Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Biochimiche e Biotecnologiche

Ciclo XXIX

**Settore Concorsuale di afferenza: 05/E1**

**Settore Scientifico disciplinare: BIO/10**

# Highly sensitive and specific method for detection of clinically relevant fusion genes across cancer

**Presentata da:** **Dr. Fabio Fuligni**

| **Coordinatore Dottorato** | **Relatore** |
|---|---|
| **Prof. Santi Mario Spampinato** | **Prof.ssa Rita Casadio** |

**Esame finale anno 2017**

# Table of contents

# Chapter 1: Thesis Objective

## 1.1 KiCS project

The Kids Cancer Sequencing program (KiCS) is a translational research program established at The Hospital for Sick Children (SickKids) in Toronto, Canada. The program aims are to improve the diagnosis and therapeutic options of paediatric cancer patients. The program enrolls patients with a newly diagnosed childhood tumour or who may show signs of a potential cancer susceptibility. Patients enrolled in the program have one of the following attributes: have a new primary or relapsed tumour, have a genetic predisposition for cancer, have a cancer with poor prognosis, or have a poor response to conventional cancer therapies.

These patients undergo genomic sequencing using next generation sequencing (NGS) technology and the resulting data is analyzed to better characterize the tumour in an attempt to identify a unique “fingerprint.” Information derived from NGS analysis are subsequently used to find variants relevant to cancer etiology and diagnosis, identify treatment options for each specific patient and follow the tumour’s response to treatment.

## 1.2 Research aims

The main objectives of this PhD thesis was to develop novel bioinformatics algorithms for the detection of clinically relevant variants from the RNA Sequencing (RNA-Seq) data. Then to use this highly accurate approach on tumors of childhood cancer patients, enrolled in KiCS. Finally, to create a

classification scheme that enables a non-specialist to interpret the functional consequences of each somatic fusion variant.

The project was divided into two different sub-projects, each attempting to answer an open questions in transcriptomic analysis applied to precision medicine oncology.

The first question relates to the accuracy of RNA sequencing: can the sensitivity and specificity of transcriptomic data improve such that he can replace standard molecular assays?. The production of an accurate and complete transcriptome makes RNA-Seq an ideal approach to improve the diagnosis and therapeutic treatment of cancer patients. However, even if clinicians could ‘read’ all of the transcripts of all the oncology patients on the day of their diagnosis, they would still have the massive challenge of interpreting the results. In fact, the majority of current bioinformatics tools achieve poor sensitivity and specificity for detecting non-canonical fusions or cryptic splicing events. They may also produce vast lists of putative fusions, which do not subsequently validate or are found in normal controls.

Sub-project 1 involved the development of a novel software package for detecting, filtering, validating and classifying driver oncogenic chimeric transcripts, to overcome the lack of sensitivity and specificity problem pf the existing fusion detection approaches.

The second question relates to transcriptional abundance in cancer**:** how does the transcriptional output of a cancer cell change as it acquires somatic mutations, becomes neoplastic, and ultimately metastasizes?. Transcriptional amplification, whereby the entire transcriptome of a cell increases in expression, represents the direct effect of somatic mutations in transcription and can be used in the development of novel therapeutic strategies for

aggressive tumours. However, knowledge of the tumour types driven by transcriptional amplification, as well as identification of the genes mediating this effect is relatively unknown. Currently there are no software tool to accurately measure the transcriptional output of a cancer, *in vivo*, from heterogeneous patient specimens that have undergone RNA-Seq.

In sub-project 2, a computational method to measure the transcriptional abundance of human cancer cells from primary tumours was created to catalogue the rules governing how somatic mutation exerts direct transcriptional effects. Results for this project were published on Cell Reports in 2016[1]

# Chapter 2: Fusion Validator: Highly accurate fusion gene detection from RNA sequencing of cancer

## 2.1 Introduction

### *2.1.1 Gene fusions detection as a diagnostic test in clinical oncology*

Accumulation of specific genomic aberrations like single nucleotide mutations and chromosomal structural rearrangement are a major cause of cancer development[2]. Chromosomal rearrangements, including genomic deletions, duplications, inversions and translocations can lead to the formation of a fusion of two genes, that would otherwise be physically separated. This resulting fusion is exclusively expressed in cancer cells[3] (Figure 1).

Recurrent gene fusions like BCR–ABL1 in chronic myeloid leukemia[4], EWSR1-FLI1 in Ewing's sarcoma[5], EML4-ALK in lung cancer[6] or FGFR-TACC in glioblastoma[7], are considered strong driver mutations and used as diagnostic markers or for therapeutic decision-making (Figure 2).

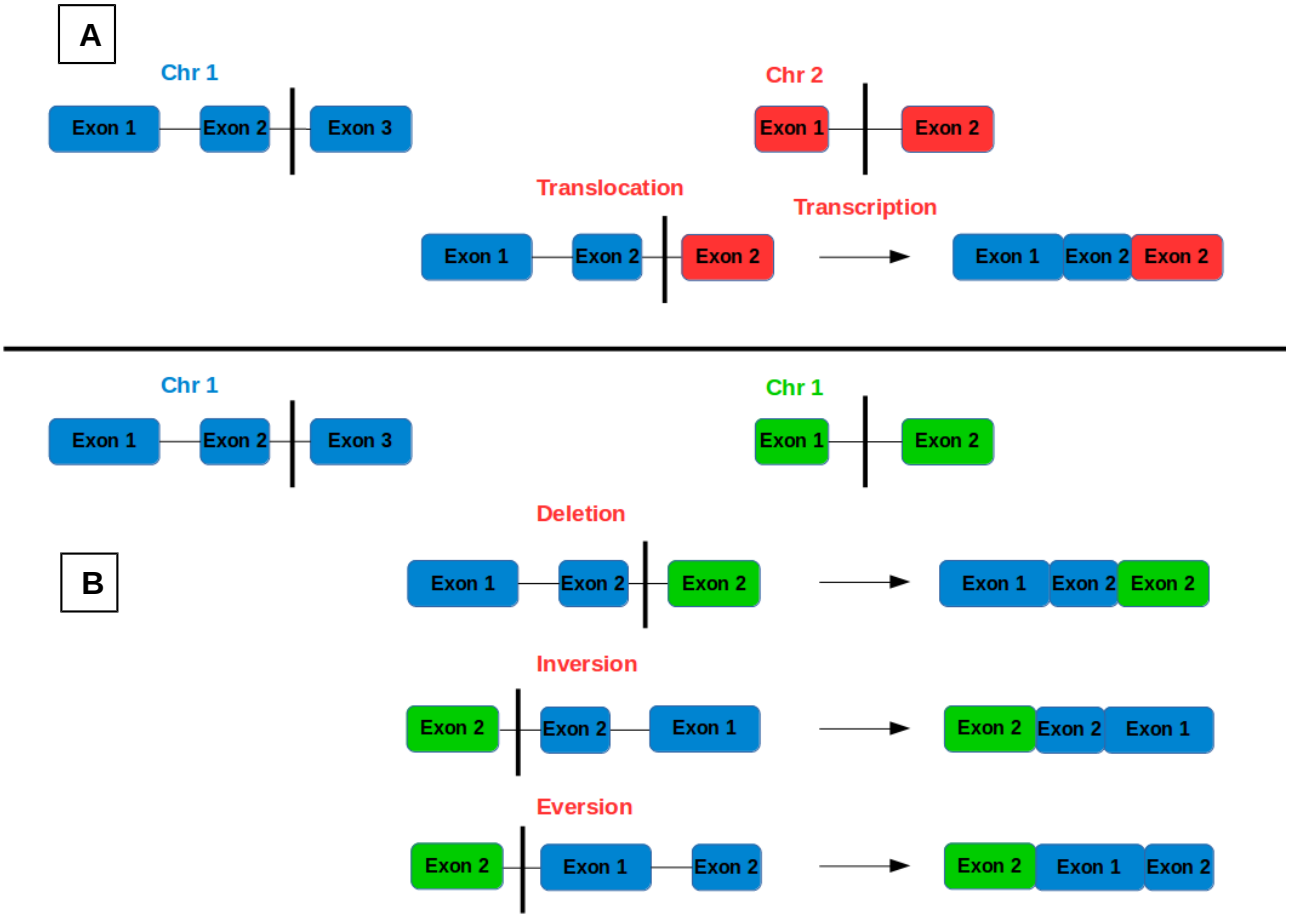


**Figure 1. Schematic representation of a gene fusion from genomic rearrangements between different chromosomes (A) or same chromosome (B).** In genomic rearrangements between different chromosomes (A) genes from different chromosomes have a translocation on the breakpoint position (vertical black line) and form a fusion gene. The product of the translocation is transcribed into a fusion transcript containing exons from gene 1 (blue) and from gene 2 (red). Rearrangement in the same chromosome (B) can be classified as deletion (when a part of chromosome is lost during replication), inversion (when a part of chromosome is reversed end to end) or eversion (when a part of chromosome is reversed end to beginning) and are transcribed into fusion transcripts containing exons from gene 1 (blue) and from gene 2 (green)

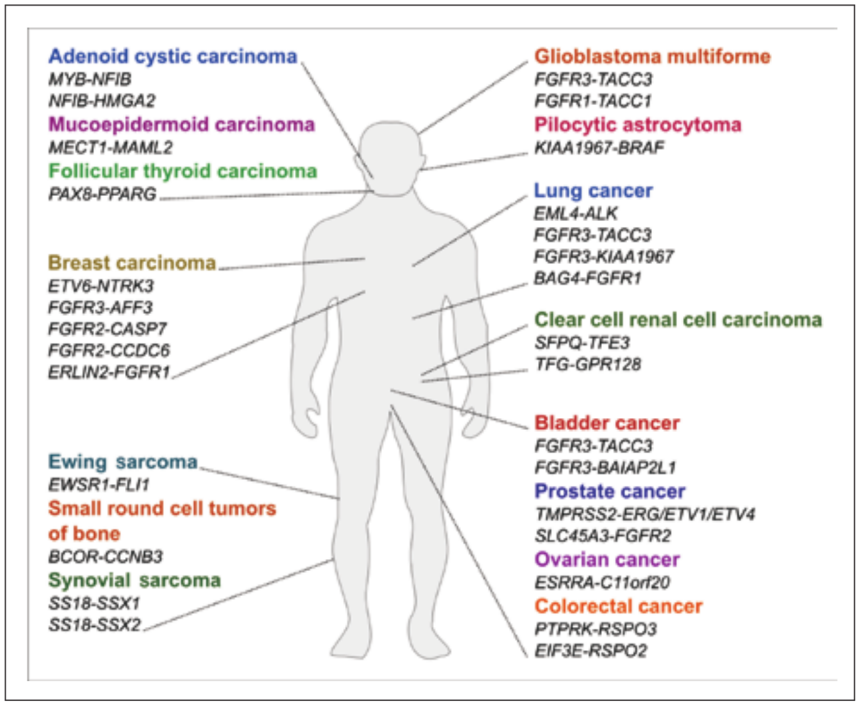


**Figure 2. List of most recurrent gene fusions in different cancer types according to Parker et al. [Parker]**

The advent of NGS platforms and the use of RNA-Seq paired-end technique in tumour studies allowed the identification of an increasing number of fusion transcripts collected in public databases[8,9] (Figure 3).

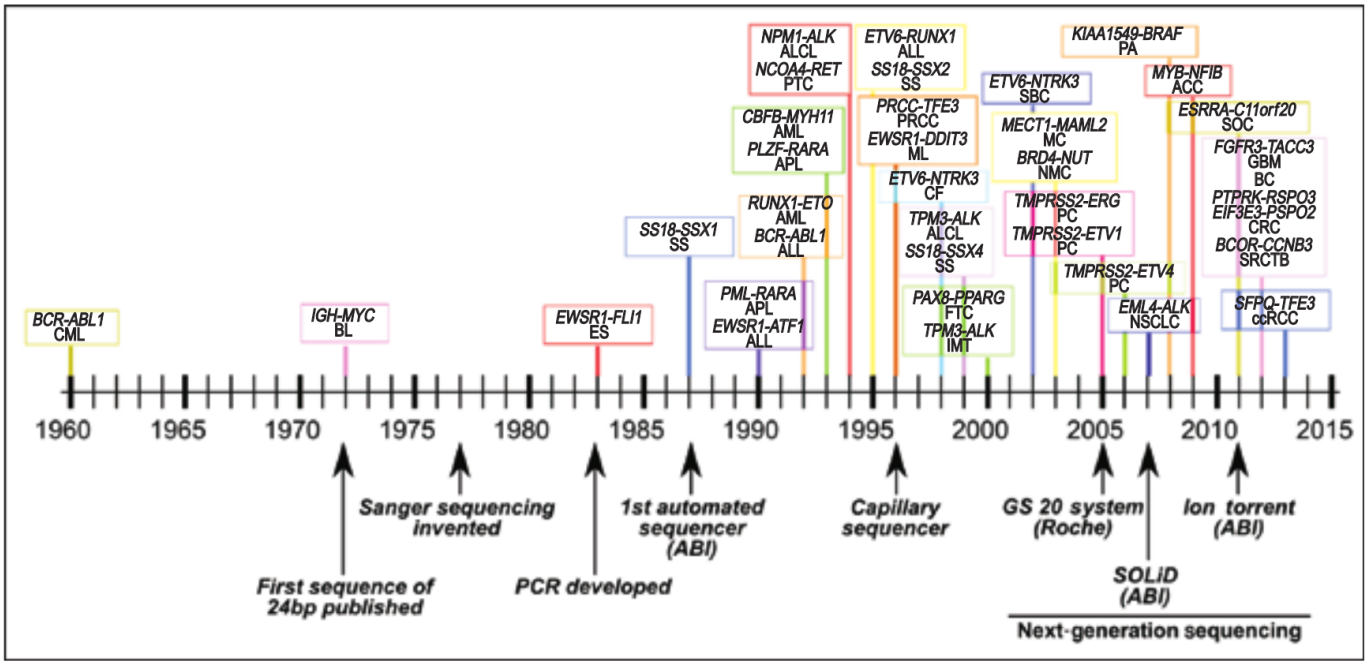


**Figure 3. Timeline showing the years in which particular driver fusions were discovered, compared to the year in which DNA sequencing technologies became available. (Figure from Parker et al. [Parker])**

The ability to characterize the entire transcriptomic profile in a precise and efficient way, at reduced cost compared to traditional techniques, and with the power to uncover novel events in a single test, makes RNA-Seq fusion detection a very suitable and attractive instrument to improve the diagnosis and therapeutic treatment of cancer patients[10]. In the last few years, different institutes have established personalized cancer medicine programs giving rise to what is called precision oncology[11]. Some institutes have included fusion transcripts in their investigation of biological driver events[12,13].

Regardless of the integration of NGS diagnostic tests within time and cost budgets, significant challenges for clinical interpretation exist. First, algorithms used to detect any targetable genomic alteration must be robust and have the ability to detect a wider range of variants with high sensitivity compared to

available methods. Several experimental design components, like sequence length, coverage and the choice of appropriate library preparation protocols, should be also taken in consideration before sequencing, to avoid missing biologically relevant events. Second, the list of putative candidate variants detected can contain false positive calls, due to technological and biological biases in NGS data, and this decrease the specificity. Correctly filtering out chimeric events is crucial to reduce the number of candidate variants to investigate. Third, the genetic variants discovered can include events not present in any database or appearing in only a single patient, may produce fusions between adjacent genes in the genome (read through), or alterations observed in normal tissues[14-15]. Therefore, a very accurate annotation and validation of results is required to best select candidate oncogenic variants[16]. Some of the approaches used to confirm RNA-Seq variants like Sanger sequencing or real time PCR assay[17] are time consuming and labor intensive, making the validation of a large number of transcriptomic event candidates not feasible. The demand of a robust analytical validation on customizable panel of gene fusion at affordable costs, and with low RNA input requirements, has led different groups to use targeted RNA deep sequencing as a NGS based diagnostic test in clinical oncology[18-19].

### *2.1.2 Different approaches for gene fusion detection in RNA-Seq Experiments*

To partially resolve gaps in RNA-Seq fusion transcripts analysis, several new software tools have been developed over the past few years. These tools differ each other by reads alignment strategies, fusion prediction algorithms, and/or filtering criteria used[16,20,21] (Table 1).

| Tool name | PMID | Citation |
|---|---|---|
| Bellerophontes | 22711792 | Abate et al. (2012) |
| BreakFusion | 22563071 | Bradeen et al. (2006) |
| BreakPointer | 22302574 | Sun et al. (2012) |
| ChildSeq-RNA | 24517889 | Qadir et al. (2014) |
| ChimeraScan | 21840877 | Iyer et al. (2011) |
| Comrad | 21478487 | McPherson et al. (2011b) |
| defuse | 21625565 | McPherson et al. (2011a) |
| Dissect | 22689759 | Yorukoglu et al. (2012) |
| EBARDenovo | 23457040 | Chu et al. (2013) |
| EricScript | 23093608 | Benelli et al. (2012) |
| FusionAnalyser | 22570408 | Piazza et al. (2012) |
| FusionCatcher | 21247443 | Edgren et al. (2011) |
| FusionFinder | 22761941 | Francis et al. (2012) |
| FusionHunter | 21546395 | Li et al. (2011) |
| FusionMap | 21593131 | Ge et al. (2011) |
| FusionSeq | 20964841 | Sboner et al. (2010) |
| Ivy Center Fusion discovery tool | 24261984 | Shah et al. (2013) |
| LifeScope | 22496636 | Sakarya et al. (2012) |
| MapSplice | 20802226 | Wang et al. (2010) |
| NFuse | 22745232 | McPherson et al. (2012) |
| Pegasus | 25183062 | Abate et al. (2014) |
| ShortFuse | 21330288 | Kinsella et al. (2011) |
| SnowShoes-FTD | 21622959 | Asmann et al. (2011) |
| SOAPFuse | 23409703 | Jia et al. (2013) |
| TopHat-Fusion | 21835007 | Kim and Salzberg (2011) |

**Table 1. List of fusion detection bioinformatic software developed in the past years (Figure from Davare et al.[16]).**

Carrara et al.[22] classified fusion detection tools according to alignment strategies: Softwares like deFuse[23], Fusionseq[24], FusionHunter[25], Ericscript[26] and SOAPfuse[27] align paired-end reads to a reference sequence and create a set of putative fusion products using discordant alignments (Whole paired-end approach Figure 4). Other tools like MapSplice[28], FusionFinder[29] and FusionMap[30] fragment reads into smaller segments and try to find fusion candidates aligning these fragments against the reference (Direct fragmentation approach Figure 5). Another strategy that combines both paired-end and fragment alignment is used by ChimeraScan[31], Bellerophontes[32] and

Tophat-Fusion[33]. Using this two-step approach, reads are first aligned as paired-end sequences against the reference to detect putative fusion products via discordant alignments. Reads that remain unaligned after first step are then fragmented and realigned to identify junction-spanning reads of the fusion transcript[21] (Figure 6).

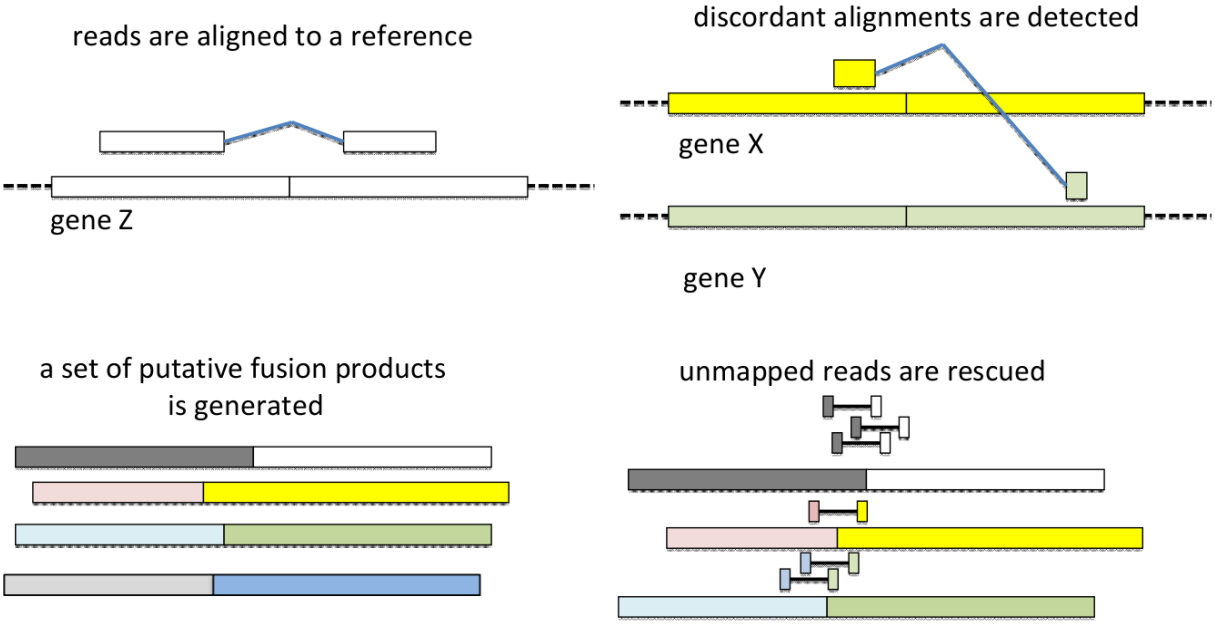


**Figure 4. Whole Paired-end fusion detection approach.**

In the whole paired-end approach, sequences are aligned to the reference genome and transcriptome. The reads, in which each mate aligns to a different gene (discordant reads), are then used to select a list of putative fusion products. All the unmapped reads from the first alignment step are then realigned locally to confirm each putative fusion product. (Figure from Raffaele Calogero's presentation "Alternative Splicing Variants and Translocation Induced Chimera detection, strength and limits of state of the art bioinformatics approaches).

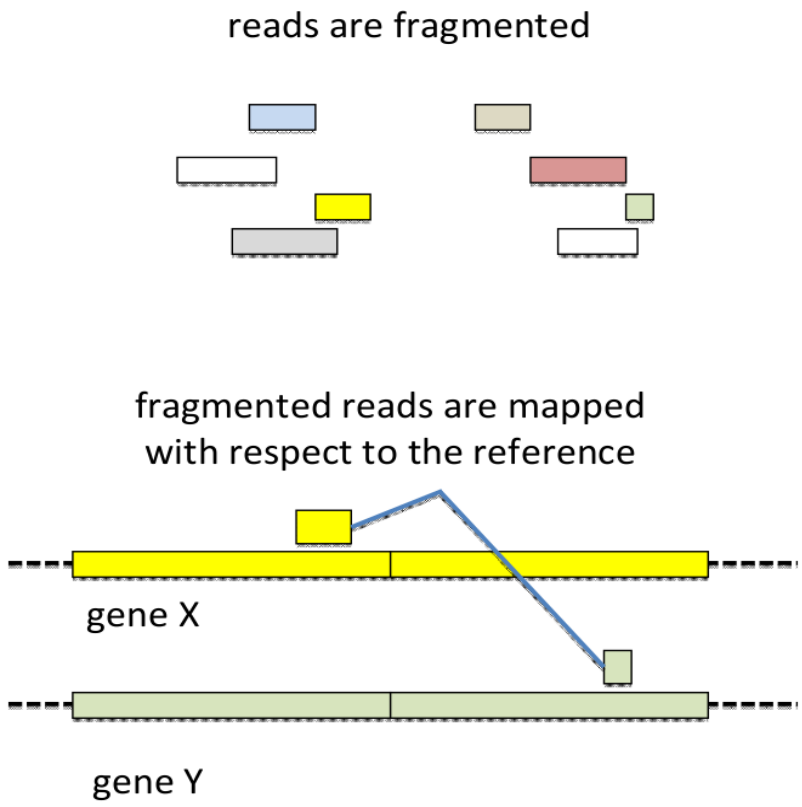


**Figure 5. Direct fragmentation fusion detection approach.**

In the direct fragmentation approach, sequences are fragmented into small segments of a user defined size and then aligned to the reference. Discordant aligned mate pairs are then used to find potential candidate fusions. (Figure from Raffaele Calogero's presentation "Alternative Splicing Variants and Translocation Induced Chimera detection, strength and limits of state of the art bioinformatics approaches).

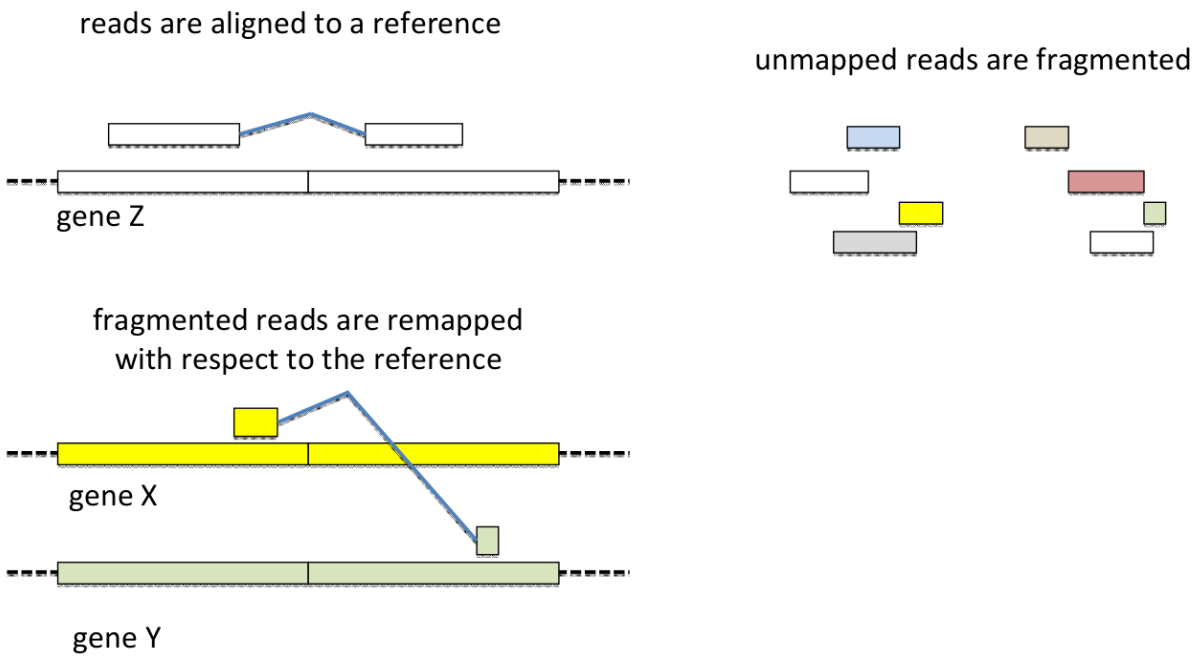


**Figure 6. Paired-end + fragmentation fusion detection approach.**

In the paired-ends + fragmentation approach the reads are first aligned in paired-ends against the reference to detect putative fusion products via discordant alignments. Then sequences unaligned at first step are fragmented and realigned to identify junction-spanning reads of the fusion transcript. (Figure from Raffaele Calogero's presentation "Alternative Splicing Variants and Translocation Induced Chimera detection, strength and limits of state of the art bioinformatics approaches).

Performance evaluation of different computational methods for gene fusion discovery, on both real and simulated datasets, revealed a consistently high number of false positive events and very little overlap between different tools[22-34].

Due to heterogeneous results and different sensitivity in identifying chimeras, a combination of various fusion finder tools was suggested to compensate individual tool errors and correctly detect driver fusions in cancer patients[35].
This combination can lead to the identification of thousands of different fusion transcripts, most of them false positives, that reduce the specificity and increase the complexity of downstream analysis and experimental validation.
The false positive rate can be partially reduced using filtering steps implemented on fusion finder algorithms, but the number of events that require a biological confirmation still remains high. Most common filtering options among different fusion detection tools includes the removal of:

- read-through transcripts
- candidate fusion mapping on homologous and repetitive regions
- PCR artifacts
- transcripts supported by poor read quality, read pair distance, number of spanning-junctions supporting sequences and number of nucleotides overlapping each side of the fusions breakpoint.

Other filtering steps are software-specific, like the comparison between chimeric transcript expression and the corresponding genes expression, used by FusionSeq, or the option to select only canonical or semi-canonical junctions, used by MapSplice. Many fusion detection programs also accept user provided blacklists of gene fusions found in normal tissues, or from existing fusion transcript databases. However, this filtering procedure can introduce additional errors, since the detection of gene fusions in normal datasets may encounter the same software-dependent sensitivity and specificity biases found in tumour samples[36].
Moreover, only few bioinformatics softwares can currently annotate, filter and prioritize fusion transcripts detected by multiple algorithms. The Chimera R package from Beccuti et al.[37] can manipulate outputs from 12 different fusion detection softwares and includes a breakpoint validation feature through de

novo assembly of reads. However, Chimera encounters issues in summarizing identical fusion genes picked by different fusion finder algorithms, especially when fusions involve genes overlapping each other on the opposite strands or results from tools that use different transcriptome annotations are compared.

Fusion Matcher (FuMa)[38] was designed to improve Chimera's functionality by comparing and matching fusion genes coming from different fusion finder algorithms, and using a unique and consistent annotation to easily summarize identical fusions. Other fusion annotation tools try to predict oncogenic potential of fusion genes using machine learning algorithms. Oncofuse[39] for example, uses a naive Bayes Network classifier, trained on features present in known oncogenic fusions, to predict the probability of a novel chimera to be classified as a driver fusion. This machine learning classifier takes into consideration protein domains maintained in fusion transcripts, but ignores interactions between functional protein domains. For this reason, recently Abate et al. developed Pegasus[40], a functional annotation software that identifies the fusion's reading frame and conserved/lost protein domains for each reconstructed chimeric transcript sequence. Pegasus then uses this information to train a classifier based on a gradient tree boosting algorithm, in order to predict fusion oncogenic potential. At present Pegasus takes in input gene fusion candidates from 3 different fusion finder algorithms, requiring the users to format results from other fusion detection tools into a common general format.

