

FuGePrior User Guide

FuGePrior is a novel tool for the prioritization of gene fusions from paired-end RNA-Seq data. It combines state of the art tools for chimeric transcript discovery and prioritization, a series of filtering and processing steps designed by considering state of the art literature on gene fusions and an analysis on functional reliability of gene fusion structure.

We report in the following on FuGePrior usage. Input/Output and configuration files provided together with FuGePrior code are relative to the Breast Cancer dataset from Edgren (SRA: SRP003186) that we widely analysed in FuGePrior paper.

Requirements

- UNIX operating system
- Python 2.6.6 (or 2.7.6)
- Reports from ChimeraScan 0.4.5 and deFuse 0.4.3 (or deFuse 0.6.1)
- Report from a chimeric transcript discovery tool which output can be formatted according to Pegasus general input file format (w run MapSplice-v2.1.8)
- Oncofuse Tool 1.0.9
- Pegasus Tool
- Human reference genome and relative index (we used hg19.fa and hg19.fa.fai repsectively)

FuGePrior folder:

- Scripts: Python source code
- Preprocessing_Scripts: Scripts necessary to modify chimeric transcript discovery and Pegasus output files as described in the following
- config_Example.txt: Detailed description on how to modify FuGePrior configuration file
- FuGePrior.sh: FuGePrior main script
- REFERENCE: Put in this folder hg19.fa and hg19.fa.fai files
- FuGePrior_config.txt: Configuration file for breast cancer dataset analysis
- ChimeraScan_out: ChimeraScan output files for breast cancer dataset. The files have been previously elaborated as described in the following
- deFuse_out: deFuse output files for breast cancer dataset. The files have been previously elaborated as described in the following
- genericTool_out: MapSplice output files for breast cancer dataset
- genericTool_out_forPegasus: MapSplice output files opportunely elaborated to run Pegasus analysis on breast cancer dataset

- Pegasus_Configuration_Files: Configuration files used to run Pegasus on breast cancer dataset
- Pegasus_output_File: Pegasus output file from breast cancer dataset analysis
- FuGePrior_out: PuGePrior output files from breast cancer dataset analysis
- FuGePriorUserGuide: FuGePrior user guide

Setup

- Extract FuGePrior folder
- Download hg19.fa reference genome and produce its index (hg19.fa.fai) using Samtools. Move these files into FuGePrior/REFERENCE folder
- If data on which you want to perform chimeric transcript discovery are in this format (SRP003186 fastq files):

```
@ERR031033.1.1 B800EKABXX:6:1:20894:1891#TAGCTTAT length=90
NCCAGTTCCTTTCTGTTACCCACCATTTGTCAACCCGGAGCCTCTTTTTTTTCTTTCCAAGAAGGCTGAGTTCTA
CATTGATGTGATTG
+ERR031033.1.1 B800EKABXX:6:1:20894:1891#TAGCTTAT length=90
```

Run programs `Elab_Mate1.py` and `Elab_Mate2.py` on `mate1` and `mate2` respectively:

```
python FuGePrior/Preprocessing_Scripts/Elab_Mate1.py -i input_mate1.fq -o
input1_elaborated.fq
python FuGePrior/Preprocessing_Scripts/Elab_Mate2.py -i input_mate2.fq -o
input2_elaborated.fq
```

to obtain fastq files opportunely formatted for gene fusion discovery tool run

- Run ChimeraScan, deFuse and a third gene fusion discovery tool. Save these outputs in separated folders according to the tool responsible for the detection (ChimeraScan_out, deFuse_out and genericTool_out). If the output is from ChimeraScan, nominate the file as `chimeras_sample#.txt`. Conversely, if it is from deFuse label it as `defuse_sample#.txt`, otherwise as `generic_sample#.txt`.

We provide in `FuGePrior/deFuse_out`, `FuGePrior/ChimeraScan_out` and `FuGePrior/gericTool_output` folders deFuse, ChimeraScan and MapSplice breast cancer output files respectively. `sample#` is the name of the sample that you will need to specify within Pegasus and FuGePrior configuration files. These files have been already processed as reported in “Gene fusion tool output processing” Subsection. You can use this data to test FuGePrior run.

