Targeted *in silico* characterization of fusion transcripts in tumor and normal tissues via FusionInspector

Brian J. Haas[1-3,*], Alexander Dobin[4], Mahmoud Ghandi[5], Anne Van Arsdale[6,7], Timothy Tickle[1,3], James T. Robinson[8], Riaz Gillani[9-12], Simon Kasif[2,13,] and Aviv Regev[1,14-16]

Affiliations

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2]Graduate Program in Bioinformatics, Boston University, Boston, MA, 02215, USA

[3]Present address: Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Cold Spring Harbor Laboratory, New York, 11724

[5]Monte Rosa Therapeutics, Boston, MA

[6]Department of Obstetrics and Gynecology and Women's Health, Albert Einstein Montefiore Medical Center, Bronx, NY

[7]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY

[8]School of Medicine, University of California San Diego, La Jolla, California

[9]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[10]Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[11]Department of Pediatrics, Harvard Medical School, Boston, MA 02215, USA

[12]Boston Children's Hospital, Boston, MA 02115, USA

[13]Department of Biomedical Engineering, Boston University, Boston, MA, 02215, USA

[14]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.

[15]Howard Hughes Medical Institute, Chevy Chase, MD, USA.

[16]Present address: Genentech, South San Francisco, CA, USA.

* corresponding author

Abstract

Background

Gene fusions play a key role as driving oncogenes in tumors, and their reliable discovery and detection is important for cancer research, diagnostics, prognostics and guiding personalized therapy. While discovering gene fusions from genome sequencing can be laborious and costly, the resulting "fusion transcripts" can be recovered from RNA-seq data of tumor and normal samples. However, alleged and putative fusion transcript can arise from multiple sources in addition to the chromosomal rearrangements yielding fusion genes, including *cis*- or *trans*-splicing events, experimental artifacts during RNA-seq or computational errors of transcriptome reconstruction methods. Understanding how to discern, interpret, categorize, and verify predicted fusion transcripts is essential for consideration in clinical settings and prioritization for further research. Here, we present FusionInspector for *in silico* characterization and interpretation of candidate fusion transcripts from RNA-seq, enabling exploration of sequence and expression characteristics of fusions and their partner genes.

Results

We applied FusionInspector to thousands of tumor and normal transcriptomes, and identified statistical and experimental features enriched among biologically impactful fusions. Through clustering and machine learning, we identified large collections of fusions potentially relevant to tumor and normal biological processes. We show that biologically relevant fusions are enriched for relatively high expression of the fusion transcript, imbalanced fusion allelic ratios, and canonical splicing patterns, and are deficient in sequence microhomologies detected between partner genes.

Conclusion

We demonstrate FusionInspector to accurately *in silico* validate fusion transcripts, and to help identify numerous understudied fusions in tumor and normal tissues samples. FusionInspector is freely available as open source for screening, characterization, and visualization of candidate fusions via RNA-seq. We believe that this work will continue driving the discipline of transparent explanation and interpretation of machine learning predictions and tracing the predictions to their experimental sources.

Keywords

RNA-seq, Fusion, Cancer, FusionInspector, STAR-Fusion, Trinity

Background

Gene fusions are intensely studied for their relevance to disease and normal cellular biology. In cancer, gene fusions typically result from chromosomal rearrangements, including well-known drivers of cancer, such as BCR--ABL1 in chronic myelogenous leukemia (CML) (1, 2), TMPRSS2--ERG in prostate cancer (3, 4), and SS18--SSX1 or SS18--SSX2 in synovial sarcoma (5, 6). Charting the diversity of fusion transcripts present in tumor and normal tissue is important for our basic understanding of the complexity and biological function of the transcriptome in normal and disease states, molecular diagnostics of cancer patients, and neoepitope discovery for targeting in personalized immunotherapy with cancer vaccines or T cell therapy (7, 8).

The structural rearrangements leading to gene fusions can be detected or inferred through whole genome sequencing or from the presence of "fusion transcripts" in whole transcriptome sequencing (RNA-seq) (9, 10). Given the easier and economical nature of RNA-seq compared to whole genome sequencing, and the effective methods for transcript assembly, RNA-seq has emerged as a leading experimental method for fusion transcript discovery and detection in both cancer research and molecular diagnostics. Fusions detected at the RNA level may also be causally responsible for functional and phenotypical changes. Dozens of computational tools have been developed to mine fusion transcripts from RNA-seq data (as referenced in (11)), and there have been multiple efforts to build catalogs of fusions across tumor and normal tissues (12-17). In general, tumor-specific fusion transcripts are presumed to derive from chromosomal rearrangements as commonly encountered in tumor samples, whereas fusions identified in normal samples are considered more likely to be derived from normal karyotypes, thus reflecting other underlying causes, such as read-through transcription and *cis*- or *trans*-spliced products.

However, predicting fusions from RNA-seq data is challenging and the various methods developed to predict fusion products from RNA-seq vary tremendously in their accuracy for fusion detection, leading to both false positives and false negatives (11, 18, 19). False positives can be driven by experimental artifacts that arise during reverse transcription or PCR amplification, and by computational mis-mapping of reads to target gene sequences (20), as well as specific differences in prediction tools. Moreover, as sequencing depth increases, the probability of detecting rare reads that support a fusion transcript prediction increases. This may be due to either lab artifacts or to real, low-rate trans-splicing of little functional relevance. Thus, there is an urgent need to understand the features that drive fusion detection and to generate high quality catalogs of well supported fusions.

Here, we describe FusionInspector, a method to assess and document the evidence for fusions. We subsequently demonstrate its applicability to understand the features of reliable fusion transcript predictions and to generate a large catalog of predicted fusions across tumor and healthy tissues. FusionInspector reassesses the read alignment evidence supporting pre-specified candidate fusion transcripts, compares the relative alignment evidence for a fusion transcript *vs.* its unfused partner transcripts. FusionInspector further evaluates the fusion transcript breakpoints in relation to sequence features considered representative of experimental and bioinformatic artifacts (21, 22), including canonical splicing sequences, reference exon gene structures, and regions of microhomology between partner genes. We integrate FusionInspector with our fusion prediction tool STAR-Fusion (11), where it can be used either for *in silico* validation of STAR-Fusion predictions, or as a standalone utility for evaluating other fusion predictions or screening a user-defined panel of fusions of interest. We show that FusionInspector's evaluation increases transparency and overall fusion prediction accuracy of an individual method or across an ensemble of methods. Finally, we apply FusionInspector to examine recurrent fusions predicted among thousands of tumor and normal samples and identify new groups of fusions potentially relevant to tumor and normal tissue biology.

Results

Development of FusionInspector for *in silico* evaluation of predicted fusion transcripts

FusionInspector (**Figure 1**) performs a supervised *in silico* evaluation of a specified set of candidate fusion transcripts, either predicted by STAR-fusion or another method from RNA-seq data, or a user-defined panel. FusionInspector captures all read alignments in the RNA-seq that provide evidence for the specified fusions or for the unfused partner genes, and further explores the candidate fusion genes for regions of microhomology (defined as short identical sequence matches of length k (here *k*=10)), and the proximity of microhomologies to putative fusion breakpoints.

To capture evidence supporting candidate fusions, FusionInspector identifies those reads that align concordantly between fusion genes as juxtaposed in their fused orientation and provide concordant alignments that span the two genes in this rearranged context. There are two types of fusion-supporting alignments: (1) split-reads that define the fusion breakpoint, and (2) spanning fragments, where each paired-end read aligns to an opposite partner gene and the fragment bridges the fusion breakpoint (**Figure 1**). FusionInspector leverages STAR aligner (23), which we enhanced here to support FusionInspector's mode of action. As input to STAR, we provide the entire reference genome along with a set of fusion contigs constructed by FusionInspector (based on the list of specified fusion candidates), and STAR aligns reads to the combined genome targets and reports those aligned to the fusion contigs for further evaluation by FusionInspector (**Methods**).

Next, FusionInspector computes a number of features associated with the fusion based on these alignments. First, it uses the number of reads exclusively supporting each fusion as a proxy for the expression of the fusion transcript (similarly, read alignments overlapping the fusion breakpoint and exclusively supporting the unfused partner genes are a proxy for the expression levels of the unfused partner genes). Second, it computes the fusion allelic ratio (FAR) for the fusion with respect to each (5’ or 3’) partner transcript (5’-FAR and 3’-FAR) as the ratio of mutually exclusive reads supporting the fusion *vs*. each unfused partner gene (**Figures 1, S1)**. Third, it examines fusion breakpoints inferred from the read alignments for canonical dinucleotide splice sites at boundaries of the breakpoints in each partner gene and for agreement with available reference gene structure annotations. When there is evidence that supports multiple fusion transcript isoforms for a given fusion gene, FusionInspector uses an expectation maximization (EM)-based algorithm to fractionally assign mutually compatible spanning fragments to the corresponding isoforms (**Methods**). It then filters fusion candidates according to minimum evidence requirements (at least one split read to define the junction breakpoint, and at least 25 aligned bases supported by at least one read on both sides of the fusion breakpoint; **Methods**). Finally, it captures microhomologies between putative fusion genes and determines the proximity of a fusion breakpoint to the nearest site of microhomology. Using these sequence and expression attributes of fusions and known characteristics of biologically relevant fusions (below), FusionInspector further predicts whether each *in silico* validated fusion candidate is likely to be biologically relevant or alternatively has features consistent with experimental or bioinformatic fusion artifacts.

We illustrate these features in the context of two contrasting examples of fusion types (**Figure 2**). Fusion EML4--ALK, a known cancer driver prevalent in lung adenocarcinoma (24, 25), has evidence of multiple transcript isoform structures, and while microhomologies are found between the EML4 and ALK genes, they tend to be distal from the fusion isoform breakpoints (**Figure 2a**). The EML4--ALK fusion breakpoints are all found at consensus dinucleotide splice sites that coincide with exon boundaries of reference gene structure annotations. In contrast, FusionInspector captures many reads supporting a putative fusion KRT13--KRT4, but the breakpoints inferred from split read alignments mostly have non-consensus dinucleotide splice sites and coincide with sites of microhomology, additionally, split reads with consensus dinucleotide splice sites mostly do not coincide with reference exon boundaries (**Figure 2b**). Because KRT13 and KRT4 are only distantly related with no easily detected nucleotide-level sequence conservation, their fusion may not be discarded by fusion transcript predictors. However, given that most fusion evidence coincided with sites of microhomology and the lack of consensus splicing at breakpoints, we infer that the putative KRT13--KRT4 fusion is artifactual. Another particularly compelling example of a similarly misleading and likely artifactual fusion is COL1A1--FN1, which is detected as prevalent in cancer-associated fibroblast cell lines (**Figure S2**). Further consideration of fusion and partner gene expression levels can aid in evaluating and prioritizing fusion candidates for further study, as we pursue below.

In silico benchmarking of FusionInspector validation accuracy

We benchmarked FusionInspector for fusion detection and analysis across 60 cancer cell line transcriptomes, using our previously established benchmarking framework (11). We assessed FusionInspector in two modes: using STAR-Fusion (current v1.9.1) predictions as the exclusive targets, or using the union of 24 different prediction methods (**Table S1**).