### *2.1.3 Project aim*

The aim of sub-project 1 is to create a bioinformatics pipeline to address the major challenges in the clinical validation of fusion transcripts from NGS experiments. This will be achieved by increasing the performance of driver fusions detection and by significantly reducing the number of false positives. To address these challenges, we developed Fusion Validator, a tool able to scan and filter a multitude of fusion genes from different fusion finder algorithms, and to validate real events through chimeric transcript sequence reconstruction and local realignment of candidate reads around fusion breakpoint.

Fusion Validator's main features are unique and yield improved results over current methods. First, the pipeline is completely algorithm agnostic, as it accepts input lists of chimeric transcripts detected by most of currently available fusion detection tools, and converts the results into a generic file format for further processing. The user then has the opportunity to select the most suitable combination of programs to use for optimal sensitivity.
Second, Fusion Validator uses the input list of chimeric transcripts to reconstruct the sequence spanning the fusion between two different genes or between two internally rearranged genes. It is designed to work with canonical fusions as well as other somatic structural changes in the transcriptome, like exon skips. Fusion Validator is also able to remove recurrent transcripts that are found in normal transcriptomes, through dynamic alignment of normal sequences around the breakpoint of aberrant transcripts. This procedure gives the user the opportunity to select a list of thousands of RNA-Seq experiments on a large number of diverse human tissues from the Genotype-Tissue Expression (GTEx) project[41] or GEUVADIS[42], and efficiently remove recurrent

false positive events, directly comparing putative junctions against a multi-tissue dataset, instead of processing the tumour and normals separately.

The dynamic local realignment approach is also used to validate the breakpoints of the chimeric transcripts and additional filtering steps are performed to significantly reduce the number of fusion candidates and increase the software's specificity. Lastly, Fusion Validator is able to annotate and assign a score to each chimeric transcript, empowering the user with the ability to rank the final validated list of fusions and quickly distinguish driver fusions for further investigation. The Fusion Validator workflow is completely automated and allows the user to merge, filter and validate thousands of fusions from different detection tools, without any additional work and in a significantly reduced computational time, using a high performance computer cluster.

## 2.2 Material and methods

### *2.2.1 Simulated datasets*

EricScript Simulator tool (Eric Script 0.5.4) was used to simulate 1000 synthetic gene fusions with breakpoints randomly chosen among all known splicing sites of involved genes (Intact exons (IE)), and the same 1000 fusion events with breakpoints randomly chosen without taking in consideration splicing sites (Broken exons (BE)). For each dataset of BE and IE fusions, approximately 13 million synthetic 125 base pair (bp) paired-end supporting reads were created. The average insert size for simulated reads was 400bp, with 50 bp standard deviation.

An additional 10 million 125bp paired-end reads were randomly generated as background noise using the BEERS simulator[43]. Synthetic reads generated with EricScript and BEERS were merged to create two starting simulated datasets: one for BE fusions and one for IE fusions.

Additional synthetic datasets containing reads of different lengths (50bp, 75bp, 100bp, 125bp) and a range of sequencing coverage (25X, 50X, 100X, 200X, 300X and 400X) were created by randomly subsampling the starting set of reads, using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and seqtk (https://github.com/lh3/seqtk). Thus creating a total of 24 different datasets with BE and 24 with IE (Figure 7) (Table 2). To ensure read quality consistency between every dataset, a Phred quality score of 25 was manually assigned to all the bases of the simulated reads generated.

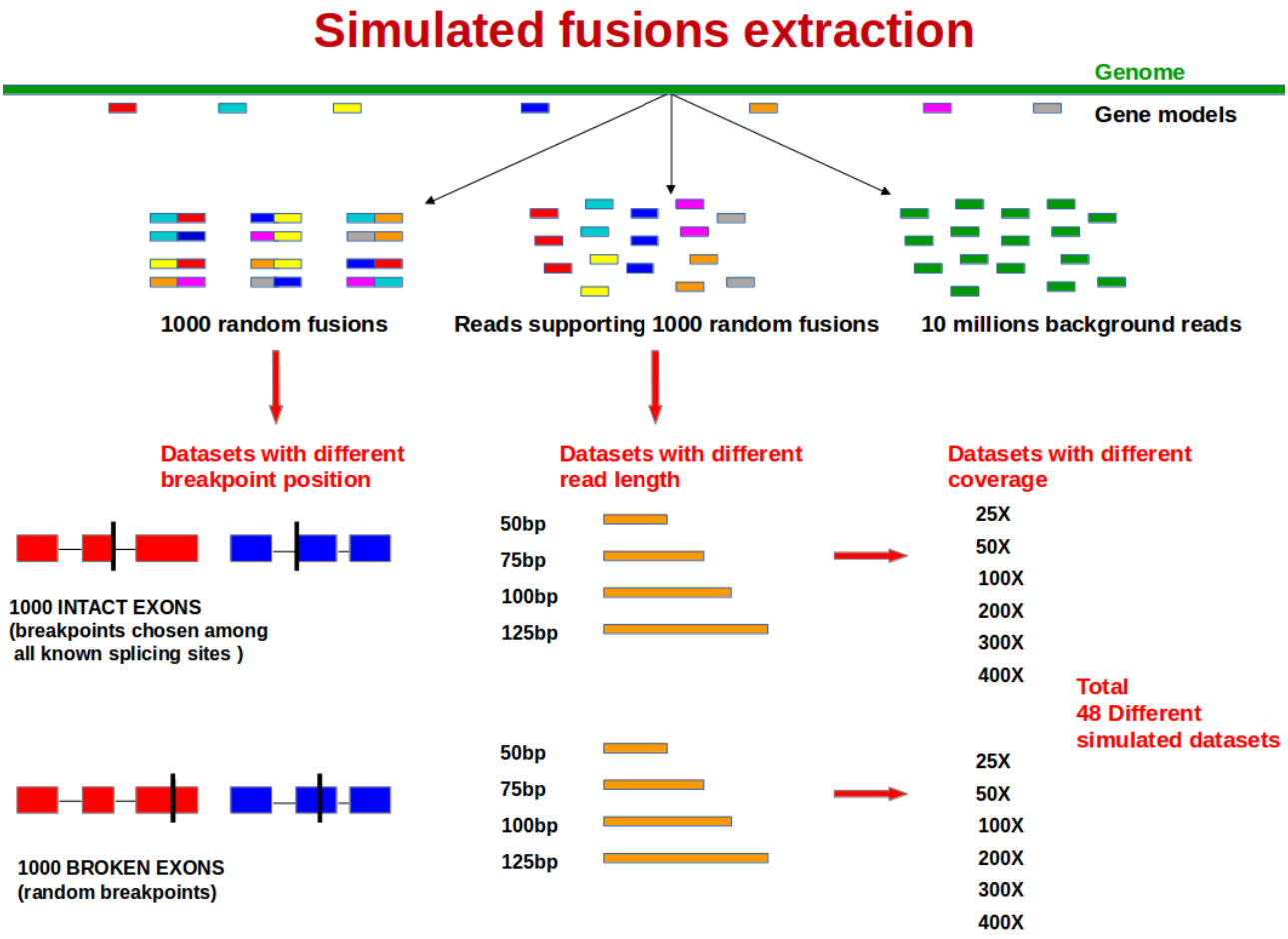


**Figure 7. Simulated fusion datasets selection steps.** 1000 random in silico chimeric transcripts were extracted from human genome and gene models, together with reads supporting each specific transcript, for a total of 48 datasets of sequences with different length (50, 75, 100, 125 base pair), sequencing coverage (25X, 50X, 100X, 200X, 300X, 400X), and breakpoint positions (on a exon junction for intact exons or random for broken exons).

| Sample | Length | Coverage | Paired-end fusion reads | Paired-end fusion read + background reads | Sample | Length | Coverage | Paired-end fusion reads | Paired-end fusion read + background reads |
|---|---|---|---|---|---|---|---|---|---|
| **BE.50bp.25X** | 50 | 25 | 726713 | 10726713 | **IE.50bp.25X** | 50 | 25 | 819372 | 10819372 |
| **BE.75bp.25X** | 75 | 25 | 484475 | 10484475 | **IE.75bp.25X** | 75 | 25 | 546248 | 10546248 |
| **BE.100bp.25X** | 100 | 25 | 363356 | 10363356 | **IE.100bp.25X** | 100 | 25 | 409686 | 10409686 |
| **BE.125bp.25X** | 125 | 25 | 290685 | 10290685 | **IE.125bp.25X** | 125 | 25 | 327749 | 10327749 |
| **BE.50bp.50X** | 50 | 50 | 1453426 | 11453426 | **IE.50bp.50X** | 50 | 50 | 1638744 | 11638744 |
| **BE.75bp.50X** | 75 | 50 | 968950 | 10968950 | **IE.75bp.50X** | 75 | 50 | 1092496 | 11092496 |
| **BE.100bp.50X** | 100 | 50 | 726713 | 10726713 | **IE.100bp.50X** | 100 | 50 | 819372 | 10819372 |
| **BE.125bp.50X** | 125 | 50 | 581370 | 10581370 | **IE.125bp.50X** | 125 | 50 | 655498 | 10655498 |
| **BE.50bp.100X** | 50 | 100 | 2906851 | 12906851 | **IE.50bp.100X** | 50 | 100 | 3277488 | 13277488 |
| **BE.75bp.100X** | 75 | 100 | 1937901 | 11937901 | **IE.75bp.100X** | 75 | 100 | 2184992 | 12184992 |
| **BE.100bp.100X** | 100 | 100 | 1453426 | 11453426 | **IE.100bp.100X** | 100 | 100 | 1638744 | 11638744 |
| **BE.125bp.100X** | 125 | 100 | 1162740 | 11162740 | **IE.125bp.100X** | 125 | 100 | 1310995 | 11310995 |
| **BE.50bp.200X** | 50 | 200 | 5813702 | 15813702 | **IE.50bp.200X** | 50 | 200 | 6554976 | 16554976 |
| **BE.75bp.200X** | 75 | 200 | 3875801 | 13875801 | **IE.75bp.200X** | 75 | 200 | 4369984 | 14369984 |
| **BE.100bp.200X** | 100 | 200 | 2906851 | 12906851 | **IE.100bp.200X** | 100 | 200 | 3277488 | 13277488 |
| **BE.125bp.200X** | 125 | 200 | 2325481 | 12325481 | **IE.125bp.200X** | 125 | 200 | 2621990 | 12621990 |
| **BE.50bp.300X** | 50 | 300 | 8720553 | 18720553 | **IE.50bp.300X** | 50 | 300 | 9832464 | 19832464 |
| **BE.75bp.300X** | 75 | 300 | 5813702 | 15813702 | **IE.75bp.300X** | 75 | 300 | 6554976 | 16554976 |
| **BE.100bp.300X** | 100 | 300 | 4360277 | 14360277 | **IE.100bp.300X** | 100 | 300 | 4916232 | 14916232 |
| **BE.125bp.300X** | 125 | 300 | 3488221 | 13488221 | **IE.125bp.300X** | 125 | 300 | 3932986 | 13932986 |
| **BE.50bp.400X** | 50 | 400 | 11627404 | 21627404 | **IE.50bp.400X** | 50 | 400 | 13109952 | 23109952 |
| **BE.75bp.400X** | 75 | 400 | 7751603 | 17751603 | **IE.75bp.400X** | 75 | 400 | 8739968 | 18739968 |
| **BE.100bp.400X** | 100 | 400 | 5813702 | 15813702 | **IE.100bp.400X** | 100 | 400 | 6554976 | 16554976 |
| **BE.125bp.400X** | 125 | 400 | 4650962 | 14650962 | **IE.125bp.400X** | 125 | 400 | 5243981 | 15243981 |

**Table 2. Number of reads supporting fusions and background reads extracted for each simulated dataset.**

### *2.2.2 Breast Cancer Cell lines*

The second dataset used for validation was from RNA-Seq of 4 Breast Cancer (BRCA) cell lines (BT-474, SK-BR-3, KPL-4 and MCF-7) containing 27 gene fusions that had already been validated[44] were downloaded from NCBI Sequence Read Archive (SRA accession number SRP003186) (Table 3).

| Sample | 5’ chromosome | 5’ gene | 3’ chromosome | 3’ gene |
|---|---|---|---|---|
| KPL-4 | 9 | NOTCH1 | 9 | NUP214 |
| KPL-4 | 19 | BSG | 19 | NFIX |
| KPL-4 | 12 | PPP1R12A | 2 | SEPT10 |
| BT-474 | 17 | ACACA | 17 | STAC2 |
| BT-474 | 3 | GLB1 | 3 | CMTM7 |
| BT-474 | 20 | VAPB | 17 | IKZF3 |
| BT-474 | 20 | DIDO1 | 20 | KIAA0406 |
| BT-474 | 20 | CPNE1 | 20 | PI3 |
| BT-474 | 20 | ZMYND8 | 20 | CEP250 |
| BT-474 | 13 | LAMP1 | 13 | MCF2L |
| BT-474 | 17 | STARD3 | 20 | DOK5 |
| BT-474 | 17 | SKA2 | 17 | MYO19 |
| BT-474 | 17 | RPS6KB1 | 17 | SNF8 |
| BT-474 | 20 | RAB22A | 19 | MYO9B |
| SK-BR-3 | 3 | SUMF1 | 3 | LRRFIP2 |
| SK-BR-3 | 8 | TATDN1 | 17 | GSDMB |
| SK-BR-3 | 14 | CCDC85C | 14 | SETD3 |
| SK-BR-3 | 17 | CYTH1 | 8 | EIF3H |
| SK-BR-3 | 5 | ANKHD1 | 5 | PCDH1 |
| SK-BR-3 | 20 | NFS1 | 20 | PREX1 |
| SK-BR-3 | 20 | CSE1L | 20 | RP4-791K14.2 |
| SK-BR-3 | 17 | RARA | 8 | PKIA |
| SK-BR-3 | 8 | WDR67 | 8 | ZNF704 |
| SK-BR-3 | 20 | DHX35 | 20 | ITCH |
| MCF-7 | 20 | BCAS4 | 17 | BCAS3 |
| MCF-7 | 20 | ARFGEF2 | 20 | SULF2 |
| MCF-7 | 17 | RPS6KB1 | 17 | TMEM49 |

**Table 3. List of fusion genes validated by Edgren et al. In 4 BRCA cell lines.**

*2.2.3 The Cancer Genome Atlas pan-cancer dataset*

Finally we used additional 50bp fastq sequences from 190 pan-cancer samples from The Cancer Genome Atlas (TCGA)[45] were downloaded from the NIH Genomic Data Commons (GDC) Data Portal [http://gdc.nci.nih.gov/], along with their respective lists of 195 recurrent fusions involving kinases (115 unique) validated by Stransky et al[46] (Table 4).

| GENE1 | GENE2 | CANCER_TYPE | TCGA_ID |
|---|---|---|---|
| SLC34A2 | ROS1 | Lung adenocarcinoma | TCGA-05-4426-01A |
| R3HDM2 | PIP4K2C | Glioblastoma multiforme | TCGA-06-0174-01A |
| EGFR | SEPT14 | Glioblastoma multiforme | TCGA-06-0750-01A |
| TMEM165 | PDGFRA | Glioblastoma multiforme | TCGA-06-2559-01A |
| NFASC | NTRK1 | Glioblastoma multiforme | TCGA-06-5411-01A |
| CEP85L | ROS1 | Glioblastoma multiforme | TCGA-06-5418-01A |
| WASF2 | FGR | Ovarian serous cystadenocarcinoma | TCGA-09-2054-01A |
| MAP2K2 | INSR | Ovarian serous cystadenocarcinoma | TCGA-13-1410-01A |
| DDR1 | PAK1 | Ovarian serous cystadenocarcinoma | TCGA-13-1477-01A |
| FGFR3 | TACC3 | Lung squamous cell carcinoma | TCGA-22-4607-01A |
| BAG4 | FGFR1 | Lung squamous cell carcinoma | TCGA-22-5480-01A |
| ARHGEF18 | INSR | Ovarian serous cystadenocarcinoma | TCGA-25-2398-01A |
| FGFR3 | TACC3 | Glioblastoma multiforme | TCGA-27-1835-01A |
| EGFR | SEPT14 | Glioblastoma multiforme | TCGA-27-1837-01A |
| EGFR | SEPT14 | Glioblastoma multiforme | TCGA-28-1747-01C |
| EGFR | SEPT14 | Glioblastoma multiforme | TCGA-28-2513-01A |
| EGFR | SEPT14 | Glioblastoma multiforme | TCGA-32-5222-01A |
| ADCY9 | PRKCB | Lung squamous cell carcinoma | TCGA-33-4533-01A |
| IGF2BP3 | PRKCA | Lung squamous cell carcinoma | TCGA-33-4587-01A |
| FGFR3 | TACC3 | Lung squamous cell carcinoma | TCGA-39-5024-01A |
| TANC2 | PRKCA | Lung squamous cell carcinoma | TCGA-43-2581-01A |
| TECR | PKN1 | Lung squamous cell carcinoma | TCGA-43-7658-01A |
| CLTC | ROS1 | Lung adenocarcinoma | TCGA-44-2665-01A |
| CLTC | ROS1 | Lung adenocarcinoma | TCGA-44-2665-01B |
| CLTC | ROS1 | Lung adenocarcinoma | TCGA-44-2665-11A |
| EML4 | ALK | Lung adenocarcinoma | TCGA-50-8460-01A |
| TRIM33 | RET | Lung adenocarcinoma | TCGA-55-6543-01A |
| EZR | ROS1 | Lung adenocarcinoma | TCGA-55-6986-01A |
| TRIM24 | NTRK2 | Lung adenocarcinoma | TCGA-55-8091-01A |
| SLC34A2 | ROS1 | Lung adenocarcinoma | TCGA-62-A46Y-01A |
| CD74 | ROS1 | Lung adenocarcinoma | TCGA-64-1680-01A |
| WASF2 | FGR | Lung squamous cell carcinoma | TCGA-66-2759-01A |
| FGFR2 | CCAR2 | Lung squamous cell carcinoma | TCGA-66-2765-01A |
| FGFR3 | TACC3 | Lung squamous cell carcinoma | TCGA-66-2786-01A |
| EML4 | ALK | Lung adenocarcinoma | TCGA-67-6215-01A |
| EML4 | ALK | Lung adenocarcinoma | TCGA-67-6216-01A |
| TUBD1 | RPS6KB1 | Lung adenocarcinoma | TCGA-69-7978-01A |
| CCDC6 | RET | Lung adenocarcinoma | TCGA-75-6203-01A |
| FGFR3 | TACC3 | Glioblastoma multiforme | TCGA-76-4925-01A |
| EML4 | ALK | Lung adenocarcinoma | TCGA-78-7163-01A |
| SPNS1 | PRKCB | Lung adenocarcinoma | TCGA-83-5908-01A |
| CD74 | ROS1 | Lung adenocarcinoma | TCGA-86-8278-01A |
| EML4 | ALK | Lung adenocarcinoma | TCGA-86-A4P8-01A |
| KIF5B | MET | Lung adenocarcinoma | TCGA-93-A4JN-01A |
| CAMK2D | ANK2 | Lung squamous cell carcinoma | TCGA-98-A53A-01A |
| DDX42 | RPS6KB1 | Breast invasive carcinoma | TCGA-A1-A0SN-01A |
| TANC2 | STRADA | Breast invasive carcinoma | TCGA-A1-A0SN-01A |
| SPINT2 | PAK1 | Breast invasive carcinoma | TCGA-A1-A0SQ-01A |
| FGFR3 | TACC3 | Kidney renal papillary cell carcinoma | TCGA-A4-7287-01A |
| TECR | PKN1 | Uterine Corpus Endometrial Carcinoma | TCGA-A5-A3LP-01A |
| XRN1 | PIP4K2A | Breast invasive carcinoma | TCGA-A8-A07C-01A |
| STK24 | PIP5K1B | Breast invasive carcinoma | TCGA-AC-A5EH-01A |
| FGFR2 | CASP7 | Breast invasive carcinoma | TCGA-AN-A0AL-01A |
| ETV6 | NTRK3 | Breast invasive carcinoma | TCGA-AO-A03U-01B |
| ZNF577 | FGFR1 | Breast invasive carcinoma | TCGA-AR-A0U3-01A |
| ZNF37A | PIP5K1B | Breast invasive carcinoma | TCGA-AR-A2LL-01A |
| ERC1 | PIK3C2G | Breast invasive carcinoma | TCGA-B6-A0IG-01A |
| PAN3 | NTRK2 | Head and Neck squamous cell carcinoma | TCGA-BB-4223-01A |
| ANXA4 | PKN1 | Liver hepatocellular carcinoma | TCGA-BC-4072-10A |
| ATG7 | BRAF | Skin Cutaneous Melanoma | TCGA-BF-A5EP-01A |
| RHOT1 | FGFR1 | Breast invasive carcinoma | TCGA-BH-A18U-01A |
| KIT | PDGFRA | Breast invasive carcinoma | TCGA-BH-A1F0-01A |
| NAP1L1 | STK38L | Breast invasive carcinoma | TCGA-BH-A1FN-01A |
| ZNF791 | FGFR1 | Breast invasive carcinoma | TCGA-BH-A209-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-BJ-A0ZJ-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-BJ-A28Z-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-BJ-A2N7-11A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-BJ-A2N8-11A |
| FNDC3B | PIK3CA | Uterine Corpus Endometrial Carcinoma | TCGA-BK-A56F-01A |
| BAIAP2L1 | MET | Kidney renal papillary cell carcinoma | TCGA-BQ-7049-01A |
| FGFR2 | TACC2 | Stomach adenocarcinoma | TCGA-BR-8080-01A |
| CASZ1 | MTOR | Stomach adenocarcinoma | TCGA-BR-8483-01A |
| ERC1 | RET | Breast invasive carcinoma | TCGA-C8-A1HJ-01A |
| TBL1XR1 | PIK3CA | Breast invasive carcinoma | TCGA-C8-A26X-01A |
| STARD3 | STRADA | Breast invasive carcinoma | TCGA-C8-A275-01A |
| CPD | ERBB2 | Stomach adenocarcinoma | TCGA-CD-5799-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-CE-A13K-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-CE-A27D-01A |
| SSBP2 | NTRK1 | Thyroid carcinoma | TCGA-CE-A3MD-01A |
| TRIM27 | RET | Thyroid carcinoma | TCGA-CE-A481-10A |
| NCOA4 | RET | Thyroid carcinoma | TCGA-CE-A482-01A |
| SPECC1L | RET | Thyroid carcinoma | TCGA-CE-A485-01A |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma | TCGA-CF-A3MF-01A |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma | TCGA-CF-A3MG-01A |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma | TCGA-CF-A3MH-01A |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma | TCGA-CF-A47S-01A |
| AGGF1 | RAF1 | Prostate adenocarcinoma | TCGA-CH-5737-01A |
| KDM7A | BRAF | Prostate adenocarcinoma | TCGA-CH-5737-01A |
| ETV6 | NTRK3 | Colon adenocarcinoma | TCGA-CK-5913-01A |
| ETV6 | NTRK3 | Colon adenocarcinoma | TCGA-CK-5916-01A |
| CCDC6 | RET | Colon adenocarcinoma | TCGA-CM-4743-01A |
| LYN | NTRK3 | Head and Neck squamous cell carcinoma | TCGA-CN-6997-01A |
| FGFR3 | TACC3 | Head and Neck squamous cell carcinoma | TCGA-CR-6473-01A |
| FGFR3 | ELAVL3 | Brain Lower Grade Glioma | TCGA-CS-6186-01A |
| FGFR3 | TACC3 | Head and Neck squamous cell carcinoma | TCGA-CV-7100-01A |
| KAZN | MTOR | Uterine Corpus Endometrial Carcinoma | TCGA-D1-A3JQ-01A |
| TAX1BP1 | BRAF | Skin Cutaneous Melanoma | TCGA-D3-A2JC-06A |

| GENE1 | GENE2 | CANCER_TYPE | TCGA_ID |
|---|---|---|---|
| FGFR2 | CCDC6 | Breast invasive carcinoma | TCGA-D8-A13Z-01A |
| ERLIN2 | FGFR1 | Breast invasive carcinoma | TCGA-D8-A1JC-01A |
| MPRIP | RAF1 | Skin Cutaneous Melanoma | TCGA-D9-A4Z6-06A |
| AGK | BRAF | Skin Cutaneous Melanoma | TCGA-DA-A1IA-06A |
| MACF1 | BRAF | Thyroid carcinoma | TCGA-DE-A0Y2-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-DE-A2OL-01A |
| NCOA4 | RET | Thyroid carcinoma | TCGA-DE-A3KN-01A |
| AGK | BRAF | Thyroid carcinoma | TCGA-DJ-A2PX-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-DJ-A2Q1-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-DJ-A2Q3-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-DJ-A2Q4-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-DJ-A2Q5-01A |
| STRN | ALK | Thyroid carcinoma | TCGA-DJ-A3US-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-DJ-A3UV-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-DJ-A3V3-01A |
| IRF2BP2 | NTRK1 | Thyroid carcinoma | TCGA-DJ-A4UP-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-DJ-A4UQ-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-DJ-A4V0-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-DJ-A4V5-01A |
| MTSS1 | ERBB2 | Bladder Urothelial Carcinoma | TCGA-DK-A2I6-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-DO-A1JZ-01A |
| CAMK2D | ANK2 | Brain Lower Grade Glioma | TCGA-DU-5855-01A |
| EGFR | SEPT14 | Brain Lower Grade Glioma | TCGA-DU-6406-01A |
| WNK1 | STK38L | Brain Lower Grade Glioma | TCGA-DU-7007-01A |
| PTPRZ1 | MET | Brain Lower Grade Glioma | TCGA-DU-7304-02A |
| SQSTM1 | NTRK2 | Brain Lower Grade Glioma | TCGA-DU-A76L-10A |
| TRIO | TERT | Sarcoma | TCGA-DX-A1L3-01A |
| RAB3B | PKN2 | Sarcoma | TCGA-DX-A23U-01A |
| TUFT1 | PKN2 | Sarcoma | TCGA-DX-A23U-01A |
| TRIO | TERT | Sarcoma | TCGA-DX-A2J0-01A |
| TPM3 | NTRK1 | Sarcoma | TCGA-DX-A3UA-01A |
| PTAR1 | PIP5K1B | Sarcoma | TCGA-DX-A48N-01A |
| SRI | PIP4K2C | Sarcoma | TCGA-DX-A6BH-01A |
| PDGFRA | FIP1L1 LNX1 | Brain Lower Grade Glioma | TCGA-E1-A7YI-01A |
| TBL1XR1 | PIK3CA | Breast invasive carcinoma | TCGA-E2-A14P-01A |
| ANK1 | FGFR1 | Breast invasive carcinoma | TCGA-E2-A15A-01A |
| WHSC1L1 | FGFR1 | Breast invasive carcinoma | TCGA-E2-A15A-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-E3-A3E0-01A |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma | TCGA-E7-A5KE-10A |
| EML4 | ALK | Thyroid carcinoma | TCGA-E8-A432-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-E8-A438-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-E8-A44M-10A |
| ETV6 | NTRK3 | Skin Cutaneous Melanoma | TCGA-EB-A51B-01A |
| LMNA | RAF1 | Skin Cutaneous Melanoma | TCGA-EB-A5SF-01A |
| TRAK1 | RAF1 | Skin Cutaneous Melanoma | TCGA-EE-A2MI-06A |
| TBL1XR1 | PIK3CA | Prostate adenocarcinoma | TCGA-EJ-5507-01A |
| FGFR3 | AES | Prostate adenocarcinoma | TCGA-EJ-A7NM-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-EL-A3CY-01A |
| TPM3 | NTRK1 | Thyroid carcinoma | TCGA-EL-A3D4-10A |
| NCOA4 | RET | Thyroid carcinoma | TCGA-EL-A3H3-01A |
| AP3B1 | BRAF | Thyroid carcinoma | TCGA-EL-A3T0-01A |
| ERC1 | RET | Thyroid carcinoma | TCGA-EL-A3T9-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-EL-A3TB-01A |
| SND1 | BRAF | Thyroid carcinoma | TCGA-EL-A3ZK-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-EL-A3ZN-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-EL-A3ZP-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-EL-A3ZS-01A |
| GTF2IRD1 | ALK | Thyroid carcinoma | TCGA-EL-A4KD-01A |
| RAF1 | AGGF1 | Thyroid carcinoma | TCGA-EM-A1CS-01A |
| NCOA4 | RET | Thyroid carcinoma | TCGA-EM-A2CU-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-EM-A3AN-01A |
| TFG | NTRK1 | Thyroid carcinoma | TCGA-EM-A3AO-10A |
| ERC1 | RET | Thyroid carcinoma | TCGA-EM-A3FQ-06A |
| CLCN6 | RAF1 | Skin Cutaneous Melanoma | TCGA-ER-A19L-06A |
| WASF2 | FGR | Skin Cutaneous Melanoma | TCGA-ER-A19W-06A |
| BCL2L11 | BRAF | Thyroid carcinoma | TCGA-ET-A2MX-01A |
| RBPMS | NTRK3 | Thyroid carcinoma | TCGA-ET-A39L-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-ET-A39R-01A |
| FAM114A2 | BRAF | Thyroid carcinoma | TCGA-ET-A3BN-01A |
| FKBP15 | RET | Thyroid carcinoma | TCGA-ET-A3DQ-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-ET-A3DR-01A |
| TBL1XR1 | RET | Thyroid carcinoma | TCGA-ET-A40R-01A |
| SQSTM1 | NTRK1 | Thyroid carcinoma | TCGA-ET-A40S-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-ET-A40T-01A |
| EFNA3 | PIK3C2G | Breast invasive carcinoma | TCGA-EW-A1PC-01B |
| TRIM24 | BRAF | Rectum adenocarcinoma | TCGA-F5-6464-01A |
| SMEK2 | ALK | Rectum adenocarcinoma | TCGA-F5-6864-01A |
| ETV6 | NTRK3 | Thyroid carcinoma | TCGA-FE-A3PD-01A |
| FGFR3 | TACC3 | Brain Lower Grade Glioma | TCGA-FG-7643-01A |
| RIMKLB | PIP4K2A | Brain Lower Grade Glioma | TCGA-FG-8185-01A |
| TFG | MET | Thyroid carcinoma | TCGA-FK-A3S3-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-FK-A3SE-01A |
| AKAP13 | RET | Thyroid carcinoma | TCGA-FK-A3SG-01A |
| CDC27 | BRAF | Skin Cutaneous Melanoma | TCGA-FS-A1ZU-06A |
| SND1 | BRAF | Thyroid carcinoma | TCGA-FY-A40N-01A |
| STRN | ALK | Kidney renal papillary cell carcinoma | TCGA-G7-6792-01A |
| C8ORF34 | MET | Kidney renal papillary cell carcinoma | TCGA-GL-7773-01A |
| NARS2 | PAK1 | Skin Cutaneous Melanoma | TCGA-GN-A26D-06A |
| TPM1 | ALK | Bladder Urothelial Carcinoma | TCGA-GV-A3QG-01A |
| PAPD7 | RAF1 | Prostate adenocarcinoma | TCGA-HC-8256-01A |
| AFAP1 | NTRK2 | Brain Lower Grade Glioma | TCGA-HT-7680-01A |
| GGA2 | PRKCB | Brain Lower Grade Glioma | TCGA-HT-A5RC-01A |
| MKRN1 | BRAF | Thyroid carcinoma | TCGA-J8-A3O1-01A |
| CCDC6 | RET | Thyroid carcinoma | TCGA-J8-A4HW-01A |
| ERBB2 | PPP1R1B | Liver hepatocellular carcinoma | TCGA-KR-A7K2-01A |
| ZC3HAV1 | BRAF | Thyroid carcinoma | TCGA-KS-A4ID-01A |
| FGFR3 | TACC3 | Brain Lower Grade Glioma | TCGA-P5-A72U-01A |
| OXR1 | MET | Liver hepatocellular carcinoma | TCGA-RC-A6M6-01A |

