Gene expression profiling

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

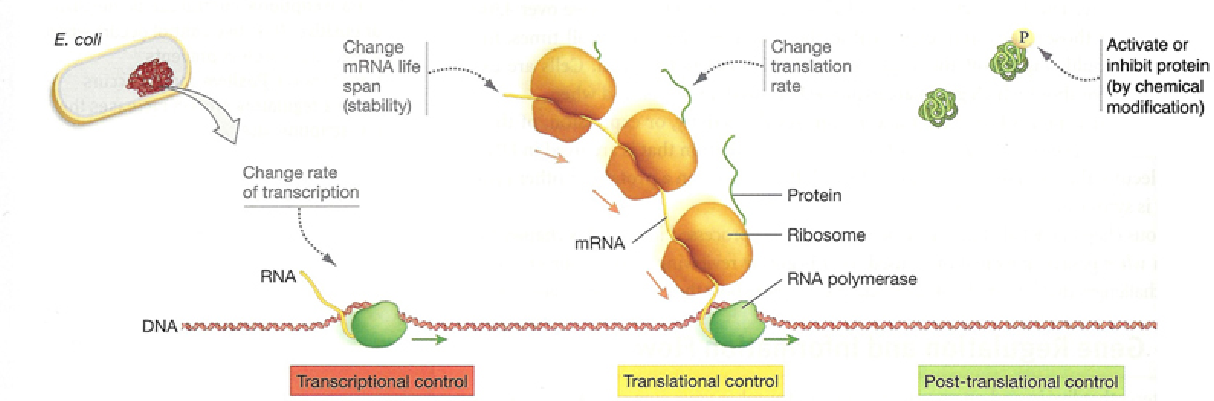
http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Overview

- Gene expression and its profiling
  - what it is and why we measure gene expression,
  - new information about what genes do under various conditions,
  - to understand gene function, discover relationships between genes,
- technologies
  - RNA sequencing
- statistical gene expression models
  - Poisson and negative binomial distribution,
  - generalized linear modesl,
- outcomes of the statistical analysis
  - focus on deregulated genes whose expression changes with changes in experimental conditions,
  - also clustering, dimensionality reduction, classification, correlation analysis.

Gene expression

- Cells must be able to respond to changes in their environment
  - regulation of transcription and translation is critical to this adaptivity,
  - genes unchanged, the changes in the abundance of particular proteins,
- gene expression
  - the process by which information from a gene is used in the synthesis of a functional gene product (protein, functional RNA),
  - its complex multi-level regulation is the basis for cellular differentiation, development, morphogenesis and the versatility and adaptability.


E. coli
Change mRNA life span (stability)
Change translation rate
P
Activate or inhibit protein (by chemical modification)
Change rate of transcription
Protein
Ribosome
mRNA
RNA polymerase
RNA
DNA
Transcriptional control
Translational control
Post-translational control

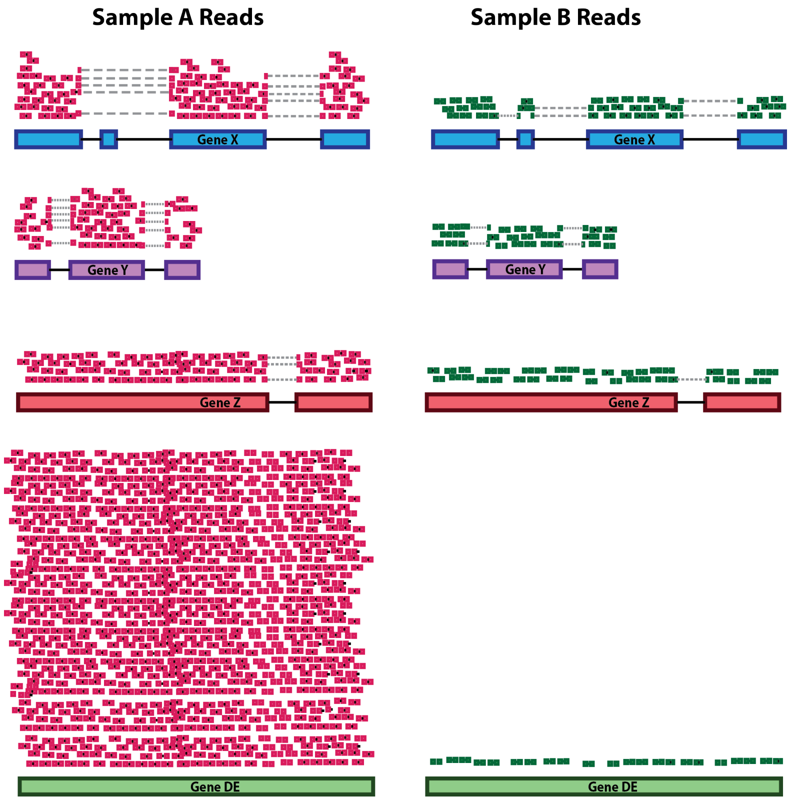
Szauter: BIOL 202 Genetics.

Gene expression profiling

- Gene expression profiling
  - the measurement of the activity (expression) of thousands of genes at once,
  - repeated many times under different experimental conditions,
  - followed by statistical analysis
    (differential expression, clustering, enrichment analysis),
  - new information about what genes do under various conditions,
- helps in gene annotation
  - sequential similarity has its limitations,
  - it cannot identify novel functions of genes/proteins.
- in this lecture, focus on **differentially expressed genes**
  - the genes with statistically significant change in expression levels between two experimental conditions (fold change/expression ratio $\neq 1$),
  - commonly diseased vs healthy, treated vs untreated, wildtype vs strain X.

