Abstract #3173

# High-throughput, systematic analysis of paired-end next-generation sequencing data to characterize the gene fusion landscape in cancer

Seth E. Sadis, Nickolay A. Khazanov, Armand Bankhead III, Dinesh Cyanam, Paul D. Williams, Sean F. Eddy, Peter J. Wyngaard, Daniel R. Rhodes
Compendia Bioscience, part of Life Technologies, Ann Arbor, MI, United States

## Abstract:

Gene fusions encode oncogenic drivers in hematological and solid tumors and are often associated with dramatic clinical responses with the appropriate targeted agents. In principle, massively parallel paired-end sequencing can identify structural rearrangements in tumor genomes and transcriptomes. However, computational methods to identify gene fusions are varied, still evolving and largely trained on cell line data. We sought to develop systematic methods to characterize known oncogenic gene fusions and to discover novel gene fusions in cancer. RNASeq data for approximately 3,400 clinical cases from 16 cancer types was obtained from the Cancer Genomics Hub (CGHub) of The Cancer Genome Atlas (TCGA). We surveyed the performance of several gene fusion callers and chose two (deFuse and TopHat) for further method development. An analysis pipeline was developed and executed in parallel on a high-performance computing cluster. Filtering and annotation was conducted on the aggregated data as a post-processing step, to enable exploratory analyses of various filters. We optimized filtering approaches on datasets that included known standards (e.g., TMPRSS2-ERG in prostate adenocarcinoma, PML-RARA in acute myeloid leukemia, etc.) to enrich for these and other gene fusions with correct 5'-3' orientation while excluding cases with ambiguous breakpoints and spanning reads, alignment errors, and read-through transcripts from adjacent genes. Predicted fusions were summarized based on the occurrence of unique genes participating in fusions with multiple partners and of unique gene pairs, each within specific diseases. Elevated expression was observed after the predicted breakpoint of the 3' gene in cases positive for predicted fusions, and added important confirmatory evidence. Thus, we characterized the incidence and distribution of several known oncogenic gene fusions including EML4-ALK and CCDC6-RET while expanding the number of gene partners identified in combination with oncogenes such as ROS1. In addition to characterizing the incidence and distribution of 31 known gene fusions, we nominated over 100 novel gene fusion pairs. One example of a novel gene fusion susceptible to available targeted therapy was FGFR3-TACC3 in 4% of bladder cancer, 2% of squamous cell lung carcinoma, and 1% each of glioblastoma and head and neck squamous cell carcinoma. Computational methods are now poised to complement biochemical approaches in the definition of the gene fusion landscape in cancer.

## Materials and Methods:

We selected two well-cited, state-of-the-art gene fusion calling packages:

- **deFuse** – developed at the Shah lab by McPherson and colleagues, with continued development by the Shah lab

  McPherson *et al.* "deFuse: an algorithm for gene fusion discovery in tumor RNASeq data" *PLoS Comp. Bio.* 2011

- **TopHat** – developed at the Salzberg lab by Kim and colleagues and currently a collaborative effort between research groups at Johns Hopkins, UC Berkley and Harvard

  Kim *et al.* "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts" *Genome Biology* 2011

With the goal of supporting both single and paired end data, we processed all single-end data using only TopHat and all paired-end data using only deFuse. TopHat has been shown to be effective with longer 75bp single-end data. Conversely, the deFuse algorithm is not compatible with single-end data and has been designed to leverage read pairs.

"Pre-processing Data" and "Detect Fusions: deFuse, TopHat" steps shown in Figure 2 were executed in parallel for all samples on a high-performance computing cluster. The filtering and annotation was conducted on the aggregated data as a post-processing step, to enable exploratory analyses of effects of various filters and annotation schemes. After summarizing filtering criteria to minimize false positive fusions, the list of Priority Fusions is validated with RNASeq Exon Expression data.

