Transcriptome Analysis Using RNA-Seq

강원대학교 동물유전체데이터분석 워크샵

Theragen

강의는 이렇게 진행될겁니다.

커본데.충실한 내용을 이해하실수 있도록 이기심을 버리고 천천히

강의 준비 자료

Link : https://github.com/baekip/KWU_Workshop

Link : https://drive.google.com/file/d/1wc7SOJP0gcEQ_nK2TI5SW2dmMVFDAsHT/view?usp=sharing

Transcriptome Sequencing

전사체 시퀀싱을 통해서는 각 샘플 간 전사체 발현 값의 차이를 확인할 수 있습니다.
모든 생명체는 자신의 유전 정보를 토대로 단백질을 만들어내기 위해서 유전자 DNA 서열을 mRNA라고 하는 중간 매개체로
전사(transcription)하는 과정을 거쳐야 합니다. 이러한 mRNA를 분석한다면 특정 시점에서 활성화된 유전자 정보를 파악할 수 있습니다.

서비스 개요

Reference Based

- Gene Expression
- DEG
- Splicing Variants
- Gene Fusions

Small-RNA Seq

- miRNA Expression
- Identification of miRNA
- Target Prediction

de novo Based

- Transcript Assembly
- Gene Expression
- DEG

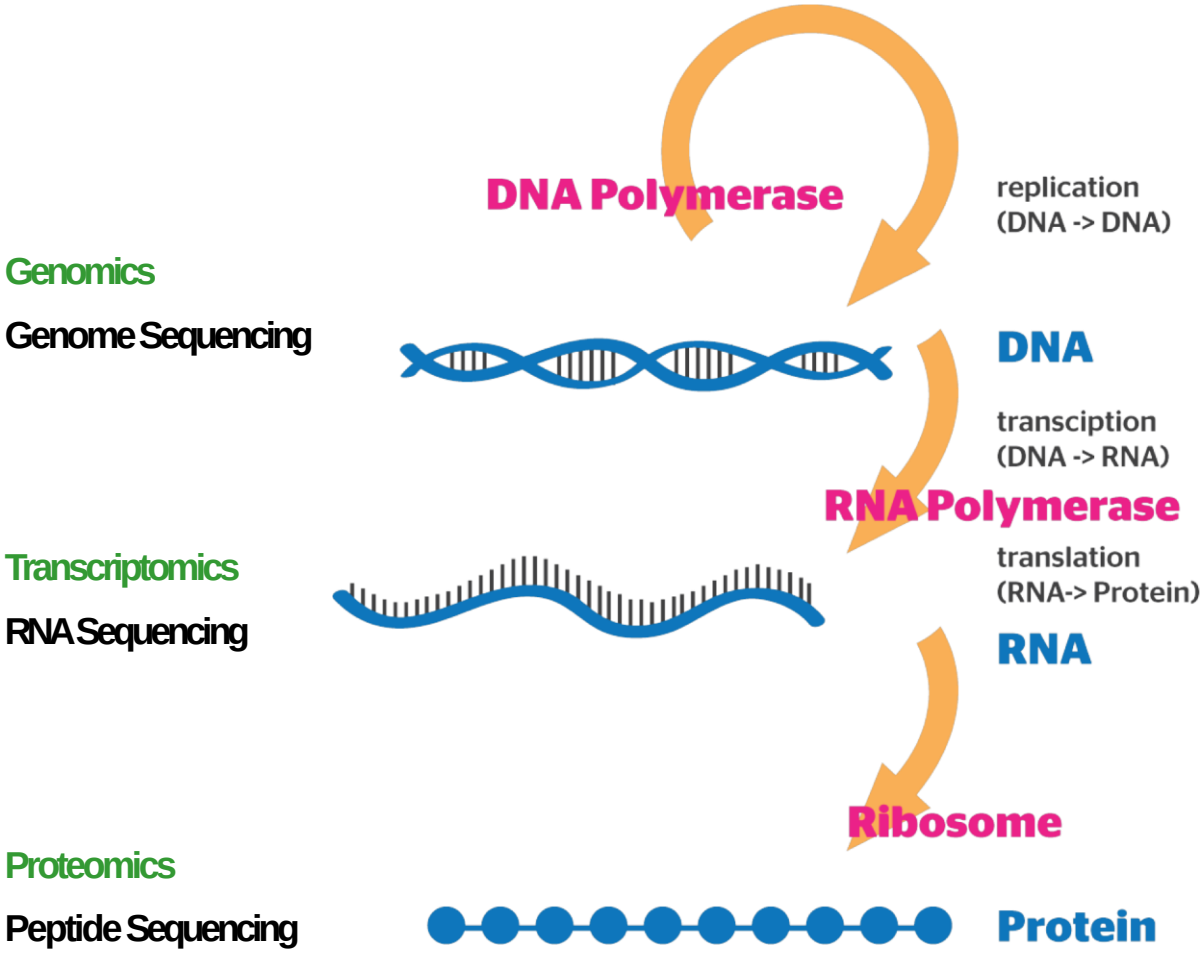
Main Capture Kit

- TruSeq Stranded mRNA Sample Prep Kit
- TruSeq Stranded Total RNA Sample Prep Kit
- NEXTflex Small RNA-Seq Kit v3

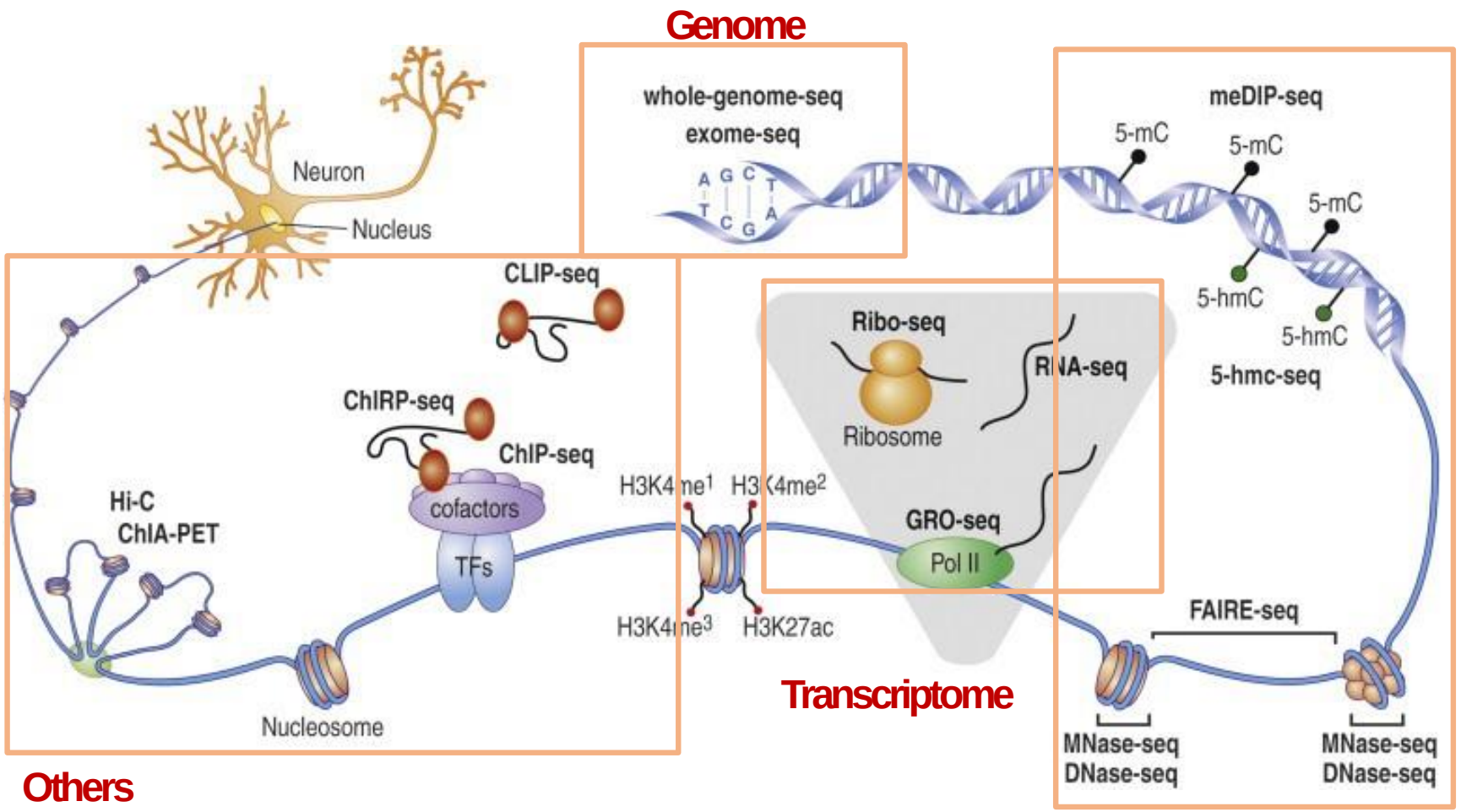
시퀀싱 플랫폼

HiSeq 2500 / HiSeq 4000 / NovaSeq 6000

Central Dogma!



Capability of Next Generation Sequencing in Molecular world

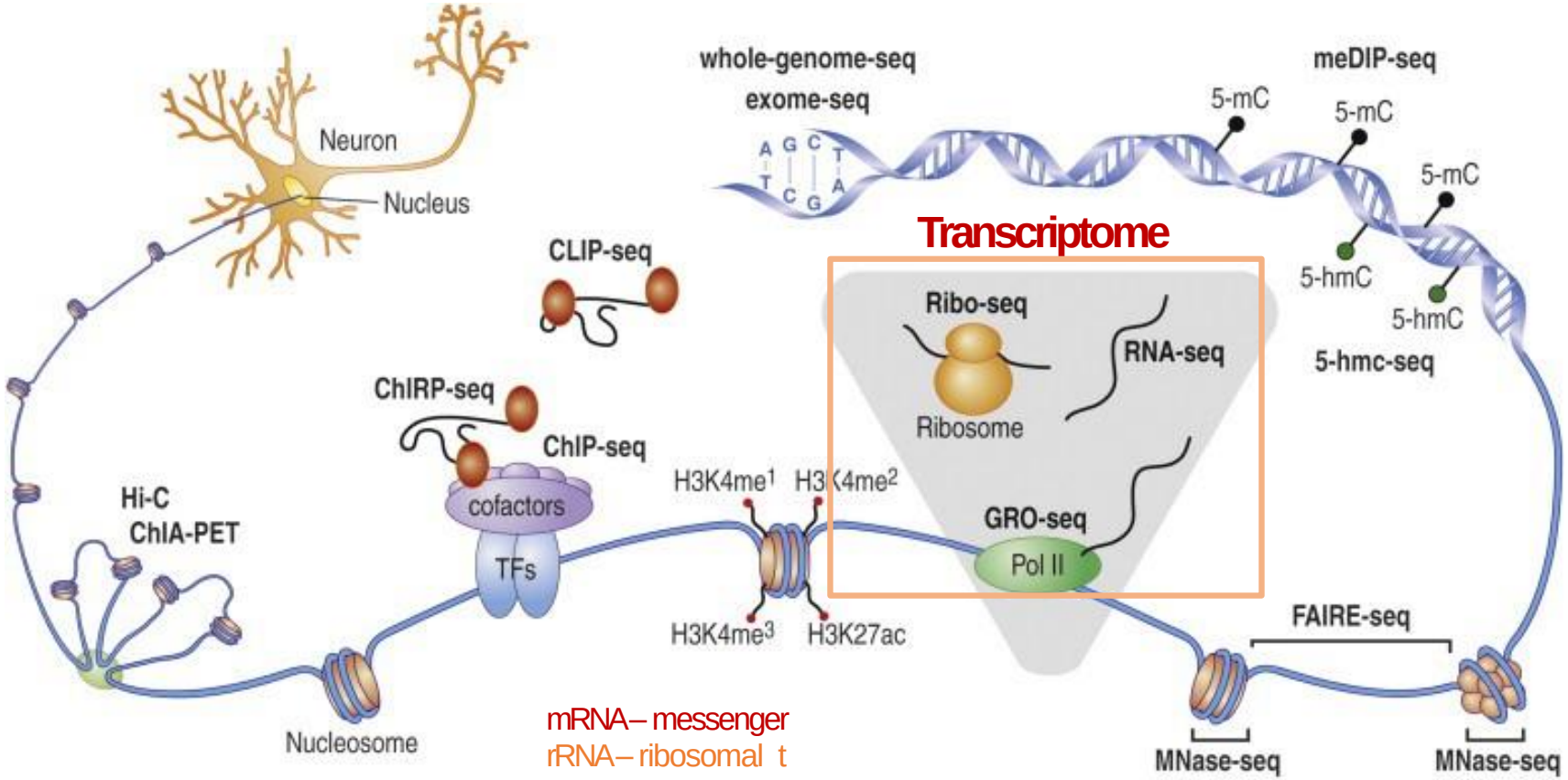


Others

TF-binding
Three-Dimensional Genome Structure
Chromatin Interaction Analysis Protein
-RNA interaction
Parallel analysis of RNAstructure

Neuron (2013) 77: 606–623

Capability of Next Generation Sequencing in Molecular world

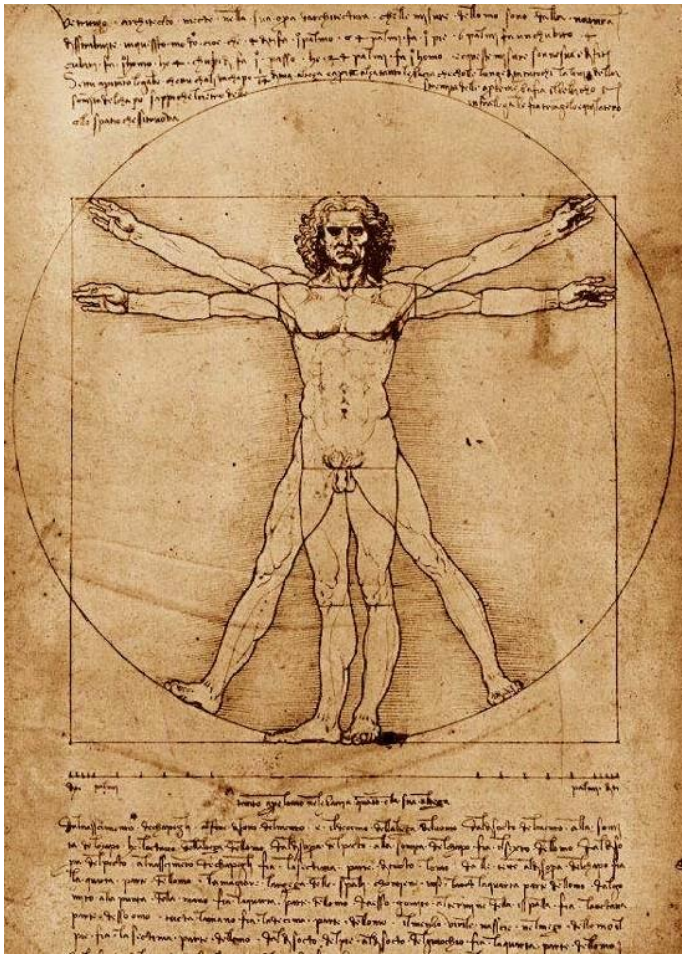


mRNA– messenger
rRNA– ribosomal t
RNA– transfer
Regulatory RNAs (siRNA, miRNA, lncRNA etc)
hnRNA– heterogeneous nuclear
scRNA– small cytoplasmic snRN
A– small nuclear
snoRNA– small nucleolar

Neuron (2013) 77: 606–623

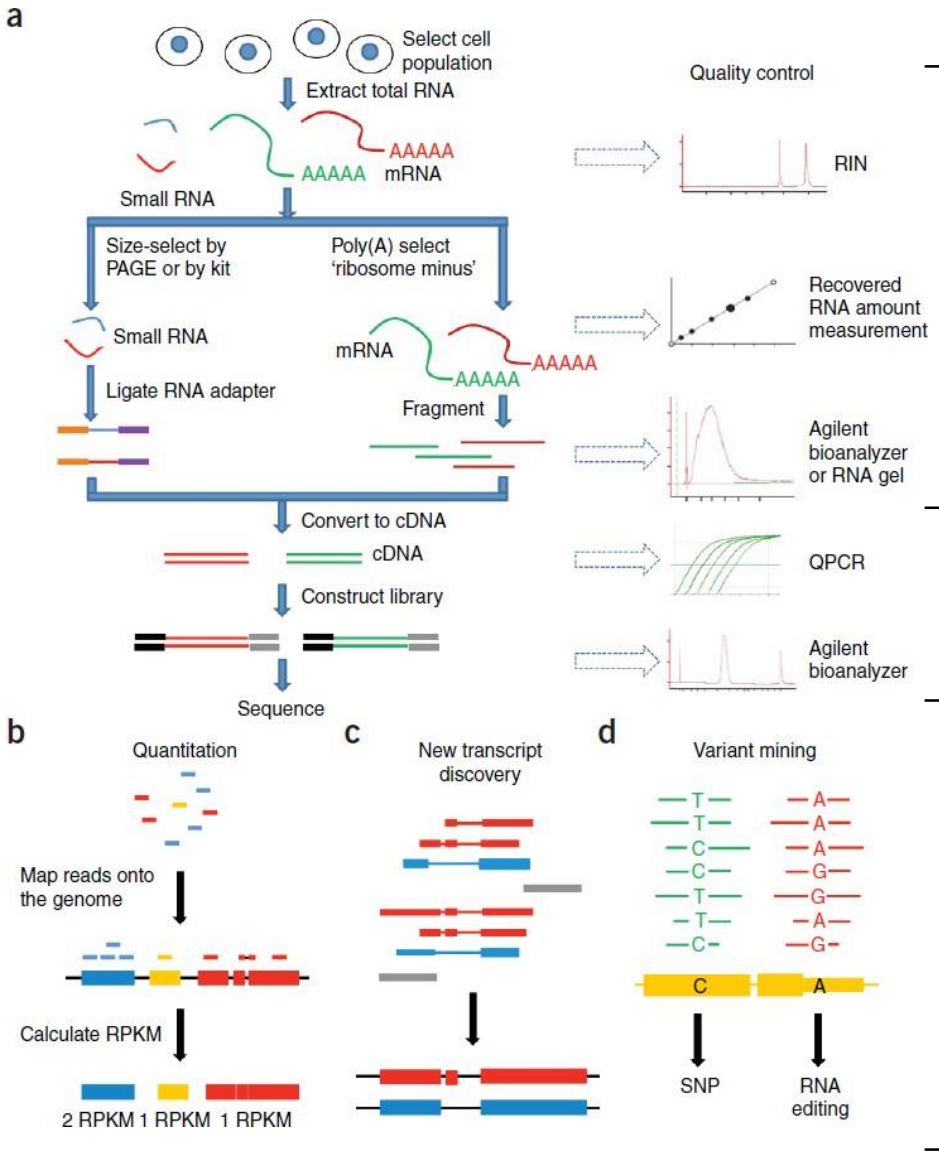
RNA-Seq : a revolutionary tool for Transcriptomics

- The complete set of transcripts in a cell, and their quantity
- The key aims of transcriptomics are:
  - To catalogue all species of transcript, including mRNAs , non-coding RNAs and small RNAs
  - To determine the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patte rns and other post-transcriptional modifications
  - To quantify the changing expression levels of each tra nscript during development and under different condi tion



Wang Z, Gerstein M, Snyder M. ***RNA-Seq: a revolutionary tool for transcriptomics****. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484.*

RNA-seq workflow



Sample Prep.

