Genomics NGS Service

Bioinformatics Analysis of RNA-seq de-novo transcriptome by Trinity

Help manual

2017

Genomics NGS Analysis Team

Genomics BioSci & Tech. Co., Ltd.

Table of Contents

Experiment Process

a.) Purify and fragment mRNA: Using poly-T oligo-attached beads to purify mRNA, which is also fragmented for cDNA synthesis.

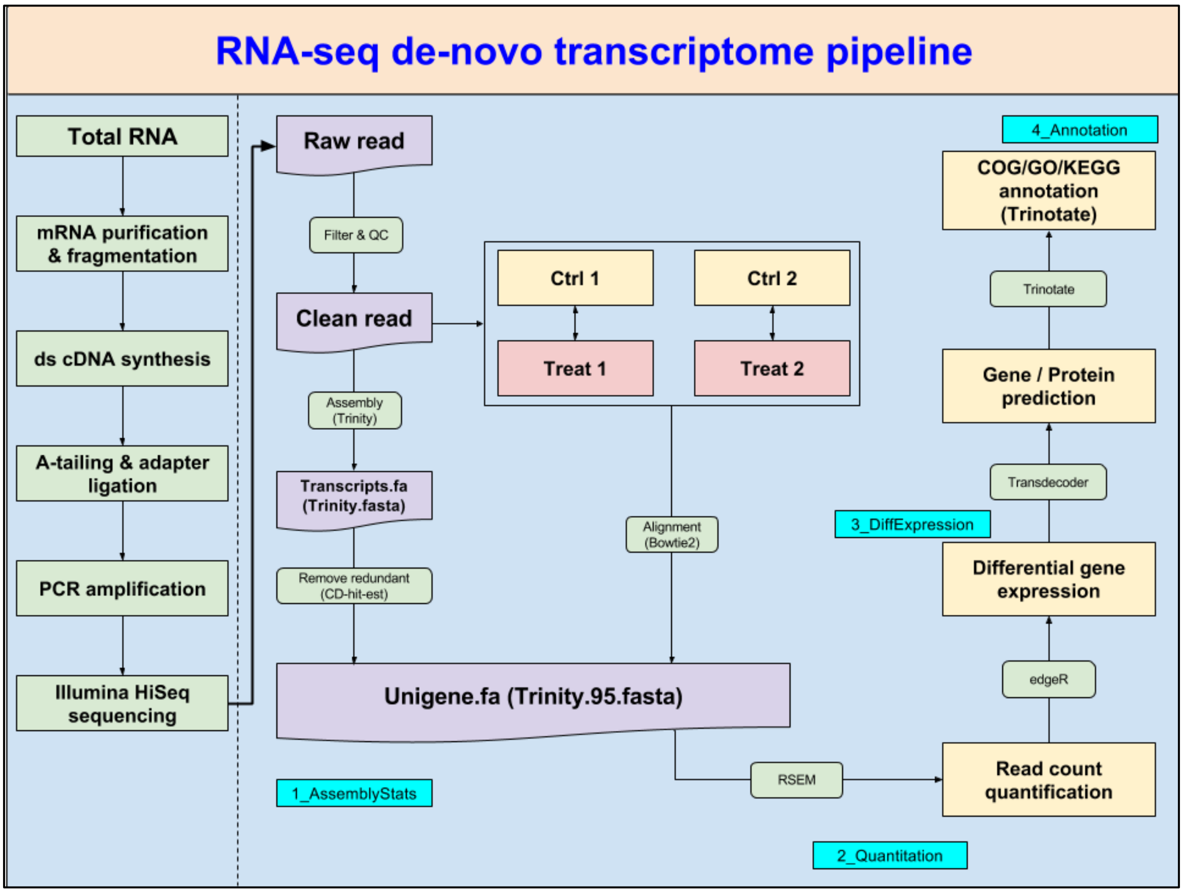
b.) Double strand cDNA synthesis: Using reverse transcriptase and random primer to synthesize first strand cDNA, and using dUTP in place of dTTP to generate double-strand cDNA.

c.) A-tailing and Adaptor Ligation: A single ‘A’ nucleotide is added to 3’ end of ds cDNAs. Then, multiple indexing adapters are ligated to 5’ and 3’ of the ends of the ds cDNA.

d.) PCR amplification Using PCR to selectively amplify those DNA fragments that have adapters on both ends.

e.) Library quality validating: Library was validated on Agilent 2100 Bio-analyzer and Real-Time PCR System.

f.) Sequencing by Illumina HiSeq platform

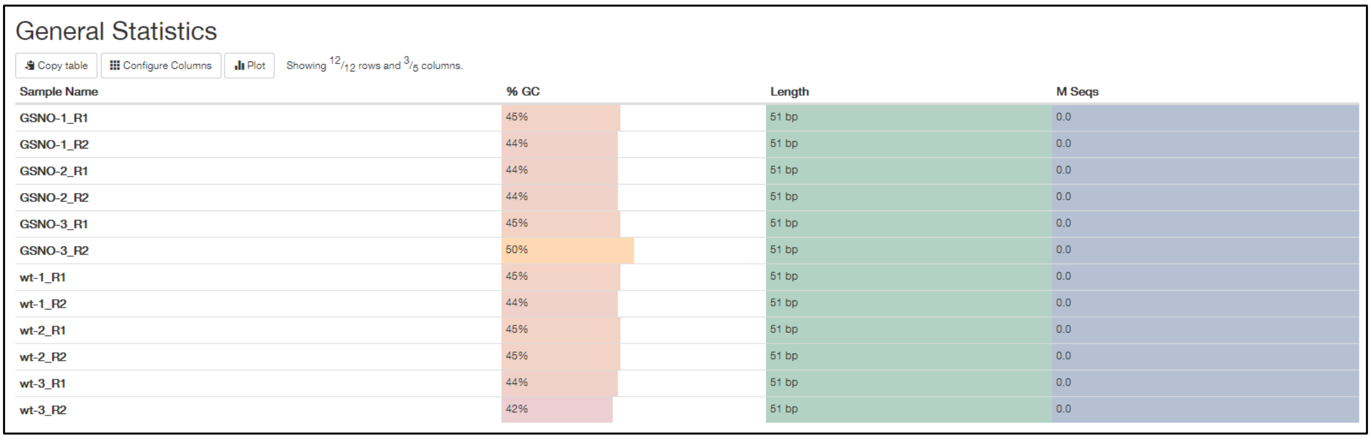

RNA-seq de-novo transcriptome pipeline
Total RNA
mRNA purification & fragmentation
ds cDNA synthesis
A-tailing & adapter ligation
PCR amplification
Illumina HiSeq sequencing
Raw read
Filter & QC
Clean read
Assembly (Trinity)
Transcripts.fa (Trinity.fasta)
Remove redundant (CD-hit-est)
Unigene.fa (Trinity.95.fasta)
1_AssemblyStats
Ctrl 1
Ctrl 2
Treat 1
Treat 2
Alignment (Bowtie2)
RSEM
Read count quantification
2_Quantitation
edgeR
Differential gene expression
3_DiffExpression
Transdecoder
Gene / Protein prediction
Trinotate
COG/GO/KEGG annotation (Trinotate)
4_Annotation

Bioinformatics analysis

0. Read QC (0_ReadQC)

We are using “**MultiQC v1.2**” for evaluating read quality. MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples [1].

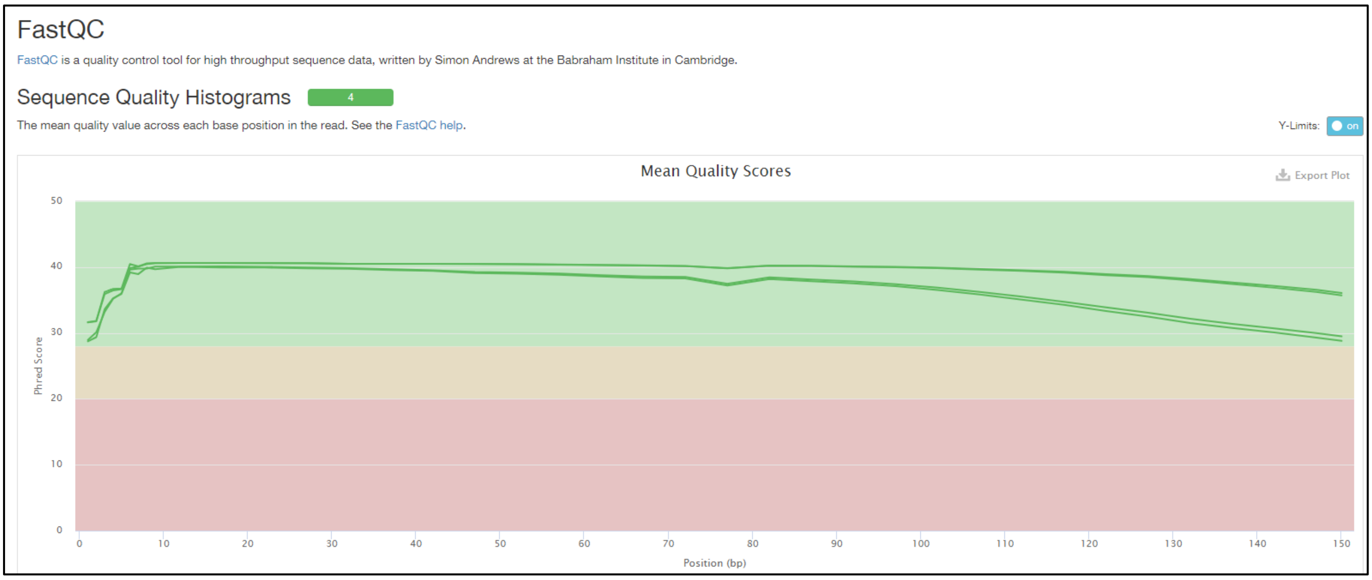
multiqc_report.html



General Statistics

Showing 12/12 rows and 3/5 columns.

Sample Name	% GC	Length	M Seqs
GSNO-1_R1	45%	51 bp	0.0
GSNO-1_R2	44%	51 bp	0.0
GSNO-2_R1	44%	51 bp	0.0
GSNO-2_R2	44%	51 bp	0.0
GSNO-3_R1	45%	51 bp	0.0
GSNO-3_R2	50%	51 bp	0.0
wt-1_R1	45%	51 bp	0.0
wt-1_R2	44%	51 bp	0.0
wt-2_R1	45%	51 bp	0.0
wt-2_R2	45%	51 bp	0.0
wt-3_R1	44%	51 bp	0.0
wt-3_R2	42%	51 bp	0.0



FastQC

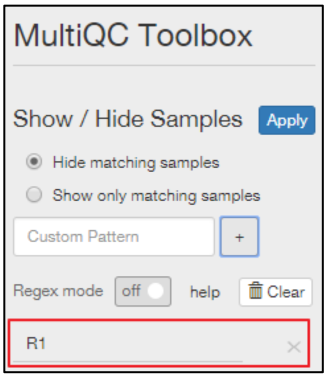
FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Quality Histograms

The mean quality value across each base position in the read. See the FastQC help.

