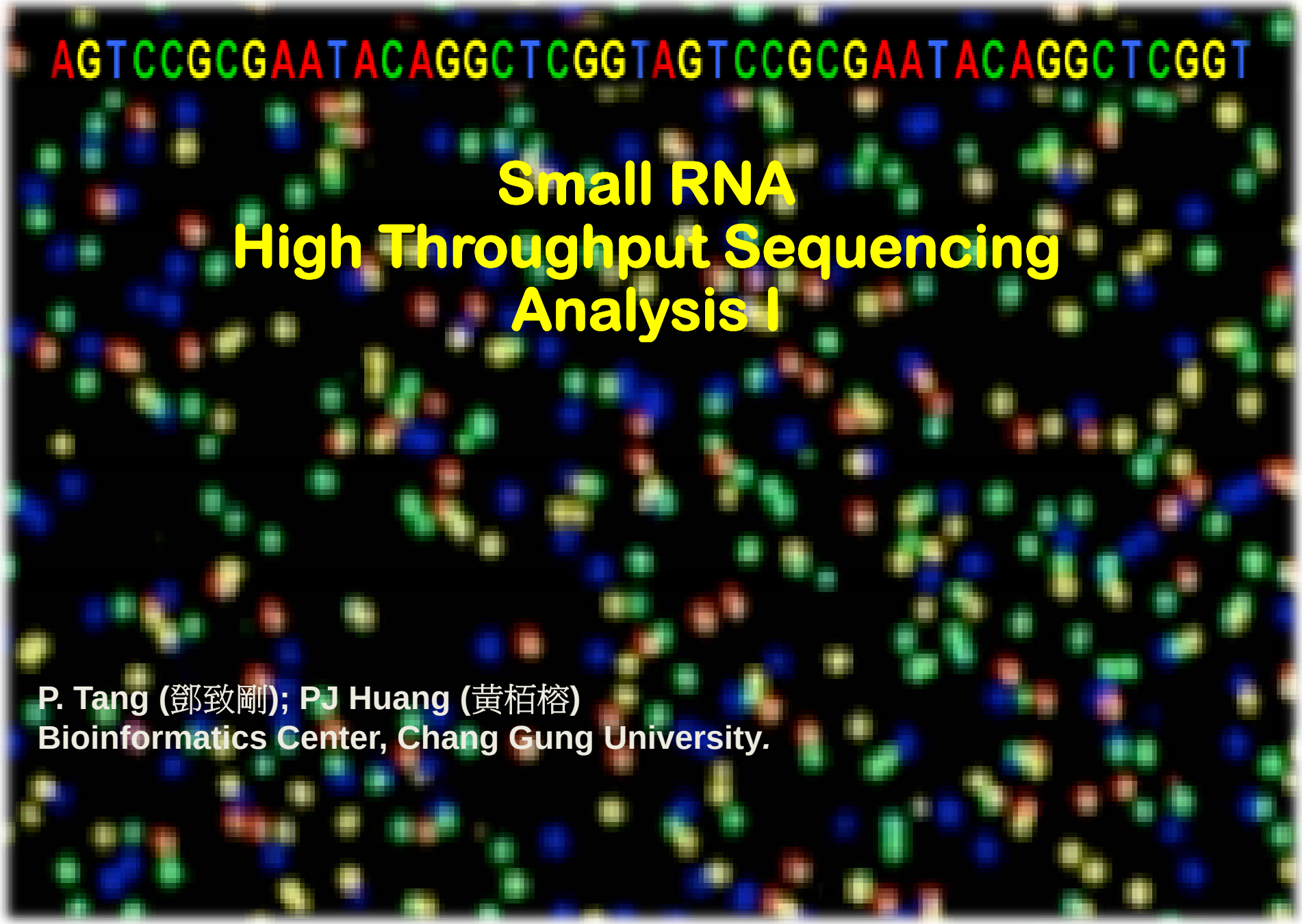


Small RNA High Throughput Sequencing Analysis I

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

Prominent members of the RNA family

Classic RNAs mediating protein synthesis

mRNAs (messenger RNAs) Transcripts of protein-coding genes that act as templates for protein synthesis.

rRNAs (ribosomal RNAs) RNA constituents of the ribonucleoprotein particles known as ribosomes, which mediate the decoding of mRNAs to the amino-acid sequences of proteins.

tRNAs (transfer RNAs) Adapter molecules carrying individual amino acids to the site of protein synthesis that recognize specific codons in mRNA.

Non-coding regulatory RNAs

siRNAs (small interfering RNAs) Small RNAs (20–25 nucleotides in length) formed through cleavage of long double-stranded RNA molecules. siRNAs are particularly important for taming the activity of transposons and combating viral infection, but they can also regulate protein-coding genes. Synthetic siRNAs can also be artificially expressed for experimental purposes.

Small RNAs (20-30 nt)

miRNAs (microRNAs) Small RNAs (20–25 nucleotides in length) that are encoded by specific genes and function in repressing mRNA translation or in mRNA degradation in plants and animals. They are processed from long, single-stranded RNA sequences that fold into hairpin structures.

piRNAs (Piwi-associated RNAs) Small RNAs (25–30 nucleotides in length) that are generated from long single-stranded precursors. They function in association with the Piwi subfamily of Argonaute proteins, and are essential for the development of germ cells.

Longer non-coding RNAs

Longer non-coding RNAs (70 - thousands nt)

RNAs of 70 to thousands of nucleotides that participate in various cellular processes, including mRNA splicing and ribosome biogenesis.

H.G. & W.F.

Nature ***451, 414-416(24 January 2008)***

Discovery of microRNA



C. elegans lin-4 (*Lee R et al. 1993*)
let-7 (*Reihhart et al. 2000*)



Victor Ambros

First miRNA - *C. elegans* lin-4 in 1993 by Victor Ambros & Lee

lin-4 causes temporal decreasing in LIN-14 protein expressed in the first larval stage (L1).

"The Caenorhabditis elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". CELL (1993)



Frank Slack

Second miRNA - *C. elegans* let-7 in 2000 by Frank Slack

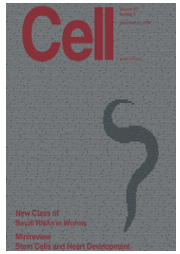
miRNA might be conservative in organic evolution and plays an important role in regulation of organic development.

"The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans" . NATURE (2000)

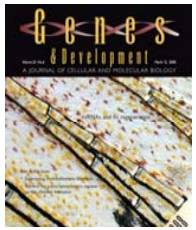




miRNAs host genes and transcription



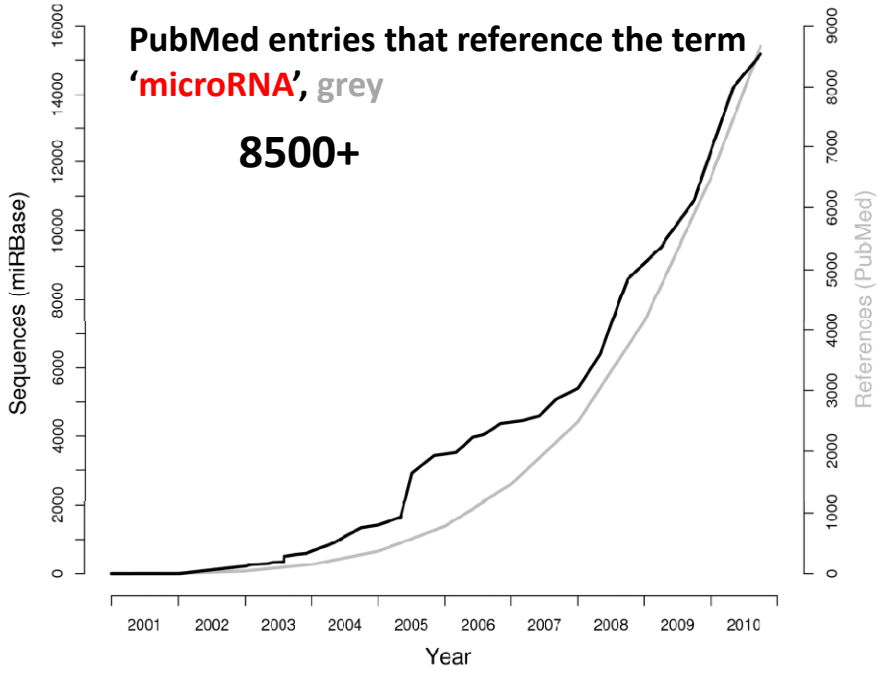
New class of small RNAs in worms



miRNAs and fin development



Stem cell division

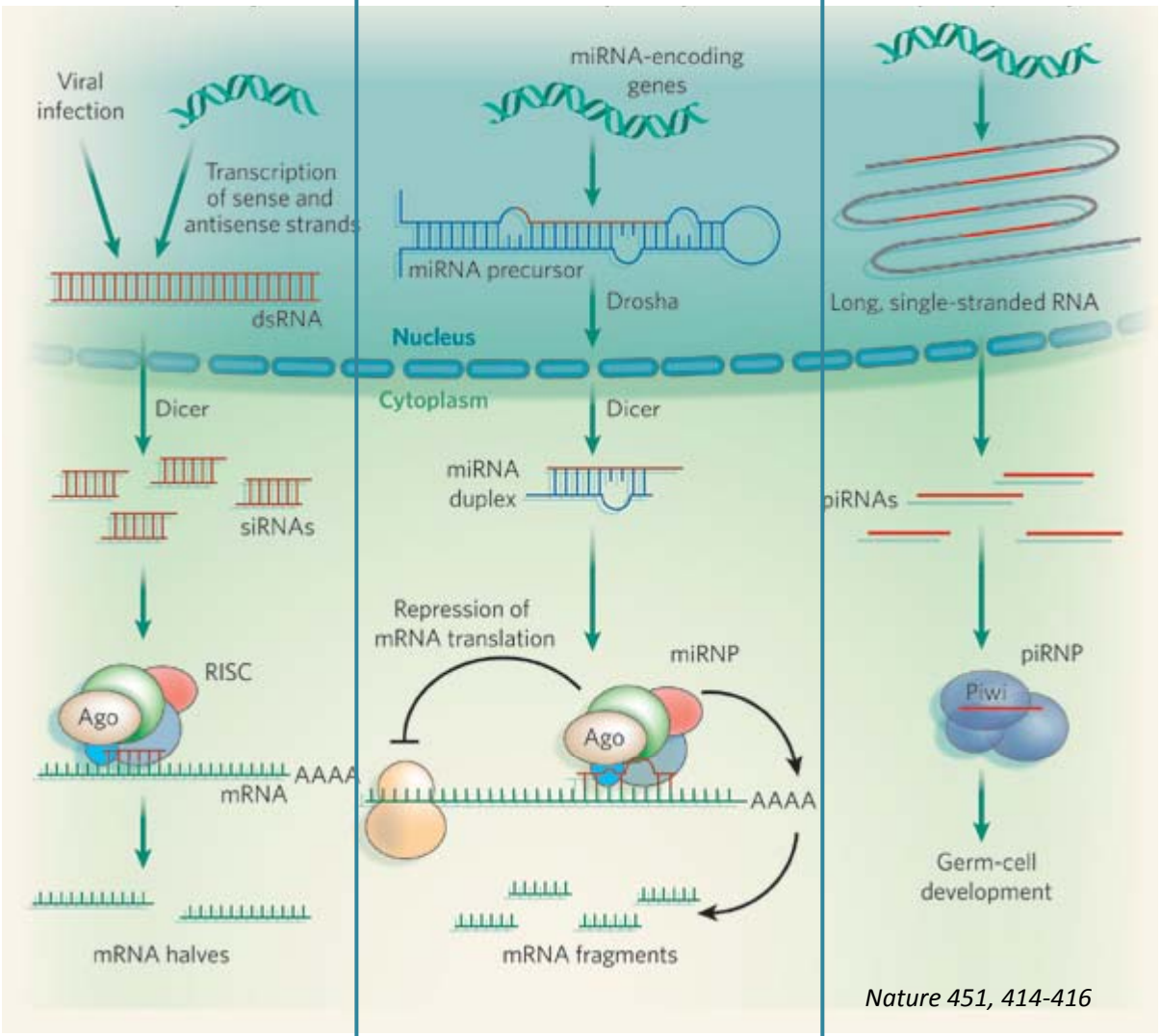


Non-coding Regulatory RNAs

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

Comparison of miRNA study methods

	Northerns	miRAGE	qRT-PCR	Microarrays	NGS
Starting material	5-25 µg	1mg	1-10ng	~5µg	1-10µg
Sensitivity	Low	Low	High	Moderate	High
Specificity	Low	Low	High	Moderate	High
Dynamic Range	~2 logs	presence/ absence	7 logs	< 4 logs	Broad range
Time to results	Several days	Several days	Hours	Several days	3 days
High throughput	No	No	No	**Yes**	**Yes**
Novel miRNA identification	Yes	Yes	Yes	No	**Yes**
Reliably distinguishes mature miRNA from precursor	Yes	No	Yes	No	Yes illumina applied biosystems

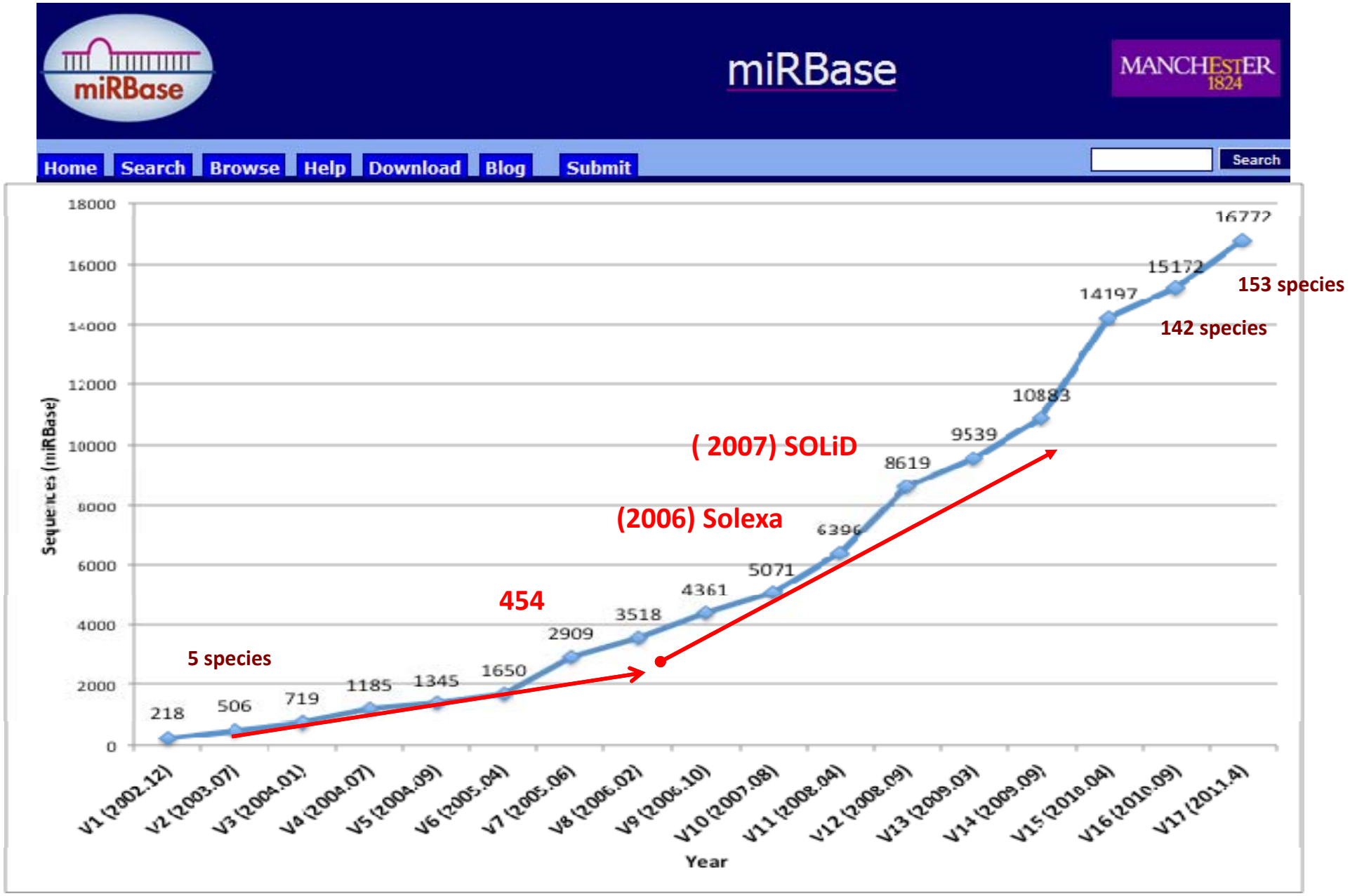
Understand the Reference Database

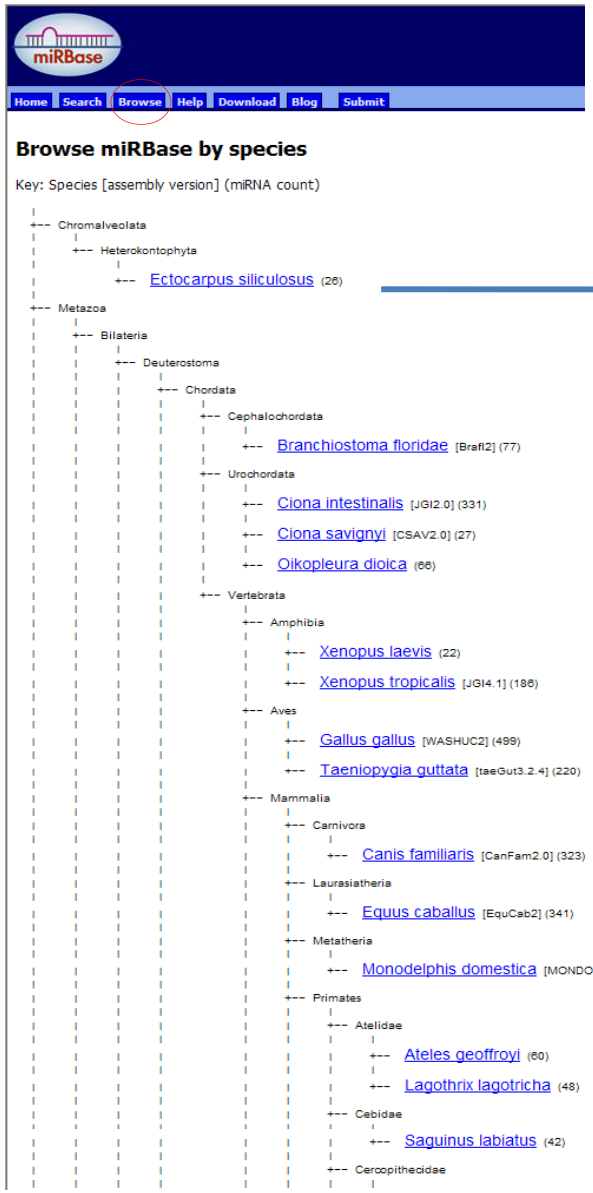
http://www.mirbase.org/



16772 entries
1220 Families
19724 mature miRNAs
12187 unique miRNAs
71 genome coordinate