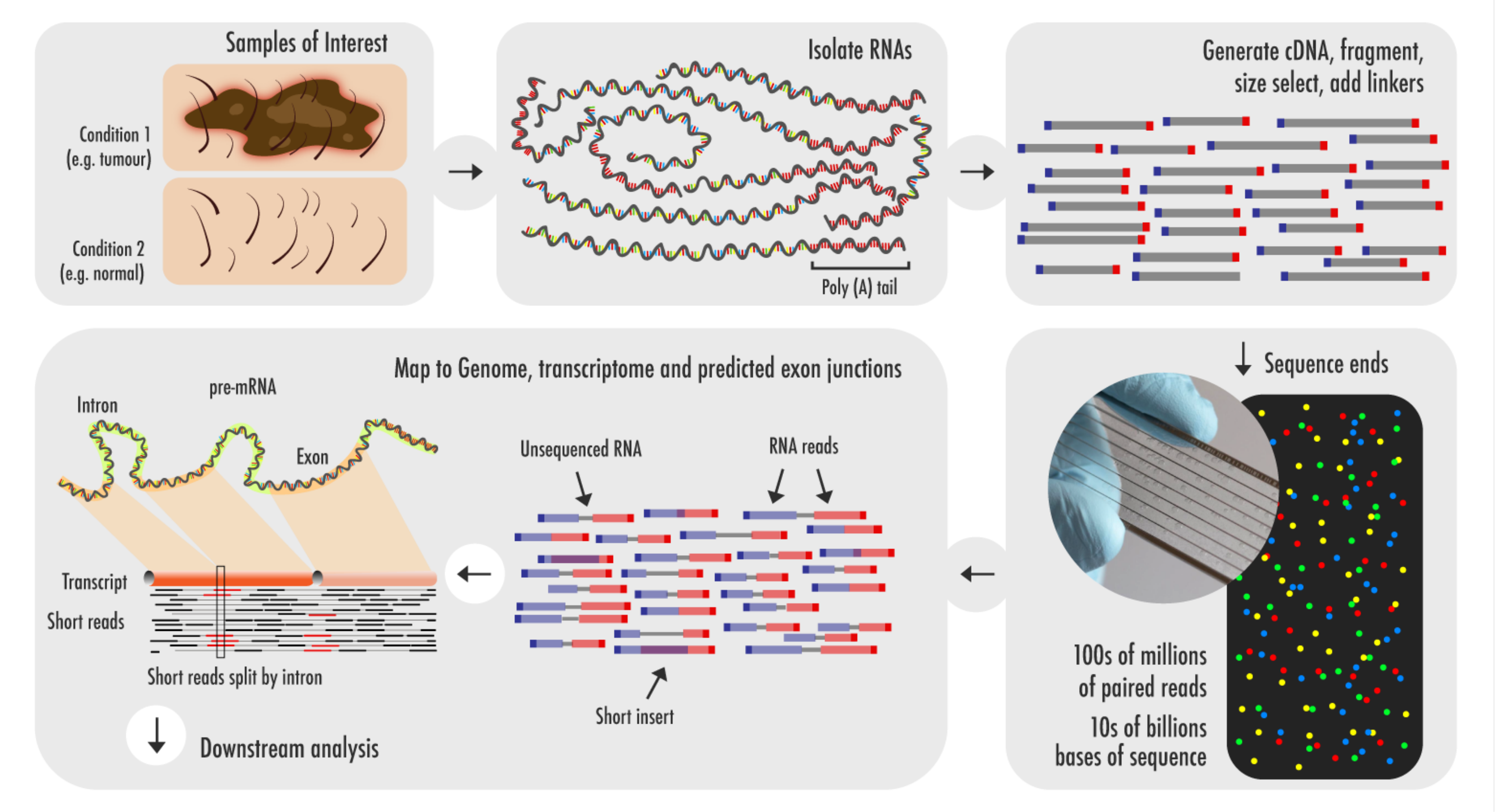
Gene expression profiling

- How to examine the RNA quantity?
  - DNA microarrays
    - ∗ dedicated probes,
    - ∗ hybridization,
  - RNA sequencing
    - ∗ next generation sequencing,
- Typical outcome:
  - a read count data table,
  - axes: transcripts/genes
    samples/libraries,
  - entries: the read counts.



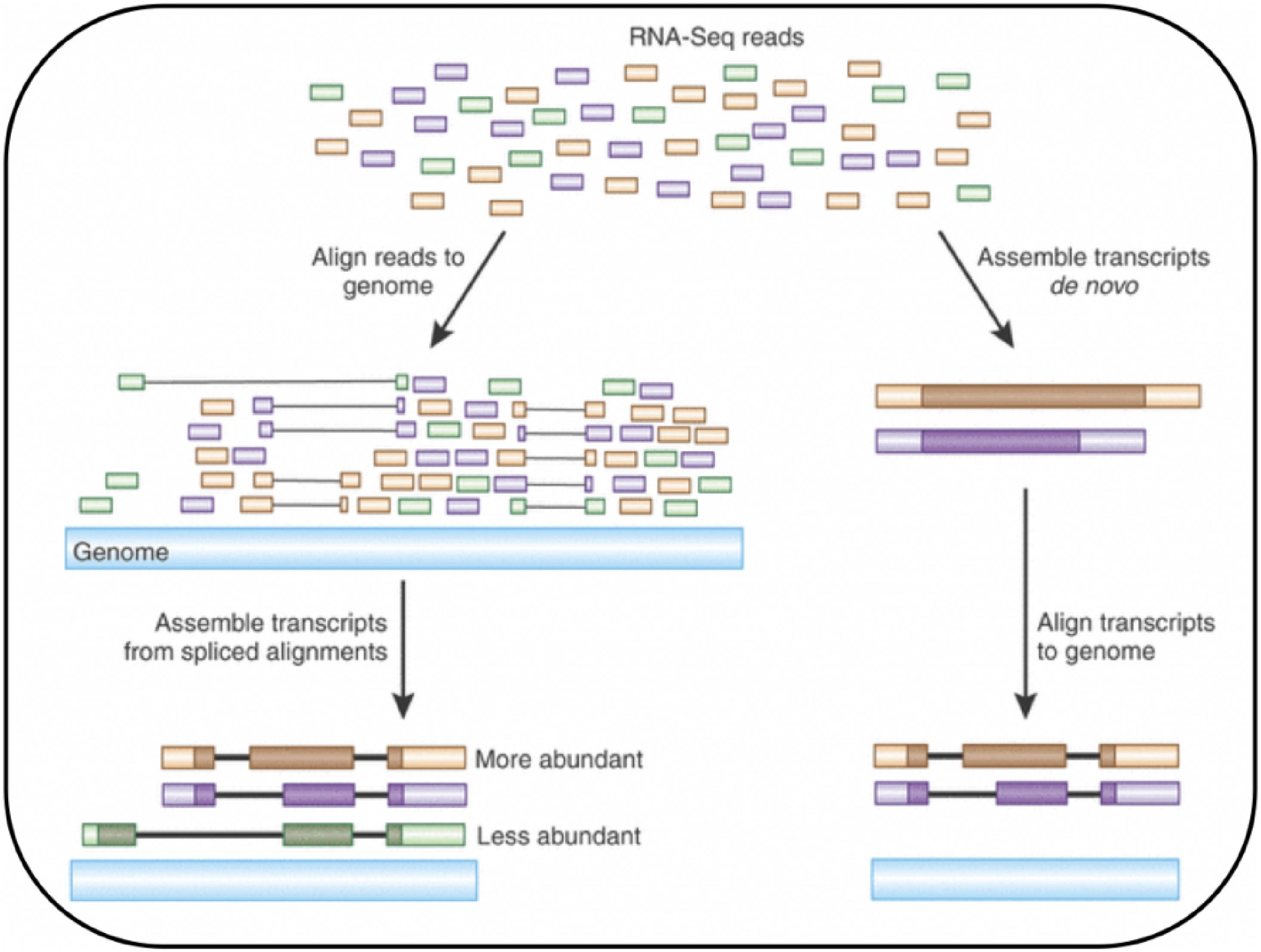
Introduction to DGE.