Y-Limits: on

[Notice]:
Using “Toolbox” in the right panel to help you show/hide samples.
Red square: mask all name containing “R1” sample



MultiQC Toolbox

Show / Hide Samples

● Hide matching samples
○ Show only matching samples

Regex mode off

R1

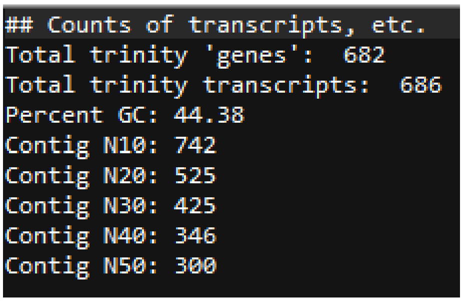
1. Assembly Stats (1_AssemblyStats)

"Trinity v2.3.2" is a well-known transcriptome de-novo assembly tool. It combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes [2].

While Trinity job has been completed, it might usually contain lots of duplicate transcripts existed in data. Thus, we commonly use another clustering tool: CD-HIT-EST [3], for processing redundant transcripts removal and try to get more specific unigenes.

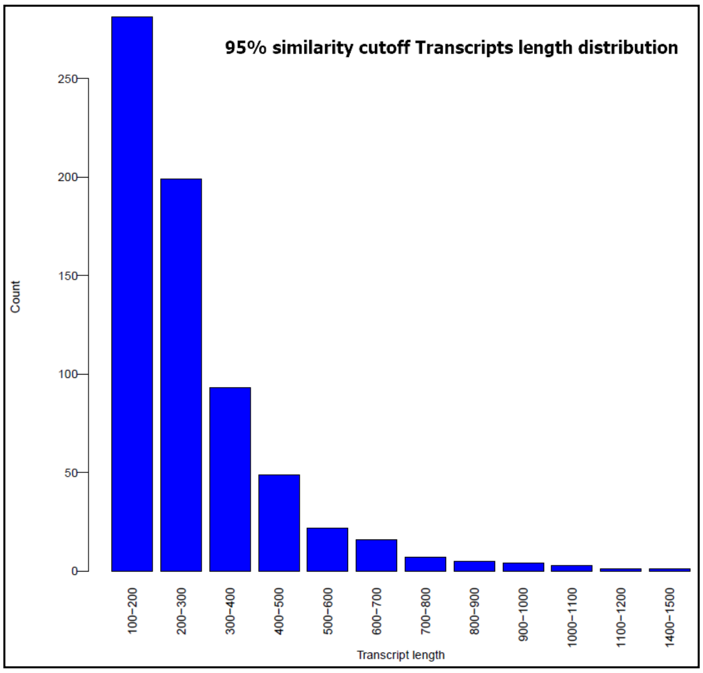
- Trinity parameters:
    - Minimum contig length => 150 bp
- CD-HIT-EST parameters:
    - sequence identity threshold => 95%

Trinity_assembled.final.stats.txt



```
## Counts of transcripts, etc.
Total trinity 'genes':  682
Total trinity transcripts:  686
Percent GC: 44.38
Contig N10: 742
Contig N20: 525
Contig N30: 425
Contig N40: 346
Contig N50: 300
```

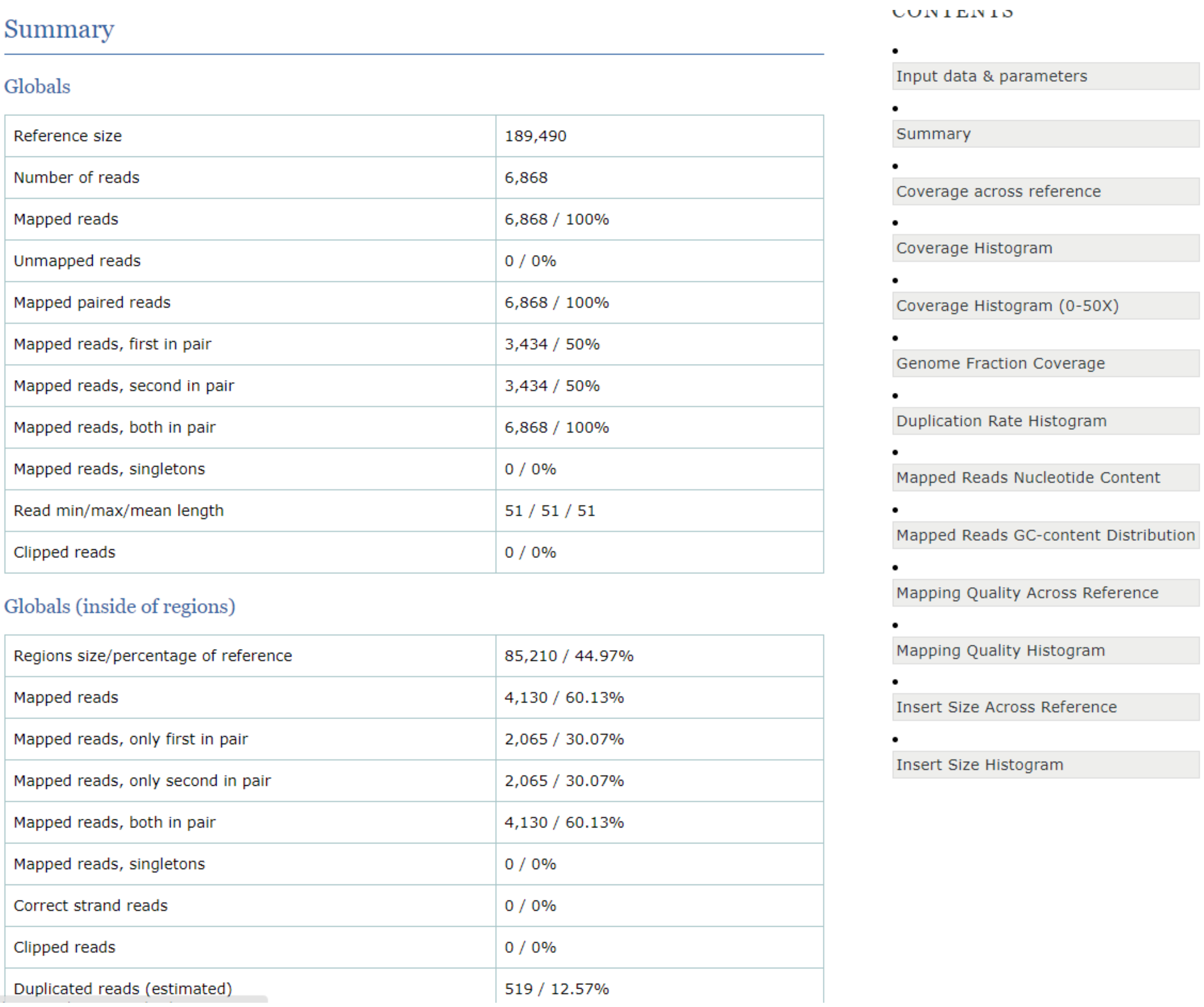
Trinity.95.dist.pdf



2. Read count quantification (2_Quantitation)

In this stage, the de-nove assembled transcriptome will be regarded as backbone reference. All of the samples are going to be aligned for calculating the abundance of read count. The alignment tool we used is “bowtie2 v2.3.2” [4], and the read count quantification tool we used is “RSEM v1.2.31” [5]. The alignment QC report we are using “Qualimap v2” for evaluation [6].

[Alignment stats]:



Summary

Globals

Reference size	189,490
Number of reads	6,868
Mapped reads	6,868 / 100%
Unmapped reads	0 / 0%
Mapped paired reads	6,868 / 100%
Mapped reads, first in pair	3,434 / 50%
Mapped reads, second in pair	3,434 / 50%
Mapped reads, both in pair	6,868 / 100%
Mapped reads, singletons	0 / 0%
Read min/max/mean length	51 / 51 / 51
Clipped reads	0 / 0%

Globals (inside of regions)

Regions size/percentage of reference	85,210 / 44.97%
Mapped reads	4,130 / 60.13%
Mapped reads, only first in pair	2,065 / 30.07%
Mapped reads, only second in pair	2,065 / 30.07%
Mapped reads, both in pair	4,130 / 60.13%
Mapped reads, singletons	0 / 0%
Correct strand reads	0 / 0%
Clipped reads	0 / 0%
Duplicated reads (estimated)	519 / 12.57%

CONTENTS

- [Notice]

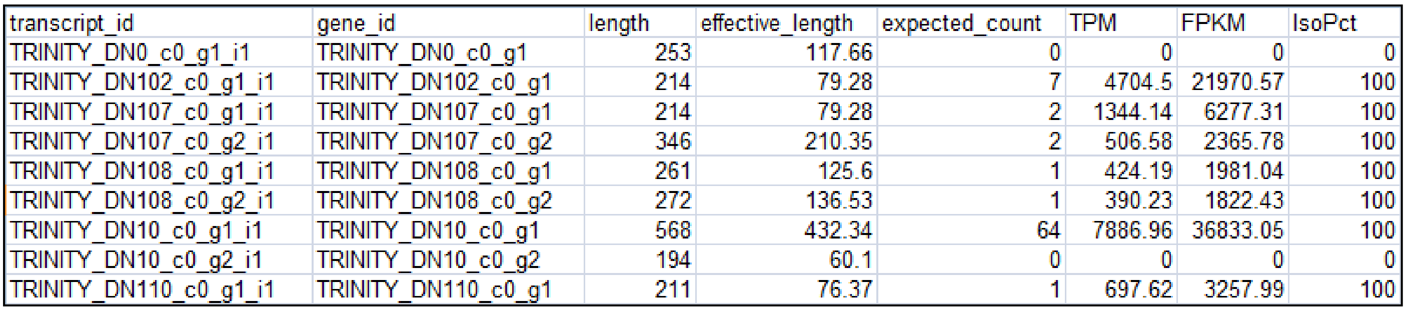
  - **Globals:** read mapping result

  - **Globals (inside of regions):** read alignment stats on transcripts

[RSEM output]:

- RSEM.isoforms.results: EM read counts per Trinity transcript (e.g. TRINITY_DN100_c0_g1_i1)
- RSEM.genes.results: EM read counts per Trinity gene (e.g. TRINITY_DN100_c0_g1)

