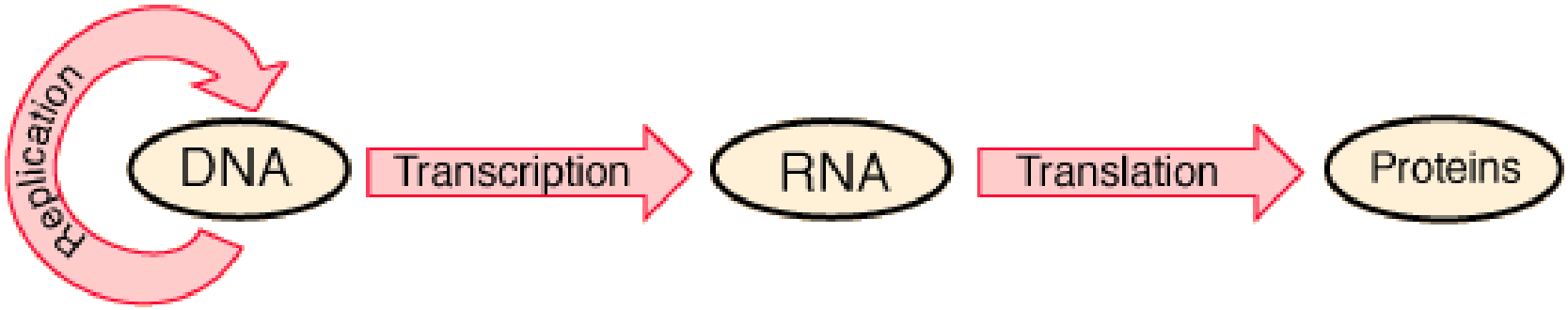
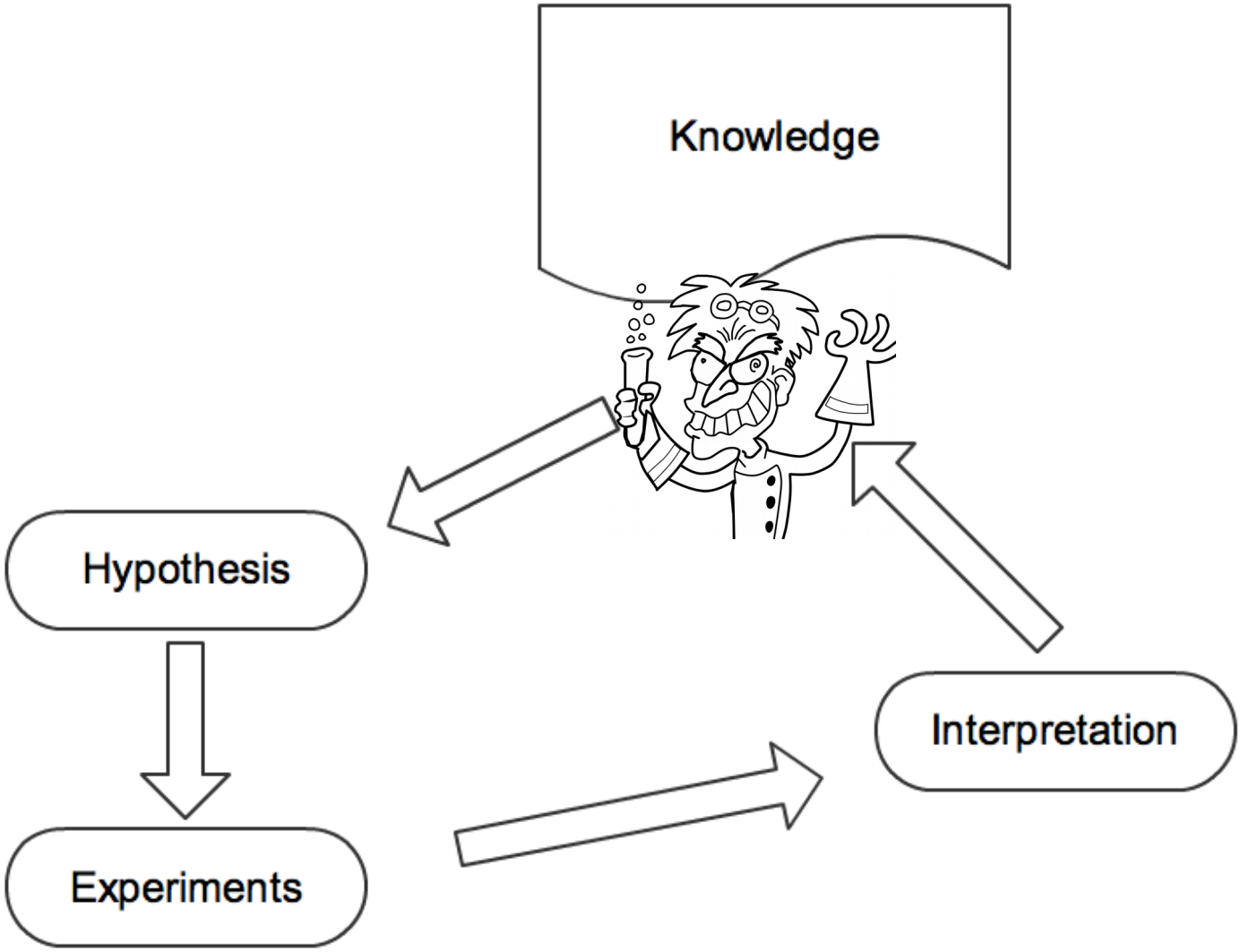
An introduction to gene expression analysis by RNA-seq:

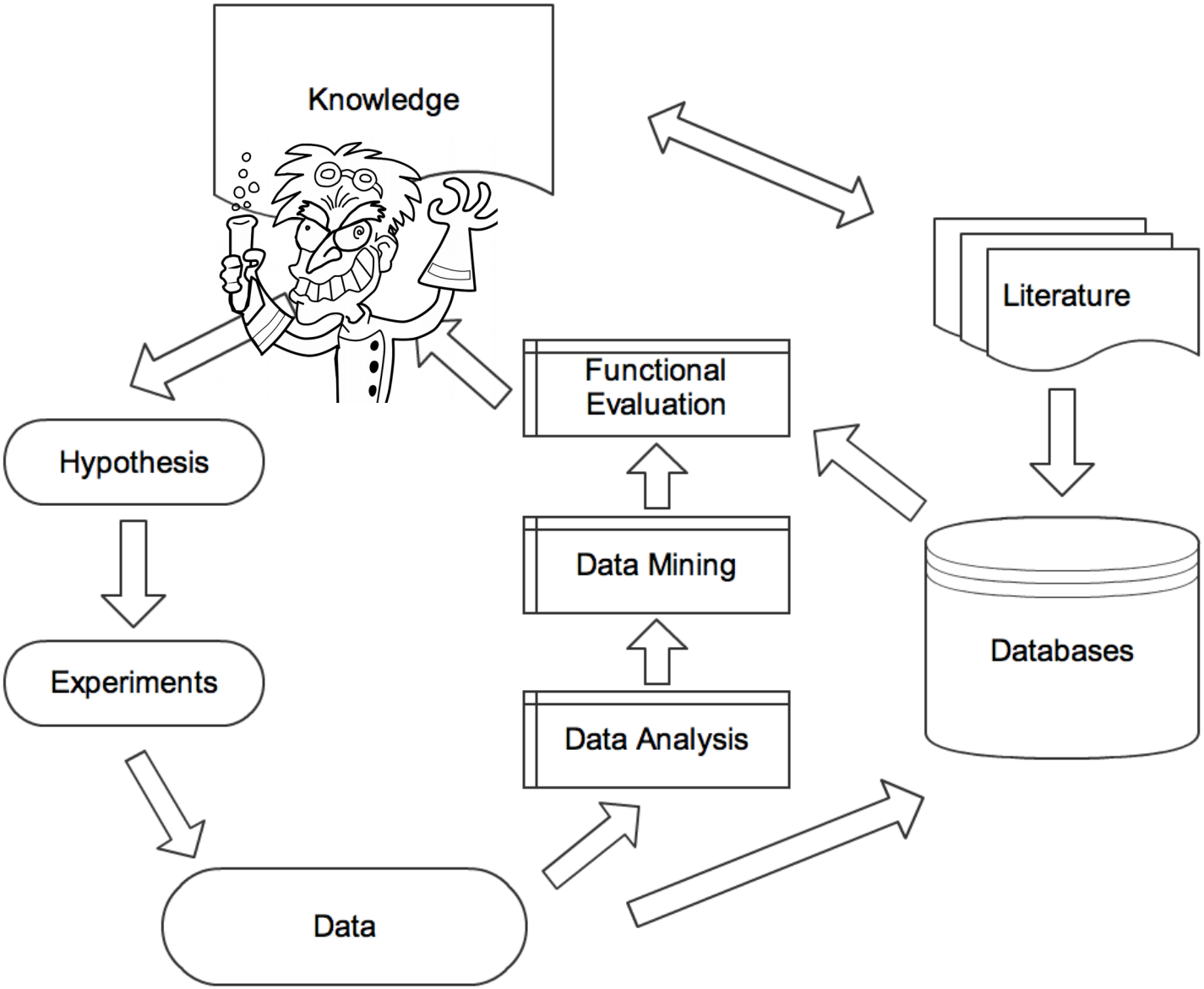
gene-level exploratory analysis and differential expression

