# C:\Users\Chengjun\Desktop\Griffin-mythology.pngDocument for GriffinDetector

Version 1.0 2014-01-04

Author: Chengjun Zhang @ the University of Chicago

[Document for GriffinDetector 1](#_Toc376689797)

[Version Changes: 2](#_Toc376689798)

[The function of the pipeline: 2](#_Toc376689799)

[Program installation: 2](#_Toc376689800)

[Data preparation: 2](#_Toc376689801)

[The advantages of the pipeline: 2](#_Toc376689802)

[Notation: 2](#_Toc376689803)

[Citation: 2](#_Toc376689804)

[Bug Report: http://longlab.uchicago.edu/sites/default/files/email.png 2](#_Toc376689805)

[Documentation of options: 3](#_Toc376689806)

[Shell parameters: 3](#_Toc376689807)

[Parameters in control file: 3](#_Toc376689808)

## Version Changes:

GriffinDetector Version 1.0

Passed the test based on the CentOS Linux release 6.0, Perl version 5.10, BLAST+ version 2.2.28, with rice tribe genome data.

## The function of the pipeline:

The main purpose of this program is to detect candidate fusion genes from annotated genomes. The fusion genes here are especially means the fusion genes that combined from two parental genes.

### ****Program installation:****

This pipeline was written by Perl script, it requires makeblastdb and blastp which belong to BLAST+ package. Before running this program, please make sure the BLAST+ package was installed. We suggest making a link at /usr/local/bin/.

### ****Data preparation:****

1. The phylogeny for the species are interested (at least three species)
2. The protein sequences for these species
3. The gff3 file for the relative species (the modification may be required)
4. The outgroup species and focus species should explicit

### ****The advantages of the pipeline:****

1. This pipeline will automatically detected the fusion genes based on the phylogeny with several species;
2. It was very easy to adjust the phylogeny, annotation and the combination of species to get fusion genes in focus species as more as possible with simple modification;
3. Since this pipeline works on several genomes, the abundance of genome data will largely relieve the bad affect caused by imperfect assembling and annotation;

### ****Notation:****

1. The hit coverage of the query gene and the hit length in some cases may not accurate;
2. The gene loss events was assume not appear in groups, so the final fusion genes may underestimated;
3. The annotation of focus species may affect the final results a lot.

### ****Citation:****

### ****Bug Report:****http://longlab.uchicago.edu/sites/default/files/email.png

**FAQs：**

## Documentation of options:

### Shell parameters:

Running GriffinDetector has two types:

1. perl GriffinDetector.pl ***control\_file***
2. perl GriffinDetector.pl ***control\_file* Y  *path\_2\_identified\_fusion\_genes***

The ***control\_file***option was let the pipeline know the informations which are required by the program; the “**Y**” option and “***path\_2\_identified\_fusion\_genes***” was used when **non-gff** required version had been finished running, then, the directory of “**identified\_chimerics**” generated by the **non-gff** version was set as a parameter, more informations will be generated compare to the **non-gff** version (four more files: chimera\_gff\_file.txt, chimera\_tree\_svg.txt, chimera\_gene\_id.txt, chimera\_gene\_detail.txt, these four files were input data for our future project, **GriffinDB**).

### Parameters in control file:

|  |
| --- |
| **SPECIES**  Spe1:path/2/protein/seq/file1  Spe2:path/2/protein/seq/file2  Spe3:path/2/protein/seq/file3  …  **SPECIES**  **SPEGFFS**  Spe1:path/2/protein/gff/file1  Spe2:path/2/protein/gff/file2  Spe3:path/2/protein/gff/file3  …  **SPEGFFS**  **TREE**:(spe3,(spe1,spe2))…;  **INGROUPS**:spe1,spe2…  **OUTGROUPS:**spe3…  **Makeblastdb:** path/2/blast+/package/makeblastdb  **Blastp:** path/2/blast+/package/blastp  **SKIP:**Y  **OUTFILE**:blast\_plus\_db |

In the control file, **SPECIES** and **SPEGFFS** are the keywords presented in pairs (in **non-gff** version,the **SPEGFFS** keywords will be ignored). Between the pair of keywords, was the information of protein sequences file and gff file. Each item was separated into two parts by “**:**”, the former one was the nick name for species; the latter one was the path to relative species.

The following key words were all presented by **bold face**, and all are separated by “**:**”, the line followed by **TREE** was the phylogeny used in the pipeline, only Newick format (<http://en.wikipedia.org/wiki/Newick_format>) with named leaf nodes are allowed. Following the key words **INGROUPS** and **OUTGROUPS**, was the species nick names separated by “**,**”, the nick names are not allowed to presented in both groups. All the species nick names used through **SPECIES**, **SPEGFFS**, **TREE**, **INGROUPS** and **OUTGROUPS** should not be altered.

The **SKIP** and **OUTFILE** key words only works when the **SKIP** was set as “**Y**”, then, the directory including the blastp results for each focus species (the species presented in the **INGROUPS**) to all other species (including itself), should be offered; this design will save times when people want to add or change the annotation for **non-INGROUPS** species—for example, spe3 was an OUTGROUPS species, it was replaced by spe4, then, if spe1 and spe2 are the **INGROUPS** species, then, people can manually running:

|  |
| --- |
| blastp –db **spe4-db** –query **spe1** –out **spe1**2**spe4**\_aa2aa\_blast+.query -evalue 0.00001 -outfmt 7 -max\_target\_seqs 10 -num\_threads 10  blastp –db **spe4-db** –query **spe2** –out **spe2**2**spe4**\_aa2aa\_blast+.query -evalue 0.00001 -outfmt 7 -max\_target\_seqs 10 -num\_threads 10 |

and then add the results to the previous document; to retain the same threshold, changing the parameters inside are discouraged.