RNA sequencing



Mackenzie: RNA-seq: Basics, Applications and Protocol.

Read mapping


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Haas and Zody: Advancing RNA-seq analysis.

Read count as a random variable

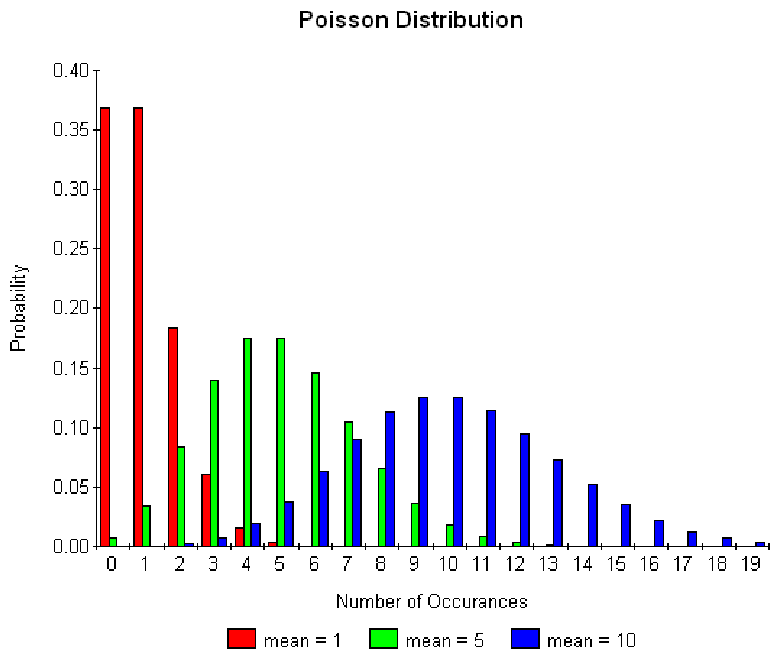
- Consider the read count for a transcript observed in a set of samples
  - the read count is a non-negative discrete variable,
  - the simplest way is to model it with the **Poisson distribution**
    - it expresses the probability of a given number of events k occurring in a fixed interval of time or space,
    - these events occur with a constant mean rate λ,
    - these events appear independently

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

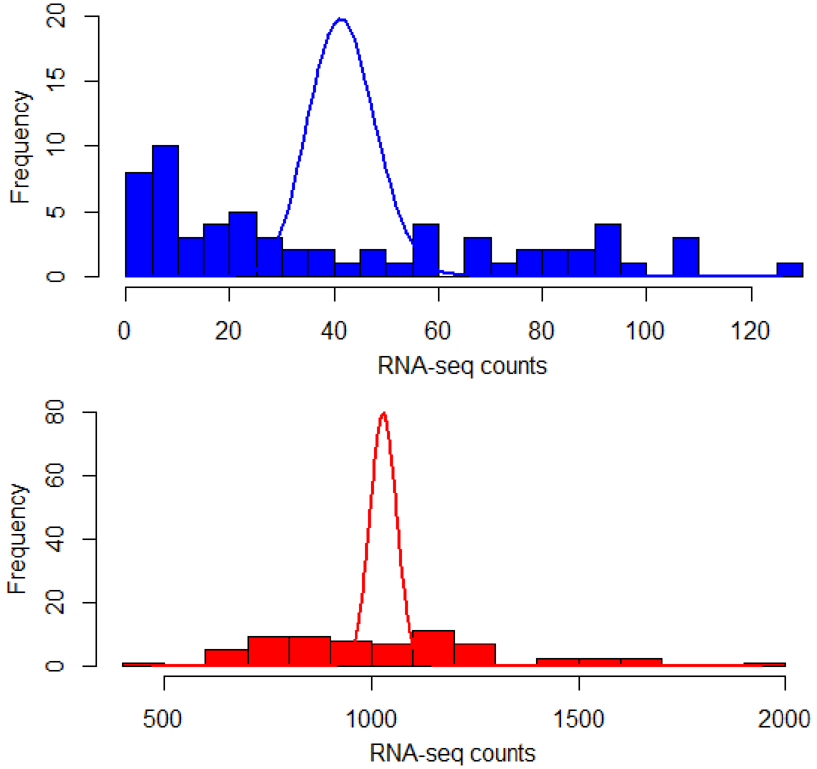
  - in the case of RNA-seq data
    - event = a read matches a transcript,
    - fixed space = transcipt, samples = realizations of random variable.

