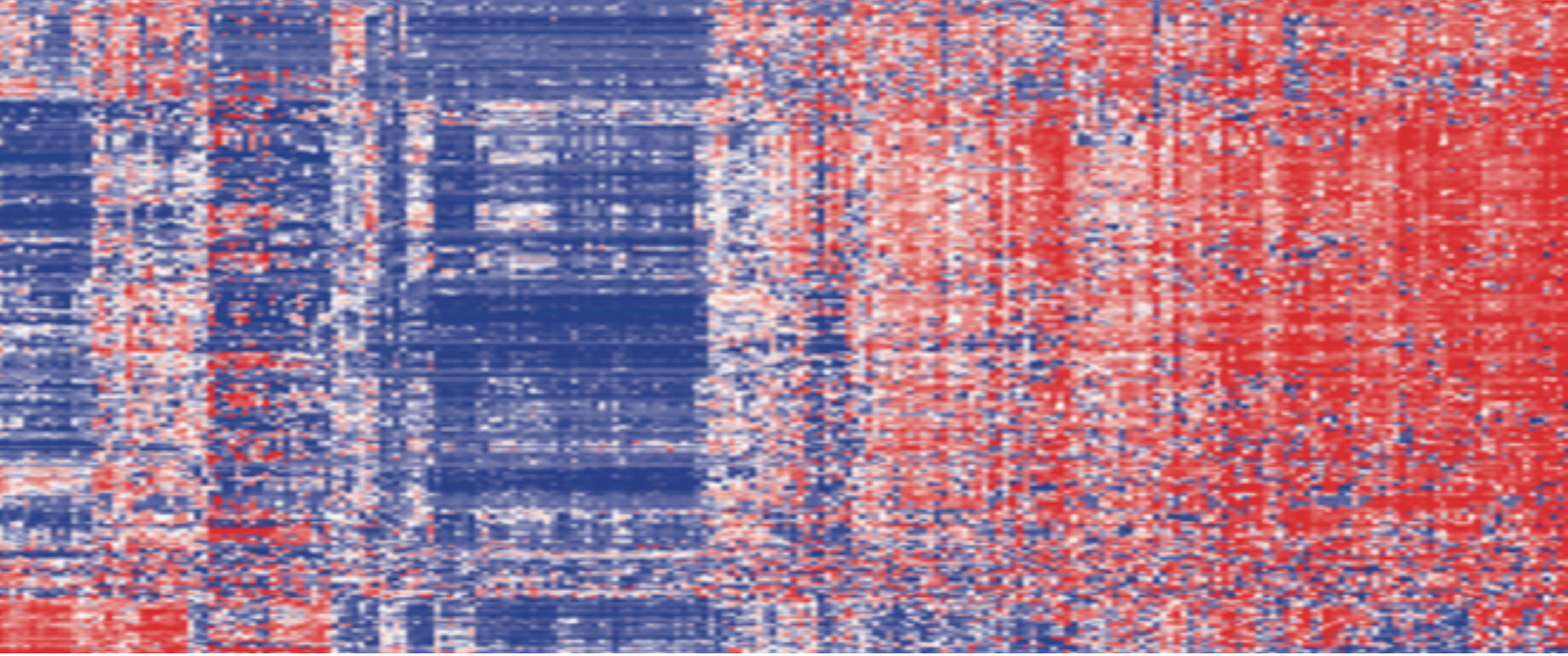


PROSPECTS OF RNA-SEQ

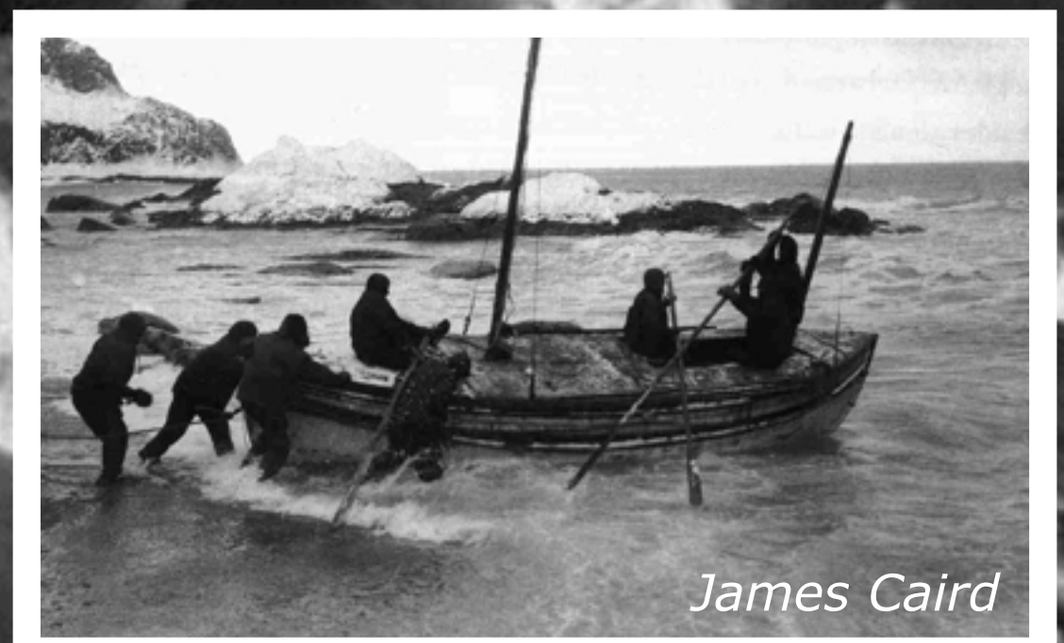
Jean-Claude Walser

GDC

Genetic Diversity Centre

Zurich

http://www.gdc.ethz.ch



James Caird

Aug 30, 1916
Yelcho ship reaches Elephant Island and rescues the rest of the crew

