape of cancers. Recent efforts have reported the importance of using systematic filtering and aggregation of multiple fusion callers to expand the number of biologically-relevant fusions in adult cancers (12,13). However, to our knowledge there are no tools or packages developed to filter, aggregate, and detect recurrent and putative oncogenic fusions in a systematic, flexible, and reproducible manner. Despite the existence of a few tools with working open-source code which can assist in fusion annotation or prioritization, only three are algorithm-agnostic with the remaining tools relying on outdated fusion algorithms, rendering them unusable on current gold standard tools to date, such as STAR-Fusion (14) and Arriba (15) (**Table 1**).

Here, we developed and applied *annoFuse* to gene fusion results from STAR-Fusion and Arriba for 1,028 pediatric brain tumor samples provided as part of the OpenPBTA Project (16). First, we used FusionAnnotator to identify and filter red flag fusions, those found in healthy tissues or in gene homology databases. Using *annoFuse*, we remove fusions known or predicted to be artifactual and retain high-quality fusion calls. Second, for the fusions that pass quality checks, fusions are annotated if previously found within TCGA and each gene partner is annotated as an oncogene, tumor suppressor, kinase, transcription factor, and/or whether it has been reported in the Catalogue of Somatic Mutations in Cancer (COSMIC). Finally, we determined the recurrence pattern for fusions across the cohort and also recurrently-fused genes within each cancer histology.

Implementation

We implemented *annoFuse* using the R programming language (R version 3.5.1 (2018-07-02). The R packages required to install and run *annoFuse* are reshape2, dplyr, tidyr, ggplot2, and plotly, with optional packages: knitr and rmarkdown.

R package overview

The *annoFuse* package was developed to provide a standardized filtering and annotation method for fusion calls from Arriba and STAR-Fusion, first and second place winners of the 2017 DREAM SMC-RNA Challenge, respectively (17). In a 2019 assessment of 23 fusion algorithms for cancer biology, both Arriba and STAR-Fusion ranked in the top three fastest and most accurate tools (18). *annoFuse* utilizes a four-step process (**Figure 1**) that is available with flexible functions to perform downstream functions such as merging, filtering, and prioritization of fusion calls from multiple fusion calling algorithms on single or batch samples.

RNA Expression and Fusion Calls

Currently, *annoFuse* is compatible with fusion calls generated from Arriba v1.1.0 (19) and/or STAR-Fusion 1.5.0 (14). Both tools utilize aligned BAM and chimeric SAM files from STAR as inputs and STAR-Fusion calls are annotated with GRCh38_v27_CTAT_lib_Feb092018.plug-n-play.tar.gz, which is provided in the STAR-fusion release. Arriba should be provided with strandedness information, or set to auto-detection for poly-A enriched libraries. Additionally, the blacklist file, blacklist_hg38_GRCh38_2018-11-04.tsv.gz contained in the Arriba release tarballs, should be used to remove recurrent fusion artifacts and transcripts present in healthy tissue. An expression matrix with FPKM or TPM values is also required; the matrix should have a column "*GeneSymbol*" following the same gene naming convention as found in fusion calls.

Fusion Call Preprocessing

We leveraged the fact that STAR-Fusion uses FusionAnnotator as its final step and thus, require all fusion calls be annotated with FusionAnnotator v. 0.2.0 to contain the additional column, *"annots"*. Finally, fusion calls for all samples should be merged into a single TSV file with

an additional column, "*tumor_id*", which will enable artifact filtering, annotation, fusion prioritization, and determination of recurrence.

annoFuse Steps:

Step 1: Fusion Standardization

To obtain a standardized format for fusion calls from multiple fusion calls we use the *fusion_standardization* function to format caller specific output files to a standardizedFusionCalls format defined in the package README. *fusion_standardization* allows users to standardize fusion calls from multiple callers, as well as annotate calls with other databases from the "annots" column, which can then be used for filtering.

Step 2: Fusion Filtering

Events such as polymerase read-throughs, mis-mapping due to gene homology, and fusions occurring in healthy normal tissue confound detection for true recurrent fusion calls and false positives for genes considered as oncogenic, tumor suppressor or kinases in some cases. In this step, we filter the standardized fusion calls to remove artifacts and false positives (**Table 2**) using the function *fusion_filtering_QC.* The parameters are flexible to allow users to annotate and filter the fusions with *a priori* knowledge of their call set. For example, since the calls are pre-annotated with FusionAnnotator, the user can remove fusions known to be red-flags as annotated with any of the following databases GTEx_recurrent_STARF2019, HGNC_GENEFAM, DGD_PARALOGS, Greger_Normal, Babiceanu_Normal, BodyMap, and ConjoinG. This is done using the parameter, *artifact_filter = "GTEx_recurrent_STARF2019* | *DGD_PARALOGS* | *Normal* | *BodyMap* | *ConjoinG"*. Of note, we decided not to remove genes annotated in HGNC_GENEFAM, as this database contains multiple oncogenes and their removal resulted in missed true fusions using our validation truth set. Read-throughs annotated by any algorithm can also be removed at this step by using parameter *"readthroughFilter=TRUE"*. During validation, we observed the real oncogenic fusion, *P2RY8-CRLF2 (20,21),* annotated as a read-through in acute lymphoblastic leukemia samples, therefore, we implemented a condition such that if a fusion is

annotated as a read-through, but is present in the Mitelman cancer fusion database, we scavenge these fusions back as true positive calls.

This function also allows users to flexibly filter out fusions predicted to be artifactual while retaining high-quality fusion calls using junction read support of ≥ 1 (default) and spanning fragment support of < 10 (default) reads compared to the junction read count, as disproportionate spanning fragment support indicates false positive calls (19). Finally, if both genes of the fusion are deemed not expressed < 1 FPKM or TPM (default), the fusion transcript calls can be removed using function *expressionFilterFusion*.

Step 3: Fusion Annotation

The *annotateFusionCalls* function annotates standardized fusion calls and performs customizable fusion annotation based on user gene lists as input. As a default setting, we provide lists of, and annotate gene partners as, oncogenes, tumor suppressor genes, and oncogenic fusions.

