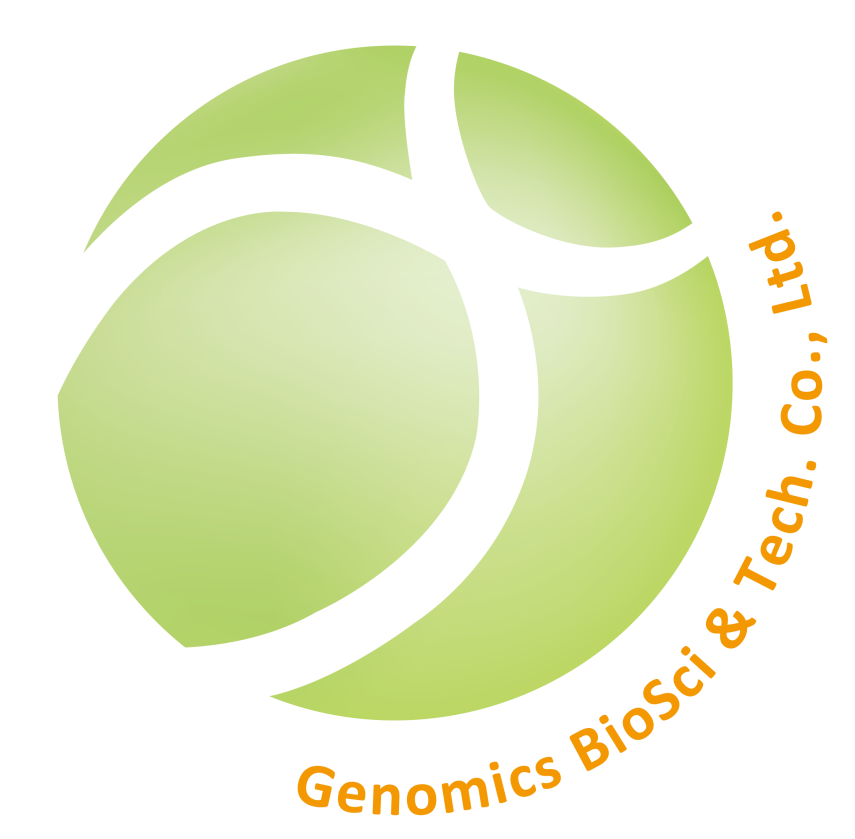
**Genomics NGS Service**

**Bioinformatics Analysis of   
Differential Gene Expression Analysis**

Help manual

2017

Genomics NGS Analysis Team



**Table of Contents**

[Experiment Process 3](#_Toc487041965)

[Bioinformatics analysis 4](#_Toc487041966)

[1. Raw data processing 4](#_Toc487041967)

[\* Fastq statistics (Fastq Stats) 4](#_Toc487041968)

[\* Alignment statistics (Alignment Stats) 5](#_Toc487041969)

[2. Calculate gene expression level 6](#_Toc487041970)

[3. DGE comparisons 7](#_Toc487041971)

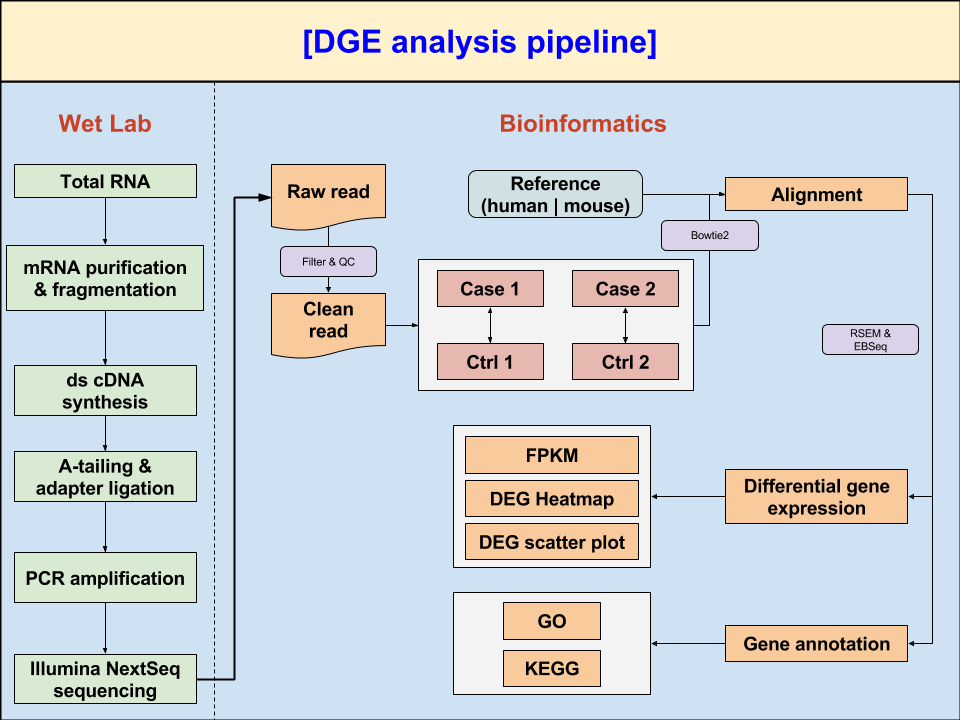
[\* Differential Gene Expression: 7](#_Toc487041972)

