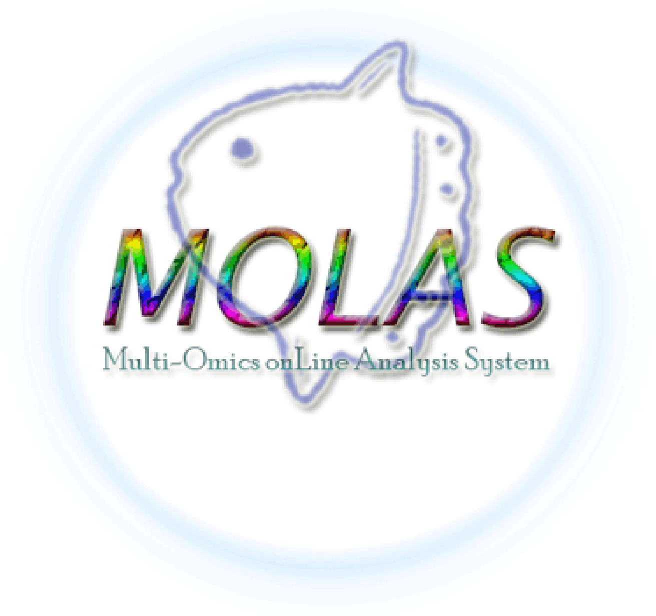
Multi-Omics onLine Analysis System for Gene Expression Profiling and Whole Methylome

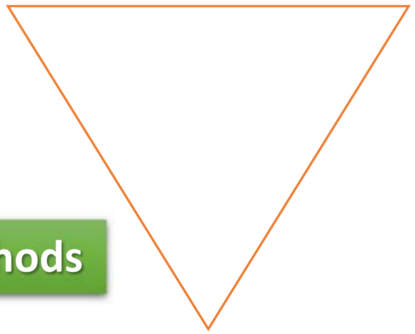


Life Science Library Training Course
2016/06/09

Chen, Shu-Hwa

IIS, Academia Sinica

High-throughput Methods



Biology Lab

Bioinfo Lab

Tech Core

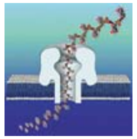
Data in public domain

Hybridization-based Methods

Microarray

Roche GS-FLX

Life Technologies SOLiD



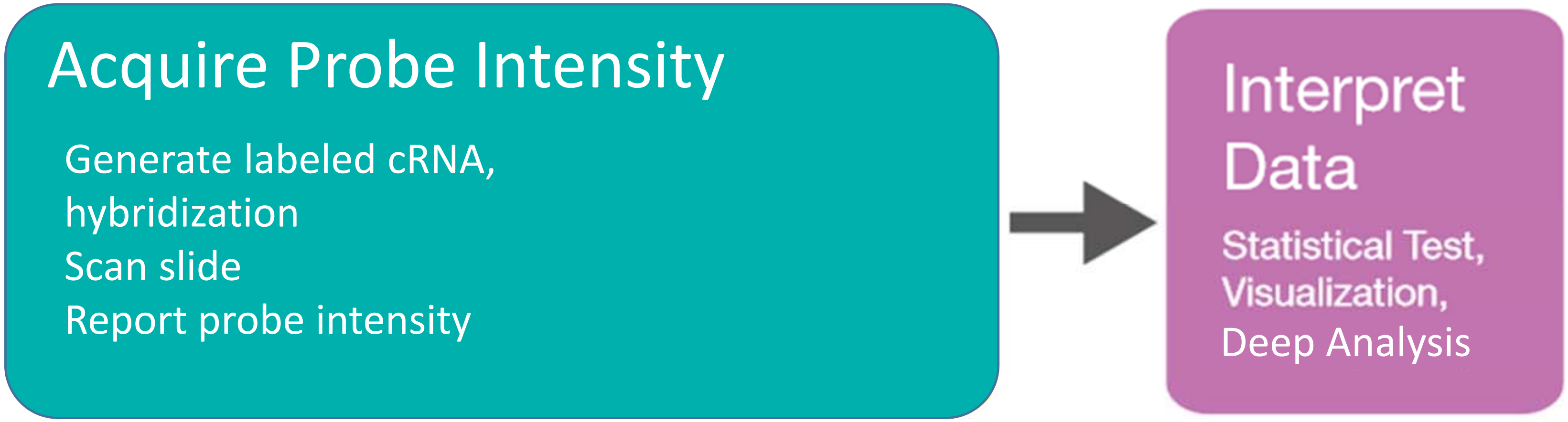
Sequencing-based Methods

Illumina HiSeq

Life Technologies Ion Torrent

PacBio

Microarray Data Analysis



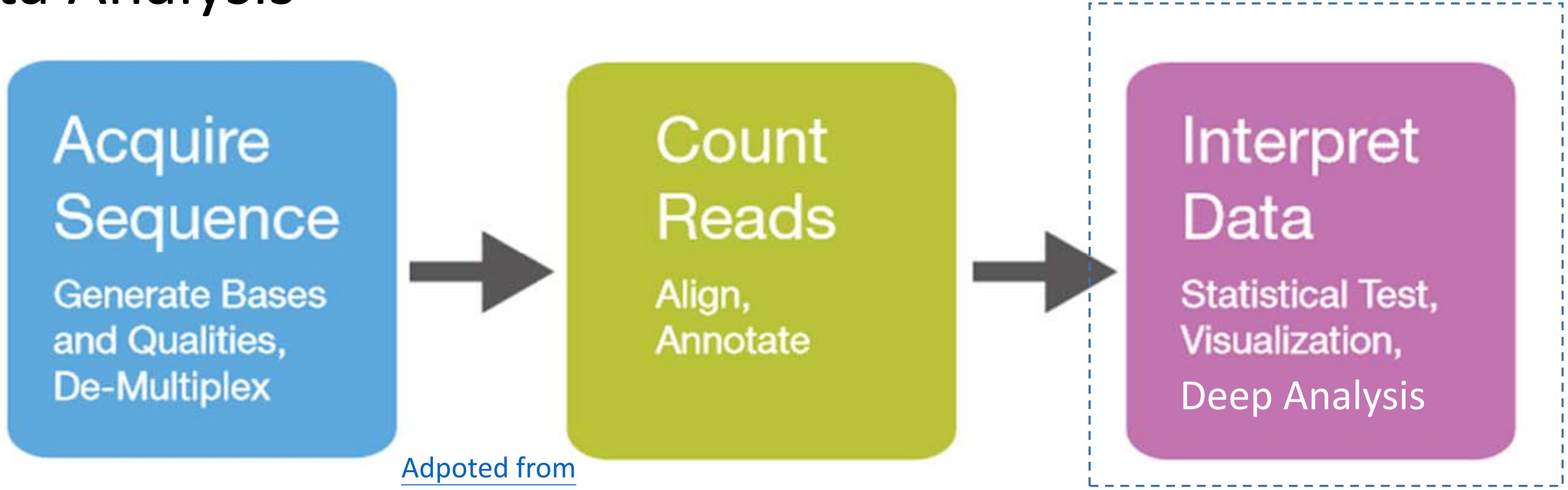
Acquire Probe Intensity

Generate labeled cRNA,
hybridization
Scan slide
Report probe intensity

Interpret Data

Statistical Test,
Visualization,
Deep Analysis

RNASeq Data Analysis



Acquire Sequence

Generate Bases
and Qualities,
De-Multiplex

Count Reads

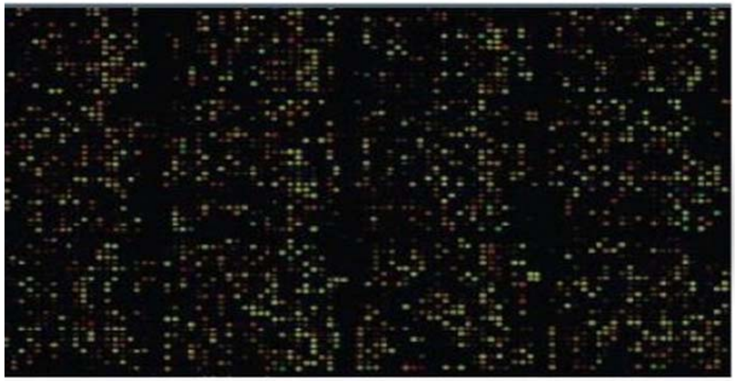
Align,
Annotate

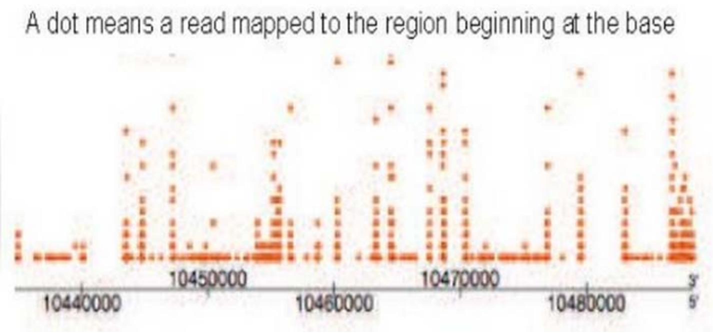
Interpret Data

Statistical Test,
Visualization,
Deep Analysis

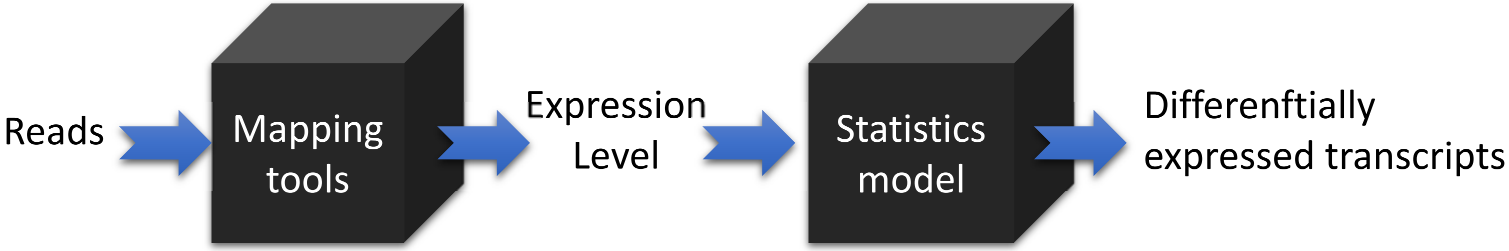
Adpoted from
http://www.illumina.com/documents/products/datasheets/datasheet_rnaseq_analysis.pdf

Analog signal vs Digital signal




A dot means a read mapped to the region beginning at the base
10440000
10450000
10460000
10470000
10480000
3'
5'

http://www.slideshare.net/ueb52/uebuat-bioinformatics-course-session-23-vhir-barcelona



A typical RNA-Seq experiment

Prepare RNA samples

↓ QC

Prepare sequencing libraries

↓ QC

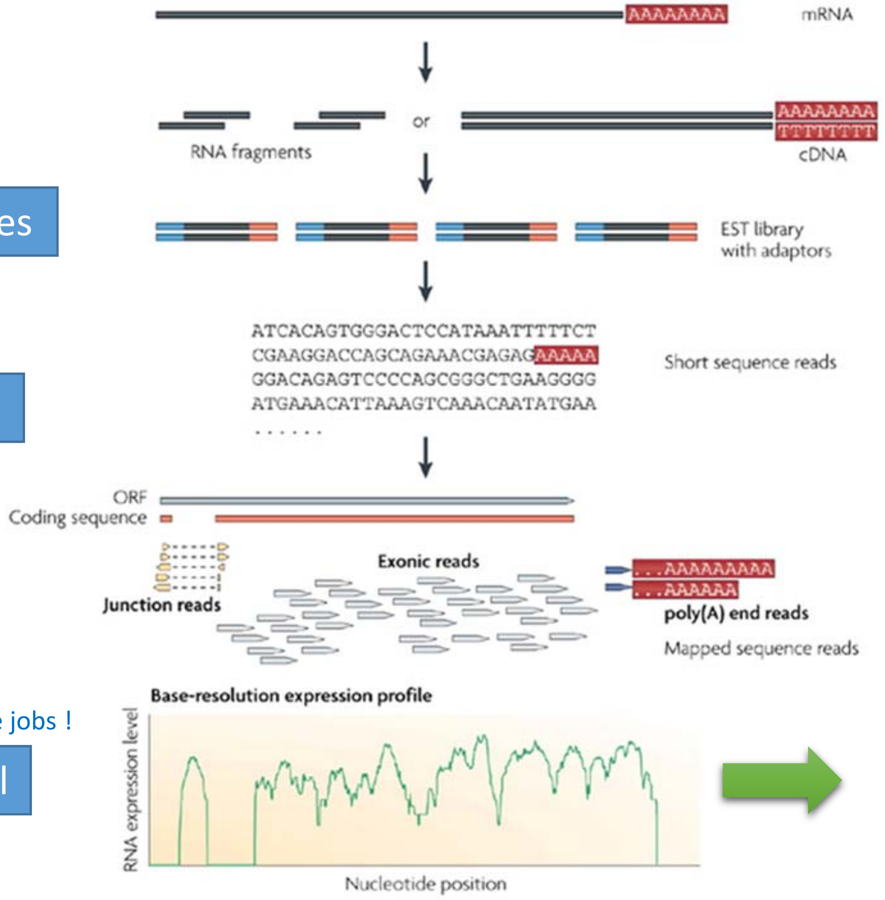
Millions of shot-gun reads

↓ QC
Data preprocessing

Map to reference

↓ Read mapping
…computing intensive jobs !

Convert to expression level



Nature Reviews | Genetics

→

Intensive analysis to Interpretate Biological Meanings

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

FastQ format

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

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

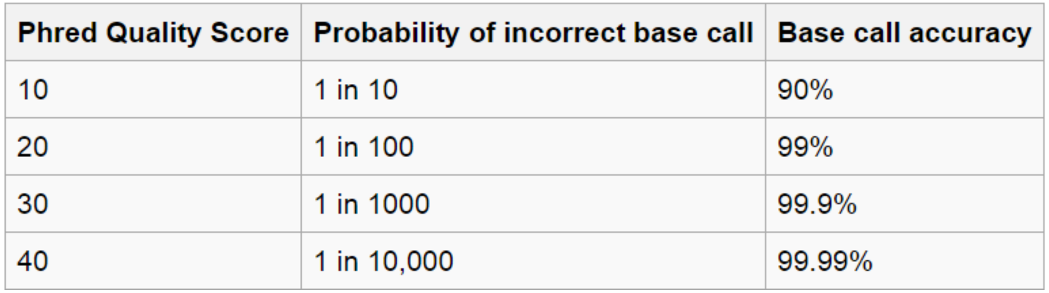
seq head	@EAS139:136:FC706VJ:2:5:1000:12850 1:N:18:ATCACG
seq letters	ACTTCAGGAGATTGTACATTTAGAGACAAAAAAAAA
+	+
quality score	BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@

FASTQ files from CASAVA-1.8 Should have the following READ-ID format:

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>

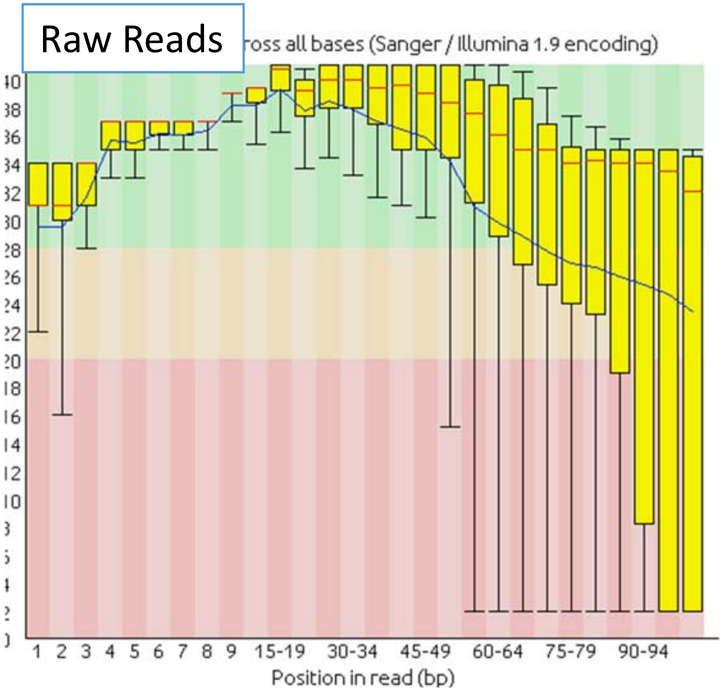
<read>:<is filtered>:<control number>:<index sequence>

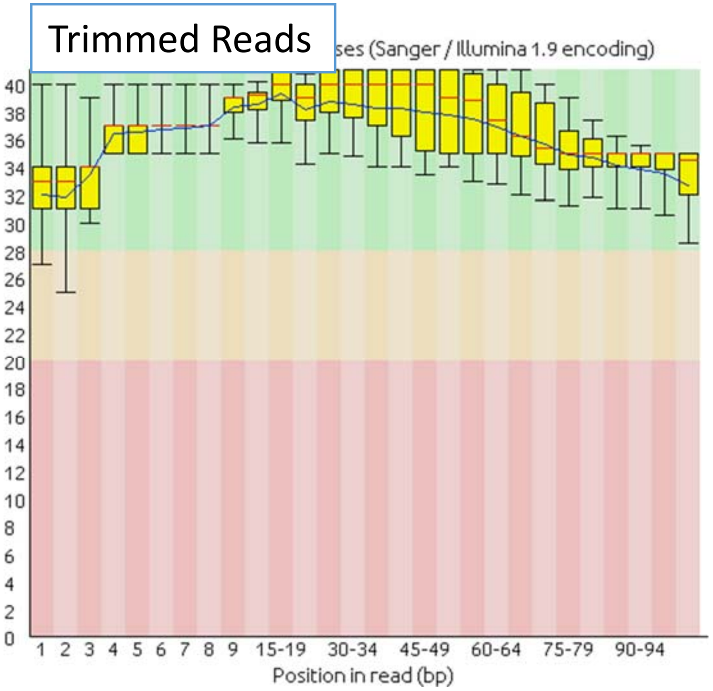
Read preprocessing



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%

- Trimming: by base quality

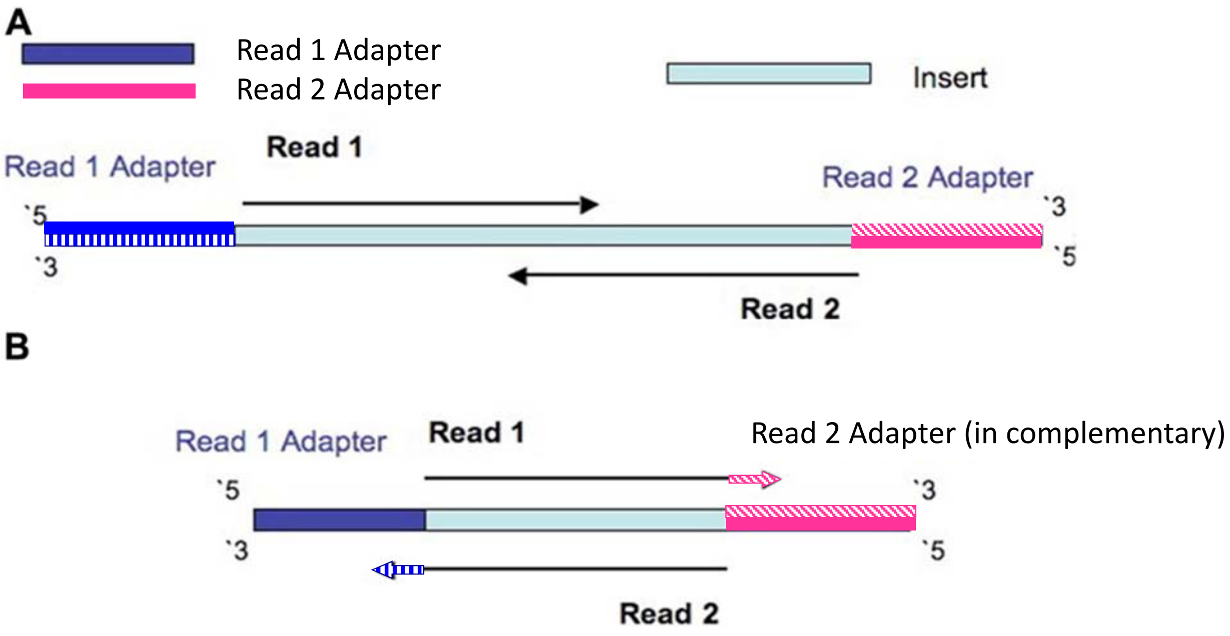

Raw Reads


Trimmed Reads

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

Read preprocessing

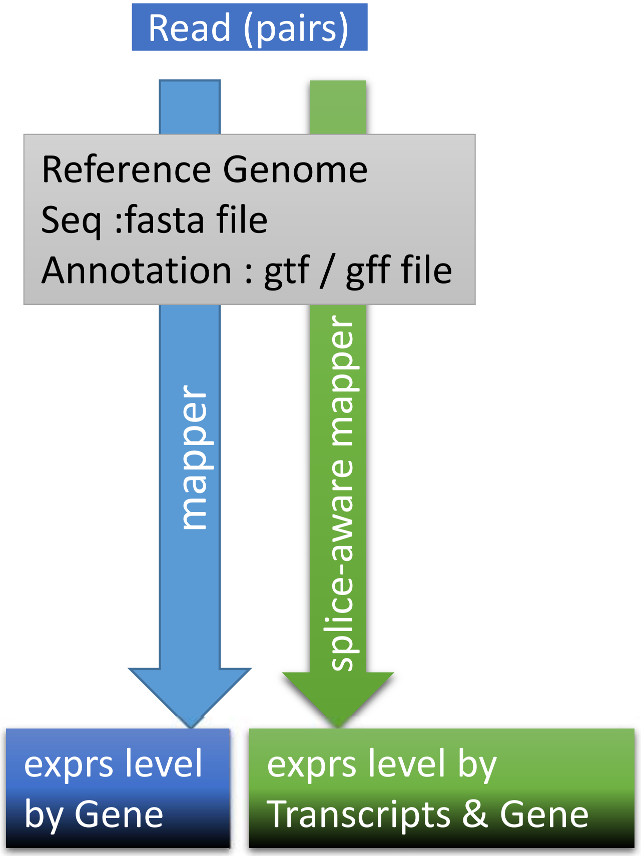
- Trimming: adapter contamination



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

Expression Level

by Gene or by Transcript?



