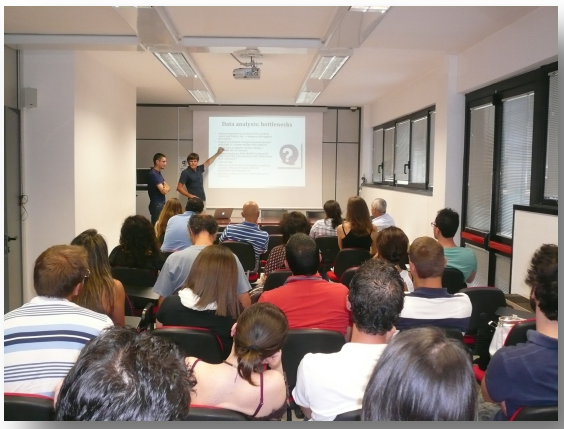
AllBio Tutorial

NGS data analysis for non-coding RNAs and small RNAs



Aim of the Tutorial

Non-coding RNA (ncRNA) are functional RNA molecule that are not translated into a protein. ncRNA genes include highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, and piRNAs and the long ncRNAs. The number of ncRNAs encoded within the genomes is unknown, however recent transcriptomic and bioinformatic studies suggest the existence of thousands of ncRNAs. Especially using the new high through-put sequencing technology (NGS) opend a door to study these ncRNA Since many of the newly identified ncRNAs have not been validated for their function, it is possible that many are non-functional.

During this Tutorial we will point out the difficulties during the sequence data analysis in finding know and unknown ncRNA. We will guide the participants through a workflow (Linux command line) developed at the Institute for Biomedical Technologies (ITB- CNR, Bari, Italy) using some modified test data.

The workflow will list at the end known and unknown miRNA and if we have a time-course of sampled data the fold changes in a graphical interface which will allow a sophisticated interrogation to extract the required information.

In an extension, the workflow will also find other ncRNA, than miRNA however only based on the list of known ncRNA. This ncRNA repository is a product if ITB which will soon be publicly available.

Tutorial

Teachers

- Andreas Gisel ITB-CNR

- Angelica Tulipano ITB-CNR
- Arianna Consiglio ITB-CNR
- Flavio Licciulli ITB-CNR

Technical Support

- Giacinto Donvito INFN - Bari
- Nicola Losito ITB-CNR
- Jarno Laitinen CSC

Data

- The data set is an Illumina sequencing of small RNA of mouse (Mus musculus)
- Small RNA analysis of wildtype Mouse embryo and Adar1 null mouse embryo at E11.0 and E11.5
- SRR361337 - Small RNAs from Mouse Embyro Day11
- SRR361338 - Small RNAs from ADAR1 KO Mouse Embyro Day11
- SRR361340 - Small RNAs from ADAR1 KO Mouse Embyro Day11.5
- http://www.ebi.ac.uk/ena/data/view/PRJNA148757

Workflow

- The workflow starts with the input of the raw data you normally get from a sequencing center and at the end will give you a list of known and unknown miRNAs.
- If you have a series of experiments you will have access to a graphical interface where you can filter the results by different parameters to find the miRNA and related target genes of your interest.

```
@SRR361337.4 unknown:2:1:6:870 length=36
GGGGAATCTGACTGTCTAANTCGTATGCCGTCTTCT
+SRR361337.4 unknown:2:1:6:870 length=36
BBCB@?@<BA;=BB>6=;3%;5434;4/.021?</=
@SRR361337.8 unknown:2:1:6:936 length=36
AGTTCTACAGTCCGACGATCTCGTATGCCGTCTTCT
+SRR361337.8 unknown:2:1:6:936 length=36
ACBB:AB@?2<>7>>553>1;3769;7#########
@SRR361337.12 unknown:2:1:6:653 length=36
TGGAGTGTGACAATGGTGTTTGTCGTATGCCGTCTT
+SRR361337.12 unknown:2:1:6:653 length=36
BCCBC?BA@A@BBBBA98;B?)>28@9@5/-);4@C
@SRR361337.16 unknown:2:1:6:238 length=36
ATACTGCATCAGGAACTGACTGGATCGTATGCCGTT
+SRR361337.16 unknown:2:1:6:238 length=36
BBA7A@:@A?>;>AA8<4:6<695>4##########
@SRR361337.20 unknown:2:1:6:1221 length=36
TATTGCACTTGTCCCGGCCTGTTCGTATGCCGTCTT
+SRR361337.20 unknown:2:1:6:1221 length=36
BBBAA<@<BB<A>>;6735?9<:.+:6@3&,)5*=A
```

