



MicroRNA Sequence Analysis of Small RNA-Seq by miRDeep2

Ya-Chi Lin, Ph.D.



Outlines

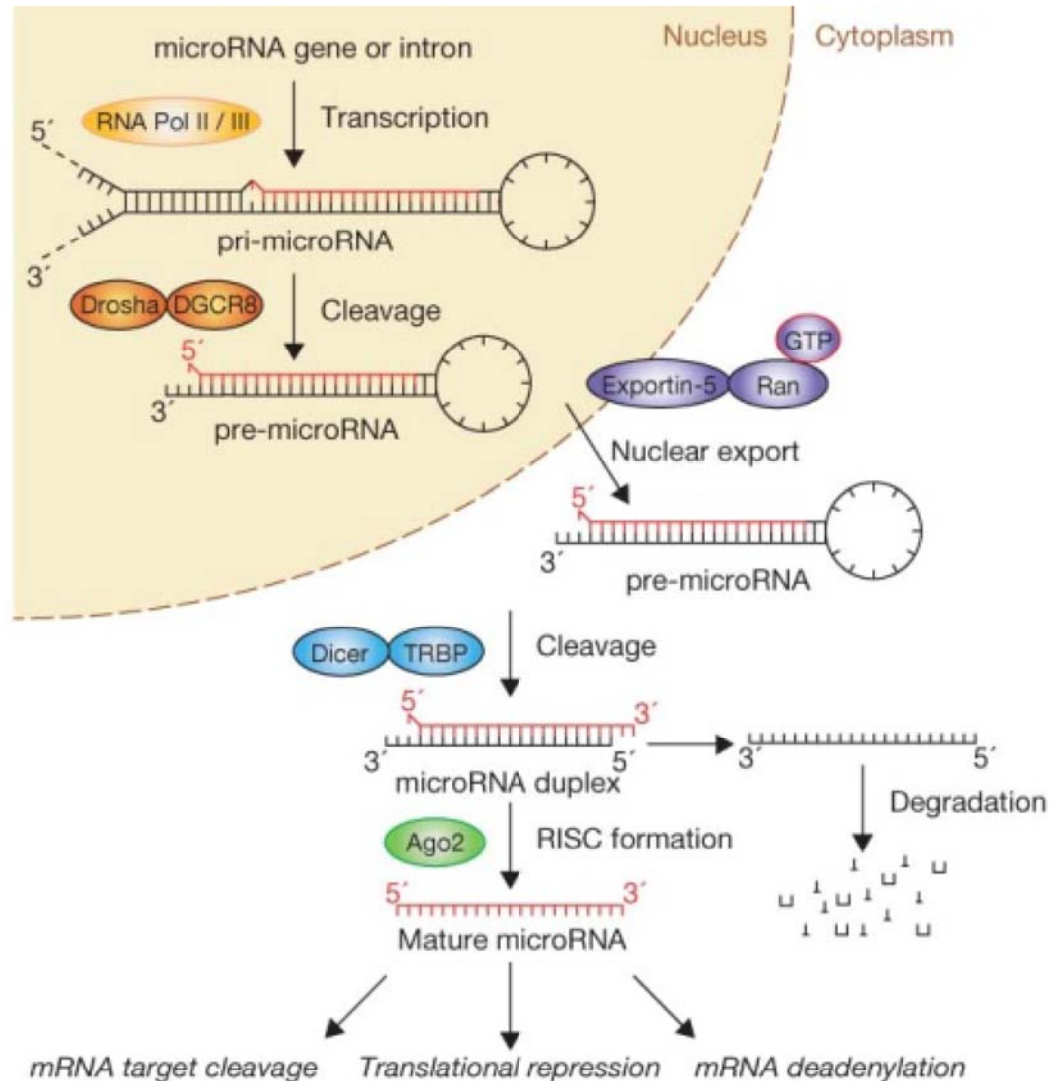
- **Background Introduction**
- **Pre-process Data**
- **Analysis & Results Interpretation**



Outlines

- **Background Introduction**
- Pre-process Data
- Analysis & Results Interpretation

MicroRNA biogenesis



Winter J et al. Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol. 2009 Mar;11(3):228-34.