### TCGA Disease Type Acronyms

**BLCA**=bladder carcinoma, **BRCA**=breast carcinoma, **CESC**=cervical squamous cell carcinoma, **COAD**=colon adenocarcinoma, **GBM**=glioblastoma multiforme, **HNSC**=head and neck squamous cell carcinoma, **KIRC**=clear cell renal cell carcinoma, **KIRP**=kidney renal papillary cell carcinoma, **LAML**=acute myeloid leukemia, **LGG**=Brain Lower Grade Glioma, **LIHC**=Liver hepatocellular carcinoma, **LUAD**=lung adenocarcinoma, **LUSC**=squamous cell lung carcinoma, **OV**=ovarian serous adenocarcinoma, **PRAD**=prostate adenocarcinoma, **READ**=rectal adenocarcinoma, **SKCM**=cutaneous melanoma, **STAD**=stomach adenocarcinoma, **THCA**=thyroid carcinoma, **UCEC**=uterine corpus endometrioid carcinoma

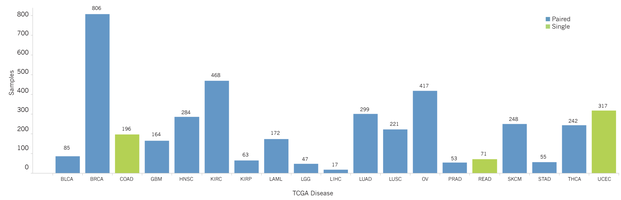
## Summary Points:

- **4,225 cancer patient samples across 19 diseases were processed with deFuse and TopHat gene fusion calling software using a cloud-based computation infrastructure.**
- **Compendia Bioscience identified filtering criteria for gene fusion events that enrich for high confidence, chemically validated gene fusion events.**
- **Pan-disease fusions (e.g. FGFR|TACC3, CCDC6|RET) and multi-partner fusion events (e.g. ERC1|RET, CCDC6|RET) broaden the clinical population scope of gene fusion events.**

## Results:

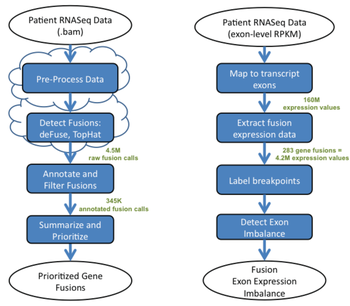
### Figure 1. 4,225 samples processed across 19 diseases

Samples were processed with Defuse and Tophat gene fusion calling software using cloud-based computing.




### Figure 2. Gene Fusion Processing Workflows

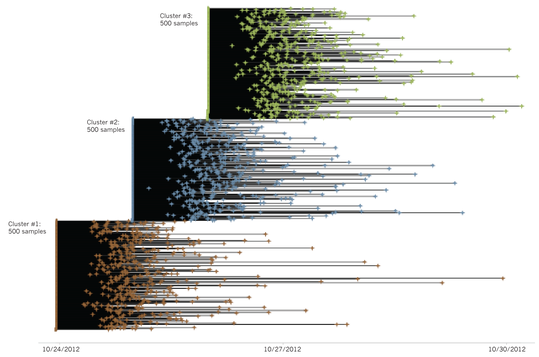
Compendia Bioscience's gene fusion processing produced 4.5 million calls that were filtered and prioritized to generate a list of high confidence "Priority" fusion calls.




| Unique Priority Fusions | Unique Priority Mitelman Fusions | |
|---|---|---|
| 7,123 | 43 | Priority Filtered |
| 283 | 17 | Recurrent |
| 20 | 7 | Exon Expression Imbalance |

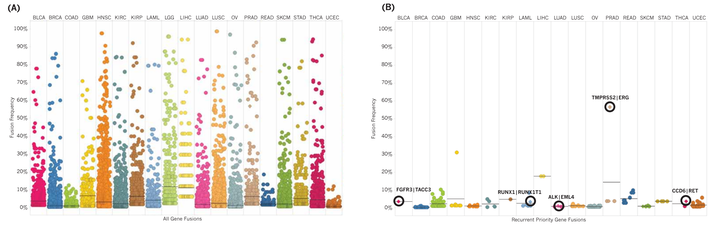
### Figure 3. Gene Fusion Detection Using Cluster Computing: 4.65 compute years in 6 days

Three 500 node clusters were used to process samples using deFuse caller software. In total 4,225 RNASeq samples were processed in 2012 by Compendia Bioscience generating 28.2 TB of fusion results data.




