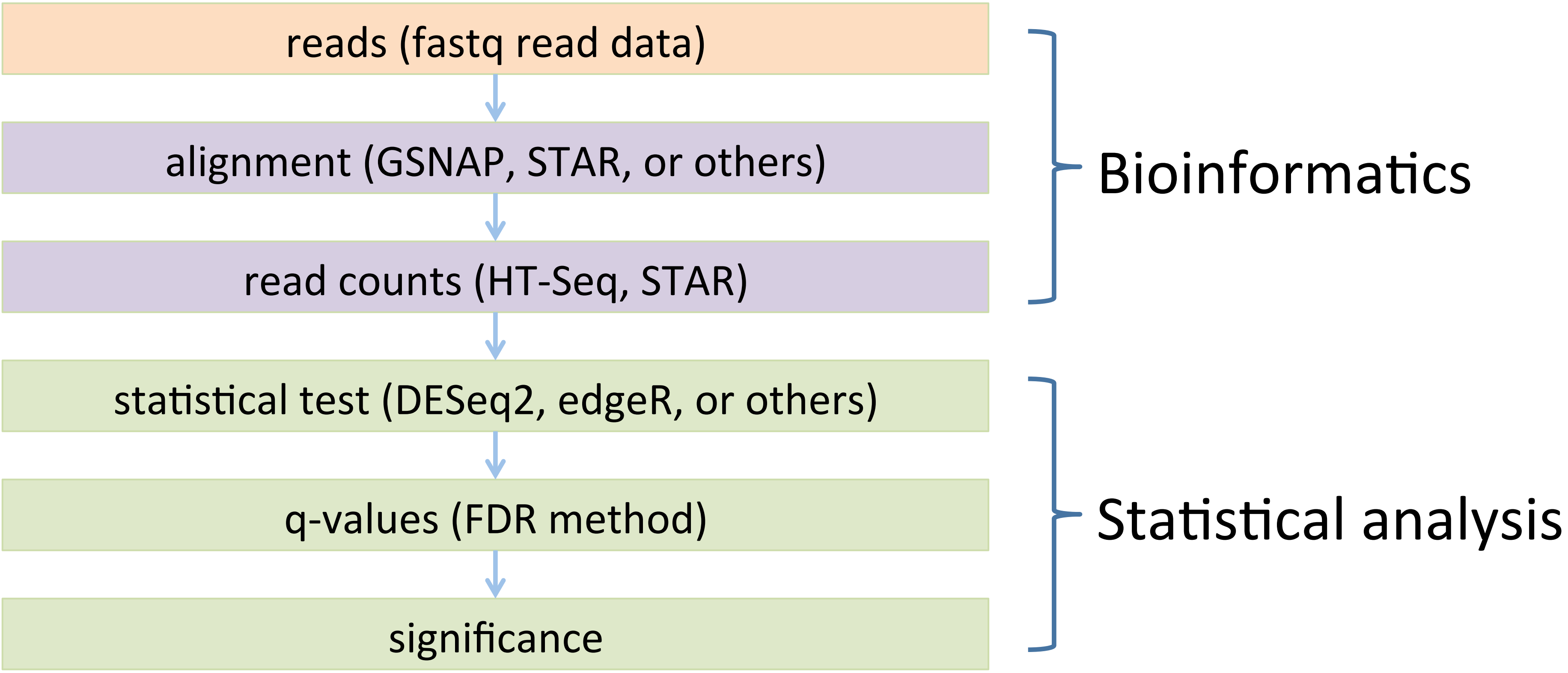
Design of RNA-Seq and Result Interpretation (II)

Sanzhen Liu
Department of Plant Pathology
Kansas State University
@K-State IGF RNA-Seq Workshop (PLPTH885)

6/7/2018

Bioinformatics and Statistics (Illumina data)



reads (fastq read data)

alignment (GSNAP, STAR, or others)

read counts (HT-Seq, STAR)

statistical test (DESeq2, edgeR, or others)

q-values (FDR method)

