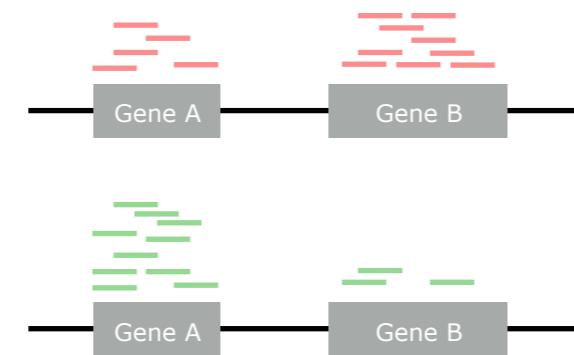
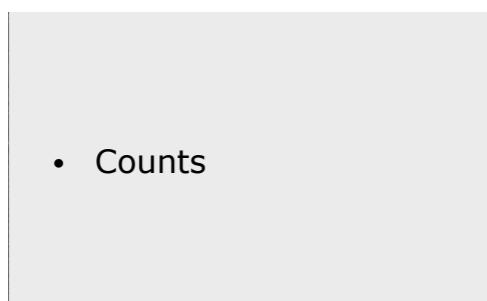
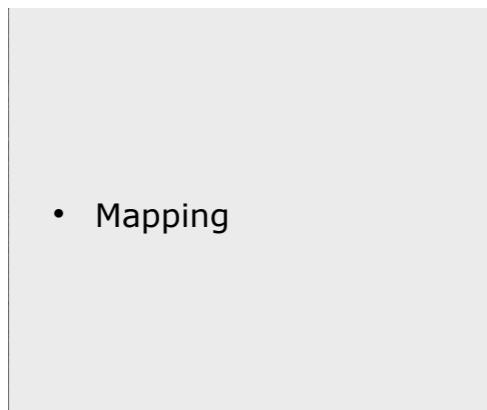
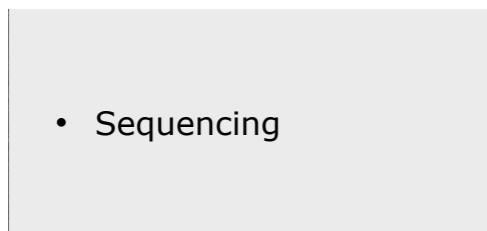
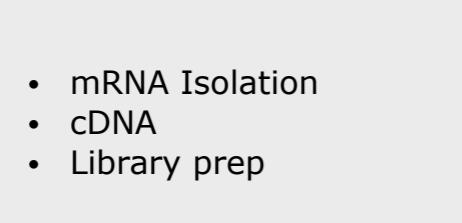




**Molecular
BioSystems****PAPER**[View Article Online](#)[View Journal](#) | [View Issue](#)Cite this: *Mol. BioSyst.*, 2016,
12, 508**Strand-specific RNA-seq analysis of the
Lactobacillus delbrueckii subsp. *bulgaricus*
transcriptome†**Huajun Zheng,^{‡^a} Enuo Liu,^{‡^a} Tao Shi,^a Luyi Ye,^a Tomonobu Konno,^b
Munehiro Oda^c and Zai-Si Ji*^{ab}

***Lactobacillus delbrueckii* subsp. *bulgaricus* 2038** (*Lb. bulgaricus* 2038) is an industrial bacterium that is used as a starter for dairy products. ... Here, we utilized RNA-seq to explore the transcriptome of *Lb. bulgaricus* 2038 from four different growth phases under whey conditions. The most abundantly expressed genes in the four stages were mainly involved in translation (for the logarithmic stage), glycolysis (for control/lag stages), lactic acid production (all the four stages), and 10-formyl tetrahydrofolate production (for the stationary stage).

Product	% expressed
Conserved hypothetical protein	16.7
Small heat shock protein	5.7
Chaperonin GroES	2.6
Conserved hypothetical protein	2.2
Chaperonin GroEL	1.3



$$\text{Gene A: } \frac{5}{10} = 0.5 \text{ fold change}$$

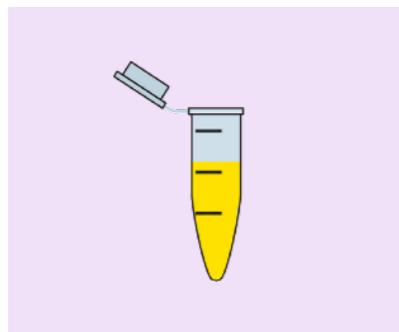
$$\text{Gene B: } \frac{9}{3} = 3 \text{ fold change}$$



<http://www.rna-seqblog.com>



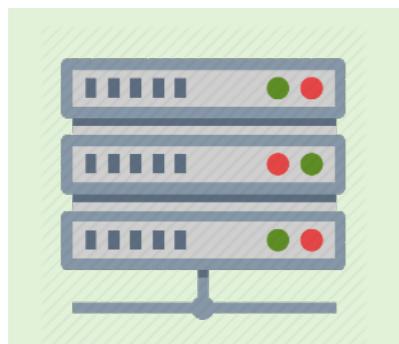
Sample Design



Sample preparation
RNA extraction
Cleaning (e.g. remove ribosomal RNA)



Library Prep
Illumina (paired-end) sequencing



QC and QF
Mapping (genome / transcriptome)
Count Tables (raw counts)



Data Analysis



What is the purpose of your RNAseq experiment?

The central purpose of an RNA-seq experiment can be:

- to quantify transcription (DE or time series)
- establish a reference (transcriptome)
- to identify the structure (exons) of transcribed genes
- explore splice junctions
- characterise small RNA
- identify novel/rare transcripts
- transcriptional start sites
- to quantify transcription



What is the purpose of your RNAseq experiment?

Purpose

Design

Preparation

Methode

Analysis

Extras



What resources are available and what is the quality?

References (e.g. genome, transcriptome)

Assembly Quality (e.g. draft, contamination)

Annotation Level (e.g. unknown function, missing)



What resources are available and what is the quality?

