FusionVet

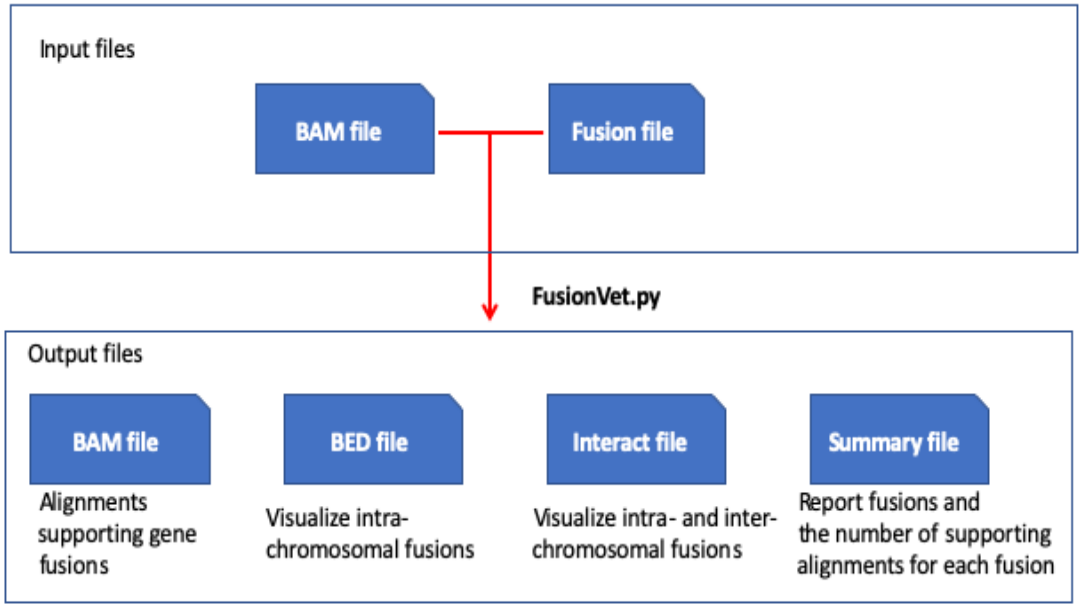
Release 1.0.1

Sep 20, 2019

FusionVet Documentation

Gene fusion is one of the most common somatic alterations that plays an important role in tumorgenesis. Well-known examples include the intra-chromosomal TMPRSS2-ERG fusions in prostate cancer, and the inter-chromosomal BCR-ABL fusions in chronic myelogenous leukemia (CML). With the advent of next generation sequencing technologies especially RNA-seq and the development of dozens of fusion detection tools, most recurrent gene fusions in common cancers have been identified. These fusion are cataloged in databases such as COSMIC , FusionGDB , FusionHub, ChimerDB and TumorFusions).

To facilitate molecular testing, we developed FusionVet (Fusion Visualization and Evaluation Tool) to quickly (and accurately) examine if a gene fusion with clinical significance exists in a particular sample or not.



CHAPTER 1

Installation

You will need **pip** to install FusionVet. Pip is already installed if your Python3 version >= 3.4. Otherwise, follow this instruction to install pip. Use this command to install FusionVet and its dependency packages.

```
$ pip3 install git+https://github.com/liguowang/fusionvet.git
```

Alternatively, it is also available on PyPI.

```
$ pip3 install fusionvet
```

CHAPTER 2

Usage

Options:

--version show program's version number and exit

-h, --help show this help message and exit

-b INPUT_BAM, --bam=INPUT_BAM Input BAM file. The BAM file should be sorted and indexed using SAMtools (http://samtools.sourceforge.net/). (mandatory)

-c INPUT_CHIMERAS, --chimeras=INPUT_CHIMERAS Fusion file. This file can be 6 columns (chr1 start1 end1 chr2 start2 end2) or 8 columns (chr1 start1 end1 name1 chr2 start2 end2 name2) separated by Tab or Space. Lines starting with '#' will be ignored. (mandatory)

-o OUTPUT_FILE, --output=OUTPUT_FILE Prefix of output files. Four files will be created including "prefix.fusion.sorted.bam", "prefix.fusion.bed", "prefix.fusion.interact.bed" and "prefix.fusion.summary.txt". (mandatory)

-q MAP_QUAL, --mapq=MAP_QUAL Mapping quality cutoff. default=30

-t, --track-header If set, add "track line" to the BED file.