RNA-seq data and Poisson distribution

- Poisson distribution assumes that mean and variance are equal (given by λ)
  - this is often not true for RNA-Seq data.



Poisson distributions with 3 different λs.
Variance grows with mean.



RNA-seq counts have higher variance than expected by Poisson dist.
Histograms of two example transcripts in about 70 samples shown.

RNA-seq data and negative binomial distribution

- Employ the **negative binomial (NB) distribution** instead
  - in a sequence of independent and identical Bernoulli trials with success probability p, we observe k success trials before the r-th failure

$$f(k; r, p) = \binom{k + r - 1}{k} (1 - p)^r p^k$$

  - mean is smaller than variance

$$\mu = \frac{pr}{1 - p} \qquad \sigma^2 = \frac{pr}{(1 - p)^2}$$

- let us **reparametrize** $NB(r, p)$ using mean μ and dispersion α instead of r and p

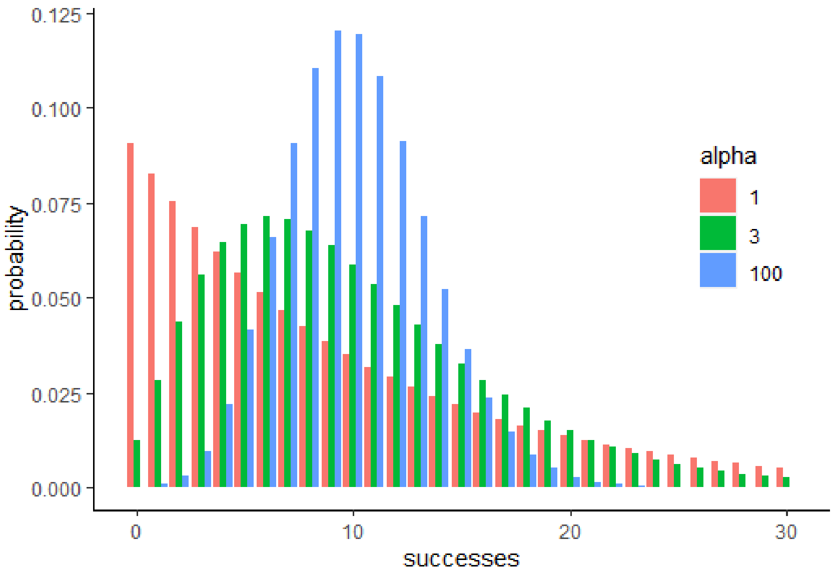
$$r = \alpha \qquad p = \frac{\mu}{\alpha + \mu}$$

- the new form of $NB(\mu, \alpha)$

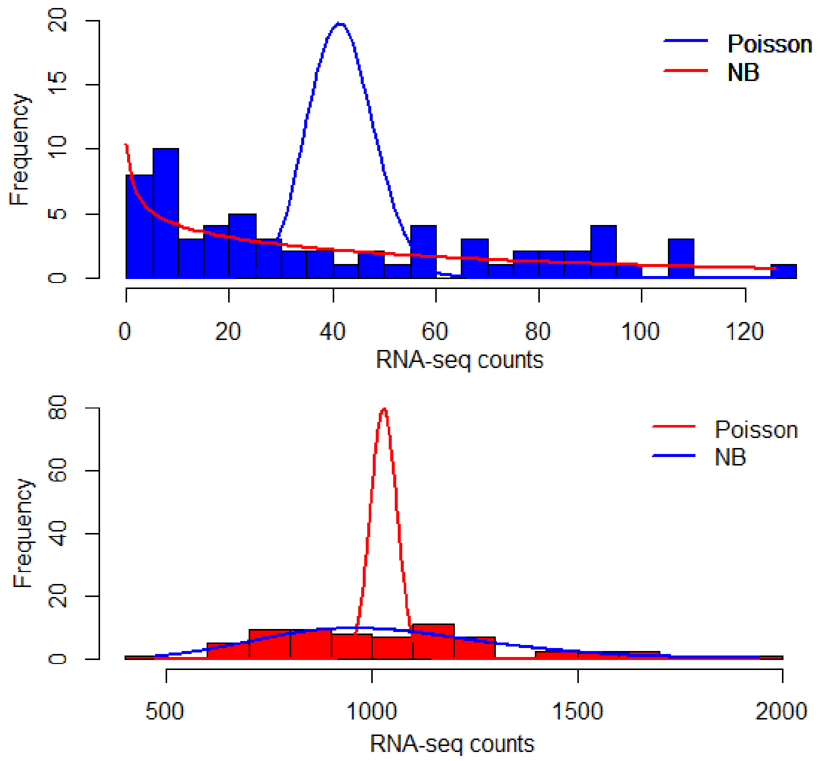
$$f(k; \mu, \alpha) = \binom{k + \alpha - 1}{k} \left(\frac{\alpha}{\alpha + \mu}\right)^{\alpha} \left(\frac{\mu}{\alpha + \mu}\right)^{k}$$

RNA-seq data and negative binomial distribution

- NB distribution allows to fit overdispersed count data
  - we can compare fits of Poisson and NB model to decide whether overdispersion occurs.

