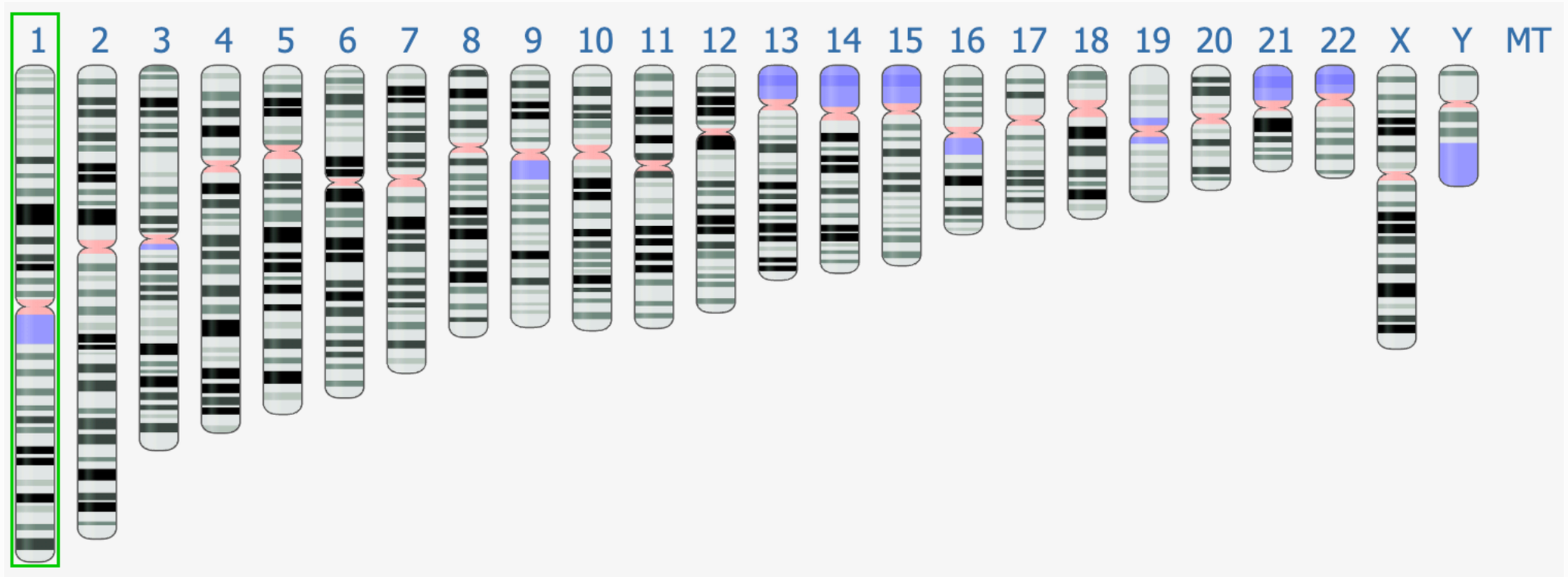


Genomics108

BIOL647
Digital Biology

Rodolfo Aramayo

CC BY NC SA

Genomics108

General Outline

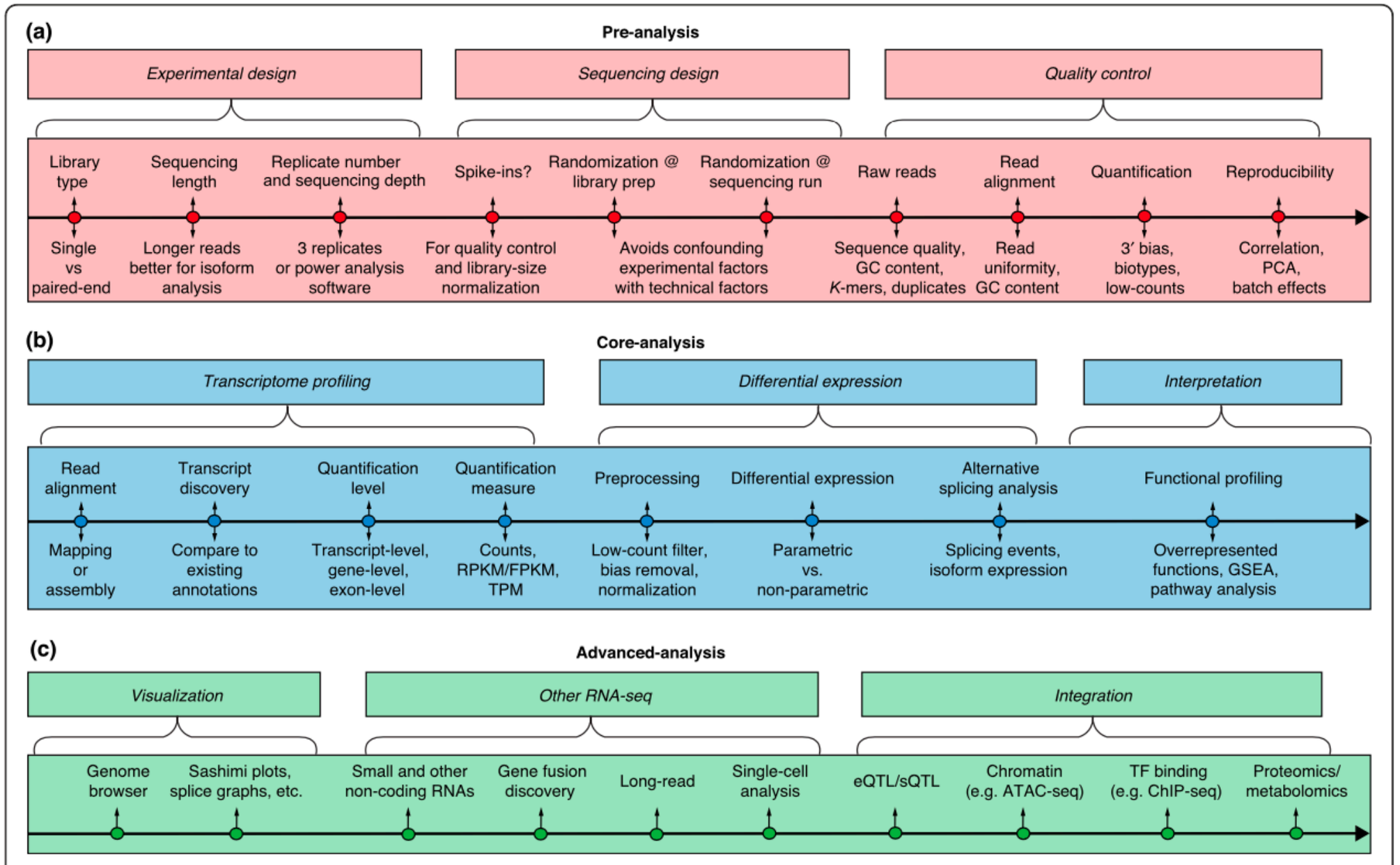
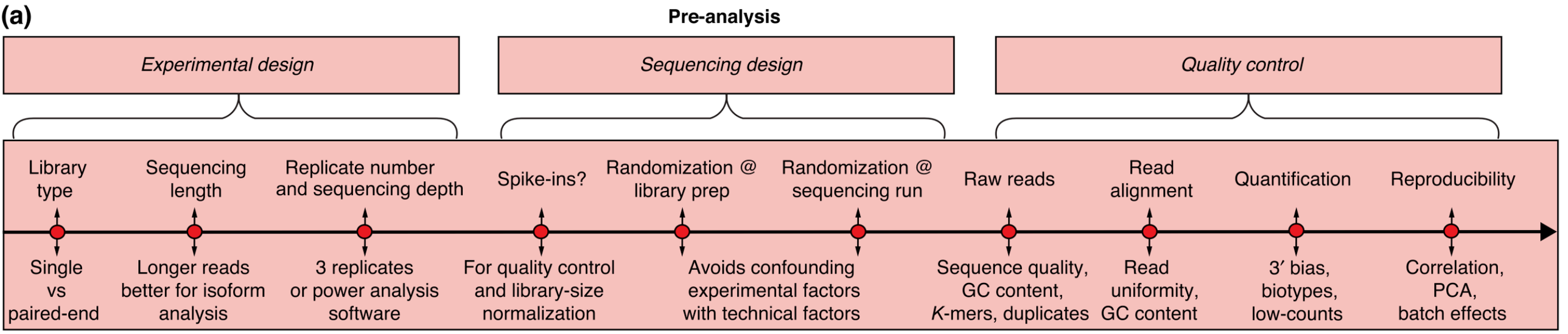


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

https://pubmed.ncbi.nlm.nih.gov/26813401/

Genomics108

General Outline



https://pubmed.ncbi.nlm.nih.gov/26813401/

Genomics108

Introduction To Experimental Design

- Keep it simple

- Classical experimental design example:
  - Developmental time course experiment (Time series)
  - Without missing values, where possible
  - Intended analysis must be feasible – can the available samples and hypothesis of interest be combined to formulate a testable statistical hypothesis?