-k, --keep-unknown-mapq If set, keep alignments with unknown mapping quality (i.e., MAPQ = 255).

CHAPTER 3

Input files

FusionVet needs two types of input files.

3.1 BAM file

BAM file must be sorted and indexed using samtools

3.2 gene fusion file

The gene fusion file is a plain text file with 8 columns separated by space or tab (The first 4 columns describe the "chrom", "transcription_start", "transcription_end" and "symbol" of gene-1, the other 4 columns describe the same information for gene-2. Below example file defines two fusions:

```
chr21   39739182        40033704        ERG     chr21   42836477        42880085        ␣
↪  TMPRSS2
chr14   38033152        38033701        EST14         chr7    13930855        ␣
↪14031050        ETV1
```

CHAPTER 4

Output files

FusionVet generates 5 files

- prefix.fusion.sorted.bam
- prefix.fusion.sorted.bam.bai
- prefix.fusion.bed
- prefix.fusion.interact.bed
- prefix.fusion.summary.txt

4.1 prefix.fusion.sorted.bam

BAM file containing the chimeric reads supporting gene fusions. Comparing to the orignal BAM file, two additional tags are added to each alignment record: FN (Fusion Name) and SR (Supporting Read)

- SR:i:1 : Fusion was supported by split read
- SR:i:2 : Fusion was supported by paired reads
- SR:I:3 : Fusion was supported by both split read and paired reads.

```
$ samtools view out.fusion.sorted.bam | head -10
UNC13-SN749:172:D101FACXX:8:1104:12580:173001/1        99      chr21   39775575
↪ 66      48M     =       42879910        -3104288
↪CTTTCACCGCCCACTCCAGCCACTGCCGCACATGGTCTGTACTCCATA
↪CCCFFFFFHHHHHJJJJIIJJIJJJJIJIIJJJJJFHGJGFHIHIJJJ        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:1104:4678:34964/2  163     chr21   39817326        66
↪ 48M     =       42879890        -3062517
↪CCTTGAGCCATTCACCTGGCTAGGGTTACATTCCATTTTGATGGTGAC        CCCFFFDFHHHHBGHIJJJJJJIIJ?
↪GIIGIJGGGIJJJJJJJJIFDG        RG:Z:120508_UNC13-SN749_0172_AD101FACXX_8_CGATGT
↪  IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:1208:10044:4367/1  99      chr21   39817340        66
↪ 48M     =       42880015        -3062628
↪CCTGGCTAGGGTTACATTCCATTTTGATGGTGACCCTGGCTGGGGGTT
↪CCCFFFFFHHHFFIJJIJJJIIJJJJIIJJHJJJJIJJJIJJJJIJI>        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
```

(continues on next page)

