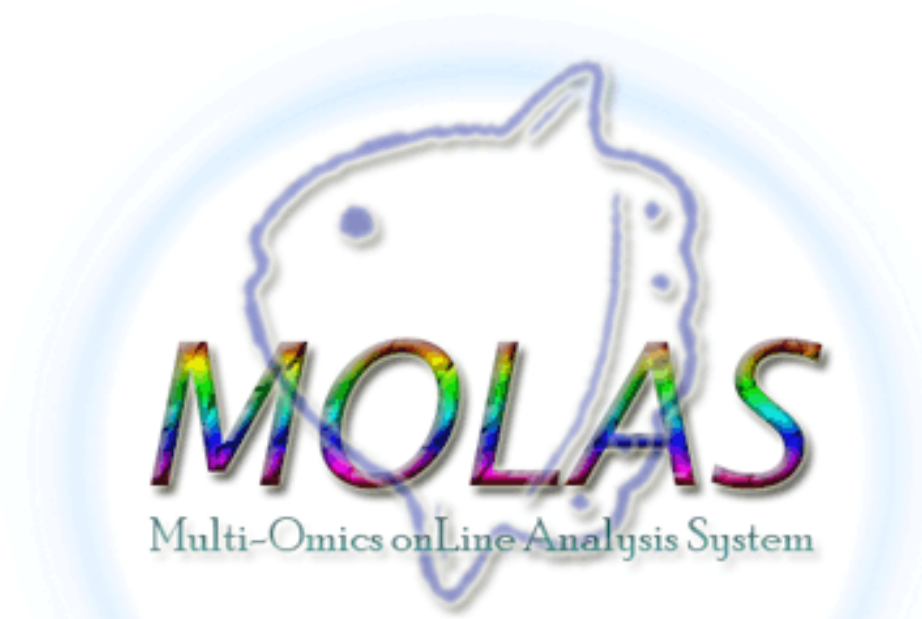


AS Life Science Library Training Course

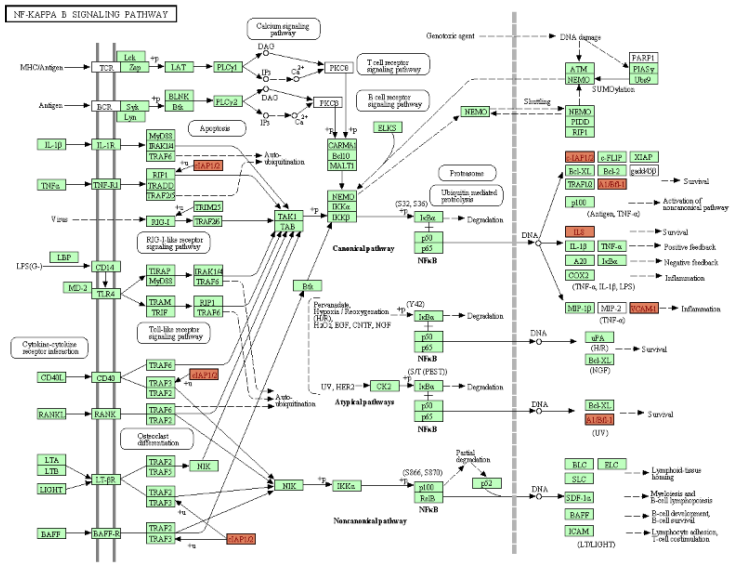
2015/06/09

Chen, Shu-Hwa Ph.D.

IIS, Academia Sinica

Outline

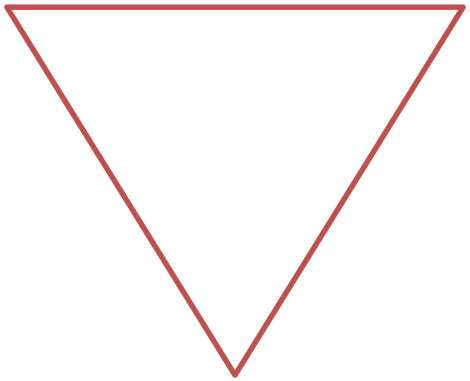
- Something about RNASeq
- Introduce MOLAS system
  - How to submit your data
  - How to view and analyze your data



High-throughput Methods

Biology Lab

Bioinfo Lab



Tech Core

Microarray Technology

NGS TECHNOLOGY

Roche GS-FLX

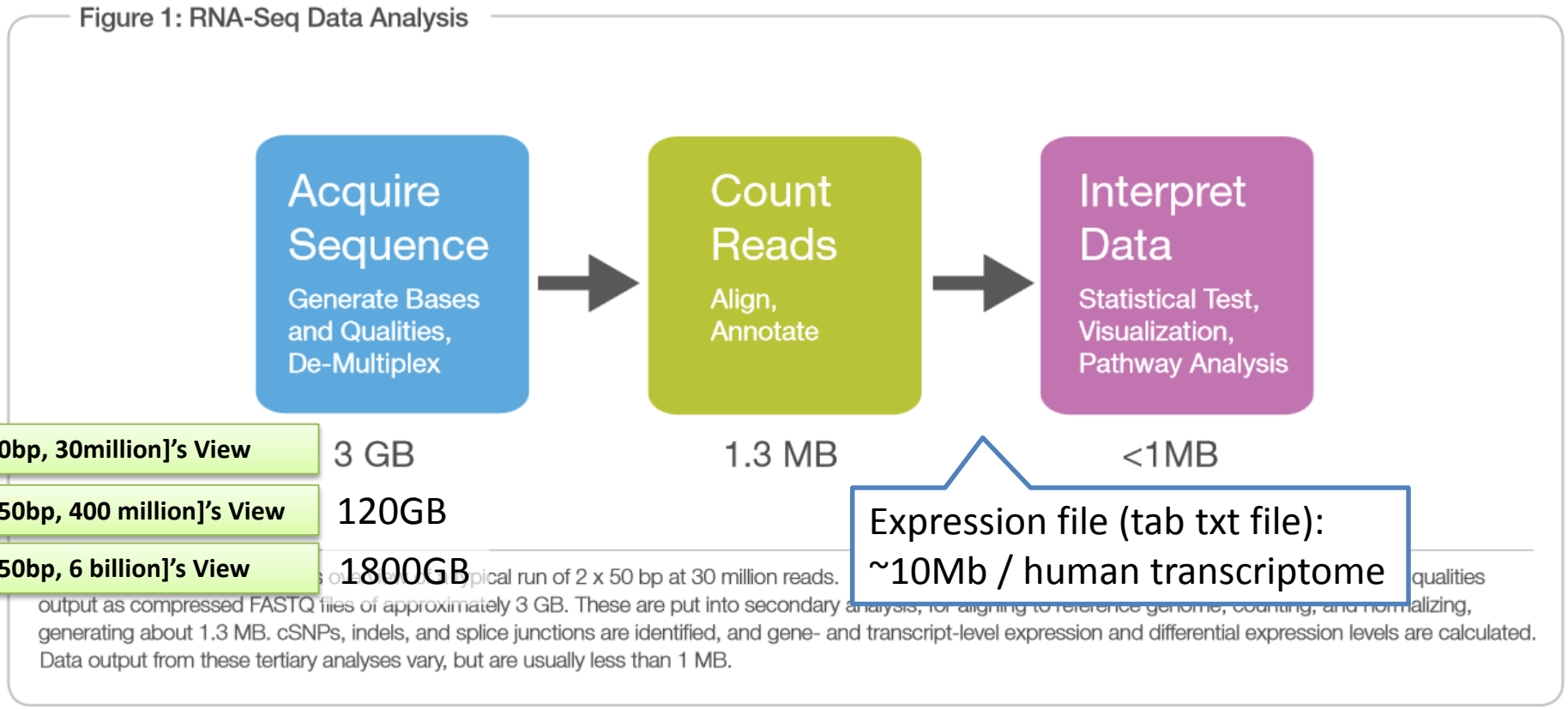
Life Technologies SOLiD

Illumina HiSeq

Life Technologies Ion Torrent

PacBio

RNASeq Data Analysis



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

Sequencing Platforms

ABI 3730xl
Sanger Sequencing

454 Life Sciences
pyrosequencing

SOLiD +
Illumina

PACIFIC BIOSCIENCES

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

- 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	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+	+
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>

http://en.wikipedia.org/wiki/FASTQ_format
http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/

FastQ format

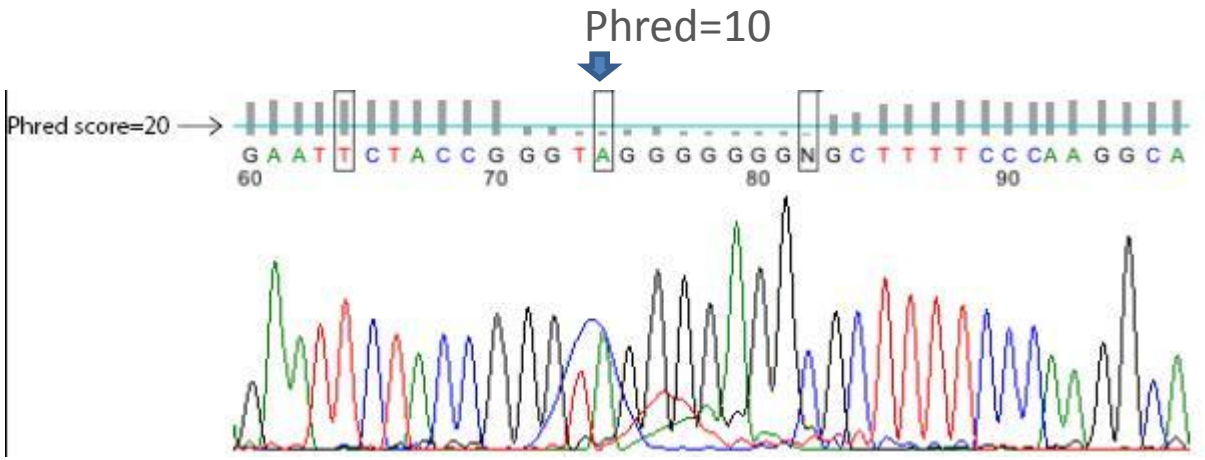
- Start with “@”
- Four lines:,, “+” w/ or w/o seq head, quality scores

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..............................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..............................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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%