Nucleic Acids Research, 2010, 1–6


miRBase
MANCHESTER 1824
Home
Search
Browse
Help
Download
Blog
Submit
Search
Sequences (miRBase)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
218
506
719
1185
1345
1650
2909
3518
4361
5071
6396
8619
9539
10883
14197
15172
16772
153 species
142 species
(2007) SOLiD
(2006) Solexa
454
5 species
V1 (2002.12)
V2 (2003.07)
V3 (2004.01)
V4 (2004.07)
V5 (2004.09)
V6 (2005.04)
V7 (2005.06)
V8 (2006.02)
V9 (2006.10)
V10 (2007.08)
V11 (2008.04)
V12 (2008.09)
V13 (2009.03)
V14 (2009.09)
V15 (2010.04)
V16 (2010.09)
V17 (2011.4)
Year



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

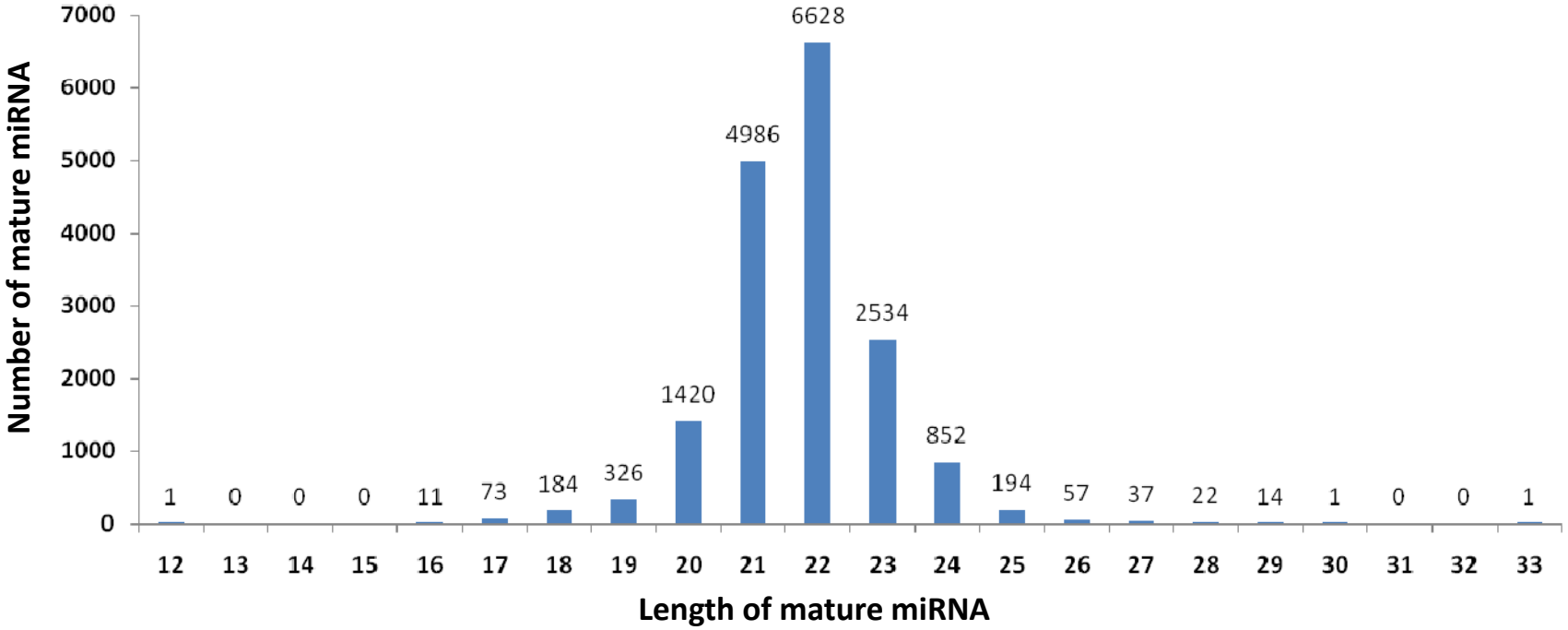
aga.gff	ame.gff	ath.gff	bmo.gff	bta.gff	cbr.gff	cel.gff
cfa.gff	cre.gff	dme.gff	dps.gff	dre.gff	ebv.gff	fru.gff
gga.gff	hcmv.gff	hsa.gff	kshv.gff	mdo.gff	mghv.gff	mml.gff
mmu.gff	osa.gff	ptc.gff	ptr.gff	rno.gff	sme.gff	tae.gff
tni.gff	vvi.gff	xtr.gff	zma.gff			

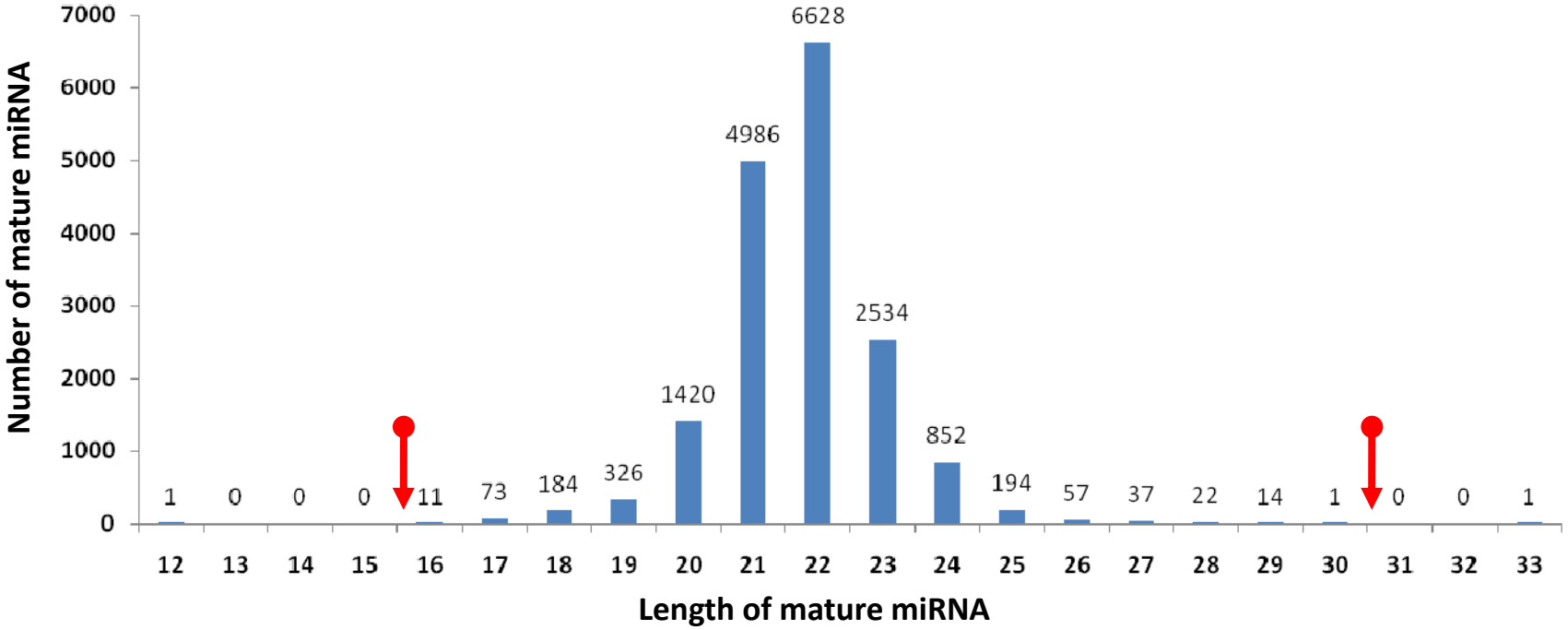
mature.fa

```
>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

Deep Sequencing of small-RNA

Vendor:	Roche			illumina				applied biosystems				
Technology:	454			Solexa GA				SOLiD				
Platform:	GS20	FLX	Ti	I	II	IIx	HiSeq	1	2	3	4	5500
Reads: (M)	0.5	0.5	1.25	28	100	250	2000	40	115	320	700	2400
Library preparation	emPCR			Bridge PCR				emPCR				
NGS chemistry	Pyrosequencing			Reversible terminator				Cleavable probe sequencing by ligation				
Pros	Longer reads improve mapping in repetitive regions; fast run times			The most widely used platform in the field				Two-base encoding provides inherent error correction				
Cons	High error rates in homopolymer repeats			Low multiplexing capability of samples				Long run times				
Fragment												
Read length:	100	200	400	35	50	100	100	25	35	50		75
Run time: (d)	0.25	0.3	0.4	3	3	5	<4.5	6	5	8		
Yield: (Gb)	0.05	0.1	0.5	1	5	25	119	1	4	16	50	
Rate: (Gb/d)	0.2	0.33	1.25	0.33	1.67	5	25	0.34	1.6	4	6.2	

Analysis Tools for small-RNA Deep Sequencing

LINUX/UNIX BASED TOOLS

miRDeep Discovering microRNAs from deep sequencing data
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

MiRNAkey: A software pipeline for the analysis of microRNA Deep Sequencing data.
Bioinformatics (2010) Aug 27

WEB-BASED TOOLS

miRCat: A toolkit for analysing large-scale plant small RNA datasets.
Bioinformatics (2008) 24 (19): 2252-2253 Application Note

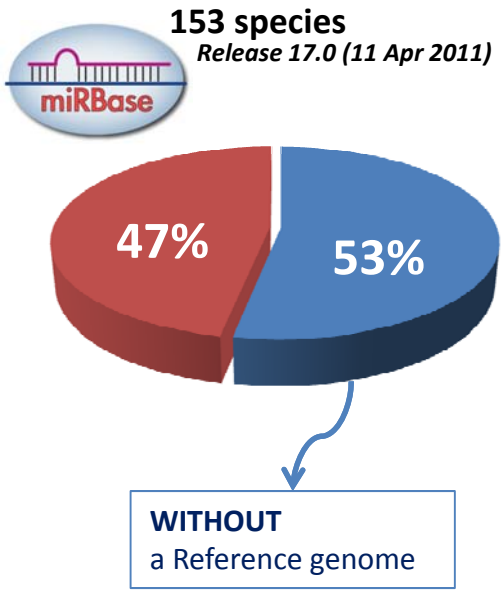
miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments
Nucleic Acids Research (2009) 37 (SUPPL. 2), pp. W68-W76

DSAP: deep-sequencing small RNA analysis pipeline
Nucleic Acids Research. (2010) 38(Web Server issue): W385–W391.

mirTools: microRNA profiling and discovery based on high-throughput sequencing
Nucleic Acids Research. (2010) 38(Web Server issue): W392–W397

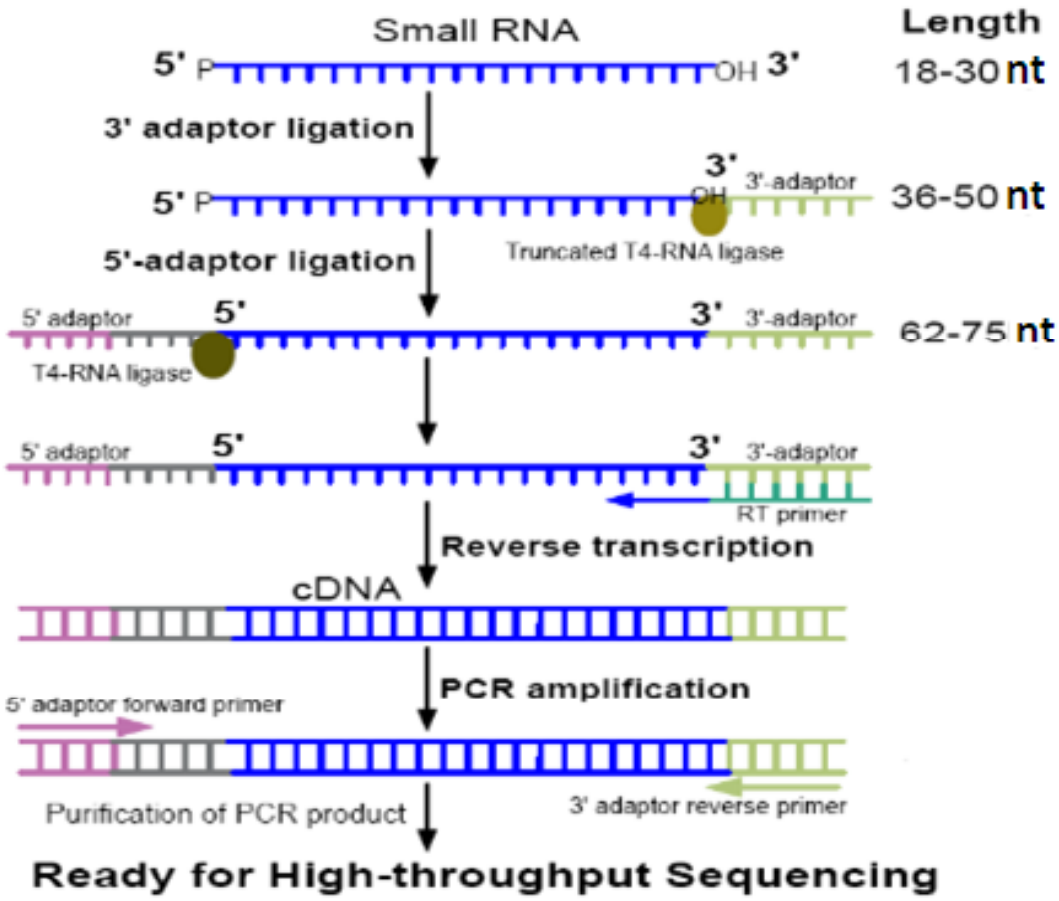
miRanalyzer2: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments
Nucleic Acids Research. (2011) 39 (11)

Comparison of Tools

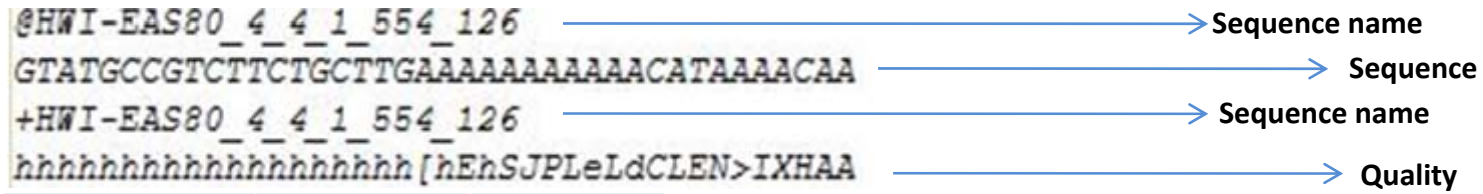


Our goal	miRanalyzer	miRCat	miRExpress	miRDeep
Web based	✔	✔		
No-limit on Organisms			✔	
Classification of small RNA				
Classification of miRNA	✔	✔	✔	
isomiR alignment	✔			
Differential miRNA expression		✔	✔	
Phylogenic distribution of miRNAs				
New miRNA prediction	✔			✔
Target prediction	✔			

Sample Preparation for small RNA NGS Sequencing



Data Processing

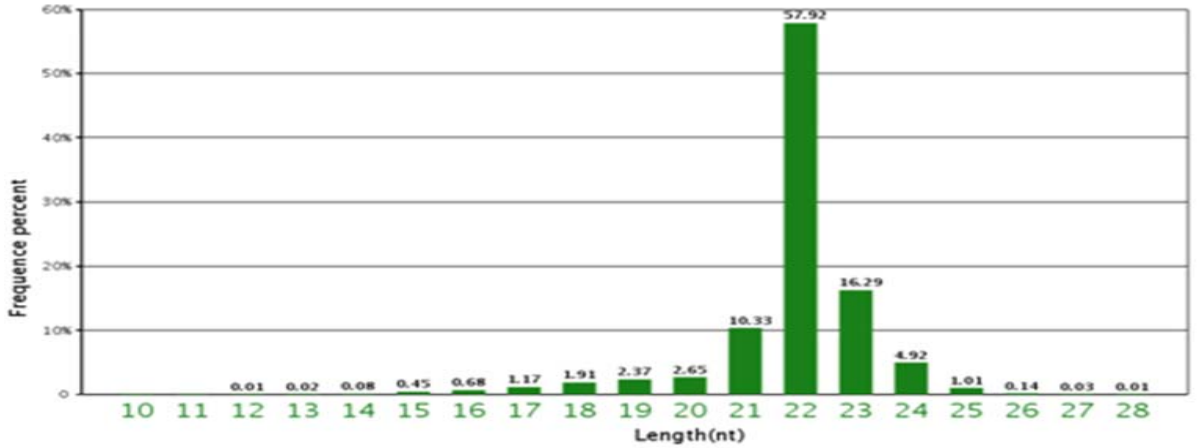


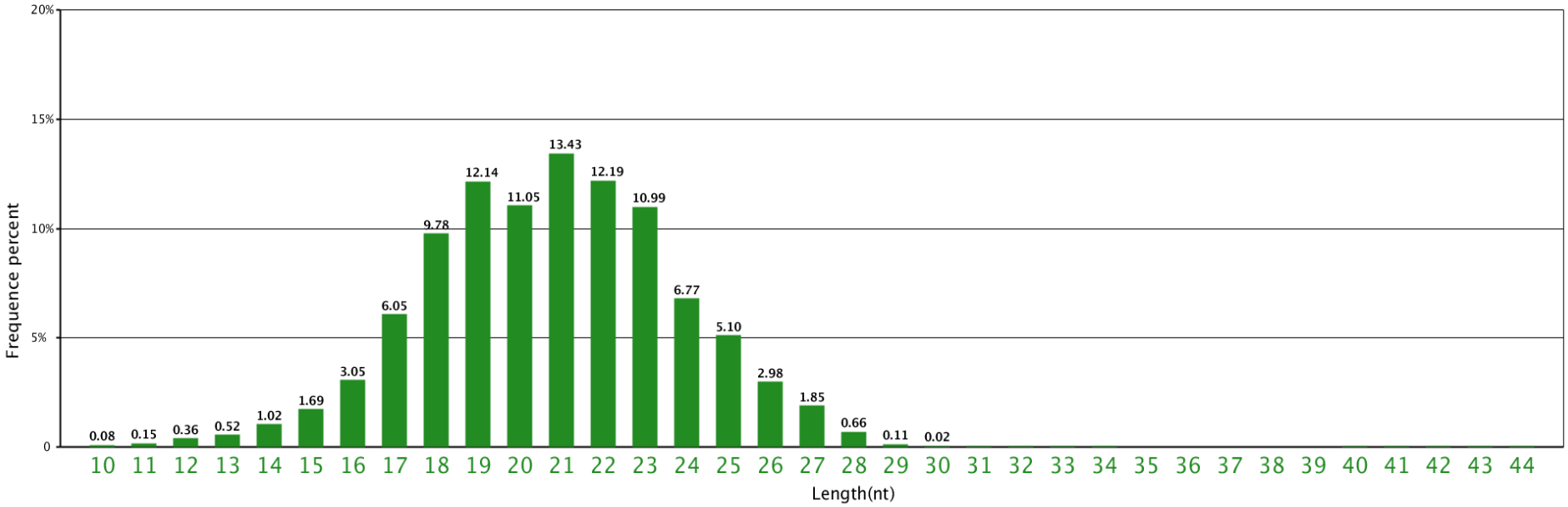
Sequencing error rate(E)	Sequencing quality	Corresponding character
5%	13	M
1%	20	T
0.1%	30	^
0.01%	40	h

The data is processed by the following steps:

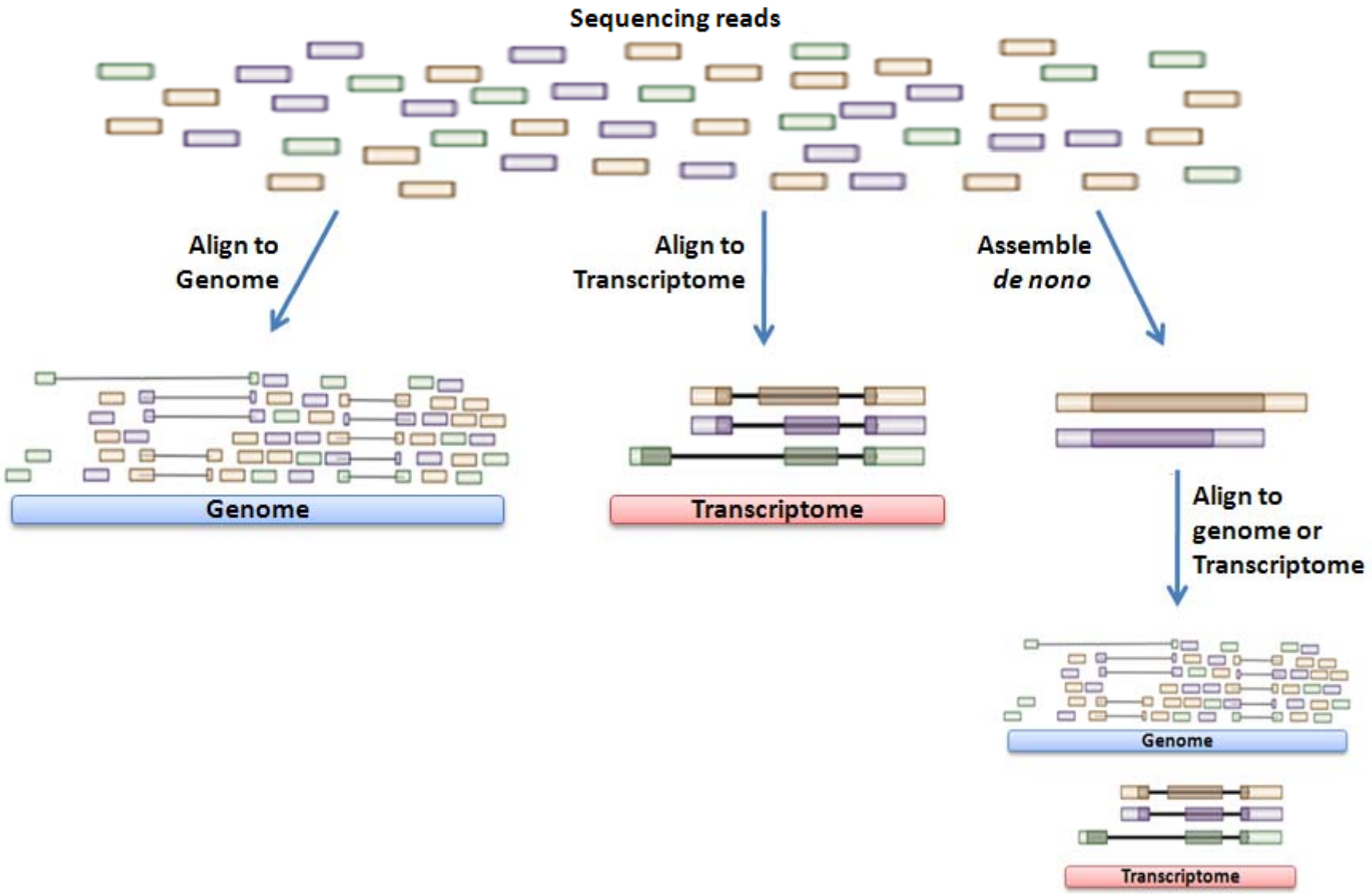
1) Getting rid of low quality reads
2) Getting rid of reads with 5' primer contaminants
3) Getting rid of reads without 3' primer
4) Getting rid of reads without the insert tag
5) Getting rid of reads with poly A/G/C/T/N
6) Getting rid of reads shorter than 16-18nt ?
7) Summarize the length distribution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

Which one is the correct distribution?





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

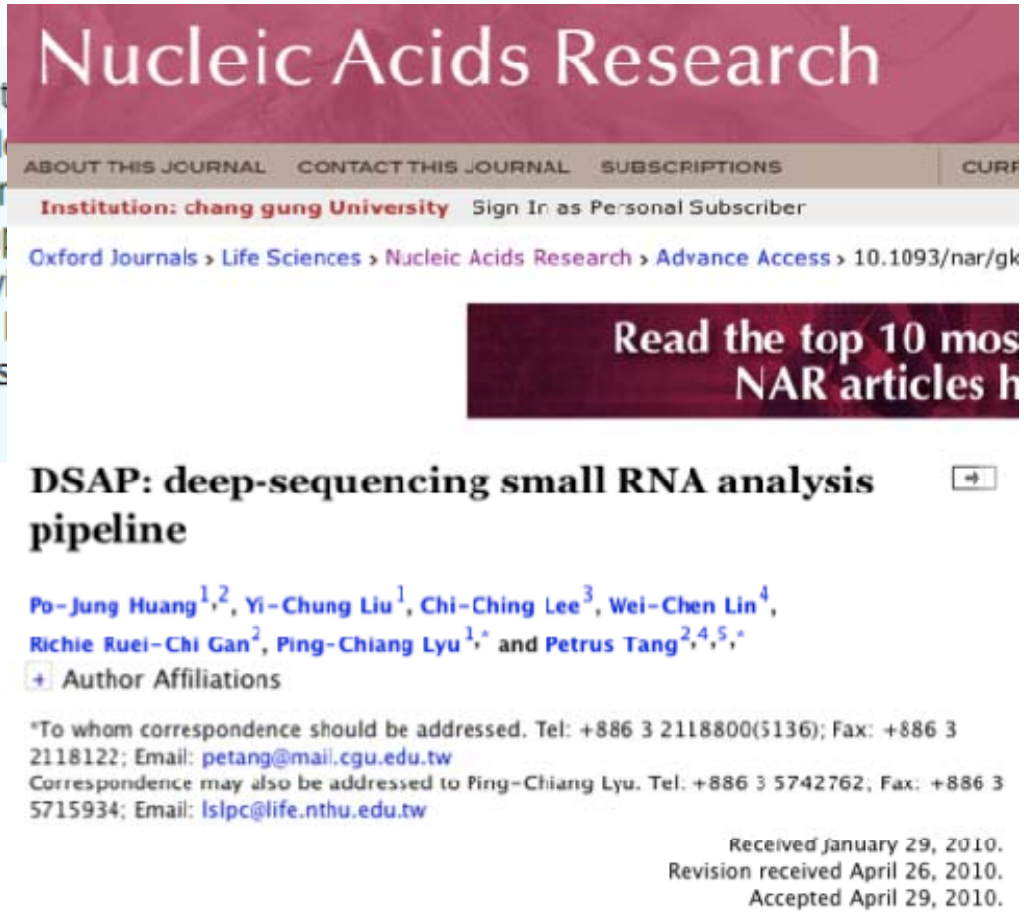
Assembly

Low expression = difficult/impossible to assemble

Misassemblies/fragmented contigs due to repeats

Requires vast amounts of memory





http://dsap.life.nthu.edu.tw
http://dsap.cgu.edu.tw

Implementation



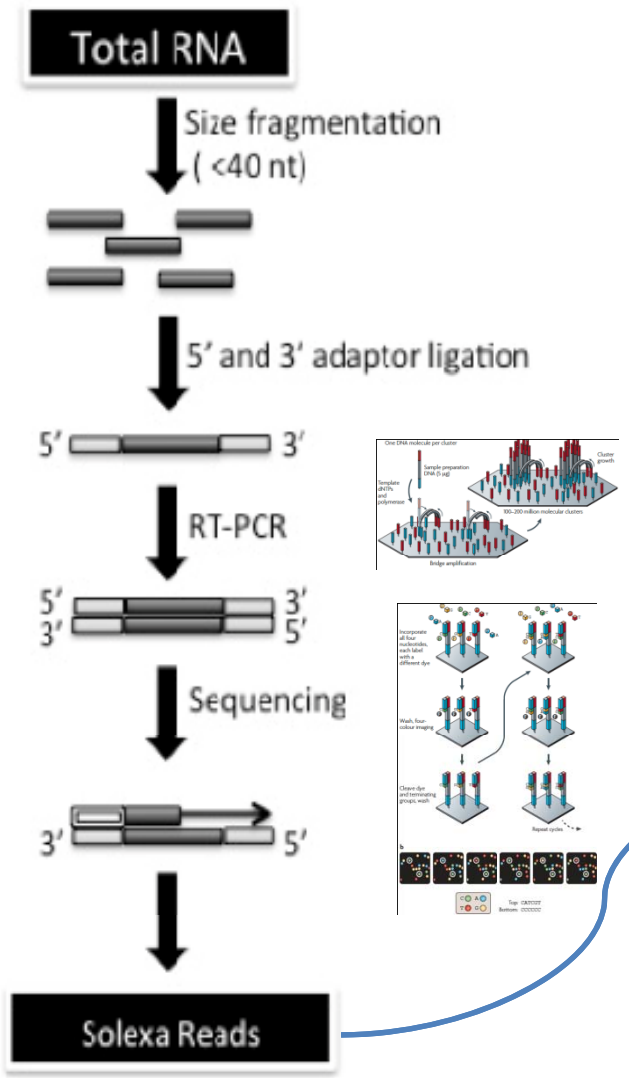
DSAP runs on a Linux CentOS 64-bit server housing two quad-core Intel® Xeon® 5300 Series Processors and 16 GB RAM.

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
```