(continued from previous page)

```
UNC13-SN749:172:D101FACXX:8:1301:12176:174226/2        163     chr21   39817361
↪ 66      48M     =       42879922        -3062514
↪TTTTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAATCCTGCTGAG
↪BCCFFFFFHFHHHJJJJJJJJJJJJFHIIIJJIIJJJJJJJJIJIJJJ        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:2201:10011:20671/1 99      chr21   39817379        66
↪ 48M     =       42879951        -3062525
↪CTGGGGGTTGAGACAGCCAATCCTGCTGAGGGACGCGTGGGCTCATCT
↪CCCFFFFDHHHGHJJJJJJJJJJJJIIJJJJJJIJIJGHHHFFFDEEEE        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:1108:17583:42031/2 163     chr21   39817384        65
↪ 48M     =       42880007        -3062576
↪GGTTGAGACAGCCAATCCTGCTGAGGGACGCGTGGGCTCATCTTGGAA        ?@;
↪BDFDABFFDHHAFHHGHIIIJGIIJGIAE?@6;FGH@DDCC@CA#        RG:Z:120508_UNC13-SN749_0172_
↪AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:2302:8715:52295/1  99      chr21   39817385        66
↪ 48M     =       42879932        -3062500
↪GTTGAGACAGCCAATCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAG
↪CCCFFFFFHHHHHJJJJJHJJJJJJJJJJJJFHIJIIJGIJIJIIJIJ        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:2305:11177:45091/1 99      chr21   39817385        66
↪ 48M     =       42880014        -3062582
↪GTTGAGACAGCCAATCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAG
↪B@CFFFFFHHHHHJJJJJJJJJJJJJIJJJJHJJJJJJJJJJJJIJJG        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:2306:12796:14838/2 163     chr21   39817391        53
↪ 48M     =       42879889        -3062451
↪ACAGCCAATCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAGTCTGTA
↪@CCFFFFFHHHGHJJJJJJJJJJJJHGIJIJJJJJJJJIIIJHHJ###        RG:Z:120508_UNC13-SN749_
↪0172_AD101FACXX_8_CGATGT        IH:i:1  HI:i:1  NM:i:1  SR:i:2  FN:Z:ERG--TMPRSS2
UNC13-SN749:172:D101FACXX:8:1308:12672:71749/1 99      chr21   39817394        66
↪ 48M     =       42880007        -3062566
↪GCCAATCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAGTCTGTCCAT        ?@@FDDDFADF?D@AAB?
↪ACGAHHEHG@BFHIGHBB=8=88@C=@@CE        RG:Z:120508_UNC13-SN749_0172_AD101FACXX_8_
↪CGATGT        IH:i:1  HI:i:1  NM:i:0  SR:i:2  FN:Z:ERG--TMPRSS2
```

4.2 prefix.fusion.sorted.bam.bai

The index file of prefix.fusion.sorted.bam

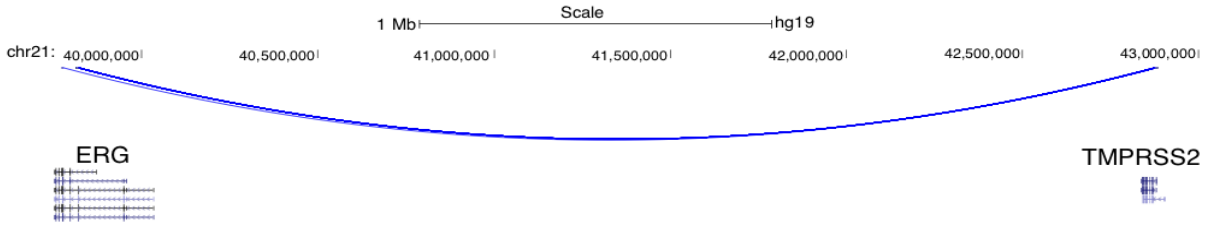
4.3 prefix.fusion.bed

This is standard BED12 format file. Paired reads are merged into a single BED entry. This file can be uploaded to UCSC genome browser to visualize intra-chromosomal fusions. This is useful to identify the **fusion point**. If this file is too large to upload to UCSC genome browser directly, you could try to convert this **BED** file into **bigBed** file (using the bedToBigBed program) following this instruction.

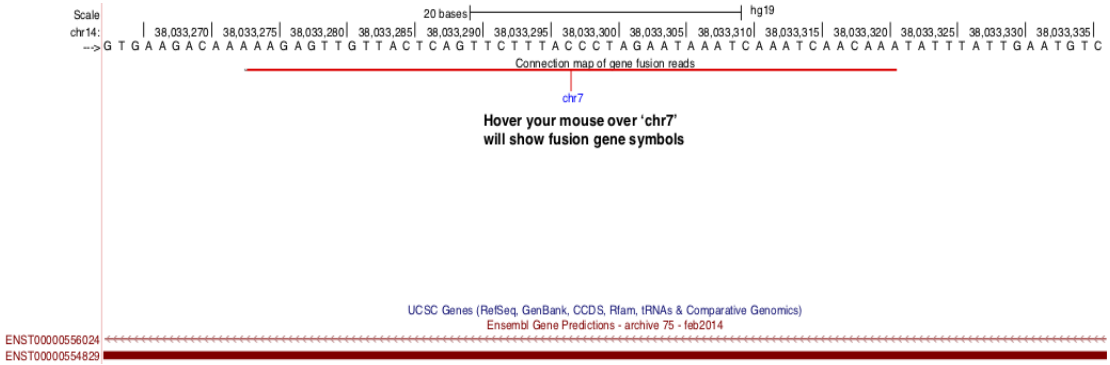
4.4 prefix.fusion.interact.bed

This is Interact format file. This file can be uploaded to UCSC genome browser to visualize both intra-chromosomal and inter-chromosomal fusions. If this file is too large to upload to UCSC genome browser directly, you could try to convert this **Interact** file into **bigInteract** file (using the bedToBigBed program) following this instruction.

