CSZ Manual

Characterization of Small RNAome for Zebrafish

Version 1.0.0
10/08/2013

Author: Yu Xue & Yuangen Yao

Contact: Dr. Yu Xue, xueyu@mail.hust.edu.cn; Yuangen Yao, yygen89@gmail.com
The CUCKOO Workgroup
The latest version of CSZ is available from http://csz.biocuckoo.org/

Index

Statement

1. **Implementation**. The CSZ of the CUCKOO Workgroup are implemented in Perl v5.10.1. Currently, only local stand-alone packages of LINUX system will be provided.
2. **Availability**. Our softwares are freely available for academic researches. For non-profit users, you can copy, distribute and use the softwares for your scientific studies. Our softwares are not free for commercial usage.
3. **Usage**. Our softwares are designed in an easy-to-use manner. Also, we invite you to read the manual before using the softwares.
4. **Updation**. Our softwares will be updated routinely based on users' suggestions and advices. Thus, your feedback is greatly important for our future updation. Please do not hesitate to contact with us if you have any concerns.
5. **Citation**. Usually, the latest published articles will be shown on the software websites. We wish you could cite the article if the software has been helpful for your work.
6. **Acknowledgements**.

This work was supported by grants from the National Basic Research Program (973 project) (2012CB910101, 2013CB933903, 2010CB945401, 2013CB945300, and 2012CB911201), Natural Science Foundation of China (31171263, 81272578, 31071154, 31000640, 31171387, 81090414, 81230052 and 91019020), and International Science & Technology Cooperation Program of China (0S2013ZR0003).

Introduction

Here we designed a zebrafish-specific algorithm of ZmirP (zebrafish miRNA prediction), with 8 new and 57 previously reported sequence and structure features. These features were combined together to construct an SVM model for validating the miRNAs predicted from MIREAP and miRDeep2. The performance and robustness of ZmirP were extensively evaluated by 4-, 6-, 8-, 10-fold and leave-one-out validation (LOO). Compared with other existing approaches, ZmirP exhibits greater sensitivity of 95.64% and specificity of 98.84%. Then we greatly improved the CPSS (1) and developed a more specific platform as CSZ (characterization of small RNAome for zebrafish) for the analysis of the high-through sequencing data.

As applications of ZmirP and CSZ, the total RNAs during eight distinct stages, including 1-cell (0.2 hpf (hours post fertilization)), 16-cell (1.5 hpf), 512-cell (2.75 hpf), oblong (3.7 hpf), 5.3 hpf, 6-somite (12 hpf), 24 hpf and 48 hpf (2), of zebrafish early embryonic development were isolated and used for the small RNA-seq. Then the expression profiles of small RNAs (sRNAs) in eight early zebrafish developmental stages were analyzed with CSZ. From the results, we observed that the expression levels of piRNAs are gradually decreased, while miRNA expressions are gradually increased during early embryonic stages. Thus, the sRNA class transition from piRNA to miRNA was confirmed in early zebrafish embryonic development. Furthermore, we observed that the diverse and complex of expression patterns and levels of 129 known miRNA families are dramatically increased as development proceeds. Moreover, 25 novel miRNA candidates were predicted by CSZ with high confidence. We randomly selected three predicted miRNAs for further experimental investigation, and two of them, m0027-5p and chr6_7844-5p, were confirmed through Northern blots. In addition, widespread expression of piRNAs before MZT suggested piRNAs may play a potential role during early development. Taken together, our studies contributed valuable clues for further investigating the sRNA regulation of embryonic development, and provided useful techniques for small RNAome analysis. The CSZ package was implemented in Perl and freely downloadable at: http://csz.biocuckoo.org/

Prerequisites

1. Linux system with perl v5.10.1 or higher installed

2. Please ensure the following software are installed before proceeding

(1) miRDeep2: a completely overhauled tool for identifying both known and novel microRNAs with deep sequencing data. The release of mirdeep2_0_0_1 is recommended to installation, which is available from https://www.mdc-berlin.de/36381670/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/older_miRDeep2_versions/mirdeep2_0_0_1.zip. Then follow the installation instructions of miRDeep2 in the README file to finish installation. More details about miRDeep2 are provided at https://www.mdc-berlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep.

(2) MIREAP: a tool for discovering both known and novel microRNAs from small RNA deep sequencing data by Solexa/454/Solid technology. You can obtain a free copy at http://sourceforge.net/projects/mireap/files/mireap/. We recommend that users could download the latest release mireap_0.2. Then follow the installation prompts of MIREAP to finish installation.

