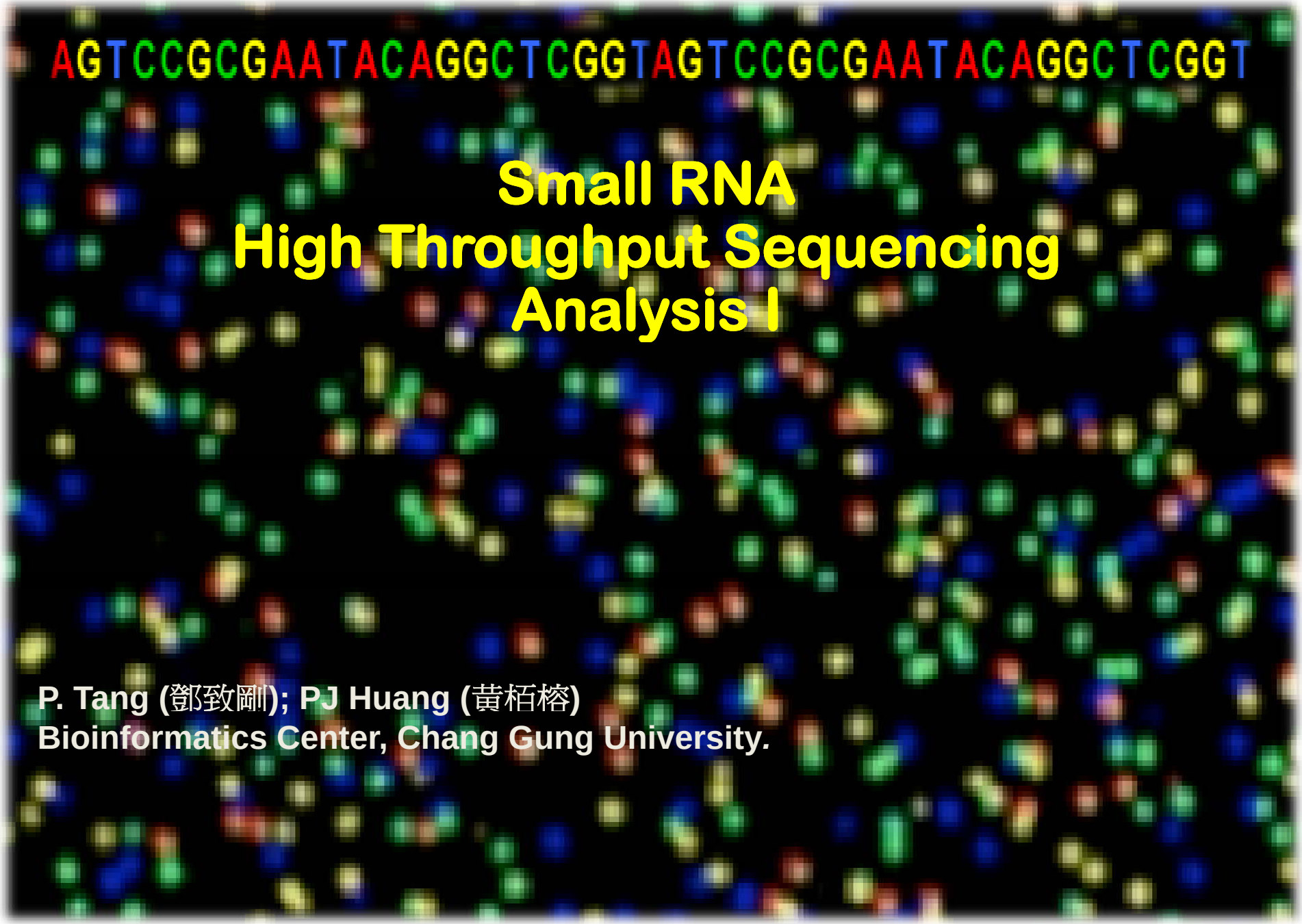


Small RNA High Throughput Sequencing Analysis I

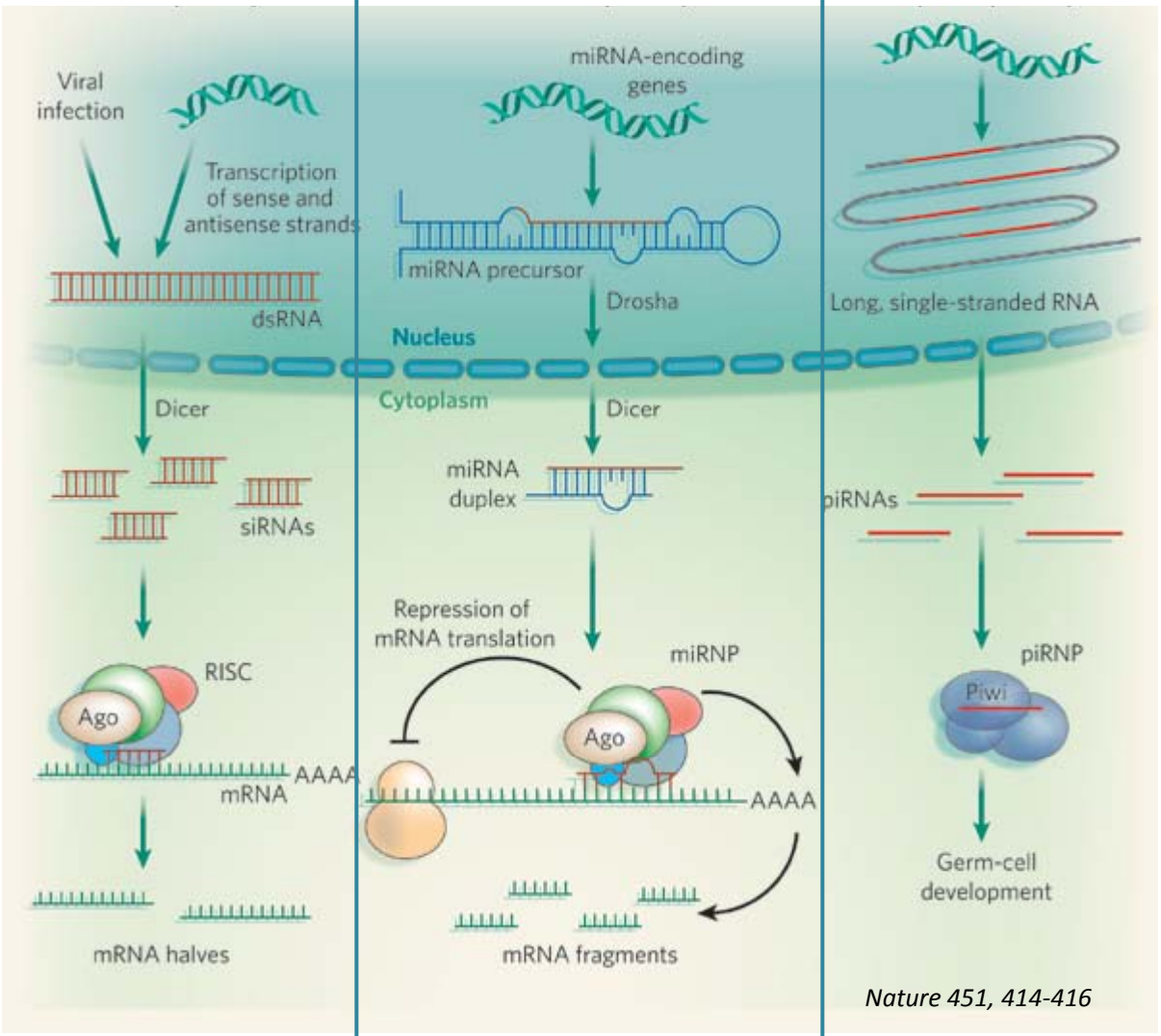
P. Tang (鄧致剛); PJ Huang (黃栢榕)
Bioinformatics Center, Chang Gung University.

Non-coding Regulatory RNAs

siRNA
Formed through cleavage of synthetic long double-stranded RNA molecules

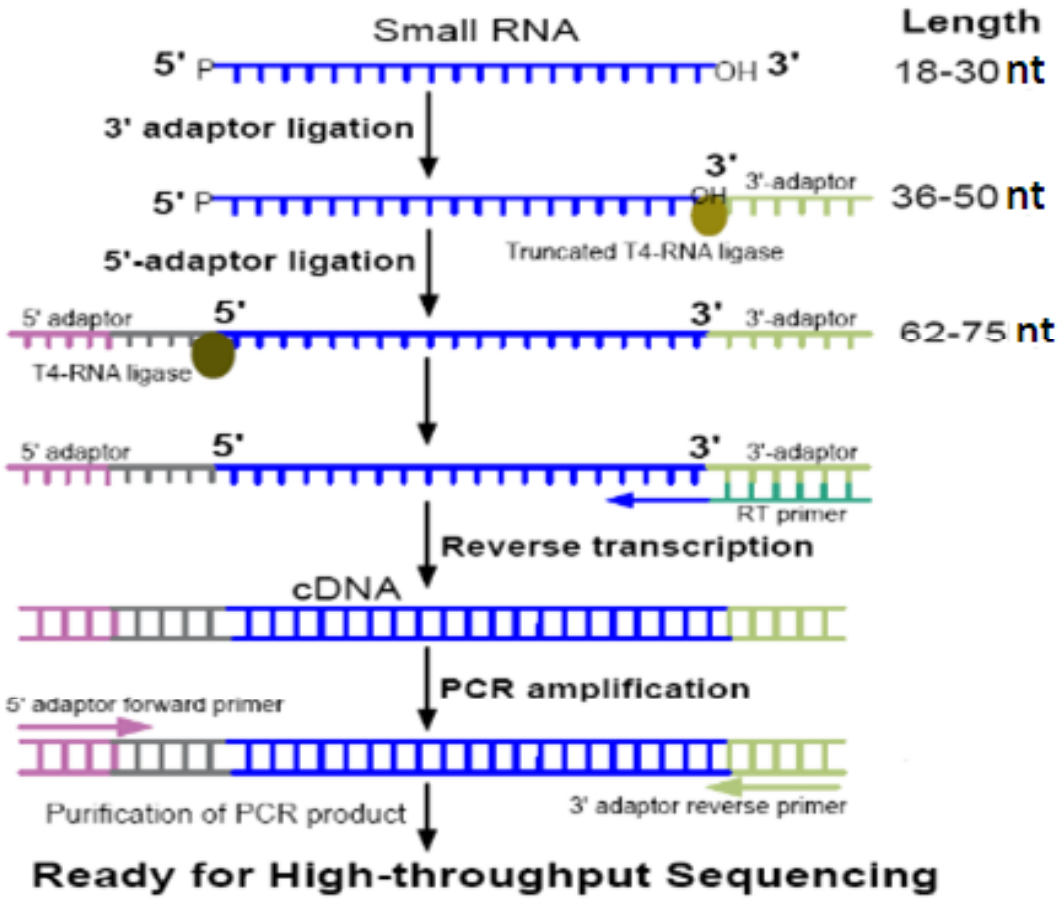
microRNA (miRNA)
Processed from long, single-stranded RNA sequences that fold into hairpin structures

Piwi-RNA (piRNA)
Generated from long single-stranded precursors.