Refresh | Quick search

Actions	Sequence	miRNA 1 ID	Fold Change Image	CTRL-T1, T1-T2 diff_exp	CTRL-T1 Fold_Change	T1-T2 Fold_Change	CTRL-T2 Fold Change	CTRL-T1 pValue	T1-T2 pValue	RPM CTRL	RPM T1	RPM T2	miRNA 2 ID	miRNA 2 Species
	GTGGGAAGGAACTACAAGACAGCT	-		over,over	4.870	2.376	7.246	1.63251e-42	0	3.73	109.06	566.14	mmu-miR-6368	Mus musculus
	GCGGGGCTGGGCGCGCGC	-		over,over	1.580	4.460	6.040	0.130443	1.21607e-21	0.00	2.99	65.78	hsa-miR-4508	Homo sapiens
	GTGAGGACTGGGGAGGTGGA	mmu-miR-1224-5p		over,over	1.463	0.104	1.567	0.0179358	0.503983	3.73	10.28	11.05	hsa-miR-1224-5p	Homo sapiens
	CGTACCGTGAGTAATAATGCG	mmu-miR-126a-5p		over,over	0.042	1.059	1.101	0.533006	0.0115424	11.91	12.27	25.57	bta-miR-126-3p	Bos taurus
	AGGAGAAAGGCCTCTCTCTC	mmu-miR-124-5p		over,under	2.612	-1.669	0.943	0.000588264	0.030519	1.68	10.28	3.23	-	
	TGGAAGACTAGTGATTTTGTTGT	mmu-miR-7a-5p		over,over	0.489	0.400	0.889	0.00335009	0.0120397	75.36	105.74	139.49	hsa-miR-7-5p	Homo sapiens
	TCCTGTACTGAGCTGCCCCGA	mmu-miR-486-5p		over,under	1.576	-0.756	0.820	2.09934e-29	0.0000000600358	82.18	244.96	145.08	hsa-miR-486-5p	Homo sapiens
	TGAGGGGCAGAGAGCGAGACT	mmu-miR-423-5p		under,over	-0.618	0.985	0.367	0.0000161822	0.000000000209085	159.26	103.75	205.39	hsa-miR-423-5p	Homo sapiens
	AAAAGCTGGGTTGAGAGGGCG	mmu-miR-320-5p		over,under	0.624	-0.567	0.057	0.000000000520234	0.000000716915	217.93	335.77	226.61	hsa-miR-320a	Homo sapiens

Steps of the workflow:

1. create work directory and load data
2. process the raw data for the mapping and miRNA search
3. calculate the fold change between the different biological sample
4. run the miRNA search using mirDeep2 (Friedländer et al, 2008)
5. run a mapping with BWA (Li and Durbin, 2009) on the mouse genome
6. combine all data
7. extract the data for the database
8. load the database (will be done for this tutorial)
9. access the graphical interface of the database

Detailed steps to follow the tutorial:

1. move to the working device
   - `cd studentX (X is your number)`

1. we create a working directory in our working ‘home’
   - `mkdir work`

1. enter into the working directory
   - `cd work`

1. copy the data from centralized directory (/Tutorial) to the working directory
   - `cp /Tutorial/SRR3613* .`

1. copy some software from centralized directory (/Tutorial) to the working directory
   - `cp /Tutorial/fold.tgz .`

1. unzip the the software
   - `tar xzvf fold.tgz`

1. assign the path of the software
   - `path=/Tutorial/software`

1. run the first step of the raw data processing - it will search for the adaptor sequence and remove it. If the sequence doesn’t have an adaptor the sequence will be removed.
   - `perl $path/step1_clean.pl SRR361337red4.fastq TCGTATGCCGTCTTCTGCT S1_-_1_C &`
   - `perl $path/step1_clean.pl SRR361338red4.fastq TCGTATGCCGTCTTCTGCT S2_-_1_C &`
   - `perl $path/step1_clean.pl SRR361340red4.fastq TCGTATGCCGTCTTCTGCT S3_-_1_C &`