- *mRNA Prep.*

Library construction

- *cDNA synthesis*

Sequencing

Bioinformatics

- *Mapping*
- *Quantification*
- *Differential expression*
- *Functional annotation*

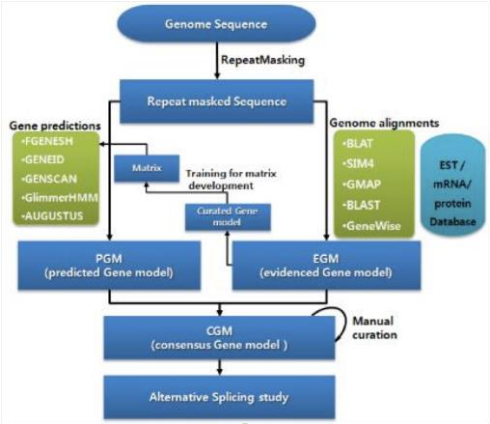
(a) *Schematic diagram of RNA-seq library construction.* ***(b)*** *Mapping programs align reads to the reference genome and map splice junctions.* ***(c)*** *New transcript discovery.* ***(d)*** *Identification of genomic variants (for example, SNPs) or candidates for RNA editing. (Zeng & Mortazavi 2012)*

Diverse Approach of Transcriptome analysis

DNA & RNA level

✻ De novo Genome assembly

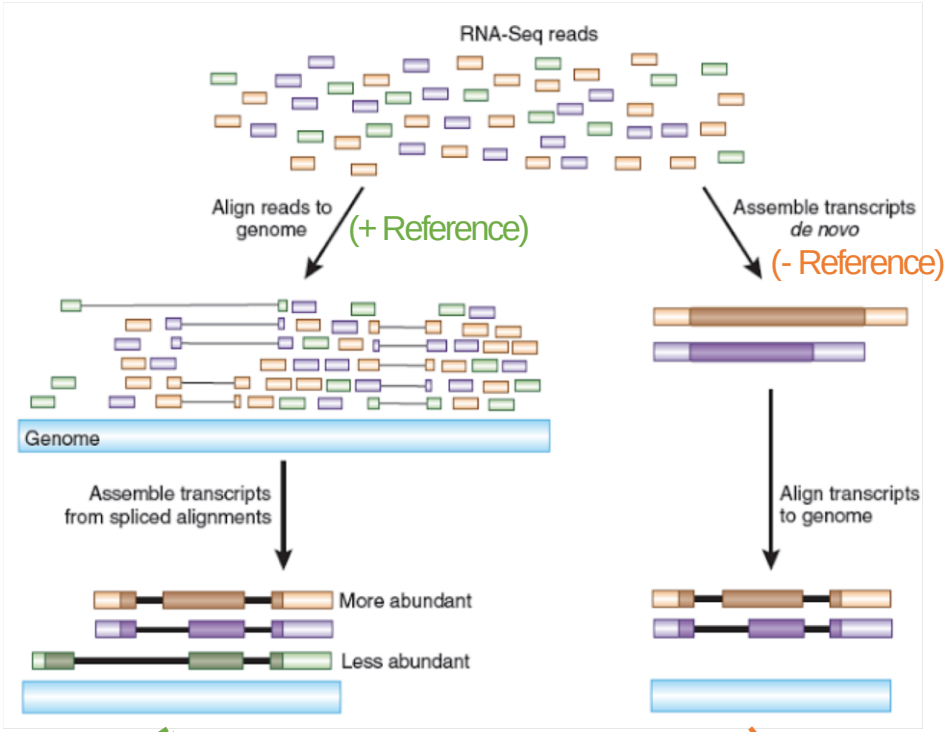
Structural Annotation
- Repeat identification
- Evidence-driven alignment
- Ab initio Gene prediction



RNA level

✻ Mapping-first approach

✻ Assembly-first method



Novel transcripts/ Alternative isoforms finding

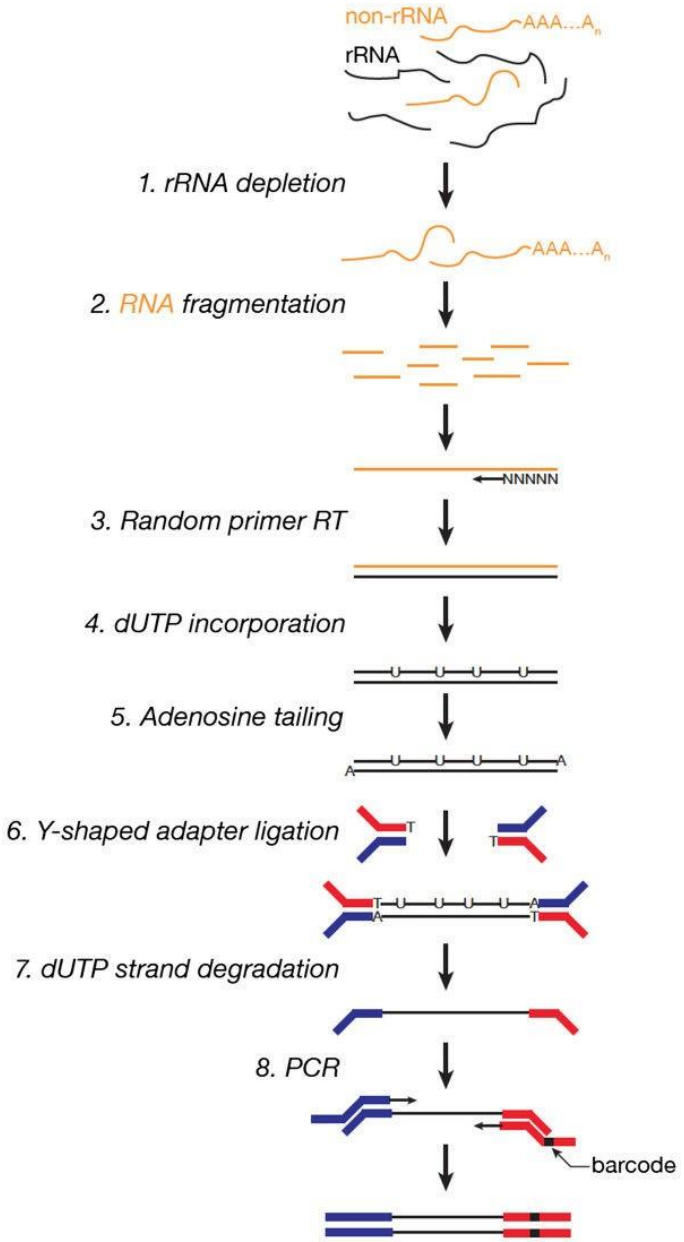
Gene expression quantification & Differential expression (DE)

Functional annotation of assembled transcriptome

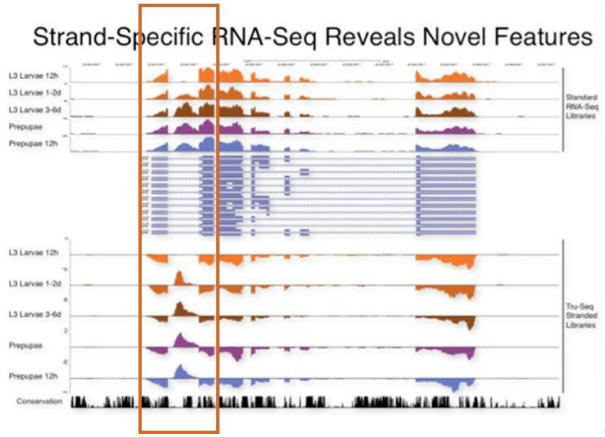
http://www.incodom.kr/Genome_annotation

Method for stranded-specific RNA-Seq

Method

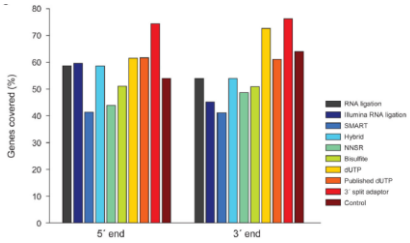


Stranded-specific vs. Non-stranded-specific



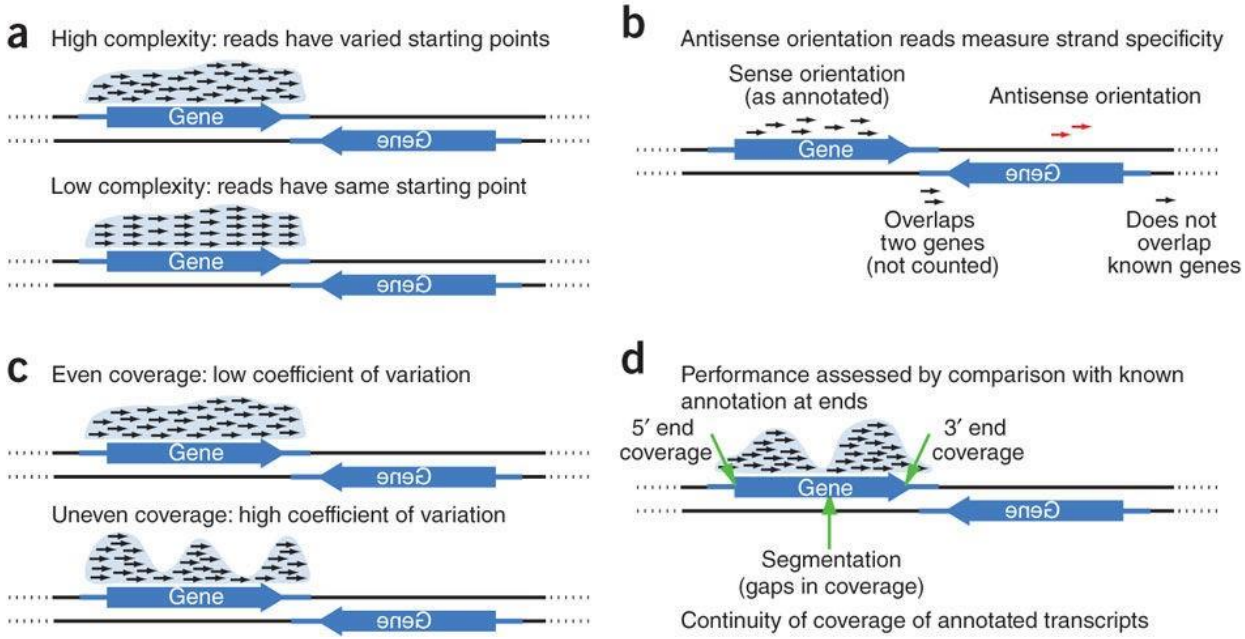
No orientation

Strandedness preserved

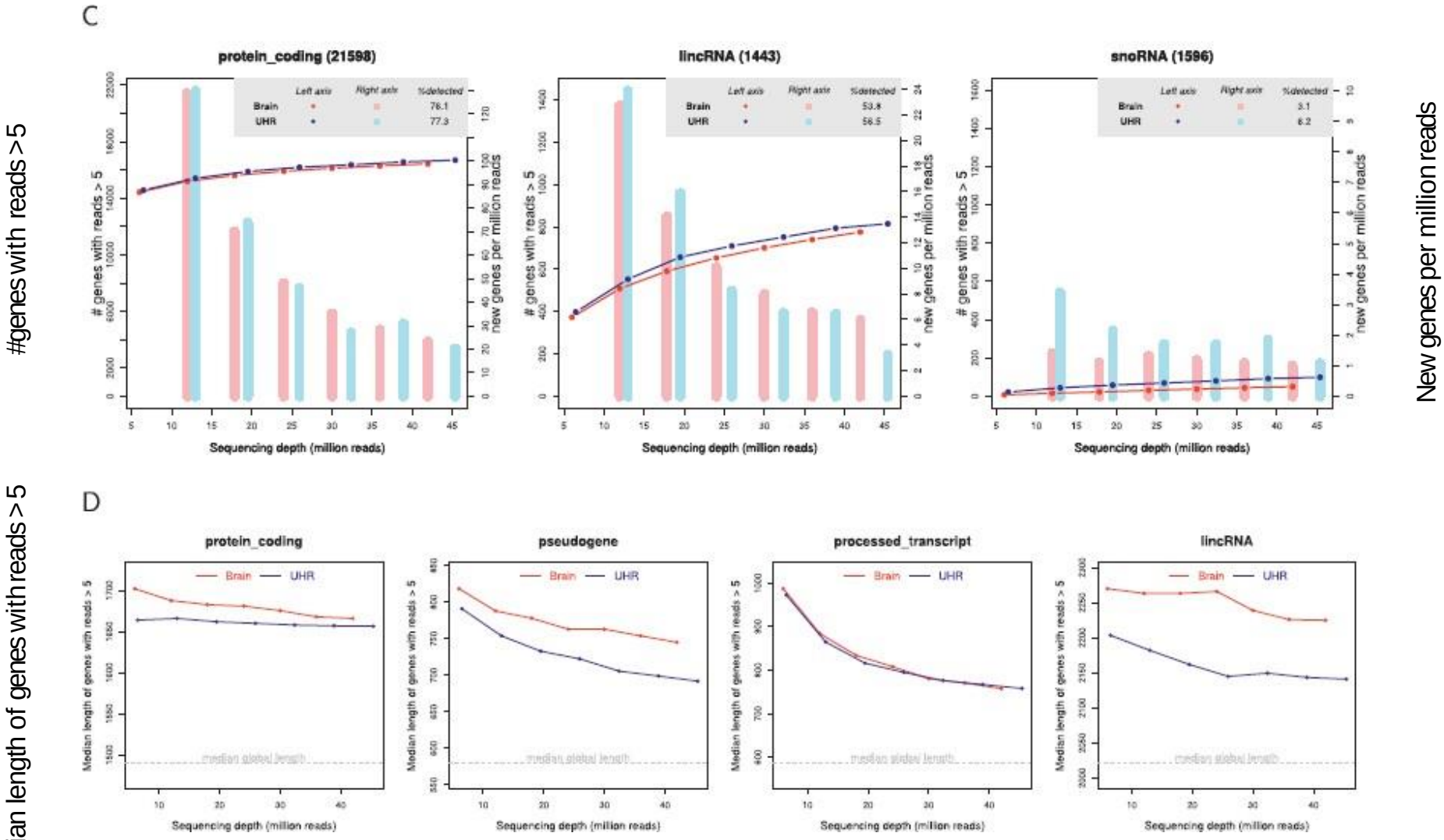


5′ and 3′ end t ranscript coverage

Key criteria for evaluation of strand-specific RNAseq libraries (Levin et al. 2010)



How does sequencing depth affects to the estimation of differential expression in RNAseq data?



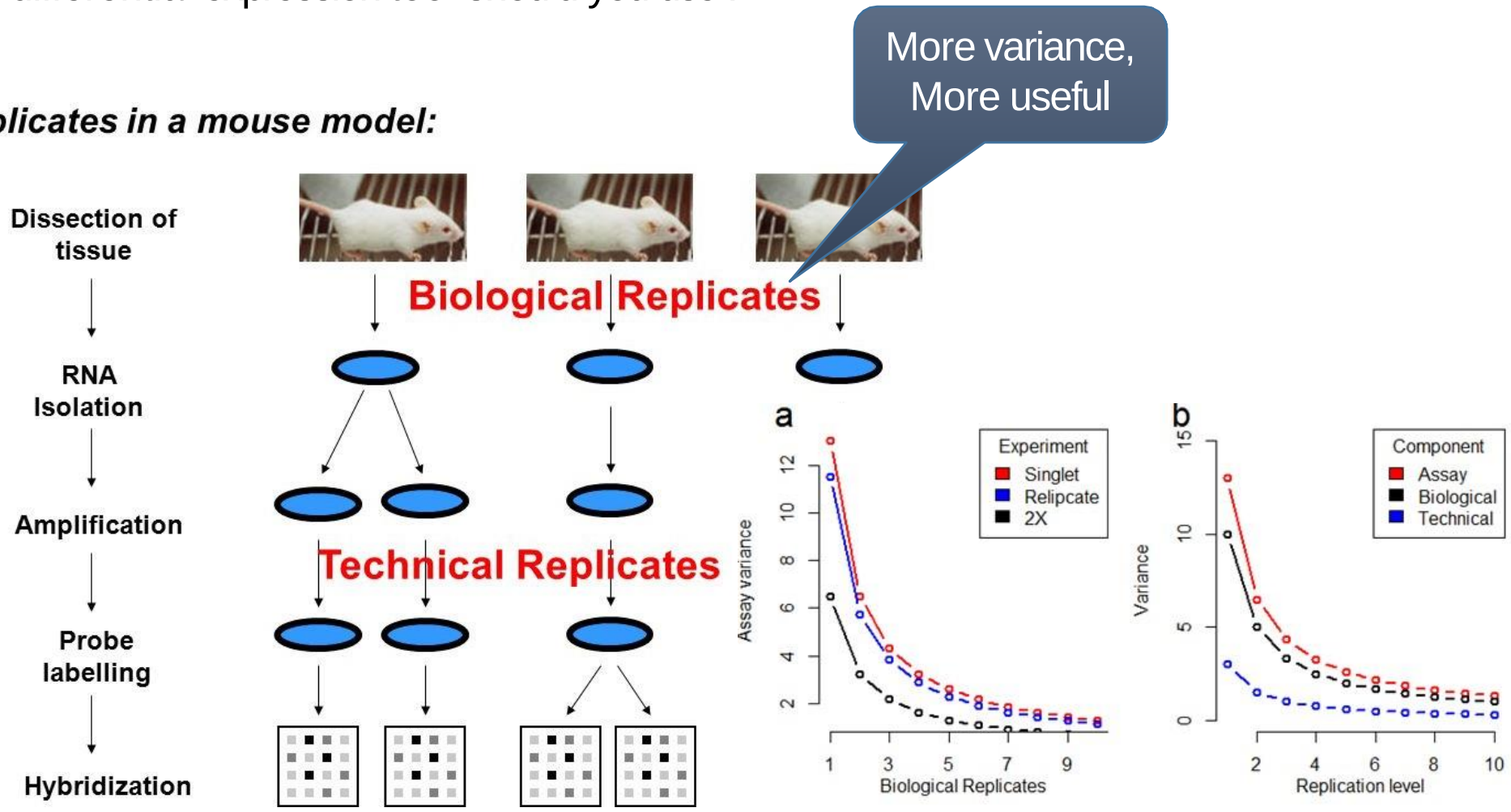
- Saturation curves and NDR bars for protein-coding, lincRNA, and snoRNA
- As more it is sequenced, small genes are easier detected.

(Differential expression in RNA-seq: a matter of depth. Tarazona et al. 2011)

Replicates

How many biological replicates are needed in an RNA-Seq experiment and which differential expression tool should you use?

Replicates in a mouse model:



Helpful for Statistical problems

Breen 2016

An overview of the Tuxedo protocol : How to basic RNA-Seq analysis (Trapnell et al. 2012)

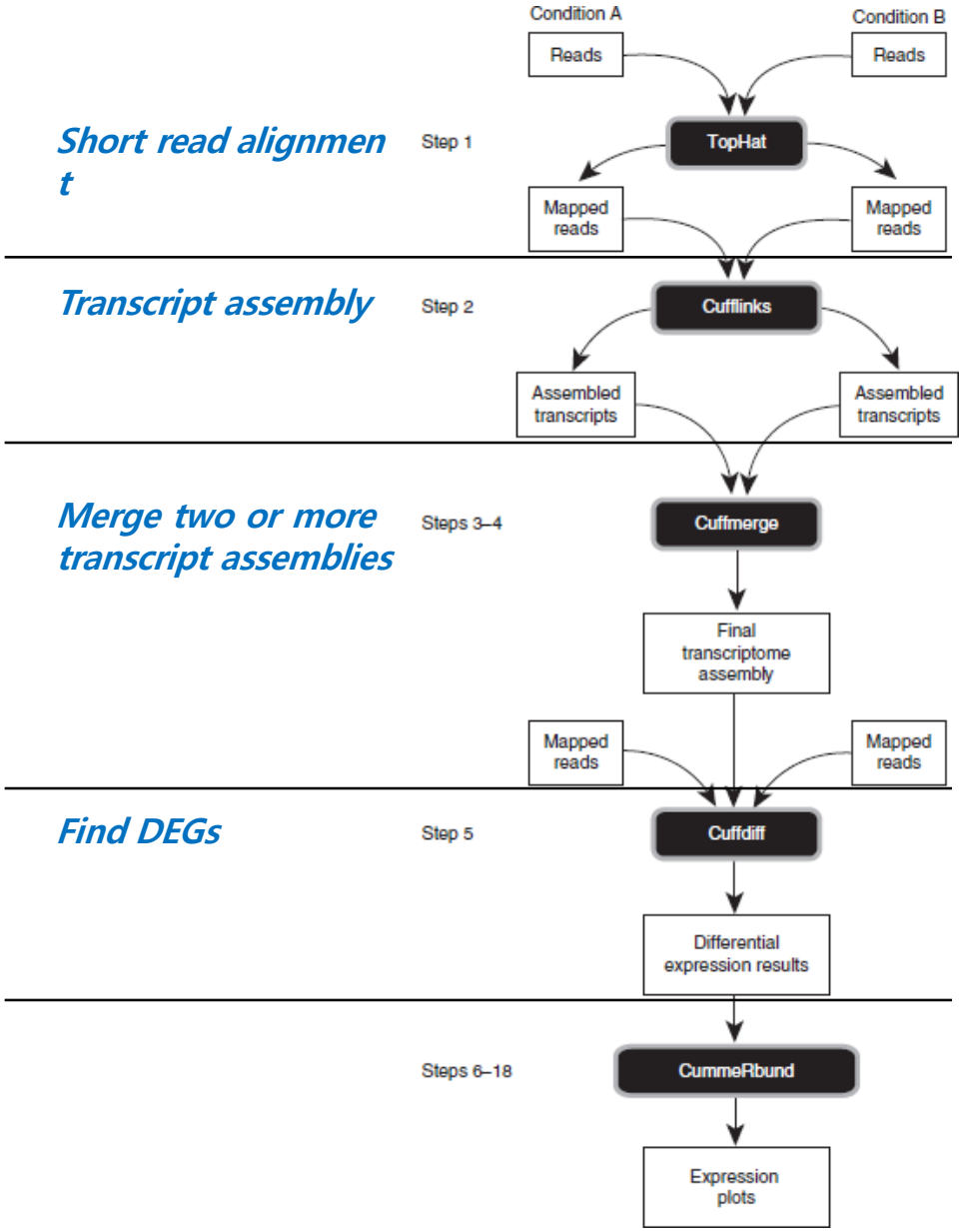


TABLE 1 | Library type options for TopHat and Cufflinks.