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

A typical RNA-Seq experiment

Prepare RNA samples

QC

Prepare sequencing libraries

QC

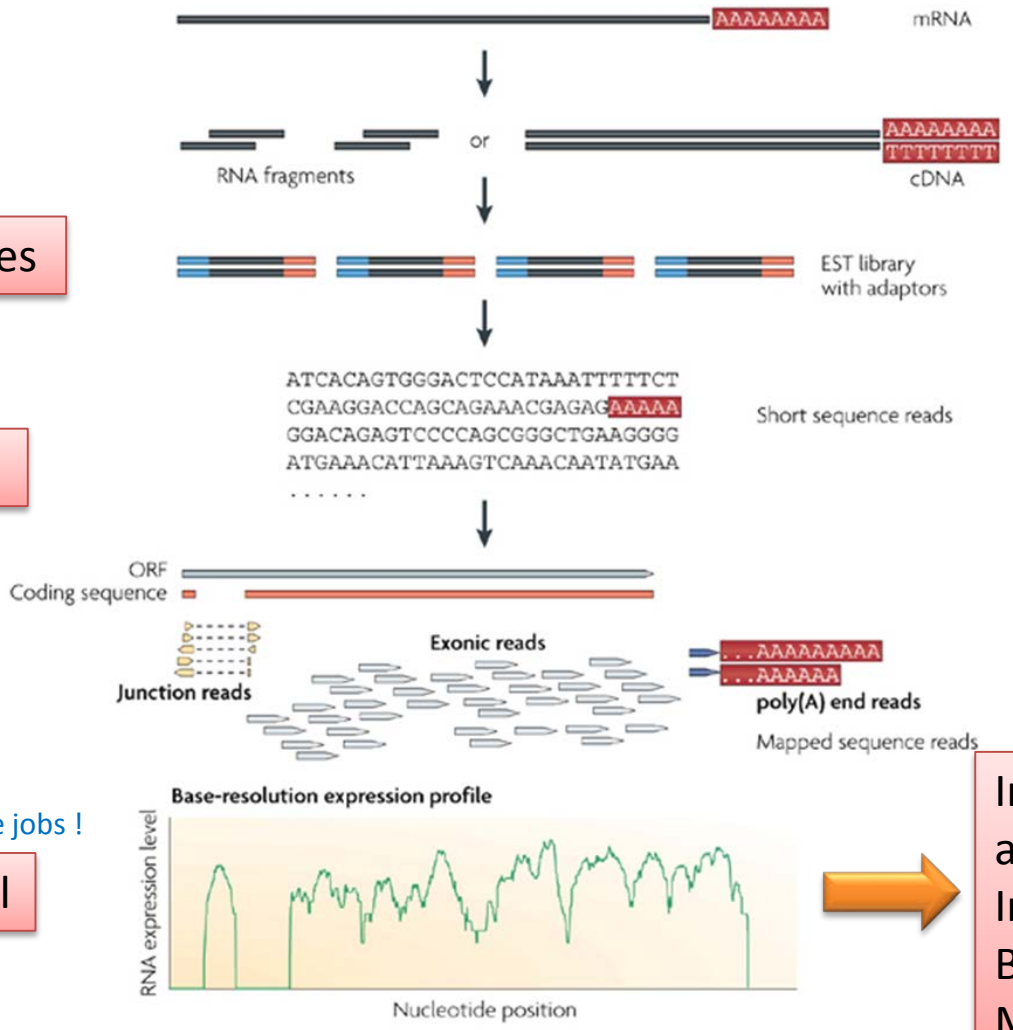
Millions of shot-gun reads

QC
Data preprocessing

Map to reference

Read mapping
…computing intensive jobs !

Convert to expression level



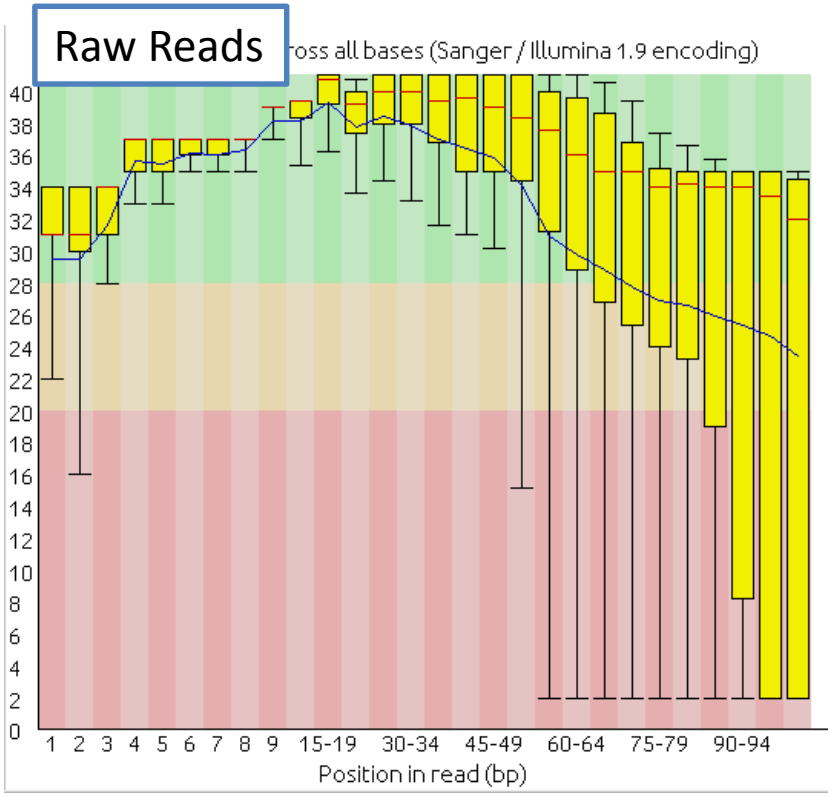
Intensive analysis to Interpretate Biological Meanings

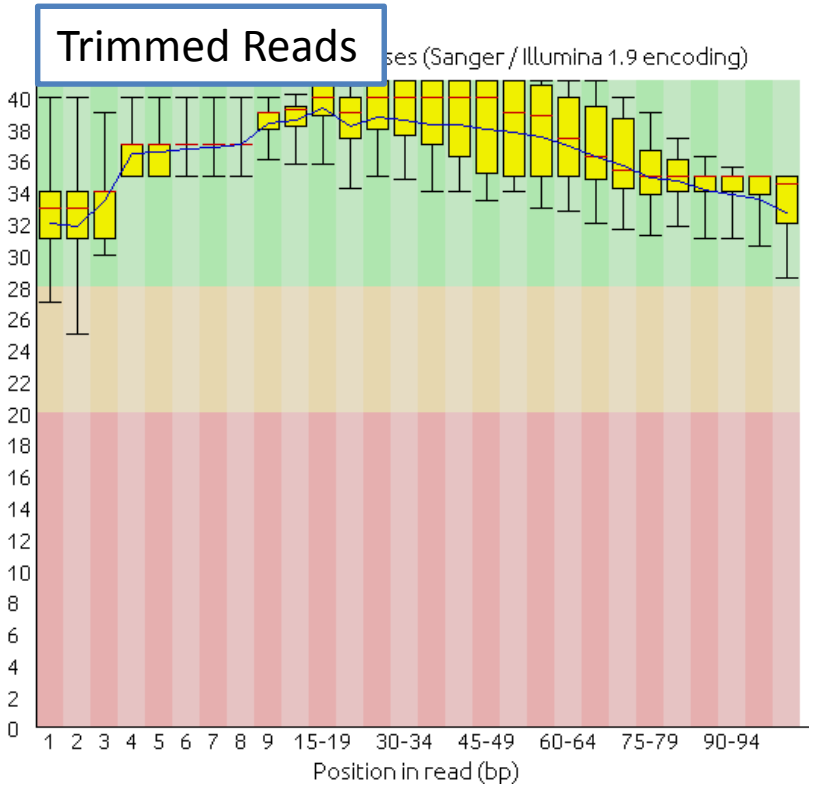
http://www.nature.com/nrg/journal/v10/n1/full/nrg2484.html