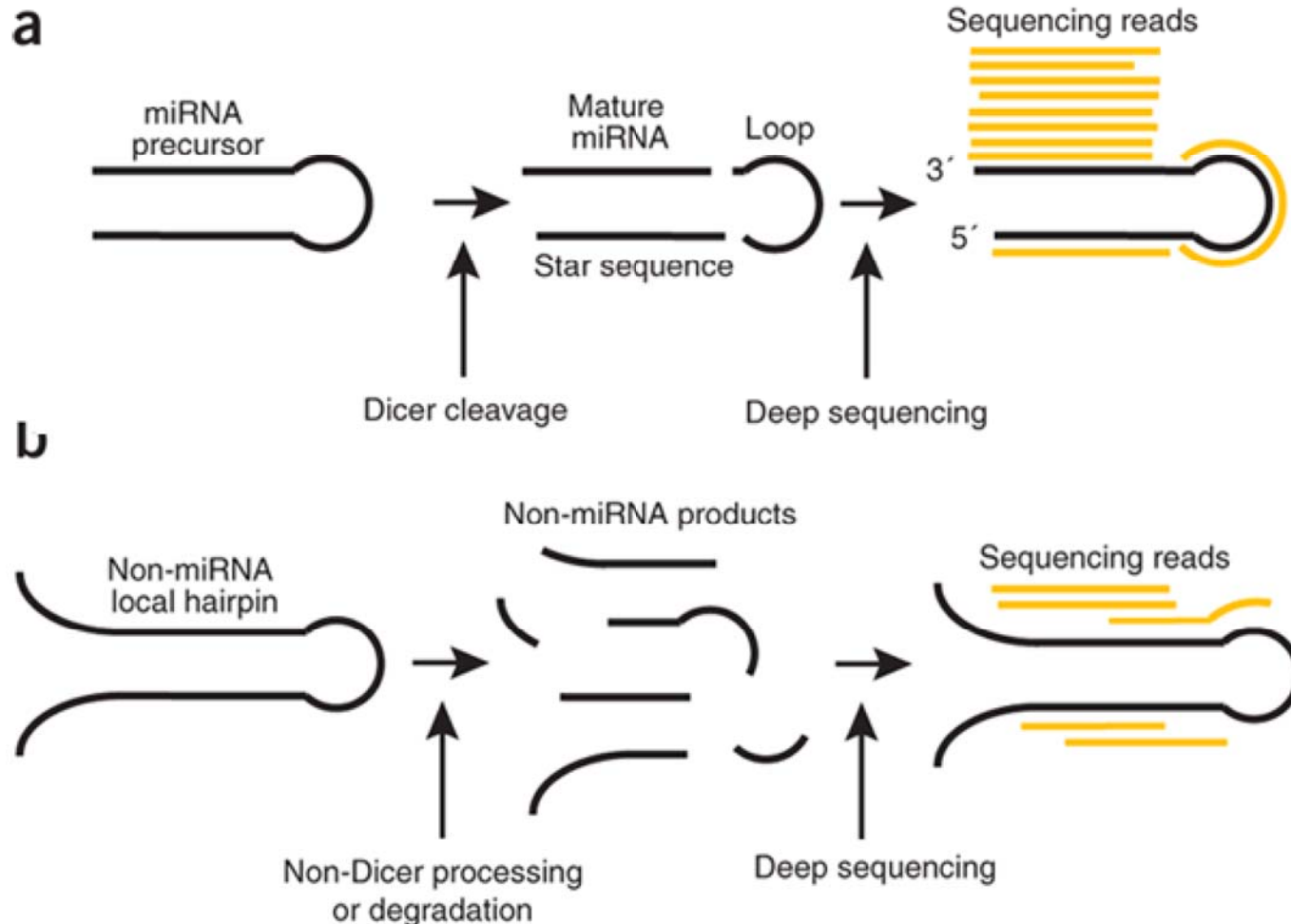


Requirements of microRNA prediction/analysis

- can **accurately** identify known and novel miRNAs in all animal major clades
- can **distinguish** miRNAs from other argonaute bound small RNAs
- reports miRNAs that can stand up to **high throughput** validation
- be **efficient** in memory and time consumption
- be **user-friendly**

Friedländer MR et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012 Jan;40(1):37-52.

Principle of miRDeep



Friedländer MR et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008 Apr;26(4):407-15.