*** Basically, we are using "RSEM.isoforms.results" for the downstream jobs.**

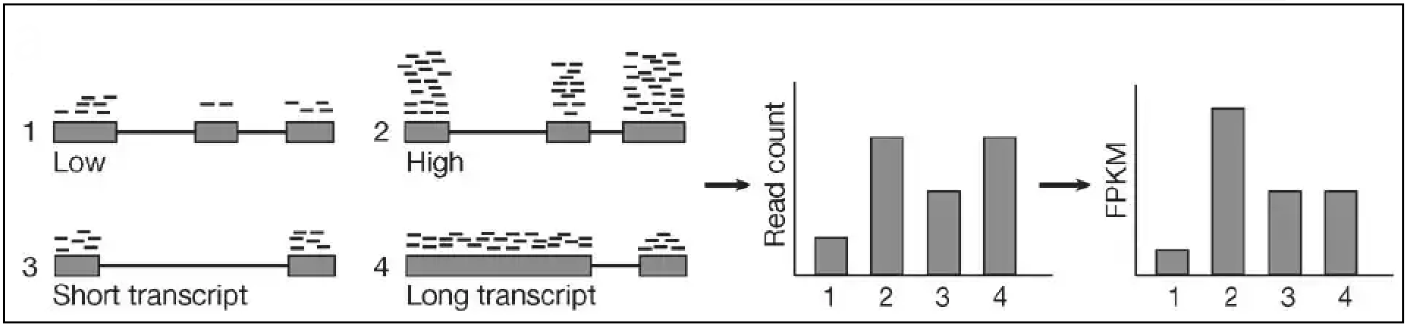


transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN0_c0_g1_i1	TRINITY_DN0_c0_g1	253	117.66	0	0	0	0
TRINITY_DN102_c0_g1_i1	TRINITY_DN102_c0_g1	214	79.28	7	4704.5	21970.57	100
TRINITY_DN107_c0_g1_i1	TRINITY_DN107_c0_g1	214	79.28	2	1344.14	6277.31	100
TRINITY_DN107_c0_g2_i1	TRINITY_DN107_c0_g2	346	210.35	2	506.58	2365.78	100
TRINITY_DN108_c0_g1_i1	TRINITY_DN108_c0_g1	261	125.6	1	424.19	1981.04	100
TRINITY_DN108_c0_g2_i1	TRINITY_DN108_c0_g2	272	136.53	1	390.23	1822.43	100
TRINITY_DN10_c0_g1_i1	TRINITY_DN10_c0_g1	568	432.34	64	7886.96	36833.05	100
TRINITY_DN10_c0_g2_i1	TRINITY_DN10_c0_g2	194	60.1	0	0	0	0
TRINITY_DN110_c0_g1_i1	TRINITY_DN110_c0_g1	211	76.37	1	697.62	3257.99	100

<u>Note</u>:

- **effective_length**: counts only the positions that can generate a valid fragment.
- **expected_count**: sum of the posterior probability of each read comes from this transcripts over all reads.
- **TPM**: Transcripts Per Million. It is a relative measure of transcript abundance. The sum of all transcripts' TPM is 1 million.
- **FPKM**: Fragment Per Kilobase of transcript per Million mapped reads. If reads are paired-end, each R1 or R2 mapped to transcript will be counted 1.
- **IsoPct:** isoform percentage. It is the percentage of expression for a given transcript compared with all expression from that Trinity component. If its parent gene has only one isoform or the gene information is not provided, this field will be set to 100.

$$FPKM = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$



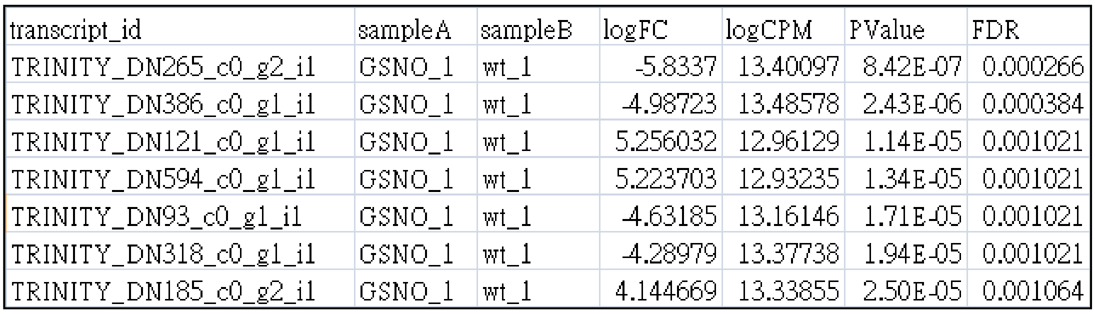
Ref: (http://dx.doi.org/10.1038/nmeth.1613)

3. DGE comparisons (3_DiffExpression)

As we got the read quantification data, various user-provided different comparisons are going to be calculated by “edgeR v3.5” [7], an R package which could process multiple differential expression analysis of RNA-seq expression profile with biological replication.

[DE output]:

- {comparisons}.edgeR.DE_results



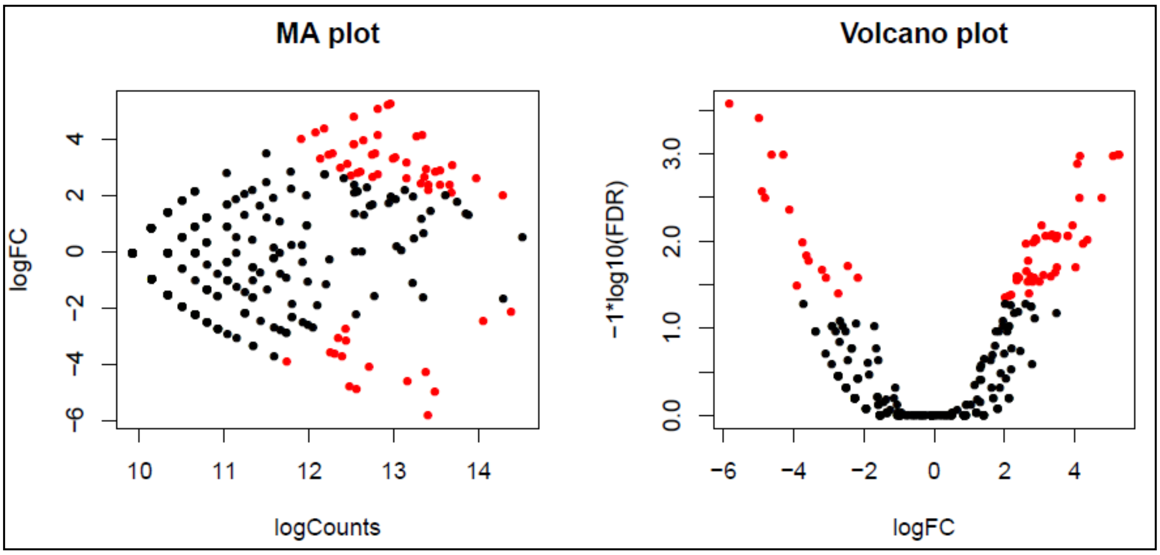
transcript_id	sampleA	sampleB	logFC	logCPM	PValue	FDR
TRINITY_DN265_c0_g2_i1	GSNO_1	wt_1	-5.8337	13.40097	8.42E-07	0.000266
TRINITY_DN386_c0_g1_i1	GSNO_1	wt_1	-4.98723	13.48578	2.43E-06	0.000384
TRINITY_DN121_c0_g1_i1	GSNO_1	wt_1	5.256032	12.96129	1.14E-05	0.001021
TRINITY_DN594_c0_g1_i1	GSNO_1	wt_1	5.223703	12.93235	1.34E-05	0.001021
TRINITY_DN93_c0_g1_i1	GSNO_1	wt_1	-4.63185	13.16146	1.71E-05	0.001021
TRINITY_DN318_c0_g1_i1	GSNO_1	wt_1	-4.28979	13.37738	1.94E-05	0.001021
TRINITY_DN185_c0_g2_i1	GSNO_1	wt_1	4.144669	13.33855	2.50E-05	0.001064

Note:

- **logFC**: log difference between sampleA and sampleB.
- **logCPM**: log counts per million, which is as similar as measuring expression level
- **FDR**: false discovery rate, which could help for validating the false positives in p-value result

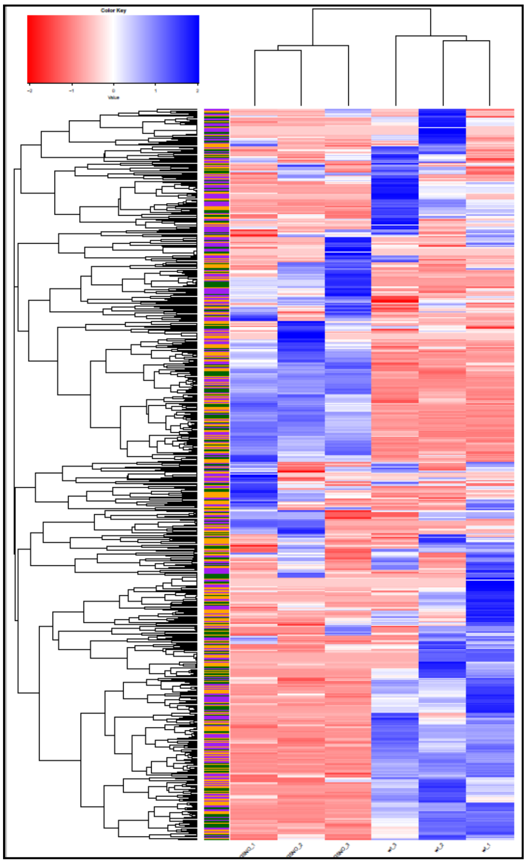
- {comparisons}.edgeR.DE_results.MA_n_Volcano