Nature Reviews | Genetics

Read preprocessing

- Trimming: by base quality score

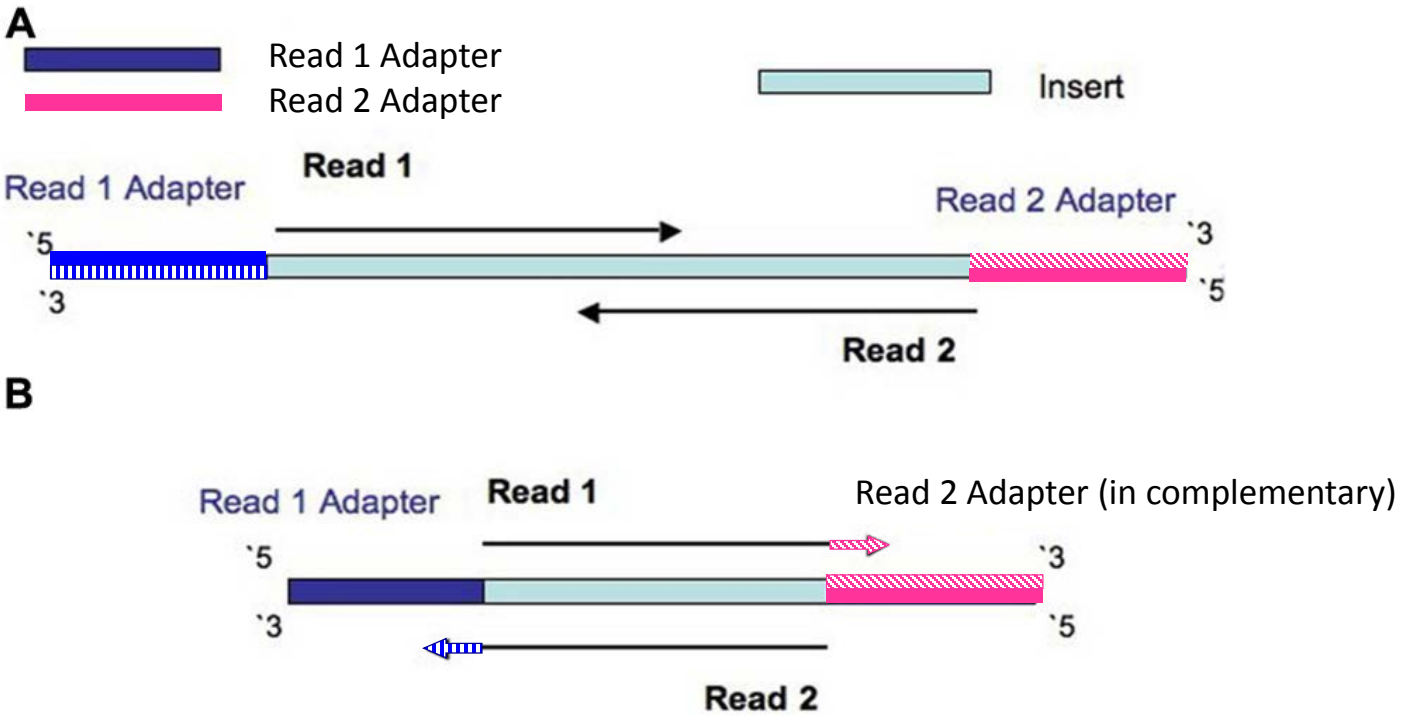




http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

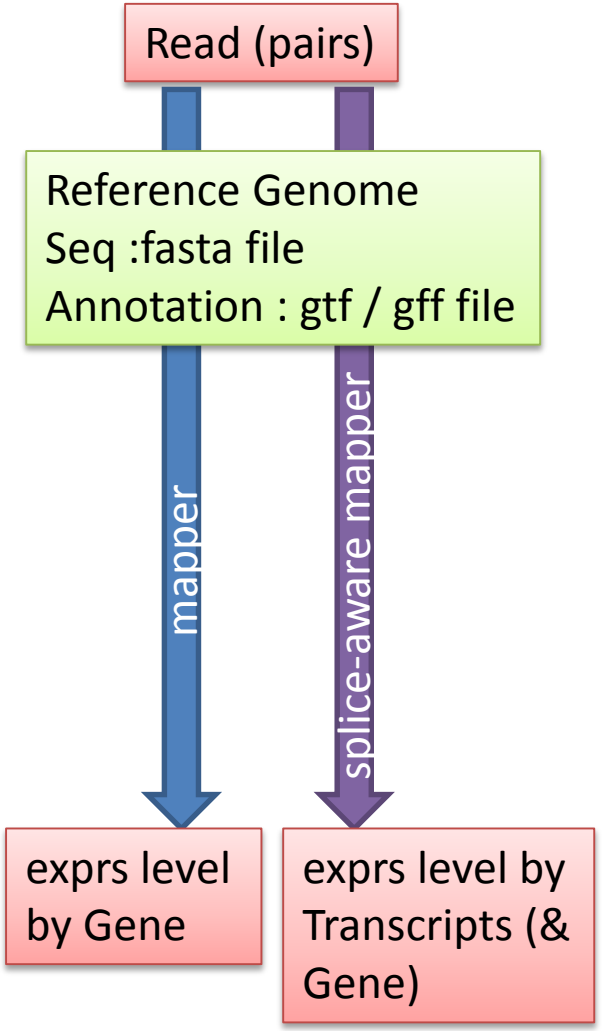
Read preprocessing

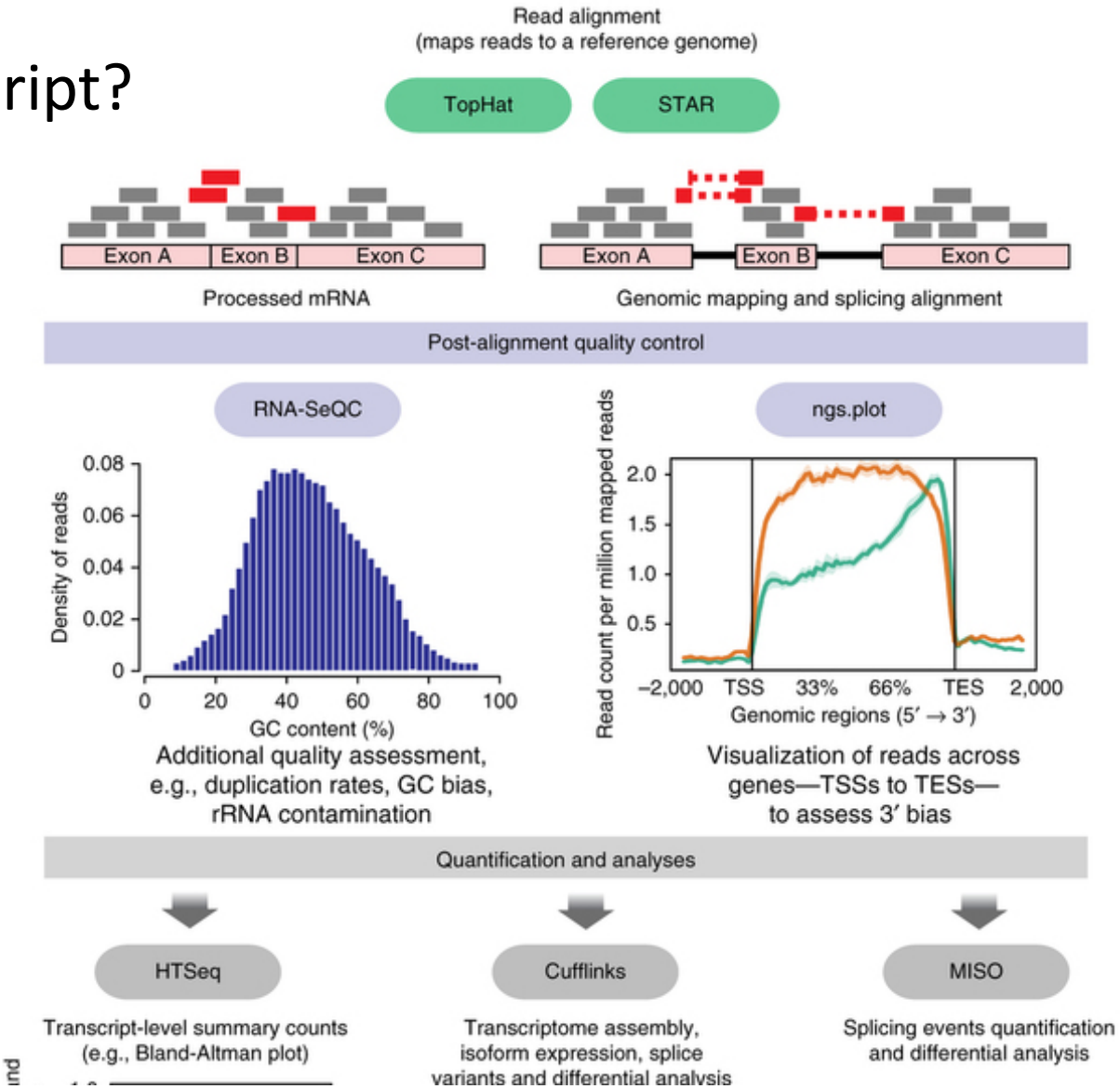
- Trimming: adapter contamination



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

Expression Level by Gene or by Transcript?





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

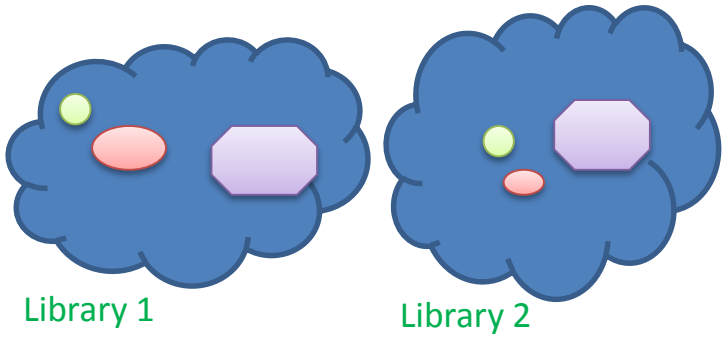
Other issues

- Stranded or not?
- PolyA tailed or rRNA depletion?
- Special protocols that need extra bioinformatical works?
- Trimmed read length? Low complexity repeats? Other sources of contamination?
- Have reference genome?
- Novel transcripts?

The Normalized Expression Level

- Between sample:

Library Size Effect



- 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

The art of Normalization

- Borrowing information across genes to get a better estimate.

- 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 retain statistical power

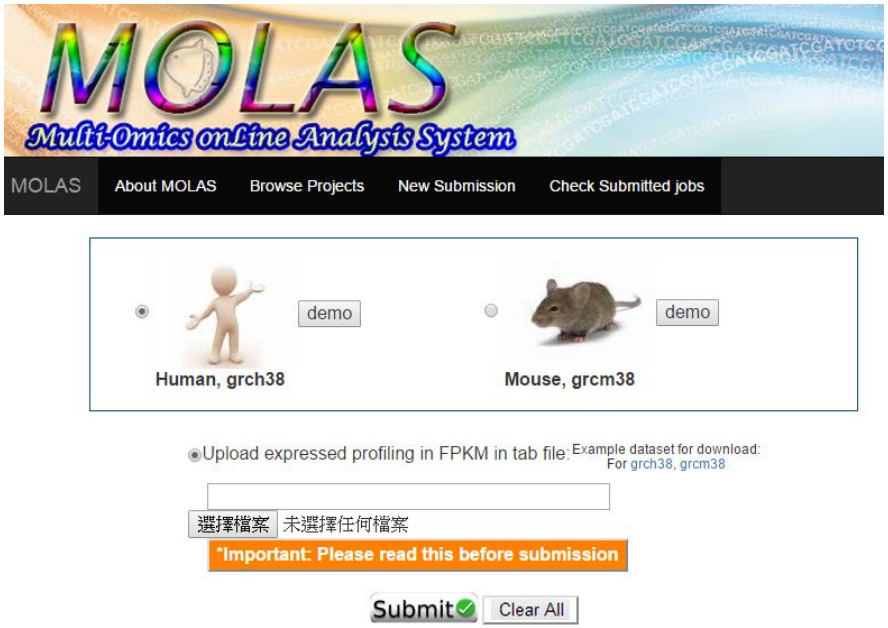
- Programs like SAMSeq (rank-based model, only applicable for large replicates) and limma are fine with continuous values (like FPKM). Limma takes more cares about weak mean-variance relationship (stabilizing variation).