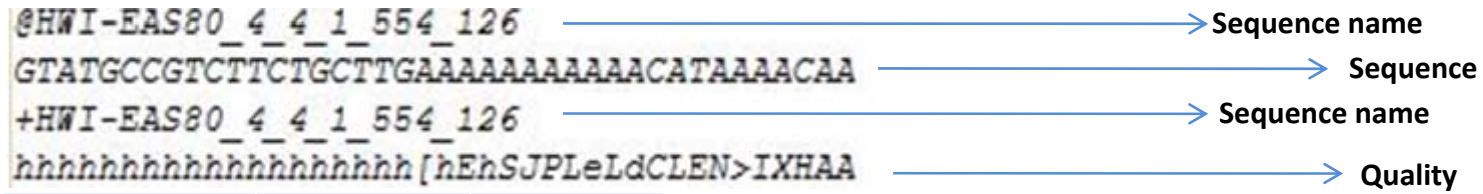


Nature 451, 414-416

Sample Preparation for small RNA Sequencing



Data Processing

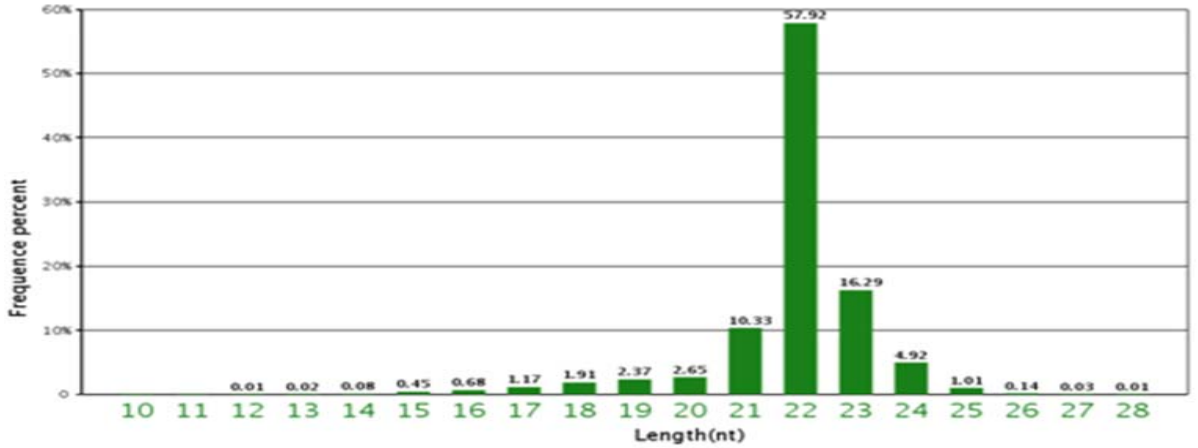


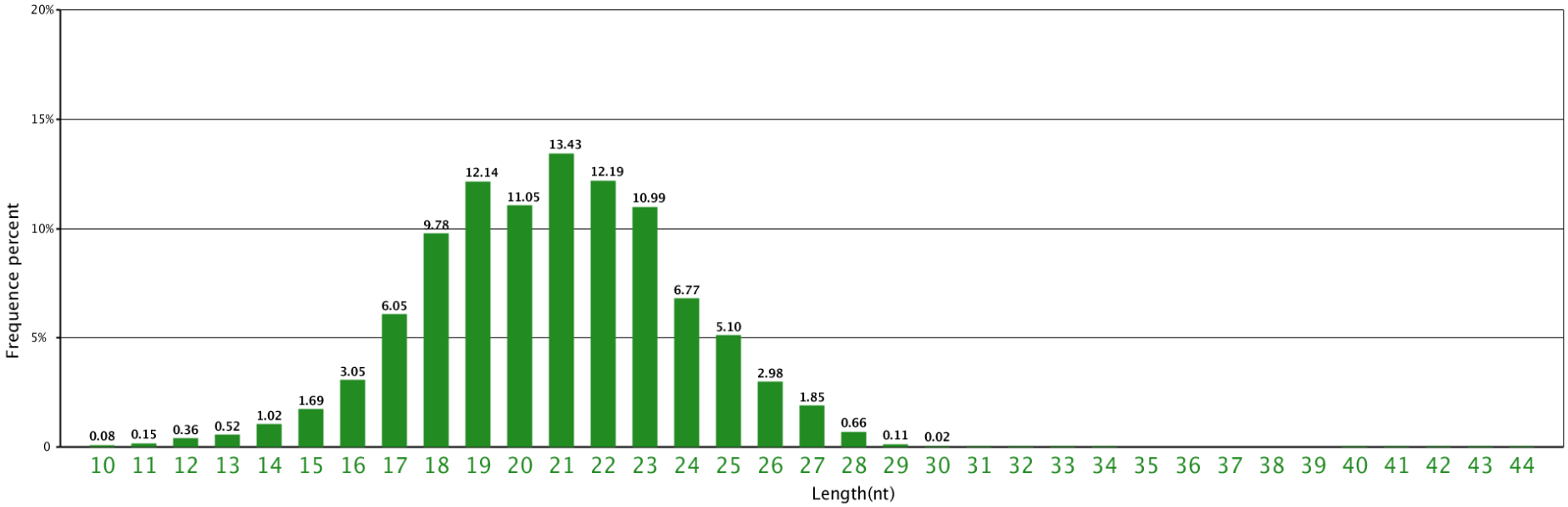
Sequencing error rate(E)	Sequencing quality	Corresponding character
5%	13	M
1%	20	T
0.1%	30	^
0.01%	40	h

The data is processed by the following steps:

1) Getting rid of low quality reads
2) Getting rid of reads with 5' primer contaminants
3) Getting rid of reads without 3' primer
4) Getting rid of reads without the insert tag
5) Getting rid of reads with poly A/G/C/T/N
6) Getting rid of reads shorter than 16-18nt ?
7) Summarize the length distribution of the clean reads

Which one is the correct distribution?

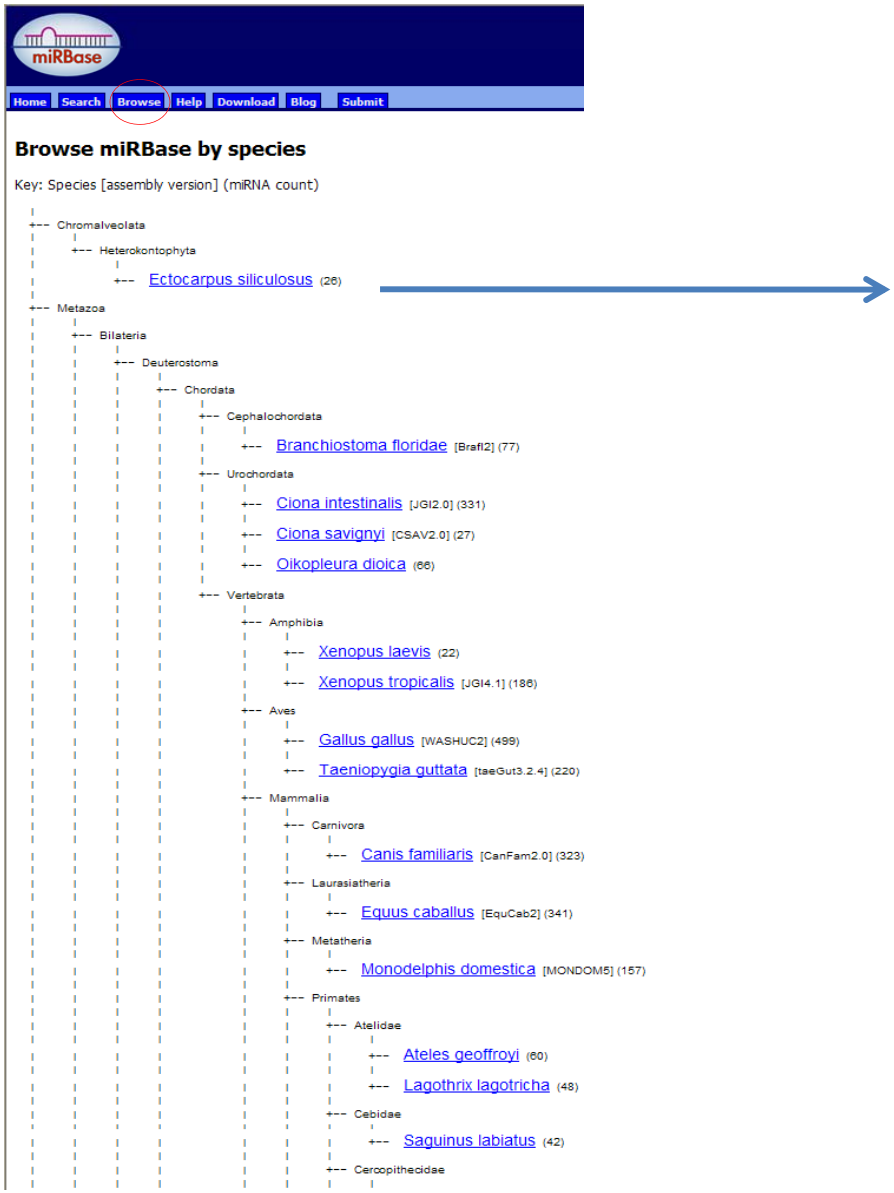




Understand the Reference Database

http://www.mirbase.org/





ID	Accession	Fetch
esi-MIR3450	MI0014677	☐
esi-MIR3451	MI0014689	☐
esi-MIR3452	MI0014692	☐
esi-MIR3453	MI0014688	☐
esi-MIR3454a	MI0014672	☐
esi-MIR3454b	MI0014678	☐
esi-MIR3454c	MI0014674	☐
esi-MIR3454d	MI0014690	☐
esi-MIR3454e	MI0014683	☐
esi-MIR3454f	MI0014675	☐
esi-MIR3454g	MI0014671	☐
esi-MIR3455	MI0014676	☐
esi-MIR3456	MI0014691	☐
esi-MIR3457	MI0014684	☐
esi-MIR3458	MI0014695	☐
esi-MIR3459	MI0014686	☐
esi-MIR3460	MI0014679	☐
esi-MIR3461	MI0014681	☐
esi-MIR3462	MI0014687	☐
esi-MIR3463	MI0014682	☐
esi-MIR3464	MI0014670	☐
esi-MIR3465	MI0014694	☐
esi-MIR3466	MI0014673	☐
esi-MIR3467	MI0014685	☐
esi-MIR3468	MI0014693	☐
esi-MIR3469	MI0014680	☐

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

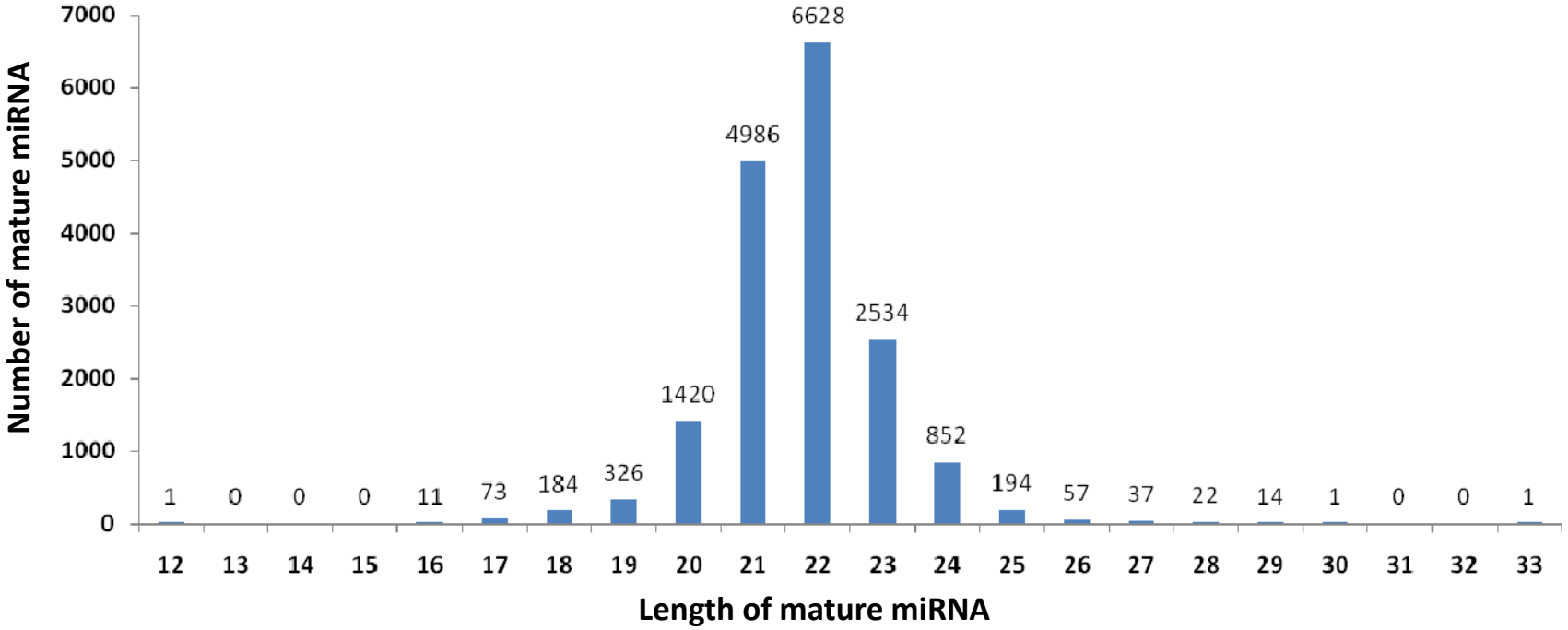
aga.gff	ame.gff	ath.gff	bmo.gff	bta.gff	cbr.gff	cel.gff
cfa.gff	cre.gff	dme.gff	dps.gff	dre.gff	ebv.gff	fru.gff
gga.gff	hcmv.gff	hsa.gff	kshv.gff	mdo.gff	mghv.gff	mml.gff
mmu.gff	osa.gff	ptc.gff	ptr.gff	rno.gff	sme.gff	tae.gff
tni.gff	vvi.gff	xtr.gff	zma.gff			