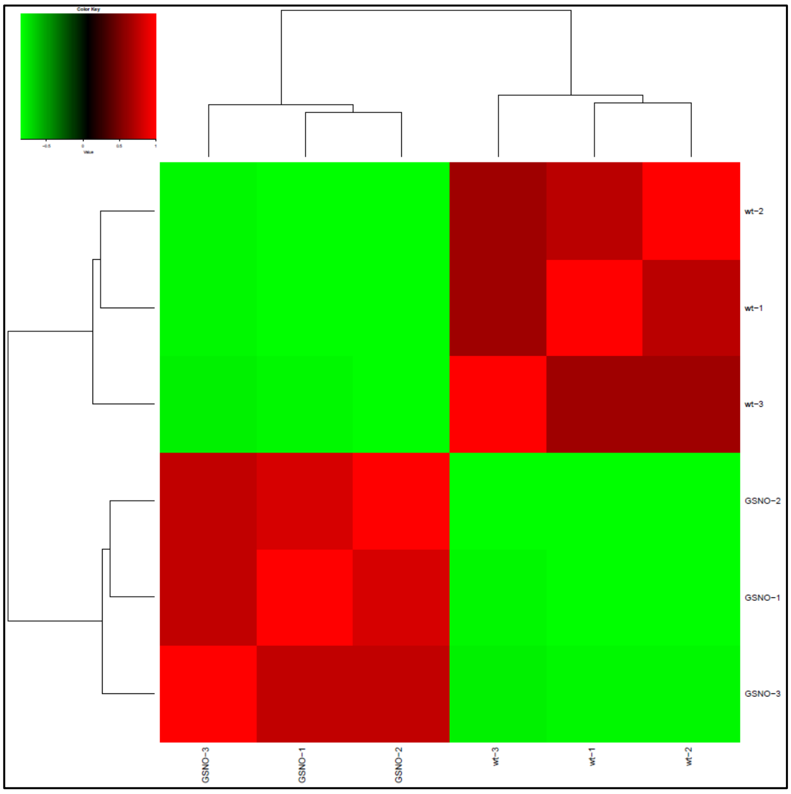
Red dot: p-value < 0.05



- All_samples_heatmap.pdf

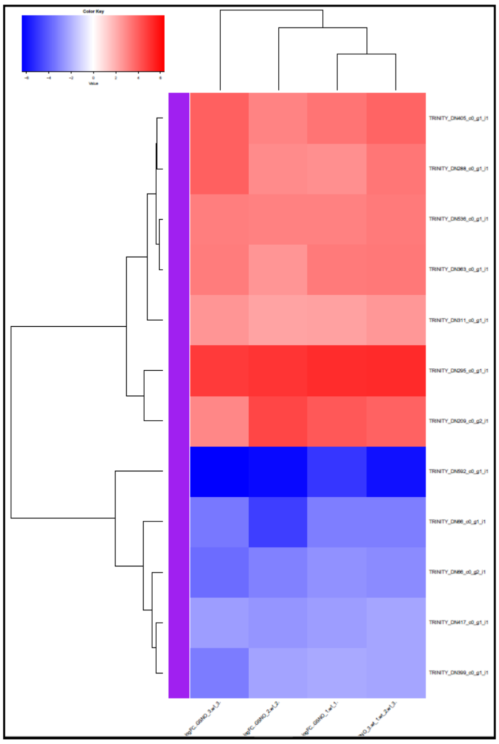
Select **TPM** value to compare DE by heatmap in each comparison.





- all_groups_heatmap.pdf (only intersection genes within groups will be shown)

Select **p-value<0.05** and 1>**logFC>-1** data to compare DE by heatmap in all comparisons, and normalized by **z-score**.



4. Annotation (4_Annotation)

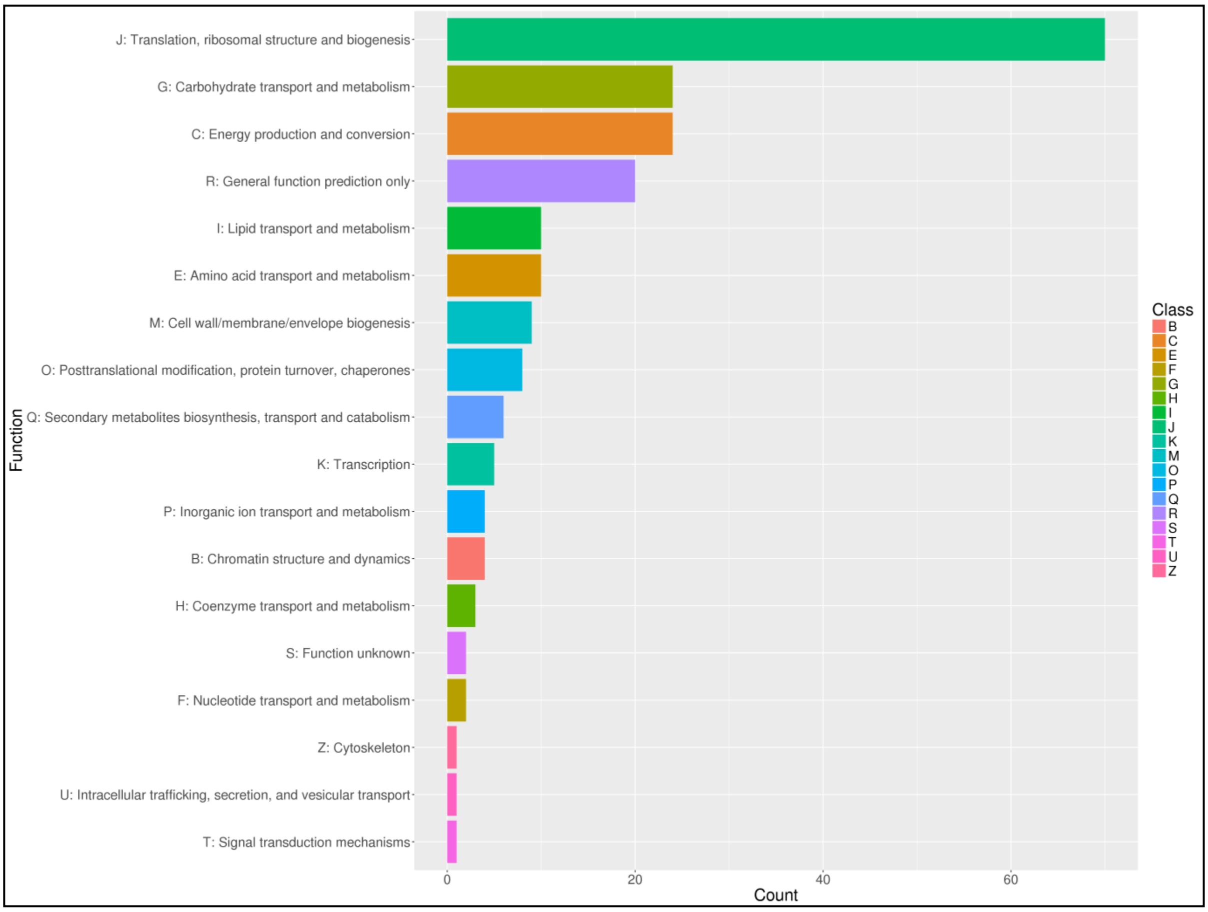
Before annotation work start, we need to parse coding regions within transcripts by gene prediction tool – “Transdecoder v3.0.1” [8] and retrieve protein sequences in the meanwhile.

“Trinotate v3.0.2” is a comprehensive annotation suite designed for functional annotation of de novo assembled transcriptomes, from model or non-model organisms [9]. Our functional annotation works including:

- blastx / blastp: homology search to known & reviewed database (UniprotKB/Swiss-Prot)
- PFAM: protein domain identification
- signalP / TmHMM protein signal peptide and transmembrane domain prediction
- COG / GO / KEGG: functional & pathway annotation

[Protein group function annotation by COG/eggNOG]

In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG.



[GO annotation of transcripts]:

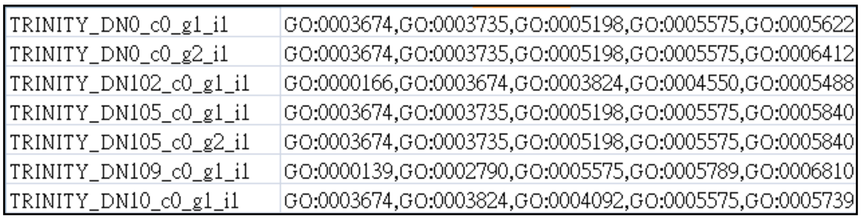
Gene ontology concern with annotation of genes and gene products and to provide centralized access to resources and tools. both GO and COG provide specific information about gene or gene products.

There are three main classes in GO database:

1. **Cellular Component:** These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
2. **Biological Process:** A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions.
3. **Molecular Function:** Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity".

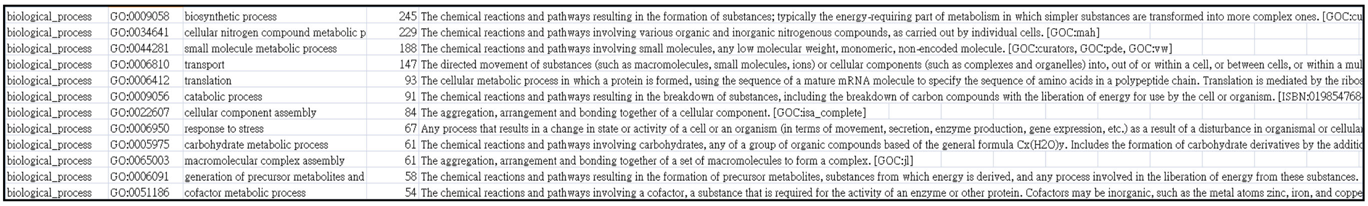
All transcripts are searched to **GO slim database** which contain a subset of the terms in the whole GO. GO slims are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required. Once the GO terms have been corresponded to the transcripts, **Map2Slim** could help us to dig out more informative annotation of transcripts' function.

Trinotate_report.xls.gene_ontology (GO terms extraction)



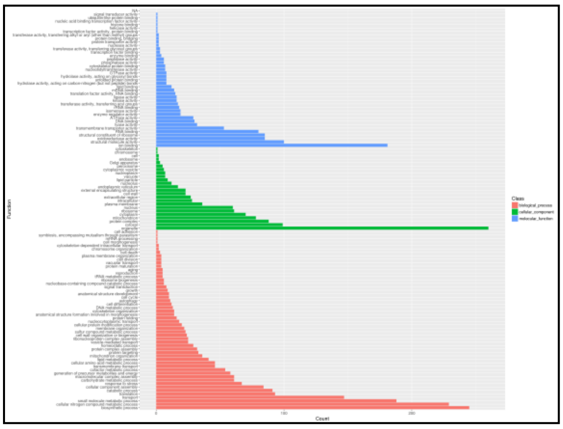
TRINITY_DN0_c0_g1_i1	GO:0003674,GO:0003735,GO:0005198,GO:0005575,GO:0005622
TRINITY_DN0_c0_g2_i1	GO:0003674,GO:0003735,GO:0005198,GO:0005575,GO:0006412
TRINITY_DN102_c0_g1_i1	GO:0000166,GO:0003674,GO:0003824,GO:0004550,GO:0005488
TRINITY_DN105_c0_g1_i1	GO:0003674,GO:0003735,GO:0005198,GO:0005575,GO:0005840
TRINITY_DN105_c0_g2_i1	GO:0003674,GO:0003735,GO:0005198,GO:0005575,GO:0005840
TRINITY_DN109_c0_g1_i1	GO:0000139,GO:0002790,GO:0005575,GO:0005789,GO:0006810
TRINITY_DN10_c0_g1_i1	GO:0003674,GO:0003824,GO:0004092,GO:0005575,GO:0005739