NGS data format



Solexa Reads(FASTQ) 35~40+

Line1: **@HWI-EAS82_3_FC204V1AAXX:2:1:900:453**
Line2: **CGAGGGCCTGGGTTCGAACCCCAGAGTTCGTATGC**
Line3: **+HWI-EAS82_3_FC204V1AAXX:2:1:900:453**
Line4: **hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh**

```
@HWI-EAS82_3_FC204V1AAXX:2:1:752:642
CGAGTTCTACAGTCCGACGATTCGTATGTCGTCTT
+HWI-EAS82_3_FC204V1AAXX:2:1:752:642
hhhhhhhhhhhhhhAhLhhUhhhhghhVChhhdhh
@HWI-EAS82_3_FC204V1AAXX:2:1:890:394
AGTTCTACAGTCCGACGGATCTCGTATGCCGTCTT
+HWI-EAS82_3_FC204V1AAXX:2:1:890:394
hhhhhhhhhhhhhhAhLhhUhhhhghhVChhhdhh
```

.
.
.
.
.
.

Line1: sequence identifier
Line2: raw sequence letters
Line3: sequence identifier
Line4: quality values

FASTQ converting and Quality filtering

10,000,000 reads (1.2 GB)

```
@HWI-EAS82_3_FC204V1AAXX:2:1:900:453
CGAGGGCCTGGGTTCGAACCCCAGAGTTCGTATGC
+HWI-EAS82_3_FC204V1AAXX:2:1:900:453
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS82_3_FC204V1AAXX:2:1:752:642
CGAGTTCTACAGTCCGACGATTCGTATGTCGTCTT
+HWI-EAS82_3_FC204V1AAXX:2:1:752:642
hhhhhhhhhhhhhhAhLhhUhhhhghhVChhhdhh
@HWI-EAS82_3_FC204V1AAXX:2:1:890:394
AGTTCTACAGTCCGACGGATCTCGTATGCCGTCTT
+HWI-EAS82_3_FC204V1AAXX:2:1:890:394
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS82_3_FC204V1AAXX:2:1:892:379
AGTTCTACAGTCCGACGATCTCGTATGCCGTCTTC
+HWI-EAS82_3_FC204V1AAXX:2:1:892:379
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS82_3_FC204V1AAXX:2:1:898:486
GGGGCTGAAGCATAATGGCATTGCTGTCGTATGCC
+HWI-EAS82_3_FC204V1AAXX:2:1:898:486
hhhhhhhhhhhhhhhhhhhhhhhhhRhh`hhhAhU
```

Emboss 6.1p1
CLC Genomics Workbench
Biopython
BioPerl
BioRuby
FASTQ-Groomer
Biopieces
FASTX

(15 MB)

Counts	Sequence tag
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	AACATTCATTGCTGTCGGTGGGTTTCGTATGCA

FASTQ variants

FASTQ-Sanger NCBI-SRA
FASTQ-Solexa
FASTQ-illumina 1.3+
FASTQ-illumina 1.5+

Nucleic Acids Research 2009, Volume38, Issue6 pp.1767-1771

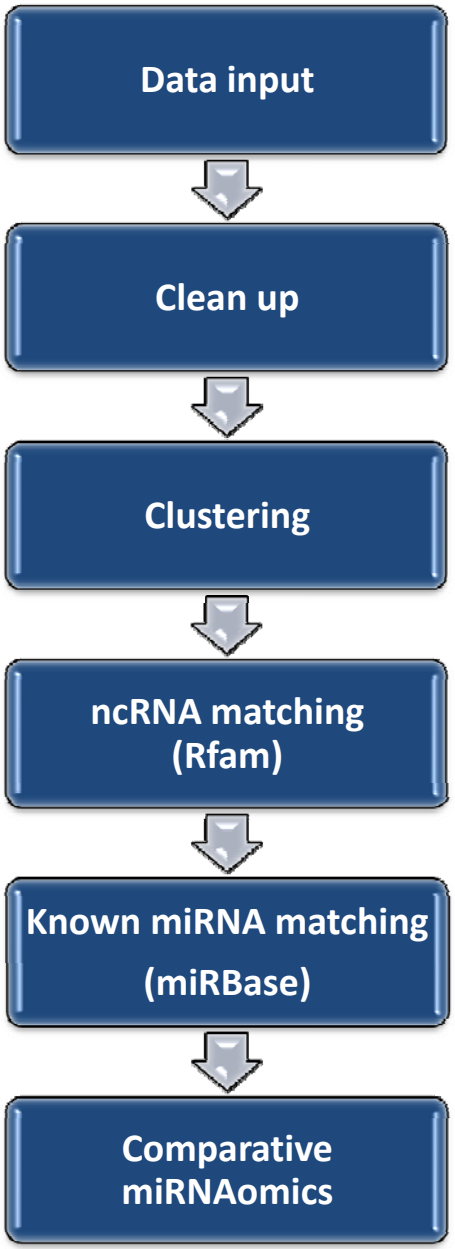
$Q_{phred} = -10 \log_{10}(e)$

40 = -10 $\log_{10}$(1/10000)
30 = -10 $\log_{10}$(1/1000)
20 = -10 $\log_{10}$(1/100)

e is the estimated probability of a base being wrong

FASTQ-Sanger
FASTQ-illumina 1.3+
FASTQ-illumina 1.5+

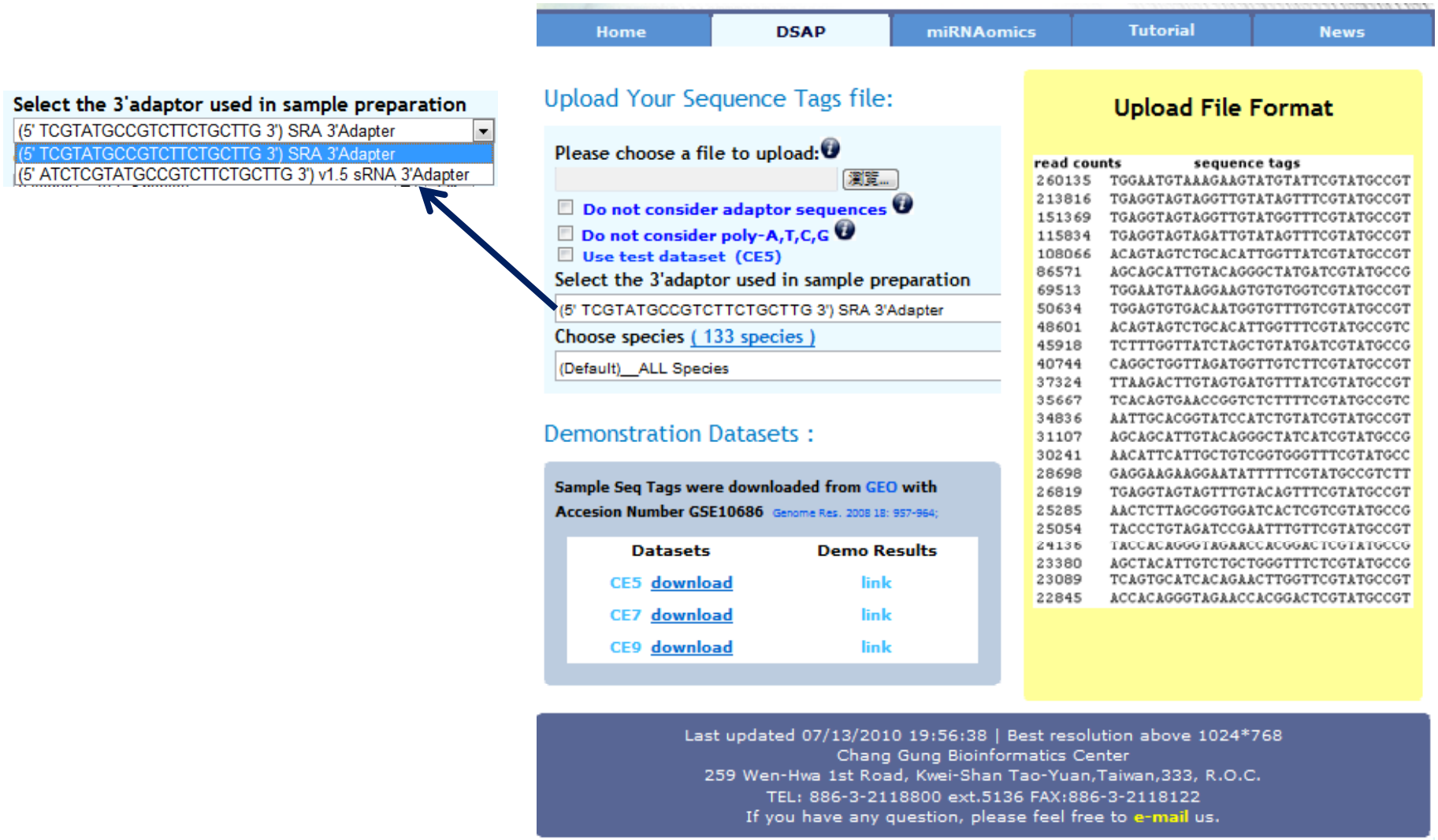
Workflow



- Sequence tags
- 3'adaptor trimming
- 5' adaptor filtering
- Poly A, T, C, G/N filtering
- Unique tags