1. create for each sample a directory for the second step of the raw data processing and move the output from step1
   - `mkdir S1`
   - `mv S1_-_1_C.fastq S1`
   - `mkdir S2`
   - `mv S2_-_1_C.fastq S2`
   - `mkdir S3`
   - `mv S3_-_1_C.fastq S3`

1. run the second step of the raw data processing - this step splits the sequences by sequence size and creates for each length a fastq and a fasta file. The last two arguments gives the possibility to merge a range of sequence sizes for further analysis. In addition step2 also create some statistics files for the downstream processing.
   - `perl $path/step2_last_12.pl $PWD/S1 $PWD/S1 18 26 &`
   - `perl $path/step2_last_12.pl $PWD/S2 $PWD/S2 18 26 &`
   - `perl $path/step2_last_12.pl $PWD/S3 $PWD/S3 18 26 &`

1. for the fold-change calculation create for each comparison a new directory and copy the necessary statistics data from step1 - important is that one data file is labeled as control (C) and the other as experiment (E) - we rename the second file as experiment
   - `mkdir S1-S2fold`
   - `cp S1/S1_-_1_C_all18-26.txt S1-S2fold`
   - `cp S2/S2_-_1_C_all18-26.txt S1-S2fold`
   - `mv S1-S2fold/S2_-_1_C_all18-26.txt S1-S2fold/S2_-_1_E_all18-26.txt`
   - `mkdir S1-S3fold`
   - `cp S1/S1_-_1_C_all18-26.txt S1-S3fold`
   - `cp S3/S3_-_1_C_all18-26.txt S1-S3fold`

- `mv S1-S3fold/S3_-_1_C_all18-26.txt S1-S3fold/S3_-_1_E_all18-26.txt`
- `mkdir S2-S3fold`
- `cp S2/S2_-_1_C_all18-26.txt S2-S3fold`
- `cp S3/S3_-_1_C_all18-26.txt S2-S3fold`
- `mv S2-S3fold/S3_-_1_C_all18-26.txt S2-S3fold/S3_-_1_E_all18-26.txt`

1. set the new path as the home path
   - `path2=/work/studentX/work`

1. enter into the software directory and run the fold-change calculation
   - `cd fold`
   - `perl scatter_last1.pl $path2/S1-S2fold $path2/S1-S2fold`
   - `perl scatter_last1.pl $path2/S1-S3fold $path2/S1-S3fold`
   - `perl scatter_last1.pl $path2/S2-S3fold $path2/S2-S3fold`

1. return to your working directory
   - `cd ..`

1. for the miRNA discovery we use mirDeep2 - mirDeep2 is divided into two steps: a) the mapping and b) the miRNA discovery.
2. we need a local miDeep2 installation
   - `cp /Tutorial/mirdeep2_0_0_5.zip .`
   - `unzip mirdeep2_0_0_5.zip`
   - `cd mirdeep2`
   - `./install.pl`
   - `logout and log in again`
   - `cd studentX/work`

1. for the mapping step (using bowtie) we need to copy the corresponding fasta file into the mirdeep directory
   - `cp S1/S1_-_1_C_all18-26.fasta mirdeep2`
   - `cp S2/S2_-_1_C_all18-26.fasta mirdeep2`
   - `cp S3/S3_-_1_C_all18-26.fasta mirdeep2`

1. set the genome path to get access to the bowtie index of the mouse genome
   - `genome=/Tutorial/genome`

1. enter into the mirdeep2 directory and run the mapping algorithm
   - `cd mirdeep2`
   - `perl mapper.pl S1_-_1_C_all18-26.fasta -c -m -p $genome/mouse -s S1_collapsed.fa -t S1_collapsed_vs_genome.arf –v &`
   - `perl mapper.pl S2_-_1_C_all18-26.fasta -c -m -p $genome/mouse -s S2_collapsed.fa -t S2_collapsed_vs_genome.arf –v &`
   - `perl mapper.pl S3_-_1_C_all18-26.fasta -c -m -p $genome/mouse -s S3_collapsed.fa -t S3_collapsed_vs_genome.arf –v &`