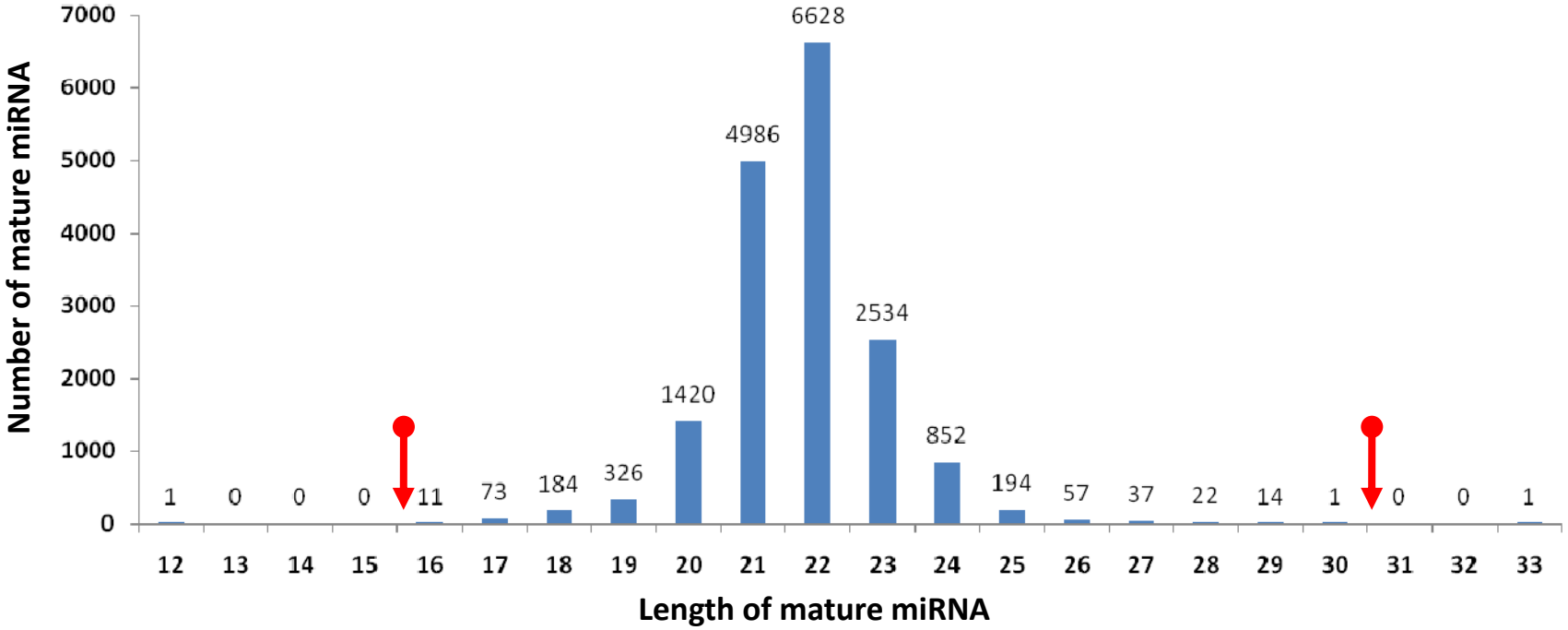
mature.fa

```
>hsa-miR-3944 MIMAT0018360 Homo sapiens miR-3944
UUCGGGCUGGCCUGCUGCUCCGG
>hsa-miR-3945 MIMAT0018361 Homo sapiens miR-3945
AGGGCAUAGGAGAGGGUUGAUAU
>far-miR159 MIMAT0018362 Festuca arundinacea miR159
UUUGGAUUGAAGGGAGCUCUG
>far-miR160 MIMAT0018363 Festuca arundinacea miR160
UGCCUGGCUCCCUGUAUGCCA
```

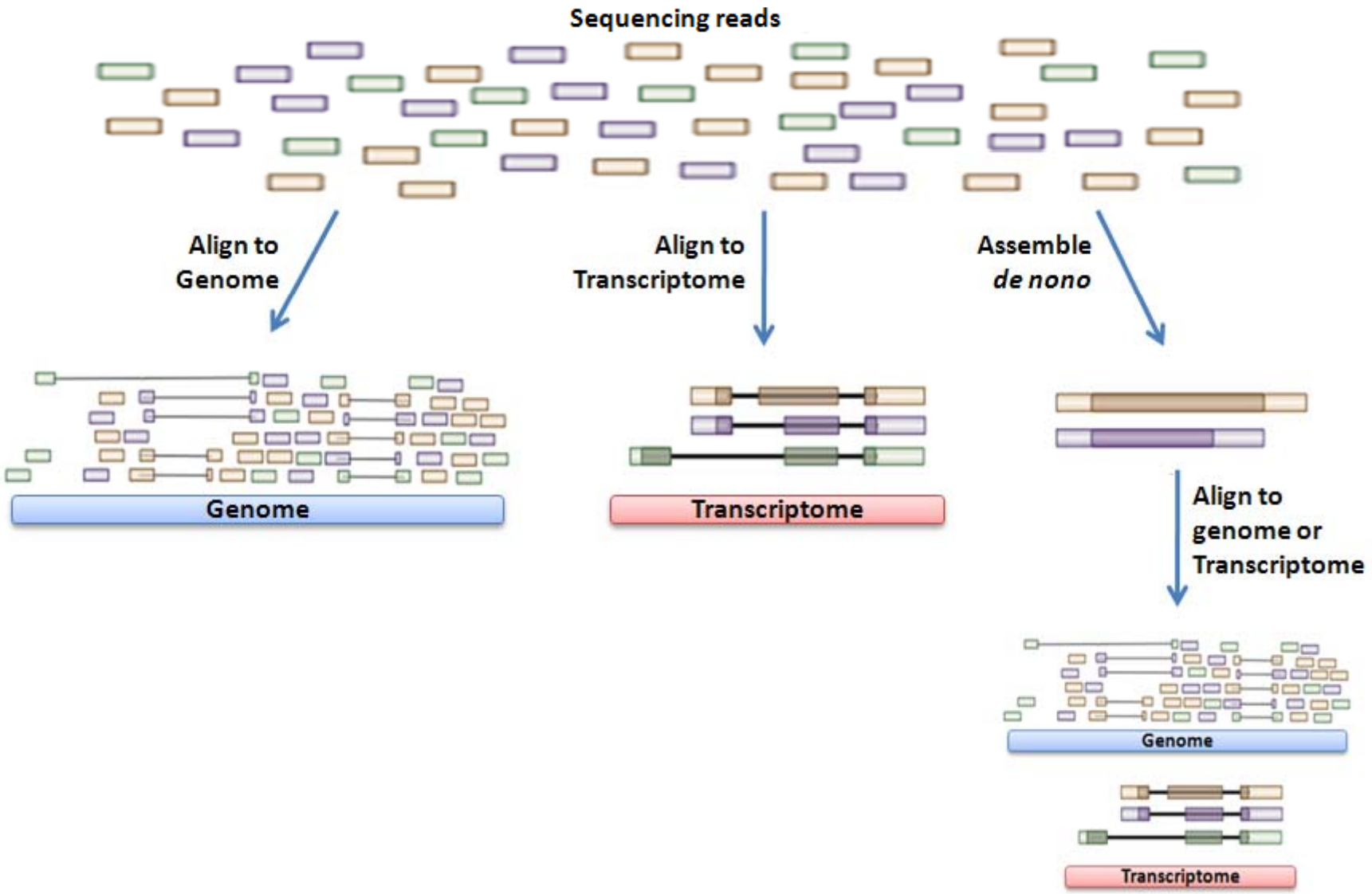
hairpin.fa

```
>cel-let-7 MI0000001 Caenorhabditis elegans let-7 stem-loop
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAAC
UAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA
>cel-lin-4 MI0000002 Caenorhabditis elegans lin-4 stem-loop
AUGCUUCCGGCCUGUUCCCUGAGACCUCAAGUGUGAGUGUACUAUUGAUGCUUCACACCU
GGGCUCUCCGGGUACCAGGACGGUUUGAGCAGAU
```


Number of mature miRNA
7000
6000
5000
4000
3000
2000
1000
0
1
0
0
0
11
73
184
326
1420
4986
6628
2534
852
194
57
37
22
14
1
0
0
1
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
Length of mature miRNA


Number of mature miRNA
7000
6000
5000
4000
3000
2000
1000
0
1
0
0
0
11
73
184
326
1420
4986
6628
2534
852
194
57
37
22
14
1
0
0
1
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
Length of mature miRNA

Sequencing Reads Assessment



Mapping Result

Map to Genome	reads number	percentage
Total Reads	27489704	100.00%
Total BasePairs	2474073360	100.00%
Total Mapped Reads	25739521	93.63%
perfect match	18872284	68.65%
<=2bp mismatch	6867237	24.98%
unique match	21717286	79.00%
multi-position match	4022235	14.63%
Total Unmapped Reads	1750183	6.37%

Map to Gene	reads number	percentage
Total Reads	27489704	100.00%
Total BasePairs	2474073360	100.00%
Total Mapped Reads	23370329	85.01%
perfect match	17137891	62.34%
<=2bp mismatch	6232438	22.67%
unique match	20387103	74.16%
multi-position match	2983226	10.85%
Total Unmapped Reads	4119375	14.99%

Drawbacks for each strategy

Alignment to genome

Computationally expensive

It is never a good idea to simply align HTS data to the genome

Need a spliced aligner or a surrogate (such as including exon-exon junction sequences in 'genome')

Alignment to transcriptome

Reads deriving from non-genic structures may be 'forcibly' (and erroneously) aligned to genes

Incorrect gene expression values

False positive SNVs

Many other potential problems

Assembly

Low expression = difficult/impossible to assemble

Misassemblies/fragmented contigs due to repeats

Requires vast amounts of memory

Analysis Tools for small-RNA Deep Sequencing

LINUX/UNIX BASED TOOLS

Discovering microRNAs from deep sequencing data using **miRDeep**
Nature Biotechnology (2008) 26 (4), pp. 407-415

MIRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression
BMC Bioinformatics (2009) 10, art. no. 1471, pp. 328 .

SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.
Nucleic acids research (2010) 38 (5), pp. e34

MIReNA: Finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data
Bioinformatics (2010) 26 (18), art. no. btq329, pp. 2226-2234

WEB-BASED TOOLS

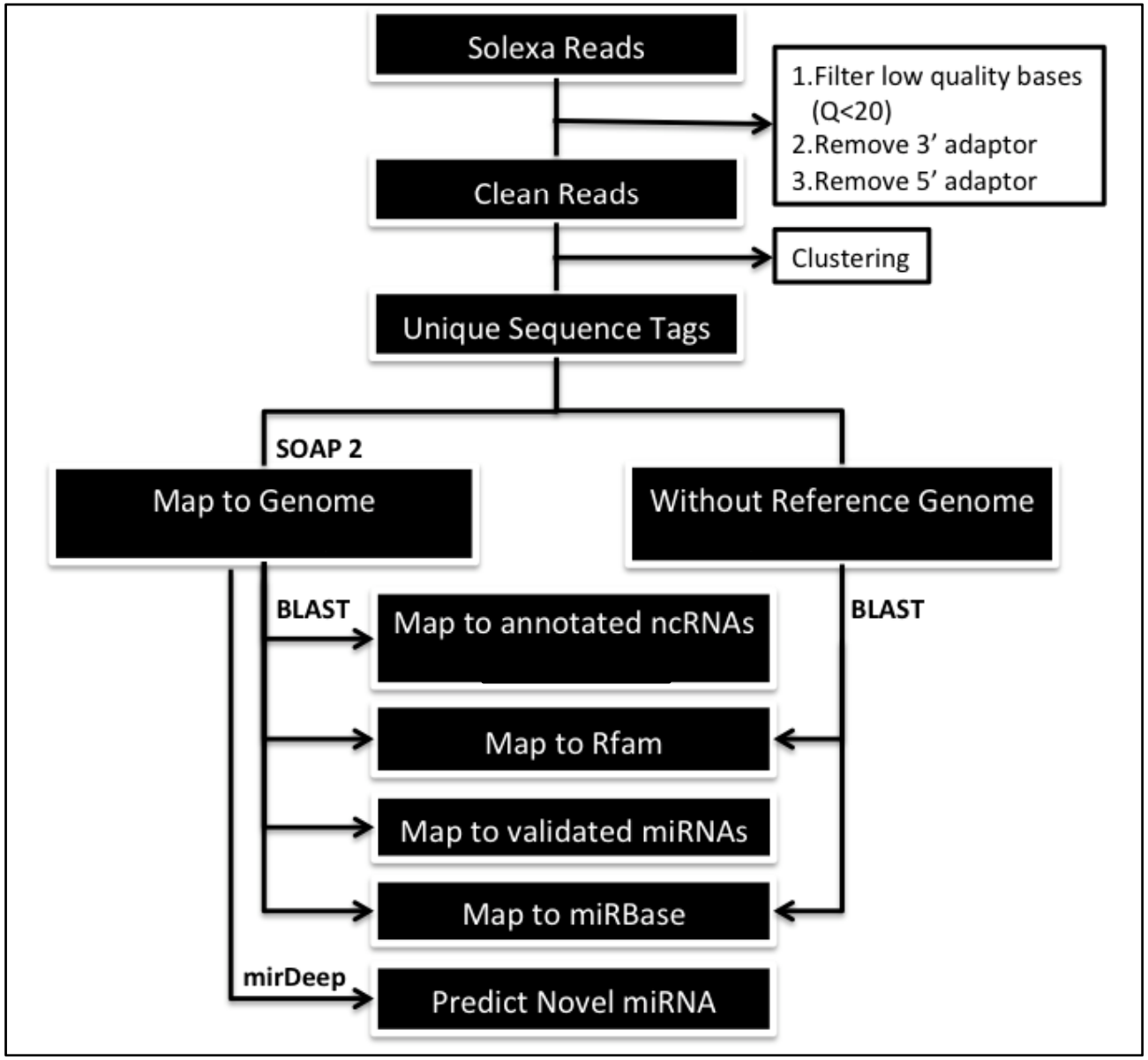
miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments
Nucleic Acids Research (2009) 37 (SUPPL. 2), pp. W68-W76 *
(http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php)

DSAP: deep-sequencing small RNA analysis pipeline
Nucleic Acids Research. (2010) 38(Web Server issue): W385–W391
(http://dsap.cgu.edu.tw/)

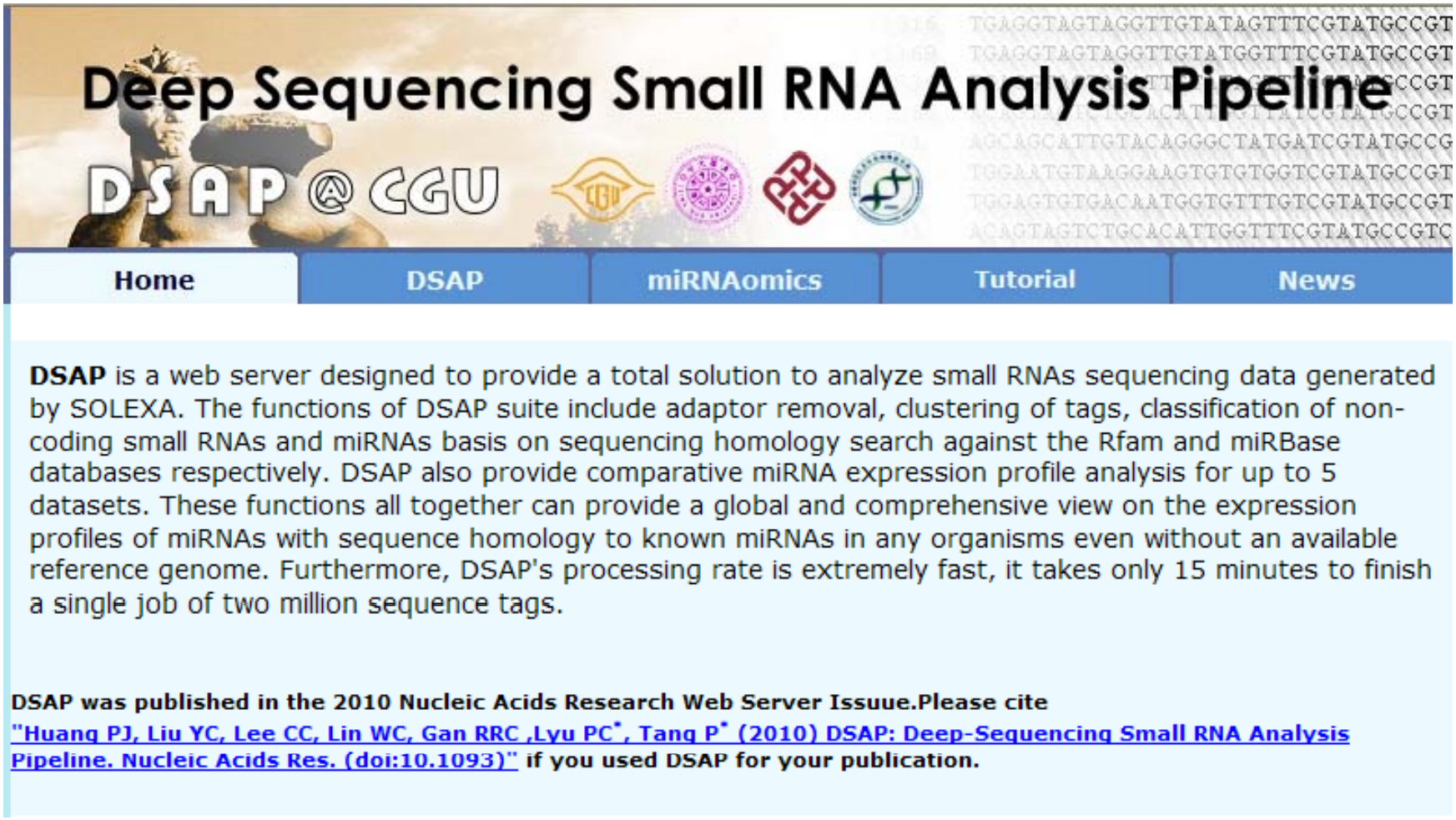
mirTools: microRNA profiling and discovery based on high-throughput sequencing
Nucleic Acids Research. (2010) 38(Web Server issue): W392–W397
(http://centre.bioinformatics.zj.cn/mirtools/)

Comparison of Tools

	DSAP	miRTools	miRAnalyzer	miRExpress
Service type	Web server	Web server	Web server	Local server (command-line)
Installation	NOT AVAILABLE	NOT AVAILABLE	NOT AVAILABLE	Local installation
Input file format	Tab-delineated	Adapter trimmed FASTA	Tab-delineated, Fasta	Tab-delineated, Fastq
miRNA Database	miRBase v16	miRBase V15	miRBase v10.1	User update
Number of organisms available	115	15	9	115
Built in additional database	Rfam	Rfam	Rfam v9.1, RepBase	NOT AVAILABLE
Classification of small RNA	YES	YES	YES	NOT AVAILABLE
Classification of miRNA	YES	YES	YES	YES
Iso-miRNA alignment	YES	YES	YES	NOT AVAILABLE
Cross-species distribution of miRNA	YES(Graphic)	NOT AVAILABLE	NOT AVAILABLE	YES(text)
Differential miRNA expression	YES (Graphic)	YES	NOT AVAILABLE	YES(text)
New miRNA prediction	NOT AVAILABLE	YES	Limited to 9 organisms	NOT AVAILABLE
Target prediction	NOT AVAILABLE	YES	Limited to 9 organisms	NOT AVAILABLE
Results output format	Table, Graphic	Table	Table	Text file
Results link to miRBase	YES	NOT AVAILABLE	NOT AVAILABLE	NOT AVAILABLE


Solexa Reads
1.Filter low quality bases (Q<20)
2.Remove 3' adaptor
3.Remove 5' adaptor
Clean Reads
Clustering
Unique Sequence Tags
SOAP 2
Map to Genome
Without Reference Genome
BLAST
Map to annotated ncRNAs
BLAST
Map to Rfam
Map to validated miRNAs
Map to miRBase
mirDeep
Predict Novel miRNA

http://dsap.cgu.edu.tw



Workflow

Clean up

↓

Clustering

↓

ncRNA matching

(Rfam V10)