Using FusionInspector following STAR-Fusion consistently yielded better performance compared to our earlier evaluated prediction methods (**Figure S3**). While FusionInspector in this execution mode cannot find additional fusions not initially predicted by STAR-Fusion, it increased its positive predictive value (**Figure S3a**) by eliminating fusion predictions based on alternative assessments of read support, while largely retaining sensitivity.

Applying FusionInspector in fusion screening mode by providing it with all fusions predicted by any of 24 different methods (as in Table S4 of (11), **Methods**), FusionInspector had high accuracy and was among the top performing methods (**Figure S3b**). This allowed FusionInspector to also capture likely true fusions not captured by STAR-Fusion but predicted by other methods (**Table S1**).

A high-quality catalog of recurrent fusion transcripts from cancer and healthy tissue using FusionInspector

We next used FusionInspector to create a vetted catalog of recurrent fusion transcripts based on RNA-Seq from tumors (from TCGA (26)) and matched healthy tissue (from TCGA and GTEx (27)). We first predicted fusion transcripts with STAR-Fusion (v1.7) across 9,426 tumor and 707 normal samples from TCGA, and 8,375 normal samples from GTEx (**Table S2**). We initially applied lenient fusion evidence requirements to maximize sensitivity (**Methods**). As a result, putative fusion transcripts were detected in nearly all tumor and normal samples. After applying a minimum expression level threshold (0.1 FFPM), we detected a significantly higher number of fusions in tumors *vs*. paired normal samples in several TCGA tumor types (**Figure S4a**), although there were similar median numbers of predicted fusions per sample type in TCGA tumor and GTEx normal samples (t-test, p=0.5, **Figure S4b**). We readily identified known cancer fusions included in the COSMIC fusion collection (28, 29) ("COSMIC-fusions", **Figure**

3a) according to known disease associations and prevalence, such as TMPRSS2--ERG identified in roughly half of prostate cancers (3), FGFR3--TACC3 in glioblastoma (30), and PML--RARA in the acute promyelocytic leukemia subtype of acute myeloid leukemia (31). COSMIC fusions were more highly expressed than most predicted fusions, which had low estimated expression levels and few supporting reads (**Figure 3b-d**).

Next, we used FusionInspector to examine the sequence and expression features of fusion transcripts that were recurrently detected across tumor and/or normal samples, in order to distinguish biologically impactful fusions (akin to the COSMIC fusions) from experimental or computational artifacts, or from low levels of *cis*- or *trans*-splicing from highly expressed genes. To this end, we analyzed 53,240 fusion isoforms (38,591 fusion occurrences and 14,649 alternatively spliced fusion isoforms) from 628 TCGA and 530 GTEx representative samples (**Methods**). For each fusion candidate, FusionInspector identified the number of reads supporting the fusion and those supporting the unfused partner genes at putative breakpoints, identified regions of microhomology between partner genes, and determined the following features: inferred fusion expression level (FFPM), 5’ and 3’ fusion allelic ratios (5’-FAR, 3’-FAR), 5’ and 3’ unfused gene expression levels (5’-counter-FFPM and 3’-counter-FFPM), presence of consensus *vs*. non-consensus dinucleotide splice sites at fusion breakpoints, agreement or disagreement with reference gene structure exon boundaries at splice junctions, number of microhomologies observed between the two partner genes, and the distance of each inferred fusion breakpoint to the nearest site of microhomology.

To distinguish fusion artifacts from those with features consistent with biologically impactful fusions, we clustered fusions by their feature profiles (**Figure 4a, Table S3, Methods**). The clustering procedure produced 61 high granularity clusters, which we further grouped by hierarchical clustering according to median fusion attribute values in each fine cluster (**Figure 4b**). We then focused on examining clusters enriched for COSMIC fusions as a proxy for biologically impactful fusions.

One fusion cluster (C4) was significantly enriched with COSMIC fusions, harboring 57% of our detected instances of COSMIC fusions among these samples, but only 4% of all called fusions ($p < 10^{-90}$, Fisher's Exact one-sided test) (**Figure 4b**). Fusions in this cluster had splice breakpoints consistent with consensus splice sites and matching known reference gene structure exon boundaries, were relatively highly expressed, and were deficient in microhomologies between fusion partner genes. Most of the COSMIC fusions in C4 also have a 3'-FAR that exceeds the 5'-FAR, consistent with the fusion transcript being driven from an active 5' partner's promoter and a 3' unfused partner expressed at lower levels (**Figure S5**). Sixteen additional clusters, all but one (C10) of which are members of one large hierarchical cluster with related features, had at least two COSMIC fusions per cluster and spanned 34% of the fusions overall.

Conversely, other fusion clusters had features indicative of experimental or computational artifacts, especially enrichment in microhomologies that could confound alignment or contribute to RT mis-priming, and had no COSMIC fusions. We thus consider those fusions as putative artifacts. These clusters encompassed 3% of all fusion occurrences (restricted to the highest expressed fusion isoform per occurrence): 2/3 (2% of all fusions) had moderately-to-highly expressed partner genes, suggesting origination from RT mis-priming, and a 1/3 with little evidence for partner gene expression, suggesting read misalignments artifacts. The low portion of such presumed artifacts is a testament to STAR-Fusion's rigorous filtering (11). Finally, another 1% of fusions involve highly expressed partner genes, where the detected fusion represented a small fraction of the total expression from these loci. These fusions may result from low levels of *cis*- or *trans*-splicing from the highly expressed partner genes.

Targeted screening for novel COSMIC-like fusions by a classifier

We reasoned that the set of 1,511 predicted fusion occurrences (835 distinct gene pairings) that were members of the COSMIC enriched cluster C4, are likely enriched for fusions of functional significance and should be prioritized for further study. Some are already known to be relevant to cancer but not yet included in the COSMIC database, such as EGFR--SEPT14 (32), PVT1--MYC (33-35), and TPM3--NTRK1 (36). Others are reciprocal fusions for COSMIC fusions that could result from balanced translocations, including reciprocal ABL1--BCR1 of COSMIC BCR1--ABL1, BRAF--SND1 of COSMIC SND1--BRAF, and PPARG--PAX8 of COSMIC PAX8--PPARG. This fusion cluster is also enriched for fusions exclusively identified in pancreatic tissue (explored below).

To gain further insights into the characteristics of the COSMIC-like fusions in C4, we screened additional TCGA and GTEx samples to characterize additional occurrences of C4 representative fusions. We refocused FusionInspector on 236 key fusions (231 C4 recurrent fusion gene pairs including 26 COSMIC fusions, and another five COSMIC fusions not included in C4, **Methods,**

Table S4 (37)). We screened each of 2,764 TCGA and 1,009 GTEx representative samples for these 236 fusions (**Methods**), using FusionInspector's screening modality, collecting FusionInspector validations and attributes for 37,211 additional fusion occurrences (**Figure 5**, **Figures S6a-e**, **Table S5**), and ranked fusions by the difference in their initial STAR-Fusion detected prevalence in tumors *vs*. normal samples. Finally, to determine whether additional occurrences of these fusions have characteristics consistent with C4 fusions, general COSMIC-like fusions, are artifact-like, or belong to another category, we trained a random forest classifier using examples of fusions from the initial analysis to predict the labels of each of 61 Leiden clusters and applied it to predict the cluster labels of fusion isoforms examined in this expanded targeted survey. Finally, we categorized each fusion occurrence by the overall category (*e.g.*, "COSMIC-like") of the hierarchical cluster to which the Leiden cluster label it was classified to belongs (**Figure 5a**, **Methods**). In this way, we could distinguish individual fusion isoforms by their characteristics in the context of all the analyzed fusions.

FusionInspector-screened occurrences of known COSMIC fusions were mostly tumor-enriched with few to no normal samples identified with evidence (noting all 1,009 GTEx normal samples were screened by FusionInspector for an identical list of the COSMIC fusions). For most (27/31) known COSMIC fusions, at least 80% of fusion occurrences were classified as COSMIC-like, with the remaining four having at least half of occurrences classified as COSMIC-like, and none were classified as artifacts (**Figure 5a, Table 6S**). All but 31 of the 236 fusions had instances classified as C4, and all but 39 had at least 50% of their occurrences classified as COSMIC-like. Only 7 fusions had at least 10% of their occurrences classified as having high counter-evidence

(C49 or C51), and only 9 fusions had any occurrences predicted as artifacts, with both categories enriched for pancreas-specific fusions (further discussed below).

This analysis highlighted intriguing, well supported fusions for further study. For example, while the top-ranked tumor-enriched fusion, FGFR3--TACC3 (rank 1, 70 tumor samples, 0 normal) is a known oncogenic driver (30), other top ranking fusions, such as CCAT1--CASC8 (rank 2, 42 tumors - mostly lung and stomach cancers, 0 normal) and VCL--ADK (rank 3, 36 tumors - also mostly lung and stomach cancers, 0 normal) have not yet been extensively studied. CCAT1--CASC8 was only recently reported in the fusion catalog generated by DEEPEST fusion (17), and VCL--ADK was only previously reported in a study of cancer cell lines (12).

Some fusion transcripts are prevalent in normal tissues and may not be oncogenic

Approximately half of the 236 C4 targeted fusions in our analysis were robustly detected in normal tissues; these may not be particularly relevant to cancer biology, but may play a role in normal biological processes. Of these, 61 fusions are broadly expressed across at least five tissues, involve intra-chromosomal pairs of genes, and can be largely explained by read-thru transcription, local rearrangements or *trans*-splicing of neighboring transcripts.

Some other putative fusions that are prevalent in normal tissues may in fact represent normal structural variation in the human genome, which is not accounted for when performing read alignment to a single human reference. For example, Fusion KANSL1--ARL17, which would require a local rearrangement in the human reference genome, is prevalent across both tumor and normal tissues (median of 31% of individuals, **Figure S7**), and is known to correspond to a common haplotype involving a locally rearranged genomic region observed in populations of European descent (38). An earlier report identified KANSL1--ARL17 in diverse tumor samples and proposed that it may be a cancer predisposition germline fusion specific to Europeans (39). Note, however, that no specific human genetic association evidence was shown for predisposition thus far, and we observe slightly higher prevalence of KANSL-ARL17 among GTEx normal samples than tumors from TCGA (**Figure S7**). Another normal fusion due to a rarer germline structural variation is TFG--GPR128, previously associated with a copy number variation and a haplotype frequency estimated at around 2% of European descent (40). Consistently, we find TFG-GPR128 broadly expressed across tumor and normal tissues and represented similarly at a median of 2% of all tissues examined (**Figure S8**). As more evidence of common structural variation becomes available, other prevalent fusions found in normal tissues may be more easily explained.

Another set of fusions that are less easily explained involve those we found only in normal pancreas and pancreatic carcinoma (**Figures 4, S5e**), involving various pairwise combination of CPA1, CPA2, CLPS, CELA2A, CELA3A, CTRB1, CTRB2, and CTRC (*e.g.*, CELA3A--CPA2, CELA3B--CELA2A, and CELA3A--CELA2A) fused to generate in-frame fusion products. These genes are among the highest expressed in pancreas and mostly on different chromosomes, suggesting trans-splicing may be the predominant underlying mechanism. While the transcripts were within the COSMIC-peak-enriched fusion cluster, the random forest based fusion classifier did not predict their newer instances as COSMIC-like, and some of them (*e.g.*, CELA3A--CTRC) have high fractions of occurrences predicted as 'high-counter-evidence' or 'artifact-like' types (**Figures 5, S6e, Table 6S**).

