Multi-Omics onLine Analysis System for Profiling Gene Expression

MOLAS

Multi-Omics onLine Analysis System

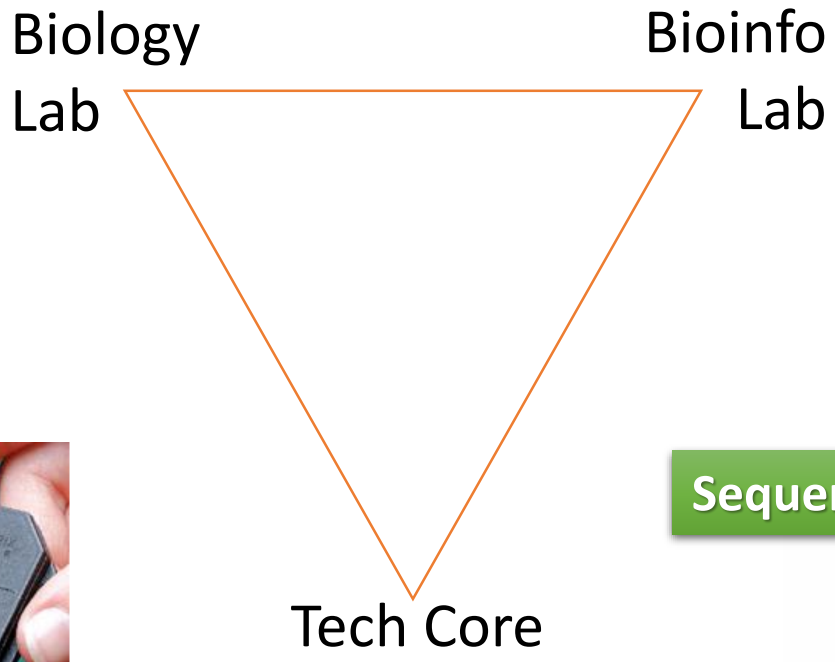
Life Science Library Training Course

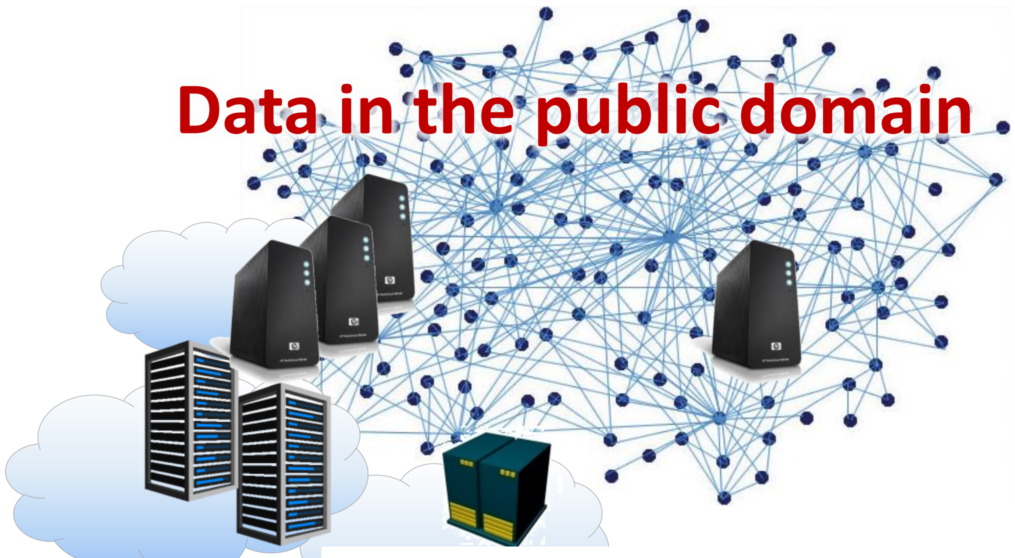
2018/12

Chen, Shu-Hwa

IIS, Academia Sinica

High-Throughput Methods





Data in the public domain

NCBI Site map | All databases | Search

Sequence Read Archive

https://www.ncbi.nlm.nih.gov/sra

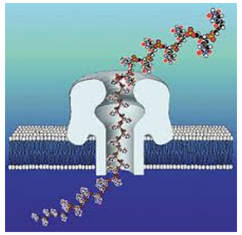
NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

https://cancergenome.nih.gov/

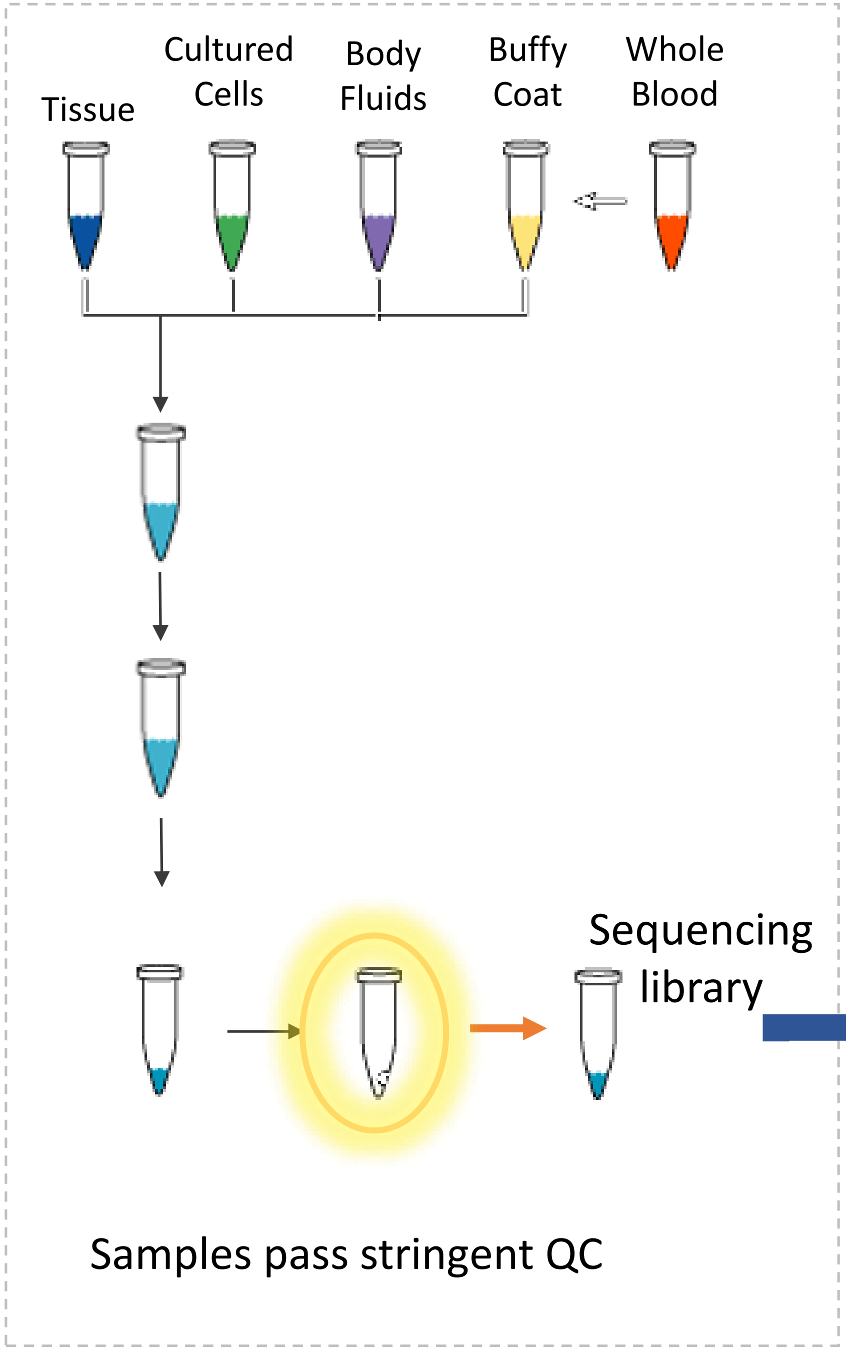
Hybridization-based Methods

Sequencing-based Methods

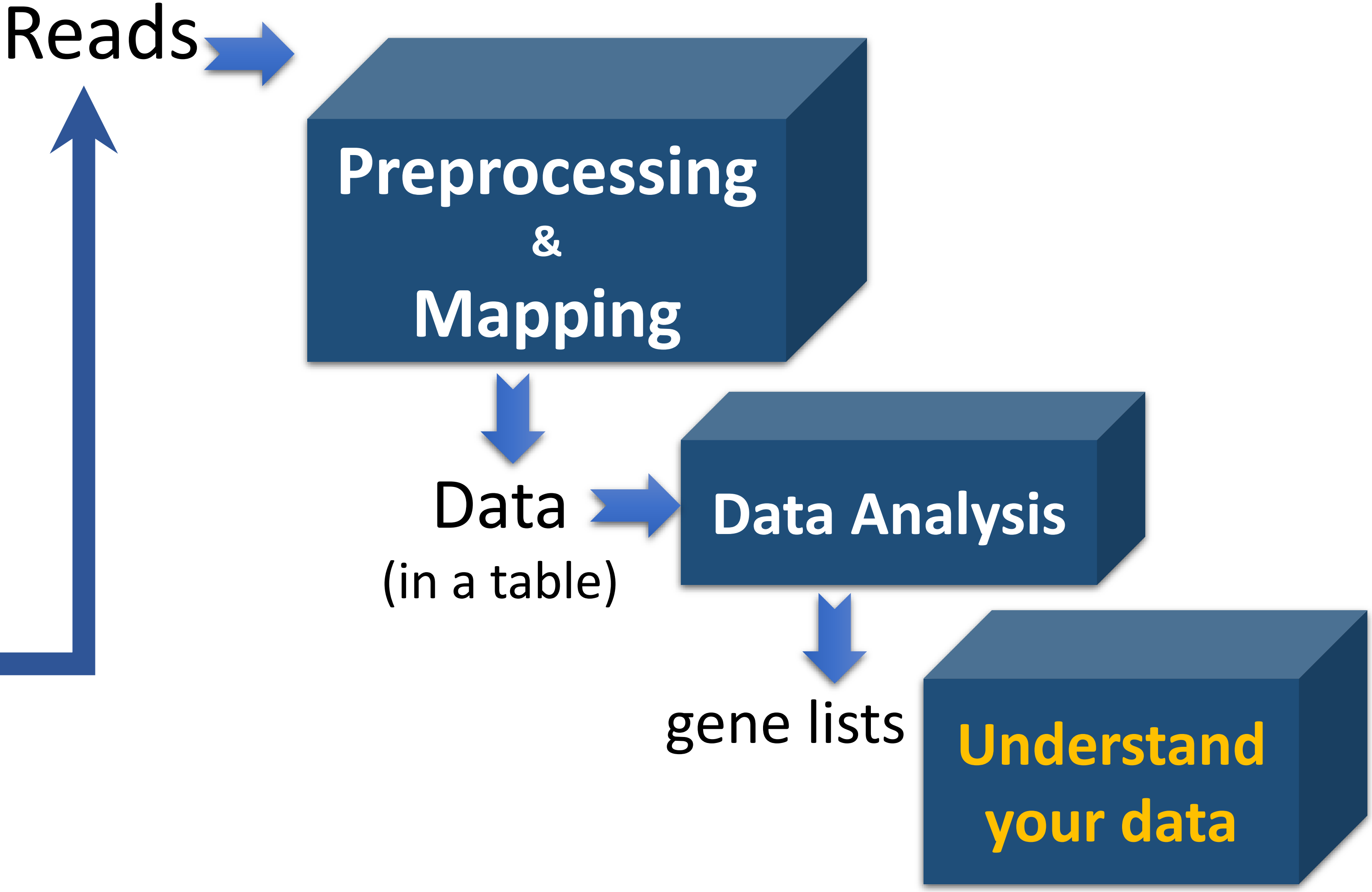
PacBio



How to ....?



Samples pass stringent QC



Reads

Preprocessing & Mapping

Data (in a table)

Data Analysis

gene lists

Understand your data

Read in FastQ format

https://en.wikipedia.org/wiki/FASTQ_format

DON’T TRY TO OPEN a fastq file on your desktop PC

- Start with “@”
- Four lines: “+” w/ or w/o seq head, quality scores

seq head	`@EAS139:136:FC706VJ:2:5:1000:12850 1:N:18:ATCACG`
seq letters	`ACTTCAGGAGATTGTACATTTAGAGACAAAAAAAAA`
+	`+`
quality score	`BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@`

R1, R2 can in separate fastq files
or sometimes in an interlanced fastq file:

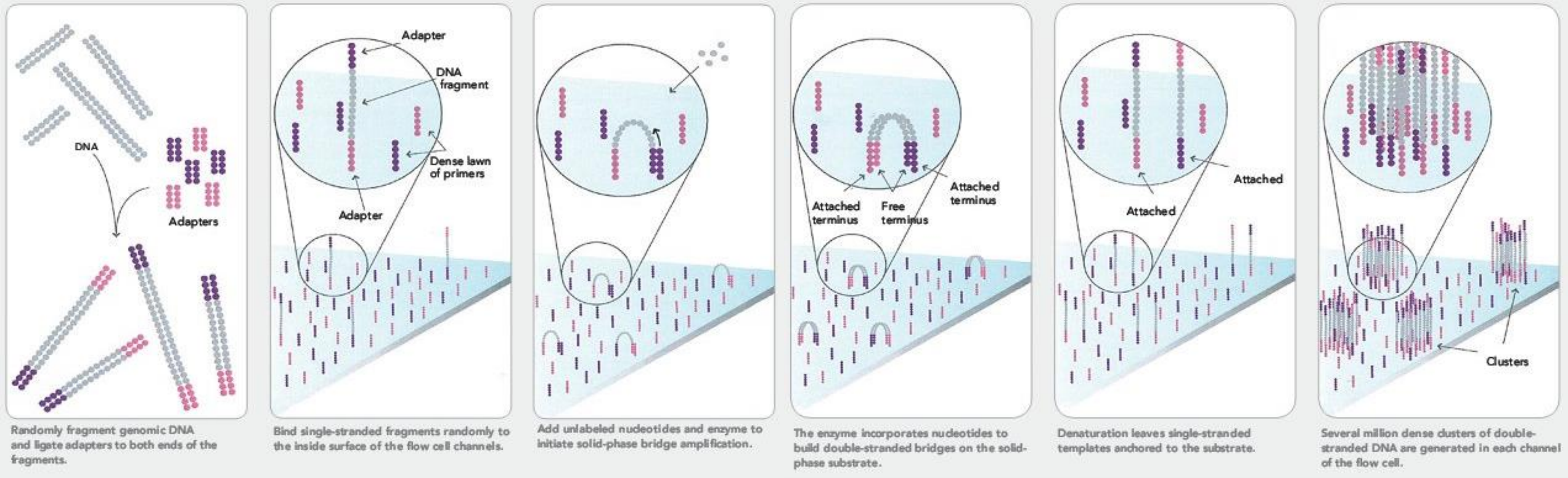
```
@xxxxx:xxxx:xxxx:..... 1:N:0
................
+
................
@xxxxx:xxxx:xxxx:..... 2:N:0
................
+
................
@xxxxx:xxxx:xxxx:..... 1:N:0
................
+
................
@xxxxx:xxxx:xxxx:..... 2:N:0
................
+
................
```

Illumina reads

fastq files from a sequencer should have the following READ-ID format:

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<index sequence>

Illumina Technology



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

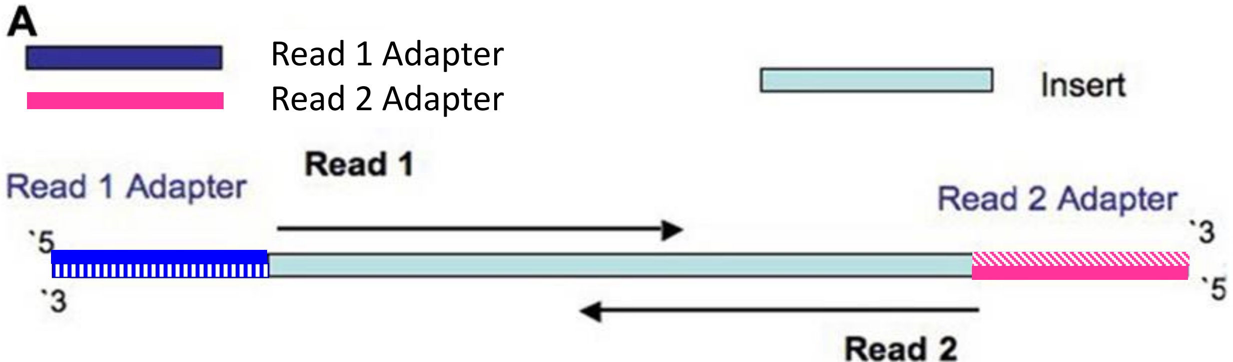
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

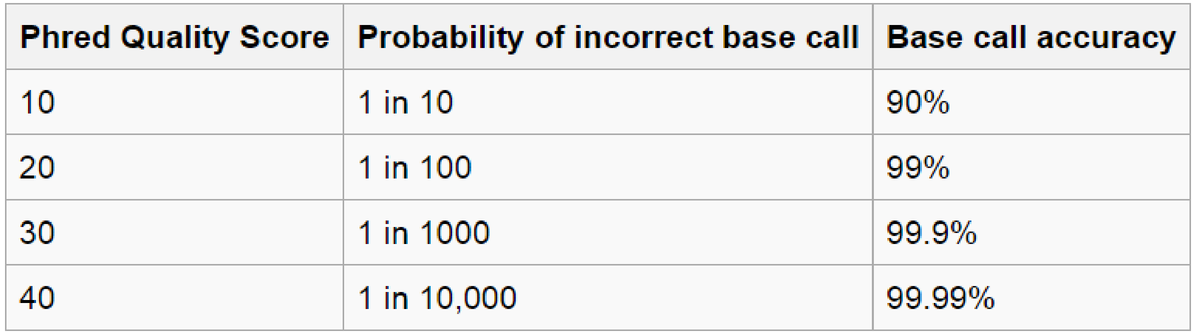
Denaturation leaves single-stranded templates anchored to the substrate.

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.



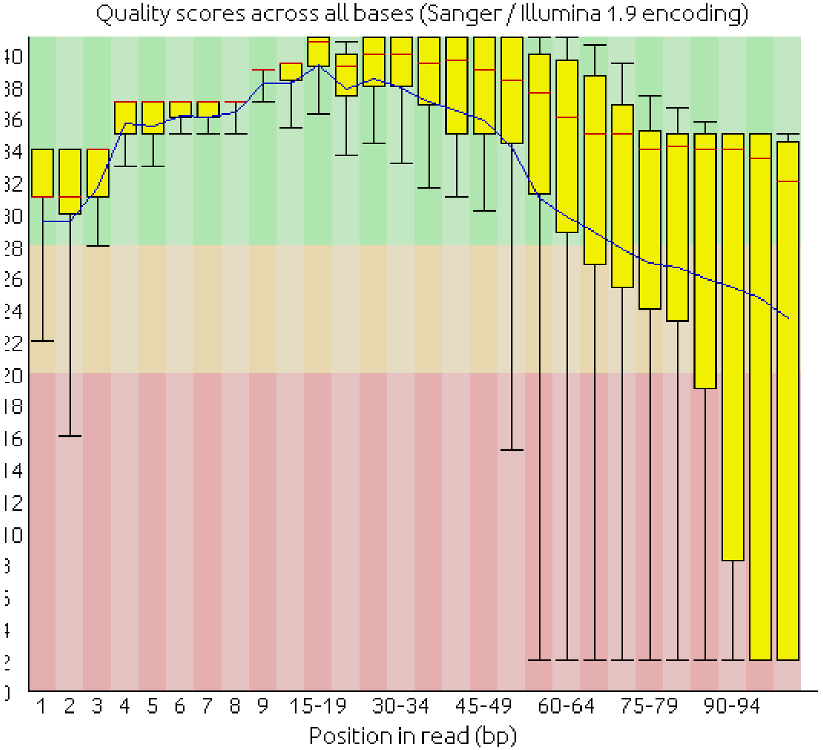
Check the Quality of Reads

Trimming for base quality

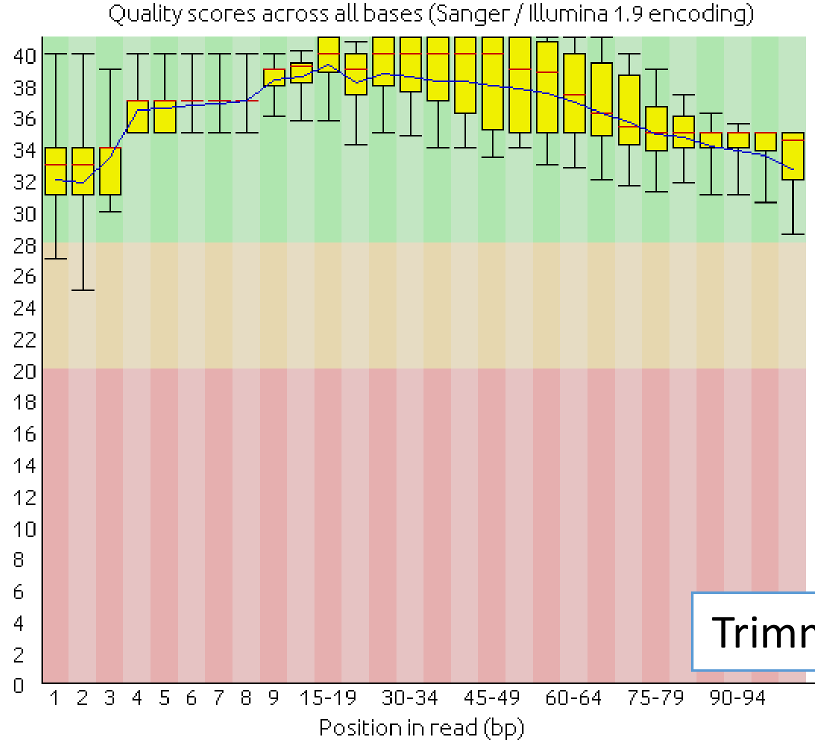


Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Raw Reads



Trimming
by base quality

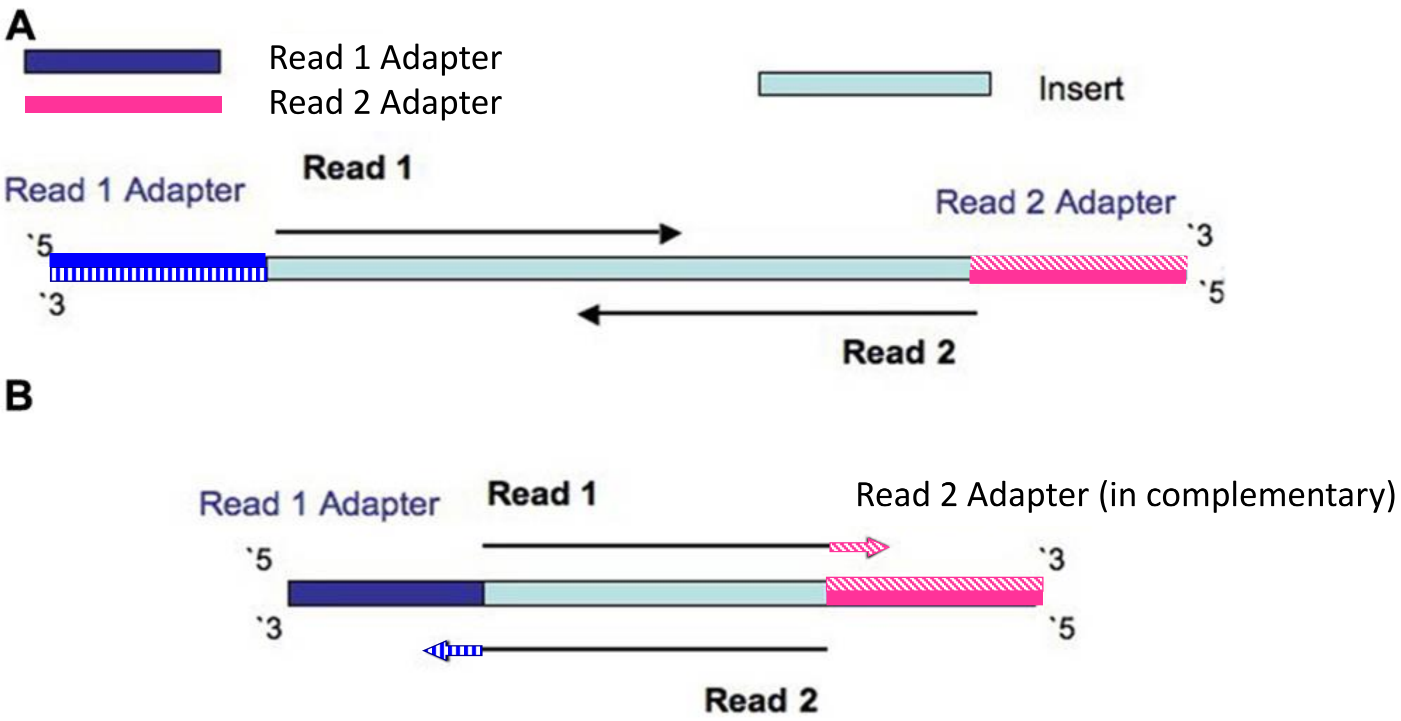


Trimmed Reads

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Read Preprocessing (optional):

Trimming for adapter contamination



Modified from figure2. http://journal.frontiersin.org/article/10.3389/fgene.2014.00005/full

Mapping: the Options

Mapping to a Known Genome

- Working on the target species: to profile the gene repertoire on a well-defined reference genome (fully sequenced and annotated) .
- Using a known genome close to your sample.

de novo Assembling + Mapping

- Create the reference (transcriptome or genome) by assembling.
- Annotate the new assembled reference.
- Map reads to the new assembled reference.