Read (pairs)

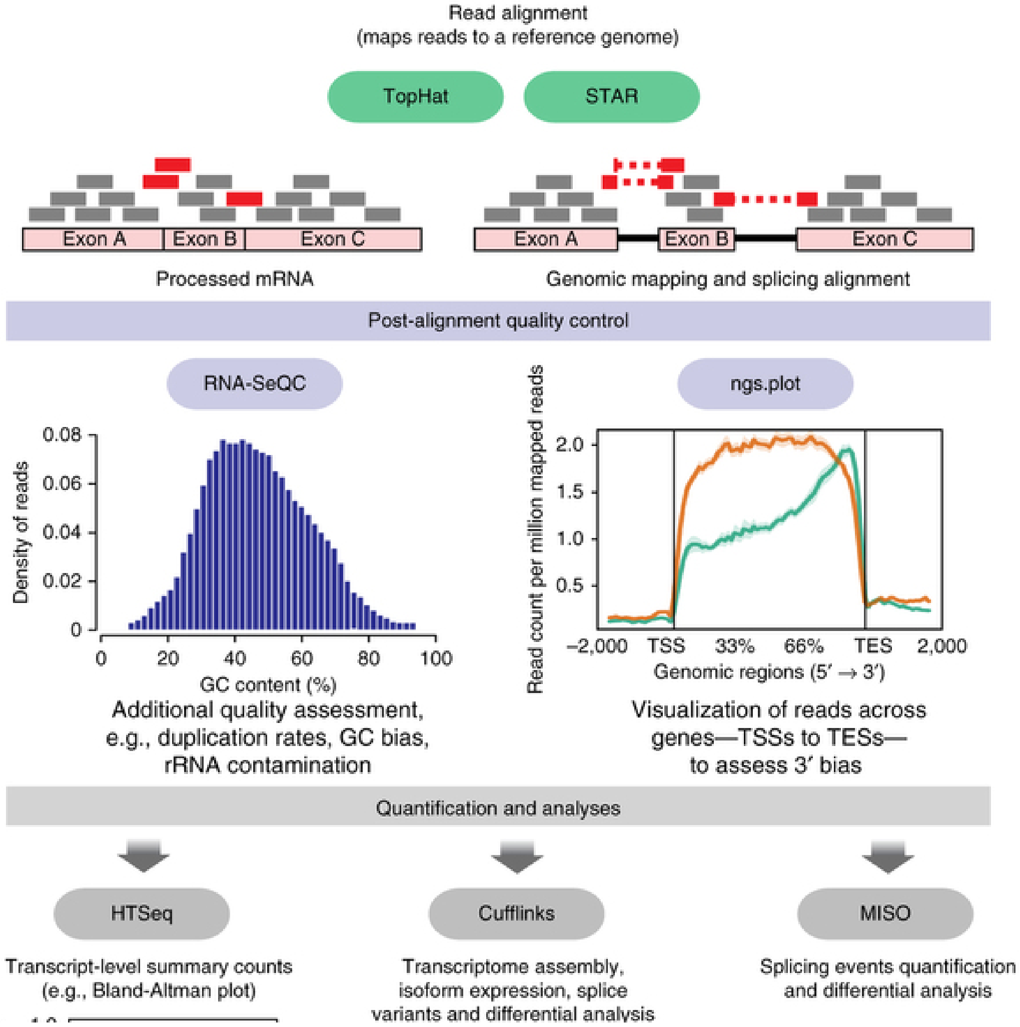
Reference Genome
Seq :fasta file
Annotation : gtf / gff file

mapper

splice-aware mapper

exprs level by Gene

exprs level by Transcripts & Gene



Read alignment
(maps reads to a reference genome)
TopHat
STAR
Exon A
Exon B
Exon C
Processed mRNA
Exon A
Exon B
Exon C
Genomic mapping and splicing alignment
Post-alignment quality control
RNA-SeQC
0.08
0.06
0.04
0.02
0
Density of reads
0
20
40
60
80
100
GC content (%)
Additional quality assessment, e.g., duplication rates, GC bias, rRNA contamination
ngs.plot
Read count per million mapped reads
2.0
1.5
1.0
0.5
−2,000
TSS
33%
66%
TES
2,000
Genomic regions (5′ → 3′)
Visualization of reads across genes—TSSs to TESs—to assess 3′ bias
Quantification and analyses
HTSeq
Transcript-level summary counts (e.g., Bland-Altman plot)
Cufflinks
Transcriptome assembly, isoform expression, splice variants and differential analysis
MISO
Splicing events quantification and differential analysis

http://www.nature.com/neuro/journal/v17/n11/fig_tab/nn.3816_F1.html

Other issues

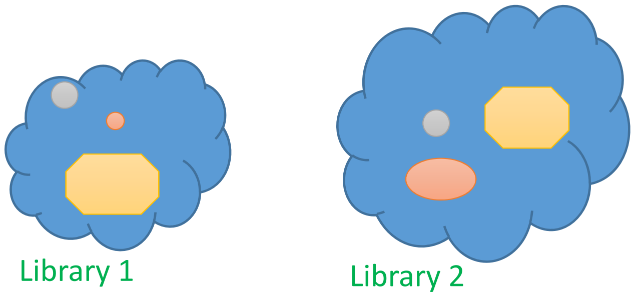
- Stranded or not?
- PolyA tailed or rRNA depletion?
- Have reference genome? Novel transcripts? Fusion transcripts?

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

Normalization is a Necessary Evil

- Between samples:

  Initial Input ; Volume of Reads



- Within sample:

  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

How many replicates does the experiment design includes?

- Theoretically………… BUT! in reality …………….
- Borrowing information among genes to get better estimates.
- Count-based model
  - edgeR, DESeq ….. etc.
  - Use "read count" (or estimated count from RSEM) and enforced a normalization model to fit data to the statistic assumption
  - Want to provide an analysis with statistic power
- Programs like SAMSeq (rank-based model, only applicable for large replicates) and limma are fine with continuous values (like FPKM). Limma takes more cares about weak mean-variance relationship (stabilizing variation).

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

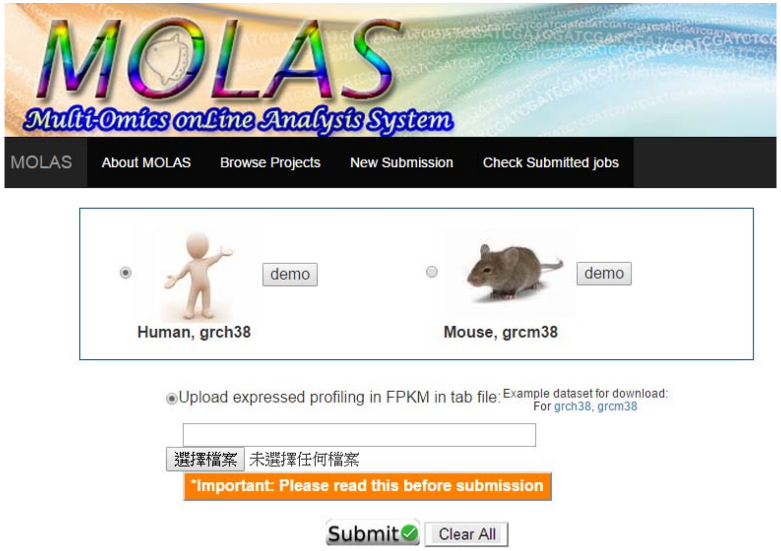
MOLAS

Multi-Omics onLine Analysis System

The Usage

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

多重體學線上分析平台
Multi-Omics onLine Annotation System (MOLAS)



● Human, grch38 ○ Mouse, grcm38

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

未選擇任何檔案

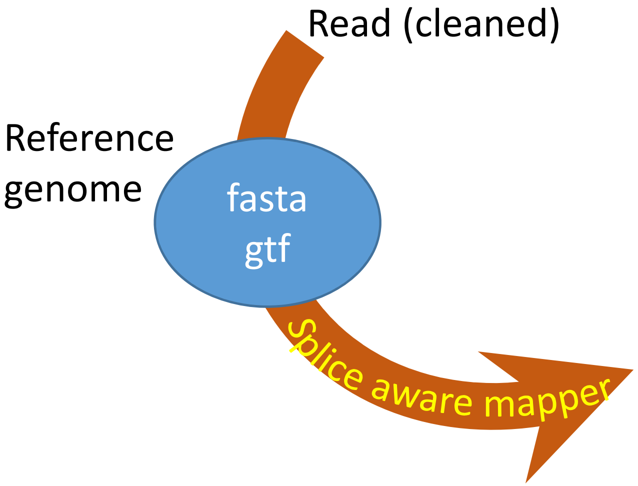
*Important: Please read this before submission

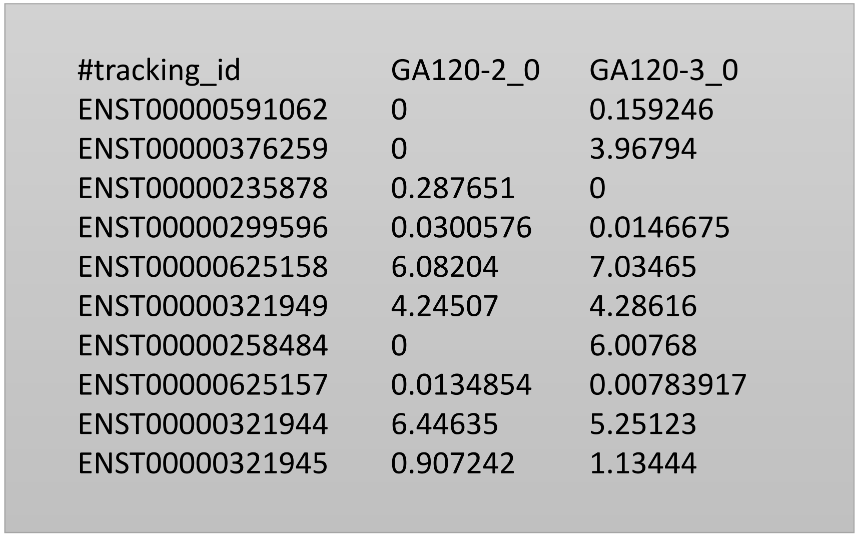
To view and analyse your RNASeq experiment

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
```

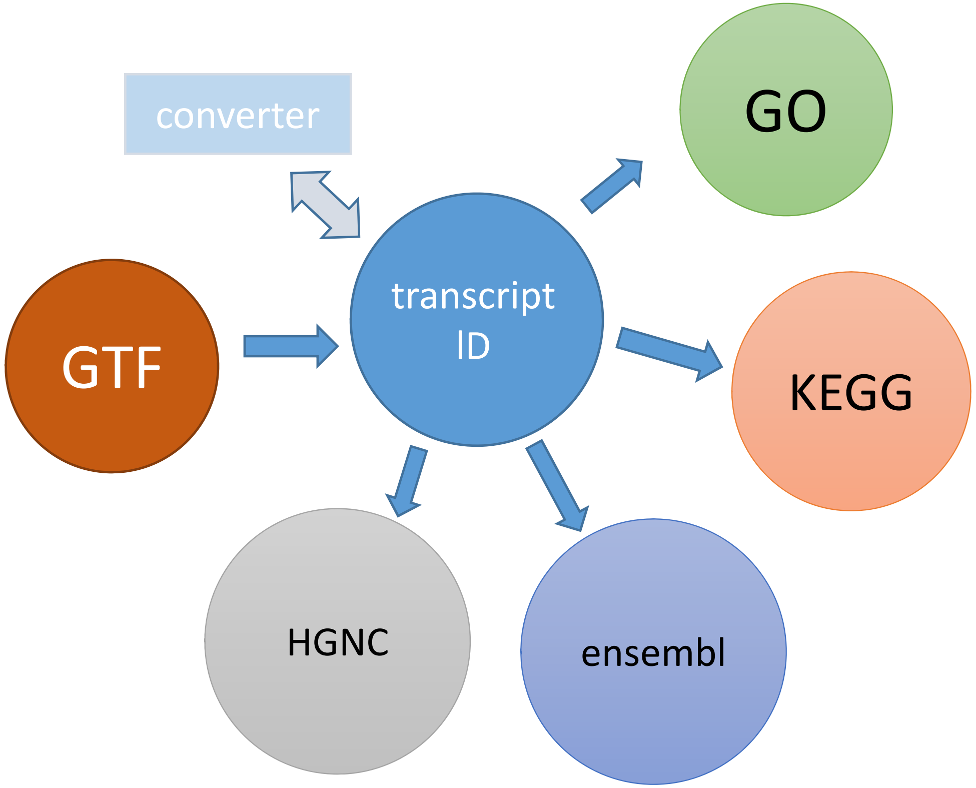
GTF: the Gene Tranfer Format

