FusionFinder

Reference Manual - v1.2 (16/03/2012)

http://bioinformatics.childhealthresearch.org.au/software/fusionfinder

Richard Francis et al, TICHR 2012

Find Fusions in RNA-Seq read data

NAME

fusionfinder.pl

DESCRIPTION

When given a single file or list of files that contain FASTQ reads and the location of two reference files, this script will run through the FusionFinder protocol in order to identify potential fusion transcripts present in the data.

This involves the following series of steps:

1. Alignment of the reads to the reference containing coding transcripts
2. Creation of pseudo paired-end reads from those reads having no match in step 1
3. Alignment of pseudo paired-end reads to the coding reference
4. Storing reads evidencing fusions and filtering those that relate to false positives
5. Output and summary of fusion candidates

For a full description of this software please see http://bioinformatics.childhealthresearch.org.au/software/fusionfinder/

SYNOPSIS

fusionfinder.pl [options]

OPTIONS

--reads

Specifies the filename of the FASTQ formatted reads file(s). More than one file can be submitted as a quoted, comma separated list. [Required]

--cref

Specifies the filename of the coding reference file. [Required]

--ncref

Specifies the filename of the noncoding reference file. [Required]

--mp_cutoff

Specifies the cutoff that gives the minimum number of read pairs that indicate the existence of a G1:G2 pair before it is considered "real" (default = 4)

--readthrough

Specifies the distance in bp to use as the cutoff for defining a potential read-through transcript (default 20,000) as per Nacu 2011 where the majority of read-through transcripts observed involved genes that were within 20kb of each other

--config

Tells the script the location of your configuration file.

At a minimum this file should contain the location of your ensembl API root directory as follows:

```
[API]
path=/full/path/to/api/root
```

By default this script will connect to the central UK Ensembl

database (ensembldb.ensembl.org), however we recommend that you use the Ensembl database closest to you. If you wish to use a local Ensembl database or alternative mirror the ensembl hostname, username and password details of the server you are connecting to should also be provided in the configuration file.

Details of how to simply install a local Ensembl database can be found on the project website

The Ensembl details should be provided as follows:
[Ensembl]
hostname=
user=
pass=

FusionFinder runs Bowtie for the alignment steps of the protocol. If Bowtie is not on your path, you can specify the path to the program as follows:
[Bowtie]
path=/full/path/to/Bowtie

--threads
Specifies the number of threads to use for the Bowtie alignment steps (default = 1)

--[no]fhitzero
Enables/Disables the filter applied if the 5' read of the pseudo pair has no hits in the provided sam file (enabled by default. use --nofhitzero to disable this option)

--[no]lhitzero
Enables/Disables the filter applied if the 3' read of the pseudo pair has no hits in the provided sam file (enabled by default. use --nolhitzero to disable this option)

--[no]flparas
Enables/Disables the filter applied if the genes the pseudo pairs hit are paralogs according to Ensembl (enabled by default. use --noflparas to disable this option)

--[no]flstrands
Enables/Disables the filter applied if the genes the pseudo pairs hit exist in the same genomic location but on opposite strands (enabled by default. use --noflstrands to disable this option)

--[no]blockrepeat
Enables/Disables the filter applied if the region covered by the alignment block on G1 contains the same class of repeat (according to RepeatMasker) as the region covered by the alignment block on G2 (enabled by default. use --noblockrepeat to disable this option)

--5ratio
Specifies the ratio to use for the size of the 5' pseudo paired-end read with respect to the parent read length. Default is 0.4 which would produce a 30mer from a 75mer parent.

--3ratio
Specifies the ratio to use for the size of the 3' pseudo paired-end read with respect to the parent read length. Default is 0.4 which would produce a 30mer from a 75mer parent.

--phred33
Tells the script that the quality scores in the FASTq file are based on a Phred score with an offset of 33, ie Sanger format. Default is 64, Illumina format

--species
Specifies the species your data relates to. Default is human.

--help
Displays this help

BUGS

Please report them to <bioinformatics@ichr.uwa.edu.au>

COPYRIGHT

Generation of multiple alignments for interesting fusion candidates

NAME

make_alignments.pl

DESCRIPTION

When given the reads output file from find_fusions.pl and the HGNC symbols of a G1:G2 pair, this script will generate three alignment files to assist in the identification of a breakpoint.

1. The sequences of all G1 transcripts containing the implicated G1 exon, the G1 exon itself and all 5' members of pseudo paired-end reads providing evidence for G1

2. The same for G2 and all 3' pseudo paired-end read members evidencing G2

3. The sequences of all parent full length reads and the implicated G1 and G2 exons

For a full description of this software please see http://bioinformatics.childhealthresearch.org.au/software/

SYNOPSIS

make_alignments.pl [options]

OPTIONS

--readsfile
Specifies the filename of the reads output file from find_fusions.pl

--g1
Specifies the HGNC symbol for G1 as per the output file from find_fusions.pl

--g2
Specifies the HGNC symbol for G2 as per the output file from find_fusions.pl

--noalign
Tells the program to only extract the sequence files and not to align them. This can be used if you have not installed Muscle locally and want to align the sequences manually, eg at www.ebi.ac.uk/muscle

--limit
Tells the program to limit the number of reads aligned. This is particularly useful if there is considerable read evidence for a

fusion candidate as alignments can take some time to complete.

--flank

Specifies the number of bases to extract from transcripts that are upstream of the G1 exon and downstream of the G2 exon.

These sequences are aligned to the 5' and 3' pseudo paired-end reads respectively. Default is 250 bases.

--config
Tells the script the location of your configuration file.

At a minimum this file should contain the location of your ensembl API root directory as follows:

[API]
path=/full/path/to/api/root

By default this script will connect to the central UK Ensembl database (ensembldb.ensembl.org), however we recommend that you use the Ensembl database closest to you. If you wish to use a local Ensembl database or alternative mirror the ensembl hostname, username and password details of the server you are connecting to should also be provided in the configuration file.

Details of how to simply install a local Ensembl database can be found on the project website

The Ensembl details should be provided as follows:
[Ensembl]
hostname=
user=
pass=

--help
Displays this help

BUGS

Please report them to <bioinformatics@ichr.uwa.edu.au>

COPYRIGHT

Creation of a reference transcriptome

NAME

make_reftrans.pl

DESCRIPTION

This script will connect to an Ensembl mirror database and create two FASTA files one containing all transcripts that are known to be protein coding and the other containing those that do not code for a protein. The tag line for each sequence in this file is formatted for use with fusionfinder.pl

For a full description of this software please see
http://bioinformatics.childhealthresearch.org.au/software/

SYNOPSIS

make_reftrans.pl [options]

OPTIONS

--species
Specifies the species to retrieve this transcriptome from. Default is human.

--basedir
Specifies the base directory where you would like to create the transcriptome file. Default is the current directory

--config
Tells the script the location of your configuration file.

At a minimum this file should contain the location of your ensembl API root directory as follows:

[API]
path=/full/path/to/api/root

By default this script will connect to the central UK Ensembl database (ensembldb.ensembl.org), however we recommend that you use the Ensembl database closest to you. If you wish to use a local Ensembl database or alternative mirror the ensembl hostname, username and password details of the server you are connecting to should also be provided in the configuration file.

Details of how to simply install a local Ensembl database can be found on the project website

The Ensembl details should be provided as follows:
[Ensembl]
hostname=
user=
pass=

--help
Displays this help

BUGS

Please report them to <bioinformatics@ichr.uwa.edu.au>

COPYRIGHT