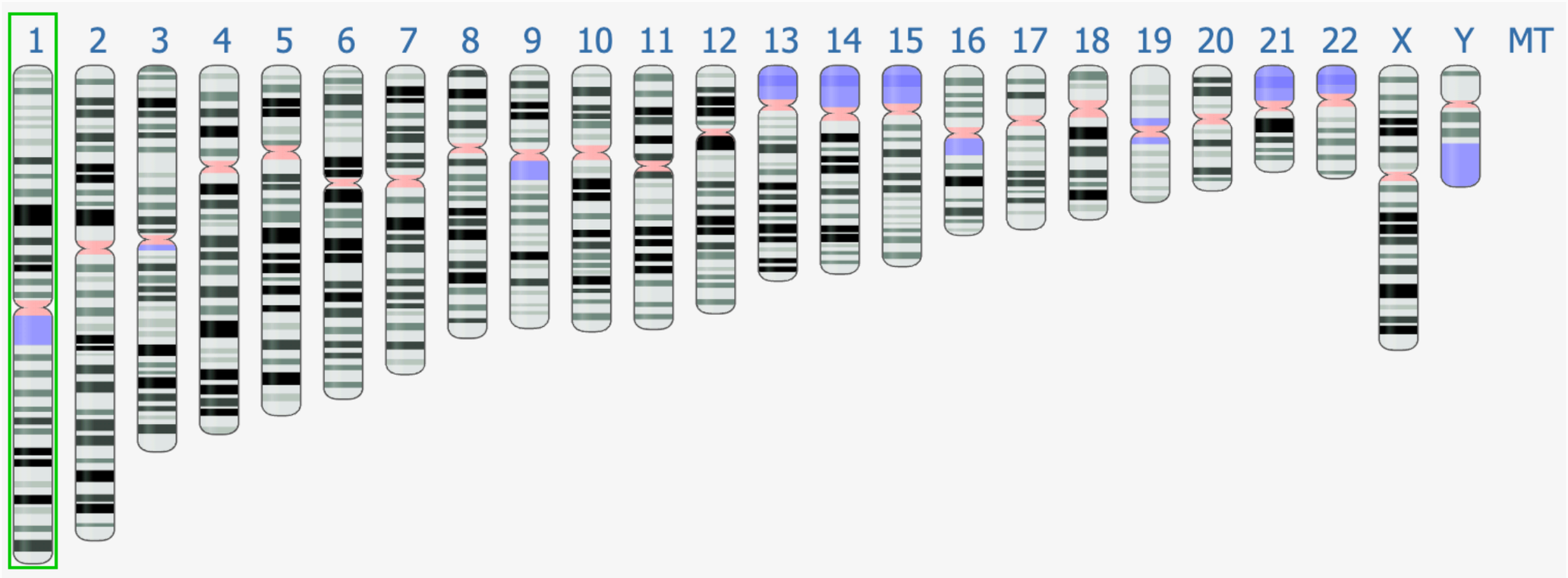



# Genomics110

## BIOL647
## Digital Biology

Rodolfo Aramayo

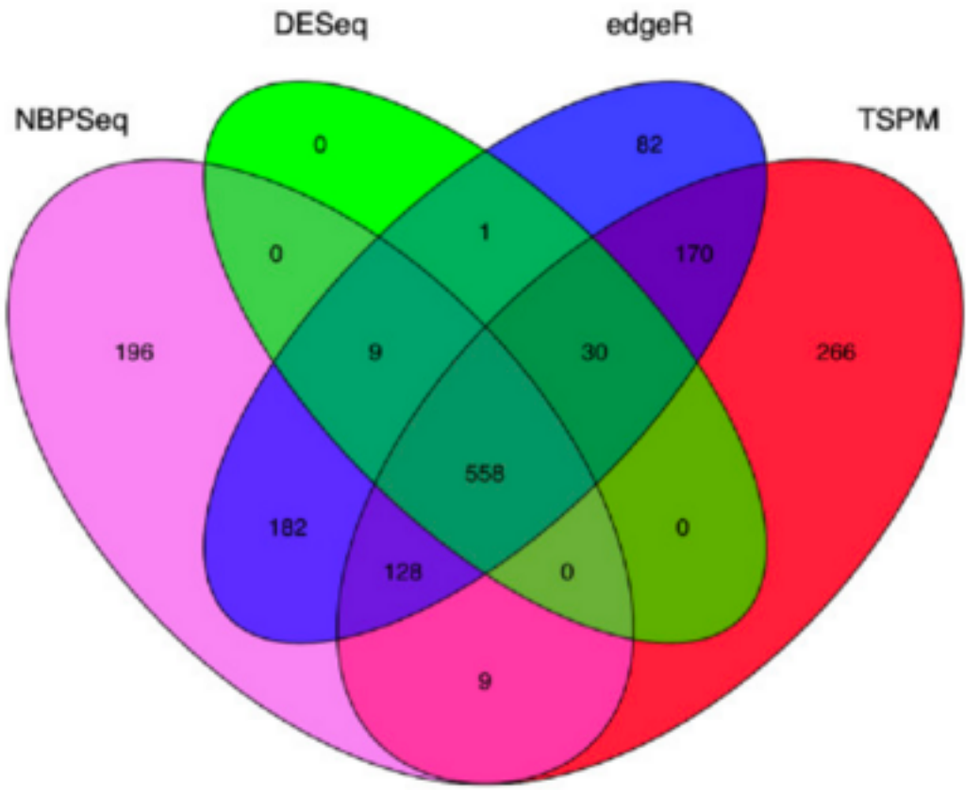
# Genomics110

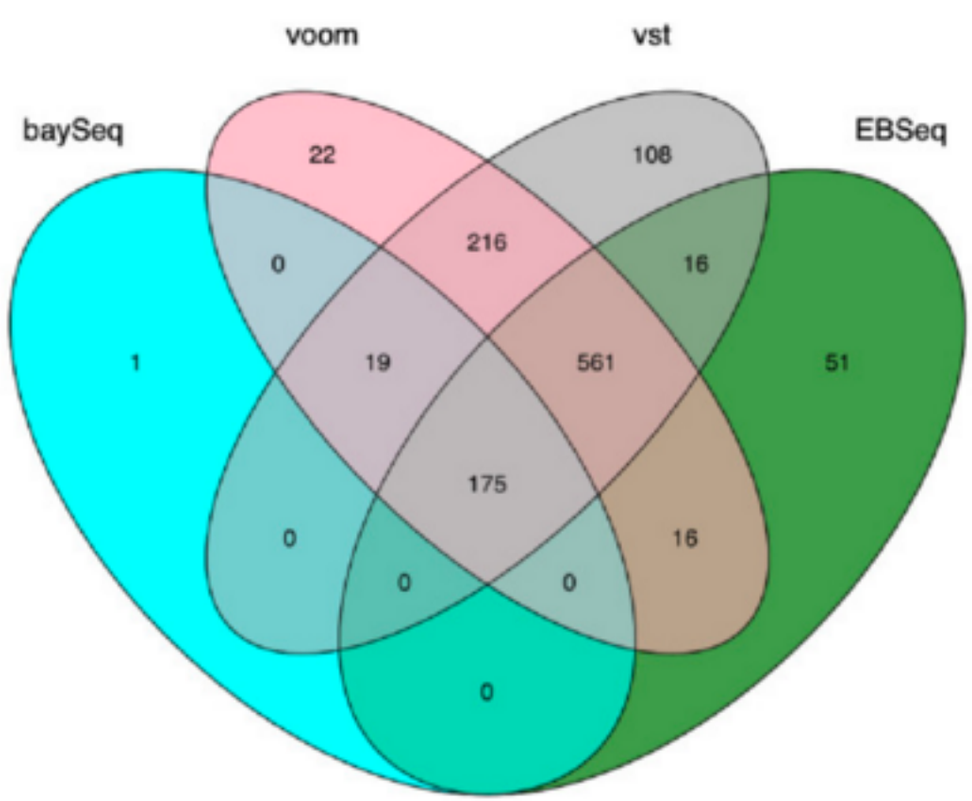
## Tools for gene-level differential expresssion analysis

- There are a number of software packages that have been developed for differential expression analysis of RNA-seq data.

- Many studies describing comparisons between these methods show that while there is some concordance in the genes that are identified as differentially expressed, there is also much variability between tools.

- Additionally, there is no one method that performs optimally under all conditions (Soneson and Dleorenzi, 2013).



# Genomics110

## Tools for transcript-level differential expression analysis

- Until this point we have focused on looking for expression changes at the gene-level.

- If you are interested in looking at splice isoform expression changes between groups, not that the previous methods (i.e DESeq2) will not work.

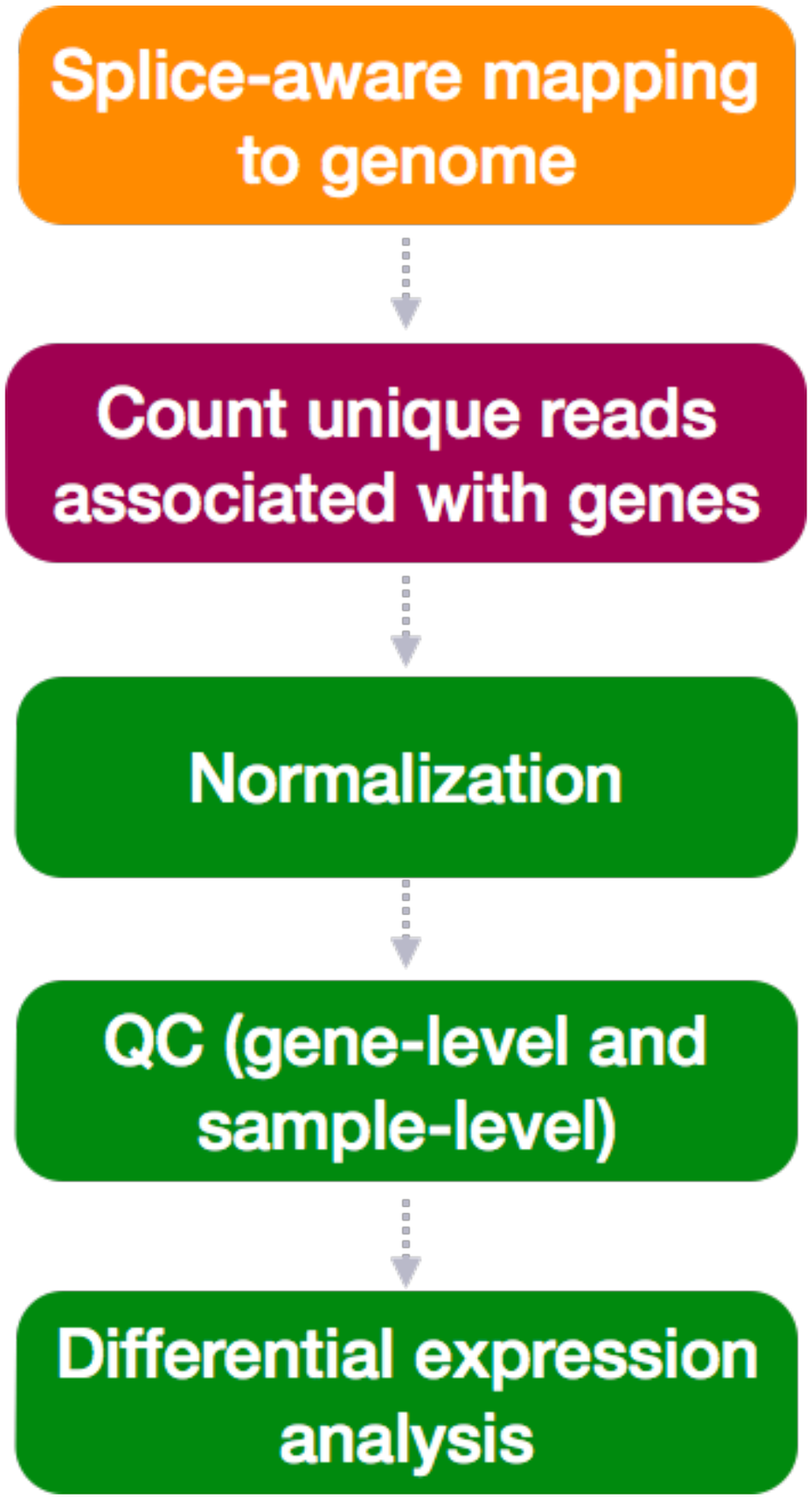
- To demonstrate how to identify transcript-level differential expression we will describe a tool called Sleuth.
  - Sleuth is a fast, lightweight tool that uses transcript abundance estimates output from pseudo-alignment algorithms that use bootstrap sampling, such as Sailfish, Salmon, and Kallisto, to perform differential expression analysis of gene isoforms.
  - Sleuth accounts for this technical variability by using bootstraps as a proxy for technical replicates, which are used to model the technical variability in the abundance estimates. Bootstrapping essentially calculates the abundance estimates for all genes using a different sub-sample of reads during each round of bootstrapping.
  - The variation in the abundance estimates output from each round of bootstrapping is used for the estimation of the technical variance for each gene.