[\* Scatter Plot 7](#_Toc487041973)

[4. Annotation 8](#_Toc487041974)

[\* GO/KEGG Plot 8](#_Toc487041975)

[5. Reference 10](#_Toc487041976)

****

# Experiment Process

1. Purify and fragment mRNA: Using poly-T oligo-attached beads to purify mRNA, which is also fragmented primed for cDNA synthesis.
2. 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.
3. 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.
4. PCR amplification Using PCR to selectively amplify those DNA fragments that have adapters on both ends.
5. Library quality validating: Library was validated on Agilent 2100 Bio-analyzer and Real-Time PCR System.
6. Sequencing by Illumina NextSeq

# Bioinformatics analysis

## 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:**

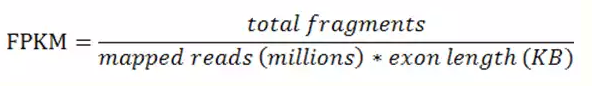
* **Aligned 0 times**: reads not mapped.
* **Aligned exactly 1 time**: reads mapped on one site
* **Aligned > 1 time**: reads mapped on multiple sites (repeat region)
* **Overall alignment rate**: total alignment rate including mapping “1 time” & “over 1 times” reads

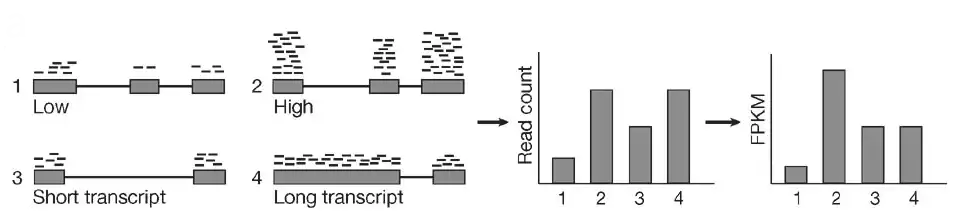
## Calculate gene expression level

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





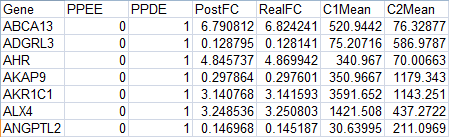
Ref: (http://dx.doi.org/10.1038/nmeth.1613)