Library type	RNA-seq protocol
fr-unstranded (default)	Illumina TruSeq
fr-firststrand	dUTP, NSR, NNSR[39]
fr-secondstrand	Directional Illumina (Ligation), Standard SOLiD

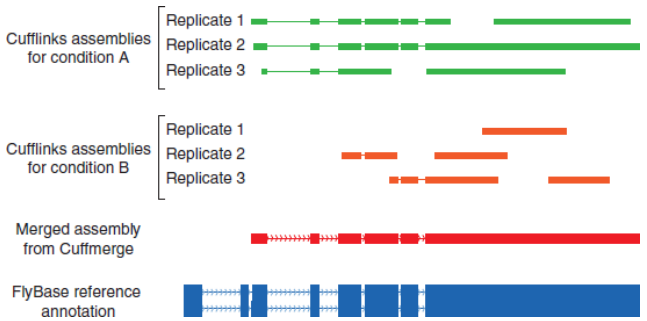
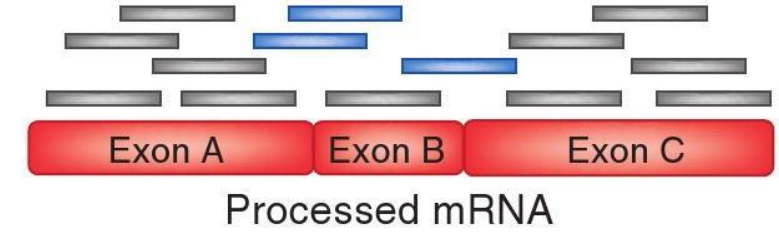
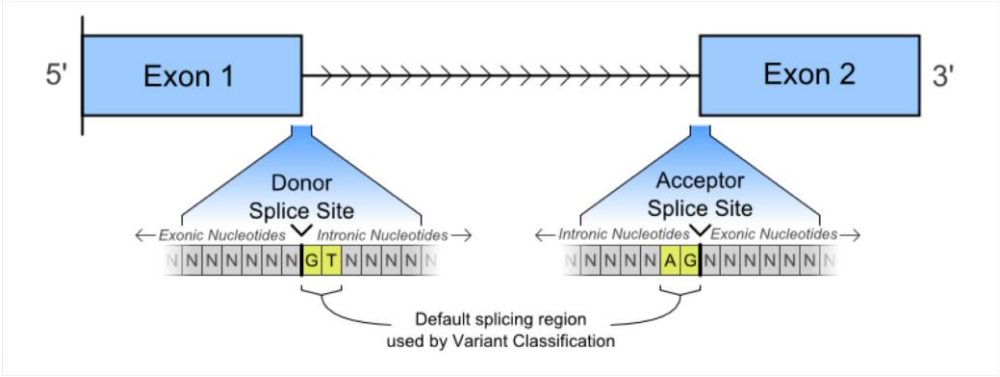
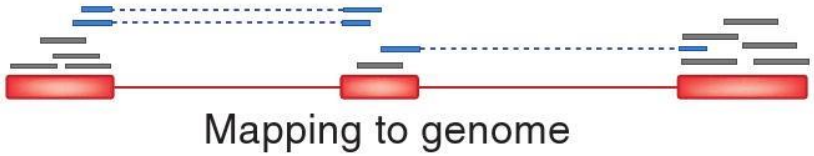


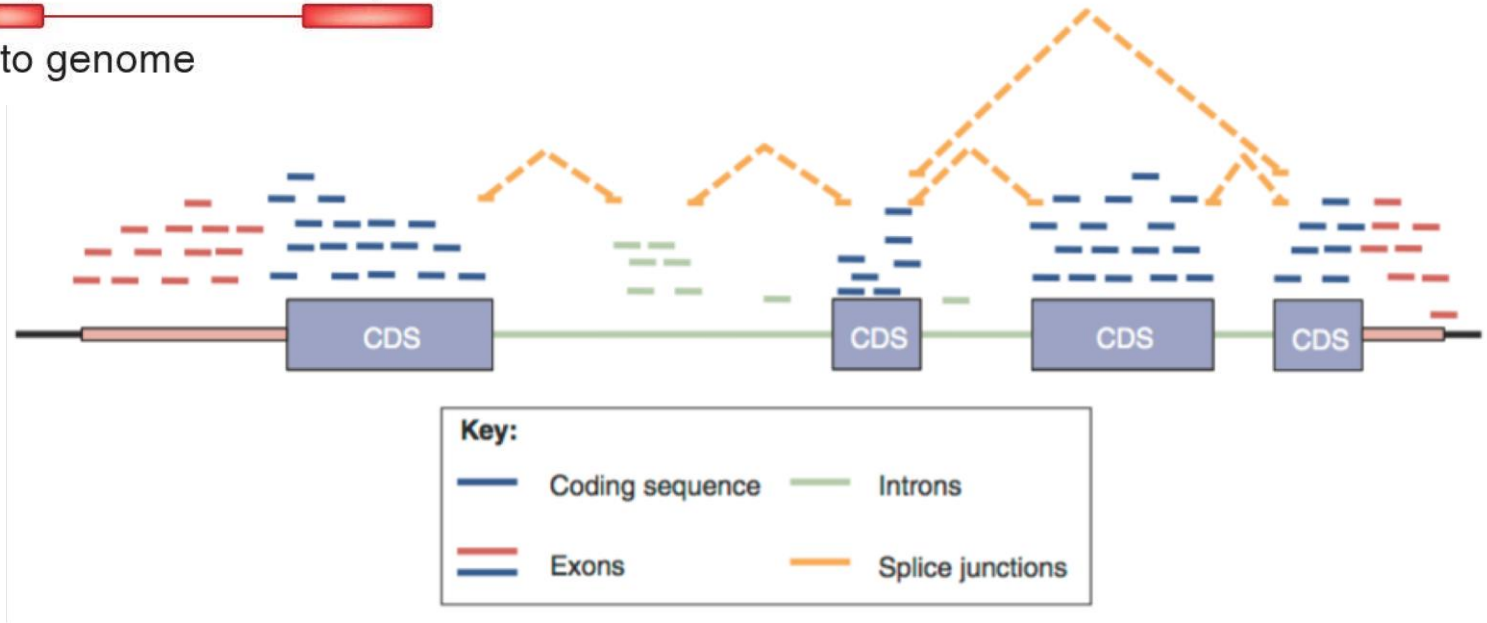
Figure 2 | An overview of the Tuxedo protocol. In an experiment involving two conditions, reads are first mapped to the genome with TopHat. The reads for each biological replicate are mapped independently. These mapped reads are provided as input to Cufflinks, which produces one file of assembled transfrags for each replicate. The assembly files are merged with the reference transcriptome annotation into a unified annotation for further analysis. This merged annotation is quantified in each condition by Cuffdiff, which produces expression data in a set of tabular files. These files are indexed and visualized with CummeRbund to facilitate exploration of genes identified by Cuffdiff as differentially expressed, spliced, or transcriptionally regulated genes. FPKM, fragments per kilobase of transcript per million fragments mapped.

Splice junction을 고려한 mapping





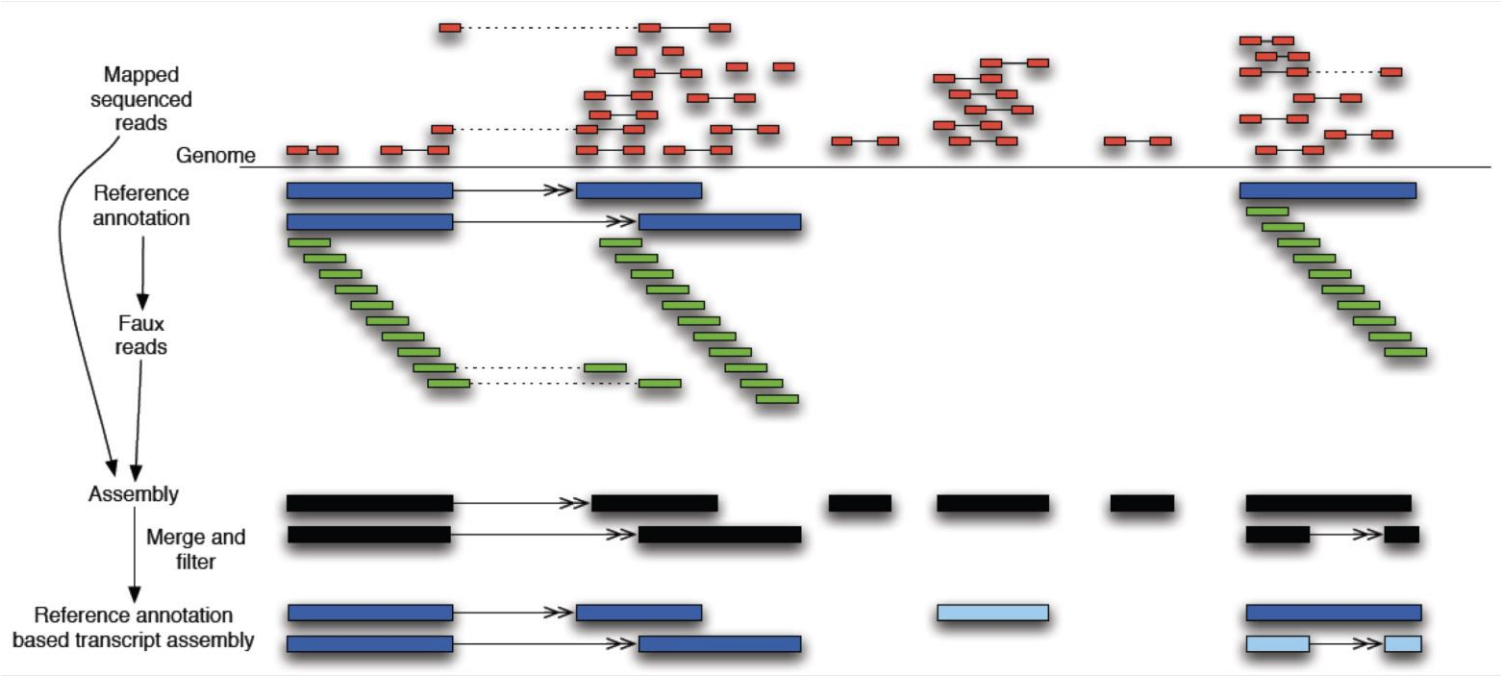




Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220.
Trapnell C, Salzberg SL., How to map billions of short reads onto genomes. Nat Biotechnol. 2009 May;27(5):455-7.

Expression & Modeling w/Cufflinks

Detecting (a) splice junctions, (b) mismatches and (c) tails (i.e. TopHat, STAR)

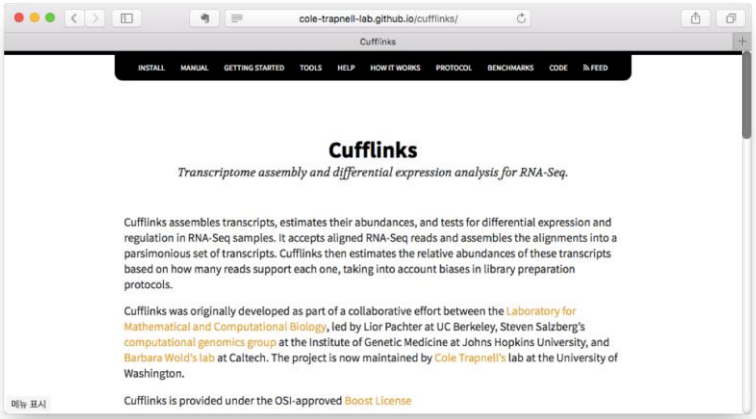


Adam Roberts et al., Identification of novel transcripts in annotated genomes using RNA9Seq. Bioinforma4cs, 2011, 27:2325–2329

Transcriptome assembly and differential expression analysis for RNA-seq

Cufflinks assembles transcripts, estimates their abundances, and tests for differential e xpression and regulation in RNA-Seqsamples.

Cufflinks constructs a parsimonious set of transcripts that “explain” the reads observe d in an RNA-seqexperiment.

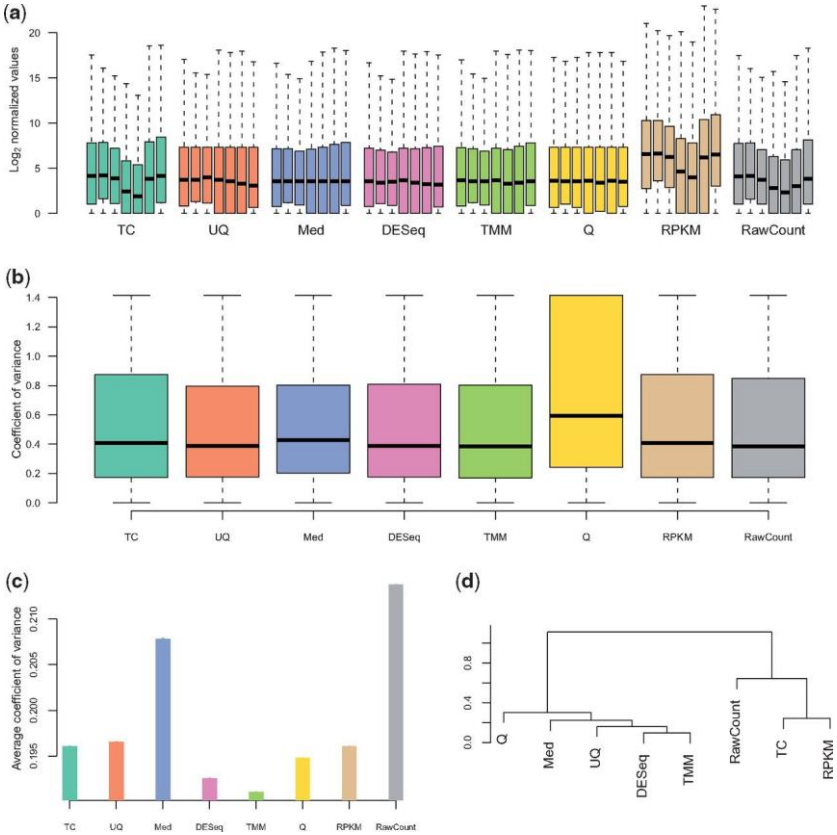


Normalization 이해하기 : 비교를 위한 최소한의 조건

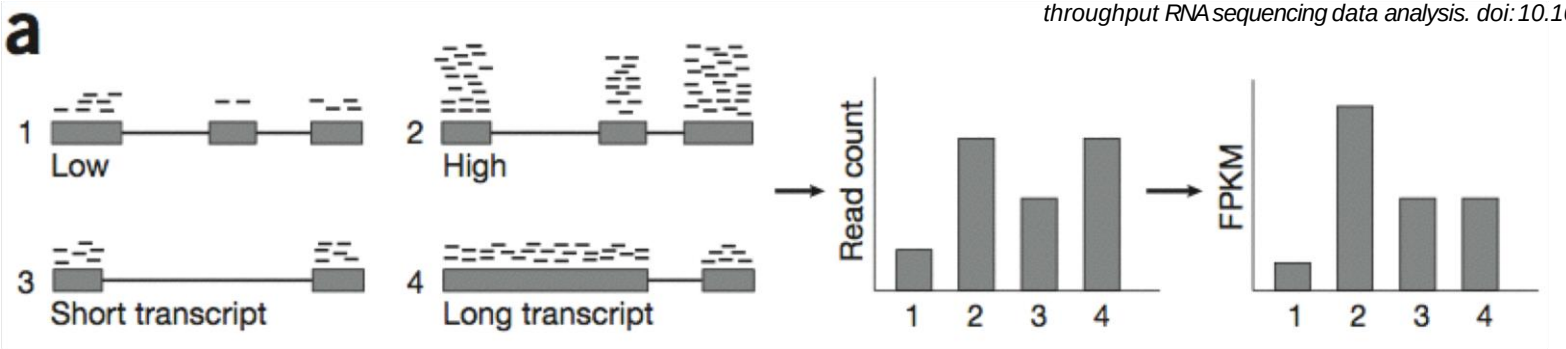
R/FPKM	# Length of gene/tr # Sequencing depth (=library size)
TPM	# Length of gene/tr # 1 (million)
TMM	# # Correction factor

- R/FPKM : Reads Per Kilobase of transcript per Million mapped reads
- TPM : Transcripts Per Million
- TMM : Trimmed Mean of M-values

Read counts need to be properly normalized to extract meaningful expression estimates



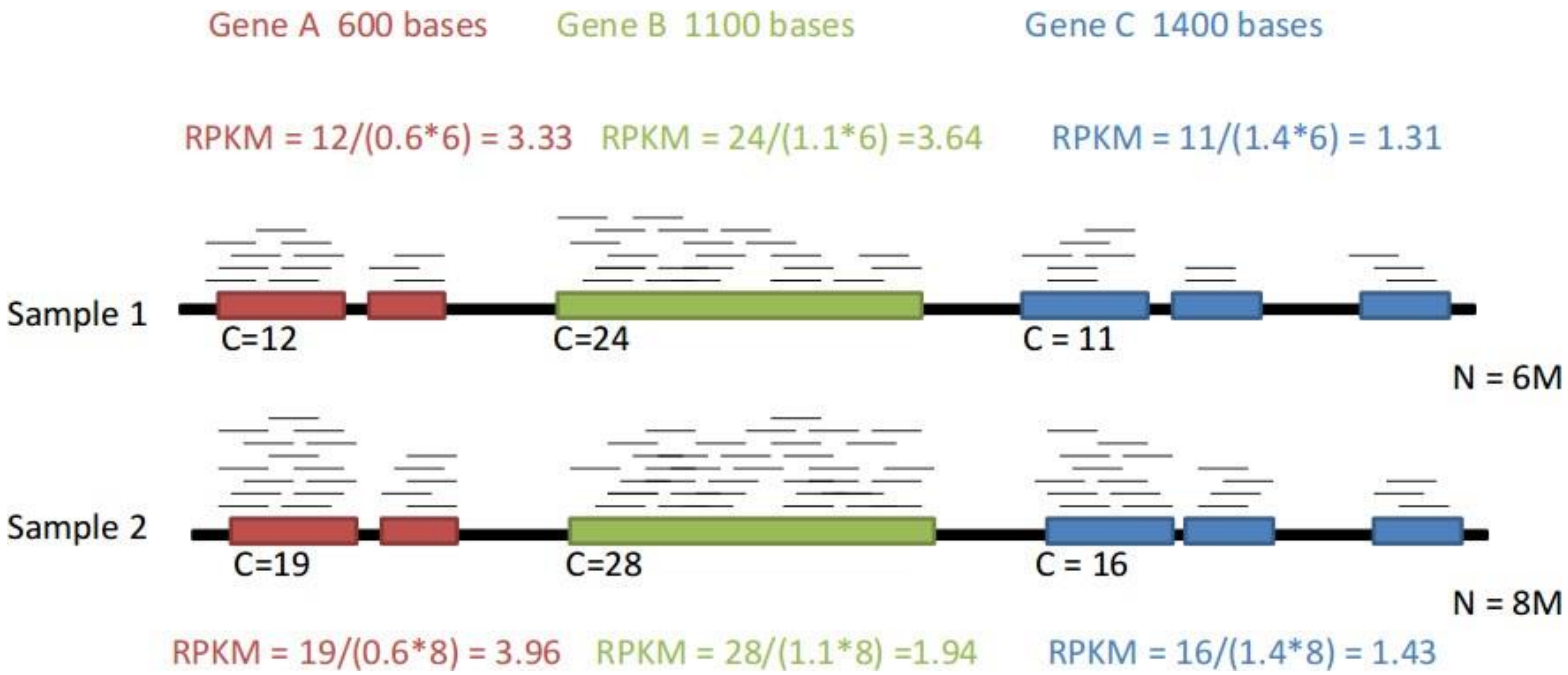
A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. doi: 10.1093/bib/bbs046.



Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011 Jun;8(6):469-77.

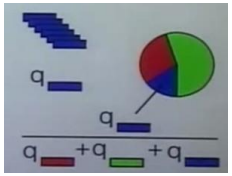
RPKM

RPKM is a method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads.



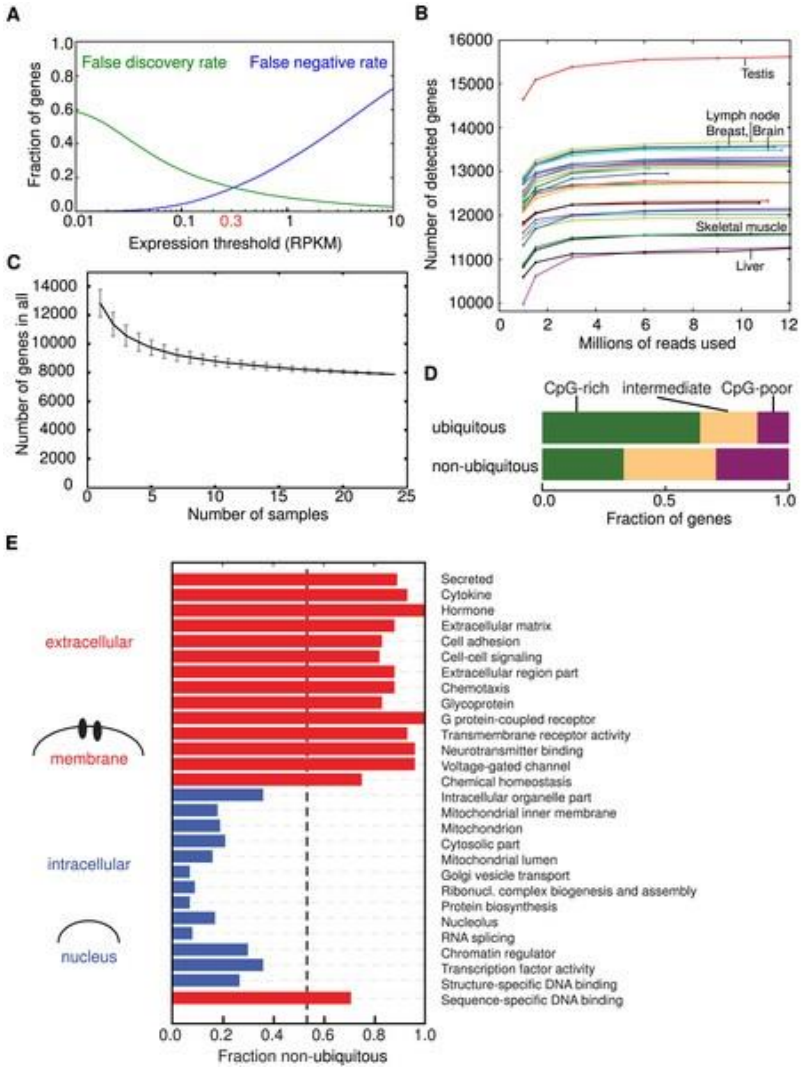
RPKM = 10^6 * C* 10^3 / L*N where,
C= number of mappable reads per each feature (in our case per gene)
L= number of the length of the feature (lenght of the gene) **in kb (10^3)**
N = Total number of mappable reads per sample **in Millions (10^6)**

$$R = \frac{10^9 C}{NL}$$



Reference:
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). ***Mapping and quantifying mammalian transcriptomes by rna-seq.*** *Nat Methods , 5(7):621-628.*

How many expressed gene or transcripts?



(A) False discovery and negative rate for the detection of genes as a function of detection threshold used, demonstrating how **a threshold of 0.3 RPKM was chosen.**

(B) The number of genes detected (>0.3 RPKM) at different sequencing depths. Each curve represents a sample. Above 3 million reads the sequence depth matters little for how many genes are detected as expressed.

(C) The number of ubiquitous genes (**expressed >0.3 RPKM in all samples**) as a function of the number of samples used. Error bars show the standard variation, black line the mean.

Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. PLoS Comput Biol 5(12): e1000598. doi:10.1371/journal.pcbi.1000598

differentially expressed genes (DEGs) analysis

The identification of genes (or other types of genomic features, such as transcripts or exons) that are expressed in significantly different quantities in distinct groups of samples, be it biological con didtions (drug-treated vs. controls), diseased vs. healthy individuals, different tissues, different stag es of development, or something else.