Some well-established oncogenic fusions are also reliably detected in normal samples

Several of the COSMIC fusions or other tumor-enriched fusions with known ties to cancer were surprisingly identified in both tumor and normal samples. For example, the prostate cancer fusion TMPRSS2—ERG, identified as our fourth most tumor-enriched fusion (182 of 465 TCGA prostate tumor samples), is also detected in six normal prostate samples (5 TCGA, 1 GTEx) (**Figure S6a**). TMPRSS2—ERG was only identified by FusionInspector in prostate tumors or normal prostate, reflecting both its high tumor-specificity and high specificity of fusion-calling.

In another example, COSMIC fusion PVT1--MYC was originally identified by STAR-Fusion in 21 samples (14 TCGA tumor, 1 TCGA normal, and 5 GTEx samples). Interactions between PVT1 and MYC including their fusion are well-known contributors to tumorigenesis (33-35). Through subsequent screening of PVT1--MYC with FusionInspector, we identify additional samples, totaling 32 samples (+9 TCGA tumor, +1 TCGA normal, and +3,-1 GTEx). Most (21/32) are expressed at low levels (below 0.1 FFPM), and we do not find strong evidence for expression to be generally higher in tumor samples than normals ($p < 0.07$, Wilcoxon rank sum test). However, five of the 32 PVT1--MYC occurrences were identified in cervical cancer tumors, and all were significantly more highly expressed than the other samples ($p < 0.02$), with the most highly expressed at 19 FFPM (**Figure S9**). PVT1 and MYC are co-localized to a proximal region in the bottom arm of chromosome 8 and a PVT1--MYC fusion would likely involve local restructuring at the locus in tumors to generate the fusion product. Interestingly, this chromosome 8 region is a known hotspot for insertion of human papilloma virus (HPV) (41), the leading cause of cervical cancer (>90% of cases). Most of the TCGA cervical cancer samples we identified with PVT1--MYC have HPV insertions at this hotspot (see Table S3 of (41)). Thus, we hypothesize that HPV insertion contributes to the formation of the PVT1--MYC fusions. We did not find evidence for HPV insertion in the breast cancer sample with similar levels of PVT1--MYC expression (data not shown).

COSMIC fusion VTI1A--TCF7L2, originally identified as an oncogenic fusion in colorectal cancer (42), was most abundant in stomach, colon, and esophageal carcinoma samples, but also detected in seven individual GTEx normal samples (brain, whole blood, tibial nerve, tibial artery, prostate, and breast) (**Table S5**). While VTI1A--TCF7L2 was not enriched for detection in tumors *vs*. normal, only those fusions in colon cancer were highly expressed (> 0.15 FFPM), whereas other tumor and normal instances were lowly expressed (<0.05 FFPM; many at the limit of detection; **Figure S10**), supported by a single split read defining the fusion breakpoint (**Table S5**). This could be consistent with a very low proportion of cells in the normal tissue expressing the fusion, compared to a large clone in the tumor.

Surprisingly, COSMIC fusion BCR--ABL1 was not tumor-enriched in our analysis, likely due to paucity of the relevant tumors in TCGA. In particular, BCR--ABL1 occurs in >95% of chronic myeloid leukemia (CML) cases (1, 2), but TCGA lacks CML samples. Indeed, the four TCGA tumors with BCR--ABL1 likely correspond to another subtype of AML defined with this fusion (43). Three of these AML tumors also have evidence of the reciprocal ABL1--BCR fusion, and the oncogenic BCR--ABL1 is expressed at higher levels than the reciprocal counterpart in each sample. Interestingly, we detected eight instances of the oncogenic BCR--ABL1 fusion (and none of the reciprocal) in seven different GTEx normal tissues (1 each of adipose, breast, nerve, prostate, and thyroid and 2 pancreas). We observed no sequence or expression features distinguishing these fusions from those we identified in AML, and fusion breakpoints for those in GTEx normal tissues are identical to those found in AML. In general, when we find COSMIC fusions in GTEx normal samples, they are found at low frequencies (< 1% prevalence in a tissue type).

Novel COSMIC-like fusion potentially relevant to breast and nasopharyngeal cancer

Other fusions clustered outside of C4 and found tumor enriched may also warrant further attention. For example, novel COSMIC-like fusion FSIP1--RP11-624L4.1 was detected in 22% of breast cancer tumors studied (240 of 1,086). While it was also detected in 14 normal breast samples (3 TCGA, 11 GTEx) and two additional samples (prostate and esophagus), its expression was significantly higher in tumors than normal tissue (**Figure 6a**, Benjamini Hochberg FDR < 0.004, Wilcoxon rank sum test). FSIP1 (fibrous sheath interacting protein 1) was previously identified as a prognostic marker for HER2-positive breast cancers and its high expression is associated with poor patient outcomes (44). Fusion partner RP11-624L4.1 is a lncRNA, which is collinear and 170kb downstream from FSIP1, and was recently identified as an oncogene relevant to nasopharyngeal carcinoma (45). FSIP1 and RP11-624L4.1 expression is positively correlated in both tumor and normal tissues (Pearson $r = 0.6$) and the fusion FSIP1--RP11-624L4.1 is found only among those samples most highly expressing both fusion partners (**Figure 6b**).

Discussion

We developed FusionInspector to enable exploration of the evidence supporting candidate fusions, flag likely artifacts, and identify those fusions with sequence and expression features similar to known biologically relevant fusion transcripts. Given a list of candidate fused gene pairs, FusionInspector captures RNA-seq read alignments that support either the fused genes or the unfused partner genes. From the fusion and partner gene expression evidence coupled with sequence features relating to the fusion breakpoint, FusionInspector helps the user to reason about the nature and quality of any target fusion transcript.

Clustering fusions by shared sequence and expression features identified a cluster of fusions highly enriched for COSMIC fusions. Fusions in this COSMIC "peak enriched" cluster had relatively high fusion expression with 3'-FAR generally exceeding 5'-FAR, suggesting oncogenic activity from the 3'-fused transcript. Analysis initiated by fusions in the COSMIC peak enriched cluster highlighted several putative novel or less appreciated oncogenic fusions, including CCAT1--CASC8 and VCL—ADK, based on their feature similarity to other well-known tumor-enriched fusions. Only ~3% of initially predicted fusions were members of clusters likely enriched for artifacts based on features such as high partner gene expression or sites of microhomology at or near the fusion breakpoint. The low artifact rate is likely due to the strong filtering of the initial input catalog from STAR-Fusion.

While we focused on fusions identified in the COSMIC-peak-enriched cluster, other COSMIC-like fusion clusters also harbor important oncogenic fusions. For example, the COSMIC fusions SS18--SSX1 and SS18--SSX2, known drivers of synovial sarcoma (5, 46), are in other clusters

(C38 & C39), due in part to their higher 5'-FAR. Another fusion of interest, FSIP1--RP11-624L4.1 is present in 240 (22%) of breast tumors analyzed and in 16 normal breast tissues samples, where it is expressed at significantly lower levels. While the individual fusion partners have cancer associations (44, 45), any role for this newly identified fusion transcript deserves consideration in further exploring the roles of both genes in disease.

In some cases, fusions that were found in both tumor and normal tissues might reflect a low level of oncogenic events. For example, hallmark driver fusions including TMPRSS2--ERG and BCR--ABL1 are also detected in GTEx normal tissue samples, which may reflect low proportion of pre-malignant or transformed cells (47). We also detect COSMIC fusion VTI1A--TCF7L2 across multiple tumor and normal tissue types (consistent with (48)), but only highly expressed in colon cancer samples where it is a postulated oncogenic driver (42, 49). Whether such a fusion could contribute to tumorigenesis in a different tissue with different cellular circuitry remains unknown.

Fusions that were prevalent among normal tissues can mostly be explained by read-through transcription and *cis*-splicing of colinear genes, but some may simply reflect natural germline structural variations that may exist in the population. With ongoing advancements in methods for detecting and cataloguing of structural variants(50, 51), we may soon better understand the structural basis for many naturally occurring fusion transcripts. Access to matched RNA-seq and whole genome sequencing of the same samples across individuals would greatly facilitate such efforts.

Pancreas stood out as a clear outlier among all normal tissues explored for fusions. While we suspect some of the putative fusions detected in pancreas are derived from RT or alignment artifacts, several did have features consistent with *trans*-splicing of highly expressed partner genes, with *trans*-spliced products yielding in-frame proteins. In general, these fusion occurrences do not have COSMIC-like sequence and expression features. *Trans*-spliced in-frame fusion transcripts have the potential to expand functional diversity from our otherwise linear genomes (52), and even if these pancreas specific candidates failed to ultimately reach our COSMIC-like prioritization status, they may be worth additional studies.

FusionInspector opens the way to further explore the biological impact of the predicted fusions and the tissues and gene expression networks in which they are phenotypically relevant. FusionInspector helps illuminate the evidence supporting fusions in RNA-seq, or to sensitively and accurately screen for relevant fusions in samples of interest. Since short reads remain limited in their capacity to represent full length fusion transcripts, FusionInspector further integrates Trinity (53, 54) for *de novo* reconstruction to optionally reconstruct more full-length fusion transcripts from RNA-seq data aligned to each fusion contig. FusionInspector is available as a stand-alone application for screening lists of candidate fusion transcripts, and is also incorporated into STAR-Fusion for *in silico* validation or visualization of STAR-Fusion predicted fusion transcripts. This facilitates analysis of fusions from both bulk and single cell RNA-Seq, as we have recently demonstrated (55).

Long read transcriptome sequencing may eventually obviate short read sequencing for fusion detection, thus removing the need for *de novo* reconstruction of full-length fusion transcripts (56, 57). Full-length single molecule direct RNA-Seq (58) should also avoid RT amplification artifacts. Conversely, other features scored by FusionInspector, such as expression characteristics of fusion transcripts with respect to partner genes, will remain relevant and easily adapted for long read RNA-seq.

Machine learning is likely to play an increasingly important role in biomedical science and its clinical applications. In this paper we emphasize an important companion direction to machine learning, namely generating transparent and interpretable predictions, loosely referred to as explanations. The area of explanations and causal interpretation is growing rapidly in AI (59). We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data.

We hope that the practical applicability of FusionInspector will help drive transparency and other explanatory efforts in predictive areas in genomics and personalized medicine more generally. FusionInspector is freely available as open source on GitHub, provided in container form via Docker and Singularity, and accessible on the Terra cloud computing framework for secure and scalable application across large compendiums of sample collections or patient-derived RNA-seq data.

Conclusions

FusionInspector helps users navigate the landscape of fusion transcript predictions, including screening and reassessment of evidence supporting fusion predictions, and visualization of the evidence via interactive reports (**Figure S11**). We demonstrated that FusionInspector can successfully explore recurrent fusions in tumor and normal samples, and used it to identify clusters of fusions with features similar to those of known biological impact and potentially relevant to cancer or normal tissue biology, highlighting fusion transcripts that should be prioritized for further investigations. FusionInspector can be easily leveraged as an add-on component to any fusion transcript prediction pipeline, and is directly incorporated into STAR-Fusion to facilitate execution as part of the Trinity Cancer Transcriptome Analysis Toolkit (60).