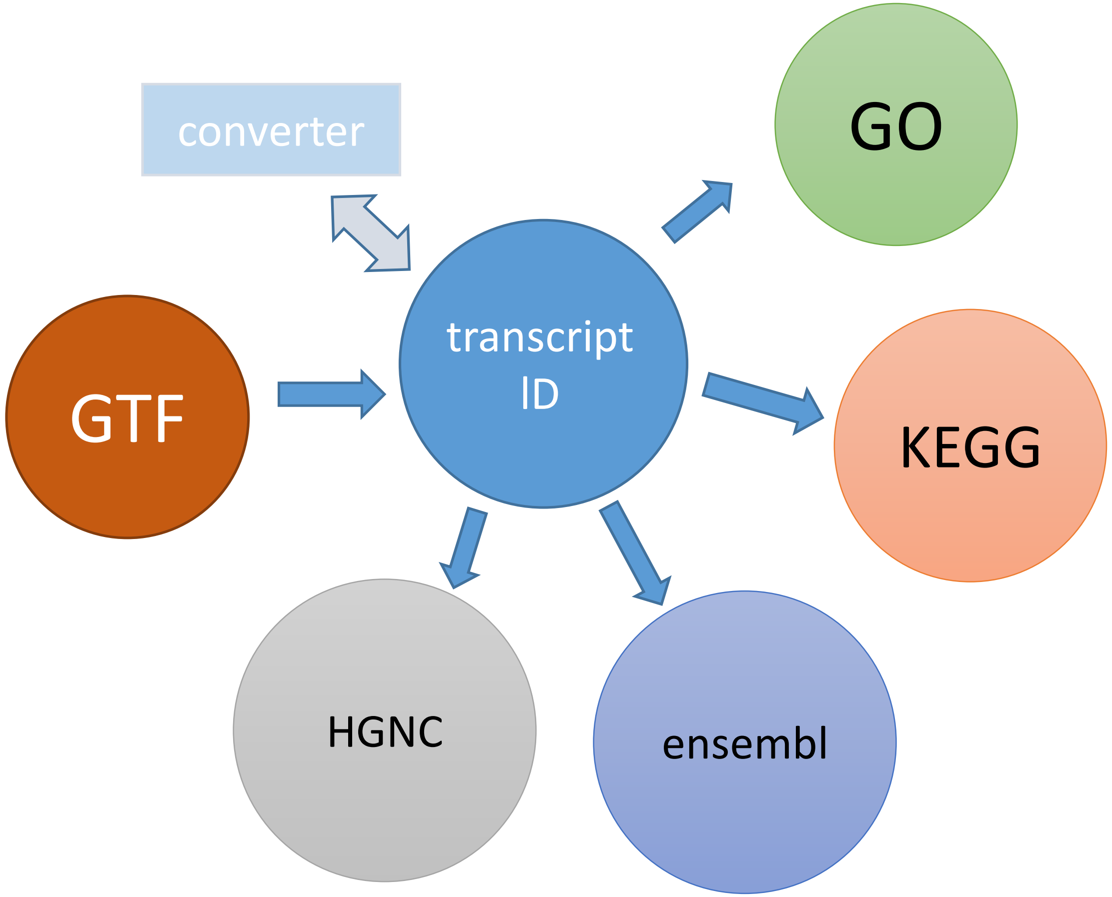
GTF: the Gene Tranfer Format

```
1       ensembl_havana  transcript      4344146 4360314 .       -       .       gene_id "ENSMUSG000000259
00"; gene_version "6"; transcript_id "ENSMUST00000027032"; transcript_version "5"; gene_name "Rp1"; gene_
source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "Rp1-001"; transcript_source "ens
embl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS14804";
```

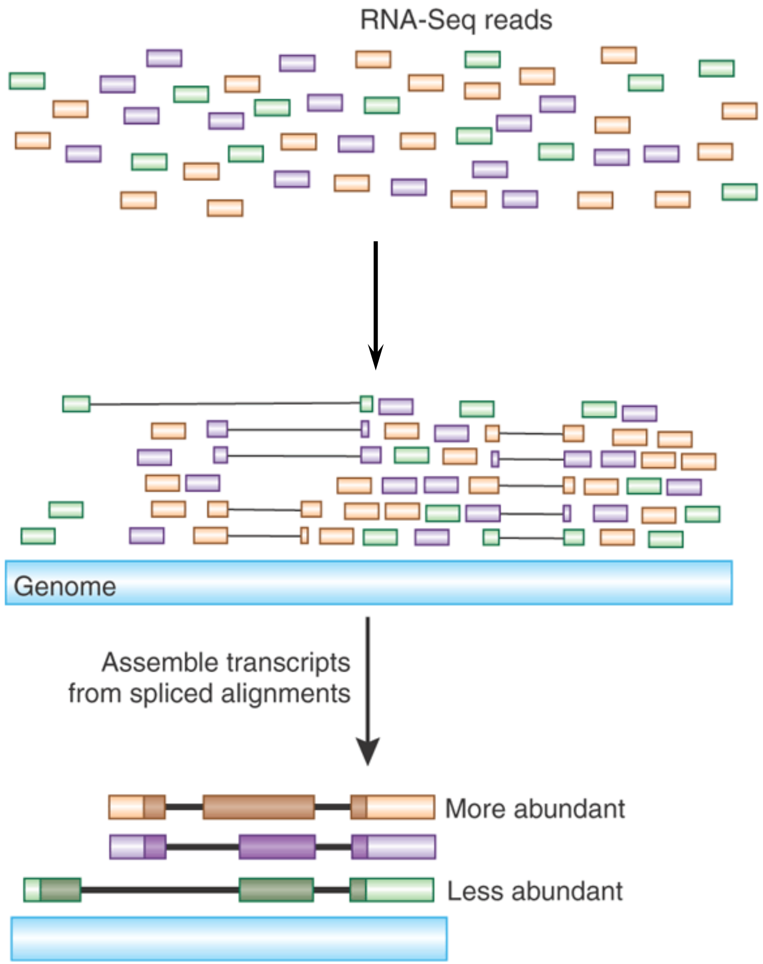
MOLAS compatible GTF

grch38
grch37

grcm38



Read Mapping

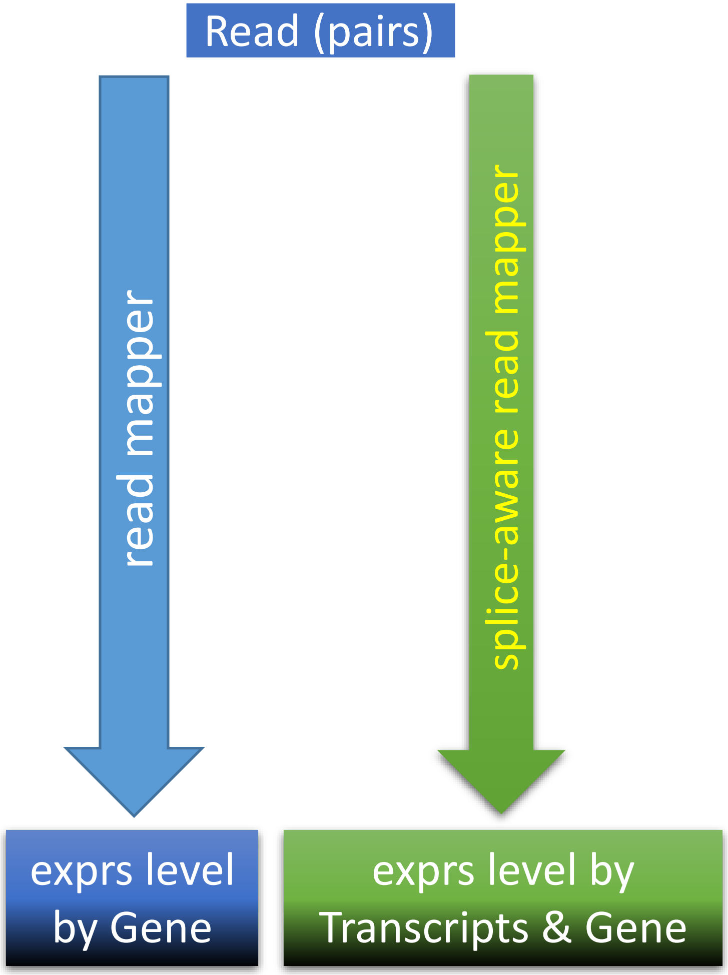


Reference Genome

- Seq: fasta file / prebuilt index
- Annotation : gtf / gff file

Expression Level

by Gene or by Transcript?



Read (pairs)

read mapper

splice-aware read mapper

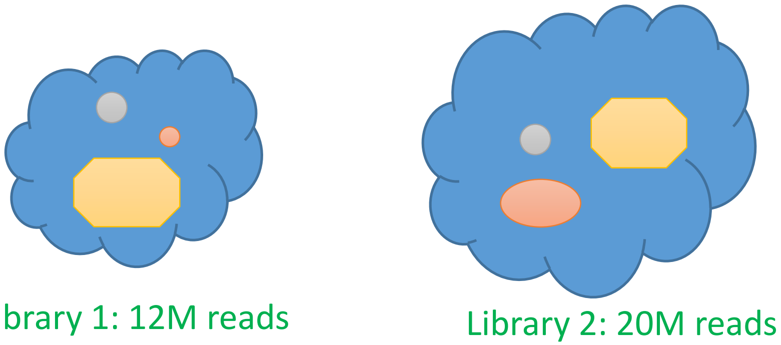
exprs level by Gene

exprs level by Transcripts & Gene

Normalization is a Necessary Evil

- Between samples:

  Initial Input ; Volume of Reads



- Within sample:

  transcript length effect

- Count the mapped read number, normalized to library size

  cpm: count per million reads

- Count the mapped read number, normalized to BOTH library size and (target seq) length

  ✓TPM: transcripts per million reads

  ✓RSEM: RNA-Seq by Expectation-Maximization

  ✓RPKM: reads Per kilobase of exon per million mapped reads

  ✓FPKM: fragments per kilobase of exon per million fragments mapped

http://www.slideshare.net/mikaelhuss/rnaseq-differential-expression-analysis

MOLAS

Multi-Omics onLine Analysis System

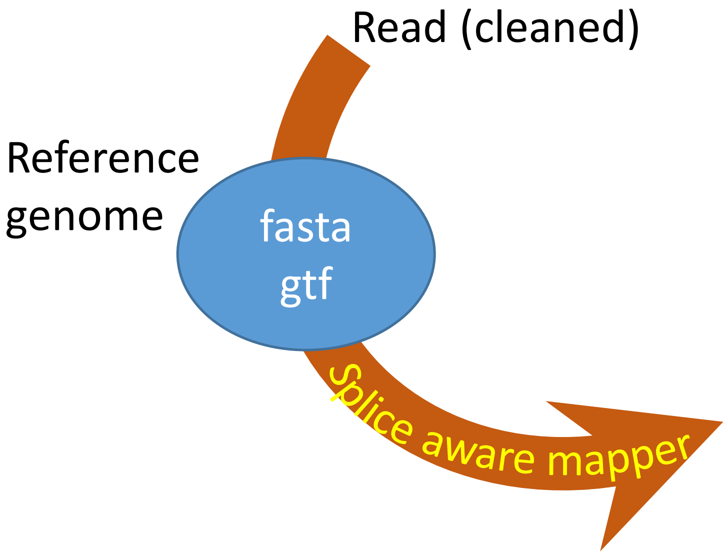
The Usage

Demo: http://molas.iis.sinica.edu.tw/grch38/

All you need is an expression file

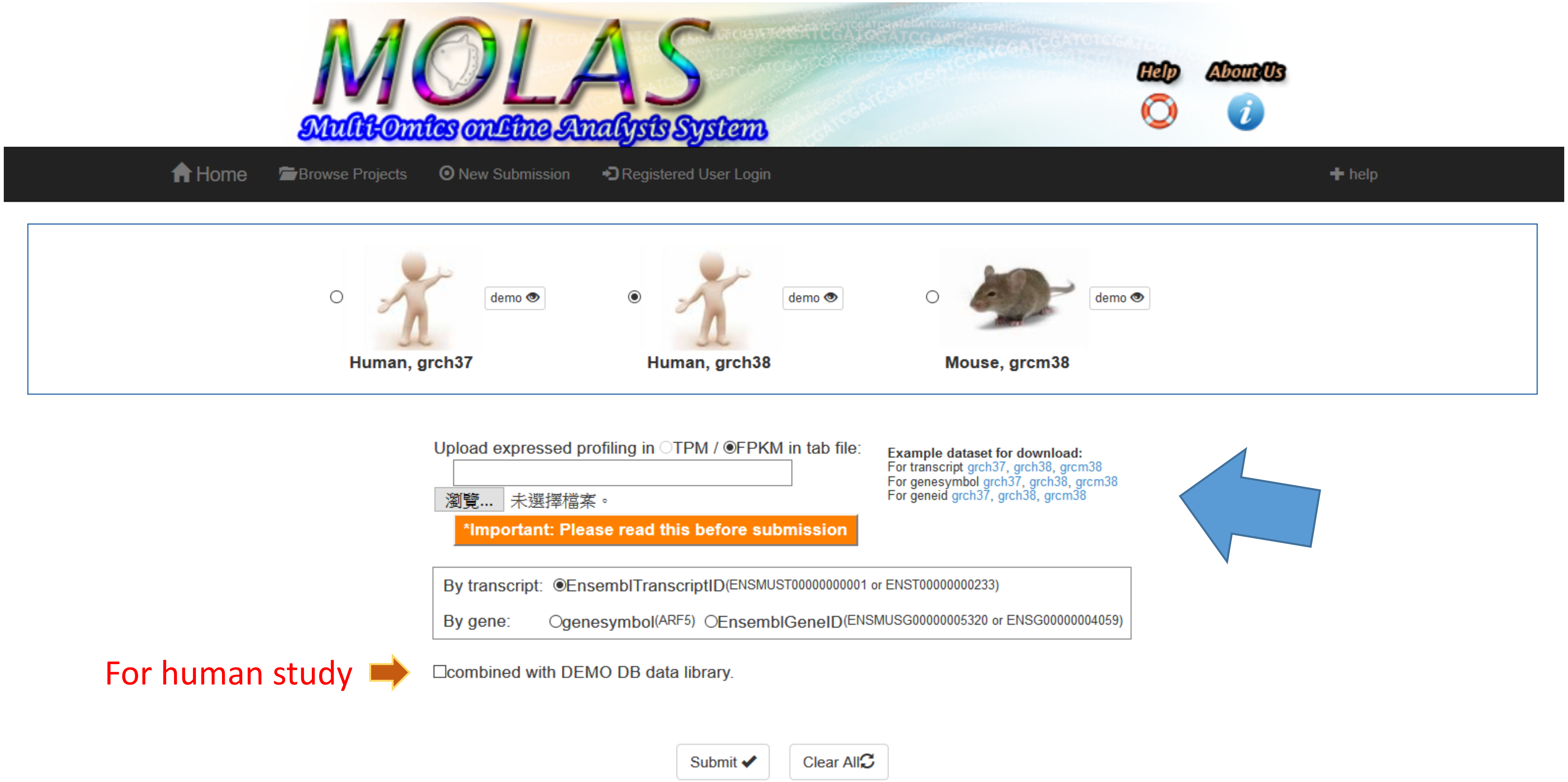
Input file

- A tab-delimited text file generated by other software (e.g. cufflink, EdgeR, RSEM) in ensembl transcript id (grch38 and grcm38)



```
#tracking_id        GA120-2_0     GA120-3_0
ENST00000591062     0             0.159246
ENST00000376259     0             3.96794
ENST00000235878     0.287651      0
ENST00000299596     0.0300576     0.0146675
ENST00000625158     6.08204       7.03465
ENST00000321949     4.24507       4.28616
ENST00000258484     0             6.00768
ENST00000625157     0.0134854     0.00783917
ENST00000321944     6.44635       5.25123
ENST00000321945     0.907242      1.13444
```

New Submission



○ Human, grch37 demo

● Human, grch38 demo

○ Mouse, grcm38 demo

Upload expressed profiling in ○TPM / ●FPKM in tab file:

瀏覽... 未選擇檔案。

*Important: Please read this before submission

Example dataset for download:
For transcript grch37, grch38, grcm38
For genesymbol grch37, grch38, grcm38
For geneid grch37, grch38, grcm38

By transcript: ●EnsemblTranscriptID(ENSMUST00000000001 or ENST00000000233)

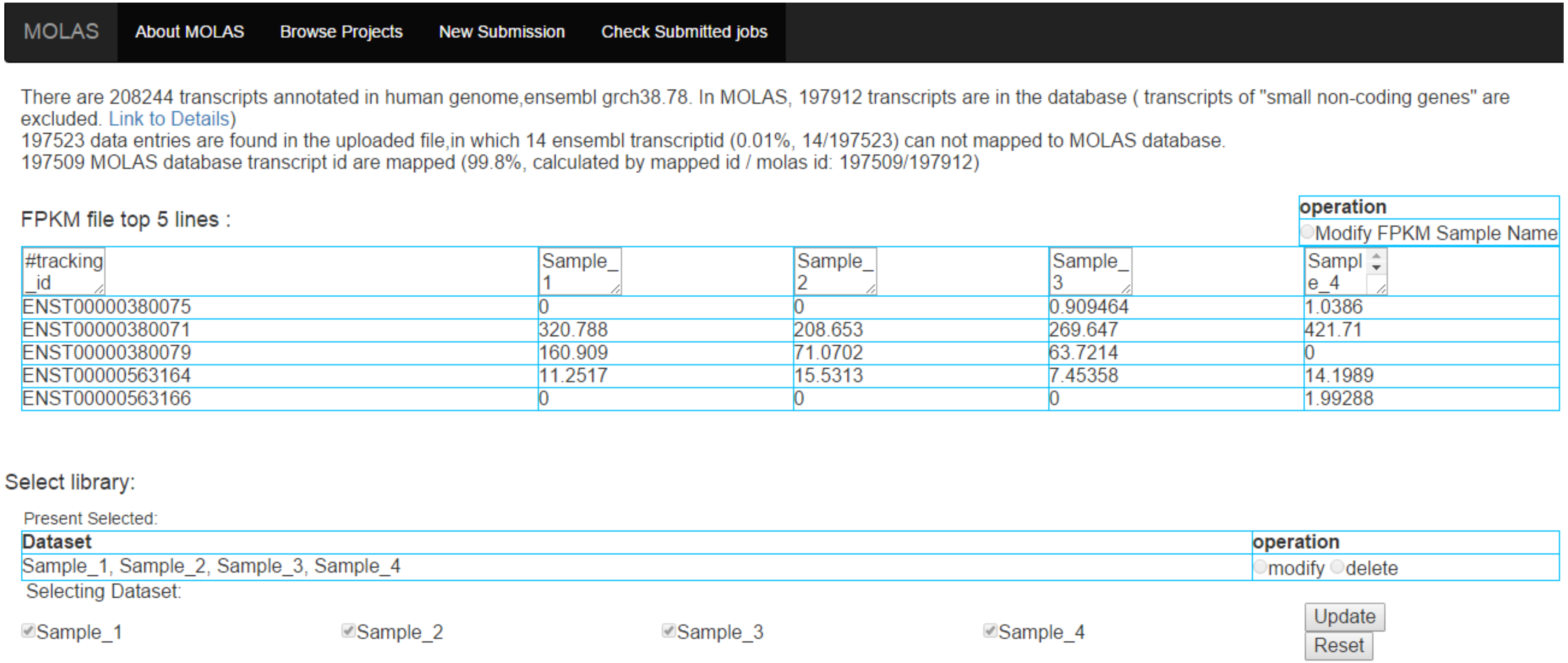
By gene: ○genesymbol(ARF5) ○EnsemblGeneID(ENSMUSG00000005320 or ENSG00000004059)

For human study

□combined with DEMO DB data library.

Submit ✔ Clear All

New Submission



There are 208244 transcripts annotated in human genome,ensembl grch38.78. In MOLAS, 197912 transcripts are in the database (transcripts of "small non-coding genes" are excluded. Link to Details)
197523 data entries are found in the uploaded file,in which 14 ensembl transcriptid (0.01%, 14/197523) can not mapped to MOLAS database.
197509 MOLAS database transcript id are mapped (99.8%, calculated by mapped id / molas id: 197509/197912)

FPKM file top 5 lines :

operation
○Modify FPKM Sample Name

#tracking_id	Sample_1	Sample_2	Sample_3	Sample_4
ENST00000380075	0	0	0.909464	1.0386
ENST00000380071	320.788	208.653	269.647	421.71
ENST00000380079	160.909	71.0702	63.7214	0
ENST00000563164	11.2517	15.5313	7.45358	14.1989
ENST00000563166	0	0	0	1.99288

Select library:

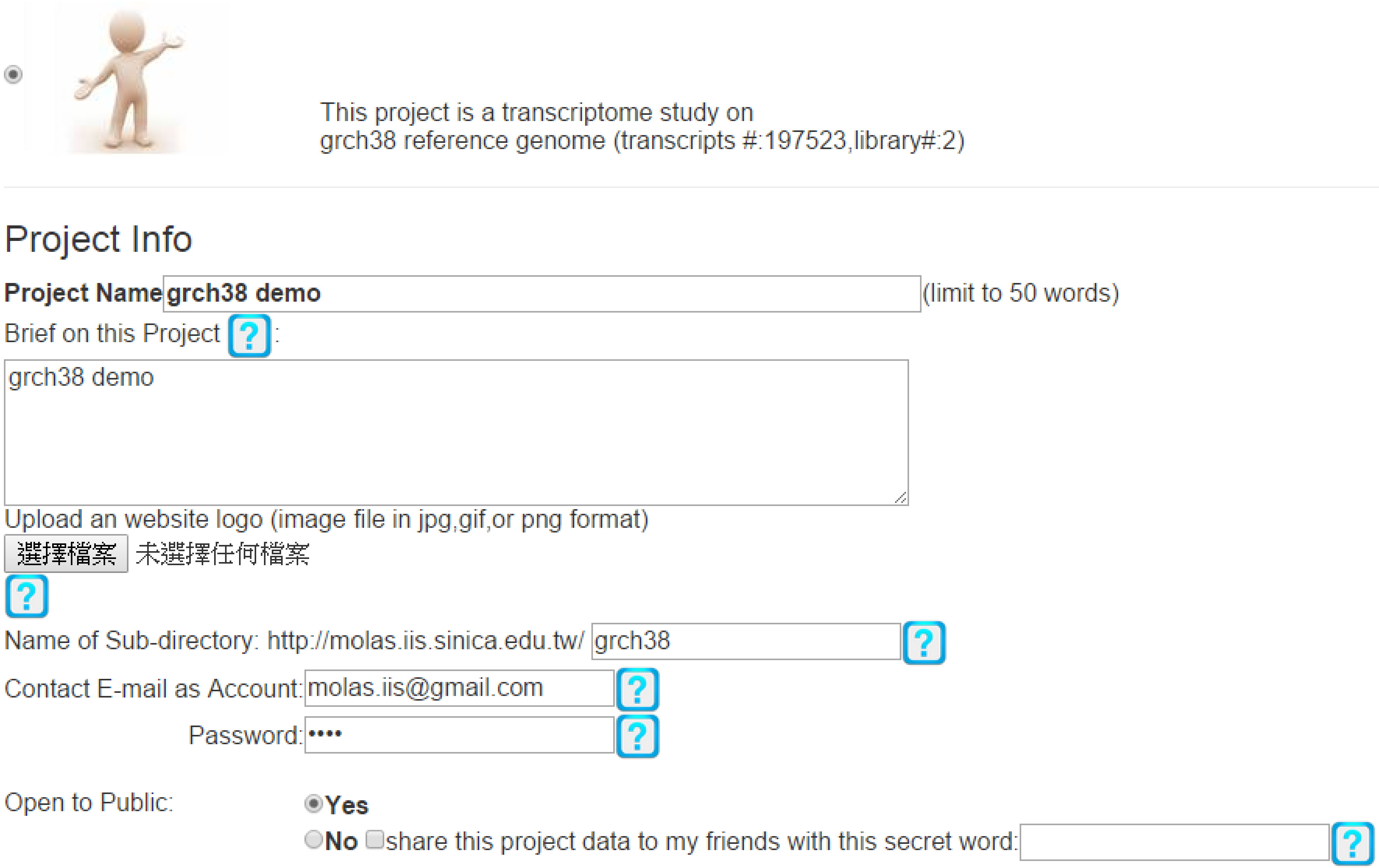
Present Selected:

Dataset	operation
Sample_1, Sample_2, Sample_3, Sample_4	○modify ○delete

Selecting Dataset:

☑Sample_1 ☑Sample_2 ☑Sample_3 ☑Sample_4

Project Profile



● This project is a transcriptome study on
grch38 reference genome (transcripts #:197523,library#:2)

Project Info

Project Name **grch38 demo** (limit to 50 words)

Brief on this Project ?:

grch38 demo

Upload an website logo (image file in jpg,gif,or png format)

選擇檔案 未選擇任何檔案

?