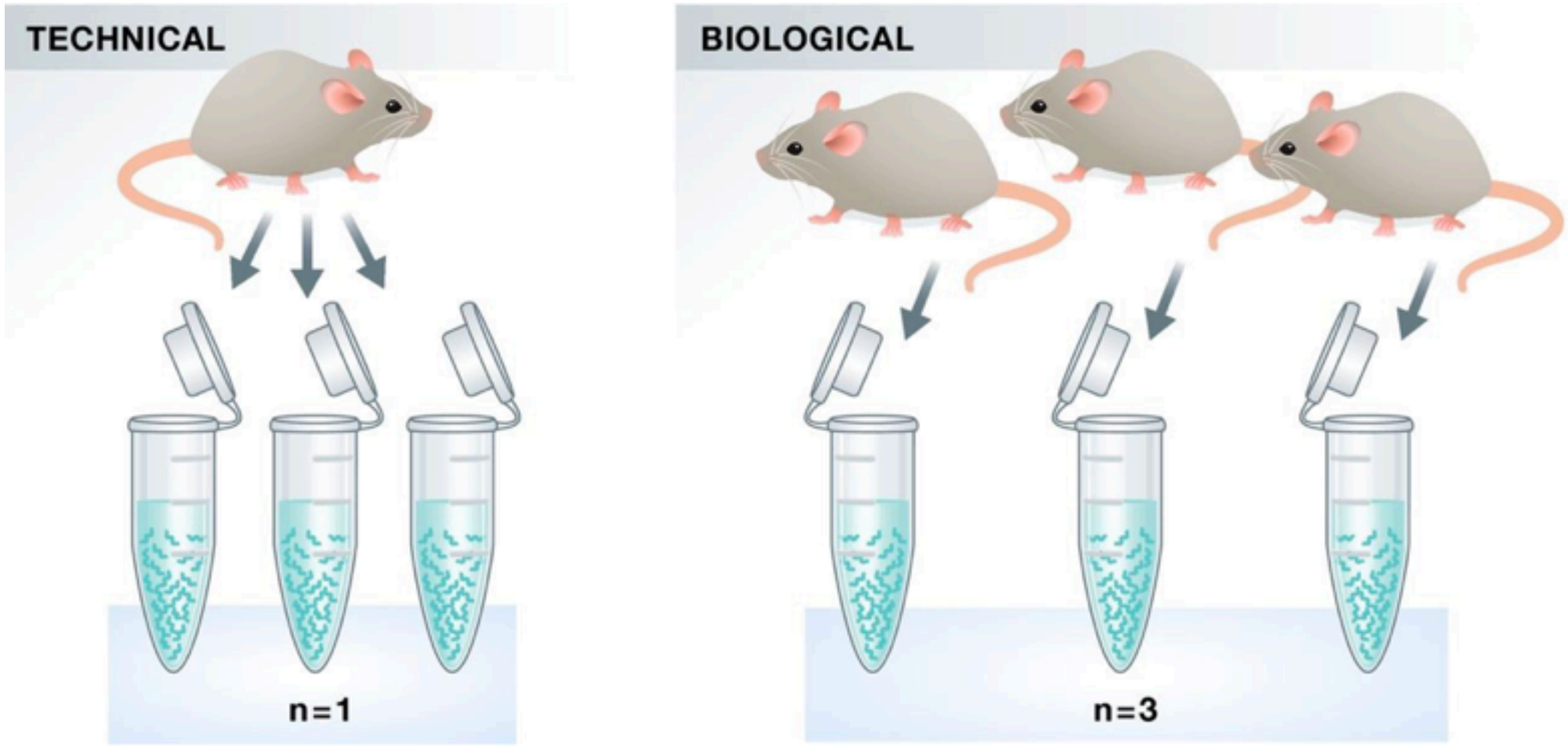
- Include Replicates
  - Extent of replication determines nuance of biological question
  - No replication (1 sample per treatment) really limits what you can say and therefore gives you limited statistical options
  - Include at least 3 if not more (6 is best) replicates per treatment
  - Would you be able to detect 2-fold change in average expression between groups?
  - Number of replicates varies by experiment:
  - 10-50 replicates per treatment for population studies, e.g., cancer cell lines
  - 1000's of replicates for prospective studies, e.g., SNP discovery

http://genomebiology.com/2009/10/3/R25

Genomics108

Introduction To Experimental Design

- **Experimental replicates can be performed as technical replicates or biological replicates**



- **Technical replicates: use the same biological sample to repeat the technical or experimental steps in order to accurately measure technical variation and remove it during analysis.**

- **Biological replicates use different biological samples of the same condition to measure the biological variation between samples.**

https://www.embopress.org/doi/full/10.15252/embj.201592958

Genomics108

Introduction To Experimental Design

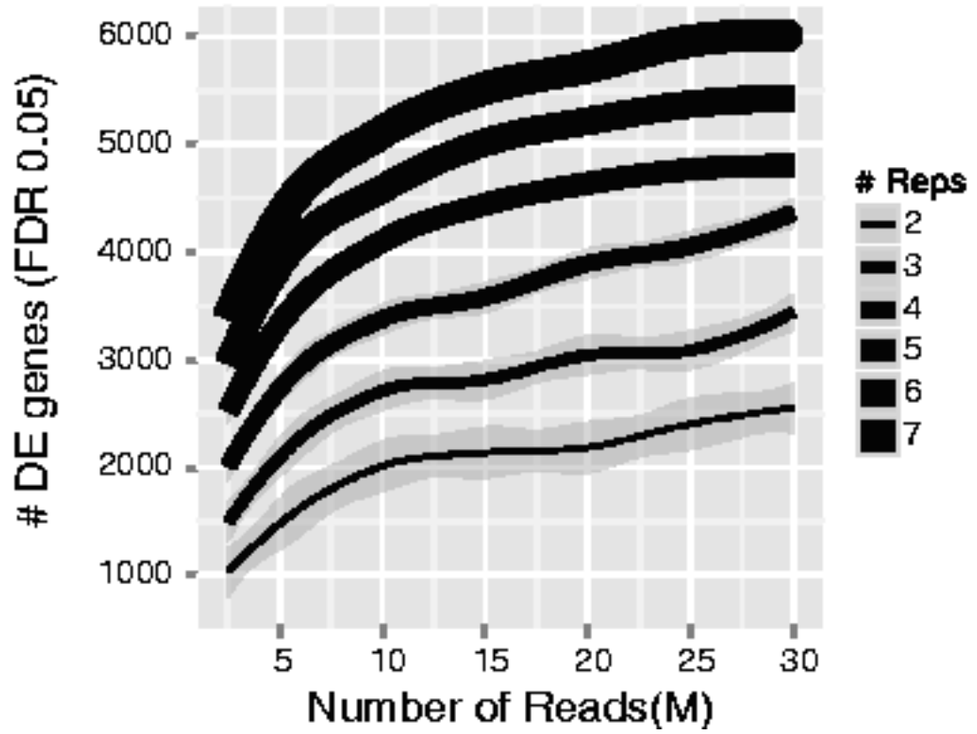
- Experimental replicates can be performed as technical replicates or biological replicates
- In the days of microarrays, technical replicates were considered a necessity; however, with the current RNA-Seq technologies, technical variation is much lower than biological variation and technical replicates are unnecessary.
- In contrast, biological replicates are absolutely essential.
    - For differential expression analysis, the more biological replicates, the better the estimates of biological variation and the more precise our estimates of the mean expression levels.
    - This leads to more accurate modeling of our data and identification of more differentially expressed genes.

https://www.embopress.org/doi/full/10.15252/embj.201592958

Genomics108

Introduction To Experimental Design

- **Experimental replicates**



- **As the figure above illustrates, biological replicates are of greater importance than sequencing depth. The figure shows the relationship between sequencing depth and number of replicates on the number of differentially expressed genes identified.**