Methods

Fusion transcript prediction for TCGA and GTEx

Fusions were predicted for TCGA and GTEx samples using STAR-Fusion (v1.7). First, the STAR (v2.6.1a) aligner was used to align RNA-seq reads from each sample to the human genome as follows:

```
"STAR --genomeDir ctat_genome_lib_build_dir/ref_genome.fa.star.idx --outReadsUnmapped None --chimSegmentMin 12 --chimJunctionOverhangMin 12 --chimOutJunctionFormat 1 --alignSJDBoverhangMin 10 --alignMatesGapMax 100000 --alignIntronMax 100000 --alignSJstitchMismatchNmax 5 -1 5 5 --runThreadN 16 --outSAMstrandField intronMotif --outSAMunmapped Within --outSAMtype BAM Unsorted --readFilesIn reads_1.fastq reqds_2.fastq --outSAMattrRGline ID:GRPundef --chimMultimapScoreRange 10 --chimMultimapNmax 10 --chimNonchimScoreDropMin 10 --peOverlapNbasesMin 12 --peOverlapMMp 0.1 --genomeLoad NoSharedMemory --twopassMode Basic".
```

The resulting Chimeric.out.junction files generated by STAR containing candidate chimeric reads were then analyzed by STAR-Fusion like so "STAR-Fusion -J Chimeric.out.junction -O $output_dir/STARF --genome_lib_dir $ctat_genome_lib --min_FFPM 0 --no_annotation_filter" leveraging CTAT genome library GRCh38_gencode_v22_CTAT_lib_Sept032019.

The parameters used here eliminated any filtering of fusions according to fusion expression levels or based on fusion annotations, so as to retain any fusions known to frequently occur in normal samples in both the normal and the tumor samples for further study. All STAR-Fusion predictions are provided in **Table S1**.

Fusion tumor enrichment was computed for each fusion according to ((# tumor with fusion + 1) / (total tumor samples)) / ((# normal with fusion + 1) / (total normal samples)).

FusionInspector method and implementation

FusionInspector takes as input a list of candidate fusions and RNA-seq files in fastq format, with each fusion formatted as "geneA--geneB" indicating a candidate fusion between geneA (5') and geneB (3'). Leveraging the companion CTAT genome library set of genomic resources (identical to that used with STAR-Fusion, including the human reference genome, gene structure annotations, and STAR genome index), FusionInspector constructs fusion contigs by extracting the genomic sequences for each geneA and geneB, and concatenating each geneA and geneB pair into a single contig in collinear transcribed orientation. Gene structure annotations for fusion genes are similarly restructured to match the position and orientation of the corresponding genes in the fusion contigs. By default, long introns are shrunk to 1 kb in length by removing central regions of intron sequences, reducing the alignment search space and simplifying downstream visualizations.

RNA-seq reads are aligned to the fusion contigs along with the whole reference genome by running STAR with both inputs, including the pre-indexed whole genome and a fasta file containing the fusion contigs. STAR first loads the whole reference genome index into RAM, then builds an index for the fusion contigs, and incorporates the fusion contig index into the whole genome index. Only those reads that align concordantly to the fusion contigs, while considering all alignments to the combined targets, are reported. Note, that in the fusion context, all fusion-supporting reads are aligned concordantly, but will align partially to one gene and partially to the adjacent gene. This functionality was implemented in STAR since version 2.5.0a to support FusionInspector functionality. STAR-Fusion directly executes STAR to align reads like so " STAR --runThreadN 4 --genomeDir ctat_genome_lib_build_dir/ref_genome.fa.star.idx --outSAMtype BAM SortedByCoordinate --twopassMode Basic --alignSJDBoverhangMin 10 --genomeSuffixLengthMax 10000 --limitBAMsortRAM 47271261705 --alignInsertionFlush Right --alignMatesGapMax 100000 --alignIntronMax 100000 --readFilesIn reads_1.fastq.gz reads_2.fastq.gz --genomeFastaFiles finspector.fa --outSAMfilter KeepAllAddedReferences --sjdbGTFfile finspector.gtf --alignSJstitchMismatchNmax 5 -1 5 5 --scoreGapNoncan -6 --readFilesCommand 'gunzip -c' ", where 'finspector.fa' and 'finspector.gtf' correspond to the fusion contigs sequence and structure annotation files.

FusionInspector examines the aligned reads output by STAR and identifies read alignments supporting fusions between gene pairs represented by the fusion contigs. Candidate fusion breakpoints are identified by split read alignments having partial alignments that anchor to exons of the neighboring fusion genes. Spanning fragments are identified as paired-end reads having each read mapping entirely on opposite sides of the breakpoint. Alignments must meet minimum evidence criteria to be counted as evidence, and require at least 96% sequence identity and no more than 10 bases unaligned at their ends (soft- or hard-clipped bases). For split reads, at least 10 bases must align adjacent to each breakpoint (anchor), and each anchor region must have sufficient sequence complexity, requiring entropy >= 1.2. For spanning fragments, each paired-end read must have sufficient sequence complexity, requiring entropy >= 1.2. Preliminary fusion predictions are defined based on candidate fusion breakpoints and sets of compatible spanning fragments. RNA-seq fragments that span a candidate breakpoint but support transcription from an unfused partner gene are captured and stored as counter-evidence and used to compute the partner gene counter FFPM and fusion allelic ratio.

There is often evidence for multiple fusion isoforms, and while the split reads are unique to and define each breakpoint, the spanning fragments are often compatible with multiple breakpoints and assigned ambiguously. We implemented an expectation maximization (EM) algorithm based on that described in kallisto (61) to fractionally assign RNA-seq evidence fragments to fusion isoforms according to maximum likelihood. Fusion expression values (FFPM) are then computed based on estimated RNA-seq fragment counts resulting from the EM.

Fusion candidates are then filtered according to defined minimum evidence requirement, with defaults set as requiring at least one split read to define the junction breakpoint, and at least 25 aligned bases supported by at least one read on both sides of the fusion breakpoint. If the breakpoint involves non-consensus dinucleotide splice sites, then at least three split reads are required to support the breakpoint. Reads must also be found to align with at least 96% (default) sequence identity. A final filter of fusion predictions to exclude those containing overly promiscuous fusion partners (maximum 10) or those involving paralogs of more dominantly supported fusions is applied identically as previously described (62).

Optionally, Trinity de novo assembly (53, 54) is integrated to de novo reconstruct candidate fusion transcripts based on reads aligning to the fusion contigs. When employed, Trinity-reconstructed fusion transcripts are identified in the final FusionInspector report and the assembled transcripts are available for further study. In addition, FusionInspector integrates IGV-reports (63) to generate an interactive web-based summary (and fully self-contained html file) of predicted fusions coupled to a web-based interactive genome viewer to examine the read alignments found as evidence for the fusions.

FusionInspector benchmarking

FusionInspector (v2.4.0) was benchmarked in both execution modes: (**1**) as a post-process to STAR-Fusion and examining only those predictions generated by STAR-Fusion (v1.9.1), and (**2**) fusion screening mode, providing FusionInspector with a list of fusion candidates that we derived from a diverse set of prediction methods. Benchmarking was performed using cancer cell line RNA-seq, as previously described (62), except STAR-Fusion and FusionInspector were both excluded from the list of methods whose intersections define the truth sets, and hence eliminating any potential bias of accuracy in favor of either STAR-Fusion or FusionInspector fusion predictions. For (2), fusion screening mode, lists of candidate fusions were generated per sample only requiring the fusion to be predicted by any of the 24 different prediction methods evaluated earlier on these samples (see Table S4 of (11)), and first filtered to remove likely read-through fusions, fusions between paralogs, and any fusions found among normal tissue types - all as previously described in the earlier benchmarking (Table S4 of (11)). FusionInspector predictions on cancer cell lines used for benchmarking are provided in **Table S1**.

Applications of FusionInspector to TCGA and GTEx

FusionInspector v2.4.0 was applied to TCGA and GTEx samples in both execution modes: (1) an in silico validation of STAR-Fusion (v1.9.1) predictions, and (2) screening a specified set of fusion candidates. FusionInspector was run on TCGA v11 and GTEx v8 via Terra/AnVIL (64). Each execution mode is detailed below.

First, FusionInspector was used to reexamine a subset of 628 TCGA and 530 GTEx samples identified as containing instances of recurrent STAR-Fusion (v1.7) predictions. Candidate samples were identified based on individual fusions (a) having minimum 0.1 FFPM and (b) found in tissue types with at least three occurrences and comprising at least 10% of samples of that tissue type, or (c) containing a COSMIC fusion. Samples were then greedily selected to maximize recurrent fusion content while minimizing numbers of selected samples, retaining up to 10 samples per fusion. These samples were reexamined by executing the current STAR-Fusion (v1.9.1) including FusionInspector (v2.4.0) as a post-process like so: "STAR-Fusion --left_fq ${sample_name}_1.fastq --right_fq ${sample_name}_2.fastq --CPU 16 --genome_lib_dir ctat_genome_lib_build_dir --output_dir ${sample_name} --FusionInspector validate --no_annotation_filter --min_FFPM 0 " leveraging companion CTAT genome library "GRCh38_gencode_v22_CTAT_lib_Apr032020". The FusionInspector abridged outputs were consolidated and presented as **Table S3**. These fusions were subsequently subject to Leiden clustering (65) (see ***Fusion Clustering and Class Prediction*** section below).

Second, FusionInspector was run in fusion screening mode to explore instances of defined COSMIC-peak-enriched fusions (Leiden cluster 4 (C4) of the 61 fusion clusters found to be heavily enriched for COSMIC fusions). There were 231 instances of C4 fusions selected according to the following criteria: found in at least 3 samples, at least one fusion occurrence found clustered to C4, and at least 30% of occurrences annotated as COSMIC-like. These were further supplemented with five recurrent COSMIC fusions that are not members of C4 (ERC1--RET, SLC34A2--ROS1, SS18--SSX1, SS18--SSX2, and VTI1A--TCF7L2), to a total of 236 fusion gene pairs (**Table S4**). The 236 fusion gene targets were provided as input to FusionInspector for screening 2,764 TCGA and 1,009 GTEx samples, each with the same list of 236 candidates. These samples were selected based on having a STAR-Fusion predicted occurrence of at least one of these fusions (from **Table S2**), and selecting a maximum of 50 samples per-fusion gene-pairing (with samples sometimes containing multiple fusion types), except for pancreatic and prostate cancer (TCGA) and normal pancreas tissue (GTEx) for which all samples were selected as targets. FusionInspector was executed like so: "FusionInspector --

fusions $Table_S4_fusions --genome_lib_dir ctat_genome_lib_build_dir -O ${sample_name} --left_fq ${sample_name}_1.fastq --right_fq ${sample_name}_2.fastq --out_prefix ${sample_name} --vis" leveraging companion CTAT genome library "GRCh38_gencode_v22_CTAT_lib_Apr032020", and results for screening of these samples are provided in **Table S5**.