significance

Bioinformatics

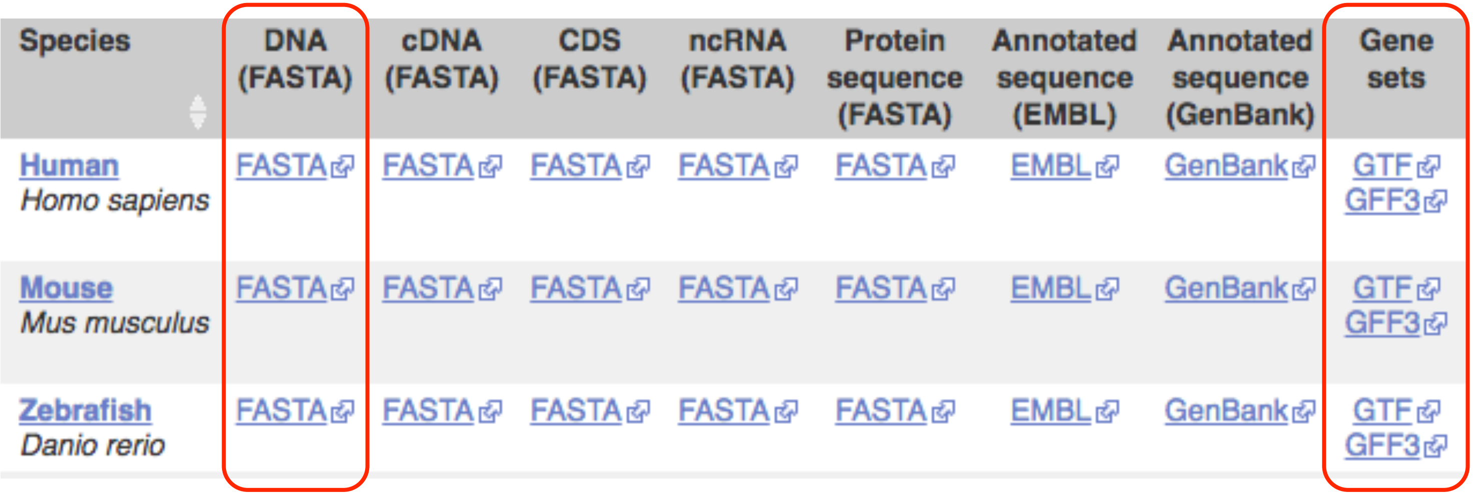
Statistical analysis

STAR pipeline – from reads to counts

Required files:

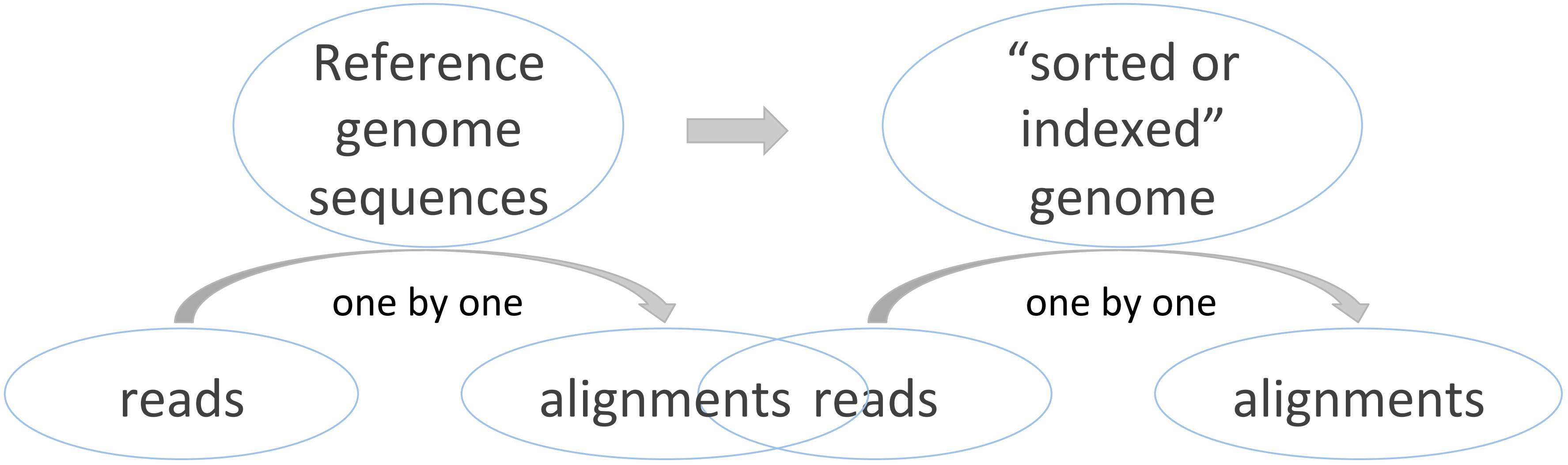
1. Reference genome (fasta file)
2. Gene information (gff or gtf gene annotation)
3. Reads (fastq files) – your own data

Many reference genomes and gff/gtf files are available at:
http://ensembl.org/info/data/ftp



Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets
Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3

Reads to counts - **reference indexing**



```
STAR --runMode genomeGenerate \
        --genomeDir . \
        --genomeFastaFiles reference.fas \
        --sjdbGTFfile genes.gtf \
        --runThreadN 48
```