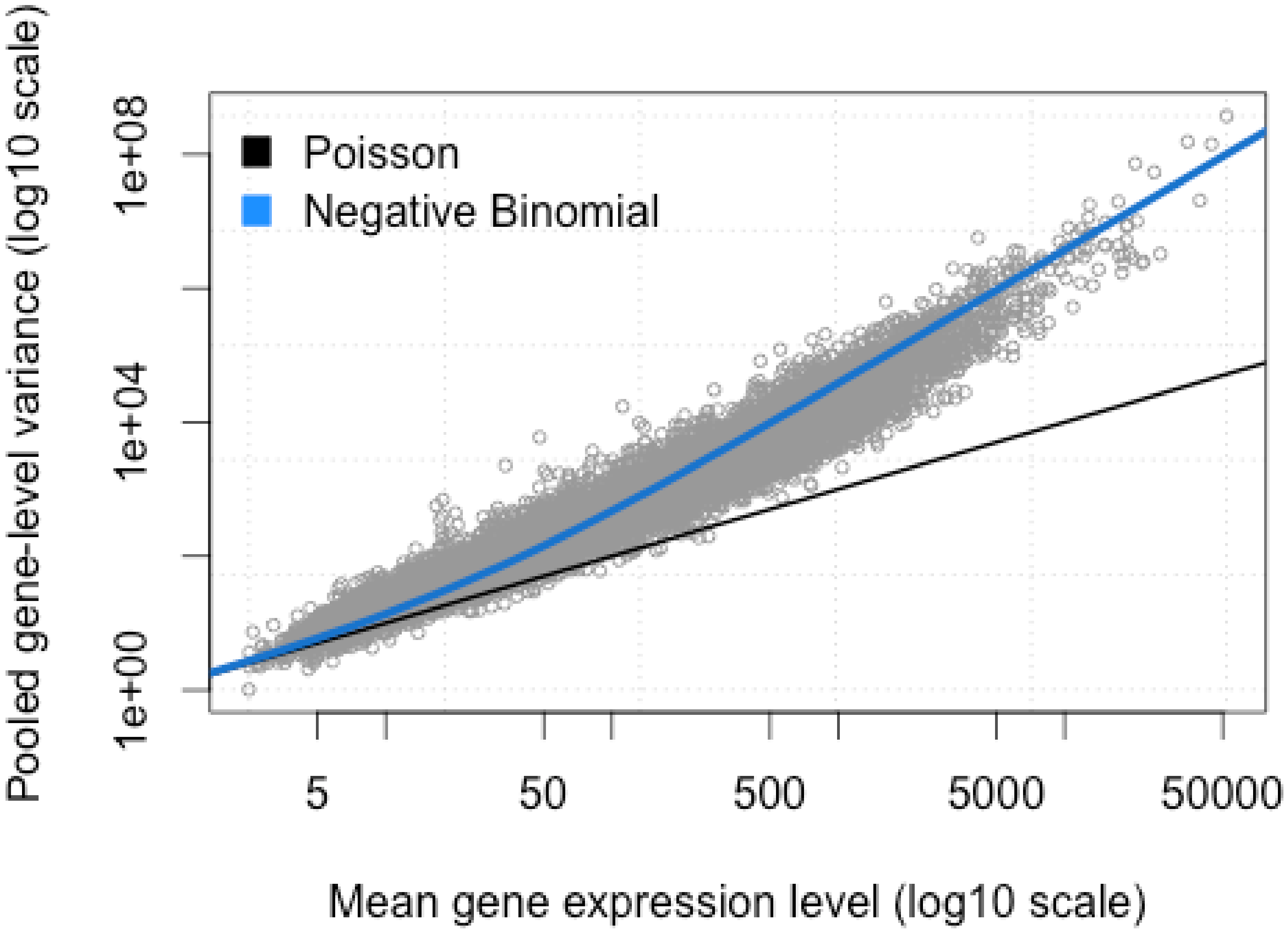


NB distributions with 3 different α values.
Variance decreases with increasing α.
μ is kept 10 in the three distributions being shown.



NB model much better fits RNA-seq counts.

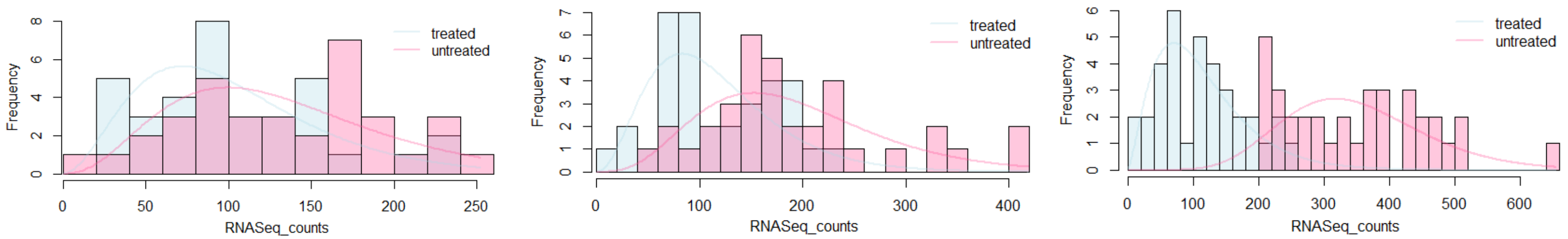
Mean and variance in RNA-seq data (ReCount project)



https://github.com/bioramble/sequencing

Differential expression

- Up to now, we have seen expression distributions in all the samples,
- there could be **variables of interest** whose influence has to be considered,
- let us have the following experiment
  - response variable: read count for a transcript t,
  - factor to study: treatment with a new drug d,
  - experimental design: 70 units/people, a randomly selected half is treated with d, the rest of people untreated/placebo,
- possible outcomes
  - d regulates the mean transcription level of t or it remains unchanged.



Three transcripts with increasing chance of differential expression.

Generalized linear model (GLM)

- There could be more experimental variables that influence the expression
  - multiple factors of interest,
  - other confounders that could not be fully controlled: age of people in the study, personnel that carries out the experiment, cell distribution in the bulk sample, etc.
  - a multivariate model is generally needed,
- linear regression model
  - its assumptions (linearity, homoscedasticity, normality) not met here,

$$E(Y) = \mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- **generalized linear model**
  - introduces a link function g, often non-linear,

$$g(E(Y)) = g(\mu_{Y|X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  - for count response models (Poisson, NB) g=log().