Fusion transcript clustering and attribute class prediction

All 53,240 fusion isoforms surveyed by FusionInspector from our initial subset of TCGA and GTEx samples were clustered according to sequence and expression characteristics. Microhomologies defined as exact *k*-mers with *k*=10 were identified between candidate fusion gene pairs as represented in the FusionInspector-constructed fusion contigs (with introns shrunk to a max of 1 kb each for simpler visualizations). The Euclidean distance of each candidate fusion breakpoint to the nearest site of microhomology was determined in the FusionInspector fusion contig coordinate system. Attributes of interest for clustering fusions were: (1) the fusion expression level (FFPM), (2,3) partner gene fusion allelic ratios (5'-FAR and 3'-FAR), (4,5) the left and right unfused partner gene expression levels expressed as 5'- and 3'-counter-FFPM and computed based on the number of counter-reads observed as aligned at each corresponding gene breakpoint site, (6,7) indicators for consensus dinucleotides and agreement with reference gene structure exon boundaries at the fusion breakpoints, and (8) the number of microhomologies and (9) distance of the breakpoint to the nearest microhomology. These numerical values were centered and scaled to Z-scores, truncated within the interval [-2,2] to remove outliers, and then rescaled so each attribute numerical vector would fill the interval [-2,2] simplifying our evaluation of metrics using a consistent low-to-high range for each attribute type.

We calcuated the distance between fusions based on vectors with these values, constructed a *k*-nearest-neighbor graph (*k*=50) of fusions, and clustered the graph by Leiden clustering (65) (resolution_parameter = 3). The impact of the resolution parameter on clustering and COSMIC fusion enrichment was examined (**Figure S12**), and the parameter with the most granular set of clusters was selected for further analysis. Clusters were manually reviewed and grouped and annotated according to median cluster attributes, with cluster annotation term assignments as "COSMIC-like" if clusters contained at least two COSMIC fusions, "COSMIC-peak-enriched" if predicted as cluster C4, "high-counter FFPM" indicating relatively high expression of the partner genes and potentially resulting from a low rate of trans-splicing, and categories "High FAR" and "Microhomology RT-induced artifact" to reflect likely bioinformatic or reverse-transcription related artifacts (as labeled in **Figure 4b**).

A random forest classifier was built to predict Leiden cluster membership based on scaled fusion attributes. The classifier was constructed by randomly selecting a maximum of 300 fusions (median cluster size) from each cluster, and leveraging 2/3 of fusions for training and 1/3 for testing, all performed using Ranger (66). Fusions predicted to be assigned to any cluster noted earlier with a fusion cluster annotation (*e.g.*, "COSMIC-like") are assigned a prediction according to that fusion cluster annotation term. Such fusion attribute cluster predictions are now incorporated into the latest FusionInspector (v2.6.0).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data are provided in supplementary tables in addition to being available on GitHub with code to demonstrate analysis methods and generation of figures (37) .

Competing interests

A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was a scientific advisory board member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until 31 July 2020. From 1 August 2020, A.R. has been an employee of Genentech. MG is a current employee and stock holder at Monte Rosa Therapeutics.

Funding

This work has been supported by National Cancer Institute grants U24CA180922 (A.R. and B.J.H.), R50CA211461 (B.J.H.), Howard Hughes Medical Institute (A.R.), Klarman Cell Observatory (A.R.), NHGRI R01HG009318 (AD), and American Association of Obstetricians and Gynecologists Foundation (AAOGF) Scholar Grant 2017-2020 (AVA).

Authors' Contributions

BJH performed analyses, developed the FusionInspector software, and wrote the initial draft of this manuscript. AD enhanced the STAR aligner software to support FusionInspector execution. JTR and TT developed the FusionInspector IGV-reports interactive fusion evidence visualization component. AVA assisted with investigations of HPV insertions and MYC--PVT1 fusion studies in CESC samples. AR and SK advised this work, and all authors made intellectual contributions towards the design of FusionInspector and to the final manuscript.

Acknowledgements

We thank Joshua Gould for assisting with software pipeline development and Terra integration, and Leslie Gaffney for help with generating publication quality figures.

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. TCGA v11 data were obtained through Terra/Anvil via dbGaP accesssion number phs000178.v11.p8.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx v8 data used for the analyses described in this manuscript were obtained from Terra/AnVIL via via dbGaP accession number phs000424.v8.p2.

References

1. Kurzrock R, Gutterman JU, Talpaz M. The molecular genetics of Philadelphia chromosome-positive leukemias. N Engl J Med. 1988;319(15):990-8.

2. Ren R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. Nat Rev Cancer. 2005;5(3):172-83.

3. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310(5748):644-8.

4. Rubin MA, Maher CA, Chinnaiyan AM. Common gene rearrangements in prostate cancer. J Clin Oncol. 2011;29(27):3659-68.

5. Clark J, Rocques PJ, Crew AJ, Gill S, Shipley J, Chan AM, et al. Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma. Nat Genet. 1994;7(4):502-8.

6. Hale R, Sandakly S, Shipley J, Walters Z. Epigenetic Targets in Synovial Sarcoma: A Mini-Review. Front Oncol. 2019;9:1078.

7. Yang W, Lee KW, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat Med. 2019;25(5):767-75.

8. Wei Z, Zhou C, Zhang Z, Guan M, Zhang C, Liu Z, et al. The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis. iScience. 2019;21:249-60.

9. 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.

10. Kumar S, Razzaq SK, Vo AD, Gautam M, Li H. Identifying fusion transcripts using next generation sequencing. Wiley Interdiscip Rev RNA. 2016;7(6):811-23.

11. 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 Biology. 2019;20(1):213.

12. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol. 2015;33(3):306-12.

13. Babiceanu M, Qin F, Xie Z, Jia Y, Lopez K, Janus N, et al. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res. 2016;44(6):2859-72.

14. 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;46(D1):D1144-D9.

15. 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.

16. Kim P, Yoon S, Kim N, Lee S, Ko M, Lee H, et al. ChimerDB 2.0--a knowledgebase for fusion genes updated. Nucleic Acids Res. 2010;38(Database issue):D81-5.

17. Dehghannasiri R, Freeman DE, Jordanski M, Hsieh GL, Damljanovic A, Lehnert E, et al. Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. Proc Natl Acad Sci U S A. 2019;116(31):15524-33.

18. Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, et al. State-of-the-art fusion-finder algorithms sensitivity and specificity. Biomed Res Int. 2013;2013:340620.

19. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep. 2016;6:21597.

20. Yu CY, Liu HJ, Hung LY, Kuo HC, Chuang TJ. Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? Nucleic Acids Res. 2014;42(14):9410-23.

21. Peng Z, Yuan C, Zellmer L, Liu S, Xu N, Liao DJ. Hypothesis: Artifacts, Including Spurious Chimeric RNAs with a Short Homologous Sequence, Caused by Consecutive Reverse Transcriptions and Endogenous Random Primers. J Cancer. 2015;6(6):555-67.

22. Shivram H, Iyer VR. Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. RNA. 2018;24(9):1266-74.

23. 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.

24. 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.

25. Sabir SR, Yeoh S, Jackson G, Bayliss R. EML4-ALK Variants: Biological and Molecular Properties, and the Implications for Patients. Cancers (Basel). 2017;9(9).

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

27. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.

28. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):D777-D83.

29. "Wellcome Sanger Institute". COSMIC Catalogue of Somatic Mutations in Cancer 2019 [Available from: https://cancer.sanger.ac.uk/cosmic.

30. Lasorella A, Sanson M, Iavarone A. FGFR-TACC gene fusions in human glioma. Neuro Oncol. 2017;19(4):475-83.

31. Liquori A, Ibanez M, Sargas C, Sanz MA, Barragan E, Cervera J. Acute Promyelocytic Leukemia: A Constellation of Molecular Events around a Single PML-RARA Fusion Gene. Cancers (Basel). 2020;12(3).

32. Frattini V, Trifonov V, Chan JM, Castano A, Lia M, Abate F, et al. The integrated landscape of driver genomic alterations in glioblastoma. Nat Genet. 2013;45(10):1141-9.

33. Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012;488(7409):49-56.

34. Jin K, Wang S, Zhang Y, Xia M, Mo Y, Li X, et al. Long non-coding RNA PVT1 interacts with MYC and its downstream molecules to synergistically promote tumorigenesis. Cell Mol Life Sci. 2019;76(21):4275-89.

35. Tolomeo D, Agostini A, Visci G, Traversa D, Storlazzi CT. PVT1: A long non-coding RNA recurrently involved in neoplasia-associated fusion transcripts. Gene. 2021;779:145497.

36. Ardini E, Bosotti R, Borgia AL, De Ponti C, Somaschini A, Cammarota R, et al. The TPM3-NTRK1 rearrangement is a recurring event in colorectal carcinoma and is associated with tumor sensitivity to TRKA kinase inhibition. Mol Oncol. 2014;8(8):1495-507.

37. Haas B. Analyses, Code, and Data Supporting the FusionInspector Paper 2021 [Available from: https://github.com/broadinstitute/FusionInspectorPaper.

38. Boettger LM, Handsaker RE, Zody MC, McCarroll SA. Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet. 2012;44(8):881-5.

39. Zhou JX, Yang X, Ning S, Wang L, Wang K, Zhang Y, et al. Identification of KANSARL as the first cancer predisposition fusion gene specific to the population of European ancestry origin. Oncotarget. 2017;8(31):50594-607.

40. Chase A, Ernst T, Fiebig A, Collins A, Grand F, Erben P, et al. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. Haematologica. 2010;95(1):20-6.

41. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, Barretos Cancer H, Baylor College of M, Beckman Research Institute of City of H, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017;543(7645):378-84.

42. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011;43(10):964-8.

43. Neuendorff NR, Burmeister T, Dorken B, Westermann J. BCR-ABL-positive acute myeloid leukemia: a new entity? Analysis of clinical and molecular features. Ann Hematol. 2016;95(8):1211-21.

44. Yan M, Wang J, Ren Y, Li L, He W, Zhang Y, et al. Over-expression of FSIP1 promotes breast cancer progression and confers resistance to docetaxel via MRP1 stabilization. Cell Death Dis. 2019;10(3):204.

45. Zhou L, Liu R, Liang X, Zhang S, Bi W, Yang M, et al. lncRNA RP11-624L4.1 Is Associated with Unfavorable Prognosis and Promotes Proliferation via the CDK4/6-Cyclin D1-Rb-E2F1 Pathway in NPC. Mol Ther Nucleic Acids. 2020;22:1025-39.

46. Gazendam AM, Popovic S, Munir S, Parasu N, Wilson D, Ghert M. Synovial Sarcoma: A Clinical Review. Curr Oncol. 2021;28(3):1909-20.

47. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science. 2019;366(6465).

48. Nome T, Hoff AM, Bakken AC, Rognum TO, Nesbakken A, Skotheim RI. High frequency of fusion transcripts involving TCF7L2 in colorectal cancer: novel fusion partner and splice variants. PLoS One. 2014;9(3):e91264.