Reads to counts – **alignment and read counting**

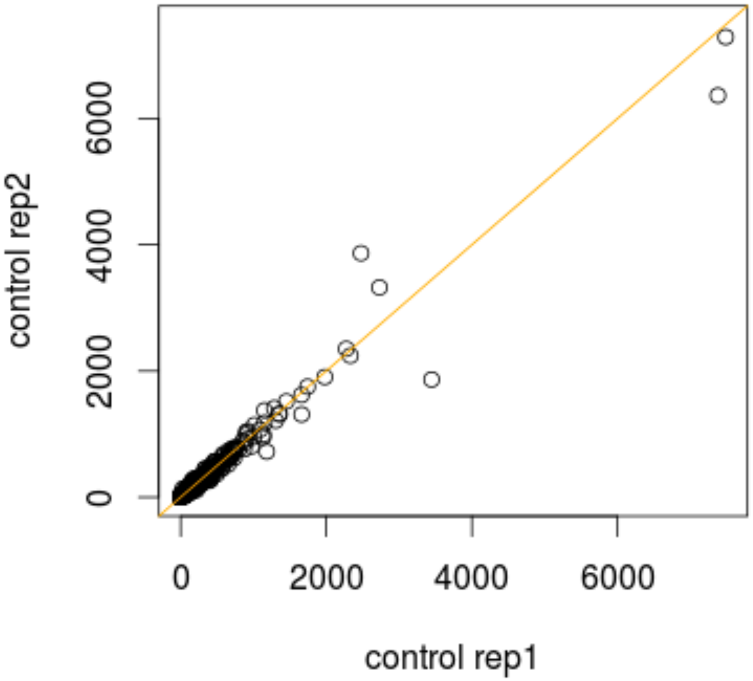
```
STAR --genomeDir reference.fas \
    --readFilesIn read1.fq read2.fq \
    --alignIntronMax 100000 \
    --alignMatesGapMax 100000 \
    --outFileNamePrefix output \
    --outSAMattrIHstart 0 \
    --outSAMmultNmax 1 \
    --outSAMstrandField intronMotif \
    --outFilterIntronMotifs RemoveNoncanonicalUnannotated \
    --outSAMtype BAM SortedByCoordinate \
    --quantMode GeneCounts \
    --outFilterMismatchNmax 5 \
    --outFilterMismatchNoverLmax 0.05 \
    --outFilterMatchNmin 50 \
    --outSJfilterReads Unique \
    --outFilterMultimapNmax 1 \
    --outSAMmapqUnique 60 \
    --outFilterMultimapScoreRange 2
```

Count matrix: Read counts (Raw) per gene

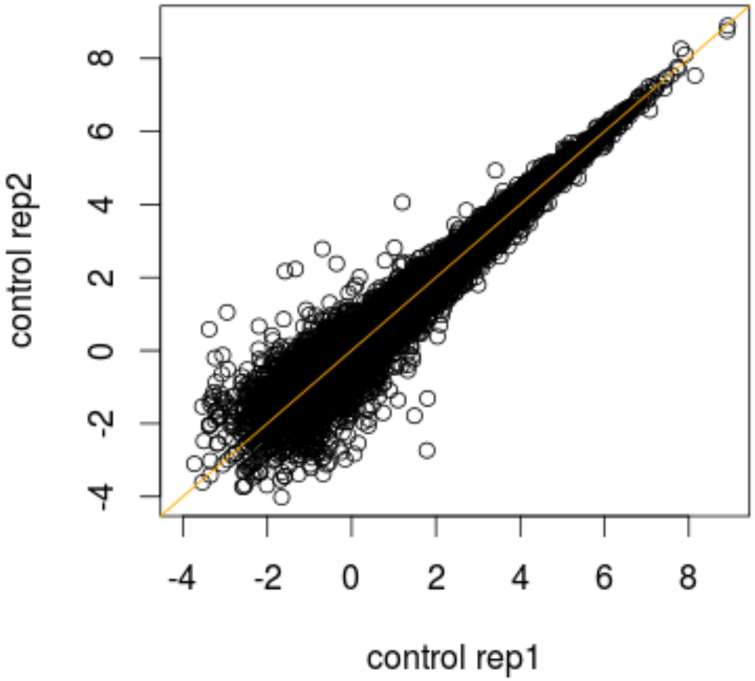
Gene	sample 1	sample 2	sample 3
gene 1	6,075	5,934	3,370
gene 2	295	377	169
...	...	...	...

