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], Miguel A. Brown[1,4], Bo Zhang[1,4], Pichai Raman[1,2,4], Phillip B. Storm[4], John M. Maris[3], Adam C. Resnick[1,2,4], Jaclyn N. Taroni[5,*], Jo Lynne Rokita[1,2,*]

[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 1028 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 using support of at least one junction read and if there is disproportionate spanning fragment support of more than 10 reads compared to the junction read count, we removed them to remove false positives from background noise. Second, we prioritized and captured known, as well as putative oncogenic driver, fusions previously reported in TCGA, or fusions containing gene partners that are known oncogenes, tumor suppressor genes, or COSMIC genes. Finally, using *annoFuse,* we determined recurrent fusions across the cohort and recurrently-fused genes within each histology.

Conclusions

annoFuse provides a standardized filtering and annotation 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 dataset. We are expanding the package to be widely-applicable to other fusion algorithms, adding functionalities, 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. 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 (1), suggesting that many predictions may not represent true events. Additionally, transcriptional read-throughs (2), 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 (3–5). For example, the recurrent fusion *KIAA1549-BRAF* is found in between 66-80% of low grade gliomas and results in a fusion transcript that has constitutive BRAF kinase activity (6). *EWSR1-FLI1* is found in nearly 100% of Ewing sarcoma and forms an oncogenic RNA complex, driving tumorigenesis (7). Thus, the fusion databases, ChimerDB (8) and TumorFusions (9), have been developed utilizing RNA fusions called in The Cancer Genome Atlas (TCGA) (10,11) 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 (11,12). 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 such as STAR-Fusion (13) and Arriba (14) (**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 Open Pediatric Brain Tumor Atlas (OpenPBTA) Project (15). 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 cosmic census genes. Finally, we determined 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). R packages used to create *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 (16). 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 (17). *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.

RNA Expression and Fusion Calls

Currently, *annoFuse* is compatible with fusion calls generated from Arriba v1.1.0 (18) and/or STAR-Fusion 1.5.0 (13). 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 STARfusion uses FusionAnnotator as its final step and thus, require all fusion calls be annotated with FusionAnnotator v. 0.2.0 tol 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 *fusion_standardization* function to format caller specific output files to a standardizedFusionCalls format defined in the package README. *fusion_standardization* allows users to standardized fusion calls from multiple callers, users have the freedom to annotate their calls with other databases as 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* (19,20), 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 (18). Finally, if both genes of the fusion are deemed not expressed < 1 FPKM (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.

Step 4: 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, as these could represent novel fusions. 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.

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. This enables *annoFuse* to scavenge back potential oncogenic fusions which may have otherwise been filtered. Both sets of fusions are then merged into a final set of putative oncogenic fusions. Finally, *samplecount_fusion_call* identifies fusions recurrently called in (default ≥ 2) samples within each group.

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.

Results and Discussion

Technical validation of annoFuse

Few gene fusion “truth” sets exist and they are comprised of simulated data or synthetic fusions spiked into breast cancer cell lines or total RNA (16,17,21). We therefore utilized a recent study in which fusions were called and high-confidence fusions reported in 244 patient-derived xenograft models from the Pediatric Preclinical Testing Consortium (PPTC) (22). A set of 27 fusions were molecularly validated from acute lymphoblastic leukemia (ALL) models in the PPTC dataset and comprise of a “truth” set. **Table 3** describes the performance of *annoFuse*, in which we achieved 100% accuracy in calling true positive fusions and an average 96% accuracy of high-confidence fusions as defined in (22). 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 and 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 (23) and fusion calls generated by Arriba v1.1.0 (18) and/or STAR-Fusion 1.5.0 (13) which were released as part of the Pediatric Brain Tumor Atlas (24). 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 (25).

Following fusion standardization, annotation, and filtering, 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 (26).

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 from 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. Here, we note that in most histologies, the most prevalent gene partners were classified as oncogenes and the least prevalent as tumor suppressor genes. Notably, many 3' fusion partners within low-grade astrocytic tumors are kinases, which follows expectations listed below.

Following project-specific filtering, we observed *KIAA1549--BRAF* fusions as the most recurrent in-frame fusion in our cohort (n = 109/943), which was expected as KIAA1549-BRAF expressing low-grade astrocytic tumors comprise the largest representative histology in the OpenPBTA cohort (n = 504/943). *C11orf95--RELA* was predominant in ependymal tumors (n = 25/173), as expected in supratentorial ependymomas (27). Other expected recurrent oncogenic fusions obtained through *annoFuse* were *EWSR1-FLI1* in CNS Ewing sarcomas (28), and *KANK1-NTRK2, MYB-QKI,* and *FAM131B-BRAF* in low-grade astrocytic tumors (3,29) (**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**).