(3) LIBSVM: an integrated tool for support vector classification, regression and distribution estimation. The version 3.16 of LIBSVM is recommended to installation and is freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/oldfiles/libsvm-3.16.zip. For more details and for information on the LIBSVM, please visit website of LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Installation

1. Download CSZ

Latest Released Version 1.0.0: 10/08/2013: http://csz.biocuckoo.org/.

2. Unpack distribution

Unpack the distribution in your specified directory (e.g. /home/yyg)
(1) cd /home/yyg
(2) unzip csz_1.0.zip

3. Installation

(1) A provided perl script can facilitate configuration of prerequisites for CSZ. With the provided install.pl script type: please run perl install.pl if user have permission to installation software.

(2) Without the install.pl script type: follow the instructions given in miRDeep2, MIREAP and LIBSVM to finish installation.

During installation, user should not get any error messages otherwise something is not correctly installed.

Usage of ZmirP

ZmirP is a module of the CSZ, but can be used independently. When invoking ZmirP for prediction of zebrafish-specific pre-miRNAs, you need three files: a FASTA file containing the putative pre-miRNA sequences you want to validate, a SVM model file containing parameters and support vectors, and a scaled model file containing parameters for scaling input data for LIBSVM. The default SVM model file and scaled model file were included in the model directory with the CSZ distribution. Before running ZmirP, make sure that sequence file and both model files can be accessed by programs.

Please make sure that the description line for each sequence in FASTA files is unique and DNA/RNA sequences do not include weird characters except A, C, G, T and U. Otherwise, the sequences will be skipped directly and not be outputted in final prediction results. More details about FASTA format is provided at http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml.

Usage:

perl zmirp_predict.pl [options] -L [LIBSVM path] -P <fasta> -M <SVM model> -H <scaled model>

Options:

--LIBSVM/-L [path] : the path to LIBSVM package. This is the optional and the default path is ./Packages/libsvm-3.16/

--posFasta/-P <fasta> : the query fasta file containing putative pre-miRNA sequences

--SVMmodel/-M <SVM model> : the SVM model file containing parameters and support vectors

--scaleModel/-H <scaled model> : the scaled model file containing parameters for scaling input data for LIBSVM

--dir/-D [dir] : the result directory containing predicted results. This is the optional.

--help/-h : print help information and quit

Usage of CSZ

The procedure of CSZ is shown in Fig. 1. First, we only reserved unique reads ranging from 18 to 35 nt. The short reads observed with at least three times were assumed to be potential sRNA molecules or degradation fragments of larger RNAs (3), and were mapped to the zebrafish reference genome using Bowtie, with only one nucleotide mismatch (4). Then the identified reads were subsequently mapped to miRBase (5), Rfam (6), repeat sequences (7), RefSeq mRNAs (7), and piRNABank (8) (Fig. 1). By this procedure, the reads were successively classified into the following categories, including miRNA, rRNA, tRNA, and snRNA/snoRNA, genomic repeat, mRNA, and piRNA. Because repeat sequences in the annotation file were pre-classified into different classes, such as rRNA, tRNA, and snRNA/snoRNA, these RNAs were removed from the repeat sequences and recalled back to the four groups (Fig. 1). Also, because piRNAs can locate in repetitive sequences, we further identified potential piRNAs by mapping other repetitive sequences to piRNABank (Fig. 1).

Subsequently, unknown sequences that could not be assigned to any of known categories, were used to detect potentially novel miRNAs using MIREAP (https://sourceforge.net/projects/mireap/) and miRDeep2 (9) with the default settings. Because too many putative results were generated by MIREAP or miRDeep2, ZmirP algorithm was adopted for further filtering potentially false positive hits.