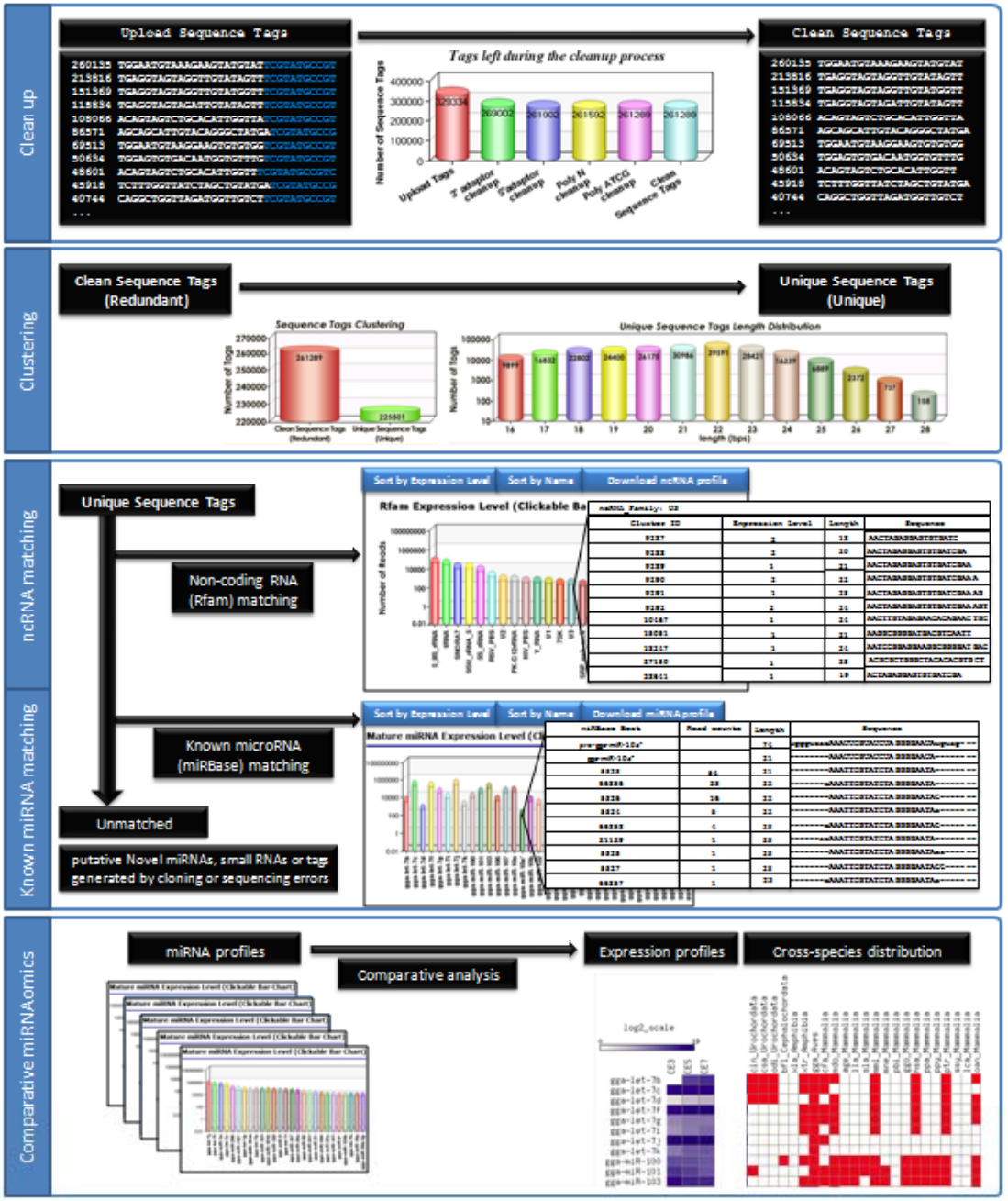
↓

Known miRNA Matching

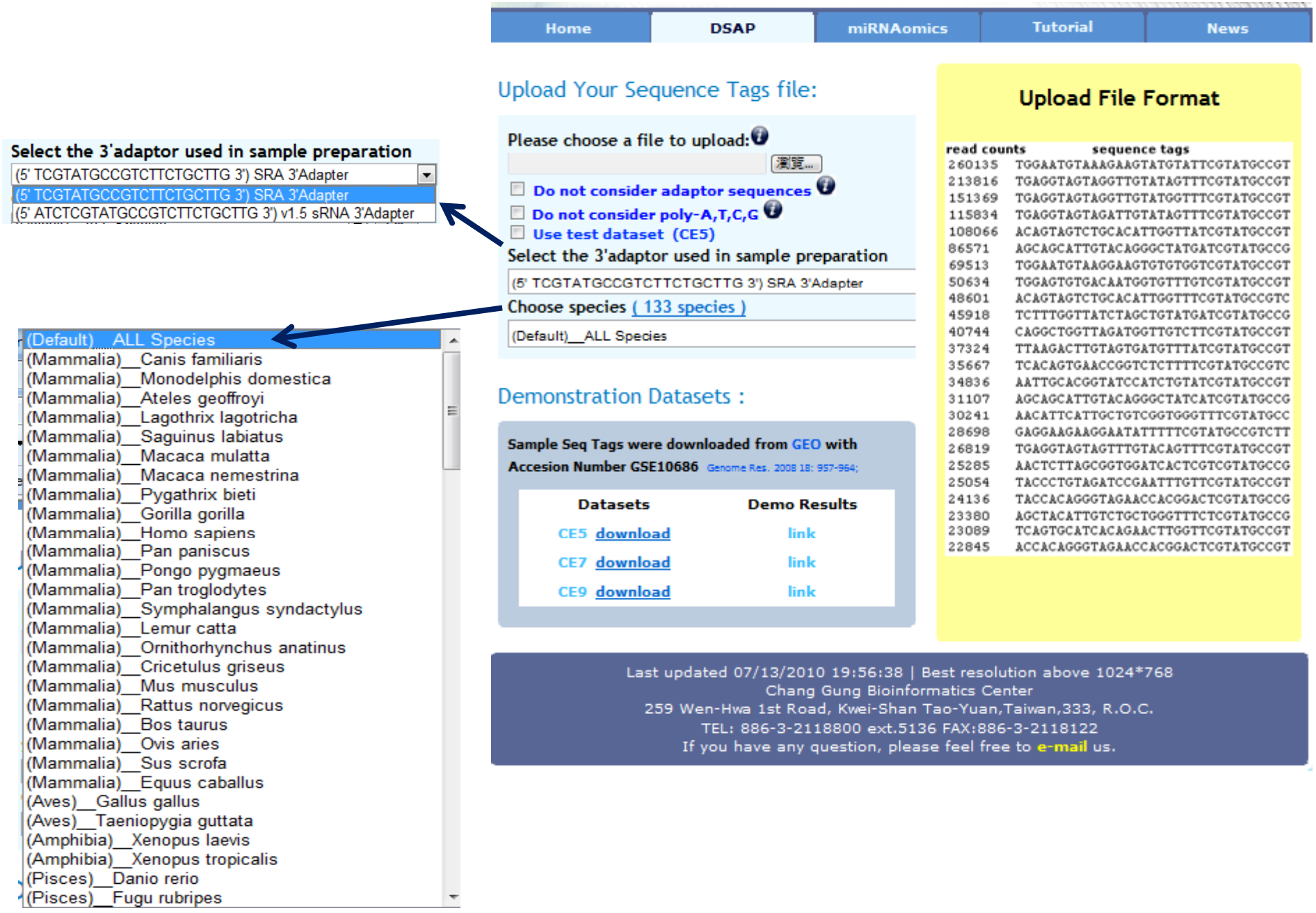
(miRBase V16)

↓

Comparative miRNAomics



DSAP Data Input Page



Implementation



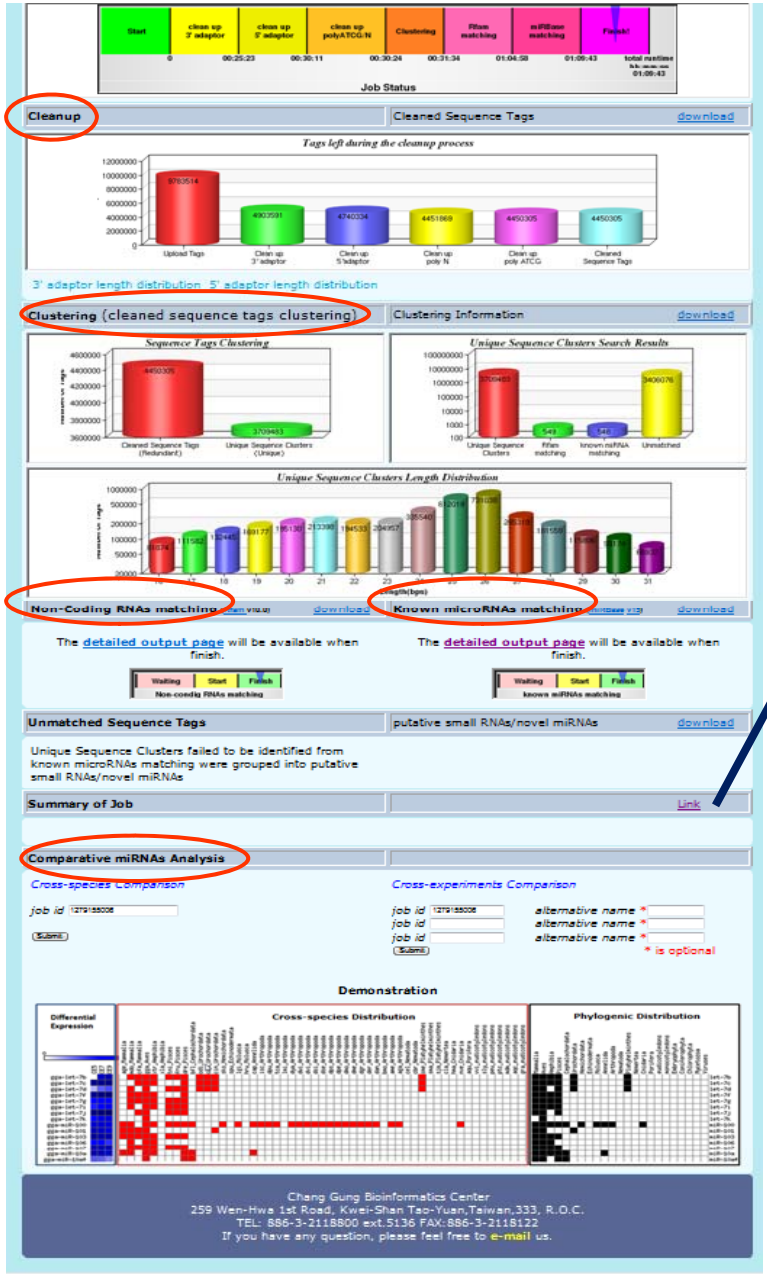
DSAP runs on a Linux CentOS 64-bit server housing two quad-core Intel® Xeon® 5300 Series Processors and 16 GB RAM.

Datasets	Number of Sequence Tags(reads)	Processing time (hh:mm:ss)
Chicken 1	153,406	00:01:20
Chicken 2	220,166	00:01:58
Protist 1	395,939	00:02:50
Protist 2	697,983	00:03:56
Chicken 3	329,334	00:03:42
Protist 3	754,059	00:04:38
Protist 4	736,939	00:05:02
Plant 1	1,643,030	00:11:34
Plant 2	2,090,730	00:13:53
Fish	2,121,561(9,460,897)	00:44:29
Human	9,783,514(38,346,007)	01:09:43

Sequence Tags File Format

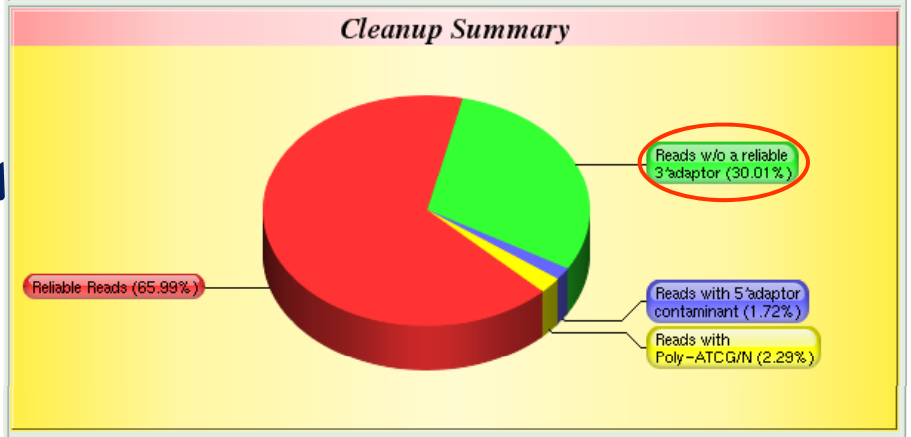
```
260135   TGGAATGTAAAGAAGTATGTATTCGTATGCCGT
213816   TGAGGTAGTAGGTTGTATAGTTTCGTATGCCGT
151369   TGAGGTAGTAGGTTGTATGGTTTCGTATGCCGT
115834   TGAGGTAGTAGATTGTATAGTTTCGTATGCCGT
108066   ACAGTAGTCTGCACATTGGTTATCGTATGCCGT
86571    AGCAGCATTGTACAGGGCTATGATCGTATGCCG
69513    TGGAATGTAAGGAAGTGTGTGGTCGTATGCCGT
50634    TGGAGTGTGACAATGGTGTTTGTCGTATGCCGT
48601    ACAGTAGTCTGCACATTGGTTTCGTATGCCGTC
45918    TCTTTGGTTATCTAGCTGTATGATCGTATGCCG
40744    CAGGCTGGTTAGATGGTTGTCTTCGTATGCCGT
37324    TTAAGACTTGTAGTGATGTTTATCGTATGCCGT
35667    TCACAGTGAACCGGTCTCTTTTCGTATGCCGTC
34836    AATTGCACGGTATCCATCTGTATCGTATGCCGT
31107    AGCAGCATTGTACAGGGCTATCATCGTATGCCG
30241    AACATTCATTGCTGTCGGTGGGTTTCGTATGCC
28698    GAGGAAGAAGGAATATTTTTCGTATGCCGTCTT
26819    TGAGGTAGTAGTTTGTACAGTTTCGTATGCCGT
25285    AACTCTTAGCGGTGGATCACTCGTCGTATGCCG
25054    TACCCTGTAGATCCGAATTTGTTCGTATGCCGT
24136    TACCACAGGGTAGAACCACGGACTCGTATGCCG
23380    AGCTACATTGTCTGCTGGGTTTCTCGTATGCCG
23089    TCAGTGCATCACAGAACTTGGTTCGTATGCCGT
```