### Figure 4. Filtering Fusion Calls

Before priority filtering, fusion calls occur at a higher frequency across patient populations **(A)**. However, previously validated, high-confidence gene fusions such as TMPRSS|ERG and ALK|EML4, are observed in recurrent priority fusions **(B)**. Diamonds represent recurrent fusions with observed exon expression.




### Figure 5. TMRSS2|ERG Observed in 57% of Prostate Samples

Filtering criteria were developed by characterizing known high confidence fusions, such as TMPRSS2|ERG observed in 57% of prostate cancer samples. Figure 6 **(A)** shows defuse breakpoint calls for each of the 23 fusion positive patients. A fused gene product exon map is shown in Figure 6 **(B)** with green blocks pertaining to exons upstream of the breakpoint for the 5' partner and with red blocks pertaining to exons downstream of the 3' partner.

**(A)**

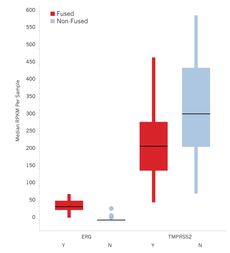
| TCGA Tumor Sample Barcode | GeneA Gene Symbol | GeneA Chromosome | GeneA Breakpoint | GeneB Gene Symbol | GeneB Chromosome | GeneB Breakpoint | Spanning Reads | Fusion Exon |
|---|---|---|---|---|---|---|---|---|
| TCGA-CH-5739-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,817,544 | 30 | T2-E4 |
| TCGA-CH-5740-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 59 | T3-E4 |
| TCGA-CH-5741-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 198 | T3-E4 |
| TCGA-CH-5746-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,817,544 | 10 | T2-E4 |
| TCGA-CH-5765-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,817,544 | 9 | T2-E4 |
| TCGA-CH-5768-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,956,869 | 17 | T2-E1 |
| TCGA-CH-5769-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 42 | T3-E4 |
| TCGA-CH-5789-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,866,283 | ERG | chr21 | 39,817,544 | 54 | T4-E4 |
| TCGA-CH-5790-01A-11R-1580-07 | TMPRSS2 | chr21 | 42,879,891 | ERG | chr21 | 39,817,518 | 23 | T2-E2 |
| TCGA-EJ-5496-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,879,884 | ERG | chr21 | 39,870,288 | 23 | T2-E4 |
| TCGA-EJ-5497-01A-02R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 52 | T3-E4 |
| TCGA-EJ-5499-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 75 | T3-E4 |
| TCGA-EJ-5507-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,880,008 | ERG | chr21 | 39,817,544 | 5 | T1-E4 |
| TCGA-EJ-5508-01A-02R-1580-07 | TMPRSS2 | chr21 | 42,880,008 | ERG | chr21 | 39,817,544 | 8 | T1-E4 |
| TCGA-EJ-5516-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,880,008 | ERG | chr21 | 39,817,544 | 5 | T1-E4 |
| TCGA-EJ-5522-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,956,869 | 25 | T2-E2 |
| TCGA-EJ-5524-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,879,891 | ERG | chr21 | 39,817,518 | 30 | T2-E4 |
| TCGA-EJ-5525-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 61 | T3-E4 |
| TCGA-EJ-5526-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 188 | T3-E4 |
| TCGA-EJ-5542-01A-01R-1580-07 | TMPRSS2 | chr21 | 42,866,406 | ERG | chr21 | 39,817,479 | 6 | T4-E4 |
| TCGA-G9-6332-01A-11R-1789-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 214 | T3-E4 |
| TCGA-G9-6365-01A-11R-1789-07 | TMPRSS2 | chr21 | 42,870,046 | ERG | chr21 | 39,817,544 | 15 | T3-E4 |
| TCGA-G9-6384-01A-11R-1789-07 | TMPRSS2 | chr21 | 42,879,877 | ERG | chr21 | 39,817,544 | 33 | T2-E4 |