GLM with a negative binomial distribution

- Transcript read count is a generalized linear function of exp. conditions
  - i – transcript index, j – sample index, r – covariate (treatment) index,
  - s_{ij} – transcript and sample specific factor,
  - x_{jr} – treatment r of sample j,
  - β_{ir} – logarithmic fold change for transcript i contributed by covariate r,
- Observed read count Y_{ij} of a transcript i in sample j

$$Y_{ij} \approx NB(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i)$$

- Mean read count proportional to the true transcript count q_{ij}

$$E(Y_{ij}) = \mu_{ij} = s_{ij} q_{ij}$$

- Nonlinear (log) link function

$$\log \frac{E(Y_{ij})}{s_{ij}} = \log q_{ij} = \sum_r x_{jr} \beta_{ir}$$

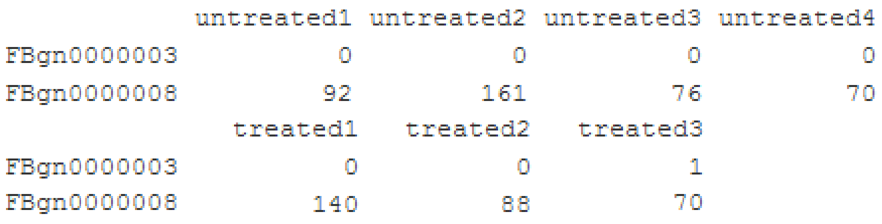
The Pasilla gene RNA-seq experiment

- Pasilla (PS) gene knock-down
  - the Drosophila melanogaster ortholog of mammalian NOVA1/2,
  - PS gene regulates alternative splicing of pre-mRNA,
- Experiment: Pasilla is depleted (treated) and RNA-seq is measured,
- Control: wild type (untreated) RNA-seq is measured,
- What genes are differentially expressed in response to Pasilla depletion?
- see Brooks et al.: Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. 2011.

Drosophila melanogaster, https://www.yourgenome.org

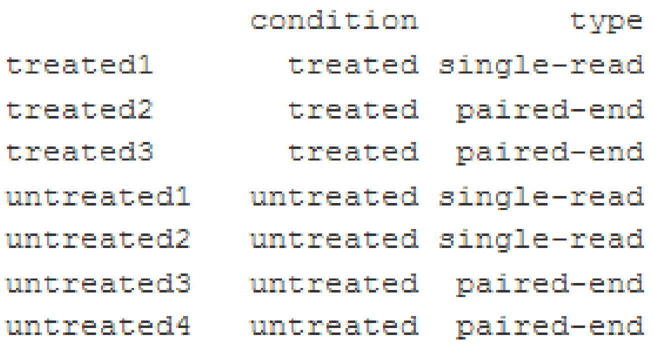
The Pasilla experiment, experimental design

- 7 samples available
  - condition: 3 of them treated (PS depleted), 4 untreated (wild type),
  - data type: 3 single-read samples and 4 paired-end read samples,
- experimental design in GLM
  - $g(Y) = X\beta$,
  - Y: (normalized) transcript counts,
  - X: covariates (condition, data type, interactions),
  - build GLM for $\approx$ 15,000 transcripts.



```
            untreated1 untreated2 untreated3 untreated4
FBgn0000003          0          0          0          0
FBgn0000008         92        161         76         70
              treated1   treated2   treated3
FBgn0000003          0          0          1
FBgn0000008        140         88         70
```

Count matrix (two transcripts shown only).