Resources

Design

Preparation

Methode

Analysis

Extras



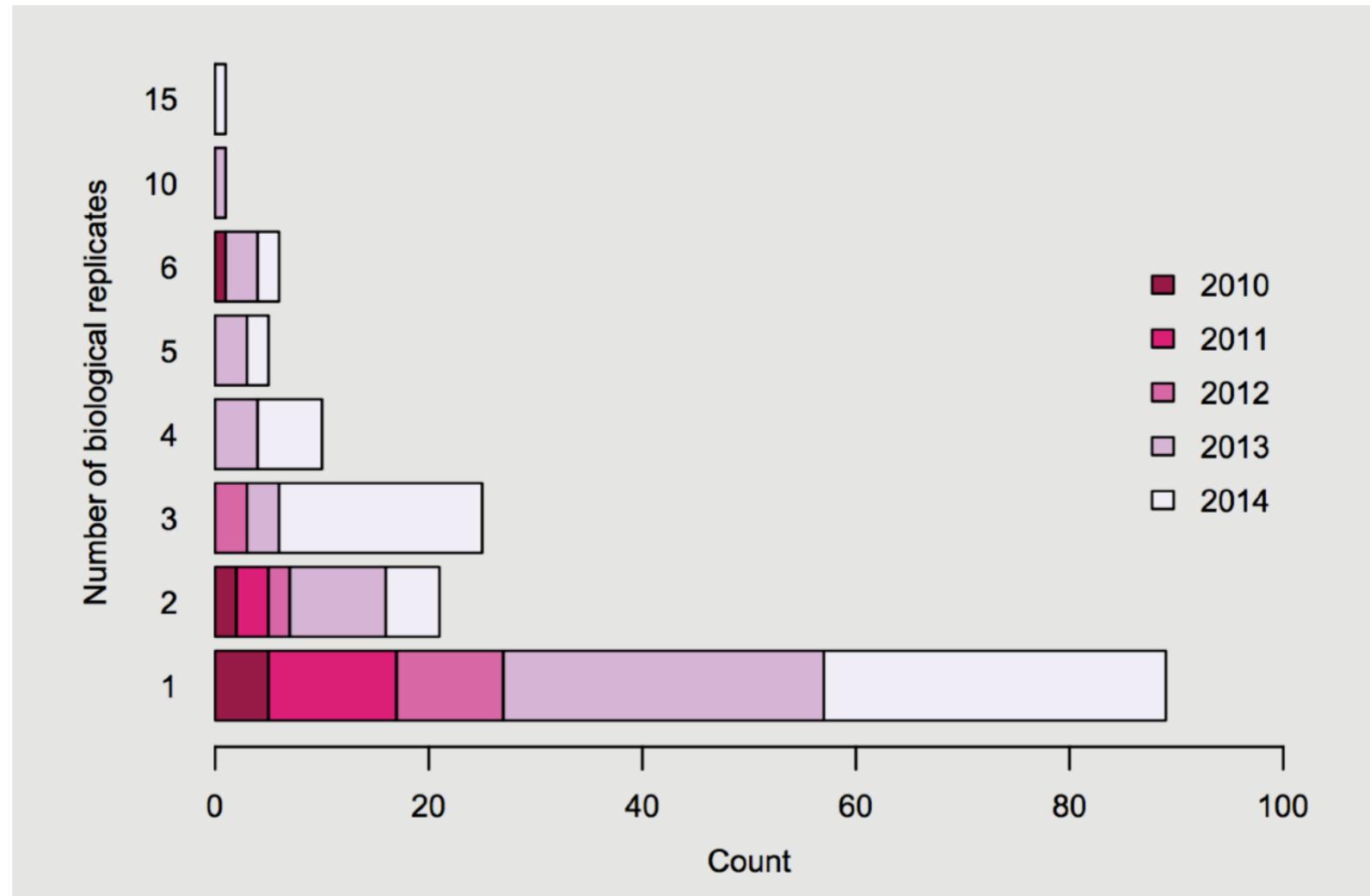
What factors influence the design?

How many **samples / replicates** are needed?

What (min) depth of sequencing **coverage** is required?

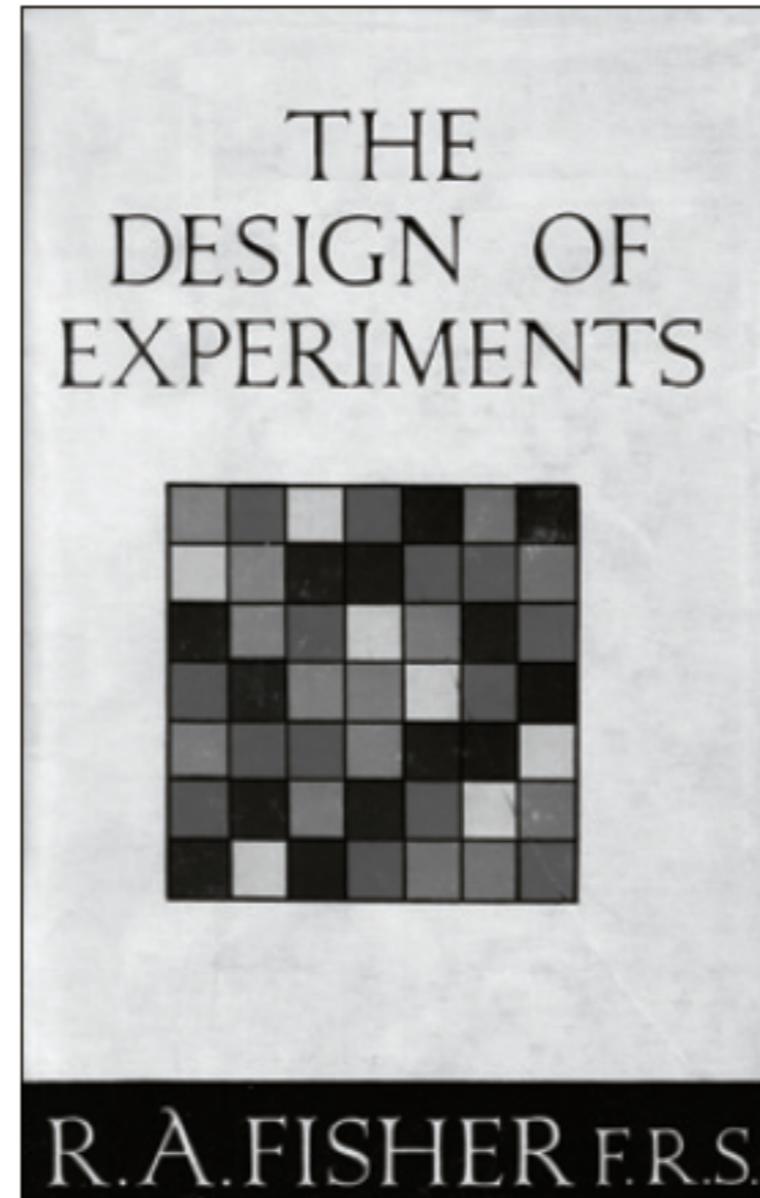
What is the **trade off** between coverage and biological replicates?