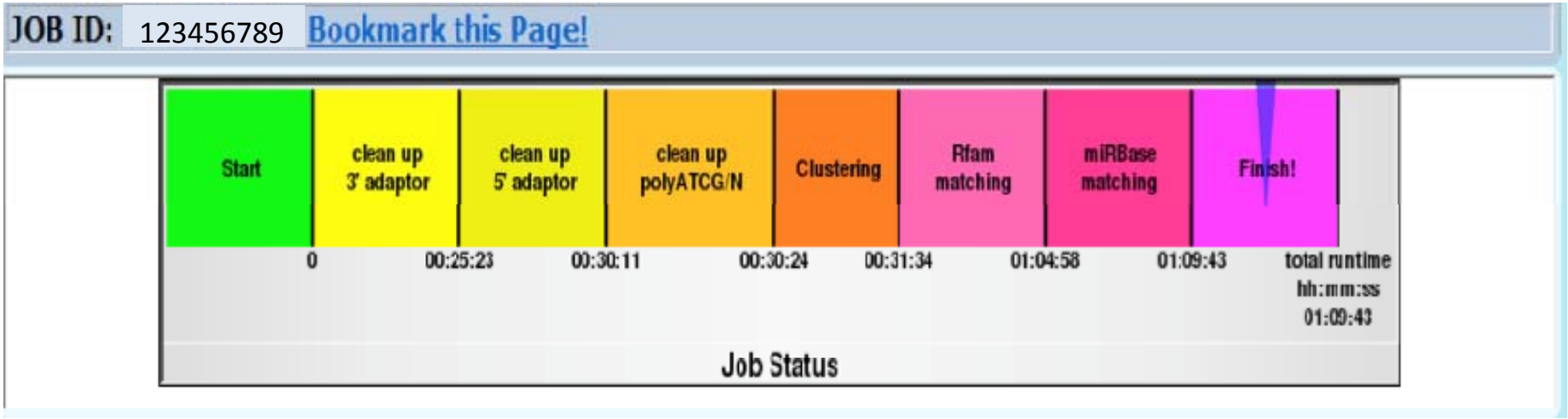
Data input

- DSAP Data Input Page



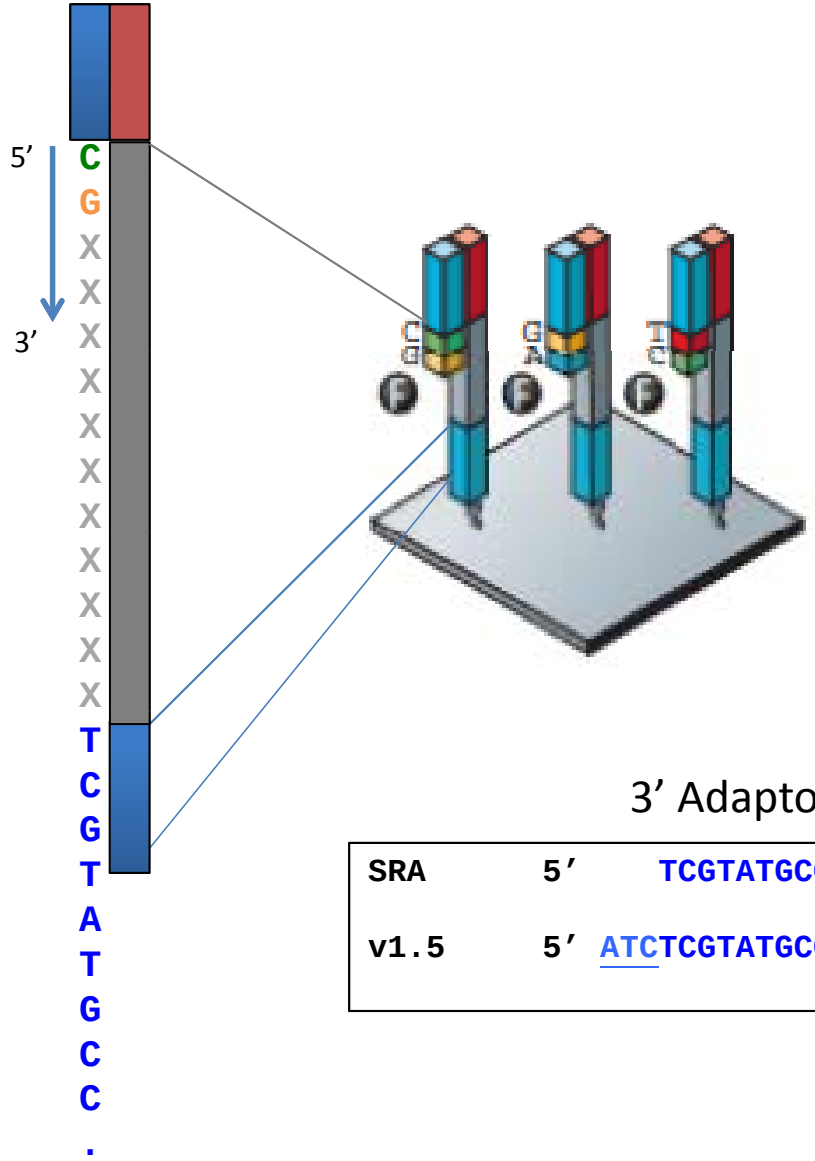
Data input

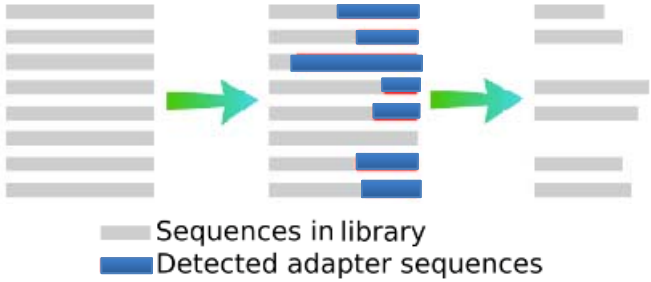
- Job status



Clean up

- 3’ adaptor trimming



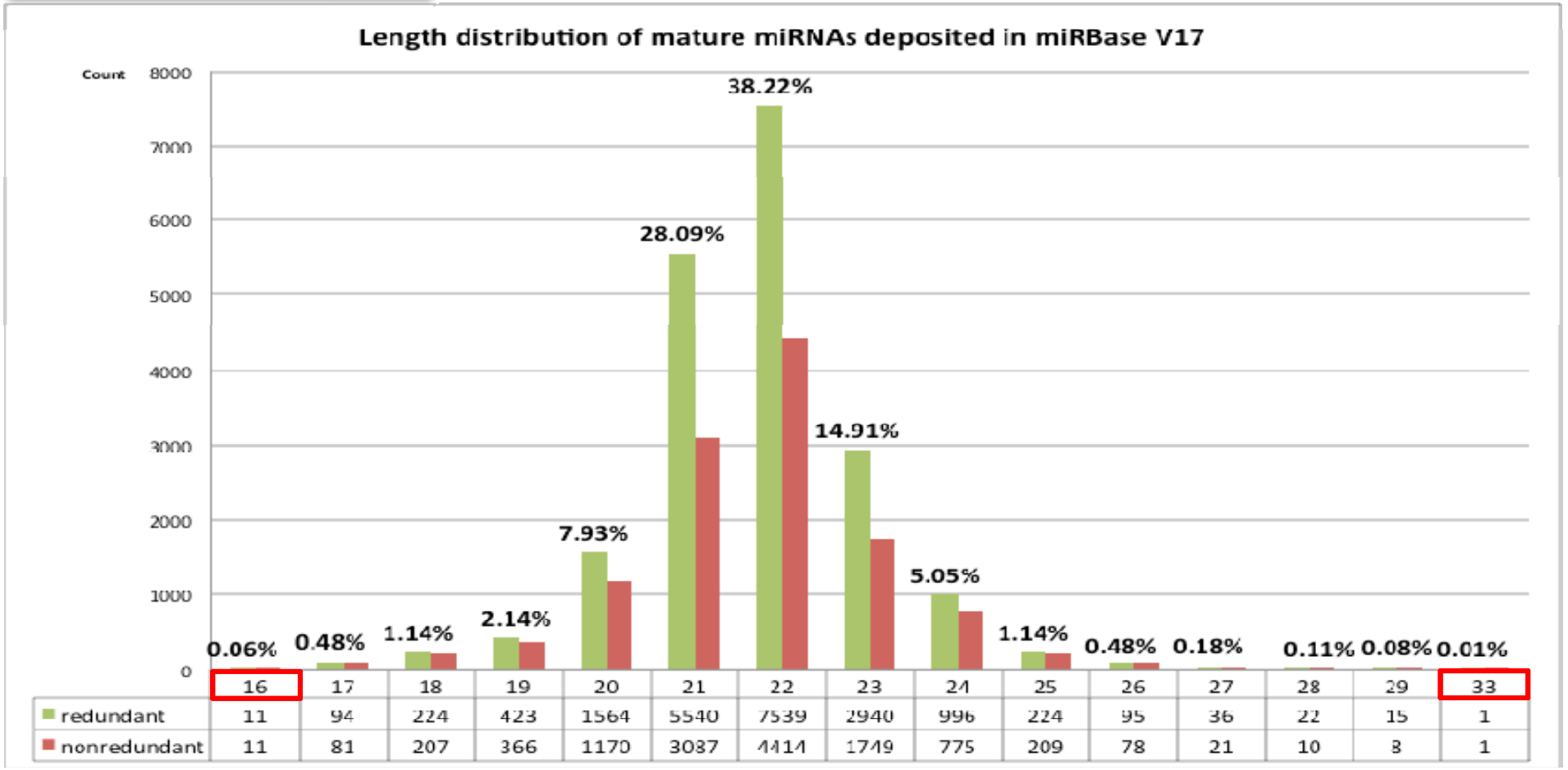


3’ Adaptor sequences

SRA	5’	TCGTATGCCGTCTTCTGCTTG 3’ (21nt)
v1.5	5’	ATCTCGTATGCCGTCTTCTGCTTG 3’ (24nt)

Clean up

- 3’ adaptor trimming



	16	17	18	19	20	21	22	23	24	25	26	27	28	29	33
redundant	11	94	224	423	1564	5540	7539	2940	996	224	95	36	22	15	1
nonredundant	11	81	207	366	1170	3087	4414	1749	775	209	78	21	10	8	1

```
(Discard)           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX    X*35

(Keep)              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXTCGTA   X*30
(Keep)              XXXXXXXXXXXXXXXXXXXXXTCGTATGCCGTCT   X*22
(Keep)              XXXXXXXXXXXXXXXTCGTATGCCGTCTTCTGCT   X*16
(Discard <16nt)     XXXXXXXXXXXXXXTCGTATGCCGTCTTCTGCTT   X*15
(?)                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTC   X*33
```

Clean up

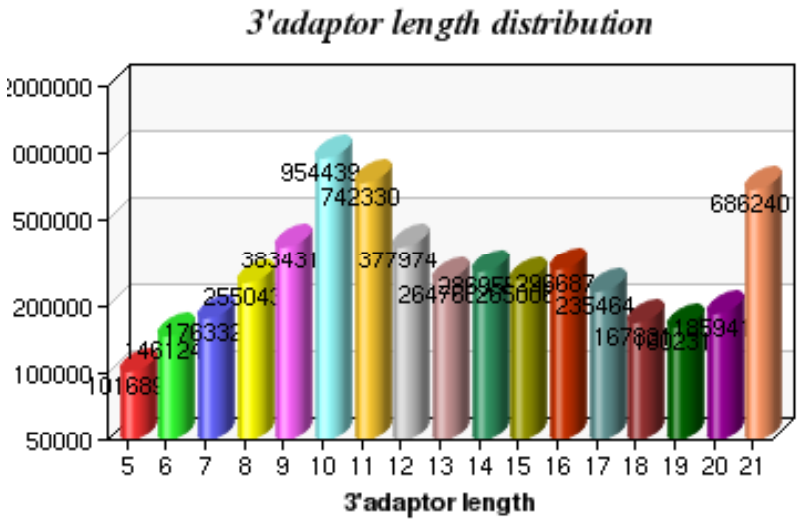
- 3’ adaptor trimming

- **BLASTN**
  - **Word size : default = 11 for nucleotides**

- **Smith-Waterman**
  - **Water (Emboss)**
  - **CLC Cube**



- **Supermatcher (Emboss)**
  - **Word-match + Smith-Waterman**
  - **40,000,000 reads (10,000,000 tags)**
    → 20 min



Clean up

➢ 5’ adaptor filtering

➢**Supermatcher (Emboss)**

```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX  X*35
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTCGTA
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX  max: X*30
XXXXXXXXXXXXXXXX                    min: X*16
```

(Discard) Detected 5’adaptor length >=16

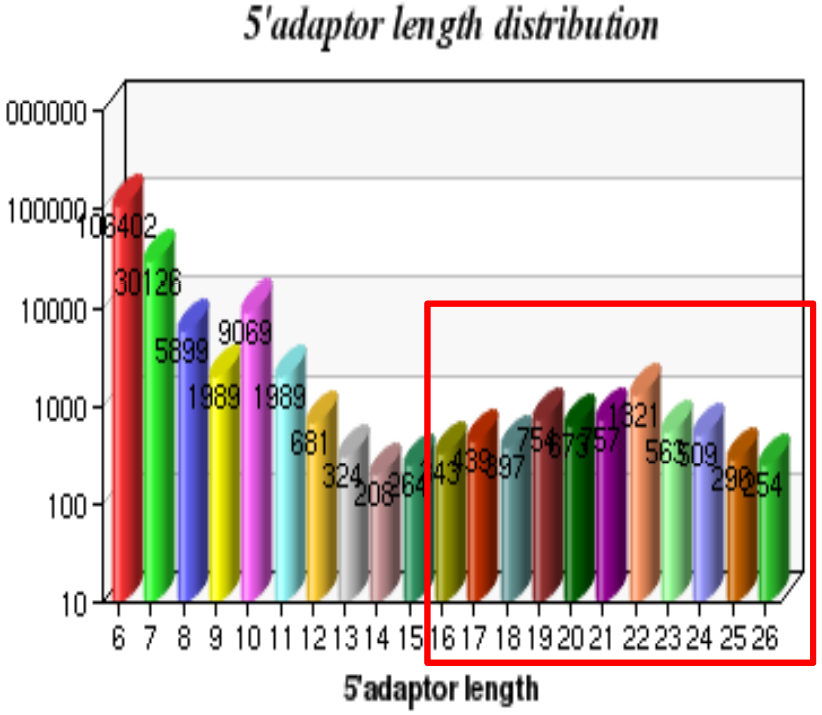
CTACAGTCCGACGATC

GTTCAGAGTTCTACAG

GTTCAGAGTTCTACAGTCCGACGATC



“ACGATC”
can map to 42 miRNAs



5’Adaptor sequence

5’GTTCAGAGTTCTACAGTCCGACGATC 3’ (26nt)

Clean up

- Poly-A/T/C/G and N filtering

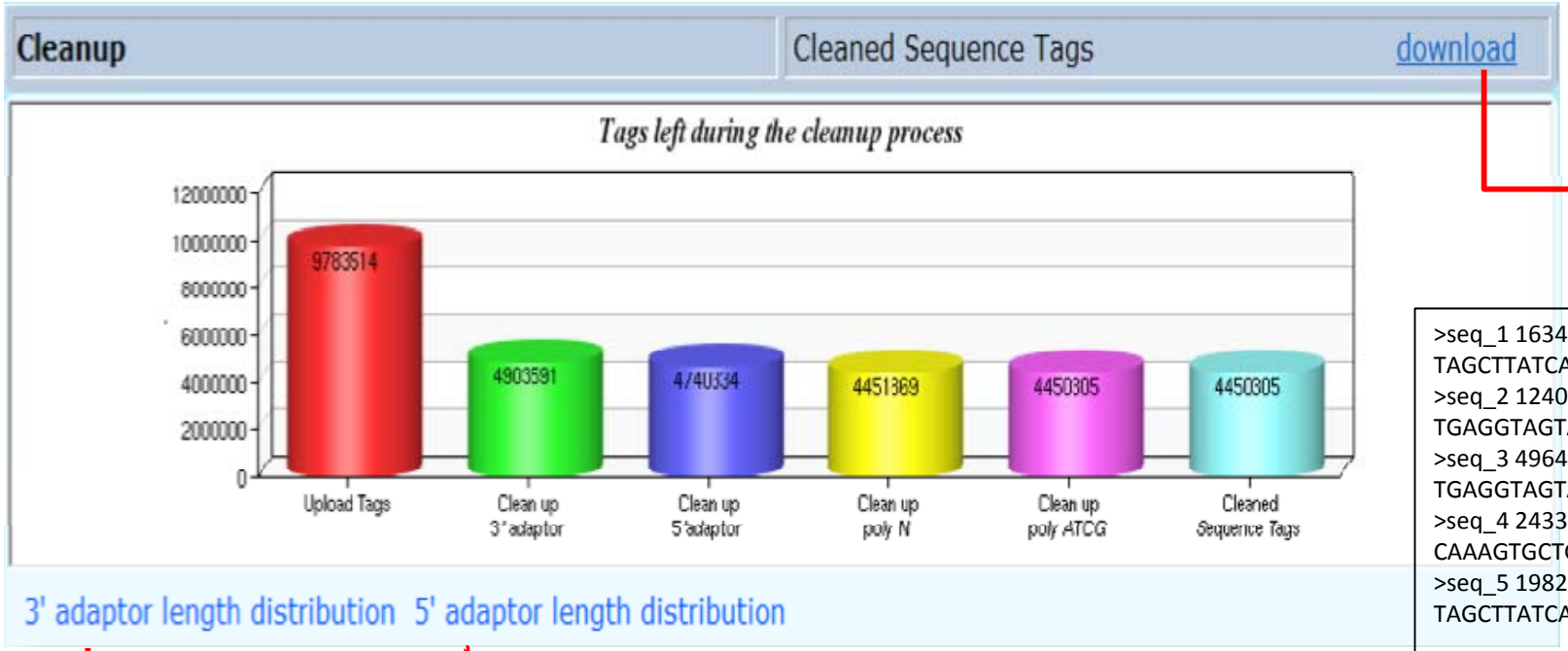
GGAGGGAGG**AAAAAAAAAAA**	gga-miR-1599	Gallus gallus
AAAAAAAAAAACUCCCCCCC	bta-miR-2391	Bos taurus
UGUUGUAC**UUUUUUUUUU**GUUC	hsa-miR-3613-5p	Homo sapiens
AA**GGGGGGGGGGGG**AAAGA	osa-miR2919	Oryza sativa
GAGGG**CCCCCCCC**AAUCCUGU	ssc-miR-296	Sus scrofa

