|  |
| --- |
|  |
| **20120316\_miRDeep** |

This Hands-on was adapted from miRDeep2 Document

Login  163.25.92.42 with your student\_id/passwd

pjhuang@NGS-course:~$ **mkdir 0316**

pjhuang@NGS-course:~$ **cd 0316**

pjhuang@NGS-course:~/0316$**cp -r /opt/ngstools/mirdeep2/tutorial\_dir/ .**

To run the tutorial please go to the tutorial subfolder.

pjhuang@NGS-course:~/0316$ **cd tutorial\_dir/**

Introduction:

The user wishes to analyze deep sequencing data mapping to a ~6 kb region on C. elegans chromosome II for known and novel miRNA genes.

---------------------------------------------------------------------------------------------------------------------------------------------------

**Preliminary files:**

**cel\_cluster.fa**:                               a fasta file with the reference genome (this file is in fact a ~6 kb region of the C. elegans chromosome II).

**mature\_ref\_this\_species.fa**:          a fasta file with the reference miRBase mature miRNAs for the species (C. elegans miRBase v.14 mature miRNAs)

**mature\_ref\_other\_species.fa**:        a fasta file with the reference miRBase mature miRNAs for related species (C. briggsae and D. melanogaster miRBase v.14 mature miRNAs)

**precursors\_ref\_this\_species.fa**:    a fasta file with the reference miRBase precursor miRNAs for the species (C. elegans miRBase v.14 precursor miRNAs)

**reads.fa**:                                       a fasta file with the deep sequencing reads.

--------------------------------------------------------------------------------------------------------------------------------------------------

**Analysis:**

**Step 1:**

**build an index of the genome (in this case the ~6 kb region):**

pjhuang@NGS-course:~/0316/tutorial\_dir$ **ls -l**

total 19216

-rw-rw-r-- 1 pjhuang pjhuang     6248 2012-03-15 10:54 cel\_cluster.fa

-rw-rw-r-- 1 pjhuang pjhuang     8864 2012-03-15 10:54 mature\_ref\_other\_species.fa

-rw-rw-r-- 1 pjhuang pjhuang     6384 2012-03-15 10:54 mature\_ref\_this\_species.fa

-rw-rw-r-- 1 pjhuang pjhuang      647 2012-03-15 10:54 precursors\_ref\_this\_species.fa

-rw-rw---- 1 pjhuang pjhuang    31707 2012-03-15 10:54 README

-rw-rw-r-- 1 pjhuang pjhuang 19570072 2012-03-15 10:54 reads.fa

-rw-rw-r-- 1 pjhuang pjhuang    32964 2012-03-15 10:54 sample\_result.html

-rw-rw-r-- 1 pjhuang pjhuang     3412 2012-03-15 10:54 TUTORIAL

****pjhuang@NGS-course:~/0316/tutorial\_dir$ **bowtie-build  cel\_cluster.fa  cel\_cluster**

pjhuang@NGS-course:~/0316/tutorial\_dir$ **ls -l**

total 27432

-rw-rw-r-- 1 pjhuang pjhuang  4196281 2012-03-15 11:03 cel\_cluster.1.ebwt

-rw-rw-r-- 1 pjhuang pjhuang      768 2012-03-15 11:03 cel\_cluster.2.ebwt

-rw-rw-r-- 1 pjhuang pjhuang       17 2012-03-15 11:03 cel\_cluster.3.ebwt

-rw-rw-r-- 1 pjhuang pjhuang     1525 2012-03-15 11:03 cel\_cluster.4.ebwt

-rw-rw-r-- 1 pjhuang pjhuang     6248 2012-03-15 10:54 cel\_cluster.fa

-rw-rw-r-- 1 pjhuang pjhuang  4196281 2012-03-15 11:03 cel\_cluster.rev.1.ebwt

-rw-rw-r-- 1 pjhuang pjhuang      768 2012-03-15 11:03 cel\_cluster.rev.2.ebwt

-rw-rw-r-- 1 pjhuang pjhuang     8864 2012-03-15 10:54 mature\_ref\_other\_species.fa

-rw-rw-r-- 1 pjhuang pjhuang     6384 2012-03-15 10:54 mature\_ref\_this\_species.fa

-rw-rw-r-- 1 pjhuang pjhuang      647 2012-03-15 10:54 precursors\_ref\_this\_species.fa