Outlines

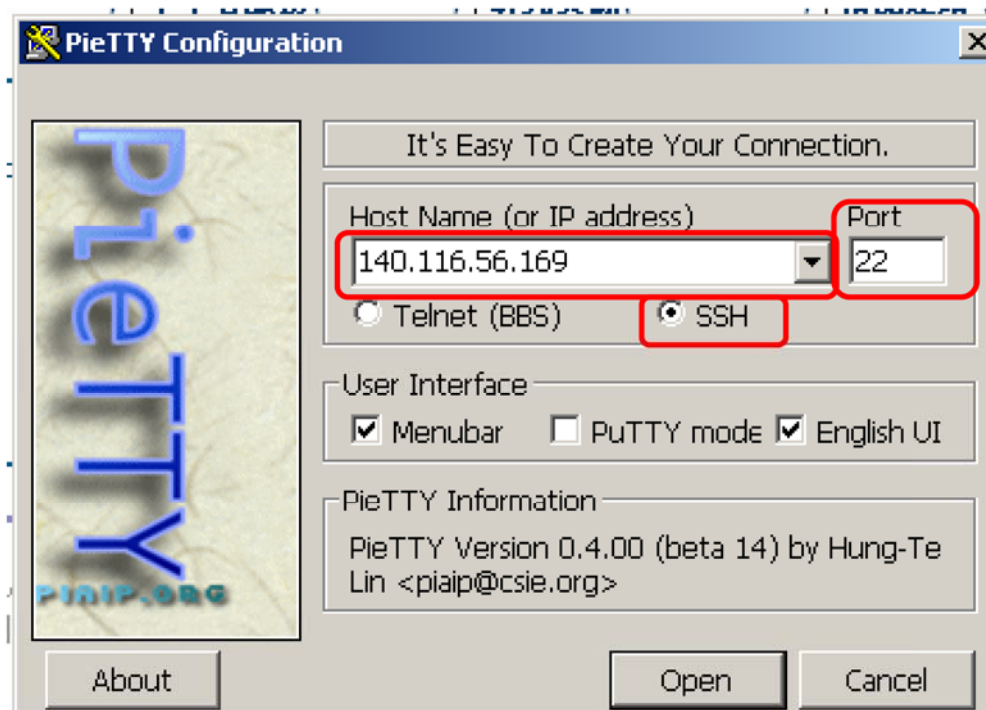
- Background Introduction
- **Pre-process Data**
- Analysis & Results Interpretation



Working Interface

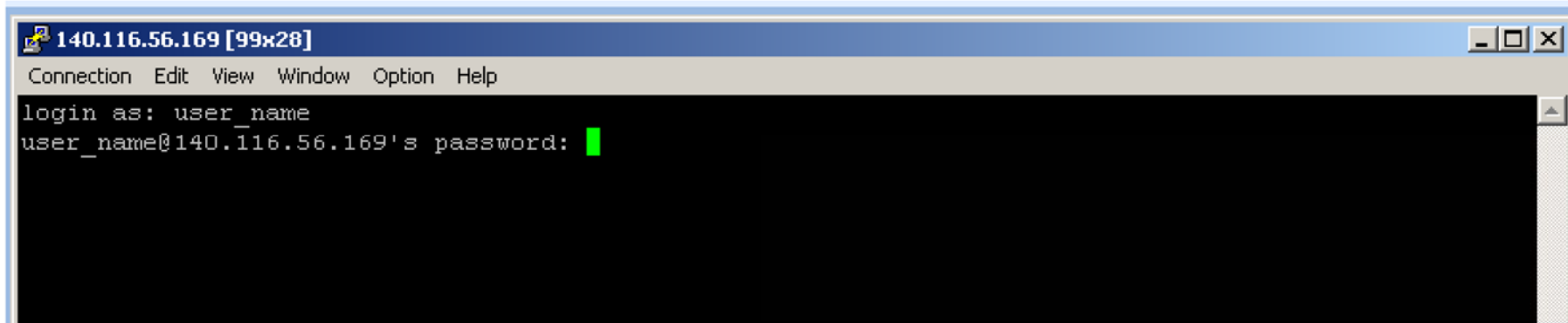
- Examples
 - Windows: PuTTY or PieTTY
 - Mac: Terminal (built-in app)
- Link to download PieTTY
 - <http://www.binfo.ncku.edu.tw/tools.php>

PieTTY [login]



- IP: 140.116.56.169
- Notes:
 - Port: 22
 - Connection type: SSH

PieTTY [Practice with Linux commands]

A screenshot of a PieTTY terminal window. The title bar shows the IP address 140.116.56.169 and the window size 99x28. The menu bar includes Connection, Edit, View, Window, Option, and Help. The terminal content shows a login prompt 'login as: user_name' and a password prompt 'user_name@140.116.56.169's password:' followed by a green cursor. The terminal background is black, and the text is white.

```
140.116.56.169 [99x28]
Connection Edit View Window Option Help
login as: user_name
user_name@140.116.56.169's password: █
```

- ls
- ls -l
- ls --help
- Q1: How many files do you see after ls command?
- Q2: Any difference between ls and ls -l?



PieTTY [Practice with Linux commands]

- less file_name
 - “q” on keyboard the to leave
- head file_name
- tail file_name



Time for practice



miRDeep2 [Files in today's practice]

- reads.fa
- genome.fa
- *.ebwt
- mature.fa
- precursor.fa
- others_mature.fa



miRDeep2 [Workflow]

Genome Index

- bowtie-build **input.fa** **output** &
- Ex: bowtie-build **genome.fa** **genome.bb** &
- Several genome.bb.x.ebwt files will also be output.