How much **money** do we have?



Library : pooled samples Replicates: n=1

Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.



Fisher, R. A., (1935) The Design of Experiments. Ed. 2. Oliver & Boyd, Edinburgh.

Copyright © 2010 by the Genetics Society of America
DOI: 10.1534/genetics.110.114983

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge¹

Department of Statistics, Purdue University, West Lafayette, Indiana 47907

Manuscript received January 31, 2010

Accepted for publication March 15, 2010

"Indisputably, the best way to ensure reproducibility and accuracy of results is to include independent **biological replicates** (technical replicates are no substitute) and to acknowledge anticipated nuisance factors (e.g., lane, batch, and flow-cell effects) in the design."

Auer & Doerge (2010) Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, 185 no. 2, 405-416-2223.

Differential expression in RNA-seq: a matter of depth

Sonia Tarazona^{1,2}, Fernando García-Alcalde¹, Joaquín Dopazo¹, Alberto Ferrer², and Ana Conesa^{1,*}

¹ Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain

² Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Valencia, Spain

* Corresponding author. Email: aconesa@cipf.es

August 29, 2011

“Our results reveal that most existing methodologies suffer from a strong dependency on **sequencing depth** for their differential expression calls and that this results in a considerable number of false positives that increases as the number of reads grows.”

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21, 2213–2223.

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,*}¹Institute of Genomics and Systems Biology, ²Committee on Development, Regeneration, and Stem Cell Biology and³Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso

“Our analysis showed that sequencing **less reads and performing more biological replication** is an effective strategy to increase power and accuracy in large-scale differential expression RNA-seq studies, and provided new insights into efficient experiment design of RNA-seq studies.”

2x10M (20M) PE-reads > 2x15M (30M) PE-reads => 6% increase
> 3x10M (30M) PE-reads => 35% increase

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

NICHOLAS J. SCHURCH,^{1,6} PIETÀ 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}

¹Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

²Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

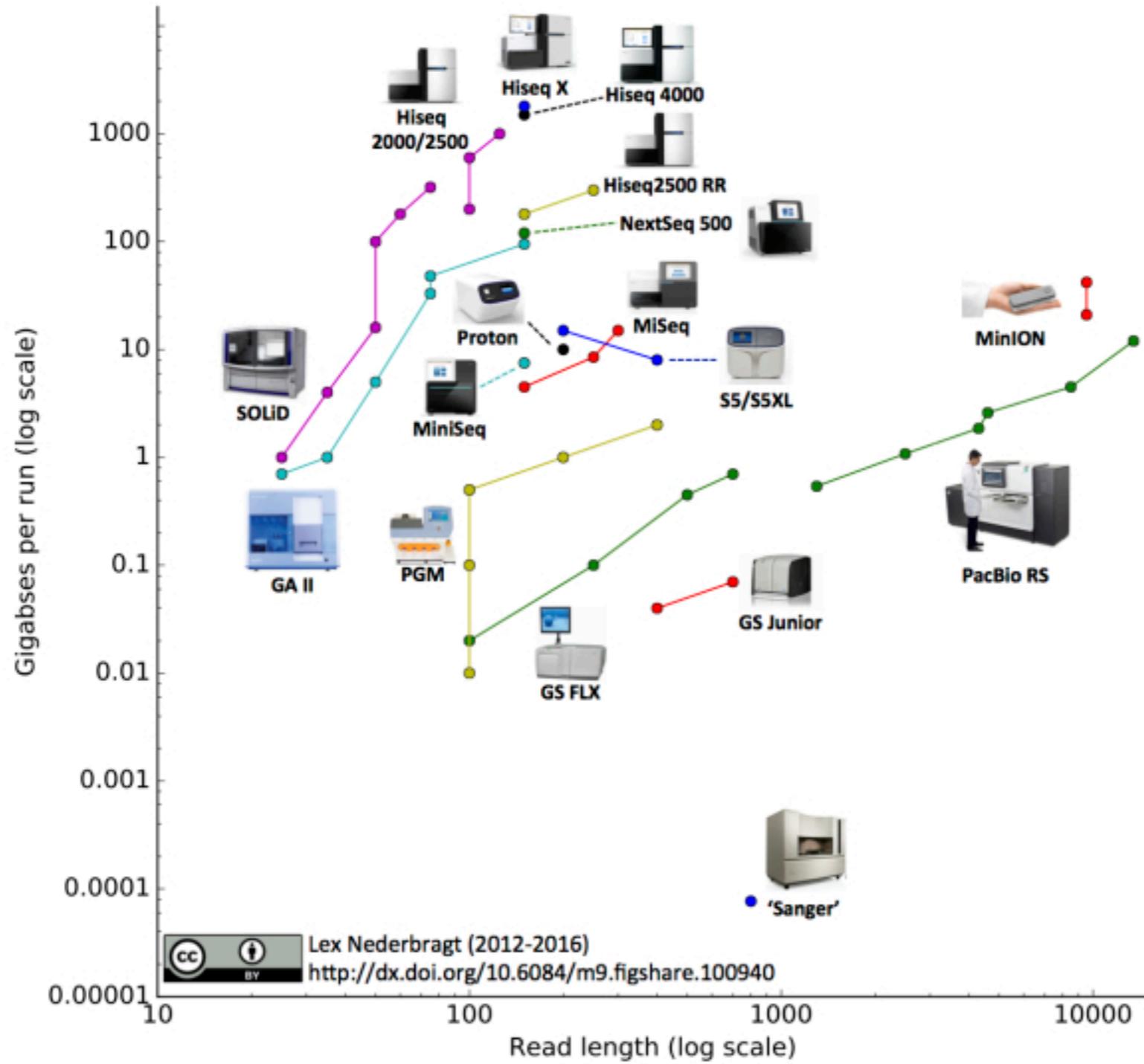
³Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

⁴Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

⁵Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

“With **three biological replicates**, nine 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**.”