**Table 4. List of recurrent kinase fusions validated by Stransky et al. In 190 pan cancer TCGA samples.**

### *2.2.4 Normal tissue control dataset*

We obtained control transcriptomes derived from the non-diseased tissue of healthy individuals from the NHGRI GTEx consortium (database version 4). Representative samples for each tissue and sub-tissue type were selected. Samples were ordered by RNA Integrity Number (RIN) in descending order and by the time with which samples were prepared after the patients decease in ascending order. We excluded samples for which the autolysis score was greater than 2. Up to 30 samples for each tissue type were then selected from the sorted list, for a total of 1277 samples from 43 different tissues (Table 5).

| # Samples | Tissue | Subtissue |
|---|---|---|
| 30 | Adipose | Adipose - Subcutaneous |
| 31 | Adipose | Adipose - Visceral (Omentum) |
| 30 | Adrenal | Adrenal Gland |
| 11 | Bladder | Bladder |
| 30 | Blood | Cells - EBV-transformed lymphocytes |
| 30 | Blood | Artery - Aorta |
| 30 | Blood | Artery - Tibial |
| 30 | Blood | Whole Blood |
| 43 | Blood | Artery - Coronary |
| 10 | Brain | Brain - Amygdala |
| 10 | Brain | Brain - Cerebellum |
| 10 | Brain | Brain - Cortex |
| 11 | Brain | Brain - Anterior cingulate cortex (BA24) |
| 13 | Brain | Brain - Frontal Cortex (BA9) |
| 13 | Brain | Brain - Putamen (basal ganglia) |
| 13 | Brain | Brain - Substantia nigra |
| 14 | Brain | Brain - Hippocampus |
| 14 | Brain | Brain - Spinal cord (cervical c-1) |
| 15 | Brain | Brain - Cerebellar Hemisphere |
| 17 | Brain | Brain - Hypothalamus |
| 18 | Brain | Brain - Caudate (basal ganglia) |
| 18 | Brain | Brain - Nucleus accumbens (basal ganglia) |
| 30 | Breast | Breast - Mammary Tissue |
| 4 | Cervix | Cervix - Endocervix |
| 6 | Cervix | Cervix - Ectocervix |
| 16 | Colon | Colon - Sigmoid |
| 30 | Colon | Colon - Transverse |
| 24 | Esophagus | Esophagus - Gastroesophageal Junction |
| 30 | Esophagus | Esophagus - Mucosa |
| 30 | Esophagus | Esophagus - Muscularis |
| 7 | Fallopian | Fallopian Tube |
| 30 | Heart | Heart - Left Ventricle |
| 41 | Heart | Heart - Atrial Appendage |
| 8 | Kidney | Kidney - Cortex |
| 35 | Liver | Liver |
| 30 | Lung | Lung |
| 30 | Muscle | Muscle - Skeletal |
| 30 | Nerve | Nerve - Tibial |
| 39 | Ovary | Ovary |
| 30 | Pancreas | Pancreas |
| 23 | Pituitary | Pituitary |
| 41 | Prostate | Prostate |
| 6 | Salivary | Minor Salivary Gland |
| 30 | Skin | Cells - Transformed fibroblasts |
| 30 | Skin | Skin - Sun Exposed (Lower leg) |
| 43 | Skin | Skin - Not Sun Exposed (Suprapubic) |
| 17 | Small | Small Intestine - Terminal Ileum |
| 36 | Spleen | Spleen |
| 30 | Stomach | Stomach |
| 30 | Testis | Testis |
| 30 | Thyroid | Thyroid |
| 36 | Uterus | Uterus |
| 34 | Vagina | Vagina |

**Table 5. List of GTEX samples used for Fusion Validator's normal filtering step stratified by tissue and subtissue.**

## 2.3 Analysis

### *2.3.1 Fusion detection analysis*

Fusion transcripts for both simulated dataset and BRCA cell lines test set were detected using Defuse 0.6.2, Chimerascan 0.4.5, STAR-fusion 0.7.0[47], MapSplice 2.1.9 and FusionCatcher 0.99.4d_beta[48] with default parameters and no filtering options activated. Defuse, Chimerascan and STAR-fusion were also selected to find gene fusion candidates on TCGA pan-cancer dataset. Gencode Release 19 (GRCh37.p13) was used as reference gene model for all the alignments.

### *2.3.2 Statistical analysis*

Fusion genes randomly selected in simulated dataset were validated using Fusion Validator. Positive Predicted Value (PPV), sensitivity, specificity, accuracy and F-Measure were used to assess the performance of the novel tool. PPV is defined as the proportion of true positive fusions divided by the positive calls. Sensitivity is computed as the ratio between true positives events and the sum of true positives and false negatives, while specificity correspond to the ratio between true negatives and the sum of true negatives and false positives. Accuracy represent the proportion of true calls (positives and

negatives) on the total number of calls and F Measure is calculated as the harmonic mean of PPV and Sensitivity.

## 2.4 Fusion Validator workflow

The Fusion Validator workflow consists of 4 different components (Figure 8):

Step 1) Integration of different fusion detection tools and reconstruction of chimeric transcripts.

Step 2) Removal of non-cancer fusions through local realignment of sequences from normal tissues on chimeric transcripts.

Step 3) Validation of fusion transcripts through local realignment and de novo assembly of tumour reads.

Step 4) Annotation of fusions transcripts and ranking score assignment.

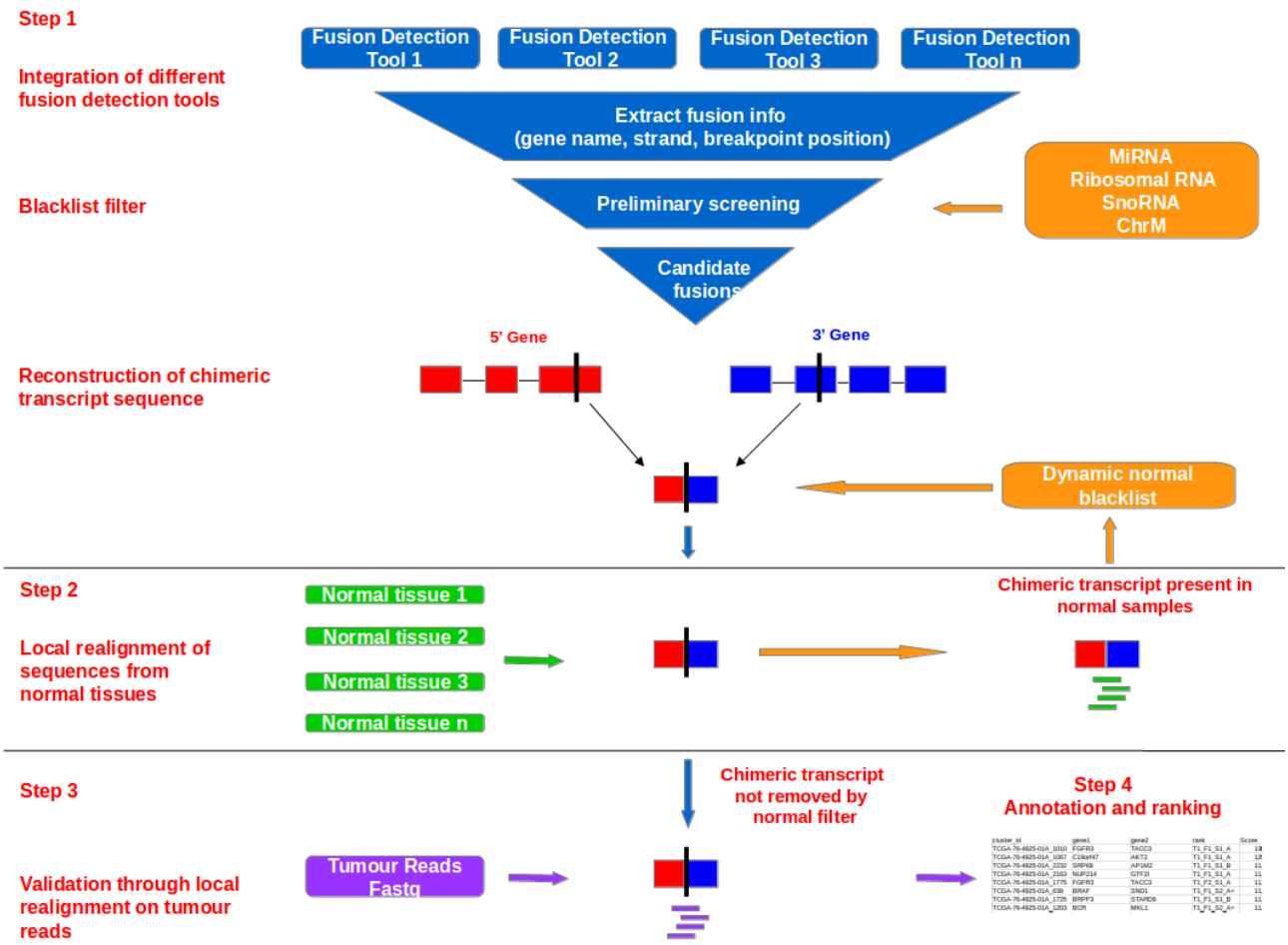


**Figure 8. Fusion Validator workflow.**

Step 1: Chimeric transcript breakpoint coordinates from several fusions or alternative splicing detection software are merged and annotated using a standardized file format. Each chimeric transcript sequence around the breakpoint is subsequently reconstructed by extracting a user defined region length upstream and downstream from the fusion breakpoint.

Step 2: Sequences from user-provided list of normal samples are locally realigned against each candidate chimeric transcript to remove aberrant junctions present in normal tissues. A blacklist with the coordinates of each event removed by normal filter is dynamically populated whenever a new sample is processed.

Step 3: Reads from the tumour sample are locally realigned against each detected and filtered chimeric transcript to select sequences spanning the breakpoint as potential candidates to reconstruct the fusion/skip.
Validation of the breakpoint is performed using the candidate reads and combining both a de novo assembly and a seed and extend algorithm for scaffold closing.
Additional filtering options are included to remove fusions mapping on homologous or highly repetitive regions.
Step 4: Chimeric transcripts retained after the validation step are annotated and a ranking score based on a linear combination of predictor variables is calculated for each event.

### *2.4.1 Integration of chimeric transcripts detected from multiple tools*

The first component of Fusion Validator creates a framework that converts results from different fusion detection software into a generic file format, generating a list of aberrant transcript junction sequences for further processing. First, Fusion Validator collects basic fusion information, like gene names, breakpoint coordinates and the number of reads supporting the chimeric event from different fusion or alternative splicing detection tools. A set of different modules are then used to extract additional chimeric transcript information from a user defined gene model GTF file, to create a consensus multi tool output with a standard annotation: The annotation includes coordinates and strand of the genes that create the chimeric transcript, splice-donor and splice-acceptor site, gene location and exonic location of the breakpoints. During the first step, a preliminary optional screening of the fusion transcripts is also performed to remove read through events, fusions/skips involving miRNAs, small nucleolar RNAs, ribosomal and mitochondrial genes. Any fusion transcripts detected by more than one algorithm involving the same genes with identical breakpoints and orientation, are collapsed into one single record and noted as having been recurrently found by different software.

After this initial screening, the multi-tool integration step reconstructs each chimeric transcript sequence around the breakpoints, extracting and merging a region upstream from the fusion breakpoint for 5’ gene, and downstream from the breakpoint for 3’ gene (default length 200bp). The chimeric sequence is retrieved at the genomic level for breakpoints falling into intronic or up/downstream regions and at transcriptomic level for exonic breakpoints. If an exonic breakpoint overlaps different transcript isoforms, the reconstruction algorithm gives priority to transcripts whose breakpoint is on an exon junction, and then to the longest length isoforms. Fusion Validator currently supports the analysis of fusions detected by Defuse, Chimerascan, STAR-Fusion, TopHat-Fusion, MapSplice, FusionCatcher and exons skips detected by Skippy.

### *2.4.2 Normal tissues filter*

The second component of Fusion Validator is responsible for removing recurrent transcripts that are also found in normal transcriptomes, using a user provided list of sequences from RNA-seq of normal tissues. This filtering step also makes use of a blacklist database of fusions already validated in normal tissues, that is dynamically updated as well as new samples are processed through the Fusion Validator pipeline. The list of chimeric transcripts created after the multi tool integration step is first scanned against the normal blacklist to remove all the events already found in normal tissues and involving the same genes with the same strands and breakpoint coordinates.

Then, sequencing reads for each user provided normal tissue are locally aligned with STAR 2.4.2a against the chimeric transcripts that remain after

blacklist screening. Every aligned read is considered as a candidate for supporting the fusion in a normal sample if it spans the breakpoint of the chimeric transcript with a perfect match and with a minimum overlap (parameter -e, default = 10 bp). Every chimeric transcript containing a minimum number of candidate reads (parameter -r , default = 3 reads) for more than 3 normal tissues (parameter -n , default = 3 samples), is discarded and its breakpoint coordinates dynamically populates the blacklist normal database. Remaining events not found in normal tissues are selected for further validation in tumour samples.

Simulated normal filtering steps were run on fusions detected on 5 random TCGA pan cancer samples of different sizes from Stransky et al., using subsets with different number of Gtex normal tissue samples, to assess the relationship between the number of normal samples screened versus the number of fusions discarded by the normal filter.

Results in Figure 9 show that the number of fusions removed by the normal filter rapidly increases when up to 250 normal samples are used. The distribution curve tends to slightly increases when the list contains over 500 normals, and starts reaching a plateau at approximately 1300 samples. This amount of normal samples represents the optimal configuration for Fusion Validator normal filtering module, since adding extra samples provides minimal benefit in terms of normal fusion removal, while significantly increasing computational time.

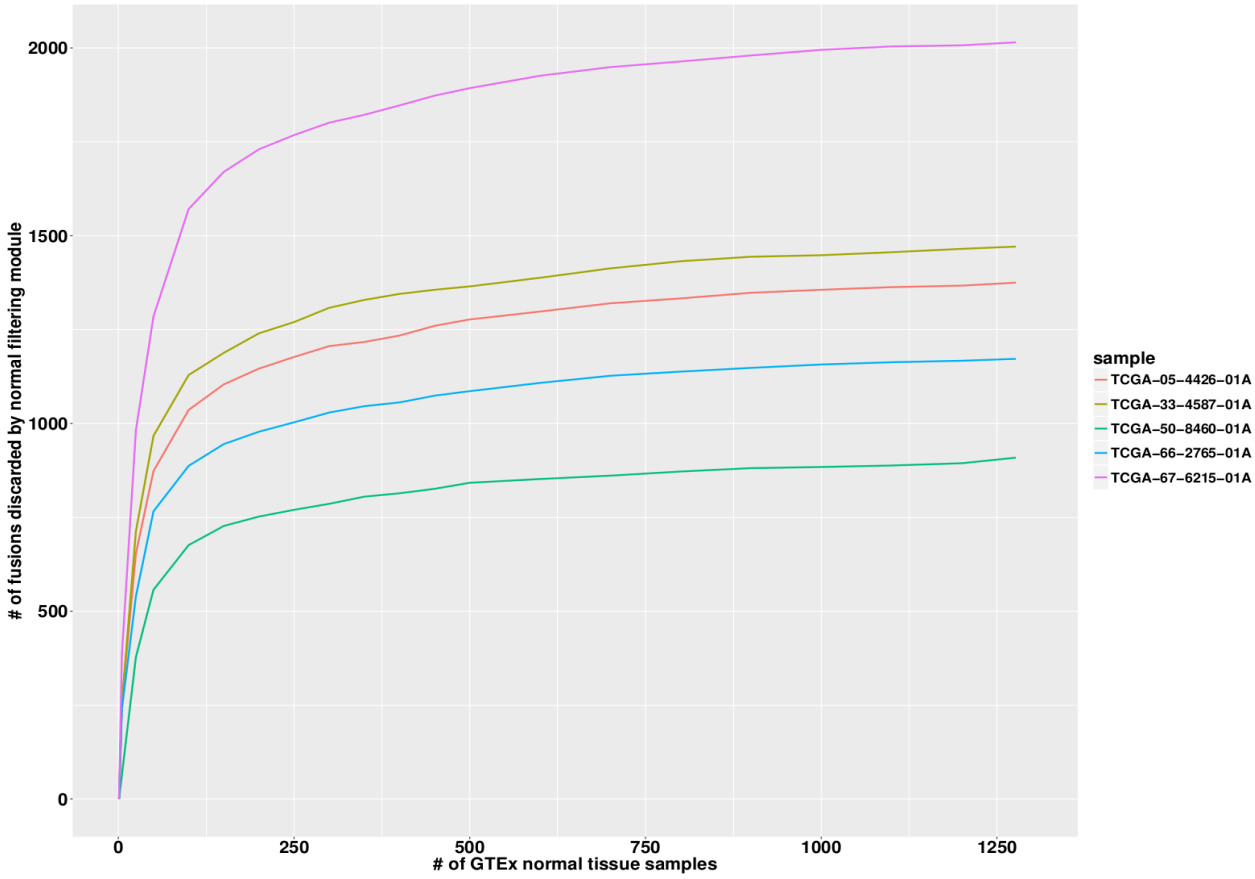


**Figure 9. Distribution of number of fusions removed by Fusion Validator normal filter varying the number of GTEx normal tissues used for 5 different pan-cancer TCGA samples.**

The number of events classified as present in normal tissues and removed by Fusion Validator's normal filter tends to increase in each of the 5 TCGA samples processed as the number of GTEx normal tissues used increases. The distribution curve tends to slightly increase when the list contains over 500 normals, until it reaches a plateau at around 1300 samples.

*2.4.3 Local realignment validation*

The third component of Fusion Validator evaluates chimeric transcripts candidates through a dynamic realignment approach, to distinguish real fusion or skip events from artifacts. RNA-seq reads from the tumour are locally aligned in single end against each chimeric transcript retrieved in the previous step, using STAR 2.4.2a. All the reads aligning on each chimeric transcript are extracted using samtools and selected as potential candidates to attempt to reconstruct the sequence spanning the breakpoint and validate the fusion/skip. Validation of the breakpoint is performed using the candidates reads and combining both a de novo assembly and a seed and extend algorithm for scaffold closing (Figure 10). De novo assembly of candidate reads is performed using Abyss 1.9.0[49] with iteration steps for different kmer size. Contigs generated by Abyss are then compared with the reconstructed chimeric transcript using BLAST. Fusions/skips are considered validated if at least one de novo assembled contig spans the breakpoint of the chimeric transcript, with at least 3 bp overlap and with maximum 1 mismatch.

The second approach used for validating the aberrant transcripts creates an artificial 5 base gap around the chimeric transcript breakpoint (the base on the breakpoint and the 2 bases at 5' and 3' of the breakpoint are replaced with Ns), and uses Gapfiller software[50] to try to close the artificially created scaffold between the two genes involved in a fusion, or the two exons involved in a skip. In brief, reads previously selected as potential candidates for chimeric transcript reconstruction are used as input sequences for Gapfiller and aligned against the artificial scaffolds for each fusion/skip using Bowtie[51] (for reads lower than 50 bp) or the Burrows-Wheeler Aligner (BWA)[52] (for reads longer than 50bp). Reads aligned on scaffold sequences are then split into shorter k-mers and used to iteratively fill the created gap, from the left and right edge, one nucleotide at a time. The k-mer size is selected as 85% of the length of the

reads; This size was tested on different Gapfiller runs on simulated and real datasets and was the one which ensured the best compromise between coverage and accuracy in gap closing. Every nucleotide incorporated step by step by Gapfiller is considered to fill a base gap if is covered by at least 2 k-mer sequences. After filling all the gapped bases, the scaffold can be considered closed and the chimeric transcript validated if an overlap of minimum 3 bp can be found in both left and right extension and the difference between the final length of the gapclosed sequence and the length of the original scaffold is not higher than 1 bp. Chimeric transcripts successful reconstructed and confirmed by de novo assembly or gap filling approach are flagged in the final output as validated.

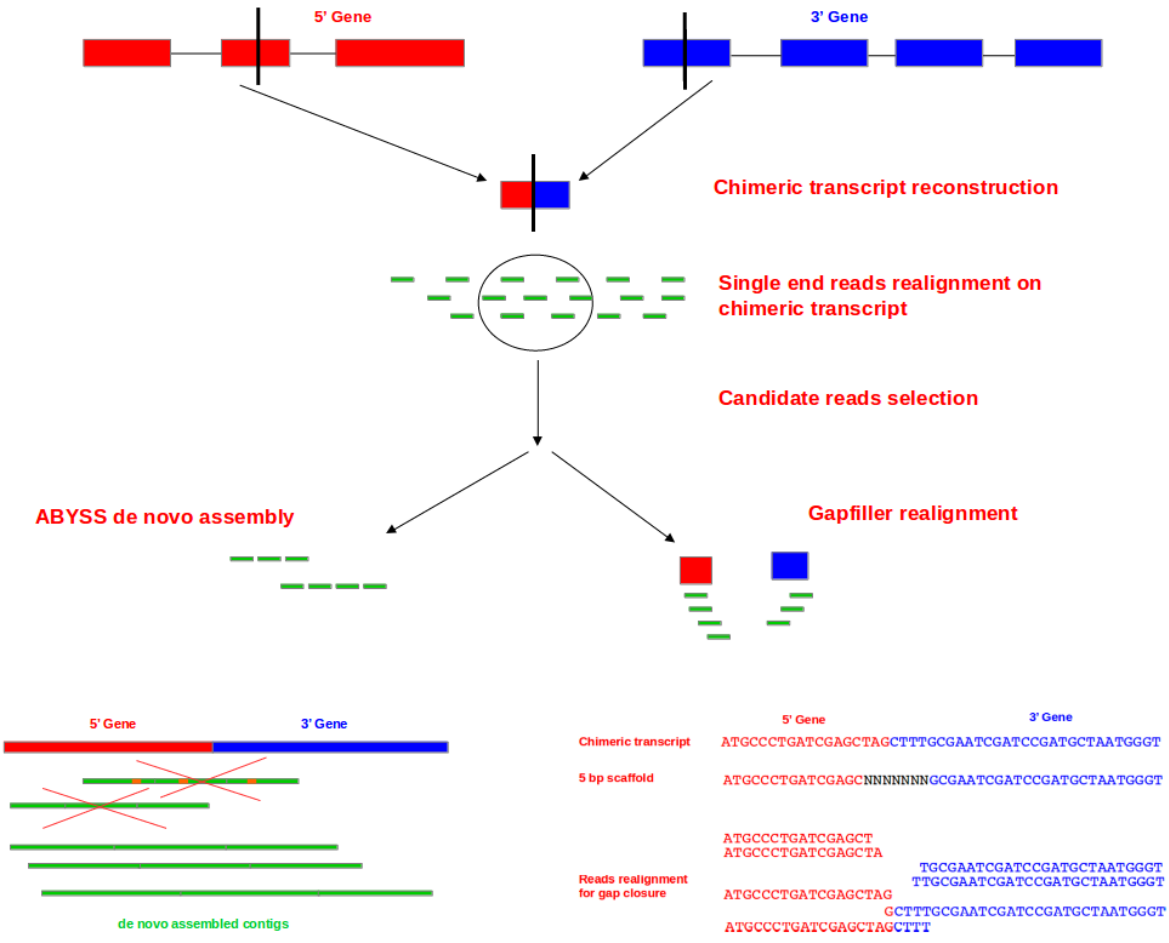

**Figure 10. Schematic representation of chimeric transcript's breakpoint validation.**
Tumour reads are realigned in single end on each reconstructed chimeric transcript and only sequences spanning the breakpoint are selected as candidate reads.
Using a de novo assembly approach (ABYSS), a contig generated by candidate reads reassembly is eligible to validate the chimeric transcript if it spans the breakpoint of the chimeric sequence, with at least 3 bp overlap and with maximum 1 mismatch.
Using a seed and extend alignment approach (Gapfiller), a chimeric transcript is validated if the candidate reads can close and extend, for minimum 3bp in both 5' and 3' direction, an artificial 5bp scaffold around the chimeric sequence breakpoint.

### *2.4.4 Genomic realignment validation*

Additional filtering options are included only for fusion genes validation to remove chimeric transcripts with sequence similarity between two regions around the breakpoint, or with other locations in the genome, as well as transcripts with breakpoint located in highly repetitive regions (Figure 11).
During this step, previously reconstructed fusion transcript sequences are aligned to the reference genome using BLAST, with DUST filtering for low complexity regions activated.

Chimeric transcripts that realign in genomic regions of low complexity are flagged by Fusion Validator as "low complexity regions". Then, for every analyzed transcript, if the sequence spanning the breakpoint aligns with more than 50bp and with 100% identity to other locations in the genome, the chimeric candidate is considered as a misalignment due to high level of homology between regions and reported by Fusion Validator with the flag "homologous region". As final step, for every BLAST alignment spanning the breakpoint of a chimeric transcript, the difference between the end of the 5' gene alignment and the breakpoint position and the difference between the start of 3' gene

alignment and the breakpoint position is computed. If the maximum of the two differences is higher than 5 bases with identity greater than 99%, there is a high similarity between the regions of the two genes surrounding the breakpoint and the fusion transcript is flagged as “similarity between genes” by Fusion Validator. The genomic realignment filter can be used to remove fusions involving pseudogenes or solve conflicts between genes with different rearrangements (so called promiscuous genes).