- **Note that an increase in the number of replicates tends to return more DE genes than increasing the sequencing depth. Therefore, generally more replicates are better than higher sequencing depth, with the caveat that higher depth is required for detection of lowly expressed DE genes and for performing isoform-level differential expression.**

https://doi.org/10.1093/bioinformatics/btt688

Genomics108

Introduction To Experimental Design

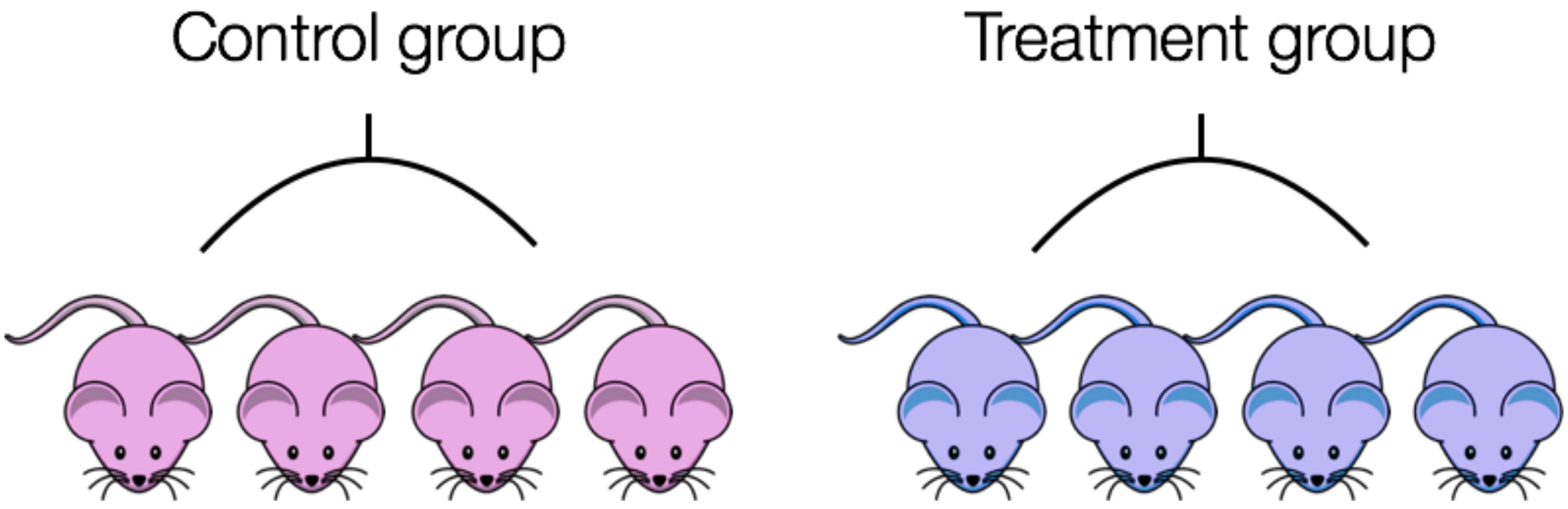
- **Replicates are almost always preferred to greater sequencing depth for bulk RNA-Seq. However, guidelines depend on the experiment performed and the desired analysis. Below we list some general guidelines for replicates and sequencing depth to help with experimental planning:**
  - **General gene-level differential expression:**
    - **ENCODE guidelines suggest 30 million SE reads per sample (stranded).**
    - **15 million reads per sample is often sufficient, if there are at least 4 replicates.**
    - **More replicates >> More sequencing depth**
  - **Gene-level differential expression with detection of lowly-expressed genes:**
    - **Sequence deeper with at least 30-60 million reads depending on level of expression (start with 30 million with a good number of replicates).**
    - **Similarly benefits from replicates more than sequencing depth.**
  - **Splice-isoform differential expression:**
    - **For known isoforms, suggested to have a depth of at least 30 million reads per sample and paired-end reads.**
    - **For novel isoforms should have more depth (> 60 million reads per sample).**
    - **Choose biological replicates over paired/deeper sequencing.**
  - **Other types of RNA analyses (intron retention, small RNA-Seq, etc.):**
    - **Different recommendations depending on the analysis.**
    - **Almost always more biological replicates are better!**

https://doi.org/10.1093/bioinformatics/btt688

Genomics108

Introduction To Experimental Design

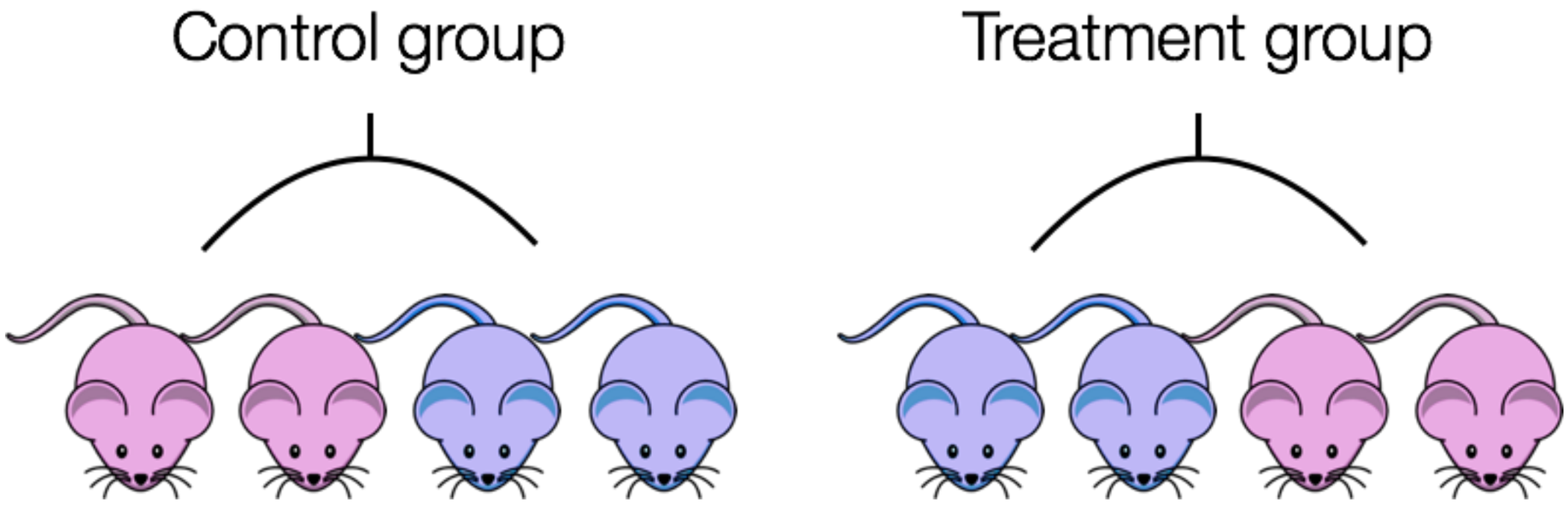
- **Confounding Effects**
  - **A confounded RNA-Seq experiment is one where you cannot distinguish the separate effects of two different sources of variation in the data.**
  - **For example, we know that sex has large effects on gene expression, and if all of our control mice were female and all of the treatment mice were male, then our treatment effect would be confounded by sex. We could not differentiate the effect of treatment from the effect of sex.**