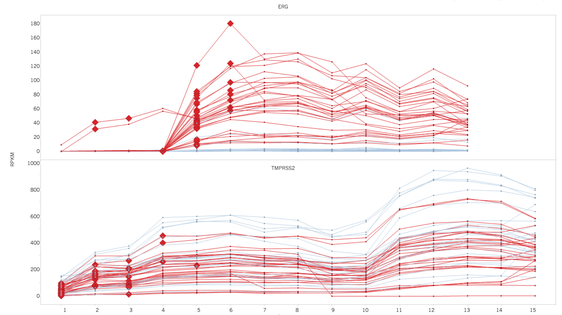
**(B)**


### Figure 6. Exon Expression Imbalance

Fused samples exhibited exon expression imbalance prior to and after predicted fusion breakpoints. 3' partner genes of fused samples had elevated expression compared to non-fused samples.

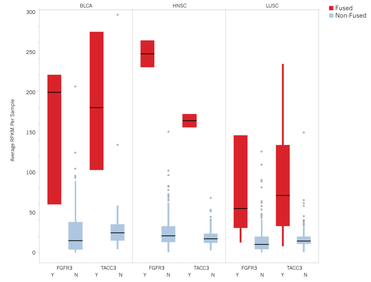
Recurrent CBI Priority Fusions were visualized using RNASeq exon expression data downloaded using the GDAC Firehose tool to provide secondary evidence of true positive fusion events by searching for exon expression imbalance before and after the breakpoint call. Specifically, if the 3' partner's expression is impacted by the 5' partner's promoter region, then exon expression should increase post the predicted breakpoint. This effect is especially visible when viewing fused versus non-fused patient samples.




### Figure 7. Up-regulation of FGFR and TACC3 in 9 fused samples across bladder, head and neck, lung squamous diseases

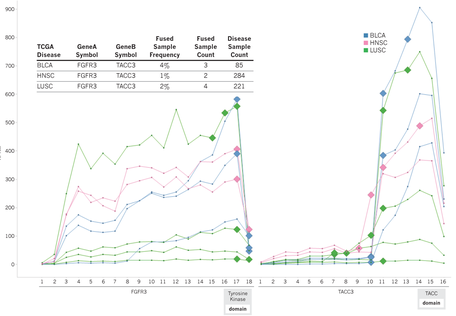
Increased expression for both fusion partners was observed in fused samples versus non-fused samples.




### Figure 8. Exon Expression Imbalance at TACC3 Exon 10 observed across 3 diseases

FGFR|TACC3 fusions were observed across bladder, head and neck, and lung squamous cancer samples with similar breakpoint mapping and exon expression imbalance in balance in both fusion partners.

| TCGA Disease | GeneA Symbol | GeneB Symbol | Fused Sample Frequency | Fused Sample Count | Disease Sample Count |
|---|---|---|---|---|---|
| BLCA | FGFR3 | TACC3 | 4% | 3 | 85 |
| HNSC | FGFR3 | TACC3 | 1% | 2 | 284 |
| LUSC | FGFR3 | TACC3 | 2% | 4 | 221 |




### Figure 9. Common Exon Expression Imbalance Prior to RET Tyrosine Kinase Domain

Gene fusions involving the RET were observed with multiple partners (e.g. CCDC6,ERC1) and across multiple diseases. Interestingly, breakpoints were observed at similar locations within RET and CCD6C6 fusion partners.

| TCGA Disease | GeneA Symbol | GeneB Symbol | Fused Sample Frequency | Fused Sample Count | Disease Sample Count |
|---|---|---|---|---|---|
| LUAD | CCDC6 | RET | 0% | 1 | 299 |
| THCA | CCDC6 | RET | 4% | 9 | 242 |
| BRCA | ERC1 | RET | 0% | 1 | 809 |