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 (30), Pegasus (31), chimera (32), and co-Fuse (33) 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 (34) 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 (35) is a web-based tool which enables annotation of fusions with 28 databases, however, is not programmatically scalable. FusionAnnotator (36) annotates fusions for presence in 15 cancer-associated databases, oncogene lists, and seven databases for fusions not relevant in cancer.

AGFusion (37) annotates protein domains, and Fusion Pathway (38) 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. 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), 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 acknowledge that protein domain annotation and retention is very important for prioritizing fusion calls and as such, we are working to add functionality from the algorithm-agnostic AGFusion tool in the near future. Likewise, we would like to integrate the recent FusionPathway tool, which is also algorithm agnostic, but depends on protein domain annotation to perform GSEA for oncogenic association. We 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 also include assessment of domain retention, combined with linkage 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 are 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 bioinformatically-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 multi-algorithm 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 (1,17) 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

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

TK : Tyrosine Kinase

TKL : Tyrosine Kinase-Like

WHO : World Health Organization

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All 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.

Competing interests

The authors declare no competing interests.

Funding

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

Authors' contributions

Conceptualization: KSG, JLR, JNT

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

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

Visualization: KSG, JLR

Supervision: JLR, JNT

Funding Acquisition: JLR, ACR, PJS, JMM

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. 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

2. 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

3. 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.

4. 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.

5. 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.

6. 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

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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

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

15. 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

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

17. 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.

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

19. 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.

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

21. 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.

22. 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.

23. 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.

24. 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

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

26. 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.

27. 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.

28. 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.

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

30. 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.

31. 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.

32. 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/

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

34. 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.

35. 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.

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

37. 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

38. 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. 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 STARFusion 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. (AGC = Protein Kinases A, G, and C; Atypical = kinases with no structural similarity to ePKs; CAMK = Calcium/Calmodulin Kinases; CK1 = Cell Kinase; CMGC = CDK, MAPK, GSK3, and CLK kinases; Other = unique kinases not belonging to any other group; STE = STE7, STE11, and STE20 genes which form the MAPK cascade; TK = Tyrosine Kinases; TKL = Tyrosine Kinase-Line (TKL) (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

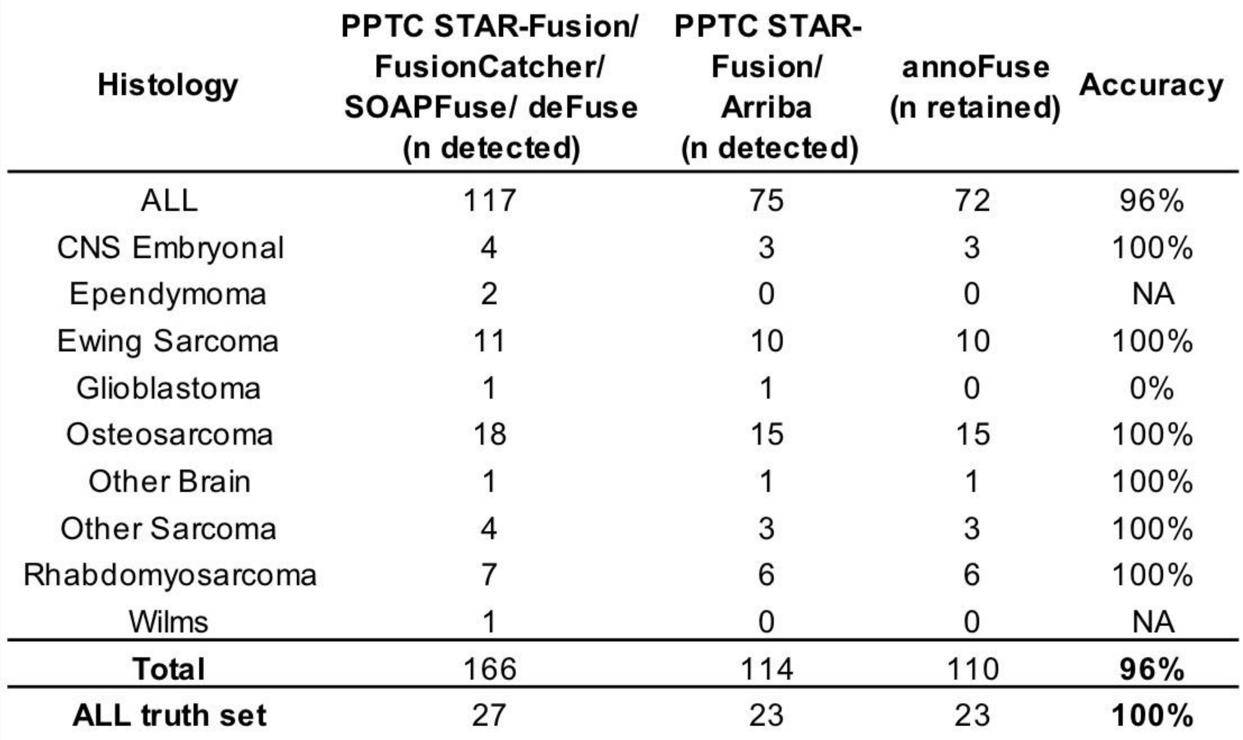
category are represented in each. In all panels except for B, fusion calls were merged from both STAR-Fusion and Arriba.