Genomics108

Introduction To Experimental Design

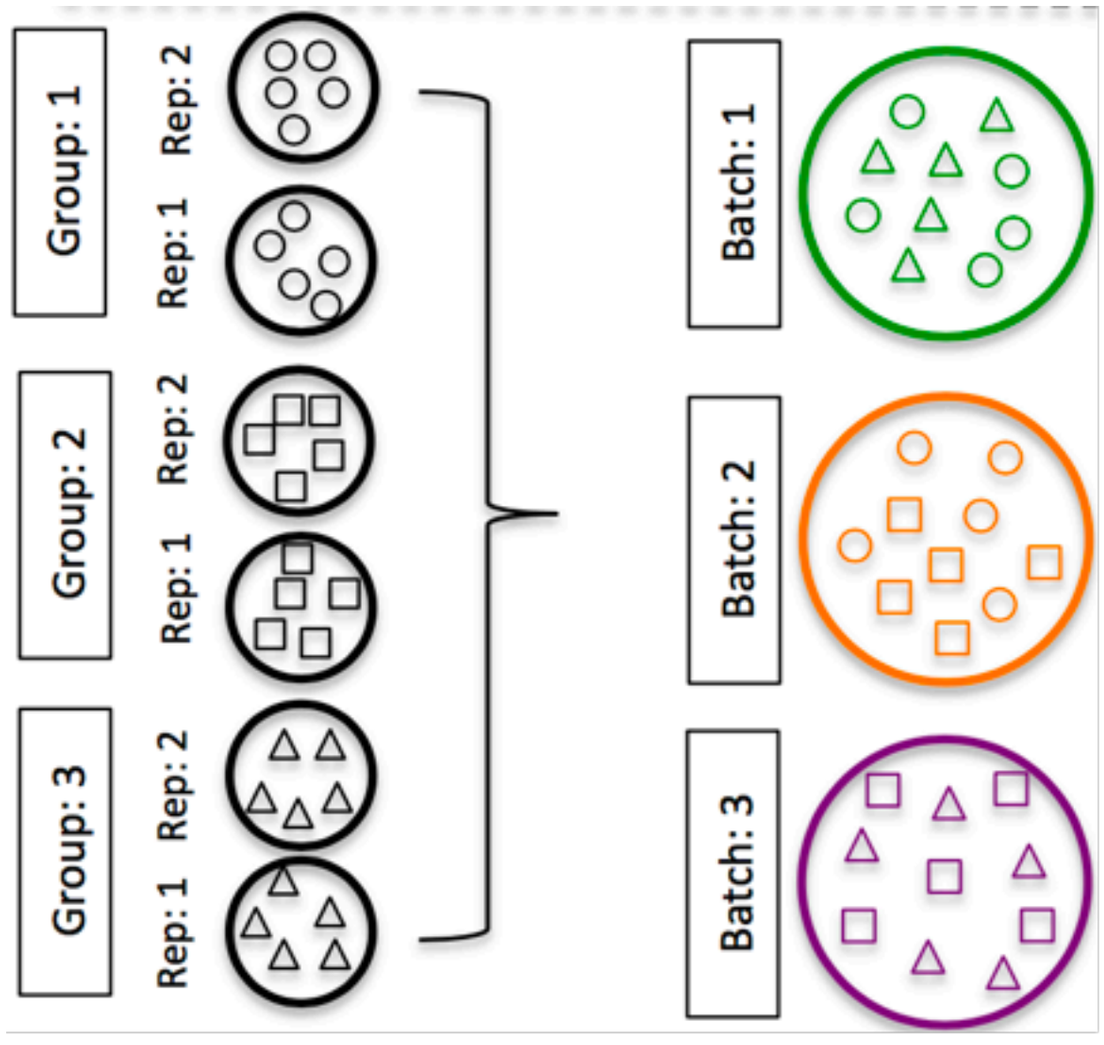
- **To AVOID confounding:**
  - **Ensure animals in each condition are all the same sex, age, litter, and batch, if possible.**
  - **If not possible, then ensure to split the animals equally between conditions**

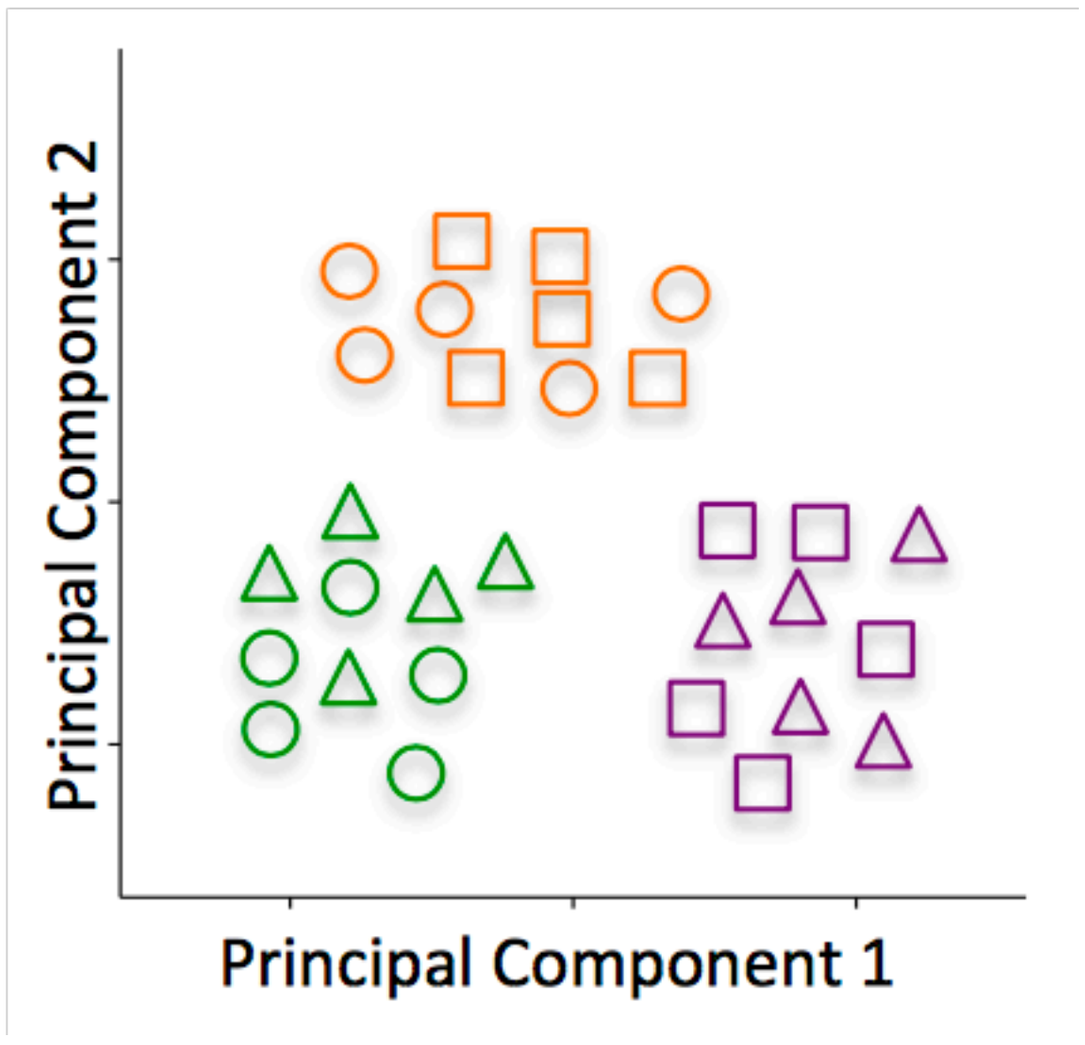


Genomics108

Introduction To Experimental Design

- **Batch effects**
  - **Batch effects are a significant issue for RNA-seq analyses, since you can see significant differences in expression due solely to batch.**





https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

- How to know whether you have batches?
  - Were all RNA isolations performed on the same day?
  - Were all library preparations performed on the same day?
  - Did the same person perform the RNA isolation/library preparation for all samples?
  - Did you use the same reagents/kits for all samples?
  - Did you perform the RNA isolation/library preparation in the same location?
- If any of the answers is No, then you have batches.