The optional *ZscoredAnnotation* function provides z-scored expression values from a user-supplied matrix such as GTEx or within cohort to compare samples with and without the fusion to look for over or under expression of fused genes compared to normal using a zscoreFilter. A cutoff of 2 (default) is set to annotate any score > 2 standard deviations away from the median as differentially-expressed. Researchers can then use this information to decide whether to perform additional downstream filtering.

Single sample run

For single samples, we developed the *annoFuseSingleSample* function which performs fusion standardization of Arriba and STAR-Fusion calls, fusion filtering, and fusion annotation with user-provided gene and fusion reference lists.

Project-Specific Filtering

Each study often requires additional downstream analyses be performed once high-quality annotated fusion calls are obtained. We developed functions to enable analyses at a cohort (or project-level) and/or group-level (eg: histologies) designed to remove cohort-specific artifactual calls while retaining high-confidence fusion calls. The function *called_by_n_callers* annotates the number of algorithms that detected each fusion. We retained fusions with genes not annotated with the gene lists above (eg: oncogene, etc) that were detected by both algorithms as inframe or frameshift but not annotated as LOCAL_INVERSION or LOCAL_REARRANGEMENT by FusionAnnotator, as these could represent novel fusions. Additionally, *samplecount_fusion_call* identifies fusions recurrently called in (default ≥ 2) samples within each group. At the group-level, we add *groupcount_fusion_calls* (default ≥ 1) to remove fusions that are present in more than one type of cancer. At the sample level, *fusion_multifused* detects fusions in which one gene partner is detected with multiple partners (default ≥ 5), and we remove these as potential false positives. This enables *annoFuse* to scavenge back potential oncogenic fusions which may have otherwise been filtered. Separately, the function *fusion_driver* retains only fusions in which a gene partner was annotated as a tumor suppressor gene, oncogene, kinase, transcription factor, and/or the fusion was previously found in TCGA. To further reduce the false positives and fusions containing pseudogenes from the cohort, we next filtered fusions using a cutoff of present in > 4 broad histologies after reviewing the fusion distributions within the OpenPBTA Project (**Additional File 1 : Figure S1**). Finally, both sets of fusions are merged into a final set of putative oncogenic fusions.

Fusion Domain Annotation

The *getPfamDomain* function in annoFuse provides domain annotation for each fused gene in standardized fusion calls. We used the UCSC pfamID Description database (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/pfamDesc.txt.gz) and domain location database (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ucscGenePfam.txt.gz), along with bioMart (22,23) gene coordinates to get genomic locations of each domain in a gene. By identifying the breakpoint within the gene coordinate, we annotate each domain as one of the following: DomainRetained=***Yes*** if domains are completely within the gene boundary and breakpoint, DomainRetained=***Partial*** if domains overlap the breakpoints, or DomainRetained=***No*** if domains are downstream of gene breakpoint. This annotation provides domain retention information which enables prioritization to generate new hypotheses, validate fusion transcript functional impact, and or identify targeted therapeutic options.

Visualization

Quick visualization of filtered and annotated fusion calls can provide information useful for review and downstream analysis. We provide the function *plotSummary* which provides distribution of intra-chromosomal and inter-chromosomal fusions, number of in-frame and frameshift calls per algorithm, and distribution of gene biotypes, kinase group, and oncogenic annotation. If project-specific filtering is utilized, barplots displaying recurrent fusion and recurrently-fused genes can be generated using *plotRecurrentFusion* and *plotRecurrentFusedGene,* respectively. *plotBreakpoints* can be used to generate all transcripts and breakpoints per gene to visualize the exon and domain retention resulting from the fusion (**Figure 4**).

Results and Discussion

Technical validation of annoFuse

Few gene fusion “truth” sets exist and those that do consist of simulated data or synthetic fusions spiked into breast cancer cell lines or total RNA (17,18,24). We therefore utilized a recent study in which high-confidence fusions were reported in 244 patient-derived xenograft models from the Pediatric Preclinical Testing Consortium (PPTC) (25). A set of 27 fusions were molecularly validated from acute lymphoblastic leukemia (ALL) models in the PPTC dataset and contains a “truth” set. We first ran Arriba on the PPTC dataset and determined which fusions were detected using only STAR-Fusiona and Arriba (**Table 3**). Next, we used *annoFuse* to filter and prioritize putative oncogenic fusions. **Table 3** shows the performance of *annoFuse*, which retained the 23 true positive ALL fusions (100%) and retained an average of 96% fusions previously defined as high-confidence (putative oncogenic fusions) in (25). Interestingly, only 114/166 total fusions were detected using STAR-Fusion and Arriba (23/27 within the “truth” set), implying gold standard algorithms alone still fail to capture the full landscape of gene fusions, reflecting that additional algorithms should be integrated into our workflow. Of the 114 fusions we detected, 110 were retained as putative oncogenic fusions using *annoFuse*. The four fusions *annoFuse* did not retain were removed with the “read-through” filter, which can be turned off as an option.

Case study with annoFuse using OpenPBTA

As proof of concept, we utilized RNA expression generated by STAR-RSEM (26) and fusion calls generated by Arriba v1.1.0 (19) and/or STAR-Fusion 1.5.0 (14) which were released as part of the Pediatric Brain Tumor Atlas (27). The algorithms were run as described in ***RNA Expression and Fusion Calls.*** The RNA expression and fusion workflows are publicly available within the Gabriella Miller KidsFirst GitHub repository (28).

Following fusion standardization, annotation, and filtration, we applied project-specific filtering to the OpenPBTA RNA-Seq cohort (n = 1,028 biospecimens from n = 943 patients). **Figure 2** is a sample summary PDF designed to give the user an overall glance of the fusion annotations and fusion characteristics within the cohort. From the OpenPBTA cohort, it is clear that there were predominantly more intra-chromosomal fusions called than inter-chromosomal fusions, even after filtering for read-through events (**Figure 2A**). While a low-grade astrocytic

tumors are the major pediatric brain tumor subtype known to be enriched for gene fusions, it was surprising to observe a large number of fusions in diffuse astrocytic and oligodendroglial tumors and the project-specific utility of *annoFuse* allows researchers to further prioritize fusions. Histologies within the OpenPBTA project were classified according to broad WHO 2016 subtypes (29).