## 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.

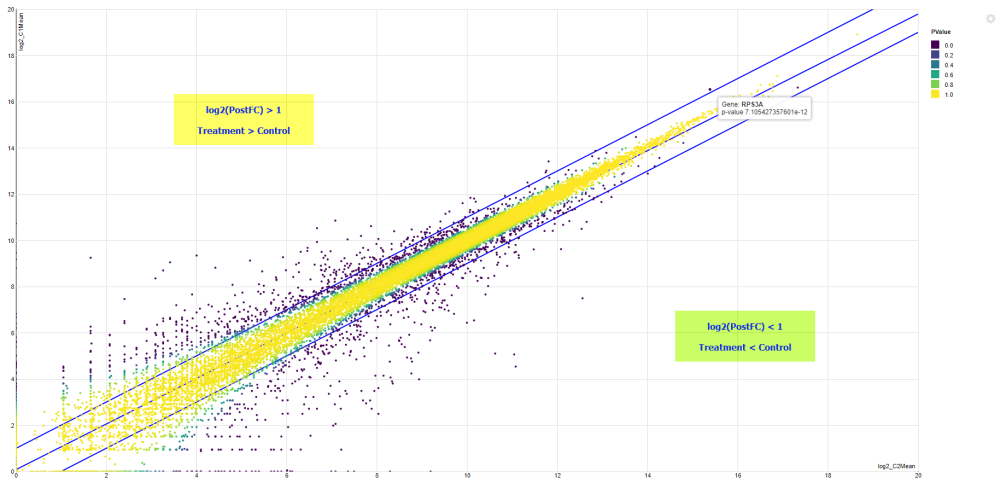
### \* Differential Gene Expression:

* “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



### \* Scatter Plot

* X axis: C2 mean (control)
* Y axis: C1 mean (treatment)
* Dots above to upper blue line are the log2(PostFC) > 1
* Dots below to lower blue line are the log2(PostFC) < 1
* If dots are more darker, the pvalues are more lower.