Reads Formatting

- collapse_reads.pl **input.fa** (3-letters species abbrev.) > **output.fa** &
- Ex: collapse_reads.pl **reads.fa** dre > **reads_col.fa** &
- After formatting: (3-letters species abbrev.) _uniqueNo_xNum Ex: spe_1_X100

Execute mapper.pl

- mapper.pl **reads_col.fa** -c -p **genome.bb** -t **reads_col_vs_genome.arf** &
- A arf file will be generated after executing.

Execute miRDeep2.pl

- miRDeep2.pl **reads_col.fa** **genome.fa** **reads_col_vs_genome.arf** **mature.fa** **others_mature.fa** **precursor.fa** 2> report.log &
- Result files will be generated.



Reads formatting

- collapse_reads.pl $\triangle$ input.fa (3-letters species abbrev.) $\triangle$ > $\triangle$ output.fa &
- collapse_reads.pl reads.fa dre > reads_col.fa &
- Use find_read_count.pl $\triangle$ reads_col.fa
to calculate how many total reads count in fasta file



Time for practice

請善用**Tab**鍵 ^^



Execute mapper.pl

- `mapper.pl reads_collapsed.fa -c -p genome -t
reads_collapsed_vs_genome.arf &`
- `mapper.pl reads_col.fa -c -p genome.bb -t
reads_col_vs_genome.arf &`

⌘ Note: The genome needs be the output filename
(**genome.bb**) of bowtie-build!



Outlines

- Background Introduction
- Pre-process Data
- **Analysis & Results Interpretation**



Excute miRDeep2.pl

- miRDeep2.pl reads_collapsed.fa genome.fa reads_col_vs_genome.arf study_species_mature.fa others_mature.fa study_species_precursor.fa 2> report.log &
- miRDeep2.pl reads_col.fa genome.fa reads_col_vs_genome.arf mature.fa others_mature.fa precursor.fa 2> report.log &

※Notes

1. new_reads_col.fa

2. genome.fa

3.

<u>mature_study_species.fa</u>	<u>others_mature.fa</u>	<u>precursor_study_species.fa</u>
Zebra fish	Fugu or others	Zebra fish
none	none	none

4. No space between 2>



Time for practice

請善用**Tab**鍵 ^^



Download results

Name	Ext	Size	Type	Changed	Attr
..			Parent directory	2013-08-30 18:06...	a
pdfs_31_08_2013_t_10_56_14			File Folder	2013-08-31 12:01...	a
controls		19,860	File	2012-12-18 14:42...	a
controls2		31,086	File	2012-12-18 17:51...	a
predictions_ori		322,035	File	2012-11-26 10:48...	a
README		17,531	File	2012-11-26 10:48...	a
TUTORIAL		3,412	File	2012-11-26 13:45...	a
miRNAs_expressed_all_samples_31_08...		32,158	Microsoft Office...	2013-08-31 10:56...	a
result_31_08_2013_t_10_56_14.csv		8,920	Microsoft Office...	2013-08-31 11:13...	a
Thumbs.db		5,120	Data Base File	2013-03-25 04:40...	ash
expression_31_08_2013_t_10_56_14....		55,988	Firefox HTML D...	2013-08-31 10:56...	a
result_31_08_2013_t_10_56_14.html		91,884	Firefox HTML D...	2013-08-31 11:13...	a
sample_result.html		33,378	Firefox HTML D...	2012-06-25 19:02...	a
Discovering microRNAs from deep sequ...		503,882	Adobe Acrobat ...	2012-11-13 14:51...	a
mirdeep2_NAR2012.pdf		1,831,985	Adobe Acrobat ...	2013-06-22 16:11...	a
filter_alignments.pl		4,862	PL File	2008-04-03 23:22...	a
~\$mirDeep2.pptx		165	Microsoft Office...	2013-08-31 10:25...	ah
Linux取代字元.pptx		81,639	Microsoft Office...	2013-08-30 15:19...	a
miRDeep.pptx		1,202,036	Microsoft Office...	2012-12-18 16:28...	a
miRDeep2.pptx		1,340,811	Microsoft Office...	2013-08-31 11:32...	a
miRDeep2_09062013.pptx		359,905	Microsoft Office...	2013-08-30 16:14...	a
O_miRDeep2_SOP.pptx		340,495	Microsoft Office...	2013-08-30 11:30...	a
CLC_RNAseq_download.rar		3,718,077	WinRAR archive	2012-12-25 14:30...	a
mirdeepNEW.rar		1,531,266	WinRAR archive	2012-12-18 16:36...	a
mirdeep.TXT		2,046	Text Document	2012-12-14 11:08...	a
demo_full.xls		426,496	Microsoft Office...	2012-12-18 16:29...	a
demo_results.xlsx		160,585	Microsoft Office...	2012-12-18 16:34...	a