```
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";
```

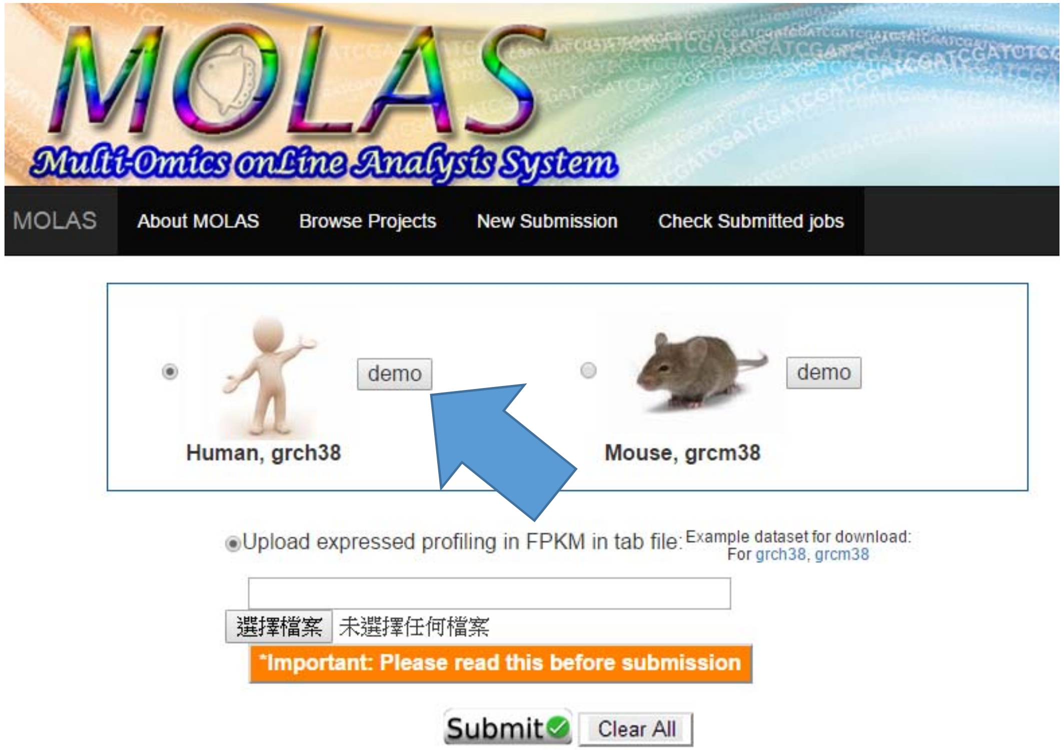
MOLAS compatible GTF

grch38
grch37

grcm38



New Submssion

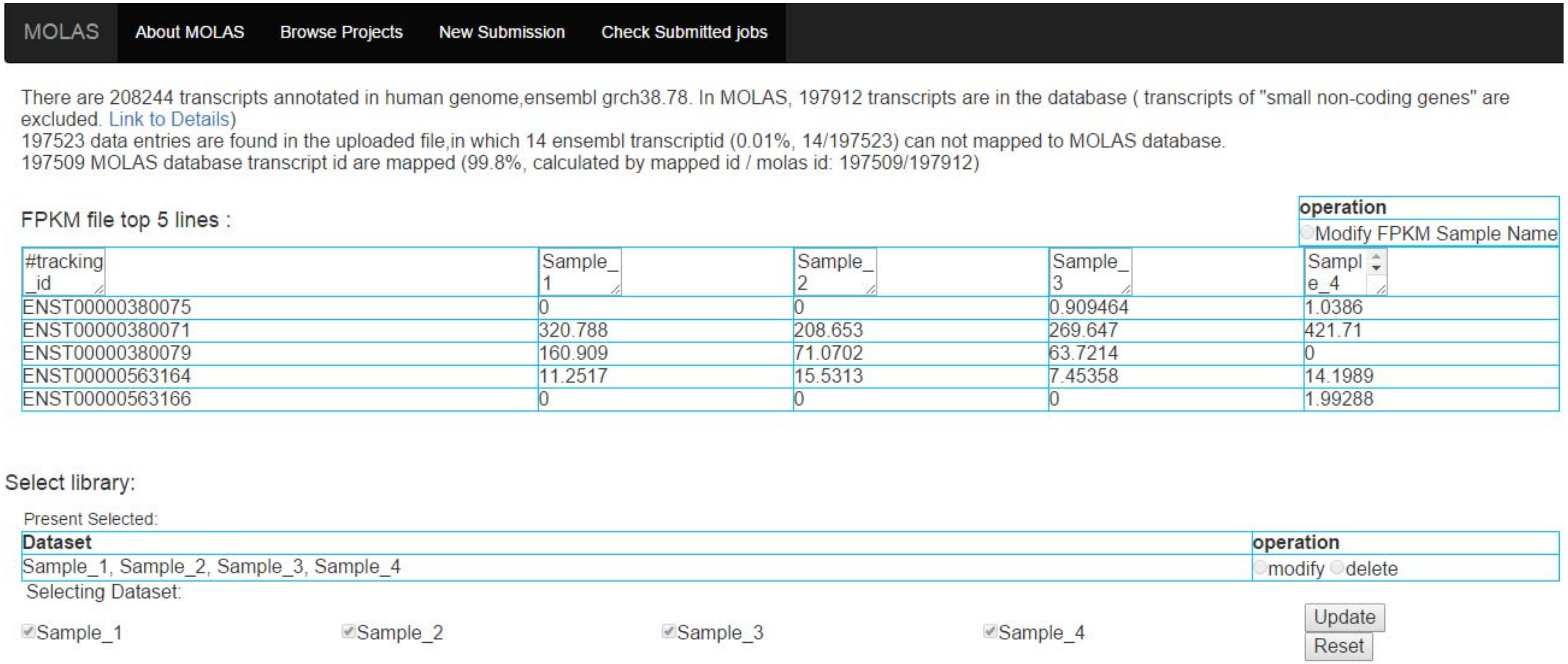


● Human, grch38

○ Mouse, grcm38

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

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:

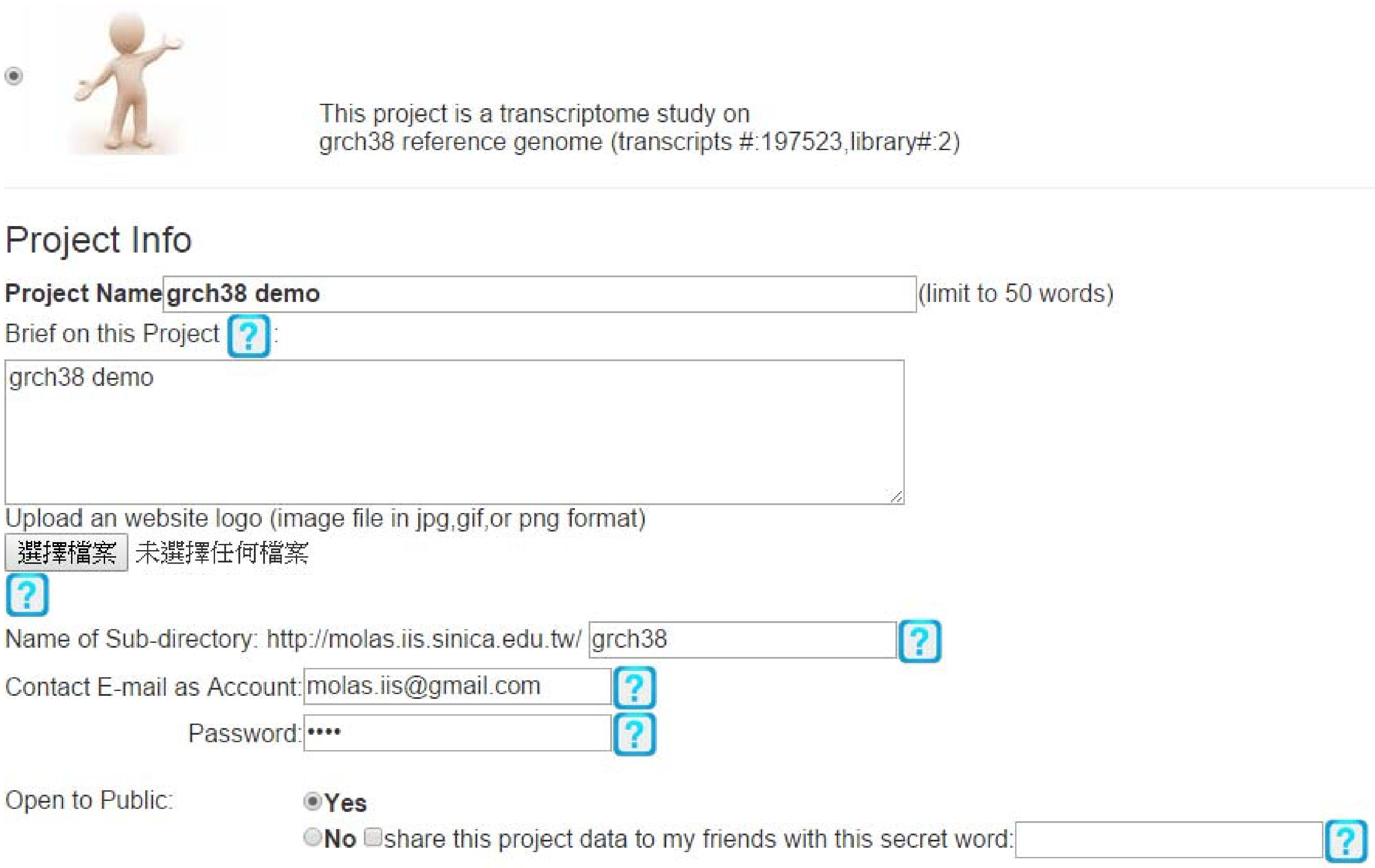
Present Selected:

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)

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

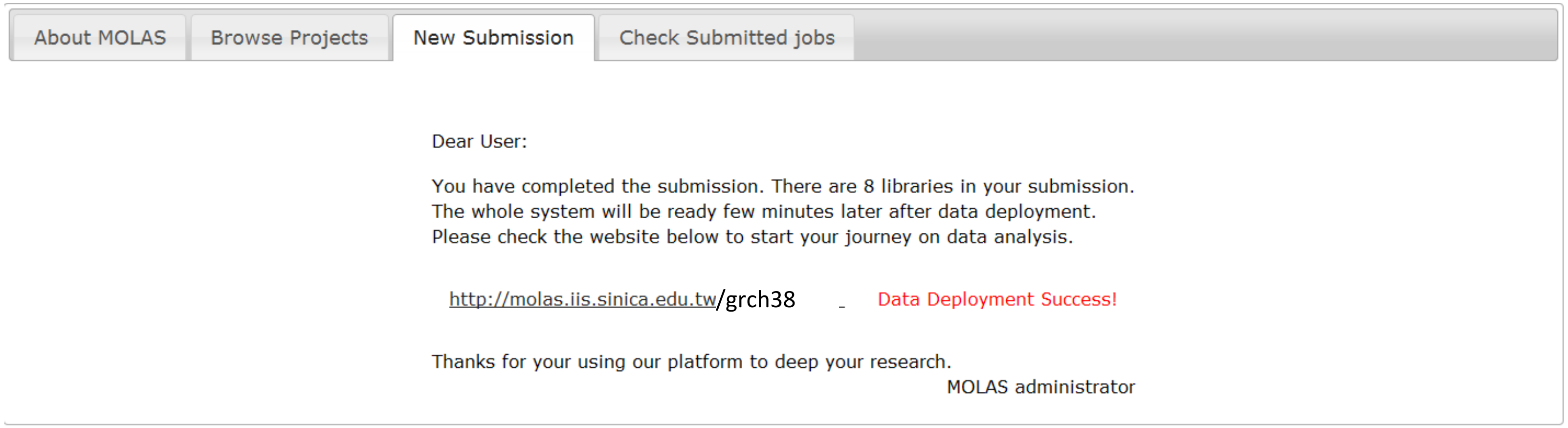
Contact E-mail as Account: molas.iis@gmail.com ?

Password: •••• ?

Open to Public: ● **Yes**

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

Deployment Success



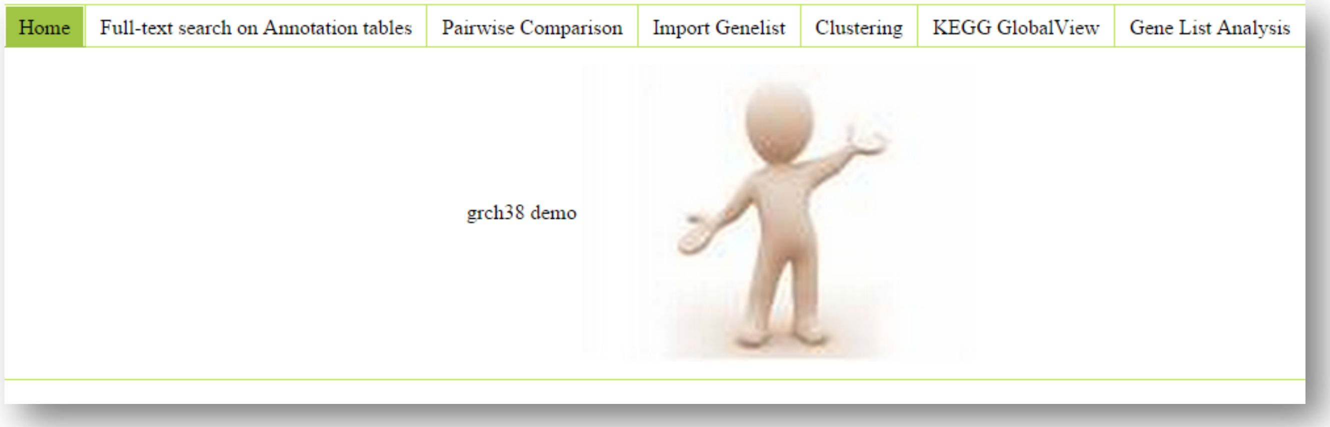
Dear User:

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!

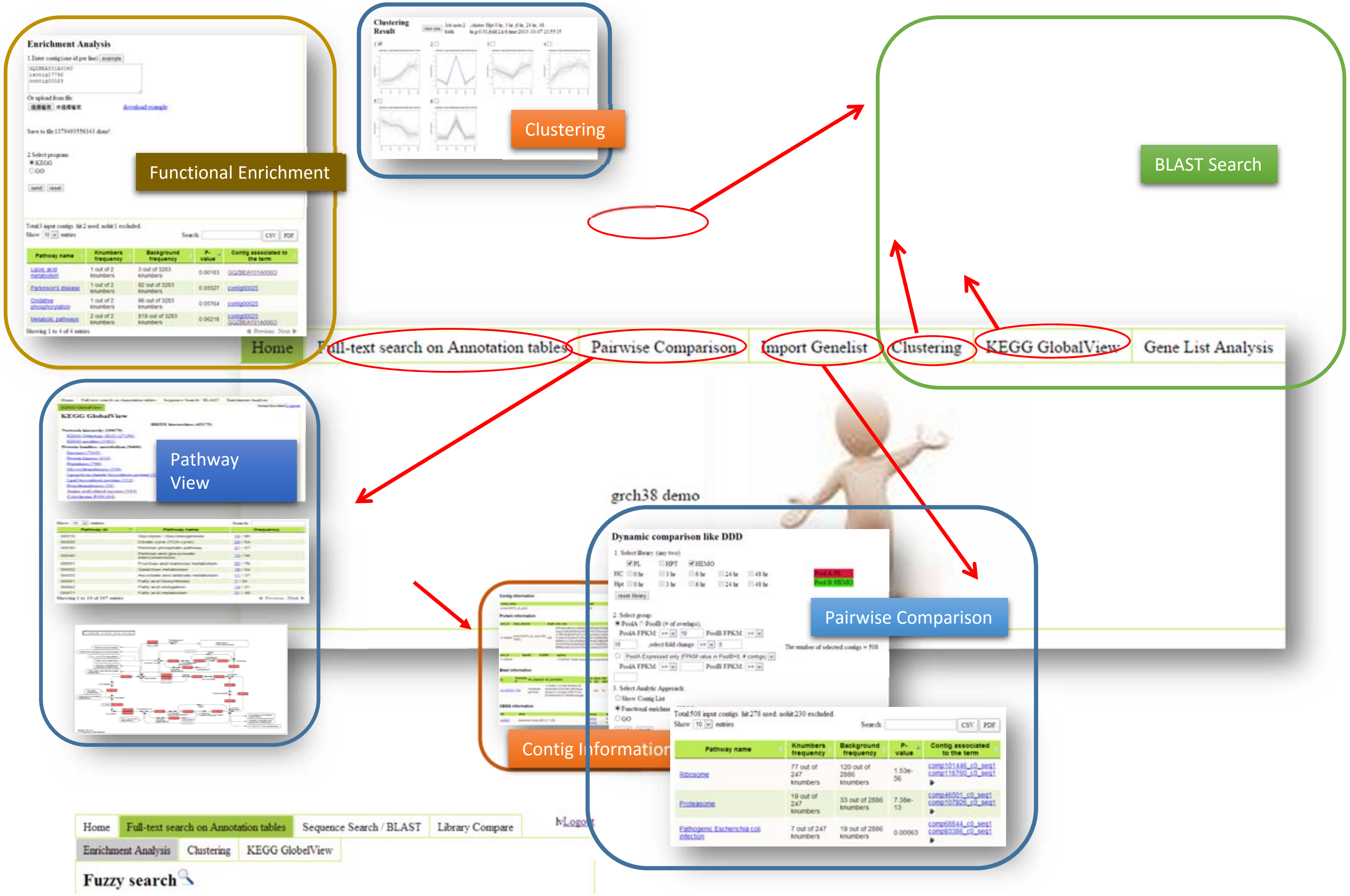
Thanks for your using our platform to deep your research.
MOLAS administrator

Browse project and …….



grch38 demo

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



Enrichment Analysis

1.Enter contig(one id per line) example

GQZBEAY01A0060
isotig17792
contig00025

Or upload from file:

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

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

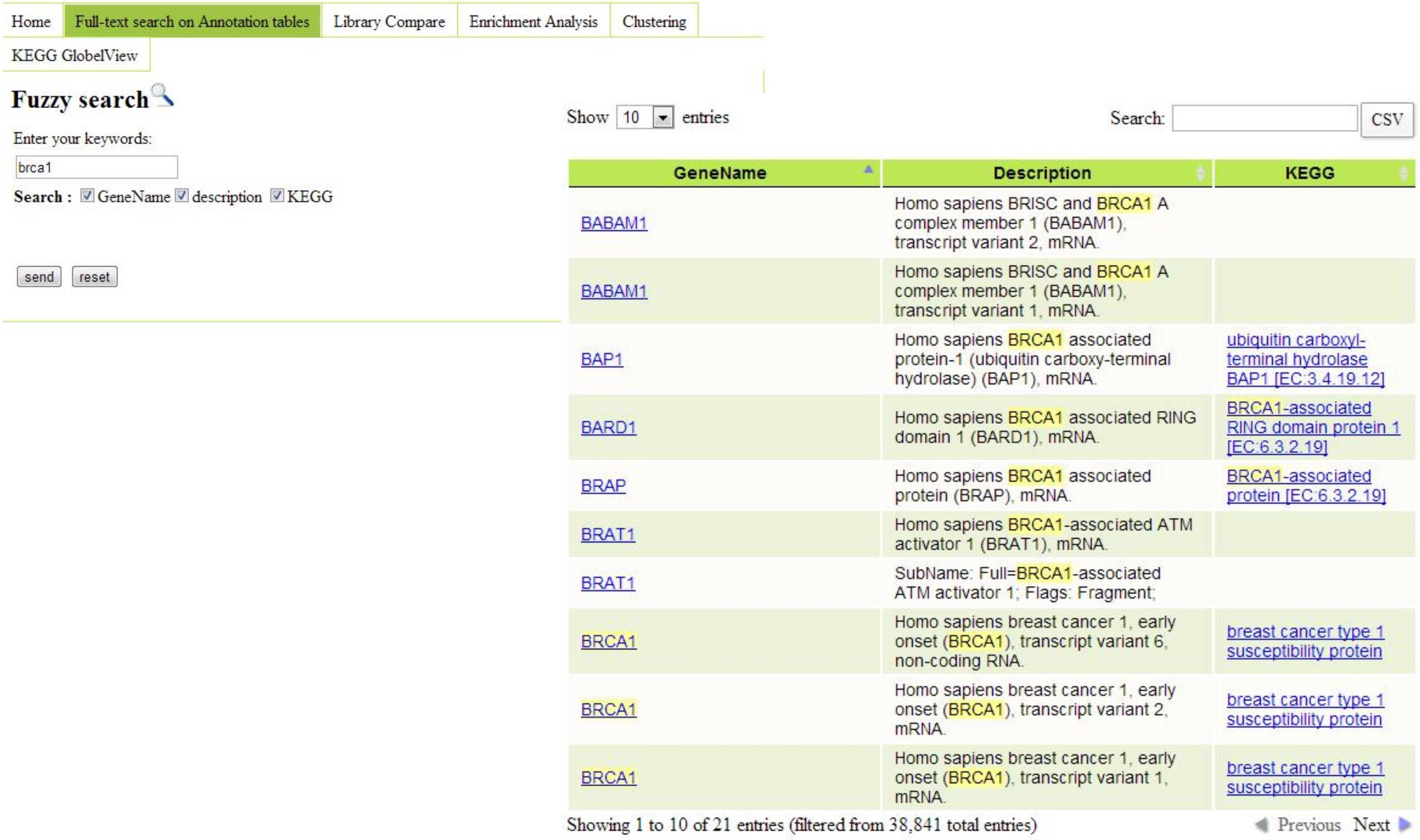
3. Select Analytic Approach:
○ Show Contig List
● Functional enrichme
○ GO

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	7.36e-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

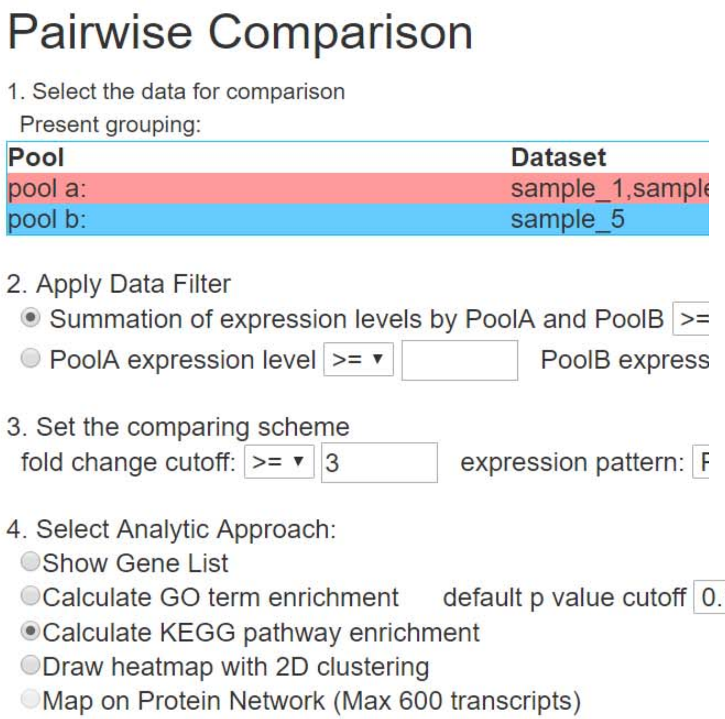
Search:

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 Comparis

Select libraries you



Pairwise Comparison

1. Select the data for comparison

Present grouping:

Pool	Dataset
pool a:	sample_1,sample
pool b:	sample_5

2. Apply Data Filter

● Summation of expression levels by PoolA and PoolB >=

○ PoolA expression level >= ▾ PoolB express

3. Set the comparing scheme

fold change cutoff: >= ▾ 3 expression pattern: F

4. Select Analytic Approach:

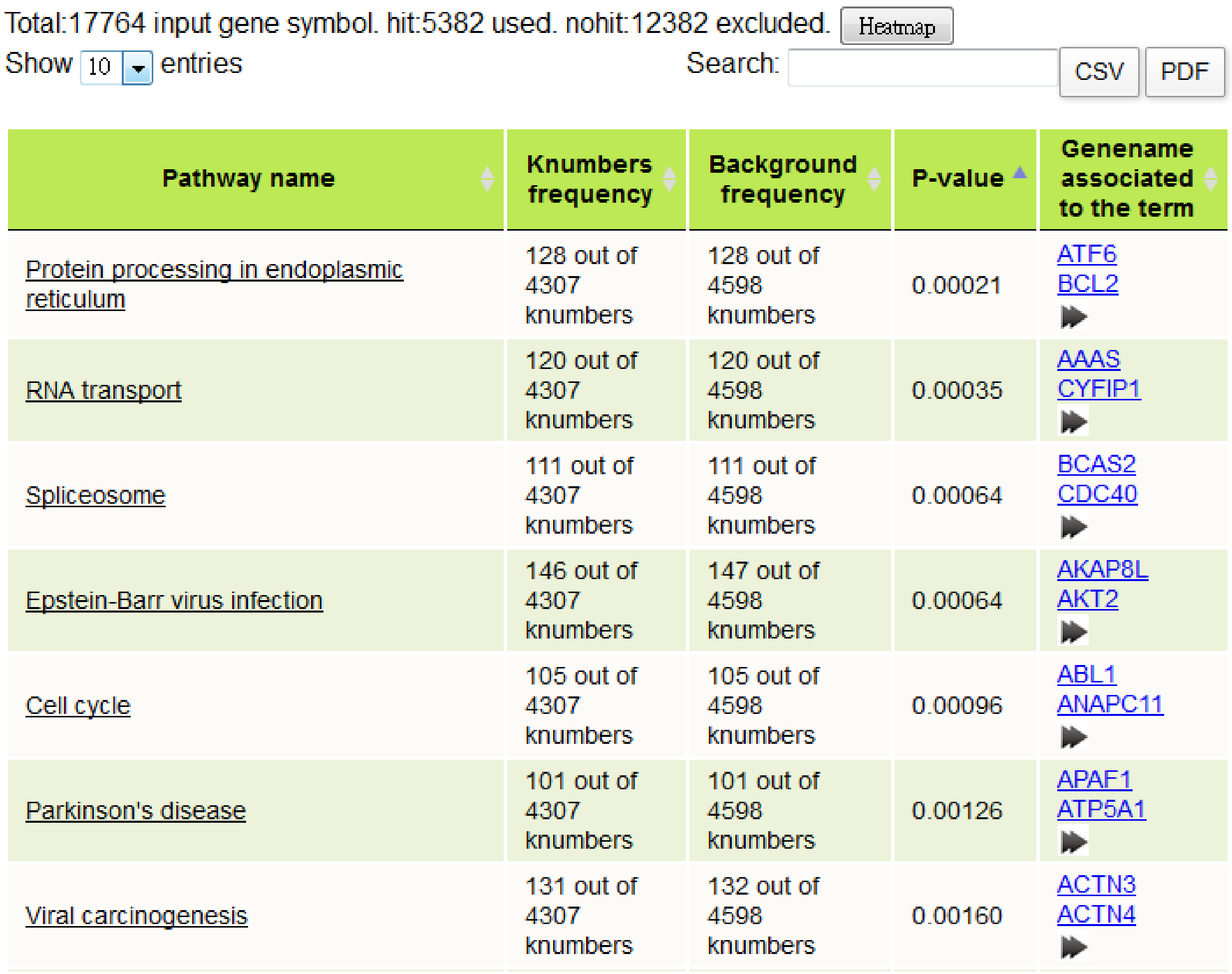
○ Show Gene List

○ Calculate GO term enrichment default p value cutoff 0.1 ▾

● Calculate KEGG pathway enrichment

○ Draw heatmap with 2D clustering

○ Map on Protein Network (Max 600 transcripts)



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

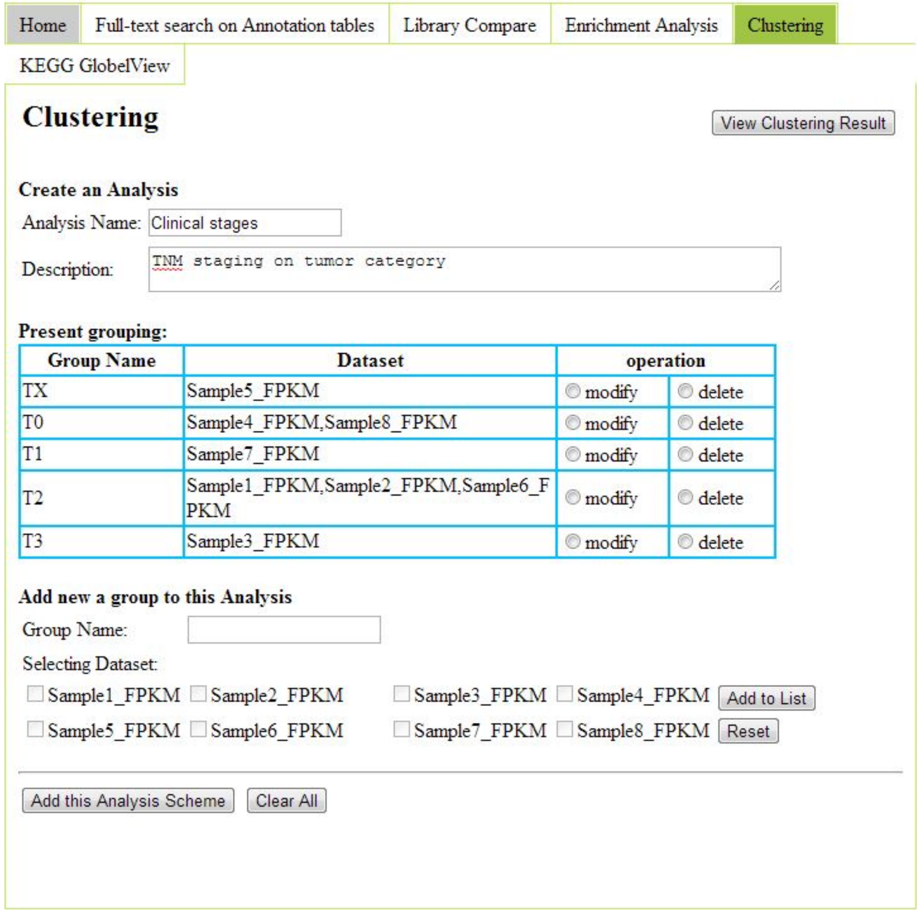
Show 10 ▾ entries

Search:

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 ▶▶

Clustering

If some samples have similar properties, clustering can help group them together and perform gene expression profile analysis.



Clustering

Create an Analysis

Analysis Name: Clinical stages

Description: TNM staging on tumor category

Present grouping:

Group Name	Dataset	operation	
TX	Sample5_FPKM	○ modify	○ delete
T0	Sample4_FPKM,Sample8_FPKM	○ modify	○ delete
T1	Sample7_FPKM	○ modify	○ delete
T2	Sample1_FPKM,Sample2_FPKM,Sample6_FPKM	○ modify	○ delete
T3	Sample3_FPKM	○ modify	○ delete

Add new a group to this Analysis

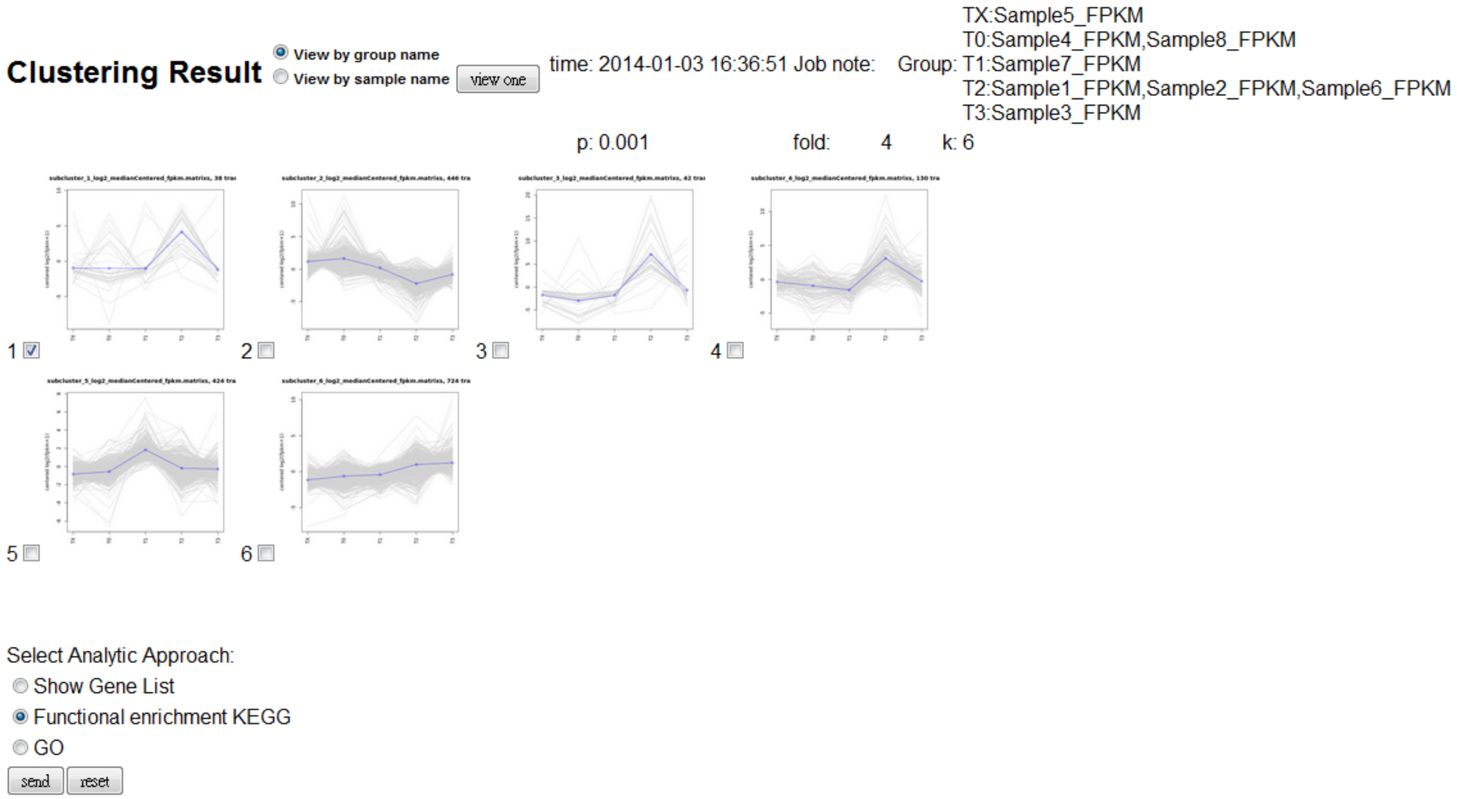
Group Name:

Selecting Dataset:

□ Sample1_FPKM □ Sample2_FPKM □ Sample3_FPKM □ Sample4_FPKM

□ Sample5_FPKM □ Sample6_FPKM □ Sample7_FPKM □ Sample8_FPKM

Clustering Results



Clustering Result ● View by group name ○ View by sample name [view one] time: 2014-01-03 16:36:51 Job note: Group: TX:Sample5_FPKM
T0:Sample4_FPKM,Sample8_FPKM
T1:Sample7_FPKM
T2:Sample1_FPKM,Sample2_FPKM,Sample6_FPKM
T3:Sample3_FPKM

p: 0.001 fold: 4 k: 6

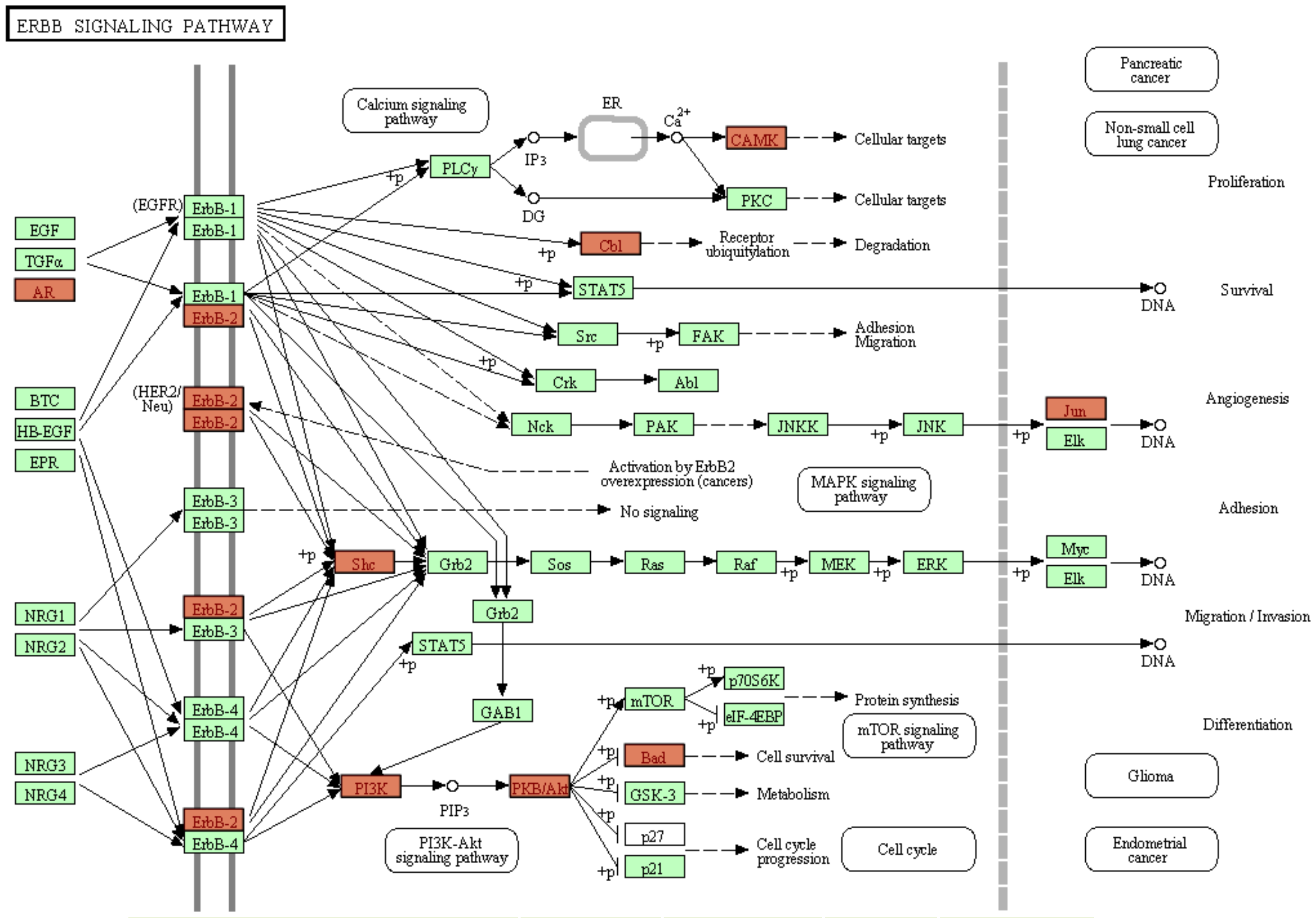
subcluster_1_log2_medianCentered_fpkm.matrix, 38 tra
subcluster_2_log2_medianCentered_fpkm.matrix, 446 tra
subcluster_3_log2_medianCentered_fpkm.matrix, 42 tra
subcluster_4_log2_medianCentered_fpkm.matrix, 130 tra
subcluster_5_log2_medianCentered_fpkm.matrix, 424 tra
subcluster_6_log2_medianCentered_fpkm.matrix, 724 tra
centered log2(fpkm+1)

1 ☑ 2 □ 3 □ 4 □
5 □ 6 □

Select Analytic Approach:
○ Show Gene List
● Functional enrichment KEGG
○ GO

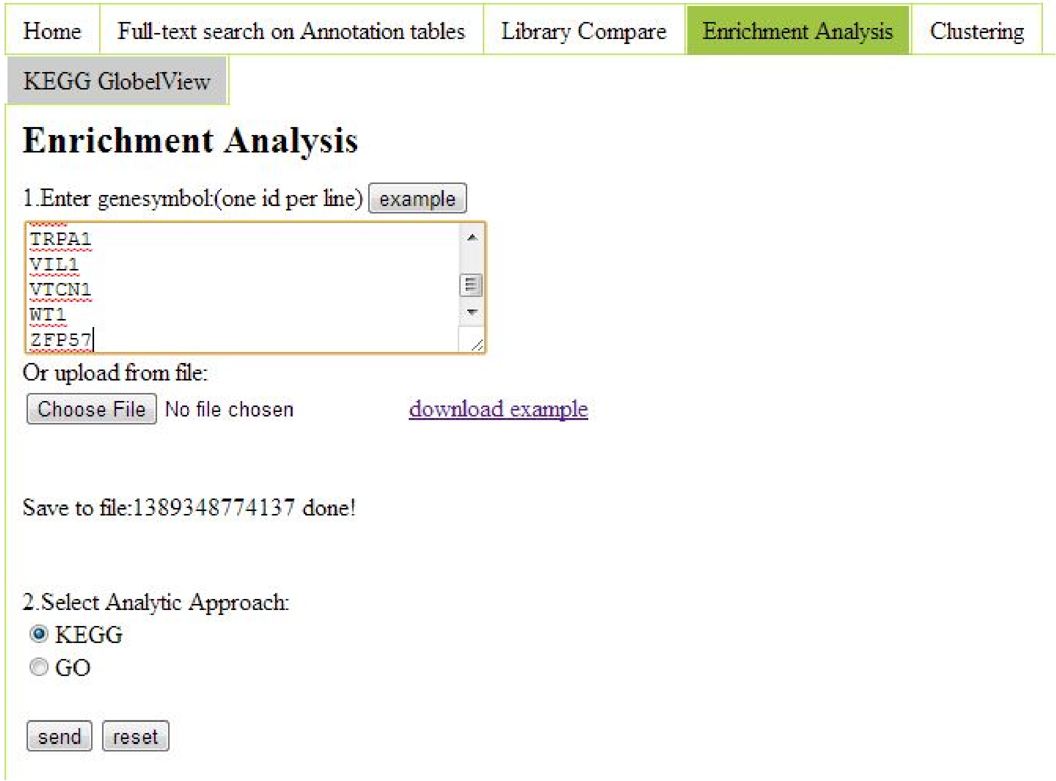
[send] [reset]

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)