DSAP Data Output



No. of total reads	38346007
No. of sequence tags	9783514
No. of tags after 3'adaptor clean up	4903591
No. of tags after 5'adaptor Clean up	4740334
No. of tags after poly N cleanup	4451869
No. of tags after poly-ATCG Clean up	4450305
No. of cleaned sequence tags	4450305
No. of reads in cleaned sequence tags	25303578
Percentage of reliable reads #	65.99

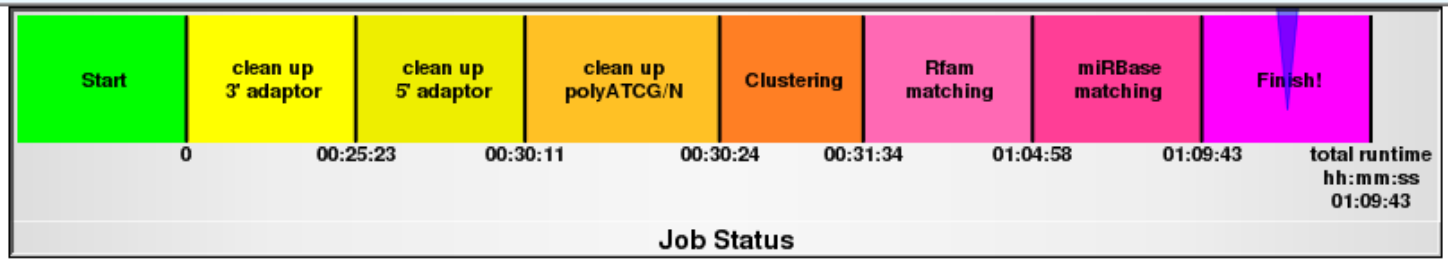
#Percentage of reliable reads=(No. of reads in cleaned sequence tags/No. of total reads)*100



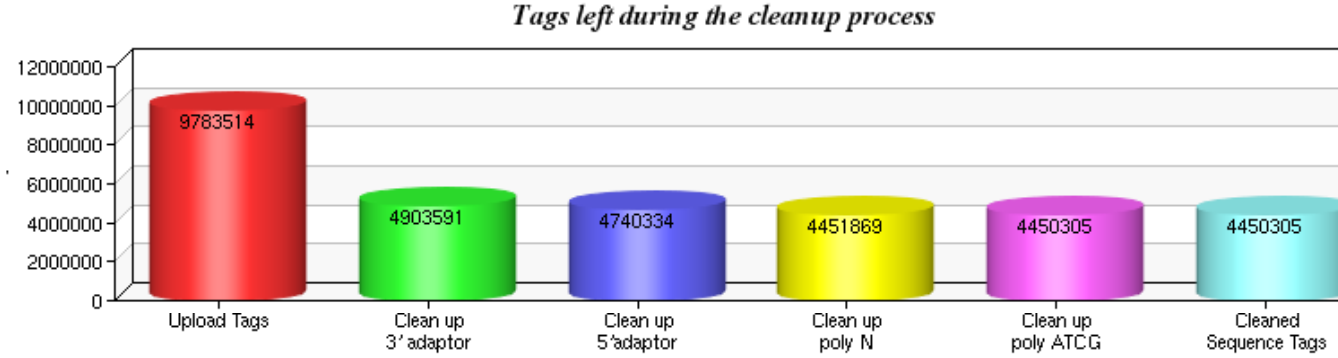
No. of Unique Sequence Clusters (USC)	3709483
No. of matched ncRNA in Rfam	549
No. of USC matched to Rfam	282658
No. of total reads matched to Rfam	2930944
Percentage of reads matched to Rfam	11.58
No. of matched miRNA in miRBase	548
No. of USC matched to miRBase	20749
No. of total reads matched to miRBase	10830075
Percentage of reads matched to miRNA	42.80
USC Unmatched	3406076

Number of reads: 38,346,007 **Number of Tags: 9,783,514**

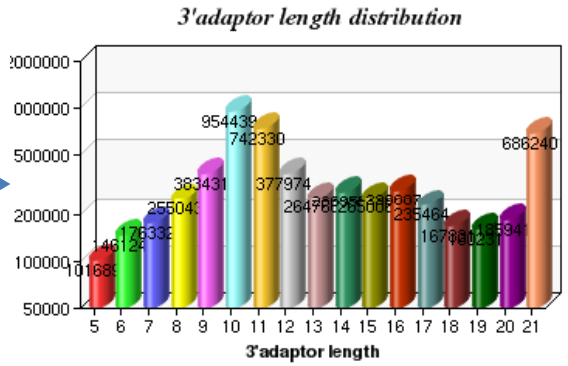
JOB ID:1234567890 Bookmark this Page!

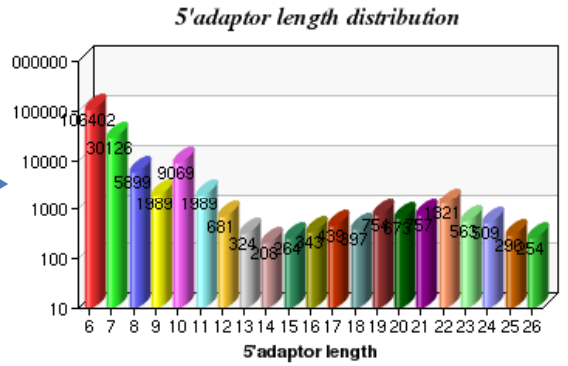


Cleanup Cleaned Sequence Tags download



3' adaptor length distribution 5' adaptor length distribution



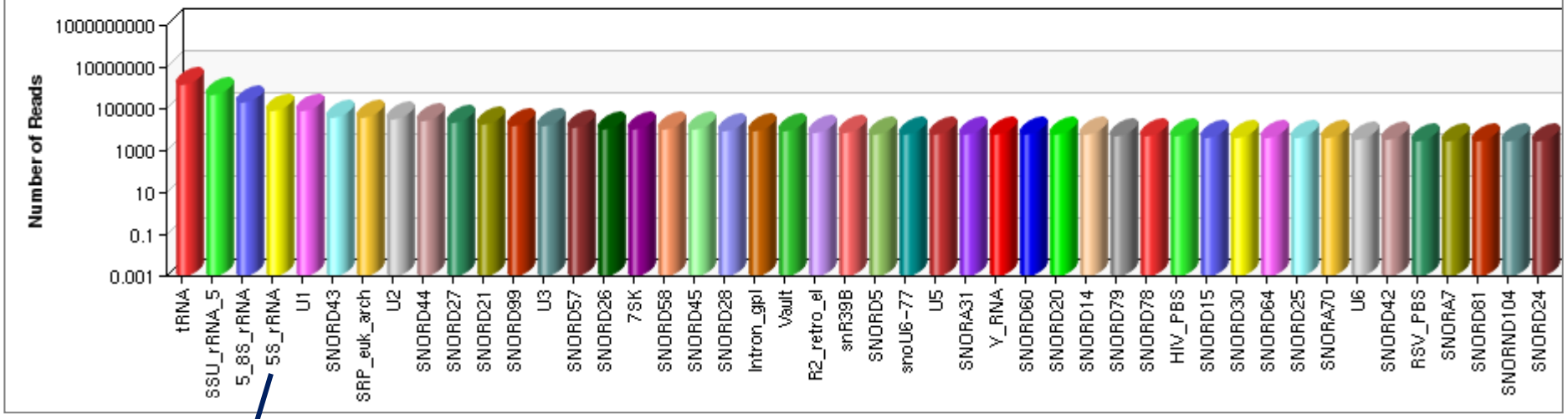


Non-Coding RNAs Searching

keyword search Go

Sort By Expression Level | Sort By Name | Download ncRNA profile

Rfam Expression Level (Clickable Bar Chart) 549 distinct ncRNAs found



ncRNA_Family: 5S_rRNA

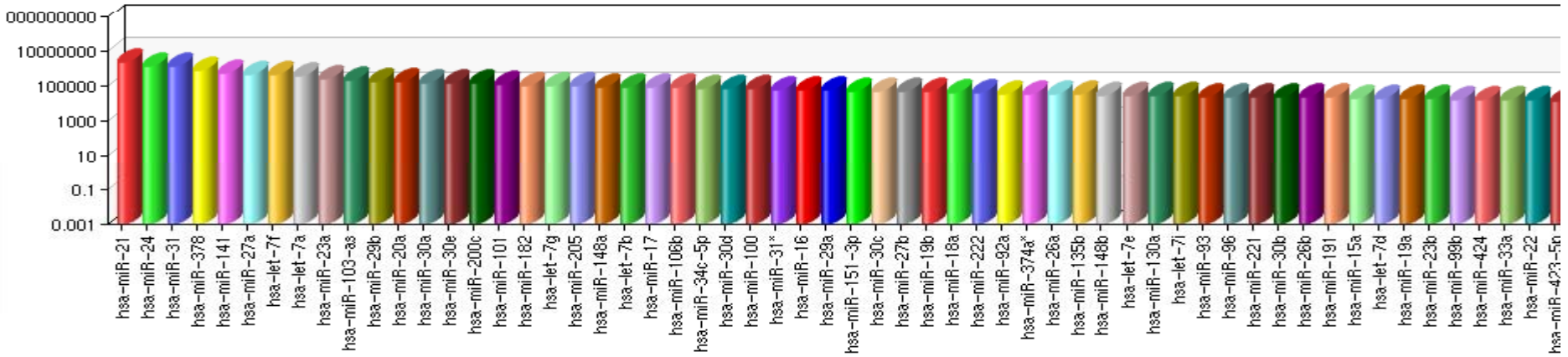
Cluster ID	Expression Level	Length	Sequence
7	1	20	AAAAAAAAAAAAAGCCAATC
4230	1	21	AAAACCGGGTGCTGTAGGATC
7225	1	29	AAAAGCATGTTAGTGTTCTGATGGCAATC
12886	1	28	AAAATTTAATTAGAATCACAAATTTATC