Overall difference of read counts among samples

Scatter plot

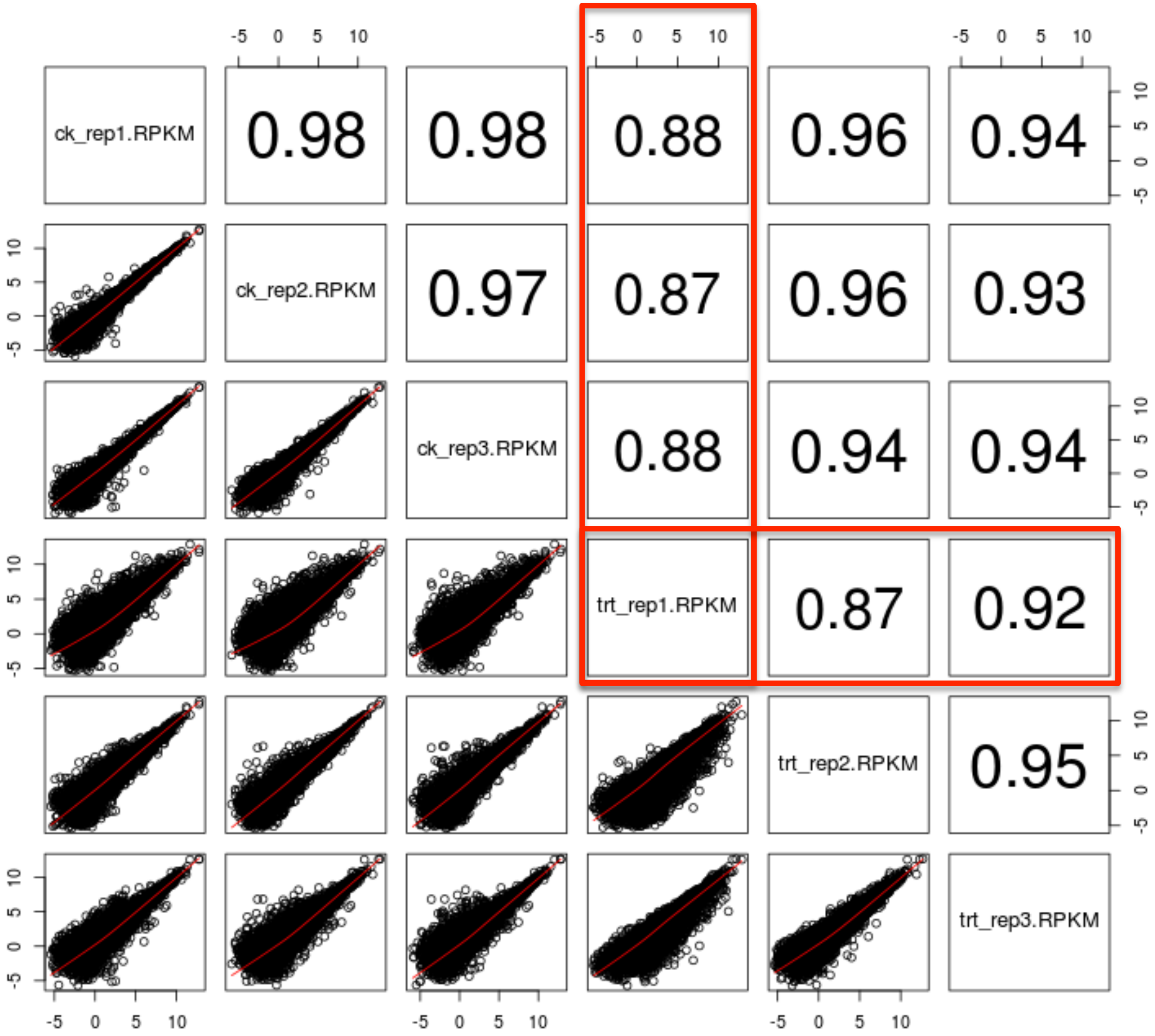


RPKM scatter plot



Log RPKM scatter plot

Pair-wise scatter plot



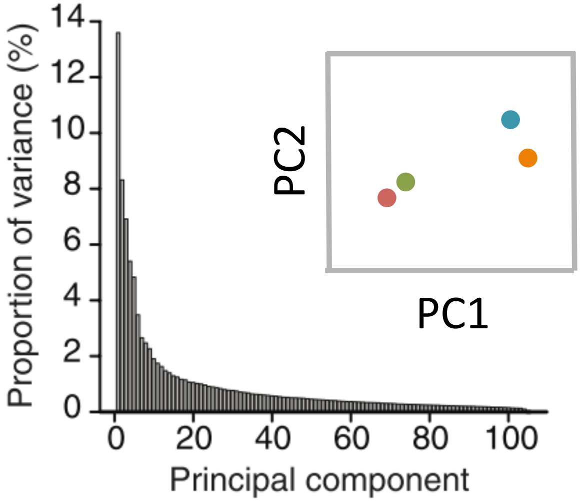
Principal Component Analysis (PCA)

PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set.