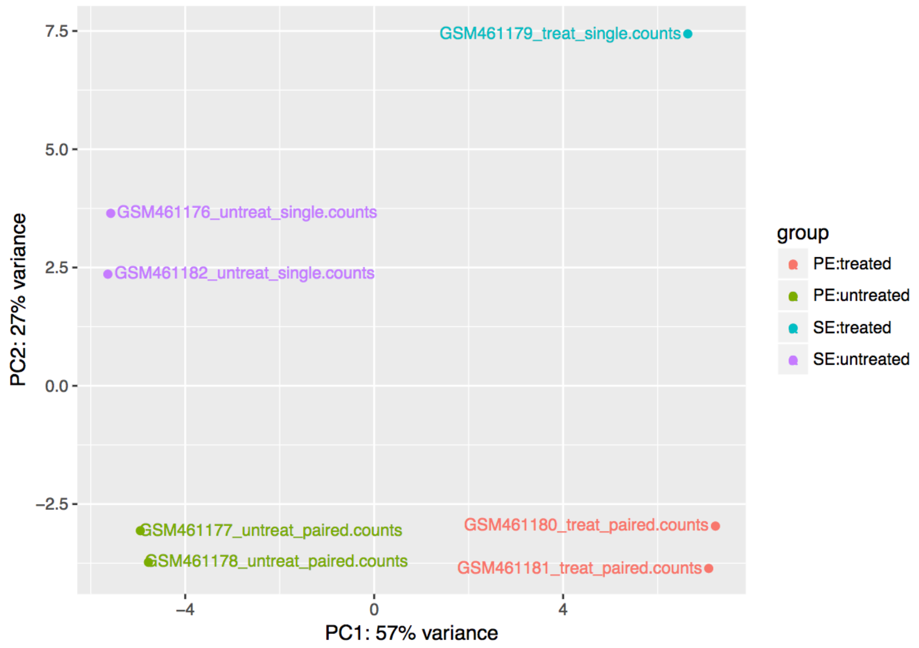


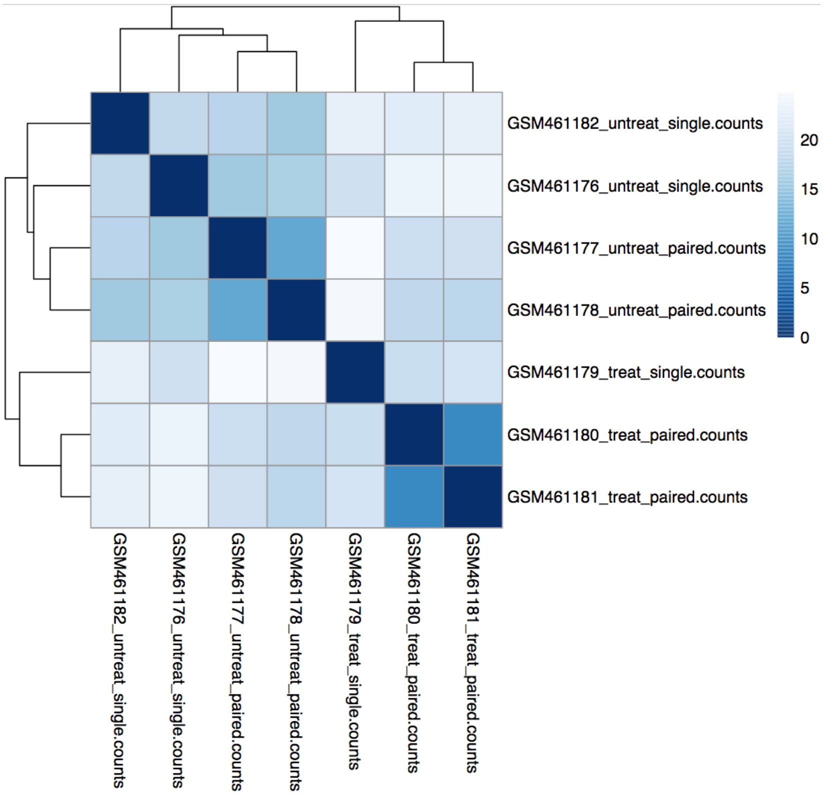
```
             condition        type
treated1       treated single-read
treated2       treated  paired-end
treated3       treated  paired-end
untreated1   untreated single-read
untreated2   untreated single-read
untreated3   untreated  paired-end
untreated4   untreated  paired-end
```

Sample information.

Pasilla, relationships between sample expression profiles

- Employ dimensionality reduction (PCA) and/or clustering (hierarchical).

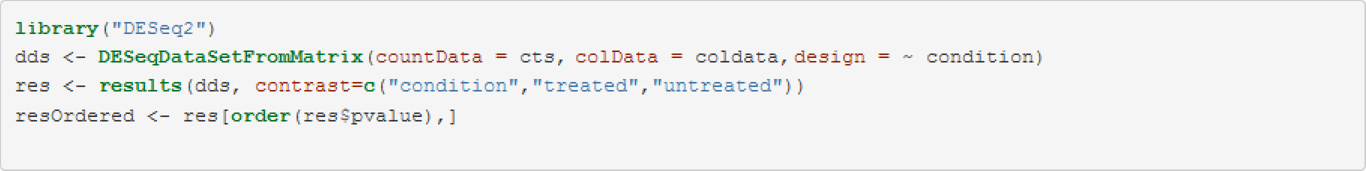




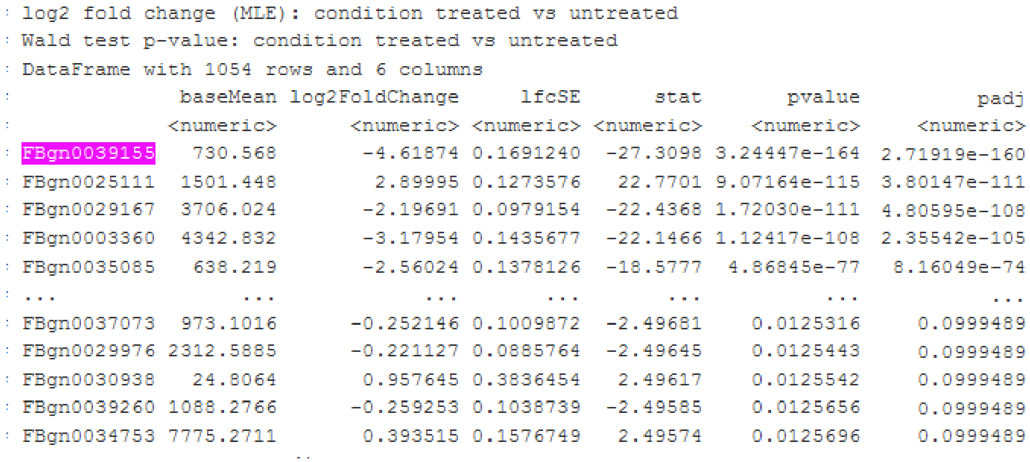
Galaxy Training, https://training.galaxyproject.org/

Pasilla, differentially expressed genes

- Can be found e.g., with DESeq2 tool (R package)
  - it implements NB GLM,
  - improved with shrinkage estimators for dispersion and fold change.