49. Davidsen J, Larsen S, Coskun M, Gogenur I, Dahlgaard K, Bennett EP, et al. The VTI1A-TCF4 colon cancer fusion protein is a dominant negative regulator of Wnt signaling and is transcriptionally regulated by intestinal homeodomain factor CDX2. PLoS One. 2018;13(7):e0200215.

50. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444-51.

51. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature. 2020;583(7814):83-9.

52. Gingeras TR. Implications of chimaeric non-co-linear transcripts. Nature. 2009;461(7261):206-11.

53. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-52.

54. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494-512.

55. Jerby-Arnon L, Neftel C, Shore ME, Weisman HR, Mathewson ND, McBride MJ, et al. Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat Med. 2021;27(2):289-300.

56. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. BMC Genomics. 2020;21(Suppl 11):793.

57. Rautiainen M, Durai DA, Chen Y, Xin L, Low HM, Göke J, et al. AERON: Transcript quantification and gene-fusion detection using long reads. bioRxiv. 2020:2020.01.27.921338.

58. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201-6.

59. Kasif S, Roberts RJ. We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data. PLoS Biol. 2020;18(11):e3000999.

60. Haas BJ. Trinity Cancer Transcriptome Analysis Toolkit 2019 [Available from: https://github.com/NCIP/Trinity_CTAT/wiki.

61. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-7.

62. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv. 2017.

63. Robinson J. igv-reports 2019 [Available from: https://github.com/igvteam/igv-reports.

64. Terra. Terra: a scalable platform for biomedical research [Available from: https://terra.bio/.

65. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.

66. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. 2017 %9 C++; classification; machine learning; R; random forests; Rcpp; recursive partitioning; survival analysis %! ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. 2017;77(1 %@ 1548-7660 %8 2017-03-31 %7 2017-03-31):17.

67. Commons" NCIGD. TCGA Study Abbreviations 2021 [Available from: https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations.

68. Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675-8.

Figure Legends

Figure 1. FusionInspector Overview. **Top:** Lists of fusion candidates derived from predictions of one or multiple fusion detection methods, or from a screening panel, are provided to FusionInspector as input along with RNA-seq in fastq format. For each candidate fusion gene, fusion contigs are generated by fusing the full-length gene candidates as collinear on a single contig. Intronic regions are by default each shrunk to 1 kb. RNA-seq reads are then aligned to a reference consisting of the entire genome supplemented with fusion contigs. Fusion-derived reads that would normally align discordantly as chimeric alignments in the reference genome (top example) instead align concordantly in the fusion contig context (bottom example). **Middle:** FusionInspector identifies split read alignments (light blue) and spanning pairs (purple) supporting the gene fusion in addition to read alignments that overlap the breakpoints and instead support the unfused fusion partners (fusion counter reads). **Bottom:** For those fusions where FusionInspector captures RNA-seq read support ("*in silico* validation"), it reports fusion sequence and expression characteristics including reference gene structure splice agreement, and renders each prediction as COSMIC-like, potential artifact, or other category (**Methods**).

Figure 2. Features of fusion genes distinguish reliable and likely artifactual fusions. Fusion isoform expression level (dot size), splice type (dot color) and splice junction dinucleotide (dot shape) at each fusion breakpoint position involving the 5' (x axis) and 3' (y axis) partners of (**a**) EML--ALK (in COSMIC) and (**b**) KRT13--KRT4 (likely artifactual) fusions. Black dots: positions of microhomology (10 base exact match). Structures of collapsed isoforms for fusion partner genes are drawn along each axis.

Figure 3: COSMIC fusions show distinctive properties among STAR-Fusion predictions across TCGA and GTEx. (**a**) Tissue and tumor composition. Percentages of TCGA tumor or GTEx normal samples (y axis) with corresponding predicted COSMIC fusions (x axis). TCGA study abbreviation codes as in (67). (**b,c**) COSMIC fusions are more highly expressed than other predicted fusions. (**b-d**) Expression levels. (**b**) Distribution of fusion expression levels (*y* axis, FFPM; right-truncated at 1 FFPM) for all fusions predicted in TCGA tumors (purple), TCGA normal (blue), GTEx (green) and in COSMIC (red). (**c**) Cumulative fraction (*y* axis) of all predicted fusions at each minimum fusion expression (*x* axis, FFPM). (**d**) Distribution of fusion expression levels (y axis, FFPM) for each predicted COSMIC fusions (x axis). For a-d, fusions are restricted to the single highest expressed fusion isoform per sample occurrence, require reference annotation splice agreement at breakpoints, and have mitochondrial, HLA, and immunoglobulin gene containing fusions filtered.

Figure 4. Fusions grouping by sequence and expression features distinguishes COSMIC-like fusions from likely artifactual ones. (**a**) Fusion clusters. Uniform Manifold Approximation and projection (UMAP) of 53,240 fusion isoforms feature profiles (dots), colored by Leiden cluster. Red label: Cluster C4. (**b**) Cluster C4 is enriched for COSMIC fusions. Features (columns, right), number of COSMIC fusions (x axis, second from left), cluster size (x axis, second from right), and fraction of COSMIC fusions (x axis, right) for each fusion cluster (rows). Heatmap shows median scaled intensity values for each feature (color bar).

Figure 5. Characteristic properties of recurrent C4 and COSMIC fusions can distinguish biologically meaningful fusions and fusion instances. 236 selected COSMIC-peak-enriched (C4) and additional COSMIC fusions (columns / x axis) rank ordered by tumor enrichment and shown with fraction of the instances of each fusion in each category based on predicted Leiden cluster labels (**a**, rows, top) or corresponding to presumed impact on coding sequence (**a**, rows, bottom); fusion structure type based on the fusion partner's chromosomal location (**b**); fraction of instances that is in each tumor or tissue type in TCGA and GTEx (**c**, rows); presence in COSMIC (**d**, purple), significantly higher expression in tumors *vs.* normal tissues (**e,** Wilcoxon rank sum test applied to FFPM, Benjamini Hochberg FDR < 0.05 and median tumor FFPM > median normal FFPM, orange), number of tumor (seagreen) or normal (light red) samples (**f,** y axis) predicted by STAR-Fusion to contain the fusion, rank ordered by tumor enrichment (**f,** x axis, (**Methods**, gray). See **Figures 6a-e** for each fusion pair.

Figure 6: FSIP1--RP11-624L4.1 fusion in breast cancer. (**a**) Expression level (FFPM, y axis) of FSIP1--RP11-624L4.1 fusion in tumor (blue) and normal (red) TCGA breast cancer samples, ranked by FSIP1--RP11-624L4.1 expression. (**b**) Expression levels of FSIP1 (x axis) and RP11-624L4.1 (y axis) in each tumor (blue) and normal (red) samples (Pearson r=0.6, p-value < 2.2e-16). Diamonds: samples where the fusion trsanscript is detected. Expression values were $\log_2$ transformed upper-quartile normalized gene FPKM measurements obtained from the Xena platform (68).

Figure 1

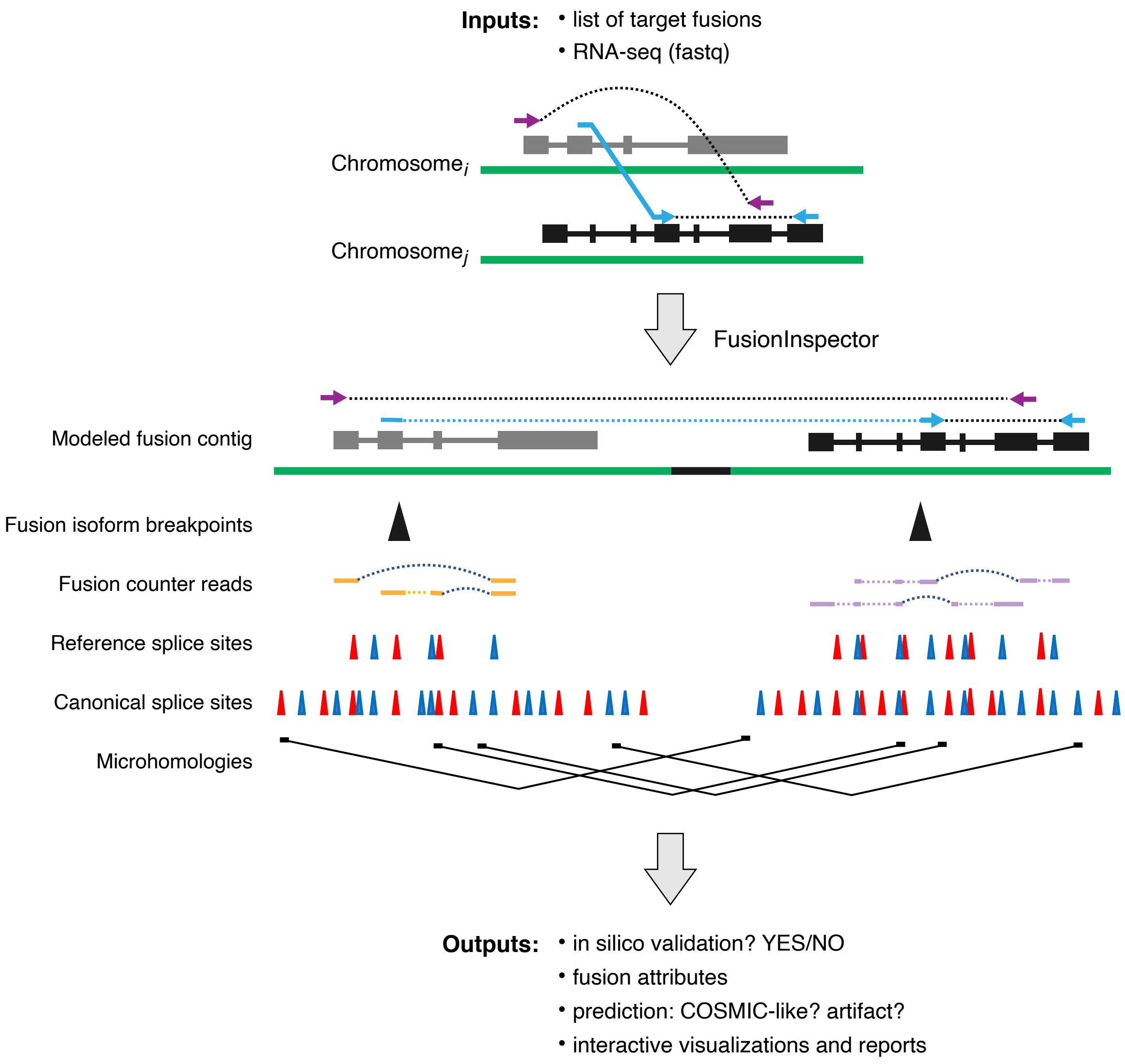

Inputs:
• list of target fusions
• RNA-seq (fastq)
Chromosome$_i$
Chromosome$_j$
FusionInspector
Modeled fusion contig
Fusion isoform breakpoints
Fusion counter reads
Reference splice sites
Canonical splice sites
Microhomologies
Outputs:
• in silico validation? YES/NO
• fusion attributes
• prediction: COSMIC-like? artifact?
• interactive visualizations and reports