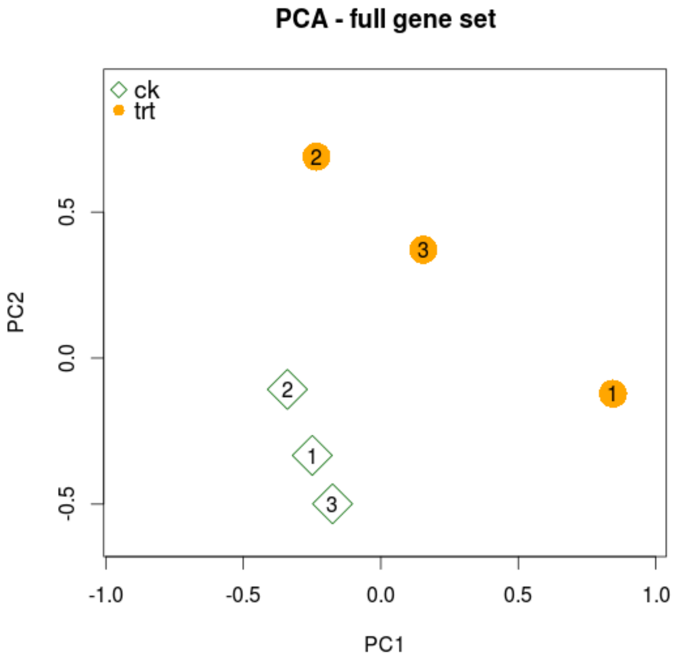
Feature/ variable	John	Mike	Jack	Justin
Weight (lb)	150	243	186	128
Height (cm)	171	190	178	175
…				

	Control			Treatment		
GeneID	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3
1	2679	2360	2573	2563	3398	3012
2	177	161	171	154	137	152
3	381	371	397	541	723	635
…						
30000	990	1073	1236	850	672	859

Normalized and standardized data



Nature Biotech, 2008, 26:303-4



Statistical test for differential expression

π_G: Proportion of transcript fragments of *gene G* among all transcripts

Sample one read, the distribution of the read from *gene G* is Bernoulli(π_G)
Pr(read from G) = π_G
Pr(read not from G) = 1 - π_G

Sample N read, the distribution of the number of reads from *gene G* is:

$Binomial(N, \pi_G) \approx Poisson(N\pi_G)$

In the Poisson distribution, mean = variance = $N\pi_G$

However, the Poisson distribution can not well explain data variance (overdispersion issue)

That is why a Negative-Binomial (NB)distribution was introduced

Counts of reads from a gene ~ $NB(mean = \ \mu, variance = \ \sigma^2)$

$$\sigma^2 = \ \mu + \ \varphi \ \mu^2$$

φ is a dispersion parameter

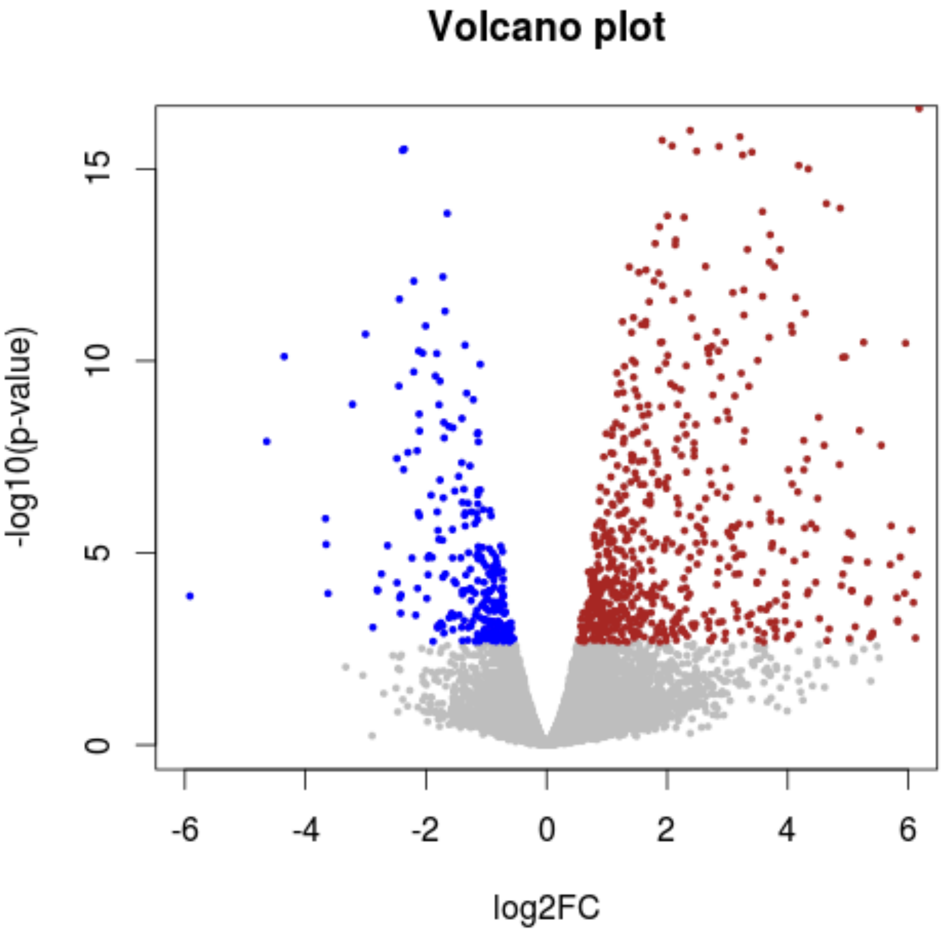
- Generalized Linear Model (GLM) to deal with count data
- **NB-GLM** to incorporate dispersion into the model