FC(Fold Change) and Statistics threshold

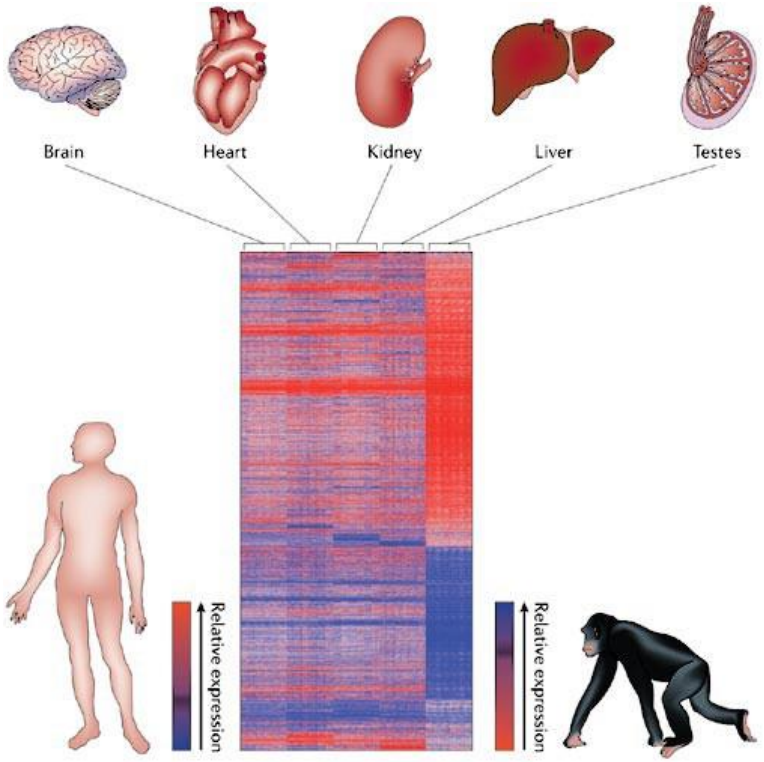
FDR correction: Benjamini–Hochberg method (Benjamini et al. 1995)

A difference in expression is only statistically significant when that difference is outside expected variation

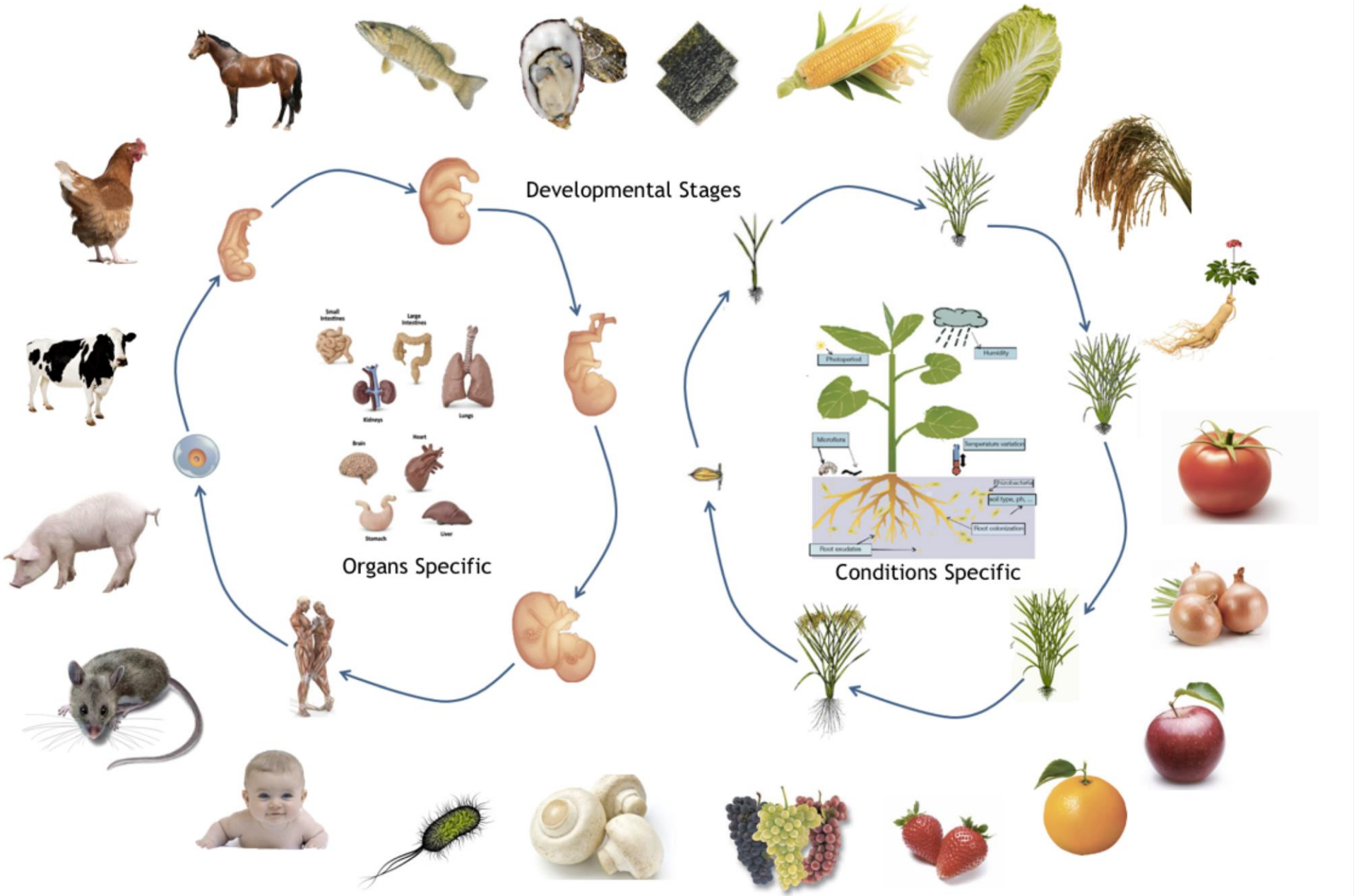
Major challenge: How to model technical and biological variance?

Negative binomial distribution is most commonly used to model sample variance (DESeq, edgeR, baySeq, Cufflinks suite)

Different types of expression quantification and normalization



Data Classification of RNA-Seq projects

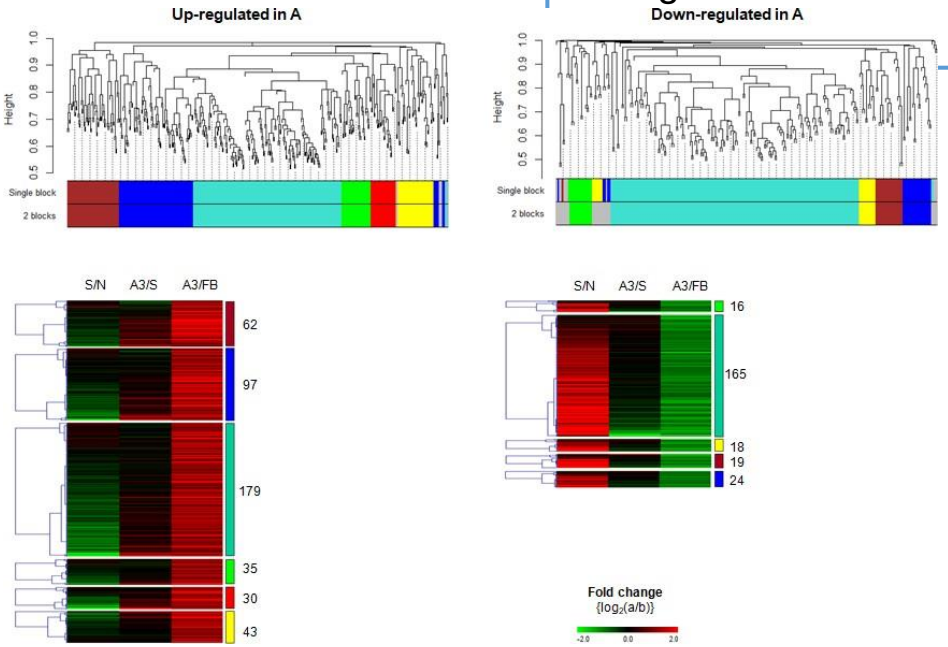


Thanks Dr.Murthi

Classification of differentially expressed genes (DEGs)

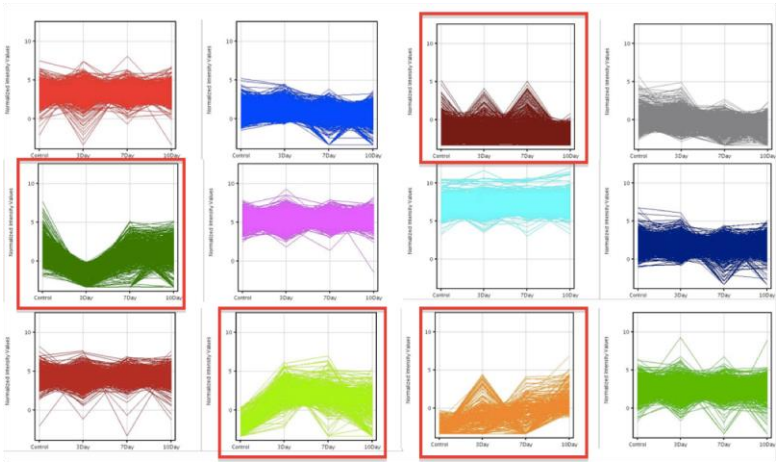
- **Pairwise**
  - CONTROL:CASE
  - …

Cuffdiff	edgeR	DESeq	TCC
Negative binomial (?)	Negative binomial	Negative binomial	Negative binomial
gene isoform	gene	gene	gene
geneset	Raw Read Count	Raw Read Count	Raw Read Count



Multi-conditions

- *Tissue-specific genes*
- *Time-series (Similar expression pattern)*

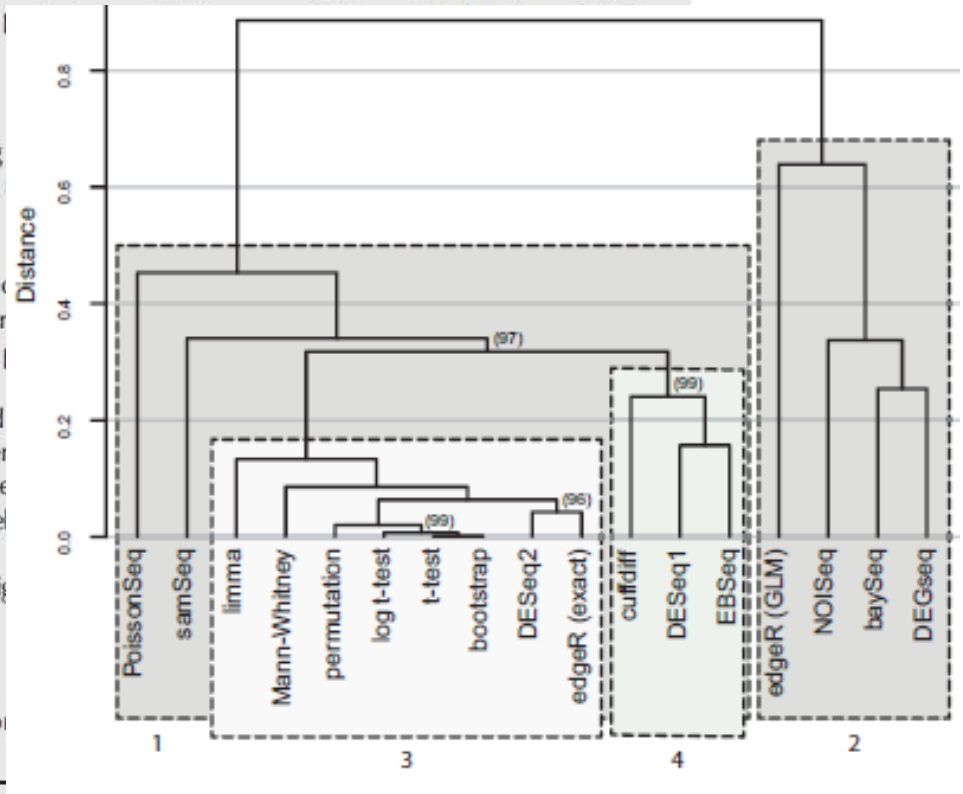


- *Clustering using Weighted Gene Co-expression Network Analysis (WGCNA)*
- *Hierarchical clustering using Euclidean distance and complete linkage clustering*

RNA-seq differential gene expression tools and statistical tests

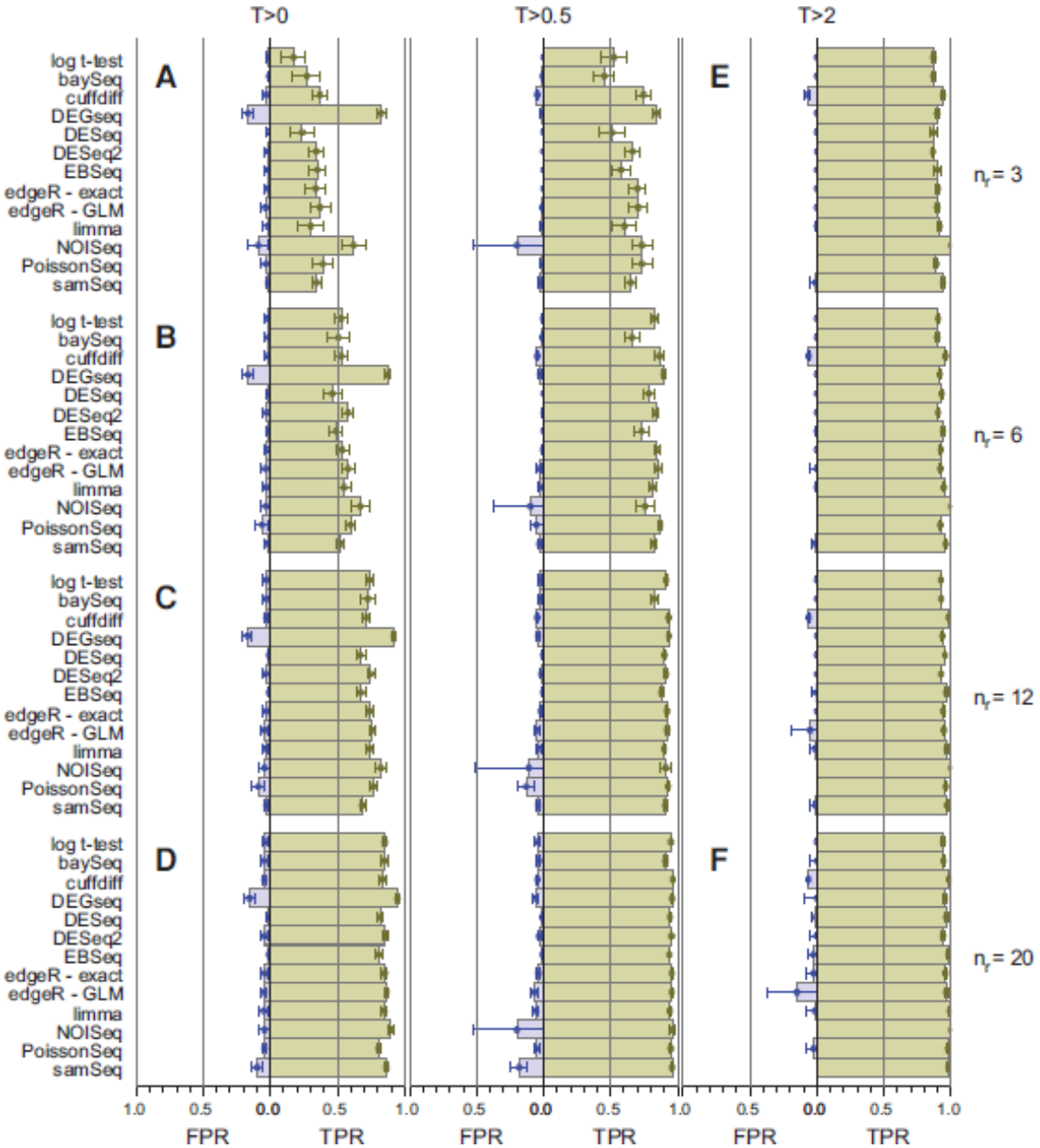
TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description	Version	Citations[d]	Reference
t-test	Normal	DEseq[a]	Two-sample *t*-test for equal variances	–	–	–
log *t*-test	Log-normal	DEseq[a]	Log-ratio *t*-test	–	–	–
Mann-Whitney	None	DEseq[a]	Mann-Whitney test	–	–	Mann and Whitney (1947)
Permutation	None	DEseq[a]	Permutation test	–	–	Efron and Tibshirani (1993a)
Bootstrap	Normal	DEseq[a]	Bootstrap test	–	–	Efron and Tibshirani (1993a)
baySeq[c]	Negative binomial	Internal	Empirical Bayesian estimate of likelihood			
Cuffdiff	Negative binomial	Internal	Unknown			
DECseq[c]	Binomial	None	Random sampling model using exact test and the likelihood			
DESeq[c]	Negative binomial	DEseq[a]	Shrinkage variance			
DESeq2[c]	Negative binomial	DEseq[a]	Shrinkage variance with varian and Cook's distance pre-filter			
EBSeq[c]	Negative binomial	DEseq[a] (median)	Empirical Bayesian estimate of likelihood			
edgeR[c]	Negative binomial	TMM[b]	Empirical Bayes estimation and exact test analogous to Fisher test but adapted to over-dispe or a generalized linear mode			
Limma[c]	Log-normal	TMM[b]	Generalized linear model			
NOISeq[c]	None	RPKM	Nonparametric test based on sig noise ratio			
PoissonSeq[c]	Poisson log-linear model	Internal	Score statistic			
SAMSeq[c]	None	Internal	Mann-Whitney test with Poisso resampling			

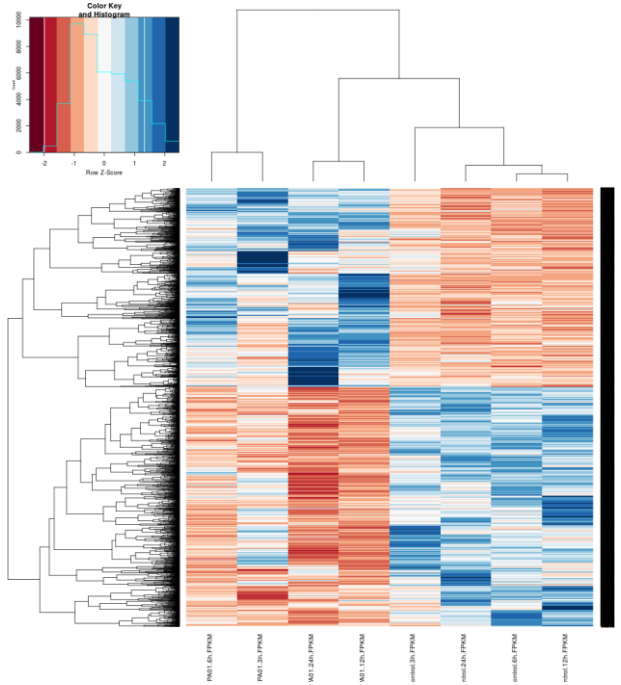


[a]See Anders and Huber (2010).
[b]See Robinson and Oshlack (2010).
[c]R (v3.2.2) and bioconductor (v3.1).
[d]As reported by PubMed Central articles that reference the listed reference (December 21, 2015).

Comparison of performance for each of the DGE tools



Comparison of the true positive rate (TPR) and false positive rate (FPR) performance for each of the DGE tools on low-, medium-, and highly replicated RNA-seq data ($nr \in \{3, 6, 12, 20\}$ —rows) for three $|\log2(FC)|$ thresholds ($T \in \{0, 0.5, 2\}$—columns).



Schurch et al. 2016

Functional annotation for DEGs: GO enrichment

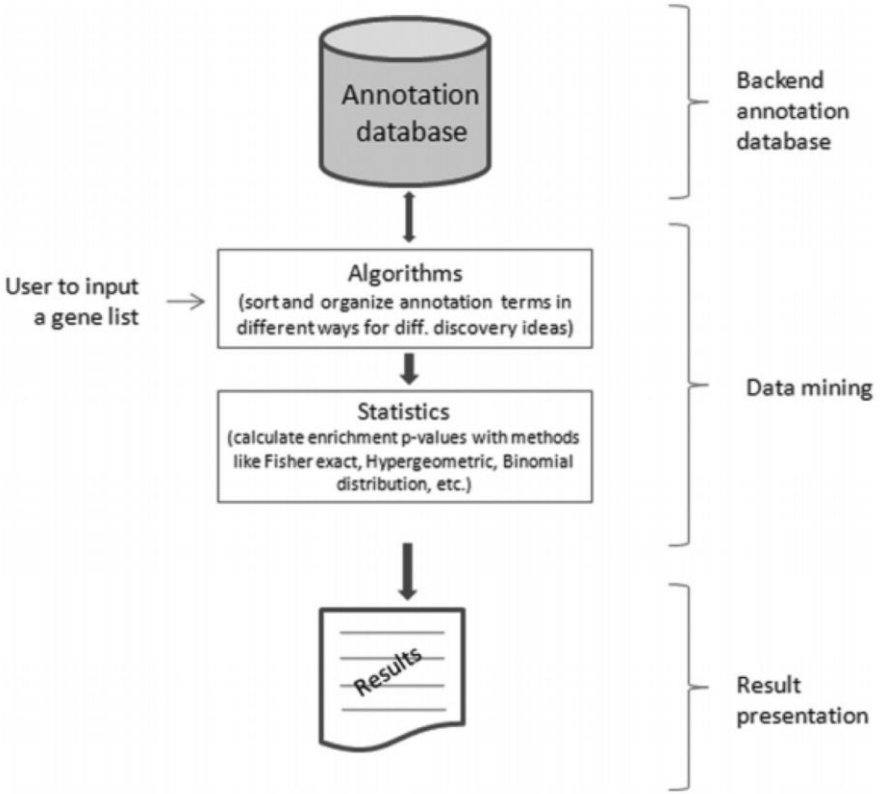
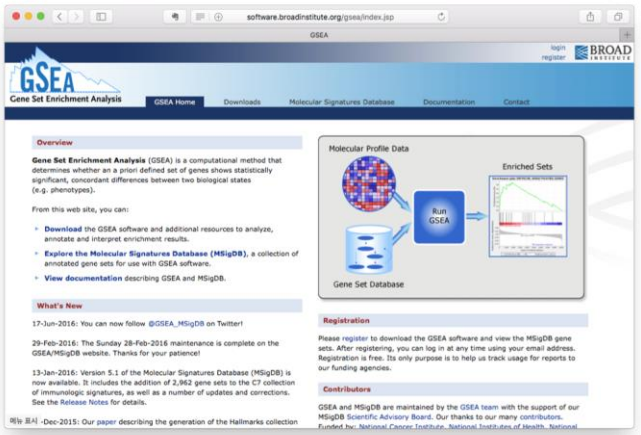


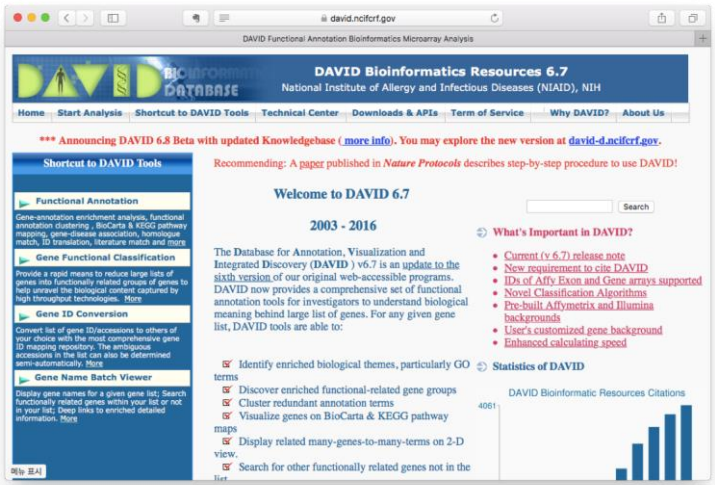
Figure 1. The infrastructure of typical enrichment tools. Even though the enrichment analysis tools have distinct features, they can be generally described as three major layers: backend annotation database; data mining; and result presentation. Each of the layers, rather than statistical methods alone, greatly influences the analytic results.