Name of Sub-directory: http://molas.iis.sinica.edu.tw/ grch38 ?

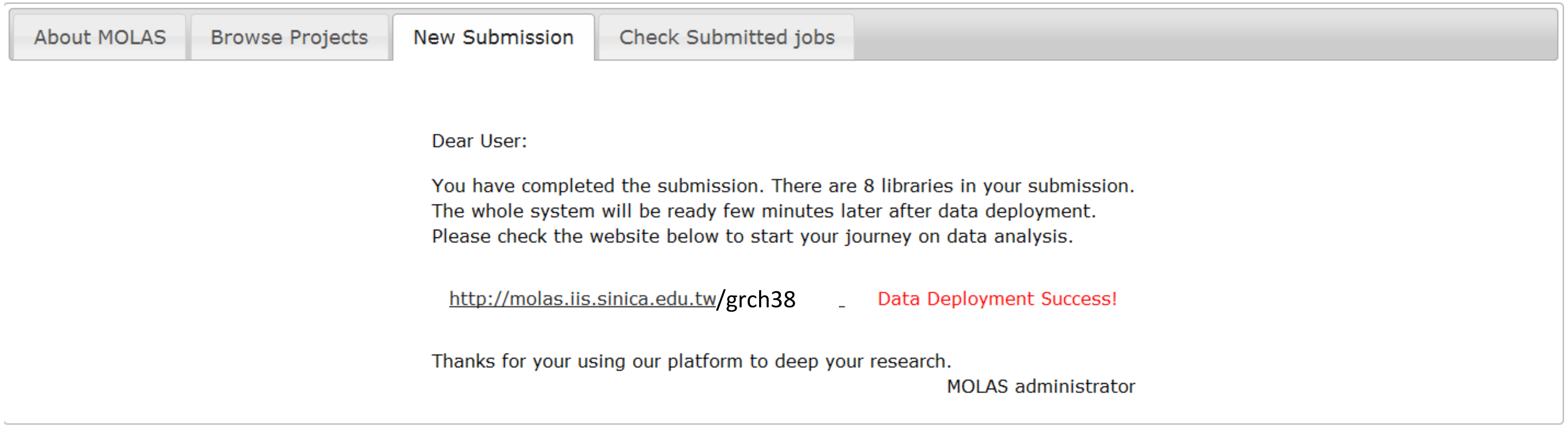
Contact E-mail as Account: molas.iis@gmail.com ?

Password: •••• ?

Open to Public: ● **Yes**

○ **No** □ share this project data to my friends with this secret word: ?

Deployment Success



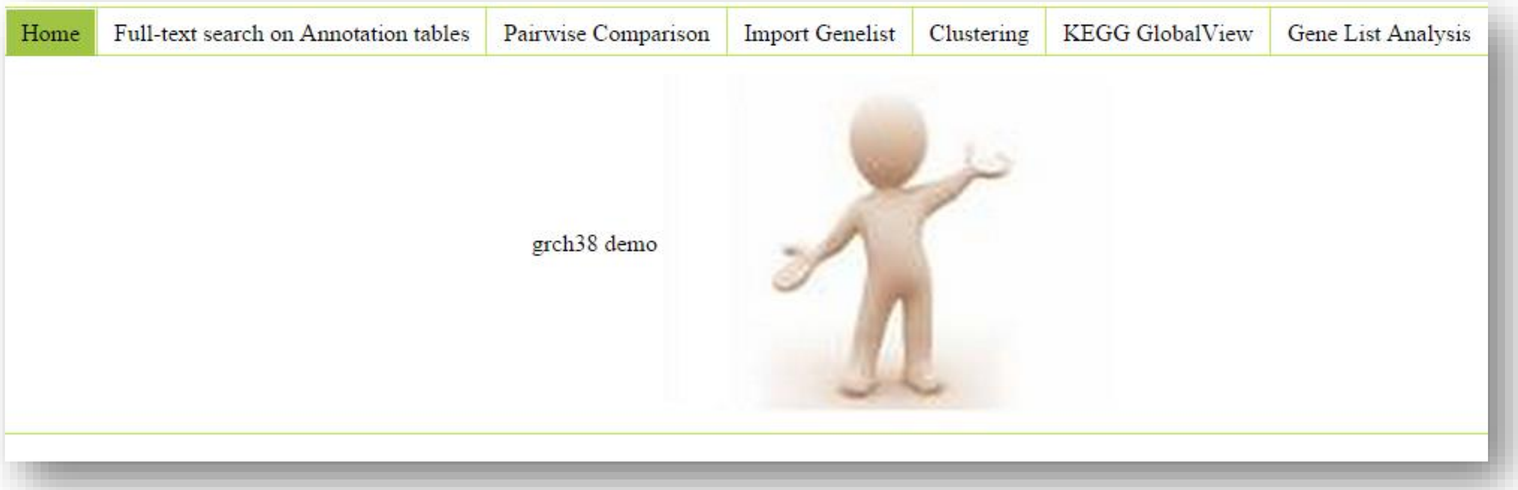
Dear User:

You have completed the submission. There are 8 libraries in your submission.
The whole system will be ready few minutes later after data deployment.
Please check the website below to start your journey on data analysis.

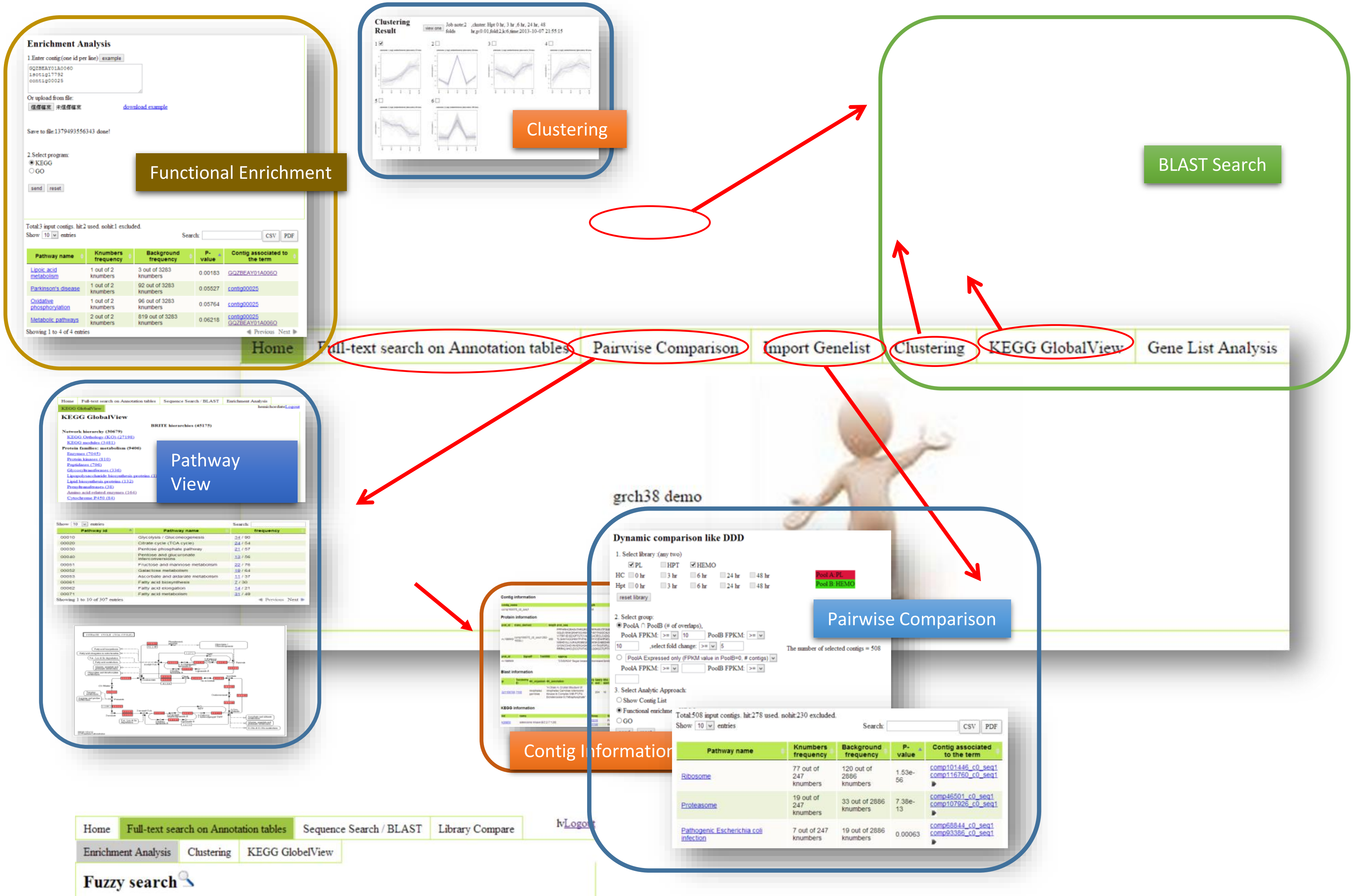
http://molas.iis.sinica.edu.tw/grch38 - Data Deployment Success!

Thanks for your using our platform to deep your research.
MOLAS administrator

Browse project and …….



http://molas.iis.sinica.edu.tw/grch38/



Enrichment Analysis

1.Enter contig(one id per line) example

```
GQZBEAY01A006O
isotig17792
contig00025
```

Or upload from file: 選擇檔案 未選擇檔案 download example

Save to file:1379493556343 done!

2.Select program:
● KEGG
○ GO

send reset

Total:3 input contigs. hit:2 used. nohit:1 excluded.
Show 10 entries Search: CSV PDF

Pathway name	Knumbers frequency	Background frequency	P-value	Contig associated to the term
Lipoic acid metabolism	1 out of 2 knumbers	3 out of 3283 knumbers	0.00183	GQZBEAY01A006O
Parkinson's disease	1 out of 2 knumbers	92 out of 3283 knumbers	0.05527	contig00025
Oxidative phosphorylation	1 out of 2 knumbers	96 out of 3283 knumbers	0.05764	contig00025
Metabolic pathways	2 out of 2 knumbers	819 out of 3283 knumbers	0.06218	contig00025 GQZBEAY01A006O

Showing 1 to 4 of 4 entries ◀ Previous Next ▶

Clustering Result view one Job note:2 folds ,cluster: Hpt 0 hr, 3 hr ,6 hr, 24 hr, 48 hr,p:0.01,fold:2,k:6,time:2013-10-07 21:55:15

1 ☑ 2 □ 3 □ 4 □ 5 □ 6 □

Home | Full-text search on Annotation tables | Sequence Search / BLAST | Enrichment Analysis
KEGG GlobalView hemichordateLogout

KEGG GlobalView

BRITE hierarchies (45175)

Network hierarchy (30679)
- KEGG Orthology (KO) (27198)
- KEGG modules (3481)

Protein families: metabolism (9406)
- Enzymes (7045)
- Protein kinases (810)
- Peptidases (796)
- Glycosyltransferases (336)
- Lipopolysaccharide biosynthesis proteins (1
- Lipid biosynthesis proteins (132)
- Prenyltransferases (38)
- Amino acid related enzymes (164)
- Cytochrome P450 (84)

Show 10 entries Search:

Pathway id	Pathway name	frequency
00010	Glycolysis / Gluconeogenesis	34 / 90
00020	Citrate cycle (TCA cycle)	24 / 54
00030	Pentose phosphate pathway	21 / 57
00040	Pentose and glucuronate interconversions	13 / 56
00051	Fructose and mannose metabolism	22 / 76
00052	Galactose metabolism	19 / 64
00053	Ascorbate and aldarate metabolism	11 / 37
00061	Fatty acid biosynthesis	7 / 30
00062	Fatty acid elongation	14 / 21
00071	Fatty acid metabolism	31 / 49

Showing 1 to 10 of 307 entries ◀ Previous Next ▶

CITRATE CYCLE (TCA CYCLE)

Contig information

Protein information

Blast information

KEGG information

Dynamic comparison like DDD

1. Select library :(any two)

	☑ PL	□ HPT	☑ HEMO		
HC	□ 0 hr	□ 3 hr	□ 6 hr	□ 24 hr	□ 48 hr
Hpt	□ 0 hr	□ 3 hr	□ 6 hr	□ 24 hr	□ 48 hr

reset library

Pool A: PL
Pool B: HEMO

2. Select group:
● PoolA ∩ PoolB (# of overlaps),
PoolA FPKM: >= 10 PoolB FPKM: >= 10 ,select fold change: >= 5
○ PoolA Expressed only (FPKM value in PoolB=0, # contigs)
PoolA FPKM: >= PoolB FPKM: >=

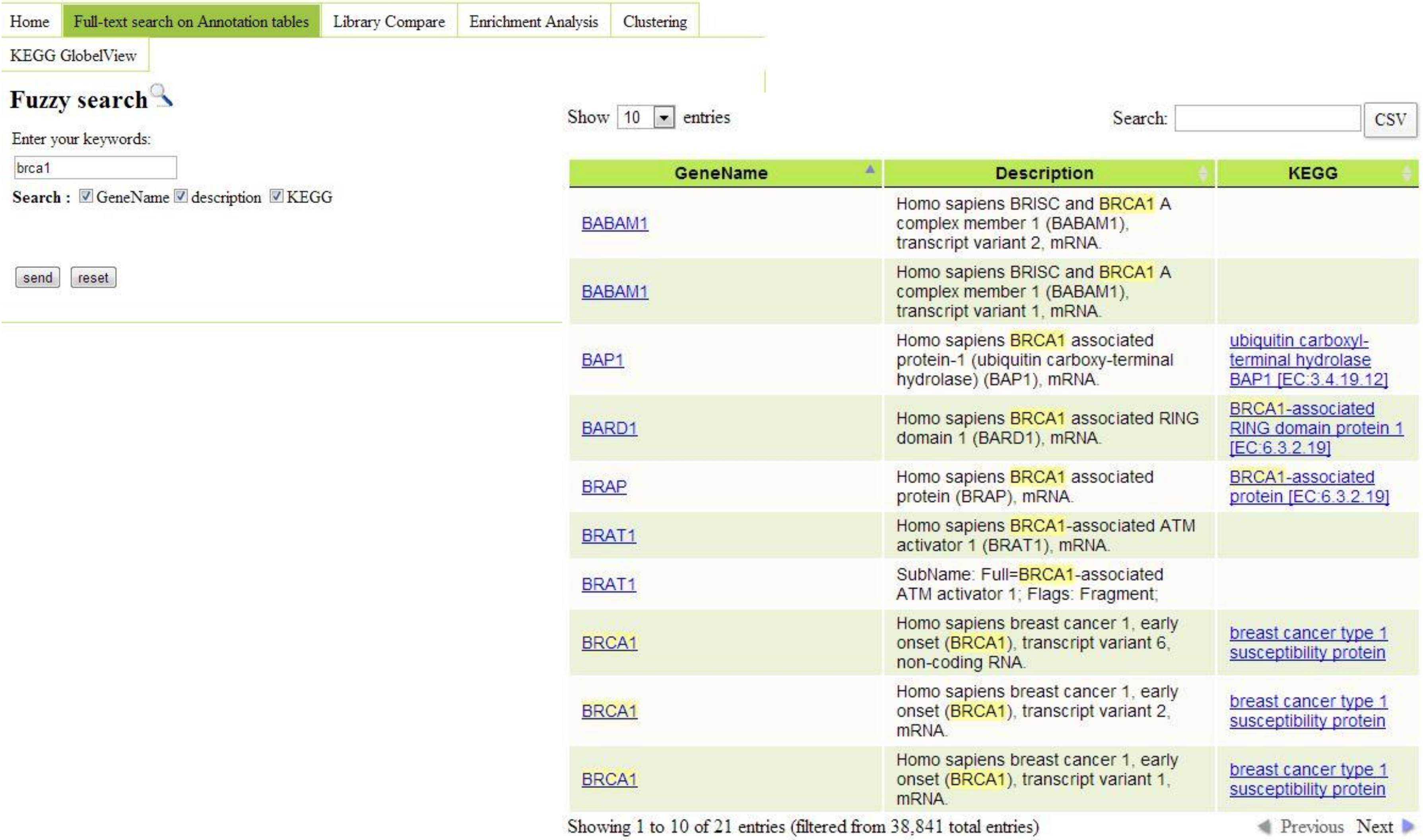
The number of selected contigs = 508

3. Select Analytic Approach:
○ Show Contig List
● Functional enrichme
○ GO

Total:508 input contigs. hit:278 used. nohit:230 excluded.
Show 10 entries Search: CSV PDF

Pathway name	Knumbers frequency	Background frequency	P-value	Contig associated to the term
Ribosome	77 out of 247 knumbers	120 out of 2886 knumbers	1.53e-56	comp101446_c0_seq1 comp116760_c0_seq1 ▶
Proteasome	19 out of 247 knumbers	33 out of 2886 knumbers	7.38e-13	comp46501_c0_seq1 comp107926_c0_seq1 ▶
Pathogenic Escherichia coli infection	7 out of 247 knumbers	19 out of 2886 knumbers	0.00063	comp68844_c0_seq1 comp93386_c0_seq1 ▶

Fuzzy Search



Fuzzy search

Enter your keywords:

brca1

Search : ☑ GeneName ☑ description ☑ KEGG

Show 10 entries

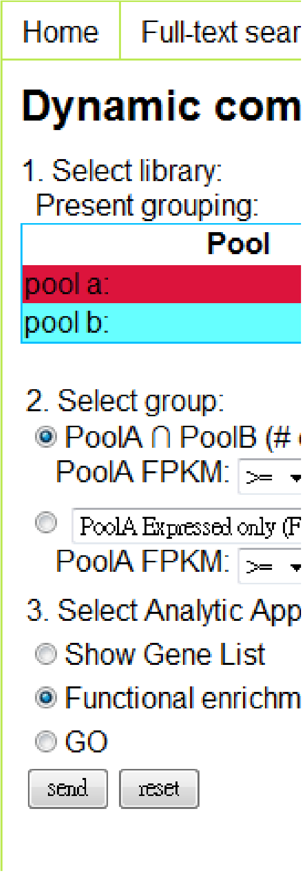
Search:

GeneName	Description	KEGG
BABAM1	Homo sapiens BRISC and BRCA1 A complex member 1 (BABAM1), transcript variant 2, mRNA.	
BABAM1	Homo sapiens BRISC and BRCA1 A complex member 1 (BABAM1), transcript variant 1, mRNA.	
BAP1	Homo sapiens BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase) (BAP1), mRNA.	ubiquitin carboxyl-terminal hydrolase BAP1 [EC:3.4.19.12]
BARD1	Homo sapiens BRCA1 associated RING domain 1 (BARD1), mRNA.	BRCA1-associated RING domain protein 1 [EC:6.3.2.19]
BRAP	Homo sapiens BRCA1 associated protein (BRAP), mRNA.	BRCA1-associated protein [EC:6.3.2.19]
BRAT1	Homo sapiens BRCA1-associated ATM activator 1 (BRAT1), mRNA.	
BRAT1	SubName: Full=BRCA1-associated ATM activator 1; Flags: Fragment;	
BRCA1	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 6, non-coding RNA.	breast cancer type 1 susceptibility protein
BRCA1	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 2, mRNA.	breast cancer type 1 susceptibility protein
BRCA1	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 1, mRNA.	breast cancer type 1 susceptibility protein

Showing 1 to 10 of 21 entries (filtered from 38,841 total entries)

Pairwise Comparison

Select libra



Home | Full-text sear

Dynamic com

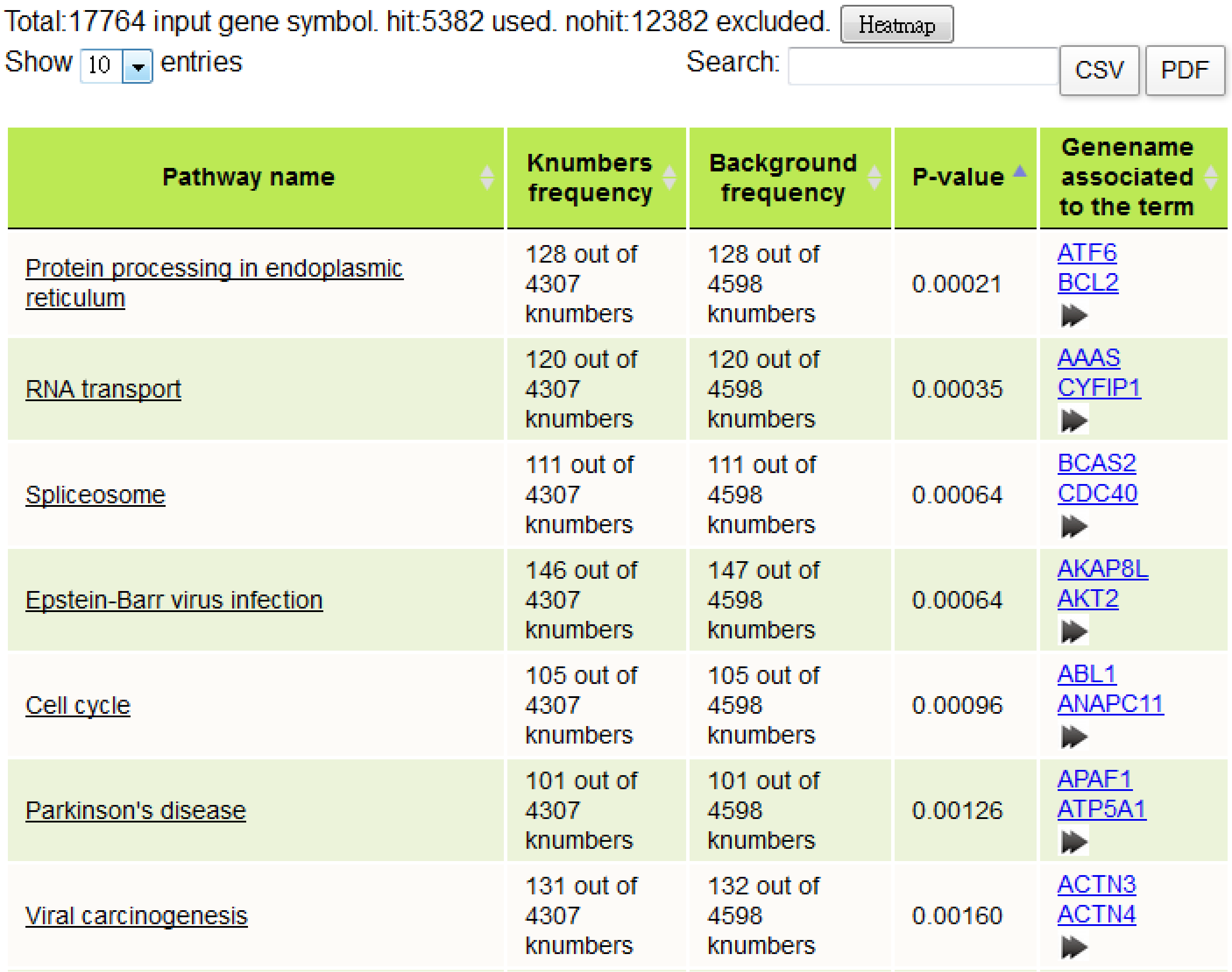
1. Select library:
Present grouping:

Pool
pool a:
pool b:

2. Select group:
● PoolA ∩ PoolB (#
PoolA FPKM: >=
○ PoolA Expressed only (F
PoolA FPKM: >=

3. Select Analytic App
○ Show Gene List
● Functional enrichm
○ GO

send reset

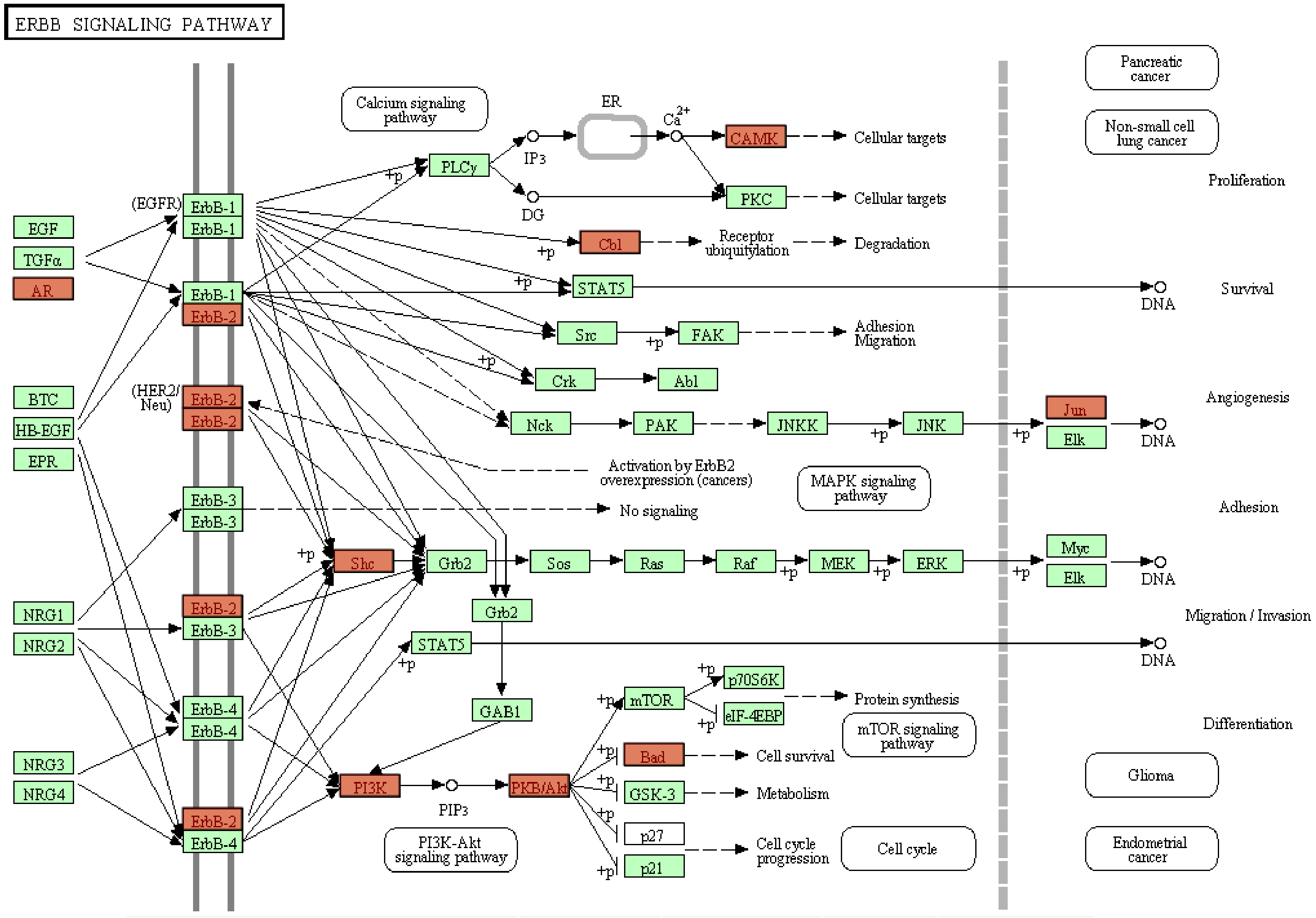


Total:17764 input gene symbol. hit:5382 used. nohit:12382 excluded. Heatmap

Show 10 entries Search: CSV PDF

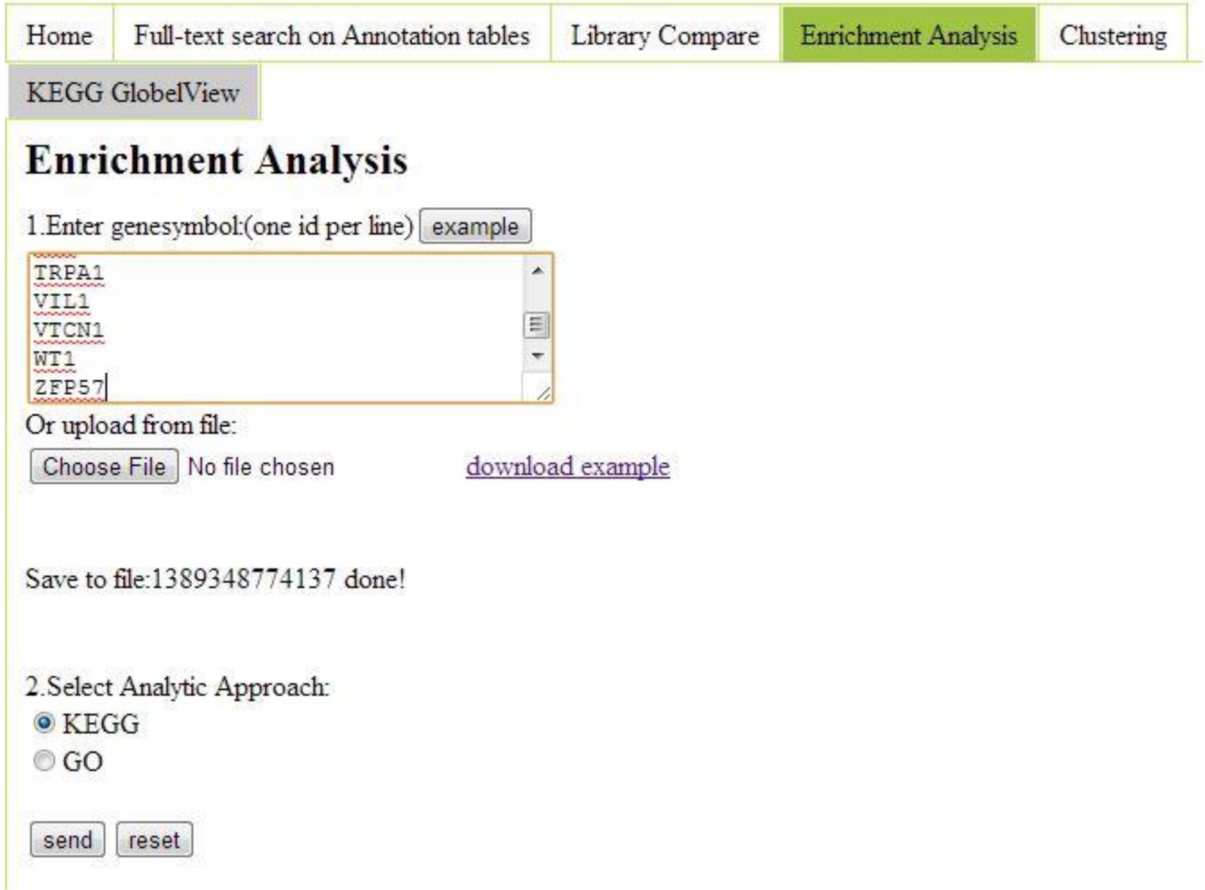
Pathway name	Knumbers frequency	Background frequency	P-value	Genename associated to the term
Protein processing in endoplasmic reticulum	128 out of 4307 knumbers	128 out of 4598 knumbers	0.00021	ATF6 BCL2 ⏩
RNA transport	120 out of 4307 knumbers	120 out of 4598 knumbers	0.00035	AAAS CYFIP1 ⏩
Spliceosome	111 out of 4307 knumbers	111 out of 4598 knumbers	0.00064	BCAS2 CDC40 ⏩
Epstein-Barr virus infection	146 out of 4307 knumbers	147 out of 4598 knumbers	0.00064	AKAP8L AKT2 ⏩
Cell cycle	105 out of 4307 knumbers	105 out of 4598 knumbers	0.00096	ABL1 ANAPC11 ⏩
Parkinson's disease	101 out of 4307 knumbers	101 out of 4598 knumbers	0.00126	APAF1 ATP5A1 ⏩
Viral carcinogenesis	131 out of 4307 knumbers	132 out of 4598 knumbers	0.00160	ACTN3 ACTN4 ⏩

KEGG Pathway



Enrichment Analysis

Insert a list of interesting genes to see which pathway they are involved.



Enrichment Analysis

1.Enter genesymbol:(one id per line) example

```
TRPA1
VIL1
VTCN1
WT1
ZFP57
```

Or upload from file:

Choose File No file chosen download example

Save to file:1389348774137 done!

2.Select Analytic Approach:

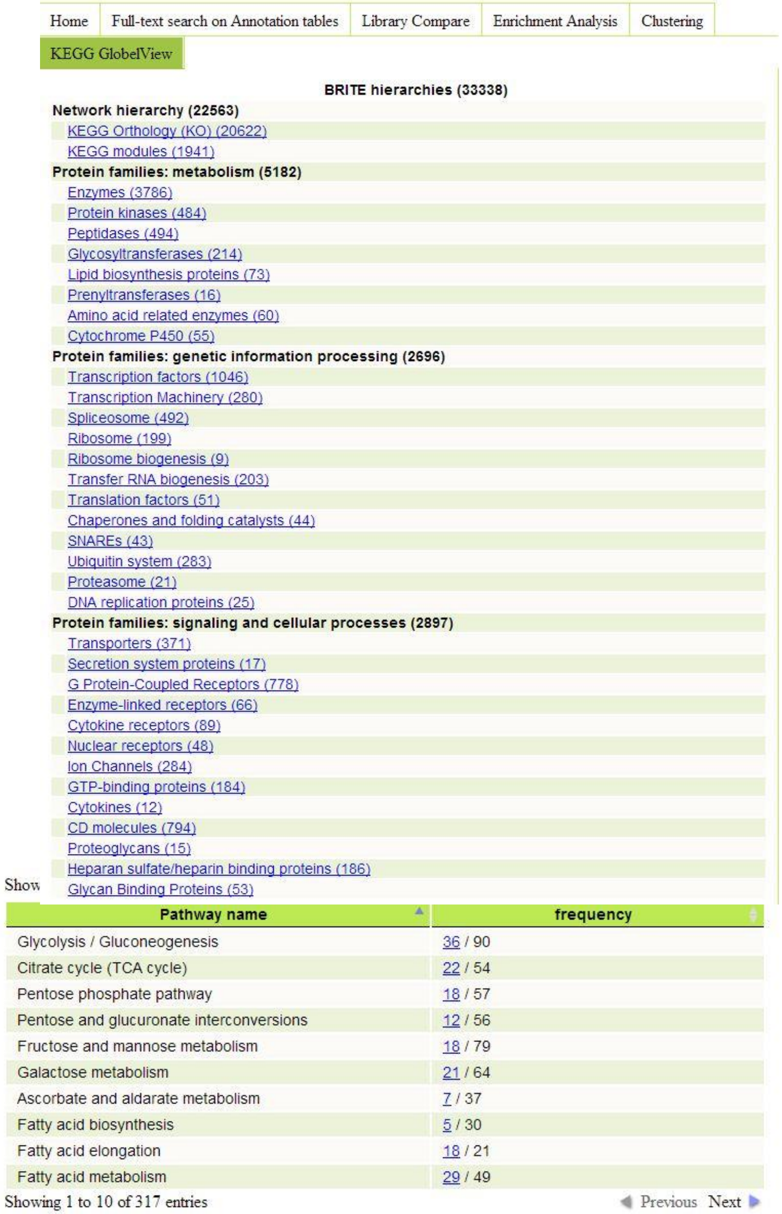
● KEGG

○ GO

send reset

KEGG Global View

KEGG Global View provide an canonical pathway-type overview of genes involved in a particular KEGG pathway.



BRITE hierarchies (33338)

Network hierarchy (22563)
KEGG Orthology (KO) (20622)
KEGG modules (1941)

Protein families: metabolism (5182)
Enzymes (3786)
Protein kinases (484)
Peptidases (494)
Glycosyltransferases (214)
Lipid biosynthesis proteins (73)
Prenyltransferases (16)
Amino acid related enzymes (60)
Cytochrome P450 (55)

Protein families: genetic information processing (2696)
Transcription factors (1046)
Transcription Machinery (280)
Spliceosome (492)
Ribosome (199)
Ribosome biogenesis (9)
Transfer RNA biogenesis (203)
Translation factors (51)
Chaperones and folding catalysts (44)
SNAREs (43)
Ubiquitin system (283)
Proteasome (21)
DNA replication proteins (25)

Protein families: signaling and cellular processes (2897)
Transporters (371)
Secretion system proteins (17)
G Protein-Coupled Receptors (778)
Enzyme-linked receptors (66)
Cytokine receptors (89)
Nuclear receptors (48)
Ion Channels (284)
GTP-binding proteins (184)
Cytokines (12)
CD molecules (794)
Proteoglycans (15)
Heparan sulfate/heparin binding proteins (186)
Glycan Binding Proteins (53)

Show

Pathway name	frequency
Glycolysis / Gluconeogenesis	36 / 90
Citrate cycle (TCA cycle)	22 / 54
Pentose phosphate pathway	18 / 57
Pentose and glucuronate interconversions	12 / 56
Fructose and mannose metabolism	18 / 79
Galactose metabolism	21 / 64
Ascorbate and aldarate metabolism	7 / 37
Fatty acid biosynthesis	5 / 30
Fatty acid elongation	18 / 21
Fatty acid metabolism	29 / 49

Showing 1 to 10 of 317 entries

Demo

Hands on practice on MOLAS

- Build your own project
- Browse project and conduct a study

http://molas.iis.sinica.edu.tw/human_grch38_demo/

MOLAS

Multi-Omics onLine Analysis System

What to do if you have no replicates?

Suggestions from edgeR authors

- Be satisfied with a descriptive analysis, that might include an MDS plot and an analysis of fold changes. Do not claim a significance statistical analysis.
    - In edgeR (empirically): Simply pick a reasonable dispersion value, based on your experience with similar data, and use that for detecting differentially expressed transcripts.

      0.4 human data (genetically "not" identical)

      0.1 for "genetically identical" strains of model organisms

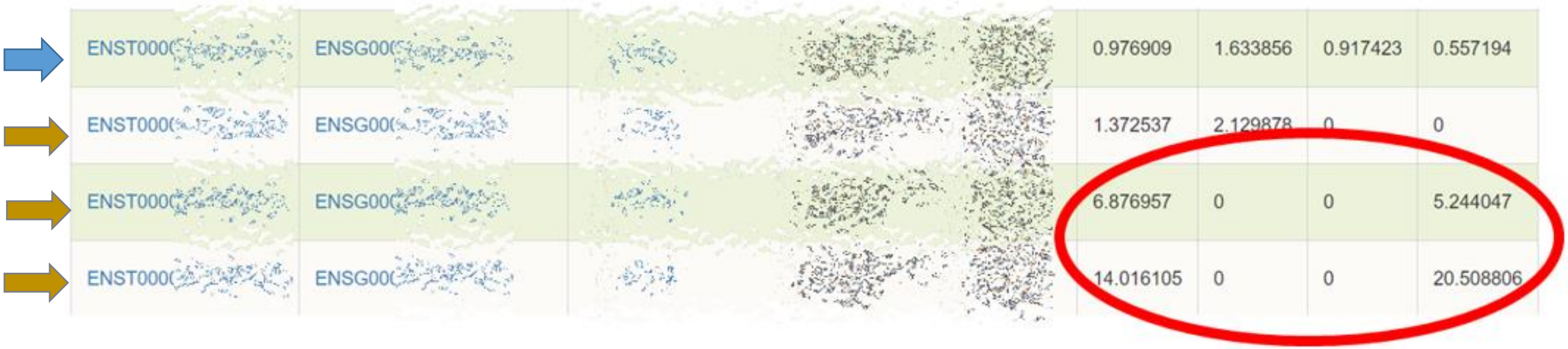
      0.01 for technical replicates
    - Simulation data: NOISeq

https://f1000research.com/articles/5-1438/v2

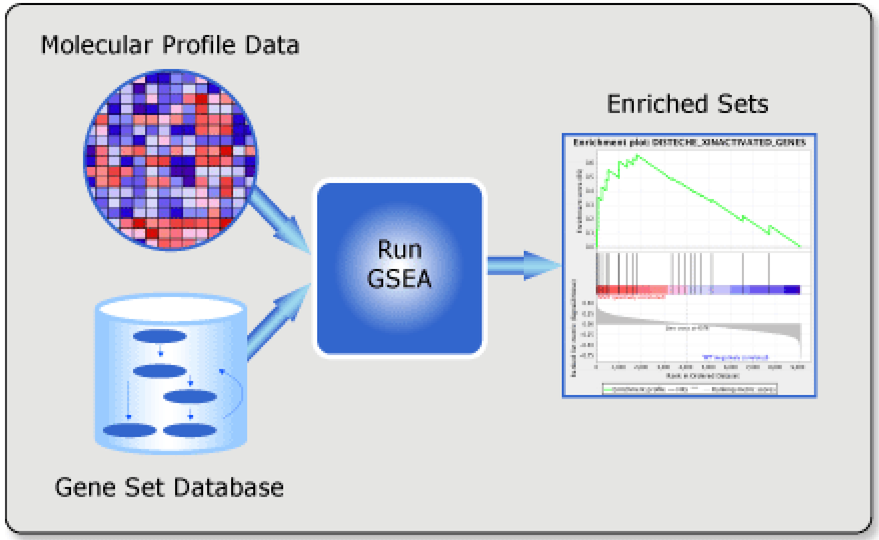
edgeR paper http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/
menu http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Genes with different transcripts…..

Protein coding

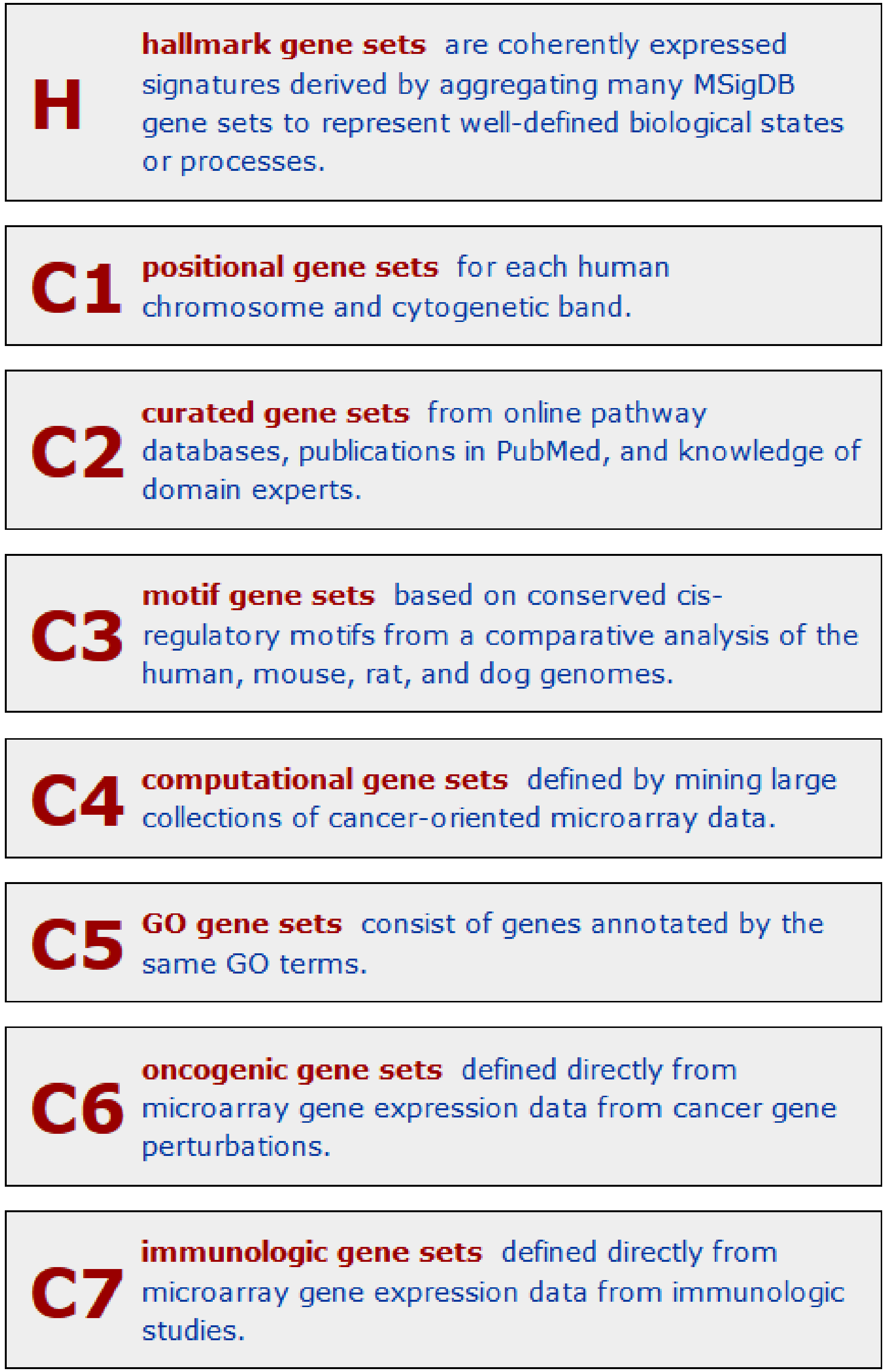


					Drug resistant			Drug resistant
ENST000[redacted]	ENSG00[redacted]	[redacted]	[redacted]	[redacted]	0.976909	1.633856	0.917423	0.557194
ENST000[redacted]	ENSG00[redacted]	[redacted]	[redacted]	[redacted]	1.372537	2.129878	0	0
ENST000[redacted]	ENSG00[redacted]	[redacted]	[redacted]	[redacted]	6.876957	0	0	5.244047
ENST000[redacted]	ENSG00[redacted]	[redacted]	[redacted]	[redacted]	14.016105	0	0	20.508806



Gene Set Enrichment Analysis: a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).
Ref: Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550)

the Molecular Signatures Database (MSigDB), a collection of annotated gene sets for use with GSEA software.
Ref: Liberzon, et al. (2011, Bionformatics), Liberzon, et al. (2015, Cell Systems)



H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

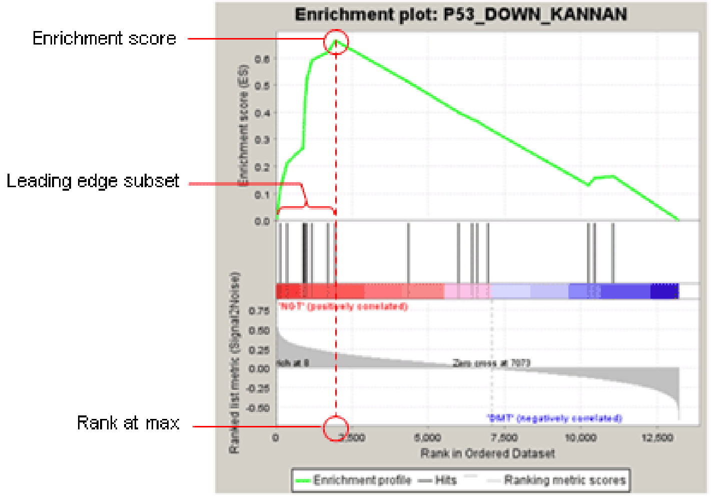
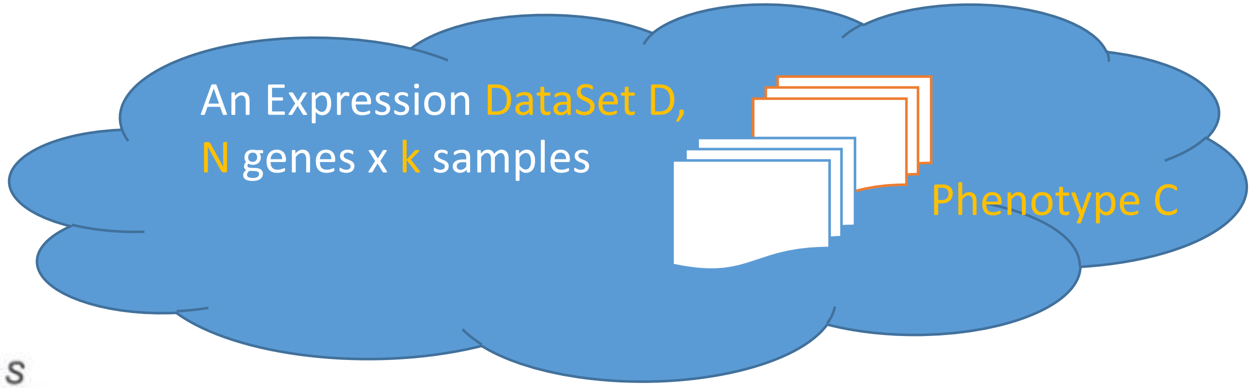


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

……. the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes.

……. Gene sets with a distinct peak at the beginning (such as the one shown here) or end of the ranked list are generally the most interesting.