Name	Ext	Size	Changed	Rights
..			2013-08-30 17:06:37	rw-xr-xr-x
pdfs_31_08_2013_t_10_56_14			2013-08-31 11:13:59	rw-xr-xr-x
dir_prepare_signature1377917852			2013-08-31 10:57:33	rw-xr-xr-x
expression_analyses			2013-08-31 10:56:36	rw-xr-xr-x
mirdeep_runs			2013-08-31 10:56:14	rw-xr-xr-x
result_31_08_2013_t_10_56_14.bed		0	2013-08-31 11:13:59	rw-r--r--
report.log		11,057	2013-08-31 11:13:59	rw-r--r--
result_31_08_2013_t_10_56_14.html		91,006	2013-08-31 11:13:14	rw-r--r--
result_31_08_2013_t_10_56_14.csv		8,920	2013-08-31 11:13:14	rw-r--r--
error_31_08_2013_t_10_56_14.log		348	2013-08-31 11:13:13	rw-r--r--
expression_31_08_2013_t_10_56_14.html		55,278	2013-08-31 10:56:41	rw-r--r--
miRNAs_expressed_all_samples_31_08_2013_t_10_56_14.csv		32,158	2013-08-31 10:56:36	rw-r--r--
genome.fa		1,376,438,282	2013-08-31 10:54:24	rw-r--r--
reads_col_vs_genome.arf		806,167	2013-08-31 10:52:30	rw-r--r--
mapper.log		1,372	2013-08-31 10:52:30	rw-r--r--
genome.bb.rev.2.ebwt		169,331,596	2013-08-31 10:50:40	rw-r--r--
genome.bb.rev.1.ebwt		391,545,252	2013-08-31 10:50:39	rw-r--r--
genome.bb.4.ebwt		338,663,181	2013-08-31 10:50:38	rw-r--r--
genome.bb.3.ebwt		230,138	2013-08-31 10:50:38	rw-r--r--
genome.bb.2.ebwt		169,331,596	2013-08-31 10:50:38	rw-r--r--
genome.bb.1.ebwt		391,545,252	2013-08-31 10:50:37	rw-r--r--
reads_col.fa		458,752	2013-08-31 10:43:58	rw-r--r--
hsa_others_mature.fa		77,747	2013-08-30 18:07:30	rw-r--r--
reads.fa		2,175,085	2013-08-30 17:32:57	rw-r--r--
others_mature.fa		931,763	2013-08-30 17:11:51	rw-r--r--
precursor.fa		51,609	2013-08-30 17:11:50	rw-r--r--
mature.fa		9,173	2013-08-30 17:11:50	rw-r--r--



miRDeep2 results [1]

- A single html that links to all results
- Performance survey
- Estimates of sensitivity and # of true positive novel microRNAs

Survey of miRDeep2 performance for score cut-offs -10 to 10

	novel miRNAs			known miRBase miRNAs				
miRDeep2 score	predicted by miRDeep2	estimated false positives	estimated true positives	in species	in data	detected by miRDeep2	estimated signal-to-noise	excision gearing
10	45	2 ± 2	43 ± 2 (95 ± 3%)	247	141	92 (65%)	36.3	2
9	46	3 ± 2	43 ± 2 (94 ± 3%)	247	141	92 (65%)	35.3	2
8	54	3 ± 2	51 ± 2 (95 ± 3%)	247	141	92 (65%)	35.3	2
7	59	3 ± 2	56 ± 2 (95 ± 3%)	247	141	92 (65%)	34.2	2
6	63	4 ± 2	59 ± 2 (94 ± 3%)	247	141	92 (65%)	31.1	2
5	135	7 ± 2	128 ± 2 (95 ± 2%)	247	141	108 (77%)	25.5	2
4	201	15 ± 4	186 ± 4 (92 ± 2%)	247	141	115 (82%)	17	2
3	237	46 ± 7	191 ± 7 (81 ± 3%)	247	141	116 (82%)	6.9	2
2	300	78 ± 8	222 ± 8 (74 ± 3%)	247	141	116 (82%)	4.9	2
1	508	126 ± 12	382 ± 12 (75 ± 2%)	247	141	117 (83%)	4.6	2
0	771	244 ± 13	527 ± 13 (68 ± 2%)	247	141	118 (84%)	3.5	2

Friedländer MR et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012 Jan;40(1):37-52.



miRDeep2 results [II]