- More information about the theory/process for sleuth is available in the Nature Methods paper, this blogpost and step-by-step tutorials are available on the sleuth website.

- NOTE: Kallisto is distributed under a non-commercial license, while Sailfish and Salmon are distributed under the GNU General Public License, version 3.

# Genomics110

## Count modeling and Hypothesis testing



# Genomics110

## Evaluating Results

There's a new RNA-seq metric on the block…

- We used to report RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million)
  - These normalized read counts for:
    - 1) The sequencing depth (that's the "Million" part)
      - Sequencing runs with more depth will have more reads mapping to each gene.
    - 2) The length of the gene (that's the "Kilobase" part)
      - Longer genes will have more reads mapping to them.
- Now they want us to use TPM – Transcripts per million

https://youtu.be/TTUrtCY2k-w

# Genomics110

## Normalization of count data

- **The first step in the DE analysis workflow is count normalization, which is necessary to make accurate comparisons of gene expression between samples.**

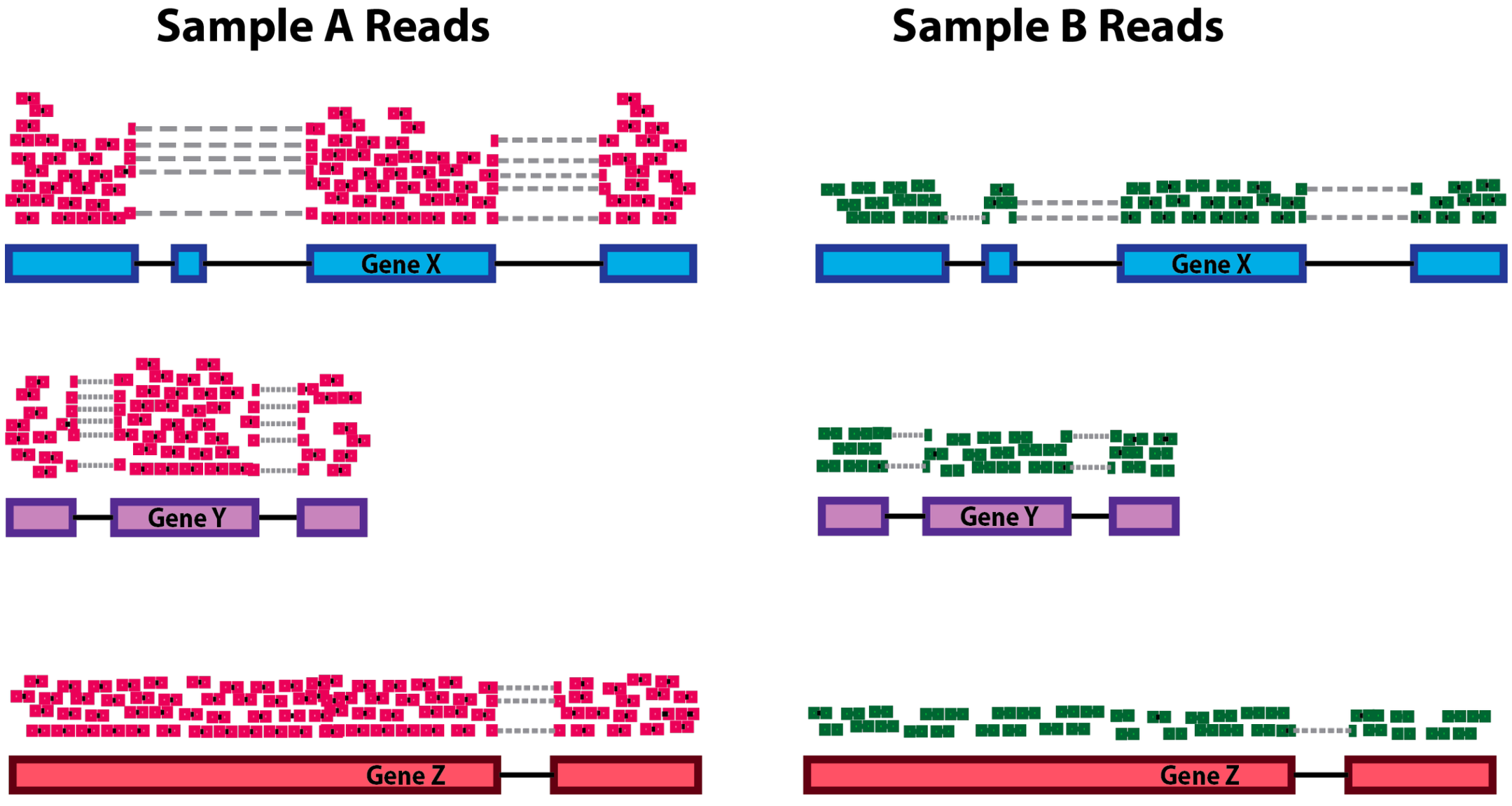
- **The counts of mapped reads for each gene is proportional to the expression of RNA (“interesting”) in addition to many other factors (“uninteresting”).**

- **Normalization is the process of scaling raw count values to account for the “uninteresting” factors. In this way the expression levels are more comparable between and within samples.**

# Genomics110

## Normalization of count data

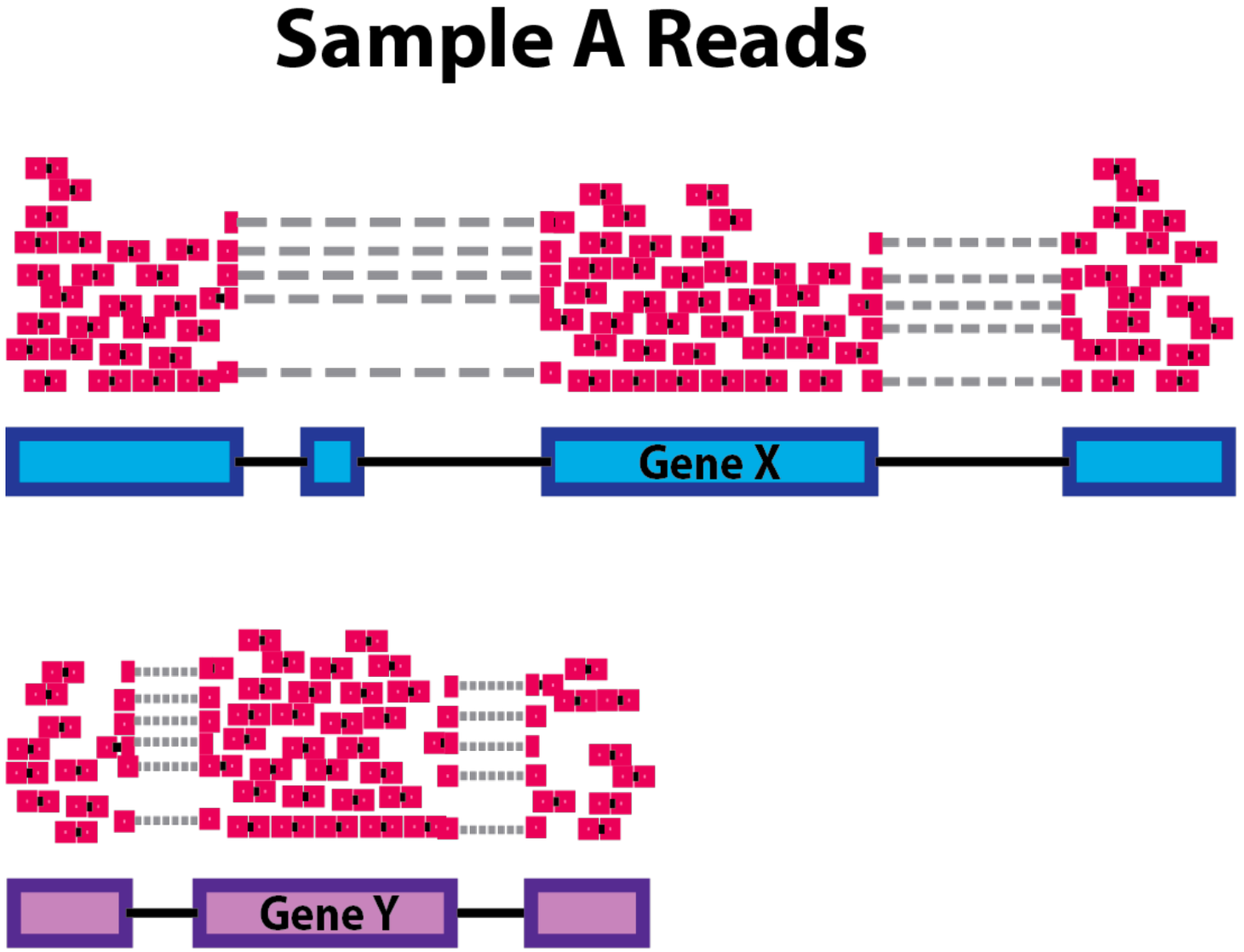
- **The main factors often considered during normalization are:**
  - **Sequencing depth: Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in Sample A relative to Sample B, however this is a consequence of Sample A having double the sequencing depth.**