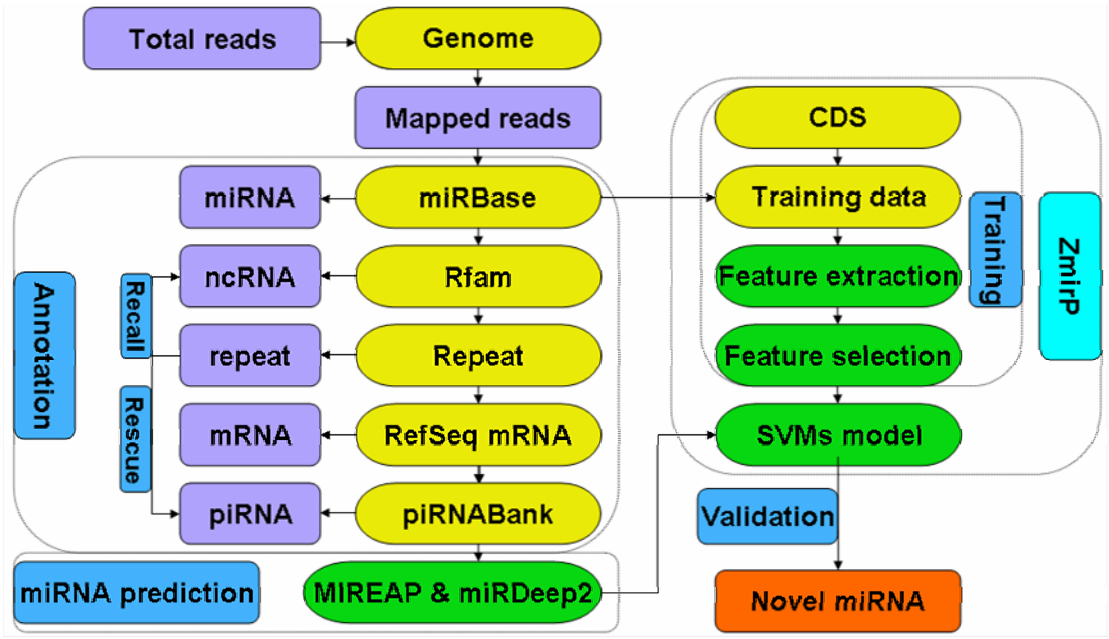

Figure 1. General workflow of CSZ

(1) Data preparation: When invoking CSZ for characterization of small RNAome from high-throughput sequencing data, you need a series of files. The first FASTA file contains the high-throughput sequencing data in the following format:
>t0000018 10357
TGGATAACTGAAAGCACCGGAAACTGGA
Or
>t0472076_x5
AAACACAACTGAAGCACATGGAAGAATG
Where t0000018 and t0472076 indicate the respective sequence names, 10357 and 5 stand for the observed frequency of each sequence in sequenced libraries. An example of sequencing data is deposited in the sRNA_seq directory with the CSZ distribution.

The second and third FASTA files containing the respective reference genome sequences and refMrna sequences are downloaded directly from UCSC database http://hgdownload.soe.ucsc.edu/downloads.html. For convenience, please name the downloaded files as XXX.genome.fa and XXX.mRNA.fa, respectively. Unless stated otherwise, XXX stands for the three letter acronym of specified species in miRBase database. More detailed about it, please read organisms file in current CSZ distribution.

The fourth file is RepeatMasker .out file deposited in UCSC and downloaded from http://hgdownload.soe.ucsc.edu/downloads.html. Then repeat sequences of specified species are extracted from reference genome according to annotations in the RepeatMasker .out file through running perl script getRepeatSeq.pl.

The fifth FASTA file containing CDS of specified species is retrieved from the Table Browser of UCSC. For convenience, please name the retrieved file as XXX.cds.fa.

The sixth FASTA file, named Rfam.fasta in Rfam database (downloaded from: ftp://ftp.sanger.ac.uk/pub/databases/Rfam/), contains all kinds of non-coding RNA. Sequences of rRNA, tRNA, snRNA and snoRNA are extracted automatically from this FASTA file when CSZ running.

The seventh and eighth FASTA files, named mature.fa and hairpin.fa in miRBase database (download from: http://www.mirbase.org/ftp.shtml), contain sequences of all mature miRNA and miRNA hairpins deposited in miRBase, respectively. Sequences of all mature miRNA and miRNA hairpins of specified species are extracted automatically from these FASTA files when CSZ running.

The ninth FASTA file containing piRNA sequences of specified species is retrieved from piRNABank database (http://pirnabank.ibab.ac.in/). This is

optional because piRNA sequences of only seven species were recorded in piRNABank database. For convenience, please name the retrieved file as XXX.piRNA.fa.

The tenth GFF3 file contains genome coordinate information of all mature miRNA and miRNA hairpins deposited in miRBase database, which is available from http://www.mirbase.org/ftp.shtml. For a mature sequence that map equally well to more than one positions on a precursor sequence, then genome coordinate information will be used for determining precisely mapping position of this miRNA on its precursor. But in reality this happens rarely. Therefore, this file is optional.

(2) Running CSZ: Before proceeding, make sure that above-mentioned required files (except files including sequencing data) must be deposited in the same directory. Furthermore, all above-mentioned required files must be accessed by CSZ.

Usage:

perl csz.pl [options] -I <data path> -T <Mireap> and/or -T <miRDeep2>
-V <SVM model > -H <scaled model > -F <sample_1.fa> -F <sample_2.fa> ...

Options:

--species/-S <species> : the three letter acronym of specified species in miRBase database. More details about it, please read organisms file in current CSZ distribution.