```
TRPA1
VIL1
VTCN1
WT1
ZFP57
```

Or upload from file:

Save to file:1389348774137 done!

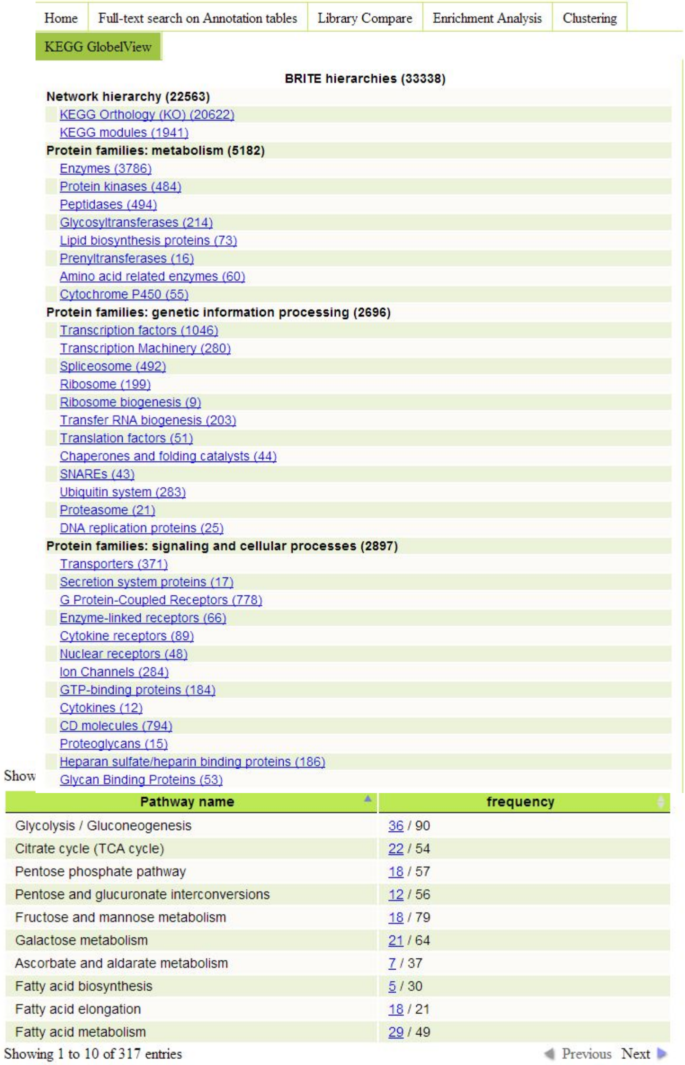
2.Select Analytic Approach:

● KEGG

○ GO

KEGG Global View

KEGG Global View provide an overview picture of KEGG pathway of human (hg19) and mouse (mm10) organisms. You can investigate specific metabolic pathway by exploring each category.



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

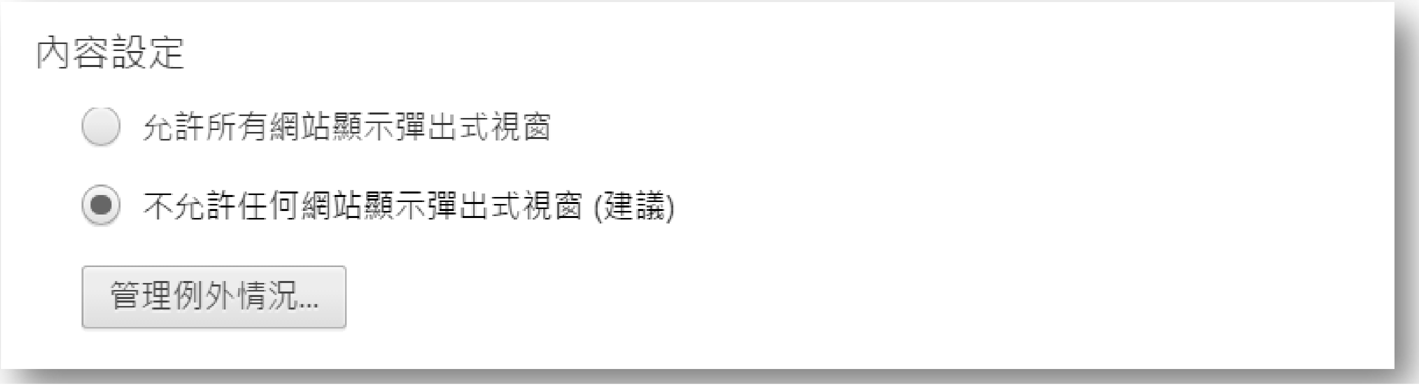
Hands on practice on MOLAS

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

MOLAS
Multi-Omics onLine Analysis System

http://molas.iis.sinica.edu.tw/human_demo_grch38/

http://molas.iis.sinica.edu.tw/mouse_demo_grcm38/



內容設定

○ 允許所有網站顯示彈出式視窗

● 不允許任何網站顯示彈出式視窗 (建議)

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 attempt a significance analysis. This may be the best advice.
- Simply pick a reasonable dispersion value, based on your experience with similar data, and use that for DE detection
  - In edgeR (empirically):

    0.4 human data (genetically “not” identical)

    0.1 for “genetically identical” strains of model organisms

    0.01 for technical replicates
- estimate dispersion from dataset reducing one (less critical) experiment factor
- estimate dispersion from a sizeable number of control transcripts that should not be DE if there exists

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

Limitations

- Assumption of “Uniformity” of all expressed transcripts may not always true
- Uncertain problems in mapping
  - Transcripts length issue
  - Redundance seq in genome
  - Reference is never a perfect match to the actual biological source of RNA being sequenced
- Reference & no Reference
- Lag in analytic tools.
- No single robotic analylsis scheme fits all kind of needs
- Cost !!

epi-MOLAS

Epi-genoMics OnLine Analysis System

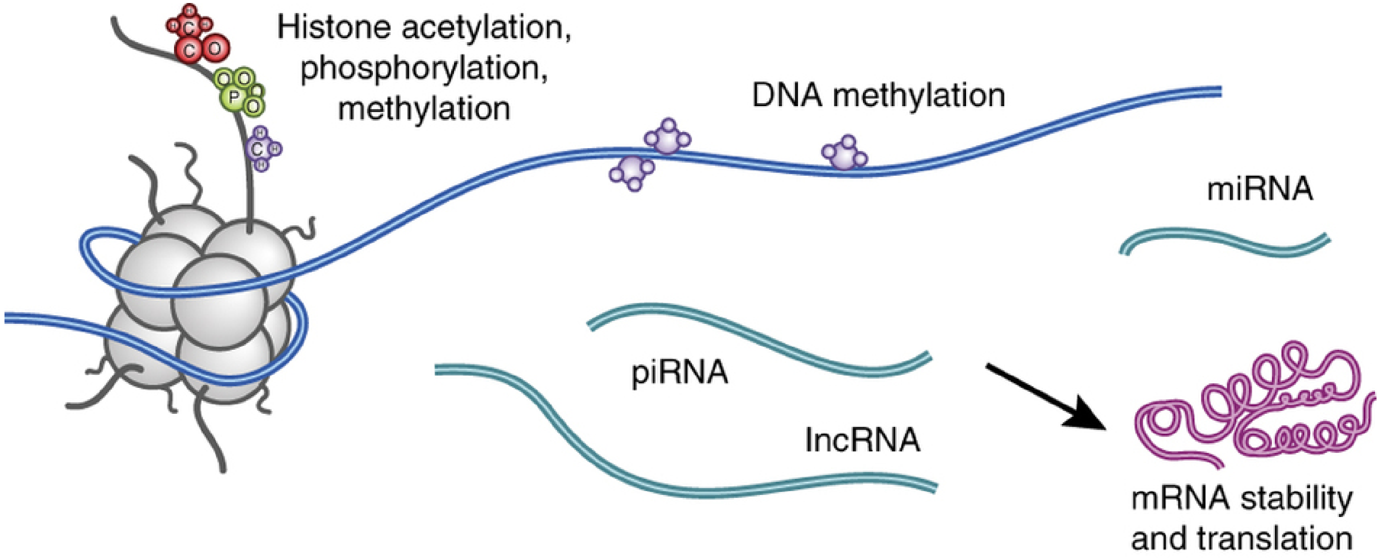
Shu-Hwa Chen

Institute of Information Science
Academia Sinica, Taiwan

2016/10/27

Epigenetic Modification

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



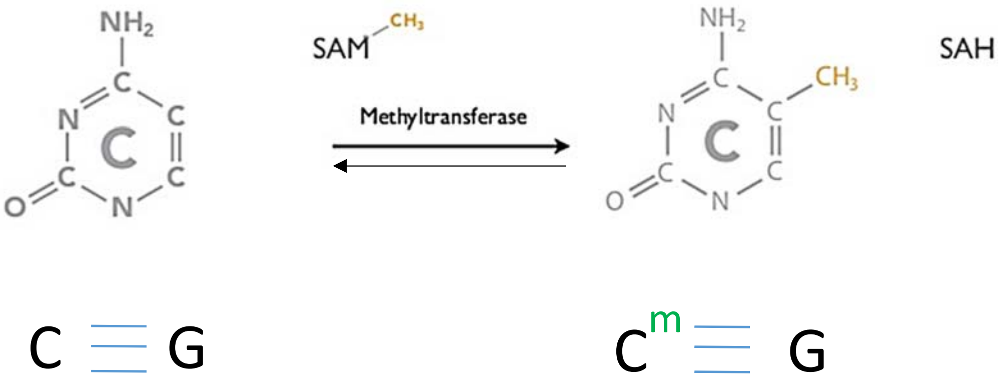
One Genome

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……..

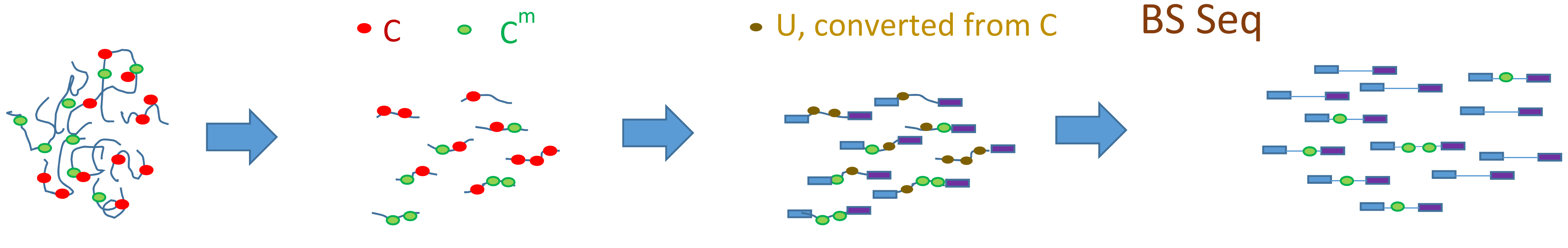
Whole Genome Shotgun Bisulfite Sequencing