Da Wei Huang et al., Nucleic Acids Research, 2009

- GSEA(Gene Set Enrichment Analysis) – Broad Institute

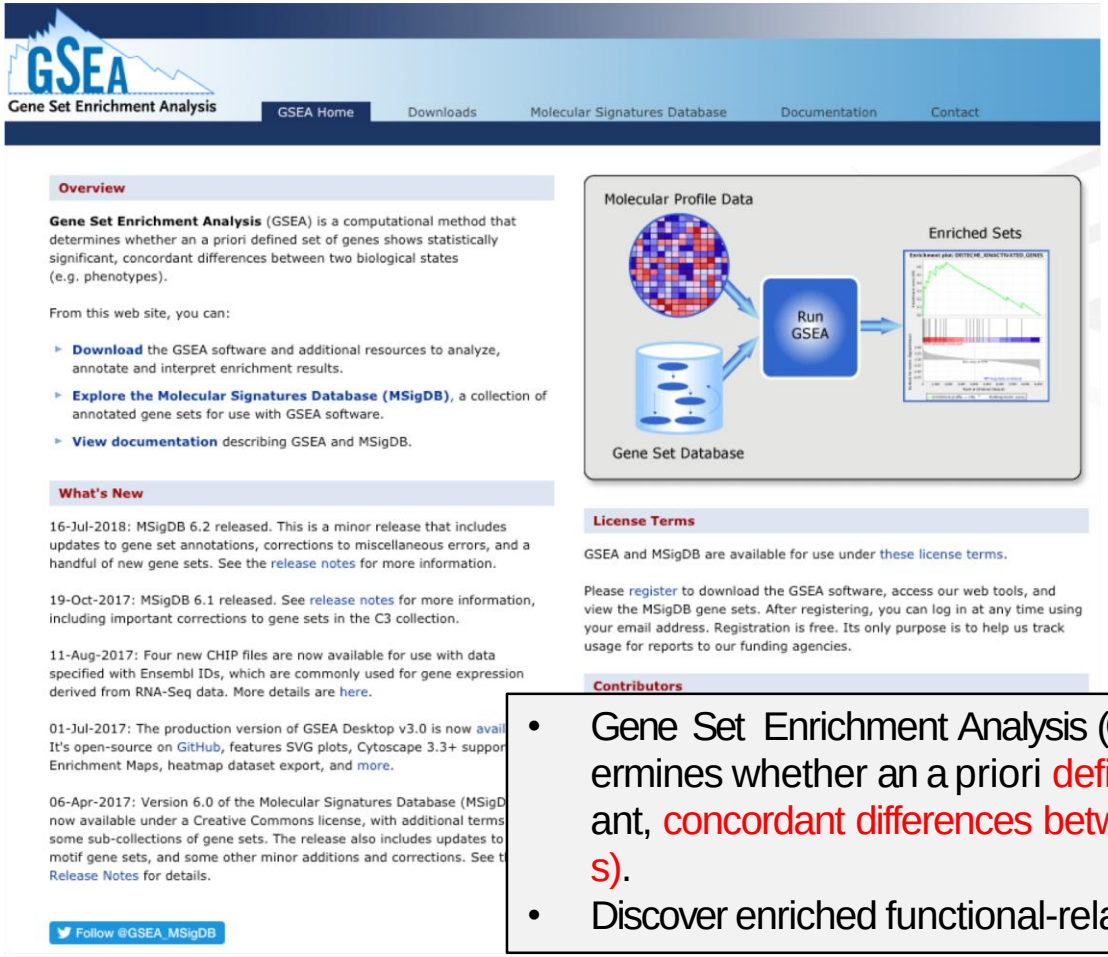


- DAVID



Gene Set Enrichment Analysis

The Broad Institute website is in cooperation with MSigDB and has a downloadable GSEAsoftware



MSigDB
Molecular Signatures Database

The 17810 gene sets in the Molecular Signatures Database (MSigDB) are divided into 8 major collections, and several sub-collections.

H: hallmark gene sets
C1: positional gene sets
C2: curated gene sets
C3: motif gene sets
C4: computational gene sets
C5: GOgene sets
C6: oncogenic gene sets C
7: immunologic gene sets

- Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).
- Discover enriched functional-related gene groups (i.e. GOterms)

http://software.broadinstitute.org/gsea http://software.broadinstitute.org/gsea/doc/desktop_tutorial.jsp

GSEA method

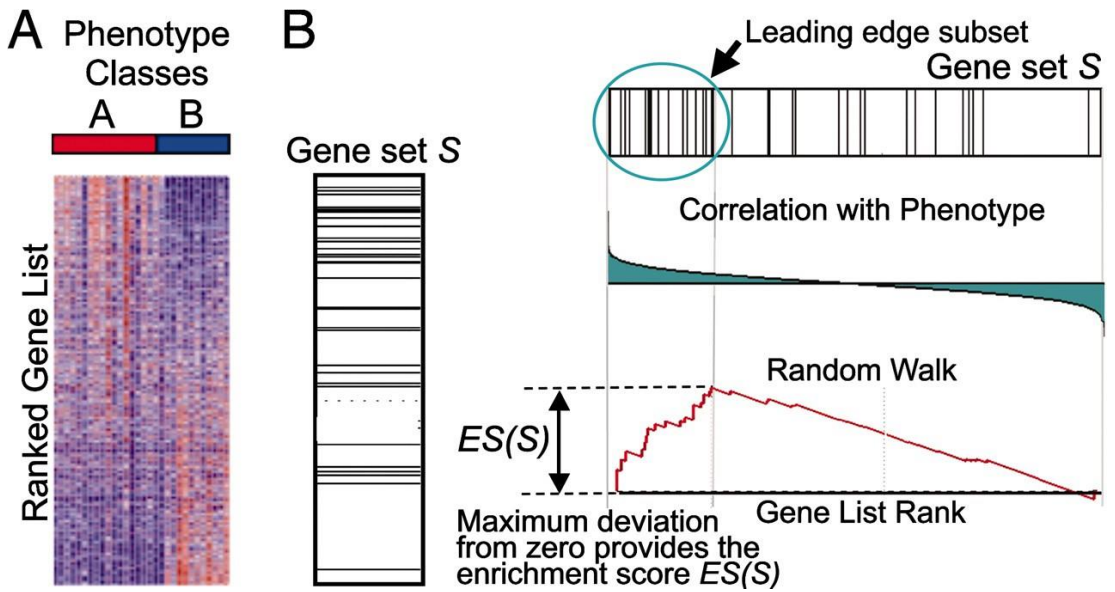
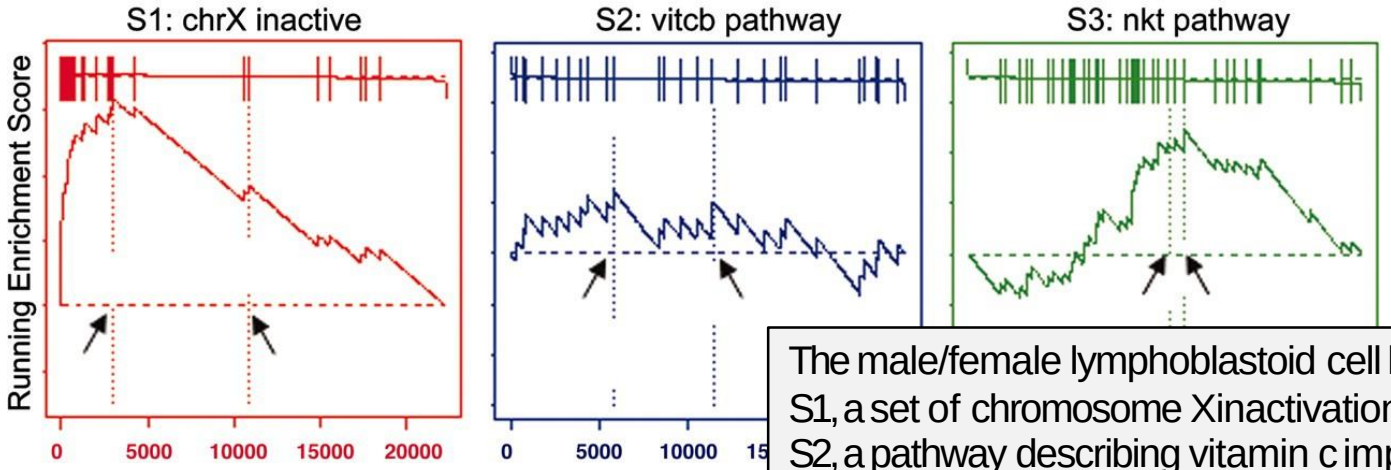


Fig. 1.
A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set S within the sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.



The male/female lymphoblastoid cell line
S1, a set of chromosome X inactivation genes
S2, a pathway describing vitamin c import into neurons
S3, related to chemokine receptors expressed by Thelper cells

Aravind Subramanian et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS October 25, 2005

Database for Annotation, Visualization and Integrated Discovery

DAVID goes beyond standard GSEA with additional functions like switching between gene and protein identifiers on the genome-wide scale, however, it is important to note that the annotations used by DAVID have not been updated since January 2010 which can have a considerable impact on practical interpretation of results. In October 2016, DAVID Knowledgebase version 6.8 was released with a complete rebuild.

PROTOCOL

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources

Da Wei Huang[1,2], Brad T Sherman[1,2] & Richard A Lempicki[1]

[1]Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to R.A.L. (rlempicki@mail.nih.gov) or D.W.H. (huangdawei@mail.nih.gov)

Published online 18 December 2008; doi:10.1038/nprot.2008.211

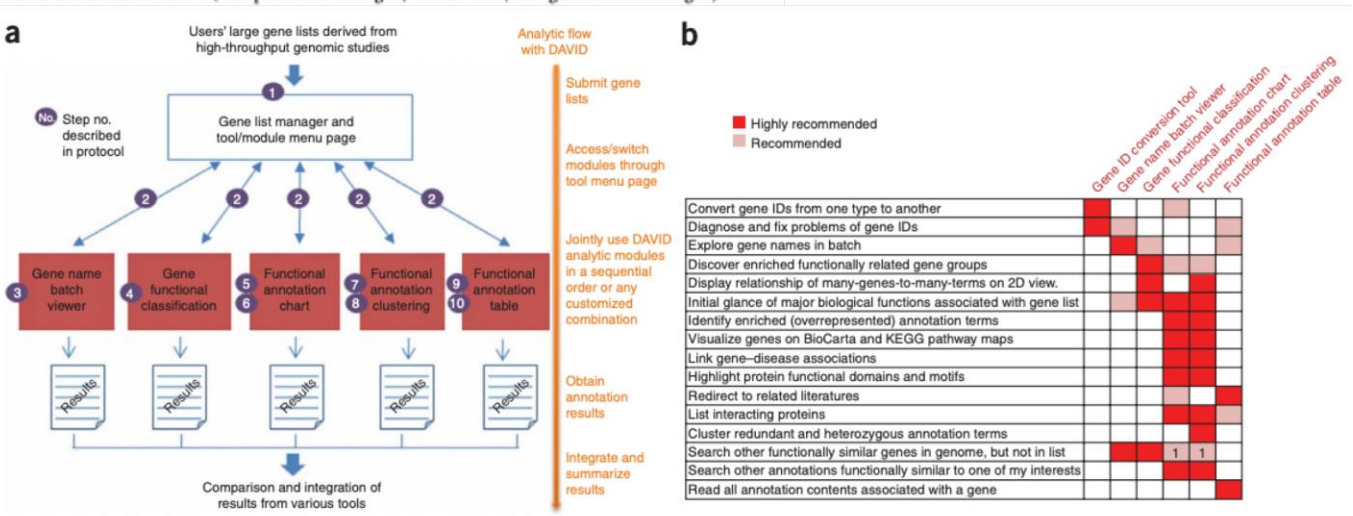
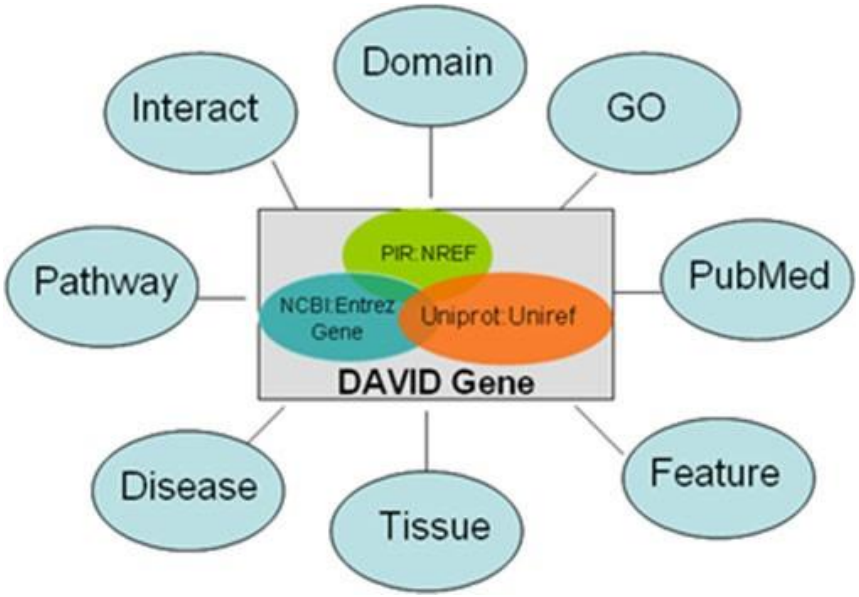
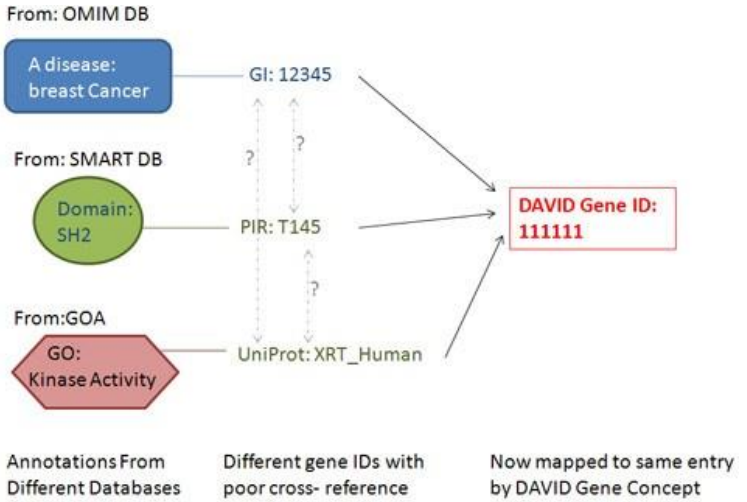


Figure 6 | Analytic tools/modules in DAVID. (**a**) After the user's gene list is submitted to DAVID, the gene list manager may be accessed by all DAVID analytic modules (red boxes) at any time. The circled numbers indicate step numbers described in PROCEDURE to facilitate reading. (**b**) DAVID analytic modules, each having different strengths and focus, can be used independently or jointly. A roadmap to help users to choose some or all DAVID analytic modules for the analysis of large gene lists.

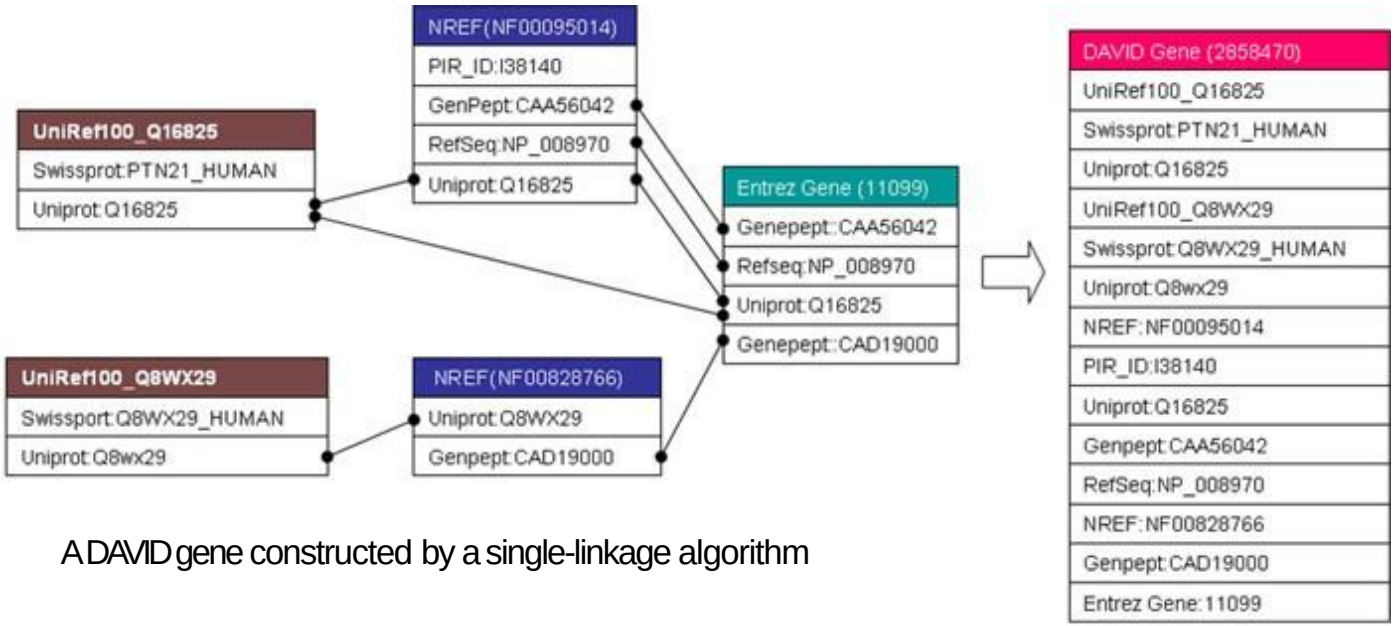
SUPER-Powerful "DAVID Knowledgebase"



Hypothetical Illustration of DAVID Knowled gebase centralized by DAVID genes



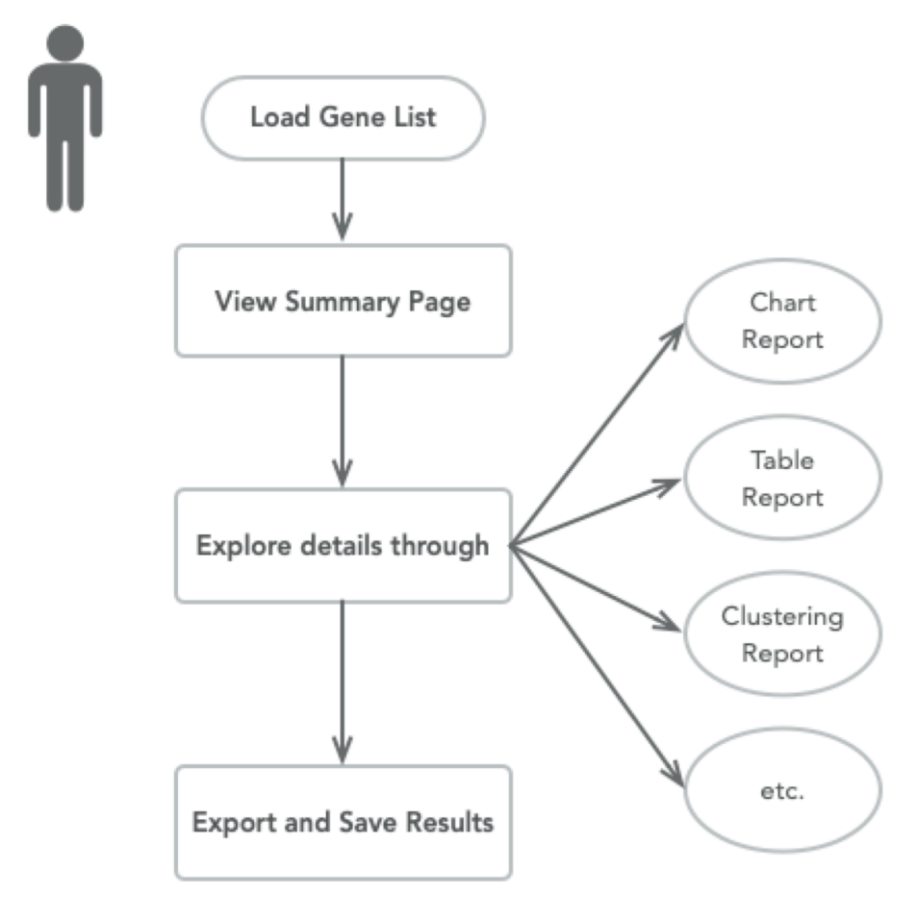
How Are Annotations Assigned to DAVIDGene?