Paolo Sonego
Computational Biology Unit, Fondazione Edmund Mach
Via E. Mach 1, 38010, S. Michele a/A, Tn, Italy
office phone: +39 0461 615 645
mail: paolo.sonego@fmach.it



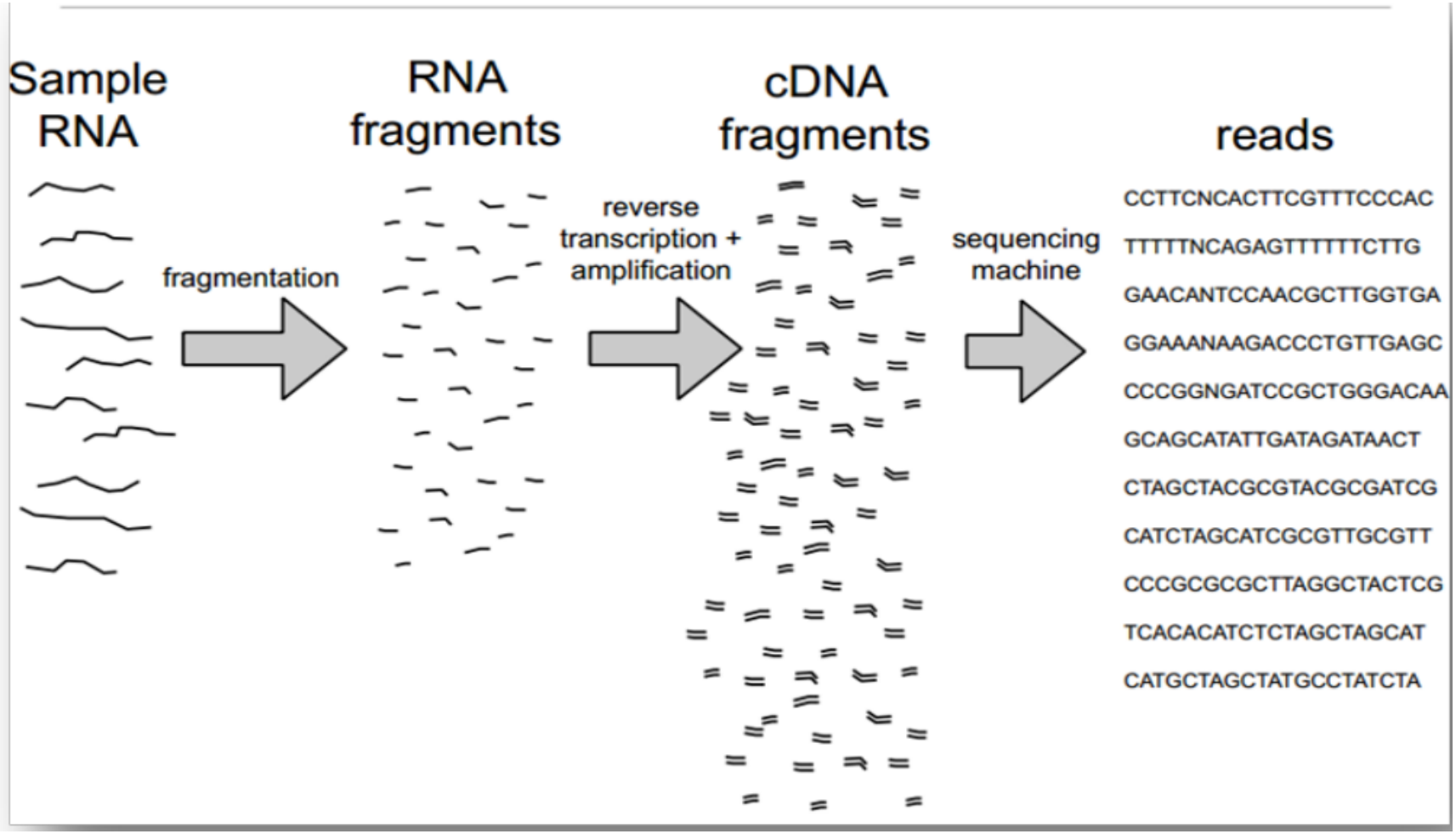
Central dogma of molecular biology


Knowledge
Hypothesis
Experiments
Interpretation

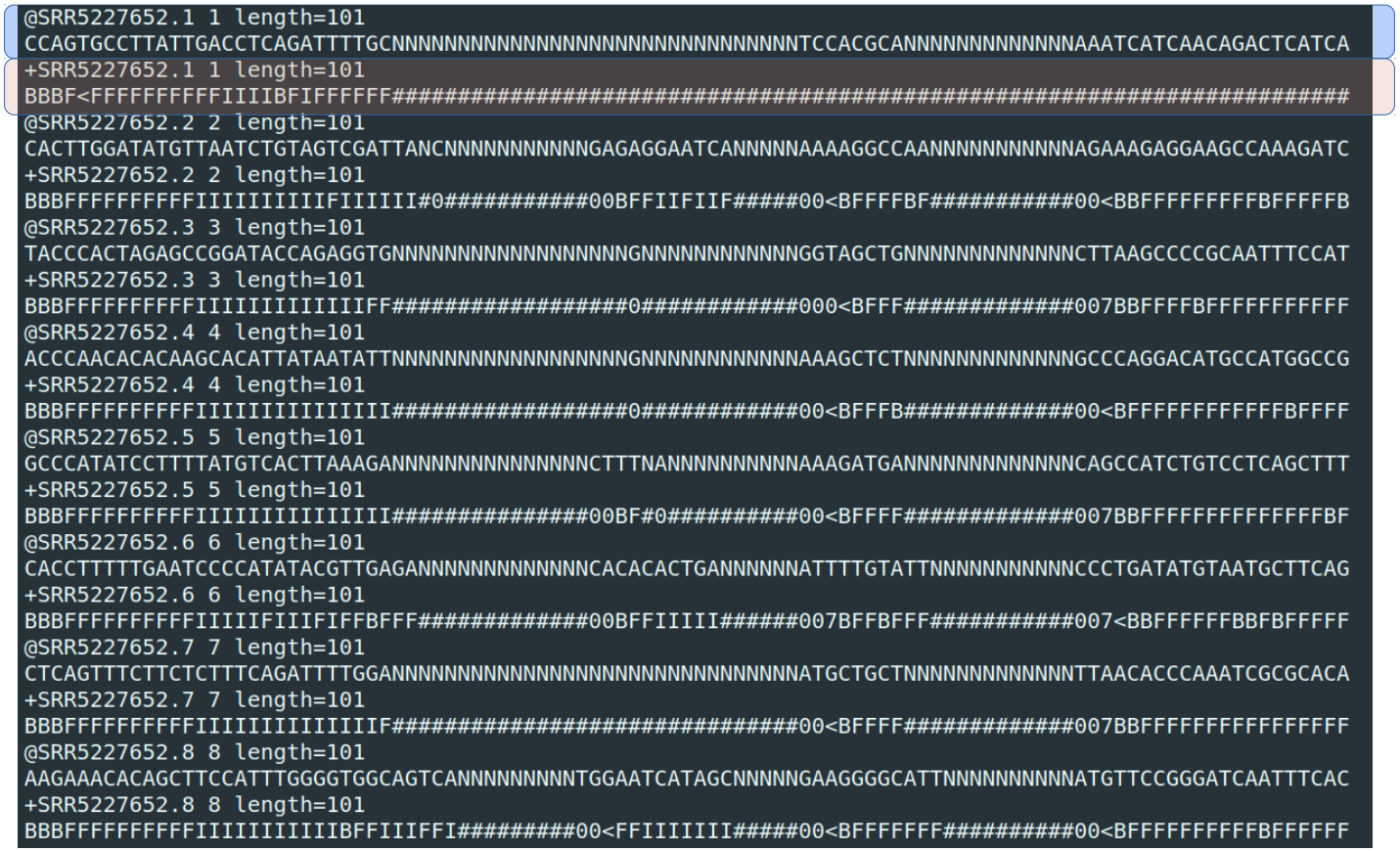

Knowledge
Literature
Functional Evaluation
Hypothesis
Data Mining
Databases
Experiments
Data Analysis
Data

RNA-Seq

- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.
- RNA or DNA is extracted from sample tissue/cells and fragmented. RNA is converted to cDNA by reverse transcription. DNA Fragments are converted into the library by ligation to sequencing adapters containing specific sequences designed to interact with the NGS platform. The next step involves clonal amplification of the library. The final step generates the actual sequence via the chemistries for each technology.