Figure 2

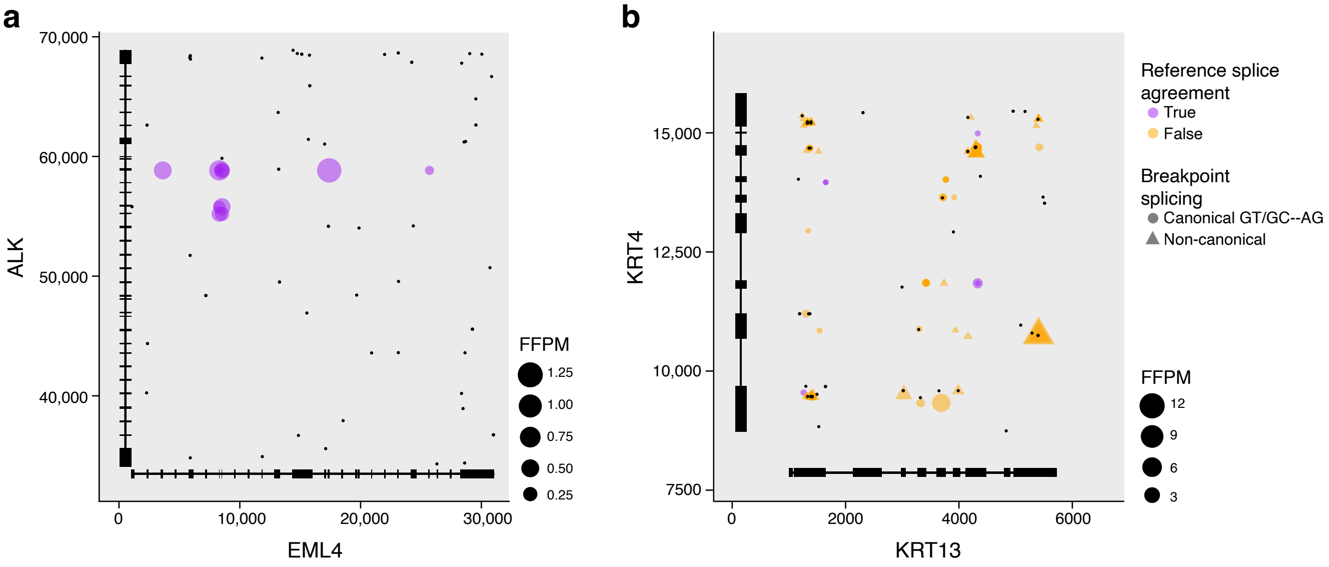

a
70,000
60,000
50,000
40,000
ALK
0
10,000
20,000
30,000
EML4
FFPM
1.25
1.00
0.75
0.50
0.25
b
15,000
12,500
10,000
7500
KRT4
0
2000
4000
6000
KRT13
Reference splice agreement
True
False
Breakpoint splicing
Canonical GT/GC--AG
Non-canonical
FFPM
12
9
6
3

Figure 3

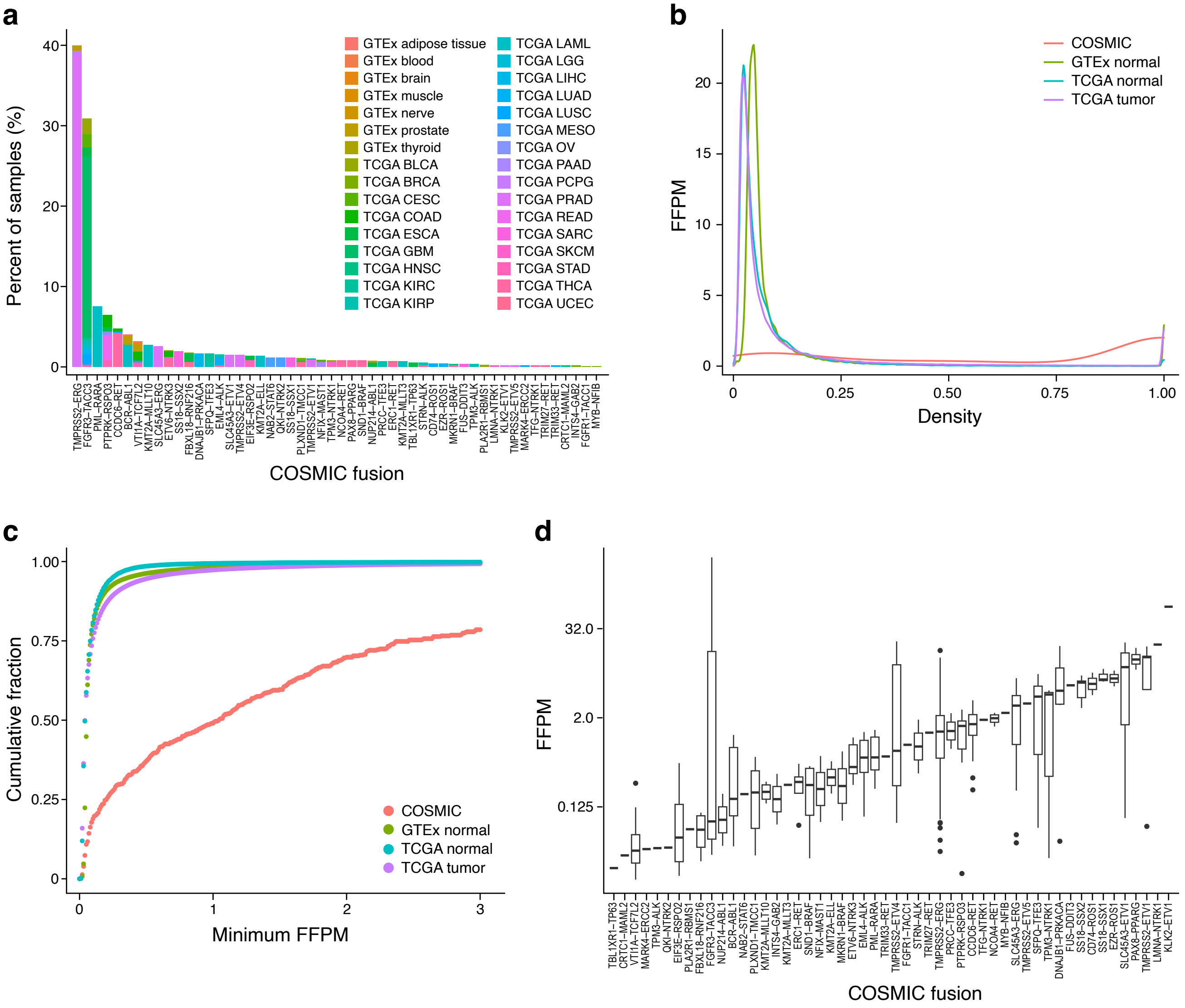

a
Percent of samples (%)
0
10
20
30
40
GTEx adipose tissue
GTEx blood
GTEx brain
GTEx muscle
GTEx nerve
GTEx prostate
GTEx thyroid
TCGA BLCA
TCGA BRCA
TCGA CESC
TCGA COAD
TCGA ESCA
TCGA GBM
TCGA HNSC
TCGA KIRC
TCGA KIRP
TCGA LAML
TCGA LGG
TCGA LIHC
TCGA LUAD
TCGA LUSC
TCGA MESO
TCGA OV
TCGA PAAD
TCGA PCPG
TCGA PRAD
TCGA READ
TCGA SARC
TCGA SKCM
TCGA STAD
TCGA THCA
TCGA UCEC
TMPRSS2–ERG
FGFR3–TACC3
PML–RARA
PTPRK–RSPO3
CCDC6–RET
BCR–ABL1
VTI1A–TCF7L2
KMT2A–MLLT10
SLC45A3–ERG
ETV6–NTRK3
SS18–SSX2
FBXL18–RNF216
DNAJB1–PRKACA
SFPQ–TFE3
EML4–ALK
SLC45A3–ETV1
TMPRSS2–ETV4
EIF3E–RSPO2
KMT2A–ELL
NAB2–STAT6
QKI–NTRK2
SS18–SSX1
PLXND1–TMCC1
TMPRSS2–ETV1
NFIX–MAST1
TPM3–NTRK1
NCOA4–RET
PAX8–PPARG
SND1–BRAF
NUP214–ABL1
PRCC–TFE3
ERC1–RET
KMT2A–MLLT3
TBL1XR1–TP63
STRN–ALK
CD74–ROS1
EZR–ROS1
MKRN1–BRAF
FUS–DDIT3
TPM3–ALK
PLA2R1–RBMS1
LMNA–NTRK1
KLK2–ETV1
TMPRSS2–ETV5
MARK4–ERCC2
TFG–NTRK1
TRIM27–RET
TRIM33–RET
CRTC1–MAML2
INTS4–GAB2
FGFR1–TACC1
MYB–NFIB
COSMIC fusion
b
FFPM
0
5
10
15
20
0
0.25
0.50
0.75
1.00
Density
COSMIC
GTEx normal
TCGA normal
TCGA tumor
c
Cumulative fraction
0
0.25
0.50
0.75
1.00
0
1
2
3
Minimum FFPM
COSMIC
GTEx normal
TCGA normal
TCGA tumor
d
FFPM
0.125
2.0
32.0
TBL1XR1–TP63
CRTC1–MAML2
VTI1A–TCF7L2
MARK4–ERCC2
TPM3–ALK
QKI–NTRK2
EIF3E–RSPO2
PLA2R1–RBMS1
FBXL18–RNF216
FGFR3–TACC3
NUP214–ABL1
BCR–ABL1
NAB2–STAT6
PLXND1–TMCC1
KMT2A–MLLT10
INTS4–GAB2
KMT2A–MLLT3
ERC1–RET
SND1–BRAF
NFIX–MAST1
KMT2A–ELL
MKRN1–BRAF
ETV6–NTRK3
EML4–ALK
PML–RARA
TRIM33–RET
TMPRSS2–ETV4
FGFR1–TACC1
STRN–ALK
TRIM27–RET
TMPRSS2–ERG
PRCC–TFE3
PTPRK–RSPO3
CCDC6–RET
TFG–NTRK1
NCOA4–RET
MYB–NFIB
SLC45A3–ERG
TMPRSS2–ETV5
SFPQ–TFE3
TPM3–NTRK1
DNAJB1–PRKACA
FUS–DDIT3
SS18–SSX2
CD74–ROS1
SS18–SSX1
EZR–ROS1
SLC45A3–ETV1
PAX8–PPARG
TMPRSS2–ETV1
LMNA–NTRK1
KLK2–ETV1
COSMIC fusion