A DAVID gene constructed by a single-linkage algorithm

Enrichment analysis by DAVID

Typical Analysis Flow



EASEScore, a modified Fisher Exact P-Value

- 2x2 contingency table

	User Genes	Genome
In Pathway	3-1*	40
Not In Pathway	297	29960

- Pathway : p53 signaling pathway
- Human genome background (30,000 genes)
- Belong to p53 signaling pathway (300 genes)
- *minimum 2 genes per Pathway

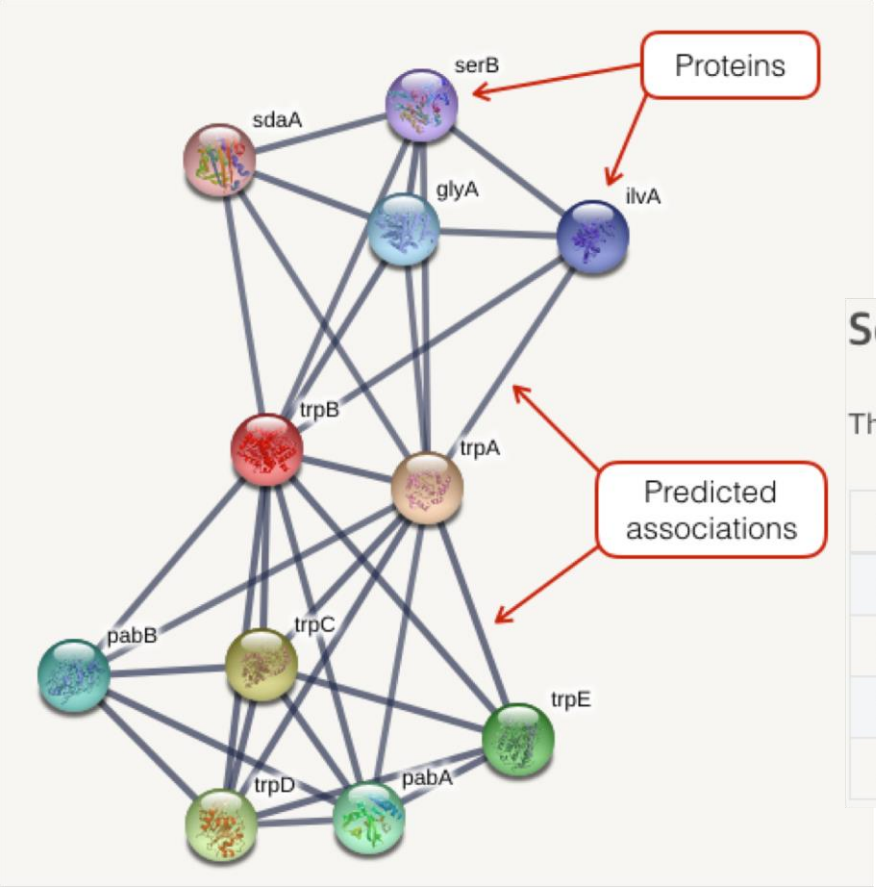
$$\text{enrichment} = \frac{m/n}{M/N}$$

N = all genes (universe)
M = all genes belonging to a pathway
n = your gene list
m = genes of your gene list that belongs to the pathway

$$\text{P-value} = \sum_{m}^{\min(K,n)} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

Functional annotation for DEGs: Network analysis

- STRING is a biological database and web resource of known and predicted protein–protein interactions (interaction score > 0.7 high confidence)



Schemas

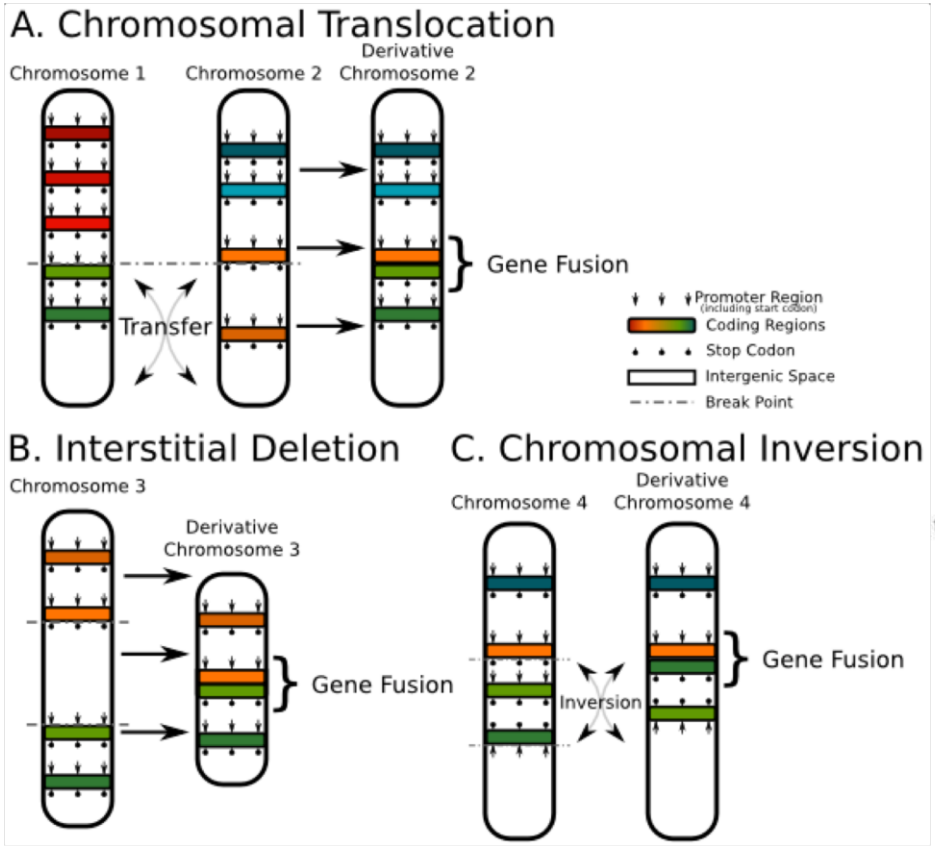
There are four schemas in STRING that describes different aspects of the content.

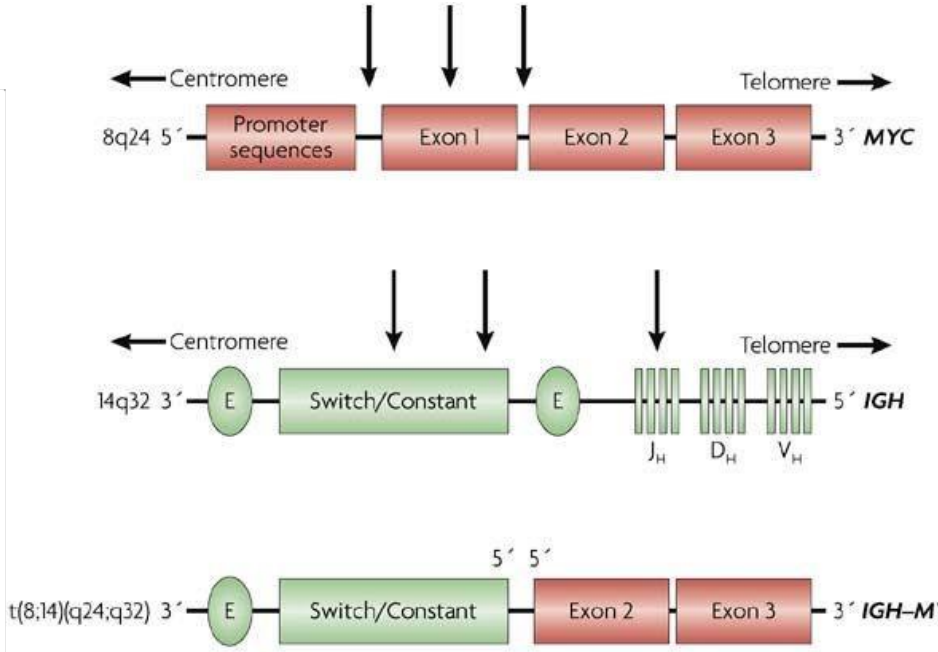
schema	description
evidence	contains info of the underlying evidence for interactions.
homology	blast hits used to propagate evidence to other species by means of homology.
items	info about entries (protein names, species, orthgroups, etc.).
network	the interactions and their scores.

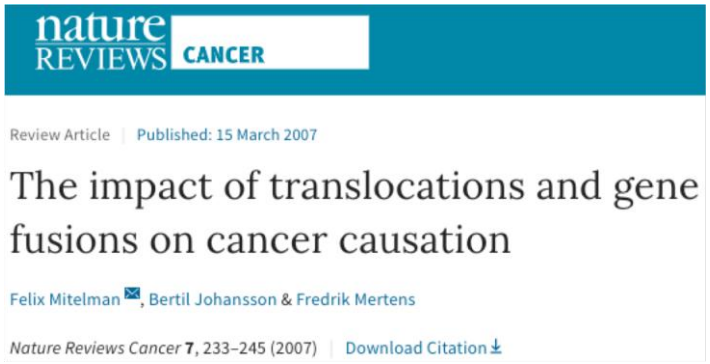
Gene fusion analysis

A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as a result of: translocation, interstitial deletion, or chromosomal inversion.

* BioTools : fusionCatcher, Defuse, EricScript, Tophat-fusion,







RNA-Seq VS scRNA-Seq (Single-cell sequencing)

Bulk RNA-Seq

Measures the average expression level for each gene across a large population of input cells

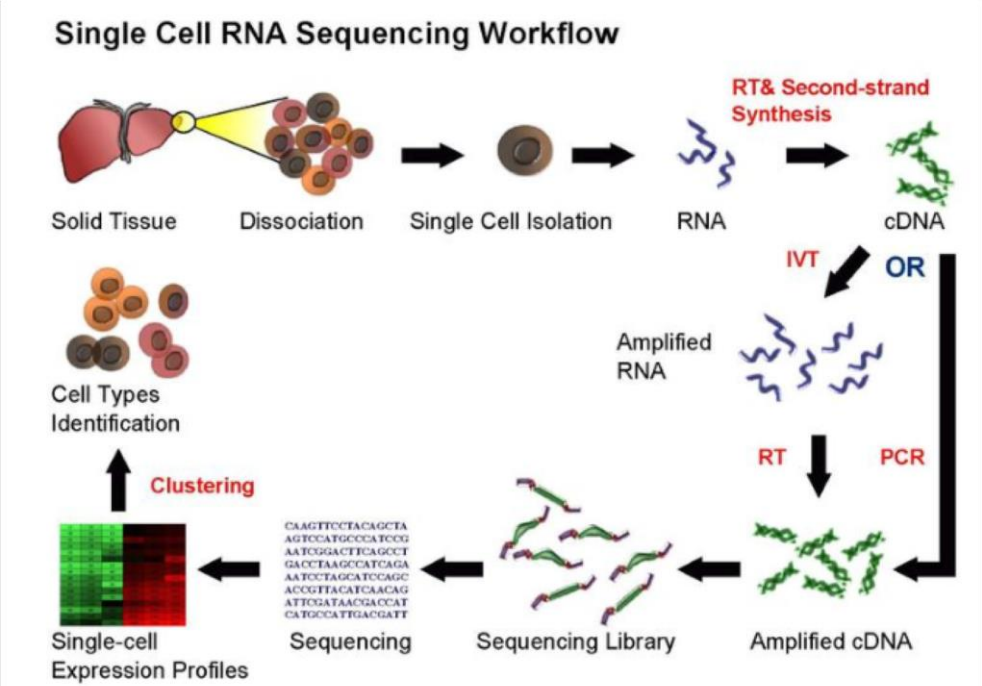
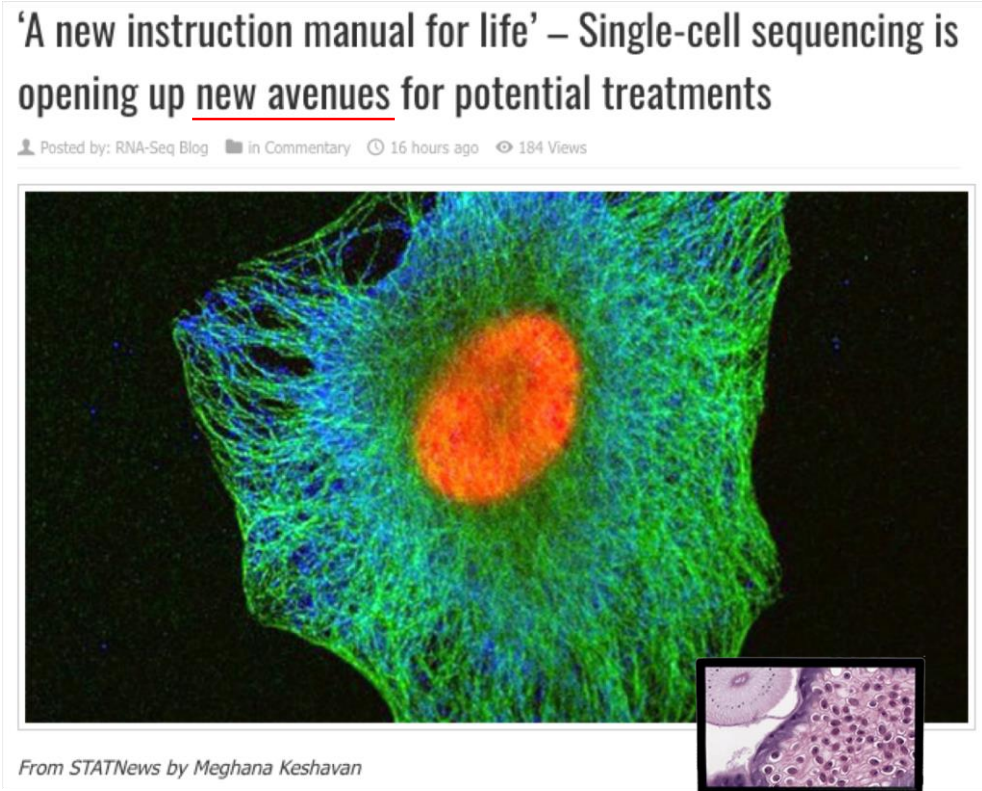
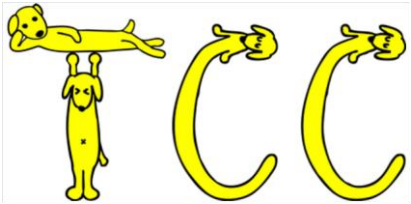


Figure 2.1: Single cell sequencing (taken from Wikipedia)



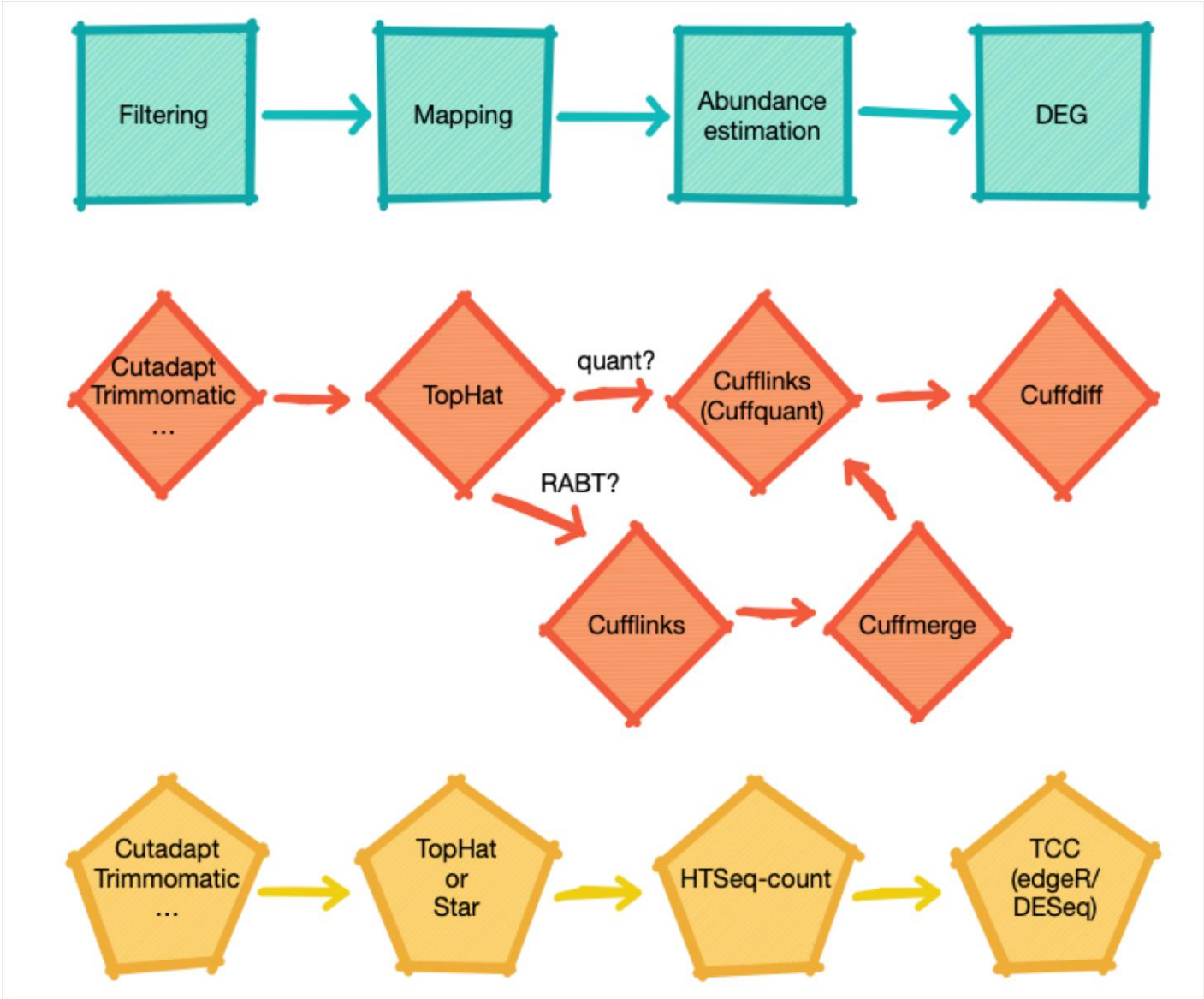
Transcriptome Analysis Using RNA-Seq

Differential Gene Expression Analysis : TCC



유승일 (Seung-il Yoo)

Basic process of RNA-seq Analysis



HTSeq

A framework to process and analyze data from high-throughput sequencing (HTS) assays.

- Development: https://github.com/simon-anders/htseq
- Documentation: http://htseq.readthedocs.io

설치

(https://pypi.org/project/HTSeq/)

```
>>> import HTSeq
>>> HTSeq.FastqReader("filename", "solexa")
```

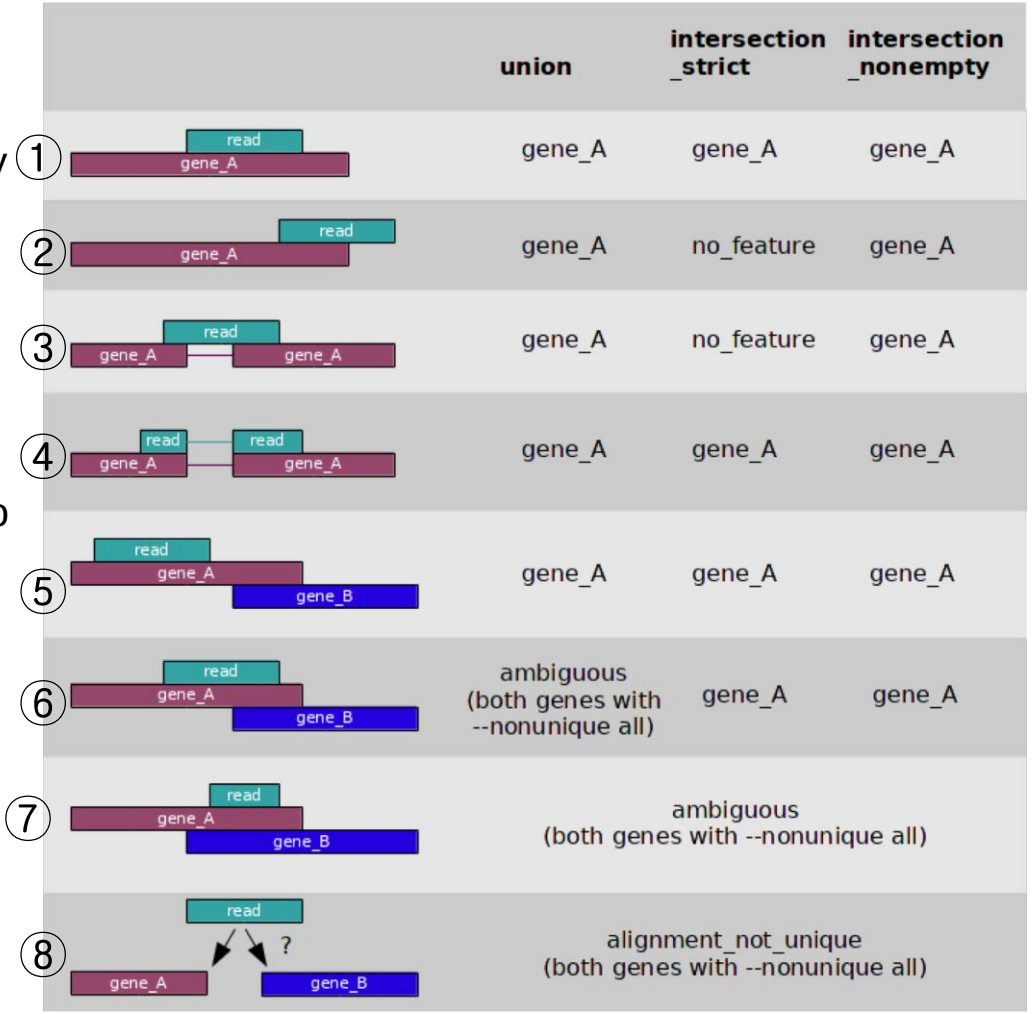
완료

```
    100% |                                | 13.8MB 63kB/s
Installing collected packages: pysam, numpy, HTseq
Successfully installed HTseq-0.11.0 numpy-1.15.4 pysam-0.15.1
```

The Python Package Index (PyPI) is a repository of software for the Python programming language.
pip installation : (https://pip.pypa.io/en/stable/installing/)

HTSeq-count

- To count how many reads map to each feature
- Not counted for any feature for various reasons, namely:
  - no_feature : reads which could not be assigned to any feature
  - ambiguous : reads which could have been assigned to more than one feature and hence were not counted for any of these
  - too_low_aQual : reads which were not counted due to the –a option
  - not_aligned : reads in the SAM file without alignment
  - alignment_not_unique : reads with more than one reported alignment. These reads are recognized from the NH optional SAM field tag.
- If you have paired-end data, you have to sort the SAM (BAM) file by read name first.