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.

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 the filtering 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. Overlap and accuracy of high-confidence fusion calls from PPTC PDX dataset using the STAR-Fusion/Arriba annoFuse workflow. Retention accuracy of high-confidence calls averaged 96% across the entire dataset and was 100% for the ALL truth set (ALL = acute lymphoblastic leukemia).



Histology	PPTC STAR-Fusion/ FusionCatcher/ SOAPFuse/ deFuse (n detected)	PPTC STAR-Fusion/ Arriba (n detected)	annoFuse (n retained)	Accuracy
ALL	117	75	72	96%
CNS Embryonal	4	3	3	100%
Ependymoma	2	0	0	NA
Ewing Sarcoma	11	10	10	100%
Glioblastoma	1	1	0	0%
Osteosarcoma	18	15	15	100%
Other Brain	1	1	1	100%
Other Sarcoma	4	3	3	100%
Rhabdomyosarcoma	7	6	6	100%
Wilms	1	0	0	NA
Total	166	114	110	**96%**
ALL truth set	27	23	23	**100%**

Figure 1

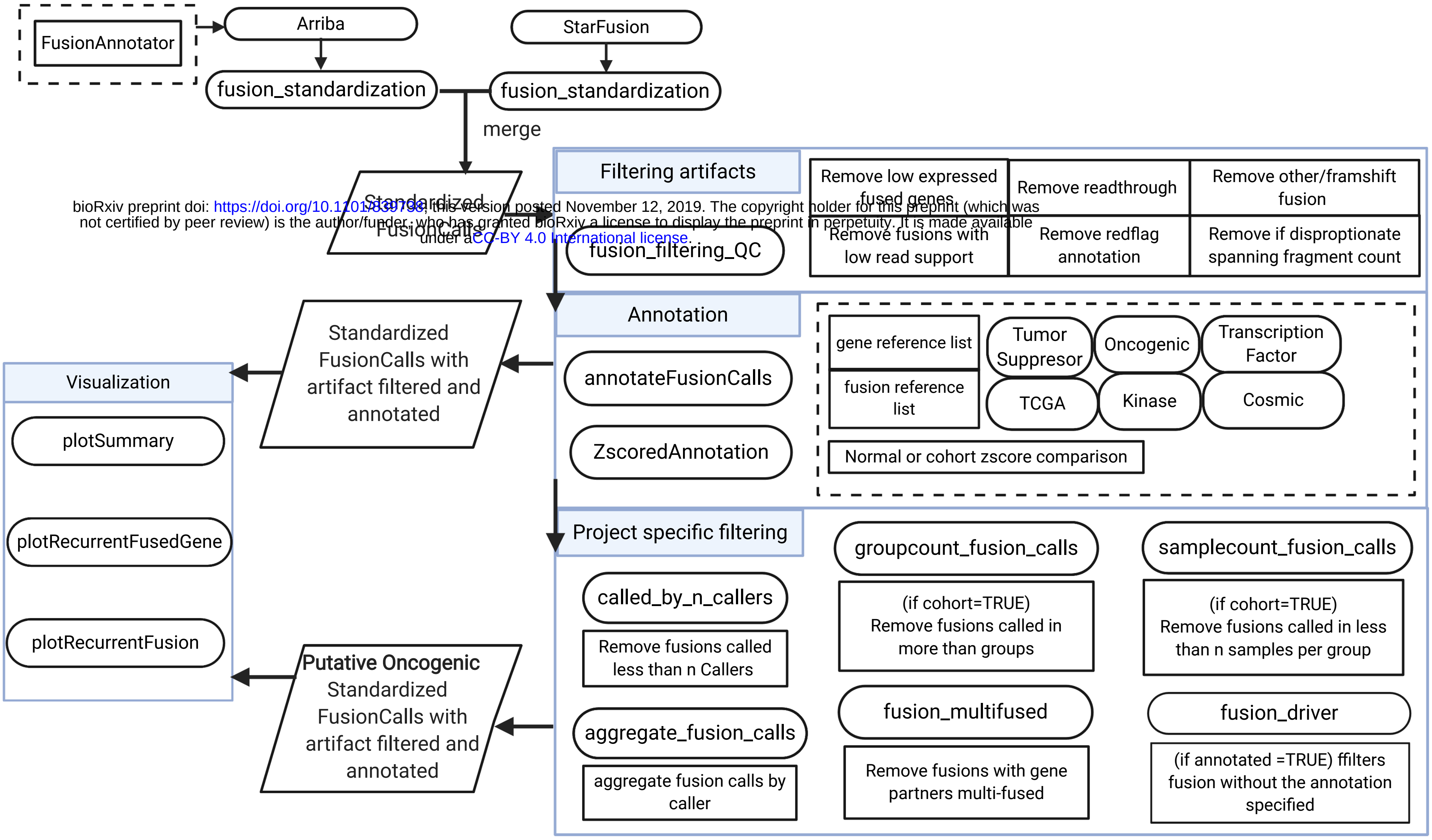

FusionAnnotator
Arriba
StarFusion
fusion_standardization
fusion_standardization
merge
Standardized FusionCalls
Filtering artifacts
fusion_filtering_QC
Remove low expressed fused genes
Remove readthrough
Remove other/framshift fusion
Remove fusions with low read support
Remove redflag annotation
Remove if disproptionate spanning fragment count
Annotation
annotateFusionCalls
ZscoredAnnotation
gene reference list
fusion reference list
Tumor Suppresor
Oncogenic
Transcription Factor
TCGA
Kinase
Cosmic
Normal or cohort zscore comparison
Standardized FusionCalls with artifact filtered and annotated
Visualization
plotSummary
plotRecurrentFusedGene
plotRecurrentFusion
Project specific filtering
called_by_n_callers
Remove fusions called less than n Callers
aggregate_fusion_calls
aggregate fusion calls by caller
groupcount_fusion_calls
(if cohort=TRUE) Remove fusions called in more than groups
fusion_multifused
Remove fusions with gene partners multi-fused
samplecount_fusion_calls
(if cohort=TRUE) Remove fusions called in less than n samples per group
fusion_driver
(if annotated =TRUE) ffilters fusion without the annotation specified
Putative Oncogenic Standardized FusionCalls with artifact filtered and annotated