Visualization of DE results

Volcano plot

DE Result

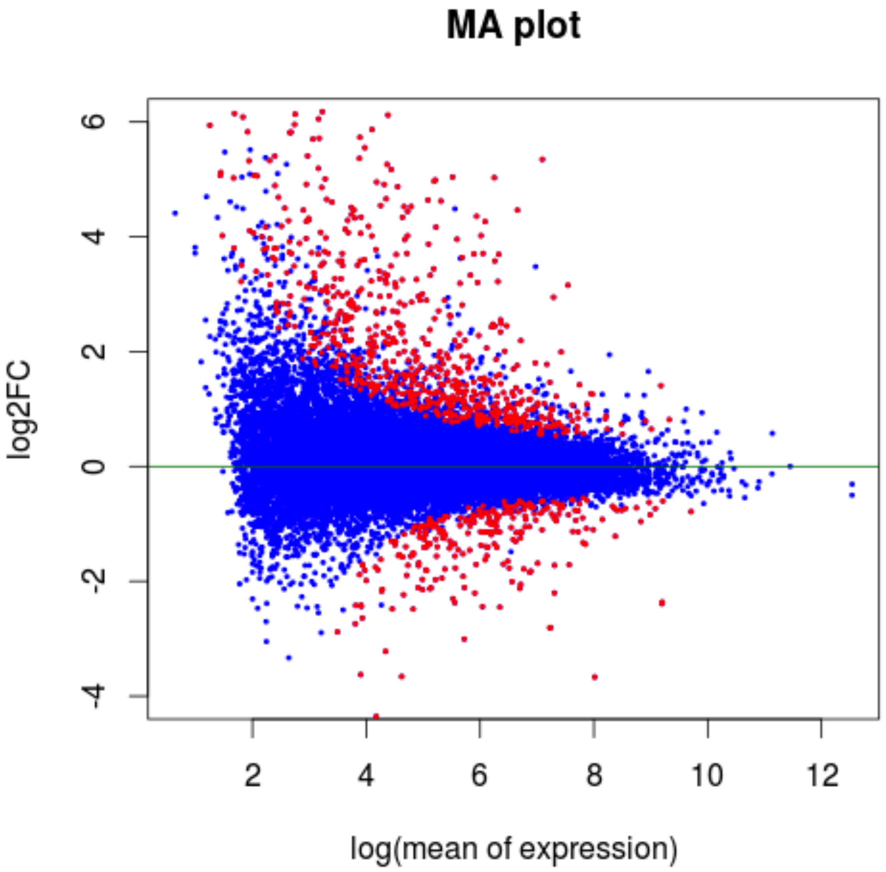
GeneID	Log2FC	p-value	-log10(pvalue)
1	-0.40	0.037	1.43
2	0.03	0.916	0.04
3	-0.89	2.42E-05	4.62
4	0.30	0.130	0.89
5	-0.36	0.140	0.85
6	-0.07	0.811	0.09
…			



MA plot

M (log ratios) and A (mean average)

GeneID	Mean RPKM	log mean	log2FC
1	0.51	-0.29	-0.40
2	1.25	0.10	0.03
3	3.52	0.55	-0.89
4	0.19	-0.72	0.30
5	2.34	0.37	-0.36
6	6.14	0.79	-0.07
…			