Sample RNA
fragmentation
RNA fragments
reverse transcription + amplification
cDNA fragments
sequencing machine
reads
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA

Fastq files (raw data)



Sequence

Quality

Machine Learning for Gene Expression Analysis

- Class Discovery (Clustering):
  - Dividing samples into reproducible classes that have similar behavior or properties
  - Find genes whose expression profile is similar
  - Find groups of experiment whose expression profile is similar
- Class Prediction (Classification):
  - In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more of these classes
- Class Comparison (Differential Expression):
  - Identify the genes which are up and down regulated among the experimental groups

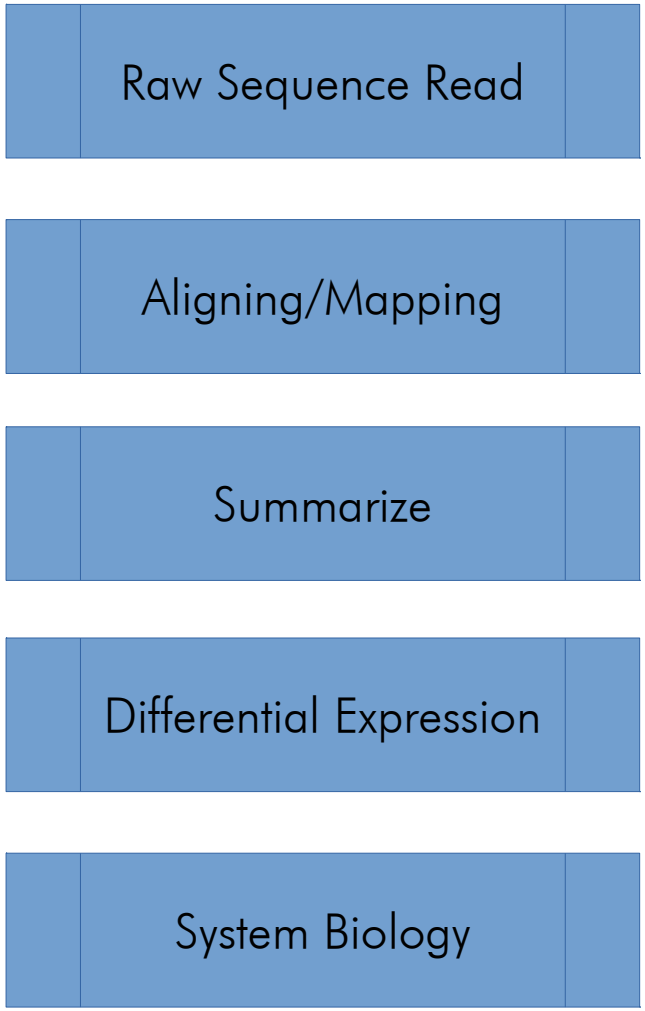
Learning Objectives

- Understand the steps of the analysis from raw reads to expression counts, Class comparison and interpretation of gene lists using R/Bioconductor.
- Check both raw and processed data.
- Select the proper normalization/transformation according to the task performed (EDA/QC, Downstream analysis)
- Make sense of the results of the Class comparison

Experimental Design

- What is the purpose of this experiment?
- How many replicates?
- Technical or biological replicates?
- More reads or more replicates?

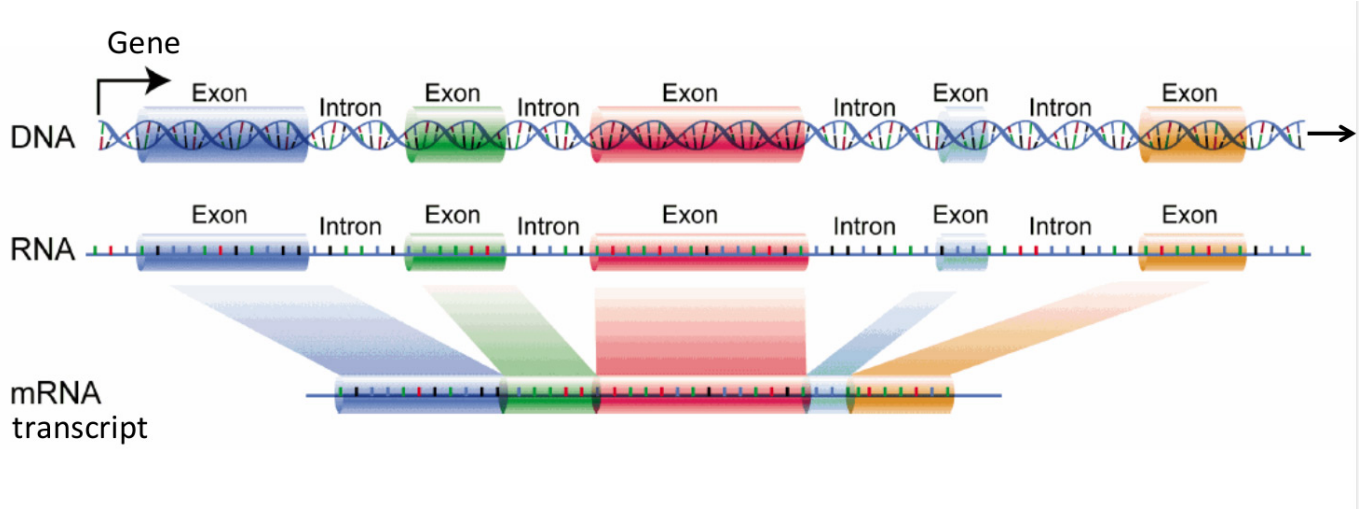
RNA-Seq Workflow



Alignment/Mapping

- There are several choices for this step, see, for example, Baruzzo et al. 2017:
- **alignement methods**
  - STAR
  - Subread/Rsubread
  - HISAT2
  - tophat
- **pseudo-alignement methods (transcript abundance quantification methods)**
  - Salmon
  - Kallisto

Summarization

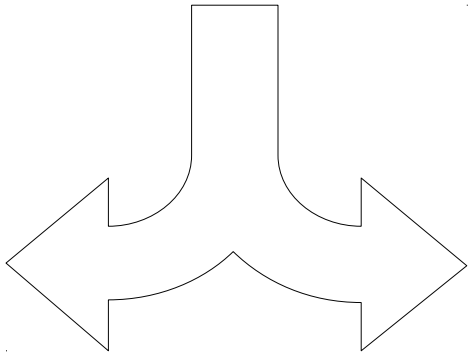


- Summarization is the process that assign mapped reads to genomic features such as genes, exons, promoters, etc. For gene-level differential expression the number of reads overlapping annotated exons of each gene can be used as a measure of the expression level of that gene.
- Typically, methods that perform this tasks take as input a set of files that contain read mapping results and an annotation file that includes genomic features and return a read count for each gene in each sample, producing a matrix of read counts (integer).

	SRR5227652_trimmed.bam	SRR5227653_trimmed.bam	SRR5227654_trimmed.bam	SRR5227655_trimmed.bam	SRR5227656_trimmed.bam
VIT_01s0010g00020	193	65	69	19	9
VIT_01s0010g00060	1016	675	356	369	430
VIT_01s0010g00240	2626	2175	770	260	255
VIT_01s0010g00330	142	126	69	196	186
VIT_01s0010g00340	32	32	23	79	93
VIT_01s0010g00360	12	10	4	7	7

Two Paths in RNA-Seq Analysis

	SRR5227652_trimmed.bam	SRR5227653_trimmed.bam	SRR5227654_trimmed.bam	SRR5227655_trimmed.bam	SRR5227656_trimmed.bam
VIT_01s0010g00020	193	65	69	19	9
VIT_01s0010g00060	1016	675	356	369	430
VIT_01s0010g00240	2626	2175	770	260	255
VIT_01s0010g00330	142	126	69	196	186
VIT_01s0010g00340	32	32	23	79	93
VIT_01s0010g00360	12	10	4	7	7



Transformations
and
Exploratory Data Analysis (EDA)

Differential Expression

Pre-processing

- Transformation for EDA
- Exploratory Data Analysis (EDA)
- Removing low expressed genes
- Normalization and transformation for DE analysis

Transformations for EDA

- The number of reads for a given gene is proportional to the expression level of the gene AND to its transcript length AND to the sequencing depth of the library.

- Counts Per Million (CPM): divide each read count per library size in millions.