The number of in-frame and frameshift fusions per algorithm were roughly equivalent within each STAR-Fusion and Arriba fusion calls (**Figure 2B**). **Figure 2C** depicts the density of genes categorized by gene biotype (biological type), and as expected, the filtered and annotated calls were enriched for biologically-functional fusions; the majority of gene partners are classified as protein-coding. The majority of gene partners were annotated as tyrosine kinase (TK) or tyrosine kinase-like (TKL) (**Figure 2D**). In **Figure 2E**, the user can explore the biological and oncogenic relevance of the fusions across histologies. Of the fusions harboring kinase domains, we found that the majority of 3' partners retained kinase domains, supporting these fusions as functionally relevant (**Additional file 1 : Figure S2**).

Following project-specific filtering, we observed *KIAA1549-BRAF* fusions as the most recurrent in-frame fusion in our cohort (n = 109/898), which was expected as *KIAA1549-BRAF* expressing low-grade astrocytic tumors comprise the largest representative histology in the OpenPBTA cohort (n = 236/898). *C11orf95-RELA* was predominant in ependymal tumors (n = 25/80), as expected in supratentorial ependymomas (30). Other expected recurrent oncogenic fusions obtained through *annoFuse* were *EWSR1-FLI1* in CNS Ewing sarcomas (31), and *KANK1-NTRK2, MYB-QKI,* and *FAM131B-BRAF* in low-grade astrocytic tumors (4,32) (**Figure 3A**). In addition to recurrent fusions, we also detect recurrently-fused genes to account for partner promiscuity. This enables us to see a broader picture of gene fusions, specifically within diffuse astrocytic and oligodendroglial tumors, in which we see fusions prevalent in *ST7, MET, FYN, REV3L, AUTS2,* and *ROS1,* and meningiomas, in which *NF2* fusions are common. (**Figure 3B**). Finally, we added functionality to visualize domain information (**Figure 4**) to quickly scan for domains retained and lost across the dataset.

The few openly-available fusion annotation and prioritization tools (**Table 1**) each have specific annotation and/or prioritization functionalities, however, the majority are no longer maintained and only work on outdated fusion algorithms. Oncofuse (33), Pegasus (34), chimera (35), and co-Fuse (36) have not been updated in two or more years, and as a result, these tools lack compatibility with newer and improved fusion algorithms. The chimeraviz R package (37) is well-maintained and compatible with nine fusion algorithms, but only performs visualizations of fusions, thus prioritization is not possible using this tool. The remaining four tools are algorithm agnostic, yet perform only specific aspects of annotation and prioritization. FusionHub (38) is a web-based tool which enables annotation of fusions with 28 databases, however, is not programmatically scalable. FusionAnnotator (39) determines the presence of fusions in 15 cancer-associated databases, oncogene lists, and seven databases for fusions not relevant in cancer. AGFusion (40) annotates protein domains, and Fusion Pathway (41) utilizes fusion and protein domain annotations in gene set enrichment analysis (GSEA) to infer oncogenic pathway association. When used alone, none of these tools flexibly perform fusion annotation and prioritization. Therefore, we leverage the algorithm agnostic capabilities of FusionAnnotator to pre-annotate fusion input from STAR-Fusion and Arriba.

By integrating FusionAnnotator with functionality of the current gold standard algorithms STAR-Fusion and Arriba, we were able to improve the aforementioned tools' capabilities by meeting the current demands of the research community. We provide the user with flexible filtering parameters and envision *annoFuse* will be used to quickly filter sequencing artifacts and false positives, as well as further annotate fusions for additional biologically functionality (eg: kinases, transcription factors, oncogenes, tumor suppressor genes) to increase the signal to noise ratio in a cohort of fusion calls. Additionally, users can opt to simply annotate and filter artifacts or use *annoFuse* to functionally prioritize fusions as putative oncogenic drivers. During the prioritization steps, we filter based on genes with cancer relevance (see biological functionality list above) and perform analysis of fusion and fused-gene recurrence to create a stringently filtered, prioritized list of fusions likely to have oncogenic potential.

As an additional feature, we plan to add expression-based comparison of genes between fused samples, normal, and within a histology or cohort. We also plan to add additional fusion algorithms currently used by the community, such as deFuse, FusionCatcher, and SOAPfuse, to further increase the applicability of *annoFuse*. Future features could link domain retention to drug databases to predict fusion-directed targeting strategies.

Conclusions

Gene fusions provide a unique mutational context in cancer in which two functionally-distinct genes could be combined to function as a new biological entity. Despite showing great promise as diagnostic, prognostic, and therapeutic targets, translation in the oncology clinic is not yet accelerated for gene fusions. This has been partly due to limited translation of the large number of computationally-derived fusion results into biologically meaningful information. In our efforts to address this, we introduce *annoFuse*, an R Package to annotate and prioritize putative oncogenic RNA fusions, providing a range of functionalities to filter and annotate fusion calls from multiple algorithms. We include a cancer-specific workflow to find recurrent, oncogenic fusions from large cohorts containing multiple cancer histologies. The filtering and annotation steps within *annoFuse* enable users to integrate calls from multiple algorithms to improve high-confidence, consensus fusion calling. The lack of concordance among algorithms as well as variable accuracy with fusion truth sets (2,18) adds analytical complexity for researchers and clinicians aiming to prioritize research or therapies based on fusion findings. Through *annoFuse*, we add algorithm flexibility and integration, to identify recurrent fusions and/or recurrently-fused genes as novel oncogenic drivers. We expect *annoFuse* to be broadly applicable to cancer datasets and to facilitate researchers to better inform preclinical studies targeting novel, putative oncogenic fusions and ultimately, aid in the rational design of therapeutic modulators of gene fusions in cancer.