# Genomics110

## Normalization of count data

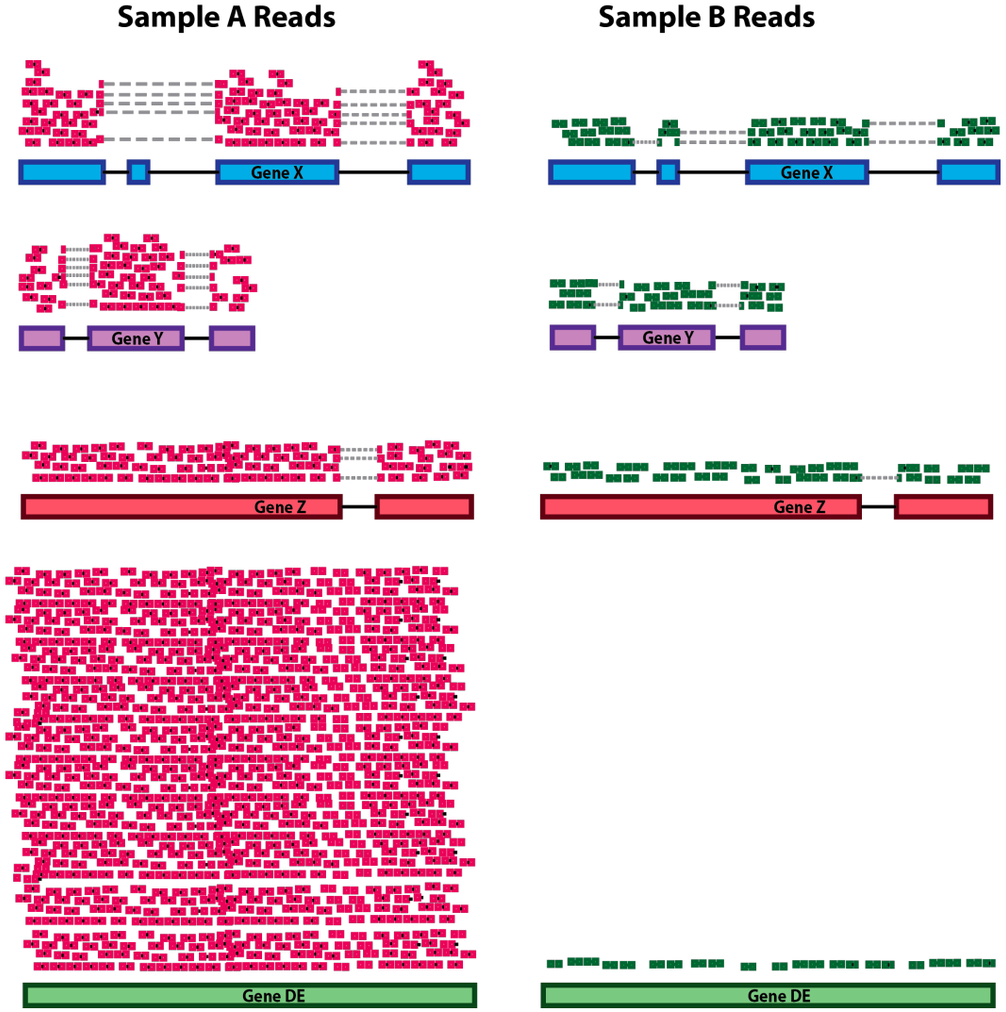
- **The main factors often considered during normalization are:**
  - **Gene length: Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, Gene X and Gene Y have similar levels of expression, but the number of reads mapped to Gene X would be many more than the number mapped to Gene Y because Gene X is longer.**



# Genomics110

## Normalization of count data

- **The main factors often considered during normalization are:**
  - **RNA composition: A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples, and is particularly important when performing differential expression analyses [10.1186].**

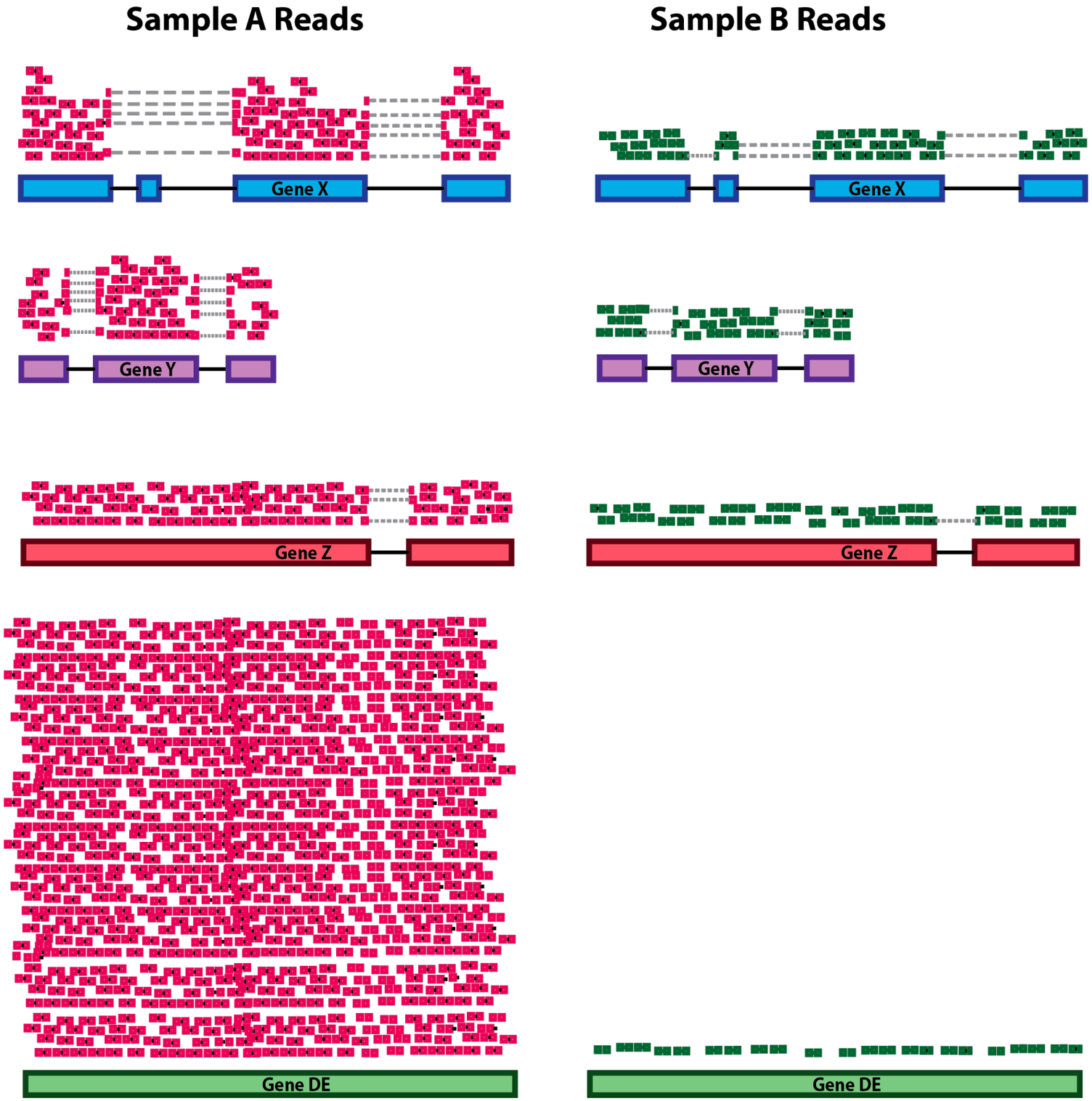


# Genomics110

## Normalization of count data

- In the example, if we were to divide each sample by the total number of counts to normalize, the counts would be greatly skewed by the DE gene, which takes up most of the counts for Sample A, but not Sample B.

- Most other genes for Sample A would be divided by the larger number of total counts and appear to be less expressed than those same genes in Sample B.



# Genomics110

## Normalization of count data

- **While normalization is essential for differential expression analyses, it is also necessary for QC, exploratory data analysis, visualization of data, and whenever you are exploring or comparing counts between or within samples.**

| *Normalization method* | *Description* | *Accounted factors* | *Recommendations for use* |
|---|---|---|---|
| *CPM (counts per million)* | *counts scaled by total number of reads* | *sequencing depth* | *gene count comparisons between replicates of the same samplegroup; NOT for within sample comparisons or DE analysis* |
| *TPM (transcripts per kilobase million)* | *counts per length of transcript (kb) per million reads mapped* | *sequencing depth and gene length* | *gene count comparisons within a sample or between samples of the same sample group; NOT for DE analysis* |
| *RPKM/FPKM (reads/ fragments per kilobase of exon per million reads/ fragments mapped)* | *similar to TPM* | *sequencing depth and gene length* | *gene count comparisons between genes within a sample; NOT for between sample comparisons or DE analysis* |
| *DESeq2's median of ratios [1]* | *counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene* | *sequencing depth and RNA composition* | *gene count comparisons between samples and for DE analysis; NOT for within sample comparisons* |
| *EdgeR's trimmed mean of M values (TMM) [2]* | *uses a weighted trimmed mean of the log expression ratios between samples* | *sequencing depth, RNA composition, and gene length* | *gene count comparisons between and within samples and for DE analysis* |

# Genomics110

## Normalization of count data

- **Raw result. Counts per gene or transcript.**

| *Gene* | *Control* | *Experimental* |
|---|---|---|
| *Gene A* | *27* | *27* |
| *Gene B* | *90* | *270* |
| *Gene C* | *280* | *640* |
| *Gene D* | *1003* | *3021* |
| *Gene E* | *3100* | *3342* |

**Wow! Gene E is expressed at a 3X higher rate than Gene D!**

**Would you reach the same conclusion if you knew that the control experiment used 10 million reads while the experimental condition used 30 million reads?**

**Oooh! The experimental condition caused genes B,C,D, and E to be over-expressed!**

# Genomics110

## Normalization of count data

- **Possible explanations:**

- **3 times as many transcripts are expressed for Gene E than Gene D.**

- **Gene E is 3X as long as Gene D.**

- **Gene D and E are the same length and produce the same number of transcripts, but Gene D has a close paralog that has acted as a "sponge" for ⅔ of its alignments.**

| *Gene* | *Control* | *Experimental Condition* |
|---|---|---|
| *Gene A* | *27* | *27* |
| *Gene B* | *90* | *270* |
| *Gene C* | *280* | *640* |
| *Gene D* | *1003* | *3021* |
| *Gene E* | *3100* | *3342* |

# Genomics110

## Normalization of count data

- **RPKM (or FPKM): reads per kilobase of exon model per million reads**

| *Gene* | *Gene Size* | *Control* | *Experimental* |
|---|---|---|---|
| *Gene A* | *1 kb* | *27* | *27* |
| *Gene B* | *2 kb* | *90* | *270* |
| *Gene C* | *6 kb* | *280* | *640* |
| *Gene D* | *30 kb* | *1003* | *3021* |
| *Gene E* | *40 kb* | *3100* | *3342* |

# Genomics110

## Normalization of count data

- **RPKM (or FPKM): reads per kilobase of exon model per million reads**
  - **Step 1: Normalize (i.e., adjust) gene counts by the total amount of sequences in the experiment**

| *Gene* | *Gene Size* | *Control* | *Experimental Condition* |
|---|---|---|---|
| *Gene A* | *1 kb* | *27* | *27* |
| *Gene B* | *2 kb* | *90* | *270* |
| *Gene C* | *6 kb* | *280* | *640* |
| *Gene D* | *30 kb* | *1003* | *3021* |
| *Gene E* | *40 kb* | *3100* | *3342* |
| *Total (w/ counts for ~19000 other genes)* | | *20,000,000* | *40,000,000* |