https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

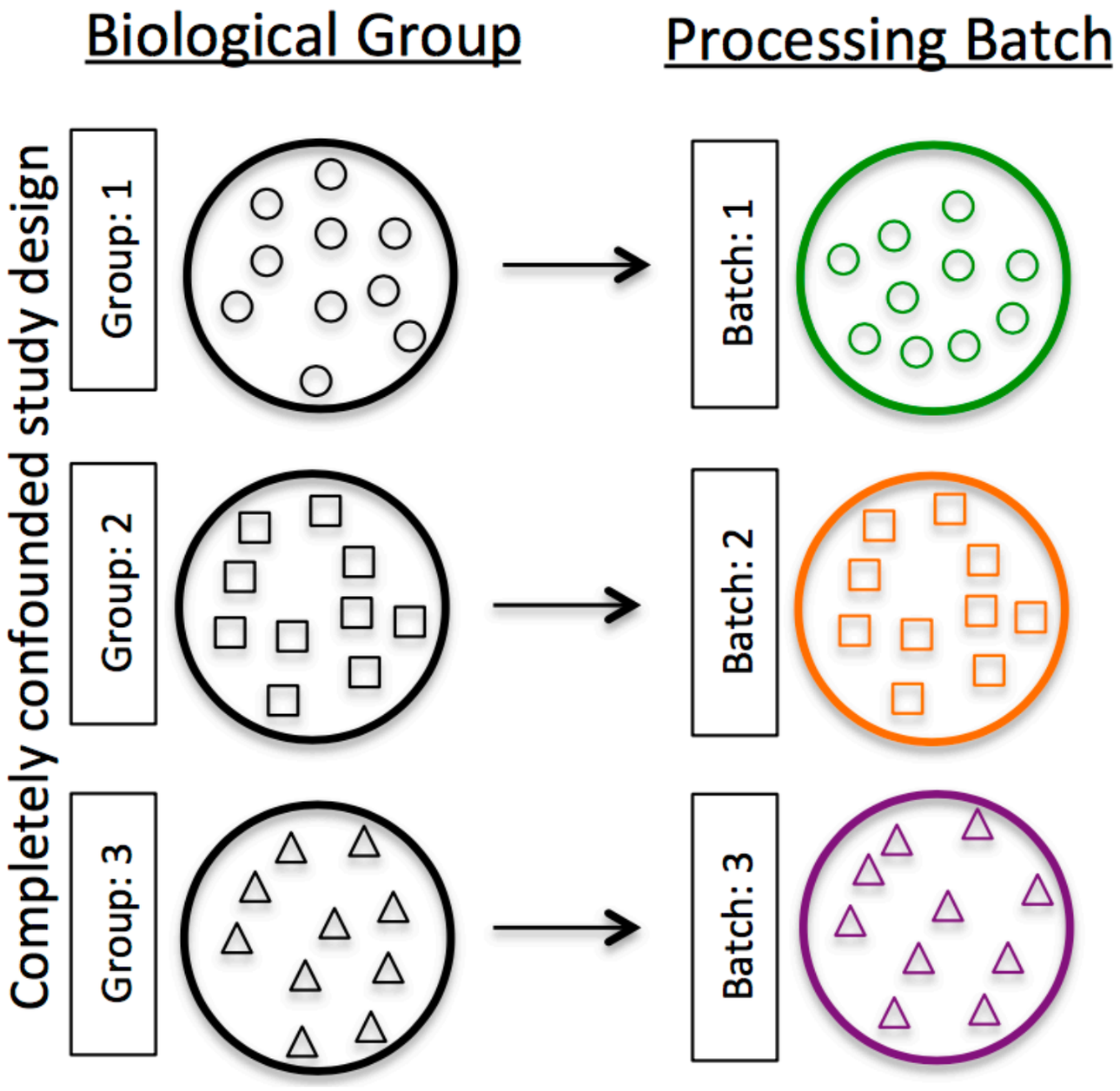
- **Best practices regarding batches:**
  - **Design the experiment from start to finish to avoid batches, if possible. If unsure of what can bring in a batch effect, talk with a biostats consultant before starting experiment.**
  - **Talking to a biostats after the experiment is complete is akin to bringing a dead person to a doctor**

https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

- If unable to avoid batches:
  - Do NOT confound your experiment by batch

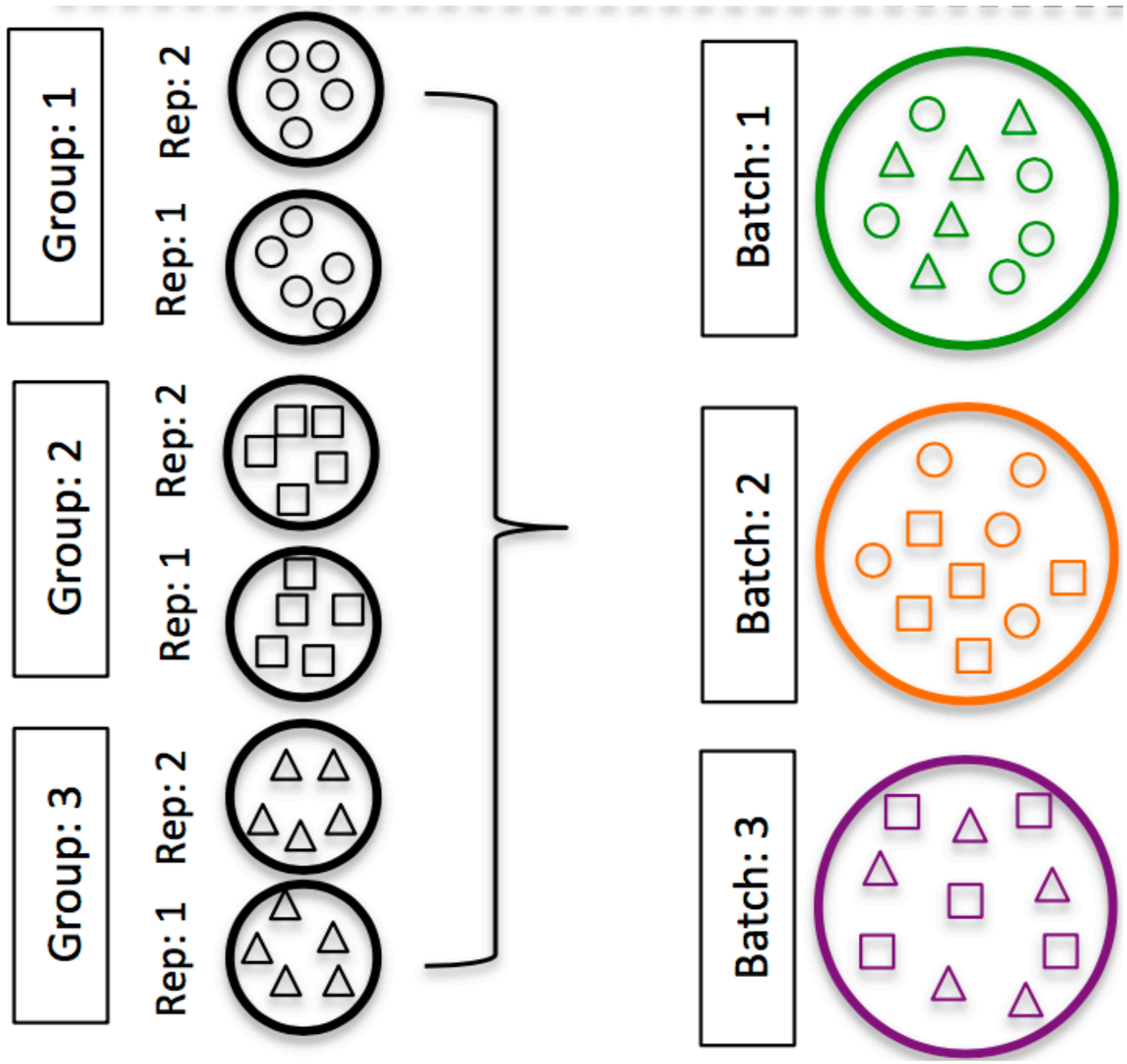


https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

- If unable to avoid batches:
  - DO split replicates of the different sample groups across batches. The more replicates the better (definitely 3 or more)



https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

- If unable to avoid batches:
  - DO include batch information in your experimental metadata. During the analysis, we can regress out the variation due to batch so it doesn't affect our results if we have that information.

sample	replicate	condition	batch
sample1	1	control	1
sample2	2	control	1
sample3	3	control	2
sample4	4	control	2
sample5	1	treatment1	1
sample6	2	treatment1	1
sample7	3	treatment1	2
sample8	4	treatment1	2
sample9	1	treatment2	1
sample10	2	treatment2	1
sample11	3	treatment2	2
sample12	4	treatment2	2

https://www.biorxiv.org/content/10.1101/025528v1

Genomics108

Introduction To Experimental Design

- **Take into consideration Surrogate Variable Analysis**

OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Abstract | High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