Availability and requirements

Project name: annoFuse: an R Package to annotate and prioritize putative oncogenic RNA fusions

Project home page: https://github.com/d3b-center/annoFuse

Operating system(s): Platform independent

Programming language: R 3.5.1

Other requirements: e.g. Java 1.3.1 or higher, Tomcat 4.0 or higher

License: MIT

Any restrictions to use by non-academics: e.g. licence needed

List of abbreviations

ALL: Acute Lymphoblastic Leukemia

BAM : Binary Alignment Map

COSMIC : Catalogue Of Somatic Mutations In Cancer

CNS : Central Nervous System

DGD_PARALOGS : Duplicated Genes Database annotated paralogs

GSEA : Gene Set Enrichment Analysis

HGNC_GENEFAM : HGNC annotated gene family

FPKM : Fragments Per Kilobase Million

OpenPBTA : Open Pediatric Brain Tumor Atlas

PI3_PI4_kinase=Phosphatidylinositol 3- and 4-kinase

Pkinase=Protein kinase domain

Pkinase_C=Protein kinase C terminal domain

Pkinase_Tyr=Protein tyrosine kinase

PPTC **:** Pediatric Preclinical Testing Consortium

RNA : Ribonucleic Acid

SAM : Sequence Alignment Map

SMC-RNA : Somatic Mutation Calling RNA DREAM Challenge (SMC-RNA)

TCGA : The Cancer Genome Atlas

TSV : Tab Separated Value

TPM : Transcripts Per Kilobase Per Million

WHO : World Health Organization

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All brain tumor raw data are available by download from the Gabriella Miller Kids First Data Resource Center with a data access agreement through the Children's Brain Tumor Tissue Consortium and processed data (release-v14-20200203) available by download through the OpenPBTA project's GitHub repository: https://github.com/AlexsLemonade/OpenPBTA-analysis. PPTC data are available by download from dbGAP (Accession Number phs001437.v1.p1) with a data access agreement.

Competing interests

The authors declare no competing interests.

Funding

Funding for this research was provided by Children's Hospital of Philadelphia Division of Neurosurgery (PBS and ACR), NIH grants U2C HL138346-03 (ACR), U24 CA220457-03 (ACR), and R35 CA220500 (JMM), and Alex's Lemonade Stand via a Young Investigator Award (JLR), Catalyst Award (JLR), and CCDL funding (JNT).

Authors' contributions

Conceptualization: KSG, JLR, JNT

Methodology: KSG, KSR, PR, JLR, JNT, MAB, BZ, YZ, NAC

Validation: KSG, JLR, JNT

Formal Analysis: KSG, JNT

Investigation: KSG, JLR

Resources: JLR, PBS, ACR

Data Curation: KSG, YZ, JLR, MAB, BZ

Writing - Original Draft: KSG, JLR

Writing - Review & Editing: KSG, JLR, PJ, ASN, JNT

Visualization: KSG, JLR

Supervision: JLR, JNT

Funding Acquisition: JLR, ACR, PBS, JMM, JNT

Acknowledgements

We acknowledge the Children's Brain Tumor Tissue Consortium and the Pacific Pediatric Neuro-Oncology Consortium for generating the RNA-Seq data from tumor tissue and the Gabriella Miller Kids First Data Resource Center for processing and hosting the data. Finally, we graciously thank the patients and families for donating tumor tissue, without which, this research would not be possible.

References

1. Wu H, Li X, Li H. Gene fusions and chimeric RNAs, and their implications in cancer. Genes Dis. 2019 Dec;6(4):385–90.

2. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data [Internet]. Vol. 6, Scientific Reports. 2016. Available from: http://dx.doi.org/10.1038/srep21597

3. 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 [Internet]. 2018 Jan 16;9(1). Available from: http://dx.doi.org/10.3390/genes9010040

4. Bandopadhayay P, Ramkissoon LA, Jain P, Bergthold G, Wala J, Zeid R, et al. MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism. Nat Genet. 2016 Mar;48(3):273–82.

5. Jain P, Fierst TM, Han HJ, Smith TE, Vakil A, Storm PB, et al. CRAF gene fusions in pediatric low-grade gliomas define a distinct drug response based on dimerization profiles. Oncogene. 2017 Nov 9;36(45):6348–58.

6. Jain P, Surrey LF, Straka J, Luo M, Lin F, Harding B, et al. Novel FGFR2-INA fusion identified in two low-grade mixed neuronal-glial tumors drives oncogenesis via MAPK and PI3K/mTOR pathway activation. Acta Neuropathol. 2018 Jul;136(1):167–9.

7. Jones DTW, Kocialkowski S, Liu L, Pearson DM, Magnus Backlund L, Ichimura K, et al. Tandem Duplication Producing a Novel Oncogenic BRAF Fusion Gene Defines the Majority of Pilocytic Astrocytomas. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-08-2097

8. Aurias A, Rimbaut C, Buffe D, Zucker JM, Mazabraud A. Translocation involving chromosome 22 in Ewing's sarcoma. A cytogenetic study of four fresh tumors. Cancer Genet Cytogenet. 1984 May;12(1):21–5.

9. Lee M, Lee K, Yu N, Jang I, Choi I, Kim P, et al. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. Nucleic Acids Res. 2017 Jan 4;45(D1):D784–9.

10. Hu X, Wang Q, Tang M, Barthel F, Amin S, Yoshihara K, et al. TumorFusions: an integrative resource for cancer-associated transcript fusions. Nucleic Acids Res. 2018 Jan 4;46(D1):D1144–9.

11. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.

12. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018 Apr 5;173(2):321–37.e10.

13. Gao Q, Liang W-W, 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 Apr 3;23(1):227–38.e3.

14. Haas BJ, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq [Internet]. bioRxiv. 2017 [cited 2019 Oct 23]. p. 120295. Available from: https://www.biorxiv.org/content/10.1101/120295v1

15. arriba [Internet]. Github; [cited 2019 Nov 6]. Available from: https://github.com/suhrig/arriba

16. Waller J. OpenPBTA: A New Collaborative Effort to Accelerate Discoveries Empowered by the Pediatric Brain Tumor Atlas - Children's Brain Tumor Tissue Consortium [Internet]. Children's Brain Tumor Tissue Consortium. 2019 [cited 2019 Nov 4]. Available from: https://cbttc.org/2019/09/29/openpbta-a-new-collaborative-effort-to-accelerate-discoveries-empowered-by-the-pediatric-brain-tumor-atlas/?fbclid=IwAR185_NzFLeEgFmhaMnKNd_cYTVUSSYcYG_ZPEbmFQLuB7sVX31CRSiWtc0

17. Bionetworks S. Synapse | Sage Bionetworks [Internet]. [cited 2019 Nov 6]. Available from: https://www.synapse.org/

18. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019 Oct 21;20(1):213.