(https://htseq.readthedocs.io/en/release_0.10.0/count.html)

HTSeq-count

```
Usage: htseq-count [options] alignment_file gff_file

Options:
  -f SAMTYPE, --format=SAMTYPE
                        type of <alignment_file> data, either 'sam' or 'bam' (default: sam)
  -r ORDER, --order=ORDER
                        'pos' or 'name'. Sorting order of <alignment_file> (default: name).
                        Paired-end sequencing data must be sorted either by position or by read name,
                        and the sorting order must be specified. Ignored for single-end data.
  -s STRANDED, --stranded=STRANDED
                        whether the data is from a strand-specific assay. Specify 'yes', 'no', or 'reverse' (default: yes).
                        'reverse' means 'yes' with reversed strand interpretation
  -a MINAQUAL, --minaqual=MINAQUAL
                        skip all reads with alignment quality lower than the given minimum value (default: 10)
  -t FEATURETYPE, --type=FEATURETYPE
                        feature type (3rd column in GFF file) to be used, all features of other type are ignored
                        (default, suitable for Ensembl GTF files: exon)
  -i IDATTR, --idattr=IDATTR
                        GFF attribute to be used as feature ID (default, suitable for Ensembl GTF files: gene_id)
  -m MODE, --mode=MODE
                        mode to handle reads overlapping more than one feature
                        (choices: union, intersection-strict, intersection-nonempty; default: union)
  -o SAMOUT, --samout=SAMOUT
                        write out all SAM alignment records into an output SAM file called SAMOUT, ann
                        otating each line with its feature assignment (as an optional field with tag 'XF')
```

HTSeq-count

```
Usage:
htseq-count [options] alignment_file gff_file
```

```
02.Align/
├── con
│   ├── Root-Control-12h
│   ├── Root-Control-24h
│   └── Root-Control-6h
└── exp
    ├── Root-PA01-12h
    ├── Root-PA01-24h
    └── Root-PA01-6h
        ├── accepted_hits.bam
        ├── align_summary.txt
        ├── deletions.bed
        ├── insertions.bed
        ├── junctions.bed
        ├── logs
        ├── prep_reads.info
        └── unmapped.bam
```

실행

```
$ samtools sort –n ./02.Align/con/Root-Control-12h/accepted_hits.bam \
                ./02.Align/con/Root-Control-12h/accepted_hits.nameSorted

$ mkdir 03.TCC
$ htseq-count –f bam –r name –s no –m union \
              –o ./02.Align/con/Root-Control-12h/accepted_hits.nameSorted.bam.SAMOUT \
              ./02.Align/con/Root-Control-12h/accepted_hits.nameSorted.bam \
              ./BioResources/Arabidopsis_thaliana.TAIR10.38.gtf \
              > ./03.TCC/Root-Control-12h.count
```

```
AT1G01010       395
AT1G01020       144
AT1G01030       4
AT1G01040       459
AT1G01046       0
AT1G01050       750
AT1G01060       1795
AT1G01070       169
AT1G01080       89
AT1G01090       1809
```

```
ENSRNA049758187 0
ENSRNA049758190 0
ENSRNA049758191 0
ENSRNA049758193 0
ENSRNA049758194 0
__no_feature    71025
__ambiguous     384882
__too_low_aQual 0
__not_aligned   0
__alignment_not_unique  333312
```

확인

```
$ less –S ./03.TCC/Root-Control-12h.count
$ head ./03.TCC/Root-Control-12h.count
$ tail ./03.TCC/Root-Control-12h.count
```

Make Read Count Matrix

실행

```
$ python htseq_count_merger.py -h
usage: htseq_count_merger.py [-h] [--infiles INFILES [INFILES ...]]
                             [--outprefix OUTPREFIX]

Merge output of HTSeq-count

optional arguments:
  -h, --help            show this help message and exit
  --infiles INFILES [INFILES ...]
                        file name must be [PATH]/[sample name].count
  --outprefix OUTPREFIX
```

input

```
03.TCC/
├── Root-Control-12h.count
├── Root-Control-24h.count
├── Root-Control-6h.count
├── Root-PA01-12h.count
├── Root-PA01-24h.count
└── Root-PA01-6h.count
```

output

```
03.TCC/
├── count.raw.mtx
└── count.norm.mtx
```

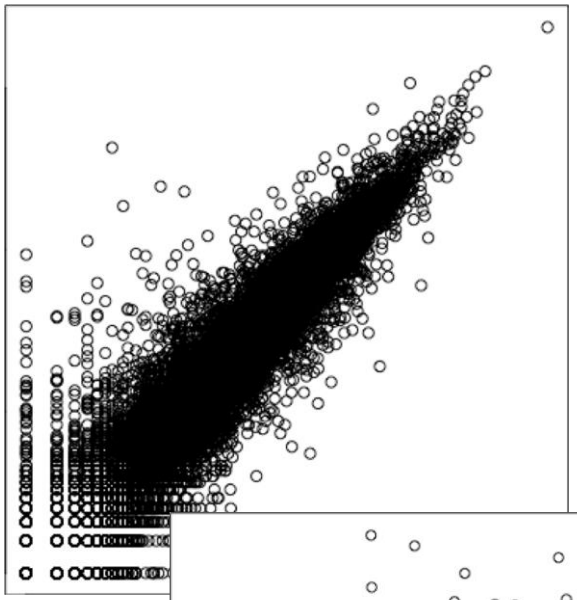
[OUTPREFIX].raw.mtx : Count matrix
[OUTPREFIX].norm.mtx : Normalized Matrix (10^9 * ReadCount / TotalReadCount)

ID	Root-Control-12h		Root-Control-24h		Root-
AT1G01010	395	295	262	490	370	375
AT1G01020	144	151	133	148	108	136
AT1G01030	4	7	5	10	4	9
AT1G01040	459	460				
AT1G01046	0	0				
AT1G01050	750	660				
AT1G01060	1795	265				
AT1G01070	169	148				
AT1G01080	89	80				

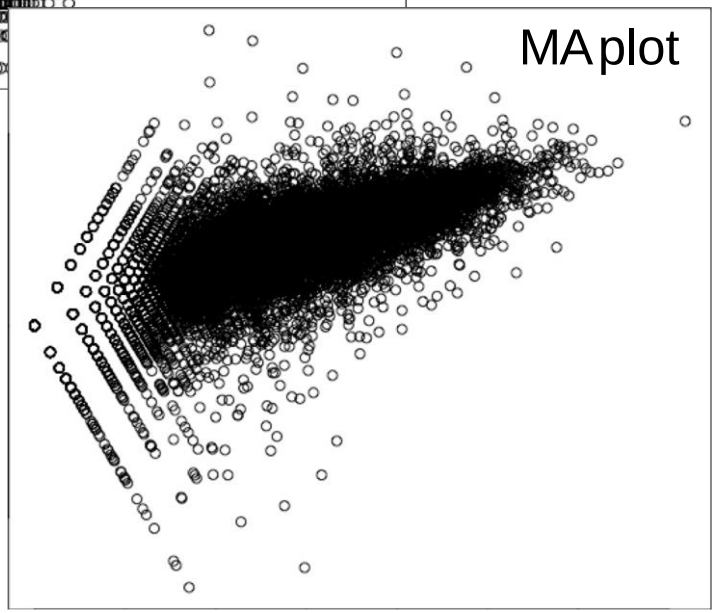
ID	Root-Control-12h		Root-Control-24h		Root-Control-6h	Root-PA01-12h	Root-PA01-24h	Root-PA01-6h
AT1G01010	0.000423		0.000324		0.00029	0.000546	0.000509	0.000409
AT1G01020	0.000155		0.000165		0.000146	0.000167	0.000147	0.000148
AT1G01030	5e-06	6e-06	5e-06	1e-05	7e-06	1.1e-05		
AT1G01040	0.000492		0.000507		0.000369	0.000458	0.000501	0.000388
AT1G01046	1e-06	1e-06	1e-06	1e-06	2e-06	1e-06		
AT1G01050	0.000801		0.000729		0.000739	0.000436	0.000561	0.000452
AT1G01060	0.001919		0.000294		0.000173	0.001486	0.000251	0.000167
AT1G01070	0.000182		0.000164		0.000154	0.000273	0.000321	0.000194
AT1G01080	9.6e-05	8.9e-05	6.6e-05	5.9e-05	6.2e-05	4.9e-05		

Gene Expression visualization

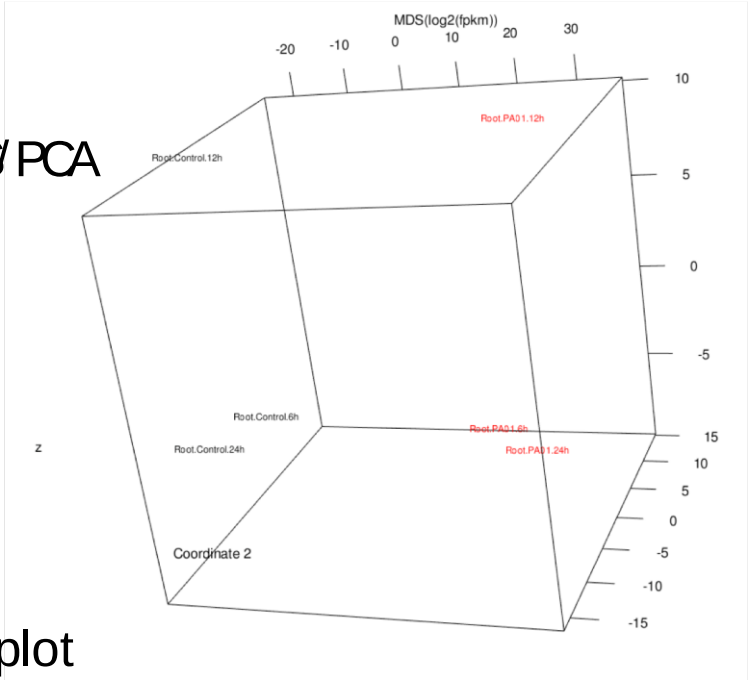
Scatter plot



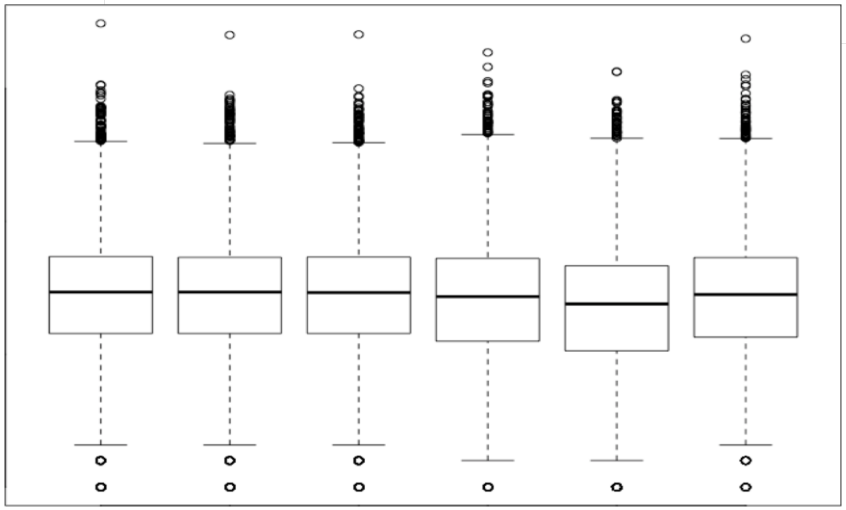
MA plot



MDS/ PCA



Boxplot



Hands on I

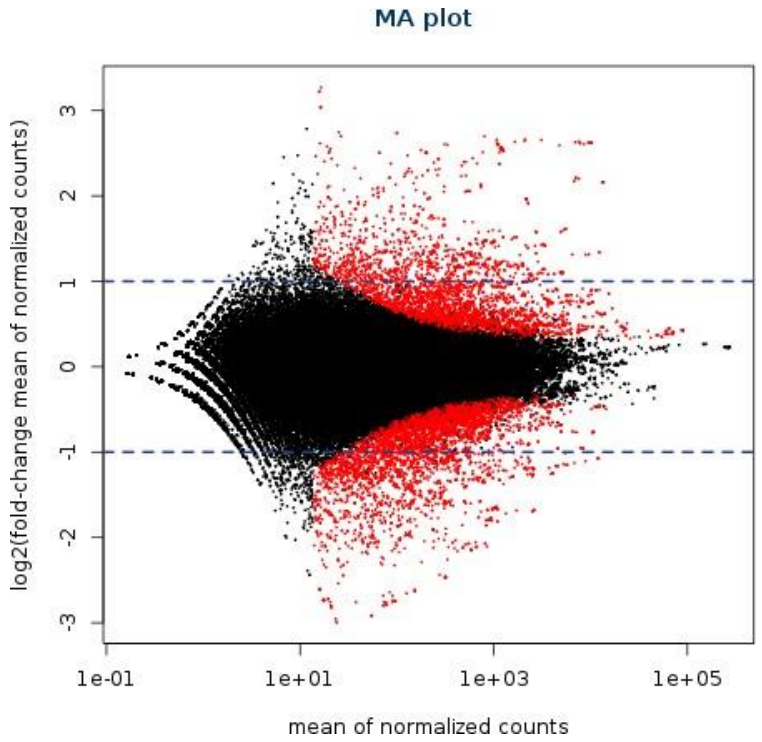
2. bash

bash ⌘1

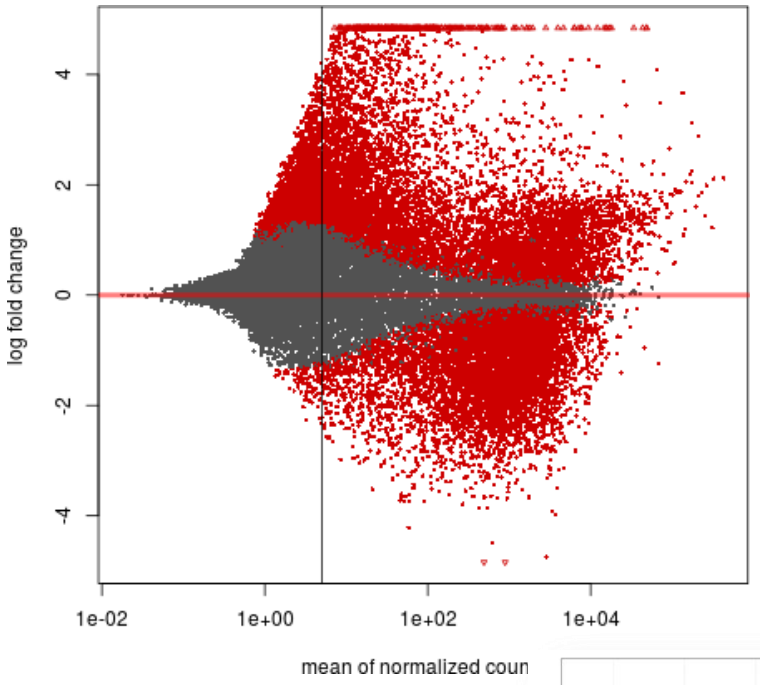
```
Seung-ilui-MacBook-Air:~ Yoo$

  # load data
  data <- read.table("03.TCC/count.raw.mtx", header = T,row.names = 1)  da
  ta2 <- read.table("03.TCC/count.norm.mtx", header = T,row.names =1)

  dim(data)  sum
  mary(data)

  # Scatter plot
  plot(data$Root.Control.6h, data$Root.PA01.6h)  plot(data2$Ro
  ot.Control.6h, data2$Root.PA01.6h)  plot(log(data$Root.Contro
  l.6h,2), log(data$Root.PA01.6h,2))  plot(log(data$Root.Control.
  24h,2), log(data$Root.PA01.24h,2))

  # boxplot
  boxplot(data, xlab="Samples", ylab="ReadCount")  boxplot(log(data,2)
  , xlab="Samples", ylab="log2(ReadCount)")  boxplot(log(data2,2), xlab
  ="Samples", ylab="log2(norm(ReadCount))")

  # MA plot   plot((log(data$Root.Control.6h,2)+log(data$Root.PA
  01.6h,2))/2,
      log(data$Root.Control.6h,2)-log(data$Root.PA01.6h,2))
```

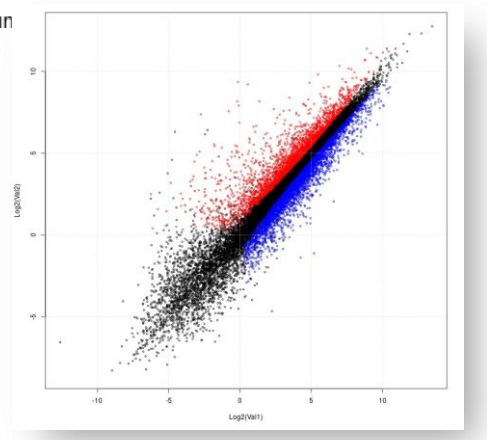
Selection of Differential Expressed Genes

Good



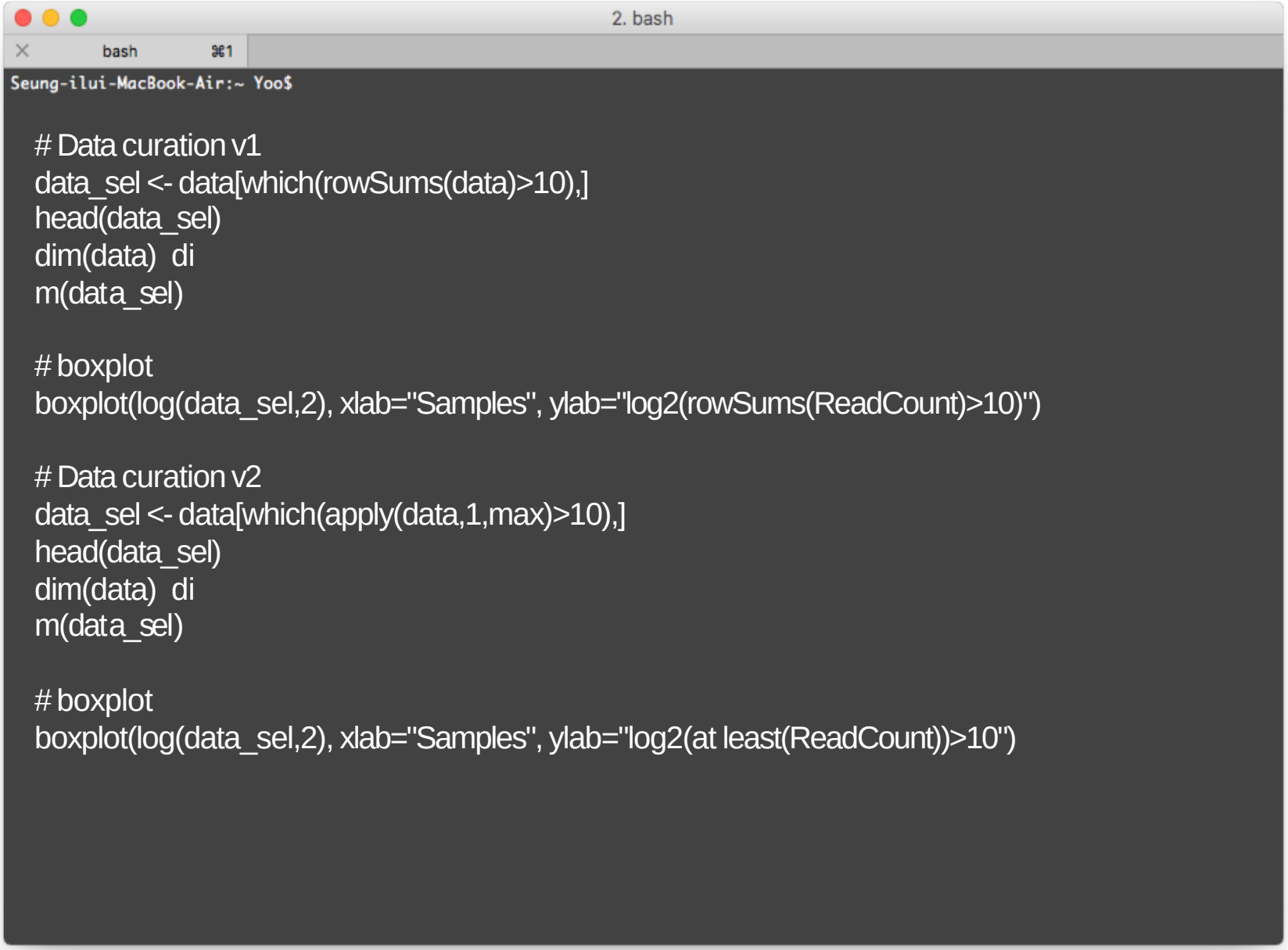
Bad





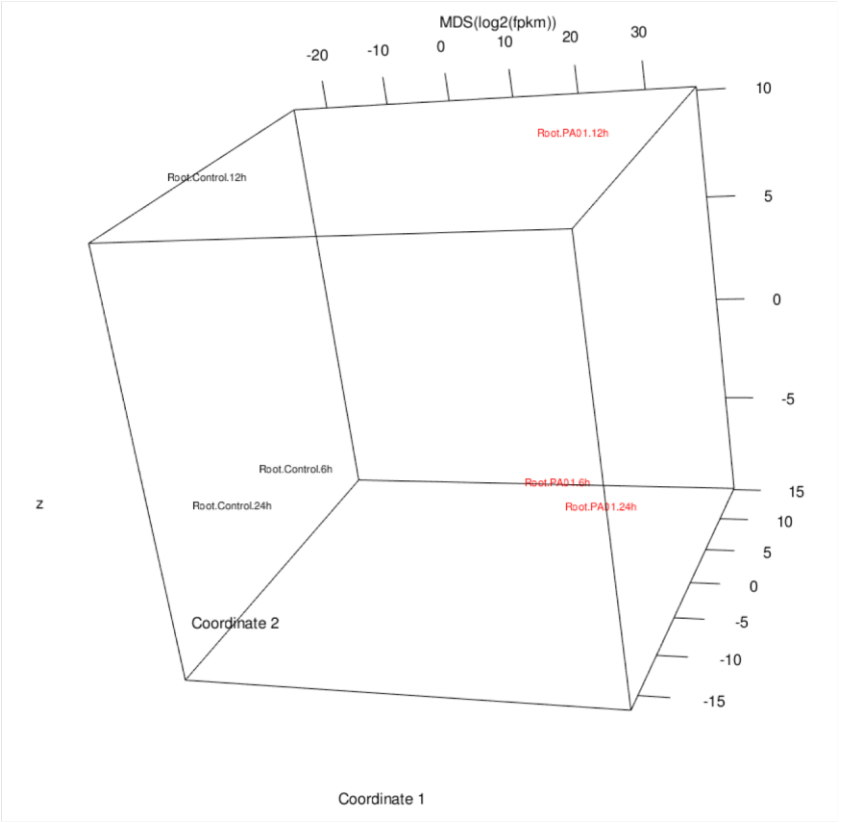
- P<0.05 or Q<0.01;
- ≥ 1.5 ~ 2 fold change;
- Min. FPKM in pairwise comparison ≥ 0.3 ~ 1 FPKM