Intra-chromosomal fusions will be visualized as below (Note the two breaking points on ERG gene). Toggle between **full** display mode and **pack/squish** display mode help identify the exact breaking point(s).



Inter-chromosomal fusions will be visualized as below. Toggle between **full** display mode and **pack/squish** display mode help identify the exact breaking point(s).



4.5 prefix.fusion.summary.txt

Report the total number of supporting RNA fragments (split reads + read pairs) for each fusion.

```
Sample_ID        ERG--TMPRSS2
Tumor_RNA_TCGA-HC-7819-01A-11R-2118-07.bam       48
```

CHAPTER 5

Performance

5.1 Speed

FusionVet is very efficient. It took about **1 second** to examine 1 fusion in a typical TCGA BAM file (7.1 Gb, 184 million reads)

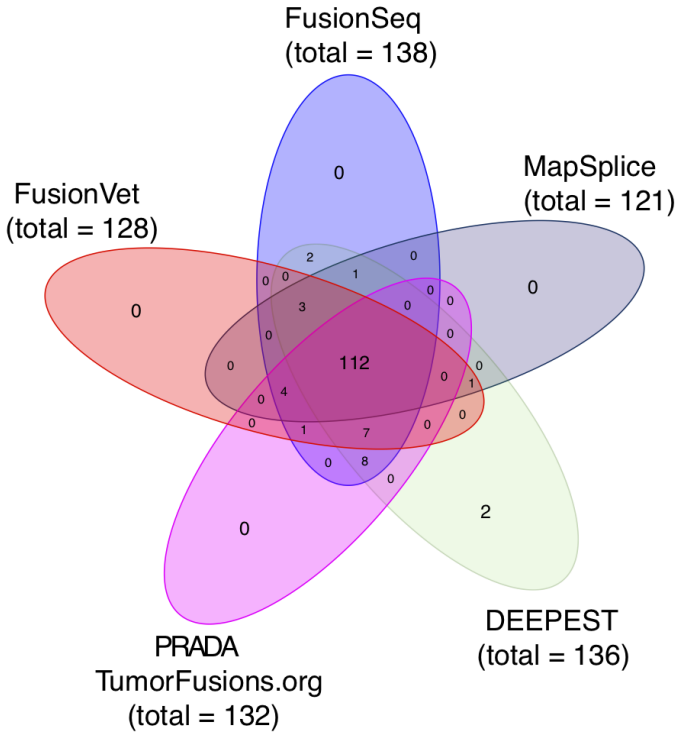
5.2 Comparison to other tools

We used FusionVet to detect ERG-TMPRSS2 fusion from the 333 TCGA prostate cancer samples. A sample is called ERG-TMPRSS2 fusion positive if it has two or more supporting fragments.

We then compare FusionVet result to:

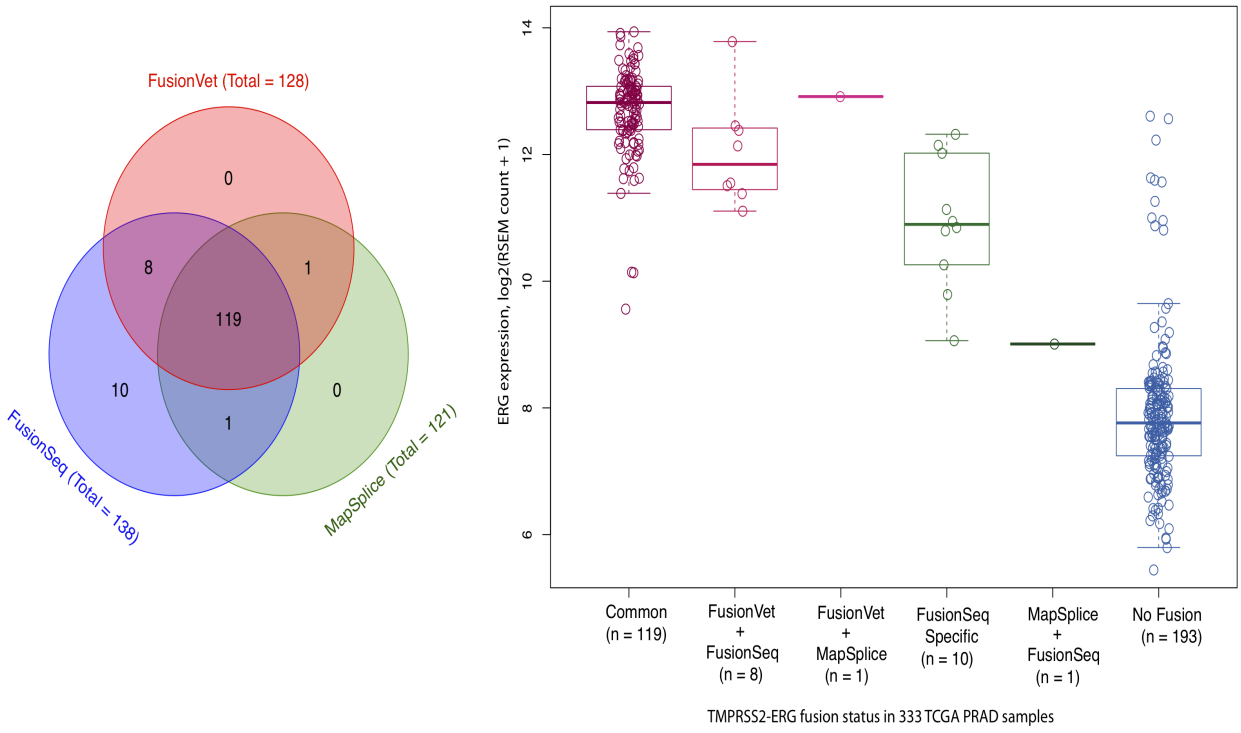
- FusionSeq-HighSens (Sboner et al., Genome Biology, 2010)
- MapSplice (Wang et al., Nucleic Acids Res. 2010)
- DEEPEST (Dehghannasiri et al., PNAS 2019)
- PRADA (tumorfusions.org) (Hu et al., Nucleic Acids Res. 2018)

We chose **FusionSeq-HighSens** and **MapSplice** because they were used in the original TCGA Cell paper. We chose **DEEPEST** and **PRADA** because they were newly developed and have demonstrated superior performance to other tools.