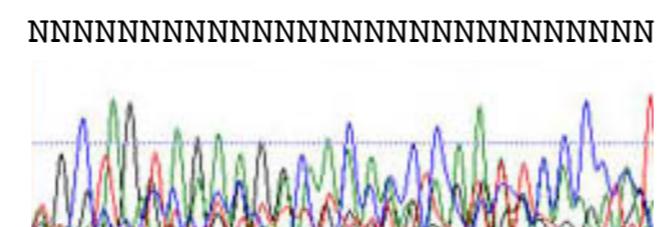
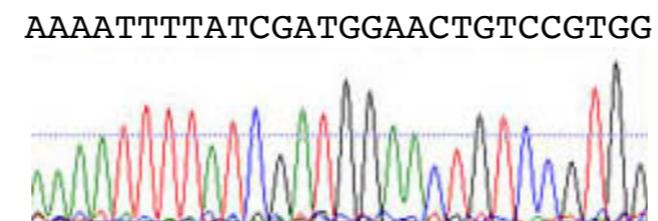
Schurch et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 22, 839–851.



Source: <https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/>

Signal-to-noise ratio

$$SNR = \frac{P_{signal}}{P_{noise}}$$



December 2011



Transitioning from Microarrays to mRNA-Seq

By providing HiSeq® 2000 mRNA-Seq data in two formats, Expression Analysis enables its customers to compare sequencing results with older array data, while gaining insight into the entire transcriptome.

Transition

Intensity ▷ Count

http://www.illumina.com/content/dam/illumina-marketing/documents/icommunity/article_2011_12_ea_rna-seq.pdf

Signal-to-noise ratio

$$SNR = \frac{P_{signal}}{P_{noise}}$$

Poisson counting errors - The uncertainty inherited in any count-based measurements.

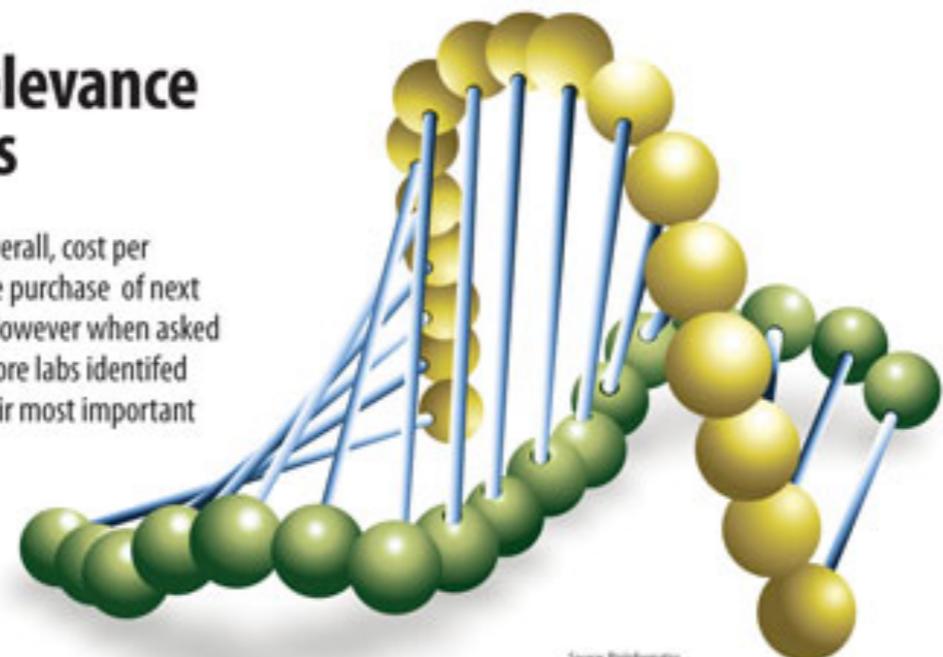
Non-Poisson technical variance - The observed imprecision between repeat measurements.

Biological variance - The natural variation in gene expression measurements.

NGS: Cost and Relevance Are Key to Buyers

Recent market research shows that, overall, cost per base was the most cited concern in the purchase of next generation sequencing instruments. However when asked to identify their top three concerns, more labs identified "Appropriate to My Application" as their most important criteria.

The 10 Most Critical Platform Attributes as Defined by Purchasers



1. Cost per base	43%	6. Read length	24%
2. Sequencing data quality	34%	7. Instrument cost	18%
3. Appropriate for my application	32%	8. Number of reads	17%
4. Reproducibility/accuracy	31%	9. Available software analysis tools	16%
5. Amount of DNA/RNA needed per experiment	25%	10. Instrument reliability	16%