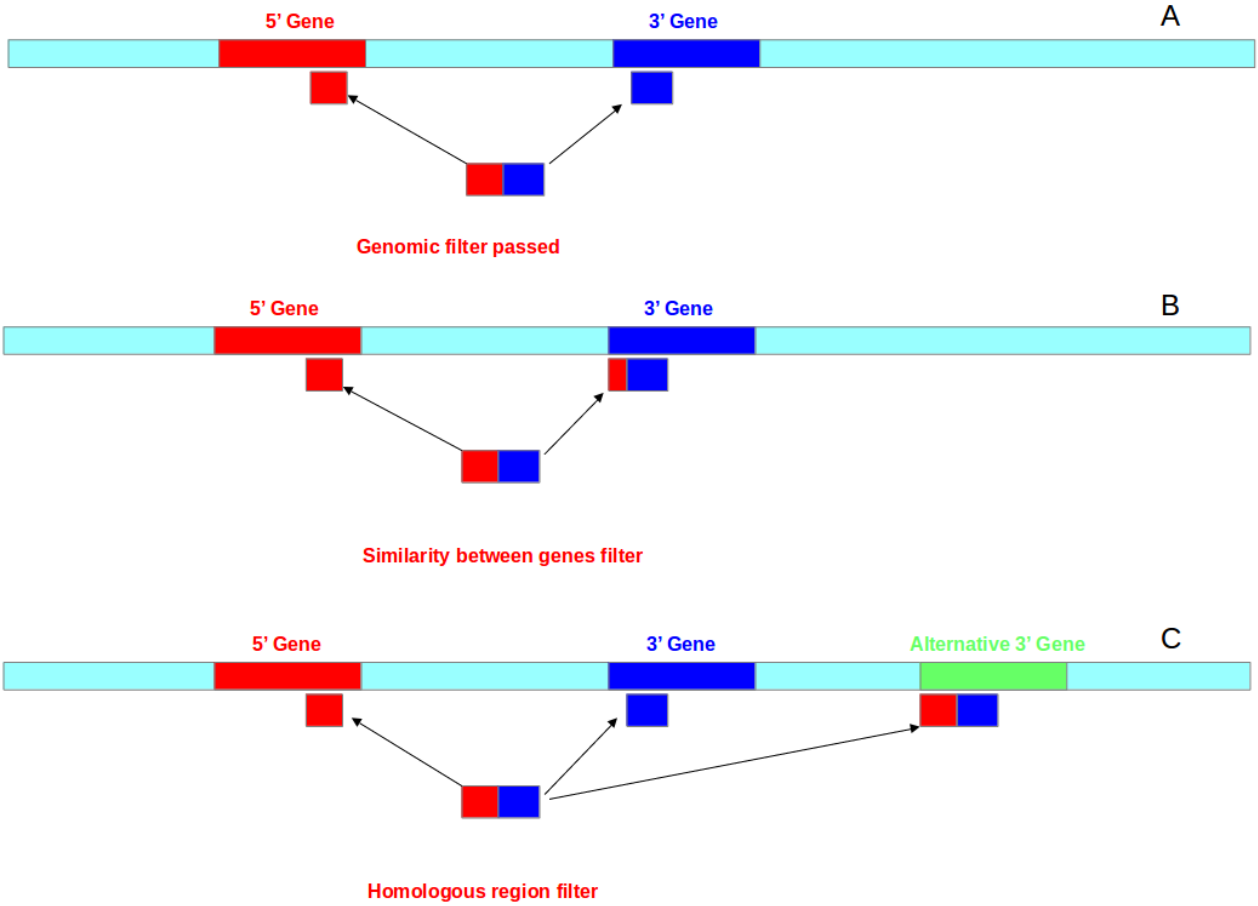

**Figure 11. Additional Fusion Validator filter based on sequence similarity on the genome.**
Realignment of candidate fusion against the genome:
a) Fusion candidate passing the genome realignment filter
b) Fusion candidate with high similarity between the regions of the two genes surrounding the breakpoint
c) Fusion candidate aligning in multiple regions with high level of homology.

### *2.4.5 Annotation and ranking*

The last component of Fusion Validator annotates each chimeric transcript retained after the validation step and calculates a score that is used to rank the final list of validated events. The ranking score is the result of a function based on a linear combination of predictor variables that best discriminate between cancer-driver validated events and false positives. The predictor variables of the function and their relative coefficients were generated using a linear discriminant analysis function as a training set on the 115 validated kinase fusions from 191 pan-cancer TCGA samples described above.

The list of annotation variables used to compute the ranking score, summarized in Table 6, are:
- Total number of reads supporting the fusion: Include mate pairs that harbour a fusion junction in the insert sequence (span reads) and pairs that harbour the fusion junction in one of the two reads (split reads).
- Breakpoint coverage: Corresponds to the number of reads locally aligning on the region 5bp upstream from the fusion breakpoint for 5’ gene, and 5bp downstream from the breakpoint for 3’ gene of the chimeric transcript. Reads that are not a primary alignment and have more than 4 mismatches are excluded from counting.

- 3’/5’ imbalance ratio: Is a measure of the read orientation distribution around the fusion transcript breakpoint and is calculated by subtracting the number of 3’ reads to the number of 5’ reads spanning the chimeric transcript breakpoint, and dividing the result by the breakpoint coverage. 3’/5’ imbalance ratio range from -1 (all the sequences spanning the breakpoint are 3’ reads) , to +1 (all the sequences spanning the breakpoint are 5’ reads). Values closer to the extremity of this interval are more likely to be artifacts caused by sequencing errors.
- Gene location: The location of the breakpoints for both the genes involved in the fusion. (intronic, up/downstream, exonic or coding sequence).
- Exonic location: The location of the breakpoint in the exons of the fusions (inside, outside or at start or end of an exon).
- Fusion recurrence in solid tumours: Annotation step check if the genes involved in the fusion are present in the Catalogue Of Somatic Mutation In Cancer (COSMIC) [Forbes] fusion database version 77 for GRCh37.
- Donor/acceptor site: The splice pattern on the fusion junction. The major canonical splice pattern GT-AG is most likely to be preserved in true fusion than the minor canonical ones (GC-AG and AT-AC) or the non-canonical (any other combination of dinucleotides).
- Number of different fusion finder tools that have identified the chimeric transcript: Recurrent driver events are usually detected by multiple fusion detection software tools.
- Fusion orientation: Indicates the strand of the fusion junction for the first and second gene involved in the fusion. Fusions with both genes on the same orientation are most likely to be true.
- Reading frame: Predicted effect of the fusion estimated using Gene Rearrangement AnalySiS (GRASS) tool [https://github.com/cancerit/grass]. Different reading frame predictions include frameshift fusions, in frame fusions, stop codon formed at breakpoint junction, UTR to UTR fusions, intronic or ambiguous events.

- Number of contigs generated by de novo reconstruction of the chimeric transcript: A high number of contigs generated by de novo assembly is due to the alignment of a large number of candidate reads with mismatches or multiple alignments. High number of de novo contigs increase the probability of a fusion transcript to be an artifact.
- Maximum percentage of chimeric transcript reconstructed by de novo aligned contigs: It's calculated for each chimeric transcript sequence as the length of the longest de novo aligned contig spanning the breakpoint junction divided by the total length of the reference chimeric transcript. A high percentage of reconstructed chimeric transcript increases the probability of a fusion to be classified as a true positive event.
- Maximum overlap around the breakpoint for de novo aligned contigs: It is calculated as the maximum difference between the end of a contig and the chimeric transcript breakpoint position, if the contig overlaps mostly on 5' gene, and as the maximum difference between the start of a contig and the breakpoint position, in the opposite case. A longer overlap around the breakpoint increases confidence in the validation of a fusion event.
- Maximum percentage of chimeric transcript reconstructed by de novo aligned contigs: It's calculated for each chimeric transcript sequence as the length of the longest de novo aligned contig spanning the breakpoint junction divided by the total length of the reference chimeric transcript. High percentage of reconstructed chimeric transcript increase the probability of a fusion to be classified as a true positive event.
- Maximum overlap around the breakpoint for de novo aligned contigs: It is calculated as the maximum difference between the end of a contig and the chimeric transcript breakpoint position, if the contig overlap mostly on 5' gene, and as the maximum difference between the start of a contig and the breakpoint position, in the opposite case. The longer is the overlap around the breakpoint the more reliable is the validation of a fusion event

| | | | |
|---|---|---|---|
| **Cancer Genes** | No genes in COSMIC census | At least 1 gene in COSMIC census | Both genes in COSMIC census |
| **Breakpoint location** | intron/intron<br>intron/utr<br>intron/up-downstream<br>up-downstream/up-downstream<br>up-downstream/utr<br>utr/utr | coding/intron<br>coding/utr<br>coding/up-downstream | coding/coding |
| **Strand** | +/-<br>-/+ | +/+<br>-/- | |
| **Frame** | Frameshift<br>Undefined<br>NULL | In Frame | |
| **Splice pattern** | Non canonical | Canonical<br>GT-AG, GC-AG, AT-AC | |
| **Supporting reads** | <=3 | 3<X<15 | >=15 |
| **Breakpoint exonic location** | No splice junction | On splice junction | |
| **Coverage around breakpoint** | Coverage = 0 | Coverage >0 | |
| **3'/5' Imbalance ratio** | Ratio <= -0.9 or >=0.9 | -0.9 < Ratio < 0.9 | |
| **Supporting de novo contigs** | >250 | <=250 | |
| **Maximum de novo breakpoint extension** | <15bp | >=15bp | |
| **% de novo reconstructed transcript** | <15% | >=15% | |

**Table 6. List of annotation variables used for ranking score prediction and relative categories.**

## 2.5 Determining the accuracy of Fusion Validator using simulated datasets

### *2.5.1 Established fusion detection tools perform differently in simulated datasets*

The performance of Fusion Validator in terms of sensitivity and PPV, compared to other fusion detection software, was assessed using a total of 48 simulated datasets with different coverages (25X, 50X, 100X, 200X, 300X and 400X), read lengths (50bp, 75bp, 100bp and 125bp) and breakpoint positions (intact or broken exons). Our comparison of different fusion detection tools demonstrated that most had widely varying performance for different breakpoint position. The only exception was defuse, which had an average sensitivity of 95.97% for IE and 96.45% for BE and an average PPV of 40.95% for IE and 41.57% for BE.

FusionCatcher, Mapsplice and STAR-fusion had higher sensitivity for BE fusions than IE (79.16% vs 57.42% for FusionCatcher, 83.52% vs 65.35% for MapSplice and 95.89% vs 84.79% for STAR-fusion). However, while FusionCatcher and Mapsplice demonstrate higher average PPV in BE datasets compared to IE (48.02% in BE fusions vs 44.07% in IE for FusionCatcher and 80.45% in BE fusions vs 76.50% in IE for MapSplice), STAR-fusion presents an opposite trend with an average PPV of 61.21% for IE fusions and 56.05% for BE ones. Chimerascan on the other hand has a significantly higher sensitivity and PPV in IE fusions compared to BE fusions (average sensitivity 88.05% in IE fusions vs 53.25% in BE and average PPV of 53.88% in IE vs 44.71% in IE) (Figure 12, Tables 7).

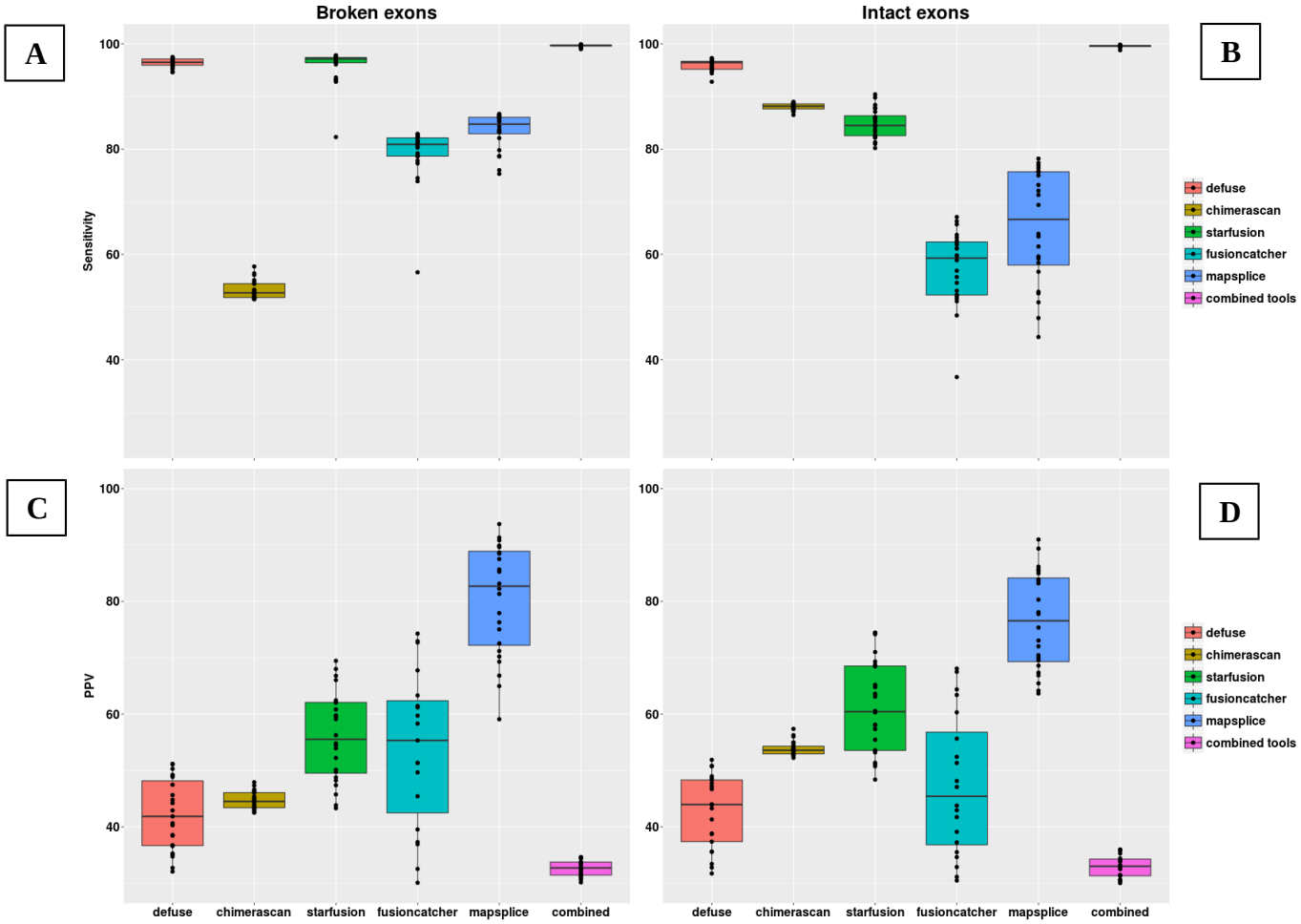


**Figure 12. Fusion transcript detection results for broken and intact exon's synthetic datasets.**

Boxplots show the distribution of Sensitivity (A, B) and PPV (C, D) in 24 broken exons (A, C) and 24 intact exons (B, D) datasets for 5 different fusion detection softwares (Defuse, Chimerascan, STAR-fusion, MapSplice and FusionCatcher) and for the multi-tools approach used by Fusion Validator.

| Sample | Sensitivity defuse | Sensitivity chimerascan | Sensitivity starfusion | Sensitivity fusioncatcher | Sensitivity mapsplice | Ppv defuse | Ppv chimerascan | Ppv starfusion | Ppv fusioncatcher | Ppv mapsplice |
|---|---|---|---|---|---|---|---|---|---|---|
| IE_25X_50bp | 95 | 87.2 | 83.3 | 36.7 | 44.3 | 48.5934 | 54.2626 | 71.0145 | 64.3860 | 90.9651 |
| IE_25X_75bp | 94.9 | 87.4 | 87.7 | 53.1 | 50.9 | 51.8863 | 53.9839 | 74.4482 | 52.4186 | 84.9750 |
| IE_25X_100bp | 94.4 | 87.3 | 87.9 | 51.9 | 52.6 | 50.8347 | 53.6241 | 74.1772 | 51.3353 | 85.6678 |
| IE_25X_125bp | 92.8 | 86.5 | 85.8 | 54.6 | 47.9 | 50.7381 | 52.9700 | 74.4146 | 42.9583 | 86.1511 |
| IE_50X_50bp | 95.1 | 87.8 | 90.4 | 48.4 | 52.9 | 43.9667 | 57.3856 | 63.6172 | 68.0731 | 89.3581 |
| IE_50X_75bp | 95.2 | 87.7 | 88.4 | 56.9 | 63.4 | 46.7125 | 53.4105 | 68.4741 | 48.1387 | 83.6412 |
| IE_50X_100bp | 94.8 | 87.1 | 87.1 | 55.7 | 63.9 | 47.7101 | 52.9483 | 68.7451 | 47.0837 | 83.2031 |
| IE_50X_125bp | 95.5 | 87.3 | 86.1 | 59.7 | 61.5 | 48.9744 | 52.9412 | 69.3237 | 39.1219 | 85.4167 |
| IE_100X_50bp | 97 | 88.2 | 89.8 | 51.1 | 56.7 | 38.7380 | 56.0356 | 58.1230 | 67.5033 | 83.8757 |
| IE_100X_75bp | 96.1 | 87.9 | 85.7 | 59.8 | 72.1 | 47.1541 | 56.2740 | 63.1075 | 43.7775 | 80.2895 |
| IE_100X_100bp | 96 | 88 | 85.2 | 58.9 | 71.3 | 48.3140 | 53.4629 | 64.7909 | 41.7434 | 77.7535 |
| IE_100X_125bp | 95.7 | 87.8 | 83.7 | 63.2 | 69.4 | 44.0000 | 53.5692 | 65.1869 | 34.6491 | 78.0652 |
| IE_200X_50bp | 96.5 | 88.7 | 86.1 | 51.6 | 58.4 | 32.7785 | 54.9907 | 53.5781 | 63.3907 | 75.3548 |
| IE_200X_75bp | 96.4 | 88.6 | 82.2 | 61.1 | 75.8 | 41.3202 | 54.9287 | 58.0508 | 37.2561 | 70.4461 |
| IE_200X_100bp | 96.6 | 88.1 | 81.3 | 61.2 | 75.7 | 43.2990 | 54.4499 | 60.3116 | 35.5194 | 72.0266 |
| IE_200X_125bp | 96.5 | 88.4 | 84.6 | 65.7 | 73.2 | 38.8330 | 52.2459 | 60.5634 | 29.6748 | 73.0539 |
| IE_300X_50bp | 96.9 | 89 | 80.2 | 52.3 | 59.2 | 29.3993 | 54.0049 | 51.1178 | 60.3230 | 68.5979 |
| IE_300X_75bp | 97 | 88.6 | 82.2 | 62.3 | 76.4 | 35.6225 | 53.9257 | 53.2183 | 32.8760 | 65.4670 |
| IE_300X_100bp | 96.7 | 88.6 | 82.6 | 61.9 | 77.4 | 37.3792 | 53.6970 | 55.4656 | 31.1525 | 69.5418 |
| IE_300X_125bp | 96.7 | 88.2 | 82.4 | 66.3 | 75 | 35.5645 | 52.8460 | 57.3611 | 26.6479 | 70.0280 |
| IE_400X_50bp | 97.3 | 88.6 | 81 | 52.2 | 59.6 | 26.1068 | 52.6128 | 48.3852 | 55.6503 | 64.1550 |
| IE_400X_75bp | 97.2 | 88.8 | 82.6 | 63.7 | 76.9 | 29.7157 | 52.9833 | 50.7519 | 30.4931 | 63.6589 |
| IE_400X_100bp | 96.6 | 88.8 | 84.3 | 62.6 | 78.2 | 31.7450 | 52.9517 | 51.3682 | 28.4934 | 67.3557 |
| IE_400X_125bp | 96.6 | 88.5 | 84.4 | 67.1 | 75.7 | 33.4256 | 52.6472 | 53.4898 | 25.0935 | 66.8728 |
| BE_25X_50bp | 95.5 | 57.7 | 82.3 | 56.6 | 76 | 49.1508 | 46.3454 | 69.4515 | 72.9381 | 93.7115 |
| BE_25X_75bp | 94.6 | 56.4 | 93.1 | 79.2 | 78.6 | 50.2924 | 44.9045 | 68.0058 | 63.3094 | 89.6237 |
| BE_25X_100bp | 95.2 | 54.4 | 93.6 | 77.3 | 78.7 | 51.0730 | 43.6948 | 66.8094 | 61.4467 | 89.8402 |
| BE_25X_125bp | 94.7 | 51.8 | 92.8 | 73.9 | 75.3 | 51.1892 | 42.5987 | 66.0498 | 51.3194 | 90.8323 |
| BE_50X_50bp | 96 | 54.9 | 93.2 | 74.5 | 79.8 | 44.7970 | 47.3276 | 60.8355 | 74.2772 | 91.3043 |
| BE_50X_75bp | 95.8 | 56.1 | 96.1 | 81.1 | 83.7 | 47.4727 | 44.4181 | 62.4431 | 59.7202 | 88.4778 |
| BE_50X_100bp | 96.5 | 55.1 | 96.7 | 80.3 | 84.3 | 48.8608 | 43.8345 | 62.0269 | 58.3152 | 88.6435 |
| BE_50X_125bp | 95.7 | 53.2 | 97.3 | 78.7 | 83.5 | 49.2284 | 43.3931 | 62.2521 | 45.4388 | 89.8816 |
| BE_100X_50bp | 96.5 | 52.8 | 96.5 | 77.8 | 82.1 | 41.8655 | 47.9129 | 56.2682 | 72.7103 | 85.2544 |
| BE_100X_75bp | 96.4 | 52.6 | 96.9 | 82.3 | 85.2 | 42.9207 | 46.6312 | 59.1214 | 51.3733 | 82.2394 |
| BE_100X_100bp | 96.4 | 54.6 | 96.7 | 81.5 | 85.9 | 44.2202 | 43.4022 | 59.7651 | 49.6648 | 85.6431 |
| BE_100X_125bp | 96.5 | 53.3 | 97.2 | 80.7 | 85.4 | 45.6481 | 43.1929 | 59.5588 | 37.3093 | 87.5000 |
| BE_200X_50bp | 97 | 52.7 | 97.2 | 78.6 | 83.2 | 35.2856 | 46.3093 | 52.2300 | 67.7586 | 72.5371 |
| BE_200X_75bp | 96.9 | 51.8 | 97.3 | 82.8 | 85.9 | 38.4524 | 46.0444 | 54.5098 | 39.5415 | 75.0218 |
| BE_200X_100bp | 97.3 | 51.5 | 97.6 | 82.1 | 86.2 | 40.3065 | 45.3744 | 54.8007 | 36.9321 | 81.3208 |
| BE_200X_125bp | 96.4 | 53.2 | 96.7 | 81.3 | 86.3 | 40.6580 | 42.9032 | 53.9320 | 29.0461 | 83.1407 |
| BE_300X_50bp | 97.2 | 52.4 | 97.4 | 78.7 | 83.3 | 32.0792 | 46.1674 | 49.7701 | 61.1975 | 64.9766 |
| BE_300X_75bp | 97.5 | 52 | 97.4 | 82.8 | 85.8 | 35.1098 | 45.1781 | 50.1287 | 32.5472 | 70.2128 |
| BE_300X_100bp | 96.4 | 51.7 | 97.6 | 82.4 | 86.5 | 36.7518 | 44.4923 | 48.7756 | 30.0950 | 76.2787 |
| BE_300X_125bp | 97.1 | 51.8 | 97.1 | 81.6 | 86.4 | 38.5624 | 42.5287 | 48.2365 | 24.8856 | 77.9080 |
| BE_400X_50bp | 97.3 | 52.7 | 97.5 | 78.8 | 83.2 | 29.7645 | 44.8129 | 47.3991 | 55.3371 | 59.0909 |
| BE_400X_75bp | 97.4 | 52.2 | 97.7 | 82.9 | 86 | 32.7395 | 44.5392 | 45.7611 | 27.9031 | 66.8221 |
| BE_400X_100bp | 97.4 | 51.5 | 97.8 | 82.4 | 86.7 | 34.7485 | 43.9795 | 43.8368 | 26.6408 | 69.3046 |
| BE_400X_125bp | 97.1 | 51.5 | 97.8 | 81.6 | 86.5 | 36.6001 | 42.9525 | 43.3127 | 22.7933 | 71.1934 |
| average_IE | 95.97916667 | 88.04583333 | 84.79166667 | 57.41666667 | 65.35 | 40.95047705 | 53.88132647 | 61.2118566 | 44.07331377 | 76.4966937 |
| average_BE | 96.45 | 53.24583333 | 95.89583333 | 79.1625 | 83.52083333 | 41.57403773 | 44.70574474 | 56.05336742 | 48.02085607 | 80.44829716 |

**Table 7. Distribution of Sensitivity, and PPV for fusion detected by 5 different softwares on 48 synthetic datasets.**

The probability of detecting a true fusion increased with the increment of the sequencing coverage for all the software, except for STAR-fusion and Chimerascan. However, adding more reads to the simulated dataset tended to inflate the number of false positive events, penalizing the PPV: This trend is observed for all the fusion detection tools, with the exception of Chimerascan, in which the PPV appear to be constant for each sequencing coverage (Figure 13).

With regards to the performance of different fusion detection software for datasets of different read length, no particular improvement in term of sensitivity was found for the inspected tools, varying the sequence length under fixed coverage. The only exception is represented by MapSplice and FusionCatcher, that show a very poor performance in term of sensitivity for datasets with reads of 50bp compared to the other lengths, particularly for IE datasets. The impact of different read lengths on the number of extra calls and, consequently, on the PPV depends on different software, breakpoint position and coverage. Tools like Defuse, for example, tend to increase the number of extra calls (and decrease the PPV), when the length of the reads decrease, in both IE and BE samples. The same trend is observed in STAR-fusion for IE samples and MapSplice for BE samples with coverage greater than 100X (Figure 13). FusionCatcher, on the other hand, shows a significantly higher PPV for shorter reads (50bp) and a progressive decrease of PPV when the read length increase. No particular changes in PPV values for different read length were found in Chimerascan.

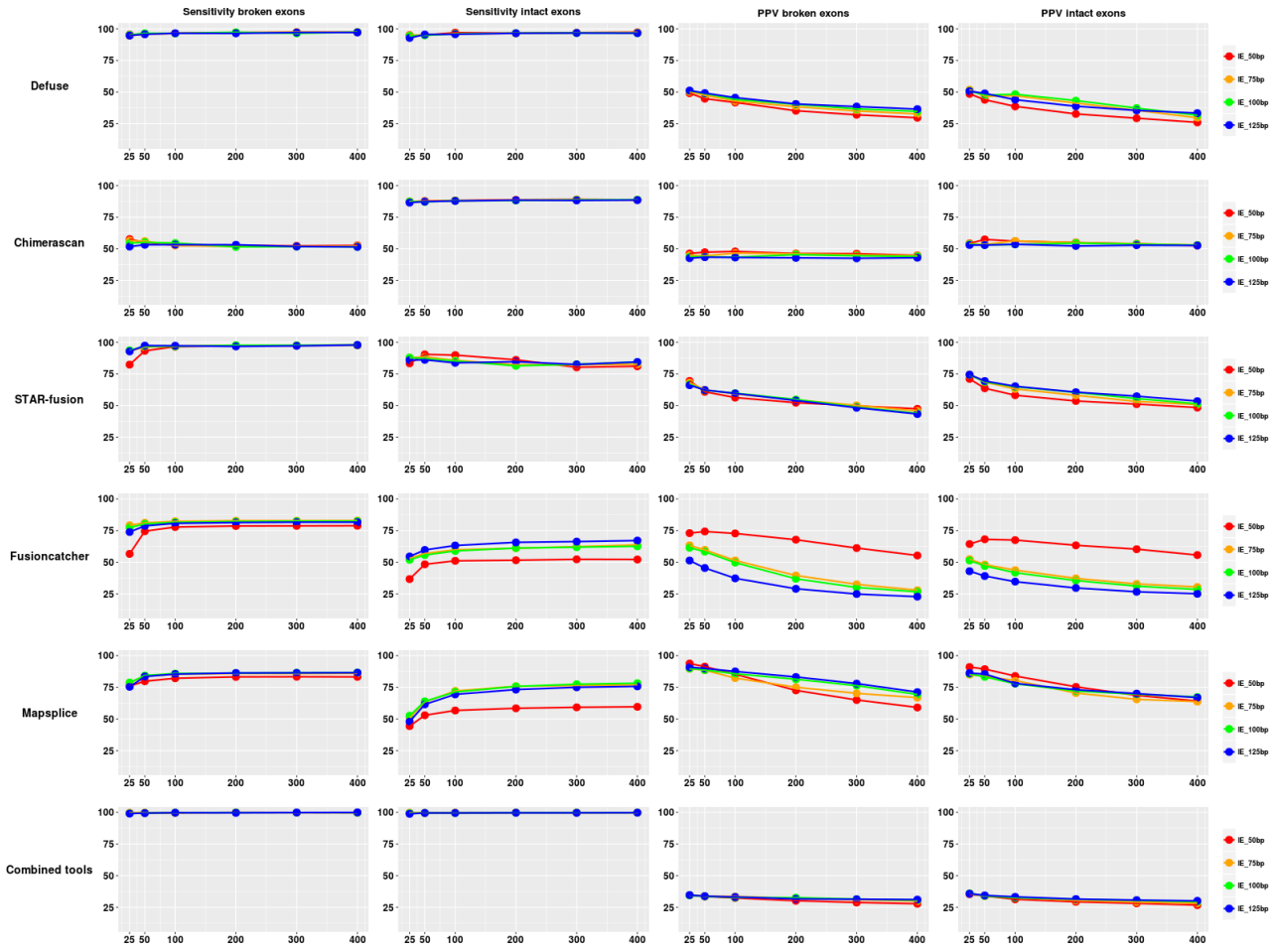


**Figure 13. Sensitivity and PPV distribution for fusion transcript detected by different software on synthetic datasets of different coverage, read length and breakpoint position.**

Each row shows the results for 5 different fusion detection softwares (Defuse, Chimerascan, STAR-fusion, MapSplice and FusionCatcher) and for the multi-tool approach used by Fusion Validator. The 4 columns of the plot panel measure distribution of Sensitivity in broken and intact exons and PPV in broken and intact exons database respectively.

X-axis shows the trend of each distribution for different read coverage.

Read lengths are represented with different colors (red for 50bp, yellow for 75bp, green for 100bp and blue for 125bp).

### *2.5.2 Fusion Validator maintains high accuracy across a range of sequence coverage and read lengths*

Using our multi-tool integrated approach implemented in Fusion Validator, chimeric transcripts detected by the 5 different fusion detection algorithms tested were merged into a single output. The multi-tools approach successfully identified an average of 99.58% of the fusions for the IE datasets and an average of 99.65% for those in the BE group, with sensitivity values quite constant across different read length and coverage datasets. Combining results from different fusion finding algorithms, however, tends to increase the number of false positive events. This can be seen from the PPV distribution, the average values of which is lower than any single software evaluated (average PPV 31.84% for IE and 32.21% for BE fusions). Average PPV for the multi-tool approach slightly decreases when the coverage increases for both BE and IE datasets and slightly increases for read length increases in BE and IE datasets with coverage greater than 100X (Table 8 and Figure 13). These results demonstrate that the multi-tool approach adopted by Fusion Validator can perform optimally for short reads and low coverage sequencing, as an increase in terms of coverage and read length does not bring any benefit on number of true fusions and extra calls detected.