Figure 2

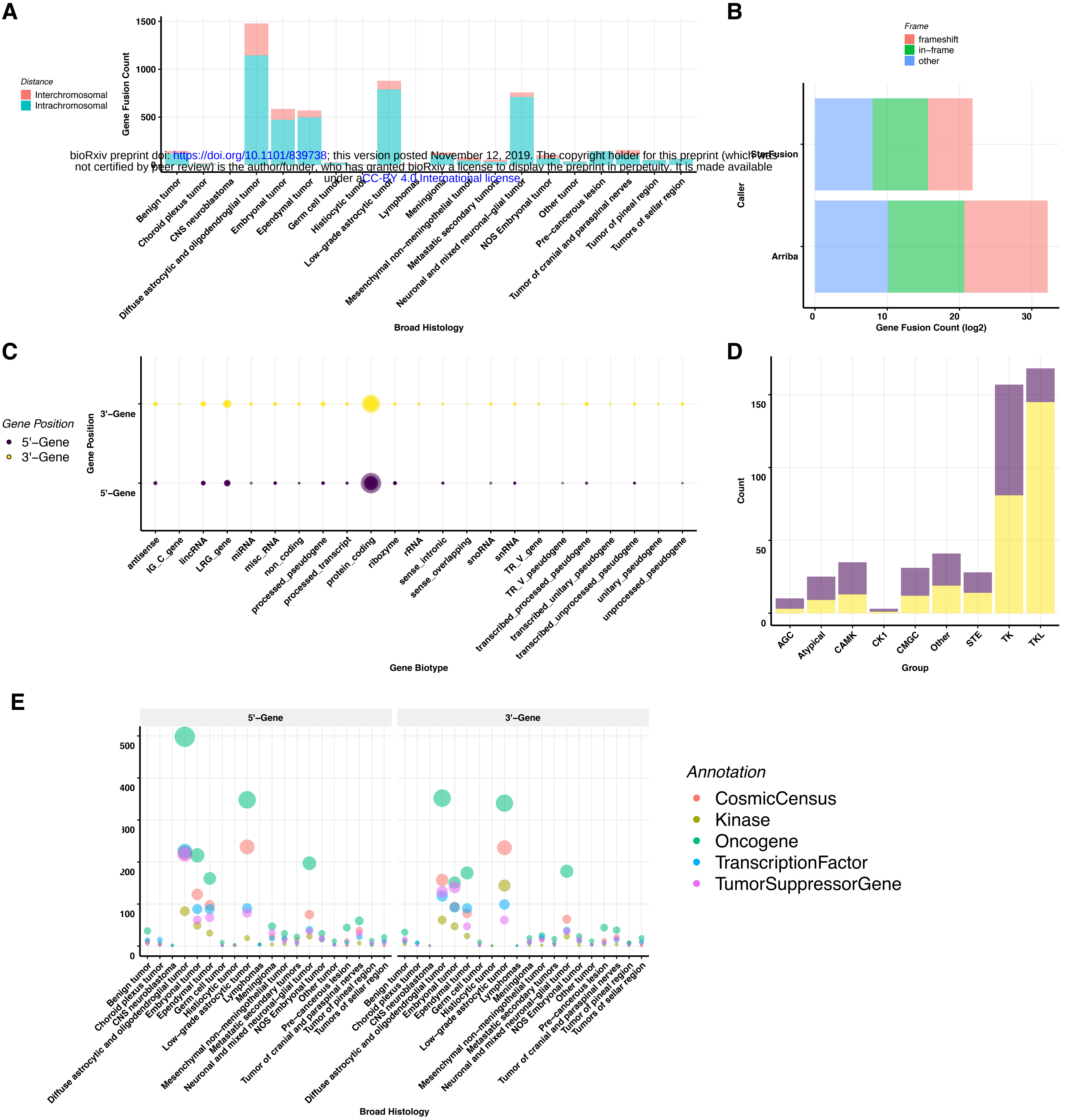