# Genomics110

## Normalization of count data

- **RPKM (or FPKM): reads per kilobase of exon model per million reads**
  - **Step 1: Normalize (i.e., adjust) gene counts by the total amount of sequences in the experiment**

| *Gene* | *Gene Size* | *Control* | *Experimental Condition* |
|---|---|---|---|
| *Gene A* | *1 kb* | *27* | *27* |
| *Gene B* | *2 kb* | *90* | *270* |
| *Gene C* | *6 kb* | *280* | *640* |
| *Gene D* | *30 kb* | *1003* | *3021* |
| *Gene E* | *40 kb* | *3100* | *3342* |
| *Total (w/ counts for ~19000 other genes)* | | *20,000,000* | *40,000,000* |
| *Millions of reads* | | *20* | *40* |

# Genomics110

## Normalization of count data

- **RPKM (or FPKM): reads per kilobase of exon model per million reads**
  - **Step 2: Normalize gene counts RPM by gene length**

| *Gene* | *Gene Size* | *Control (RPM)* | *Experimental (RPM)* |
|---|---|---|---|
| *Gene A* | *1 kb* | *1.34* | *0.675* |
| *Gene B* | *2 kb* | *4.5* | *6.75* |
| *Gene C* | *6 kb* | *14.0* | *16.0* |
| *Gene D* | *30 kb* | *50.15* | *75.525* |
| *Gene E* | *40 kb* | *155.0* | *83.55* |
| *Total (w/ counts for ~19000 other genes)* | | *20,000,000* | *40,000,000* |
| *Millions of reads* | | *20* | *40* |

# Genomics110

## Normalization of count data

- **RPKM (or FPKM): reads per kilobase of exon model per million reads**
  - **Step 2: Normalize gene counts RPM by gene length**

| *Gene* | *Gene Size* | *Control (RPM)* | *Experimental (RPM)* |
|---|---|---|---|
| *Gene A* | *1 kb* | *1.34* | *0.675* |
| *Gene B* | *2 kb* | *2.25* | *3.375* |
| *Gene C* | *6 kb* | *2.33* | *2.67* |
| *Gene D* | *30 kb* | *1.67* | *2.52* |
| *Gene E* | *40 kb* | *3.875* | *2.09* |
| *Total (w/ counts for ~19000 other genes)* | | *20,000,000* | *40,000,000* |
| *Millions of reads* | | *20* | *40* |

# Genomics110

## Normalization of count data

- **Which gene expressed the most transcripts in each condition?**

| *Gene* | *Gene Size* | *Control (RPM)* | *Experimental (RPM)* |
|---|---|---|---|
| *Gene A* | *1 kb* | *1.34* | *0.675* |
| *Gene B* | *2 kb* | *2.25* | *3.375* |
| *Gene C* | *6 kb* | *2.33* | *2.67* |
| *Gene D* | *30 kb* | *1.67* | *2.52* |
| *Gene E* | *40 kb* | *3.875* | *2.09* |
| *Total (w/ counts for ~19000 other genes)* | | *20,000,000* | *40,000,000* |
| *Millions of reads* | | *20* | *40* |

# Genomics110

## Poisson in differential (bulk) RNA expression

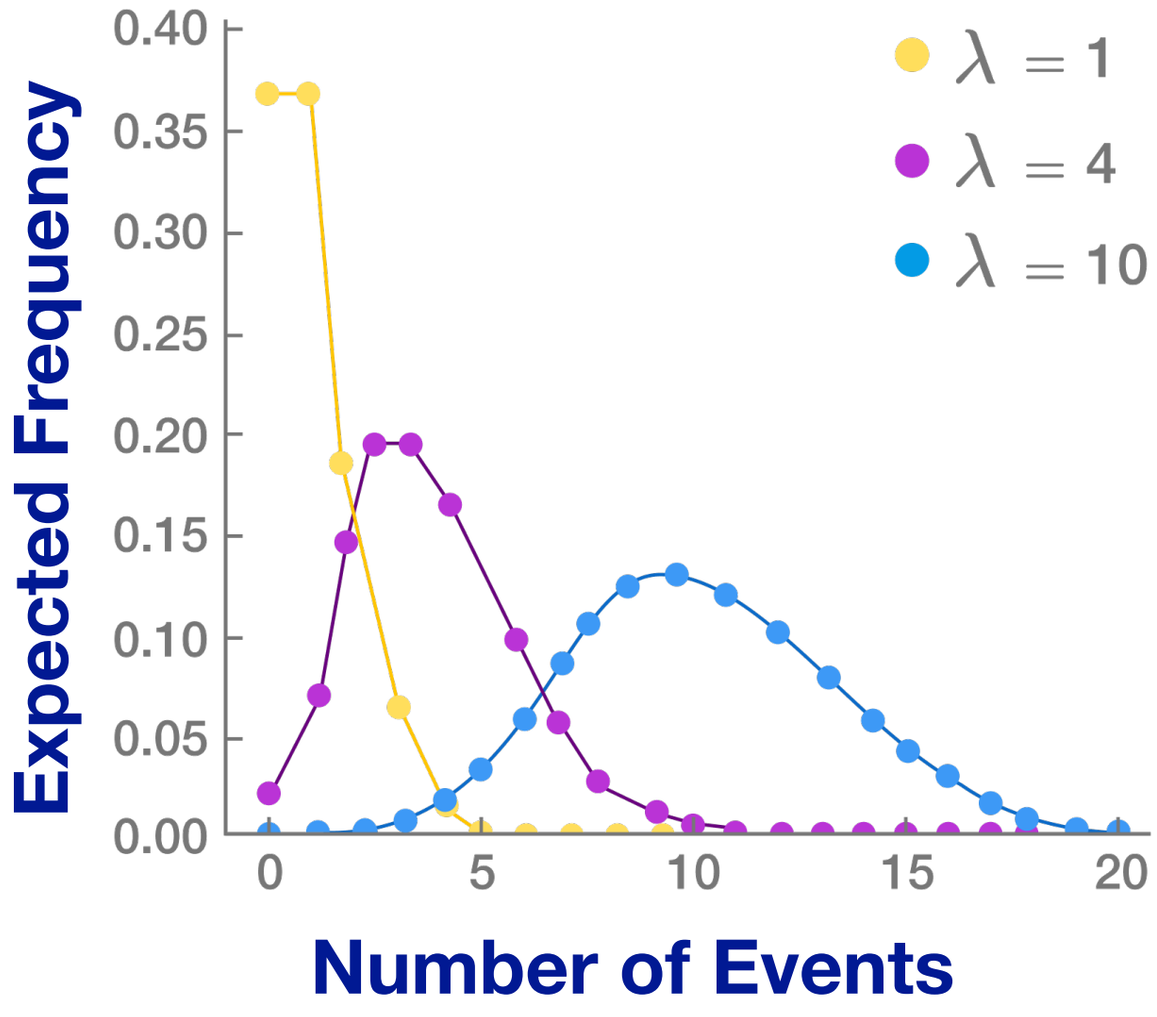
```
# Simulate 10,000 genes
N_genes = 10000

# Simulate a range of 10000 lambdas to reflect mean
# expression of each of the 10000 genes
lambdas = seq(1,1024,len=N_genes) #creates 10000 evenly-spaced numbers

# between 1 and 1024
# Simulate two technical replicates for an RNA-seq
# experiment. The observed expression for each of the
# 10,000 genes will be sampled from a Poisson distribution
# using the same lambda assigned to the gene for each
# replicate.
rep1 = rpois(N_genes, lambdas)
rep2 = rpois(N_genes, lambdas)

# remove genes where the expression was 0 in either rep.
non_zero = which(rep1>0 & rep2>0)
lambdas = lambdas[non_zero]
rep1 = rep1[non_zero]
rep2 = rep2[non_zero]

# make a data frame of lambdas and expression from replicates
rna_sim = data.frame(lambdas, rep1, rep2)
```

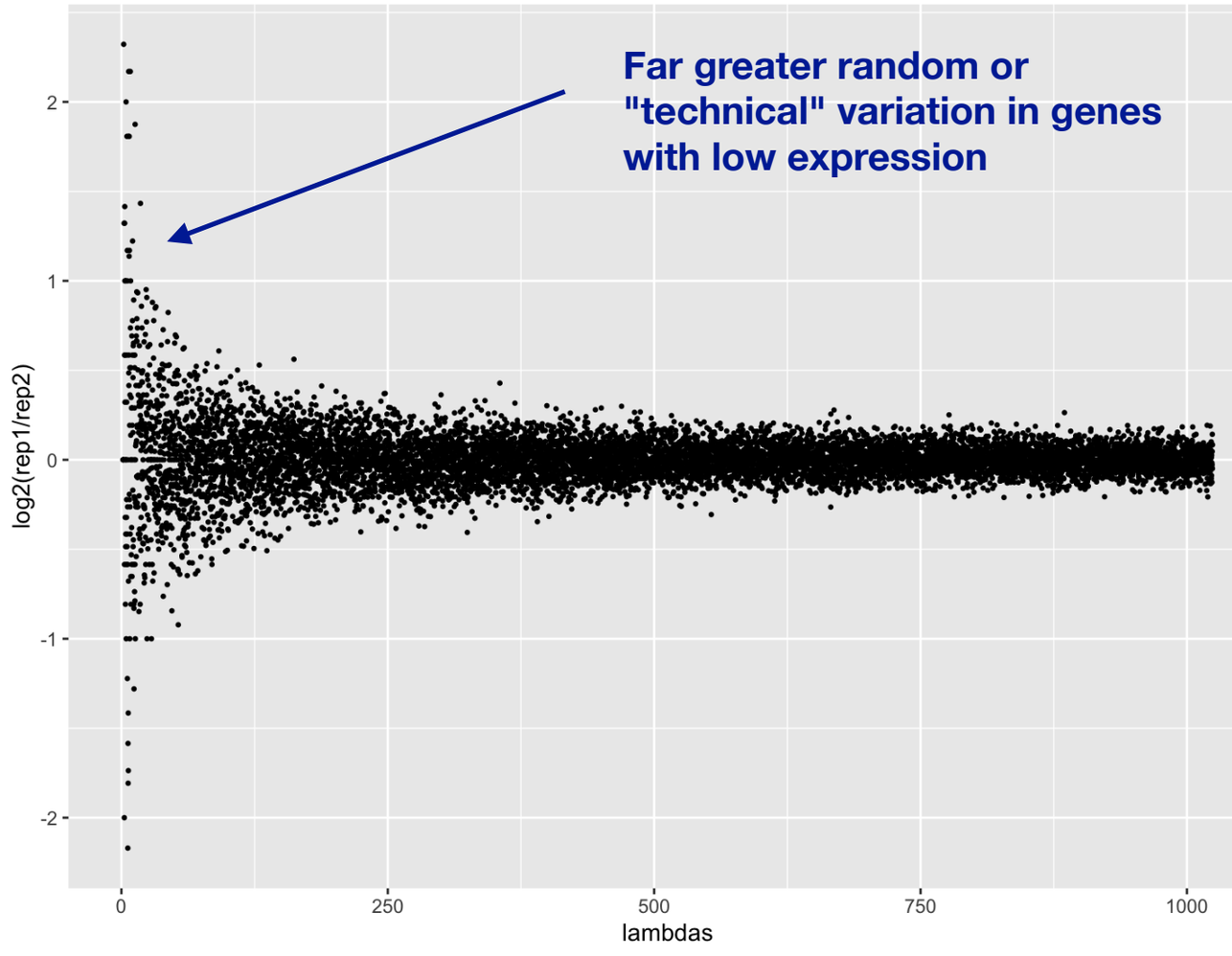
https://www.youtube.com/watch?v=KuxRYdG1EdQ

# Genomics110

## Poisson in differential (bulk) RNA expression

```
# plot the expression ratio from the two replicates as a function of # the mean expression
for the gene (lambda)
# plot in log2 space for better separation of the data
library(ggplot2)
ggplot(rna_sim, aes(x=lambdas, y=log2(rep1/rep2))) + geom_point(size=0.5)
```




https://www.youtube.com/watch?v=KuxRYdG1EdQ

# Genomics110

## Quality Control

- The next step in the differential expression workflow is QC, which includes sample-level and gene-level steps to perform QC checks on the count data to help us ensure that the samples/replicates look good and to help identify problematic expression trends and outliers.