Statistical Power of RNA-seq Experiments

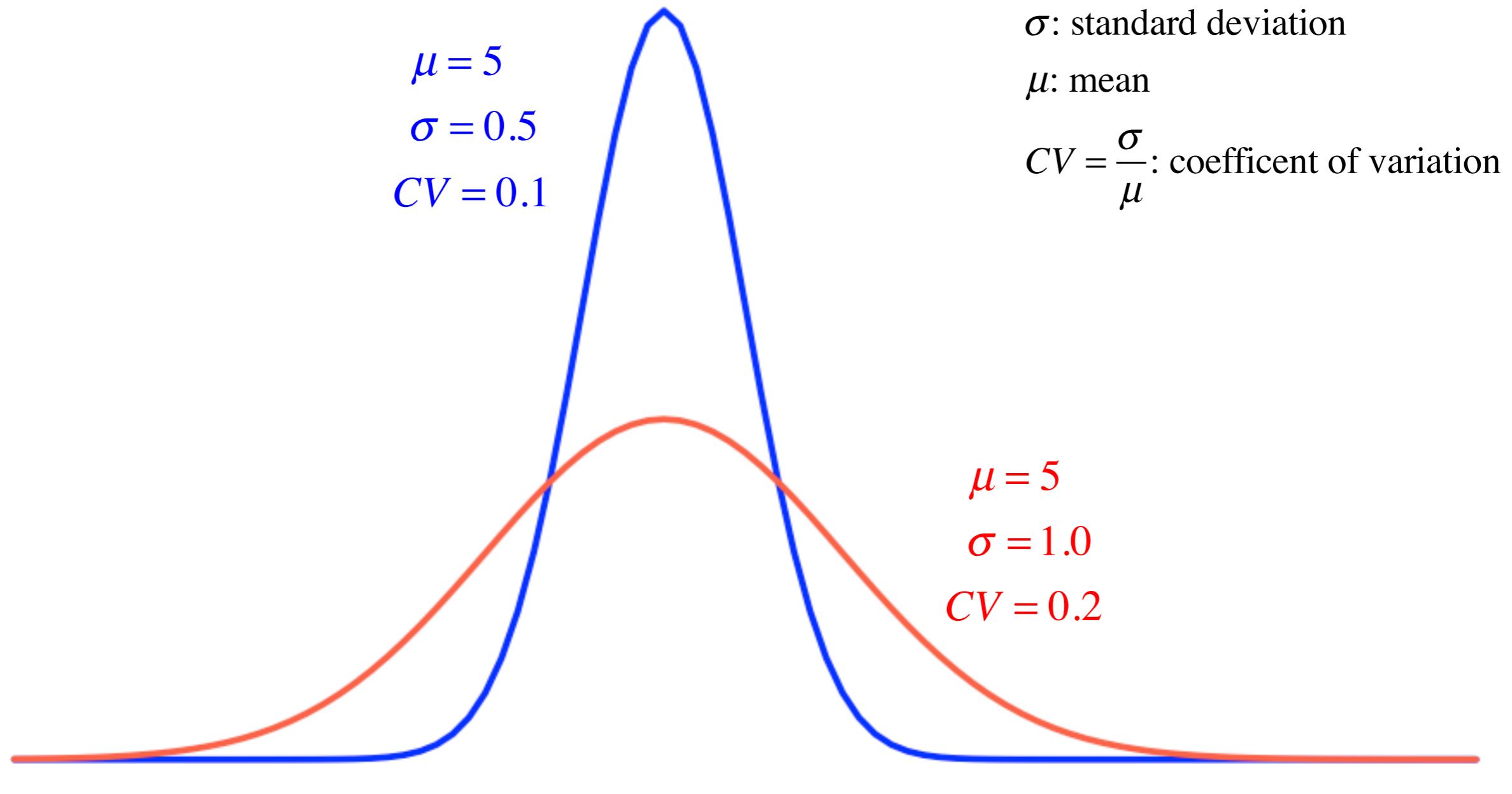
Power analysis is an important aspect of **experimental design**. It allows us to **determine the sample size required** to detect an effect of a given size with a given degree of confidence. Conversely, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints. If the probability is unacceptably low, we would be wise to alter or abandon the experiment.

The following four quantities have an intimate relationship:

- **Sample size**
- **Effect size** (two-fold, three-fold, ...)
- **Significance level** = $P_{(\text{Type I error})}$ (probability of finding an effect that is not there)
- **Power** = $1 - P_{(\text{Type II error})}$ (probability of finding an effect that is there)

Given any three, we can determine the fourth.

Source: <http://www.statmethods.net/stats/power.html>



Inbred vs. Wild
Model Organisms vs. Non-Model Organisms

$$CV = \frac{\sigma}{\mu}$$

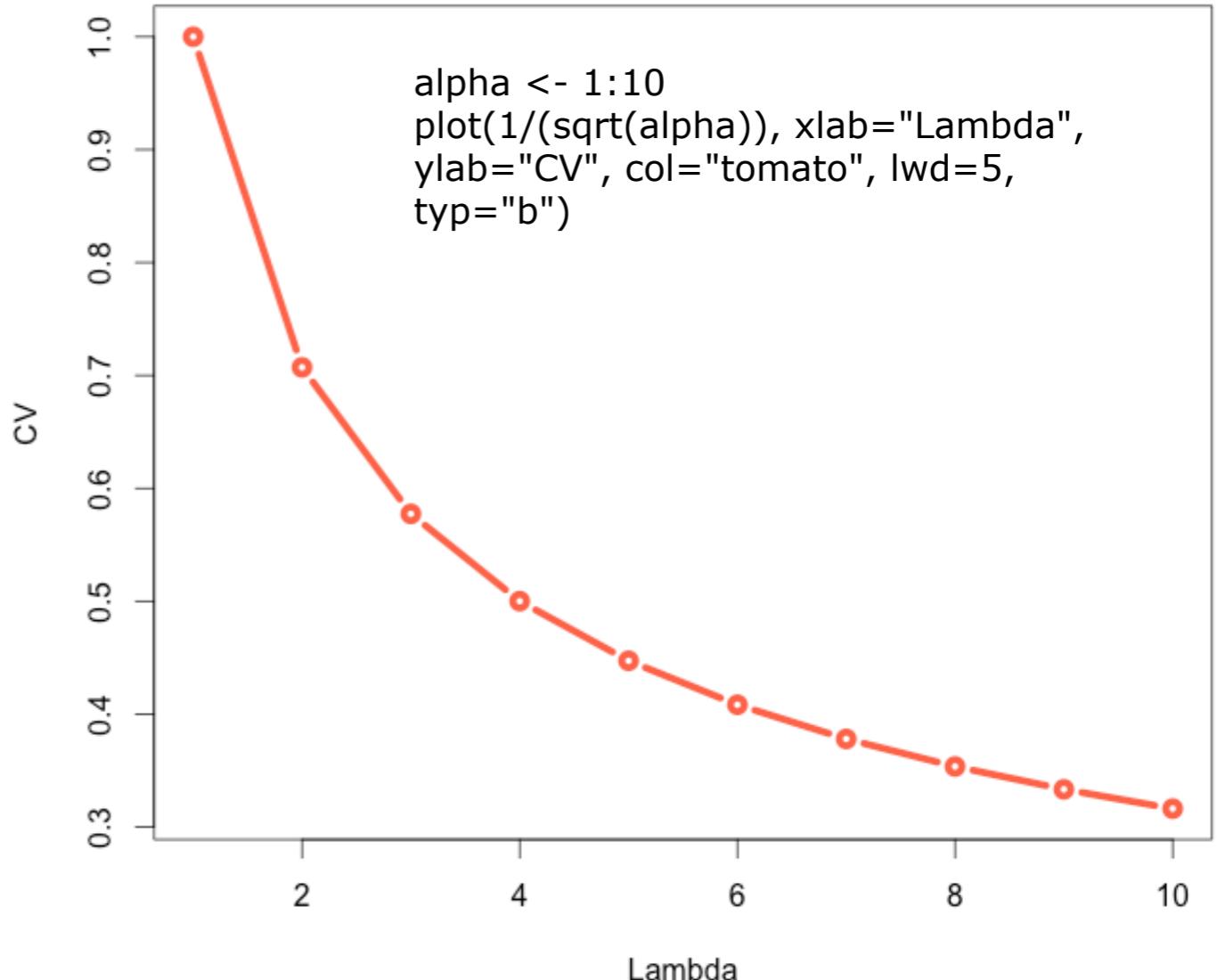
CV : coefficient of variation
 σ : standard deviation
 μ : mean

