Genomics NGS Service

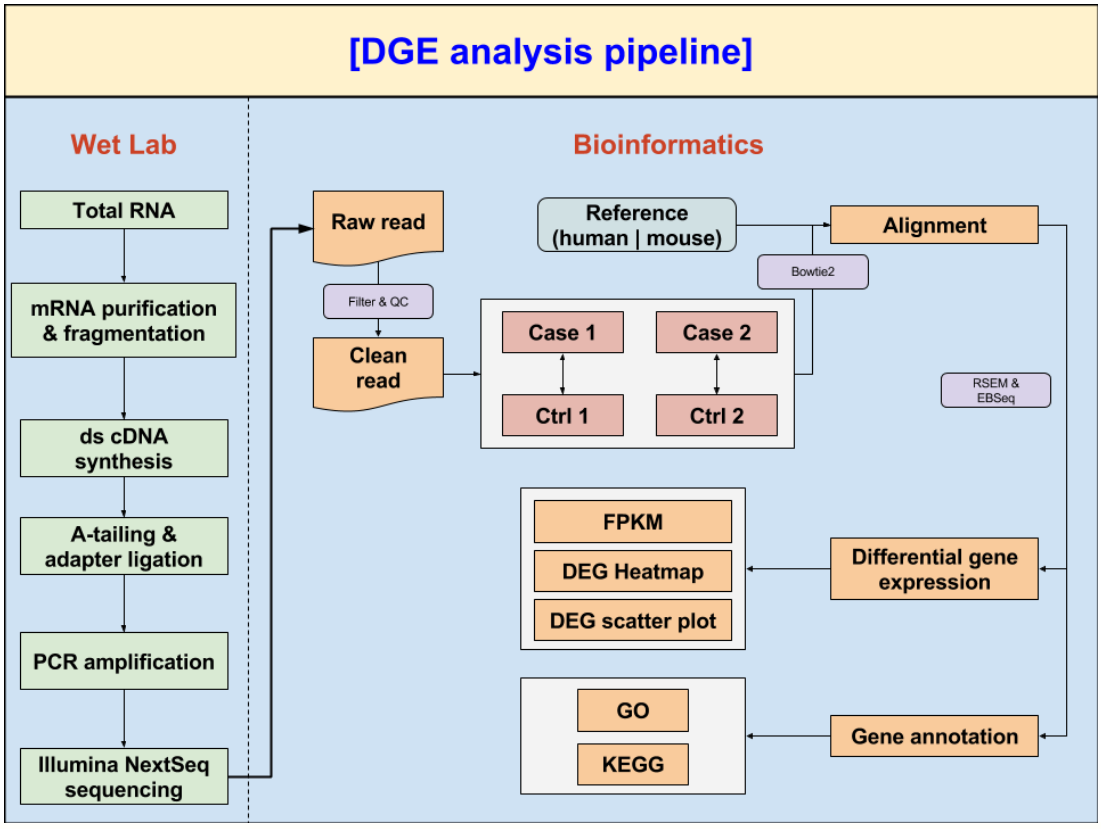
Bioinformatics Analysis of Differential Gene Expression Analysis

Help manual

2017

Genomics NGS Analysis Team

Table of Contents



Experiment Process

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

b.) First and second 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 NextSeq

Bioinformatics analysis

1. Raw data processing

Raw sequencing reads are generated by Illumina NextSeq. We used the following criteria to remove adapters and low quality bases:

- Remove raw reads with polluted-adapter.
- Trimming options (tool: “Trimmomatic v0.33”)
  - Trim off low quality end sequences by sliding windows (5 nt) with average quality value under 10.
  - Read length > 20 nt.
  - At least 55% of bases are Q20 above in both one pair reads.

* Fastq statistics (Fastq Stats)

Sample1 (treatment)	
	Sample1_R1
Reads	35,682,454
Len	76
Pct_dup	28.8936
QV_mean	34.4725
Total_base	2,684,509,315
Sample2 (control)	
	Sample2_R1
Reads	30,336,001
Len	76
Pct_dup	27.7512
QV_mean	34.3977
Total_base	2,281,051,610

Note:

- **Reads**: reads number in the fastq file.
- **Len**: read length
- **Pct_dup**: Pct reads that are duplcate
- **QV_mean**: Mean of QV
- **Total_base**: total number of bases

* Alignment statistics (Alignment Stats)

According to user-provided comparison table, we selected corresponded clean reads mapped to transcriptome by “bowtie2 v2.2.6” and the alignment data would be prepared for the following quantification stage.