Figure 4

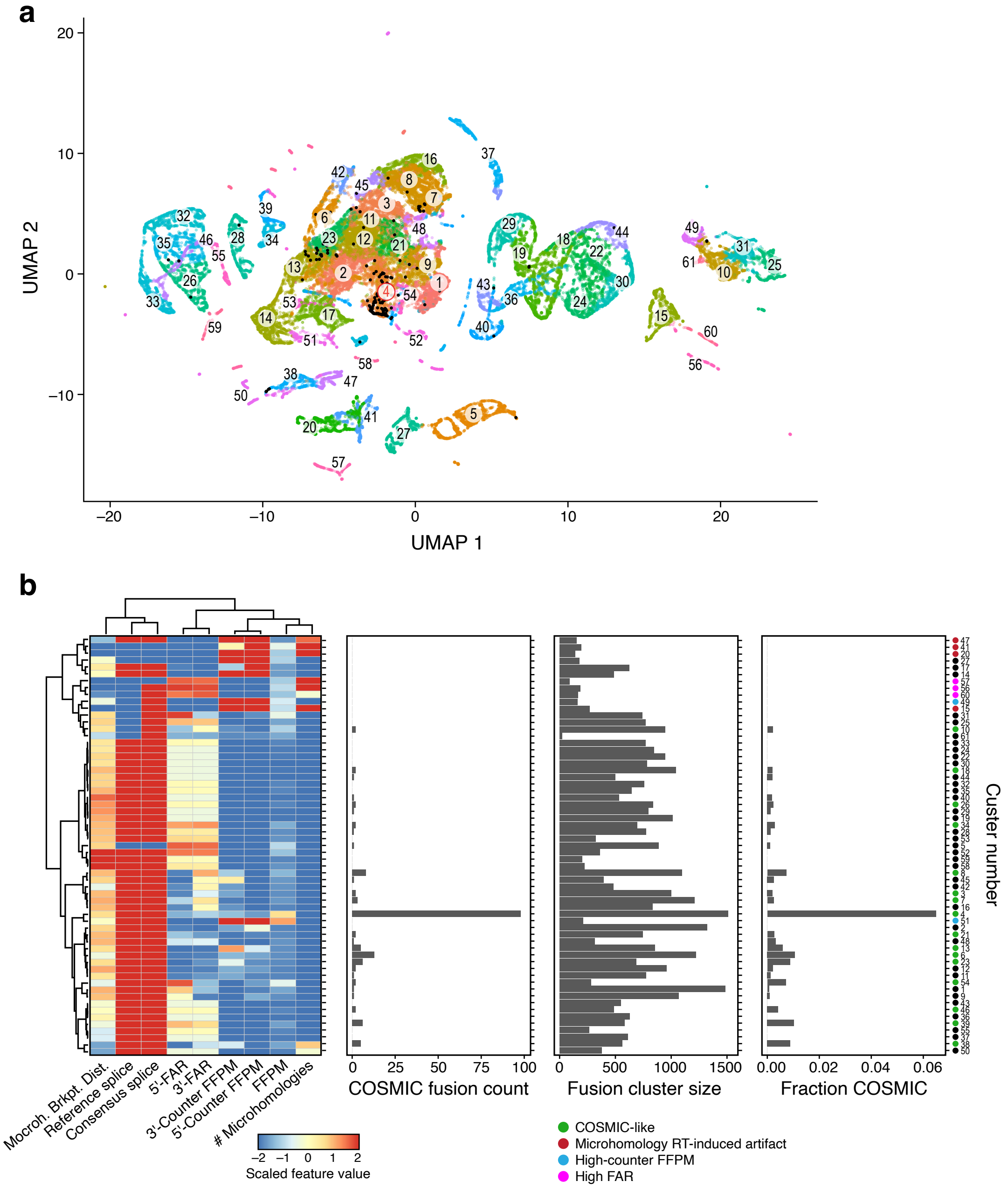

a
UMAP 2
UMAP 1
20
10
0
−10
−20
−10
0
10
20
b
Mocroh. Brkpt. Dist.
Reference splice
Consensus splice
5'-FAR
3'-FAR
3'-Counter FFPM
5'-Counter FFPM
FFPM
# Microhomologies
−2 −1 0 1 2
Scaled feature value
0 25 50 75 100
COSMIC fusion count
0 500 1000 1500
Fusion cluster size
0.00 0.02 0.04 0.06
Fraction COSMIC
Custer number
● COSMIC-like
● Microhomology RT-induced artifact
● High-counter FFPM
● High FAR

Figure 5

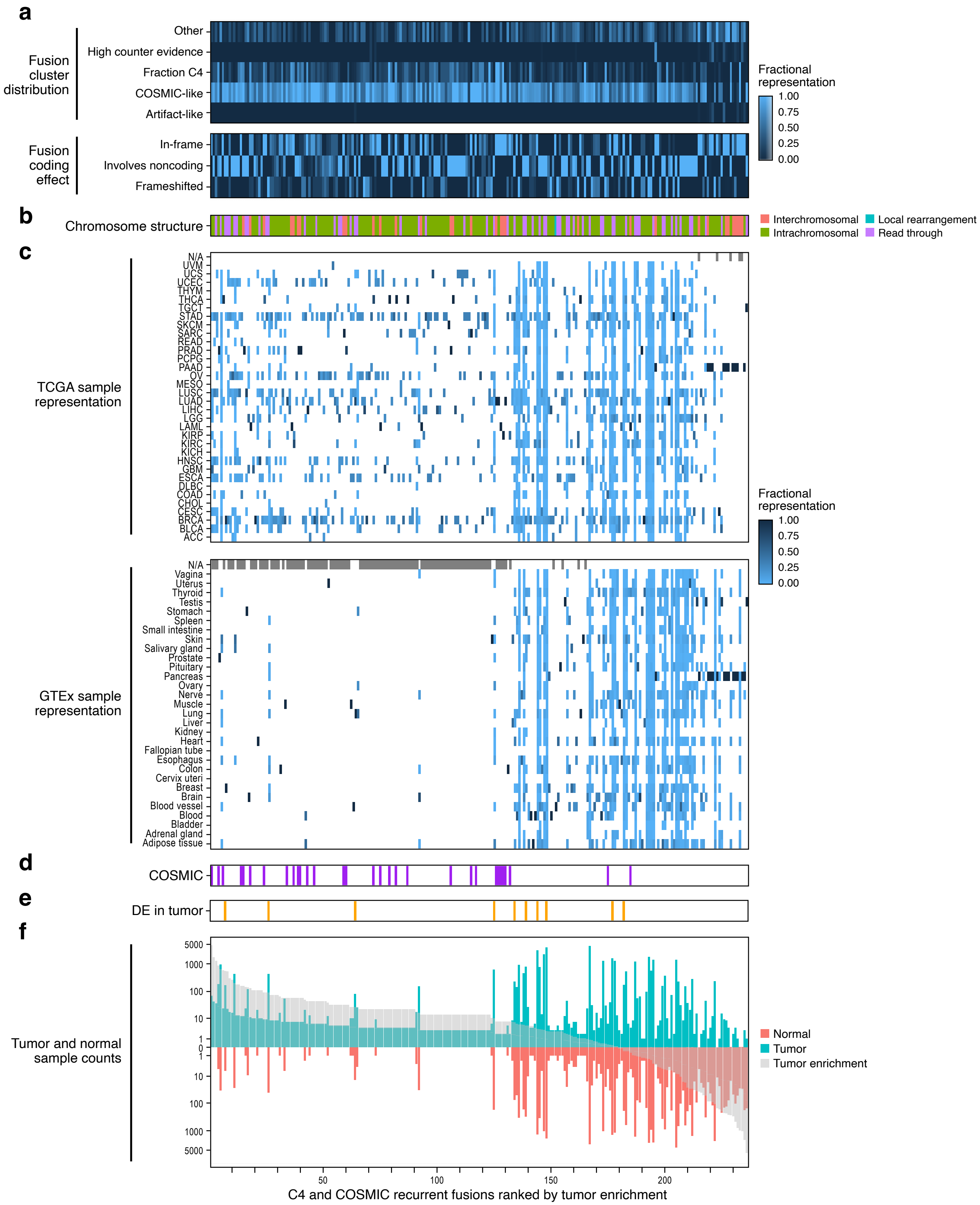
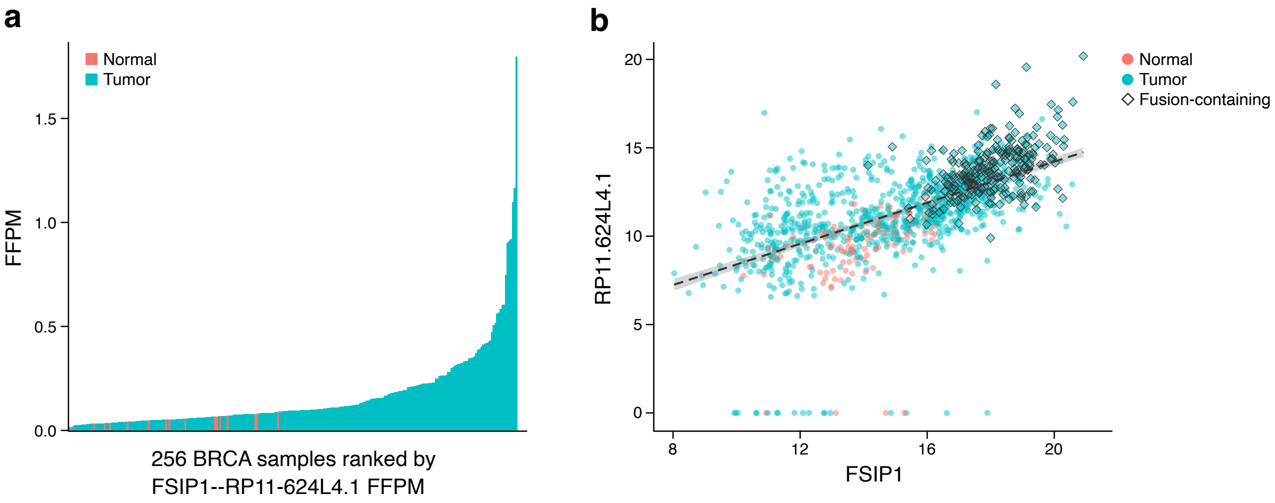

a
Fusion cluster distribution
Other
High counter evidence
Fraction C4
COSMIC-like
Artifact-like
Fusion coding effect
In-frame
Involves noncoding
Frameshifted
Fractional representation
1.00
0.75
0.50
0.25
0.00
b
Chromosome structure
Interchromosomal
Local rearrangement
Intrachromosomal
Read through
c
TCGA sample representation
N/A
UVM
UCS
UCEC
THYM
THCA
TGCT
STAD
SKCM
SARC
READ
PRAD
PCPG
PAAD
OV
MESO
LUSC
LUAD
LIHC
LGG
LAML
KIRP
KIRC
KICH
HNSC
GBM
ESCA
DLBC
COAD
CHOL
CESC
BRCA
BLCA
ACC
Fractional representation
1.00
0.75
0.50
0.25
0.00
GTEx sample representation
N/A
Vagina
Uterus
Thyroid
Testis
Stomach
Spleen
Small intestine
Skin
Salivary gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian tube
Esophagus
Colon
Cervix uteri
Breast
Brain
Blood vessel
Blood
Bladder
Adrenal gland
Adipose tissue
d
COSMIC
e
DE in tumor
f
Tumor and normal sample counts
5000
1000
100
10
1
0
1
10
100
1000
5000
Normal
Tumor
Tumor enrichment
50
100
150
200
C4 and COSMIC recurrent fusions ranked by tumor enrichment

Figure 6


a
Normal
Tumor
FFPM
0.0
0.5
1.0
1.5
256 BRCA samples ranked by
FSIP1--RP11-624L4.1 FFPM
b
Normal
Tumor
Fusion-containing
RP11.624L4.1
0
5
10
15
20
FSIP1
8
12
16
20