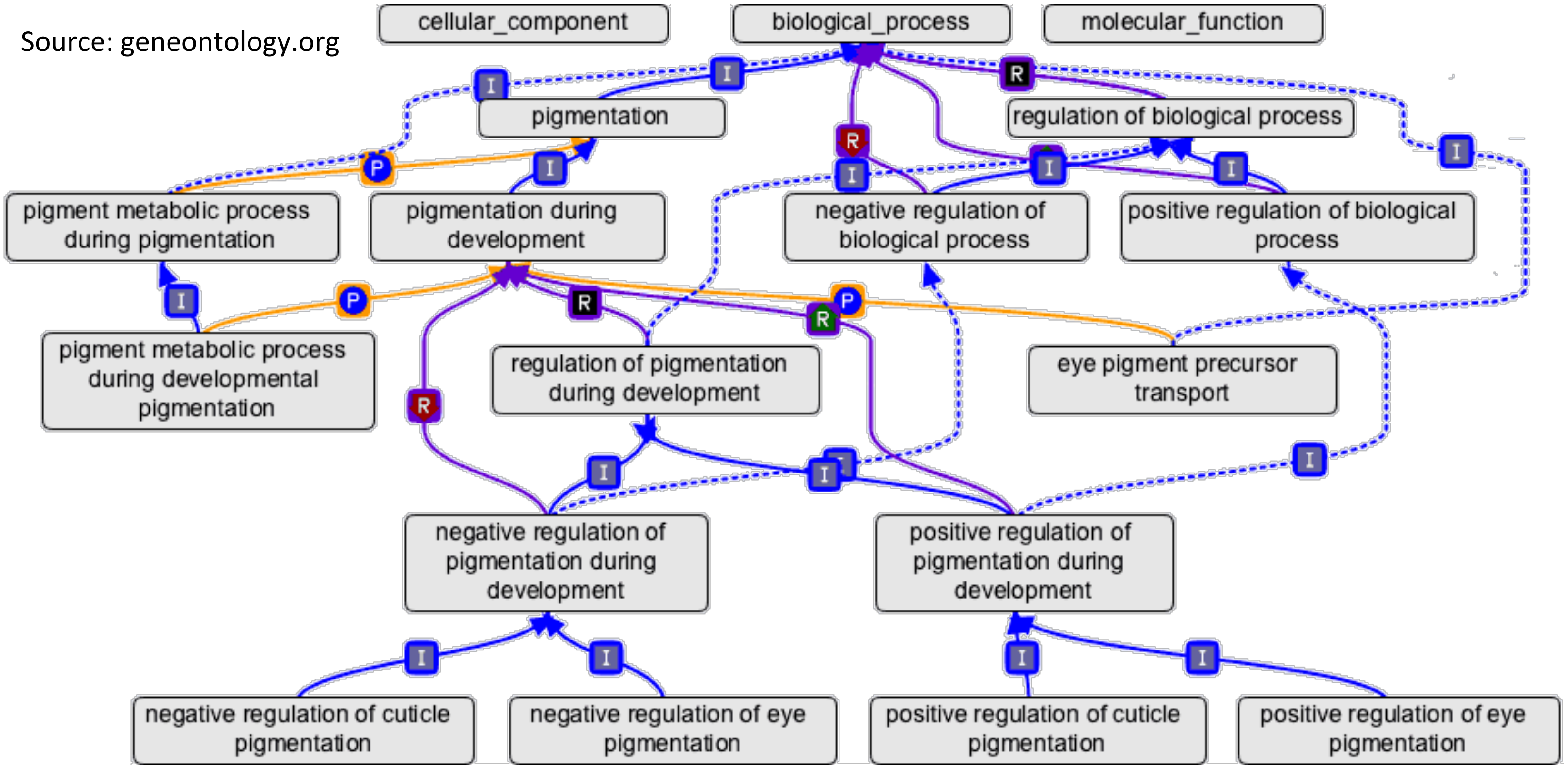
More at: en.wikipedia.org/wiki/MA_plot

Functional interpretation

Gene ontology (GO)

An ontology is a representation of a body of knowledge, within a given domain. Ontologies usually consist of a set of classes or terms with relations that operate between them.

Source: geneontology.org



cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
I
P
R

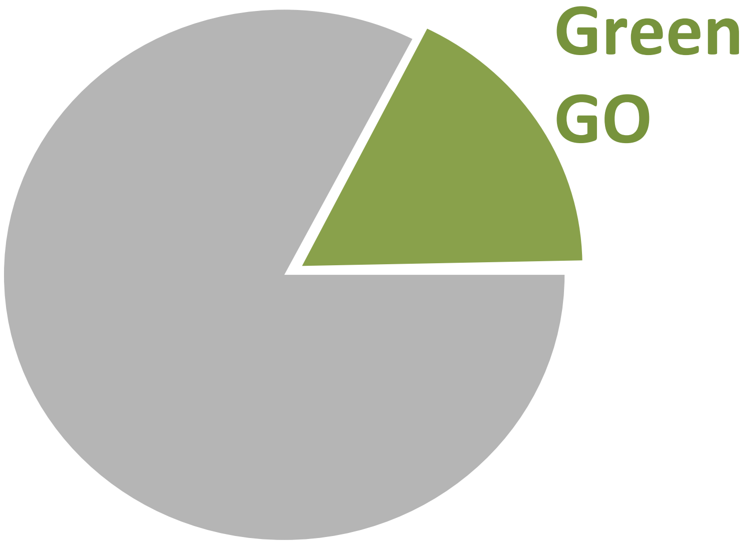
Three domains, three roots

Node: GO term (e.g., cell growth, GO:0016049, biological process)

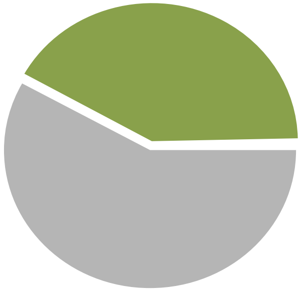
Edge: term-term connection

Each GO term can be traced back to a root

Category enrichment



All genes



significant
gene set

Is **Green GO** enriched in the significant gene set ?