Hands on II

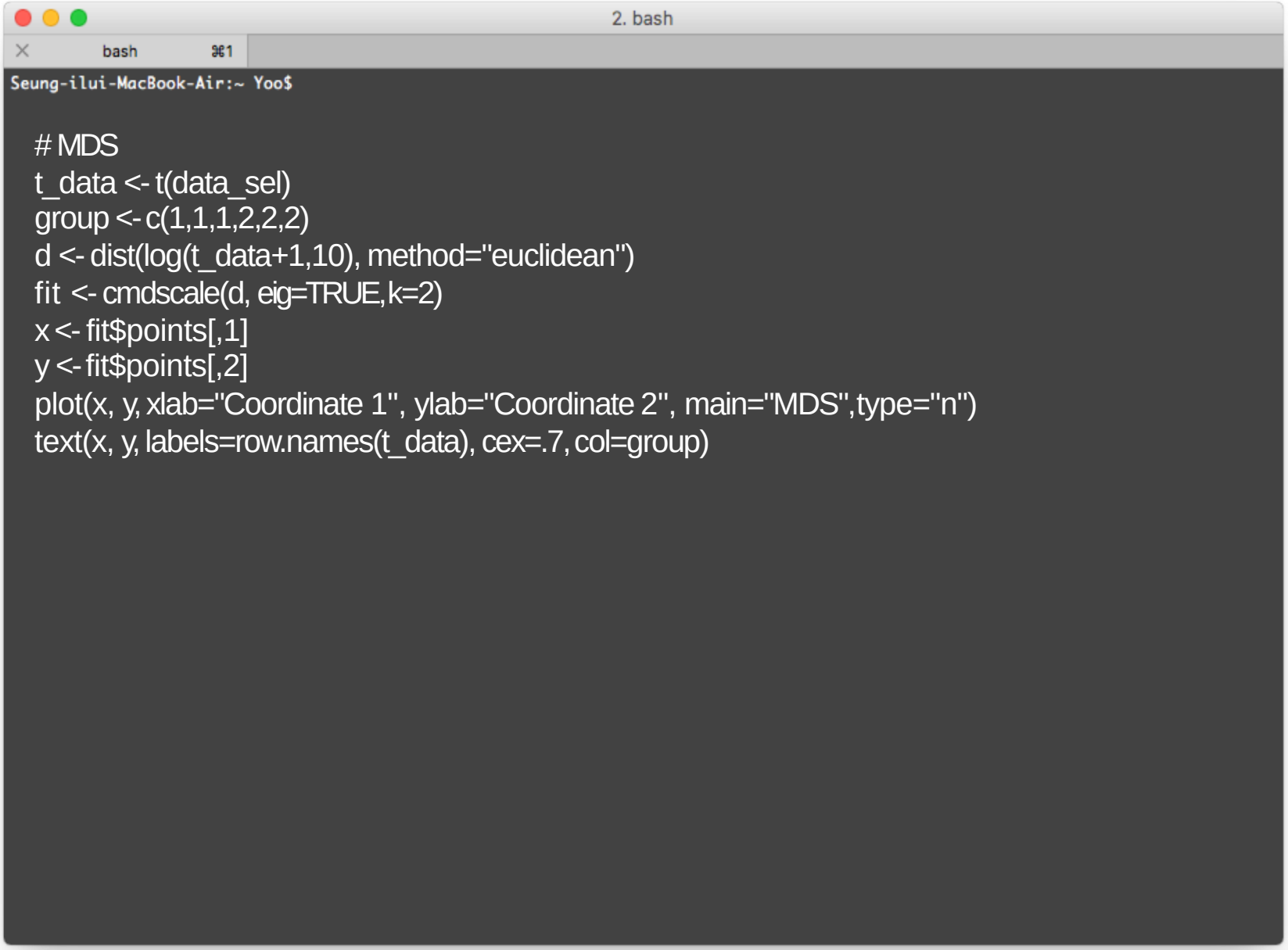

2. bash
bash ⌘1
Seung-ilui-MacBook-Air:~ Yoo$
# Data curation v1
data_sel <- data[which(rowSums(data)>10),]
head(data_sel)
dim(data) di
m(data_sel)
# boxplot
boxplot(log(data_sel,2), xlab="Samples", ylab="log2(rowSums(ReadCount)>10)")
# Data curation v2
data_sel <- data[which(apply(data,1,max)>10),]
head(data_sel)
dim(data) di
m(data_sel)
# boxplot
boxplot(log(data_sel,2), xlab="Samples", ylab="log2(at least(ReadCount))>10")

MDS (multidimensional scaling)

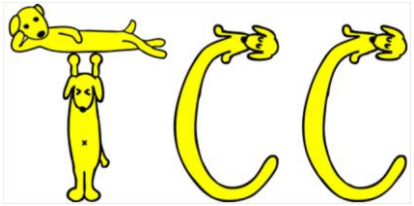
- 다차원척도법
- 여러 대상간의 객관적 또는 주관적 관계에 관한 수 치적 자료들을 처리하여 다차원 공간상에서 그 대 상들을 위치적으로 표시하여 주는 일련의 통계기법
- 수치적 자료만을 가지고는 알 수 없는 전체적인 관 계구조를 공간상의 그림을 통해 쉽게 파악할 수 있 게 한다.
- 참고자료
  - http://blog.naver.com/fox4361/20155261330
  - 장익진, 1998



Hands on III


2. bash
bash ⌘1
Seung-ilui-MacBook-Air:~ Yoo$
# MDS
t_data <- t(data_sel)
group <- c(1,1,1,2,2,2)
d <- dist(log(t_data+1,10), method="euclidean")
fit <- cmdscale(d, eig=TRUE,k=2)
x <- fit$points[,1]
y <- fit$points[,2]
plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2", main="MDS",type="n")
text(x, y, labels=row.names(t_data), cex=.7, col=group)

TCC: Differential expression analysis for tag count data with robust normalization strategies



- TCC(Tag Count Comparison)
- edgeR, DESeq 등의 기존 방법을 사용
- Replication을 고려한 통계 처리

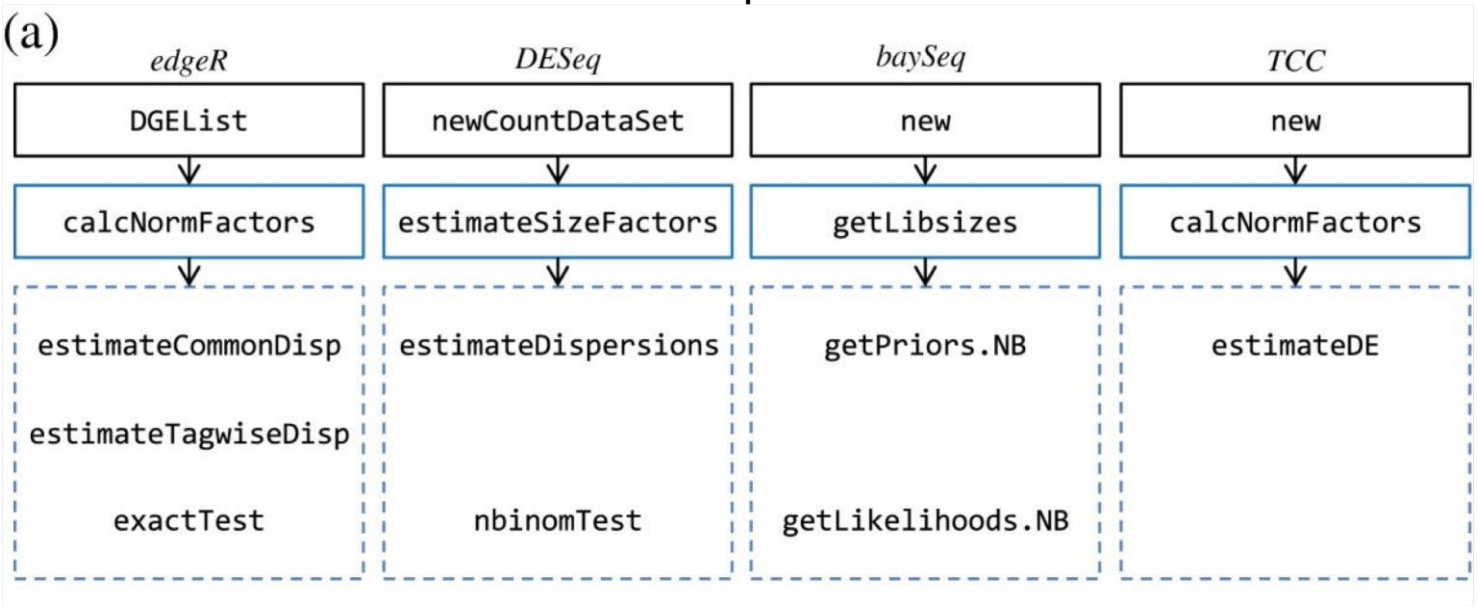


Figure 1
DEGES-based analysis pipelines in TCC. (a) Main functions for obtaining DE results from tag count data in individual packages (edgeR, DESeq, baySeq, and TCC). The analysis pipelines of the packages can be roughly divided into two steps after importing the input data (black squares):
(i) calculating normalization factors (blue solid squares)
(ii) estimating degrees of DE for each gene (blue dashed squares)

DOI: 10.18129/B9.bioc.TCC Bioconductor version: Release (3.8), R(version "3.5"), update: 2018.11.21 Sun et al. BMC Bioinformatics 2013

TCC: Differential expression analysis for tag count data with robust normalization strategies

- DEGES: DEG elimination strategy (ranking system)
- 6 combination (2 normalization methods, 3 DEG identification methods)
  - DEGES/TbT: TMM-(baySeq-TMM)n pipeline
  - DEGES/edgeR: TMM-(edgeR-TMM)n pipeline

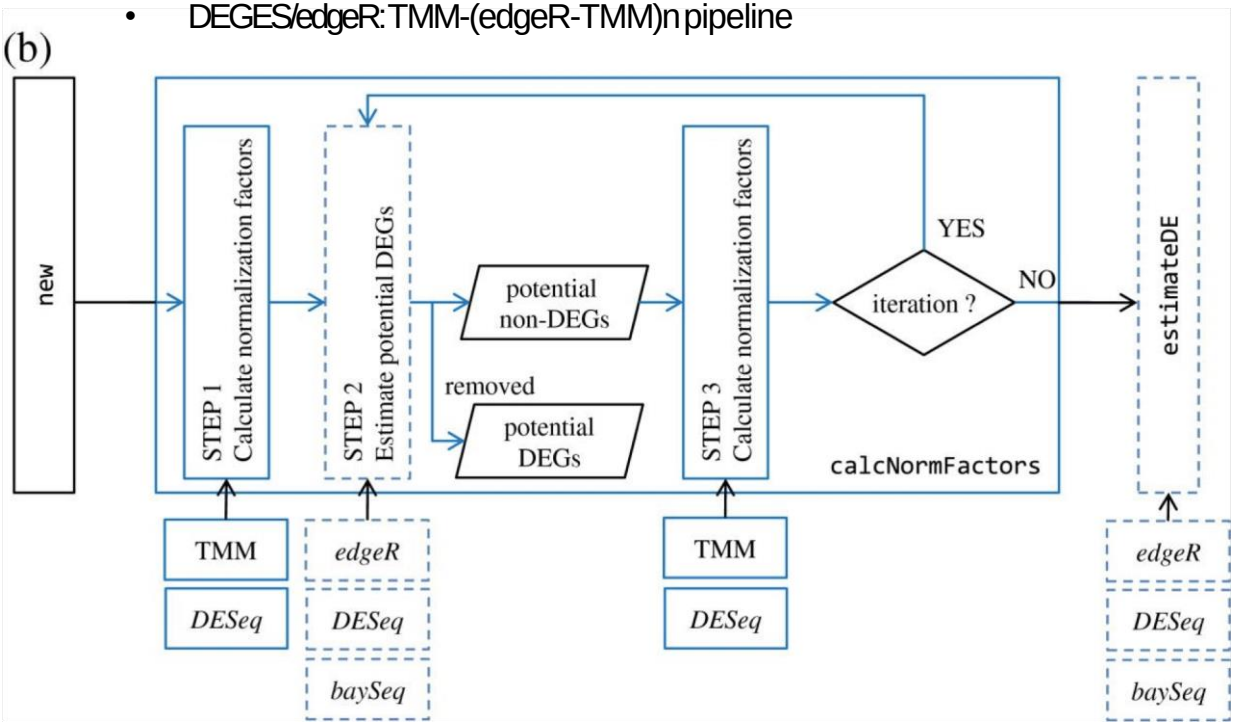


Figure 1
DEGES-based analysis pipelines in TCC. (b) Outline of the DEGES-based normalization methods implemented in TCC. The key concept of DEGES in our calcNormFactors function is to remove data flagged as potential DEGs in step 2 before calculating normalization factors in step 3. Note that steps 2 and 3 in DEGES can be repeatedly performed in order to obtain more robust normalization factors and the function accepts many iterations n (i.e., n = 0 ~ 100).

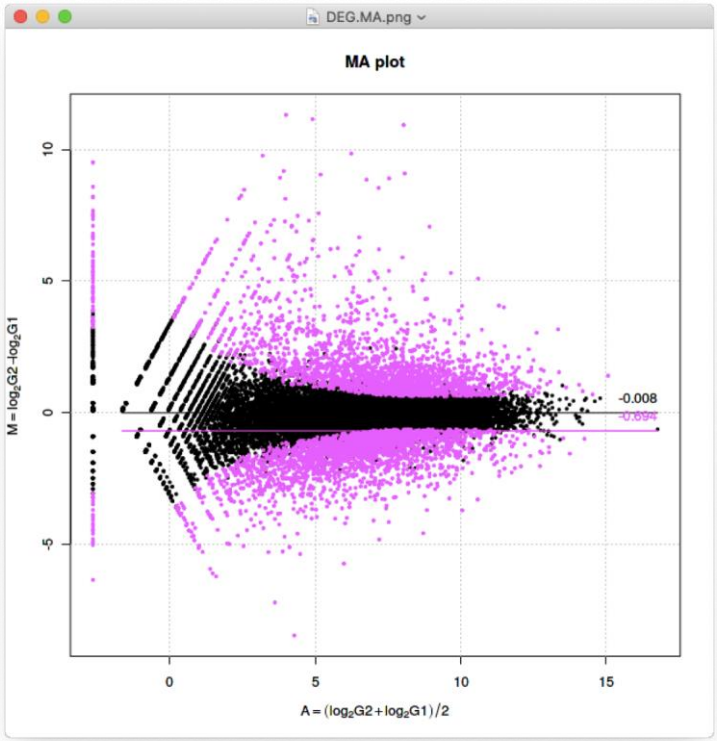
DEG Analysis with TCC

```
03.TCC/
├── count.raw.mtx
├── count.norm.mtx
├── Root-Control-12h.count
├── Root-Control-24h.count
├── Root-Control-6h.count
├── Root-PA01-12h.count
├── Root-PA01-24h.count
└── Root-PA01-6h.count
```

Rscript (TCC : iDEGES/degeR)

```
library(TCC)

raw_count <- read.table("03.TCC/count.raw.mtx", header = T,row.names =1)
sel_col <- c(1,2,3,4,5,6)
group <- c(1,1,1,2,2,2)

tcc <- new("TCC", raw_count, group)
tcc <- filterLowCountGenes(tcc)

tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger", iteration=3, FDR=0.1, floorPDEG=0.05)
tcc <- estimateDE(tcc, test.method="edger", FDR=0.1)

eff_count <- getNormalizedData(tcc)
write.table(round(eff_count,3), "03.TCC/count.tmm.mtx", sep="\t", quote=F, col.names=T, row.names=T)

result <- getResult(tcc, sort=TRUE)
write.table(result, "03.TCC/DEG.xls", sep="\t", quote=F, col.names=T, row.names=T)

png("03.TCC/DEG.MA.png", width=640, height = 640)
plot(tcc, median.lines = TRUE, cex=0.4)
dev.off()
```

Results of TCC



DEG

시트 검색

홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 공유

붙여넣기 글꼴 맞춤 숫자 조건부 서식 표 서식 셀 스타일 셀 편집

R7

	A	B	C	D	E	F	G	H	I
1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG		
2	8410	AT2G30750	8.04015479	10.9461345	1.56E-137	3.77E-133	1	1	
3	17646	AT4G31970	4.90903142	11.1692029	9.33E-82	1.13E-77	2	1	
4	2576	AT1G26380	7.53945116	8.91198906	1.90E-71	1.53E-67	3	1	
5	5021	AT1G66700	6.51341955	6.66896877	2.29E-68	1.38E-64	4	1	
6	15658	AT4G11170	5.11670594	7.58543714	7.45E-68	3.59E-64	5	1	
7	12843	AT3G28510	3.9971231	11.3291907	1.45E-64	5.85E-61	6	1	
8	6627	AT2G04050	6.23991222	9.85721532	1.33E-59	4.58E-56	7	1	
9	20875	AT5G26920	6.51798174	6.13647867	2.11E-58	6.35E-55	8	1	
10	16522	AT4G19970	-2.6394919	9.52796919	1.25E-56	3.36E-53	9	1	
11	17317	AT4G28460	4.8955265	6.22840735	1.10E-52	2.65E-49	10	1	
12	21438	AT5G40990	3.19638491	9.78015908	7.04E-52	1.54E-48	11	1	
13	2565	AT1G26240	8.0771913	9.10033482	1.63E-51	3.28E-48	12	1	
14	16630	AT4G21380	4.39684298	7.49124158	3.20E-51	5.93E-48	13	1	
15	20519	AT5G22800	8.83888489	4.33388534	1.10E-48	1.73E-45	14	1	
16	18178	AT4G37390	8.9302226	7.07334019	1.12E-48	1.73E-45	15	1	
17	7547	AT2G20800	4.18175217	7.07149633	1.15E-48	1.73E-45	16	1	
18	7758	AT2G23270	3.92075144	9.19264293	2.06E-48	2.92E-45	17	1	
19	23404	AT5G62340	5.98593236	-5.741755	3.76E-48	5.04E-45	18	1	
20	20753	AT5G25250	7.07624607	5.2927006	5.46E-47	6.93E-44	19	1	
21	2577	AT1G26390	7.1773698	8.55094203	2.22E-43	2.68E-40	20	1	
22	3452	AT1G47130	6.31077982	4.75270787	2.70E-43	3.10E-40	21	1	
23	6255	AT1G79680	4.35480925	6.89083356	4.89E-41	5.37E-38	22	1	
24	6476	AT2G02010	4.91931852	6.551152	1.18E-40	1.24E-37	23	1	
25	3089	AT1G32350	3.78760389	8.9362555	3.66E-40	3.68E-37	24	1	
26	12702	AT3G26830	8.6828265	5.04191262	1.32E-38	1.28E-35	25	1	
27	19706	AT5G13320	3.95767042	7.27638856	1.18E-36	1.10E-33	26	1	
28	5323	AT1G69920	4.78067824	7.30758969	1.78E-36	1.59E-33	27	1	
29	16660	AT4G21680	6.28867524	5.6041426	2.00E-35	1.72E-32	28	1	
30	2579	AT1G26410	6.40835375	5.98703043	2.55E-35	2.09E-32	29	1	
31	18169	AT4G37290	6.20978421	5.28039862	2.60E-35	2.09E-32	30	1	
32	9586	AT2G43000	7.18900992	6.21473463	3.43E-35	2.67E-32	31	1	
33	17085	AT4G26120	6.45880999	4.18672166	6.83E-34	5.15E-31	32	1	
34	10158	AT3G01600	5.66525101	5.18195648	3.32E-33	2.43E-30	33	1	
35	7309	AT2G18190	2.46042338	8.25579125	9.02E-33	6.40E-30	34	1	

DEG

준비 100%

Comparison of DEG Methods

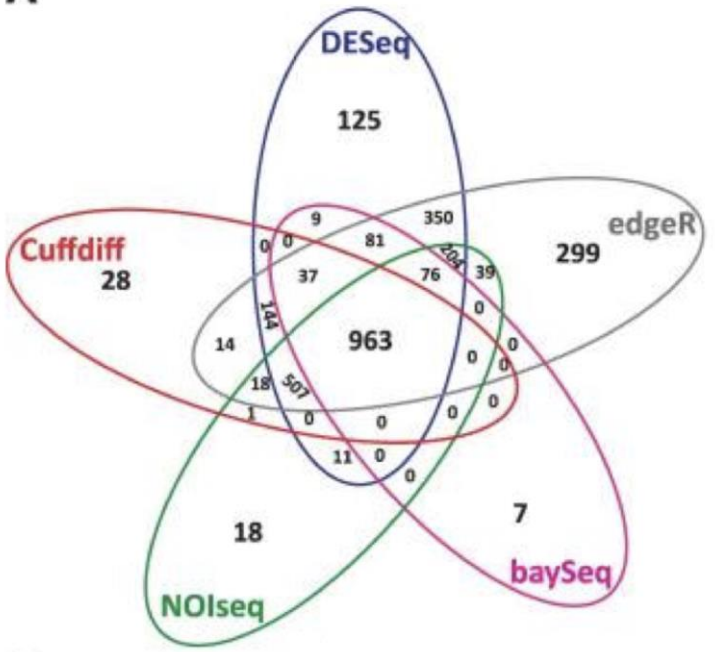
TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description	Version	Citations[d]	Reference
t-test	Normal	DEseq[a]	Two-sample *t*-test for equal variances			
log *t*-test	Log-normal	DEseq[a]	Log-ratio *t*-test			
Mann-Whitney	None	DEseq[a]	Mann-Whitney test			
Permutation	None	DEseq[a]	Permutation test			
Bootstrap	Normal	DEseq[a]	Bootstrap test			
baySeq[c]	Negative binomial	Internal	Empirical Bayesian estimate of posteri likelihood			
Cuffdiff	Negative binomial	Internal	Unknown			
DECseq[c]	Binomial	None	Random sampling model using Fisher' exact test and the likelihood ratio te			
DESeq[c]	Negative binomial	DEseq[a]	Shrinkage variance			
DESeq2[c]	Negative binomial	DEseq[a]	Shrinkage variance with variance base and Cook's distance pre-filtering			
EBSeq[c]	Negative binomial	DEseq[a] (median)	Empirical Bayesian estimate of posteri likelihood			
edgeR[c]	Negative binomial	TMM[b]	Empirical Bayes estimation and either exact test analogous to Fisher's exa test but adapted to over-dispersed d			

(2013)

As reported by PubMed Central articles that reference the listed reference (December 21, 2015).

- Venn's diagram of the comparison of differential gene expression based on RNA-seq data (result fr om Stampy aligner) through five different statist ical methods: Cuffdiff, DESeq, NOISeq, edgeR and baySeq.



(Nookaew et al., Nucleic Acids Res. 2012)

IN DISCUSSION…

In order to address the impact of different statistical methods on the i dentification of DGE, we found that Cuffdiff, baySeq, DESeq, edgeR and NOISeq generated consistent results. Additionally, the results obtained based on RNA-seq data were in good agreement with micro array data. Interestingly, edgeR identified more DGE than the other m ethods at the same cut-off, which might infer less control of type 1 err or with this method.

(Nookaew et al., Nucleic Acids Res. 2012)