Bisulfite Conversion

DNA:

$$5' – \text{T C G C C G G T A C -} \xrightarrow{\text{bisulfite}} \text{T U G C U G G T A U} \xrightarrow{\text{PCR}} \begin{array}{cccccccccc} T & \boxed{T} & G & C & \boxed{T} & G & G & T & A & \boxed{T} \\ a & \boxed{a} & c & g & \boxed{a} & c & c & a & t & \boxed{a} \end{array}$$

$$5' – \text{T } C^{m}\text{G C } C^{m}\text{G G T A C -} \xrightarrow{\text{bisulfite}} \text{T } C^{m}\text{G C } C^{m}\text{G G T A U} \xrightarrow{\text{PCR}} \begin{array}{cccccccccc} T & C & G & C & C & G & G & T & A & \boxed{T} \\ a & g & c & g & g & c & c & a & t & \boxed{a} \end{array}$$



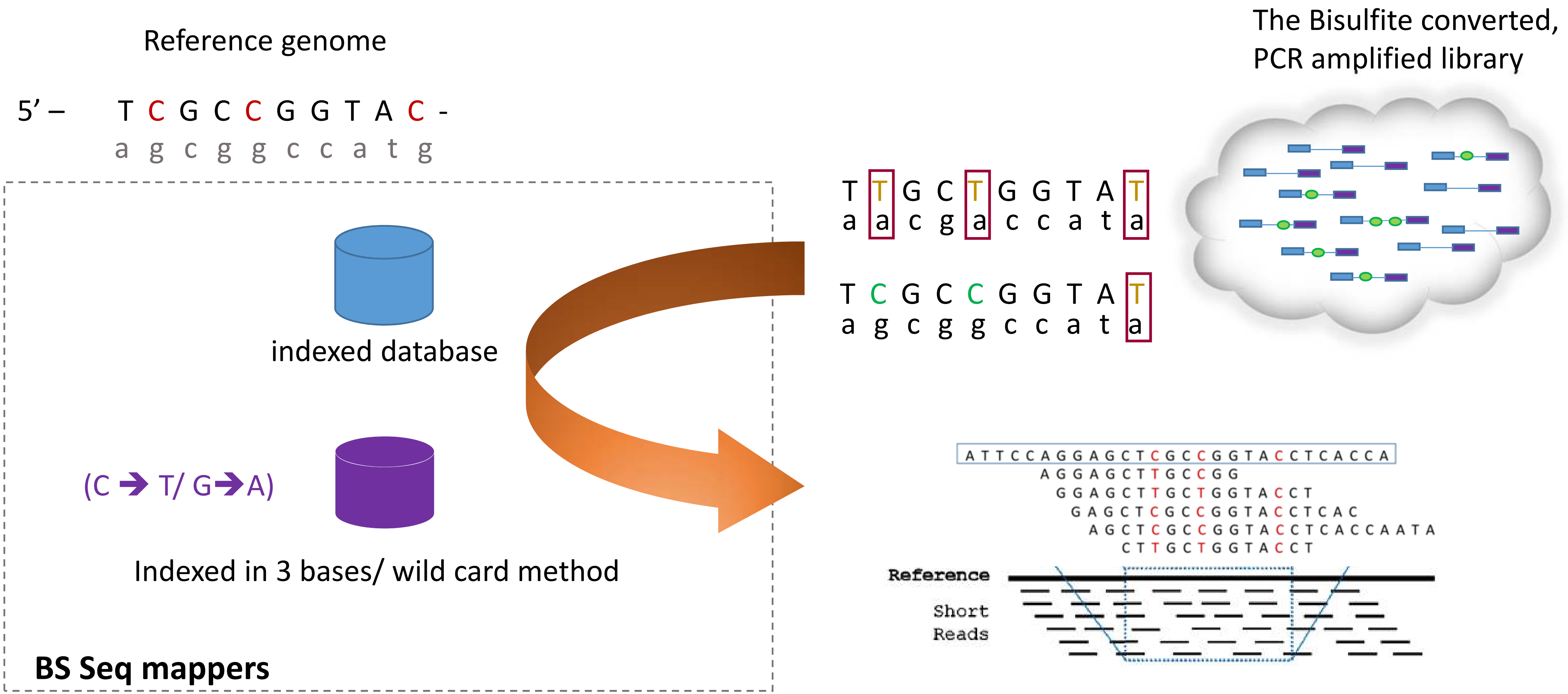
Genomic DNA

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

The Bisulfite converted, PCR amplified library

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

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

Difficulty to Access BS Seq Data/ Methylome

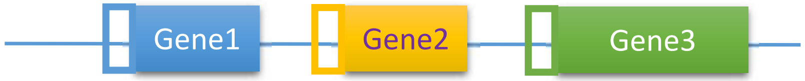
- Complicated Contents

By Context

-CG- -CHG- -CHH-

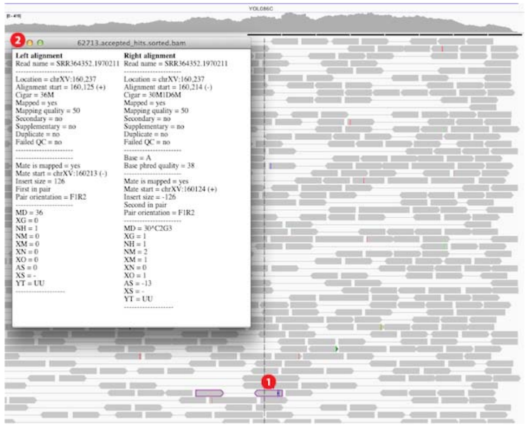
H=A, T or C

By Location



- Promoter
- Gene Body

- Visualization

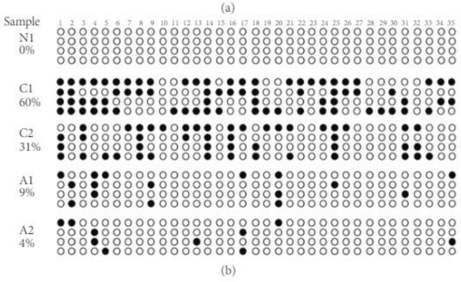


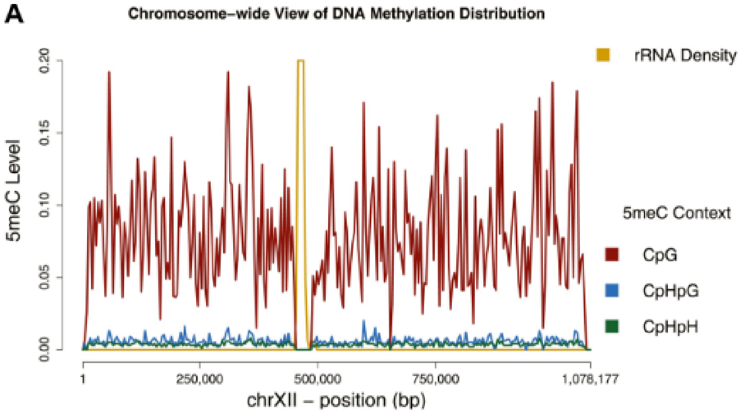
62713.accepted_hits.sorted.bam

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

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

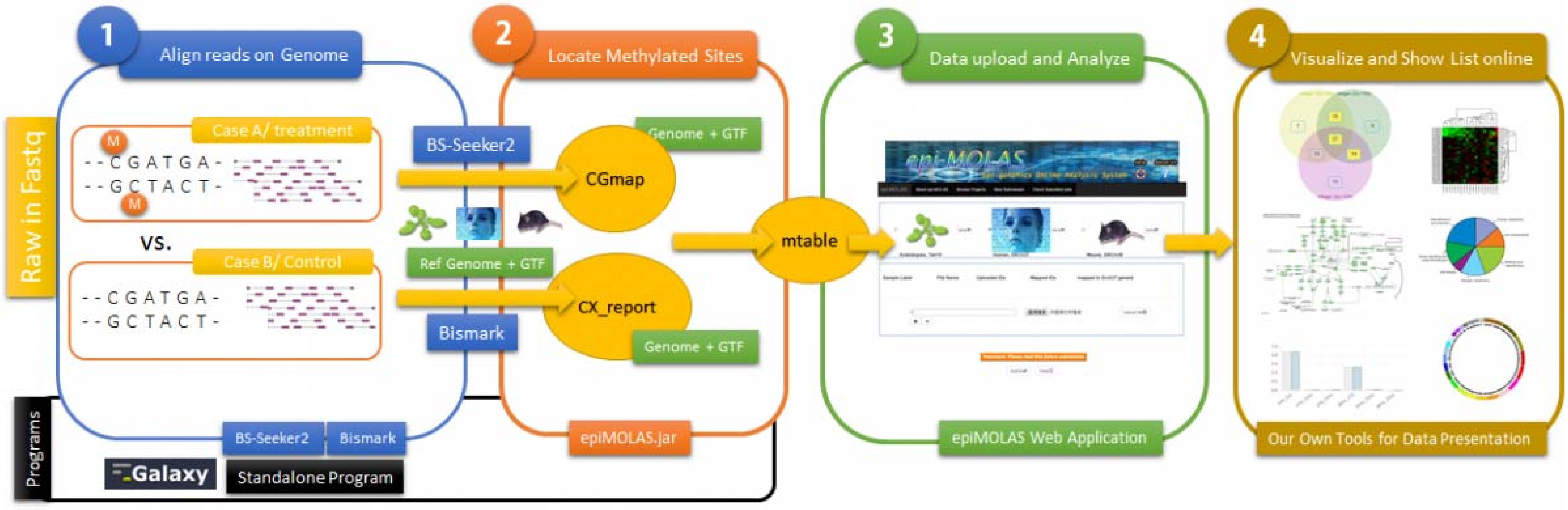
Methylated CG island





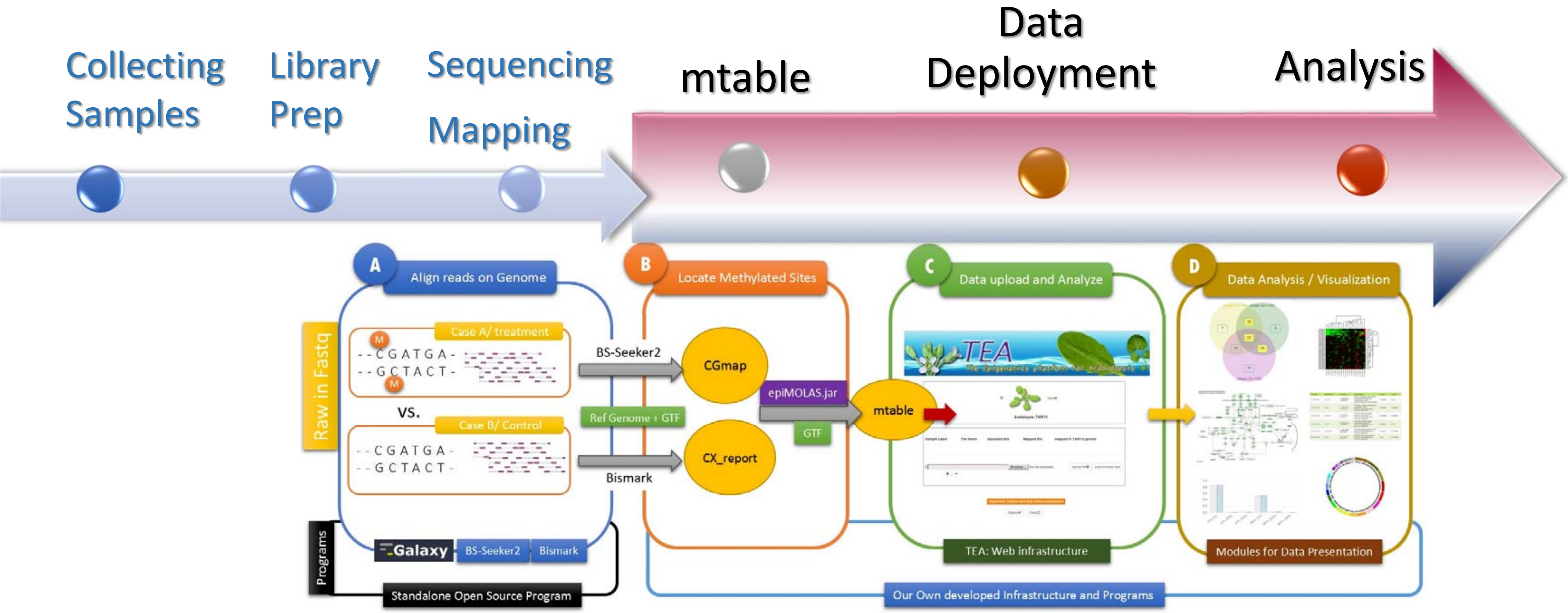
http://symbiosis.iis.sinica.edu.tw/epimolas/molas.html

The Workflow



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

TEA
The epigenomic platform for Arabidopsis


Collecting Samples

Library Prep

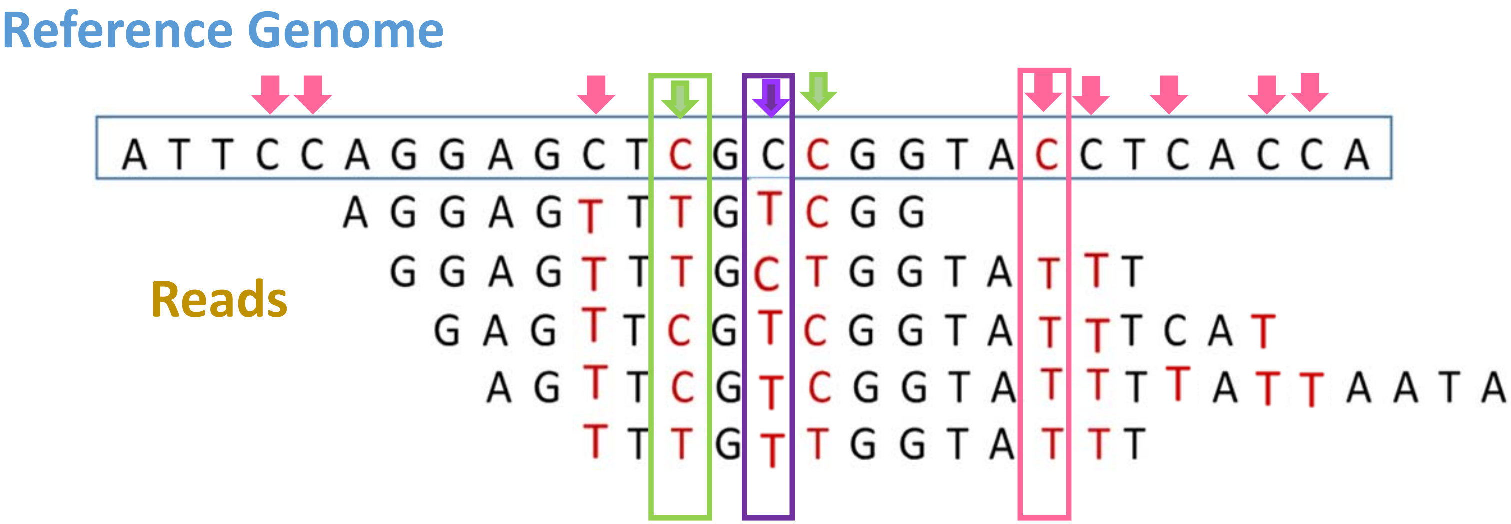
Sequencing

Mapping

mtable

Data Deployment

Analysis



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

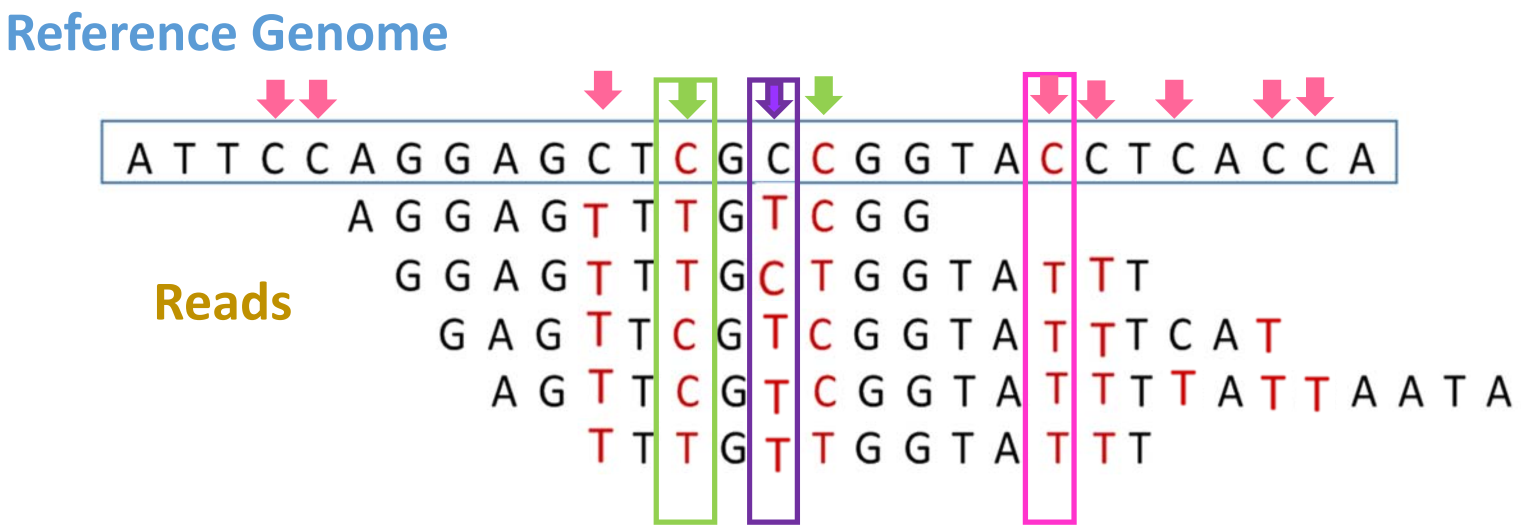
➔ score of this C: 2/5 = 0.4

- 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

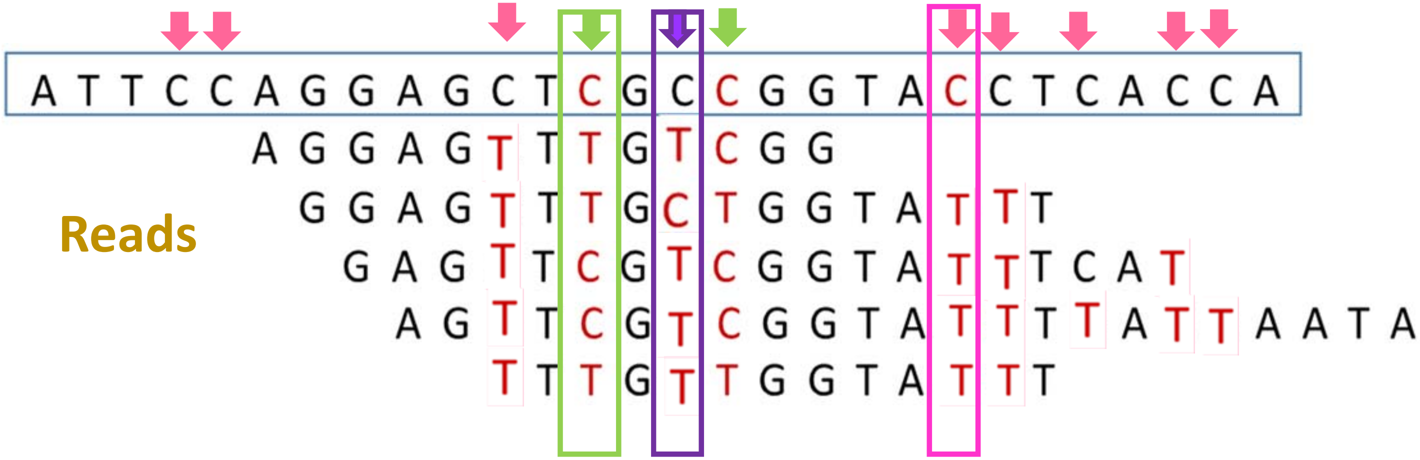
$$\text{Average DNA methylation level in promoter or gene body} = \frac{\sum_{i\in X} c_i}{\sum_{i\in X} 1} \qquad (1.2)$$

X = promoter or gene body

The Methylation Landscape

Reference Genome

Reads

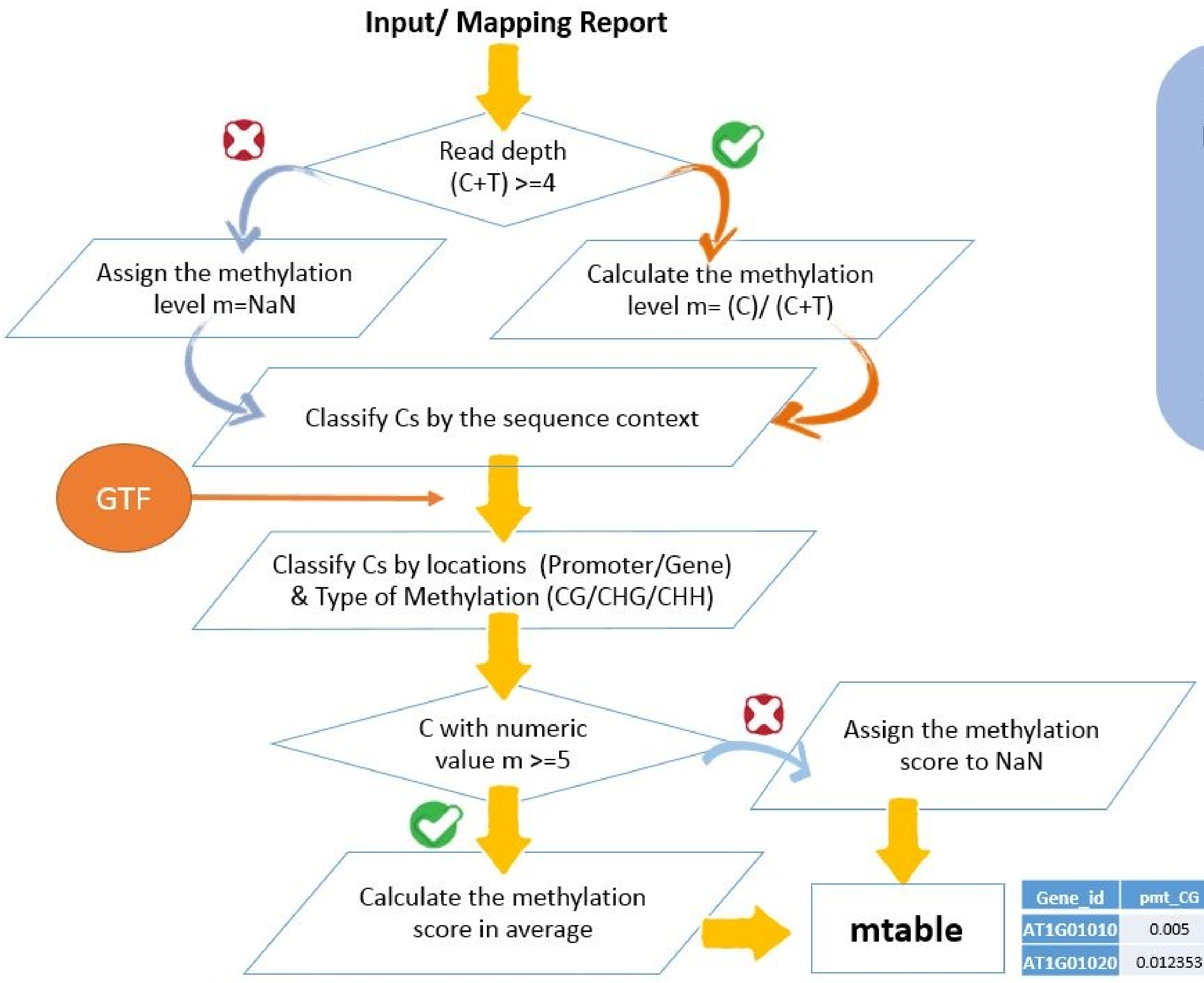