Gene fusion tool outputs processing:

- If you run deFuse 0.6.1, the relative output files need to be corrected using `correct_defuse_output.py` script

```
python FuGePrior/Preprocessing_Scripts/correct_defuse_output.py -i
results.tsv -o FuGePrior/deFuse_out/defuse_sample#.txt
```

- If you run deFuse 0.4.3, no output elaboration is required
 - Process ChimeraScan output using `correct_chimerascan_output.py` script
- ```
python FuGePrior/Preprocessing_Scripts/correct_chimerascan_output.py -i
chimeras.bedpe -o FuGePrior/ChimeraScan_out/chimeras_sample#.txt
```

### **Pegasus run:**

- If you run MapSplice-v2.1.8, process the file containing annotated gene fusions using `correct_mapsplice_output.py` to obtain Pegasus general format input file.
- ```
python FuGePrior/Preprocessing_Scripts/correct_mapsplice_output.py -i
fusions_well_annotated.txt -o mapsplice_sample#.txt
```
- You can find these files in `FuGePrior/genericTool_out_forPegasus`
- Run Pegasus tool on results from gene fusion detection tools. In `FuGePrior/Pegasus_Configuration_Files` folder you can find an example of Pegasus configuration files for breast cancer dataset.
 - Put Pegasus output files in `FuGePrior/Pegasus_out` folder. You can find this file for breast cancer analysis in `FuGePrior/Pegasus_out_file` folder

FuGePrior run:

- Create an output folder where results will be stored. You can find FuGePrior output files relative to breast cancer dataset in the folder `FuGePrior/FuGePrior_out`
 - Produce an opportune configuration file as detailed in `config_Example.txt`. `FuGePrior_config.txt` is the configuration file we used for breast cancer analysis
 - Modify line 14 in `FuGePrior.sh` main script according to your analysis
 - Run `FuGePrior.sh` script
- ```
sh FuGePrior/FuGePrior.sh
```

### **FuGePrior Output Files:**

For each of the sample under investigation FuGePrior final output file is named `Sample#_OUT`. The output file includes, for each fusion, the following information:

- 1-Oncofuse\_DriverScore: Driver probability score from Oncofuse. A value equal to '-5' means no score reported for the fusion by Oncofuse
- 2-Pegasus\_DriverScore: Driver probability score from Pegasus.
- 3-Pegasus\_FusionID: Unique identifier attributed by Pegasus to the fusion
- 4-Sample: Name of the sample under investigation
- 5-Tool: Name of the tool responsible for gene fusion detection. If the fusion has been reported by more than a tool, this column reports on the name of the tool that has been selected for gene fusion structure analysis

6-#Spanning\_reads: Number of spanning reads supporting the fusion

7-#Split\_reads: Number of split reads supporting the fusion

8-Chr1: Chromosome of 5' partner gene (according to Pegasus output)

9-Chr2: Chromosome of 3' partner gene (according to Pegasus output)

10-Gene1\_Start: Genomic starting position of 5' partner gene (according to Pegasus output)

11-Gene1\_End: Genomic ending position of 5' partner gene (according to Pegasus output)

12-Gene2\_Start: Genomic starting position of 3' partner gene (according to Pegasus output)

13-Gene2\_End: Genomic ending position of 3' partner gene (according to Pegasus output)

14-Strand1: Strand of 5' partner gene (according to Pegasus output)

15-Strand2: Strand of 3' partner gene (according to Pegasus output)

16-Gene1\_name: Gene symbol of 5' partner gene (according to Pegasus output)

17-Gene2\_name: Gene symbol of 3' partner gene (according to Pegasus output)

18-Gene1\_BP: Genomic coordinate of the 5' partner gene breakpoint (according to Pegasus output)

19-Gene2\_BP: Genomic coordinate of the 3' partner gene breakpoint (according to Pegasus output)

20-Gene1\_ID: Official id of the 5' partner gene (according to Pegasus output)

21-Gene2\_ID: Official id of the 3' partner gene (according to Pegasus output)

22-Sample\_Type: Sample type as indicated in Pegasus configuration file

23-Sample\_Occurrence\_list: List of sample names where the fusion exactly occurs (according to Pegasus output)

24-Kinase\_Info: Information relative to the presence of kinases in the 5', 3' or both partner genes (according to Pegasus output)

25-Transcript1\_ID: Transcript ids of the 5' partner gene (according to Pegasus output)

26-Transcript2\_ID: Transcript ids of the 3' partner gene (according to Pegasus output)

27-Reading\_Frame: Information about the reading frame (according to Pegasus output)

28-Exon\_Gene1: Exon number where the breakpoint of the 5' partner gene falls (according to Pegasus output)

29-Exon\_Gene2: Exon number where the breakpoint of the 3' partner gene falls (according to Pegasus output)

30-Breakpoint1\_region: Region where the breakpoint of the 5' partner gene falls (according to Pegasus output)

31-Breakpoint2\_region: Region where the breakpoint of the 3' partner gene falls (according to Pegasus output)

32-InSample\_occurrence: Number of fusions detected by the tool of Col.5 with same breakpoints but different supporting reads

33-FusionDetection\_Tools: List of gene fusion detection tools that reported on the fusion

34-Consensus Sequence: According to Col. 5 it reports on ChimeraScan split reads or third tool/deFuse consensus sequence

35-Biological\_Mechanism: Information relative to the gene fusion structure and the relative FuGePrior reconstructed virtual reference