19. arriba [Internet]. Github; [cited 2019 Nov 4]. Available from: https://github.com/suhrig/arriba

20. Cario G, Zimmermann M, Romey R, Gesk S, Vater I, Harbott J, et al. Presence of the P2RY8-CRLF2 rearrangement is associated with a poor prognosis in non--high-risk precursor B-cell acute lymphoblastic leukemia in children treated according to the ALL-BFM 2000 protocol. Blood. 2010;115(26):5393–7.

21. Panzer-Grümayer R, Köhrer S, Haas OA. The enigmatic role(s) of P2RY8-CRLF2. Oncotarget. 2017 Nov 14;8(57):96466–7.

22. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439–40.

23. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009 Jul 23;4(8):1184–91.

24. Tembe WD, Pond SJK, Legendre C, Chuang H-Y, Liang WS, Kim NE, et al. Open-access synthetic spike-in mRNA-seq data for cancer gene fusions. BMC Genomics. 2014 Sep 30;15:824.

25. Rokita JL, Rathi KS, Cardenas MF, Upton KA, Jayaseelan J, Cross KL, et al. Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design. Cell Rep. 2019 Nov 5;29(6):1675–89.e9.

26. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017 Apr 11;35(4):314–6.

27. Ijaz H, Koptyra M, Gaonkar KS, Rokita JL, Baubet VP, Tauhid L, et al. Pediatric High Grade Glioma Resources from the Children's Brain Tumor Tissue Consortium (Cbttc). Neuro Oncol [Internet]. 2019 Oct 15; Available from: http://dx.doi.org/10.1093/neuonc/noz192

28. kf-rnaseq-workflow [Internet]. Github; [cited 2019 Nov 4]. Available from: https://github.com/kids-first/kf-rnaseq-workflow

29. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016 Jun 1;131(6):803–20.

30. Fukuoka K, Kanemura Y, Shofuda T, Fukushima S, Yamashita S, Narushima D, et al. Significance of molecular classification of ependymomas: C11orf95-RELA fusion-negative supratentorial ependymomas are a heterogeneous group of tumors. Acta Neuropathol Commun. 2018 Dec 4;6(1):134.

31. Brohl AS, Solomon DA, Chang W, Wang J, Song Y, Sindiri S, et al. The genomic landscape of the Ewing Sarcoma family of tumors reveals recurrent STAG2 mutation. PLoS Genet. 2014;10(7):e1004475.

32. Venneti S, Huse JT. The evolving molecular genetics of low-grade glioma. Adv Anat Pathol. 2015 Mar;22(2):94–101.

33. Shugay M, Ortiz de Mendíbil I, Vizmanos JL, Novo FJ. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics. 2013 Oct 15;29(20):2539–46.

34. Abate F, Zairis S, Ficarra E, Acquaviva A, Wiggins CH, Frattini V, et al. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC Syst Biol. 2014 Sep 4;8:97.

35. Calogero RA, Carrara M, Beccuti M, Cordero F. chimera: A package for secondary analysis of fusion products version 1.28.0 from Bioconductor [Internet]. 2019 [cited 2019 Nov 6]. Available from: https://rdrr.io/bioc/chimera/

36. co-fuse [Internet]. Github; [cited 2019 Nov 6]. Available from: https://github.com/sakrapee/co-fuse

37. Lågstad S, Zhao S, Hoff AM, Johannessen B, Lingjærde OC, Skotheim RI. chimeraviz: a tool for visualizing chimeric RNA. Bioinformatics. 2017 Sep 15;33(18):2954–6.

38. 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 May 1;13(5):e0196588.

39. FusionAnnotator [Internet]. Github; [cited 2019 Nov 6]. Available from: https://github.com/FusionAnnotator

40. Murphy C, Elemento O. AGFusion: annotate and visualize gene fusions [Internet]. bioRxiv. 2016 [cited 2019 Nov 6]. p. 080903. Available from: https://www.biorxiv.org/content/10.1101/080903v1.abstract

41. Wu C-C, Beird HC, Zhang J, Andrew Futreal P. FusionPathway: Prediction of pathways and therapeutic targets associated with gene fusions in cancer [Internet]. Vol. 14, PLOS Computational Biology. 2018. p. e1006266. Available from: http://dx.doi.org/10.1371/journal.pcbi.1006266

Figures, Tables, and Additional Files

Figure 1. Graphical representation of pipeline. The *fusion_standardization* function standardizes calls from fusion callers to retain information regarding fused genes, breakpoints, reading frame information, as well as annotation from FusionAnnotator. Standardized fusion calls use *fusion_filtering_QC* to remove false positives such as fusions with low read support, annotated as read-throughs, found in normal and gene homolog databases and remove non-expressed fusions using *expressionFilterFusion*. Calls are annotated with *annotateFusionCalls* to include useful biological features of interest (eg. Kinase, Tumor suppressor etc.) Project-specific filtering captures recurrent fused genes using functions to filter (shown in boxes) as well

as putative driver fusion. Outputs available from *annoFuse* include TSV files of annotated and prioritized fusions, a PDF summary of fusions, and recurrently-fused gene/fusion plots.

Figure 2. Fusion annotations generated by annoFuse (A) Distribution of intra- and inter-chromosomal fusions across histologies. (B) Transcript frame distribution of fusions detected by Arriba and STAR-Fusion algorithms. (C) Bubble plot of gene partner distribution with respect to ENSEMBL biotype annotation (Size of circle proportional to number of genes). (D) Barplots representing the distribution of kinase groups represented in the PBTA cohort annotated by gene partner. (Alpha_kinase=Alpha-kinase family, Hexokinase_2=Hexokinase, PI3_PI4_kinase=Phosphatidylinositol 3- and 4-kinase, Pkinase=Protein kinase domain, Pkinase_C=Protein kinase C terminal domain,Pkinase_Tyr=Protein tyrosine kinase) (E) Bubble plot representing the distribution of fused genes as oncogenes, tumor suppressor genes, kinases, COSMIC, predicted and curated transcription factors (Size of circle proportional to number of genes). Genes belonging to more than one category are represented in each. In all panels except for B, fusion calls were merged from both STAR-Fusion and Arriba.