```
ATTCCAGGAGCTCGCCGGTACCTCACCA
      AGGAGTTTGTCGG
       GGAGTTTGCTGGTATTT
        GAGTTCGTCGGTATTTCAT
         AGTTCGTCGGTATTTTATTAATA
            TTTGTTGGTATTT
```

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

mtable

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	[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



Input/ Mapping Report

Read depth (C+T) >=4

Assign the methylation level m=NaN

Calculate the methylation level m= (C)/ (C+T)

Classify Cs by the sequence context

GTF

Classify Cs by locations (Promoter/Gene) & Type of Methylation (CG/CHG/CHH)

C with numeric value m >=5

Assign the methylation score to NaN

Calculate the methylation score in average

mtable

Gene_id	pmt_CG	gene_CG	pmt_CHG	gene_CHG	pmt_CHH	gene_CHH
AT1G01010	0.005	0.068448	0.00375	0.028333	0.004739	0.024981
AT1G01020	0.012353	0.092468	0.013182	0.015667	0.013614	0.019084

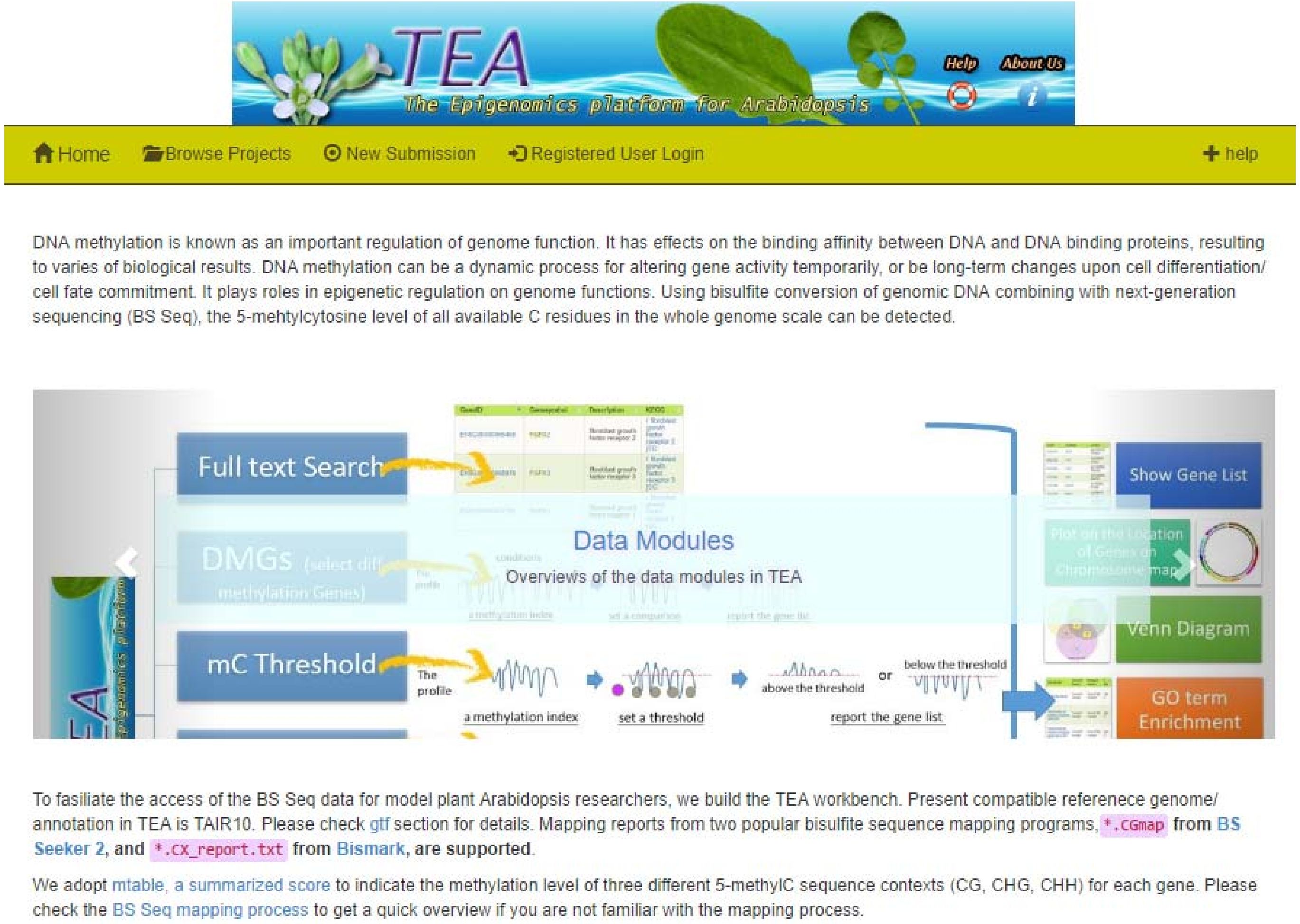
Inputs

- **BS-Seq mapping report**
  - **CGmap** from BS- Seeker2 /
  - **CX_report.txt** from Bismark
  - Or an equivalent from other BS-Seq mappers
- **GTF of the reference genome**

TEA Website

Demo site: http://tea.iis.sinica.edu.tw/

Demo site: http://symbiosis.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.

Data Modules

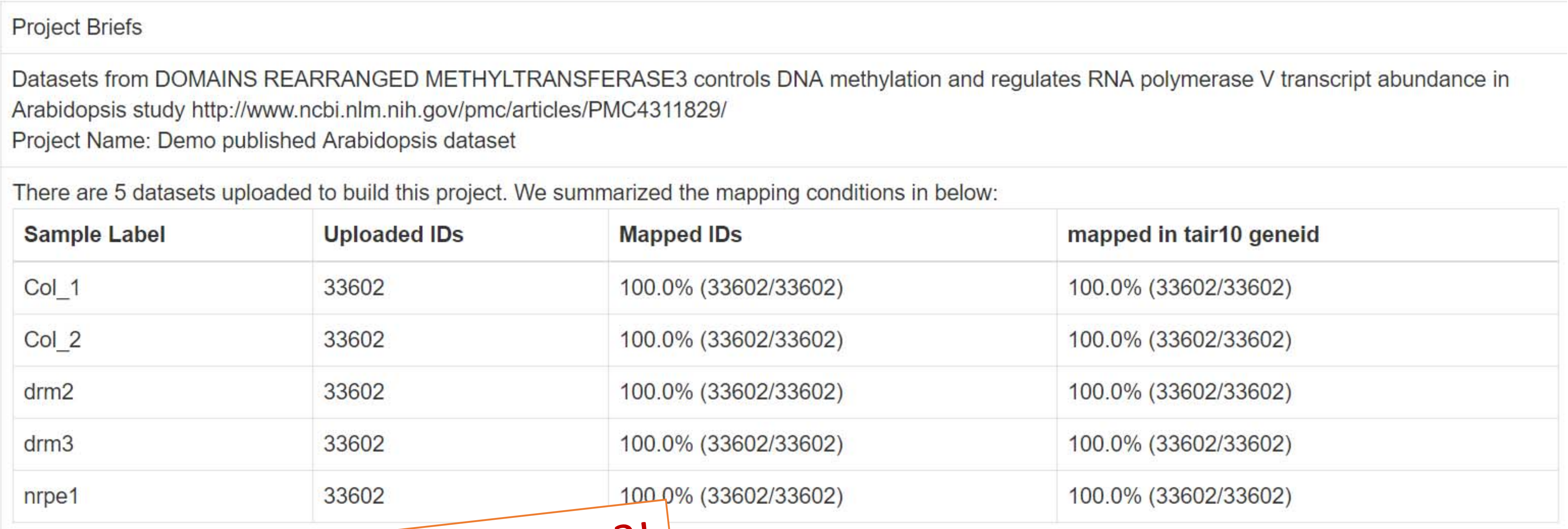
Overviews of the data modules in TEA

To fasiliate the access of the BS Seq data for model plant Arabidopsis researchers, we build the TEA workbench. Present compatible referenece 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:

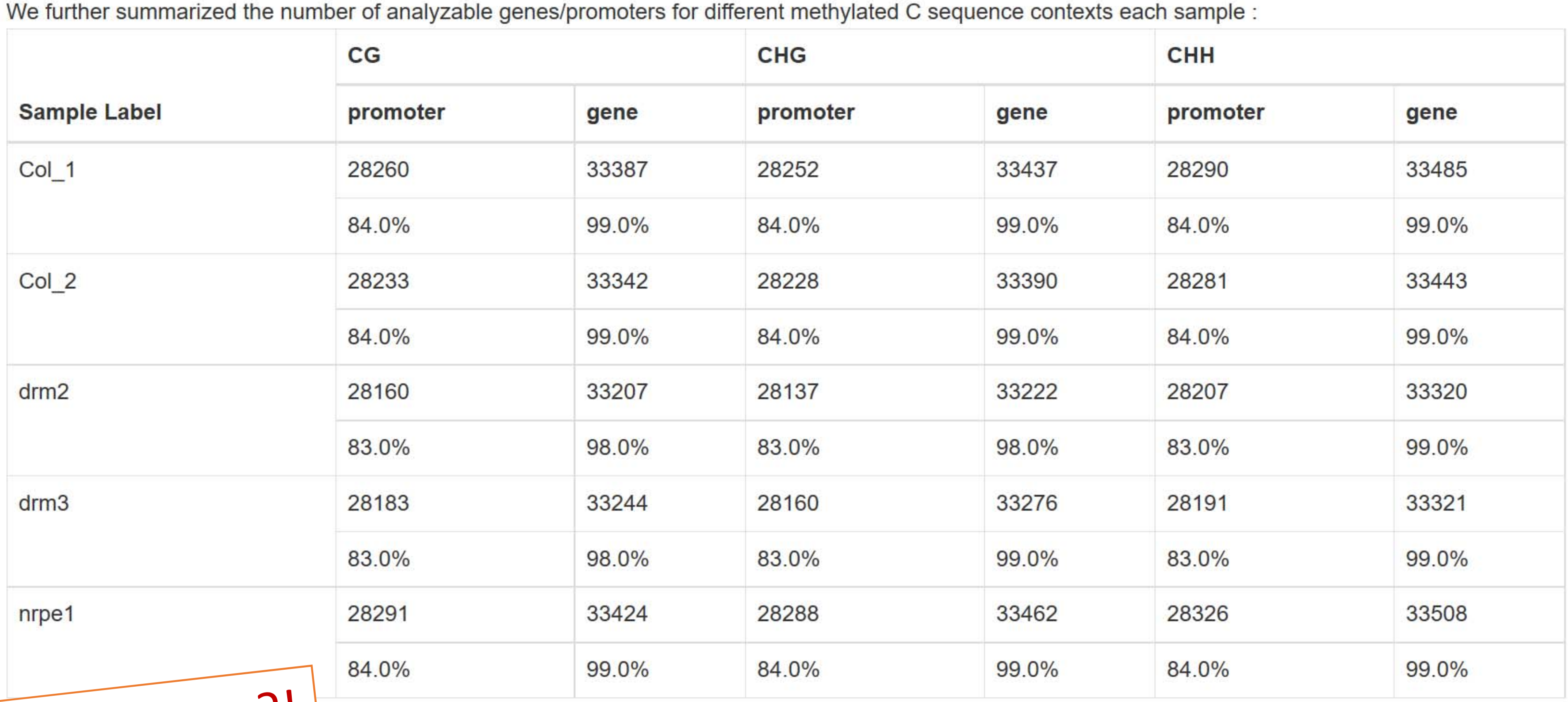
Sample Label	Uploaded IDs	Mapped IDs	mapped in tair10 geneid
Col_1	33602	100.0% (33602/33602)	100.0% (33602/33602)
Col_2	33602	100.0% (33602/33602)	100.0% (33602/33602)
drm2	33602	100.0% (33602/33602)	100.0% (33602/33602)
drm3	33602	100.0% (33602/33602)	100.0% (33602/33602)
nrpe1	33602	100.0% (33602/33602)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

Poor ID mapping rate?!

Check the gtf version

Project Summary

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

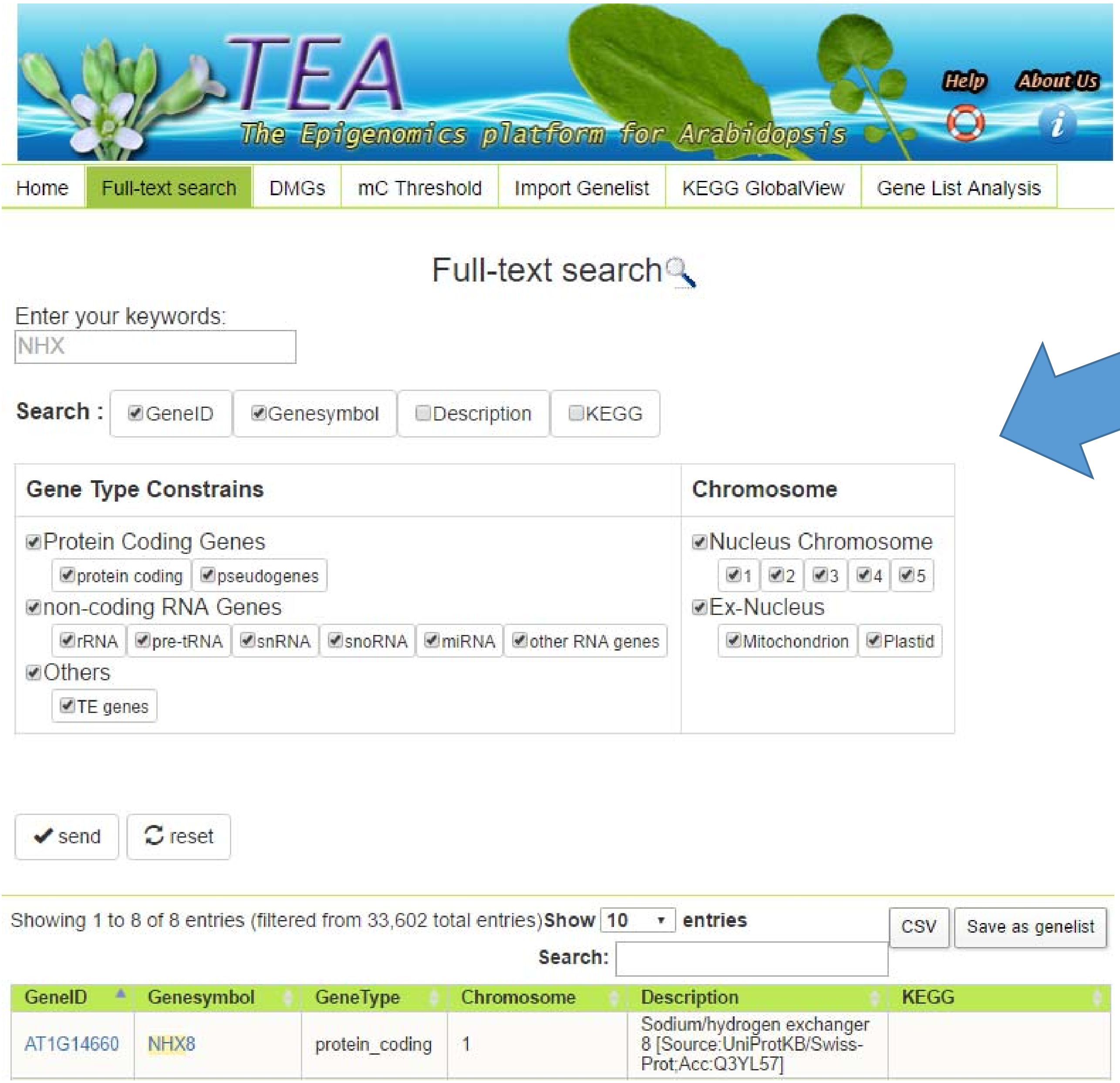


We further summarized the number of analyzable genes/promoters for different methylated C sequence contexts each sample :

	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:

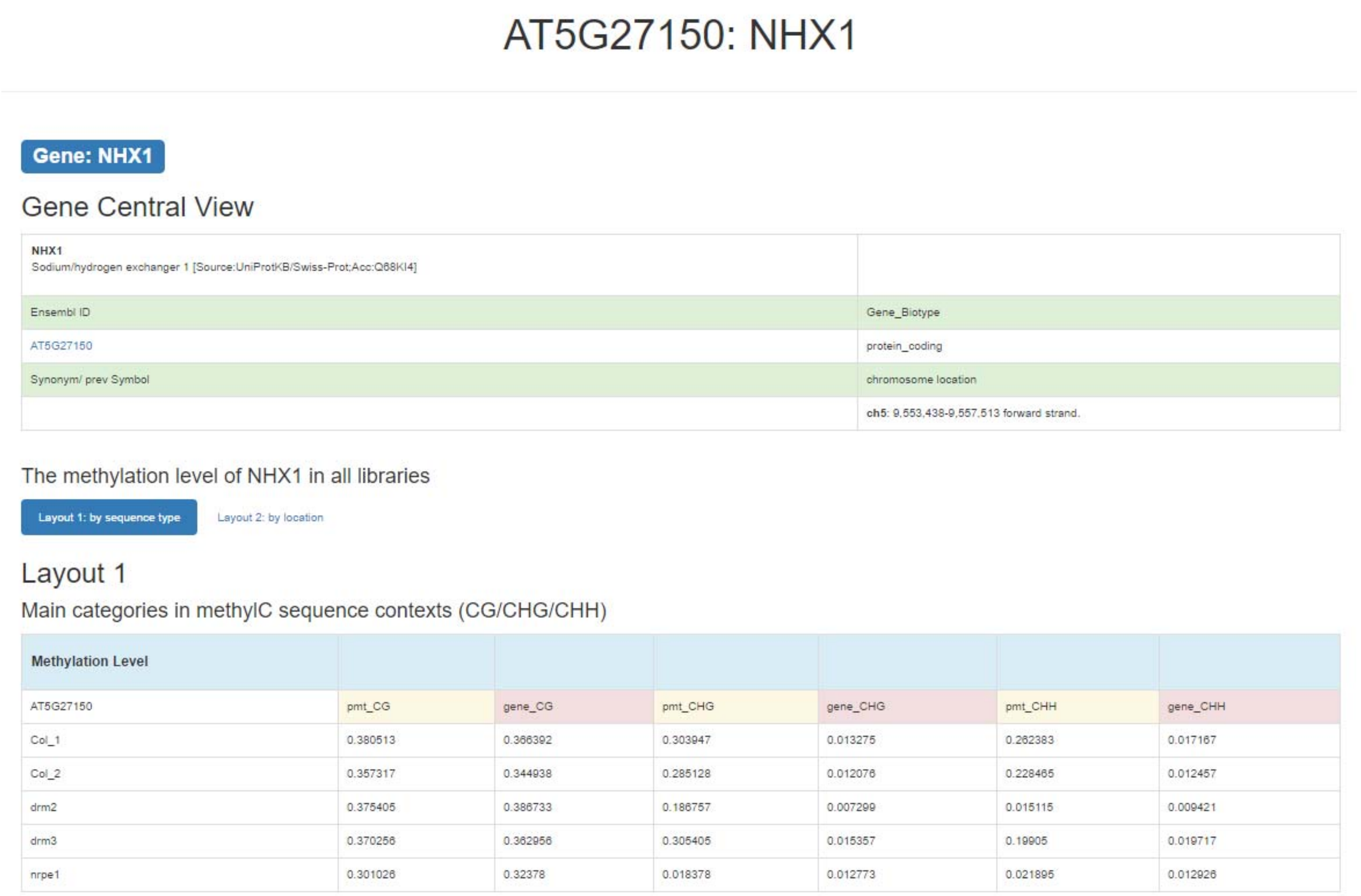
GeneID	Genesymbol	GeneType	Chromosome	Description	KEGG
AT1G14660	NHX8	protein_coding	1	Sodium/hydrogen exchanger 8 [Source:UniProtKB/Swiss-Prot;Acc:Q3YL57]	

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

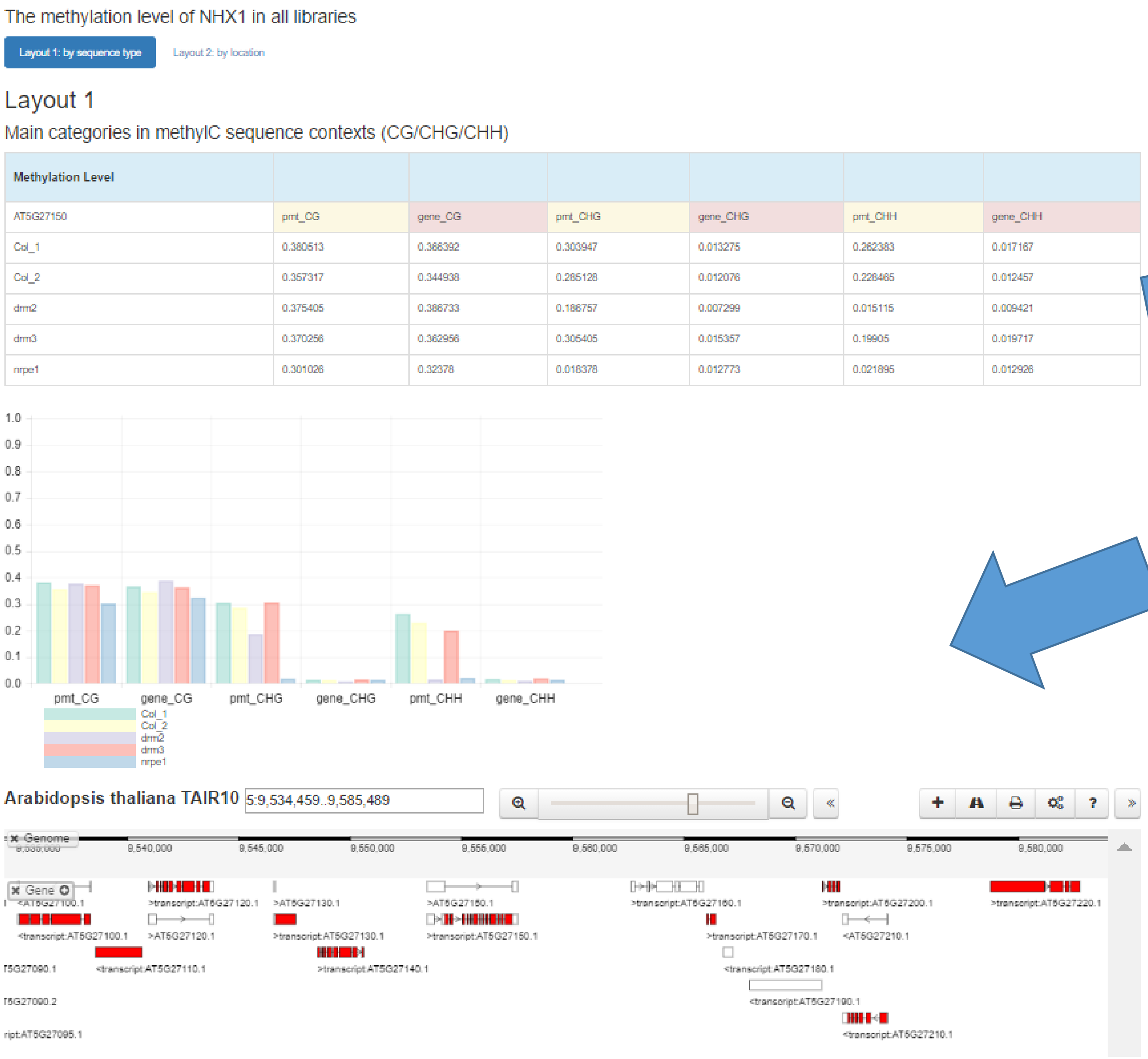
NHX1 Sodium/hydrogen exchanger 1 [Source:UniProtKB/Swiss-Prot;Acc:Q68KI4]	
Ensembl ID	Gene_Biotype
AT5G27150	protein_coding
Synonym/ prev Symbol	chromosome location
	ch5: 9,553,438-9,557,513 forward strand.

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