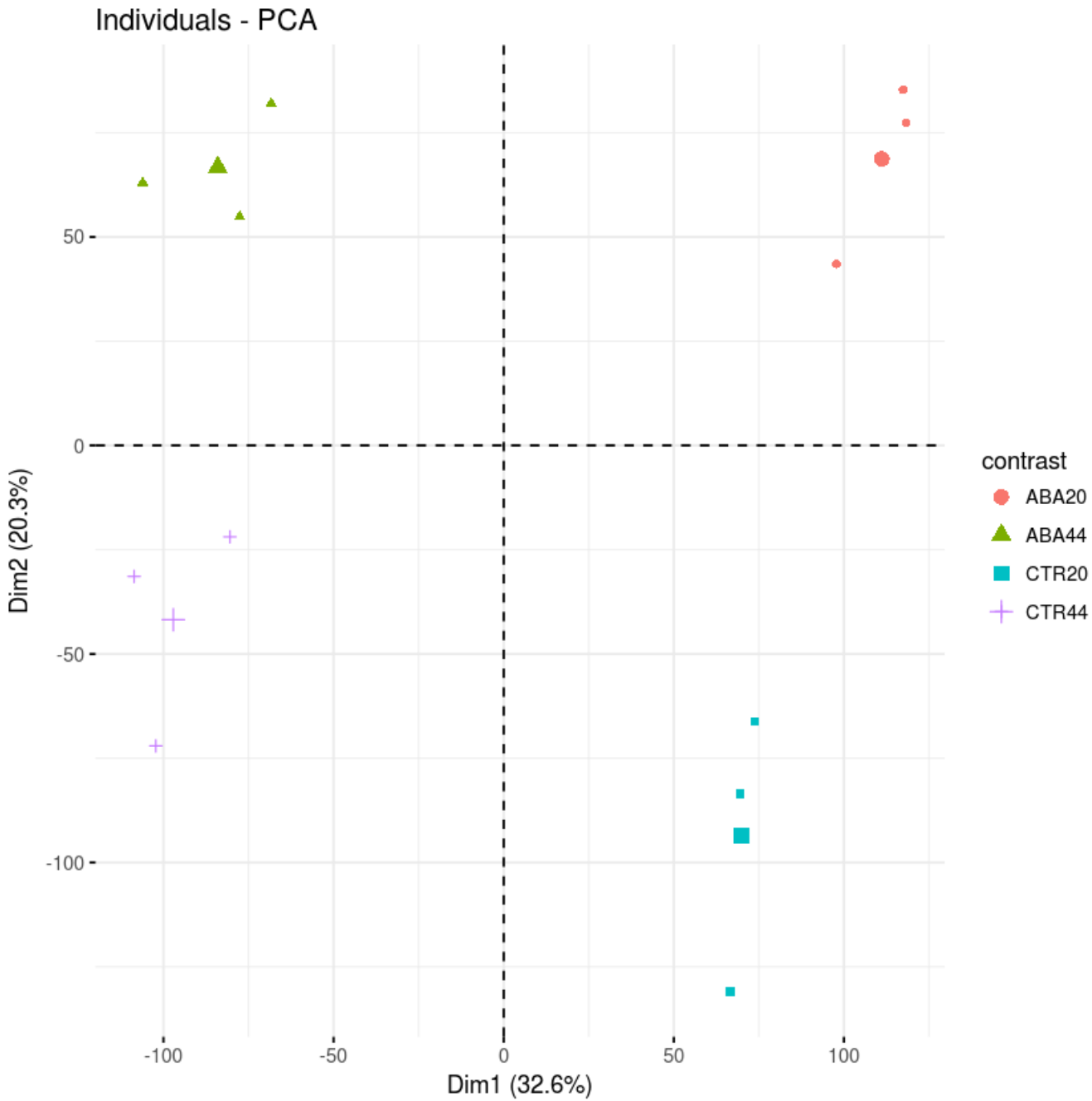
$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

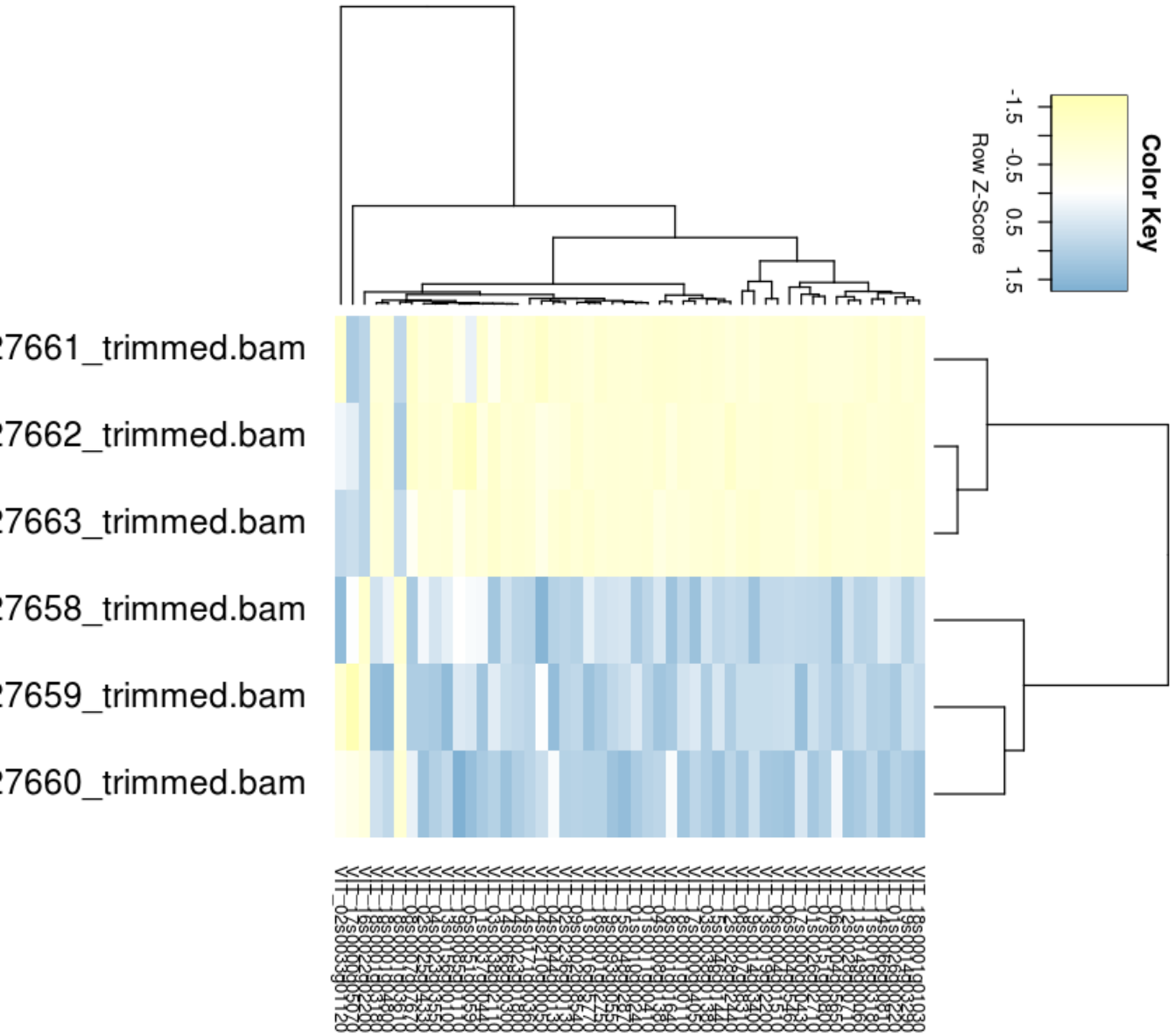
- Reads/Fragments Per Kilobase per Million (RPKM/FPKM): counts are divided by the transcript length (kb) times the total number of millions of mapped reads.