GO enrichment test – Fisher's Exact test

Gene	GO accession
GRMZM2G001475	**GO:0006519**
GRMZM2G001475	GO:0016831
GRMZM2G001500	GO:0005524
GRMZM2G001500	GO:0006457
GRMZM2G001500	GO:0051082
GRMZM2G001508	GO:0003993
GRMZM2G001514	GO:0003677
GRMZM2G001514	GO:0004879
GRMZM2G001514	GO:0005634
GRMZM2G001514	GO:0006355
…	…

GRMZM2G001475	1
GRMZM2G002652	2
GRMZM2G006480	3
…	…
GRMZM5G868038	40

Gene	Significant?
GRMZM2G001475	no
GRMZM2G002652	no
GRMZM2G006480	**yes**
…	…
GRMZM5G868038	no

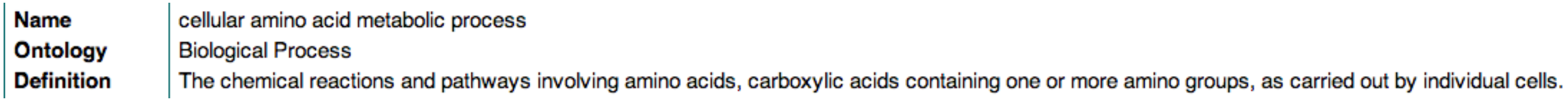
Question: Are the genes of this GO term enriched in the significant gene set?

Assumption: all genes are independent and equally likely to be selected as DEs.

2x2 Table for GO:0006519

	GO:0006519	Others
Significant	5	210
Not significant	35	39416

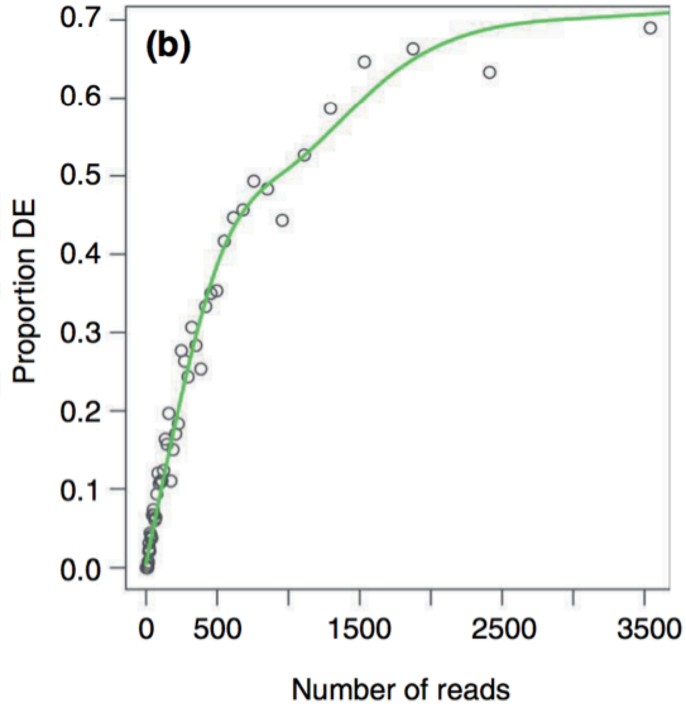
Fisher's Exact Test:
p-value = 2.518e-06



Name	cellular amino acid metabolic process
Ontology	Biological Process
Definition	The chemical reactions and pathways involving amino acids, carboxylic acids containing one or more amino groups, as carried out by individual cells.

GOSeq

Not all genes are equally likely to be selected as DEs.



Young MD, et al., (2010). Genome Biology, 11: R14.

1. The likelihood of DE as a function of number of reads is quantified through fitting a monotonic function to “proportion of DE” versus “number of reads”.

2. The function is incorporated into the enrichment statistical test

Gene	Significant?
GRMZM2G001475	no
GRMZM2G002652	no
GRMZM2G006480	**yes**
…	…
GRMZM5G868038	no

Read counts	Proportion
224	0.16
51	0.05
536	**0.38**
…	…
0	0

3. Weighted sampling to perform enrichment test

GO:0006519	# DE
Obs (from the DE analysis)	5
1st weighted sampling	1
2nd weighted sampling	0
3rd weighted sampling	2
…	…

→ p-value

Summary

- R is an excellent tool for DE analysis and data visualization.
- Many bioinformatics pipelines and statistical methods have been developed. Methods and parameters need to be carefully selected.
- A proper GO enrichment test needs to be used.

- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
- Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14.