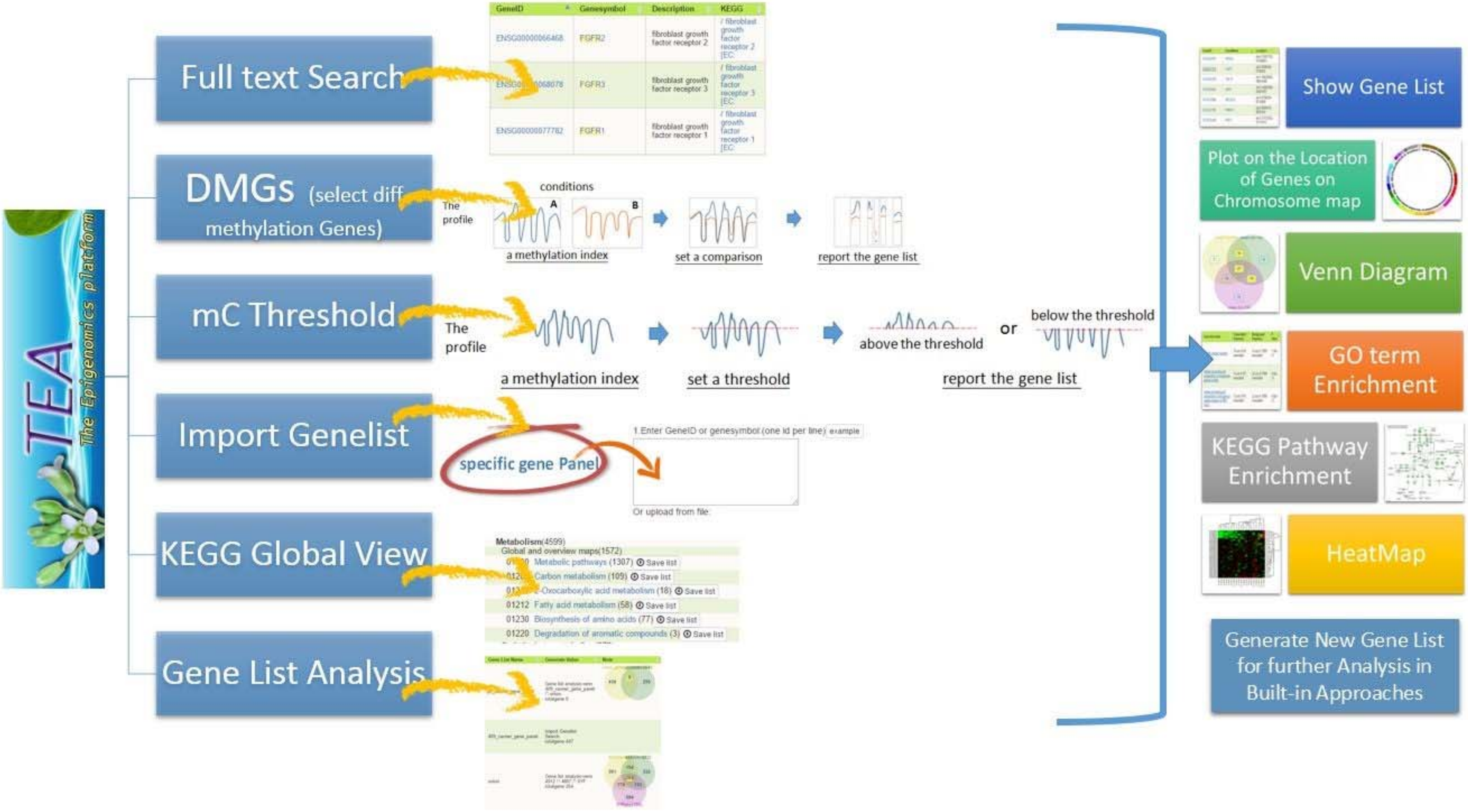
Arabidopsis thaliana TAIR10 5:9,534,459..9,585,489

Measures of Methylation

Genome Browser

Data Analysis Modules

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



TEA The Epigenomics platform

Full text Search

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:

DMGs (select diff methylation Genes)

The profile

conditions

a methylation index → set a comparison → report the gene list

mC Threshold

The profile

a methylation index → set a threshold → above the threshold or below the threshold

report the gene list

Import Genelist

specific gene Panel

1.Enter GeneID or genesymbol (one id per line) example

Or upload from file

KEGG Global View

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

Gene List Analysis

Show Gene List

Plot on the Location of Genes on Chromosome map

Venn Diagram

GO term Enrichment

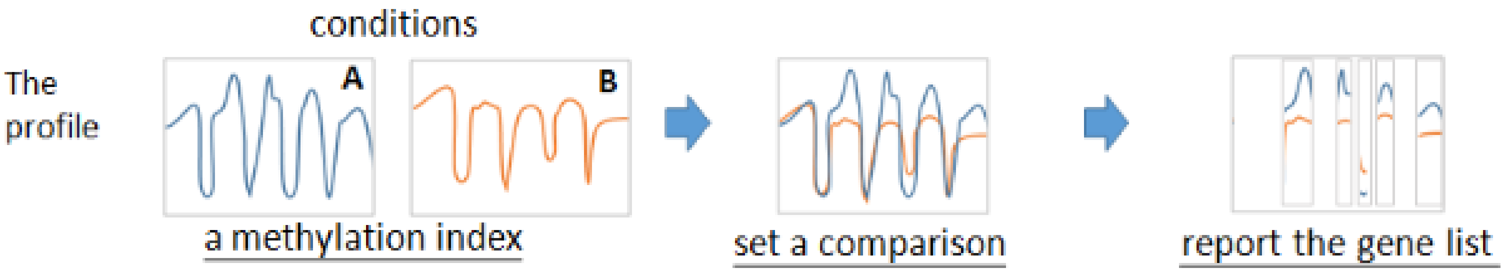
KEGG Pathway Enrichment

HeatMap

Generate New Gene List for further Analysis in Built-in Approaches

Find Genes by Value

DMGs : Select differentially methylated genes by the interested methylation score



conditions

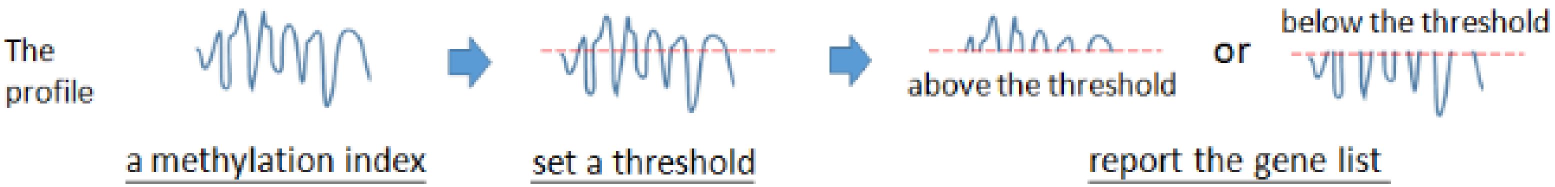
The profile

a methylation index

set a comparison

report the gene list

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



The profile

a methylation index

set a threshold

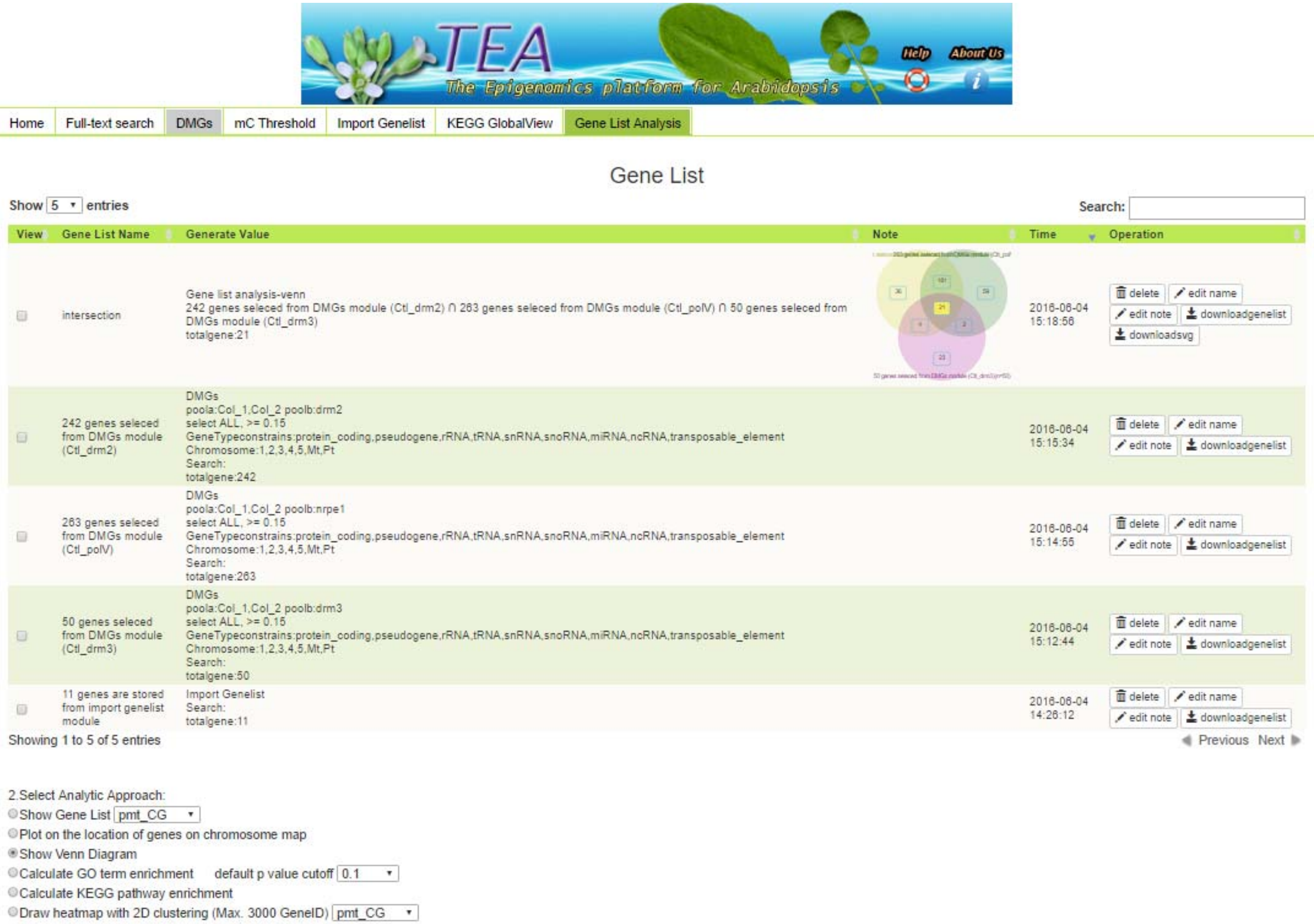
above the threshold

or

below the threshold

report the gene list

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		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

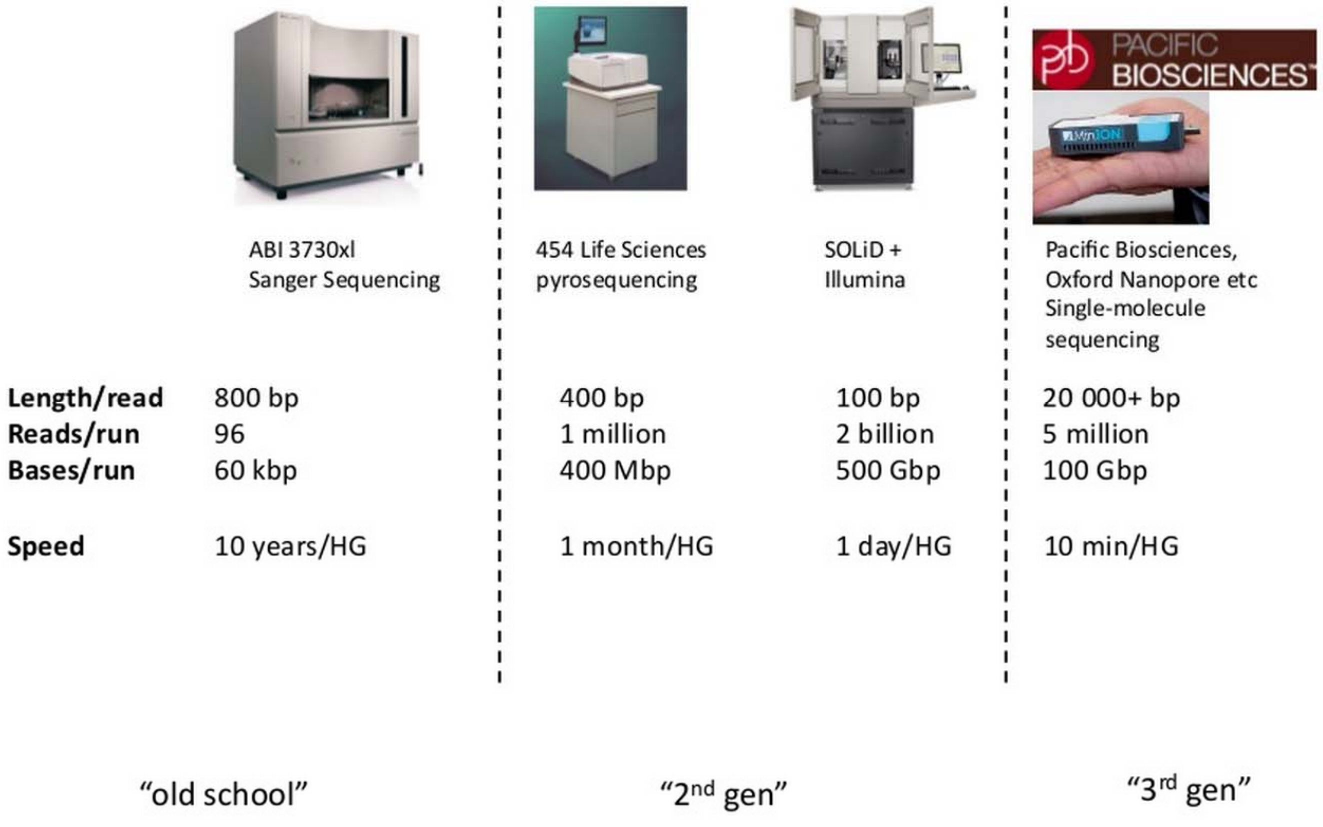
○ Draw heatmap with 2D clustering (Max. 3000 GeneID) pmt_CG ▾

Questions?

Future Works

- A more sophisticate measures that highlight the pattern of methylation
- Multi-Omic Integration

Sequencing Platforms


PACIFIC BIOSCIENCES
MinION

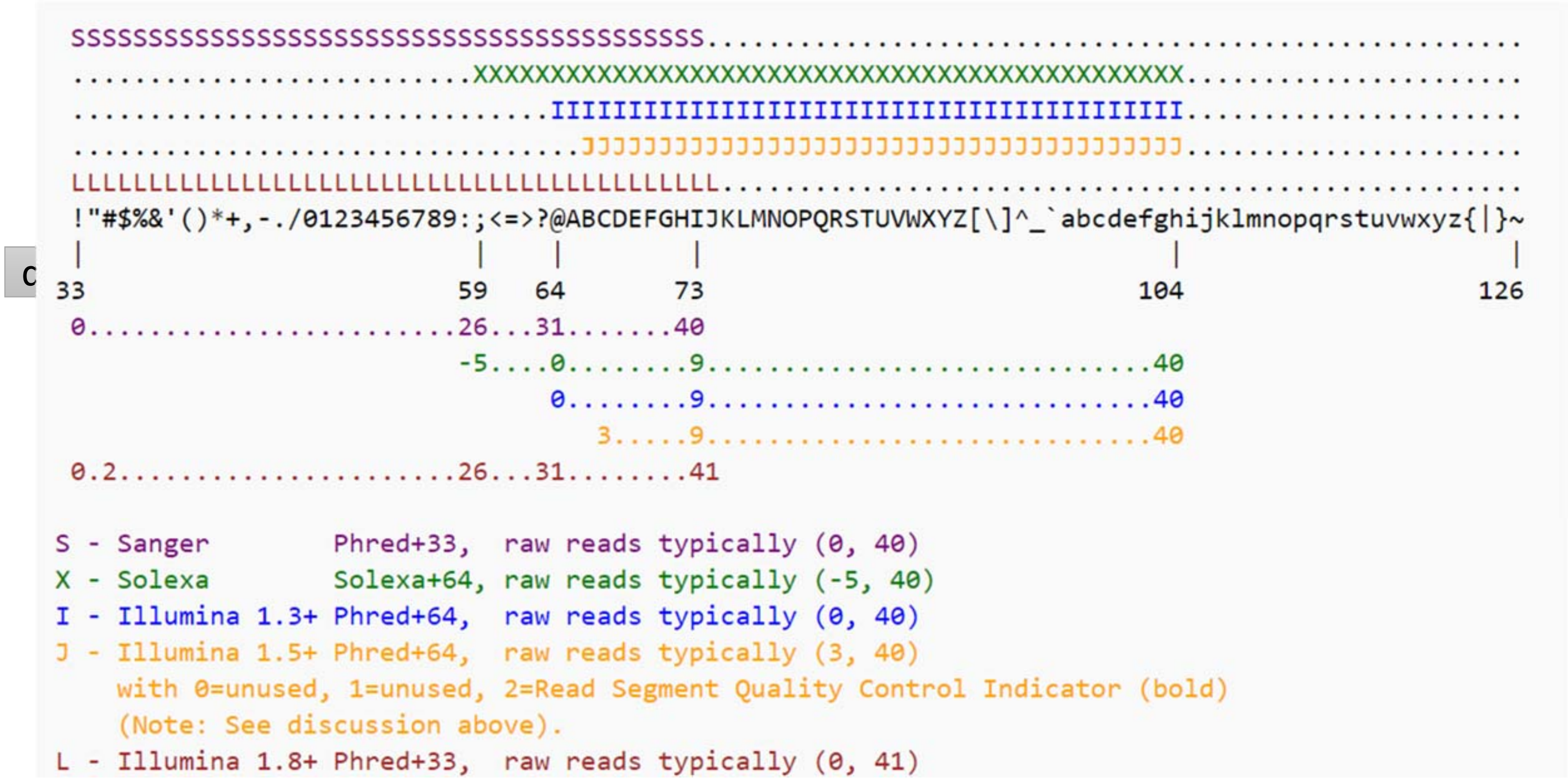
	ABI 3730xl Sanger Sequencing	454 Life Sciences pyrosequencing	SOLiD + Illumina	Pacific Biosciences, Oxford Nanopore etc Single-molecule sequencing
Length/read	800 bp	400 bp	100 bp	20 000+ bp
Reads/run	96	1 million	2 billion	5 million
Bases/run	60 kbp	400 Mbp	500 Gbp	100 Gbp
Speed	10 years/HG	1 month/HG	1 day/HG	10 min/HG
	“old school”	“2nd gen”		“3rd gen”

http://www.slideshare.net/COST-events/rnaseq-analysis-17037153?next_slideshow=1

FastQ format

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

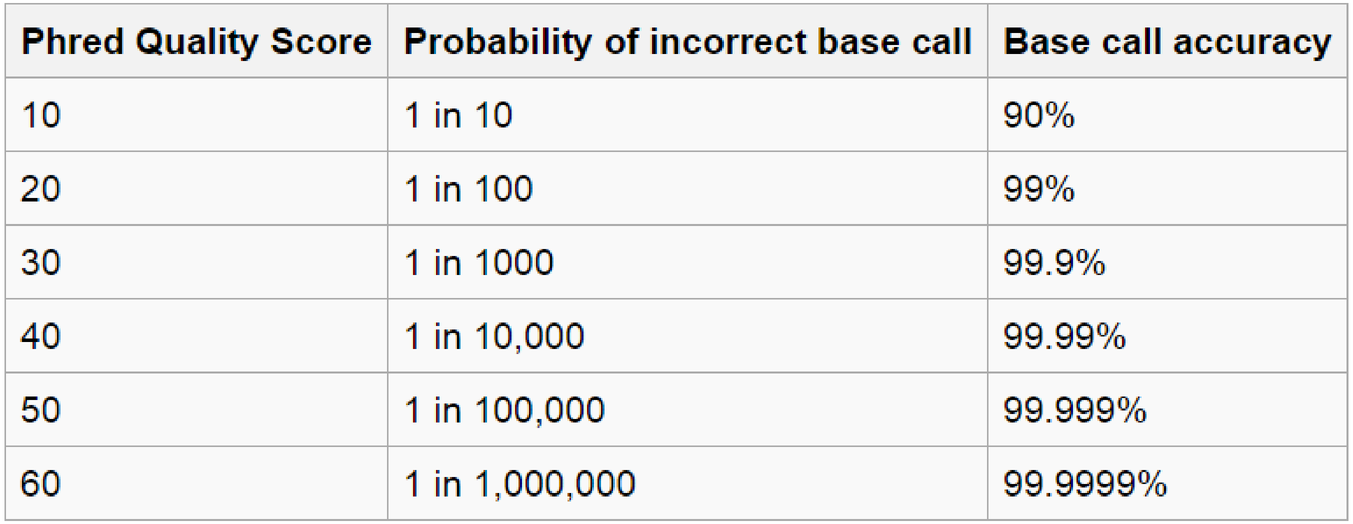
- Start with “@”
- Four lines:,, “+” w/ or w/o seq head, quality scores



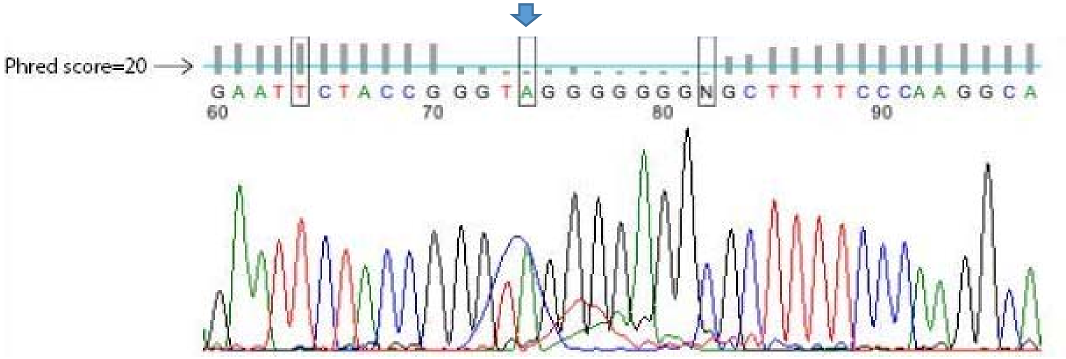
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Phred quality scores are logarithmically linked to error probabilities



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%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



Phred=10

http://en.wikipedia.org/wiki/Phred_quality_score