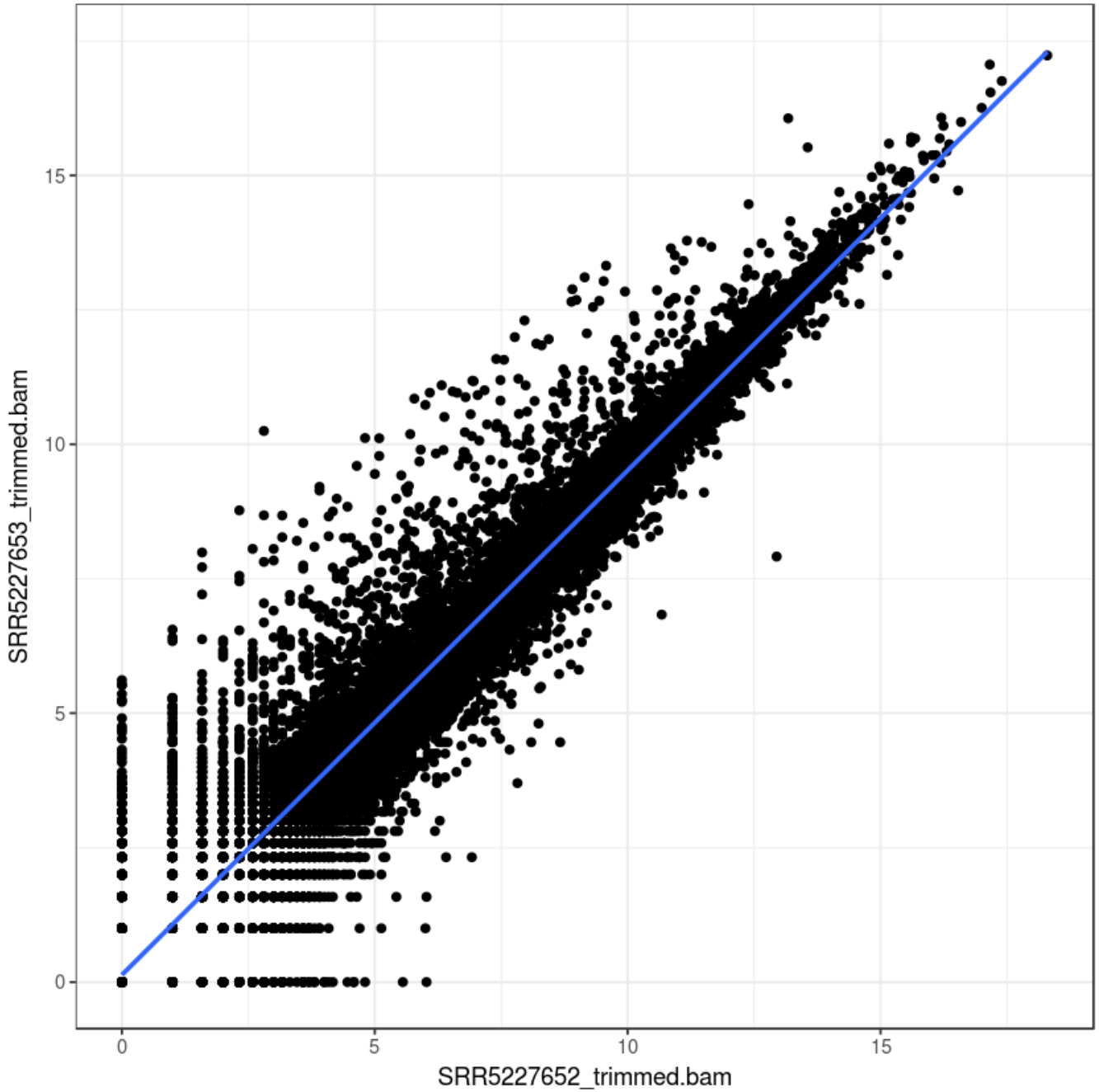
$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

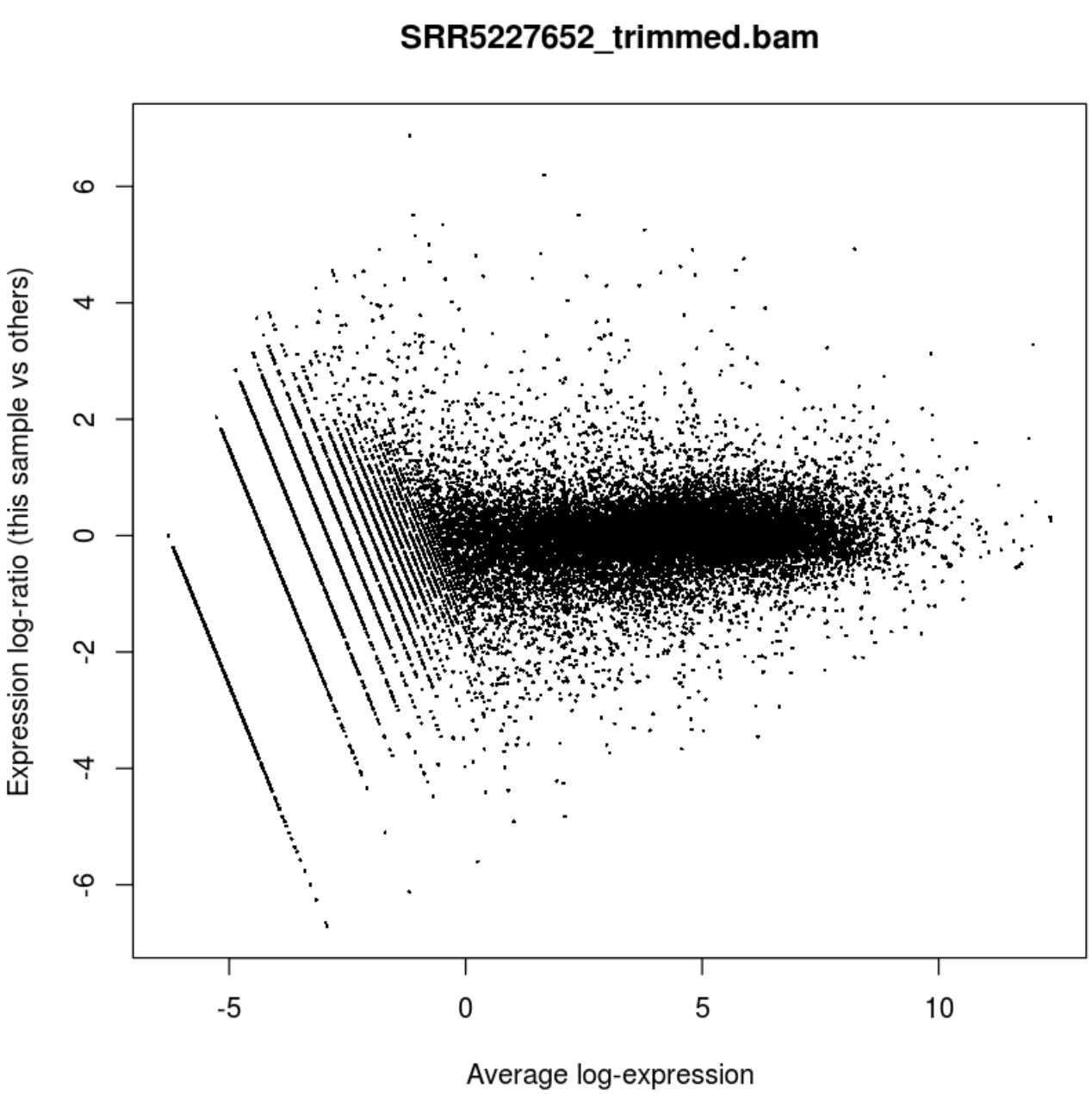
EDA Visualizations

- PCA/MDS
- Clustering
- Scatterplot
- Box-plot
- MA-plot (Mean-Difference MD-plot)


Individuals - PCA
Dim2 (20.3%)
Dim1 (32.6%)
50
0
-50
-100
-100
-50
0
50
100
contrast
ABA20
ABA44
CTR20
CTR44


Color Key
-1.5
-0.5
0.5
1.5
Row Z-Score
27661_trimmed.bam
27662_trimmed.bam
27663_trimmed.bam
27658_trimmed.bam
27659_trimmed.bam
27660_trimmed.bam


SRR5227653_trimmed.bam
SRR5227652_trimmed.bam
0
5
10
15
0
5
10
15


SRR5227652_trimmed.bam
Expression log-ratio (this sample vs others)
6
4
2
0
-2
-4
-6
-5
0
5
10
Average log-expression

Remove low expressed genes

- Genes that are not expressed at a biologically meaningful level in any condition should be discarded to reduce the subset of genes to those that are of interest, and to reduce the number of tests carried out downstream when looking at differential expression.

- Although any sensible value can be used as the expression cutoff, typically a CPM value of 1 is used in analyses as it separates expressed genes from unexpressed genes well for most datasets. Here, a CPM value of 1 means that a gene is “expressed” if it has at least 20 counts in the sample with library size ≈20 million .

Differential Expression Analysis

- Design matrix
- Contrast matrix
- Trimmed mean of M-values (TMM) Normalization
- Voom transformation
- Limma

Design

Design matrix

CTRL20	ABA20	CTRL44	ABA44
1	0	0	0
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1