The validation performed by Fusion Validator's local realignment step, on fusion transcripts detected by 5 different software, significantly reduced the number of false positive fusion calls by half in all the simulated datasets, increasing the average PPV to 66.01% for IE and 64.56% for BE subsamples, and maintaining a very high average sensitivity (94.79% and 96.75% for IE and BE respectively), specificity (77.20% and 74.71% for IE and BE respectively) and accuracy (82.78% and 81.75% for IE and BE respectively) (Table 9).

| Sample | Fusion tools detected | Fusion tools not detected | Fusion tools extra calls | Sensitivity fusion tools | Ppv fusion tools | F_measure fusion_tools |
|---|---|---|---|---|---|---|
| IE_25X_50bp | 995 | 5 | 1823 | 99.5 | 35.3087 | 52.1215 |
| IE_25X_75bp | 995 | 5 | 1775 | 99.5 | 35.9206 | 52.7851 |
| IE_25X_100bp | 993 | 7 | 1766 | 99.3 | 35.9913 | 52.8332 |
| IE_25X_125bp | 988 | 12 | 1774 | 98.8 | 35.7712 | 52.5253 |
| IE_50X_50bp | 995 | 5 | 1941 | 99.5 | 33.8896 | 50.5589 |
| IE_50X_75bp | 995 | 5 | 1934 | 99.5 | 33.9706 | 50.6490 |
| IE_50X_100bp | 995 | 5 | 1937 | 99.5 | 33.9359 | 50.6104 |
| IE_50X_125bp | 995 | 5 | 1896 | 99.5 | 34.4172 | 51.1437 |
| IE_100X_50bp | 996 | 4 | 2184 | 99.6 | 31.3208 | 47.6555 |
| IE_100X_75bp | 996 | 4 | 2040 | 99.6 | 32.8063 | 49.3558 |
| IE_100X_100bp | 995 | 5 | 2066 | 99.5 | 32.5057 | 49.0027 |
| IE_100X_125bp | 995 | 5 | 1998 | 99.5 | 33.2442 | 49.8372 |
| IE_200X_50bp | 998 | 2 | 2409 | 99.8 | 29.2926 | 45.2916 |
| IE_200X_75bp | 997 | 3 | 2281 | 99.7 | 30.4149 | 46.6106 |
| IE_200X_100bp | 997 | 3 | 2181 | 99.7 | 31.3719 | 47.7262 |
| IE_200X_125bp | 996 | 4 | 2168 | 99.6 | 31.4791 | 47.8386 |
| IE_300X_50bp | 998 | 2 | 2552 | 99.8 | 28.1127 | 43.8681 |
| IE_300X_75bp | 997 | 3 | 2427 | 99.7 | 29.1180 | 45.0723 |
| IE_300X_100bp | 997 | 3 | 2279 | 99.7 | 30.4335 | 46.6324 |
| IE_300X_125bp | 996 | 4 | 2251 | 99.6 | 30.6745 | 46.9037 |
| IE_400X_50bp | 998 | 2 | 2736 | 99.8 | 26.7274 | 42.1631 |
| IE_400X_75bp | 997 | 3 | 2542 | 99.7 | 28.1718 | 43.9304 |
| IE_400X_100bp | 997 | 3 | 2405 | 99.7 | 29.3063 | 45.2976 |
| IE_400X_125bp | 997 | 3 | 2322 | 99.7 | 30.0392 | 46.1681 |
| BE_25X_50bp | 995 | 5 | 1884 | 99.5 | 34.5606 | 51.3019 |
| BE_25X_75bp | 996 | 4 | 1884 | 99.6 | 34.5833 | 51.3402 |
| BE_25X_100bp | 992 | 8 | 1892 | 99.2 | 34.3967 | 51.0814 |
| BE_25X_125bp | 990 | 10 | 1867 | 99 | 34.6517 | 51.3352 |
| BE_50X_50bp | 994 | 6 | 1968 | 99.4 | 33.5584 | 50.1767 |
| BE_50X_75bp | 998 | 2 | 1957 | 99.8 | 33.7733 | 50.4678 |
| BE_50X_100bp | 994 | 6 | 1964 | 99.4 | 33.6038 | 50.2274 |
| BE_50X_125bp | 994 | 6 | 1950 | 99.4 | 33.7636 | 50.4057 |
| BE_100X_50bp | 996 | 4 | 2083 | 99.6 | 32.3482 | 48.8355 |
| BE_100X_75bp | 998 | 2 | 1977 | 99.8 | 33.5462 | 50.2138 |
| BE_100X_100bp | 997 | 3 | 2050 | 99.7 | 32.7207 | 49.2711 |
| BE_100X_125bp | 997 | 3 | 2013 | 99.7 | 33.1229 | 49.7257 |
| BE_200X_50bp | 999 | 1 | 2313 | 99.9 | 30.1630 | 46.3358 |
| BE_200X_75bp | 998 | 2 | 2126 | 99.8 | 31.9462 | 48.3996 |
| BE_200X_100bp | 998 | 2 | 2064 | 99.8 | 32.5931 | 49.1384 |
| BE_200X_125bp | 996 | 4 | 2152 | 99.6 | 31.6391 | 48.0231 |
| BE_300X_50bp | 998 | 2 | 2469 | 99.8 | 28.7857 | 44.6832 |
| BE_300X_75bp | 998 | 2 | 2229 | 99.8 | 30.9266 | 47.2203 |
| BE_300X_100bp | 998 | 2 | 2175 | 99.8 | 31.4529 | 47.8313 |
| BE_300X_125bp | 998 | 2 | 2183 | 99.8 | 31.3738 | 47.7398 |
| BE_400X_50bp | 997 | 3 | 2585 | 99.7 | 27.8336 | 43.5181 |
| BE_400X_75bp | 998 | 2 | 2342 | 99.8 | 29.8802 | 45.9908 |
| BE_400X_100bp | 997 | 3 | 2255 | 99.7 | 30.6581 | 46.8956 |
| BE_400X_125bp | 999 | 1 | 2215 | 99.9 | 31.0828 | 47.4134 |
| average_IE | 995.75 | 4.25 | 2153.625 | 99.575 | 31.84266476 | 48.19087403 |
| average_BE | 996.458333 | 3.54166667 | 2108.208333 | 99.6458333 | 32.20685319 | 48.64881707 |

**Table 8. Distribution of Sensitivity, PPV and F-measure for Fusion Validator multi-tool approach on 48 synthetic datasets.**

| Sample | Ppv validator | Sensitivity Validator | Specificity Validator | Accuracy Validator | F_measure Validator |
|---|---|---|---|---|---|
| IE_25X_50bp | 74.0770 | 94.7739 | 81.8482 | 86.2669 | 83.1570 |
| IE_25X_75bp | 70.4733 | 94.2714 | 77.4397 | 82.5632 | 80.6535 |
| IE_25X_100bp | 72.4138 | 90.9366 | 80.4989 | 84.1972 | 80.6250 |
| IE_25X_125bp | 72.3748 | 90.6883 | 81.2910 | 86.2419 | 80.5031 |
| IE_50X_50bp | 70.9941 | 96.1809 | 79.8141 | 85.2520 | 81.6901 |
| IE_50X_75bp | 68.3644 | 95.7789 | 76.8382 | 82.4855 | 79.7823 |
| IE_50X_100bp | 70.1368 | 92.7638 | 79.7004 | 84.1064 | 79.8788 |
| IE_50X_125bp | 71.2519 | 92.6633 | 80.7653 | 85.9218 | 80.5592 |
| IE_100X_50bp | 67.2498 | 96.4859 | 78.7176 | 84.6541 | 79.2577 |
| IE_100X_75bp | 65.6797 | 97.9920 | 74.2684 | 80.6324 | 78.6463 |
| IE_100X_100bp | 69.2598 | 92.1608 | 80.1850 | 83.7635 | 79.0858 |
| IE_100X_125bp | 68.2504 | 90.9548 | 79.6225 | 85.1988 | 77.9836 |
| IE_200X_50bp | 61.4997 | 97.7956 | 74.6473 | 81.4500 | 75.5126 |
| IE_200X_75bp | 60.9619 | 97.8937 | 71.8468 | 78.4320 | 75.1347 |
| IE_200X_100bp | 65.5914 | 91.7753 | 78.0421 | 82.4733 | 76.5050 |
| IE_200X_125bp | 67.7229 | 92.2691 | 80.3587 | 85.6827 | 78.1130 |
| IE_300X_50bp | 59.6119 | 98.4970 | 73.9233 | 80.8732 | 74.2728 |
| IE_300X_75bp | 57.5439 | 98.6961 | 69.4058 | 76.8400 | 72.7004 |
| IE_300X_100bp | 64.3599 | 93.2798 | 77.4518 | 82.3871 | 76.1671 |
| IE_300X_125bp | 66.3824 | 93.7751 | 79.4258 | 85.0015 | 77.7362 |
| IE_400X_50bp | 56.5842 | 98.5972 | 72.4251 | 79.4590 | 71.9035 |
| IE_400X_75bp | 55.8390 | 98.7964 | 68.6771 | 76.0949 | 71.3510 |
| IE_400X_100bp | 62.2907 | 93.2798 | 76.7163 | 81.8636 | 74.6988 |
| IE_400X_125bp | 65.3953 | 94.5838 | 78.9983 | 84.9654 | 77.3268 |
| BE_25X_50bp | 72.4368 | 97.9899 | 80.3079 | 86.4189 | 83.2977 |
| BE_25X_75bp | 68.8421 | 98.4940 | 76.4331 | 84.0625 | 81.0409 |
| BE_25X_100bp | 70.0893 | 94.9597 | 78.7526 | 84.3273 | 80.6507 |
| BE_25X_125bp | 70.8995 | 94.7475 | 79.3787 | 84.7042 | 81.1068 |
| BE_50X_50bp | 69.1114 | 98.5915 | 77.7439 | 84.7400 | 81.2604 |
| BE_50X_75bp | 67.8596 | 98.7976 | 76.1369 | 83.7902 | 80.4570 |
| BE_50X_100bp | 68.6063 | 96.0765 | 77.7495 | 83.9080 | 80.0503 |
| BE_50X_125bp | 69.1691 | 95.4728 | 78.3077 | 84.1033 | 80.2198 |
| BE_100X_50bp | 64.3461 | 99.2972 | 73.6918 | 81.9747 | 78.0892 |
| BE_100X_75bp | 64.6714 | 99.5992 | 72.5341 | 81.6134 | 78.4221 |
| BE_100X_100bp | 67.0723 | 93.9819 | 77.5610 | 82.9340 | 78.2790 |
| BE_100X_125bp | 67.0066 | 92.2768 | 77.4963 | 82.3920 | 77.6371 |
| BE_200X_50bp | 61.2585 | 99.3994 | 72.8491 | 80.8575 | 75.8015 |
| BE_200X_75bp | 60.7186 | 99.8998 | 69.6613 | 79.3214 | 75.5303 |
| BE_200X_100bp | 63.8167 | 93.4870 | 74.3702 | 80.6009 | 75.8537 |
| BE_200X_125bp | 65.9957 | 92.1687 | 78.0204 | 82.4968 | 76.9166 |
| BE_300X_50bp | 58.2598 | 99.2986 | 71.2434 | 79.3193 | 73.4346 |
| BE_300X_75bp | 58.2746 | 99.4990 | 68.1023 | 77.8122 | 73.5011 |
| BE_300X_100bp | 61.6037 | 94.6894 | 72.9195 | 79.7668 | 74.6445 |
| BE_300X_125bp | 62.8859 | 93.8878 | 74.6679 | 80.6979 | 75.3215 |
| BE_400X_50bp | 56.1265 | 99.4995 | 69.9187 | 78.1686 | 71.7690 |
| BE_400X_75bp | 57.3241 | 99.5992 | 68.4031 | 77.7246 | 72.7672 |
| BE_400X_100bp | 60.8142 | 95.8877 | 72.6829 | 79.7970 | 74.4258 |
| BE_400X_125bp | 62.1622 | 94.3944 | 74.0858 | 80.3983 | 74.9603 |
| average_IE | 66.01287068 | 94.78705568 | 77.20449662 | 82.78359092 | 77.63518196 |
| average_BE | 64.55629669 | 96.74979308 | 74.70909081 | 81.74708031 | 77.30988154 |

**Table 9. Distribution of Sensitivity, PPV, Specificity, Accuracy and F-measure for Fusion Validator on 48 synthetic datasets.**

When comparing the performance of Fusion Validator across subsamples of different coverage and read length, a slight increase in sensitivity can be observed for each coverage increment in both IE and BE datasets, with 50bp and 75bp sequences showing higher sensitivity than 100bp and 125bp datasets. This difference is due to the method used to subsample different simulated datasets, as, for fixed coverage, an increase in read length decreases the total number of reads in the dataset and, consequently, the number of candidate reads used by Fusion Validator to reconstruct the chimeric transcript. The average sensitivity of Fusion Validator for shorter reads datasets tends to converge to the one calculated for longer reads samples when there is an increase in coverage (Figure 14). In contrast to what happens to the sensitivity, PPV and specificity for Fusion Validator tends to decrease slightly for each increment of coverage, with longer sequences showing better PPV and specificity than the shorter ones in both IE and BE datasets. Accuracy for Fusion Validator appear to be constant across samples of different coverage, as a result of a balance between the increase of TP events and sensitivity and the decrease of TN and specificity related to the increase of read coverage.

The overall constant accuracy across different datasets makes Fusion Validator an ideal tool with a good combination of sensitivity and specificity, that can perform extremely well also with short reads and coverage under 100X.

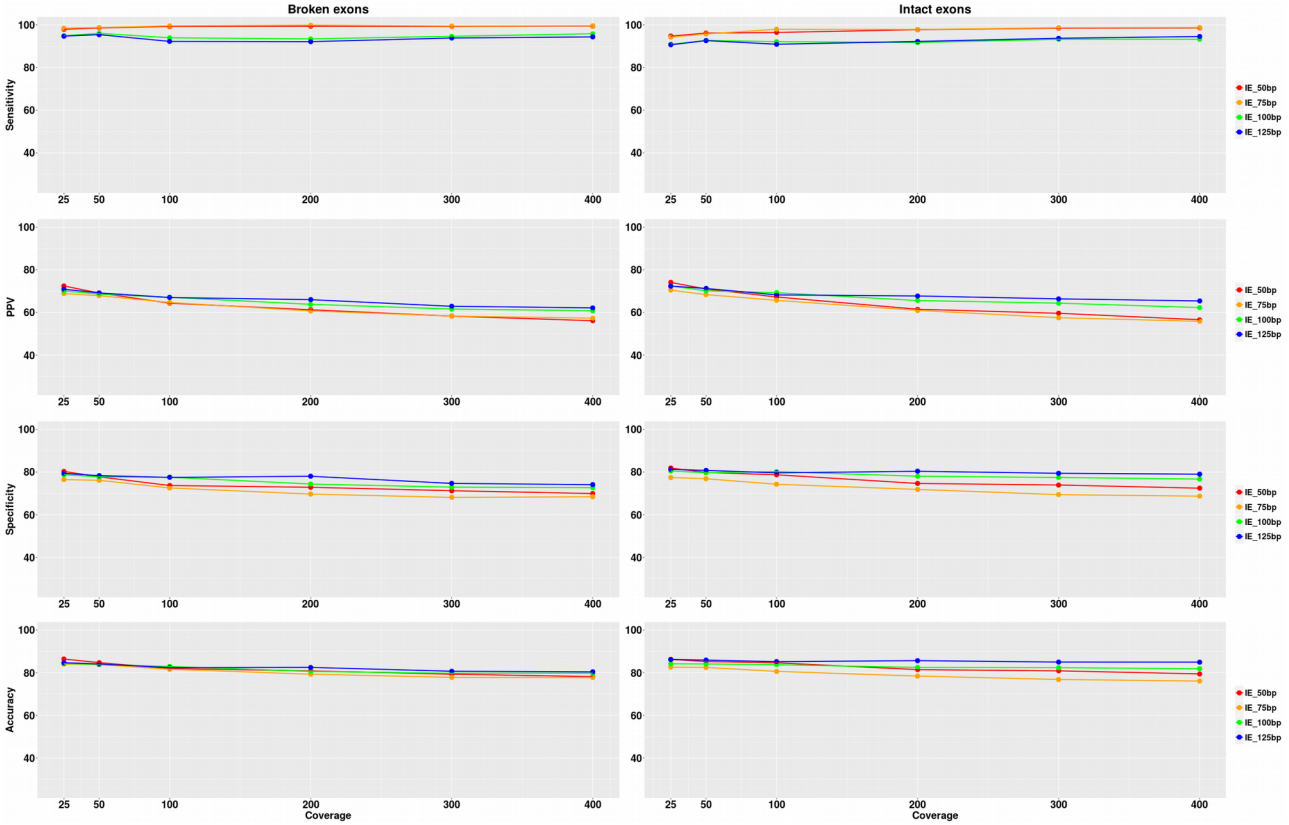


**Figure 14. Sensitivity, PPV, Specificity and Accuracy distribution for Fusion Validator on synthetic datasets of different coverage, read length and breakpoint position.**

The rows of the plot panel shows distribution of Sensitivity, PPV, Specificity and Accuracy.

The columns of the panel measure distribution of different metrics in broken and intact exons database respectively.

X-axis shows the trend of each distribution for different read coverage.

Read lengths are represented with different colors (red for 50bp, yellow for 75bp, green for 100bp and blue for 125bp).

### *2.5.3 Fusion Validator demonstrates a better combination of sensitivity and precision compared to other fusion detection tools*

Fusion Validator showed the best average sensitivity among different fusion detection tools in BE datasets and the second best after Defuse in IE subsets, with only a 1.19% difference in average sensitivity (94.79% for Fusion Validator vs 95.98% for Defuse). Fusion Validator was the second best performer in terms of average PPV for both IE and BE dataset, after MapSplice (66.01% vs 76.50% in IE datasets and 64.50% vs 80.45% in BE datasets for Fusion Validator and MapSplice respectively). However, MapSplice was able to detect only an average of 65.35% of known fusions for IE subsets and an average of 83.52% for BE, with large range of variability among different coverage and read length datasets (Figure 15A-B).

To compare the overall performance of Fusion Validator with results from other fusion detection software, the F measure was calculated for each simulated dataset to summarize sensitivity and PPV with a single standardized index. Distribution of F measure for different fusion finder tools showed that Fusion Validator achieved the best performance for IE datasets, with an average F measure of 77.64%, and an average F measure of 77.31% for BE dataset, that represent the second best score after MapSplice (average F measure for MapSplice 81.51%) (Figure 15C-D). However, as reported above, despite the high PPV, MapSplice's performance in term of sensitivity is not in the same range of Fusion Validator, with a probability of correctly identify a real fusion that is 13.23% less than Fusion Validator.

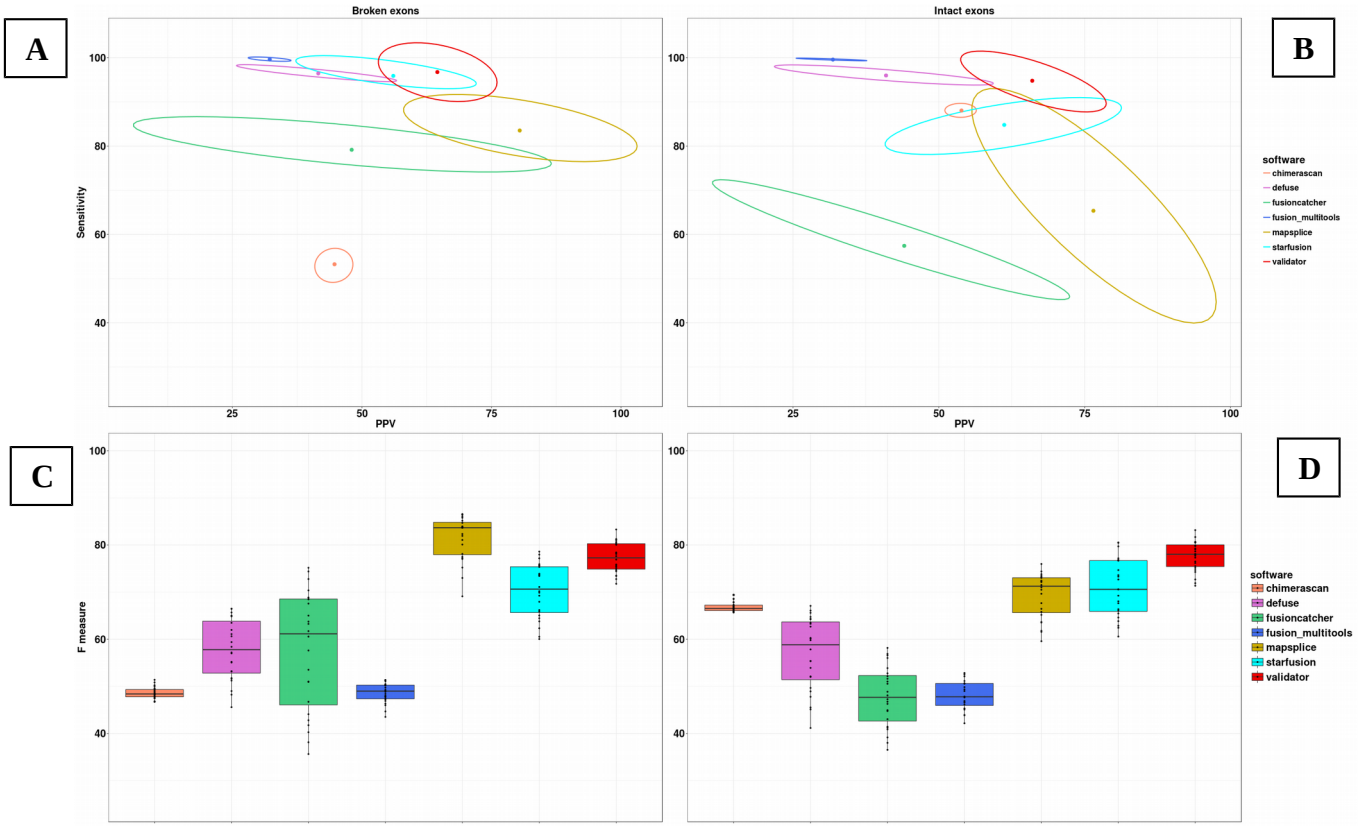


**Figure 15. Fusion detection performance evaluation across different softwares for synthetic datasets.**

The scatterplots display the relation between Sensitivity and PPV in broken exons (A) and intact exons (B) databases for 5 different fusion detection tools (Defuse, Chimerascan, STAR-fusion, MapSplice and FusionCatcher), multi-tools approach and Fusion Validator.

Colored dots point out the average Sensitivity and PPV for each software, while 95% confidence intervals for Sensitivity and PPV is represented by colored ellipses.

Boxplots show the distribution of F-measure in broken exons (C) and intact exons (D) datasets.

## 2.6 Test set analysis

### *2.6.1 Fusion detection in breast cancer Cell Lines*

To assess the ability of Fusion Validator to validate real known fusion transcripts from publicly available data sets, and evaluate its abilityto reduce the number of candidate events detected by different fusion finder algorithms, 27 experimentally validated fusions from 4 different breast cancer cell lines (BT-474, SK-BR-3, KPL-4, and MCF-7), identified by Edgren et al., were used as a true positive set.

Results in table 10 show that none of the 5 fusion detection tools used (Defuse, Chimerascan, STAR-fusion, MapSplice, FusionCatcher) identified all 27 known fusions. The number of true positive calls ranged from a minimum of 19/27 (70.37% sensitivity) in MapSplice to a maximum of 24/27 (88.89% sensitivity) using Chimerascan. To detect all 27 experimentally validated chimeric transcripts, a combination of fusion calls from the 5 different software were performed by the first component of Fusion Validator. This step merged 5,494 fusion candidates, which were subsequently reduced to 3,190 after the normal filtering, and finally further reduced to 1,134 after Fusion Validator's final step. This represent a striking 79.36% reduction in the number of false positive events (PPV for Fusion Validator 2.29% vs 0.47% for combined tools).

Strikingly, Fusion Validator was able to correctly validate 26 out of 27 known fusions, with a sensitivity significantly higher than all the other fusion finder algorithms (sensitivity 96.30%). Chimerascan and MapSplice show a PPV slightly higher to that of Fusion Validator (2.29% for Fusion Validator vs 2.88% for Chimerascan and 2.97% for MapSplice) but these tools were able to

correctly identify only 24/27 and 19/27 known fusions respectively. FusionCatcher is the software with the higher PPV (43.14%), but it shows a lack in sensitivity, with only 22/27 fusions correctly identified (sensitivity 81.48%) (Table 11).

| Cell line | Defuse | Chimerascan | STAR-fusion | FusionCatcher | MapSplice | Combined tools | Fusion Validator | Total fusions detected | Total fusions after Normal filter | Total fusions after Validator filter |
|---|---|---|---|---|---|---|---|---|---|---|
| BT-474 | 10/11 | 10/11 | 10/11 | 9/11 | 9/11 | 11/11 | 11/11 | 2556 | 1461 | 538 |
| SK-BR-3 | 7/10 | 8/10 | 7/10 | 7/10 | 5/10 | 10/10 | 9/10 | 1930 | 1111 | 403 |
| KPL-4 | 2/3 | 3/3 | 3/3 | 3/3 | 2/3 | 3/3 | 3/3 | 456 | 261 | 73 |
| MCF-7 | 2/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 552 | 357 | 120 |
| Total | 21/27 | 24/27 | 23/27 | 22/27 | 19/27 | 27/27 | 26/27 | 5494 | 3190 | 1134 |

**Table 11. Number of fusion transcript detected by 5 different softwares and filtered by Fusion Validator on 4 BRCA Cell lines.**

| Software | Sensitivity | PPV |
|---|---|---|
| Defuse | 77.78% | 0.74% |
| Chimerascan | 88.89% | 2.88% |
| STAR-fusion | 85.18% | 1.67% |
| FusionCatcher | 81.48% | 43.14% |
| MapSplice | 70.37% | 2.97% |
| Combined tools | 100% | 0.47% |
| Fusion Validator | 96.30% | 2.29% |

**Table 12. Average Sensitivity and PPV of 5 different fusion detection softwares, multi-tool approach and Fusion Validator on 4 BRCA Cell lines.**

### *2.6.2 Fusion Validator easily identified driver kinase fusions in TCGA pan cancer data*

To see if our method was able to correctly detect fusions from a larger cohort of different tumour types and validate additional chimeric transcripts not found or not reviewed in previous studies, we extended the analysis to 190 TCGA pan-cancer samples carrying 195 validated recurrent kinase fusions (115 unique) from Stransky et al.. This dataset was reanalyzed using a combination of 3 fusion detection tools (Defuse, Chimerascan and STAR-fusion) and processed through the Fusion Validator pipeline.

The multi-tool chimeric transcript detection approach was able to select an average of 4904 fusion candidates per sample (95% confidence interval 4635-5174), that were reduced to an average of 2945 (95% confidence interval 2755-3136) and 983 (95% confidence interval 900-1067) per sample after normal filtering and local realignment validation step respectively, with a significant reduction of 79.95% of the candidate fusions. 191 out of 195 (97.95%) recurrent validated fusions were also confirmed in silico by Fusion Validator, with the four missing chimeric transcripts not found by any of the 3 fusion finder algorithms used, and, therefore not been processed by the Fusion Validator pipeline (list of fusions in Appendix A). One of the 4 missed transcripts is a complex fusion affecting *PDGFRA* and overlapping genes *FIP1L1* and *LNX1* in Brain Lower Grade Glioma sample TCGA-E1-A7YI-01A. For this sample STAR-fusion detected a chimeric transcript involving *FIP1L1*, but with a different partner *CHIC2*. The *CHIC2* gene overlap with the longer isoform of *PDGFRA,* which is present only in UCSC gene models (uc003haa.3) and not on the Gencode19 annotation used for the analysis: The *FIP1L1-CHIC2* fusion can thus be considered a product of different annotation databases rather than an event missed by fusion finder algorithms (Figure 16).