--samples/-F <fasta> : fasta files containing the deep sequencing data in the specified format

--data_path/-I <path> : the path to directory containing all required data.

--tools/-T <Mireap and/or miRDeep2> : the names of tools using to identify novel miRNAs in CSZ

--mode/-M [union/intersection] : the way to combine predicted results of MIREAP and miRDeep2. This is optional and the default value is union.

--prepro/-r [1/0] : preprocess reads before mapping it against reference genome if this function is activated (default: 1) where 1 stand for activation

--known/-K [1/0] : quit immediately after detecting all known miRNAs if this function is activated (default: 0)

--probT/-P [float] : the probability threshold. If ZmirP score is less than this

probability threshold, then this miRNA is filtered directly out from downstream analysis (default: 0.5).

--minLen/-l [int] : the minimum of read length. If the length of reads is less than the lower bound of read length, then this reads will be directly filtered out (default: 18 nt).

--maxLen/-L [int] : the maximum of read length. If the length of reads is greater than the upper limit of read length, this reads will be directly filtered out (default: 35 nt).

--minFrq/-q [int] : the observed frequency of read. If the occurrence of read is less than this lower bound of observed frequency of read, then it will be removed out (default: 3).

--Bowtie_path/-B [path] : the path to Bowtie (default: ./Packages/mirdeep2/essentials /bowtie-0.12.5)

--Mireap_path/-E [path] : the path to Mireap (default: ./Packages/mireap_0.2)

--miRDeep2_path/-D [path] : the path to miRDeep2 (default: ./Packages/mirdeep2)

--LIBSVM_path/-C [path] : the path to LIBSVM (default: ./Packages/libsvm-3.16)

--SVMmodel/-M <SVM model> : the SVM model file containing parameters and support vectors

--scaleModel/-H <scaled model> : the scaled model file containing parameters for scaling input data for LIBSVM

--genome_fasta [fasta] : FASTA file containing reference genome sequences downloaded from UCSC database

--mature_fasta [fasta] : FASTA file containing sequences of all mature miRNA sequences deposited in miRBase database

--hairpin_fasta [fasta] : FASTA file containing sequences of all miRNA hairpins deposited in miRBase database

--Rfam_fasta [fasta] : FASTA file containing all kinds of non-coding RNA deposited in Rfam database

--repeat_fasta [fasta] : FASTA file containing repeat sequences extracted from reference genome

--mRNA_fasta [fasta] : FASTA file containing sequences of RefSeq mRNA

downloaded from UCSC database

--piRNA_fasta [fasta] : FASTA file containing sequences of piRNAs extracted from piRNABank database

--dir/-d [dir] : the result directory. This is the optional.

--help/-h : print help information and quit

Optional PARAMETERS for Bowtie:

--threads/-p [int] : launch <int> parallel search threads (default: 1). For more details about it, please read manual of Bowtie

--mismatches/-v [int] : report alignments with at most <int> mismatches (default: 1) For more details about it, please read manual of Bowtie

--genome_cmd [cmd] : the command of Bowtie for reads mapping against reference genome, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches") where threads and mismatches stand for respective the number of threads and mismatches

--miRBase_cmd [cmd] : the command of Bowtie for reads mapping against miRBase, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches -a --best --strata --norc")

--Rfam_cmd [cmd] : the command of Bowtie for reads mapping against ncRNAs deposited in Rfam, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches -a --best --strata --norc")

--repeat_cmd [cmd] : the command of Bowtie for reads mapping against repeat sequences, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches")

--mRNA_cmd [cmd] : the command of Bowtie for reads mapping against mRNAs, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches -a --best --strata --norc")

--piRNA_cmd [cmd] : the command of Bowtie for reads mapping against piRNAs, enclosed in "" if more than one word (default: "bowtie -p threads -t -f -v mismatches -a --best --strata --norc")

References

1. Zhang, Y., Xu, B., Yang, Y., Ban, R., Zhang, H., Jiang, X., Cooke, H.J., Xue, Y. and Shi, Q. (2012) CPSS: a computational platform for the analysis of small RNA deep sequencing data. *Bioinformatics*, **28**, 1925-1927.
2. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. and Schilling, T.F. (1995) Stages of embryonic development of the zebrafish. *Developmental dynamics : an official publication of the American Association of Anatomists*, **203**, 253-310.
3. Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research*, **18**, 610-621.
4. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
5. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, **39**, D152-157.
6. Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P. and Bateman, A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, **41**, D226-232.
7. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic acids research*, **39**, D876-882.
8. Sai Lakshmi, S. and Agrawal, S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research*, **36**, D173-177.
9. Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. and Rajewsky, N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, **40**, 37-52.