- Normalized counts are utilized for this step.

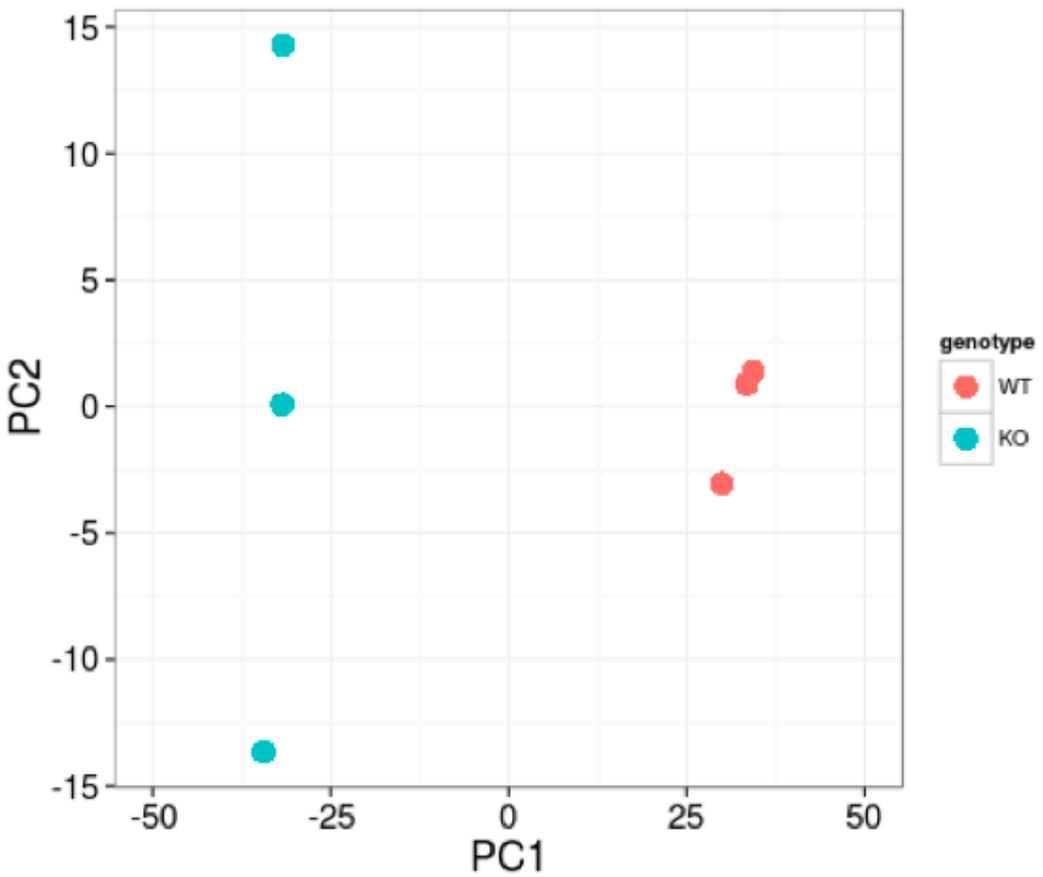
- Sample-level QC
  - A useful initial step in an RNA-seq analysis is often to assess overall similarity between samples:
    - Which samples are similar to each other, which are different?
    - Does this fit to the expectation from the experiment's design?
    - What are the major sources of variation in the dataset?

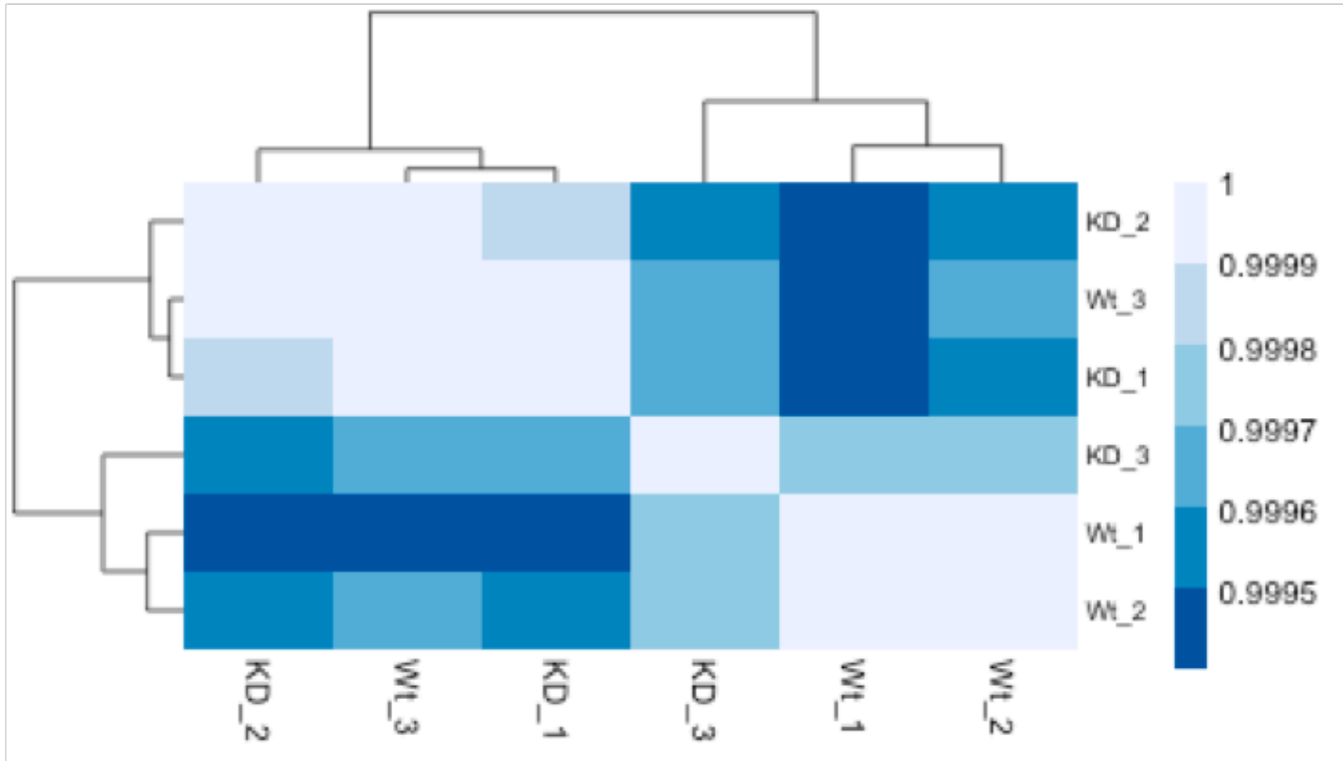
- Sample-level QC allows us to see how well our replicates cluster together, as well as, observe whether our experimental condition represents the major source of variation in the data. Performing sample-level QC can also identify any sample outliers, which may need to be explored to determine whether they need to be removed prior to DE analysis.

# Genomics110

## Quality Control

- **Sample-level QC**
  - **The 2 main methods utilized for this type of QC are Principal Component Analysis (PCA) and Hierarchical Clustering.**



# Genomics110

## Quality Control

- Gene-level QC
  - In addition to examining how well the samples/replicates cluster together, there are a few more QC steps.

  - Prior to differential expression analysis it is beneficial to omit genes that have little or no chance of being detected as differentially expressed.

  - This will increase the power to detect differentially expressed genes.

- The genes omitted fall into three categories:

  - Genes with zero counts in all samples

  - Genes with an extreme count outlier

  - Genes with a low mean normalized counts

# Genomics110

## Quality Control

- **Gene-level QC**
  - **Some statistical tools, e.g. DESeq2, used for identifying differentially expressed genes will perform this filtering by default; however other tools, e.g. EdgeR, will not.**

| | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 |
|---|---|---|---|---|---|
| ENSG00000000003 | 67 | 44 | 87 | 40 | 1138 |
| ENSG00000000005 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000000419 | 467 | 515 | 621 | 365 | 587 |
| ENSG00000000457 | 260 | 211 | 263 | 164 | 245 |
| ENSG00000000460 | 2 | 5 | 1 | 0 | 1 |

Genes with extreme count outlier

Genes with zero counts

Genes with low mean normalized counts ('Independent filtering')

# Genomics110

## Count modeling and statistical analysis

- **The final step in the differential expression analysis workflow is fitting the counts to a model and performing the statistical test for differentially expressed genes. In this step we essentially want to determine whether the mean expression levels of different sample groups are significantly different.**