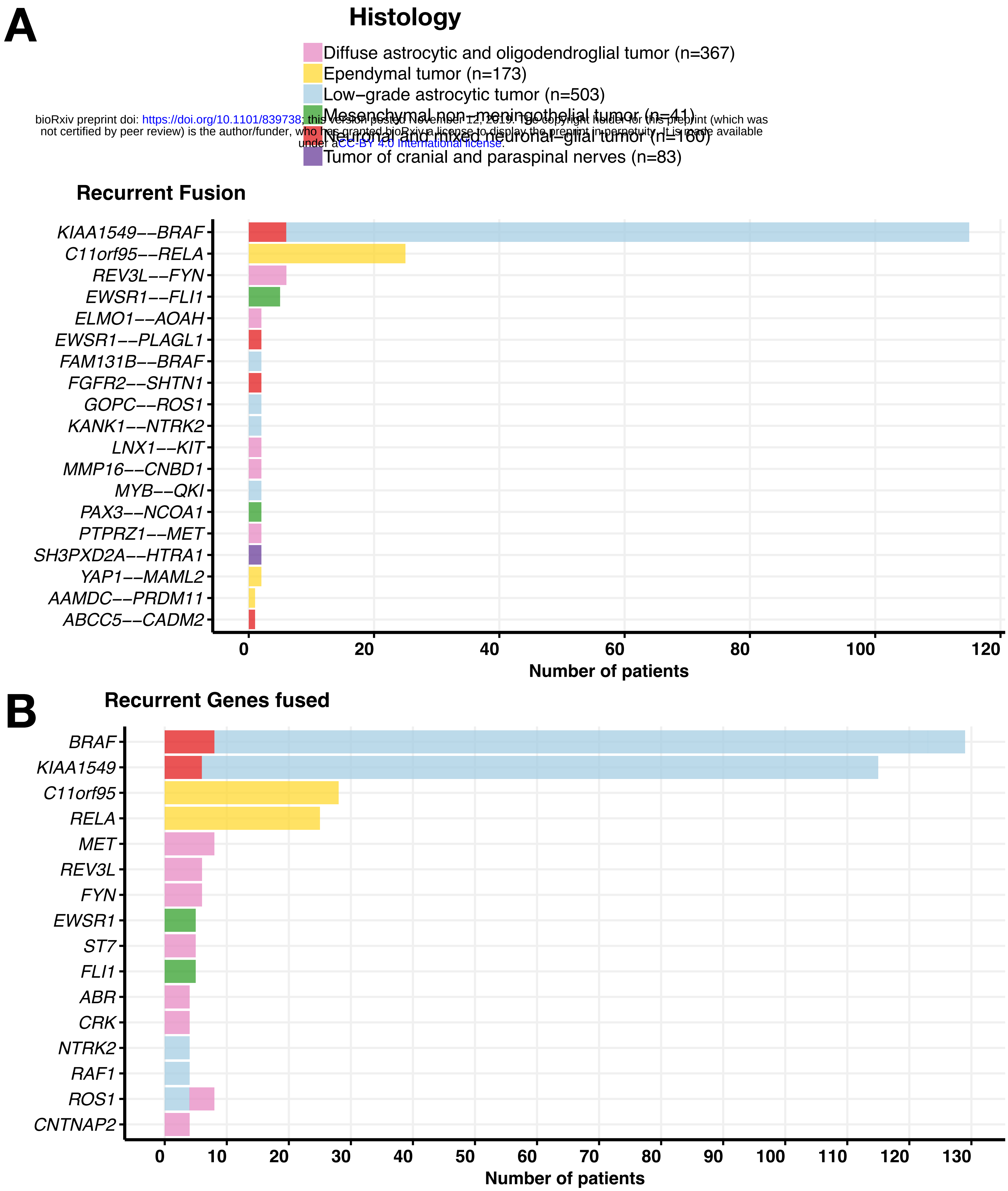