An Expression DataSet D,
N genes x k samples

Phenotype C

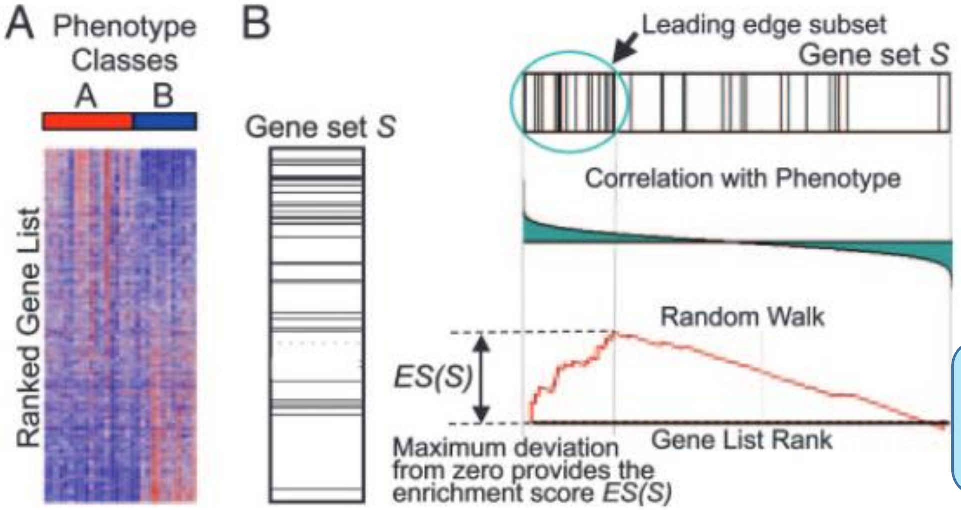
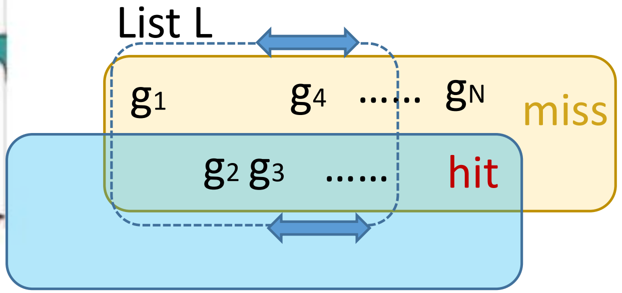


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Tamayo, et al. (2005, PNAS 102, 15545-15550)

Gene List L➔ in ranked order L={g_1,g_2,g_3, …….g_N} and the correlation to phenotype C, i.e., r(g_1)=r_1



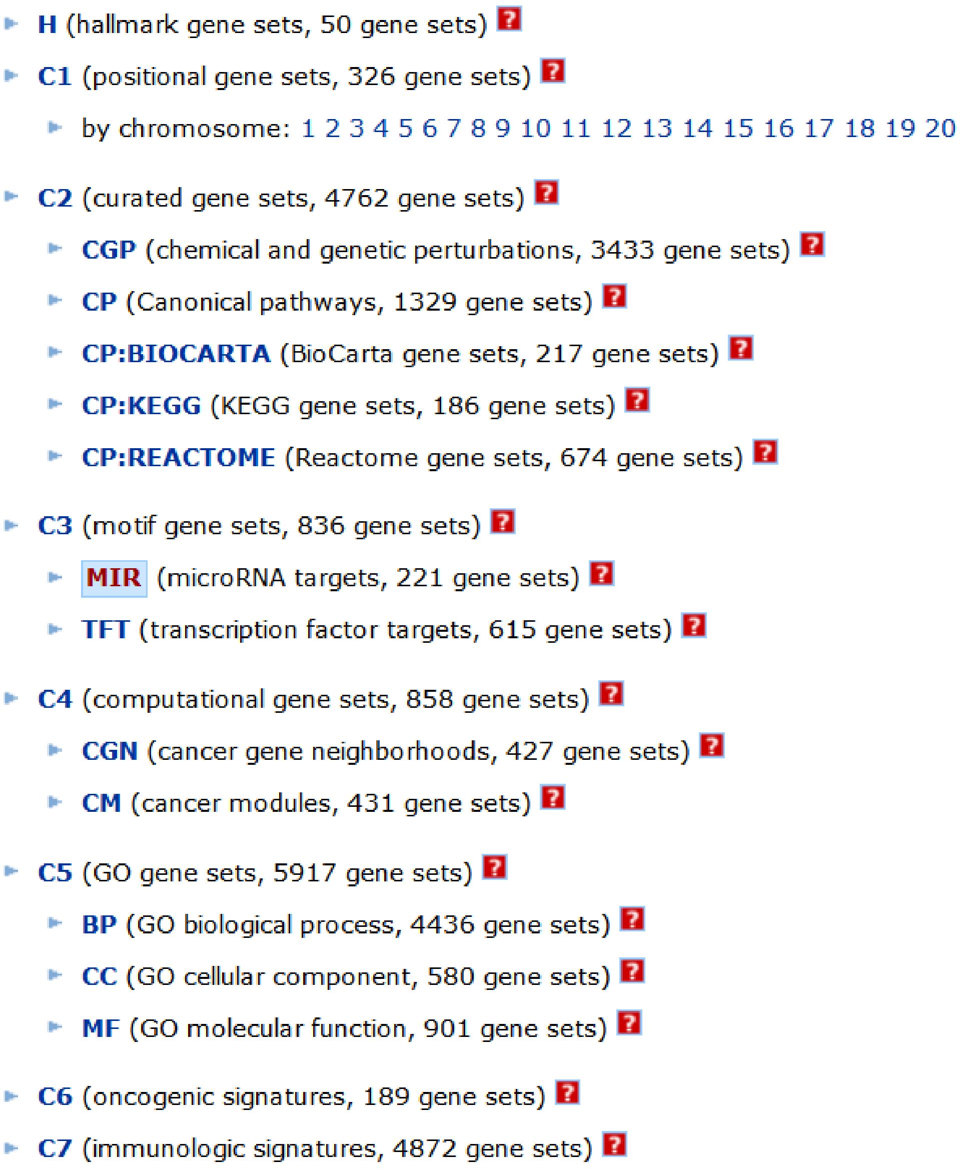
Gene Set S
(at least 15 members observed in the DataSet D)

$$P_{\text{hit}}(S,i)=\sum_{\substack{g_j\in S\\ j\le i}}\frac{|r_j|^p}{N_R},\quad \text{where } N_R=\sum_{g_j\in S}|r_j|^p$$

$$P_{\text{miss}}(S,i)=\sum_{\substack{g_j\notin S\\ j\le i}}\frac{1}{(N-N_H)}.$$

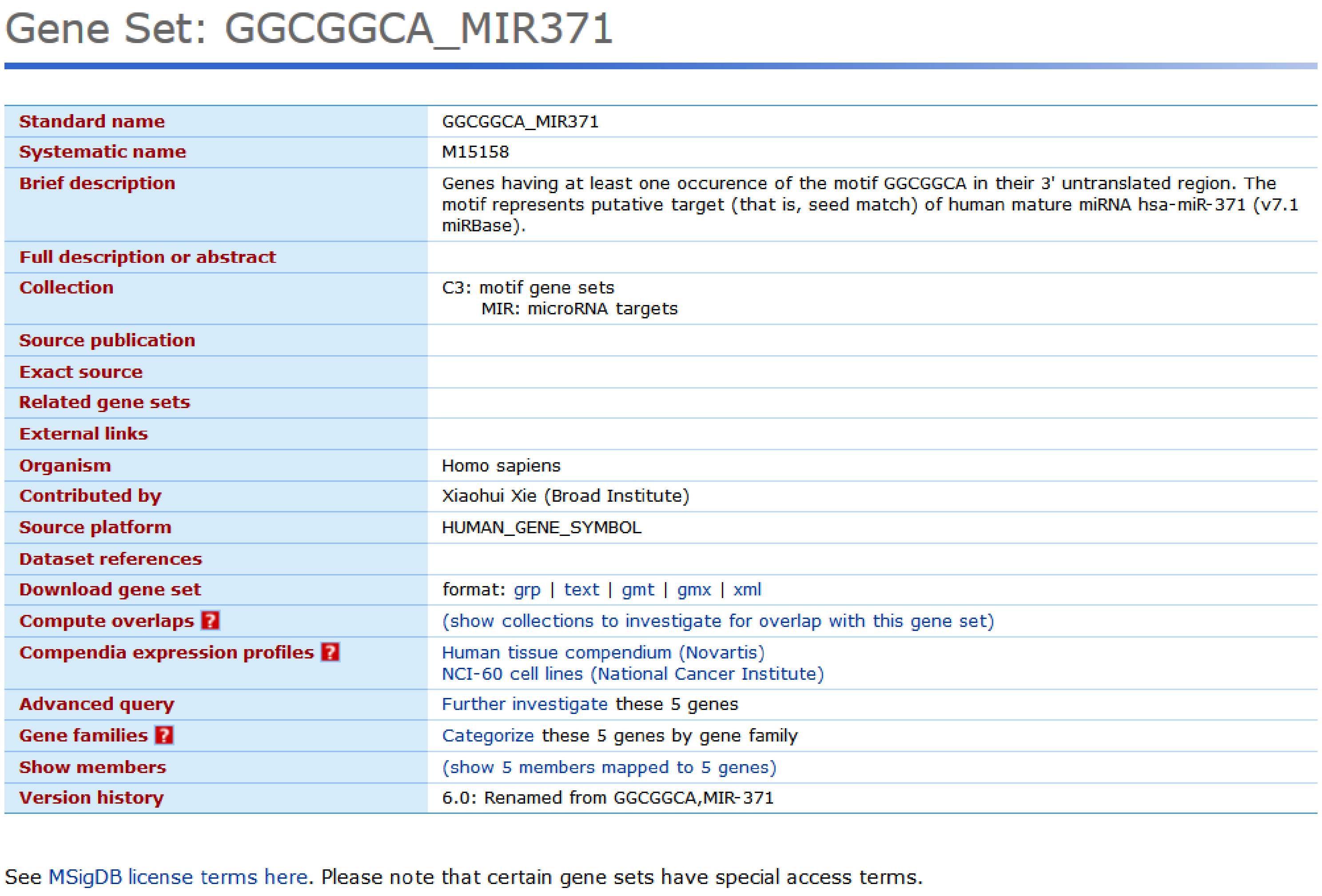
ES: max (P_{hit} - P_{miss}) from zero

a running-sum statistic



- **H** (hallmark gene sets, 50 gene sets)
- **C1** (positional gene sets, 326 gene sets)
  - by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- **C2** (curated gene sets, 4762 gene sets)
  - **CGP** (chemical and genetic perturbations, 3433 gene sets)
  - **CP** (Canonical pathways, 1329 gene sets)
  - **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets)
  - **CP:KEGG** (KEGG gene sets, 186 gene sets)
  - **CP:REACTOME** (Reactome gene sets, 674 gene sets)
- **C3** (motif gene sets, 836 gene sets)
  - **MIR** (microRNA targets, 221 gene sets)
  - **TFT** (transcription factor targets, 615 gene sets)
- **C4** (computational gene sets, 858 gene sets)
  - **CGN** (cancer gene neighborhoods, 427 gene sets)
  - **CM** (cancer modules, 431 gene sets)
- **C5** (GO gene sets, 5917 gene sets)
  - **BP** (GO biological process, 4436 gene sets)
  - **CC** (GO cellular component, 580 gene sets)
  - **MF** (GO molecular function, 901 gene sets)
- **C6** (oncogenic signatures, 189 gene sets)
- **C7** (immunologic signatures, 4872 gene sets)

http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2



Gene Set: GGCGGCA_MIR371

Standard name	GGCGGCA_MIR371
Systematic name	M15158
Brief description	Genes having at least one occurence of the motif GGCGGCA in their 3' untranslated region. The motif represents putative target (that is, seed match) of human mature miRNA hsa-miR-371 (v7.1 miRBase).
Full description or abstract	
Collection	C3: motif gene sets MIR: microRNA targets
Source publication	
Exact source	
Related gene sets	
External links	
Organism	Homo sapiens
Contributed by	Xiaohui Xie (Broad Institute)
Source platform	HUMAN_GENE_SYMBOL
Dataset references	
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps	(show collections to investigate for overlap with this gene set)
Compendia expression profiles	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 5 genes
Gene families	Categorize these 5 genes by gene family
Show members	(show 5 members mapped to 5 genes)
Version history	6.0: Renamed from GGCGGCA,MIR-371

See MSigDB license terms here. Please note that certain gene sets have special access terms.

http://software.broadinstitute.org/gsea/msigdb/cards/GGCGGCA_MIR371.html

epi-MOLAS

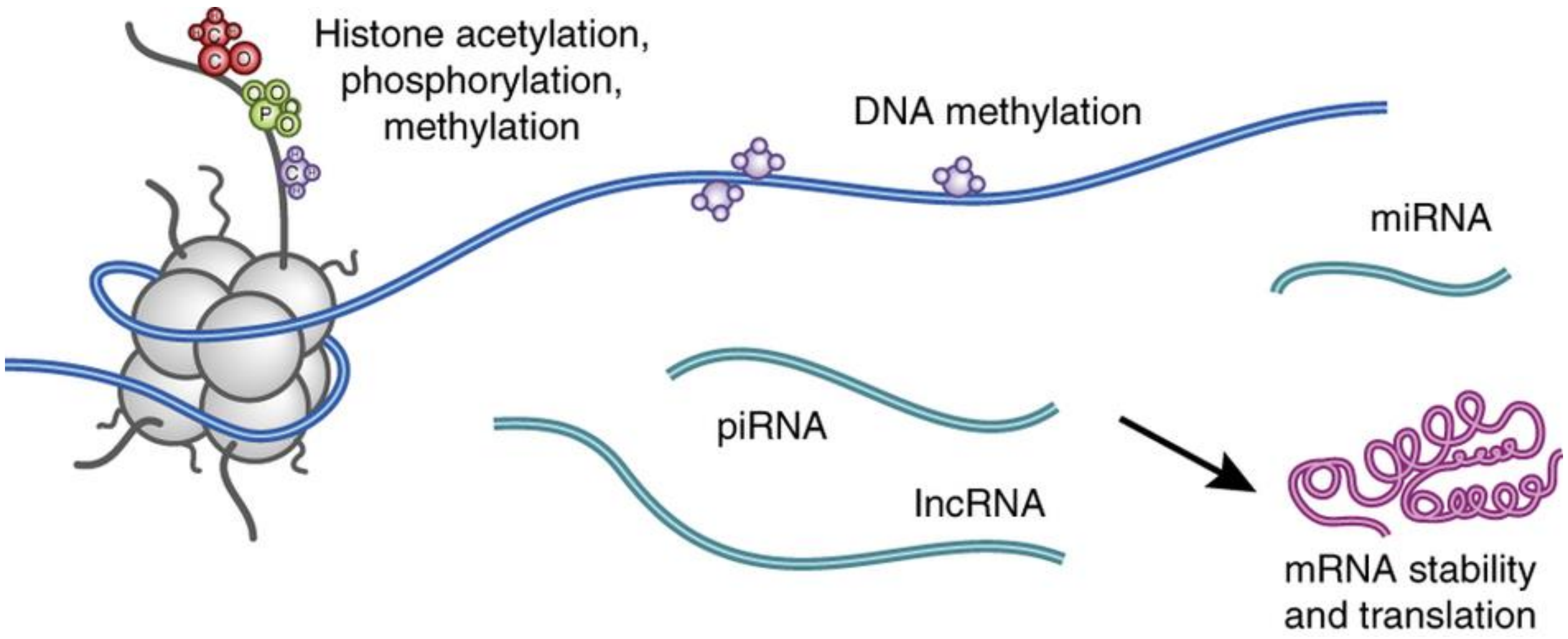
Epi-genoMics OnLine Analysis System

Shu-Hwa Chen

Institute of Information Science
Academia Sinica, Taiwan
2018/12

Epigenetic Modification

- Epigenetic Modification: Reversible modifications on genome components to affect gene expression without changing the DNA sequence



One Genome

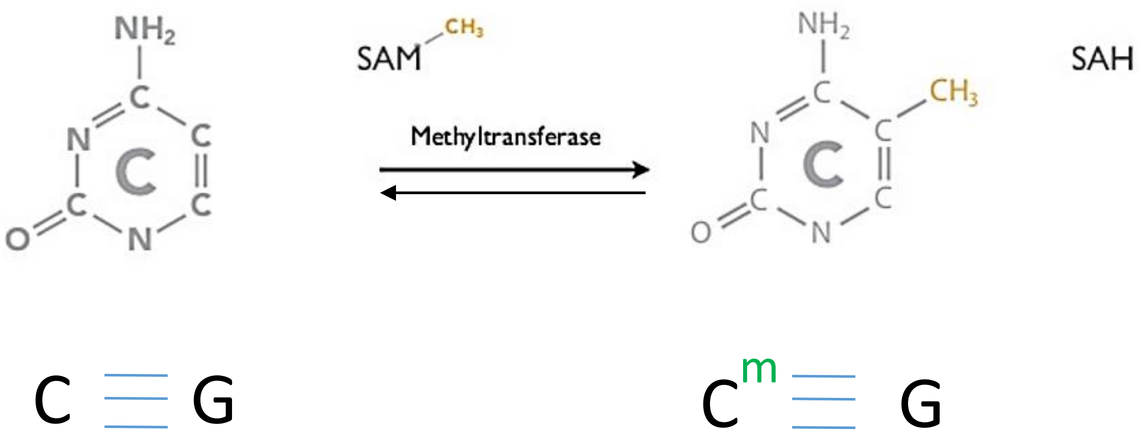
↓

Many Phenotypes

Adopted from McEwen BS et al., Nature Neuroscience 18, 1353–1363 (2015)

Methylated Cytosine: the Fifth Base

The most common and stable epigenetic marks in nucleotide level



- Involved in
  - Genomic imprinting
  - Cell Fate Determination / Reprogramming
  - Transposon genes silencing

- In vertebrates, 1-6% of genomic cytosine are methylated
- In plants, the proportion of methylated cytosine is even higher
- But……..

Whole Genome Shotgun Bisulfite Sequencing

Bisulfite Conversion

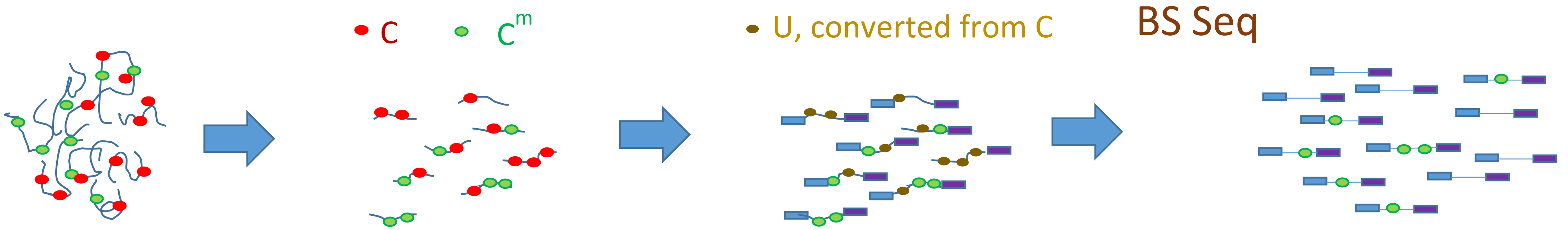
DNA:

5’ – T C G C C G G T A C - $\xrightarrow{\text{bisulfite}}$ T U G C U G G T A U $\xrightarrow{\text{PCR}}$ T $\boxed{\begin{matrix}T\\a\end{matrix}}$ G C $\boxed{\begin{matrix}T\\a\end{matrix}}$ G G T A $\boxed{\begin{matrix}T\\a\end{matrix}}$

a c g c c a t

5’ – T C^m G C C^m G G T A C - $\xrightarrow{\text{bisulfite}}$ T C^m G C C^m G G T A U $\xrightarrow{\text{PCR}}$ T C G C C G G T A $\boxed{\begin{matrix}T\\a\end{matrix}}$

a g c g g c c a t



Genomic DNA

Fragmentation/bisulfite conversion/Adapter ligation

Sequencing Library

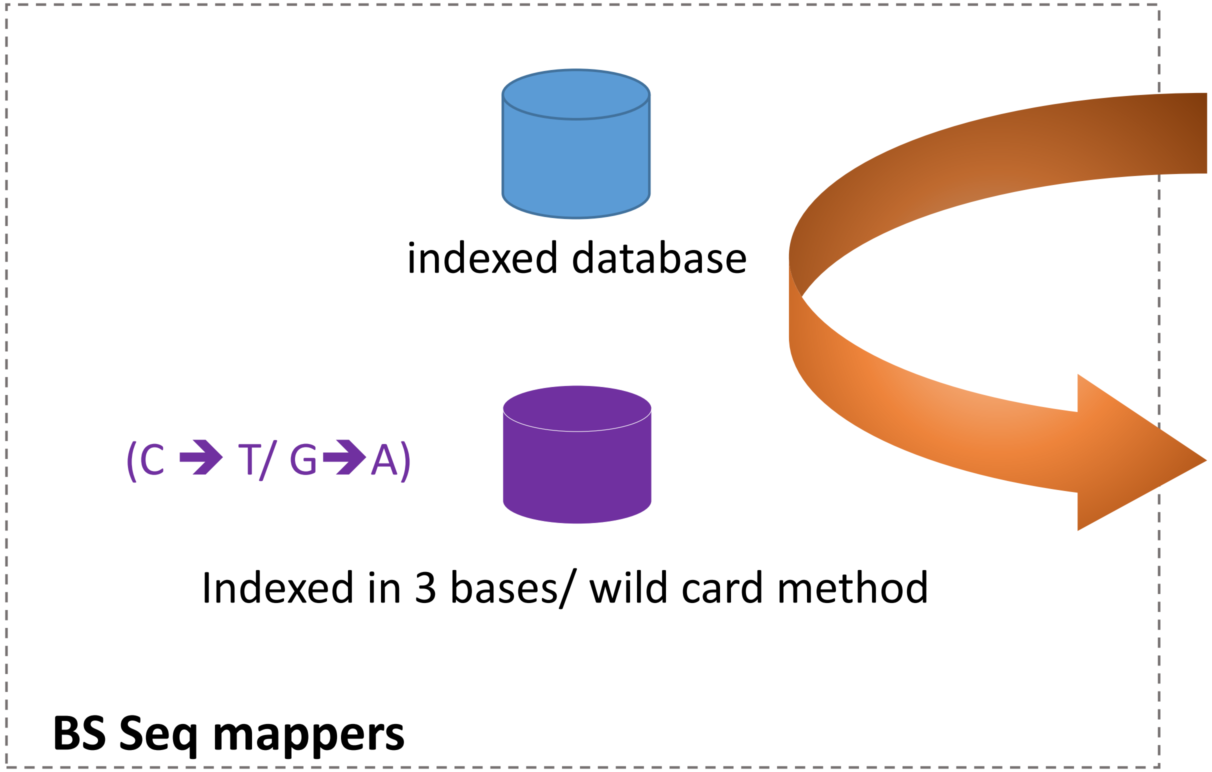
Reproduced and modified from Fig 1 in Curr Protoc Nucleic Acid Chem (2008) Chapter 6:Unit 6.10.