Sample1 (treatment)	
Total Reads Pair	35,682,454
aligned 0 times	4,241,854
aligned exactly 1 time	12,510,109
aligned >1 times	18,930,491
overall alignment rate	88.11%
Sample2 (control)	
Total Reads Pair	30,336,001
aligned 0 times	3,759,278
aligned exactly 1 time	10,410,638
aligned >1 times	16,166,085
overall alignment rate	87.61%

Note:

- **Aligned 0 times**: reads not mapped.
- **Aligned exactly 1 time**: reads mapped on one site
- **Aligned > 1 time**: reads mapped on multiple sites (repeat region)
- **Overall alignment rate**: total alignment rate including mapping “1 time” & “over 1 times” reads

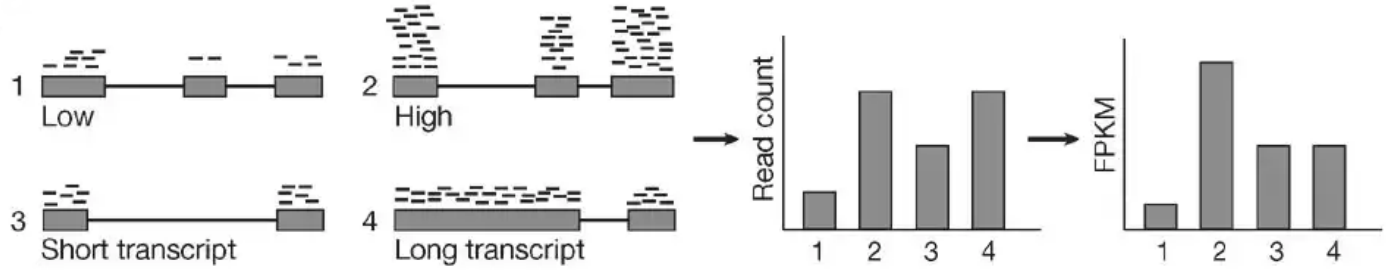
2. Calculate gene expression level

Once all of the RNA-seq reads mapped to reference transcriptome, we're using "RSEM (RNA-seq by Expectation Maximization)" for calculating read raw count and normalized quantification from each sample.

[**Gene Expression**]: raw count of mapped reads

[**FPKM**]: normalized count of mapped reads

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



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

3. DGE comparisons

As we got the read quantification data, we could continued various different comparisons for which user would like to look into. The statistic tool we selected is “EBSeq” which may be used to identify differential expressed gene and isoforms according to your given groups.

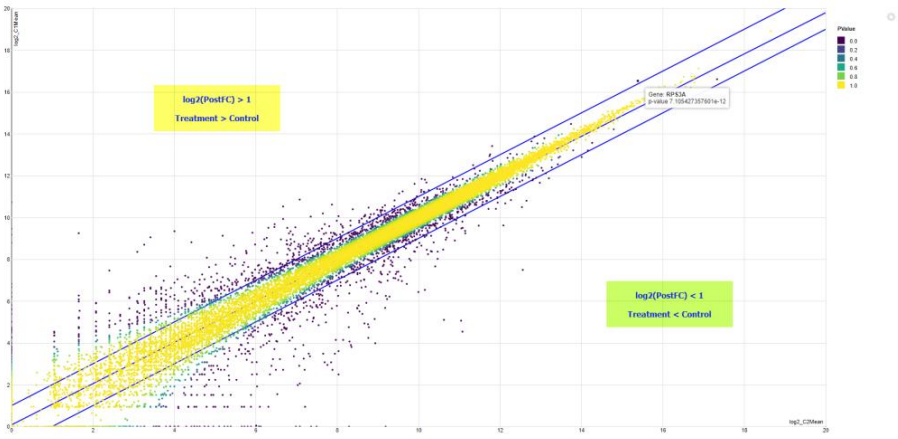
* Differential Gene Expression:

- “PPEE”: the posterior probability that a gene/transcript is **equally expressed**.
- “PPDE”: the posterior probability that a gene/transcript is **differential expressed**.
  - **Notice**: the values of PPEE and PPDE are calculated by statistical algorithm. The sum of the 2 values will be ‘1’. **Generally, you could just regard PPEE as pvalue.**
- “PostFC”: posterior fold change of condition 1 over condition 2
- “RealFC”: real fold change of condition 1 over condition 2
  - **Notice**: **PostFC is recommended over the RealFC** due to statistical concern.
- “C1Mean”: mean count of condition 1
- “C2Mean”: mean count of condition 2

Gene	PPEE	PPDE	PostFC	RealFC	C1Mean	C2Mean
ABCA13	0	1	6.790812	6.824241	520.9442	76.32877
ADGRL3	0	1	0.128795	0.128141	75.20716	586.9787
AHR	0	1	4.845737	4.869942	340.967	70.00663
AKAP9	0	1	0.297864	0.297601	350.9667	1179.343
AKR1C1	0	1	3.140768	3.141593	3591.652	1143.251
ALX4	0	1	3.248536	3.250803	1421.508	437.2722
ANGPTL2	0	1	0.146968	0.145187	30.63995	211.0969

* Scatter Plot

- X axis: C2 mean (control)
- Y axis: C1 mean (treatment)
- Dots above to upper blue line are the $\log_2$(PostFC) > 1
- Dots below to lower blue line are the $\log_2$(PostFC) < 1
- If dots are more darker, the pvalues are more lower.