https://www.nature.com/articles/nrg2825

Genomics108

Introduction To Experimental Design

- **Take into consideration Surrogate Variable Analysis**

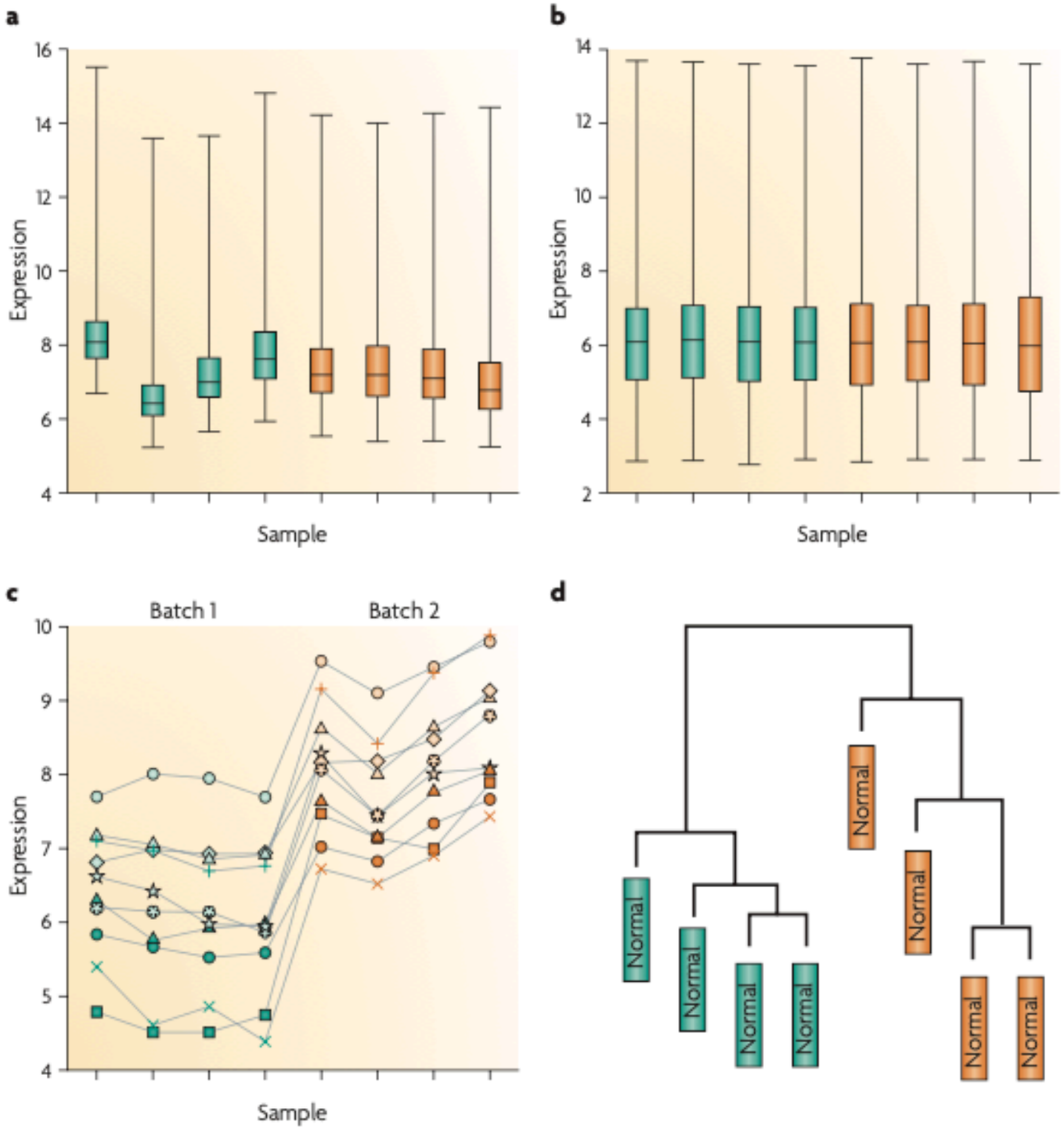


Figure 1 | **Demonstration of normalization and surviving batch effects.** For a published bladder cancer microarray data set obtained using an Affymetrix platform[9], we obtained the raw data for only the normal samples. Here, green and orange represent two different processing dates. **a** | Box plot of raw gene expression data (log base 2). **b** | Box plot of data processed with RMA, a widely used preprocessing algorithm for Affymetrix data[27]. RMA applies quantile normalization — a technique that forces the distribution of the raw signal intensities from the microarray data to be the same in all samples[28]. **c** | Example of ten genes that are susceptible to batch effects even after normalization. Hundreds of genes show similar behaviour but, for clarity, are not shown. **d** | Clustering of samples after normalization. Note that the samples perfectly cluster by processing date.

https://www.nature.com/articles/nrg2825

Genomics108

Introduction To Experimental Design

- **Take into consideration Surrogate Variable Analysis**

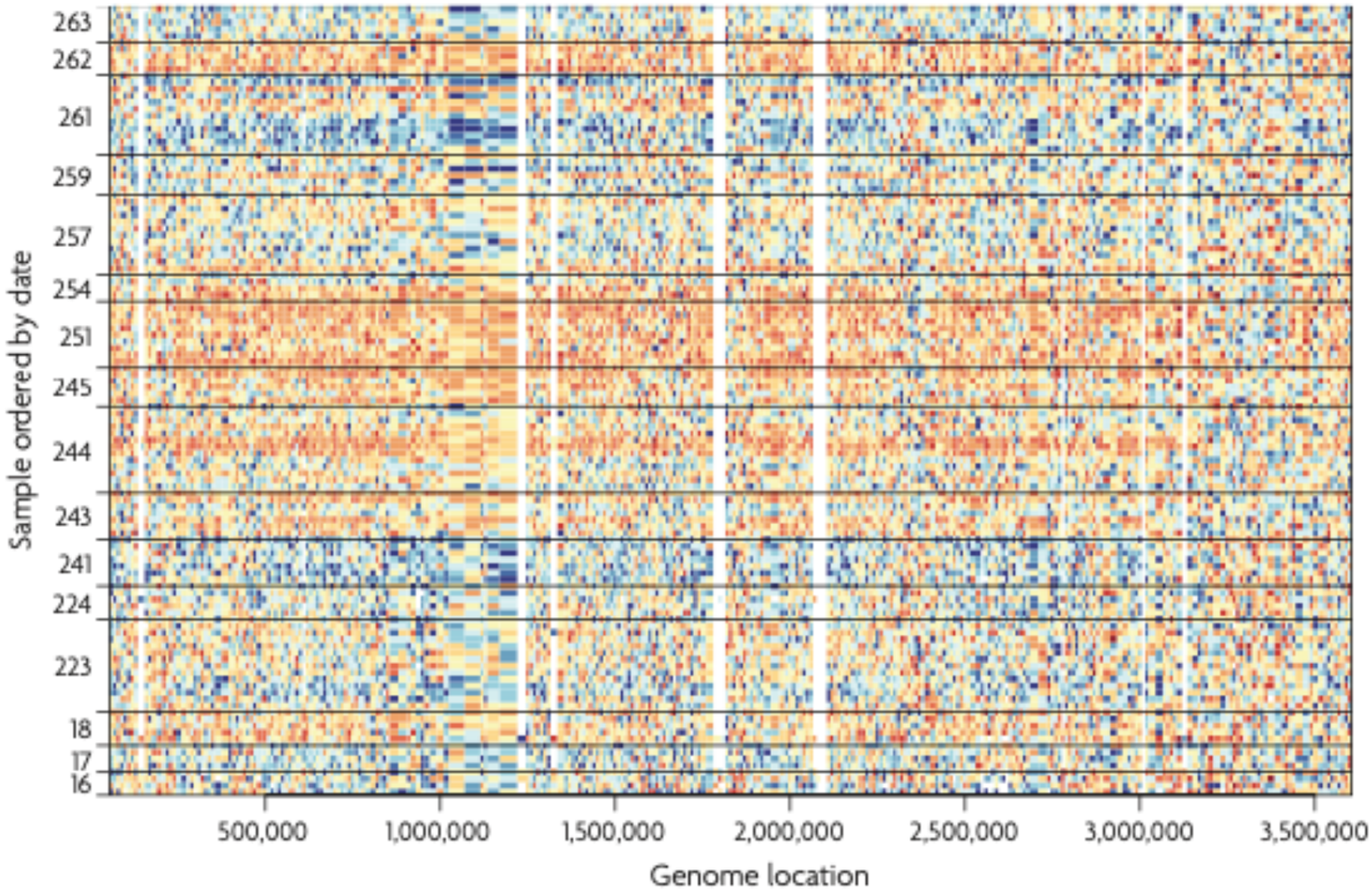


Figure 2 | **Batch effects for second-generation sequencing data from the 1000 Genomes Project.** Each row is a different HapMap sample processed in the same facility with the same platform. See Supplementary information S1 (box) for a description of the data represented here. The samples are ordered by processing date with horizontal lines dividing the different dates. We show a 3.5 Mb region from chromosome 16. Coverage data from each feature were standardized across samples: blue represents three standard deviations below average and orange represents three standard deviations above average. Various batch effects can be observed, and the largest one occurs between days 243 and 251 (the large orange horizontal streak).

https://www.nature.com/articles/nrg2825

Genomics108

Introduction To Experimental Design

- **Take into consideration Surrogate Variable Analysis**

OPEN ACCESS Freely available online

PLoS GENETICS

Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

Jeffrey T. Leek[1], John D. Storey[1,2*]

1 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

It has unambiguously been shown that genetic, environmental, demographic, and technical factors may have substantial effects on gene expression levels. In addition to the measured variable(s) of interest, there will tend to be sources of signal due to factors that are unknown, unmeasured, or too complicated to capture through simple models. We show that failing to incorporate these sources of heterogeneity into an analysis can have widespread and detrimental effects on the study. Not only can this reduce power or induce unwanted dependence across genes, but it can also introduce sources of spurious signal to many genes. This phenomenon is true even for well-designed, randomized studies. We introduce "surrogate variable analysis" (SVA) to overcome the problems caused by heterogeneity in expression studies. SVA can be applied in conjunction with standard analysis techniques to accurately capture the relationship between expression and any modeled variables of interest. We apply SVA to disease class, time course, and genetics of gene expression studies. We show that SVA increases the biological accuracy and reproducibility of analyses in genome-wide expression studies.