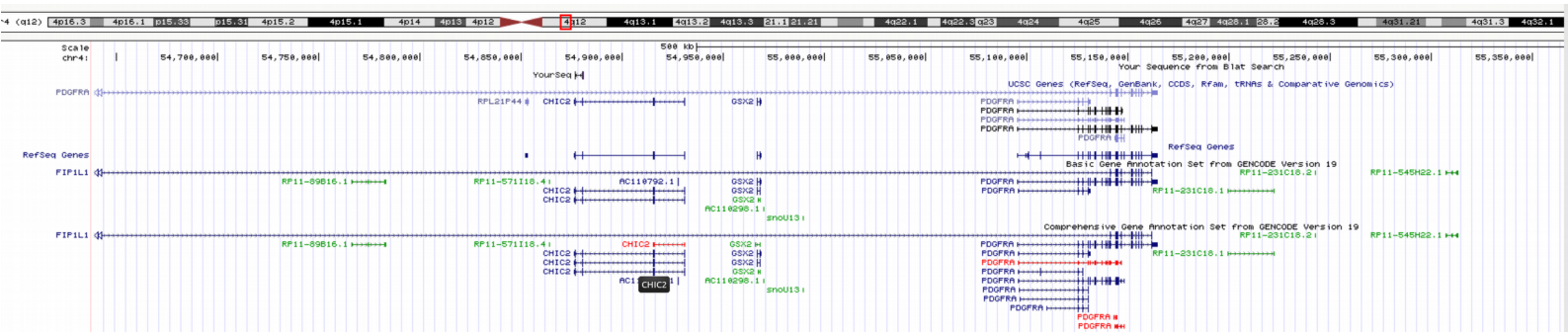


**Figure 16. UCSC Genome Browser snapshot showing overlap between gene *CHIC2* and the longest isoform of *PDGFRA* on the fusion transcript breakpoint position.**

Considering the list of non-recurring chimeric transcripts detected by Stransky et al. in the 190 samples, 2/2 validated fusions, and 22/30 non reviewed events were also confirmed by Fusion Validator (Appendix A). Fusion Validator was also able to confirm and validate an additional 5384 kinase fusions removed by the different filtering steps applied by Stransky et al., 791 of them recurrent in more than 1 of the 190 analyzed samples. This list includes recurrent driver fusions like *FGFR3-TACC3* and *RAF1-AGGF1*, that were validated by Stransky et al. In 15 (5 Bladder Urothelial Carcinoma, 3 Lung squamous cell carcinoma, 2 Glioblastoma multiforme, 2 Head and Neck squamous cell carcinoma, 2 Brain Lower Grade Glioma and 1 Kidney renal papillary cell carcinoma) and 7 samples respectively (7 Thyroid carcinoma) and were confirmed by Fusion Validator in additional 2 (2 Glioblastoma multiforme) and 1 samples (1 Thyroid carcinoma) respectively.

Sorting the final annotated list of fusions for each sample by ranking score, in descending order, we found that 151 out of 191 (79.06%) recurrent fusion transcripts validated by Fusion Validator were classified by the annotation

module with the best ranking score, thus demonstrating a high probability of being classified as driver fusions. Since Stransky et al. focused their analysis only on kinase fusions, the percentage of fusions classified with the best ranking score by Fusion Validator is affected by the presence of no kinase driver fusions, as, for example, in sample TCGA-33-4587-01A, where the highest annotation score is assigned to a non-kinase activating fusion between *PRSS33* and *CREBBP*, while the chimeric transcript involving *IGF2BP3* and *PRKCA*, validated by Stransky et al., has only the second best score. Considering only kinase fusions, the percentage of chimeric transcripts classified with the best ranking score by Fusion Validator increases to 81.67%.
For some of the samples in Stransky's dataset, more than one validated kinase fusion was found. For example in sample TCGA-CH-5737-01A, both the experimental validated fusions, *KDM7A-BRAF* and *AGGF1-RAF1*, were in confirmed in silico by Fusion Validator and classified with the best and second best annotation score, respectively.

To correctly evaluate the performance of Fusion Validator's ranking score on samples with multiple recurrent kinase fusions, the percentage of validated fusions in the top 3 and top 5 rank was considered as a more accurate metric.
180 out of 191 (94.24%) and 190 out of 191 (99.48%) validated kinase fusions received a ranking score in the top 3 and top 5 respectively (Figure 17) (Appendix A). This strong correlation between ranking score and probability of a fusion to be validated, support the idea that, to quickly identify fusion transcripts for downstream analysis, it's sufficient to sort a Fusion Validator output by ranking score and select the top 3 validated candidates. In doing so, one has a ~95% sensitivity in detecting a real kinase driver fusion.

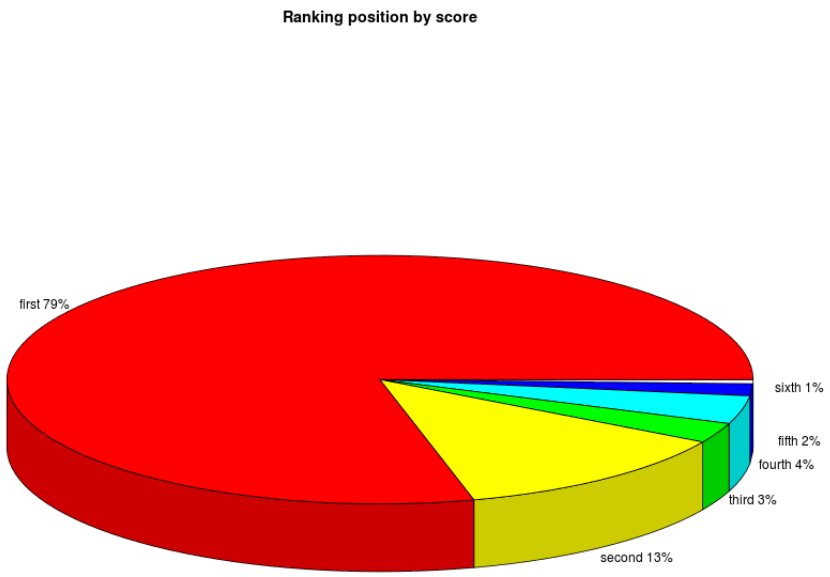


**Figure 17. Pie chart with distribution of Fusion Validator ranking positions of 191 recurrent validated kinase fusions.**

## 2.7 Discussion

In the last few years, the incorporation of RNA-Seq molecular tools into clinical diagnostics has turned out to be an attractive instrument to identify disease and patient specific biomarkers that are tractable targets for therapy. The need to detect driver gene fusions on large scale datasets have stimulated a remarkable effort to develop new bioinformatics algorithms for identifying gene

fusions from sequencing data, each of them showing different levels of sensitivity. Meta caller or so called ‘ensemble approaches’ proposed in the literature are useful to explore a wide-ranging panel of potential real fusions, but dramatically increase the number of false positive events. The choice of proper standardized filtering steps that maximize sensitivity and specificity, and the use of algorithms for candidate fusion prioritization, is crucial to quickly identify driver oncogenic fusions for downstream analysis and wet-lab validation.

To solve the main challenges in clinical validation of fusion transcripts coming from NGS experiments, we developed Fusion Validator, an optimized pipeline tool designed to collect thousands of fusion transcripts detected by different fusion finder algorithms and to identify clinically relevant fusion genes, with a significant 79.95% reduction of the number of candidates that need to be subsequently assessed through experimental validation.

Fusion Validator recreates the chimeric transcript sequence around the fusion breakpoint and performs a number of filtering steps, including the removal of read through transcripts and genes from user defined list of invalidated events, local realignment of normal tissues sequences on fusion transcripts and flagging of fusions in homologous and repetitive regions around the breakpoint. Validation of each fusion transcript sequences is carried out through a local realignment of reads around the fusion breakpoint and a combination of both a de novo assembly and a seed and extend mapping of read candidates to reconstruct the breakpoint. A ranking score based on annotation is assigned to the final list of filtered and validated transcripts. Fusion Validator can work with canonical fusions as well as other somatic structural changes in the transcriptome, like exon skips, and can take advantage of different new features that provide major accuracy improvements over traditional fusion detection and prioritization algorithms:

The combined tool approach implemented by Fusion Validator, can take in input lists of fusions from the most common fusion detection tools and easily merge results, by simply aggregating basic standard information like genes, breakpoint locations and strand orientations. This approach demonstrated a highly significant increase in sensitivity for both real and simulated datasets analyzed compared to other fusion detection algorithms. Moreover, Fusion Validator can extract additional fusion information not given by the majority of fusion detection tools. Each fusion transcript sequence is retrieved using one set of rules, independent of the software from which it was selected, in an attempt to eliminate any bias related to different references and annotations. Another distinguishing feature of Fusion Validator is the use of a dynamic normal filtering step, that realigns all of the control samples from human normal tissues against every putative aberrant junction, instead of processing the tumour and normals separately: Comparing putative junctions against a multi-tissue dataset, using raw reads, was demonstrated to be the most comprehensive way of reducing false positives. The use of a dynamic blacklist, to store normal chimeric transcripts detected by Fusion Validator, provides a big advantage in term of computational costs, as candidate fusions that have been found in normal tissues before are not re-processed, and in terms of specificity, as the normal filter's performance continues to improve as additional samples are processed. Ultimately, the use of a blacklist has allowed us to quickly discard more than a half of the initial fusions candidates. The local realignment of tumour reads to reconstruct the chimeric transcript provides a robust and reliable validation of each candidate event, since reads that perfectly encompass the breakpoint are selected first, and then reassembled with different independent methods (de novo assembly and seed and extend scaffold reconstruction). The use of a genomic realignment filter option ensures an additional reduction of false positive events caused by alignment errors or reads mapping on low complexity or homologous regions. The final ranking score assigned by the annotation module of Fusion Validator to each chimeric

transcript that passed the filtering steps, can be used by clinicians and researchers to quickly identify events of interest for downstream analysis: It's sufficient to look at the top 3 fusions for each sample, and without any experimental validation, to have a 95% probability to find a real driver oncogenic fusion.

Fusion Validator performs well in terms of overall accuracy across different experimental conditions, showing an excellent combination of sensitivity and specificity, even for low coverage and short reads sequencing. This advantage leads to reducing the cost of sequencing per patient, since it has been shown that a coverage increase does not bring particular advantage to the number of fusions successfully detected, and lack in coverage can be easily compensated by using additional fusion detection tools as input for Fusion Validator to retain a high level of sensitivity. The characteristic of maintaining a very high accuracy at reduced sequencing costs, make Fusion Validator an ideal diagnostic tool in precision medicine research applied to oncology patients, where gene fusions are critical as diagnostic and prognostic factors.

## 2.8 Software characteristics

Fusion Validator is implemented in Perl and Unix and It has been tested on a Linux high performance compute cluster (Centos distribution 6.8) with job-scheduling software Portable Batch System (PBS).

# Chapter 3: Direct transcriptional consequences of somatic mutation in cancer

## 3.1 Introduction

*3.1.1 Somatic mutations can amplify the transcriptional output in cancer cells.*

Somatic mutation underpins the development of cancer, and most solid tumours have thousands to tens of thousands of point mutations, coupled with tens to hundreds of genomic rearrangements and copy number changes[2,53,54]. Small numbers of these, known as 'driver mutations', dysregulate the fundamental cellular processes involved in normal tissue homeostasis, and confer a selective advantage to the clone. A critical point is that Darwinian selection acts on *phenotype* and so, for a somatic mutation to drive cancer, it must manifest a phenotypic effect. Transcription is the primary conduit by which changes in the genomic code are translated into cellular phenotype, with the corollary that it is a necessary criterion of driver mutations that they directly induce a change in transcript structure. Altered transcript structure can take many forms, including the creation of fusion genes by genomic rearrangement, interference with RNA splicing at mutated splice sites, alteration of the codon sequence for missense substitutions and over or under-expression of genes through copy number alterations or mutation in regulatory regions.

Beyond the primary and direct effects of somatic mutation on transcript structure, there may be a series of downstream, secondary alterations in the transcriptome occurring as a consequence of the primary effect. Most studies of the transcriptome in cancer, including those from large-scale efforts such as TCGA[45,55], have evaluated these second-order effects, concentrating predominantly on the magnitude of gene expression using microarray tecnologies[56-58] or RNA sequencing[59,60]. They have revealed large-scale disturbances of transcriptional regulation in most cancers, with expression profiles for many hundreds of genes differing from profiles of normal cellular

counterparts. Within a tumour type, similarities in transcriptional profiles across individuals allow the disease to be sub-classified into several groups, many of which have biological, therapeutic and prognostic significance. In some cases, these changes can be correlated with underlying driver mutations, such as *ERBB2* amplification in Breast Cancer[58] or specific fusion genes in acute myeloid leukaemia[61]. While these studies have concentrated on mRNA profiles, similar observations are beginning to emerge from studies of MicroRNA transcription[62], long non-coding RNA levels and even expression of pseudogenes[63]. While it is a necessary criterion for a driver mutation to directly induce modification of transcript structure, it is not sufficient. Many mutations that do not confer selective advantage, so-called passenger mutations, will also generate phenotypic consequences, but consequences of no benefit to the cell. Initial studies correlating RNA-sequencing data with genomic change in cancer have reported some of these direct effects, especially for coding point mutations or canonical fusion transcripts[59] but there has been little systematic effort to describe, measure and quantify first-order transcriptional consequences across all classes of somatic mutation found in well-annotated cancer genomes.

### *3.1.2 Transcriptional Amplification as target therapy*

Transcriptional amplification is a phenomenon by which certain oncogenes contribute to tumour progression by increasing a cell's global production of RNA and, consequently, the entire transcriptome of a cell increases in expression. Cancer cells have been shown to become heavily reliant on elevated levels of transcriptional activity, making them highly sensitive to targeted therapies that selectively inhibit transcription. For this reason transcriptional amplification represents a novel, and demonstrably targetable

feature of cancer, where precision therapies often fail, because cancer cells have the unique ability to adapt to changing environmental pressures, and targeting oncogenic pathways can ultimately select for cancer cell subpopulations that have evolved to obsolesce those particular pathways for survival. However, the understanding of transcriptional amplification's prevalence and its implications for malignant progression is still underdeveloped.

The second sub-project of this thesis was focused on developing a bioinformatics method to measure the transcriptional output of human primary tumour cells, using a validation set of nearly a thousand of TCGA Breast Cancer patients, in order to have a comprehensive understanding of transcriptional amplification in cancer cell biology.

## 3.2 Material and Methods

### *3.2.1 TCGA Breast Cancer dataset*

Aligned BAM files for 980 breast cancer samples with both RNA-Seq and exome sequencing were downloaded from CGHUB (https://cghub.ucsc.edu/) using GeneTorrent. PCR duplicates for both exome and transcriptome were removed using SAMtools. The position of somatic mutations, in MAF file format, and gene expression values (using the RSEM method) were obtained from https://tcga-data.nci.nih.gov/. Additional clinical covariates were obtained from cBioPortal (http://www.cbioportal.org/) (Figure 18). All putative mutations were re-annotated using Annovar (release 2013Aug23) and all potential germline variants were removed (present in NCBI dbSNP Human build 142). Finally, 70,071 exonic/splicing substitutions present in the 980 RNA-Seq and WES paired samples were considered for further analysis. Mutations in the 5' or 3' UTRs were excluded.

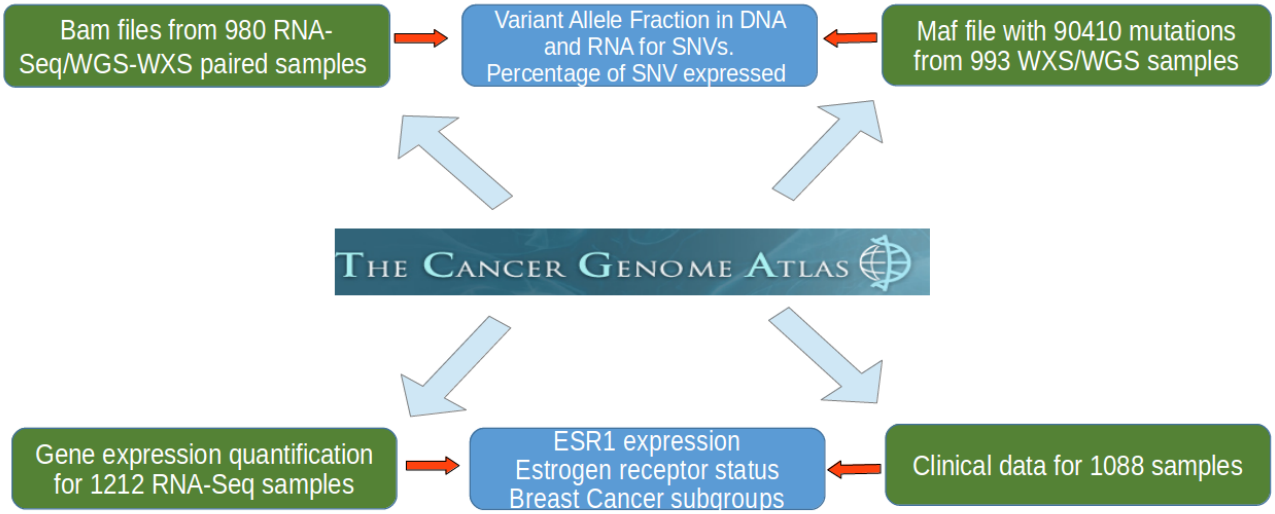


**Figure 18. Schematic overview of information downloaded from TCGA dataset for 980 Breast Cancer samples (green boxes) and extracted from starting raw data (light blue boxes).**

### *3.2.2 Analysis of variant allele fraction differences between the transcriptome and genome in TCGA data*

A bioinformatcs software was developed to accurately measure the number of bases supporting each mutation in the genome (or exome) and in the corresponding transcriptome for each of the 70,071 somatic mutations selected. The tool parses the sequencing alignment file for both DNA and RNA samples, extracts all the sequence bases aligning in each of the selected somatic point mutations and calculates the Variant Allele Fraction (VAF) for

each position as the frequency of variant alleles in a given locus, using the formulas

$$VAFdna = \frac{nVARdna}{nREFdna + nVARdna} \qquad VAFrna = \frac{nVARrna}{nREFrna + nVARrna}$$

,where nVARdna is the number of variant calls for the specific base in the DNA alignment, nREFdna is the number of reference calls for the same base, and nVARrna and nREFrna are the number of variant calls and the number of reference calls for the specific base in the RNA alignment respectively. To estimate the effect of somatic mutations on transcription, the proportion of sequencing reads supporting the mutant allele in the transcriptome was compared to that expected from the genome. This proportion was measured as the difference between VAF in the transcriptome and in the genome ($VAF_{difference} = VAF_{transcriptome} - VAF_{genome}$) (Figure 19). Mutated loci were considered as not expressed, and therefore excluded from the analysis, if the total coverage was less than five reads, or the number of reads supporting the mutated base was less than five reads. The total number of somatic variants considered for further analysis after filtering steps is 25,177 , corresponding to 955 patients (Figure 20). Information about variant allele fraction differences and percentage of expressed mutations were merged with clinical data and gene expression quantification. Linear regression was used to model the relationship between the amount of ESR1 expressed by a tumor and the VAF difference of its mutations. Survival data was analyzed using the Kaplan-Meier and log-rank Mantle-Cox methods. The limit of significance for all analyses was defined as $P < .05$. TCGA Breast Cancers were classified into known subtypes (Luminal B, Luminal A, HER2-related and triple negative) by immunohistochemistry according to Blows et al[63] (Figure 21).

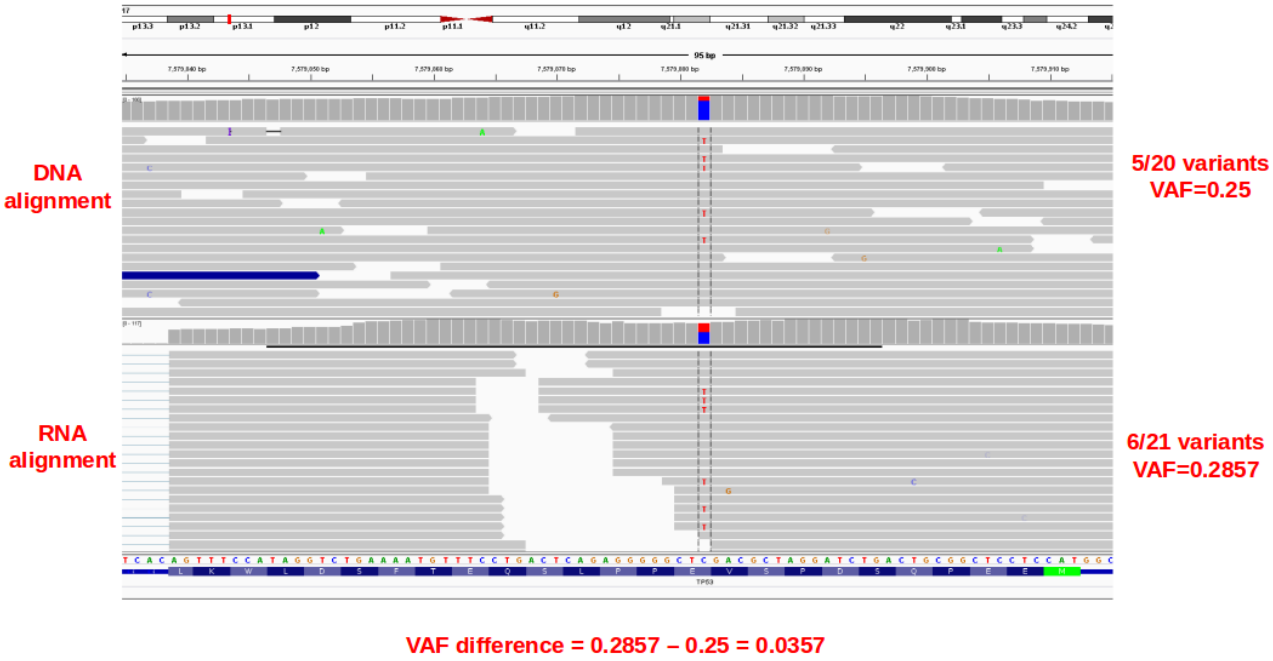


**Figure 19. Example of somatic C -> T mutation in TP53 gene alignments for DNA (on the top) and RNA (on the bottom).** For both DNA and RNA Variant Allele Fraction is calculated as the number of variants (base T in orange) divided by the number of total reads. Difference between VAF in RNA and DNA is used as a measure of transcriptional amplification for the specific mutation. Grey bars represent each aligned read and variant base T is highlighted in red.

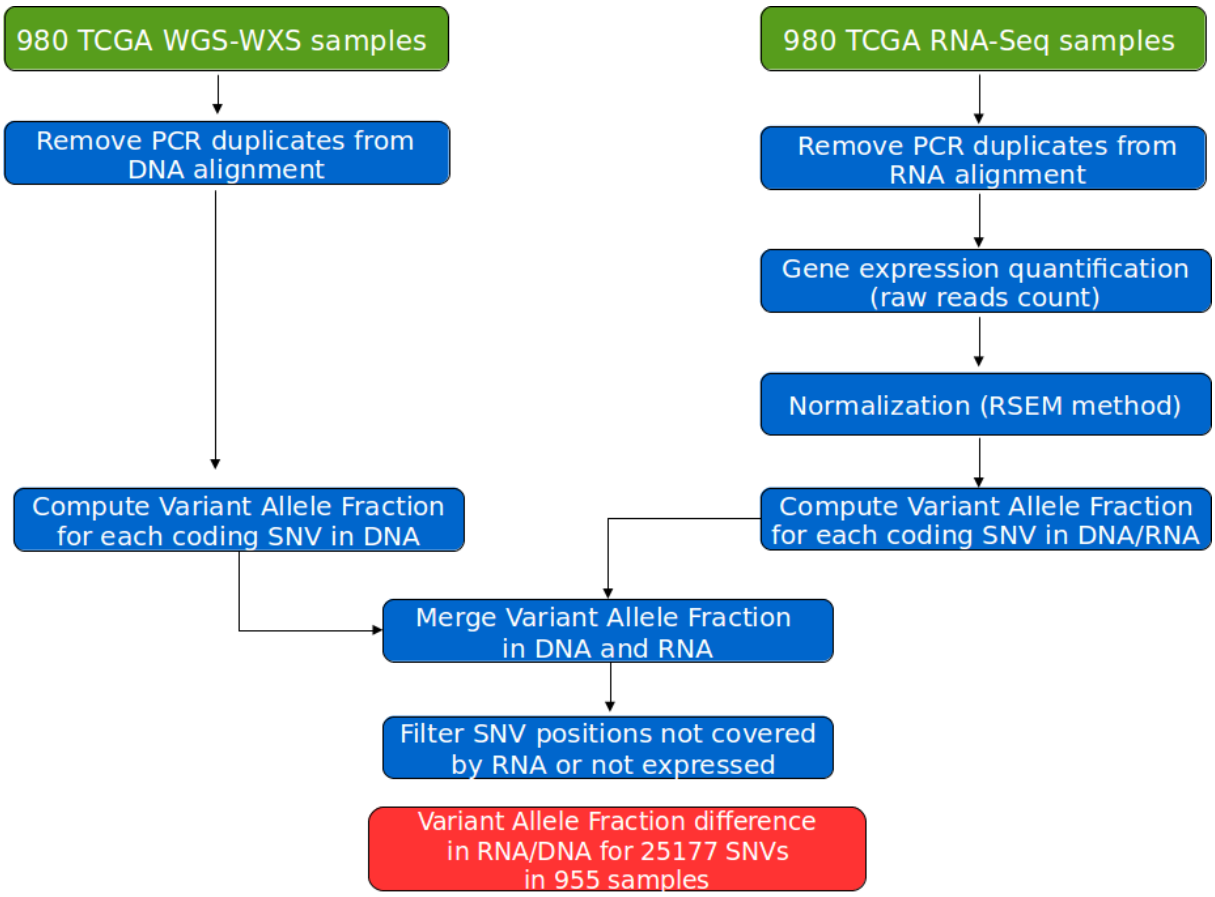


**Figure 20. Transcriptional amplification analysis workflow for 980 TCGA DNA and RNA sequencing from Breast Cancer patients.** 980 samples for which both DNA and RNA sequencing was available were analyzed. PCR duplicates were removed from both DNA and RNA alignments and variant allele fraction for each coding SNV was computed in both datasets. The proportion of sequencing reads supporting the mutant allele in the transcriptome compared to that expected from the genome was estimated as the difference between VAF in RNA and DNA. 25,177 SNVs belonging to 955 samples were selected after filtering positions not covered or not expressed in RNA.

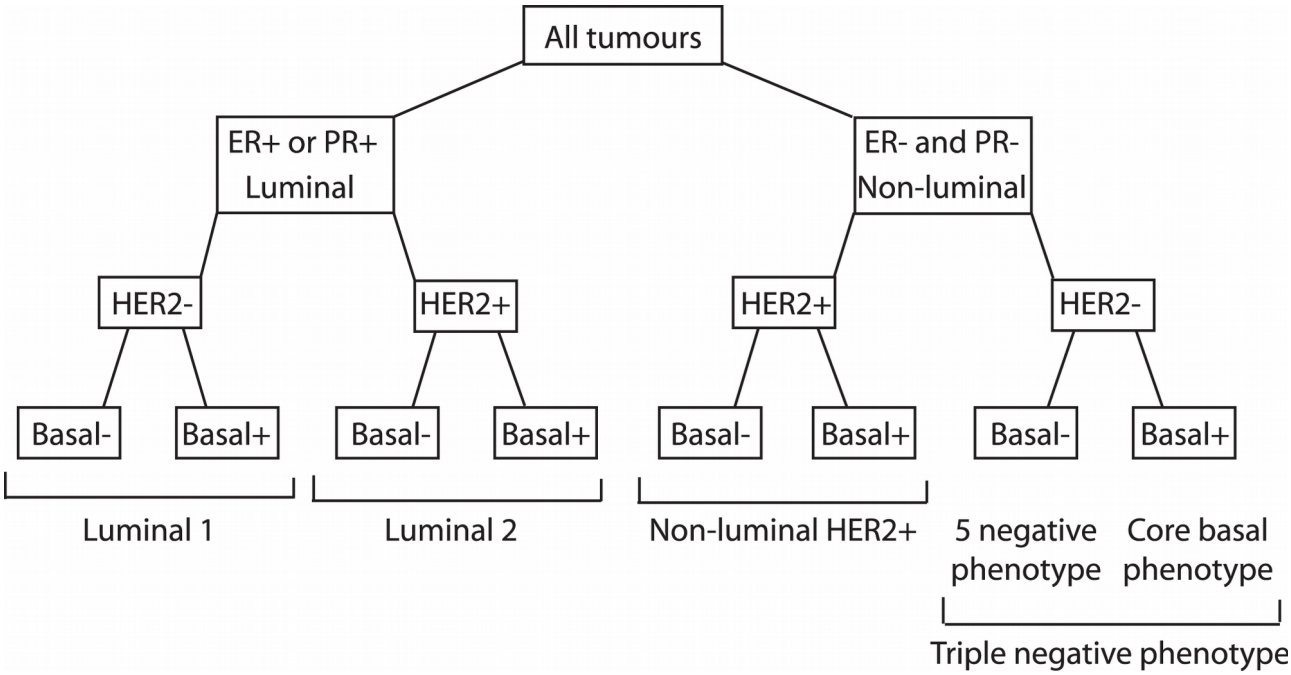
All tumours
ER+ or PR+
Luminal
ER- and PR-
Non-luminal
HER2-
HER2+
HER2+
HER2-
Basal-
Basal+
Basal-
Basal+
Basal-
Basal+
Basal-
Basal+
Luminal 1
Luminal 2
Non-luminal HER2+
5 negative
phenotype
Core basal
phenotype
Triple negative phenotype

**Figure 21. Classification of breast cancer subtypes according to IHC marker profile (Figure from Blows et .al)[64].** Each Breast Cancer subtype is characterized by the expression of immunohistochemistry markers Estrogen receptors (ER), Progesterone receptors (PR) and Human epidermal grow factor receptor-2 (HER2)

## 3.3 Results

### *3.3.1 On average exonic point mutations are expressed to the level of would expect from their prevalence in the genome.*

25,177 somatically acquired base substitutions in 955 breast cancer samples were covered in both the genome and the transcriptome (>4 reads) and were found to be expressed (>4 variant reads present in the transcriptome). The average percentage of mutations expressed in the transcriptome is about 60%. Variant allele fraction (VAF) in the genome and the transcriptome are strongly positively correlated (Pearson's correlation coefficient =0.6439, p-value<0.0001 Figure 22). To estimate the effect of somatic mutations on transcription, the proportion of sequencing reads supporting the mutant allele in the transcriptome was compared to that expected from the genome. This proportion was measured as the difference between VAF in the transcriptome and in the genome.