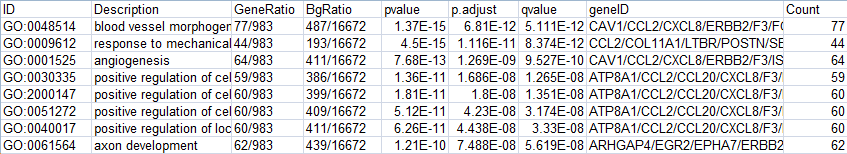
## Annotation

### \* GO/KEGG Plot

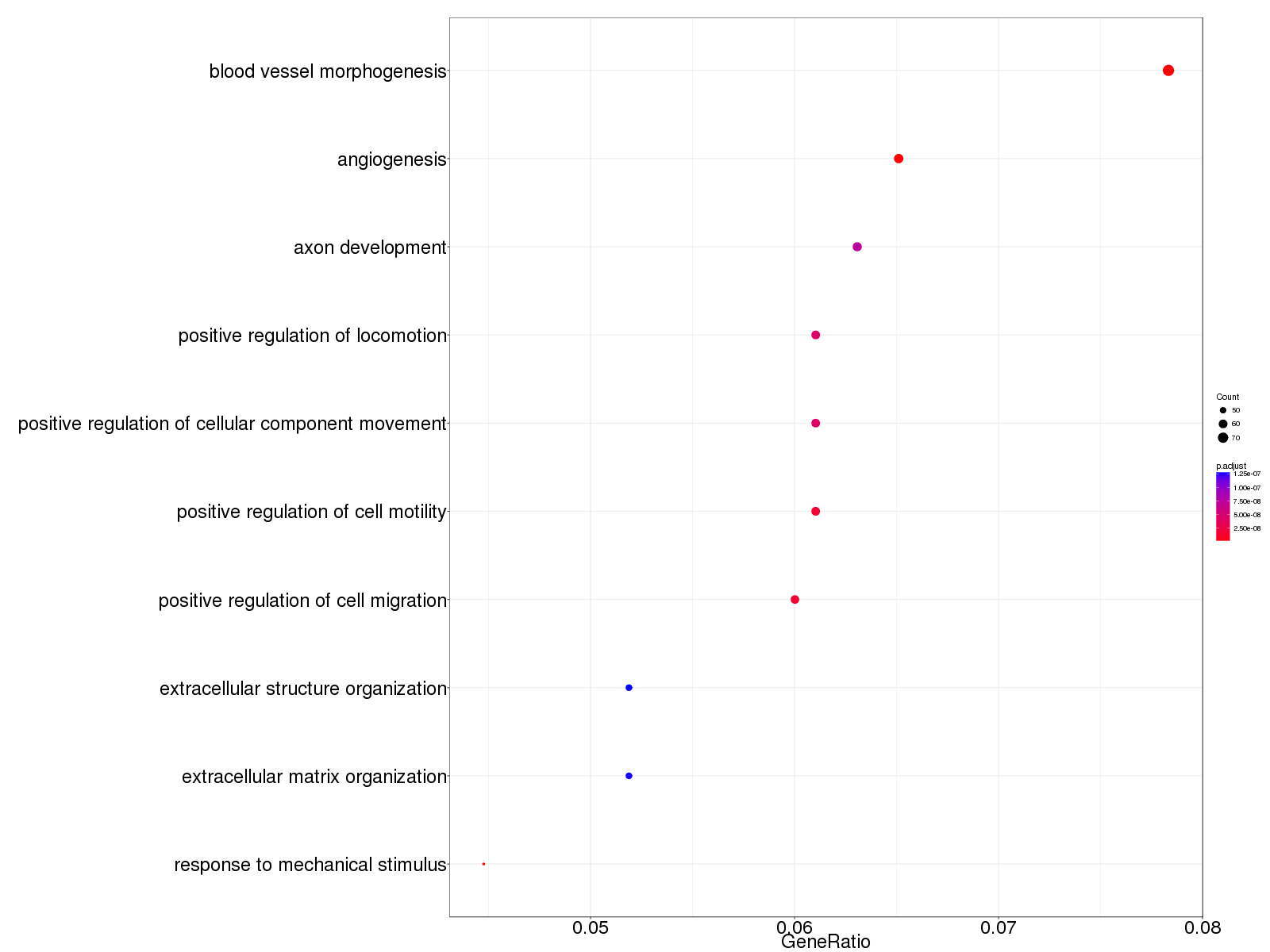
#### GO enrichment table

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**:

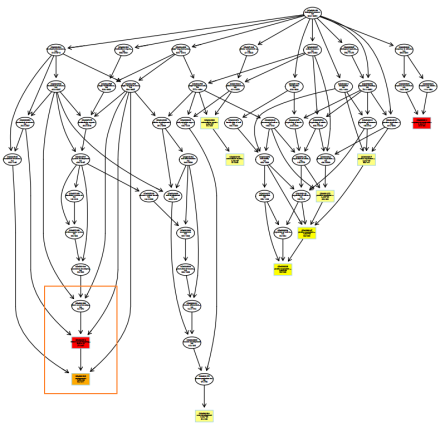
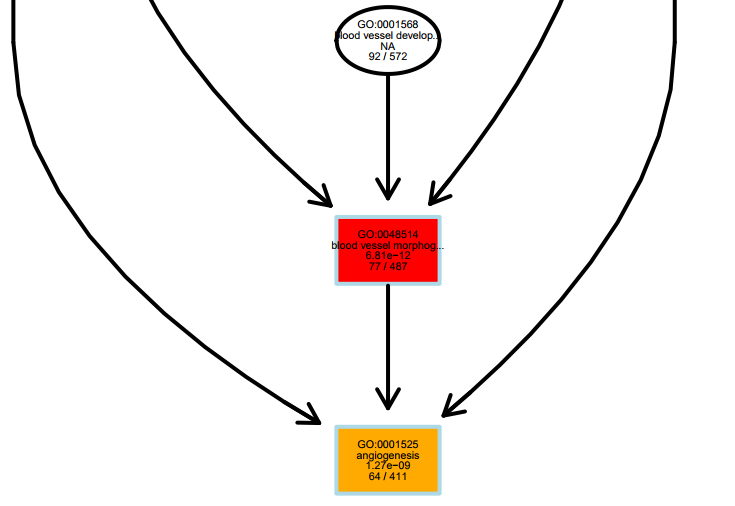
* “GeneRatio”: this bp GO term found in this case / total bp GO terms found in this case
* “BgRatio”: this bp GO term count in whole database / total bp GO terms existed in whole database so far
* “p.adjust”: calculate from GeneRatio and BgRatio

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



#### plotGOgraph (Graphical representation of GO)

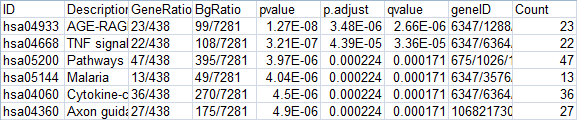
According to all of the bp GO terms we found, a cause-effect relation could be gotten from the GO enrichment table and also generate network graph.

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.



## 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](http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html). [*Nature Methods*](http://www.nature.com/nmeth). 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](http://doi.org/10.1089/omi.2011.0118).