Figure 3. Recurrent fusion plots generated by annoFuse. Bar plots as representative of histology showing recurrent fusion calls by number of patients (A) and recurrently-fused genes by number of patients (B) after filtering and annotation.

Figure 4. Breakpoint distribution for *KIAA1549-BRAF* fusion. Displayed are each fusion gene, all known transcripts, and their genomic coordinates. Red dotted lines in each gene panel are the locations of breakpoints detected for *KIAA1549* (3') and *BRAF* (5') compiled from both Arriba and STAR-Fusion. Strand information is depicted with an arrow in the gene and domain as colored boxes. Black boxes represent exons for each transcript.

Additional file 1

Figure S1. Fusions found in more than 1 histology. Barplots represent the number of histologies in which each fusion was observed. Dotted line represents the cut off (> 4 counts) used to remove potential false positives and fusions containing pseudogenes**.**

Figure S2. Distribution of kinase genes fused in 5' and 3' genes per histologies. For each broad histology, pie charts represent the percentage of fusions in which Gene1A (5') or Gene1B (3') retain their kinase domains.

Table 1. Available fusion annotation and prioritization tools. List of nine openly-available fusion annotation and prioritization software tools. Only AGFusion, FusionAnnotator, Fusion Pathway, and certain functions of FusionHub are algorithm agnostic, and most algorithms require outdated fusion algorithm input.

Table 2. Fusion filtering and annotation criteria. Fusion filtering criteria were developed to gather high quality recurrent fusion calls while retaining fusions containing oncogenes and/or tumor suppressor genes. Filtering is divided into 3 types 1) QC: filters known causes of false positives. 2) Gene-list: retains additional fusions in genes and fusions of interest list. 3) Recurrence: filters out non-recurrent fusions in genes not annotated as putative oncogenic. Annotation lists are also described.

Table 3. Validation of *annoFuse* prioritization using PPTC PDX fusion calls. Retention of high-confidence, putative oncogenic calls averaged 96% across the entire PPTC PDX dataset and was 100% for the ALL truth set (ALL = acute lymphoblastic leukemia). Column 1 = PPTC histology, Column 2 = fusion calls from STAR-Fusion, FusionCatcher, deFuse, and SOAPFuse which were filtered and reported as high-confidence in the PPTC dataset, Column 3 = PPTC reported fusions detected from STAR-Fusion and Arriba, Column 4 = Fusions retained following annoFuse filtering, Column 5 = Percent of fusions retained after applying *annoFuse*.