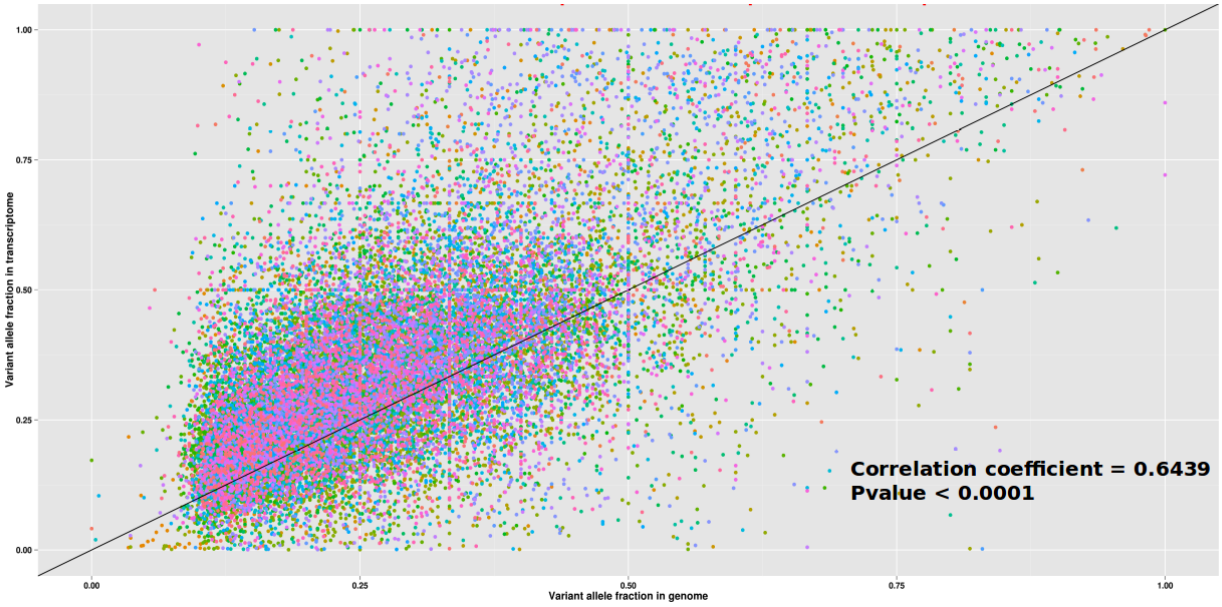


**Figure 22. Variant Allele Fraction comparison in RNA-Seq and DNA for all protein coding mutations.** X and Y axis shows the distribution of VAF in genome and transcriptome respectively. Each dot of the scatterplot represent a single nucleotide variant and is colored by sample. Black regression line represent the theoretical condition of maximum linear correlation between VAF in DNA and VAF in RNA

There were some differences in the transcription levels of base substitutions according to the predicted consequence on the protein. Silent, missense and UTR mutations have the same strong correlation between variant allele fractions in the genome and transcriptome, whereas nonsense mutations have a weaker relationship. Indeed, nonsense mutations had a significantly lower expression than predicted from the genome compared to other classes of mutation ($p<0.0001$). Several reasons could explain the lower expression of nonsense mutations. Nonsense-mediated decay could selectively target transcripts with nonsense mutations for degradation. Nonsense-mediated decay depends upon the cell distinguishing a premature termination codon from a proper termination codon. Generally, stop signals in the last exon are

considered proper, whereas those appearing more than 50-55bp upstream of the last exon-exon junction, and therefore upstream of the exon-junction complex, are more likely to be targeted for nonsense-mediated decay . Another possible explanation for the low expression of nonsense mutations is that they are tolerated only in genes not expressed in the cancer cells, those occurring in important genes would be subject to negative selection.

To explore this possibility, the expression levels from the organoids of normal breast epithelium for genes mutated in the cancer samples was compared. No clear-cut differences across the mutation categories for whether the mutated genes were found to be expressed in normal breast epithelial cells (Figure 23), suggesting that this reason does not explain the lower expression levels of nonsense mutations. Therefore, it appears as if only nonsense mediated decay explains the lower expression of these mutations.

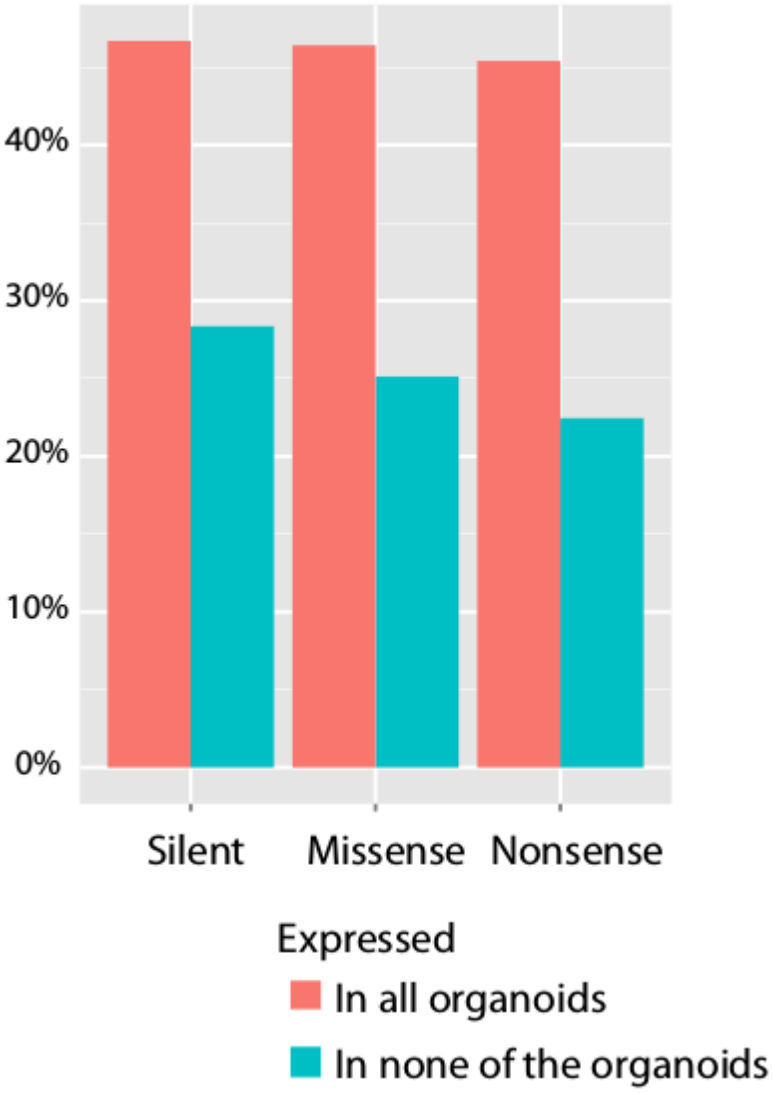


**Figure23. Absence of negative selection in nonsense mutations.**

Comparison of expression levels from the organoids of normal breast epithelium for genes mutated in the cancer samples.

*3.3.2 Expressed mutations are significantly inversely correlated to the Estrogen receptor levels*

The average VAF in the transcriptome relative to the genome is greater and significantly different in estrogen receptor positive patients (0.12669) compared to estrogen receptor negative ones (0.06353) (p-value <0.0001, Figure 24). The average VAF difference between the transcriptome and the genome is significantly negatively correlated with the expression level of gene that encodes the estrogen receptor, ESR1 (Pearson's correlation =-0.2669, p-value<0.0001), and significantly positively correlated with the percentage of mutations expressed in transcriptome (Pearson's correlation =0.1074, p-value=0.0009). The percentage of mutations expressed is also significantly positively correlated with ESR1 gene expression (Pearson's correlation=0.0725, p-value=0.0251) (Table 13).

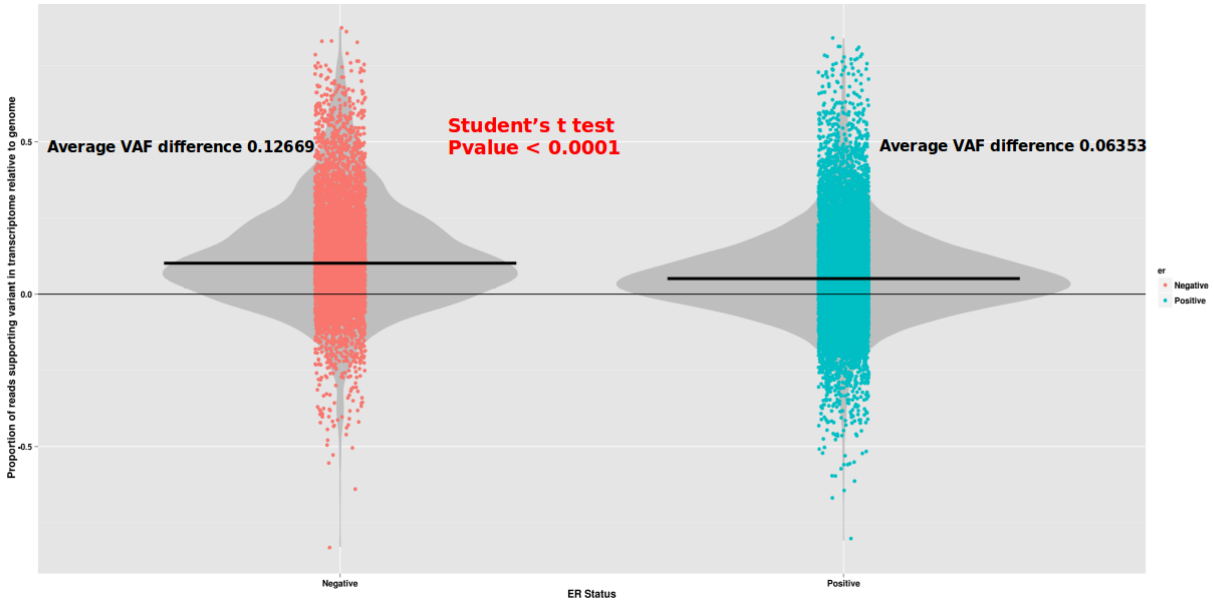


**Figure 24. Variant allele fraction in transcriptome relative to genome distribution for ER negative (red) and ER positive (light blue) samples.**

| | Average VAF Difference | ESR1 expression | # of mutations | % Mutation Expressed |
|---|---|---|---|---|
| **Average VAF Difference** | 1.0000 | -0.2669 (p<0.0001) | -0.0182 (p=0.5735) | 0.1074 (p=0.0009) |
| **ESR1 expression** | | 1.0000 | -0.0571 (p=0.0778) | 0.0725 (p=0.0251) |
| **# of mutations** | | | 1.0000 | -0.0619 (p=0.0558) |
| **% Mutation Expressed** | | | | 1.0000 |

**Table 13. Pairwise Pearson correlation for VAF difference, ESR1 gene expression, number of mutations and percentage of mutation expressed in 955 Breast Cancer samples.**

The VAF difference between the transcriptome and the genome can be estimated with a linear discriminant function of 2 variables: ESR1 expression and percentage of mutation expressed. In Figure 25, if ESR1 expression decreases, the VAF difference increases proportionally and the percentage of mutations expressed increase as well. That is, tumours with high levels of ER express fewer mutations than cancers with low ER. This relationship can be formally modelled as: for every 1% decrease in ESR1 expression, 15 more mutations are expressed in breast cancer.

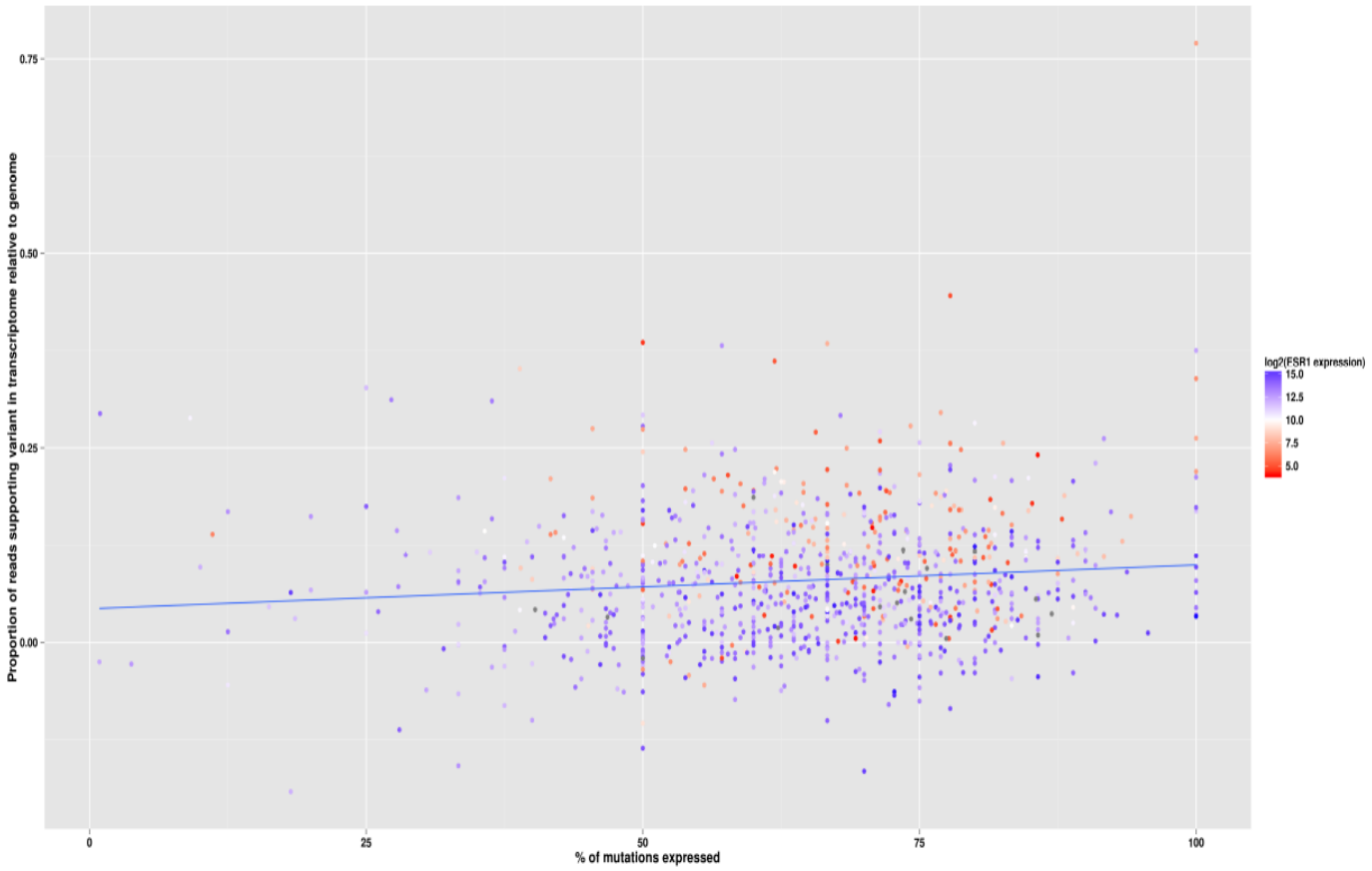


**Figure 25 Regression model for average Variant Allele Fraction difference and percentage of expressed mutations, stratified by ESR1 expression.**

Expression of ESR1 gene is represented in log scale from lower (red) to higher expression values (blue). VAF difference is directly proportional to the percentage of mutations expressed and inversely proportional to ESR1 expression

### *3.3.3 Transcriptional amplification differs in the most common breast cancer subtypes*

We stratified breast cancer patients into different subgroups by their immunohistochemistry staining for estrogen, progesterone and HER2 receptors. The distribution of percentage of mutation expressed and VAF difference between the transcriptome and the genome was compared for the four main breast cancer subgroups: Triple Negative Breast Cancer (TNBC), HER2 positive, Luminal A and Luminal B. The VAF difference in the transcriptome relative to the genome, as well as percentage of mutations expressed tends to increase if the breast cancer subgroups with better clinical outcome (Luminal A/B) are compared to subgroups with the worst prognosis (HER2 positive and TNBC). Average percentage of mutations expressed range from 61.28% of the Luminal B subgroup to 70.31% of TNBC (Figure 26).

Breast cancer patients were divided in two groups according to a threshold of 0.045504 for VAF difference, selected as the value that maximizes the sum of sensitivity and specificity in the ROC curve. Patients with high VAF difference between the transcriptome and the genome had a better prognosis than the ones with the low VAF difference group with median overall survival of 114 and 90.8 months, respectively. Survival curves computed using the Kaplan-Meier method shows a statistically significant survival difference between high and low VAF difference groups, with a 0.05 p-value for log rank Mantel-Cox test (Figure 27).

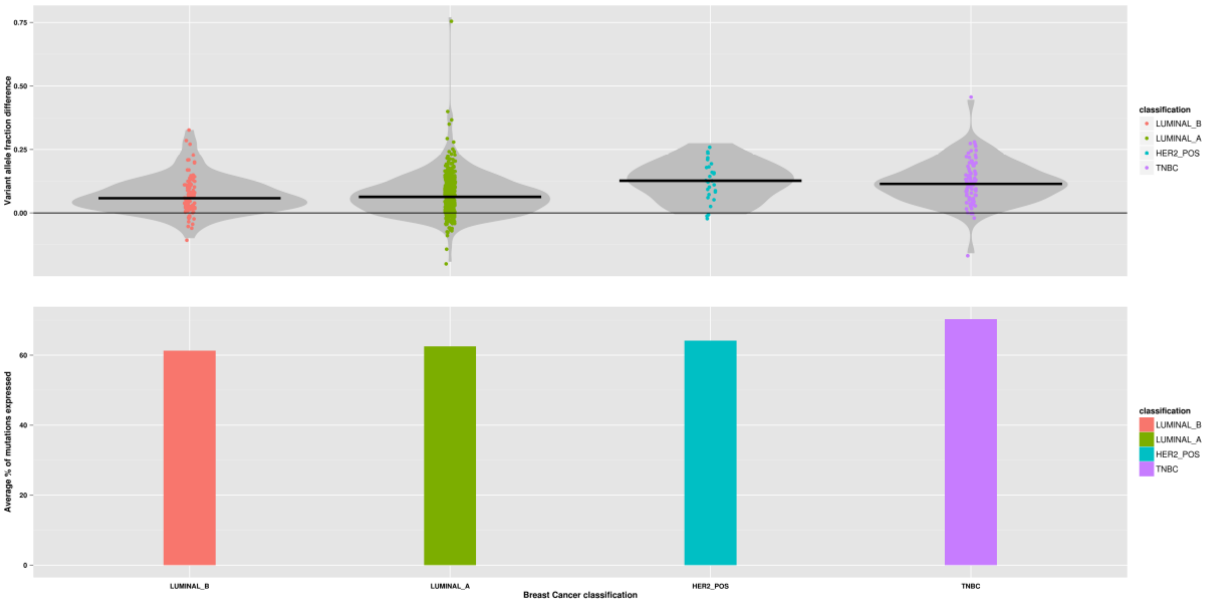


**Figure 26. Variant Allele Fraction difference (violin plots on the top) and percentage of mutated samples (histogram on the bottom) distribution according to 4 main Breast Cancer subgroups (Luminal B in red, Luminal A in green, HER2 positive in light blue and Triple Negative Breast Cancer in purple).**

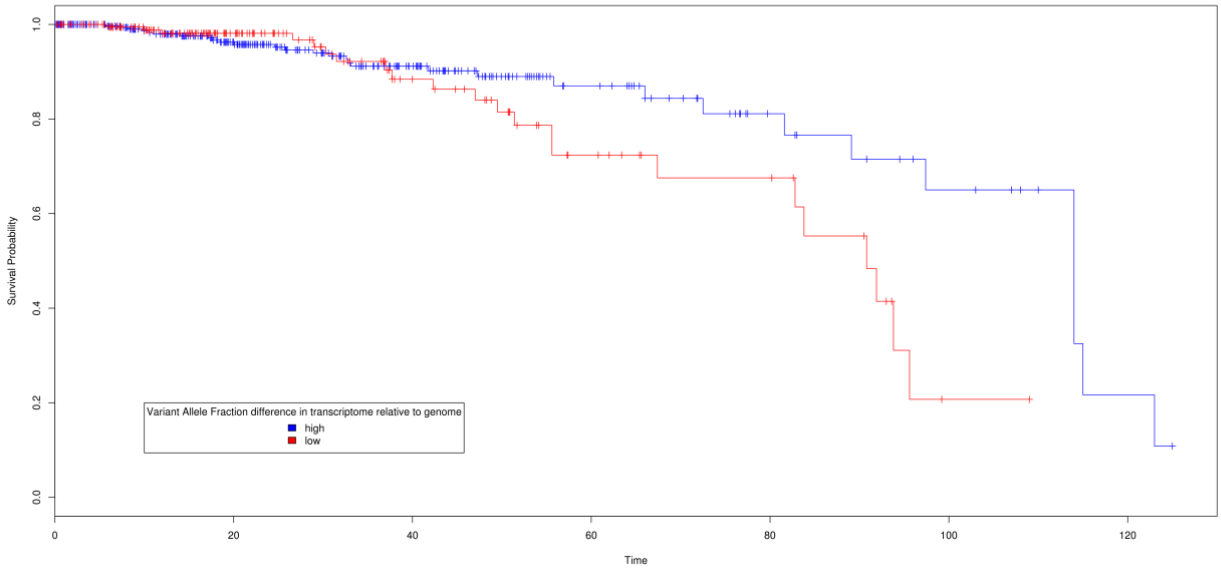


**Figure 27. Overall Survival curve for Breast Cancer patients stratified by low (red) and high (blue) Variant Allele Fraction difference.**

## 3.4 Discussion

The disturbed transcriptional landscape of cancer cells results from three main forces: (1) direct, primary consequences of somatic mutation; (2) co-ordinated, secondary gene expression changes resulting from altered cellular signaling, transcriptional regulation and chromatin landscape; and (3) general loss of transcriptional fidelity, manifesting as shorter 3' UTRs , retained introns, trans-splicing and so on.

This sub-project was focused on dissecting the immediate impact that the repertoire of somatic mutations has on the transcriptome in breast cancer,

exploring the rules that govern how the transcriptional machinery interprets somatic mutation. Integration of transcriptomic with genomic can determine which mutations are expressed and their effect on gene expression profile.

An exhaustive bioinformatics tool for analyzing RNA sequencing data combined with whole genome/exome sequencing was developed and applied to a huge dataset of 980 breast cancers. The software computes the proportion of reads supporting a given variant in the DNA (VAF DNA), that is reflective of that variant's concentration within a sample, and then models the transcriptional output of human cancer cells ,by measuring the deviation in RNA allelic fraction (VAF DNA) from the DNA allelic fraction. This measure was called VAF difference.

Using the VAF difference as a method to account for differences in tumor purity, we found that about 60% of Breast Cancer exonic point mutations are expressed and induce some transcriptional consequence, and a striking anti-correlation between ER levels and the number of expressed mutations. That is, tumors with high levels of ER express fewer mutations than cancers with low ER. This relationship determined that, for every 1% decrease in ER expression, 15 more mutations are expressed. As a breast cancer loses estrogen receptor expression, and becomes more transcriptionally active, it is more likely to actually express its complement of somatic mutations. While speculative, this is of interest to researchers in the field of immunotherapy, since somatic mutations can act as neoantigens that could trigger host immune responses. There are studies reporting strong associations between the number of neoantigens and response to immunotherapy, and these data suggest that such mutations are more likely to be expressed in ER-negative or ER-low tumors.

Stratifying Breast Cancer patients by clinical factors, we found that the difference in VAF between the transcriptome and the genome, as well as percentage of mutations expressed differs in the most common Breast Cancer subgroups and can be used to predict better or poorer clinical outcome in terms of overall survival. In fact, the VAF difference in the transcriptome relative to the genome and the percentage of exonic mutations expressed tends to increase their values going from subtypes with better (Luminal A/B) to worst (TNBC) prognosis. Breast cancer subgroups can be inferred using a model that include data from the genomic and transcriptomic profile of the patient when information about immunohistochemistry for the Estrogen Receptor, Progesterone Receptor and HER2 are unavailable or uncertain.

Having developed a method to feasibly measure differences in transcriptional amplification between individual cells on sporadic breast cancer, a well-studied tumour type, and having demonstrated that transcriptional amplification is a distinguishing feature between known cancer subtypes and may correlate with those subtype's clinical prognosis, different future works can be developed from this study. One of these includes a complete characterization of transcriptional amplification process across thousands of different tumours, in order to predict and classify different cancer subtypes using a signature based on transcriptional output measure from DNA and Rna sequencing. This signature can help clinicians to identify cancer subtype diagnosis, to select the most suitable treatment option and predict patient's clinical outcome.

# Chapter 4: Conclusions

Two bioinformatics software for RNA sequencing analysis applied to precision medicine oncology were developed:

The first, Fusion Validator is able to collect thousands of fusion transcripts detected by different fusion finder algorithm and discriminate real fusions from false positive ones. The tool recreates the chimeric transcript sequence around the fusion breakpoint and significantly reduces the number of fusion candidates for validation, performing different filtering steps like local realignment of normal tissues sequences on fusion transcripts, *de novo* and seed and extend realignment of tumour candidates reads, search for homologous and repetitive regions around the breakpoint and use of a ranking score based on fusion product annotation.

Analysis of simulated synthetic showed an overall better performance of Fusion Validator in terms of Sensitivity and PPV, when compared to 5 different fusion detection tools, with a constant F-measure across samples of different coverage, read length and breakpoint positions. Fusion Validator was able to successfully detect and 96.30% of the chimeric transcripts validated in literature on 4 Breast Cancer Cell lines and 97.95% of the recurrent kinase fusions validated in 190 pan cancer samples, with a significant 79.95% reduction of false positive events. 94.24% of the validated fusions were predicted by Fusion Validator with a ranking score on the top 3.

Fusion Validator can be used as a very quick and efficient diagnostic tool for KiCS program to increase the performance in detecting driver fusions and significantly reduce the number of false positives in particular disorders, where gene fusions are critical as diagnostic and prognostic factors.

The second software integrates transcriptomic with genomic data to measure transcriptional amplification in primary tumours and to determine which mutations are expressed. It also determine their effect on gene expression profile. The software measures the proportion of sequencing reads supporting the mutant allele in the transcriptome, compared to that expected from the genome as the Variant Allele Fraction in the transcriptome related the genome.

Analysis of 25177 somatic variants in 955 Breast Cancer samples showed that 60% of exonic point mutations are expressed and induce some transcriptional consequence, and that number of expressed mutations are significantly inversely correlated to the Estrogen Receptor levels. Variant Allele Fraction differences between the transcriptome and genome, as well as percentage of mutations expressed, differs in the most common Breast Cancer subgroups and can be used as a diagnostic tool to infer tumour subgroup when information about immunohistochemistry for the Estrogen Receptor, Progesterone Receptor and HER2 are unavailable or uncertain, or as prognostic tool to predict better or poorer clinical outcome.