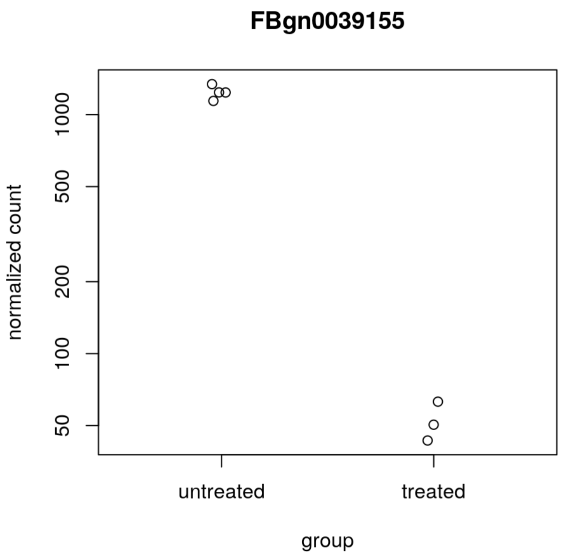


```
library("DESeq2")
dds <- DESeqDataSetFromMatrix(countData = cts, colData = coldata,design = ~ condition)
res <- results(dds, contrast=c("condition","treated","untreated"))
resOrdered <- res[order(res$pvalue),]
```



```
: log2 fold change (MLE): condition treated vs untreated
: Wald test p-value: condition treated vs untreated
: DataFrame with 1054 rows and 6 columns
:              baseMean log2FoldChange     lfcSE      stat       pvalue         padj
:             <numeric>      <numeric> <numeric> <numeric>    <numeric>    <numeric>
: FBgn0039155   730.568       -4.61874 0.1691240  -27.3098 3.24447e-164 2.71919e-160
: FBgn0025111  1501.448        2.89995 0.1273576   22.7701 9.07164e-115 3.80147e-111
: FBgn0029167  3706.024       -2.19691 0.0979154  -22.4368 1.72030e-111 4.80595e-108
: FBgn0003360  4342.832       -3.17954 0.1435677  -22.1466 1.12417e-108 2.35542e-105
: FBgn0035085   638.219       -2.56024 0.1378126  -18.5777  4.86845e-77  8.16049e-74
: ...               ...            ...       ...       ...          ...          ...
: FBgn0037073  973.1016      -0.252146 0.1009872  -2.49681    0.0125316    0.0999489
: FBgn0029976 2312.5885      -0.221127 0.0885764  -2.49645    0.0125443    0.0999489
: FBgn0030938   24.8064       0.957645 0.3836454   2.49617    0.0125542    0.0999489
: FBgn0039260 1088.2766      -0.259253 0.1038739  -2.49585    0.0125656    0.0999489
: FBgn0034753 7775.2711       0.393515 0.1576749   2.49574    0.0125696    0.0999489
```

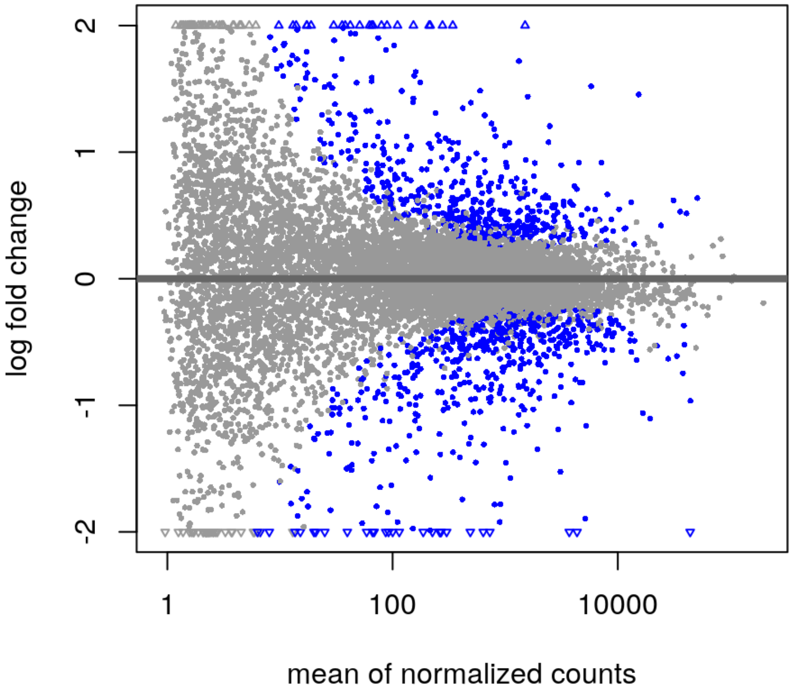
DESeq2 outcome, Love et al.: RNA-seq workflow.



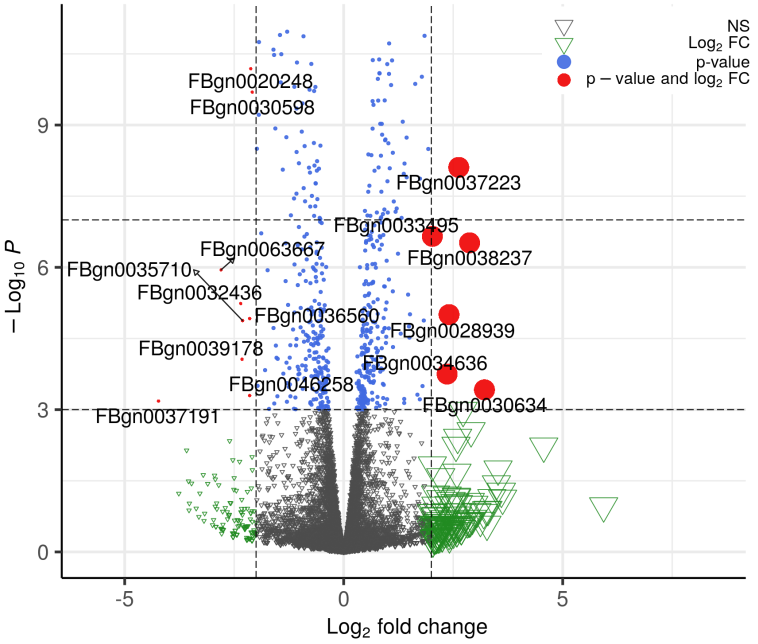
An example of DEG,

Pasilla, differentially expressed genes

- Can be visualised with MA plot or Volcano plot
  - the dots correspond to transcripts,
  - differential expression supported by a high fold change and small p-value.



Love et al.: RNA-seq workflow.



Blighe et al.: EnhancedVolcano.

Summary

- RNA-sequencing
  - NGS technique that examines quantity and sequences of RNA in a sample,
  - can be used for gene expression profiling between samples,
  - also to study alternative splicing events associated with diseases,
  - identification of allele-specific expression, etc.
- negative binomial generalized linear models (NB GLMs)
  - case studies show their usefulness on datasets with different characteristics,
  - find more differentially expressed genes with statistical evidence,
  - the genes truly biologically relevant (could be validated e.g., by qPCR),
- other issues
  - RNA-seq quality control
    - FASTQ raw reads, the read numbers, GC content, base quality scores,
  - feature count normalization
    - sequencing depth, gene length, RNA composition.