Documentation for GFP

Contact

wclee47@gmail.com for reporting errors and addressing problems.

Alignment with GSNAP

GFP reads "GSNAP output format" explained here: http://research-pub.gene.com/gmap/src/README.
When you align read sequence files using GSNAP, please do not forget to specify following 3 options.

1. -s, --use-splices=<STRING>
You might need to prepare required files for <STRING>. Please refer to "Detecting known and novel splice sites in GSNAP" section from here: http://research-pub.gene.com/gmap/src/README.

2. -N, --novelsplicing=<INT>
Please set <INT> to 1 so that GSNAP looks for novel splicing.

3. -E, --distant-splice-penalty=<INT>
Please set <INT> to 0, otherwise GSNAP might not report distance splices naturally generated by fusion events.

Prerequisites

1. "pygr", a Python module (can be downloaded from http://code.google.com/p/pygr/).
2. "bl2seq" program in the BLAST package.

Installation

1. Download GFP from ftp://ftp.gmi.ac.kr/pub/GFP-v0.7/.

2. Unpack the downloaded file.

```
tar xvzf GFP-v0.7.tar.gz
```

3. Enter the directory "GFP-v0.7" and install. Users can change the installation directory by specifying "setup.py" options (please refer to the "setup.py" manual)

```
cd GFP-v0.7
python setup.py install
```

Preparation (building exon index directory)

1. Preparing preliminary files

Before building exon index directory, three types of files are necessary.

(1) BED files (e.g. chr1.bed, chr2.bed …) for exon information. Each bed file should be first sorted by exon's start position and then by exon's end position. Tab-delimited column names of the files are:

① chromosome

② exon's start position

③ exon's end position

④ transcript's accession . gene name . exon number

⑤ 0 (always)

⑥ transcribed strand

```
…
chr1   850983    851043    NM_152486.SAMD11.exon1   0   +
chr1   851164    851256    NM_152486.SAMD11.exon2   0   +
chr1   855397    855579    NM_152486.SAMD11.exon3   0   +
…
```

(2) "transcript.bed" for transcript information. Tab-delimited column names of the file are:

① chromosome

② transcript's start position

③ transcript's end position

④ transcript's accession

⑤ 0 (always)

⑥ transcribed strand

```
chr1   67051161 67163158 NM_024763     0   -
chr1   67075872 67163158 NM_207014     0   -
chr1   8335051   8800286   NM_001042681     0   -
chr1   8335051   8406334   NM_001042682     0   -
…
```

(3) "transcript.fasta" for transcript nucleotide sequence information (FASTA format).

```
>NM_001005484
ATGGTGACTGAATTCATTTTTCTGGGTCTCTCTGATTCTCAGGAACTCCAGACCTTCCTATTTA
TGTTGTTTTTTGTATTCTATGGAGGAATCGTGTTTGGAAACC…
>NM_001005224
ATGGATGGAGAGAATCACTCAGTGGTATCTGAGTTTTTGTTTCTGGGACTCACTCATT…
…
```

If a user plans to use "refGene,txt" (can be downloaded from UCSC database (http://http://genome.ucsc.edu/) as a gene annotation file, all the necessary files are automatically created by running "refGene2bed.py" included in GFP package. Otherwise, users should make a directory containing all the necessary files formatted as above.

2. Building exon index directory

Once a directory with all the preliminary files is prepared, building the exon index directory is straightforward. Please run "build_idxDir.py" also included in the package.

Running GFP

```
GFP --- A tool to detect fusion genes using RNA-Seq

Required parameters
        -i <string>             GSNAP result file.
        -d <string>             Pre-built exon index directory.
        -o <string>             Output prefix.
        --bl2seq <string>       bl2seq excutable path.

Optional parameters
        --mpair <integer>       Minimum # of discordant read-pairs, DEFAULT: 1.
        --mspan <integer>       Minumum # of fusion spanning reads, DEFAULT: 2.
        --mcov <integer>        Minimum # of base-pairs for both genes, DEFAULT: 10.
        --mshift <integer>      Minimum # of shifting pattern(bp), DEFAULT: 1.
```

Once building the exon index directory is completed, you are ready to run GFP. The parameter setting is shown above.

<Optional parameters>
(1) --mpair : only gene fusions with ≥ (value) will be reported.
(2) --mspan: only gene fusions with ≥ (value) will be reported.
(3) --mcov: fusion-spanning reads which cover any of the exons from the two genes by < (value) will be discarded.
(4) --mshift: a fusion point (exon-exon boundary) is considered a genuine fusion point when fusion-spanning reads around the fusion point show at least (value) shifting pattern.

Output Files

GFP generates three output files:

(1) "<output prefix>_raw.txt" – shows raw fusion evidence extracted from GSNAP alignment results.

(2) "<output prefix>_fusionList.txt" – shows gene fusions which satisfied user-defined parameter settings and passed through filtering steps implemented by the program and its format is described below.

Name	Type	Description
ID	String	Serial number for each gene fusion discovered.
donor	String	Donor gene located at 5' position in fusion context.
acceptor	String	Acceptor gene located at 3' position in fusion context.
context	String	Fusion context: either "INTRA" or "INTER".
dist	Integer	Genomic distance between fusion genes.
num_pair	Integer	# of fusion-supporting discordant read-pairs.
num_span	Integer	# of fusion-supporting fusion-spanning reads.

(3) "<output prefix>_fusionEvidence.txt" – contains more detailed information on gene fusions listed in the gene fusion list file including genomic positions and exon numbers for supporting fusion evidence.

Name	Type	Description
ID	String	Serial number for each gene fusion discovered.
evidence_type	String	Evidence type: either "read-pair" or "spanning_read".
donor_pos	String	Genomic position in donor gene.
acceptor_pos	String	Genomic position in acceptor gene.
donor_exon	String	Exon number(s) in donor gene.
acceptor_exon	String	Exon number(s) in acceptor gene.