Citation: Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3(9): e161. doi:10.1371/journal.pgen.0030161

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030161

Genomics108

Introduction To Experimental Design

- **Know when your numbers are significant**

Know when your numbers are significant

Experimental biologists, their reviewers and their publishers must grasp basic statistics, urges **David L. Vaux**, or sloppy science will continue to grow.

https://www.nature.com/articles/492180a

Genomics108

Introduction To Experimental Design

- **Know when your numbers are significant**

STATISTICS GLOSSARY

Some common statistical concepts and their uses in analysing experimental results.

Term	Meaning	Common uses
Standard deviation (s.d.)	The typical difference between each value and the mean value.	Describing how broadly the sample values are distributed. s.d. $=\sqrt{(\sum (x-\text{mean})^2/(N-1))}$
Standard error of the mean (s.e.m.)	An estimate of how variable the means will be if the experiment is repeated multiple times.	Inferring where the population mean is likely to lie, or whether sets of samples are likely to come from the same population. s.e.m. $=$ s.d. $/\sqrt{N}$
Confidence interval (CI; 95%)	With 95% confidence, the population mean will lie in this interval.	To infer where the population mean lies, and to compare two populations. CI $=$ mean $\pm$ s.e.m. $\times t_{(N-1)}$
Independent data	Values from separate experiments of the same type that are not linked.	Testing hypotheses about the population.
Replicate data	Values from experiments where everything is linked as much as possible.	Serves as an internal check on performance of an experiment.
Sampling error	Variation caused by sampling part of a population rather than measuring the whole population.	Can reveal bias in the data (if it is too small) or problems with conduct of the experiment (if it is too big). In binomial distributions (such as live and dead cell counts) the expected s.d. is $\sqrt{(N\times p\times(1-p))}$; in Poisson distributions (for example, cells per field) the expected s.d. is $\sqrt{\text{mean}}$.

N, number of independent samples; *t*, the *t*-statistic; *p*, probability.

https://www.nature.com/articles/492180a

Genomics108

Introduction To Experimental Design

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

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

NICHOLAS J. SCHURCH,[1,6] PIETÁ SCHOFIELD,[1,2,6] MAREK GIERLIŃSKI,[1,2,6] CHRISTIAN COLE,[1,6] ALEXANDER SHERSTNEV,[1,6] VIJENDER SINGH,[2] NICOLA WROBEL,[3] KARIM GHARBI,[3] GORDON G. SIMPSON,[4] TOM OWEN-HUGHES,[2] MARK BLAXTER,[3] and GEOFFREY J. BARTON[1,2,5]

[1]Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[2]Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[3]Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom
[4]Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[5]Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

https://rnajournal.cshlp.org/content/early/2016/03/28/rna.053959.115

Genomics108

Introduction To Experimental Design

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

ABSTRACT

RNA-seq is now the technology of choice for genome-wide differential gene expression experiments, but it is not clear how many biological replicates are needed to ensure valid biological interpretation of the results or which statistical tools are best for analyzing the data. An RNA-seq experiment with 48 biological replicates in each of two conditions was performed to answer these questions and provide guidelines for experimental design. With three biological replicates, eight of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates. This rises to >85% for the subset of SDE genes changing in expression by more than fourfold. To achieve >85% for all SDE genes regardless of fold change requires more than 20 biological replicates. The same eight tools successfully control their false discovery rate at ≲5% for all numbers of replicates, while the remaining three tools fail to control their FDR adequately, particularly for low numbers of replicates. For future RNA-seq experiments, these results suggest that more than six biological replicates should be used, rising to more than 12 when it is important to identify SDE genes for all fold changes. If less than 12 replicates are used, a superior combination of true positive and false positive performances makes *edgeR* the leading tool. For higher replicate numbers, minimizing false positives is more important and *DESeq* marginally outperforms the other tools.

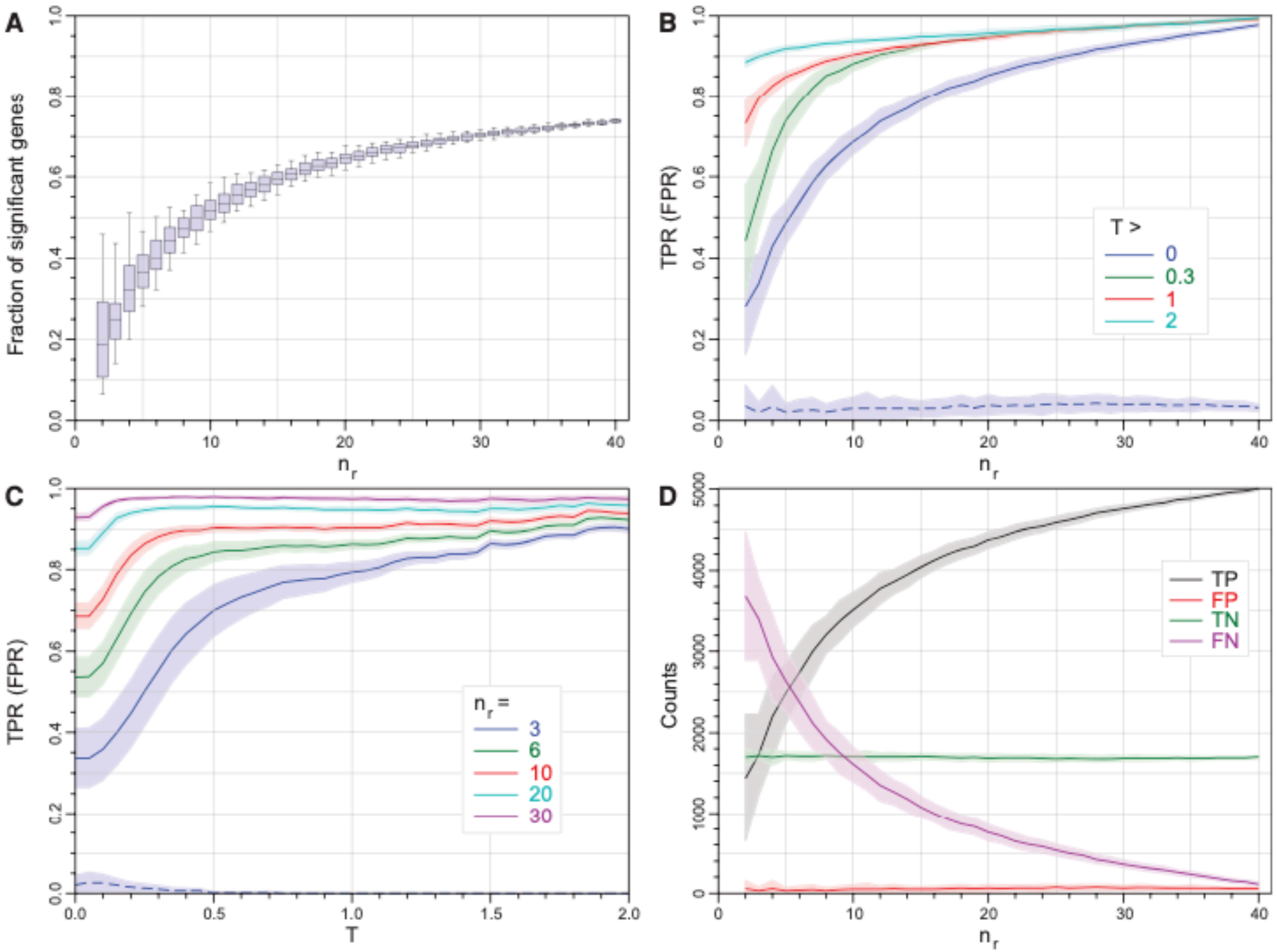
Keywords: RNA-seq; benchmarking; differential expression; replication; yeast; experimental design; statistical power

https://rnajournal.cshlp.org/content/early/2016/03/28/rna.053959.115

Genomics108

Introduction To Experimental Design

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



https://rnajournal.cshlp.org/content/early/2016/03/28/rna.053959.115

Genomics108

Introduction To Experimental Design

- **RNA-Seq differential expression studies: more sequence or more replication?**