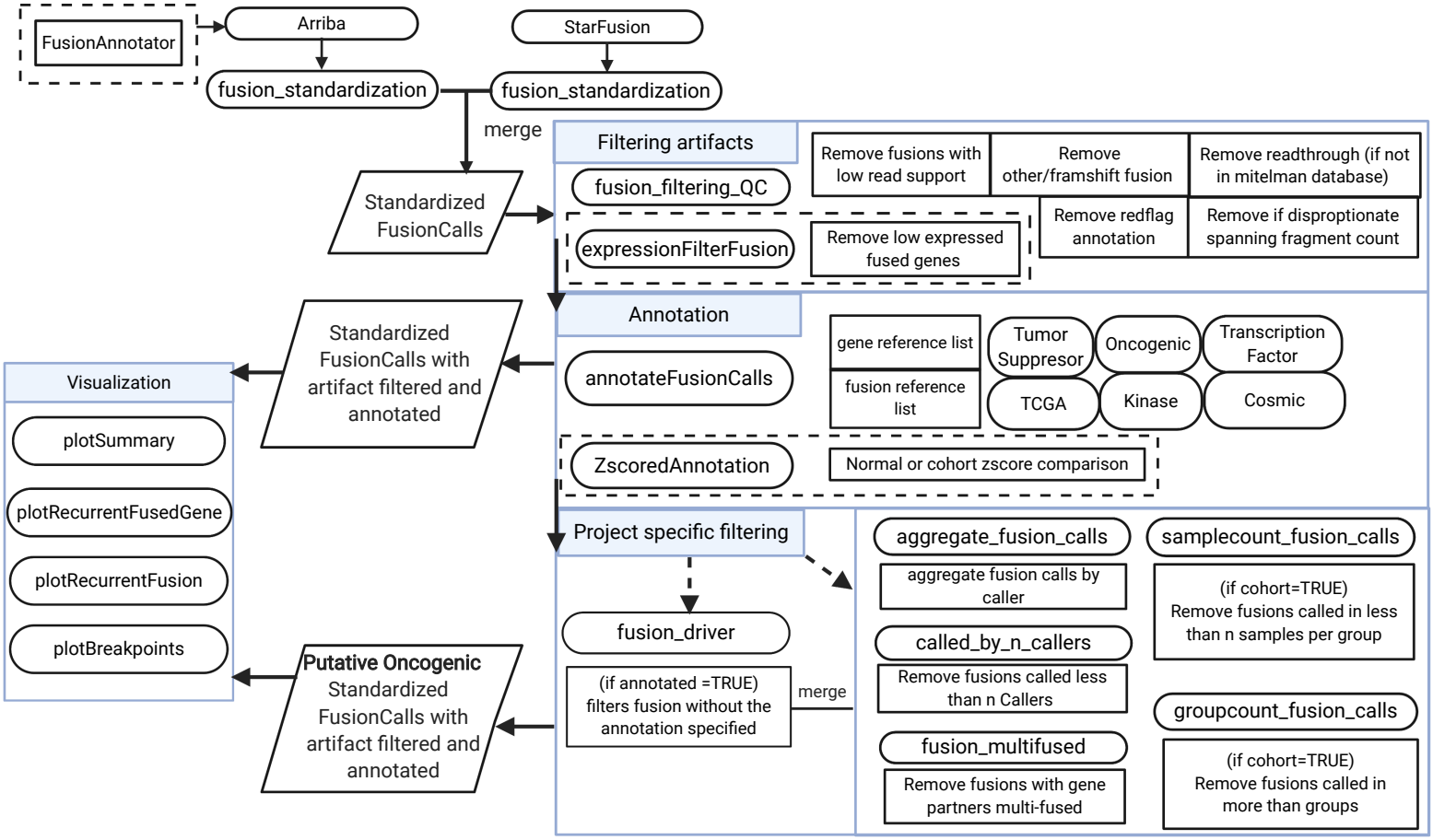

Figure 1: Graphical representation of pipeline

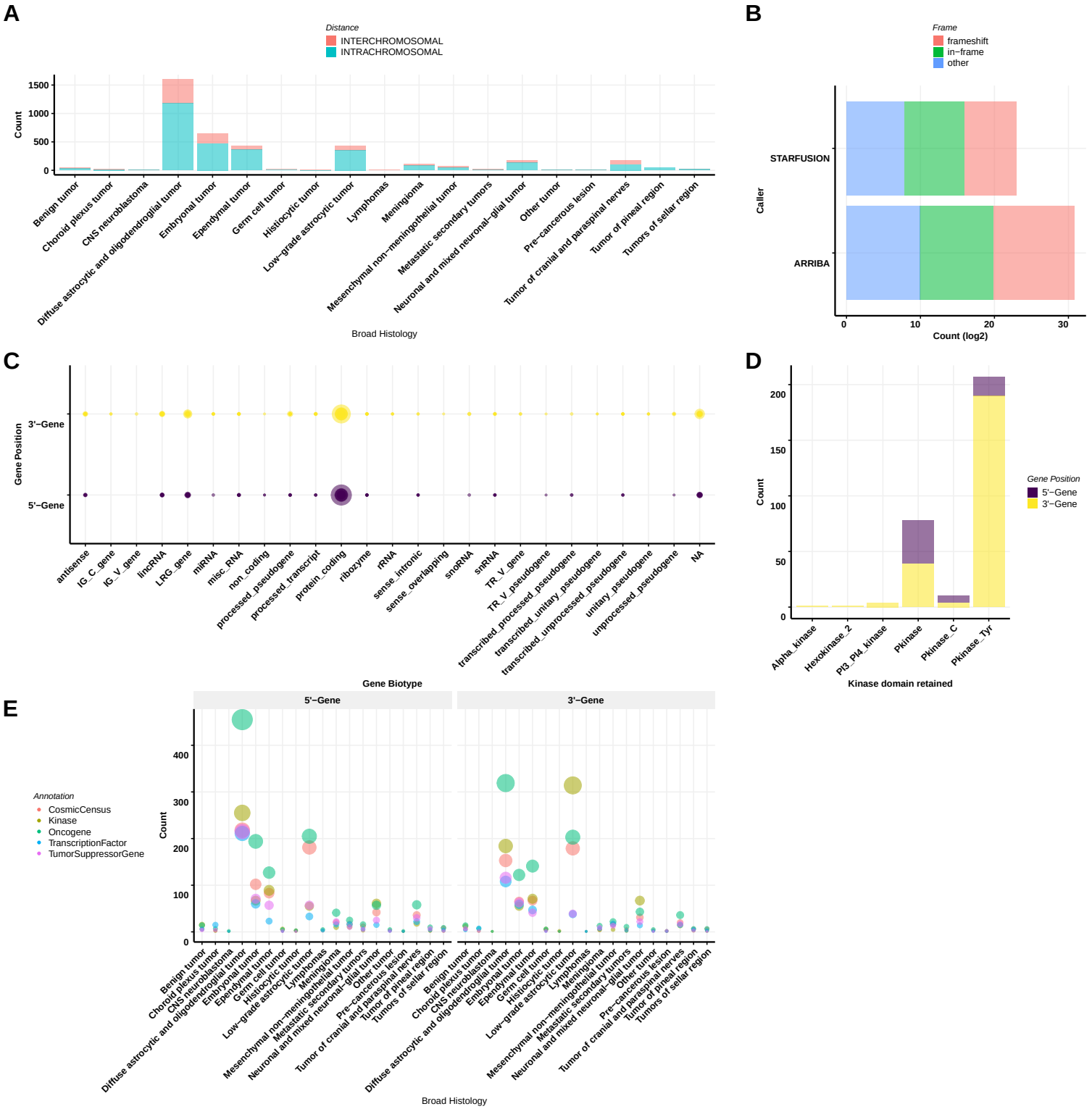

Figure 2: Fusion summary generated by annoFuse