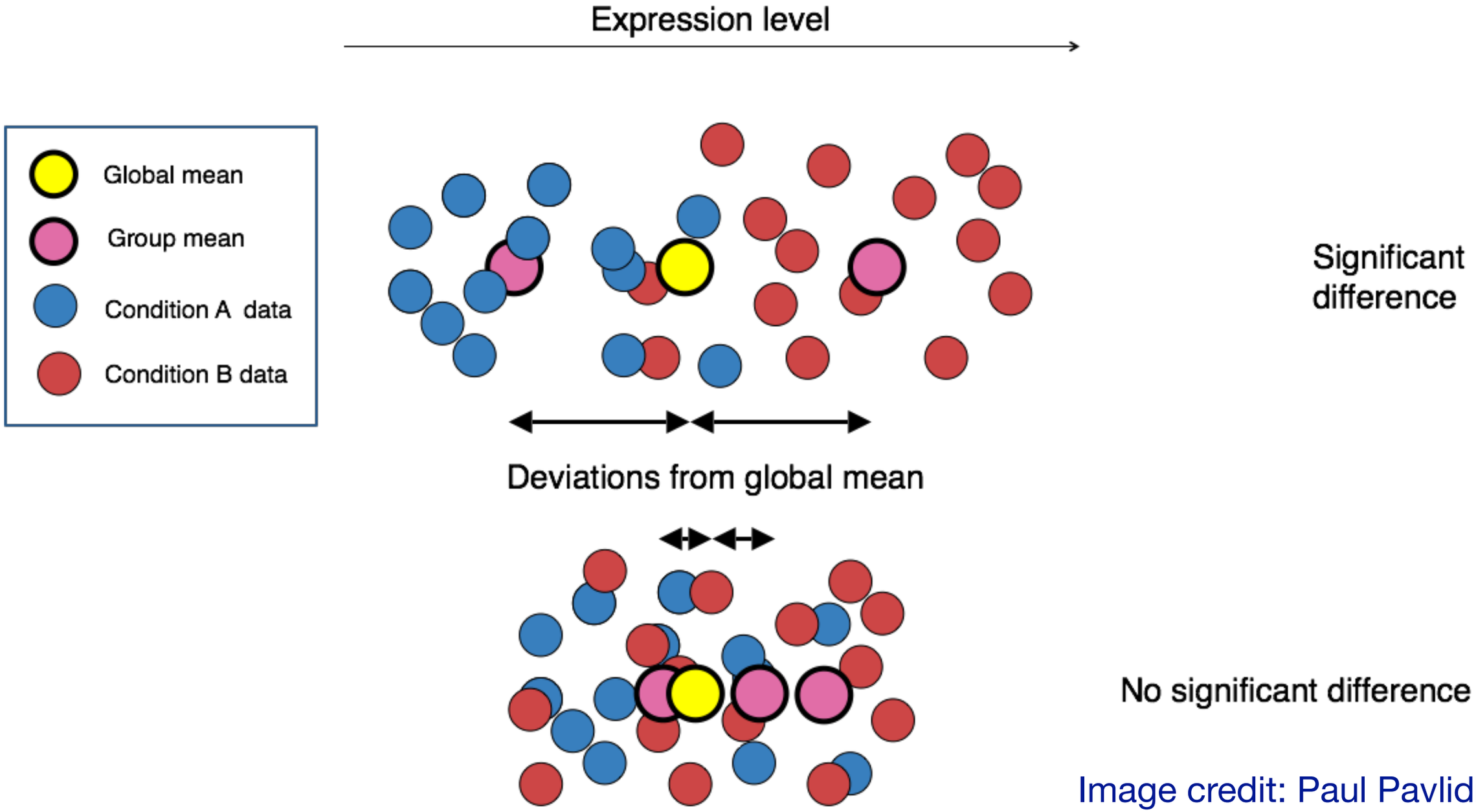



Image credit: Paul Pavlidis, UBC

# Genomics110

## Some highlights of RNA-seq count data

- There are a low number of counts associated with a large proportion of genes
- There is no upper limit for expression (large dynamic range)
- The negative binomial model has been determined to be the best fit for the count distribution for RNA-seq data where there is a lot of variance between the replicates and mean < variance

# Genomics110

## Tools for statistical analysis

- There are a number of software packages that have been developed for differential expression analysis of RNA-seq data.
- A few tools are generally recommended as best practice, e.g. DESeq2 and EdgeR.
- Both these R packages use the negative binomial model, employ similar methods, and typically, yield similar results.
- They are pretty stringent, and have a good balance between sensitivity and specificity (reducing both false positives and false negatives).
- Limma-Voom is another set of tools often used together for DE analysis, but this method may be less sensitive for small sample sizes. This method is recommended when the number of biological replicates per group grows large (> 20).
- Further reading about DGE tool comparisons.

# Genomics110

## Multiple test correction

- The output of any of these analysis methods is a p-value as well as a value assigning statistical significance after multiple test correction, and the second value is what should be used when creating lists of genes that are differentially expressed.

- Each p-value returned is the result of a single test (single gene). If we used the p-value directly with a significance cut-off of $p < 0.05$, that means there is a 5% chance it is a false positive and the more genes we test, the more we inflate the false positive rate.

    - For example, if we test 20,000 genes for differential expression, at $p < 0.05$ we would expect to find 1,000 genes by chance.

- If we found 3000 genes to be differentially expressed total, roughly one third of our genes are false positives.

- We would not want to sift through our “significant” genes to identify which ones are true positives.

# Genomics110

## Multiple test correction

- A few common methods to correct for multiple testing are listed below:
  - Bonferroni: The adjusted p-value is calculated by: p-value * m (m = total number of tests).
    - This is a very conservative approach with a high probability of false negatives, so is generally not recommended.
  - FDR/Benjamini-Hochberg: Benjamini and Hochberg (1995) defined the concept of FDR and created an algorithm to control the expected FDR below a specified level given a list of independent p-values.
    - An interpretation of the BH method for controlling the FDR is implemented in DESeq2 in which we rank the genes by p-value, then multiply each ranked p-value by m/rank.
  - Q-value / Storey method: The minimum FDR that can be attained when calling that feature significant.
    - For example, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

# Genomics110

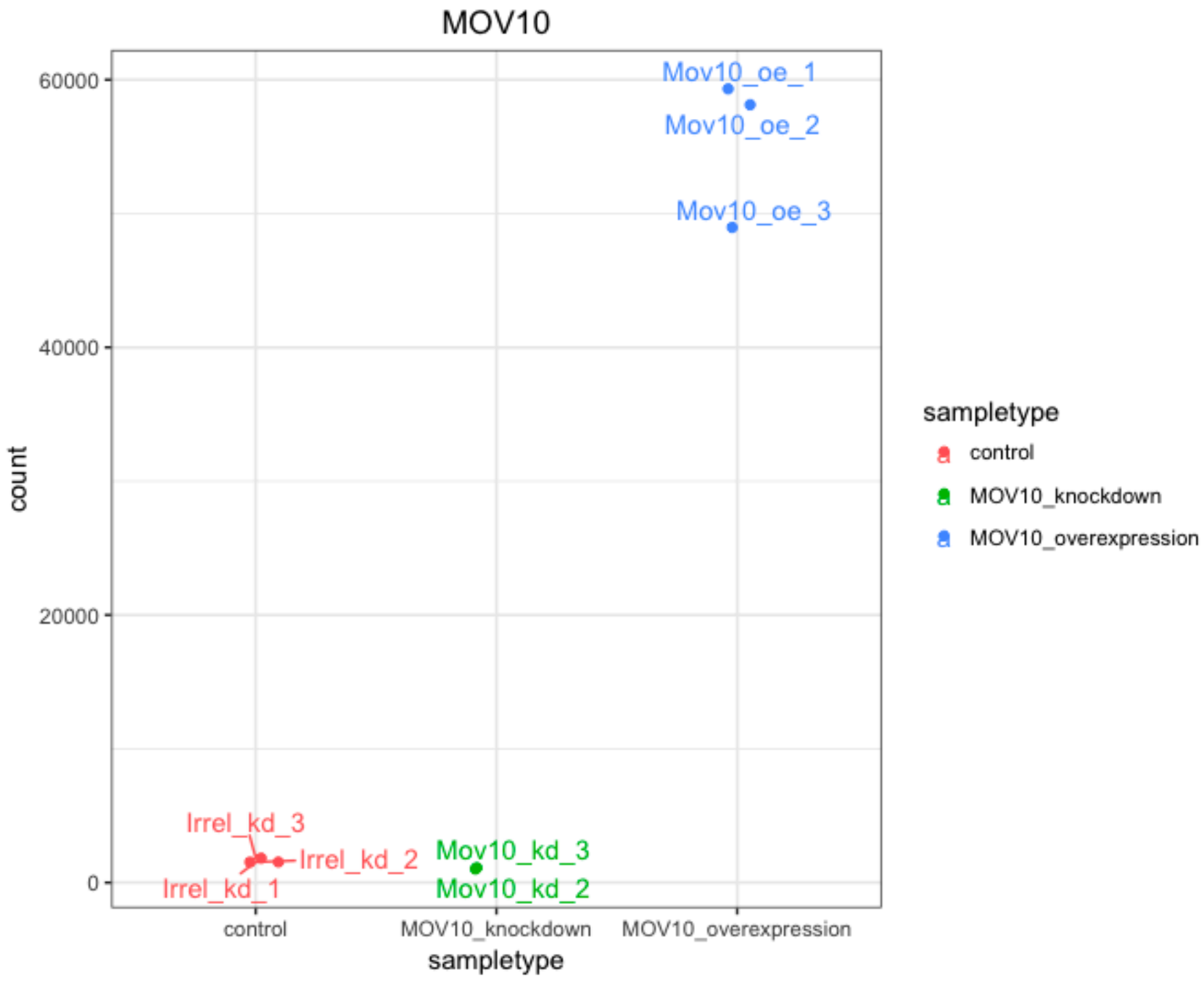
## Visualizing the Results of a DGE Experiment

- Plotting signicantly differentially expressed genes
  - One way to visualize results would be to simply plot the expression data for a handful of genes across the various sample groups.
  - This can be implemented in R (usually) for multiple genes of interest or a single gene using functions associated with
    - The package used to perform the statistical analysis (e.g. DESeq2's plotCounts() function) or
    - an external package created for this purpose (e.g. pheatmap, DEGreport) or
    - using the ggplot2 package.

# Genomics110

## Visualizing the Results of a DGE Experiment

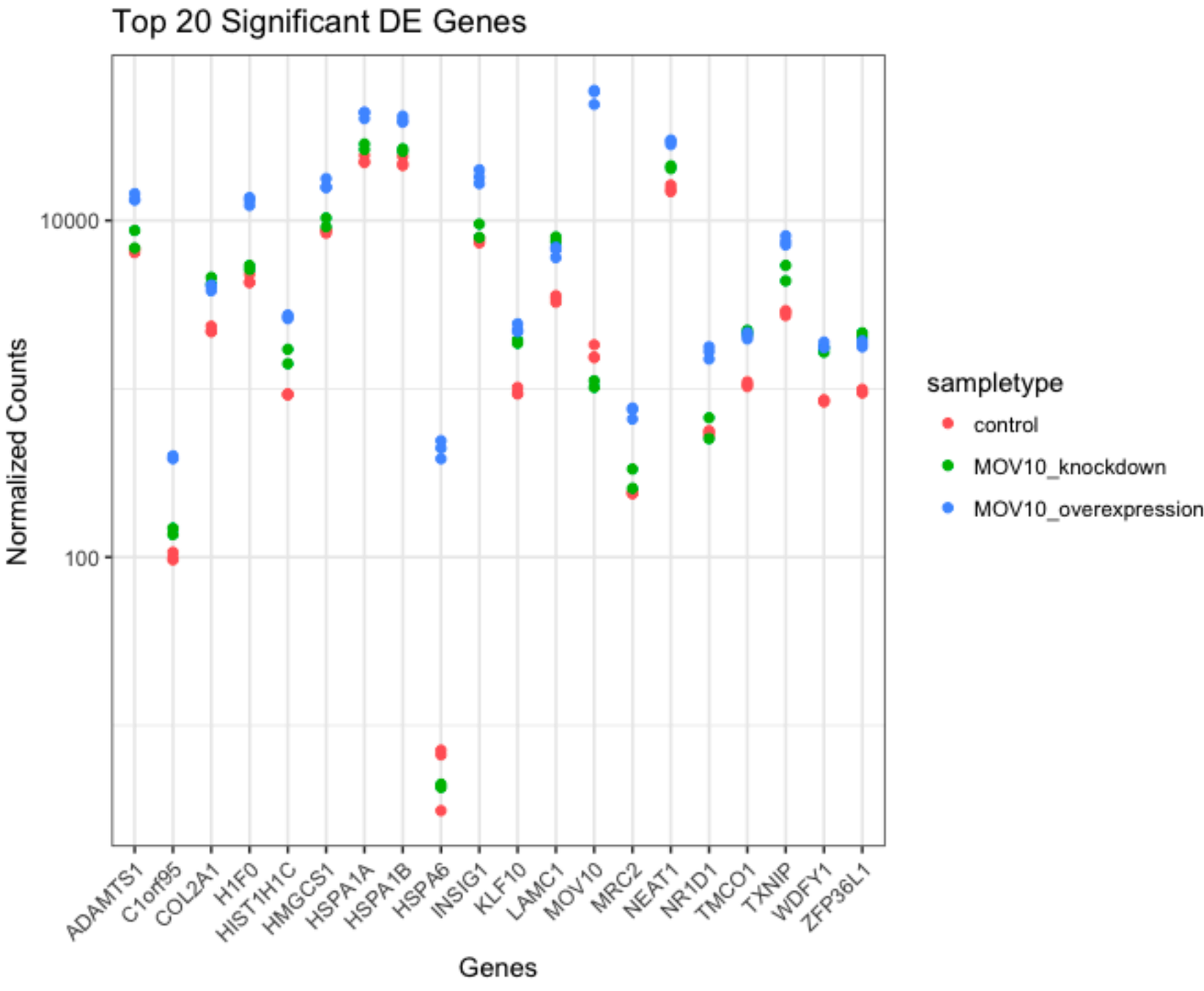
- **Plotting expression of a single gene across sample groups:**