BIOINFORMATICS **DISCOVERY NOTE** *Vol. 30 no. 3 2014, pages 301–304*
doi:10.1093/bioinformatics/btt688

Gene expression

Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

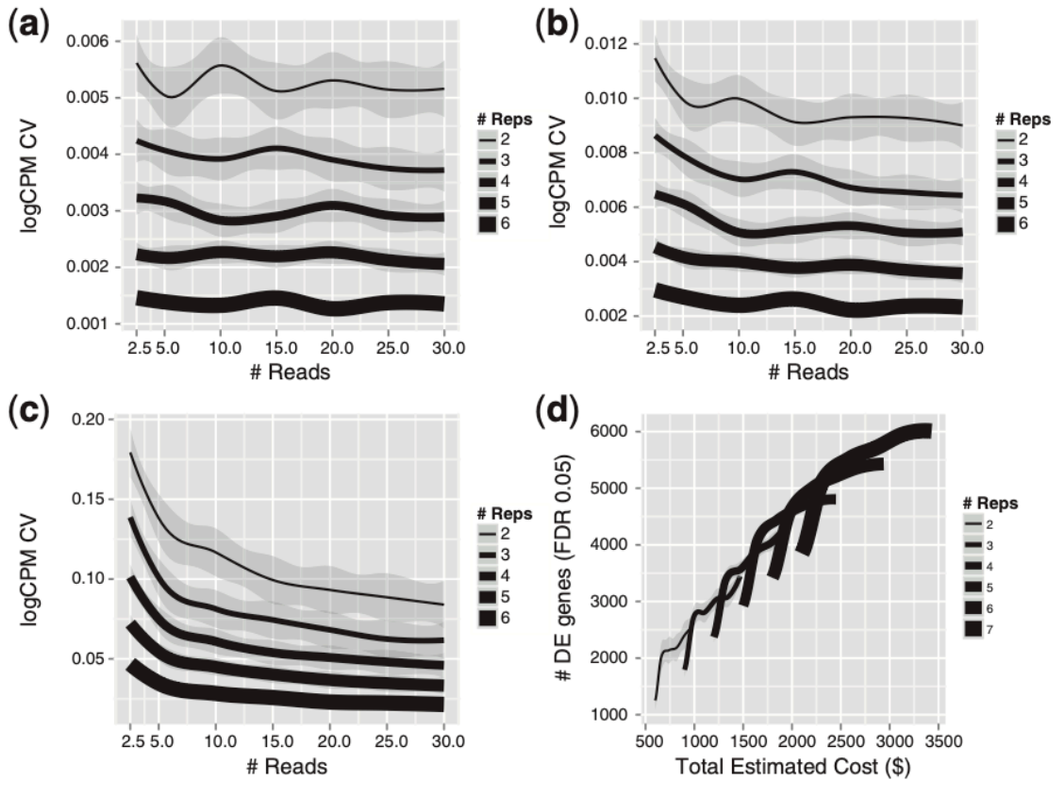
Associate Editor: Janet Kelso

https://academic.oup.com/bioinformatics/article/30/3/301/228651

Genomics108

Introduction To Experimental Design

- **RNA-Seq differential expression studies: more sequence or more replication?**



CV = Coefficient of Variation

Fig. 2. (**a–c**) The CV of logCPM for high expression level genes (a), medium expression level genes (b) and low expression level genes (c) (see Section 2 for definition). High/medium expression level genes have low CV for expression level estimates. Adding sequencing depth did not have significant effect on accuracy of estimation, whereas adding biological replicates improved accuracy significantly. For low expression level genes, both adding sequencing depth and adding biological replication level improved expression level estimation accuracy. (**d**) Number of DE genes plotted against the total estimated sequencing cost. If higher numbers of DE genes are needed, increased biological replication should be used

https://academic.oup.com/bioinformatics/article/30/3/301/228651

Genomics108

Introduction To Experimental Design

- **StatQuest: RNA-seq - the problem with technical replicates**

How much slower is convergence to zero?

$$\text{Average} = \mu + \frac{5+5-1-1-1}{5} + \frac{-2+5+2-2-1}{5}$$

We did 3 technical replicates of mouse #2, and thus...

This is what we had with 5 biological replicates.

With 3 technical reps, we'd need 15 samples total to get the same term as with 5 biological reps.

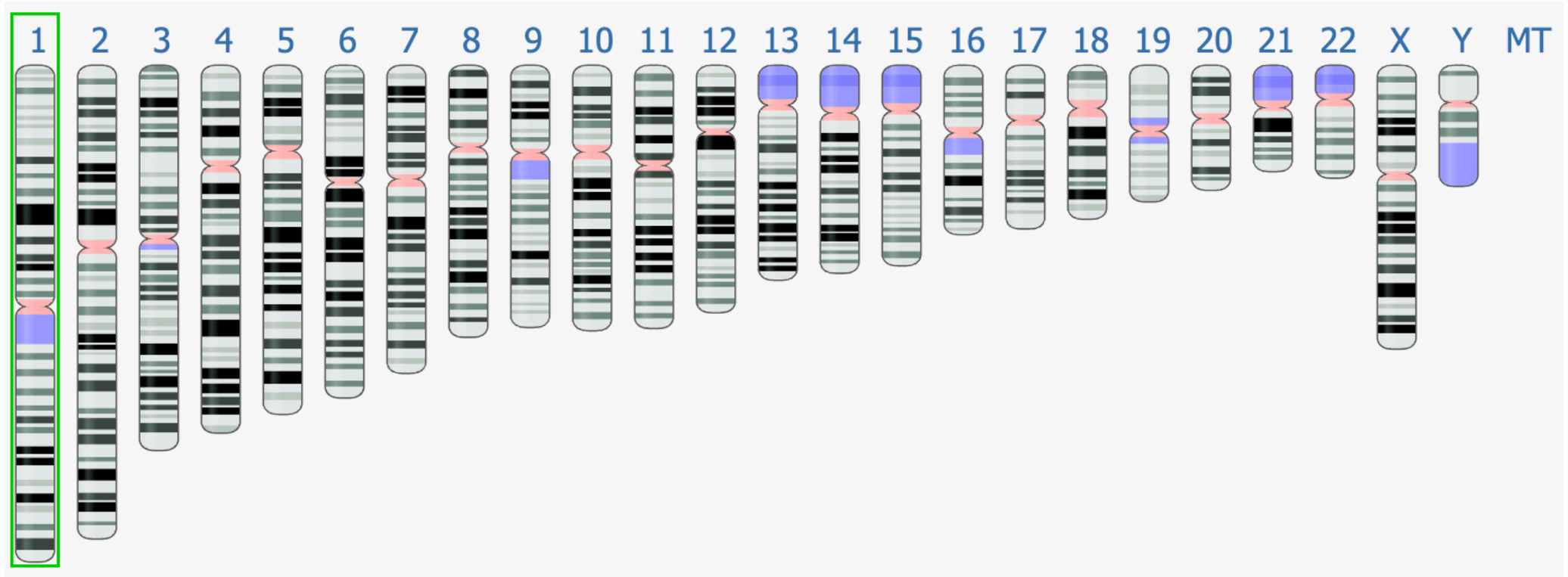
$$\frac{5-1+4+2-5}{5} = \frac{5+5+5-1-1-1+4+4+4+2+2+2-5-5-5}{15}$$

https://youtu.be/gKnfP2_Xdpo

Genomics108

References

- **RNASeq Tutorials Prepared by the Harvard Chan Bioinformatics Core**
  - **This lesson has been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). 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.**
  - **NOTE01: The Single-cell RNA-Seq Analysis Workflow lesson was adapted from Dr. Mary Piper's presentation for the Boston-area Women's Bioinformatics Meetup.**
  - **NOTE02: The Visualizing the Results of a DGE Experiment Materials and hands-on activities were adapted from RNA-seq workflow on the Bioconductor website.**
- ***Other Bibliographic References are:***
  - ***Tackling the widespread and critical impact of batch effects in high-throughput data***
  - ***Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis***
  - ***Know when your numbers are significant***
  - ***How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?***
  - ***RNA-seq differential expression studies: more sequence or more replication?***
  - ***A survey of best practices for RNA-seq data analysis***



Genomics108

BIOL647
Digital Biology

Rodolfo Aramayo

CC BY NC SA