4. Annotation

* GO/KEGG Plot

GO enrichment table

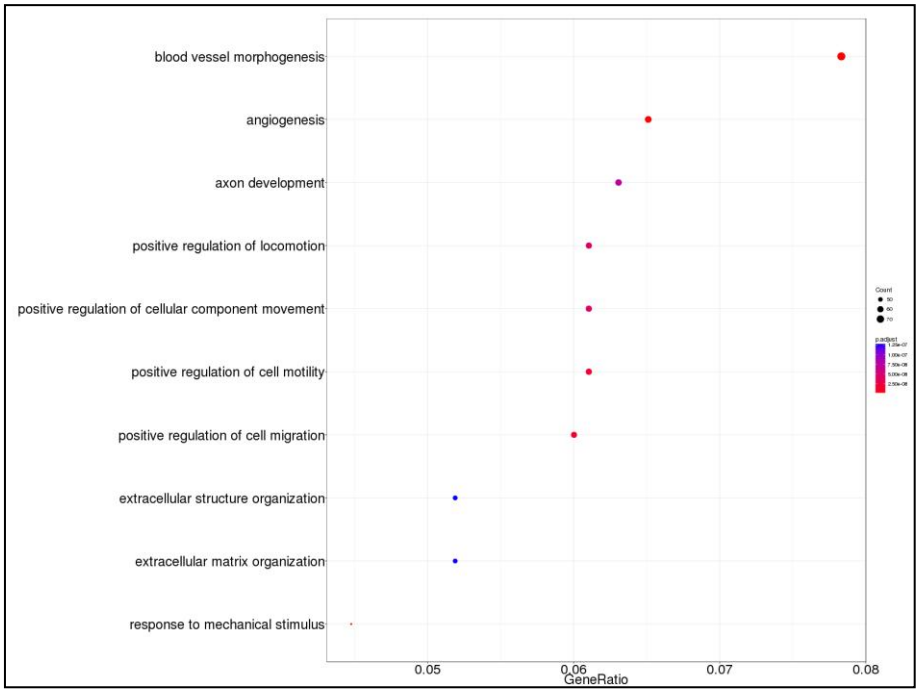
A set of genes which are up-regulated (down-regulated) under certain conditions, an enrichment analysis will find which GO terms are over-represented (under-represented) using annotations for that gene set.

For example of **biological process**:

- "GeneRatio": this bp GO term found in this case / total bp GO terms found in this case
- "BgRatio": this bp GO term count in whole database / total bp GO terms existed in whole database so far
- "p.adjust": calculate from GeneRatio and BgRatio