inbred animal strains: $CV \leq 0.2$

unrelated individuals: $CV > 0.3$

Poisson Distribution

$$CV = \frac{\sigma}{\mu} = \lambda^{-\frac{1}{2}} = \frac{1}{\sqrt{\lambda}}$$



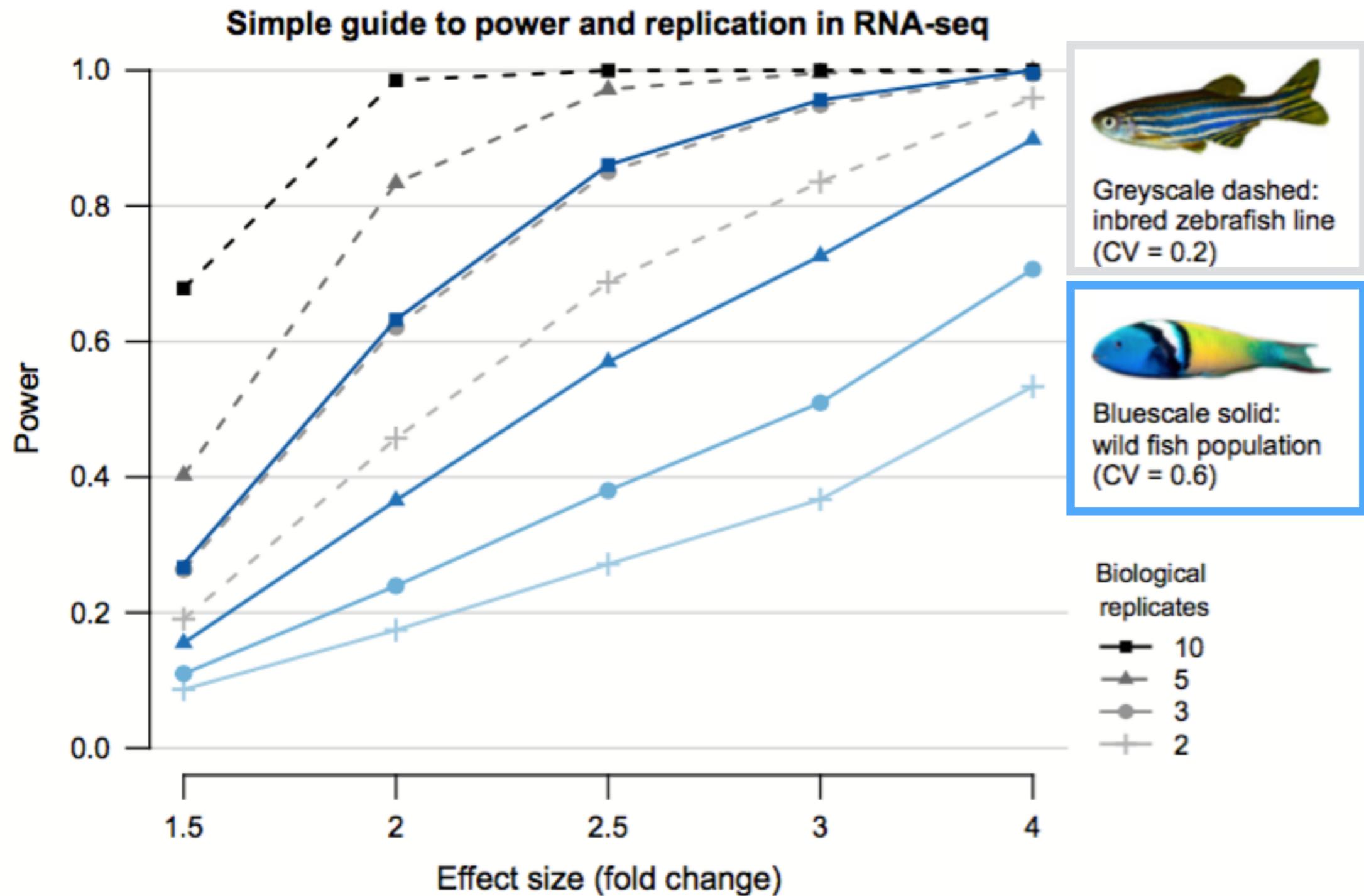
CV: coefficient of variation

λ : average number of event per interval (= mean)

σ : standard deviation ($= \sqrt{Variance}$)

μ : mean

! The expected value and variance of a Poisson-distributed random variable are both equal to λ .



Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.

```
source("https://bioconductor.org/biocLite.R")
biocLite("RNASeqPower")

library(RNASeqPower); packageVersion("RNASeqPower")
rnapower(depth=100, cv=0.4, effect=c(1, 2), alpha= .05, power=c(.8, .9))
  0.8      0.9
1     Inf      Inf
2 5.554381 7.43574
```

The **depth** of sequencing and consequent expected count μ for a given transcript.
The coefficient of variation (**cv**) of counts within each of the two groups.
The relative expression (**effect**) that we wish to detect Δ .
The target false positive rate α and false negative rate β desired (or power = $1 - \beta$).
The number of samples **n** in each group.