Mapping BS-Seq Reads to Reference Genome

Reference genome

5’ – T C G C C G G T A C -

a g c g g c c a t g



indexed database

(C → T/ G→A)

Indexed in 3 bases/ wild card method

BS Seq mappers

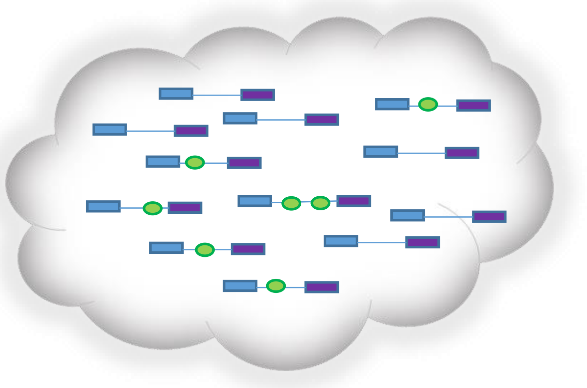
T T G C T G G T A T

a a c g a c c a t a

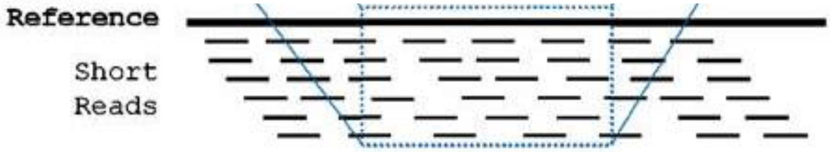
T C G C C G G T A T

a g c g g c c a t a

The Bisulfite converted, PCR amplified library



```
ATTCCAGGAGCTCGCCGGTACCTCACCA
     AGGAGCTTGCCGG
      GGAGCTTGCTGGTACCT
       GAGCTCGCCGGTACCTCAC
        AGCTCGCCGGTACCTCACCAATA
          CTTGCTGGTACCT
```



Difficulty to Access BS Seq Data/ Methylome

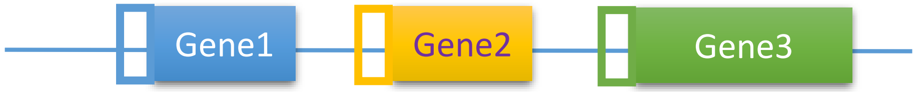
- Complicated Contents

By Context

-CG- -CHG- -CHH-

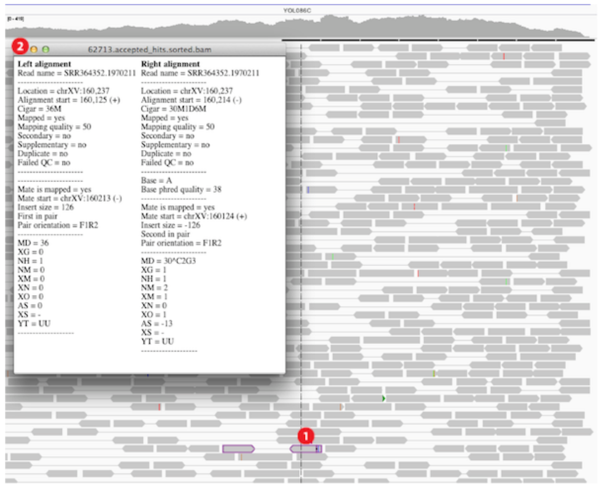
H=A, T or C

By Location



- Promoter
- Gene Body

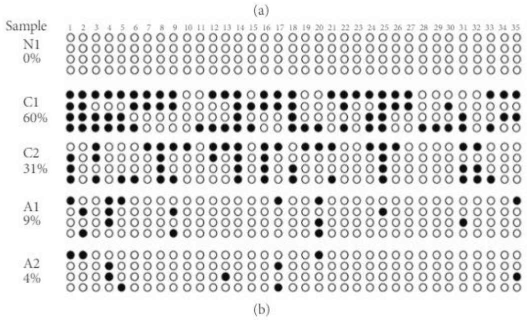
- Visualization

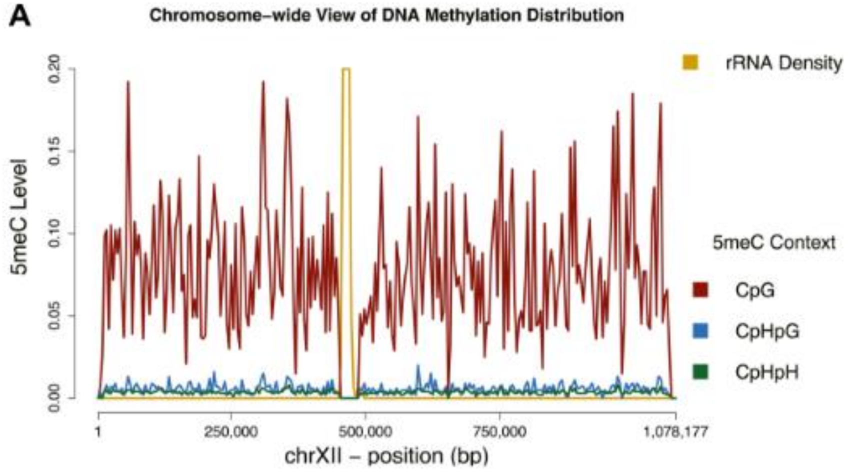


62713.accepted_hits.sorted.bam

```
Left alignment                         Right alignment
Read name = SRR364352.1970211          Read name = SRR364352.1970211
----------------------                 ----------------------
Location = chrXV:160,237               Location = chrXV:160,237
Alignment start = 160,125 (+)          Alignment start = 160,214 (-)
Cigar = 36M                            Cigar = 30M1D6M
Mapped = yes                           Mapped = yes
Mapping quality = 50                   Mapping quality = 50
Secondary = no                         Secondary = no
Supplementary = no                     Supplementary = no
Duplicate = no                         Duplicate = no
Failed QC = no                         Failed QC = no
----------------------                 ----------------------
----------------------                 Base = A
Mate is mapped = yes                   Base phred quality = 38
Mate start = chrXV:160213 (-)          ----------------------
Insert size = 126                      Mate is mapped = yes
First in pair                          Mate start = chrXV:160124 (+)
Pair orientation = F1R2                Insert size = -126
----------------------                 Second in pair
MD = 36                                Pair orientation = F1R2
XG = 0                                 ----------------------
NH = 1                                 MD = 30^C2G3
NM = 0                                 XG = 1
XM = 0                                 NH = 1
XN = 0                                 NM = 2
XO = 0                                 XM = 1
AS = 0                                 XN = 0
XS = -                                 XO = 1
YT = UU                                AS = -13
--------------------                   XS = -
                                       YT = UU
                                       --------------------
```

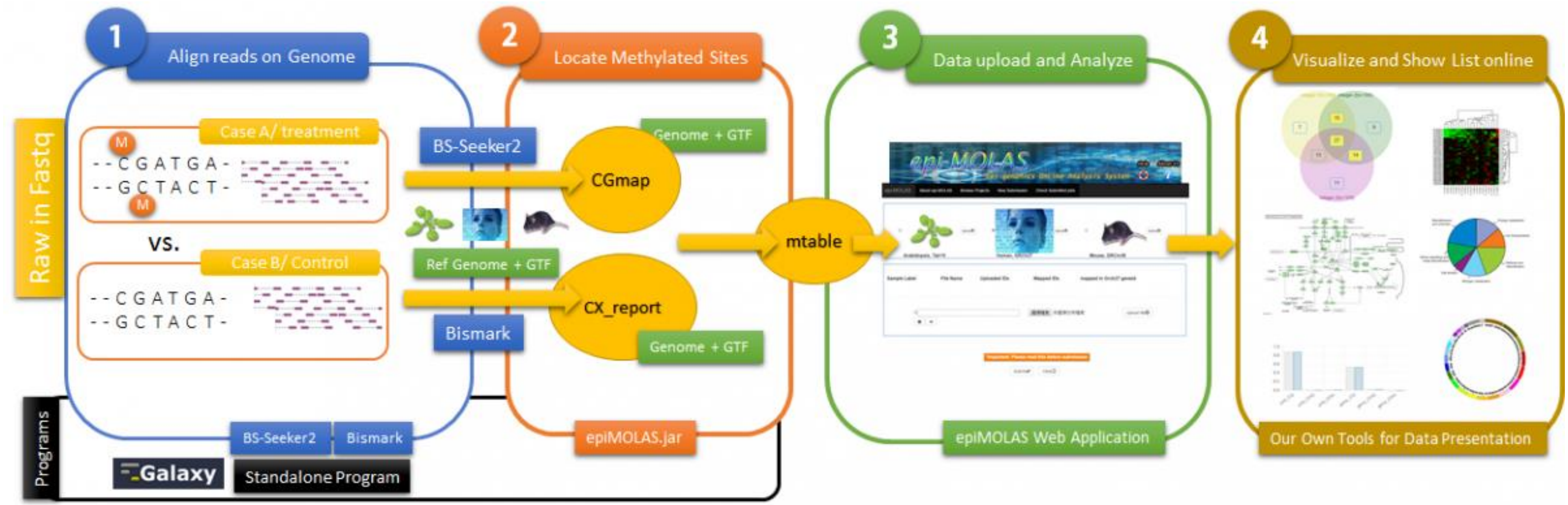
Methylated CG island





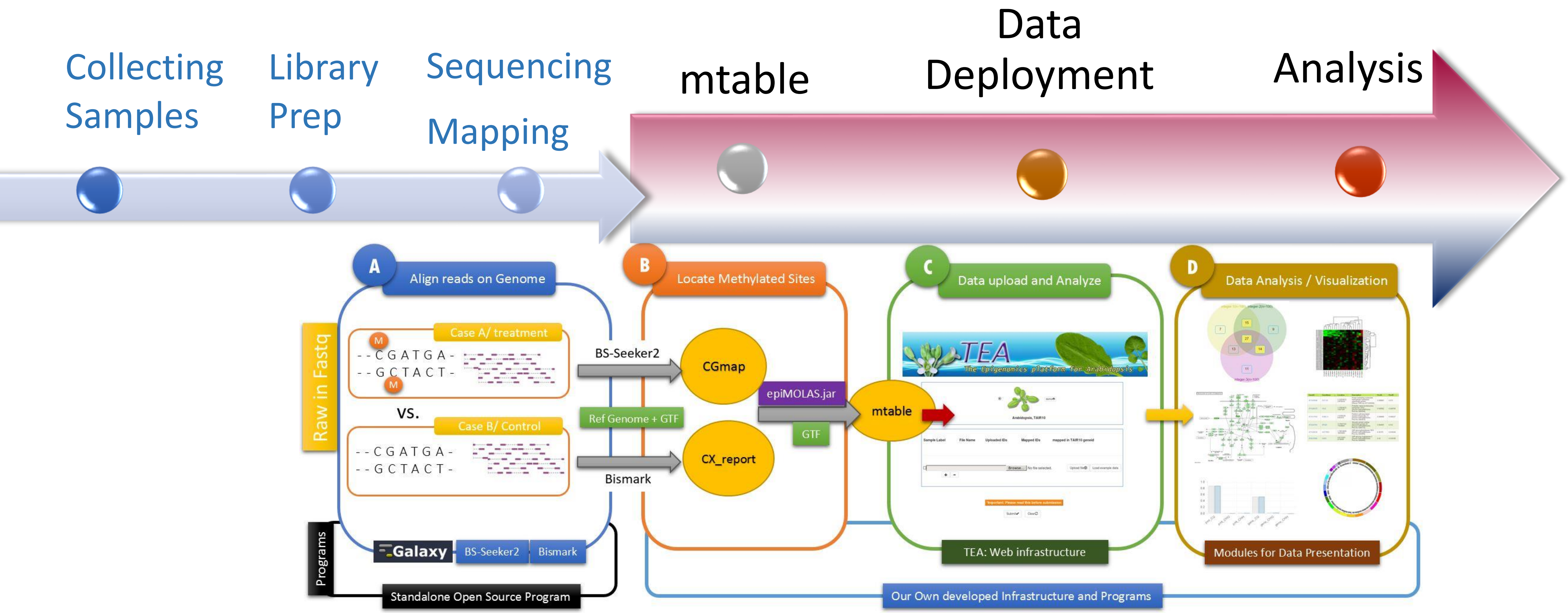
http://symbiosis.iis.sinica.edu.tw/epimolas/molas.html

The Workflow



http://tea.iis.sinica.edu.tw/tea/molas.html

TEA
The epigenomic platform for Arabidopsis



Collecting Samples

Library Prep

Sequencing

Mapping

mtable

Data Deployment

Analysis

Raw in Fastq

A Align reads on Genome

Case A/ treatment

M

- - C G A T G A -

- - G C T A C T -

M

vs.

Case B/ Control

- - C G A T G A -

- - G C T A C T -

BS-Seeker2

Ref Genome + GTF

Bismark

Programs

Galaxy BS-Seeker2 Bismark

Standalone Open Source Program

B Locate Methylated Sites

CGmap

CX_report

epiMOLAS.jar

GTF

mtable

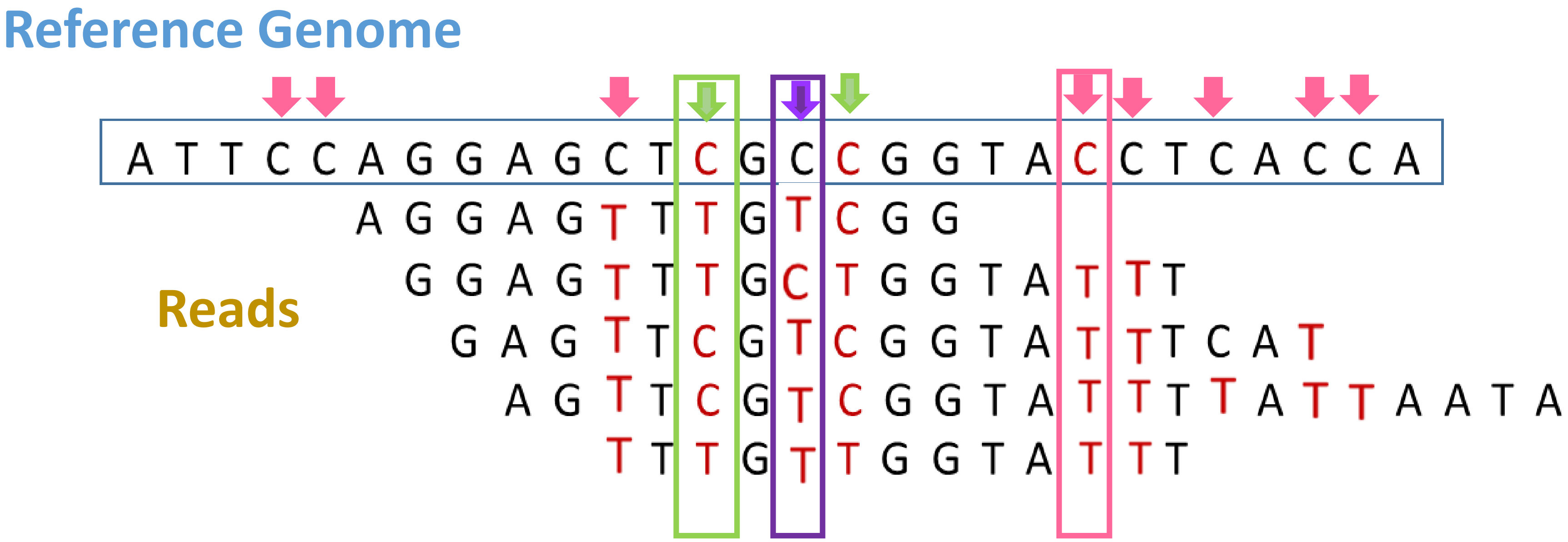
C Data upload and Analyze

TEA: Web infrastructure

D Data Analysis / Visualization

Modules for Data Presentation

Our Own developed Infrastructure and Programs



- Type: **CG**
- Total observation (Read depth): 5
- Methylated C: 2, Unmethylated C: 3

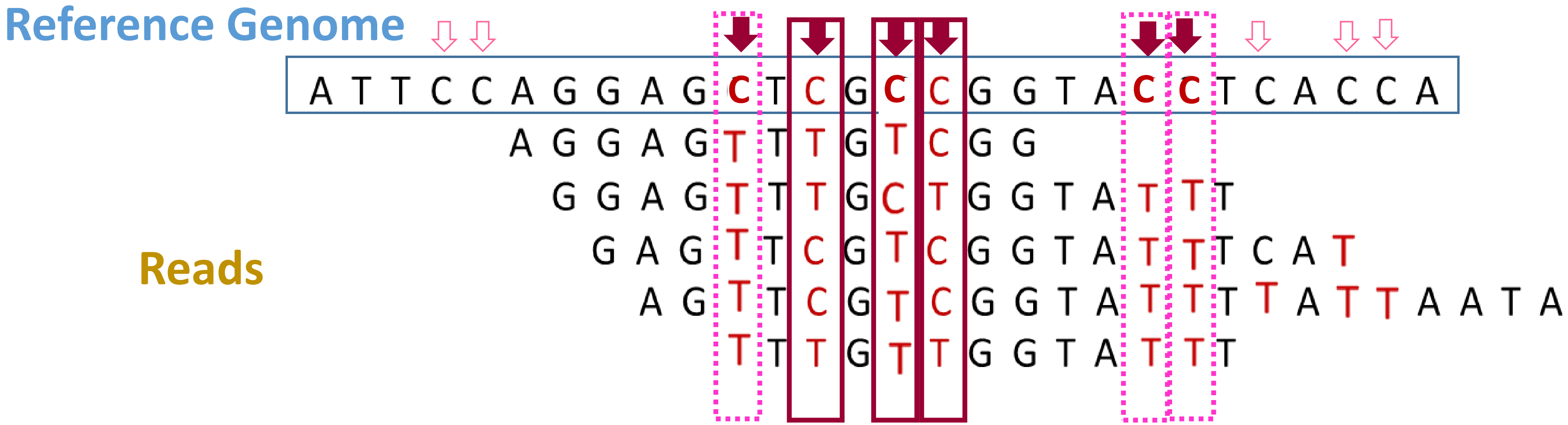
➔ score of this C: 2/5 = 0.4

- Type: **CHH**
- Total observation (Read depth): 4
- Methylated C: 0, Unmethylated C: 4

➔ score of this C: 0

- Type: **CHG**
- Total observation (Read depth): 5
- Methylated C: 3, Unmethylated C: 2

➔ score of this C: 0.6


Reference Genome
A T T C C A G G A G C T C G C C G G T A C C T C A C C A
Reads
A G G A G T T T G T C G G
G G A G T T T G C T G G T A T T T
G A G T T C G T C G G T A T T T C A T
A G T T C G T C G G T A T T T T A T T A A T A
T T T G T T G G T A T T T

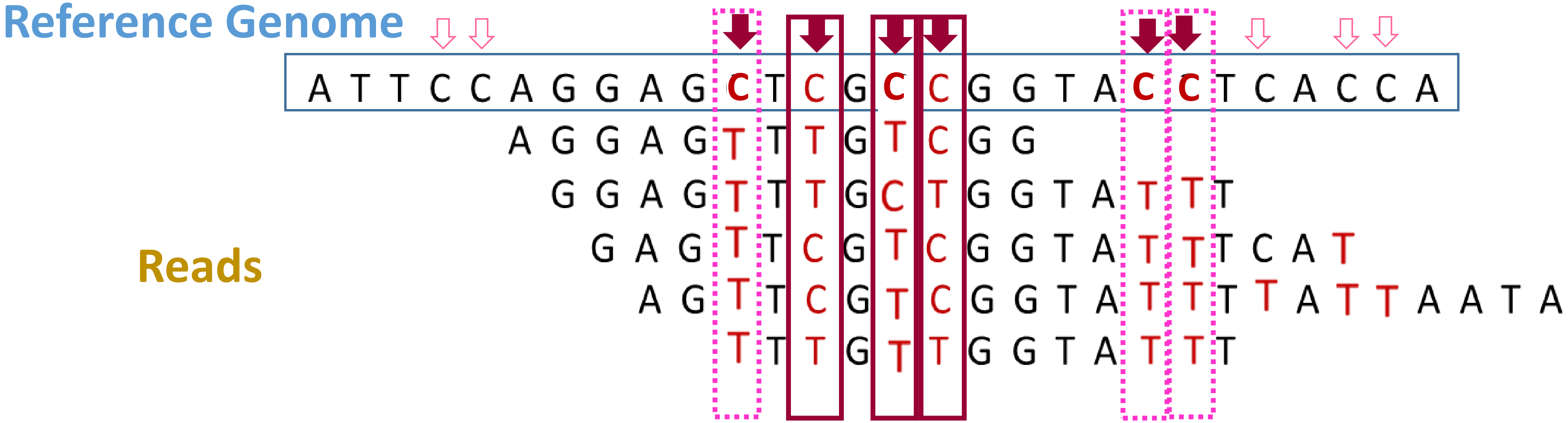
- Scored gene / promoter: # observed bases >=5

By Context

By Location

$$\text{Average DNA methylation level in promoter or gene body} = \frac{\sum_{i\in X} c_i}{\sum_{i\in X} 1} \qquad (1.2)$$

X = promoter or gene body



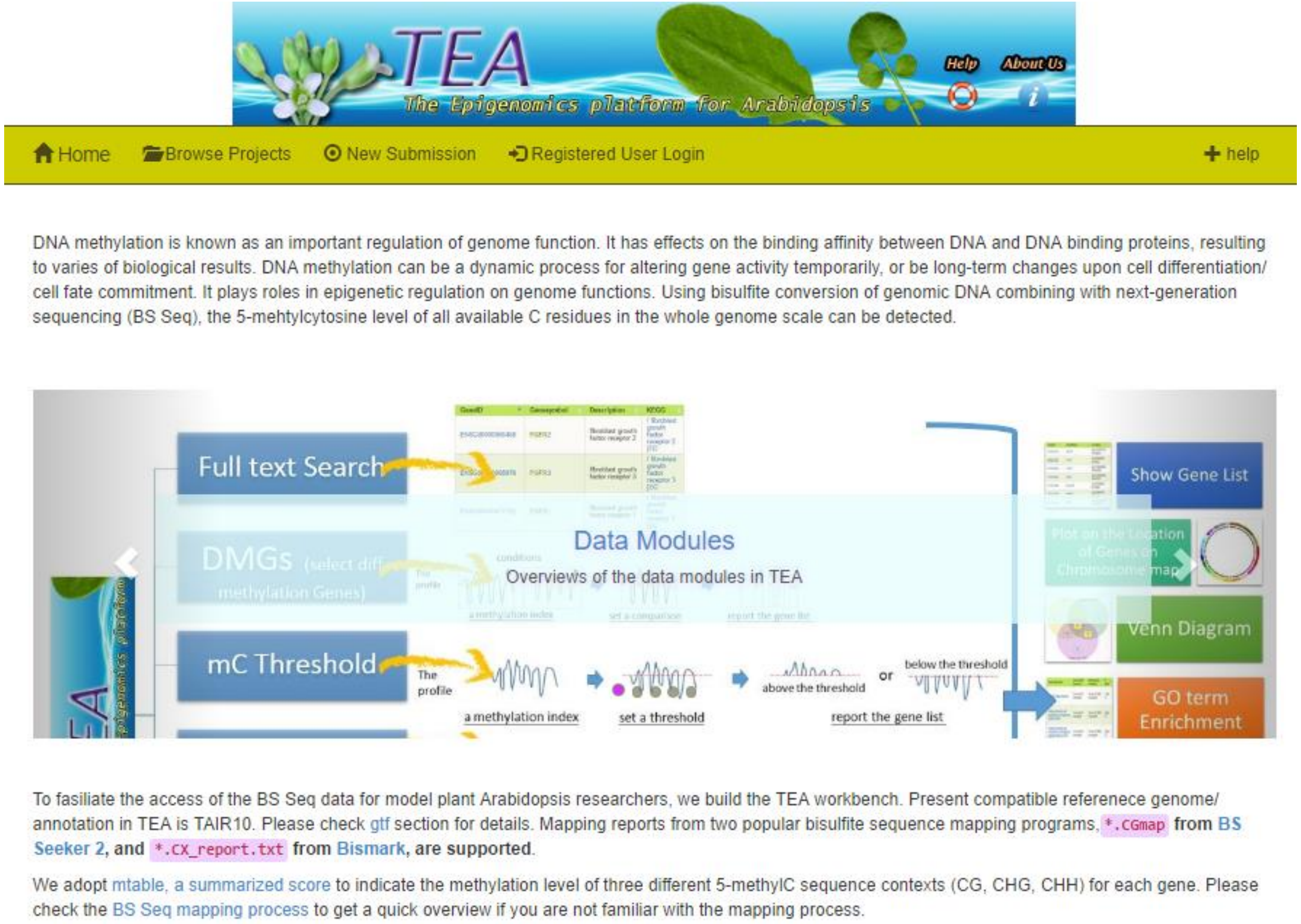
- Observed event for each C: >=4
- Scored gene / promoter: # observed bases >=5
- Supporting Mapper: BS-Seeker2 and Bismark

gene_id	pmt_CG	gene_CG	pmt_CHG	gene_CHG	pmt_CHH	gene_CHH
AT1G01010	0.011463	0.053009	0.010000	0.011635	0.021765	0.012631
AT1G01020	0.000000	0.081519	0.006957	[illegible]	0.003614	0.007521
AT1G01030	0.005385	0.0128[illegible]	[illegible]	[illegible]	0.003116	0.016939
AT1G01040	0.011200	[illegible]	[illegible]	0.015773	0.016944	0.011699
AT1G01046	0.765250	0.385000	0.022500	0.058750	0.014325	0.047727

The Methylation Landscape

TEA Website

http://tea.iis.sinica.edu.tw/tea/molas.html



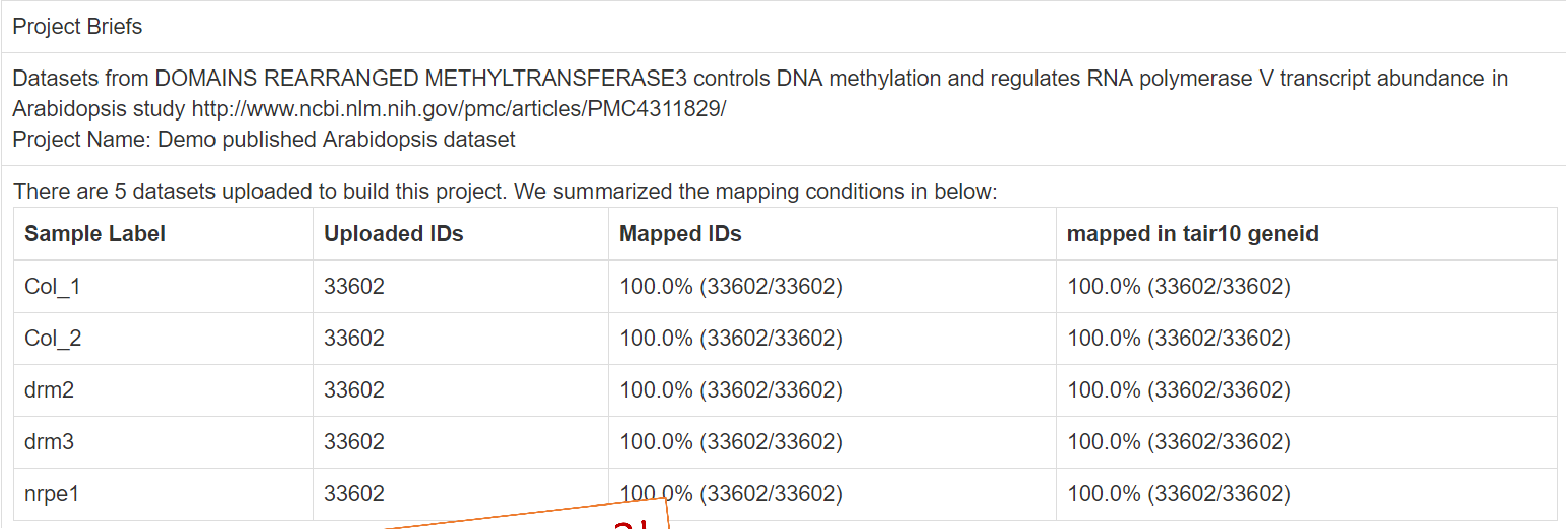
DNA methylation is known as an important regulation of genome function. It has effects on the binding affinity between DNA and DNA binding proteins, resulting to varies of biological results. DNA methylation can be a dynamic process for altering gene activity temporarily, or be long-term changes upon cell differentiation/ cell fate commitment. It plays roles in epigenetic regulation on genome functions. Using bisulfite conversion of genomic DNA combining with next-generation sequencing (BS Seq), the 5-mehtylcytosine level of all available C residues in the whole genome scale can be detected.