A*11
U*10
G*12
C*7

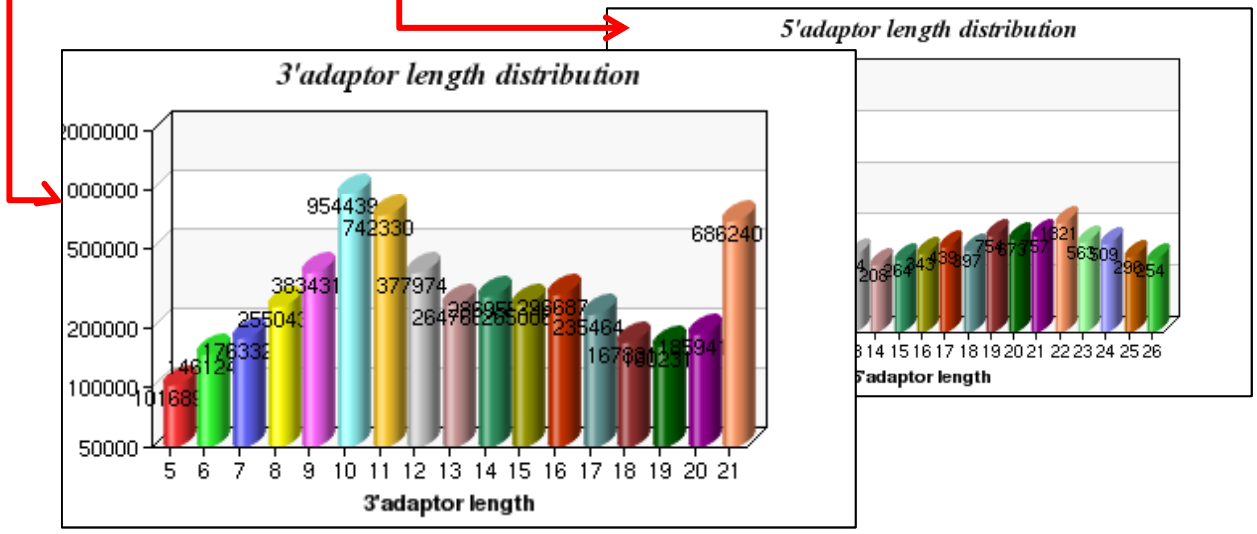
+2~3

Clean up

➢ OUTPUT



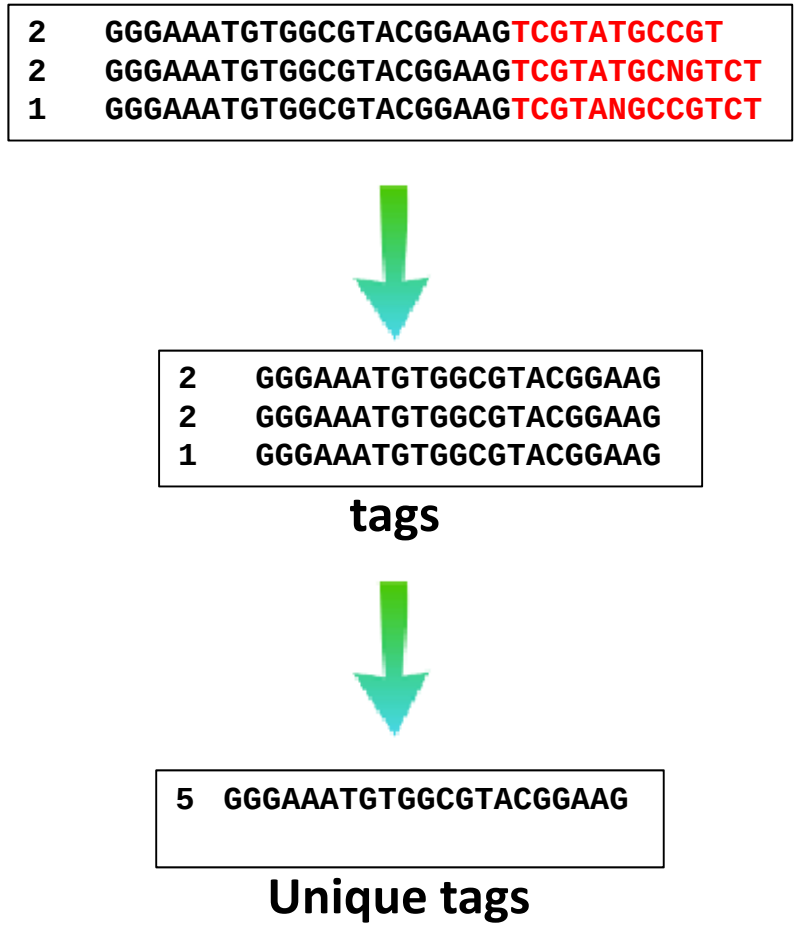

```
>seq_1 1634277
TAGCTTATCAGACTGATGTTGAC
>seq_2 1240975
TGAGGTAGTAGGTTGTATAGTT
>seq_3 496442
TGAGGTAGTAGGTTGTGTGGTT
>seq_4 243345
CAAAGTGCTGTTCGTGCAGGTAG
>seq_5 198215
TAGCTTATCAGACTGATGTTGA
                      :
                      :
```



Clustering

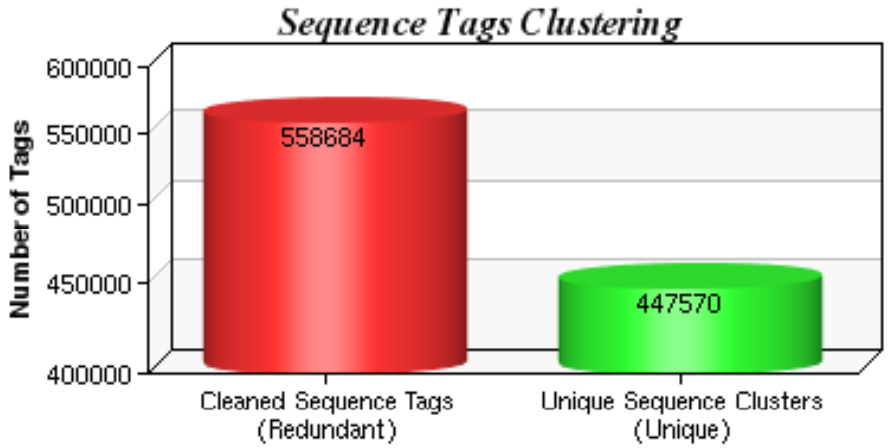
- Clustering software

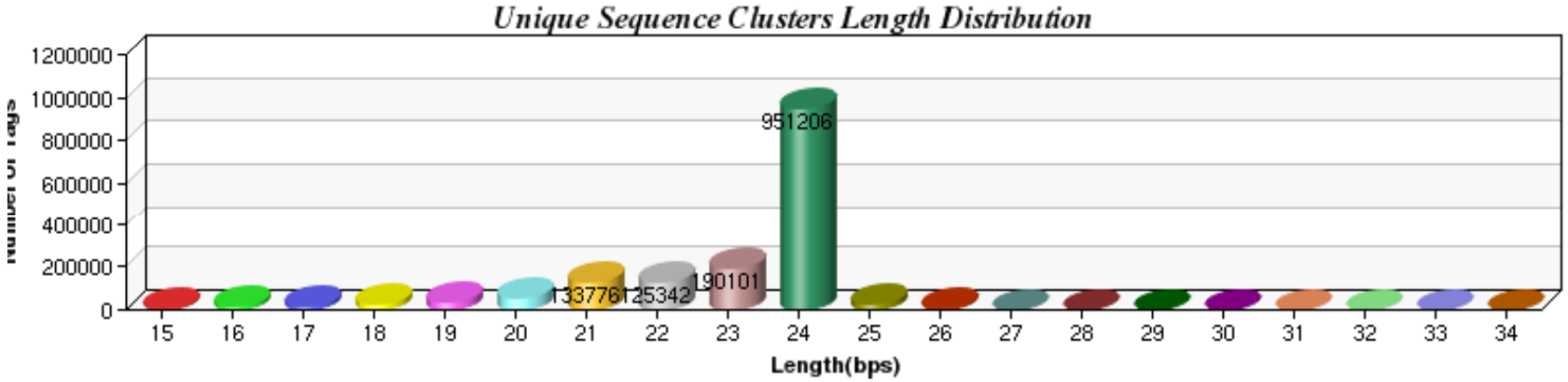
- **CD-HIT**
- **BlustClust**
- **Perl script**



Clustering

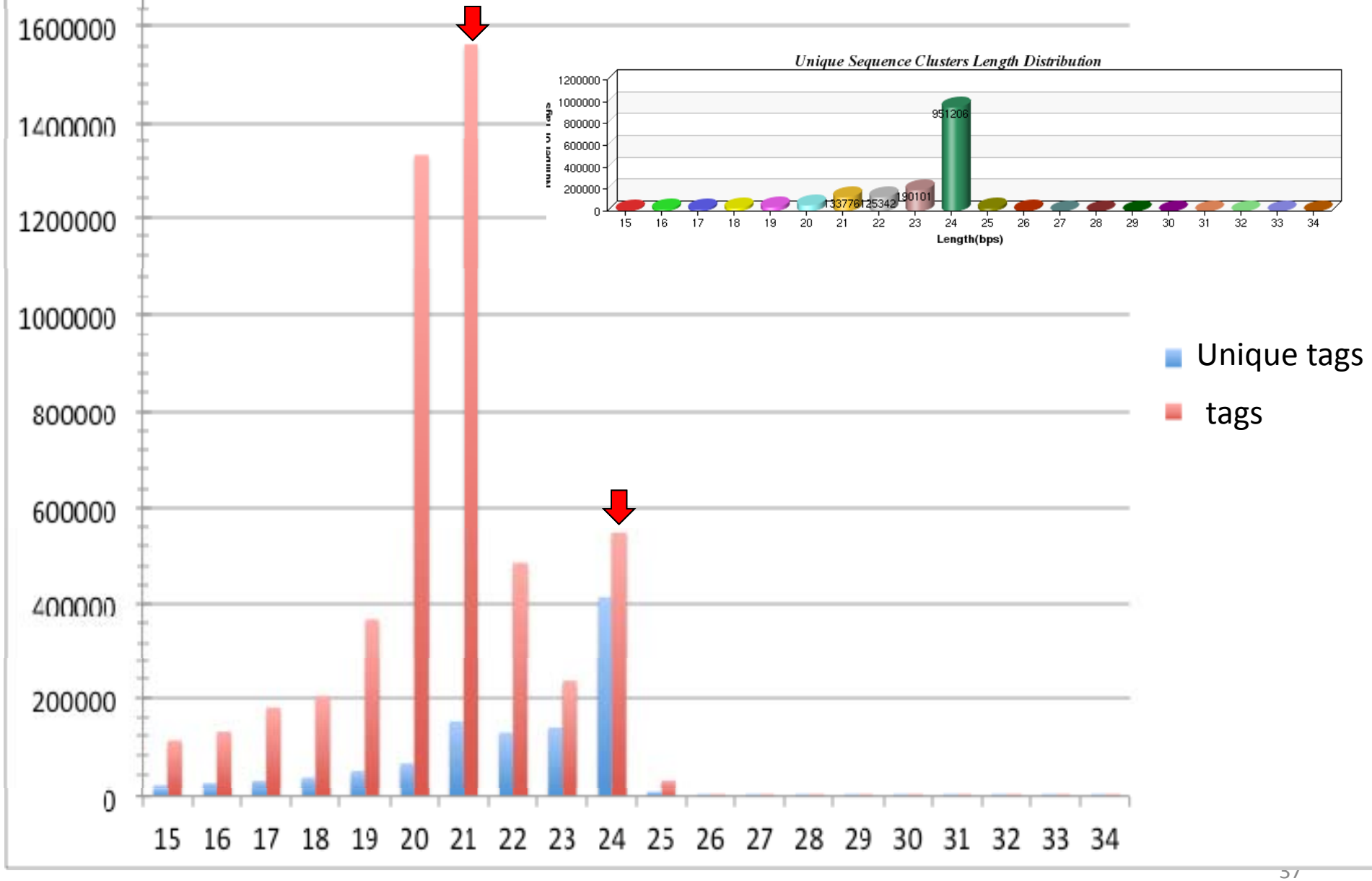
- OUTPUT





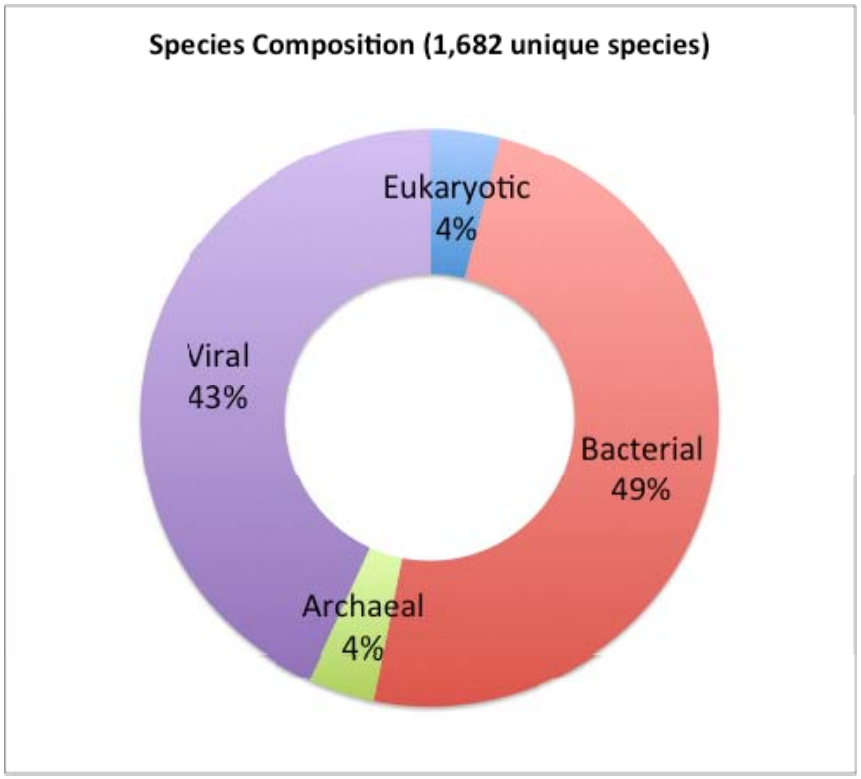
Clustering

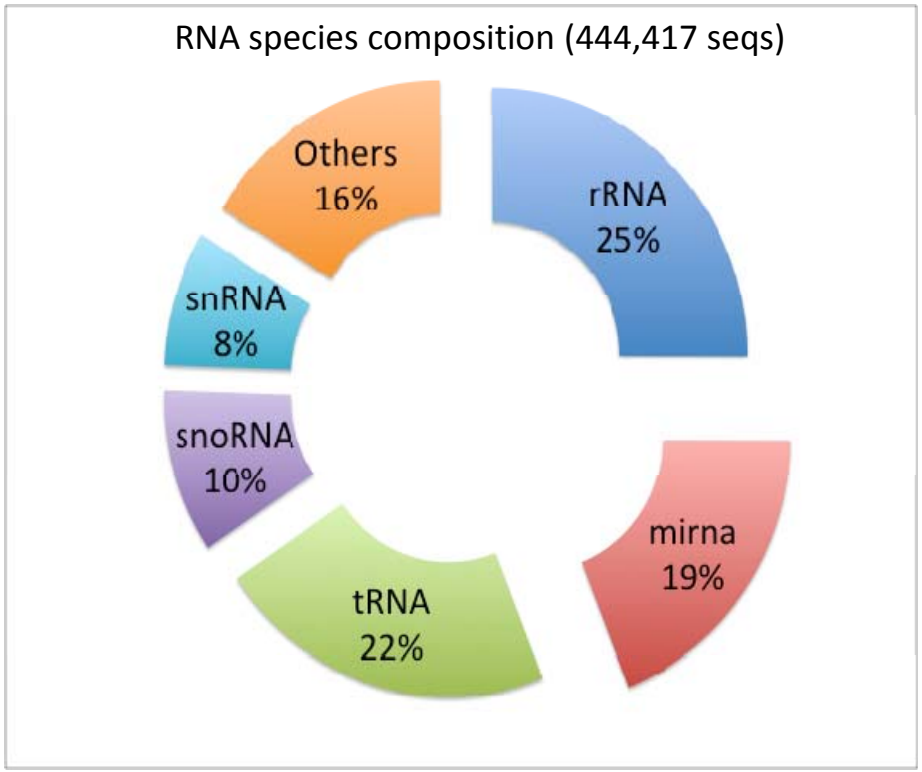
- A suggestion from our user



ncRNA matching

- Non-coding RNA database





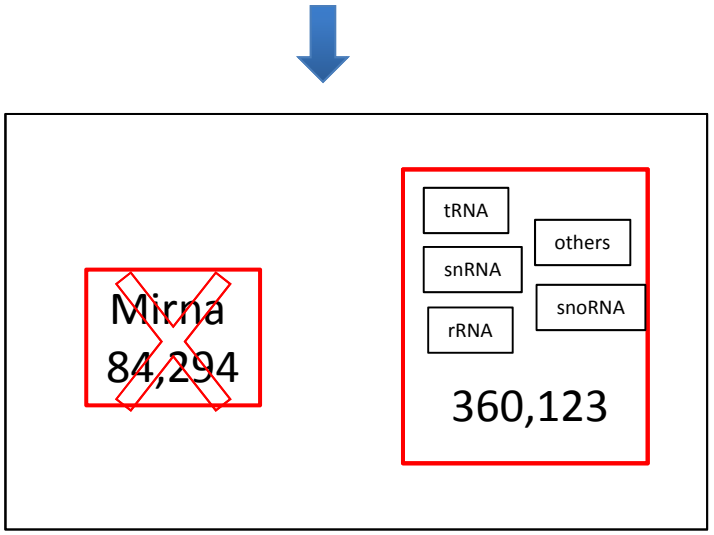
ncRNA matching

- Alignment parameter



- **BLASTN**

Blastall –F F –W 16
Allowing 2 nt mismatches



- **Rfam 9.1_mir_removed 170,020**
- **Rfam_10_mir_removed 360,123**
  (Bottleneck)

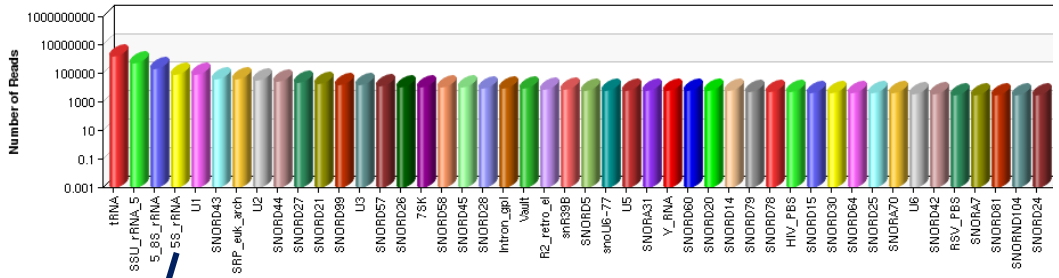
- SOAP
- BOWTIE
- BWA

ncRNA matching

➢ OUTPUT

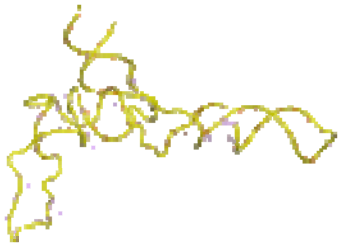
Sort by Expression level, Name; Download