Contrast matrix

	ABA20vsCTRL20	ABA44vsCTRL44
CTRL20	-1	0
ABA20	1	0
CTRL44	0	-1
ABA44	0	1

Normalization for DE Analysis

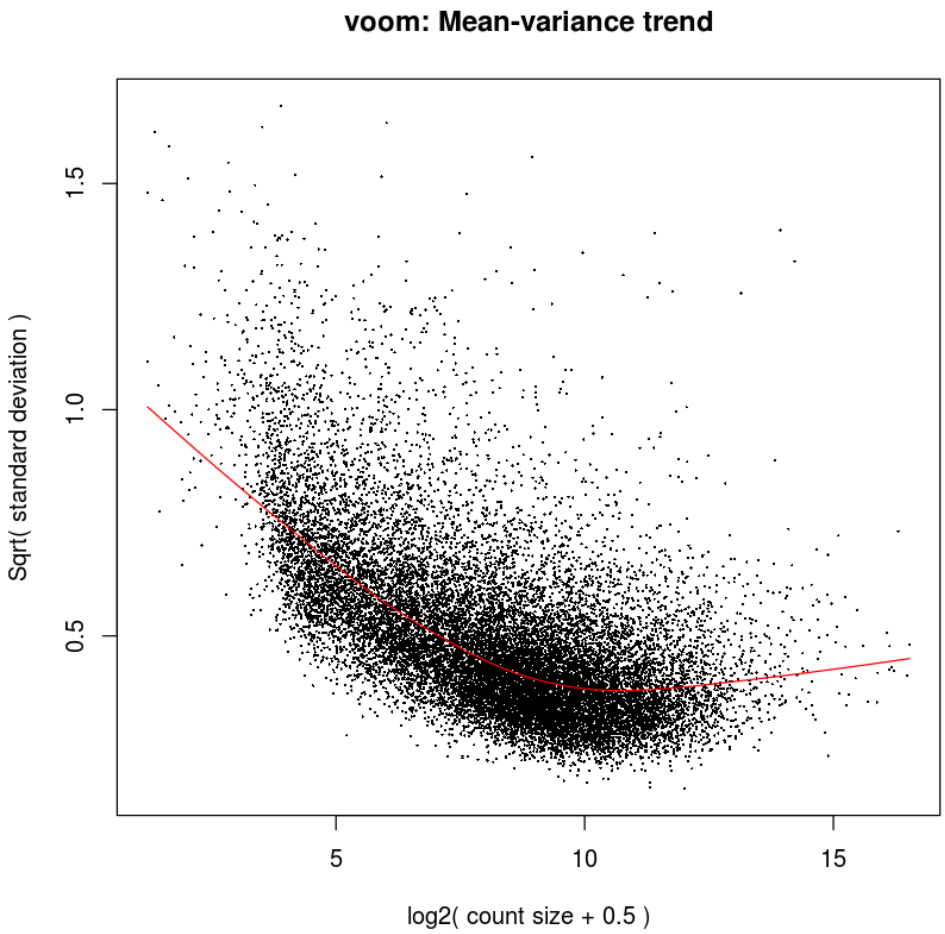
- Trimmed mean of M-Values (TMM) normalization is a normalization method commonly performed to eliminate composition biases between libraries. This generates a set of normalization factors, where the product of these factors and the library sizes defines the effective library size.
- A normalization factor below one indicates that the library size will be scaled down, as there is more suppression (i.e., composition bias) in that library relative to the other libraries. This is also equivalent to scaling the counts upwards in that sample. Conversely, a factor above one scales up the library size and is equivalent to downscaling the counts.

Limma: voom function

- Unlike methods such as edgeR or DESeq(2) that model counts using a Negative Binomial distribution, limma performs linear modelling on the log-CPM values assumed to be normally distributed and the relationship between mean and variance is taken care using precision weights calculated by the voom function.

- The read counts of gene g in sample i do not follow a Normal distribution however the transformed (logCPM) response variable converges quickly to normality. voom() models the log counts per million and fits a loess trend line to the scatterplot of variance vs. mean to create weight that are then fed into a standard limma analysis. The variance model is at the observation level and the loess trend is robust against highly variable genes.

limma/voom: mean-variance trend

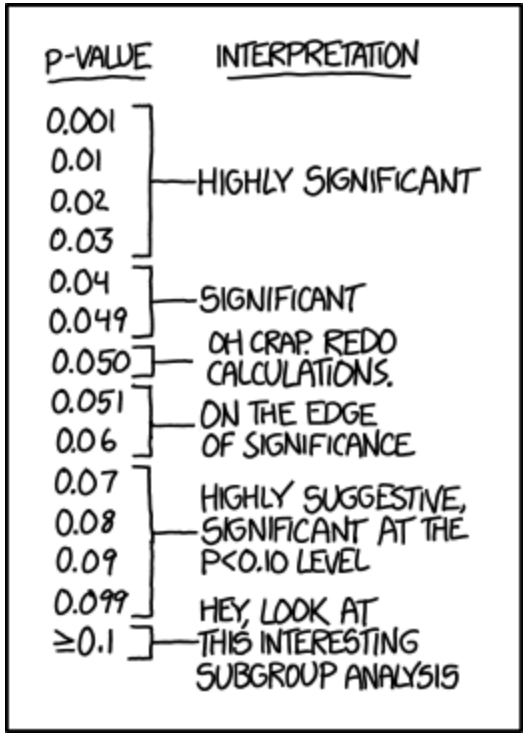


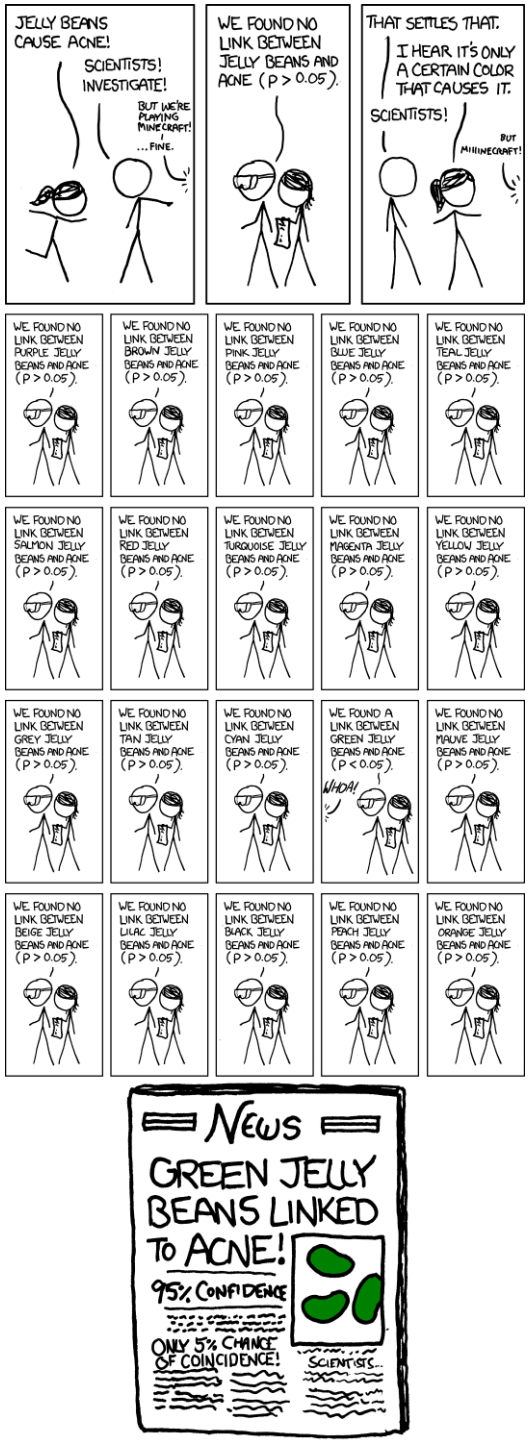
Limma: Fitting linear models for comparisons of interest

- Fitting a separate linear model to the expression values for each gene.
- ebayes moderated t-test which borrow information across all genes to obtain precise estimate of gene-wise variability.
- The method aims to remove the dependency of the variance from the mean expression level.

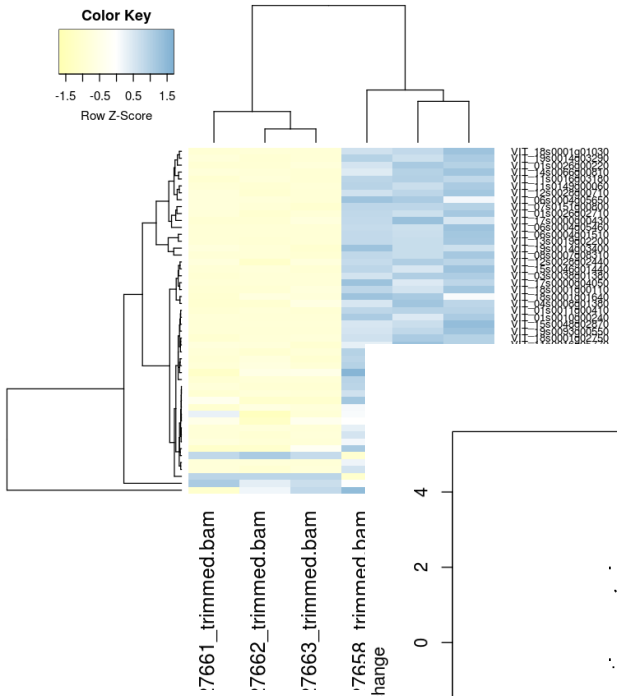
	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
VIT_02s0025g04330	VIT_02s0025g04330	-5.967939	3.903416	-26.01021	1.311512e-11	2.419346e-07	16.32686
VIT_06s0004g05460	VIT_06s0004g05460	-3.438538	6.257815	-21.98922	8.794390e-11	3.955772e-07	15.26717
VIT_19s0093g00550	VIT_19s0093g00550	-6.354018	3.811761	-24.29892	2.839847e-11	2.619333e-07	15.26463
VIT_13s0019g02200	VIT_13s0019g02200	-2.411133	7.400294	-21.56152	1.097765e-10	3.955772e-07	15.09002
VIT_04s0044g00130	VIT_04s0044g00130	-3.106525	5.974479	-20.83956	1.611715e-10	3.955772e-07	14.71222
VIT_01s0026g00220	VIT_01s0026g00220	-2.121529	7.314954	-20.77744	1.666848e-10	3.955772e-07	14.68831
VIT_17s0000g00430	VIT_17s0000g00430	-3.373297	7.214778	-20.72442	1.715519e-10	3.955772e-07	14.64948
VIT_03s0038g01380	VIT_03s0038g01380	-3.059084	6.422147	-20.45446	1.988447e-10	3.992256e-07	14.51462
VIT_14s0171g00360	VIT_14s0171g00360	-4.849779	4.405474	-21.29683	1.261908e-10	3.955772e-07	14.45467
VIT_02s0236g00130	VIT_02s0236g00130	-1.799988	5.688626	-20.14180	2.364765e-10	3.992256e-07	14.34860