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

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)

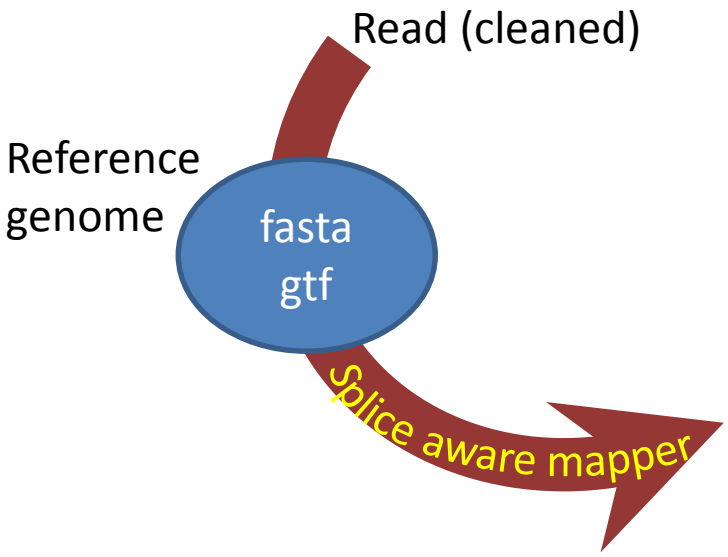


To view and analyse your RNASeq experiment

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

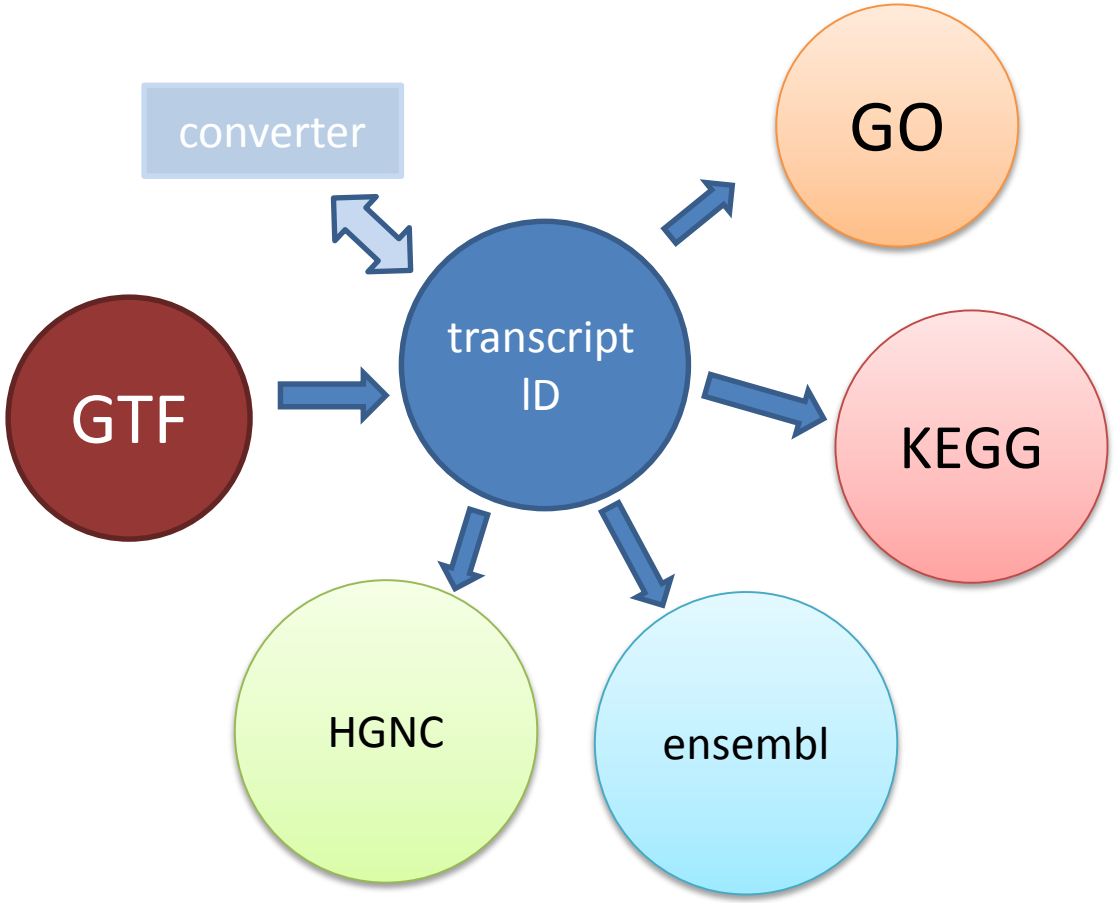
GTF: the Gene Tranfer Format

1 ensembl_havana transcript 4344146 4360314 . - . gene_id "ENSMUSG00000025900"; 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 "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS14804";

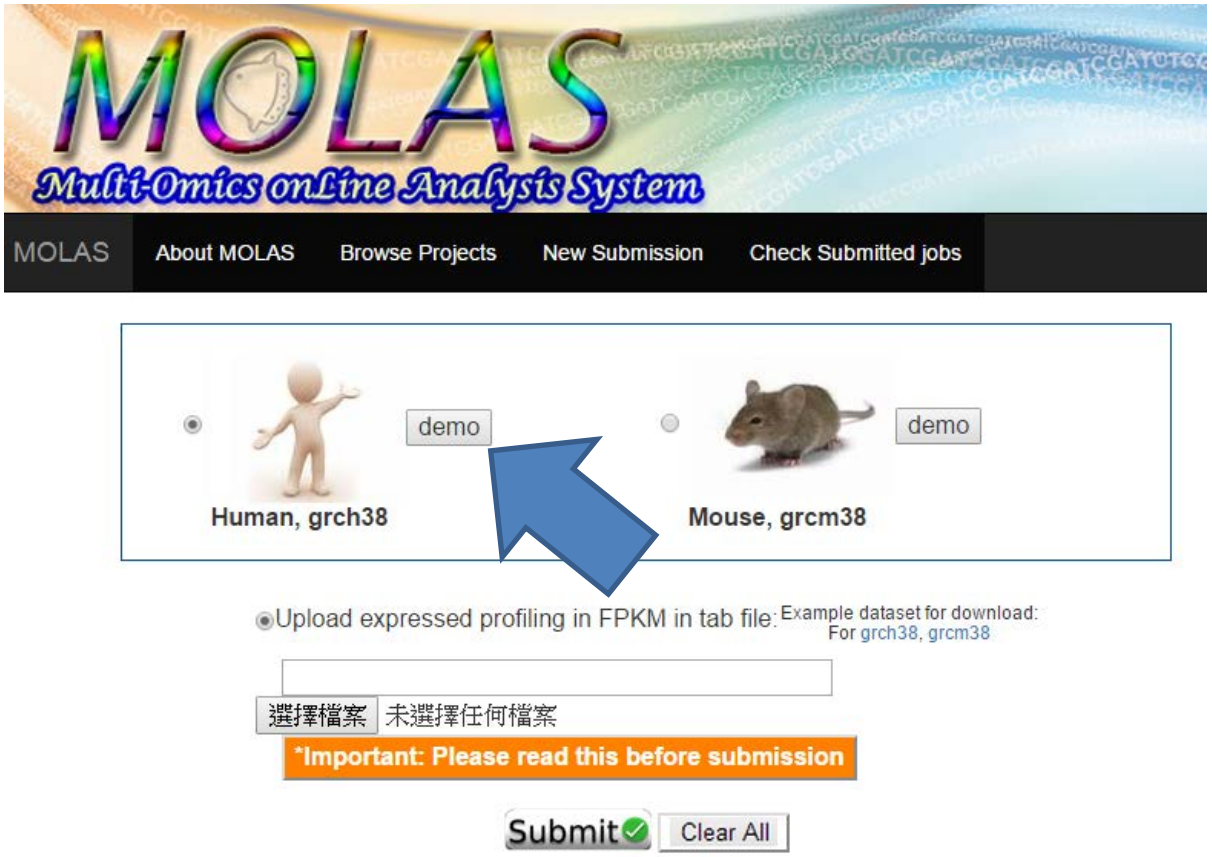
MOLAS compatible GTF

grch38
grch37

grcm38



New Submssion



New Submission

MOLAS | About MOLAS | Browse Projects | New Submission | Check Submitted jobs

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 Update Reset

Create New Project (Provide a static link for submission revised and shared for 12 months)

Just a try without Project creation (Just a dynamic link available for a week)

Clear All

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

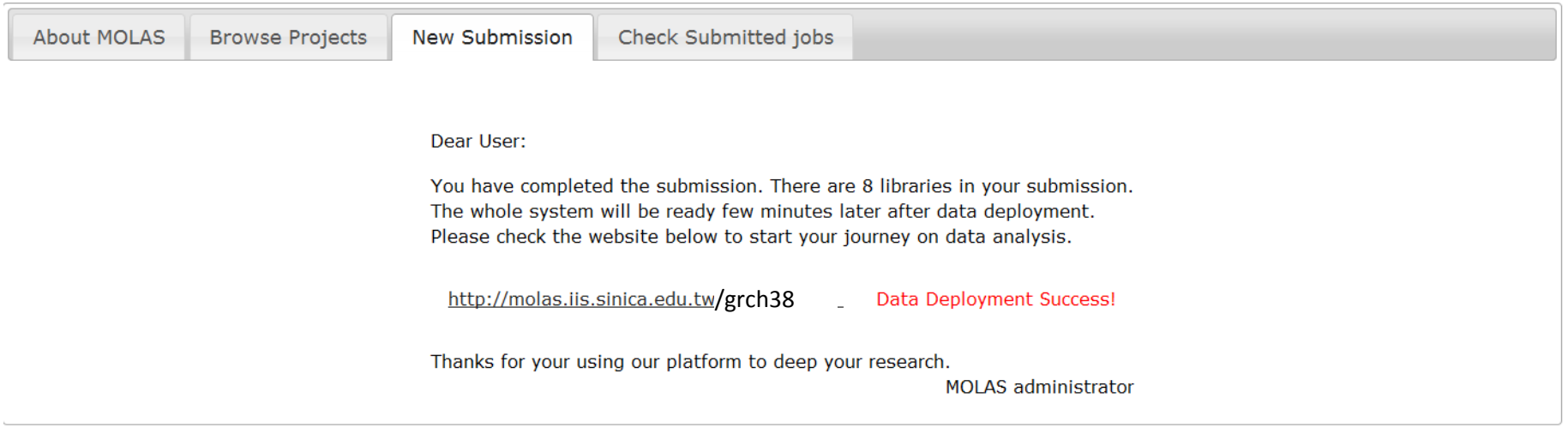
Contact E-mail as Account: molas.iis@gmail.com