To fasiliate the access of the BS Seq data for model plant Arabidopsis researchers, we build the TEA workbench. Present compatible refereence genome/ annotation in TEA is TAIR10. Please check gtf section for details. Mapping reports from two popular bisulfite sequence mapping programs, `*.CGmap` **from BS Seeker 2, and** `*.CX_report.txt` **from Bismark, are supported.**

We adopt mtable, a summarized score to indicate the methylation level of three different 5-methylC sequence contexts (CG, CHG, CHH) for each gene. Please check the BS Seq mapping process to get a quick overview if you are not familiar with the mapping process.

Project Summary

http://tea.iis.sinica.edu.tw/tair10_demo_new/



Project Briefs

Datasets from DOMAINS REARRANGED METHYLTRANSFERASE3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis study http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4311829/
Project Name: Demo published Arabidopsis dataset

There are 5 datasets uploaded to build this project. We summarized the mapping conditions in below:

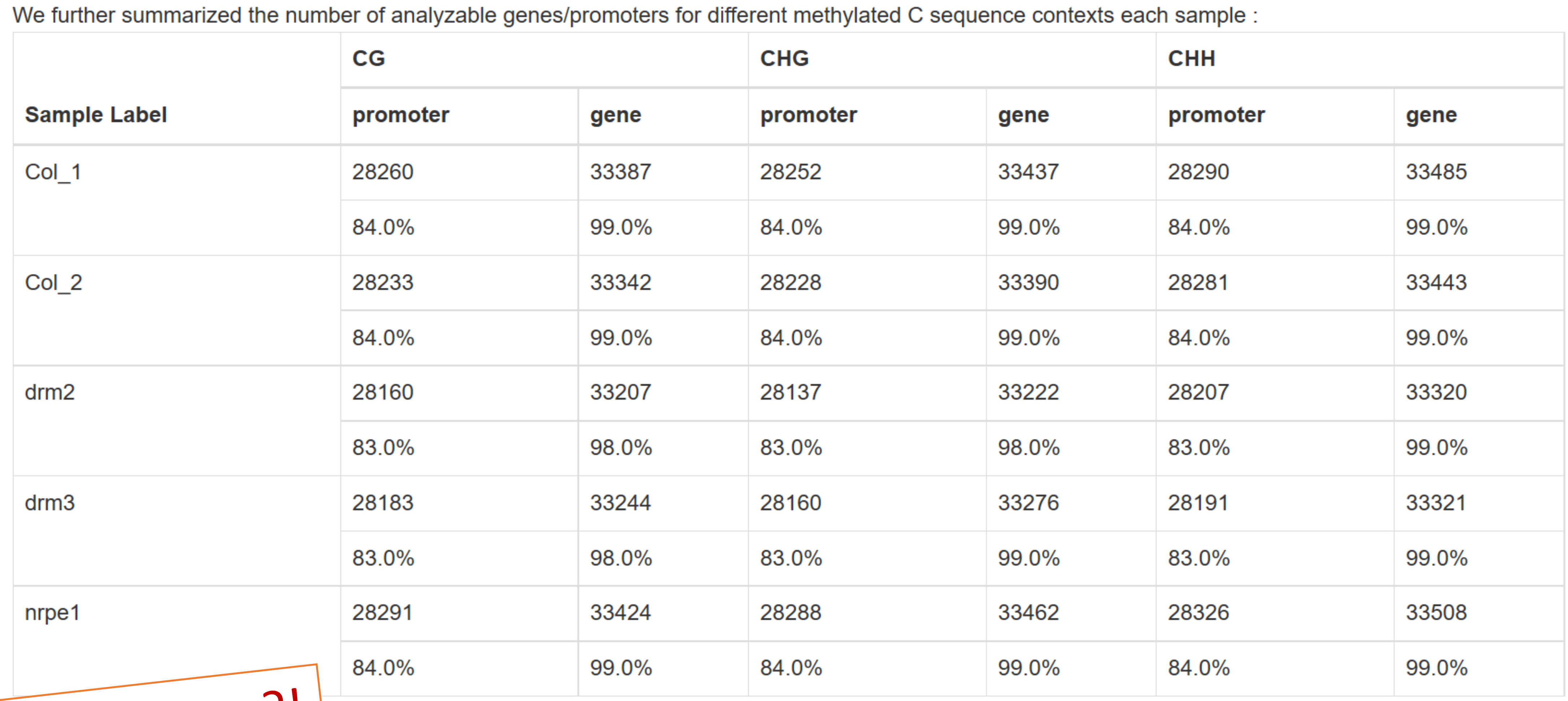
Sample Label	Uploaded IDs	Mapped IDs	mapped in tair10 geneid
Col_1	33602	100.0% (33602/33602)	100.0% (33602/33602)
Col_2	33602	100.0% (33602/33602)	100.0% (33602/33602)
drm2	33602	100.0% (33602/33602)	100.0% (33602/33602)
drm3	33602	100.0% (33602/33602)	100.0% (33602/33602)
nrpe1	33602	100.0% (33602/33602)	100.0% (33602/33602)

Poor ID mapping rate?!

Check the gtf version

Project Summary

http://tea.iis.sinica.edu.tw/tair10_demo_new/

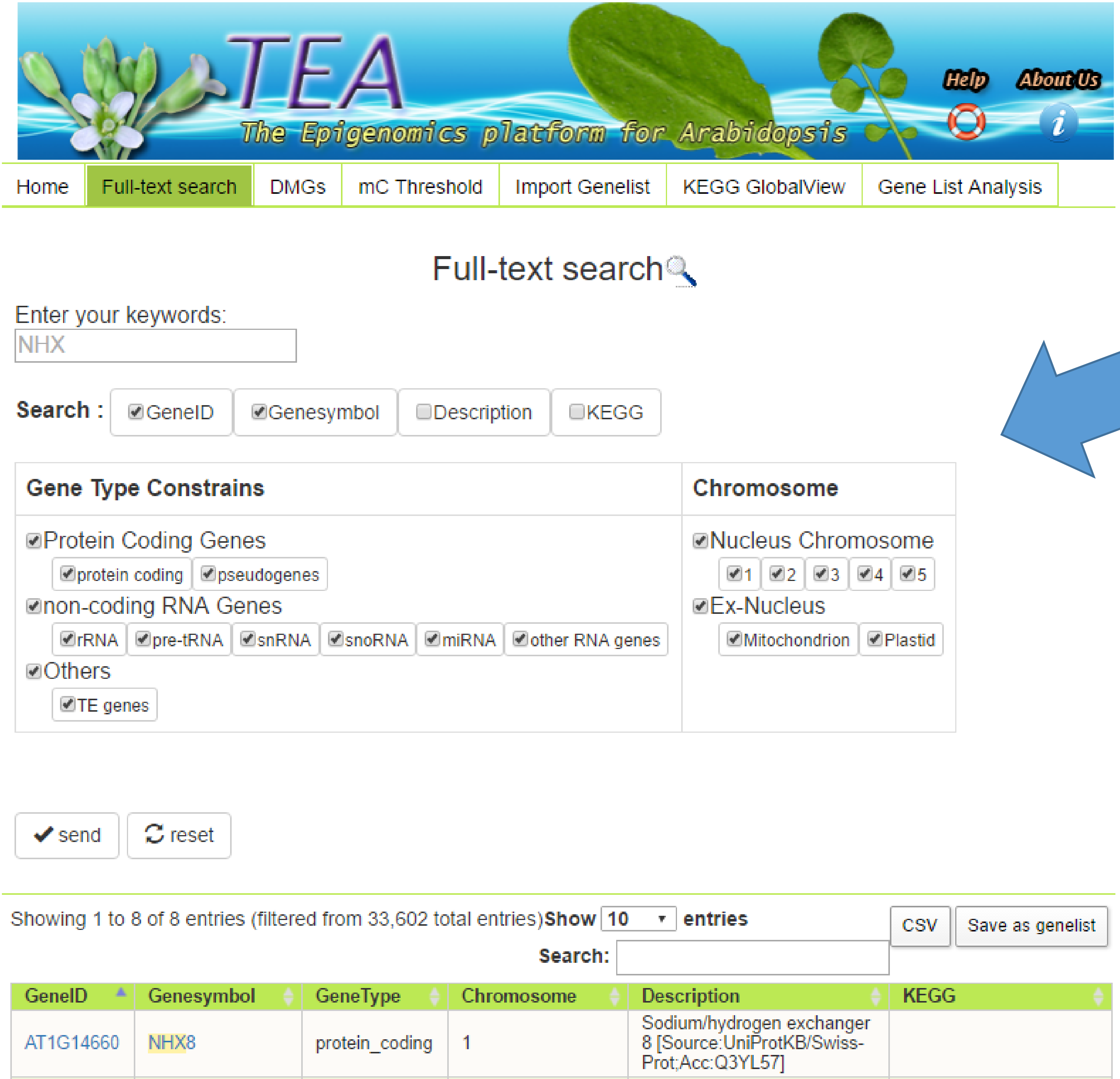


We further summarized the number of analyzable genes/promoters for different methylated C sequence contexts each sample :

	CG		CHG		CHH	
Sample Label	**promoter**	**gene**	**promoter**	**gene**	**promoter**	**gene**
Col_1	28260	33387	28252	33437	28290	33485
	84.0%	99.0%	84.0%	99.0%	84.0%	99.0%
Col_2	28233	33342	28228	33390	28281	33443
	84.0%	99.0%	84.0%	99.0%	84.0%	99.0%
drm2	28160	33207	28137	33222	28207	33320
	83.0%	98.0%	83.0%	98.0%	83.0%	99.0%
drm3	28183	33244	28160	33276	28191	33321
	83.0%	98.0%	83.0%	99.0%	83.0%	99.0%
nrpe1	28291	33424	28288	33462	28326	33508
	84.0%	99.0%	84.0%	99.0%	84.0%	99.0%

Missing Data ?!

Check the (1) read mapping rate (2) throughput



Full-text search

Enter your keywords: NHX

Search : ☑GeneID ☑Genesymbol ☐Description ☐KEGG

Gene Type Constrains	Chromosome
☑Protein Coding Genes ☑protein coding ☑pseudogenes ☑non-coding RNA Genes ☑rRNA ☑pre-tRNA ☑snRNA ☑snoRNA ☑miRNA ☑other RNA genes ☑Others ☑TE genes	☑Nucleus Chromosome ☑1 ☑2 ☑3 ☑4 ☑5 ☑Ex-Nucleus ☑Mitochondrion ☑Plastid

Showing 1 to 8 of 8 entries (filtered from 33,602 total entries)**Show** 10 **entries**

Search:

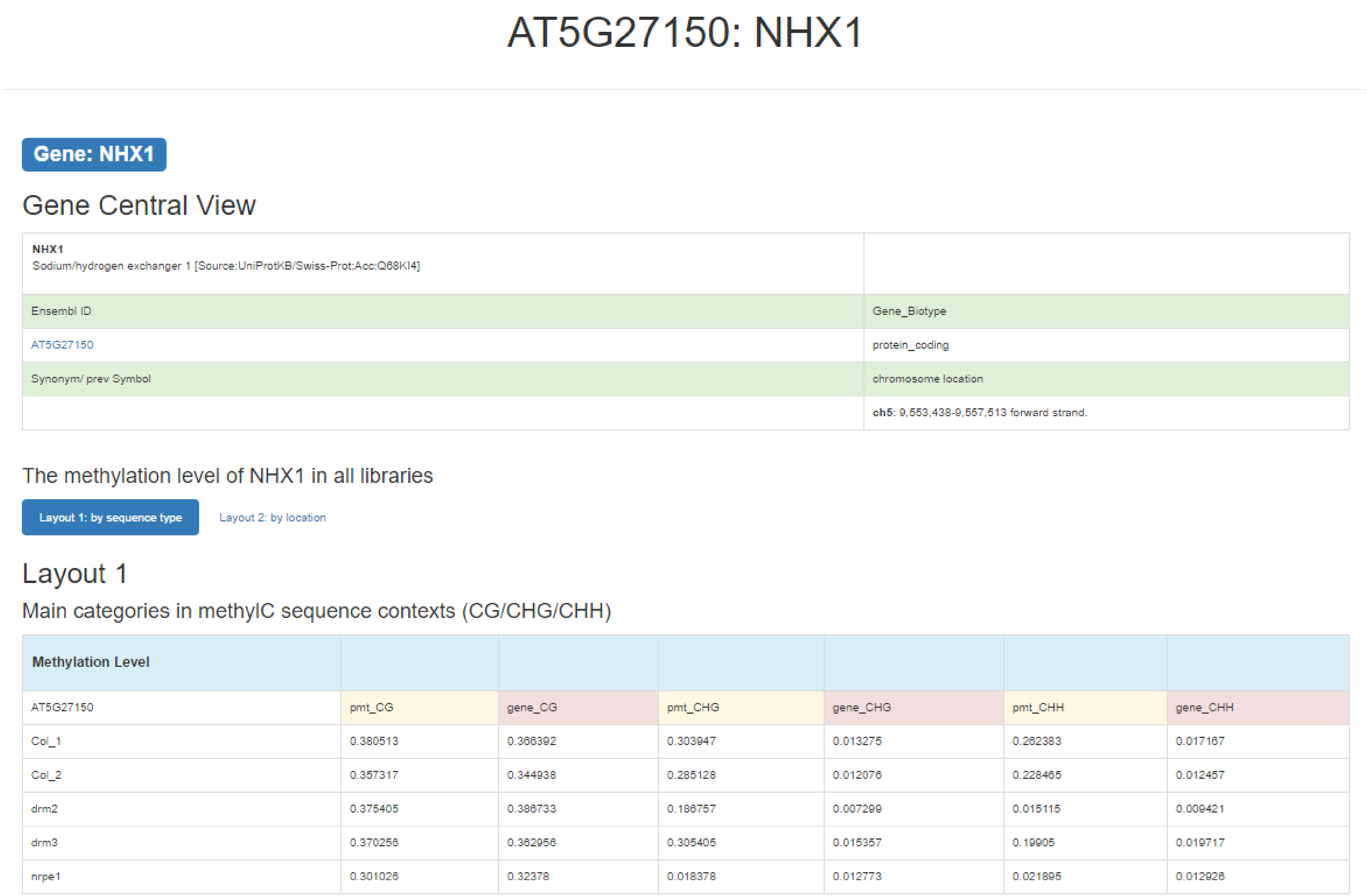
GeneID	Genesymbol	GeneType	Chromosome	Description	KEGG
AT1G14660	NHX8	protein_coding	1	Sodium/hydrogen exchanger 8 [Source:UniProtKB/Swiss-Prot;Acc:Q3YL57]	

Set the searching criteria

Get the result

http://tea.iis.sinica.edu.tw/sophia_test/gdetail.php?gid=AT5G27150

Gene Central View



AT5G27150: NHX1

Gene Central View

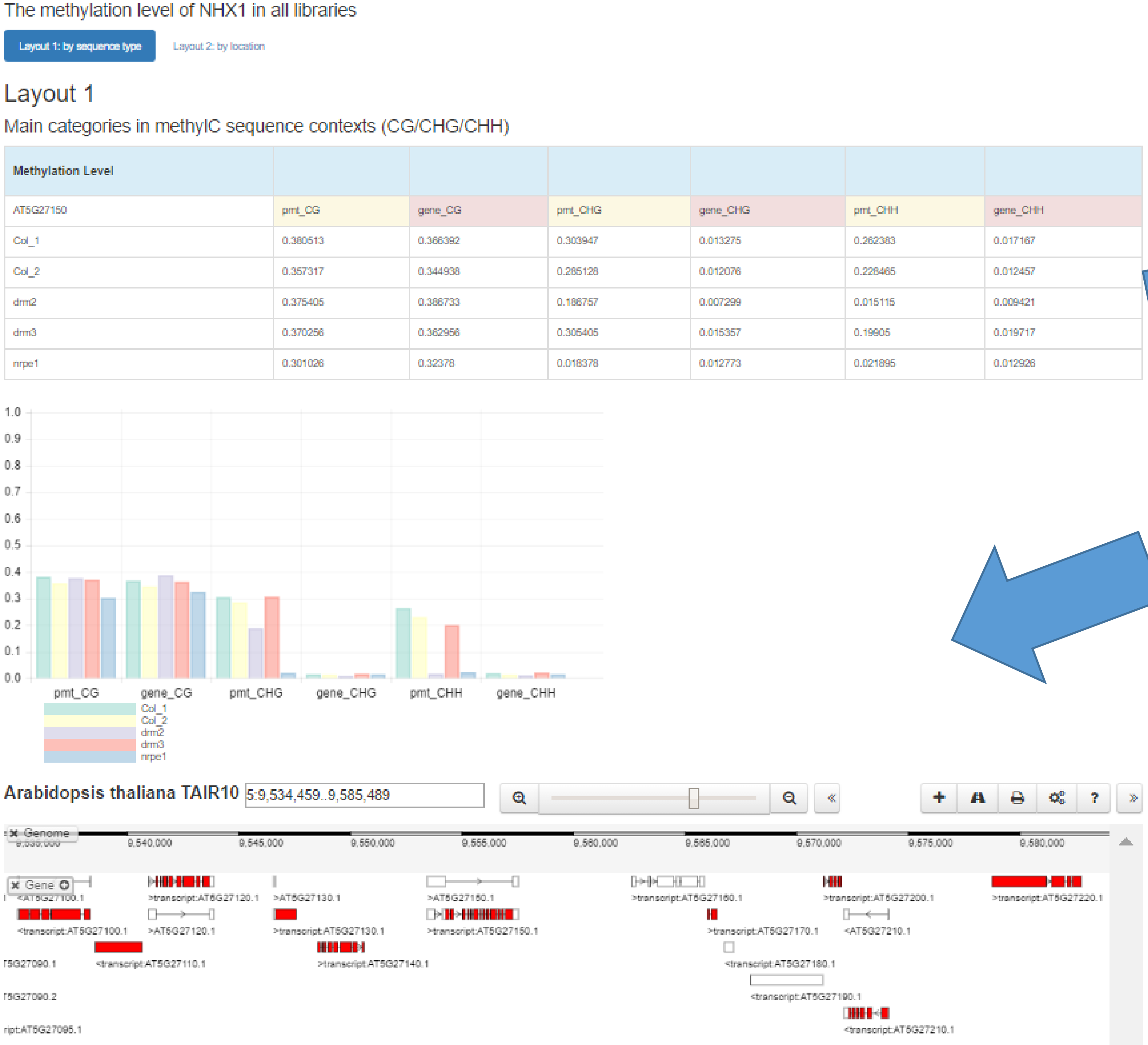
NHX1 Sodium/hydrogen exchanger 1 [Source:UniProtKB/Swiss-Prot;Acc:Q68KI4]	
Ensembl ID	Gene_Biotype
AT5G27150	protein_coding
Synonym/ prev Symbol	chromosome location
	ch5: 9,553,438-9,557,513 forward strand.

The methylation level of NHX1 in all libraries

Layout 1

Main categories in methylC sequence contexts (CG/CHG/CHH)

Methylation Level						
AT5G27150	pmt_CG	gene_CG	pmt_CHG	gene_CHG	pmt_CHH	gene_CHH
Col_1	0.380513	0.366392	0.303947	0.013275	0.262383	0.017167
Col_2	0.357317	0.344938	0.285128	0.012076	0.228465	0.012457
drm2	0.375405	0.386733	0.186757	0.007299	0.015115	0.009421
drm3	0.370256	0.362956	0.305405	0.015357	0.19905	0.019717
nrpe1	0.301026	0.32378	0.018378	0.012773	0.021895	0.012926



The methylation level of NHX1 in all libraries

Layout 1

Main categories in methylC sequence contexts (CG/CHG/CHH)

Methylation Level						
AT5G27150	pmt_CG	gene_CG	pmt_CHG	gene_CHG	pmt_CHH	gene_CHH
Col_1	0.380513	0.366392	0.303947	0.013275	0.262383	0.017167
Col_2	0.357317	0.344938	0.285128	0.012076	0.228465	0.012457
drm2	0.375405	0.386733	0.186757	0.007299	0.015115	0.009421
drm3	0.370256	0.362956	0.305405	0.015357	0.19905	0.019717
nrpe1	0.301026	0.32378	0.018378	0.012773	0.021895	0.012926

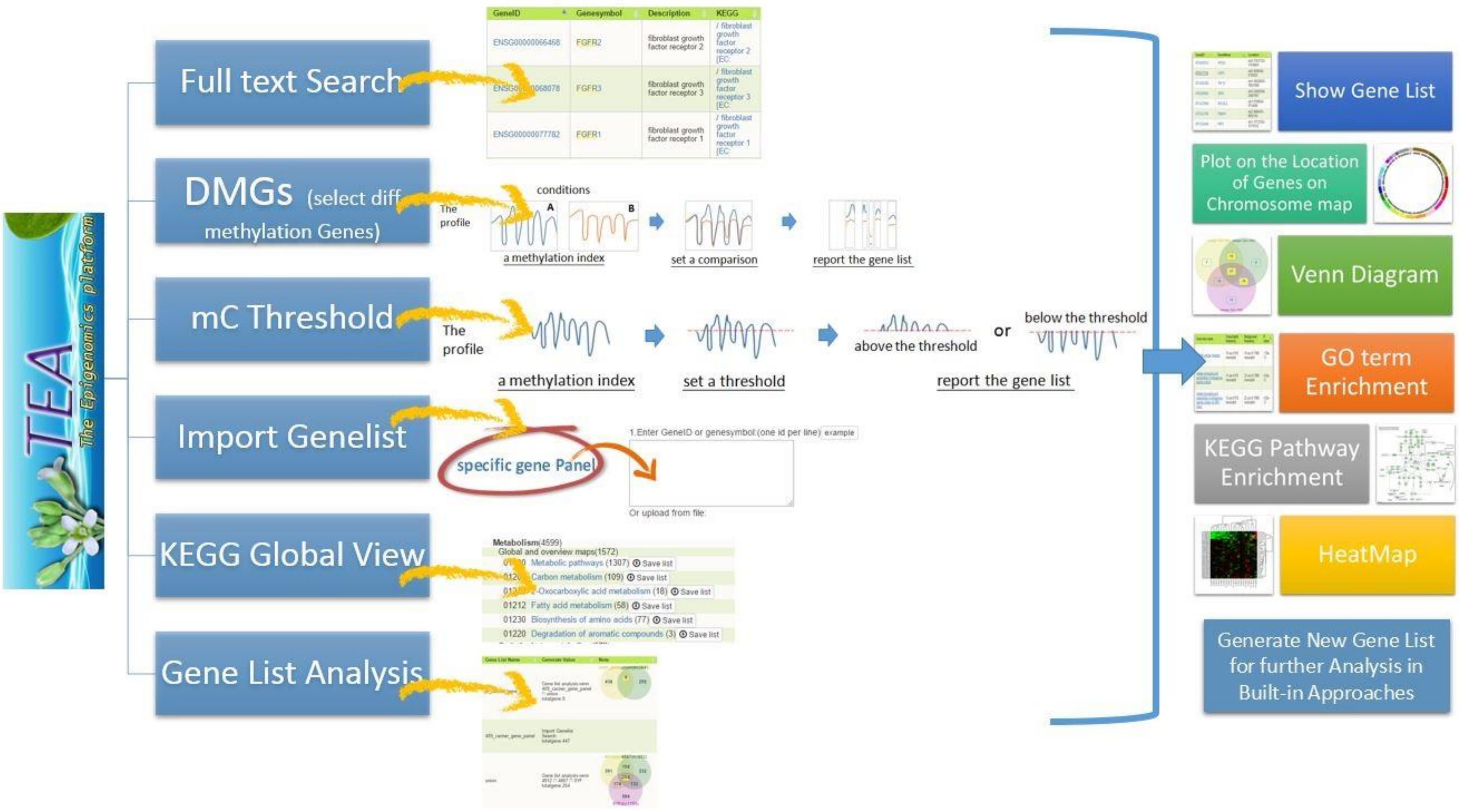
Arabidopsis thaliana TAIR10 5:9,534,459..9,585,489