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



wellcome trust sanger institute

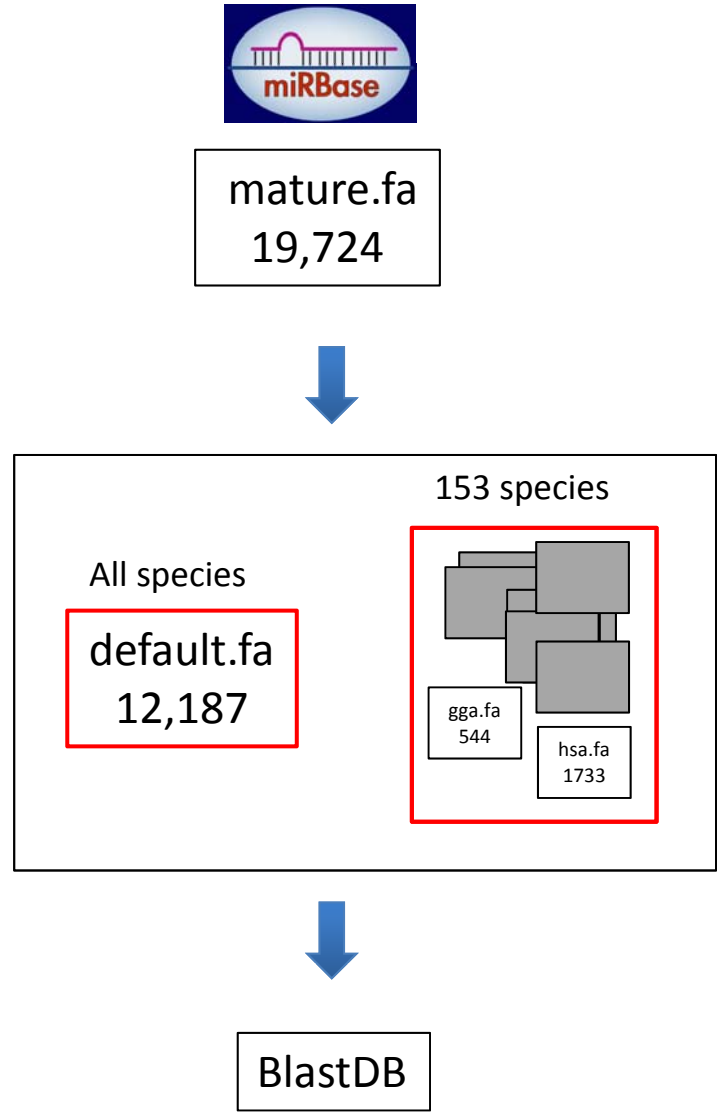
Family: *5S_rRNA* (RF00001)
Description: *5S ribosomal RNA*

Known miRNA Matching

- **BLASTN**

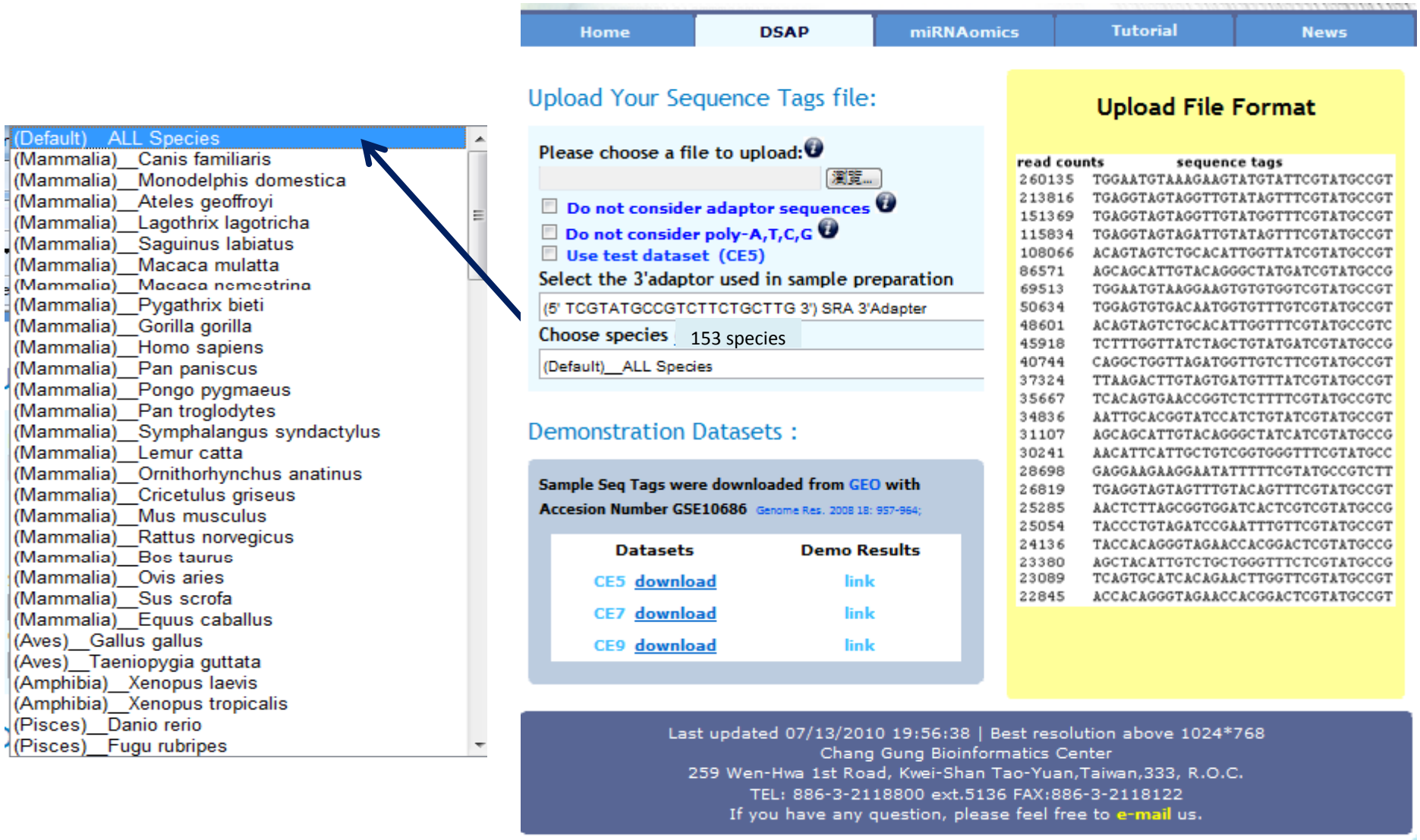
  Blastall –F F –W 16

- 100% sequence identity and cover full length of known miRNA are retained
  - Alignment parameters
  - Blastdb preparation



Known miRNA Matching

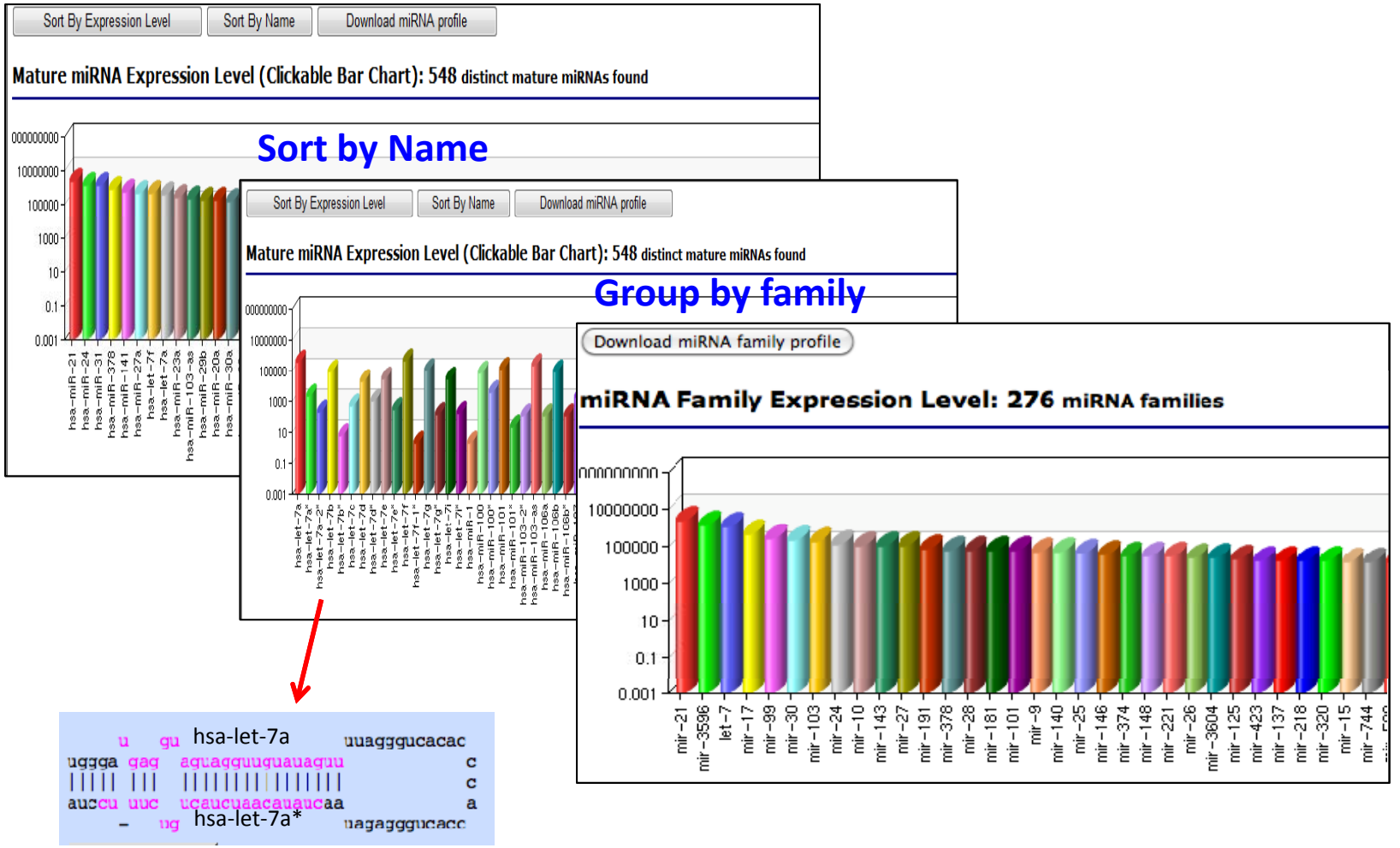
- DSAP Data input page



Known miRNA Matching

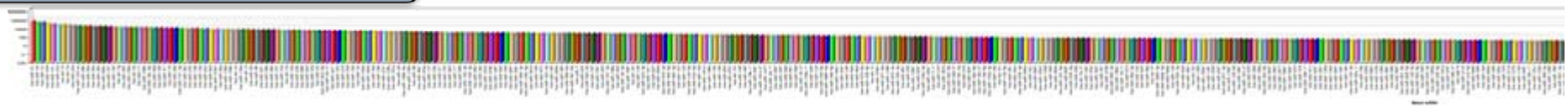
➢ OUTPUT

Sort by Expression level



Known miRNA Matching

- isomiR



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

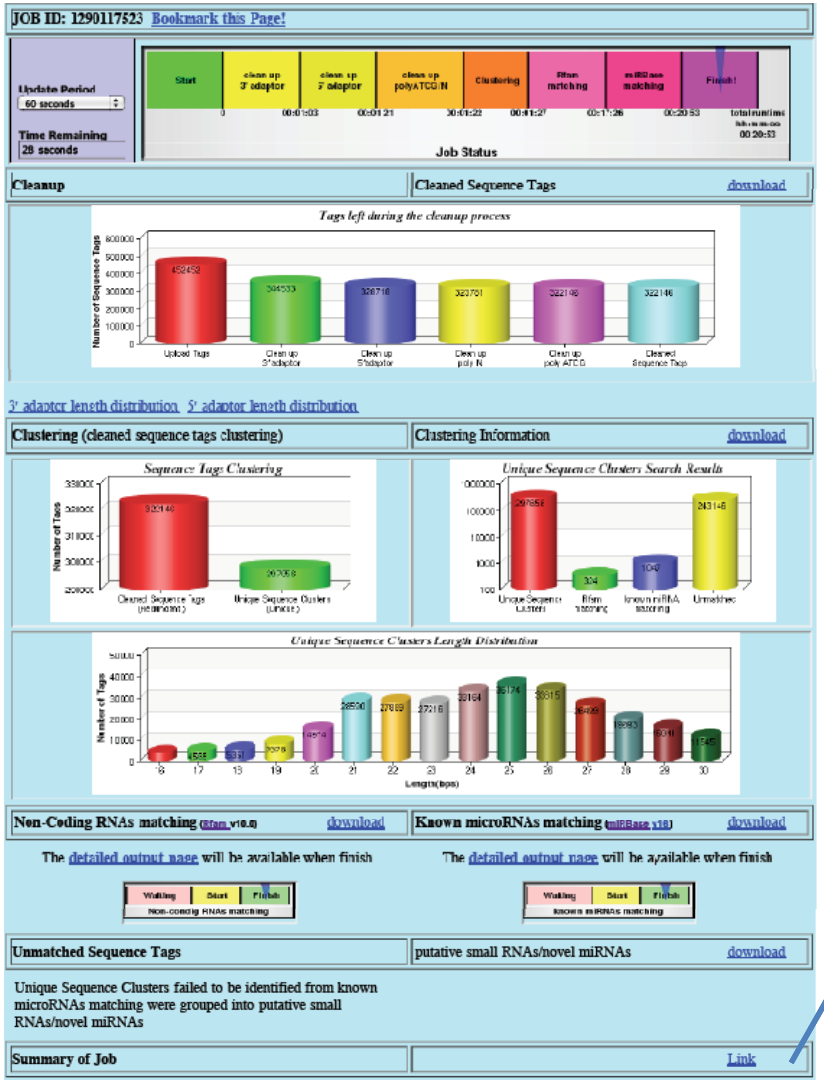
MANCHESTER 1824

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

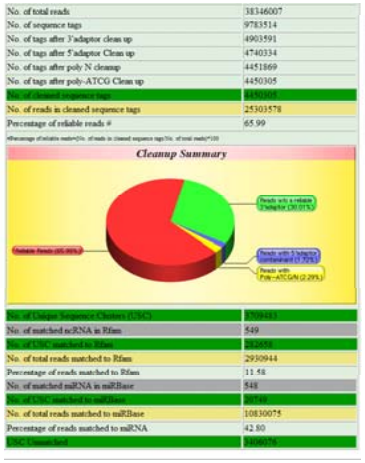
Stem-loop sequence MI0001950

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

DSAP Data Output



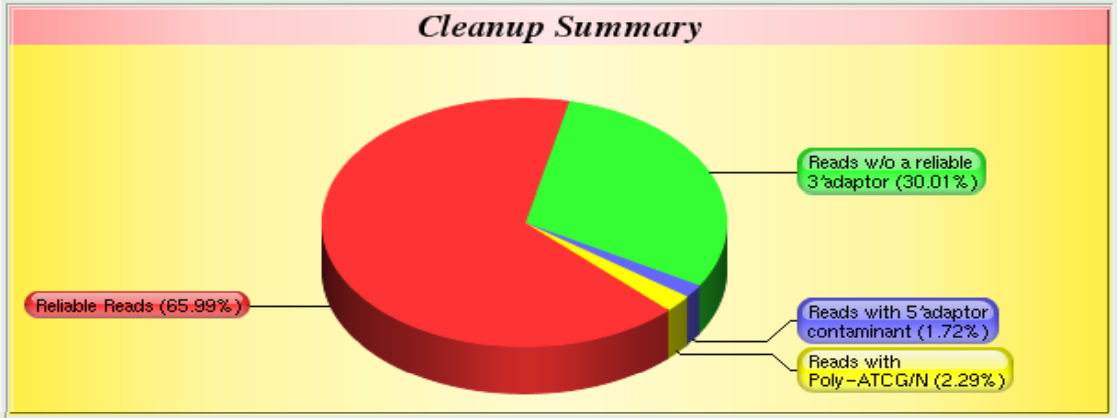
Summary of job



Summary 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



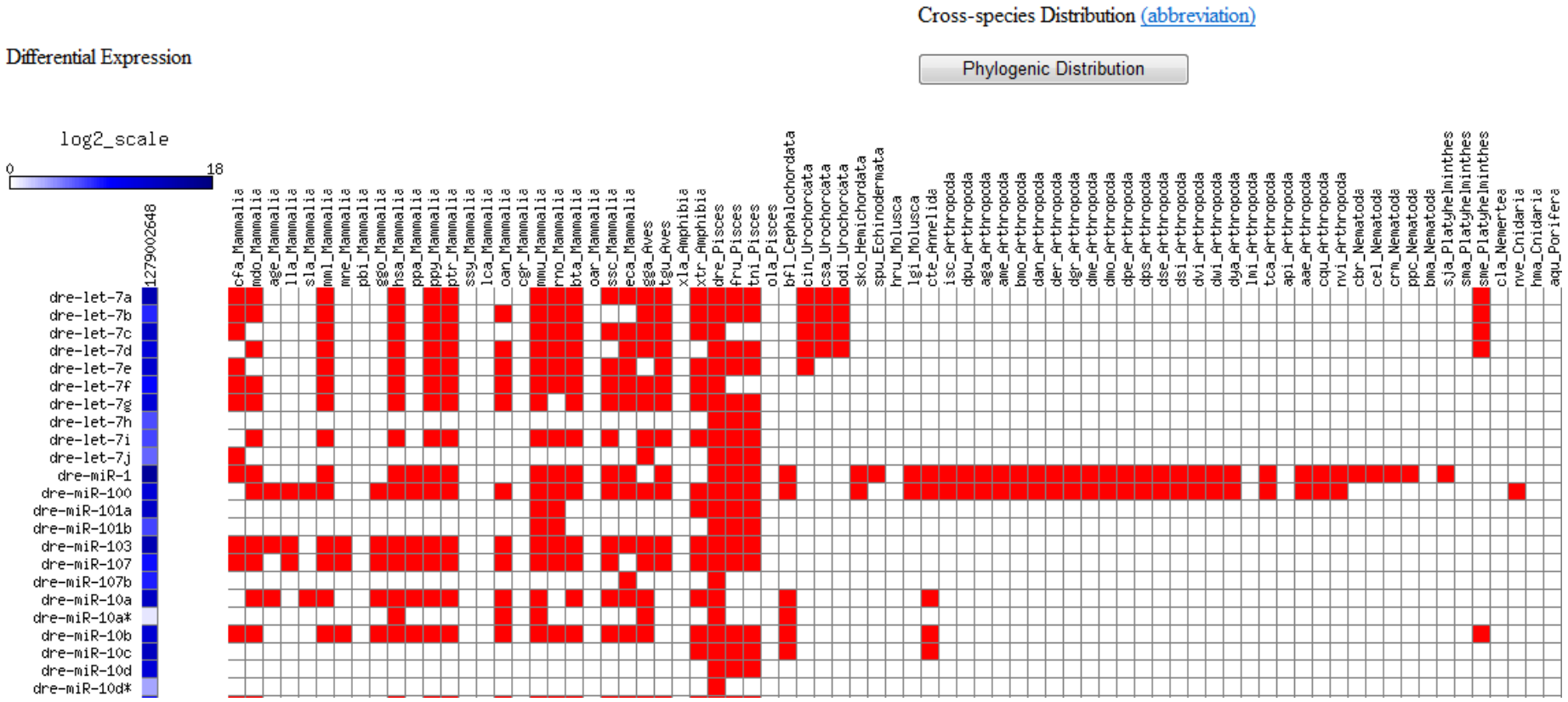
No. of Unique Sequence Clusters (USC)	3709483
No. of matched ncRNA in Rfam	549
No. of USC matched to Rfam	282658
No. of total reads matched to Rfam	2930944
Percentage of reads matched to Rfam	11.58
No. of matched miRNA in miRBase	548
No. of USC matched to miRBase	20749
No. of total reads matched to miRBase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