Password: ••••

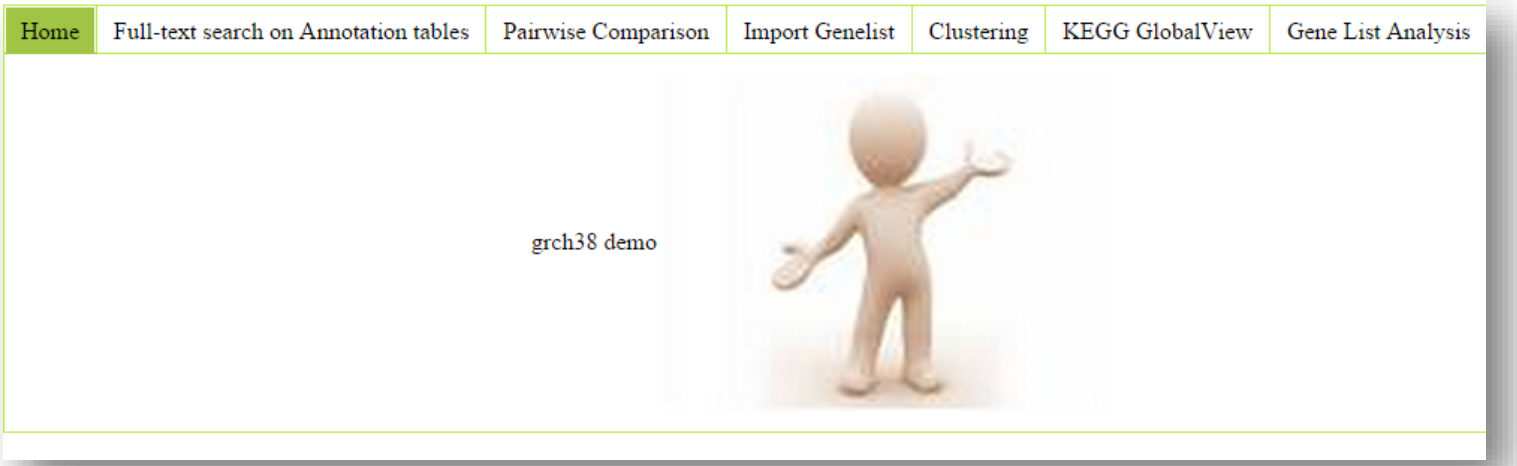
Open to Public: **Yes**

No share this project data to my friends with this secret word:

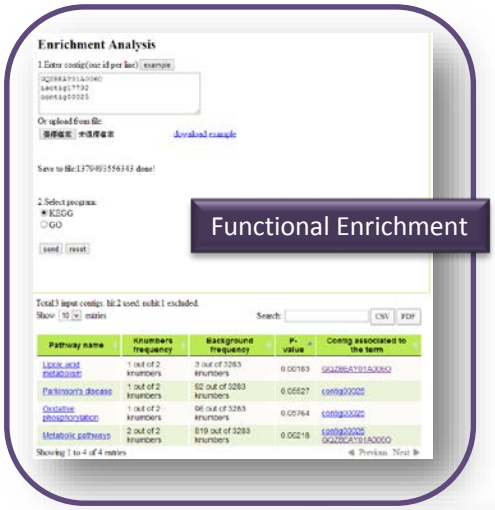
Deployment Success

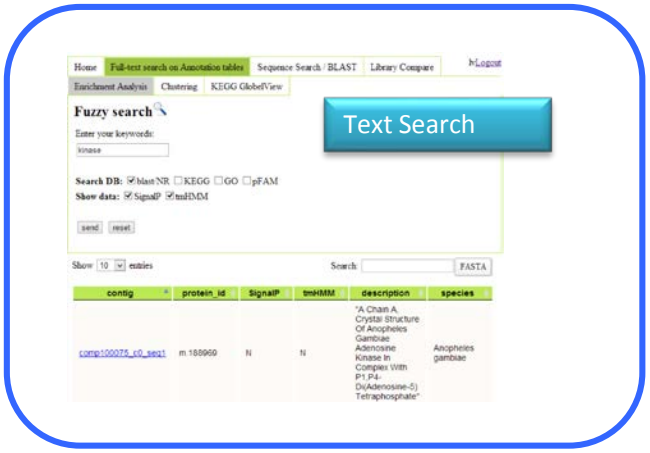


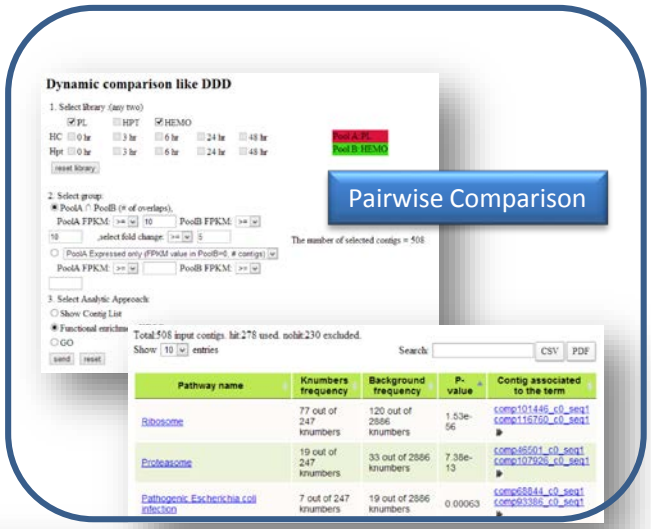
Browse project and …….

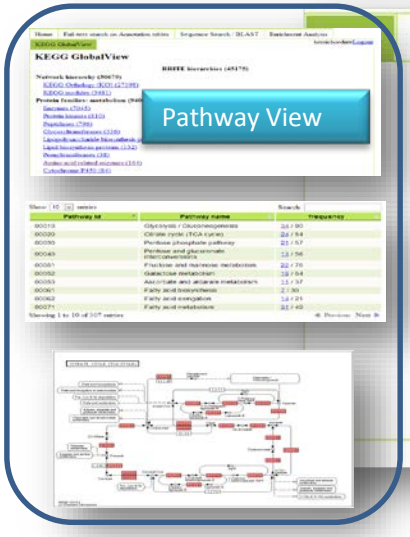


http://molas.iis.sinica.edu.tw/human_grch38_demo/

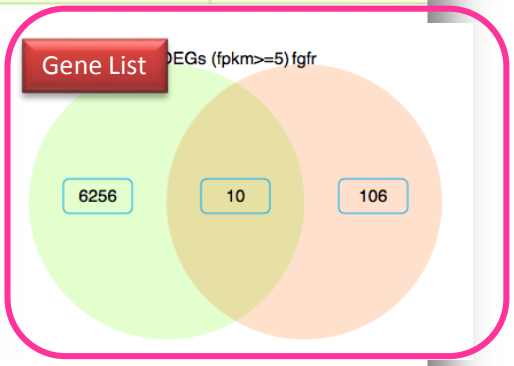

Enrichment Analysis
Functional Enrichment

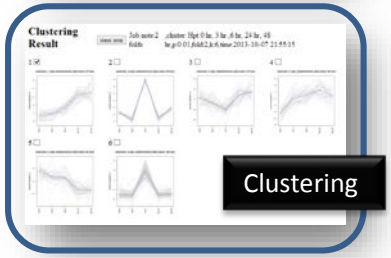

Text Search

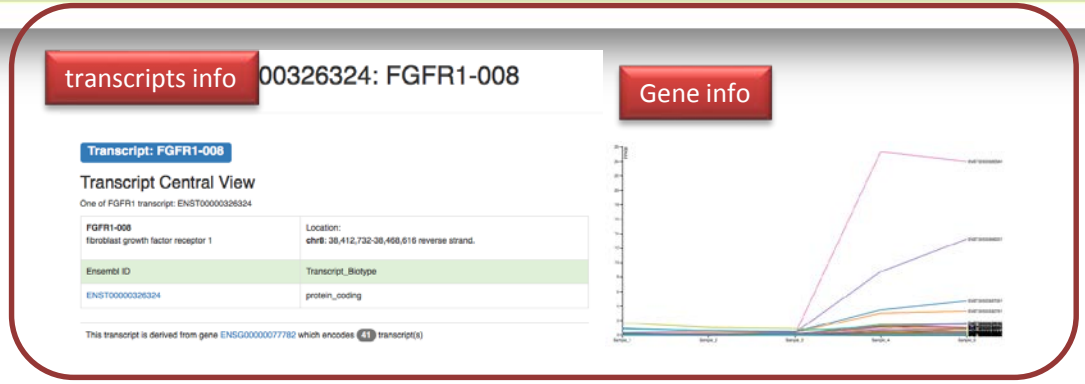

Dynamic comparison like DDD
Pairwise Comparison


Pathway View

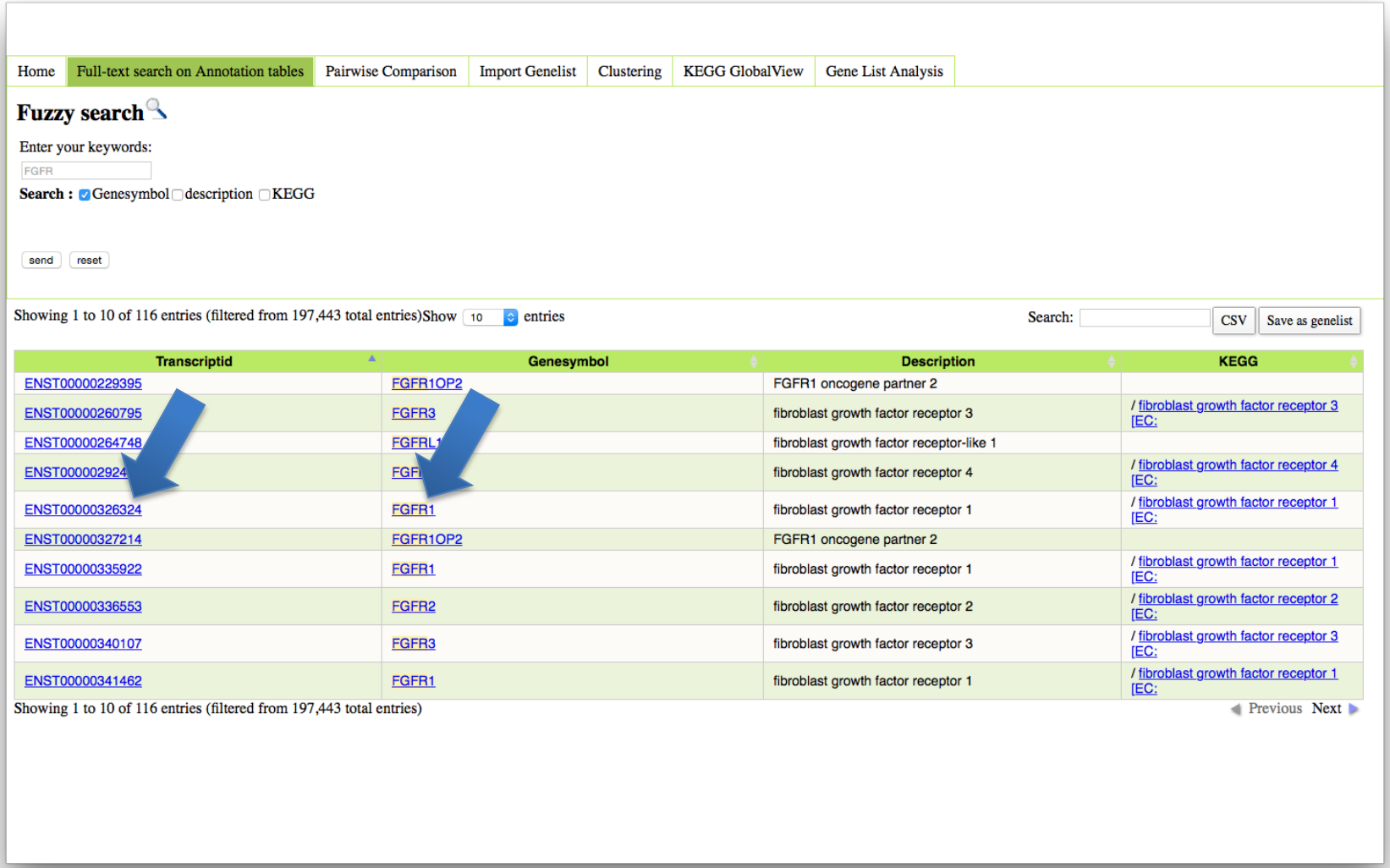
Full-text search on Annotation tables
Pairwise Comparison
Import Genelist
Clustering
KEGG GlobalView
Gene List Analysis
grch38 demo