A
Distance
Interchromosomal
Intrachromosomal
Gene Fusion Count
1500
1000
500
0
Benign tumor
Choroid plexus tumor
CNS neuroblastoma
Diffuse astrocytic and oligodendroglial tumor
Embryonal tumor
Ependymal tumor
Germ cell tumor
Histiocytic tumor
Low−grade astrocytic tumor
Lymphomas
Meningioma
Mesenchymal non−meningothelial tumor
Metastatic secondary tumors
Neuronal and mixed neuronal−glial tumor
NOS Embryonal tumor
Other tumor
Pre−cancerous lesion
Tumor of cranial and paraspinal nerves
Tumor of pineal region
Tumors of sellar region
Broad Histology
B
Frame
frameshift
in−frame
other
Caller
StarFusion
Arriba
0
10
20
30
Gene Fusion Count (log2)
C
Gene Position
5'−Gene
3'−Gene
antisense
IG_C_gene
lincRNA
LRG_gene
miRNA
misc_RNA
non_coding
processed_pseudogene
processed_transcript
protein_coding
ribozyme
rRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene
Gene Biotype
D
Count
150
100
50
0
AGC
Atypical
CAMK
CK1
CMGC
Other
STE
TK
TKL
Group
E
5'−Gene
3'−Gene
500
400
300
200
100
0
Broad Histology
Annotation
CosmicCensus
Kinase
Oncogene
TranscriptionFactor
TumorSuppressorGene

Figure 3


A
Histology
Diffuse astrocytic and oligodendroglial tumor (n=367)
Ependymal tumor (n=173)
Low−grade astrocytic tumor (n=503)
Mesenchymal non−meningothelial tumor (n=41)
Neuronal and mixed neuronal−glial tumor (n=160)
Tumor of cranial and paraspinal nerves (n=83)
Recurrent Fusion
KIAA1549−−BRAF
C11orf95−−RELA
REV3L−−FYN
EWSR1−−FLI1
ELMO1−−AOAH
EWSR1−−PLAGL1
FAM131B−−BRAF
FGFR2−−SHTN1
GOPC−−ROS1
KANK1−−NTRK2
LNX1−−KIT
MMP16−−CNBD1
MYB−−QKI
PAX3−−NCOA1
PTPRZ1−−MET
SH3PXD2A−−HTRA1
YAP1−−MAML2
AAMDC−−PRDM11
ABCC5−−CADM2
0
20
40
60
80
100
120
Number of patients
B
Recurrent Genes fused
BRAF
KIAA1549
C11orf95
RELA
MET
REV3L
FYN
EWSR1
ST7
FLI1
ABR
CRK
NTRK2
RAF1
ROS1
CNTNAP2
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Number of patients