GO_mapping.txt (informative GO annotation)



biological_process	GO:0009058	biosynthetic process	245	The chemical reactions and pathways resulting in the formation of substances; typically the energy-requiring part of metabolism in which simpler substances are transformed into more complex ones. [GOC:cu
biological_process	GO:0034641	cellular nitrogen compound metabolic p	229	The chemical reactions and pathways involving various organic and inorganic nitrogenous compounds, as carried out by individual cells. [GOC:mah]
biological_process	GO:0044281	small molecule metabolic process	188	The chemical reactions and pathways involving small molecules, any low molecular weight, monomeric, non-encoded molecule. [GOC:curators, GOC:pde, GOC:vw]
biological_process	GO:0006810	transport	147	The directed movement of substances (such as macromolecules, small molecules, ions) or cellular components (such as complexes and organelles) into, out of or within a cell, or between cells, or within a mul
biological_process	GO:0006412	translation	93	The cellular metabolic process in which a protein is formed, using the sequence of a mature mRNA molecule to specify the sequence of amino acids in a polypeptide chain. Translation is mediated by the ribos
biological_process	GO:0009056	catabolic process	91	The chemical reactions and pathways resulting in the breakdown of substances, including the breakdown of carbon compounds with the liberation of energy for use by the cell or organism. [ISBN:019854768
biological_process	GO:0022607	cellular component assembly	84	The aggregation, arrangement and bonding together of a cellular component. [GOC:isa_complete]
biological_process	GO:0006950	response to stress	67	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a disturbance in organismal or cellular
biological_process	GO:0005975	carbohydrate metabolic process	61	The chemical reactions and pathways involving carbohydrates, any of a group of organic compounds based of the general formula Cx(H2O)y. Includes the formation of carbohydrate derivatives by the additio
biological_process	GO:0065003	macromolecular complex assembly	61	The aggregation, arrangement and bonding together of a set of macromolecules to form a complex. [GOC:jl]
biological_process	GO:0006091	generation of precursor metabolites and	58	The chemical reactions and pathways resulting in the formation of precursor metabolites, substances from which energy is derived, and any process involved in the liberation of energy from these substances.
biological_process	GO:0051186	cofactor metabolic process	54	The chemical reactions and pathways involving a cofactor, a substance that is required for the activity of an enzyme or other protein. Cofactors may be inorganic, such as the metal atoms zinc, iron, and coppe

GO_barchart.png (according to GO_mapping.txt)



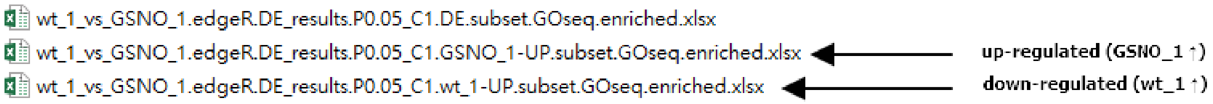
[GO enrichment basic analysis]

One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set.

In go enrichment analysis, we are using “GOseq v3.6” to finished this work. [10]

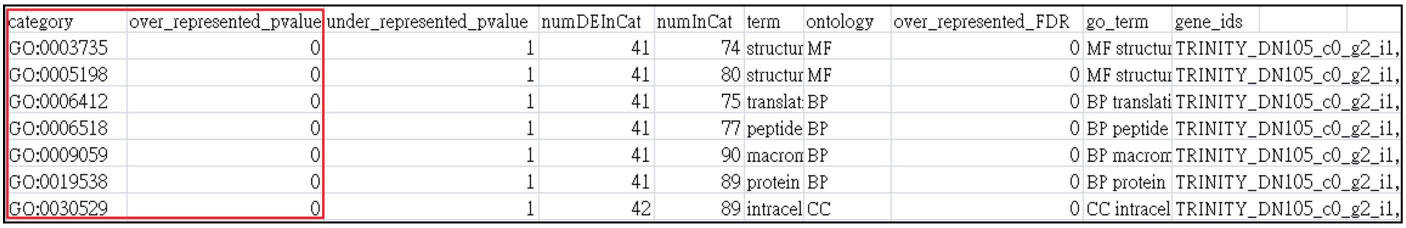
[GO enrichment dataset]:

e.g. <wt_1> v.s. <GSNO_1>: wt_1 is control & GSNO_1 is treatment



wt_1_vs_GSNO_1.edgeR.DE_results.P0.05_C1.DE.subset.GOseq.enriched.xlsx
wt_1_vs_GSNO_1.edgeR.DE_results.P0.05_C1.GSNO_1-UP.subset.GOseq.enriched.xlsx ← up-regulated (GSNO_1 ↑)
wt_1_vs_GSNO_1.edgeR.DE_results.P0.05_C1.wt_1-UP.subset.GOseq.enriched.xlsx ← down-regulated (wt_1 ↑)

[GSNO_1.UP.subset.GOseq.enrichment]

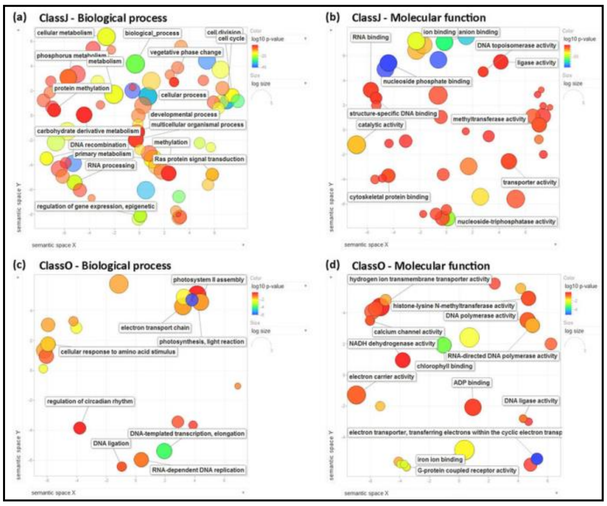


category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontology	over_represented_FDR	go_term	gene_ids
GO:0003735	0	1	41	74	structur	MF	0	MF structur	TRINITY_DN105_c0_g2_i1,
GO:0005198	0	1	41	80	structur	MF	0	MF structur	TRINITY_DN105_c0_g2_i1,
GO:0006412	0	1	41	75	translat	BP	0	BP translati	TRINITY_DN105_c0_g2_i1,
GO:0006518	0	1	41	77	peptide	BP	0	BP peptide	TRINITY_DN105_c0_g2_i1,
GO:0009059	0	1	41	90	macrom	BP	0	BP macrom	TRINITY_DN105_c0_g2_i1,
GO:0019538	0	1	41	89	protein	BP	0	BP protein	TRINITY_DN105_c0_g2_i1,
GO:0030529	0	1	42	89	intracel	CC	0	CC intracel	TRINITY_DN105_c0_g2_i1,

Note:

- **Over-represented (enrichment)**: lots of transcripts support certain GO term.
- **Under-represented (depletion)**: few of transcripts could be found in certain GO term.
- **NumDEInCat**: number of searched DE transcripts matched with the GO term.
- **NumInCat**: number of total transcripts existed in the GO term.

If user would like to be more visualized your Gene Ontology terms which are derived from gene enrichment analysis, we recommend you this online tool – **REVIGO!** (http://revigo.irb.hr) [11]
You just need to copy red square columns like above mentioned (“**category**” and “**over_represented_pvale**”).



Reference graph:

Forestan C, Aiese Cigliano R, Farinati S, Lunardon A, Sanseverino W, Varotto S. Stress-induced and epigenetic-mediated maize transcriptome regulation study by means of transcriptome reannotation and differential expression analysis. *Scientific Reports*. 2016;6:30446. doi:10.1038/srep30446.

[KEGG pathway annotation]:

- Method_A: Transcript pathway annotate by EC number

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:

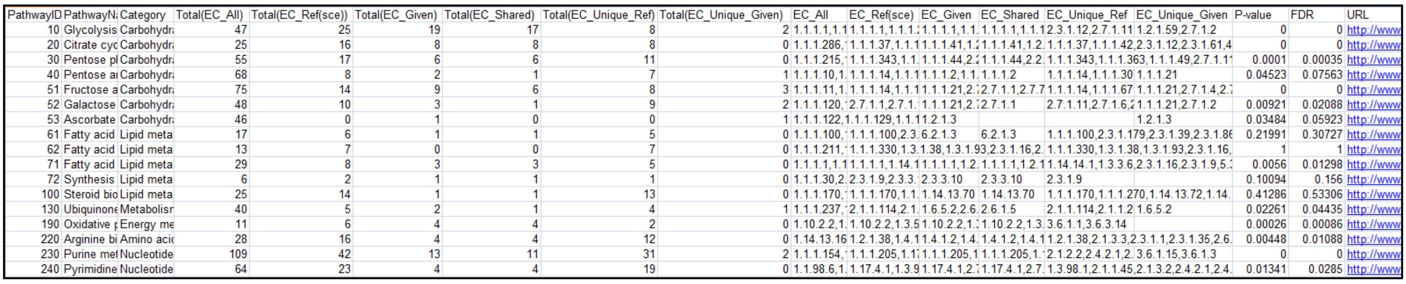
1. Metabolism

   Global/overview, Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Other amino, Glycan, Cofactor/vitamin, Terpenoid/PK, Other secondary metabolite, Xenobiotics, Chemical structure

2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development

One of our pathway results is generated from **<u>ec number (enzyme)</u>** data.