Gene List
6256
10
106


Clustering


transcripts info
00326324: FGFR1-008
Gene info
Transcript: FGFR1-008
Transcript Central View

Fuzzy Search



Home | Full-text search on Annotation tables | Pairwise Comparison | Import Genelist | Clustering | KEGG GlobalView | Gene List Analysis

Fuzzy search

Enter your keywords:

FGFR

Search : ☑ Genesymbol ☐ description ☐ KEGG

send reset

Showing 1 to 10 of 116 entries (filtered from 197,443 total entries) Show 10 entries

Search: CSV Save as genelist

Transcriptid	Genesymbol	Description	KEGG
ENST00000229395	FGFR1OP2	FGFR1 oncogene partner 2	
ENST00000260795	FGFR3	fibroblast growth factor receptor 3	/ fibroblast growth factor receptor 3 [EC:
ENST00000264748	FGFRL1	fibroblast growth factor receptor-like 1	
ENST00000292[illegible]	FGF[illegible]	fibroblast growth factor receptor 4	/ fibroblast growth factor receptor 4 [EC:
ENST00000326324	FGFR1	fibroblast growth factor receptor 1	/ fibroblast growth factor receptor 1 [EC:
ENST00000327214	FGFR1OP2	FGFR1 oncogene partner 2	
ENST00000335922	FGFR1	fibroblast growth factor receptor 1	/ fibroblast growth factor receptor 1 [EC:
ENST00000336553	FGFR2	fibroblast growth factor receptor 2	/ fibroblast growth factor receptor 2 [EC:
ENST00000340107	FGFR3	fibroblast growth factor receptor 3	/ fibroblast growth factor receptor 3 [EC:
ENST00000341462	FGFR1	fibroblast growth factor receptor 1	/ fibroblast growth factor receptor 1 [EC:

Showing 1 to 10 of 116 entries (filtered from 197,443 total entries)

Previous Next

ENST00000326324: FGFR1-008

Transcript Central View

Transcript: FGFR1-008

Transcript Central View

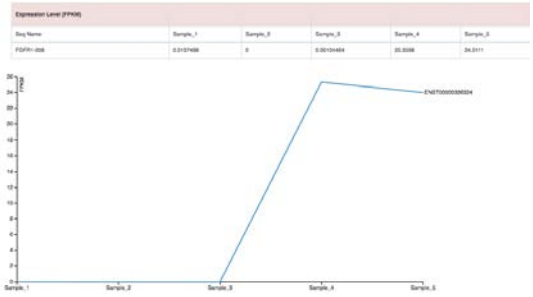
One of FGFR1 transcript: ENST00000326324

FGFR1-008 fibroblast growth factor receptor 1	Location: **chr8**: 38,412,732-38,468,616 reverse strand.
Ensembl ID	Transcript_Biotype
ENST00000326324	protein_coding

descriptions about this transcript

This transcript is derived from gene ENSG00000077782 which encodes 41 transcript(s)

Short cut to
the Gene Central View



The expression level
(the user uploaded data)

External Links for further info

Link to ensembl for
- the translated product ENSP00000327229 of transcript FGFR1-008/ENST00000326324
- transcript summary of FGFR1-008/ENST00000326324
- gene summary of FGFR1/ENSG00000077782 or Viewing the gene / transcript structure

Link to NCBI for
- gene centered view of FGFR1 in Entrez Gene (Gene ID:)
- the transcript cluster of human FGFR1 in UniGene

External Links to get more info

GO Information

GO information		
Biological Process		
	GO:0010518	positive regulation of phospholipase activity
	GO:0010863	positive regulation of phospholipase C activity
	GO:0014068	positive regulation of phosphatidylinositol 3-kinase signaling
	GO:0001501	skeletal system development
	GO:0016477	cell migration
	GO:0000165	MAPK cascade
	GO:0001764	neuron migration
	GO:0018108	peptidyl-tyrosine phosphorylation
	GO:0002376	immune system process
	GO:0034641	cellular nitrogen compound metabolic process
	GO:0038095	Fc-epsilon receptor signaling pathway
	GO:0040011	locomotion

GO terms tagged in this gene

Gene Central View

Gene: FGFR1

Gene Central View

FGFR1 fibroblast growth factor receptor 1	
Ensembl ID	Gene_Biotype
ENSG00000077782	protein_coding
Synonym/ prev Symbol	chromosome location
	ch8: 38,411,138-38,468,834 reverse strand.

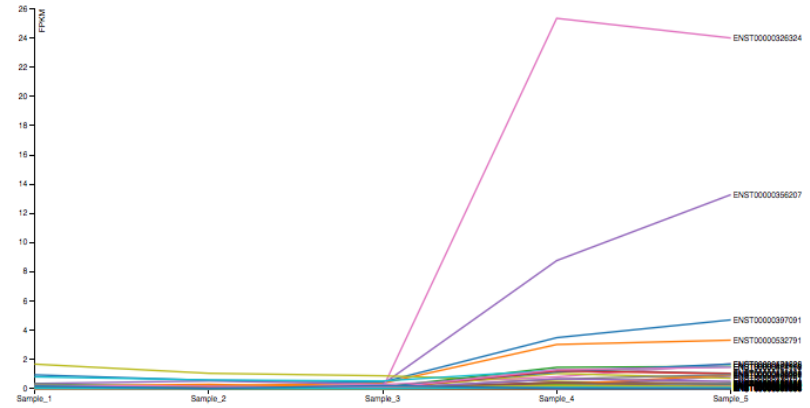
descriptions about this gene

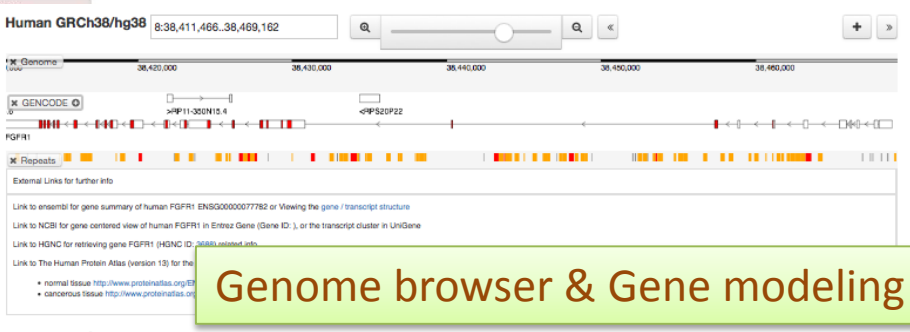
This gene is known with **41** transcript(s):

The expression level of all the possible transcripts:

Expression Level (FPKM)						
Seq Name Transcript ID	Transcript Biotype	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5
ENST00000425967 | [Ensembl]	protein_coding	0	0	0	0.0404075	0.378136
ENST00000434187 | [Ensembl]	protein_coding	0				
ENST00000531196 | [Ensembl]	nonsense_mediated_decay	0				
ENST00000526688 | [Ensembl]	processed_transcript	0				
ENST00000524528 | [Ensembl]	retained_intron	0.152759				
ENST00000326324 | [Ensembl]	protein_coding	0.0137498	0	0.00104464	25.3598	
ENST00000530668 | [Ensembl]	protein_coding	0	0	0.0290441	0	
		0	0	0	0.251281	

The expression level of all transcripts of this gene (the user uploaded data)