# Genomics110

## Visualizing the Results of a DGE Experiment

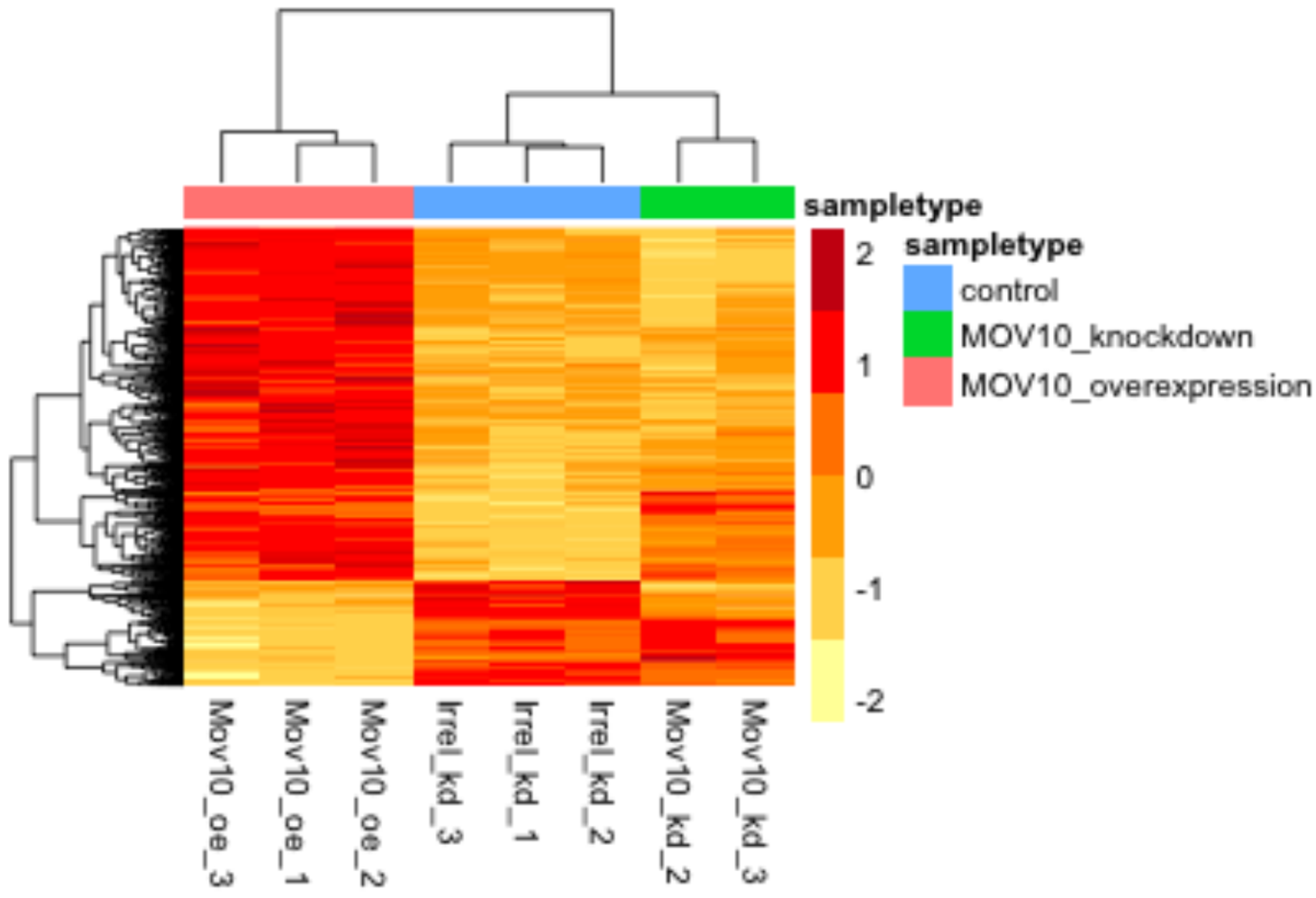
- **Plotting expression of multiple genes across sample groups :**
  - **One way to visualize results would be to simply plot the expression data for a handful of genes across the various sample groups.**
  - **The plot displays the top 20 significantly differentially expressed genes. Please note that the normalized counts on the Y axis are logged (log10) to ensure that the any large differences in expression are plotted without compromising the quality of the visualization.**



# Genomics110

## Visualizing the Results of a DGE Experiment

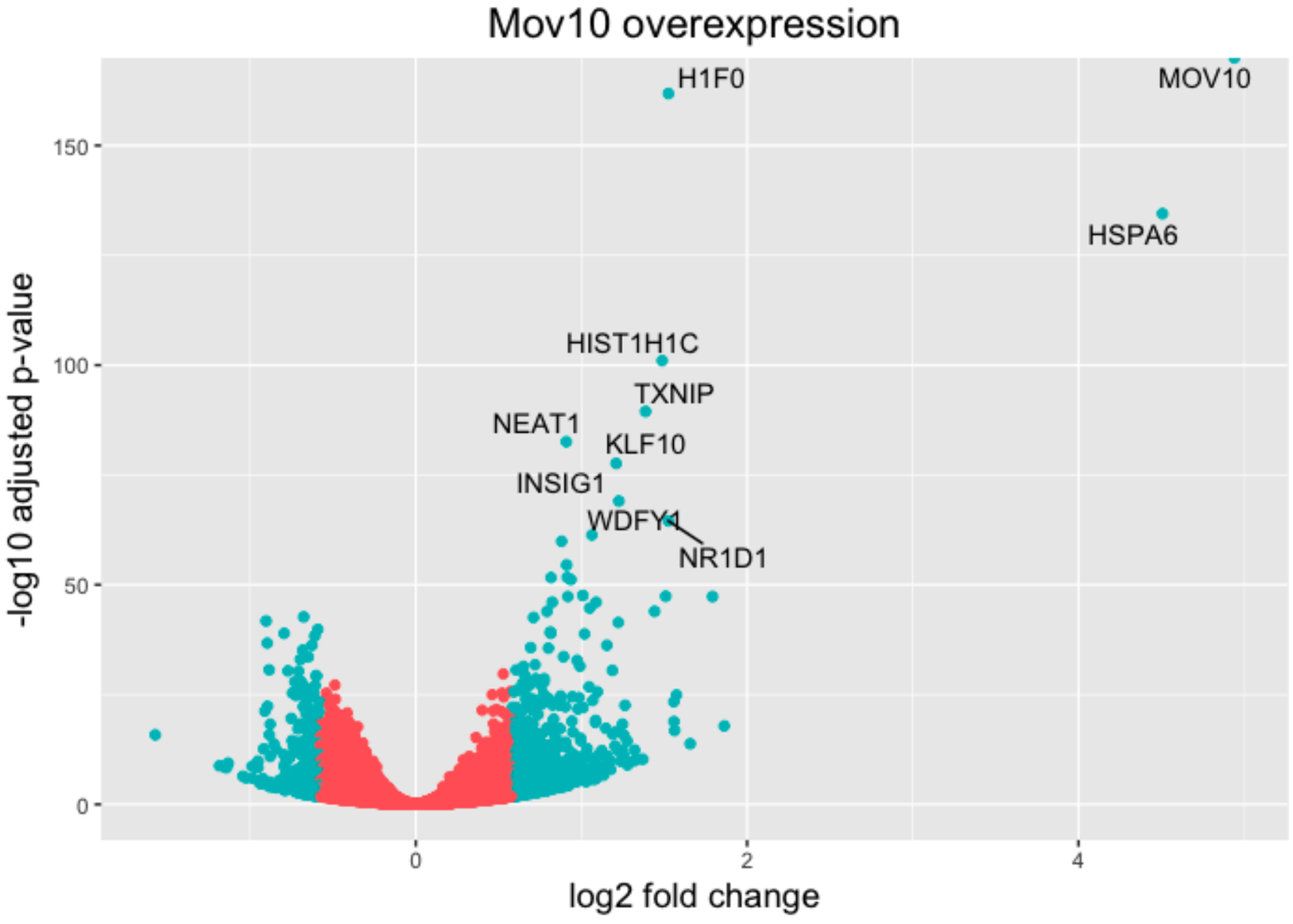
- **Heatmaps**
  - **In addition to plotting subsets, we could also extract the normalized values of all the significant genes and plot a heatmap of their expression using pheatmap().**
  - **In this heatmap Z-scores are calculated for each row (each gene) and these are plotted instead of the normalized expression values; this ensures that the expression patterns/trends that we want to visualize are not overwhelmed by the expression values.**
  - **Z-scores are computed on a gene-by-gene basis by subtracting the mean and then dividing by the standard deviation. The Z-scores are computed after the clustering, so that it only affects the graphical aesthetics and the color visualization is improved.**



# Genomics110

## Visualizing the Results of a DGE Experiment

- **Volcano plots**
  - **The plot would be great to look at the expression levels of a good number of genes, but for more of a global view there are other plots. A commonly used one is a volcano plot; in which you have the log transformed adjusted p-values are plotted on the y-axis and log2 fold change values on the x-axis.**