ec2kegg.xls

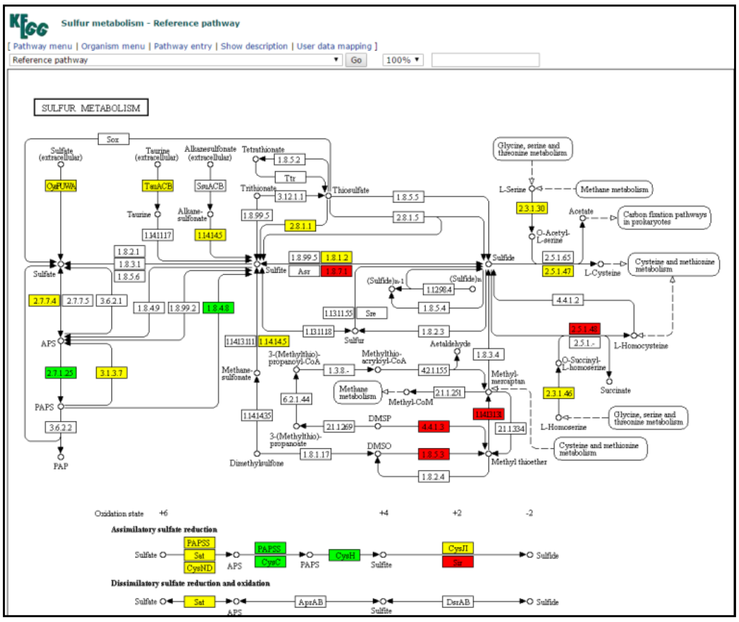


PathwayID	PathwayN	Category	Total(EC_All)	Total(EC_Ref(sce))	Total(EC_Given)	Total(EC_Shared)	Total(EC_Unique_Ref)	Total(EC_Unique_Given)	EC_All	EC_Ref(sce)	EC_Given	EC_Shared	EC_Unique_Ref	EC_Unique_Given	P-value	FDR	URL
10	Glycolysis	Carbohydr	47	25	19	17	8	2	1.1.1.1,1.1	1.1.1.1,1.1.1.	1.1.1.1,1.1.	1.1.1.1,1.1.1	2.3.1.12,2.7.1.11	1.2.1.59,2.7.1.2	0	0	http://www
20	Citrate cyc	Carbohydr	25	16	8	8	8	0	1.1.1.286,	1.1.1.37,1.1.1	1.1.1.41,1.2	1.1.1.41,1.2.	1.1.1.37,1.1.1.42,2.3.1.12,2.3.1.61,4		0	0	http://www
30	Pentose pl	Carbohydr	55	17	6	6	11	0	1.1.1.215,	1.1.1.343,1.1.	1.1.1.44,2.2	1.1.1.44,2.2.	1.1.1.343,1.1.1.363,1.1.1.49,2.7.1.11		0.0001	0.00035	http://www
40	Pentose a	Carbohydr	68	8	2	1	7	1	1.1.1.10,1.	1.1.1.14,1.1.1	1.1.1.2,1.1.	1.1.1.2	1.1.1.14,1.1.1.30	1.1.1.21	0.04523	0.07563	http://www
51	Fructose a	Carbohydr	75	14	9	6	8	3	1.1.1.11,1.	1.1.1.14,1.1.1	1.1.1.21,2.7	2.7.1.1,2.7.7	1.1.1.14,1.1.1.67	1.1.1.21,2.7.1.4,2.7	0	0	http://www
52	Galactose	Carbohydr	48	10	3	1	9	2	1.1.1.120,	2.7.1.1,2.7.1.	1.1.1.21,2.7	2.7.1.1	2.7.1.11,2.7.1.6,2	1.1.1.21,2.7.1.2	0.00921	0.02088	http://www
53	Ascorbate	Carbohydr	46	0	1	0	0	1	1.1.1.122,1.1.1.129,1.1.1		1.2.1.3			1.2.1.3	0.03484	0.05923	http://www
61	Fatty acid	Lipid meta	17	6	1	1	5	0	1.1.1.100,	1.1.1.100,2.3.	6.2.1.3	6.2.1.3	1.1.1.100,2.3.1.179,2.3.1.39,2.3.1.86		0.21991	0.30727	http://www
62	Fatty acid	Lipid meta	13	7	0	0	7	0	1.1.1.211,	1.1.1.330,1.3.1.38,1.3.1.93,2.3.1.16,2.			1.1.1.330,1.3.1.38,1.3.1.93,2.3.1.16,		1	1	http://www
71	Fatty acid	Lipid meta	29	8	3	3	5	0	1.1.1.1,1.1	1.1.1.1,1.14.1	1.1.1.1,1.2.	1.1.1.1,1.2.1	1.14.14.1,1.3.3.6,2.3.1.16,2.3.1.9,5.3		0.0056	0.01298	http://www
72	Synthesis	Lipid meta	6	2	1	1	1	0	1.1.1.30,2.	2.3.1.9,2.3.3.	2.3.3.10	2.3.3.10	2.3.1.9		0.10094	0.156	http://www
100	Steroid bio	Lipid meta	25	14	1	1	13	0	1.1.1.170,	1.1.1.170,1.1.	1.14.13.70	1.14.13.70	1.1.1.170,1.1.1.270,1.14.13.72,1.14.		0.41286	0.53306	http://www
130	Ubiquinone	Metabolism	40	5	2	1	4	1	1.1.1.237,	2.1.1.114,2.1.	1.6.5.2,2.6.	2.6.1.5	2.1.1.114,2.1.1.2	1.6.5.2	0.02261	0.04435	http://www
190	Oxidative p	Energy me	11	6	4	4	2	0	1.10.2.2,1.	1.10.2.2,1.3.5	1.10.2.2,1.3	1.10.2.2,1.3.	3.6.1.1,3.6.3.14		0.00026	0.00086	http://www
220	Arginine bi	Amino acid	28	16	4	4	12	0	1.14.13.16	1.2.1.38,1.4.1	1.4.1.2,1.4.	1.4.1.2,1.4.1	1.2.1.38,2.1.3.3,2.3.1.1,2.3.1.35,2.6.		0.00448	0.01088	http://www
230	Purine met	Nucleotide	109	42	13	11	31	2	1.1.1.154,	1.1.1.205,1.1	1.1.1.205,1	1.1.1.205,1.	2.1.2.2,2.4.2.1,2.	3.6.1.15,3.6.1.3	0	0	http://www
240	Pyrimidine	Nucleotide	64	23	4	4	19	0	1.1.98.6,1.	1.17.4.1,1.3.9	1.17.4.1,2.7	1.17.4.1,2.7.	1.3.98.1,2.1.1.45,2.1.3.2,2.4.2.1,2.4.		0.01341	0.0285	http://www

[**Column definition**]

- Total(EC_All) = number of ECs associated with the KEGG pathway;
- Total(EC_Ref(ead)) = number of ECs in reference genome ead (*E. adhaerens OV14*) associated with the KEGG pathway;
- Total(EC_Given) = number of tested ECs found to be associated with the KEGG pathway;
- Total(EC_Shared) = number of tested ECs that are shared with reference genome;
- Total(EC_Unique_Ref) = number of ECs that are unique to the reference genome;
- Total(EC_Unique_Given) = number of ECs that are unique to the tested genome.

Click URL and get the pathway information

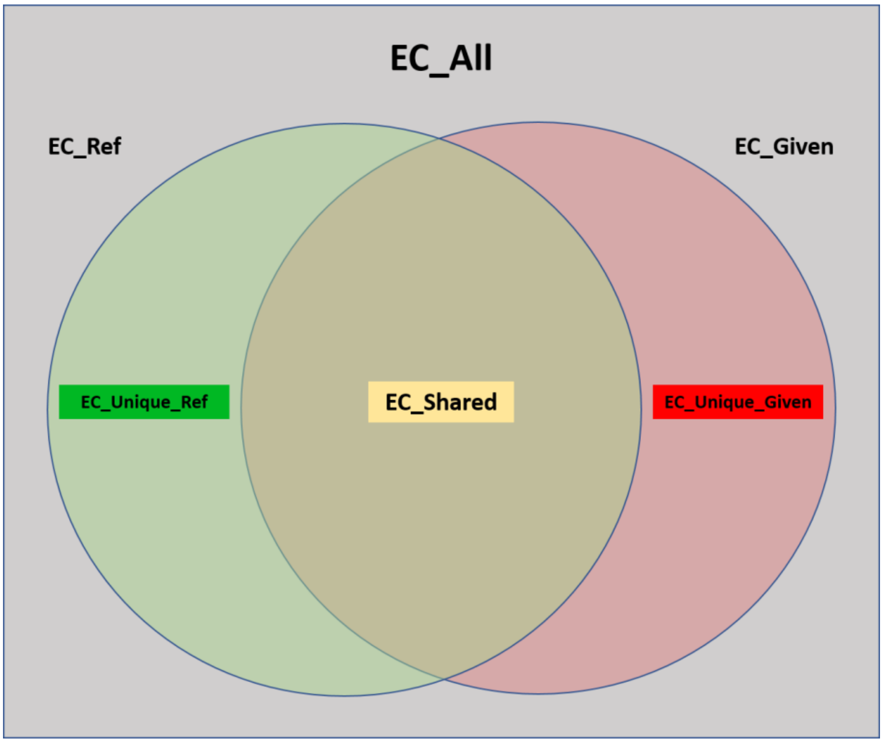

Sulfur metabolism - Reference pathway
[Pathway menu | Organism menu | Pathway entry | Show description | User data mapping]
Reference pathway
Go
100%
SULFUR METABOLISM
Assimilatory sulfate reduction
Dissimilatory sulfate reduction and oxidation

[**Pathway map color definition**]

green – an enzyme unique to a reference organism (EC_Unique_Ref)

red – an enzyme unique to a given list, (EC_Unique_Given) ← important if you would like to search novel enzymes which are not shown in ref!

yellow – a shared enzyme. (EC_Shared) ← important, shown that the searched enzyme intersection between your sample & ref!



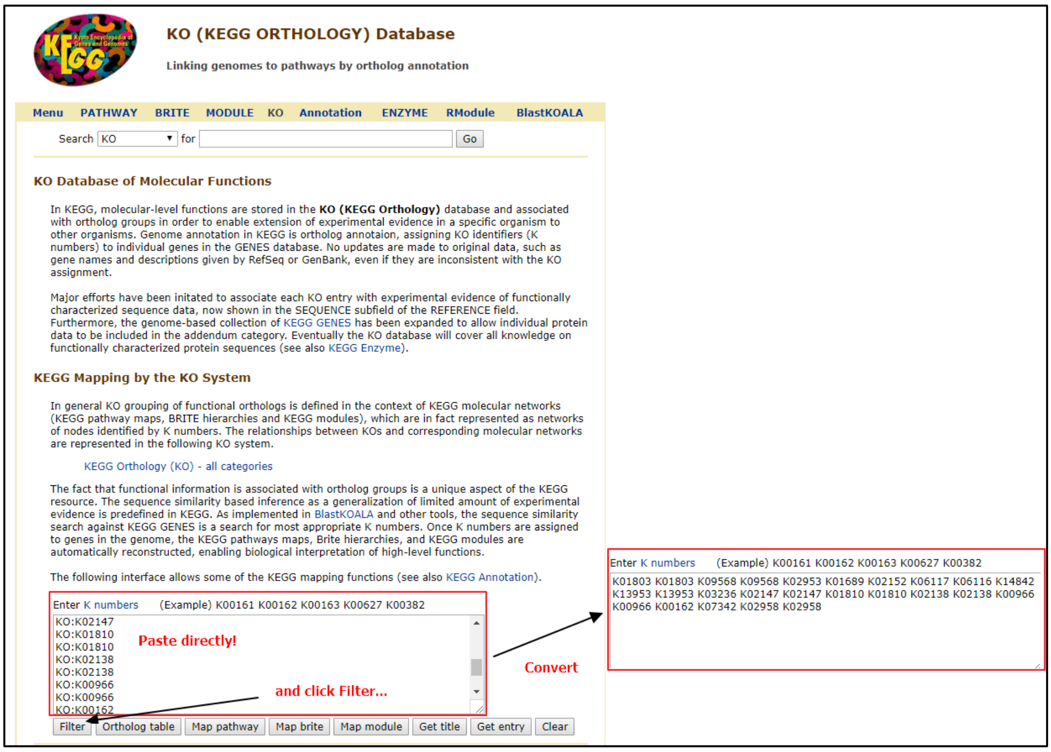
- **Method_B: Transcript pathway annotate by KO terms**

Another of our pathway results is generated from **KO terms (KEGG Orthology)** data. Genome annotation in KEGG is ortholog annotation, assigning KO identifiers (K numbers) to individual genes in the GENES database. All of the annotated KO terms are put together with the final_report.xlsx.