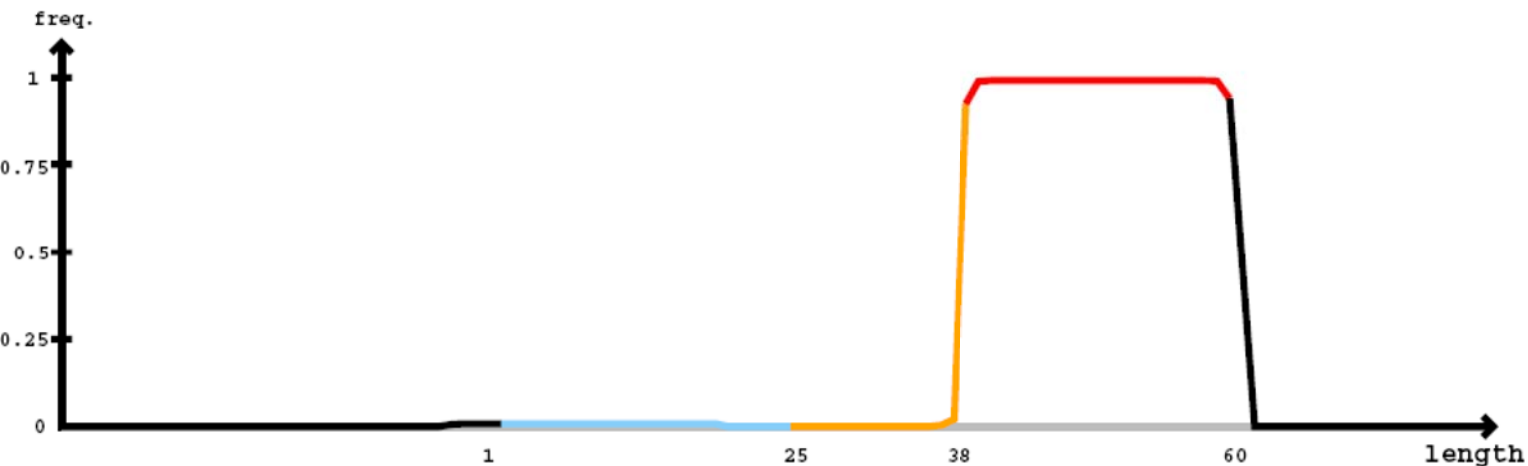
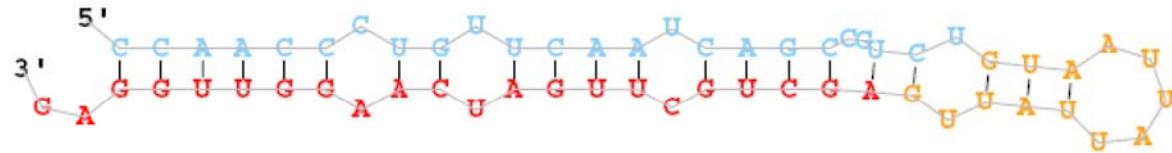
- A single html that links to all results
- Performance survey in today's example

Survey of miRDeep2 performance for score cut-offs -10 to 10

	novel miRNAs			known miRBase miRNAs				
miRDeep2 score	predicted by miRDeep2	estimated false positives	estimated true positives	in species	in data	detected by miRDeep2	estimated signal-to-noise	excision gearing
10	0	0 ± 0	0 ± 0 (0 ± 0%)	247	14	5 (36%)	71.4	1
9	0	0 ± 0	0 ± 0 (0 ± 0%)	247	14	5 (36%)	71.4	1
8	0	0 ± 0	0 ± 0 (0 ± 0%)	247	14	5 (36%)	71.4	1
7	0	0 ± 0	0 ± 0 (0 ± 0%)	247	14	5 (36%)	71.4	1
6	0	0 ± 0	0 ± 0 (0 ± 0%)	247	14	6 (43%)	85.7	1
5	1	0 ± 0	1 ± 0 (97 ± 17%)	247	14	7 (50%)	62.5	1
4	2	0 ± 0	2 ± 0 (94 ± 16%)	247	14	9 (64%)	40	1
3	3	1 ± 1	2 ± 1 (78 ± 26%)	247	14	9 (64%)	14.3	1
2	5	1 ± 1	4 ± 1 (75 ± 22%)	247	14	9 (64%)	9.5	1
1	10	2 ± 1	8 ± 1 (80 ± 13%)	247	14	10 (71%)	8.5	1
0	17	4 ± 2	13 ± 2 (77 ± 12%)	247	14	10 (71%)	6.3	1

miRDeep2 results [IV]