Comparative miRNAomics

- **Single dataset**
  - Cross-species distribution of identified miRNAs
  - Phylogenic Distribution of Identified miRNAs
- **Multiple datasets**
  - User's own profile
  - mature miRNA
  - miRNA family
  - Pairwise comparison
  - Summary output

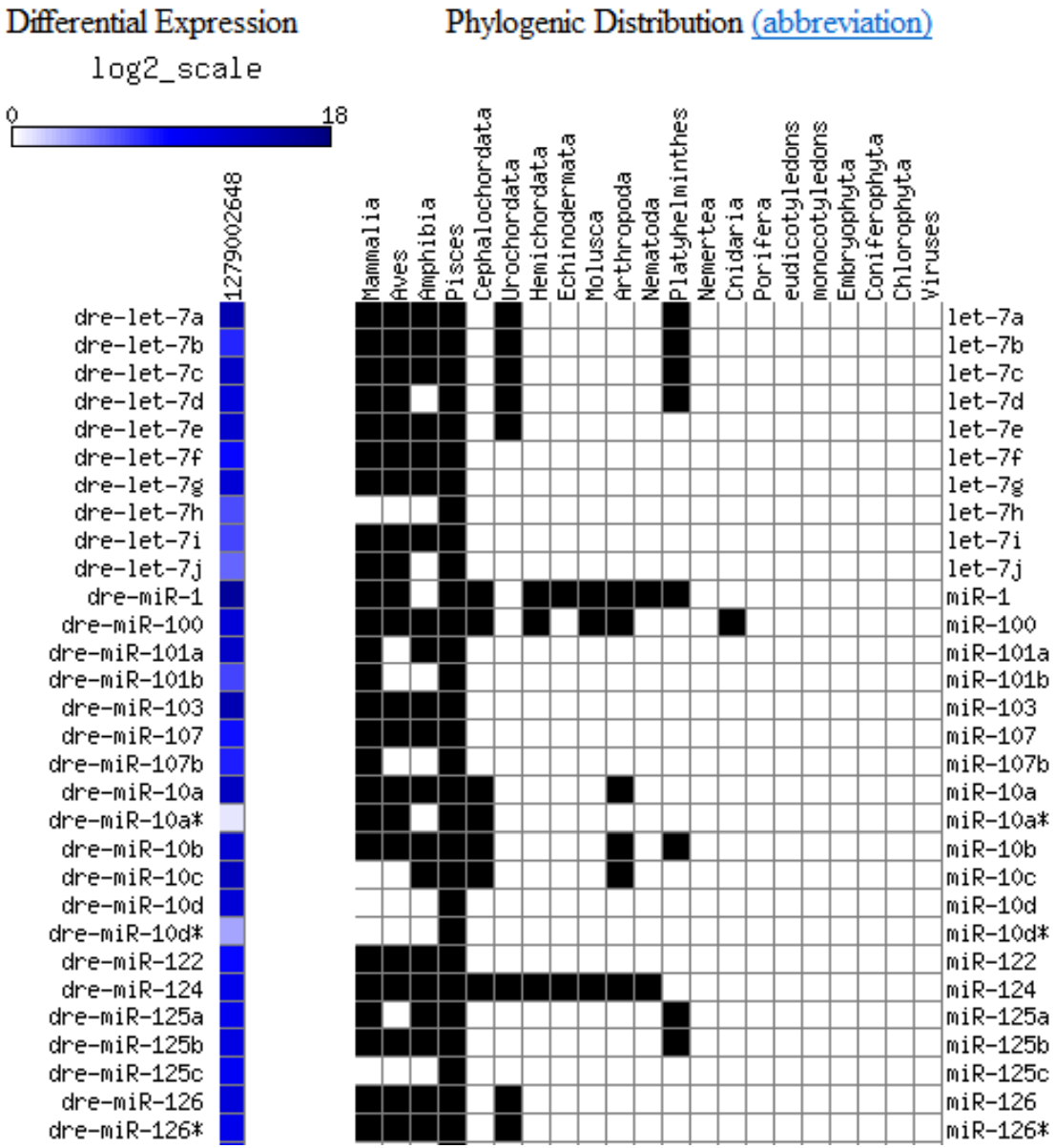
Comparative miRNAomics

- **Single dataset**
  - Cross-species distribution of identified miRNAs



Comparative miRNAomics

- **Single dataset**
  - Phylogenic Distribution of Identified miRNAs



Comparative miRNAomics

- **Multiple datasets**
  - User's own profile

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.

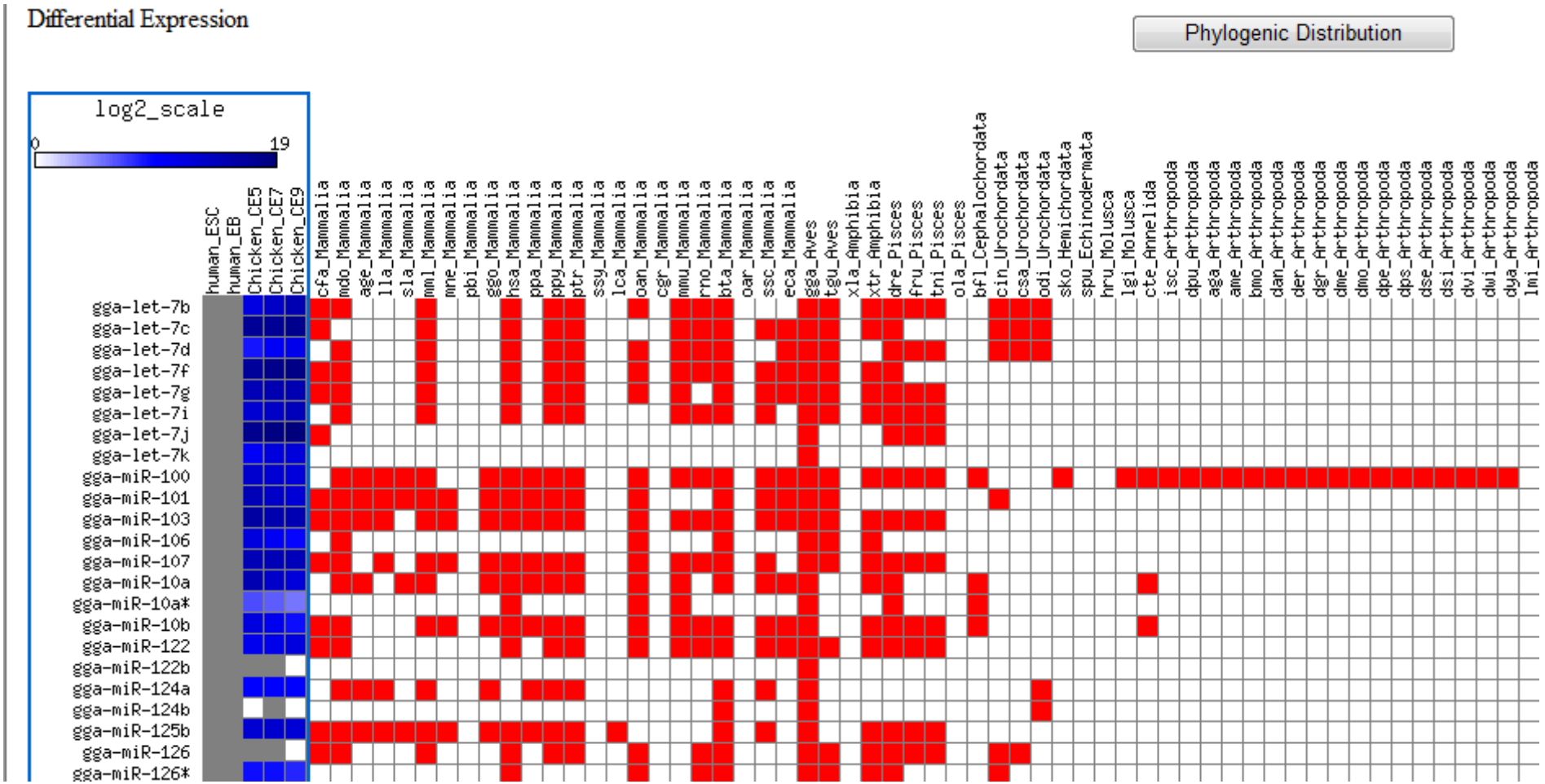
Paste your own expression profile in the textarea

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

Submit Reset

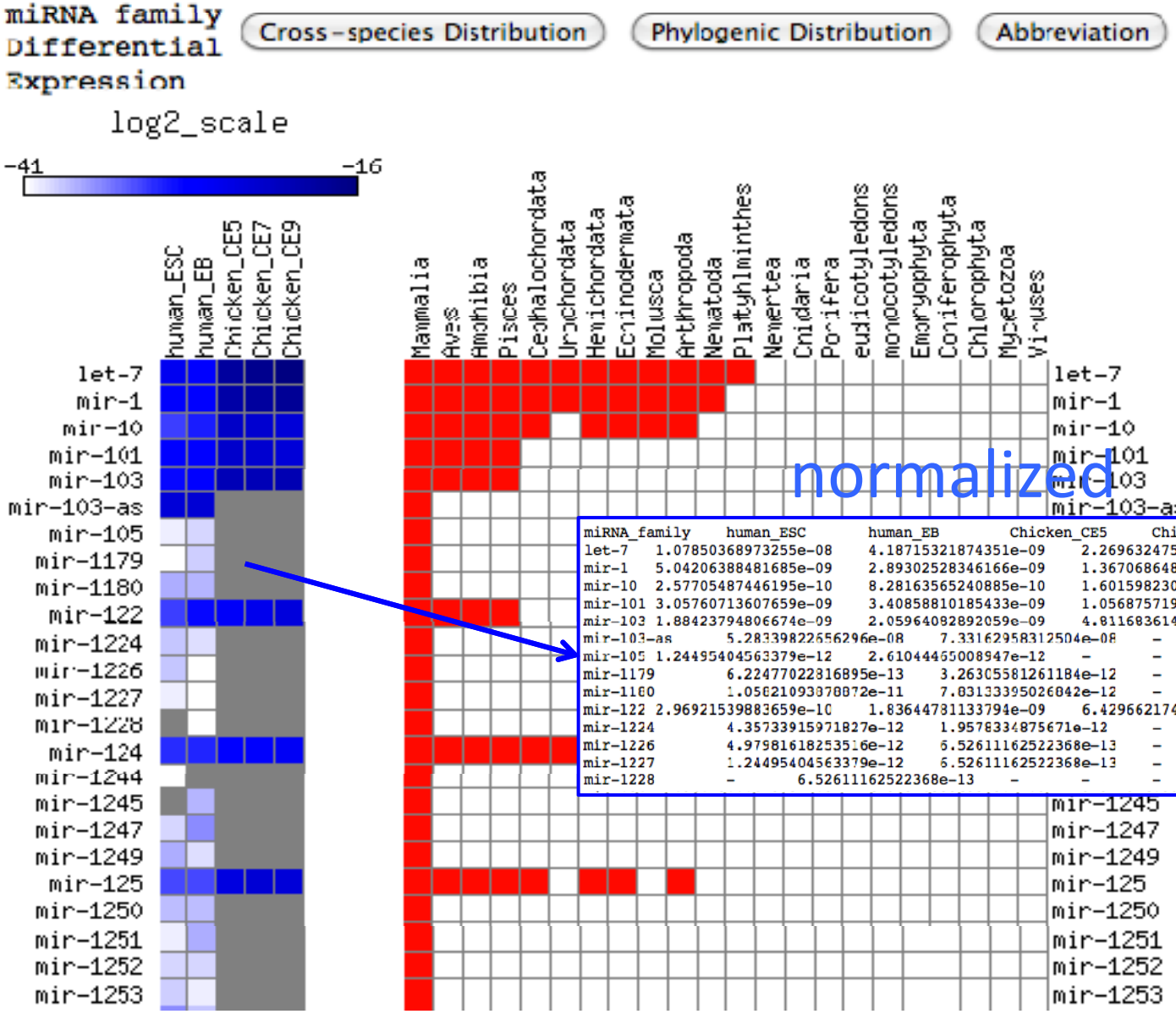
Comparative miRNAomics

- **Multiple datasets**
  - mature miRNA



Comparative miRNAomics

- **Multiple datasets**
  - miRNA family

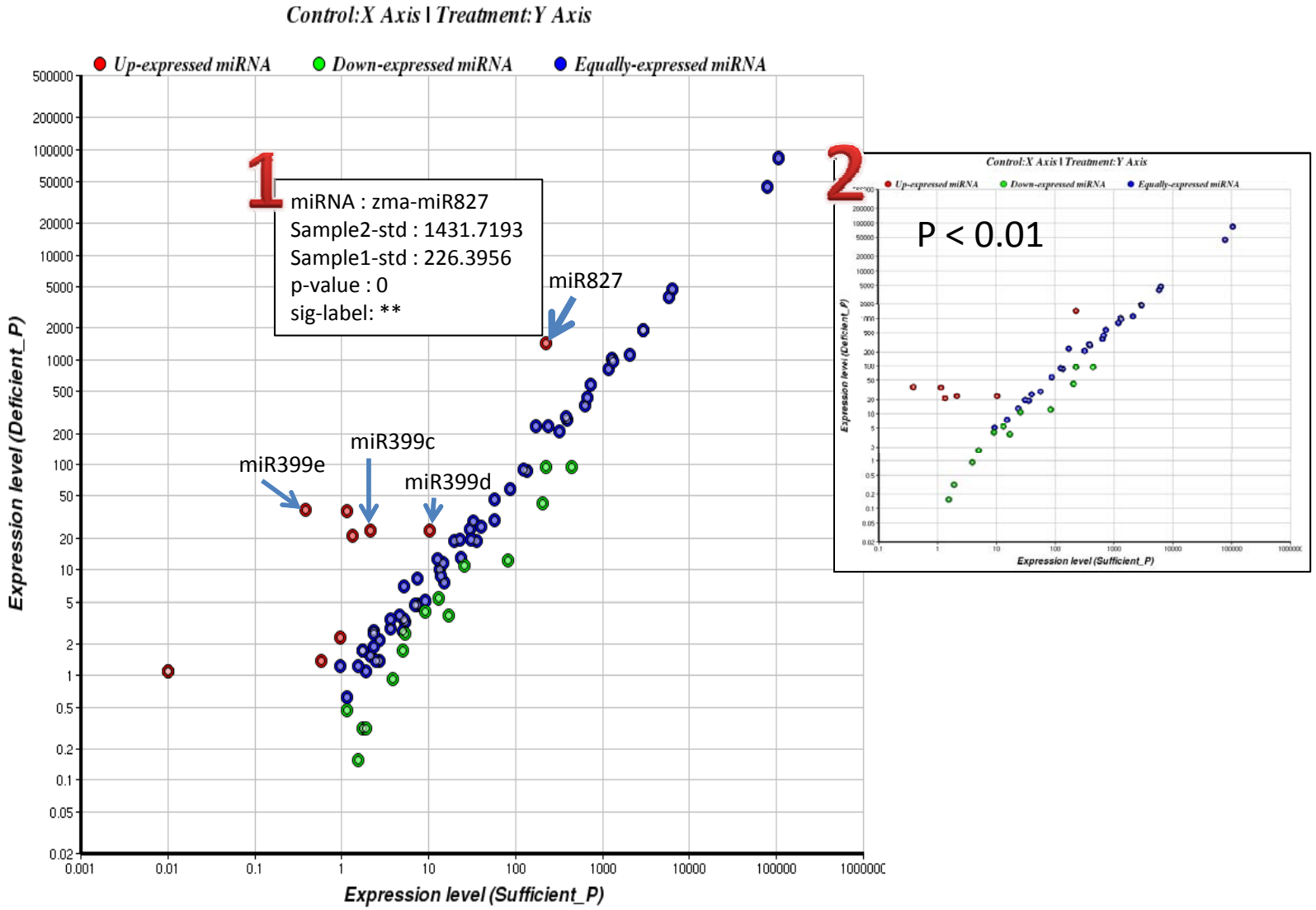


TPM(Tag Per Million)

Normalized expression = Actual miRNA count/Total count of clean reads*1,000,000

Comparative miRNAomics

- **Multiple datasets**
  - Pairwise comparison



Fold change: Fold_change=log2(treatment/control)

p-vaule: $p(y|x) = (\frac{N_2}{N_1})^y \frac{(x+y)!}{x!y!(1+\frac{N_2}{N_1})^{(x+y+1)}}$ $C(y \le y_{min}|x) = \sum_{y=0}^{y \le y_{min}} p(y|x)$ $D(y \ge y_{max}|x) = \sum_{y=y_{max}}^{\infty} p(y|x)$

The significance of digital gene expression profiles, Genome Res 7, 986-995. (1997)

Comparative miRNAomics

➢ Multiple datasets

➢ Summary output

Attached file is a summary for the analysis. A JOB ID was used to locate the result of DSAP.