KO database is larger than EC number. So generally, using KO to search could be found a bit more detail than EC for digging pathway.

We recommend user could utilize them by following steps:

1. Copy targeted KO terms from final_report.xlsx
2. Go to KEGG pathway by KO annotation: http://www.genome.jp/kegg/ko.html
3. Paste targeted KO terms and convert



KO (KEGG ORTHOLOGY) Database

Linking genomes to pathways by ortholog annotation

KO Database of Molecular Functions

In KEGG, molecular-level functions are stored in the **KO (KEGG Orthology)** database and associated with ortholog groups in order to enable extension of experimental evidence in a specific organism to other organisms. Genome annotation in KEGG is ortholog annotaion, assigning KO identifiers (K numbers) to individual genes in the GENES database. No updates are made to original data, such as gene names and descriptions given by RefSeq or GenBank, even if they are inconsistent with the KO assignment.

Major efforts have been initated to associate each KO entry with experimental evidence of functionally characterized sequence data, now shown in the SEQUENCE subfield of the REFERENCE field. Furthermore, the genome-based collection of KEGG GENES has been expanded to allow individual protein data to be included in the addendum category. Eventually the KO database will cover all knowledge on functionally characterized protein sequences (see also KEGG Enzyme).

KEGG Mapping by the KO System

In general KO grouping of functional orthologs is defined in the context of KEGG molecular networks (KEGG pathway maps, BRITE hierarchies and KEGG modules), which are in fact represented as networks of nodes identified by K numbers. The relationships between KOs and corresponding molecular networks are represented in the following KO system.

KEGG Orthology (KO) - all categories

The fact that functional information is associated with ortholog groups is a unique aspect of the KEGG resource. The sequence similarity based inference as a generalization of limited amount of experimental evidence is predefined in KEGG. As implemented in BlastKOALA and other tools, the sequence similarity search against KEGG GENES is a search for most appropriate K numbers. Once K numbers are assigned to genes in the genome, the KEGG pathways maps, Brite hierarchies, and KEGG modules are automatically reconstructed, enabling biological interpretation of high-level functions.

The following interface allows some of the KEGG mapping functions (see also KEGG Annotation).

Enter K numbers (Example) K00161 K00162 K00163 K00627 K00382

KO:K02147
KO:K01810
KO:K01810
KO:K02138
KO:K02138
KO:K00966
KO:K00966
KO:K00162

Paste directly!

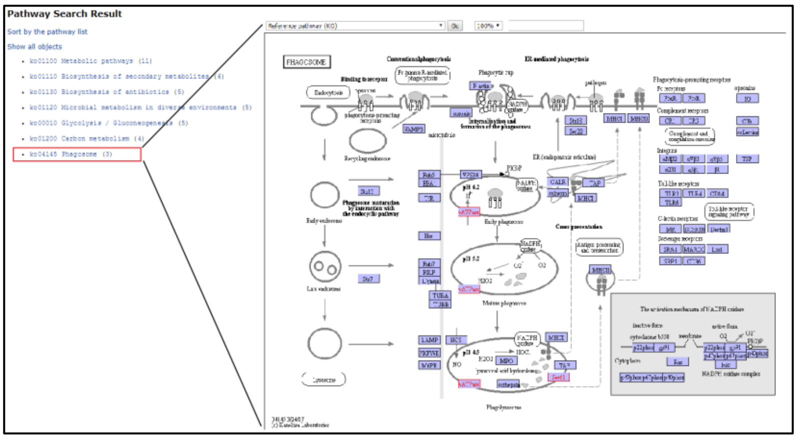
and click Filter...

Convert

Enter K numbers (Example) K00161 K00162 K00163 K00627 K00382

K01803 K01803 K09568 K09568 K02953 K01689 K02152 K06117 K06116 K14842 K13953 K13953 K03236 K02147 K02147 K01810 K01810 K02138 K02138 K00966 K00966 K00162 K07342 K02958 K02958

4. Get the pathway



Pathway Search Result

Sort by the pathway list

Show all objects

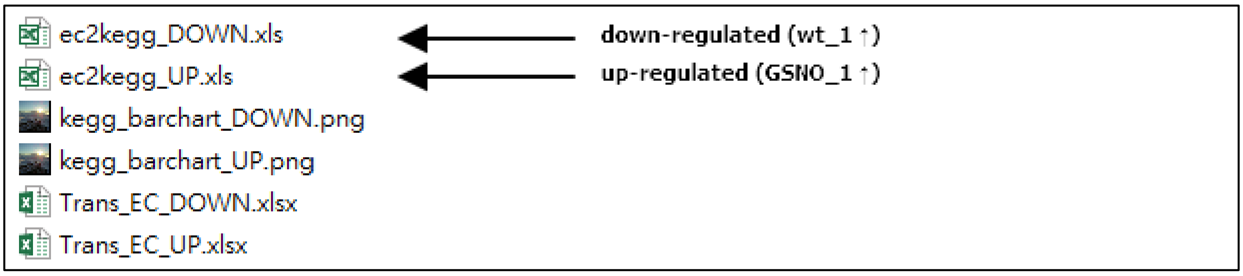
- ko01100 Metabolic pathways (11)
- ko01110 Biosynthesis of secondary metabolites (6)
- ko01130 Biosynthesis of antibiotics (5)
- ko01120 Microbial metabolism in diverse environments (5)
- ko00010 Glycolysis / Gluconeogenesis (5)
- ko01200 Carbon metabolism (4)
- ko04145 Phagosome (3)

- **KEGG pathway classified by up/down regulation**

Default, our EC number KEGG pathways are according to all of the mapped transcripts. But for user-friendly concern, custom might want to get specific up- or down-regulated pathway either. Thus, we also tried to parse this data by in-house script for you.

[KEGG up/down classified dataset]:

e.g. <wt_1> v.s. <GSNO_1>: wt_1 is control & GSNO_1 is treatment



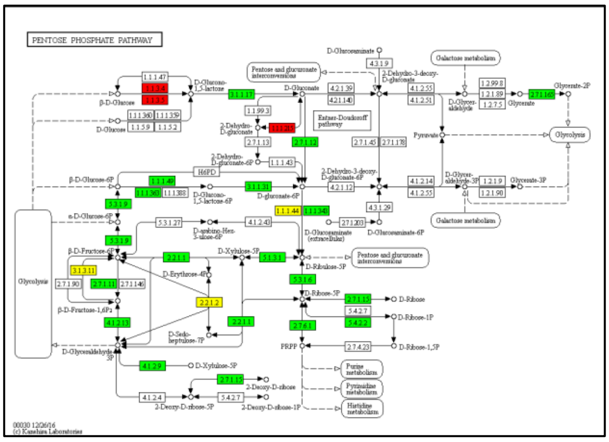
ec2kegg_DOWN.xls ← down-regulated (wt_1 ↑)
ec2kegg_UP.xls ← up-regulated (GSNO_1 ↑)
kegg_barchart_DOWN.png
kegg_barchart_UP.png
Trans_EC_DOWN.xlsx
Trans_EC_UP.xlsx

<u>Trans_EC_DOWN</u> contains all of **<u>down-regulated</u>** mapped EC number transcripts.

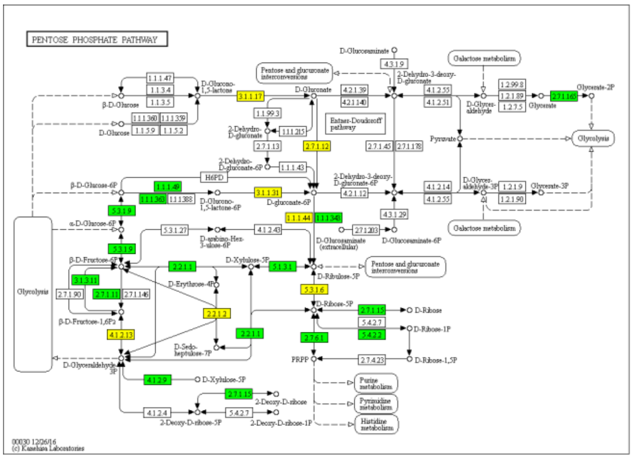
<u>Trans_EC_UP</u> contains all of **<u>up-regulated</u>** mapped EC number transcripts.

ec2kegg_DOWN and **ec2kegg_UP** are the corresponded pathway table

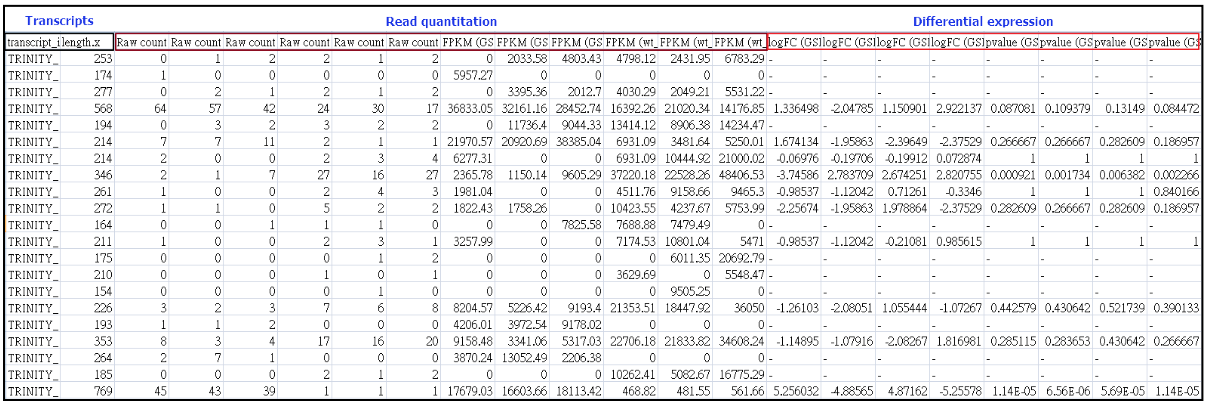
ec2kegg_UP (found 3 unique given enzyme)