Measures of Methylation

Genome Browser

Data Analysis Modules

http://tea.iis.sinica.edu.tw/tea/access_project.html

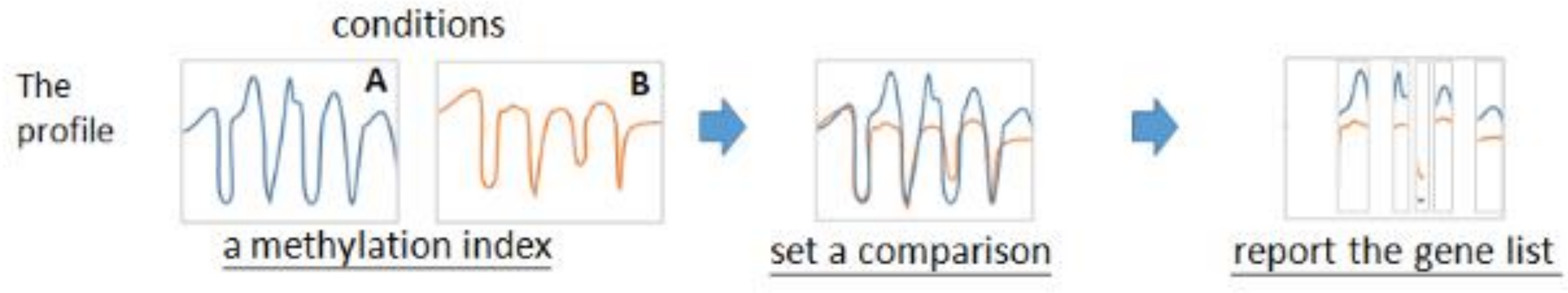


GeneID	Genesymbol	Description	KEGG
ENSG00000066468	FGFR2	fibroblast growth factor receptor 2	/ fibroblast growth factor receptor 2 [EC:
ENSG0[illegible]068078	FGFR3	fibroblast growth factor receptor 3	/ fibroblast growth factor receptor 3 [EC:
ENSG00000077782	FGFR1	fibroblast growth factor receptor 1	/ fibroblast growth factor receptor 1 [EC:

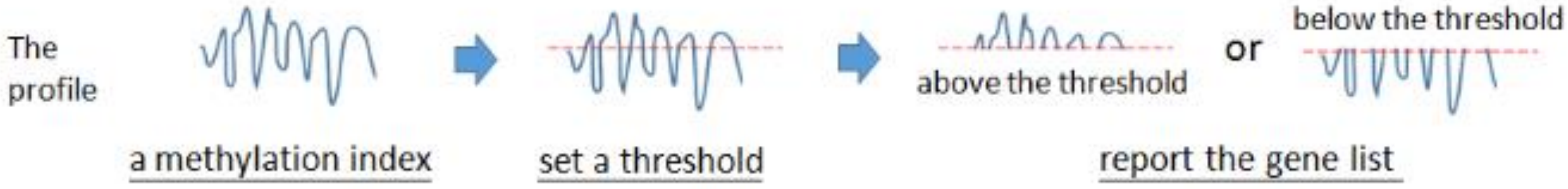
Metabolism(4599)
Global and overview maps(1572)
01[illegible]0 Metabolic pathways (1307) Save list
012[illegible] Carbon metabolism (109) Save list
012[illegible] 2-Oxocarboxylic acid metabolism (18) Save list
01212 Fatty acid metabolism (58) Save list
01230 Biosynthesis of amino acids (77) Save list
01220 Degradation of aromatic compounds (3) Save list

Find Genes by Value

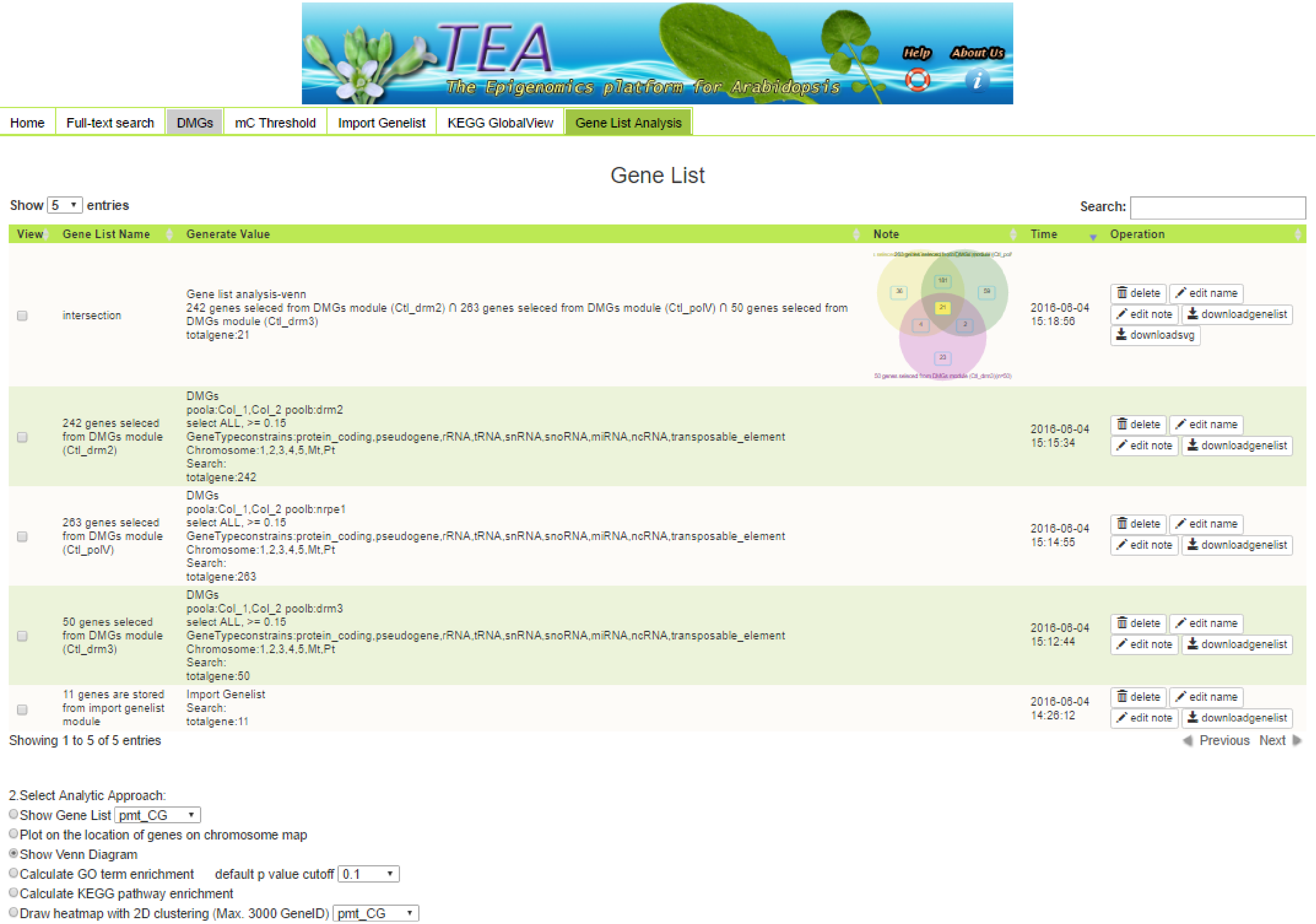
DMGs : Select differentially methylated genes by the interested methylation score



Threshold : Select genes by a cutoff value on the methylation score



Gene List and Data Visualization



Gene List

Show 5 entries

Search:

View	Gene List Name	Generate Value	Note	Time	Operation
□	intersection	Gene list analysis-venn 242 genes seleced from DMGs module (Ctl_drm2) ∩ 263 genes seleced from DMGs module (Ctl_polV) ∩ 50 genes seleced from DMGs module (Ctl_drm3) totalgene:21	36 181 59 21 4 2 23 50 genes seleced from DMGs module (Ctl_drm3)(n=50)	2016-06-04 15:18:56	delete edit name edit note downloadgenelist downloadsvg
□	242 genes seleced from DMGs module (Ctl_drm2)	DMGs poola:Col_1,Col_2 poolb:drm2 select ALL, >= 0.15 GeneTypeconstrains:protein_coding,pseudogene,rRNA,tRNA,snRNA,snoRNA,miRNA,ncRNA,transposable_element Chromosome:1,2,3,4,5,Mt,Pt Search: totalgene:242		2016-06-04 15:15:34	delete edit name edit note downloadgenelist
□	263 genes seleced from DMGs module (Ctl_polV)	DMGs poola:Col_1,Col_2 poolb:nrpe1 select ALL, >= 0.15 GeneTypeconstrains:protein_coding,pseudogene,rRNA,tRNA,snRNA,snoRNA,miRNA,ncRNA,transposable_element Chromosome:1,2,3,4,5,Mt,Pt Search: totalgene:263		2016-06-04 15:14:55	delete edit name edit note downloadgenelist
□	50 genes seleced from DMGs module (Ctl_drm3)	DMGs poola:Col_1,Col_2 poolb:drm3 select ALL, >= 0.15 GeneTypeconstrains:protein_coding,pseudogene,rRNA,tRNA,snRNA,snoRNA,miRNA,ncRNA,transposable_element Chromosome:1,2,3,4,5,Mt,Pt Search: totalgene:50		2016-06-04 15:12:44	delete edit name edit note downloadgenelist
□	11 genes are stored from import genelist module	Import Genelist Search: totalgene:11		2016-06-04 14:26:12	delete edit name edit note downloadgenelist

Showing 1 to 5 of 5 entries

Previous Next

2.Select Analytic Approach:

○ Show Gene List pmt_CG

○ Plot on the location of genes on chromosome map

● Show Venn Diagram

○ Calculate GO term enrichment default p value cutoff 0.1

○ Calculate KEGG pathway enrichment

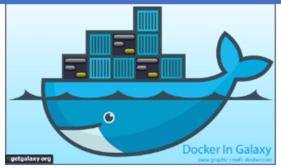
○ Draw heatmap with 2D clustering (Max. 3000 GeneID) pmt_CG

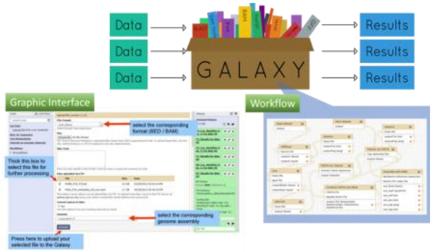
線上基因概況分析平台

http:// molas.iis.sinica.edu.tw/

Raw Reads in a dozen of GB

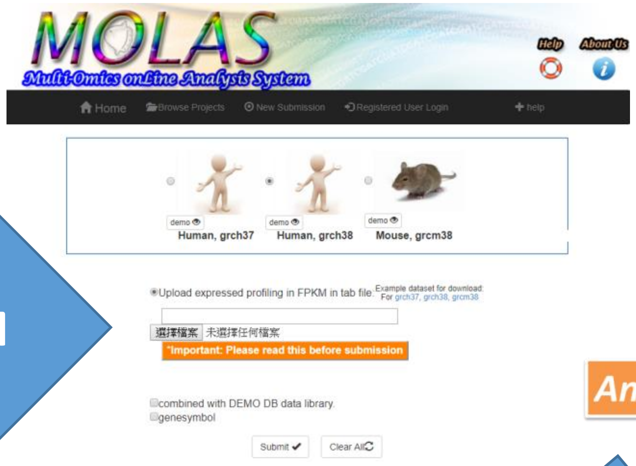
DOCEXPRESS





Estimate Expression Profiling in DOCEXPRESS

FPKM/TPM



Human, grch37 ○ Human, grch38 ● Mouse, grcm38 ○

● Upload expressed profiling in FPKM in tab file: Example dataset for download: For grch37, grch38, grcm38

未選擇任何檔案

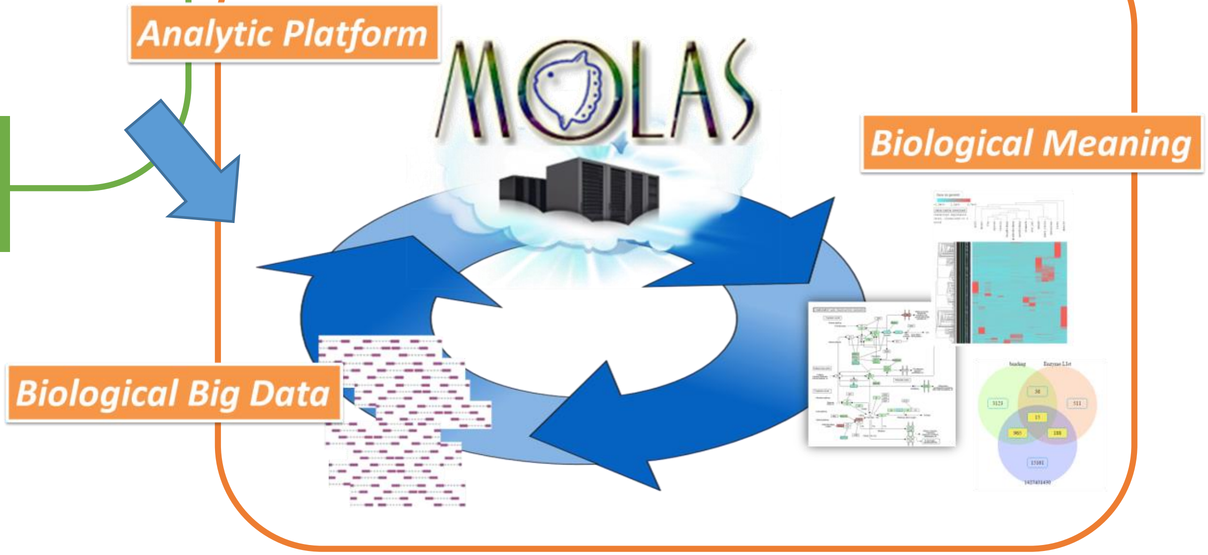
*Important: Please read this before submission

□ combined with DEMO DB data library.

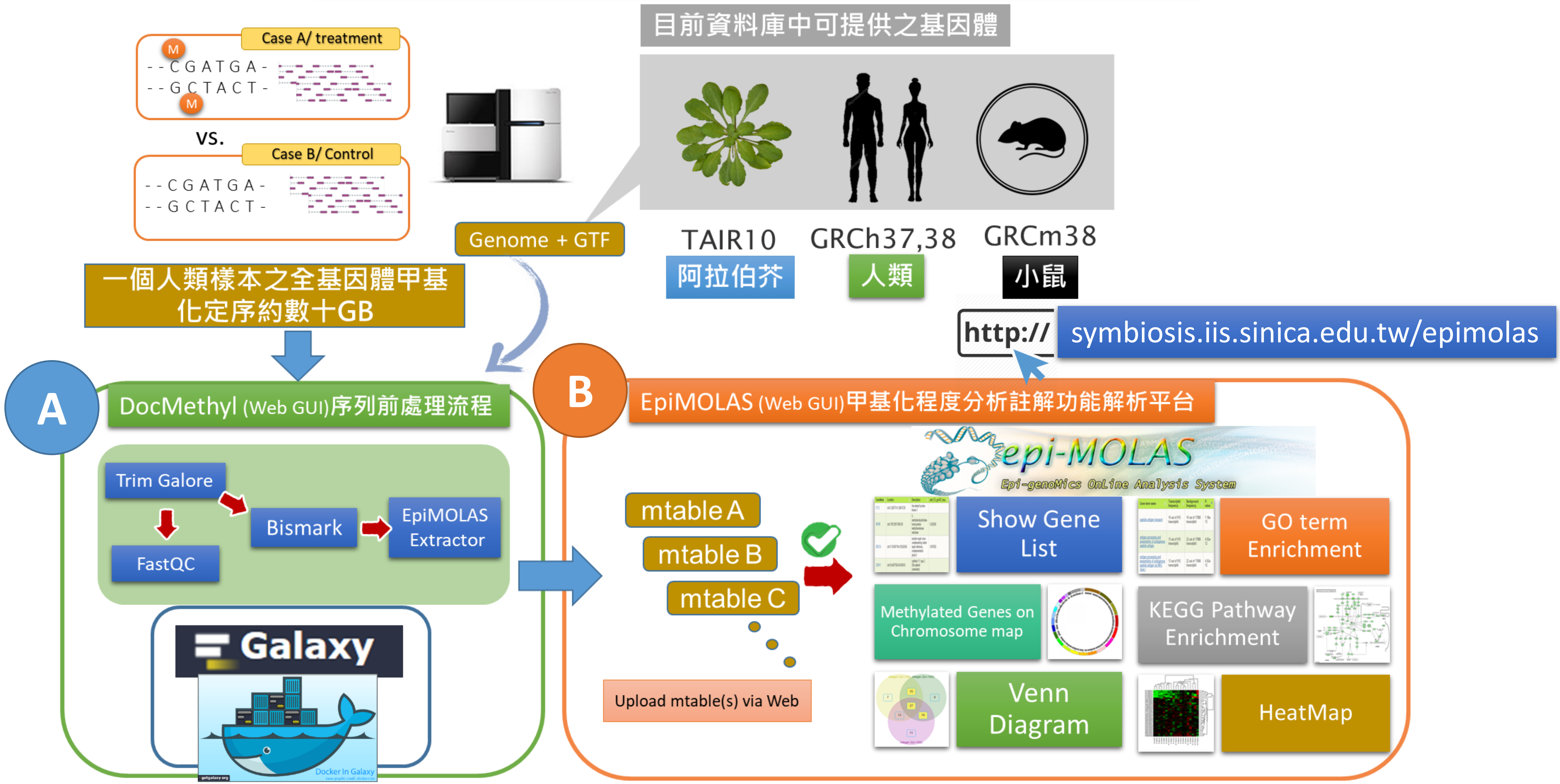
□ genesymbol

Upload to MOLAS

Unveil the biological secrets hidden behind the big biological data online



全基因體甲基化分析平台



目前資料庫中可提供之基因體

TAIR10 阿拉伯芥

GRCh37,38 人類

GRCm38 小鼠

http:// symbiosis.iis.sinica.edu.tw/epimolas

Case A/ treatment

- - C G A T G A -
- - G C T A C T -

vs.

Case B/ Control

- - C G A T G A -
- - G C T A C T -

Genome + GTF

一個人類樣本之全基因體甲基化定序約數十GB

A DocMethyl (Web GUI)序列前處理流程

Trim Galore → FastQC

Trim Galore → Bismark → EpiMOLAS Extractor

Galaxy

Docker In Galaxy

B EpiMOLAS (Web GUI)甲基化程度分析註解功能解析平台

epi-MOLAS

Epi-genoMics OnLine Analysis System

mtable A

mtable B

mtable C

Upload mtable(s) via Web

Show Gene List

GO term Enrichment

Methylated Genes on Chromosome map

KEGG Pathway Enrichment

Venn Diagram

HeatMap

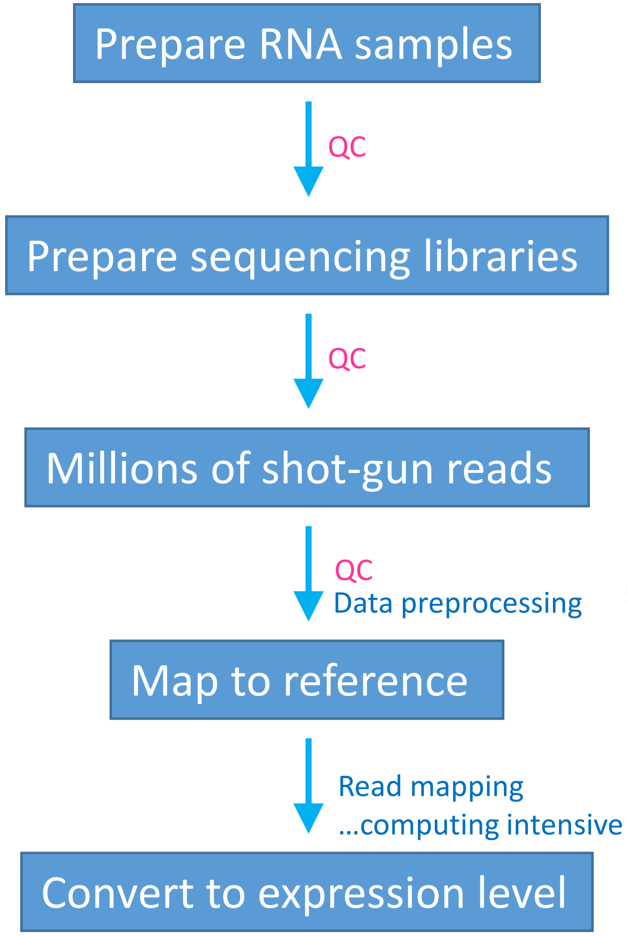
Questions?

Other Issues ?

- Experiment Design? Biological Replicates >>> Technical Replicates
- Library Protocols:
  - Stranded or not?
  - PolyA tailed or rRNA depletion?
  - Have reference genome? Novel transcripts? Fusion transcripts?

- Special protocols that need extra bioinformatical works?
- Trimmed read length? Low complexity repeats? Other sources of contamination?

A Typical RNA-Seq Experiment



Prepare RNA samples

QC

Prepare sequencing libraries

QC

Millions of shot-gun reads

QC

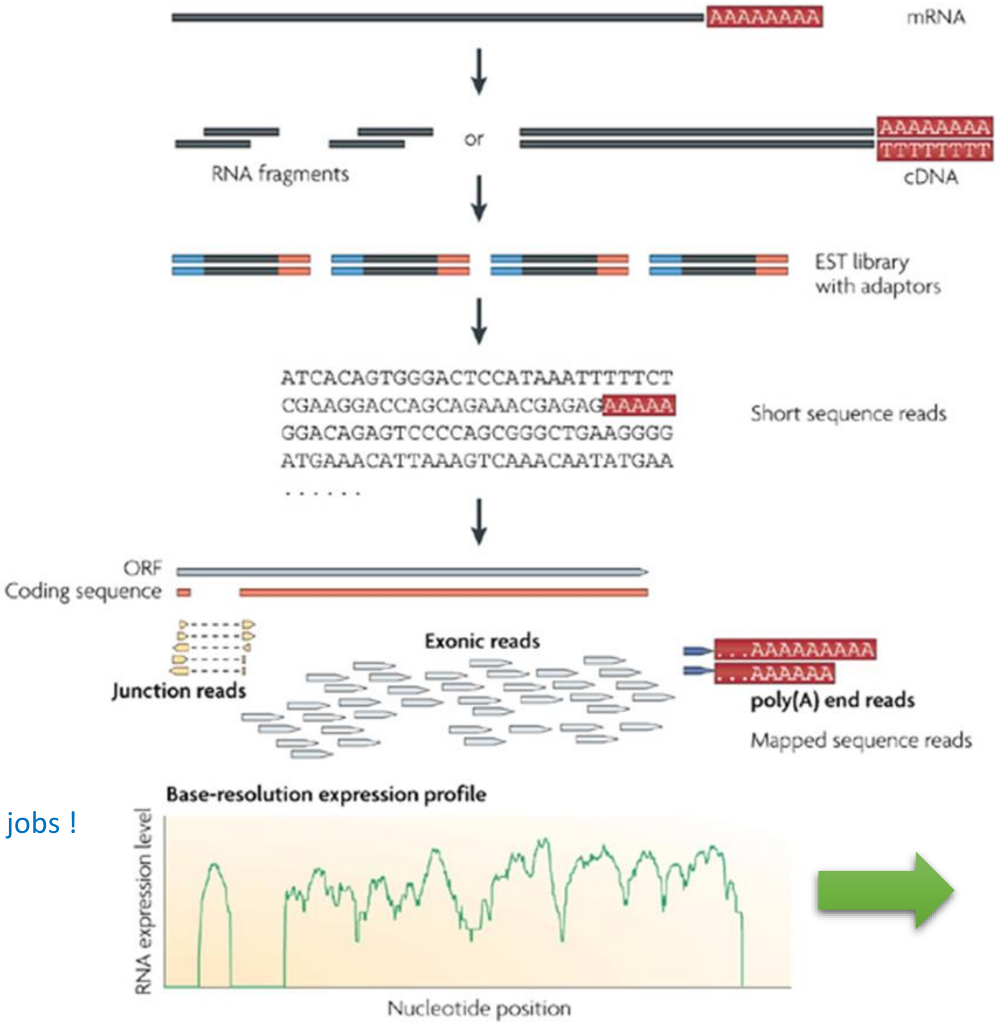
Data preprocessing

Map to reference

Read mapping

…computing intensive jobs !

Convert to expression level



Intensive analysis to Interpretate Biological Meanings

http://www.nature.com/nrg/journal/v10/n1/full/nrg2484.html

Nature Reviews | Genetics