Condition1: 130642XXXX (http://dsap.cgu.edu.tw/tmp/1306425777/130642XXXX.html)
Condition2: 130642XXXX (http://dsap.cgu.edu.tw/tmp/1306425839/130642XXXX.html)
Condition3: 130642XXXX (http://dsap.cgu.edu.tw/tmp/1306425881/130642XXXX.html)
Condition4: 130642XXXX(http://dsap.cgu.edu.tw/tmp/1306425988/130642XXXX.html)
Condition5: 130642XXXX (http://dsap.cgu.edu.tw/tmp/1306425947/130642XXXX.html)

Detail Information will be provided in a zip file for download. For each dataset , you will have 12
1. boxplot_on_*.png : quality of reads
2. clean.fasta.zip : clean reads after adaptor....
3. adaptor2.png:
4. cluster2.png: (unique sequence tags search results)
5. cluster_length.png: (Unique sequence tags length distribution)
6. Rfam.txt.zip: (ncRNA profile)
7.Rfam.png: (classification of ncRNAs)
8. mir.txt.zip: (miRNA profile)
9.miRNA_sort_by_expression.png
10.miRNA_sort_by_name.png
11.miRNA_group_by_family.png
12. smallrna.fasta.zip: (un-annotated sequences)

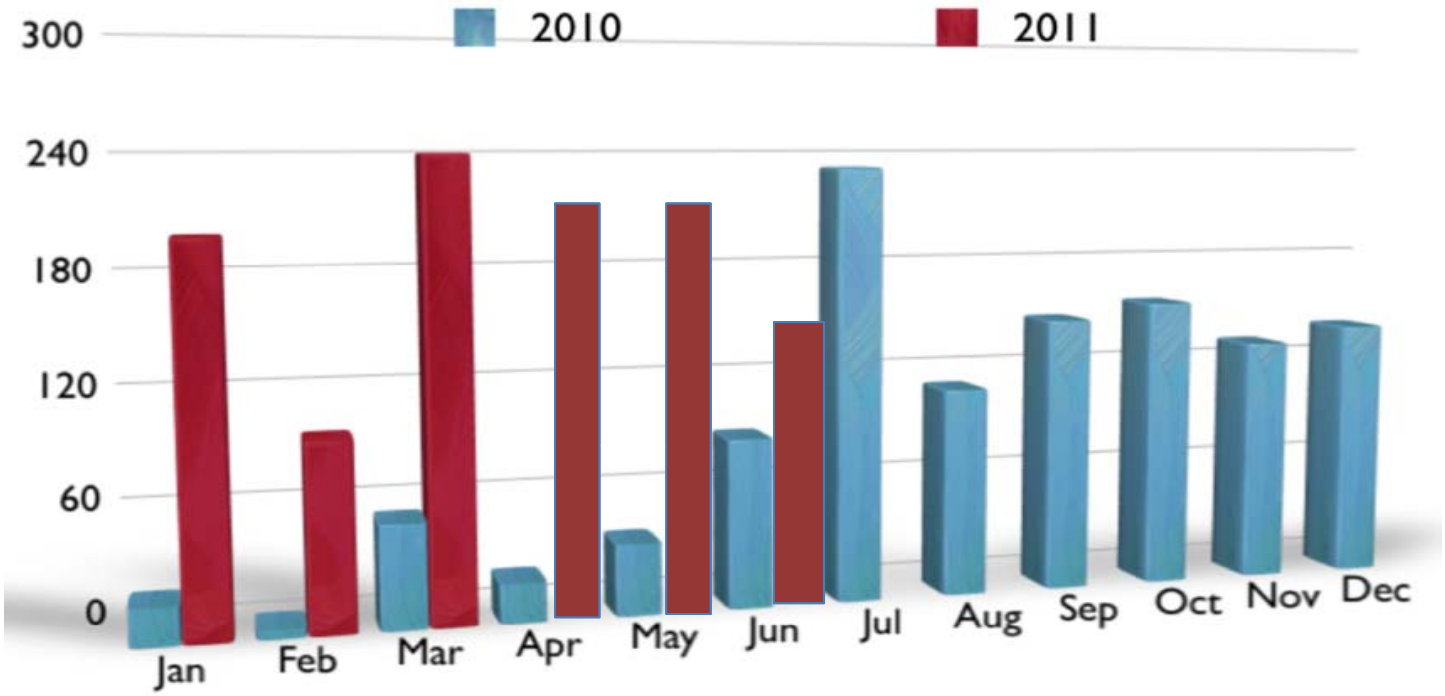
Differential expression of miRNA (normalized/non-normalized) among different conditions was group by
(1) miRNA (normalized, non-normalized)
(2) miRNA family (normalized, non-normalized)
PS: If you want to change the datasets for comparsion, you can use this web page (http://dsap.cgu.edu.tw/miRNAomics.html) and fill in the job IDs.

Differential expression of ncRNAs was group by ncRNA:
(normalized, non-normalized)

Usage Statistics

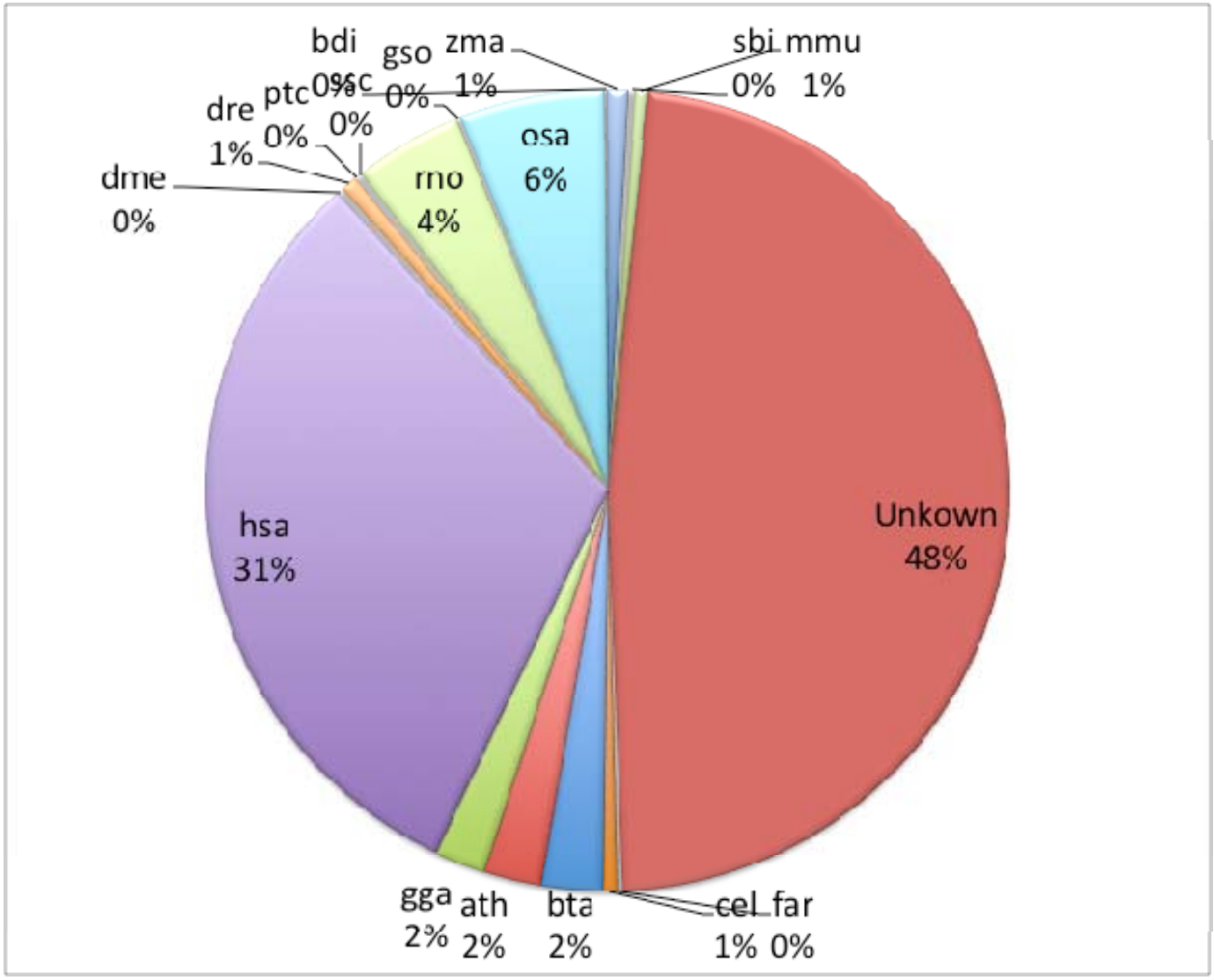
1,000 jobs mark in October, 2010

> 2,000 Datasets uploaded

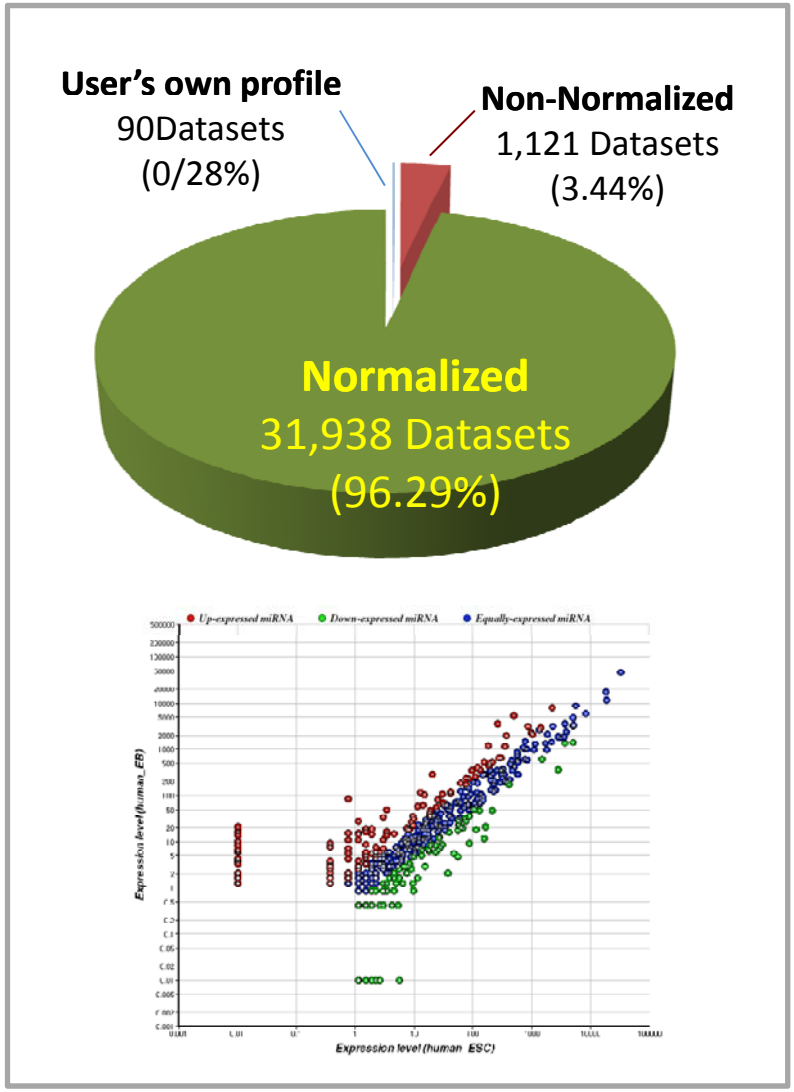


Species Statistics

48% of the datasets are Unknown species



Job Statistics of Comparative miRNAomics



Over 40,000 requests for Comparative miRNAomics

New Concepts on Bioinformatics Analysis Pipeline



Pick & Analysis

All You can Do

DSAP²

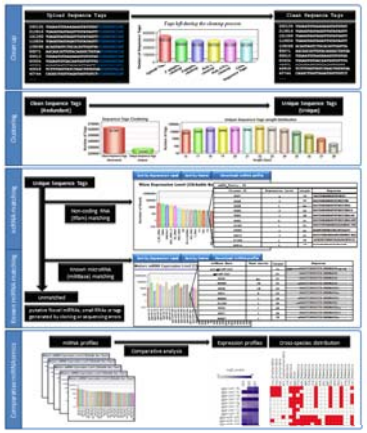
RSAP

(RNAseq Analysis Pipeline)

SVAP

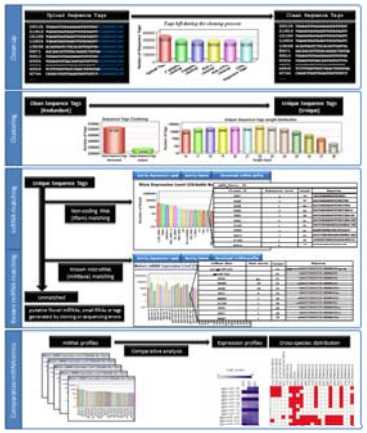
(Splice-Variant Analysis Pipeline)

DSAP² Modulized Data Processing



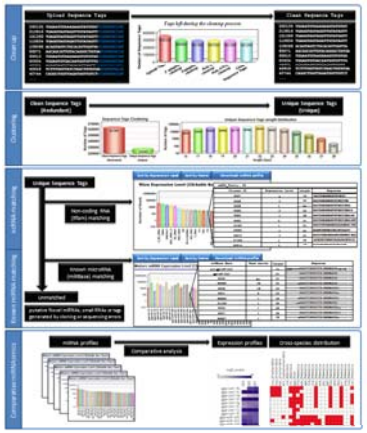
Upload Sequence	*Tag Format*
Species	*Any combination of 142 species*
Clean-up	*Adaptor 1, Adaptor 2, Custom Adaptor, No*
Clustering	*Mismatch: 0, 1, 2*
Mapping	*Blast, SOAP, Bowtie, BLAT* *Mismatch: 0, 1, 2*
Database	*Rfam, mature miRNA, miRNA Hairpin*
IsomiR	*Yes, No*
Output	*Excel, Figure (by small RNA, Family, Haiprin)*
Comparsion	*Pairwise, Multiple Datasets* *Normalized, non-normalized* *by small RNA, Family, Haiprin*

DSAP² Modulized Data Processing



- **Upload Sequence** — *Tag Format*
- **Species** — *Any combination of 142 species*
- **Clean-up** — *Adaptor 1, Adaptor 2, Custom Adaptor, No*
- **Clustering** — *Mismatch: 0, 1, 2*
- **Mapping** — *Blast, SOAP, Bowtie*
  Mismatch: 0, 1, 2
- **Database** — *Rfam, mature miRNA, miRNA Hairpin*
- **IsomiR** — *Yes, No*
- **Output** — *Excel, Figure (by small RNA, Family, Haiprin)*
- **Comparsion** — *Pairwise, Multiple Datasets*
  Normalized, non-normalized
  by small RNA, Family, Haiprin

DSAP² Modulized Data Processing



- **Upload Sequence** — *Tag Format*
- **Species** — *Any combination of 142 species*
- **Clean-up** — *Adaptor 1, Adaptor 2, Custom Adaptor, No*
- **Clustering** — *Mismatch: 0, 1, 2*
- **Mapping** — *Blast, SOAP, Bowtie*
  Mismatch: 0, 1, 2
- **Database** — *Rfam, mature miRNA, miRNA Hairpin*
- **IsomiR** — *Yes, No*
- **Output** — *Excel, Figure (by small RNA, Family, Haiprin)*
- **Comparsion** — *Pairwise, Multiple Datasets*
  Normalized, non-normalized
  by small RNA, Family, Haiprin

IsomiRs & miRNA Editing

>hsa-let-7a-1					
TGGGATGAGGTAGTAGGTTGTATAGTTTTAGGGTCACACCCACCACTGGGAGATAACTATACAATCTACTGTCTTTCCTA	hsa-let-7a-1	80			
(((((.(((..((((((((((((((((...(((.....))).((....)).....)))))))))))))))..))))))))	structure	-35.6			
*****TGAGGTAGTAGGTTGTATAGTT**********	hsa-let-7a	22	#Reads	isomiRs	
**********CTATACAATCTACTGTCTTTC***	hsa-let-7a*	21		5'	3'
..GGATGAGGTAGTAGGTTGTATAG.....	t0227878	23	1	-3	-2
....ATGAGGTAGTAGGTTGTATAGTT...	t0006845	23	42	-1	0
....ATGAGGTAGTAGGTTGTATAGT....	t0021457	22	9	-1	-1
....ATGAGGTAGTAGGTTGTATAG.....	t0039011	21	4	-1	-2
....ATGAGGTAGTAGGTTGTATAGTTT..	t0165920	24	1	-1	1
.....TGAGGTAGTAGGTTGTATAGTT...	t0000002	22	1043926	0	0
.....TGAGGTAGTAGGTTGTATAGT....	t0000015	21	96011	0	-1
.....TGAGGTAGTAGGTTGTATAGTTT..	t0000044	23	36942	0	1
.....TGAGGTAGTAGGTTGTATAG.....	t0000082	20	13528	0	-2

Chang Gung Bioinformatics Core Laboratory

長庚生物資訊中心

bioinfo.cgu.edu.tw

The Bioinformatics Core Laboratory is the central repository of all the information relevant to "The Study of Bioinformatics Technologies on Medical Applications" and "Cancer Biomarker Discovery". This core is the medium for information exchange and dissemination, providing a virtual platform for communication between the participating laboratories. The core is responsible for the acquisition, storage, analysis, and distribution of information arising from the activities of the participating investigators and from other public databases.

Missions

(1) Provide centralized services to members of the "Chang Gung Molecular Medicine Research Center" to facilitate research in bioinformatics and provide access to advanced bioinformatics technologies.

(2) Maintain and integrate databases generated from genomics, transcriptomics and proteomics research platforms for biomarker discovery.

(3) Develop and maintain a core information sharing structure to enable efficient acquisition, cataloging, storage and exchange of data between the program participants with appropriate interfaces to track progress on each of the specific aims for a particular project or investigator.

(4) Provide an environment to test the analytic techniques, hypotheses and algorithms developed from different projects.

Hardwares

Servers:

IBM eServer 336	14
Sun V60X server	10
1U Twin Server	3
IBM X3850 (128G)	1
IBMX3850 (256G)	1
IBM X3550 (512G)	1
Dell R815	1

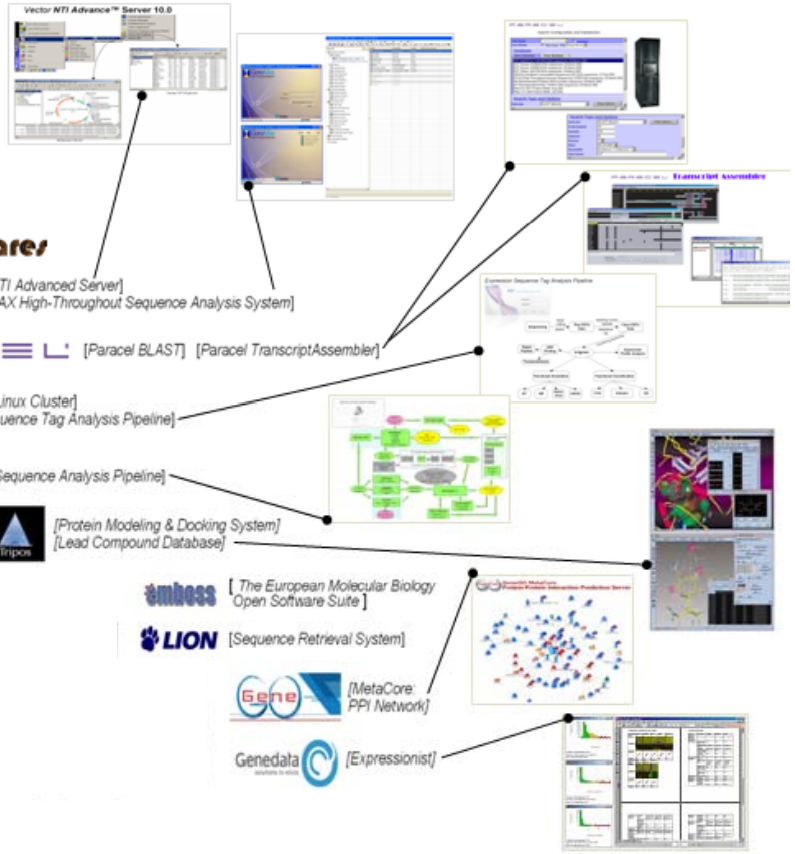
Storage:

Disk Array (6T-23T)	11

Classroom:

PC	30

Softwares



Softwares

High-Throughput Sequencing

Galaxy

DSAP