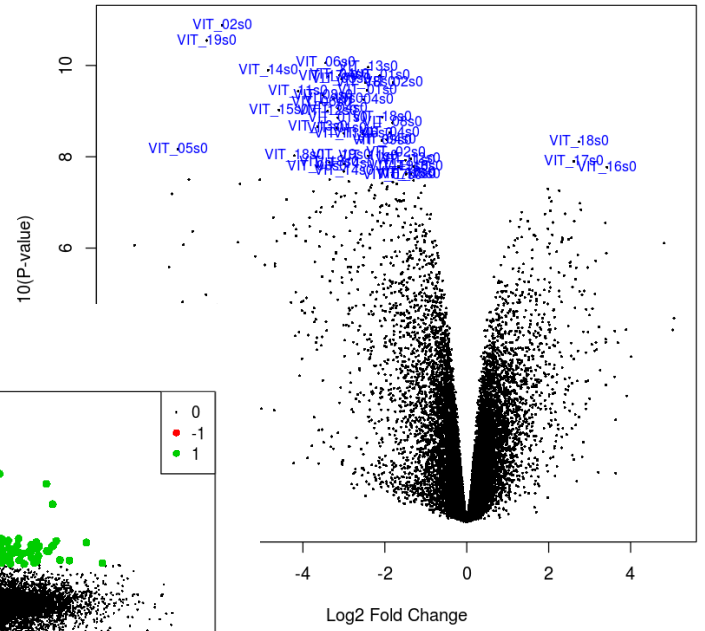
logFC	log2 fold change between compared groups
AveExpr	average log2 expression (across complete dataset)
t	moderated t-statistics
P.Value	raw p-value
adj.P.Val	adjusted p-value for multiple testing issue (FDR correction by default)
B	B-statistics, log-odds that the gene is differentially expressed