Provisional ID	: chr10_947
Score total	: 5
Score for star read(s)	: -1.3
Score for read counts	: 0
Score for mfe	: 1.7
Score for randfold	: 1.6
Score for cons. seed	: 3
Total read count	: 1480
Mature read count	: 1480
Loop read count	: 0
Star read count	: 0



Star

Mature

5'	reads	mm	sample
agaucuaaccuacauguaccugcagaguucacuuccaaccugguucaauacagccgucuguaauuuuuuuuagagcugcugaucaagguuggagaugaacuuugagaaaggaag	9	0	dre
.....acuuccaaccugguucaauacag.....	3	0	dre
.....cuuccaaccugguucaauacag.....	1	1	dre

[illegible]



Time for practice



Download fasta from miRBase [I]

- <http://www.mirbase.org/ftp.shtml>



miRBase Sequence Download

► [Go to the FTP site](#)

► [Previous releases](#)

► [README](#)

Release notes - read these first!

► [miRNA.dat](#)

all published miRNA data in EMBL format

► [hairpin.fa](#)

Fasta format sequences of all miRNA hairpins

► [mature.fa](#)

Fasta format sequences of all mature miRNA sequences

► [miRNA.diff](#)

Changes between the last release and this

► [miRNA.dead](#)

List of entries that have been removed from the database

► [miFam.dat](#)

Family classification of miRNA hairpin sequences

Genome coordinates

[aae.gff3](#)

[ame.gff3](#)

[ath.gff3](#)

[bmo.gff3](#)

[bta.gff3](#)

[cbr.gff3](#)

[cel.gff3](#)

[cfa.gff3](#)

[cre.gff3](#)

[dme.gff3](#)

[dps.gff3](#)

[dre.gff3](#)

[ebv.gff3](#)

[fru.gff3](#)

[gga.gff3](#)

[hcmv.gff3](#)

[hsa.gff3](#)

[kshv.gff3](#)

[mdo.gff3](#)

[mghv.gff3](#)

[mml.gff3](#)

[mmu.gff3](#)

[osa.gff3](#)

[ptc.gff3](#)

[ptr.gff3](#)

[rno.gff3](#)

[sme.gff3](#)

[tae.gff3](#)

[tni.gff3](#)

[vvi.gff3](#)

[xtr.gff3](#)

[zma.gff3](#)



Download fasta from miRBase [II]

- <http://www.mirbase.org/ftp.shtml>
- Right click to download mature.fa and hairpin.fa
- Rename mature.fa to all_mature.fa
- After uploading all_mature.fa:
 - `catch_fasta_with_keyword.pl hsa all_mature.fa`
 - `catch_fasta_with_keyword.pl hsa hairpin.fa`



Find species names in miRBase

- <ftp://mirbase.org/pub/mirbase/CURRENT/>

Index of <ftp://mirbase.org/pub/mirbase/CURRENT/>

Up to higher level directory

Name	Size	Last Modified	
0_THIS_IS_RELEASE_20		2013-06-24	13:35:00
aliases.txt.gz	306 KB	2013-06-24	16:57:00
aliases.txt.zip	306 KB	2013-06-24	16:57:00
database_files		2013-06-25	08:32:00
genomes		2013-06-24	17:02:00
hairpin.fa.gz	993 KB	2013-06-24	16:57:00
hairpin.fa.zip	994 KB	2013-06-24	16:57:00
LICENSE	1 KB	2013-06-24	13:36:00
mature.fa.gz	488 KB	2013-06-24	16:57:00
mature.fa.zip	488 KB	2013-06-24	16:57:00
miFam.dat.gz	111 KB	2013-06-24	16:57:00
miFam.dat.zip	111 KB	2013-06-24	16:57:00
miRNA.dat.gz	2731 KB	2013-06-24	16:57:00
miRNA.dat.zip	2731 KB	2013-06-24	16:57:00
miRNA.dead.gz	18 KB	2013-06-24	16:57:00
miRNA.dead.zip	18 KB	2013-06-24	16:57:00
miRNA.diff.gz	56 KB	2013-06-24	16:57:00
miRNA.diff.zip	56 KB	2013-06-24	16:57:00
miRNA.str.gz	2167 KB	2013-06-24	16:57:00
miRNA.str.zip	2167 KB	2013-06-24	16:57:00
miRNA.xls.gz	1859 KB	2013-06-24	16:57:00
miRNA.xls.zip	1859 KB	2013-06-24	16:57:00
organisms.txt.gz	5 KB	2013-06-24	16:57:00
organisms.txt.zip	5 KB	2013-06-24	16:57:00
README	5 KB	2013-07-03	15:29:00



Application for miRDeep2

- <http://www.binfo.ncku.edu.tw/>



首頁 (Home) | 簡介 (Introduction) | 本中心服務 (Service) | 研究資源 (Resources) | 相關連結 (Links) | 意見留言 (Contact)