# Appendix

# Appendix A

## List of kinase fusion transcripts from TCGA pan-cancer datasets processed through the Fusion Validator pipeline.

| GENE1 | GENE2 | CANCER_TYPE | TCGA_ID | RECURRENCE | VALIDATION | FUSION VALIDATOR CONFIRMATION | NOTES | GLOBAL RANKING | KINASE FUSIONS RANKING |
|---|---|---|---|---|---|---|---|---|---|
| SLC34A2 | ROS1 | Lung adenocarcinoma: TCGA | TCGA-05-4426-01A | Recurrent | Confirmed | Y | | TOP3 | TOP3 |
| R3HDM2 | PIP4K2C | Glioblastoma multiforme: TCGA | TCGA-06-0174-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| EGFR | SEPT14 | Glioblastoma multiforme: TCGA | TCGA-06-0750-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TMEM165 | PDGFRA | Glioblastoma multiforme: TCGA | TCGA-06-2559-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NFASC | NTRK1 | Glioblastoma multiforme: TCGA | TCGA-06-5411-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CEP85L | ROS1 | Glioblastoma multiforme: TCGA | TCGA-06-5418-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| WASF2 | FGR | Ovarian serous cystadenocarcinoma: TCGA | TCGA-09-2054-01A | Recurrent | Confirmed | N | Missed by fusion detection tools | | |
| MAP2K2 | INSR | Ovarian serous cystadenocarcinoma: TCGA | TCGA-13-1410-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| DDR1 | PAK1 | Ovarian serous cystadenocarcinoma: TCGA | TCGA-13-1477-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Lung squamous cell carcinoma: TCGA | TCGA-22-4607-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| BAG4 | FGFR1 | Lung squamous cell carcinoma: TCGA | TCGA-22-5480-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ARHGEF18 | INSR | Ovarian serous cystadenocarcinoma: TCGA | TCGA-25-2398-01A | Recurrent | Confirmed | N | Missed by fusion detection tools | | |
| FGFR3 | TACC3 | Glioblastoma multiforme: TCGA | TCGA-27-1835-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EGFR | SEPT14 | Glioblastoma multiforme: TCGA | TCGA-27-1837-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| EGFR | SEPT14 | Glioblastoma multiforme: TCGA | TCGA-28-1747-01C | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EGFR | SEPT14 | Glioblastoma multiforme: TCGA | TCGA-28-2513-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| NBPF3 | EPHB2 | Glioblastoma multiforme: TCGA | TCGA-28-2513-01A | Non recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EGFR | SEPT14 | Glioblastoma multiforme: TCGA | TCGA-32-5222-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| ADCY9 | PRKCB | Lung squamous cell carcinoma: TCGA | TCGA-33-4533-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| IGF2BP3 | PRKCA | Lung squamous cell carcinoma: TCGA | TCGA-33-4587-01A | Recurrent | Confirmed | Y | | TOP2 | TOP1 |
| FGFR3 | TACC3 | Lung squamous cell carcinoma: TCGA | TCGA-39-5024-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TANC2 | PRKCA | Lung squamous cell carcinoma: TCGA | TCGA-43-2581-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TECR | PKN1 | Lung squamous cell carcinoma: TCGA | TCGA-43-7658-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| CLTC | ROS1 | Lung adenocarcinoma: TCGA | TCGA-44-2665-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CLTC | ROS1 | Lung adenocarcinoma: TCGA | TCGA-44-2665-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| CLTC | ROS1 | Lung adenocarcinoma: TCGA | TCGA-44-2665-01B | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EML4 | ALK | Lung adenocarcinoma: TCGA | TCGA-50-8460-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| MINK1 | NQO2 | Lung adenocarcinoma: TCGA | TCGA-50-8460-01A | Non recurrent | Not reviewed | Y | | TOP3 | TOP3 |
| TRIM33 | RET | Lung adenocarcinoma: TCGA | TCGA-55-6543-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EZR | ROS1 | Lung adenocarcinoma: TCGA | TCGA-55-6986-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TRIM24 | NTRK2 | Lung adenocarcinoma: TCGA | TCGA-55-8091-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SLC34A2 | ROS1 | Lung adenocarcinoma: TCGA | TCGA-62-A46Y-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CD74 | ROS1 | Lung adenocarcinoma: TCGA | TCGA-64-1680-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| WASF2 | FGR | Lung squamous cell carcinoma: TCGA | TCGA-66-2759-01A | Recurrent | Confirmed | Y | | TOP5 | TOP5 |
| LATS1 | LACE1 | Lung squamous cell carcinoma: TCGA | TCGA-66-2759-01A | Non recurrent | Not reviewed | Y | | TOP1 | TOP1 |
| MAP3K5 | NKAIN2 | Lung squamous cell carcinoma: TCGA | TCGA-66-2759-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| FGFR2 | CCAR2 | Lung squamous cell carcinoma: TCGA | TCGA-66-2765-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Lung squamous cell carcinoma: TCGA | TCGA-66-2786-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EML4 | ALK | Lung adenocarcinoma: TCGA | TCGA-67-6215-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EML4 | ALK | Lung adenocarcinoma: TCGA | TCGA-67-6216-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TUBD1 | RPS6KB1 | Lung adenocarcinoma: TCGA | TCGA-69-7978-01A | Recurrent | Confirmed | Y | | TOP4 | TOP3 |
| CCDC6 | RET | Lung adenocarcinoma: TCGA | TCGA-75-6203-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Glioblastoma multiforme: TCGA | TCGA-76-4925-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PEAK1 | TM9SF3 | Glioblastoma multiforme: TCGA | TCGA-76-4925-01A | Non recurrent | Not reviewed | Y | | TOP4 | TOP4 |
| EML4 | ALK | Lung adenocarcinoma: TCGA | TCGA-78-7163-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SPNS1 | PRKCB | Lung adenocarcinoma: TCGA | TCGA-83-5908-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| CD74 | ROS1 | Lung adenocarcinoma: TCGA | TCGA-86-8278-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EML4 | ALK | Lung adenocarcinoma: TCGA | TCGA-86-A4P8-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| KIF5B | MET | Lung adenocarcinoma: TCGA | TCGA-93-A4JN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CAMK2D | ANK2 | Lung squamous cell carcinoma: TCGA | TCGA-98-A53A-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| DDX42 | RPS6KB1 | Breast invasive carcinoma: TCGA | TCGA-A1-A0SN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TANC2 | STRADA | Breast invasive carcinoma: TCGA | TCGA-A1-A0SN-01A | Recurrent | Confirmed | Y | | TOP6 | TOP6 |
| TEX14 | CLTC | Breast invasive carcinoma: TCGA | TCGA-A1-A0SN-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| TLK2 | KCNJ16 | Breast invasive carcinoma: TCGA | TCGA-A1-A0SN-01A | Non recurrent | Not reviewed | Y | | TOP3 | TOP3 |
| SPINT2 | PAK1 | Breast invasive carcinoma: TCGA | TCGA-A1-A0SQ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| MARK4 | LYPD5 | Breast invasive carcinoma: TCGA | TCGA-A1-A0SQ-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| FGFR3 | TACC3 | Kidney renal papillary cell carcinoma: TCGA | TCGA-A4-7287-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TECR | PKN1 | Uterine Corpus Endometrial Carcinoma: TCGA | TCGA-A5-A3LP-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| XRN1 | PIP4K2A | Breast invasive carcinoma: TCGA | TCGA-A8-A07C-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PI4KB | SELENBP1 | Breast invasive carcinoma: TCGA | TCGA-A8-A07C-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| MOK | ANKRD66 | Breast invasive carcinoma: TCGA | TCGA-A8-A07C-01A | Non recurrent | Not reviewed | Y | | TOP5 | TOP5 |
| STK24 | PIP5K1B | Breast invasive carcinoma: TCGA | TCGA-AC-A5EH-01A | Recurrent | Confirmed | Y | | TOP3 | TOP3 |
| FGFR2 | CASP7 | Breast invasive carcinoma: TCGA | TCGA-AN-A0AL-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Breast invasive carcinoma: TCGA | TCGA-AO-A03U-01B | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ZNF577 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-AR-A0U3-01A | Recurrent | Confirmed | Y | | TOP4 | TOP3 |
| ZNF37A | PIP5K1B | Breast invasive carcinoma: TCGA | TCGA-AR-A2LL-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ERC1 | PIK3C2G | Breast invasive carcinoma: TCGA | TCGA-B6-A0IG-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PAN3 | NTRK2 | Head and Neck squamous cell carcinoma: TCGA | TCGA-BB-4223-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| OBSCN | CASZ1 | Head and Neck squamous cell carcinoma: TCGA | TCGA-BB-4223-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| ANXA4 | PKN1 | Liver hepatocellular carcinoma: TCGA | TCGA-BC-4072-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ATG7 | BRAF | Skin Cutaneous Melanoma: TCGA | TCGA-BF-A5EP-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RHOT1 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-BH-A18U-01A | Recurrent | Confirmed | Y | | TOP3 | TOP2 |
| KIT | PDGFRA | Breast invasive carcinoma: TCGA | TCGA-BH-A1F0-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SGK1 | PDSS2 | Breast invasive carcinoma: TCGA | TCGA-BH-A1F0-01A | Non recurrent | Not reviewed | Y | | TOP9 | TOP7 |
| NAP1L1 | STK38L | Breast invasive carcinoma: TCGA | TCGA-BH-A1FN-01A | Recurrent | Confirmed | Y | | TOP4 | TOP4 |
| ZNF791 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-BH-A209-01A | Recurrent | Confirmed | Y | | TOP4 | TOP4 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-BJ-A0ZJ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-BJ-A28Z-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-BJ-A2N7-11A | Recurrent | Confirmed | Y | | TOP2 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-BJ-A2N8-11A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |

# Appendix A

# Appendix A

| GENE1 | GENE2 | CANCER_TYPE | TCGA_ID | RECURRENCE | VALIDATION | FUSION VALIDATOR CONFIRMATION | NOTES | GLOBAL RANKING | KINASE FUSIONS RANKING |
|---|---|---|---|---|---|---|---|---|---|
| FNDC3B | PIK3CA | Uterine Corpus Endometrial Carcinoma: TCGA | TCGA-BK-A56F-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| BAIAP2L1 | MET | Kidney renal papillary cell carcinoma: TCGA | TCGA-BQ-7049-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR2 | TACC2 | Stomach adenocarcinoma: TCGA | TCGA-BR-8080-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NF1 | TEX14 | Stomach adenocarcinoma: TCGA | TCGA-BR-8080-01A | Non recurrent | Not reviewed | Y | | TOP3 | TOP3 |
| CASZ1 | MTOR | Stomach adenocarcinoma: TCGA | TCGA-BR-8483-01A | Recurrent | Confirmed | N | Missed by fusion detection tools | | |
| PRKCI | GPR160 | Stomach adenocarcinoma: TCGA | TCGA-BR-8483-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| DSTYK | NUCKS1 | Stomach adenocarcinoma: TCGA | TCGA-BR-8483-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| BMPR2 | SPATS2L | Stomach adenocarcinoma: TCGA | TCGA-BR-8483-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| ERC1 | RET | Breast invasive carcinoma: TCGA | TCGA-C8-A1HJ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TBL1XR1 | PIK3CA | Breast invasive carcinoma: TCGA | TCGA-C8-A26X-01A | Recurrent | Confirmed | Y | | TOP4 | TOP3 |
| NLK | BCAS3 | Breast invasive carcinoma: TCGA | TCGA-C8-A26X-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| STARD3 | STRADA | Breast invasive carcinoma: TCGA | TCGA-C8-A275-01A | Recurrent | Confirmed | Y | | TOP5 | TOP5 |
| CPD | ERBB2 | Stomach adenocarcinoma: TCGA | TCGA-CD-5799-01A | Recurrent | Confirmed | Y | | TOP5 | TOP5 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-CE-A13K-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-CE-A27D-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SSBP2 | NTRK1 | Thyroid carcinoma: TCGA | TCGA-CE-A3MD-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TRIM27 | RET | Thyroid carcinoma: TCGA | TCGA-CE-A481-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NCOA4 | RET | Thyroid carcinoma: TCGA | TCGA-CE-A482-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SPECC1L | RET | Thyroid carcinoma: TCGA | TCGA-CE-A485-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma: TCGA | TCGA-CF-A3MF-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma: TCGA | TCGA-CF-A3MG-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma: TCGA | TCGA-CF-A3MH-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma: TCGA | TCGA-CF-A47S-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| KDM7A | BRAF | Prostate adenocarcinoma: TCGA | TCGA-CH-5737-01A | Recurrent | Confirmed | Y | JHDM1D-BRAF found. JHDM1D is an alias of KDM7A | TOP1 | TOP1 |
| AGGF1 | RAF1 | Prostate adenocarcinoma: TCGA | TCGA-CH-5737-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| ETV6 | NTRK3 | Colon adenocarcinoma: TCGA | TCGA-CK-5913-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Colon adenocarcinoma: TCGA | TCGA-CK-5916-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Colon adenocarcinoma: TCGA | TCGA-CM-4743-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| LYN | NTRK3 | Head and Neck squamous cell carcinoma: TCGA | TCGA-CN-6997-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Head and Neck squamous cell carcinoma: TCGA | TCGA-CR-6473-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | ELAVL3 | Brain Lower Grade Glioma: TCGA | TCGA-CS-6186-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Head and Neck squamous cell carcinoma: TCGA | TCGA-CV-7100-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| WNK1 | GIPC1 | Uterine Corpus Endometrial Carcinoma: TCGA | TCGA-D1-A3JQ-01A | Non recurrent | Not reviewed | Y | | TOP3 | TOP3 |
| KAZN | MTOR | Uterine Corpus Endometrial Carcinoma: TCGA | TCGA-D1-A3JQ-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| TAX1BP1 | BRAF | Skin Cutaneous Melanoma: TCGA | TCGA-D3-A2JC-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| GGA3 | VRK2 | Skin Cutaneous Melanoma: TCGA | TCGA-D3-A2JC-06A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| FGFR2 | CCDC6 | Breast invasive carcinoma: TCGA | TCGA-D8-A13Z-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ERLIN2 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-D8-A1JC-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| MPRIP | RAF1 | Skin Cutaneous Melanoma: TCGA | TCGA-D9-A4Z6-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| AGK | BRAF | Skin Cutaneous Melanoma: TCGA | TCGA-DA-A1IA-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| MACF1 | BRAF | Thyroid carcinoma: TCGA | TCGA-DE-A0Y2-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-DE-A2OL-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NCOA4 | RET | Thyroid carcinoma: TCGA | TCGA-DE-A3KN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| AGK | BRAF | Thyroid carcinoma: TCGA | TCGA-DJ-A2PX-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-DJ-A2Q1-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-DJ-A2Q3-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-DJ-A2Q4-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-DJ-A2Q5-01A | Recurrent | Confirmed | Y | | TOP2 | TOP1 |
| STRN | ALK | Thyroid carcinoma: TCGA | TCGA-DJ-A3US-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-DJ-A3UV-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-DJ-A3V3-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| IRF2BP2 | NTRK1 | Thyroid carcinoma: TCGA | TCGA-DJ-A4UP-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-DJ-A4UQ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-DJ-A4V0-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-DJ-A4V5-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CTSB | PXK | Thyroid carcinoma: TCGA | TCGA-DJ-A4V5-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| MTSS1 | ERBB2 | Bladder Urothelial Carcinoma: TCGA | TCGA-DK-A2I6-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-DO-A1JZ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CAMK2D | ANK2 | Brain Lower Grade Glioma: TCGA | TCGA-DU-5855-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EGFR | SEPT14 | Brain Lower Grade Glioma: TCGA | TCGA-DU-6406-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| WNK1 | STK38L | Brain Lower Grade Glioma: TCGA | TCGA-DU-7007-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PTPRZ1 | MET | Brain Lower Grade Glioma: TCGA | TCGA-DU-7304-02A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SQSTM1 | NTRK2 | Brain Lower Grade Glioma: TCGA | TCGA-DU-A76L-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TRIO | TERT | Sarcoma: TCGA | TCGA-DX-A1L3-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| C5ORF22 | MAP2K5 | Sarcoma: TCGA | TCGA-DX-A1L3-01A | Non recurrent | Not reviewed | N | Filtered by Validator | | |
| TUFT1 | PKN2 | Sarcoma: TCGA | TCGA-DX-A23U-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAB3B | PKN2 | Sarcoma: TCGA | TCGA-DX-A23U-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| TRIO | TERT | Sarcoma: TCGA | TCGA-DX-A2J0-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| ELK3 | CDK17 | Sarcoma: TCGA | TCGA-DX-A2J0-01A | Non recurrent | Confirmed | Y | | TOP3 | TOP3 |
| TPM3 | NTRK1 | Sarcoma: TCGA | TCGA-DX-A3UA-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PTAR1 | PIP5K1B | Sarcoma: TCGA | TCGA-DX-A48N-01A | Recurrent | Confirmed | Y | | TOP2 | TOP1 |
| C12ORF45 | CDK7 | Sarcoma: TCGA | TCGA-DX-A48N-01A | Non recurrent | Not reviewed | Y | | TOP8 | TOP7 |
| NUAK1 | UHRF1BP1L | Sarcoma: TCGA | TCGA-DX-A48N-01A | Non recurrent | Not reviewed | Y | | TOP3 | TOP2 |
| SRI | PIP4K2C | Sarcoma: TCGA | TCGA-DX-A6BH-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PDGFRA | FIP1L1 LNX1 | Brain Lower Grade Glioma: TCGA | TCGA-E1-A7YI-01A | Recurrent | Confirmed | N | FIP1L1-CHIC2 fusion found: CHIC2 overlap PDGFRA | | |
| TBL1XR1 | PIK3CA | Breast invasive carcinoma: TCGA | TCGA-E2-A14P-01A | Recurrent | Confirmed | Y | | TOP4 | TOP4 |
| GSK3B | FSTL1 | Breast invasive carcinoma: TCGA | TCGA-E2-A14P-01A | Non recurrent | Not reviewed | Y | | TOP1 | TOP1 |
| WHSC1L1 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-E2-A15A-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ANK1 | FGFR1 | Breast invasive carcinoma: TCGA | TCGA-E2-A15A-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| DYRK1A | KDM4B | Breast invasive carcinoma: TCGA | TCGA-E2-A15A-01A | Non recurrent | Not reviewed | Y | | TOP4 | TOP4 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-E3-A3E0-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Bladder Urothelial Carcinoma: TCGA | TCGA-E7-A5KE-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EML4 | ALK | Thyroid carcinoma: TCGA | TCGA-E8-A432-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-E8-A438-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |

| GENE1 | GENE2 | CANCER_TYPE | TCGA_ID | RECURRENCE | VALIDATION | FUSION VALIDATOR CONFIRMATION | NOTES | GLOBAL RANKING | KINASE FUSIONS RANKING |
|---|---|---|---|---|---|---|---|---|---|
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-E8-A44M-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Skin Cutaneous Melanoma: TCGA | TCGA-EB-A51B-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| LMNA | RAF1 | Skin Cutaneous Melanoma: TCGA | TCGA-EB-A5SF-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TRAK1 | RAF1 | Skin Cutaneous Melanoma: TCGA | TCGA-EE-A2MI-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TBL1XR1 | PIK3CA | Prostate adenocarcinoma: TCGA | TCGA-EJ-5507-01A | Recurrent | Confirmed | Y | | TOP3 | TOP2 |
| FGFR3 | AES | Prostate adenocarcinoma: TCGA | TCGA-EJ-A7NM-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3CY-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TPM3 | NTRK1 | Thyroid carcinoma: TCGA | TCGA-EL-A3D4-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NCOA4 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3H3-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| AP3B1 | BRAF | Thyroid carcinoma: TCGA | TCGA-EL-A3T0-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ERC1 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3T9-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3TB-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SND1 | BRAF | Thyroid carcinoma: TCGA | TCGA-EL-A3ZK-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-EL-A3ZN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3ZP-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-EL-A3ZS-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| GTF2IRD1 | ALK | Thyroid carcinoma: TCGA | TCGA-EL-A4KD-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RAF1 | AGGF1 | Thyroid carcinoma: TCGA | TCGA-EM-A1CS-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NCOA4 | RET | Thyroid carcinoma: TCGA | TCGA-EM-A2CU-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-EM-A3AN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TFG | NTRK1 | Thyroid carcinoma: TCGA | TCGA-EM-A3AO-10A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ERC1 | RET | Thyroid carcinoma: TCGA | TCGA-EM-A3FQ-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CLCN6 | RAF1 | Skin Cutaneous Melanoma: TCGA | TCGA-ER-A19L-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| WASF2 | FGR | Skin Cutaneous Melanoma: TCGA | TCGA-ER-A19W-06A | Recurrent | Confirmed | Y | | TOP4 | TOP3 |
| BCL2L11 | BRAF | Thyroid carcinoma: TCGA | TCGA-ET-A2MX-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RBPMS | NTRK3 | Thyroid carcinoma: TCGA | TCGA-ET-A39L-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-ET-A39R-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| GRID1 | BAZ1B | Thyroid carcinoma: TCGA | TCGA-ET-A39R-01A | Non recurrent | Not reviewed | N | Missed by fusion detection tools | | |
| FAM114A2 | BRAF | Thyroid carcinoma: TCGA | TCGA-ET-A3BN-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FKBP15 | RET | Thyroid carcinoma: TCGA | TCGA-ET-A3DQ-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-ET-A3DR-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TBL1XR1 | RET | Thyroid carcinoma: TCGA | TCGA-ET-A40R-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SQSTM1 | NTRK1 | Thyroid carcinoma: TCGA | TCGA-ET-A40S-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-ET-A40T-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| EFNA3 | PIK3C2G | Breast invasive carcinoma: TCGA | TCGA-EW-A1PC-01B | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| TRIM24 | BRAF | Rectum adenocarcinoma: TCGA | TCGA-F5-6464-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SMEK2 | ALK | Rectum adenocarcinoma: TCGA | TCGA-F5-6864-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CAMK2D | TLL2 | Rectum adenocarcinoma: TCGA | TCGA-F5-6864-01A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| ETV6 | NTRK3 | Thyroid carcinoma: TCGA | TCGA-FE-A3PD-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Brain Lower Grade Glioma: TCGA | TCGA-FG-7643-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| RIMKLB | PIP4K2A | Brain Lower Grade Glioma: TCGA | TCGA-FG-8185-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| TFG | MET | Thyroid carcinoma: TCGA | TCGA-FK-A3S3-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-FK-A3SE-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| AKAP13 | RET | Thyroid carcinoma: TCGA | TCGA-FK-A3SG-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CDC27 | BRAF | Skin Cutaneous Melanoma: TCGA | TCGA-FS-A1ZU-06A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| SND1 | BRAF | Thyroid carcinoma: TCGA | TCGA-FY-A40N-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| STRN | ALK | Kidney renal papillary cell carcinoma: TCGA | TCGA-G7-6792-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| MAP4K3 | SMC6 | Kidney renal papillary cell carcinoma: TCGA | TCGA-G7-6792-01A | Non recurrent | Not reviewed | Y | | TOP4 | TOP3 |
| C8ORF34 | MET | Kidney renal papillary cell carcinoma: TCGA | TCGA-GL-7773-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| NARS2 | PAK1 | Skin Cutaneous Melanoma: TCGA | TCGA-GN-A26D-06A | Recurrent | Confirmed | Y | | TOP3 | TOP3 |
| MARK2 | BATF2 | Skin Cutaneous Melanoma: TCGA | TCGA-GN-A26D-06A | Non recurrent | Not reviewed | Y | | TOP2 | TOP2 |
| TBK1 | GRIP1 | Skin Cutaneous Melanoma: TCGA | TCGA-GN-A26D-06A | Non recurrent | Not reviewed | Y | | TOP1 | TOP1 |
| PAK1 | PDGFD | Skin Cutaneous Melanoma: TCGA | TCGA-GN-A26D-06A | Non recurrent | Not reviewed | Y | | TOP4 | TOP4 |
| TPM1 | ALK | Bladder Urothelial Carcinoma: TCGA | TCGA-GV-A3QG-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| PAPD7 | RAF1 | Prostate adenocarcinoma: TCGA | TCGA-HC-8256-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| AFAP1 | NTRK2 | Brain Lower Grade Glioma: TCGA | TCGA-HT-7680-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| GGA2 | PRKCB | Brain Lower Grade Glioma: TCGA | TCGA-HT-A5RC-01A | Recurrent | Confirmed | Y | | TOP2 | TOP2 |
| MKRN1 | BRAF | Thyroid carcinoma: TCGA | TCGA-J8-A3O1-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| CCDC6 | RET | Thyroid carcinoma: TCGA | TCGA-J8-A4HW-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ERBB2 | PPP1R1B | Liver hepatocellular carcinoma: TCGA | TCGA-KR-A7K2-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| ZC3HAV1 | BRAF | Thyroid carcinoma: TCGA | TCGA-KS-A4ID-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| FGFR3 | TACC3 | Brain Lower Grade Glioma: TCGA | TCGA-P5-A72U-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |
| OXR1 | MET | Liver hepatocellular carcinoma: TCGA | TCGA-RC-A6M6-01A | Recurrent | Confirmed | Y | | TOP1 | TOP1 |

# Bibliography


1. Shlien A, Raine K, Fuligni F, et al. Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. *Cell Reports*. 2016;16(7):2032-2046.

2. Stratton, M.R., Campbell, P.J., and Futreal, P.A. The cancer genome. Nature 2013. *458*, 719-724.

3. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007 Apr;7(4):233-45. PMID: 17361217

4. Barnes,DJ and Melo,JV. Cytogenetic and molecular genetic aspects of chronic myeloid leukaemia. Acta Haematol. 2002, 108, 180–202.

5. Delattre O, Zucman J, Plougastel B, et al. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. Nature 1992;359:162-165.

6. Soda,M, Choi,YL, Enomoto,M, Takada,S, Yamashita,Y, Ishikawa,S, Fujiwara,S, Watanabe,H, Kurashina,K, Hatanaka,H et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007, 448, 561–566.

7. Singh D, Chan JM, Zoppoli P, Niola P, Sullivan R, Castano A, Liu EM, Reichel J, Porrati P, Pellegatta S, et al. Transforming fusions of FGFR

and TACC genes in human glioblastoma. Science 2012, 337, pp. 1231–1235

8. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015, 43, pp. D805–D811

9. Parker, B. C., & Zhang, W. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chinese Journal of Cancer 2013*, *32*(11), 594–603.

10. Macconaill LE, Van Hummelen P, Meyerson M, Hahn WC. Clinical implementation of comprehensive strategies to characterize cancer genomes: opportunities and challenges. Cancer Discov 2011. 1: 297.

11. Garraway LA, Verweij J and Ballman KV. "Precision Oncology: An Overview". J. Clinical Oncology 2013. 31 (15): 1803–1805.

12. Wu YM et al. Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 2013 3, 636–647.

13. Roychowdhury, S. et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci. Transl. Med. 2011 3, ra121.

14. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, Peters BA, Modrusan Z,Jung K, Seshagiri S, et al: Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC medical genomics 2011, 4:11.

15. Babiceanu M, Qin F, Jia Z, Lopez K, Janus N, Facemire L, Kumar S, Pang Y, Qi Y, Lazar IM. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Research, 2016.

16. Davare, M.A. and Tognon, C.E. Detecting and targetting oncogenic fusion proteins in the genomic era. Biol. Cell 2015. 107, 111–12.

17. Mody RJ, Wu YM, Lonigro RJ, et al. Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. JAMA. 2015;314:913-925.

18. Yelensky R, Donahue A, Otto G, He Jie, Juhn F, Dowing S, Frampton GM. Analytical validation of solid tumor fusion gene detection in a comprehensive NGS-based clinical cancer genomic test. Cancer Res October 1, 2014 74:4699;

19. Scolnick JA, Dimon M, Wang I-C, Huelga SC, Amorese DA. An Efficient Method for Identifying Gene Fusions by Targeted RNA Sequencing from Fresh Frozen and FFPE Samples. Xie K, ed. PLoS ONE. 2015;10(7):e0128916.

20. Wang Q, Xia J, Jia P, Pao W and Zhao Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. Brief. Bioinform. 2013, 14,506–519.

21. Beccuti M, Carrara M, Cordero F, Donatelli S and Calogero RA. The structure of state-of-art gene fusion-finder algorithms. Genome Bioinform. 2013, 1, 2.

22. 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 Feb;2013(2013):340620.

23. McPherson A, Hormozdiari F, Zayed A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Computational Biology. 2011;7(5)e1001138

24. Sboner A, Habegger L, Pflueger D, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biology. 2010;11(10):R104.

25. Li Y, Chien J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics. 2011;27(12):1708–1710.btr265

26. Benelli M, Pescucci C, Marseglia G, Severgnini M, Torricelli F and Magi A. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics 2012, 28, 3232–3239.

27. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, Yu Y, Zhu D, Nickerson ML, Wan S et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 2013, 14, R12.

28. Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research. 2010;38(18):p. e178

29. Francis RW, Thompson-Wicking K, Carter KW, Anderson D, Kees UR, Beesley AH. FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. PloS One. 2012;7(6)e39987

30. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. FusionMap: Detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011;27(14):1922–1928.btr310

31. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27(20):2903–2904

32. Abate F, Acquaviva A, Paciello G, et al. Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics. 2012;28(16):2114–2121

33. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology. 2011;12(8):R72.

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

35. Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, Kim S, Ma T, Chang TY, Priedigkeit NM, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-Seq data. Nucleic Acid Research 2005.

36. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, et al. Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. Genome Research. 2012;22(7):1231-1242.

37. Beccuti M, Carrara M, Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A and Calogero RA. “Chimera: a Bioconductor package for secondary analysis of fusion products.” Bioinformatics 2014. 0, pp. 3.

38. Hoogstrate, Y., Böttcher, R., Hiltemann, S., van der Spek, P. J., Jenster, G., & Stubbs, A. P. . FuMa: reporting overlap in RNA-seq detected fusion genes. Bioinformatics 2015. btv721.

39. Shugay M, Ortiz de Mendíbil I, Vizmanos JL and Novo FJ. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics. 16 Aug 2013.

40. Abate F, Zairis S, Ficarra E, et al. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC Syst Biol. 2014;8:97

41. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013 45, 580-5.

42. Lappalainen T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013 501, 506-11.

43. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 2011, 27(18):2518-2528.

44. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL et al. Identification of

fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. (2011), 12, R6.

45. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 490(7418):61-70.

46. Stransky N, Cer\ami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nat Commun. 2014;5:4846.

47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013, 29: 15-21.

48. Nicorici D, Satalan A, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, Virtanen S, Kilkku O, FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data, bioRxiv, Nov. 2014.

49. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009. 19:1117–1123.

50. Boetzer, M. and Pirovano, W., Towards (almost) closed genomes with GapFiller, Genome Biology 2012, 13(6).

51. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

52. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 2009, 25:1754-60.

53. Garraway, L.A., and Lander, E.S. Lessons from the cancer genome. Cell *153*, 17-37 2013.

54. Kasper, L.H., Lerach, S., Tang, H., Ma, J.*, et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. Nature *2011. 471*, 189-195.

55. Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C.*, et al.* Integrated genomic characterization of endometrial carcinoma. Nature 2013. *497*, 67-73.

56. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.*, et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012. *486*, 346-352.

57. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A.*, et al*. Molecular portraits of human breast tumours. Nature 2000 *406*, 747-752.

58. Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S.*, et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci 2001. U S A *98*, 10869-10874.

59. Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J.*, et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *2009. 461*, 809-813.

60. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012. *489*, 519-525.

61. Valk, P.J., Verhaak, R.G., Beijen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J., Lowenberg, B.*, et al*. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med. 2004. *350*, 1617-1628.

62. Dvinge, H., Git, A., Graf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A.*, et al*. The shaping and functional consequences of the microRNA landscape in breast cancer. Nature *2013. 497*, 378-382.

63. Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J.*, et al.* Expressed pseudogenes in the transcriptional landscape of human cancers. Cell *2012. 149*, 1622-1634.

64. Blows FM, Driver KE, Schmidt MK, et al. Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies. Marincola FM, ed. *PLoS Medicine*. 2010;7(5):e1000279.