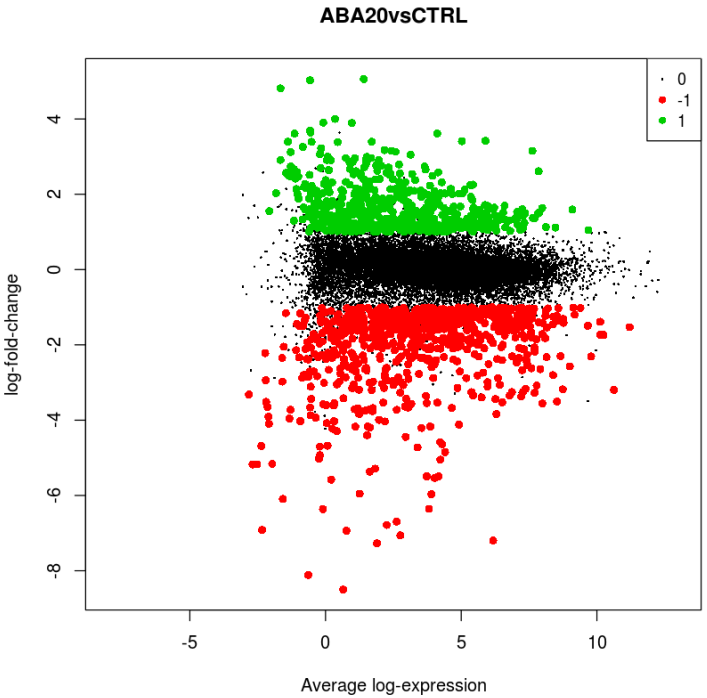

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

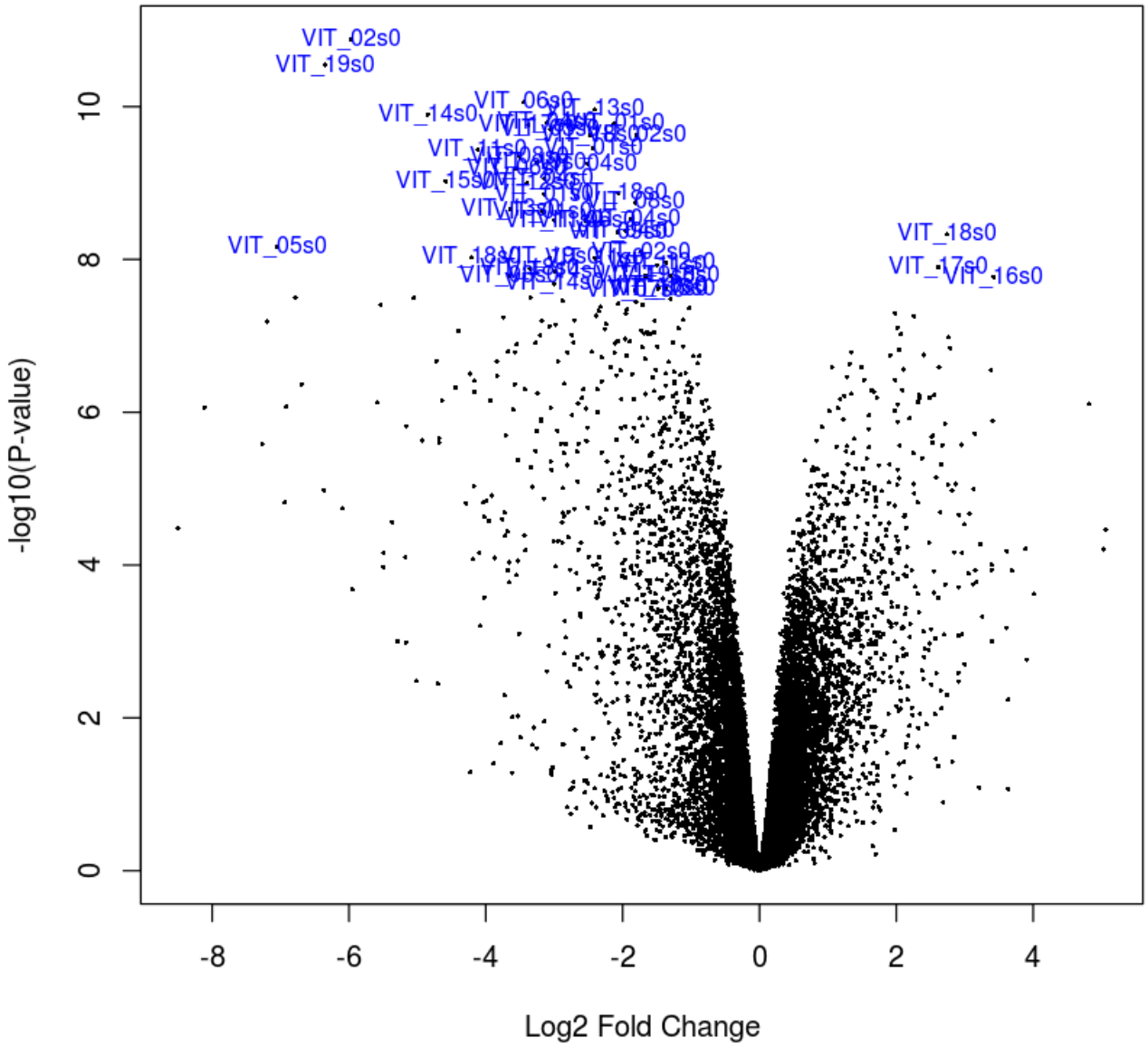

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

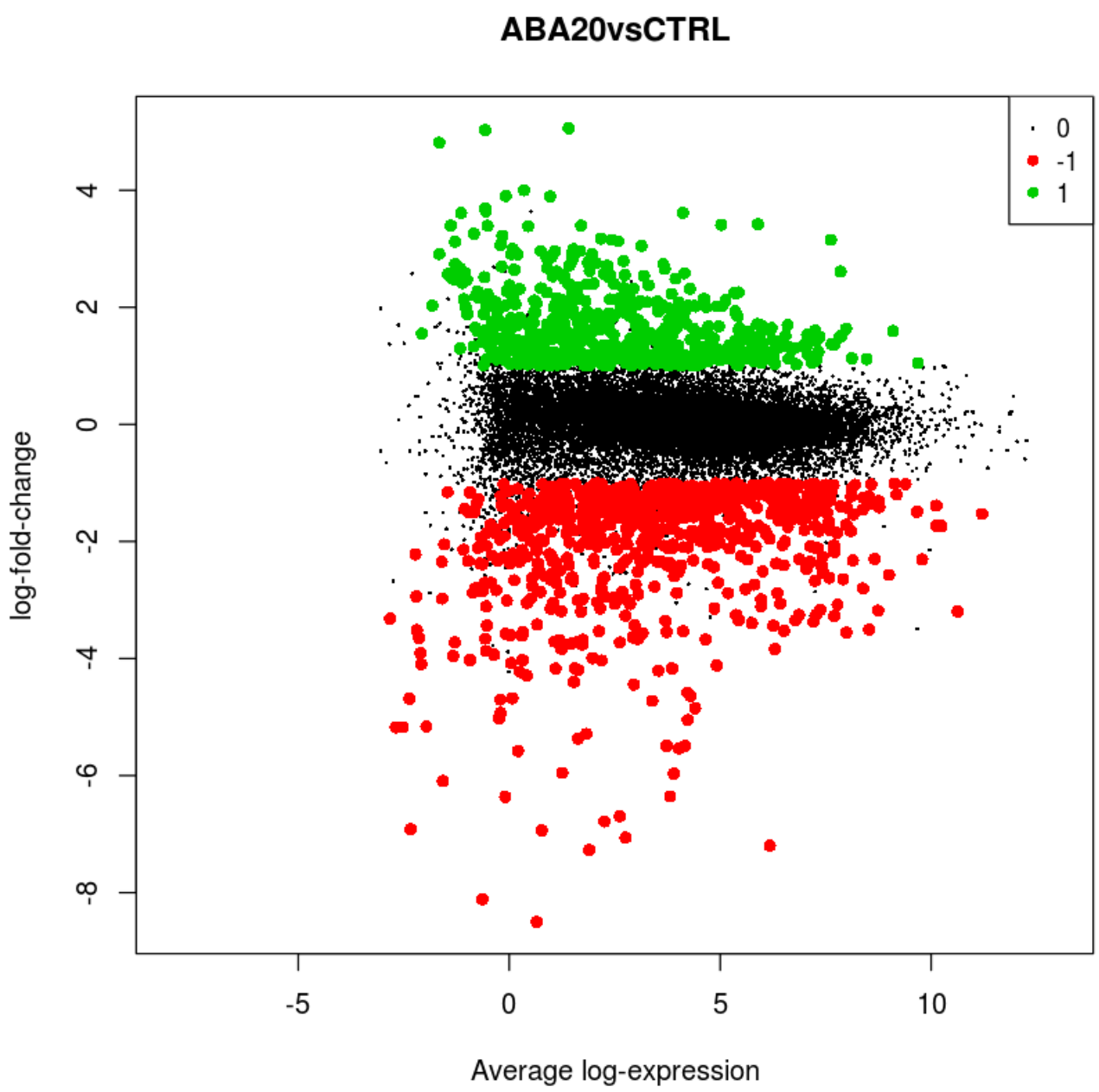
DE genes visualizations

- 





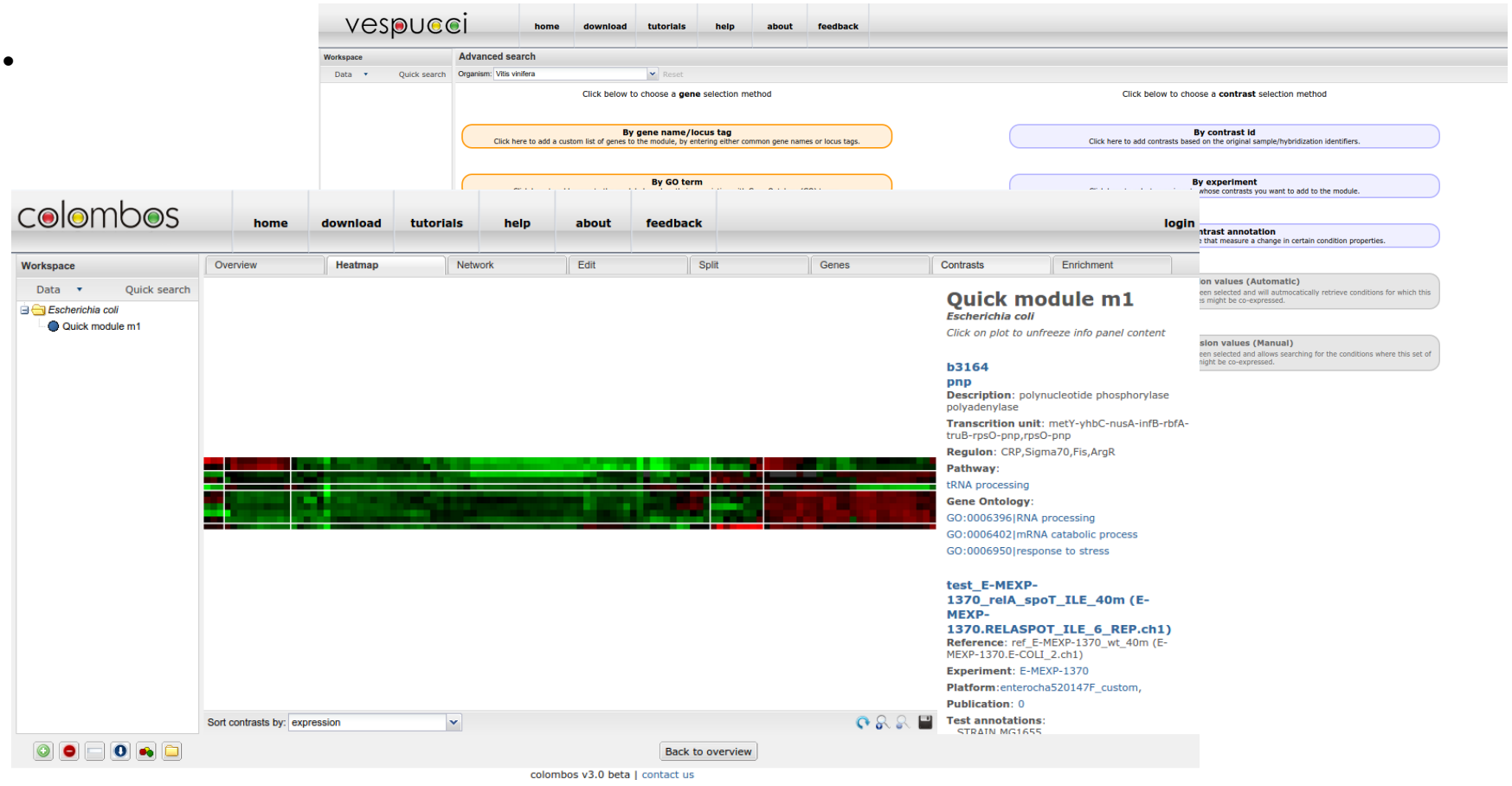

VIT_02s0
VIT_19s0
VIT_14s0
VIT_05s0
VIT_18s0
VIT_17s0
VIT_16s0
-log10(P-value)
0
2
4
6
8
10
-8
-6
-4
-2
0
2
4
Log2 Fold Change


ABA20vsCTRL
0
-1
1
log-fold-change
4
2
0
-2
-4
-6
-8
-5
0
5
10
Average log-expression

Beyond the gene list

- GO enrichment(overrepresentation) analysis:
  - Gene Ontology (BP,CC,MF) categorizes genes into functional groups.
  - Investigate whether gene sets associated with particular biological functions as represented by Gene Ontology annotations are statistically overrepresented in the identified gene groups: statistical test are used to determine which GO terms appear more frequently than would be expected by chance when examining the set of terms annotated to the input genes.
- Integration of transcriptomics data (microarray and RNA-Seq) with other -omics.
- One more thing …

Gene Expression Compendia

- 

http://colombos.net/

http://vespucci.colombos.fmach.it/

Reproducibility

- Bioconductor (16 years of development)
- (R)Markdown
- Python Notebook
- github

References

- Papers:
  - Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
  - Liao Y, Smyth GK and Shi W (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108
  - Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30.
  - MD Robinson, A Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data, Genome biology,
  - Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. 2015. "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, January, gkv007. doi:10.1093/nar/gkv007.
  - Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biology 15 (2): R29. doi:10.1186/gb-2014-15-2-r29.
- Web sites:
  - http://bioconductor.org
  - https://f1000research.com/articles/5-1408/
  - https://dockflow.org/workflow/rnaseq-gene/
  - http://robpatro.com/blog/?p=235
  - https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/