Family: *5S_rRNA* (RF00001)
Description: *5S ribosomal RNA*

miRNAs Searching

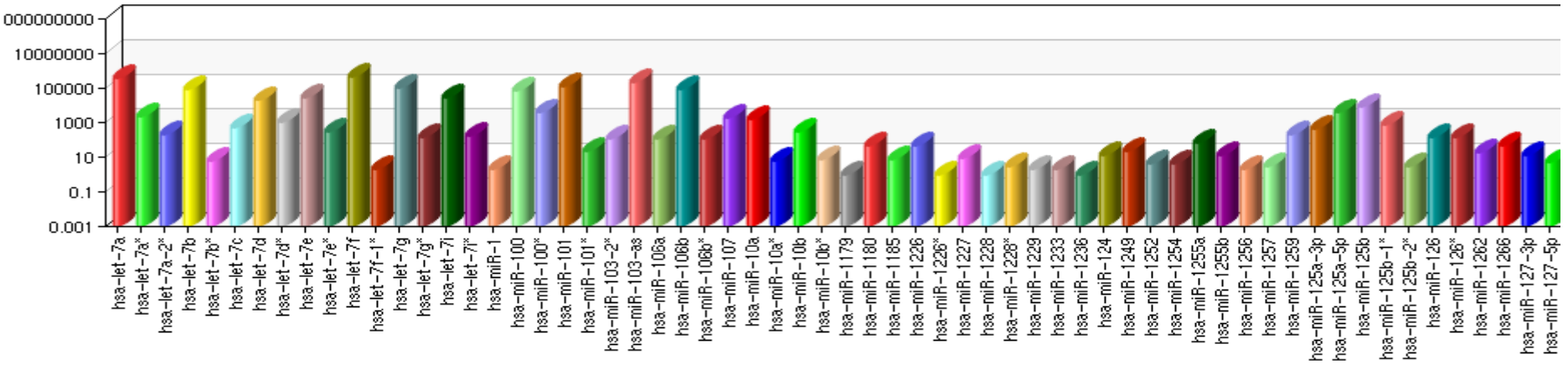
Sort By Expression Level | Sort By Name | Download miRNA profile

Mature miRNA Expression Level (Clickable Bar Chart): 548 distinct mature miRNAs found

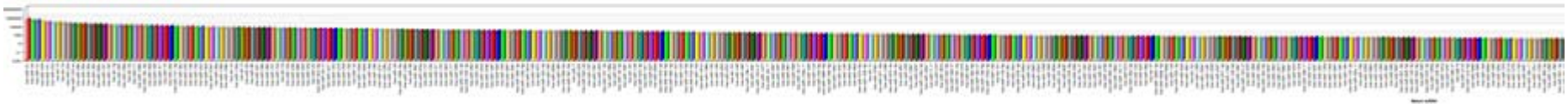


Sort By Expression Level | Sort By Name | Download miRNA profile

Mature miRNA Expression Level (Clickable Bar Chart): 548 distinct mature miRNAs found



Iso-miR



miRBase Best Hit/Cluster ID	Sort by Read Counts	Sort by Length	Sort by Sequence
pre-dre-miR-30e		80	cgggcuacUGUAAACAUCCUUGACUGGAAGcuggugca
dre-miR-30e		22	--------UGUAAACAUCCUUGACUGGAAG--------
1174487	149530	23	--------TGTAAACATCCTTGACTGGAAGc-------
1174500	21044	24	--------TGTAAACATCCTTGACTGGAAGct------
1174484	4170	22	--------TGTAAACATCCTTGACTGGAAG--------

miRBase

miRBase

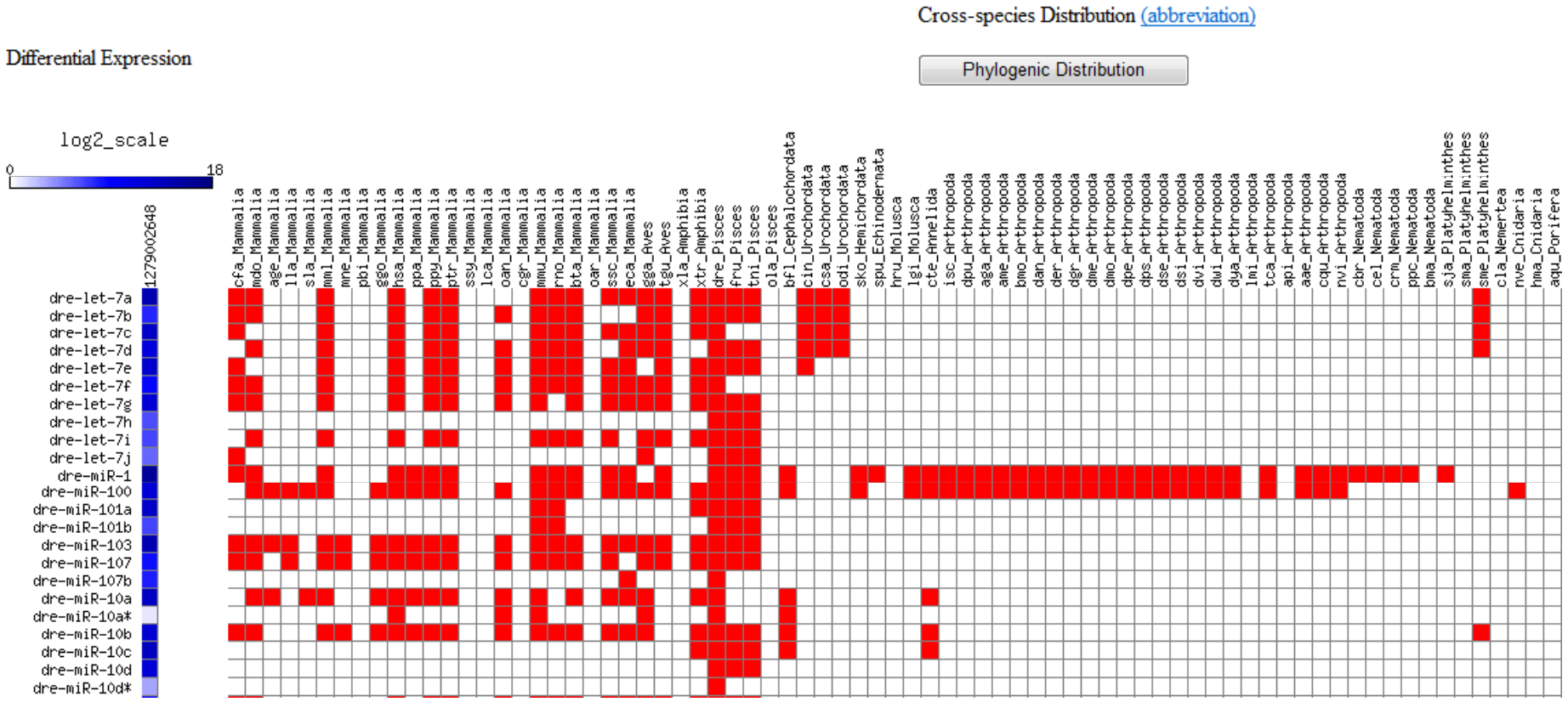
MANCHESTER 1824

Home | Search | Browse | Genomics | Help | Download | Submit | dre-mir-30e-2 | Search

Stem-loop sequence MI0001950

Accession	MI0001950
ID	dre-mir-30e-2
Description	Danio rerio miR-30e-2 stem-loop
Stem-loop	uacg a uu ggug a ggcu cuguaaacaucc gacuggaagcu c c |||| ||||||||||| ||||||||||| | ccga gacguuuguagg cugacuuucga g a ucaa c -- ggua u Get sequence

Cross-species Distribution of Identified miRNAs

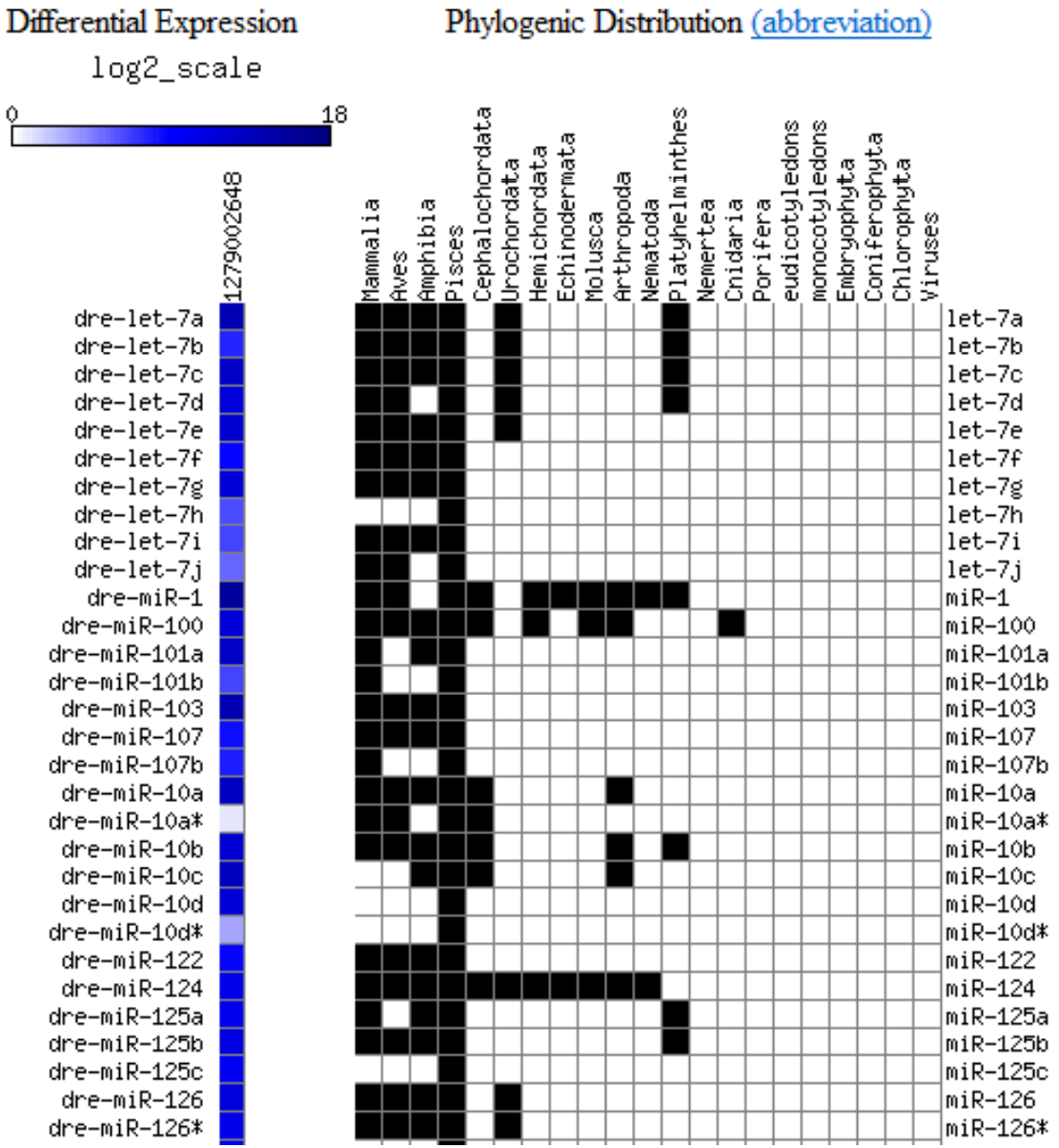


Cross-species Distribution of Identified miRNAs

Abbreviation	Organism	Classification							
cfa	Canis familiaris	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Carnivora	
mdo	Monodelphis domestica	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Metatheria	
age	Ateles geoffroyi	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Atelidae
lla	Lagothrix lagotricha	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Atelidae
sla	Saguinus labiatus	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Cebidae
mml	Macaca mulatta	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Cercopithecidae
mne	Macaca nemestrina	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Cercopithecidae
pbi	Pygathrix bieti	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Cercopithecidae
ggo	Gorilla gorilla	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Hominidae
hsa	Homo sapiens	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Hominidae
ppa	Pan paniscus	Metazoa	Bilateria	Deuterostoma	Chordata	Vertebrata	Mammalia	Primates	Hominidae



Phylogenic Distribution of Identified miRNAs



Comparative miRNAomics

1.Paste your own miRNAs expression profile

DSAP also accept experimental results from other miRNA expression analysis such as stem-loop real-time PCR, microarray or SOLiD sequencing in a tab-delimited format as shown below.