-rw-rw---- 1 pjhuang pjhuang    31707 2012-03-15 10:54 README

-rw-rw-r-- 1 pjhuang pjhuang 19570072 2012-03-15 10:54 reads.fa

-rw-rw-r-- 1 pjhuang pjhuang    32964 2012-03-15 10:54 sample\_result.html

-rw-rw-r-- 1 pjhuang pjhuang     3412 2012-03-15 10:54 TUTORIAL

**Step 2:**

**process reads and map them to the genome.**

pjhuang@NGS-course:~/0316/tutorial\_dir$ [**mapper.pl**](http://mapper.pl)

/opt/ngstools/mirdeep2/mapper.pl input\_file\_reads

This script takes as input a file with deep sequencing reads (these can be in

different formats, see the options below). The script then processes the reads

and/or maps them to the reference genome, as designated by the options given.

Options:

Read input file:

-a              input file is seq.txt format

-b              input file is qseq.txt format

**-c              input file is fasta format**

**-e              input file is fastq format**

-d              input file is a config file (see miRDeep2 documentation).

                options -a, -b or -c must be given with option -d.

Preprocessing/mapping:

-g              three-letter prefix for reads (by default 'seq')

-h              parse to fasta format

-i              convert rna to dna alphabet (to map against genome)

-j              remove all entries that have a sequence that contains letters

                other than a,c,g,t,u,n,A,C,G,T,U,N

-k seq          clip 3' adapter sequence

**-l int          discard reads shorter than int nts**

**-m              collapse reads**

**-p genome       map to genome (must be indexed by bowtie-build).** The 'genome'

                string must be the prefix of the bowtie index. For instance, if

                the first indexed file is called 'h\_sapiens\_37\_asm.1.ebwt' then

                the prefix is 'h\_sapiens\_37\_asm'.

-q              map with one mismatch in the seed (mapping takes longer)

-r int          a read is allowed to map up to this number of positions in the genome

                default is 5

Output files:

**-s file         print processed reads to this file**

-t file         print read mappings to this file

Other:

-u              do not remove directory with temporary files

**-v              outputs progress report**

-n              overwrite existing files

**-o              number of threads to use for bowtie**

Example of use:

/opt/ngstools/mirdeep2/mapper.pl reads\_seq.txt -a -h -i -j -k TCGTATGCCGTCTTCTGCTTGT  -l 18 -m -p h\_sapiens\_37\_asm -s reads.fa -t reads\_vs\_genome.arf -v

The **-c** option designates that the input file is a fasta file (for other input formats, see the README file). The **-j** options removes entries with

non-canonical letters (letters other than a,c,g,t,u,n,A,C,G,T,U,N). The **-k** option clips adapters. The **-l** option discards reads shorter than 18 nts.

The **-m** option collapses the reads. The **-p** option maps the processed reads against the previously indexed genome (cel\_cluster). The **-s** option

designates the name of the output file of processed reads and the **-t** option designates the name of the output file of the genome mappings. Last,

**-v** gives verbose output to the screen.

[****](http://mapper.pl)pjhuang@NGS-course:~/0316/tutorial\_dir$ mapper.pl reads.fa -c -j -k TCGTATGCCGTCTTCTGCTTGT  -l 18 -m -p cel\_cluster -s reads\_collapsed.fa -t reads\_collapsed\_vs\_genome.arf -v

discarding sequences with non-canonical letters

clipping 3' adapters

discarding short reads

collapsing reads

mapping reads to genome index

# reads processed: 1609

# reads with at least one reported alignment: 470 (29.21%)

# reads that failed to align: 1139 (70.79%)

Reported 480 alignments to 1 output stream(s)

trimming unmapped nts in the 3' ends

**Step 3:**

**fast quantitation of reads mapping to known miRBase precursors.**

(This step is not required for identification of known and novel miRNAs in the deep sequencing data when using [**miRDeep2.pl**](http://miRDeep2.pl).)

pjhuang@NGS-course:~/0316/tutorial\_dir$ [**quantifier.pl**](http://quantifier.pl)

usage:

       perl [**quantifier.pl**](http://quantifier.pl) [options] -p precursor.fa -m mature.fa -r reads.fa -s star.fa -t species -y [timestamp] -d [pdfs] -o [sort] -k [stringent] -c config.txt -g [number of mismatches in reads vs precursor mappings]

[options]

[mandatory parameters]

       -u     list all values allowed for the species parameter that have an entry at UCSC

       **-p precursor.fa      miRNA precursor sequences from miRBase**

       **-m mature.fa      miRNA sequences from miRBase**

       **-r reads.fa      your read sequences**

[optional parameters]