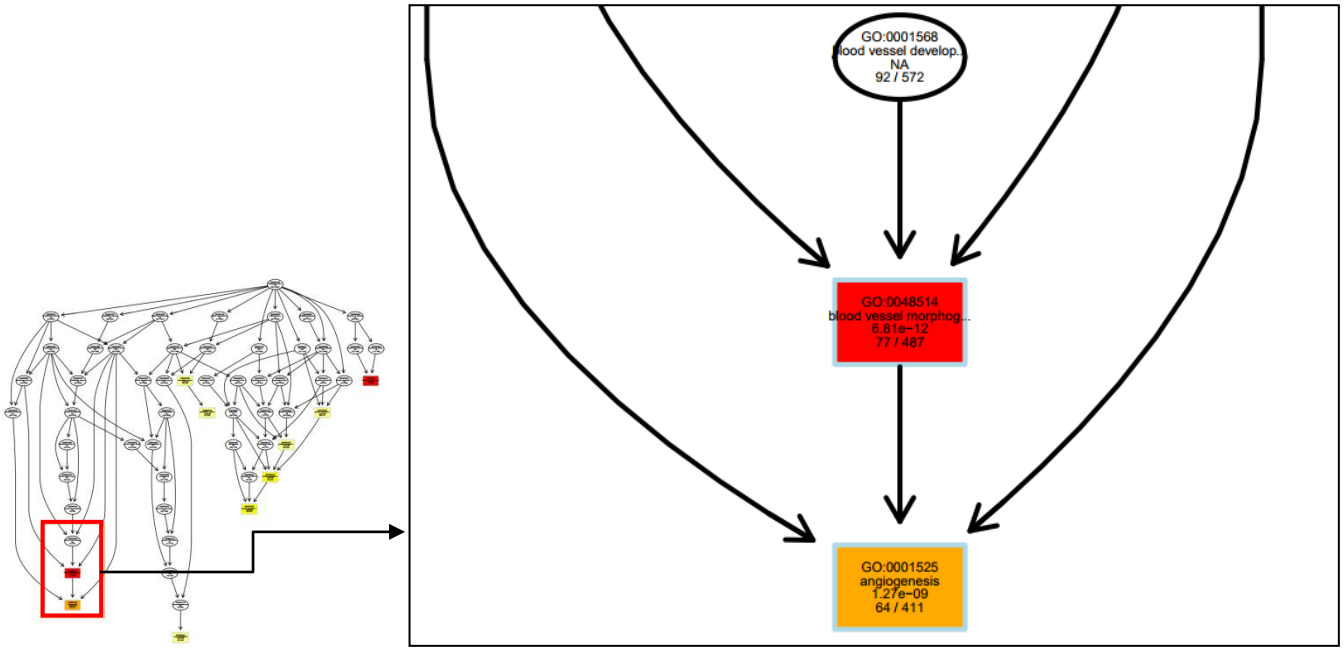
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0048514	blood vessel morphoger	77/983	487/16672	1.37E-15	6.81E-12	5.111E-12	CAV1/CCL2/CXCL8/ERBB2/F3/F(	77
GO:0009612	response to mechanical	44/983	193/16672	4.5E-15	1.116E-11	8.374E-12	CCL2/COL11A1/LTBR/POSTN/SE	44
GO:0001525	angiogenesis	64/983	411/16672	7.68E-13	1.269E-09	9.527E-10	CAV1/CCL2/CXCL8/ERBB2/F3/IS	64
GO:0030335	positive regulation of cel	59/983	386/16672	1.36E-11	1.686E-08	1.265E-08	ATP8A1/CCL2/CCL20/CXCL8/F3/	59
GO:2000147	positive regulation of cel	60/983	399/16672	1.81E-11	1.8E-08	1.351E-08	ATP8A1/CCL2/CCL20/CXCL8/F3/	60
GO:0051272	positive regulation of cel	60/983	409/16672	5.12E-11	4.23E-08	3.174E-08	ATP8A1/CCL2/CCL20/CXCL8/F3/	60
GO:0040017	positive regulation of loc	60/983	411/16672	6.26E-11	4.438E-08	3.33E-08	ATP8A1/CCL2/CCL20/CXCL8/F3/	60
GO:0061564	axon development	62/983	439/16672	1.21E-10	7.488E-08	5.619E-08	ARHGAP4/EGR2/EPHA7/ERBB2	62

Color means pvalue, and dot size means count of this GO (GeneRatio).



plotGOgraph (Graphical representation of GO)

According to all of the bp GO terms we found, a cause-effect relation could be gotten from the GO enrichment table and also generate network graph.



For example of Angiogenesis, we could find the cause-effect relation from up-stream to down-stream of GO data. The angiogenesis is the final result stage, it was related with the blood vessel morphogenesis, and also related with blood vessel development function.

Gene KEGG

Columns contain the same explanation with GO. Please refer to above GO enrichment table.

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
hsa04933	AGE-RAG	23/438	99/7281	1.27E-08	3.48E-06	2.66E-06	6347/1288	23
hsa04668	TNF signal	22/438	108/7281	3.21E-07	4.39E-05	3.36E-05	6347/6364	22
hsa05200	Pathways	47/438	395/7281	3.97E-06	0.000224	0.000171	675/1026/	47
hsa05144	Malaria	13/438	49/7281	4.04E-06	0.000224	0.000171	6347/3576	13
hsa04060	Cytokine-c	36/438	270/7281	4.5E-06	0.000224	0.000171	6347/6364	36
hsa04360	Axon guid	27/438	175/7281	4.9E-06	0.000224	0.000171	10682173(	27

5. Reference

1. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
2. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.
3. Li, Bo & Dewey, Colin N. (2011). *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome*, BioMed Central
4. Leng N and Kendziorski C (2015). *EBSeq: An R package for gene and isoform differential expression analysis of RNA-seq data*. R package version 1.16.0.
5. Yu G, Wang L, Han Y and He Q (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters.” *OMICS: A Journal of Integrative Biology*, **16**(5), pp. 284-287. doi: 10.1089/omi.2011.0118.