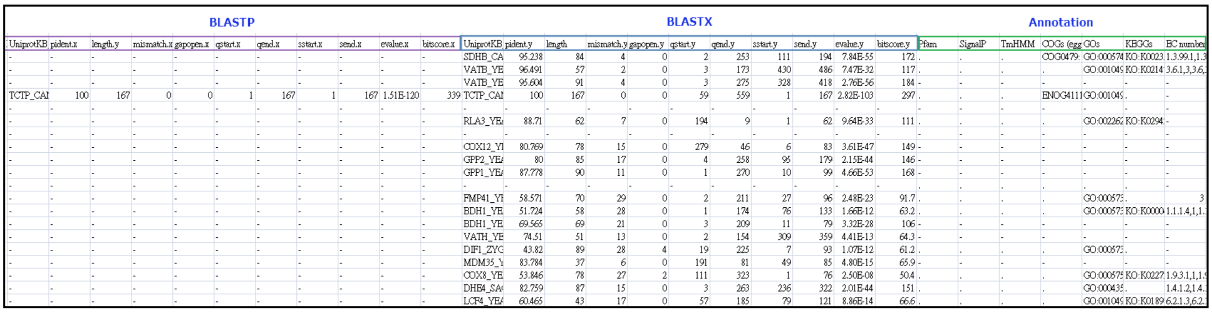
ec2kegg_DOWN (not found unique given)



***** All of the data including ‘transcript ID’, ‘read quantification’, ‘differential expression’ and functional annotation report is merged in “final_report.xlsx” *****



Transcripts		Read quantitation												Differential expression							
transcript_i	length.x	Raw count	Raw count	Raw count	Raw count	Raw count	Raw count	FPKM (GS	FPKM (GS	FPKM (GS	FPKM (wt_	FPKM (wt_	FPKM (wt_	logFC (GS	logFC (GS	logFC (GS	logFC (GS	pvalue (GS	pvalue (GS	pvalue (GS	pvalue (GS
TRINITY_	253	0	1	2	2	1	2	0	2033.58	4803.43	4798.12	2431.95	6783.29	-	-	-	-	-	-	-	-
TRINITY_	174	1	0	0	0	0	0	5957.27	0	0	0	0	0	-	-	-	-	-	-	-	-
TRINITY_	277	0	2	1	2	1	2	0	3395.36	2012.7	4030.29	2049.21	5531.22	-	-	-	-	-	-	-	-
TRINITY_	568	64	57	42	24	30	17	36833.05	32161.16	28452.74	16392.26	21020.34	14176.85	1.336498	-2.04785	1.150901	2.922137	0.087081	0.109379	0.13149	0.084472
TRINITY_	194	0	3	2	3	2	2	0	11736.4	9044.33	13414.12	8906.38	14234.47	-	-	-	-	-	-	-	-
TRINITY_	214	7	7	11	2	1	1	21970.57	20920.69	38385.04	6931.09	3481.64	5250.01	1.674134	-1.95863	-2.39649	-2.37529	0.266667	0.266667	0.282609	0.186957
TRINITY_	214	2	0	0	2	3	4	6277.31	0	0	6931.09	10444.92	21000.02	-0.06976	-0.19706	-0.19912	0.072874	1	1	1	1
TRINITY_	346	2	1	7	27	16	27	2365.78	1150.14	9605.29	37220.18	22528.26	48406.53	-3.74586	2.783709	2.674251	2.820755	0.000921	0.001734	0.006382	0.002266
TRINITY_	261	1	0	0	2	4	3	1981.04	0	0	4511.76	9158.66	9465.3	-0.98537	-1.12042	0.71261	-0.3346	1	1	1	0.840166
TRINITY_	272	1	1	0	5	2	2	1822.43	1758.26	0	10423.55	4237.67	5753.99	-2.25674	-1.95863	1.978864	-2.37529	0.282609	0.266667	0.282609	0.186957
TRINITY_	164	0	0	1	1	1	0	0	0	7825.58	7688.88	7479.49	0	-	-	-	-	-	-	-	-
TRINITY_	211	1	0	0	2	3	1	3257.99	0	0	7174.53	10801.04	5471	-0.98537	-1.12042	-0.21081	0.985615	1	1	1	1
TRINITY_	175	0	0	0	0	1	2	0	0	0	0	6011.35	20692.79	-	-	-	-	-	-	-	-
TRINITY_	210	0	0	0	1	0	1	0	0	0	3629.69	0	5548.47	-	-	-	-	-	-	-	-
TRINITY_	154	0	0	0	0	1	0	0	0	0	0	9505.25	0	-	-	-	-	-	-	-	-
TRINITY_	226	3	2	3	7	6	8	8204.57	5226.42	9193.4	21353.51	18447.92	36050	-1.26103	-2.08051	1.055444	-1.07267	0.442579	0.430642	0.521739	0.390133
TRINITY_	193	1	1	2	0	0	0	4206.01	3972.54	9178.02	0	0	0	-	-	-	-	-	-	-	-
TRINITY_	353	8	3	4	17	16	20	9158.48	3341.06	5317.03	22706.18	21833.82	34608.24	-1.14895	-1.07916	-2.08267	1.816981	0.285115	0.283653	0.430642	0.266667
TRINITY_	264	2	7	1	0	0	0	3870.24	13052.49	2206.38	0	0	0	-	-	-	-	-	-	-	-
TRINITY_	185	0	0	0	2	1	2	0	0	0	10262.41	5082.67	16775.29	-	-	-	-	-	-	-	-
TRINITY_	769	45	43	39	1	1	1	17679.03	16603.66	18113.42	468.82	481.55	561.66	5.256032	-4.88565	4.87162	-5.25578	1.14E-05	6.56E-06	5.69E-05	1.14E-05



BLASTP											BLASTX											Annotation						
UniprotKB	pident.x	length.y	mismatch.x	gapopen.x	qstart.x	qend.x	sstart.x	send.x	evalue.x	bitscore.x	UniprotKB	pident.y	length	mismatch.y	gapopen.y	qstart.y	qend.y	sstart.y	send.y	evalue.y	bitscore.y	Pfam	SignalP	TmHMM	COGs (egg	GOs	KEGGs	EC number
-	-	-	-	-	-	-	-	-	-	-	SDHB_CA	95.238	84	4	0	2	253	111	194	7.84E-55	172	.	.	.	COG0479:	GO:000574	KO:K0023	1.3.99.1,1.3
-	-	-	-	-	-	-	-	-	-	-	VATB_YE	96.491	57	2	0	3	173	430	486	7.47E-32	117	.	.	.	.	GO:001049	KO:K0214	3.6.1,3,3.6,
-	-	-	-	-	-	-	-	-	-	-	VATB_YE	95.604	91	4	0	3	275	328	418	2.76E-56	184	-	-	-	-	-	-	-
TCTP_CAI	100	167	0	0	1	167	1	167	1.51E-120	339	TCTP_CAI	100	167	0	0	59	559	1	167	2.82E-103	297	.	.	.	ENOG4111	GO:001049	.	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	RLA3_YE	88.71	62	7	0	194	9	1	62	9.64E-33	111	.	.	.	.	GO:002262	KO:K0294	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	COX12_YI	80.769	78	15	0	279	46	6	83	3.61E-47	149	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	GPP2_YE	80	85	17	0	4	258	95	179	2.15E-44	146	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	GPP1_YE	87.778	90	11	0	1	270	10	99	4.66E-53	168	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.	.	.	.	.	.	-
-	-	-	-	-	-	-	-	-	-	-	FMP41_YI	58.571	70	29	0	2	211	27	96	2.48E-23	91.7	.	.	.	.	GO:000573	.	3
-	-	-	-	-	-	-	-	-	-	-	BDH1_YE	51.724	58	28	0	1	174	76	133	1.66E-12	63.2	.	.	.	.	GO:000573	KO:K0000	1.1.1.4,1,1.
-	-	-	-	-	-	-	-	-	-	-	BDH1_YE	69.565	69	21	0	3	209	11	79	3.32E-28	106	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	VATH_YE	74.51	51	13	0	2	154	309	359	4.41E-13	64.3	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	DIF1_ZYG	43.82	89	28	4	19	225	7	93	1.07E-12	61.2	.	.	.	.	GO:000573	.	-
-	-	-	-	-	-	-	-	-	-	-	MDM35_Y	83.784	37	6	0	191	81	49	85	4.80E-15	65.9	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	COX8_YE	53.846	78	27	2	111	323	1	76	2.50E-08	50.4	.	.	.	.	GO:000575	KO:K0227	1.9.3.1,1,1.
-	-	-	-	-	-	-	-	-	-	-	DHE4_SA	82.759	87	15	0	3	263	236	322	2.01E-44	151	.	.	.	.	GO:000435	.	1.4.1.2,1.4.
-	-	-	-	-	-	-	-	-	-	-	LCF4_YE	60.465	43	17	0	57	185	79	121	8.86E-14	66.6	.	.	.	.	GO:001049	KO:K0189	6.2.1.3,6.2.

5. Reference & Useful tools

1. MultiQC: Summarize analysis results for multiple tools and samples in a single report; Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller; Bioinformatics (2016); doi: 10.1093/bioinformatics/btw354; PMID: 27312411
2. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494-512. Open Access in PMC doi: 10.1038/nprot.2013.084.
3. Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu and Weizhong Li, CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics, (2012), 28 (23): 3150-3152. doi: 10.1093/bioinformatics/bts565.
4. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
5. Li, Bo & Dewey, Colin N. (2011). *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome*, BioMed Central
6. Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). “Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.” Bioinformatics, btv566
7. McCarthy, J. D, Chen, Yunshun, Smyth and K. G (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), pp. -9.
8. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature protocols. 2013;8(8):10.1038/nprot.2013.084. doi:10.1038/nprot.2013.084.
9. https://trinotate.github.io/
10. Gene ontology analysis for RNA-seq: accounting for selection bias Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, Alicia Oshlack Genome Biology 2010, 11:R14 (4 February 2010)
11. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800 (2011).

- Useful tools:

- **Comma separator:** https://delim.co/ (分隔符號轉換行)
- **Venny diagram:** http://bioinfogp.cnb.csic.es/tools/venny/ (例如：GO enrichment BP 上下調取交集)
- **NaviGO:** http://kiharalab.org/web/navigo/views/goset.php (搜尋這些具有上下調功能的 GO 的關聯性並繪製網絡圖)
- **ClustVis:** http://biit.cs.ut.ee/clustvis/ (客製化 heatmap & PCA 繪圖網站)
- **Uniprot database:** http://www.uniprot.org/ (世界三大基因/蛋白質資料庫)
- **Uniprot ID mapping:** http://www.uniprot.org/mapping/ (Transform Uniprot gene ID to what you want)
- **KEGG ko database:** http://www.genome.jp/kegg/ko.html (使用篩選過的 KO 來搜尋 pathway)
- **KEGG mapping:** http://www.genome.jp/kegg/tool/map_pathway1.html (透過所提供的 enzyme 或 KO 來搜尋資料庫當中已註解的 pathway)