Steven N Hart, Terry M Therneau, Yuji Zhang and Jean-Pierre Kocher (2013) Calculating Sample Size Estimates for RNA Sequencing Data, J Comput Biol. 20(12):970-8

```
rnapower(depth=100, cv=0.4, effect=c(1, 2), alpha= .05, power=c(.8, .9))
```

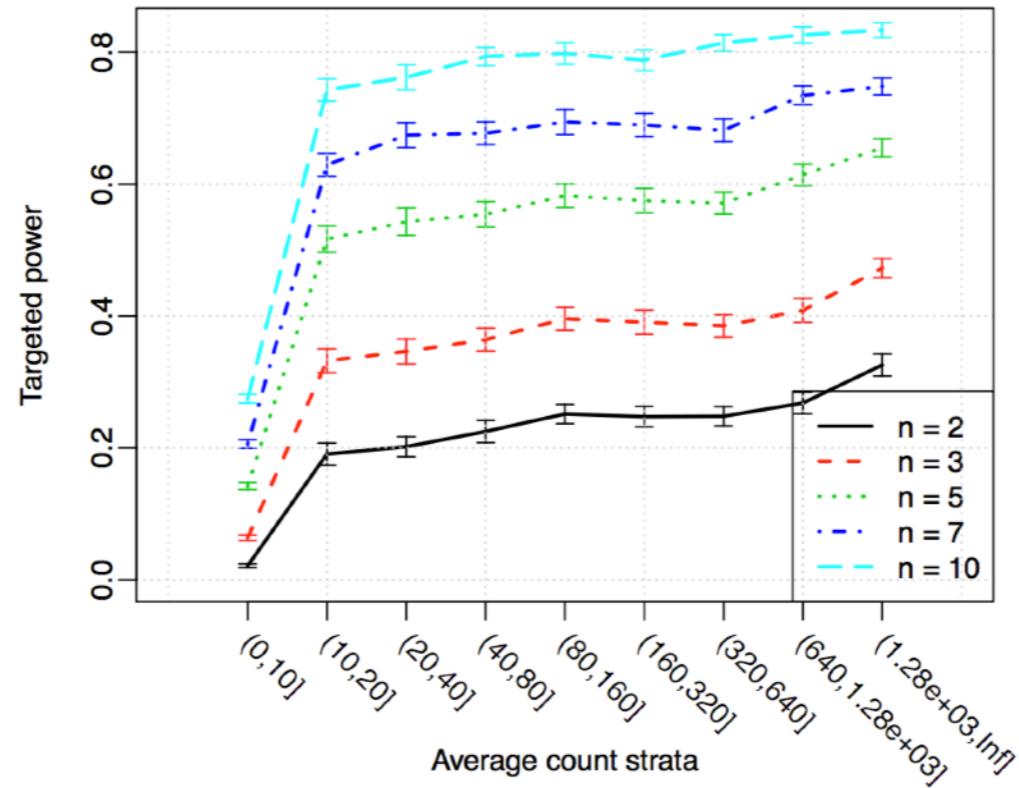
```
0.8      0.9  
1       Inf      Inf  
2 5.554381 7.43574
```

```
rnapower(depth=100, cv=0.2, effect=c(1, 2), alpha= .05, power=c(.8, .9))
```

```
0.8      0.9  
1       Inf      Inf  
2 1.633641 2.186982
```

Steven N Hart, Terry M Therneau, Yuji Zhang and Jean-Pierre Kocher (2013) Calculating Sample Size Estimates for RNA Sequencing Data, J Comput Biol. 20(12):970-8

```
source("https://bioconductor.org/biocLite.R")  
  
biocLite("PROPER")  
  
library(PROPER); packageVersion("PROPER")  
  
sim.opts <- RNaseq.SimOptions.2grp(ngenes = 5000, p.DE=0.05, lOD=1, lBaselineExpr=1)  
  
sim.res <- runSims(Nreps = c(3, 7, 11), sim.opts=sim.opts, DEMethod="edgeR", nsims=20)  
  
sim.powers <- comparePower(sim.res, alpha.type="fdr", alpha.nominal=0.1,  
stratify.by="expr", delta=0.5)  
  
plotPower(sim.powers)
```





- Expression landscape?
- Library complexity?
- Read distribution?

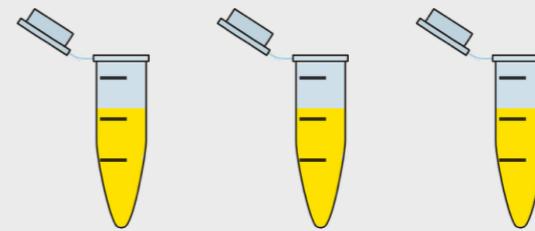
=> **Pilot sequencing** approach*

*Todd et al. (2016) The power and promises of RNA-seq in ecology and evolution. Molecular Ecology. 25, 1224-1241