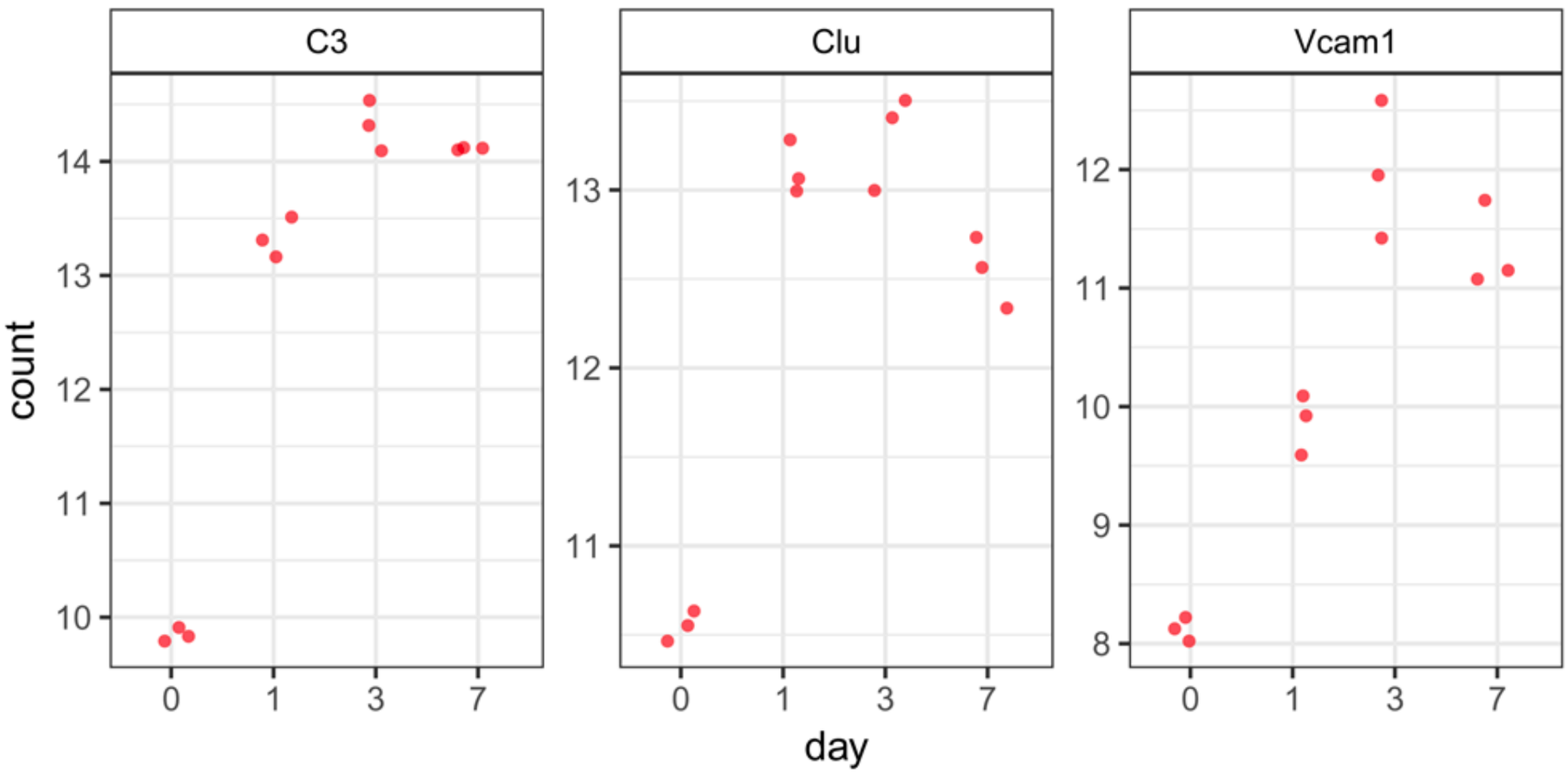


# Genomics110

## Visualizing the Results of a DGE Experiment

- **DEGreport**
  - **If you do use the DESeq2 package for differential expression analysis, the package 'DEGreport' has a lot of great functions to draw a lot of the above plots in addition to several others. Some examples are available in this vignette, and some of them are shown next.**

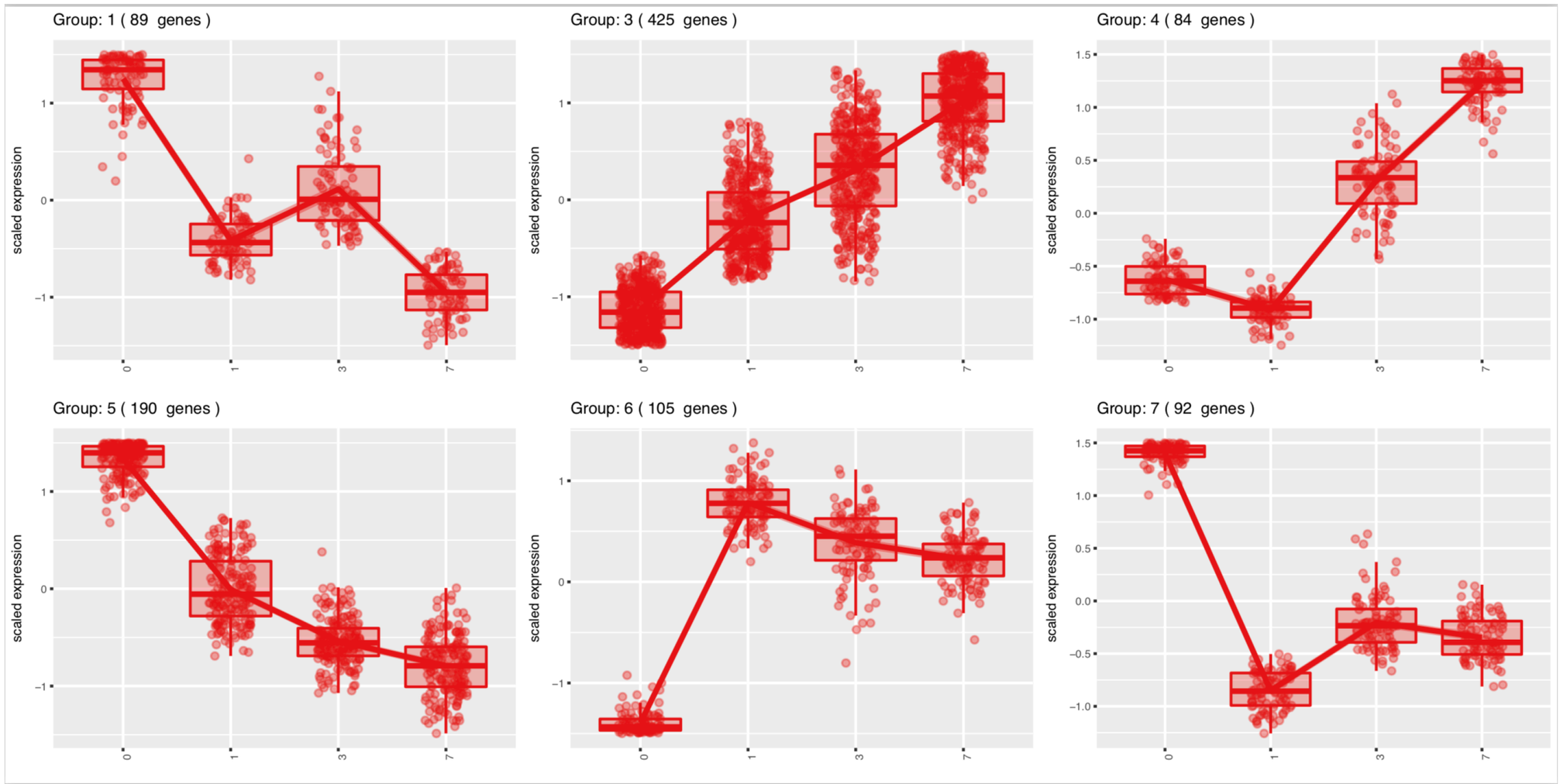
  - **Plot 1: An easy and clean way to visualize expression of genes of interest.**



# Genomics110

## Visualizing the Results of a DGE Experiment

- **DEGreport**
  - **Plot 2: When performing DE analysis on several groups, e.g. a time course experiment, grouping together genes that have similar patterns of expression and visualizing these patterns can be very helpful. The degPatterns() function in the DEGreport package performs the analysis and creates a display with this information**
  - **In addition to displaying the patterns, degPatterns() outputs a list to enable the user to extract the genes in each grouping.**




Images adapted from doi: 10.12688/f1000research.12093.2.

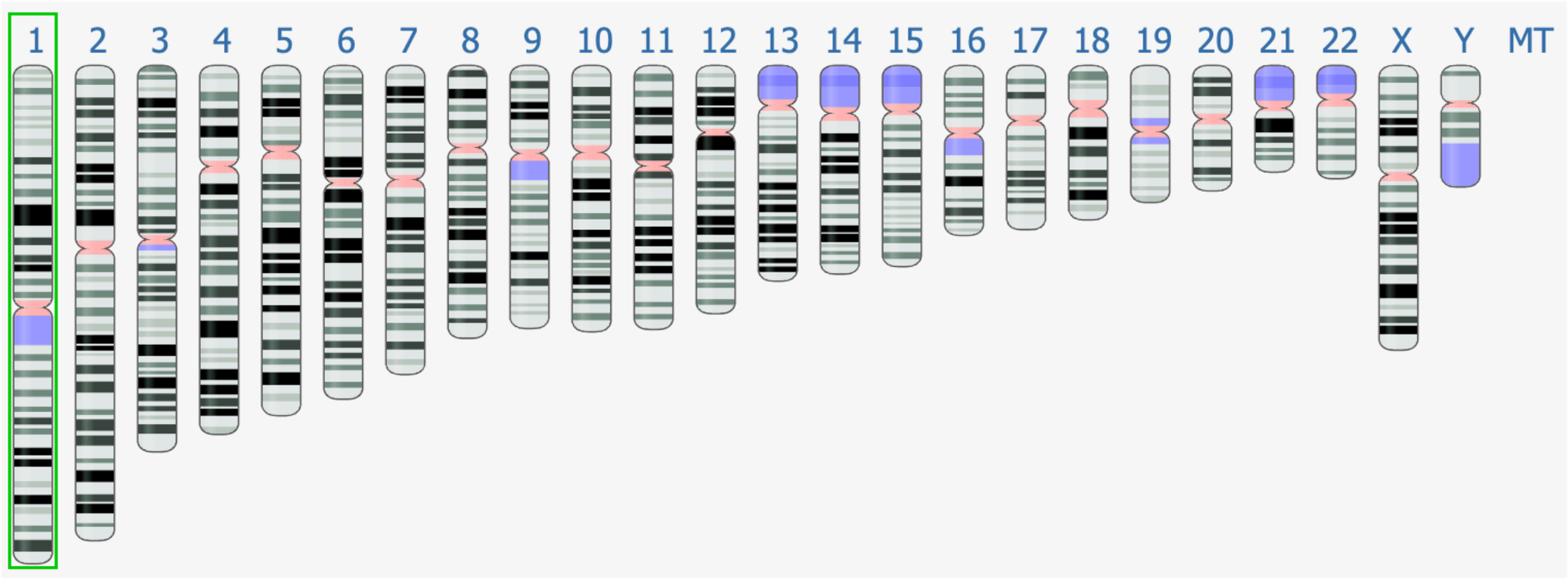
# Genomics110

## References

- ***This lesson has been developed using materials from various sources, that include, but are not restricted to training tutorials developed by the Galaxy Project team. These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.***

- ***Other Bibliographic References are:***


  - ***RSeQC: quality control of RNA-seq experiments***
  - ***Comprehensive comparative analysis of strand-specific RNA sequencing methods***
  - ***Transcript assembly and quantification by RNA-seq..***
  - ***Identification of novel transcripts in annotated genomes using…***
  - ***Differential analysis of gene regulation at transcript…***
  - ***TopHat: discovering splice junctions with RNA-Seq***
  - ***TopHat-Fusion: an algorithm for discovery…***
  - ***Differential gene and transcript expression analysis…***
  - ***TopHat2: accurate alignment of transcriptomes…***
  - ***HISAT: a fast spliced aligner with low memory…***
  - ***Stringtie enables improved reconstruction…***
  - ***Count-based differential expression analysis of RNA…***
  - ***Transcript-level expression analysis of RNA-seq…***
  - ***STAR: ultrafast universal RNA-seq aligner…***
  - ***Near-optimal probabilistic RNA-Seq quantification…***
  - ***Differential analysis of RNA-Seq…***
  - ***Salmon provides fast…***
  - ***GMAP: a genomic mapping and alignment…***
  - ***Spatially resolved transcriptomics…***
  - ***Simulation-Based omprehensive benchmarking…***
  - ***A survey of best practices for RNA-seq data Analysis***




# Genomics110

## BIOL647
## Digital Biology

Rodolfo Aramayo