Paste your own expression profile in the textarea

miR	CE5	CE7	CE9
gga-let-7b	1	12352	29903
gga-let-7c	163164	156621	198263
gga-let-7d	6	44	81
gga-let-7f	128541	255562	346444
gga-let-7g	532	1156	1075
gga-let-7i	250	936	1877
gga-let-7j	232980	373519	446116
gga-let-7k	837	3414	4844
gga-miR-100	7358	6871	4168
gga-miR-101	23497	12104	4654
gga-miR-103	89271	41329	34030
gga-miR-106	996	493	167
gga-miR-107	32021	27360	23079
gga-miR-10a	26443	9883	4075
gga-miR-10a*	54	44	17

non-normalized

Submit Reset

2.Fill in the job identifiers provided by DSAP: (max. 5)

JOB ID1	1258095940	**Alternative Name** (optional)	human ESC
JOB ID2	1258096199	**Alternative Name** (optional)	human EB
JOB ID3	1256192842	**Alternative Name** (optional)	Chicken CE5
JOB ID4	1256195990	**Alternative Name** (optional)	Chicken CE7
JOB ID5	1256196016	**Alternative Name** (optional)	Chicken CE9

normalized

Submit

2.Comparative miRNAomics by mature miRNA: (max. 5)

JOB ID1 1258095940 **Alternative Name** (optional) human ESC
JOB ID2 1258096199 **Alternative Name** (optional) human EB
JOB ID3 1256192842 **Alternative Name** (optional) Chicken CE5
JOB ID4 1256195990 **Alternative Name** (optional) Chicken CE7
JOB ID5 1256196016 **Alternative Name** (optional) Chicken CE9
Submit

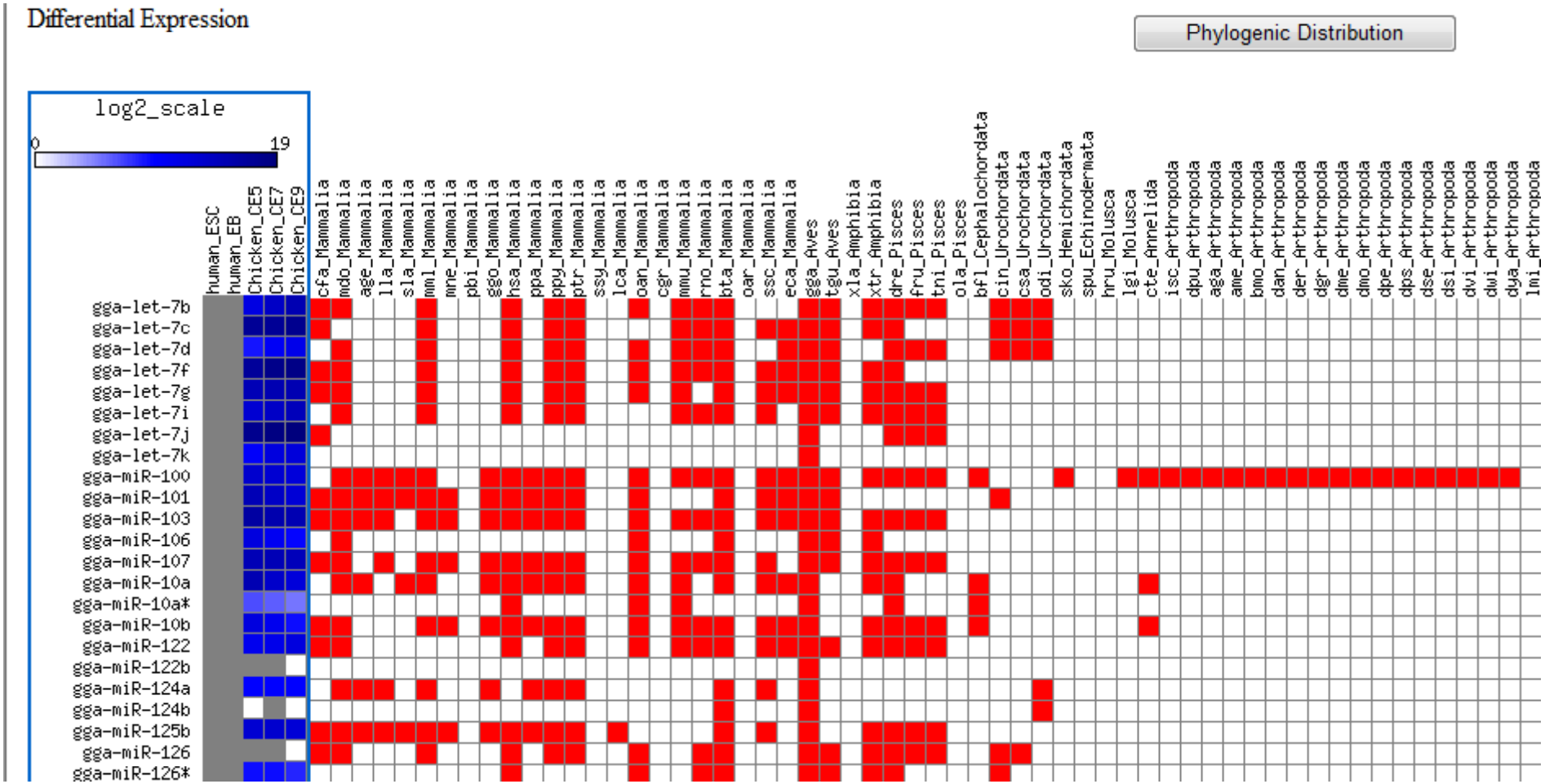
3.Comparative miRNAomics by miRNA Family: (max. 5)

JOB ID1 1258095940 **Alternative Name** (optional) human ESC
JOB ID2 1258096199 **Alternative Name** (optional) human EB
JOB ID3 1256192842 **Alternative Name** (optional) Chicken CE5
JOB ID4 1256195990 **Alternative Name** (optional) Chicken CE7
JOB ID5 1256196016 **Alternative Name** (optional) Chicken CE9
Submit

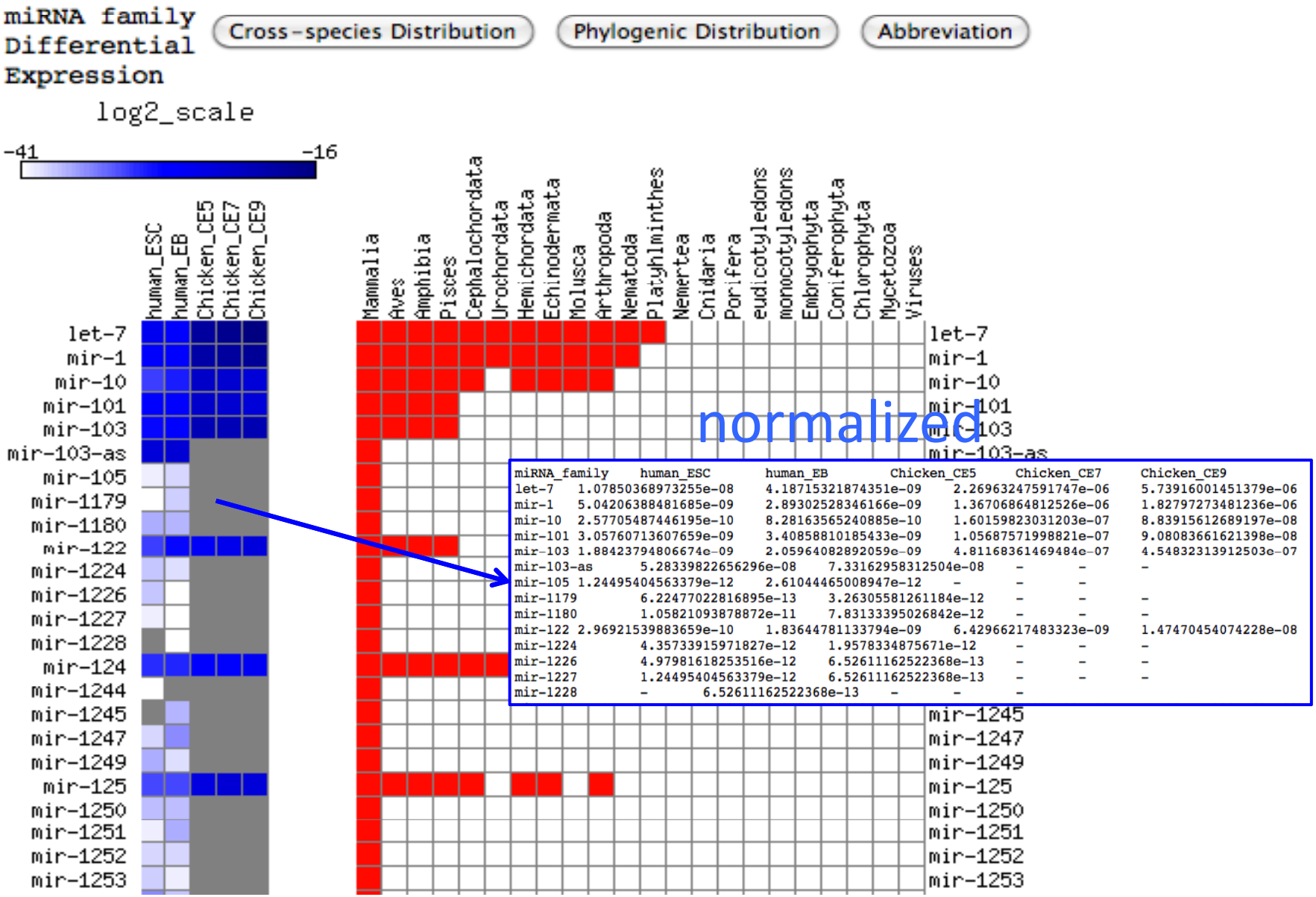
4.Pairwise Comparison: (Calculate fold-change and P-value from the normalized expression profile)

JOB ID1 1258095940 **Alternative Name** (optional) human ESC
JOB ID2 1258096199 **Alternative Name** (optional) human EB
Submit

Comparative miRNAomics by mature miRNA



Comparative miRNAomics by miRNA family



miRNA_family	human_ESC	human_EB	Chicken_CE5	Chicken_CE7	Chicken_CE9
let-7	1.07850368973255e-08	4.18715321874351e-09	2.26963247591747e-06	5.73916001451379e-06	
mir-1	5.04206388481685e-09	2.89302528346166e-09	1.36706864812526e-06	1.82797273481236e-06	
mir-10	2.57705487446195e-10	8.28163565240885e-10	1.60159823031203e-07	8.83915612689197e-08	
mir-101	3.05760713607659e-09	3.40858810185433e-09	1.05687571998821e-07	9.08083661621398e-08	
mir-103	1.88423794806674e-09	2.05964082892059e-09	4.81168361469484e-07	4.54832313912503e-07	
mir-103-as	5.28339822656296e-08	7.33162958312504e-08	-	-	-
mir-105	1.24495404563379e-12	2.61044465008947e-12	-	-	
mir-1179	6.22477022816895e-13	3.26305581261184e-12	-	-	-
mir-1180	1.05821093878872e-11	7.83133395026842e-12	-	-	-
mir-122	2.96921539883659e-10	1.83644781133794e-09	6.42966217483323e-09	1.47470454074228e-08	
mir-1224	4.35733915971827e-12	1.9578334875671e-12	-	-	-
mir-1226	4.97981618253516e-12	6.52611162522368e-13	-	-	-
mir-1227	1.24495404563379e-12	6.52611162522368e-13	-	-	-
mir-1228	-	6.52611162522368e-13	-	-	-

Pairwise Comparsion

No limit
P<0.01
P<0.05