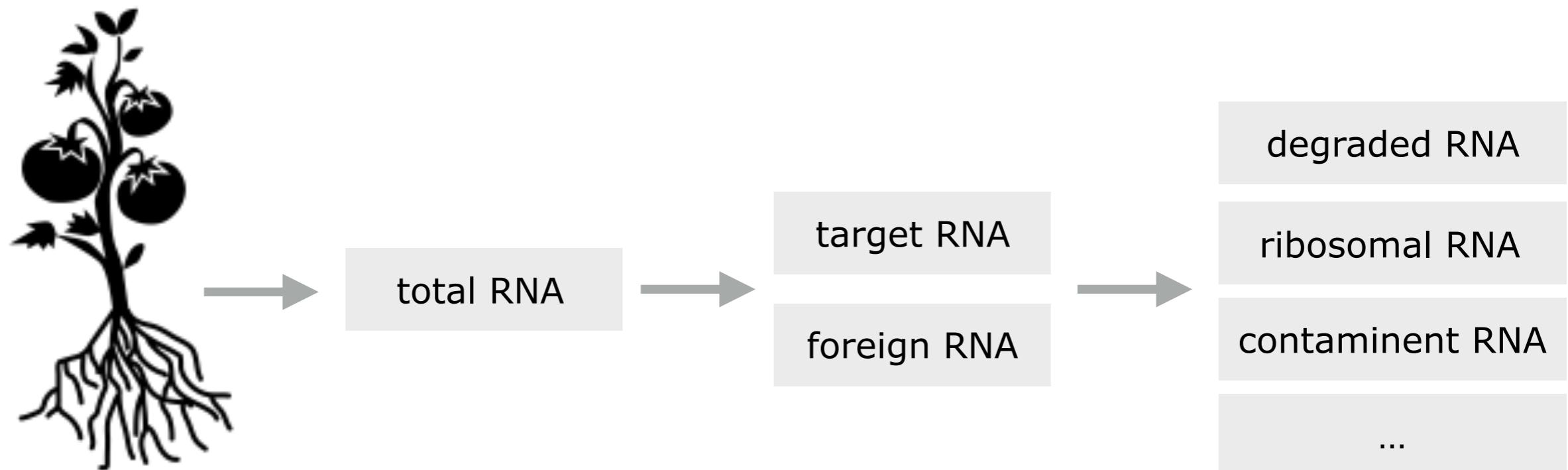
Pooling Samples

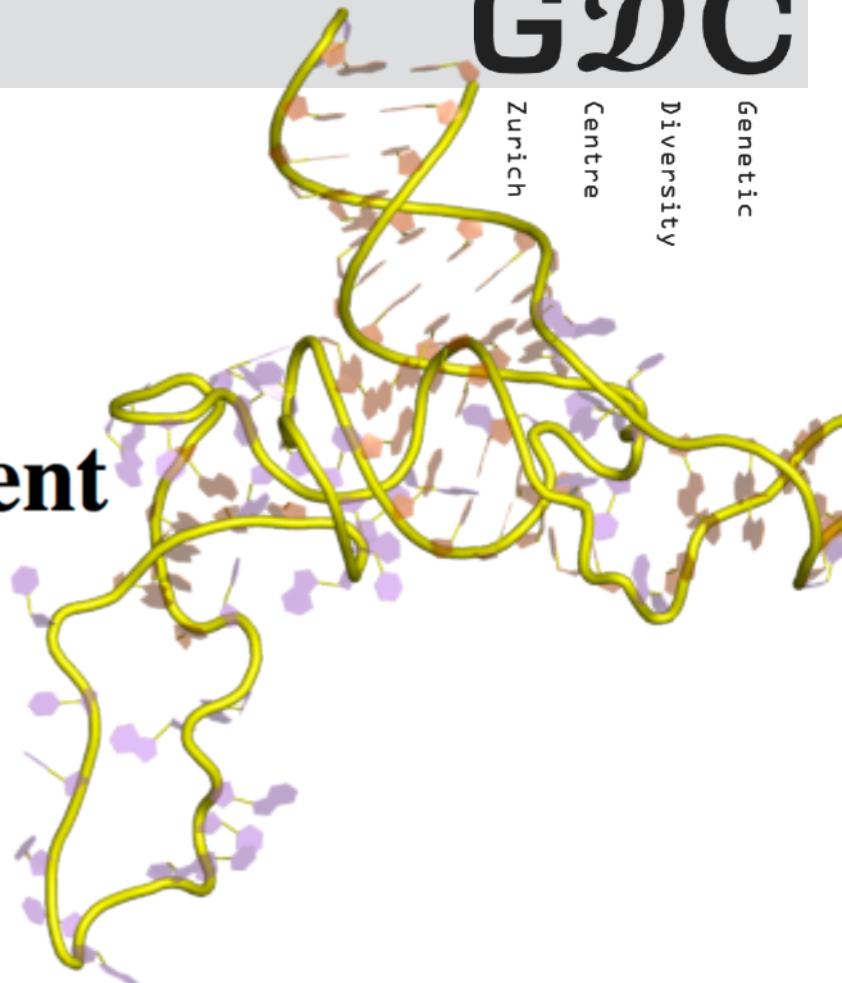


Independent Samples



- amount of starting material
- money
- sequencing depth (coverage)
- question
- sample





Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity

Dominic O'Neil,¹ Heike Glowatz,¹ and Martin Schlumpberger¹

¹Qiagen, Hilden, Germany

ABSTRACT

Ribosomal RNA (rRNA) is the most highly abundant component of RNA, comprising the majority (>80% to 90%) of the molecules present in a total RNA sample. Depletion of this rRNA fraction is desirable prior to performing an RNA-seq reaction, so that sequencing capacity can be focused on more informative parts of the transcriptome. This unit describes an rRNA depletion method based on selective hybridization of oligonucleotides to rRNA, recognition with a hybrid-specific antibody, and removal of the antibody-hybrid complex on magnetic beads. *Curr. Protoc. Mol. Biol.* 103:4.19.1-4.19.8. © 2013 by John Wiley & Sons, Inc.

Keywords: rRNA depletion • sample preparation • RNA-seq • next generation sequencing • transcriptome

MOLECULAR ECOLOGY

Molecular Ecology (2015) 24, 710–725

doi: 10.1111/mec.13055

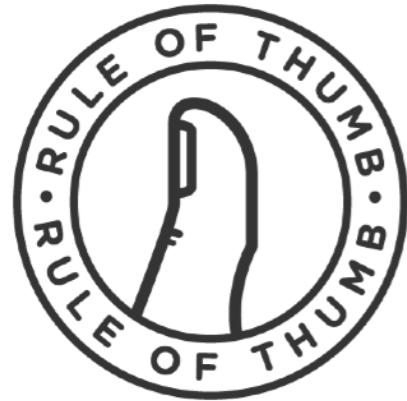
INVITED REVIEWS AND SYNTHESES

Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution?

MARIANO ALVAREZ,* AARON W. SCHREY† and CHRISTINA L. RICHARDS*

**Department of Integrative Biology, University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620, USA, †Department of Biology, Science Center, Armstrong State University, 11935 Abercorn Street, Savannah, GA 31419, USA*

“Ideally, biological validation of gene function uses independent biological samples to confirm the up- or down-regulation of genes in response to a given treatment or condition of interest. Therefore, although we did not include studies that relied solely on qPCR in our survey of transcriptomics, the use of **qPCR for confirmation of the expression of genes of interest is essential.**”



1. CLEAR SCIENTIFIC QUESTION - EXPRESSION DIFFERENCE
2. SAMPLE QUALITY AND STRINGENT QC MEASURES
3. RIBOSOMAL REMOVAL
4. USE SPIKE-IN CONTROLS (External RNA Controls Consortium - ERCC)
5. ALIGN TO THE GENE SET (TRANSCRIPTOM) AND GENOME
6. BIOLOGICAL REPLICATES (MIN 3) - MORE REPLICATES THAN DEPTH
7. 10-20M MAPPED READS PER SAMPLE - MEAN READ DEPTH 10 PER TRANSCRIPT
8. NOISE THRESHOLD AND REDUCTION
9. PILOT SEQUENCING EXPERIMENTS > *DE NOVO* ASSEMBLY