```
###############################################################################################
#                                                                                             #
# fishintron.pl                                                                               #
#                                                                                             #
# Author: Chenhong Li                                                                         #
#         School of Biological Sciences                                                       #
#         University of Nebraska - Lincoln                                                    #
#         NE 68588-0118 USA                                                                   #
#                                                                                             #
#                                                                                             #
# Created by:   Sept. 27, 2008                                                                #
# Last modified by :   , 2012                                                                 #
#                                                                                             #
###############################################################################################
```

INTRODUCTION

This is a program for developing universal exon-primed intron-crossing (EPIC) nuclear markers. Because protein-coding sequences (CDS) are the true conserved parts of exons, we screen for CDS instead of exons themselves. It searches genome sequences to find introns that are less than a given size (< $Li). The flaking CDS shall be "single-copy" and conserved among interested taxa. So primers can be desinged on the flanking region to amplify the homologous intronic fragments in a wide range of taxa. The universal EPIC markers are useful in population genetics and genomic studies for species with no genome sequence avaliable.

There are three steps in this project. First, finding the single-copy conserved CDS; second, screening for CDS-bounding introns of an appropriate size; third, designing primers for the EPIC markers. The first step, finding single-copy conserved CDS could be very useful itself for developing high-level phylogenetic markers, so this program can be run for the first step only as an option. The last step of sequence alignment and primer desiging have not been automated, since a lot of "check by eyes" are needed still.

INSTALLATION

This program is written in PERL, so it is very easy to set it up. Just download the folder containing all scripts. If you are reading this README file, you have done that already. There is only one thing you need to do is to change the path in fishintron.pl to your current perl directory, which you can find by typing "which perl" in a terminal. Then substitute the path in the first line of fishintron.pl with the one you find.

There are five scripts or modules in this program:

- makeindex.pl
  - used to make the index file for subject sequences to accelerate search.
- fishintron.pl
  - the main script taking input and execute the tasks.
- makequery.pm
  - the module used to screen for large and merged CDS.
- blastandparse.pm
  - the module used to perform within-genome and between-genome blast; parse the results and select single-copy conserved CDS markers.
- EPIC.pm

the module used to screen for and print qualified EPIC markers.

External program:
BLAST should be installed (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The directory for BLAST also need to be changed to your directory in the perl scripts.

Data files:
Genome sequencies in FASTA format for both query and reference taxa (e.g., I downloaded ENSEMBL toplevel.fa file). The files should be saved in folder for each species separately (e.g. ../Danio_rerio.db/, you can substitute Danio_rerio with your taxon name).

Genome sequencies in EMBL format for query species (e.g., I downloaded the EMBL files from ENSEMBL). The EMBL files are used to parse information of CDS, since the CDS information given by BioMart in ENSEMBL has errors. All EMBL data file should be saved in the folder for each species separatly (e.g., ../Danio_rerio.db/embl/, you can substitute Danio_rerio with your taxon name).

Format database:
The genome sequence file should be formated into local blast database using formatdb (http://www.ncbi.nlm.nih.gov/blast/docs/formatdb.html) (e.g. formatdb -p F -i Danio_rerio.fasta -n Danio_rerio.db). Index files should also be made to facilitate extracting sequences using makeindex.pl before running the main program, see the running fishintron.pl.

BEFORE RUNNING fishintron.pl

Before running the main program, first make index file using makeindex.pl, example usage:
./makeindex.pl ../Danio_rerio.db/Danio_rerio.ZFISH7.50.dna.toplevel.fa ../Danio_rerio.db/Danio_rerio.genome

You should substitute Danio_rerio with your taxon. The input file can be any name, but the output file must be yourtaxonname.genome. Please note the genus name and species name are linked by a low dash.

RUNNING fishintron.pl

Input:
The scientific name of the query species and the reference species (note, use a low dash to link the genus name and species name, separate reference taxa by a space), Values for minimum CDS size, ($Lc), maximum intron size ($Li), blast E-value ($E), identity within genome ($Iwg), coverage within genome ($Cwg), identity between genome ($Ibg), coverage between genome ($Cbg) and $CDS (if "yes", searches for CDS markers only). For more options, see OPTIONS.

Output:
The major output files include a text file listing queries that have exactly one hit in each of the reference genome when the query CDS was blasted against them, fasta files of aligned CDS markers (optional), information of selected EPIC markers (optional), and fasta files of aligned EPIC markers (optional) for designing primers to amplify the introns.

Example usage:
Change present working directory to the folder contain all scripts, then type:
./fishintron -query "Danio_rerio" -reference "Gasterosteus_aculeatus Oryzias_latipes Takifugu_rubripes Tetraodon_nigroviridis" -CDS y -Lc 800

For more options, see OPTIONS.

OPTIONS

-h: print a simple version of help information;

-query: variable for store query species name, use a low dash to link your genus and species name;

-reference: variable for store the reference species name, use a low dash to link your genus and species name, separate reference taxa by a space;

-Lc: parameter for minimum CDS size, the default is 100. If you search for CDS markers only, you might want to increase it to 700;

-Li: parameter for maximum intron size, the default is 1000, which mean the intron size is less than 1000bp in at least one of your taxa;

-Iwg: within-genome blast parameter for identity, the default is 40, smaller number results in stricter "single-copy", but maybe fewer markers;

-Cwg: within-genome blast paramter for coverage, the default is 20;

-Ibg: between-genome blast parameter for identity, the default is 55;

-Cbg: between-genome blast paramter for coverage, the default is 70;

-E = 0.000001: E-value for blast, the default is 0.000001;

-pid = 85: identity for printing EPIC. If average identity of flanking cds > -pid, print the squence alignment, because you don't want to print out too many FASTA files. Use 85 as a start;

-CDS: whether search for CDS markers only, the default is "N"ot. If you are aimed to search for CDS markers but don't care about EPIC markers, type -CDS y

DISCLAIMER

This program has been successfully used in finding CDS and EPIC markers for ray-finned fishes and CDS markers for chondrichthyans. However, it has not been tested in other taxanomic groups, so I cannot ensure it is free of bugs.