Genome browser & Gene modeling

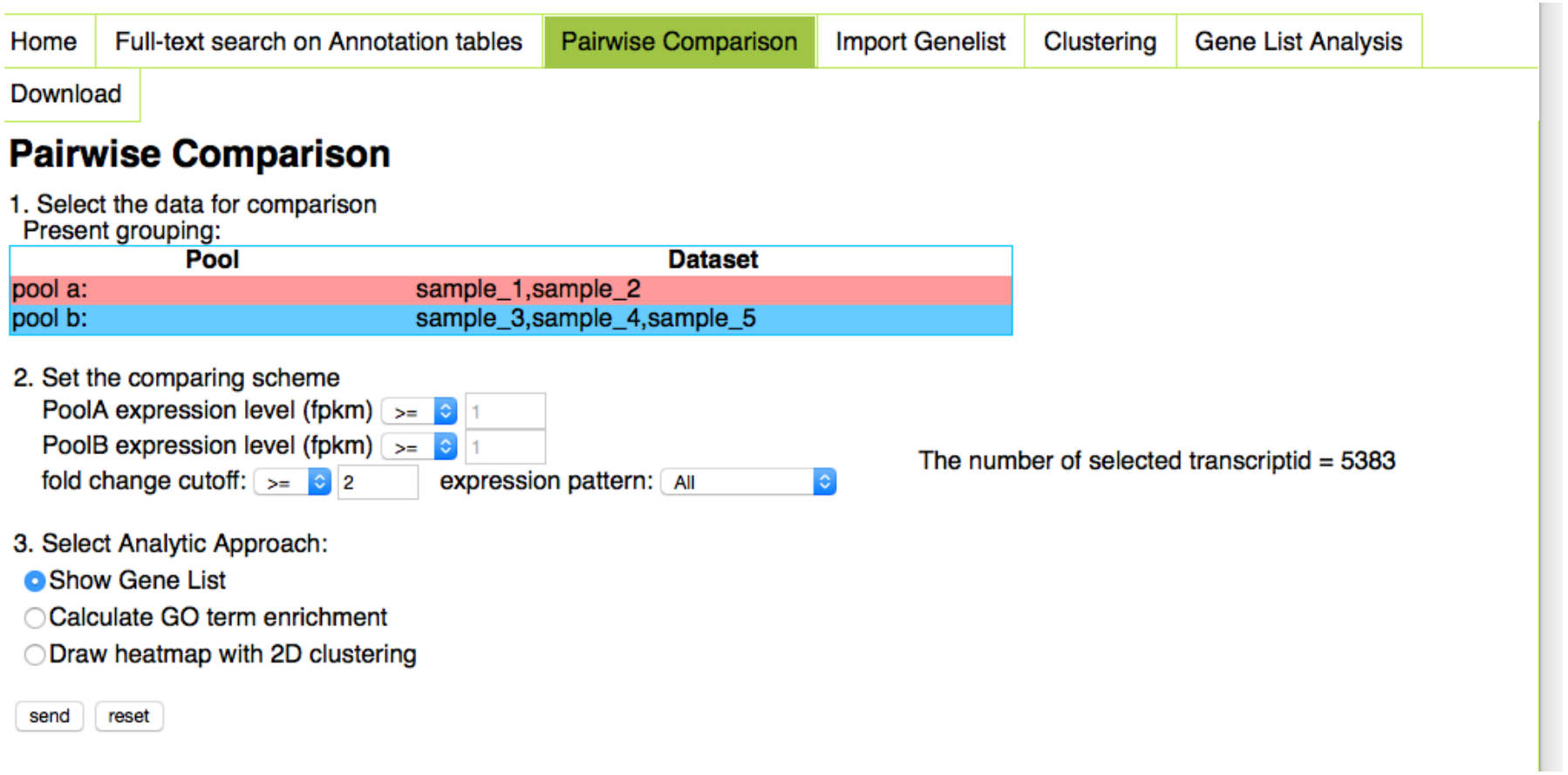
Neighboring Genes

Drop down the list to select a range. Given this gene, its downstream (e.g. - 1Kb) and upstream (e.g. + 1Kb) flanking genes are extracted as a list.

10 Kb

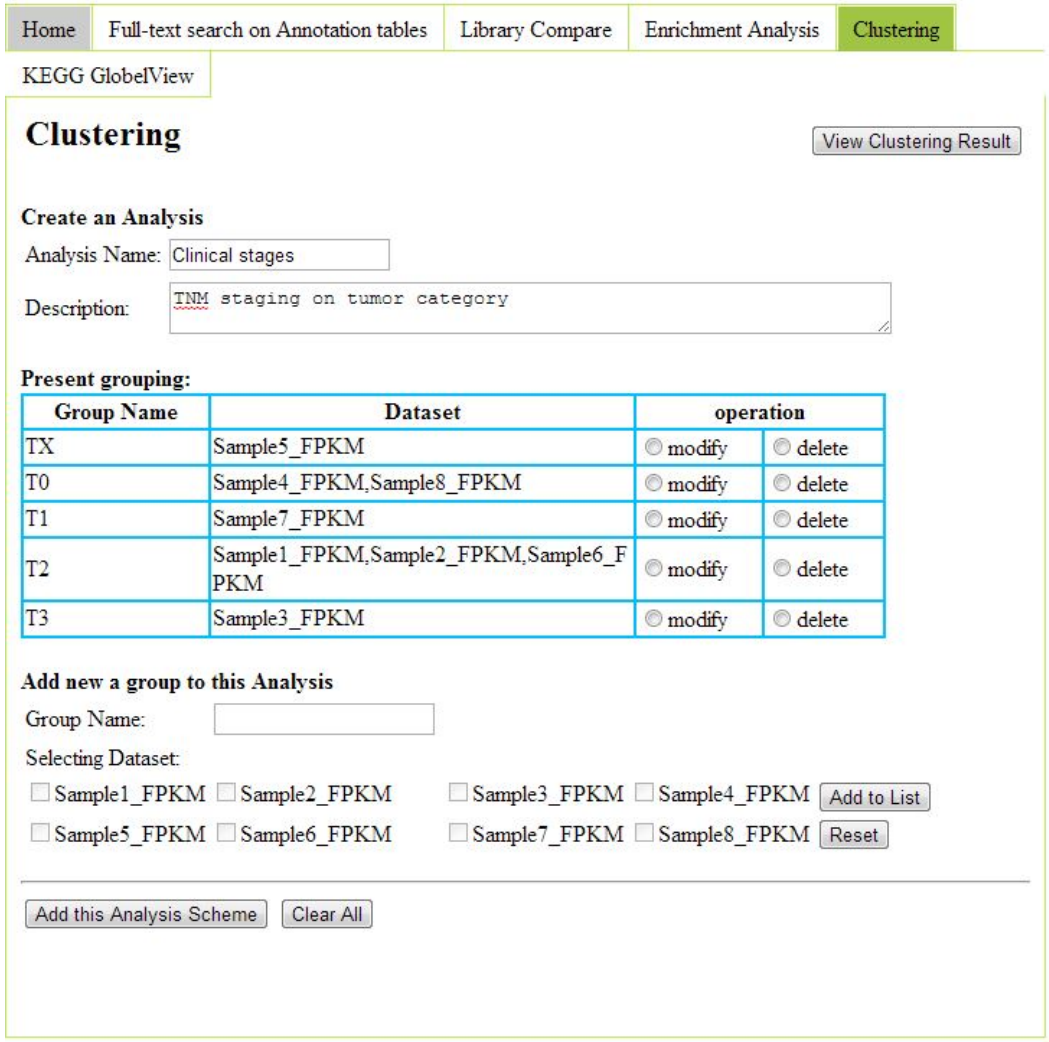
Neighboring Genes (-10000 bps ~ +10000 bps)				Save list
1	ENSG00000165046 | [Ensembl]	LETM2	protein_coding	8: 38386207 - 38409527 , forward strand
2	ENSG00000254981 | [Ensembl]	RP11-350N15.3	antisense	8: 38400536 - 38401683 , reverse strand
3	ENSG00000272159 | [Ensembl]	RP11-350N15.6	antisense	8: 38406048 - 38408742 , reverse strand
4	ENSG00000077782 | [Ensembl]	FGFR1	protein_coding	8: 38411138 - 38468834 , reverse strand
5	ENSG00000255201 | [Ensembl]	RP11-350N15.4	antisense	8: 38421889 - 38426096 , forward strand
6	ENSG00000239218 | [Ensembl]	RPS20P22	transcribed_processed_pseudogene	8: 38434347 - 38435664 , reverse strand

Pairwise Comparison



Clustering

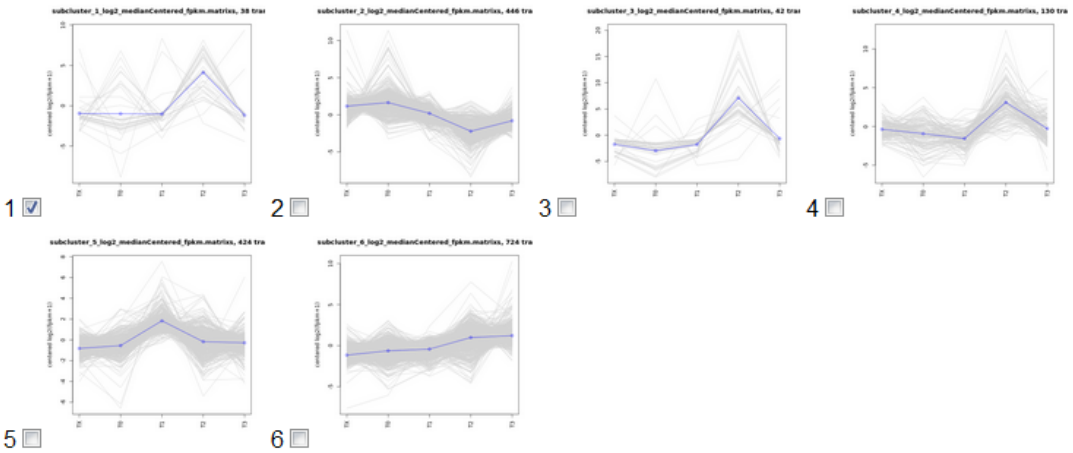
If some samples have similar properties, clustering can help group them together and perform gene expression profile analysis.



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



Select Analytic Approach:

- ○ Show Gene List
- ◉ Functional enrichment KEGG
- ○ GO

[send] [reset]

KEGG Pathway

3. Select Analytic Approach:

- ○ Show Gene List
- ○ Calculate GO term enrichment
- ◉ Calculate KEGG pathway enrichment
- ○ Draw heatmap with 2D clustering