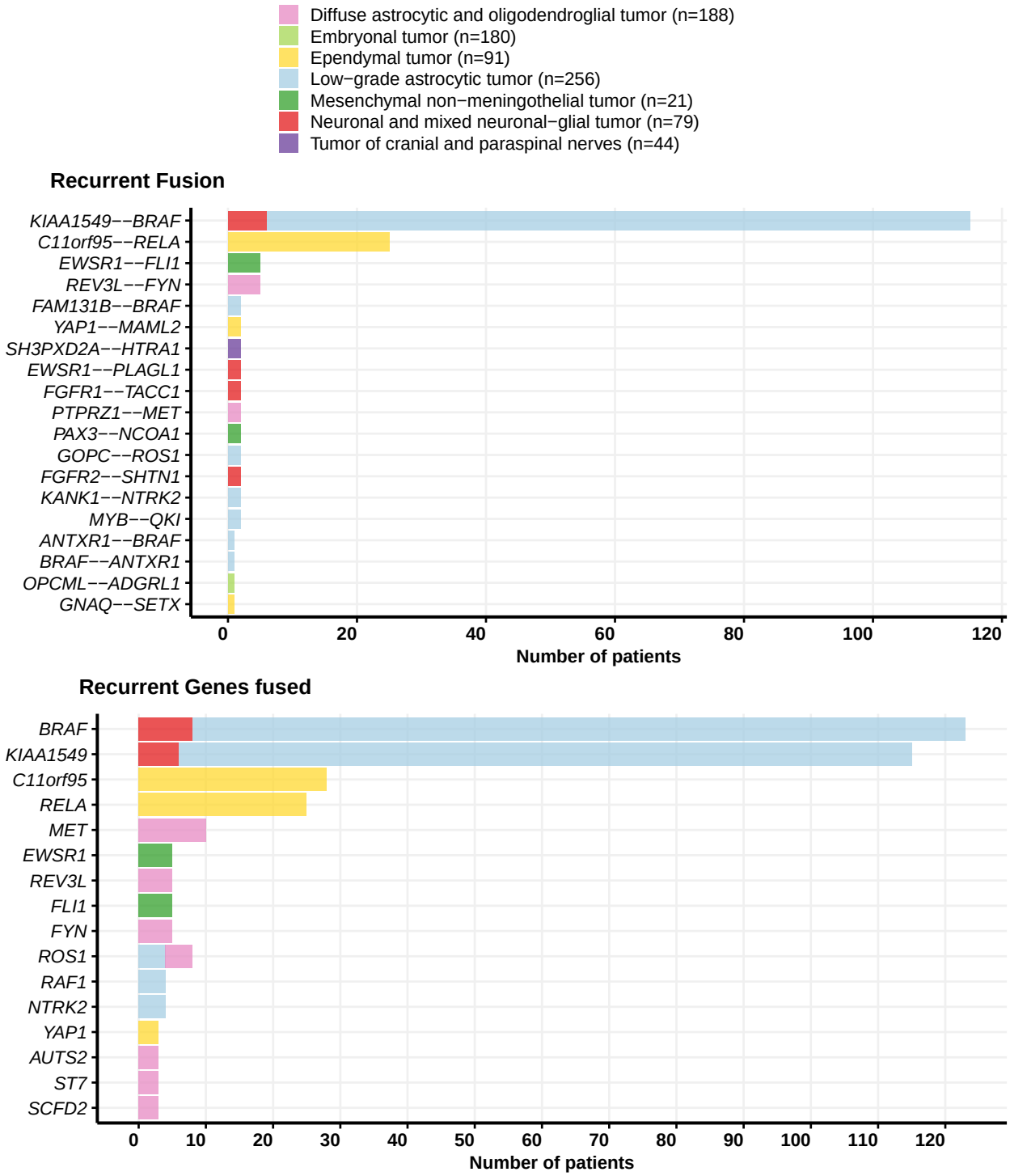

Figure 3: Recurrent fusion plots generated by annoFuse

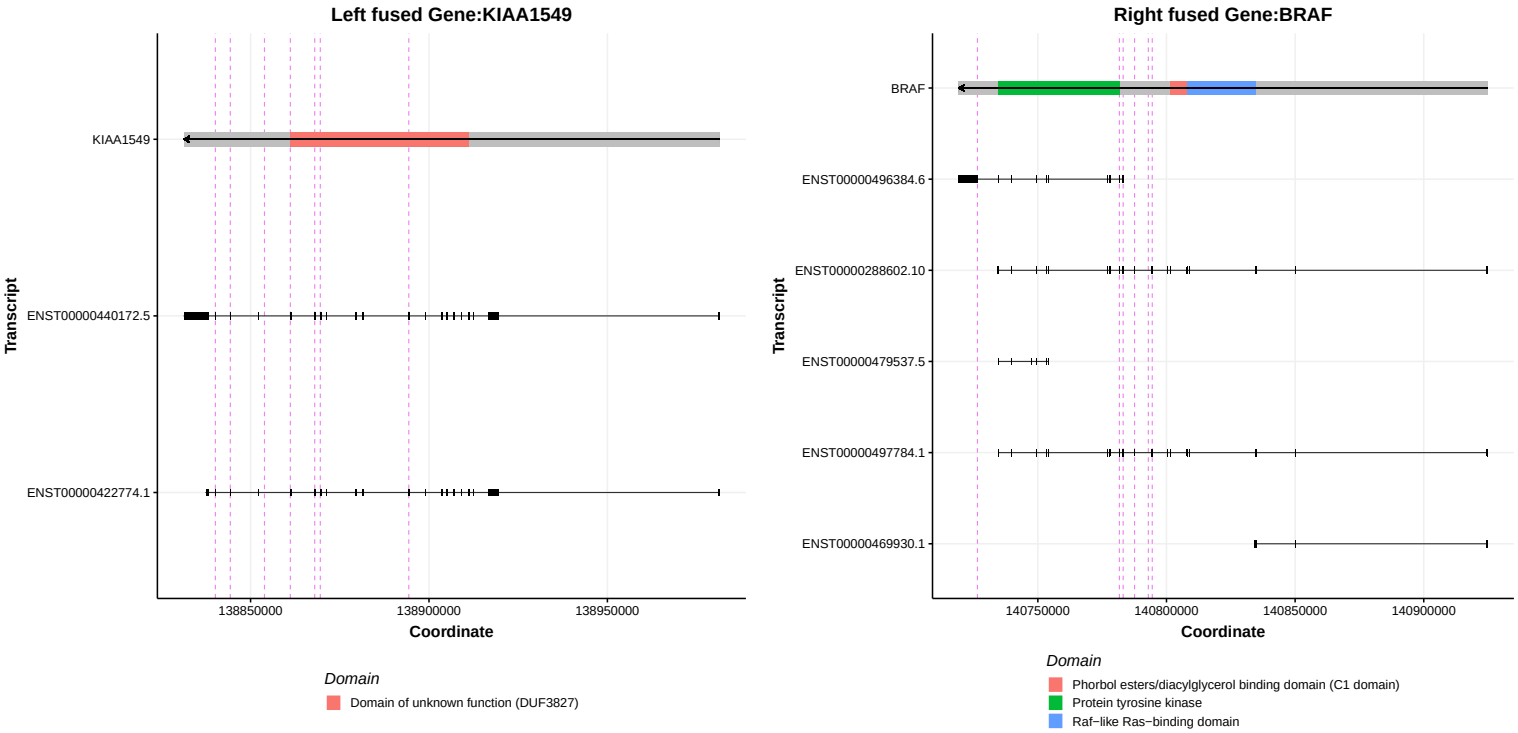

Figure 4: Breakpoint distribution for each gene in KIAA1549–BRAF fusion