1. run the miRNA discovery algorithm - here we add also the mirBase reference for mouse and for all mammals - we also rename the csv output for further processing
   - `perl miRDeep2.pl S1_collapsed.fa $genome/Mus_musculus_genomeGRCm38.71.fa S1_collapsed_vs_genome.arf $genome/mature.mmu.fa $genome/mature_19-june_Mammalia_RNA.fa none 2>log1.log`
   - `mv result*.csv mirdeep_S1.csv`
   - `perl miRDeep2.pl S2_collapsed.fa $genome/Mus_musculus_genomeGRCm38.71.fa S2_collapsed_vs_genome.arf $genome/mature.mmu.fa $genome/mature_19-june_Mammalia_RNA.fa none 2>log2.log`
   - `mv result*.csv mirdeep_S2.csv`

- `perl miRDeep2.pl S3_collapsed.fa $genome/ Mus_musculus_genomeGRCm38.71.fa S3_collapsed_vs_genome.arf $genome/mature.mmu.fa $genome/mature_19-june_Mammalia_RNA.fa none 2>log3.log`
- `mv result*.csv mirdeep_S3.csv`

1. we return to the working directory and create a directory for the genome mapping results
   - `cd ..`
   - `mkdir mapping`

1. for the mapping agains the mouse non-coding RNA repository we use BWA - we run the first step - note that we are using the BWA index for this mapping
   - `bwa aln -n 2 -t 3 $genome/ncmouse.fa S1/S1_-_1_C_18-26.fastq > mapping/S1_mapping.sai &`
   - `bwa aln -n 2 -t 3 $genome/ncmouse.fa S2/S2_-_1_C_18-26.fastq > mapping/S2_mapping.sai &`
   - `bwa aln -n 2 -t 3 $genome/ncmouse.fa S3/S3_-_1_C_18-26.fastq > mapping/S3_mapping.sai &`

1. in the second BWA step we create the sam file
   - `bwa samse $genome/ncmouse.fa mapping/S1_mapping.sai S1/S1_-_1_C_18-26.fastq > mapping/S1_-_1_C_18-26_mapping.sam &`
   - `bwa samse $genome/ncmouse.fa mapping/S2_mapping.sai S2/S2_-_1_C_18-26.fastq > mapping/S2_-_1_C_18-26_mapping.sam &`
   - `bwa samse $genome/ncmouse.fa mapping/S3_mapping.sai S3/S3_-_1_C_18-26.fastq > mapping/S3_-_1_C_18-26_mapping.sam &`

1. to combine the created data we need to create files defining the path of the files
   - `cat <<EOF > stat_files.txt`
   - `S1/S1_-_1_C_all18-26.txt`

```
- S2/S2_-_1_C_all18-26.txt
- S3/S3_-_1_C_all18-26.txt
- EOF


- cat <<EOF > foldchange_files.txt
- S1-S2fold/fisher_S2_-+S1_-_overexpression.txt
- S1-S3fold/fisher_S3_-+S1_-_overexpression.txt
- S2-S3fold/fisher_S3_-+S2_-_overexpression.txt
- EOF


- cat <<EOF > mirdeep_files.txt
- mirdeep2/mirdeep_S1.csv
- mirdeep2/mirdeep_S2.csv
- mirdeep2/mirdeep_S3.csv
- EOF


- cat <<EOF > sam_files.txt
- mapping/S1_-_1_C_18-26_mapping.sam
- mapping/S2_-_1_C_18-26_mapping.sam
- mapping/S3_-_1_C_18-26_mapping.sam
- EOF
```

1. run the algorithm to combine all the data using the create location files - we have the first data as csv and con be loaded into an Excel sheet or used for down stream analysis
   - `perl $path/miRNAInteger.pl stat_files.txt mirdeep_files.txt sam_files.txt foldchange_files.txt`

First set of data for down-stream analysis:

- foldchangeFORallSequences.txt
- foldchangeFORsequencespval0.05.txt
- foldchangeFORncmousepval0.05.txt
- foldchangeFORmirDeep.txt
- miRNA-Sample.txt

1. extract the data for the upload into the database - we will have two output files for the database
   - `perl $path/Load_miRNA_Reads.pl miRNA-Sample.txt mirDeep.txt miRNA_Reads.txt`

Second set of data for the graphical interface:

- miRNA_Reads.txt
- mirDeep.txt

THE DATA ARE ALREADY LOADED INTO THE DATABASE AND THE GRAPHICAL INTERFACE READY FOR DATA FILTERING

http://86.50.168.84/MouseTutorial/

http://86.50.168.85/MouseTutorial/

http://86.50.168.86/MouseTutorial/