send reset

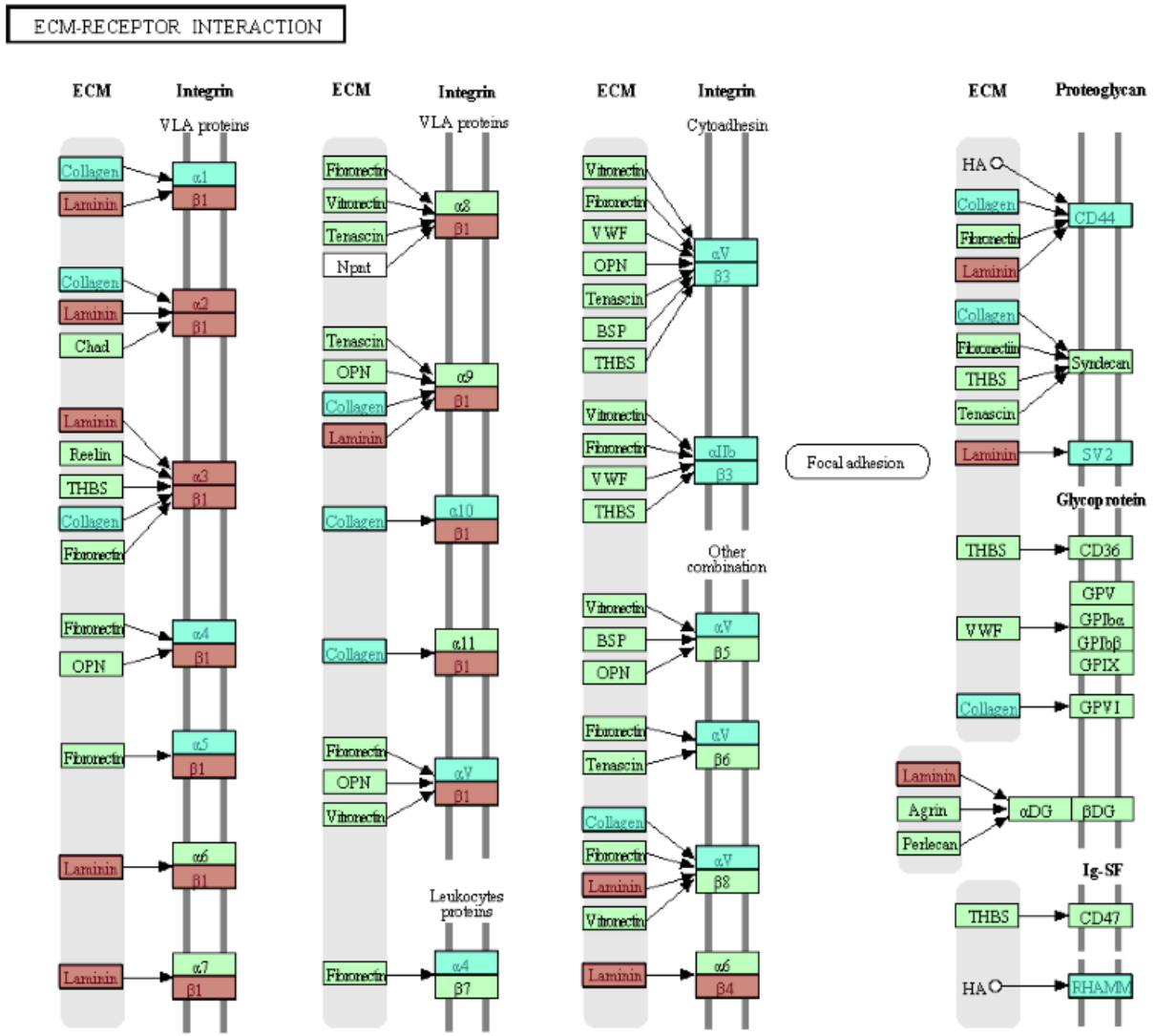
Total:2348 input gene symbol. hit:843 used. nohit:1505 excluded.

Show 10 entries

Search: CSV PDF

Pathway name	Knumbers frequency	Background frequency	P-value	transcriptid associated to the term
PI3K-Akt signaling pathway	55 out of 588 knumbers	243 out of 8706 knumbers	4.50e-16	ENST00000205386 ENST00000222139
Pathways in cancer	62 out of 588 knumbers	322 out of 8706 knumbers	1.89e-14	ENST00000205386 ENST00000222462
ECM-receptor interaction	25 out of 588 knumbers	64 out of 8706 knumbers	1.10e-13	ENST00000205386 ENST00000225964
Transcriptional misregulation in cancers	40 out of 588 knumbers	158 out of 8706 knumbers	1.10e-13	ENST00000222462 ENST00000247182
Focal adhesion	39 out of 588 knumbers	151 out of 8706 knumbers	1.12e-13	ENST00000205386 ENST00000225964
Regulation of actin cytoskeleton	40 out of 588 knumbers	162 out of 8706 knumbers	2.67e-13	ENST00000257290 ENST00000261799

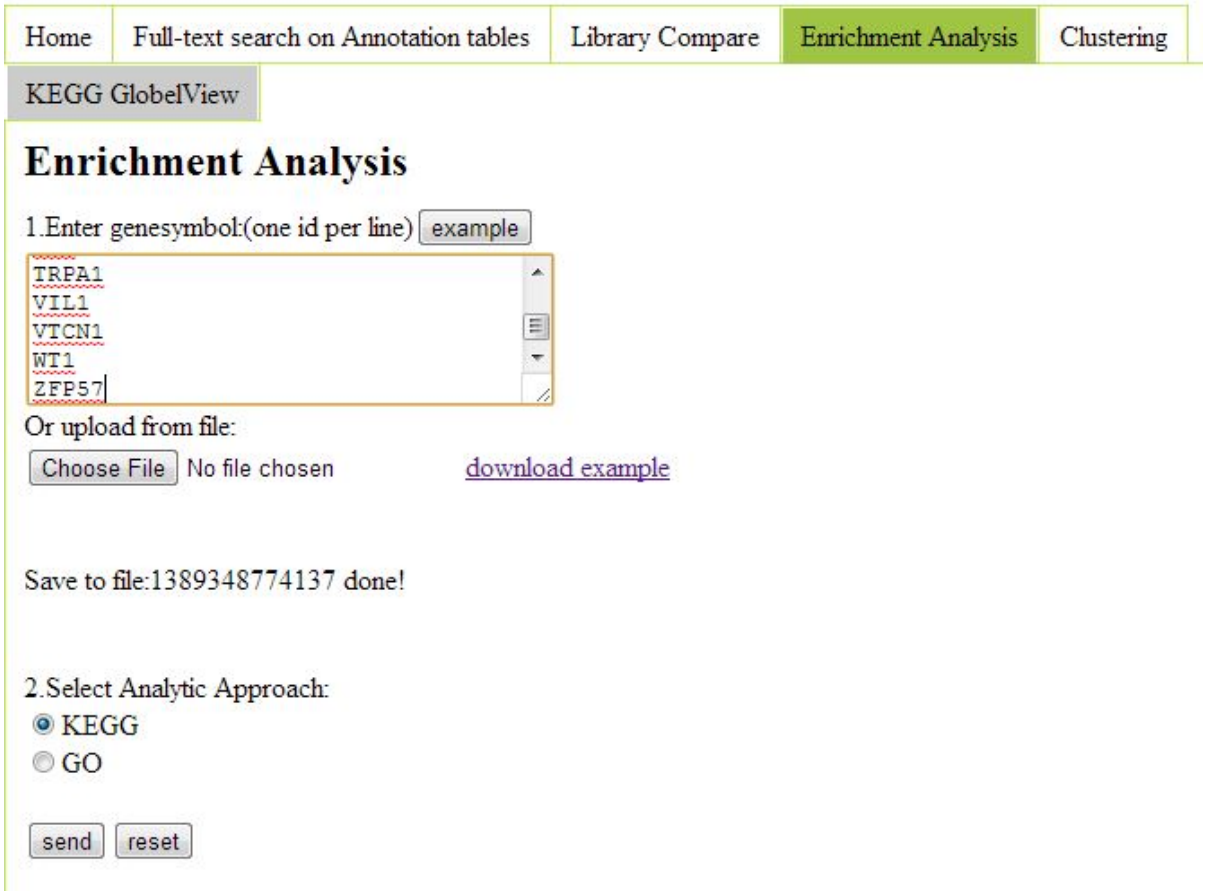
KEGG Pathway



4512 3/7/13
c) Kanehisa Laboratories

Enrichment Analysis

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



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.

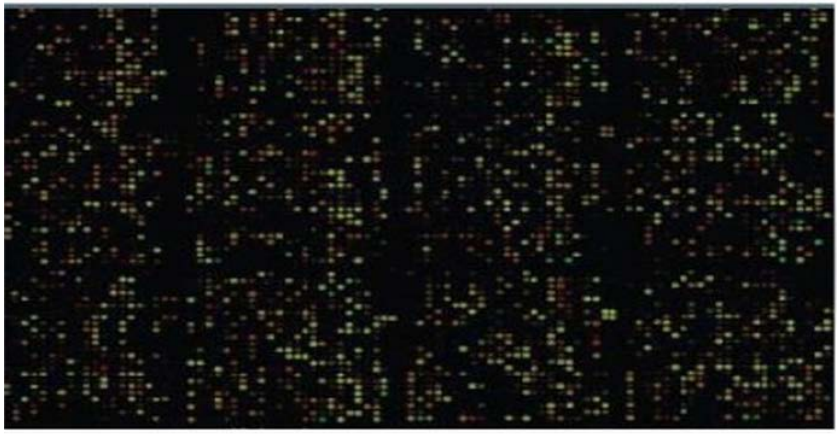
What to do if you have no replicates?

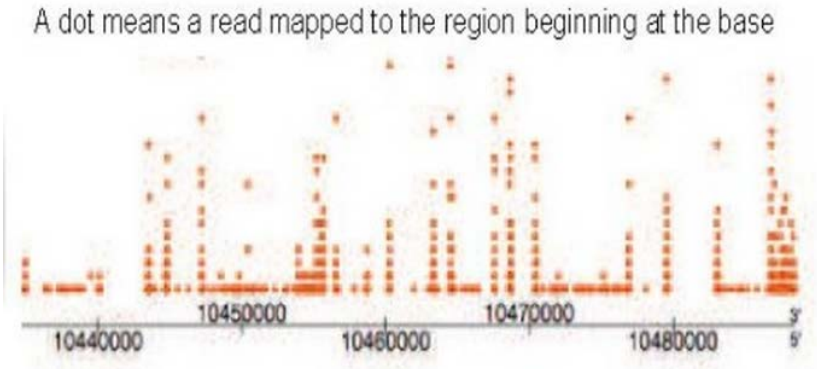
Suggestions from edgeR authors

- 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 unidentical)
    0.1 for genetically identical 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

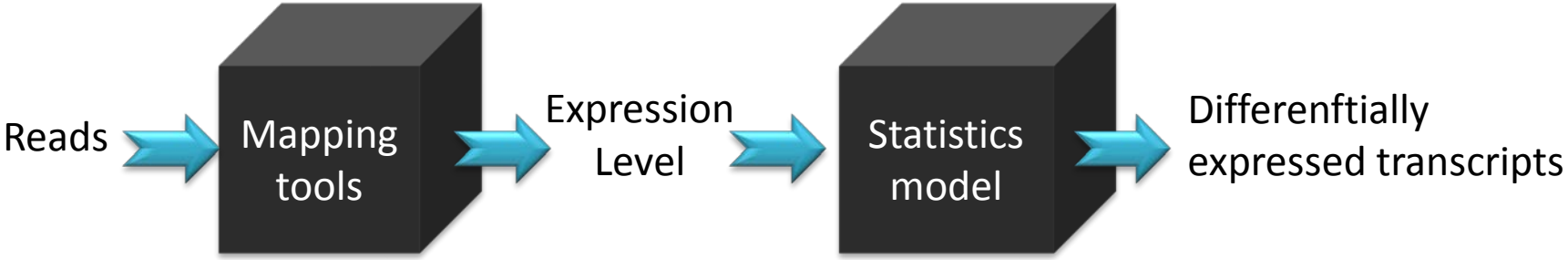
edgeR paper http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/
menu http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Analog signal vs Digital signal





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



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 !!

Thanks for your Attention