       -c [file]    config.txt file with different sample ids... or just the one sample id

       -s [star.fa] optional star sequences from miRBase

       **-t [species] e.g. Mouse or mmu**

                    if not searching in a specific species all species in your files will be analyzed

                    else only the species in your dataset is considered

     **-y [time]    optional otherwise its generating a new one**

       -d           if parameter given pdfs will not be generated, otherwise pdfs will be generated

       -o           if parameter is given reads were not sorted by sample in pdf file, default is sorting

       -k           also considers precursor-mature mappings that have different ids, eg let7c

                    would be allowed to map to pre-let7a

       -n           do not do file conversion again

       -x           do not do mapping against precursor again

       -g [int]     number of allowed mismatches when mapping reads to precursors, default 1

       -e [int]     number of nucleotides upstream of the mature sequence to consider, default 2

       -f [int]     number of nucleotides downstream of the mature sequence to consider, default 5

       -j           do not create an output.mrd file and pdfs if specified

       -w           considers the whole precursor as the 'mature sequence'

[****](http://quantifier.pl)pjhuang@NGS-course:~/0316/tutorial\_dir$ [**quantifier.pl**](http://quantifier.pl) -p precursors\_ref\_this\_species.fa -m mature\_ref\_this\_species.fa -r reads\_collapsed.fa -t cel -y 16\_19

getting samples and corresponding read numbers

seq     374333 reads

Converting input files

building bowtie index

mapping mature sequences against index

# reads processed: 174

# reads with at least one reported alignment: 6 (3.45%)

# reads that failed to align: 168 (96.55%)

Reported 6 alignments to 1 output stream(s)

mapping read sequences against index

# reads processed: 1505

# reads with at least one reported alignment: 1088 (72.29%)

# reads that failed to align: 417 (27.71%)

Reported 1099 alignments to 1 output stream(s)

analyzing data

6 mature mappings to precursors

Expressed miRNAs are written to expression\_analyses/expression\_analyses\_16\_19/miRNA\_expressed.csv

not expressed miRNAs are written to expression\_analyses/expression\_analyses\_16\_19/miRNA\_not\_expressed.csv

Creating miRBase.mrd file

after READS READ IN thing

[**make\_html2.pl**](http://make_html2.pl) -q expression\_analyses/expression\_analyses\_16\_19/miRBase.mrd -k mature\_ref\_this\_species.fa -z -t C.elegans -y 16\_19  -o -i expression\_analyses/expression\_analyses\_16\_19/mature\_ref\_this\_species\_mapped.arf  -l -m cel miRNAs\_expressed\_all\_samples\_16\_19.csv

miRNAs\_expressed\_all\_samples\_16\_19.csv file with miRNA expression values

parsing miRBase.mrd file finished

creating PDF files

creating pdf for cel-mir-39 finished

creating pdf for cel-mir-40 finished

creating pdf for cel-mir-37 finished

creating pdf for cel-mir-36 finished

creating pdf for cel-mir-38 finished

creating pdf for cel-mir-41 finished

The **miRNA\_expressed.csv** gives the read counts of the reference miRNAs in the data in tabular format. The results can also be browsed by opening

**expression\_16\_19.html** with an internet browser.

**Step 4:**

**identification of known and novel miRNAs in the deep sequencing data:**

pjhuang@NGS-course:~/0316/tutorial\_dir$ [**miRDeep2.pl**](http://miRDeep2.pl) reads\_collapsed.fa cel\_cluster.fa reads\_collapsed\_vs\_genome.arf mature\_ref\_this\_species.fa mature\_ref\_other\_species.fa precursors\_ref\_this\_species.fa -t C.elegans 2> report.log

#####################################

#                                   #

# miRDeep2                          #

#                                   #

# last change: 11/01/2011           #

#                                   #

#####################################

miRDeep2 started at 11:50:46

#Starting miRDeep2

#testing input files

#Quantitation of known miRNAs in data

#parsing genome mappings

#excising precursors

#preparing signature

#folding precursors

#computing randfold p-values

#running miRDeep core algorithm

#running permuted controls

#doing survey of accuracy

#producing graphic results

miRDeep runtime:

started: 11:50:46

ended: 11:51:40

total:0h:0m:54s

Step 5:

browse the results.

open the results.html using an internet browser. Notice that cel-miR-37 is predicted twice, since both potential precursors excised from this locus

can fold into hairpins. However, the annotated hairpin scores much higher than the non-annotated one (miRDeep2 score 6.1e+4 vs. -0.2)