◎常用工具下

◎應用資源

Vector NTI
IBM軟體下載
NCKU BLAST
NCKU SNP Database
Repeat Masker Server
TAG Database
TheBEST
Metacore
GO Finder
LCR Download
GeneSpring
GBRED
CLC Genomics Workbench
miRNA prediction

基因體分析研習會(

「台灣生物資訊核
/6假"國立成功大學
為介紹TBI Core Fac
課程均搭配電腦實地
關資料庫以及研究平
miRDeep2、TAG、
情請連至活動網頁！

lity)，將於2013/9
研習會，本課程
析網路工具，所有
門快速熟悉各種相
RED、
13/8/22~9/2，詳

【暫停服務】生物資
2013-08-21)

告日期

生物資訊中心將於2013/8/27(二)進行網路換線工程，於2013/8/27(二)8:00停止各系統服務，2013/8/27(二)18:00開始恢復各系統運作，若有不便敬請見諒！



102年度「形態及生理醫學學門」成果報告暨學術交流研討會(9月27-28日)，於中興大學惠蓀林場舉辦。(公告日期 2013-07-24)



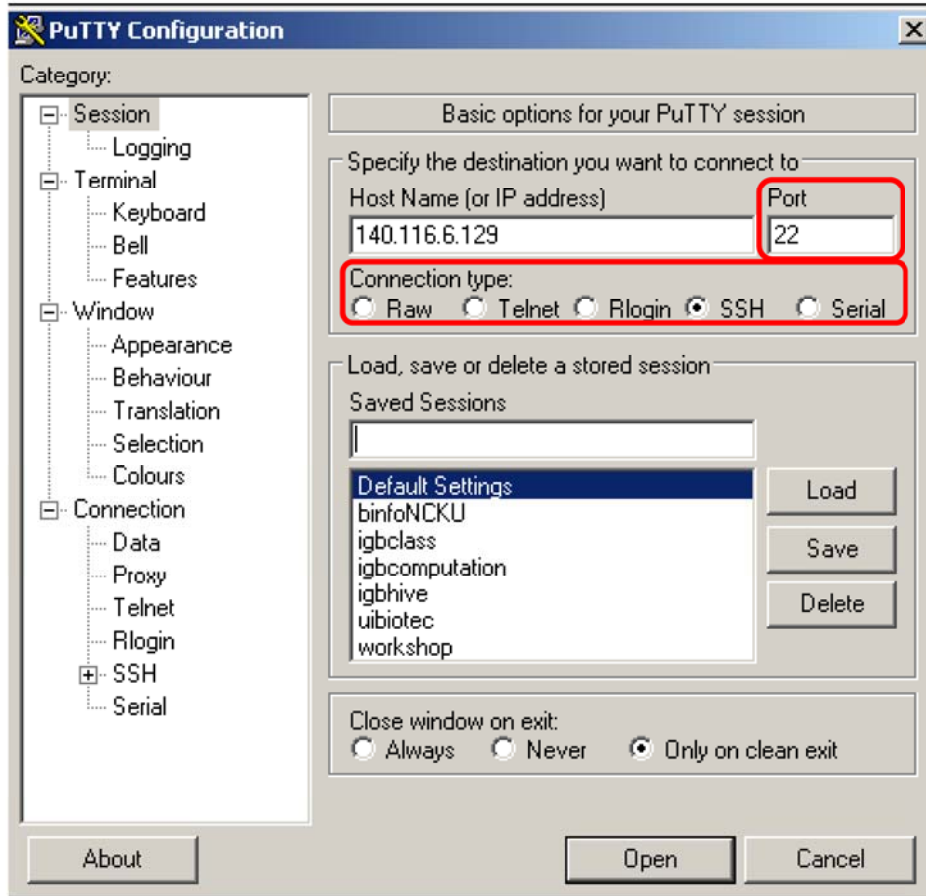
Thank you



miRDeep2 documentation

- https://www.mdc-berlin.de/36105849/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/documentation

PuTTY [login]



- IP: 140.116.56.169
- Notes:
 - Port: 22
 - Connection type: SSH

PuTTY [Practice with Linux commands]

```
login as: user_name  
user_name@140.116.6.129's password: █
```

- ls
- ls -l
- ls --help
- Q1: How many files do you see after ls command?
- Q2: Any difference between ls and ls -l?



Reads formatting [extra work]

- **cat filename | tr -d '\r' > new_filename &**
- **cat reads_col.fa | tr -d '\r' > new_reads_col.fa &**
- **Use ls -l to check the file sizes of reads_col.fa and new_reads_col.fa**