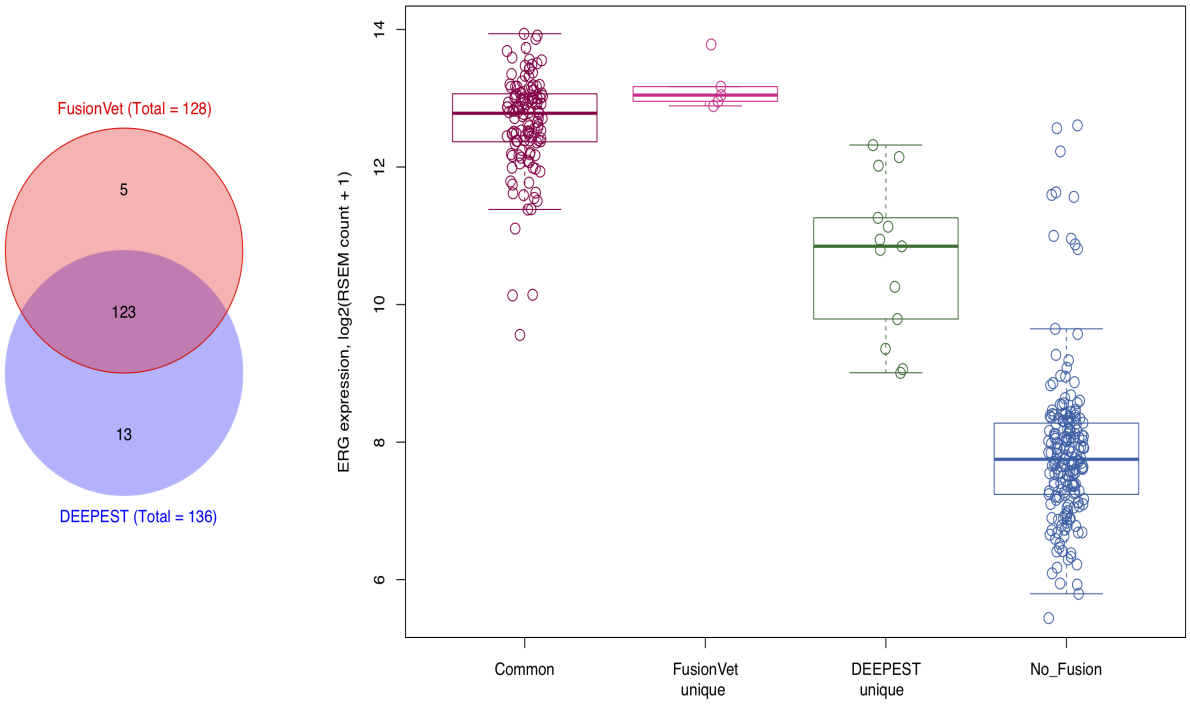


It has been found that ERG expression is significantly increased in fusion positive samples through the TMPRSS2 (an androgen responsive gene) mediated over expression (Tomlins et al., Science, 2005). Therefore, we used ERG expression as an **indirect** measurement of the authenticity of ERG-TMPRSS2 fusions.

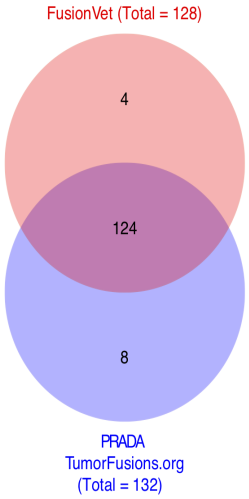
FusionVet vs FusionSeq/MapSplice

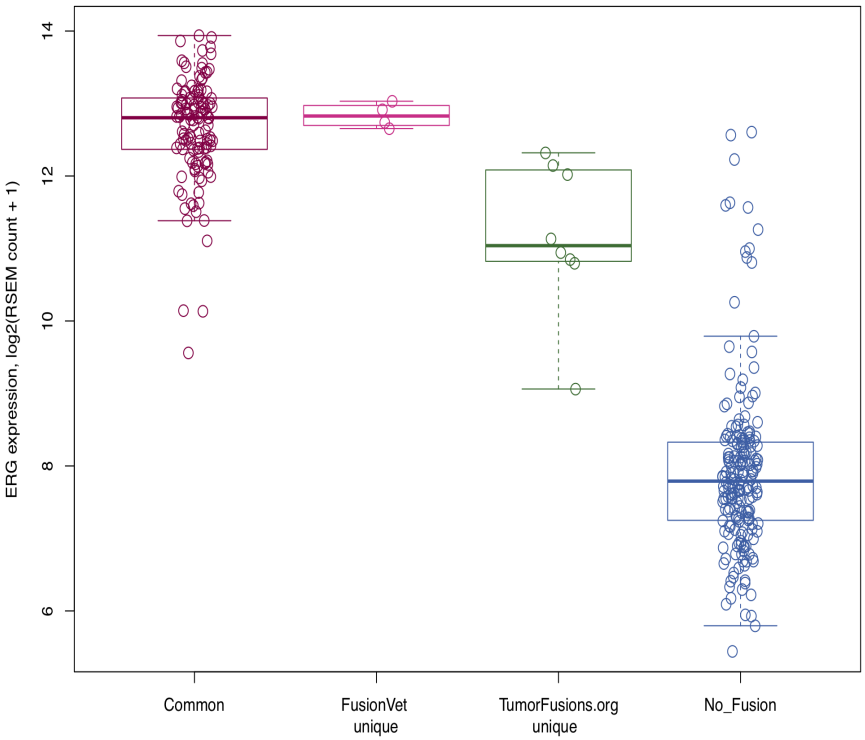


FusionVet vs DEEPEST



FusionVet vs PRADA(TumorFusions.org)


FusionVet (Total = 128)
4
124
8
PRADA
TumorFusions.org
(Total = 132)


ERG expression, log2(RSEM count + 1)
14
12
10
8
6
Common
FusionVet unique
TumorFusions.org unique
No_Fusion