May 19-20, 1916
Shackleton, Worsley and Crean cross the mountains and reach Stromness whaling station

May 10, 1916
Arrival to South Georgia
54°15'S, 36°45'W

Apr 24, 1916
Shackleton's departure on James Caird

Apr 15, 1916
Landfall on Elephant Island
61°08'S, 55°07'W

Apr 9, 1916
Launched boats for Elephant Island

1916

Dic 21, 1915
Patience Camp established

Nov 21, 1915
Endurance sank

Oct 30, 1915
Crew starts the march

Oct 27, 1915
Endurance crushed
69°05'S, 51°30'W

Jan 18, 1915
Endurance trapped in ice
76°58'S, 31°30'W

1915

100km 500km

STROMNESS BAY

SOUTH GEORGIA

Die 6,85 Meter lange *James Caird* wurde am 24. April 1916 zu Wasser gelassen. Für die folgende Fahrt hing alles von der absolut exakten Navigation Worsleys ab, die auf Beobachtungen basierten, die unter den ungeeignetsten Bedingungen gemacht werden mussten.

ELEPHANT ISLAND

WEDDEL SEA

RONNE ICE SHELF

60°

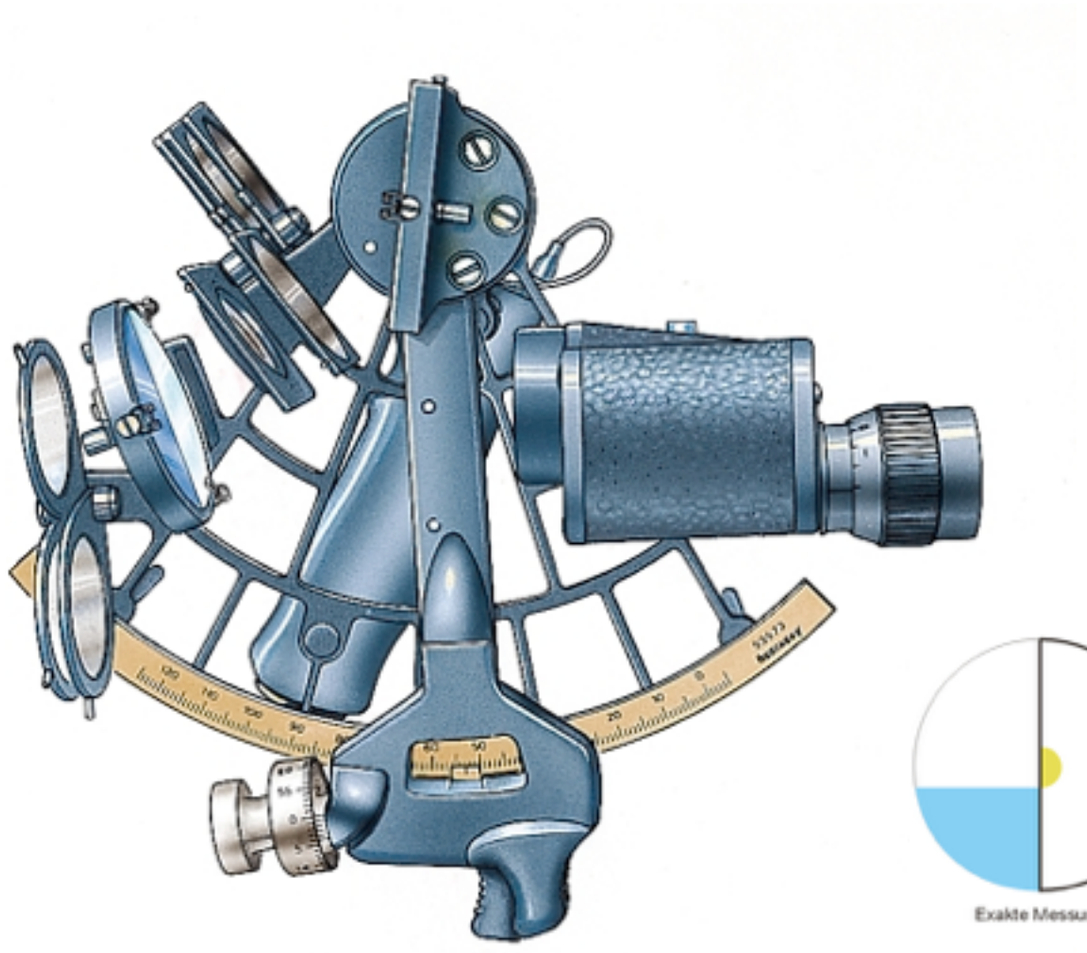
70°

80°

Circle

IMPERIAL TRANS-ANTARCTIC EXPEDITION

FROM *ENDURANCE* BESET TO FINAL RESCUE



A **sextant** is a doubly reflecting navigation instrument that measures the angle between two visible objects. The primary use of a sextant is to measure the angle between an **astronomical object and the horizon** for the purposes of celestial navigation (astronavigation).

Signal-to-noise ratio

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

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

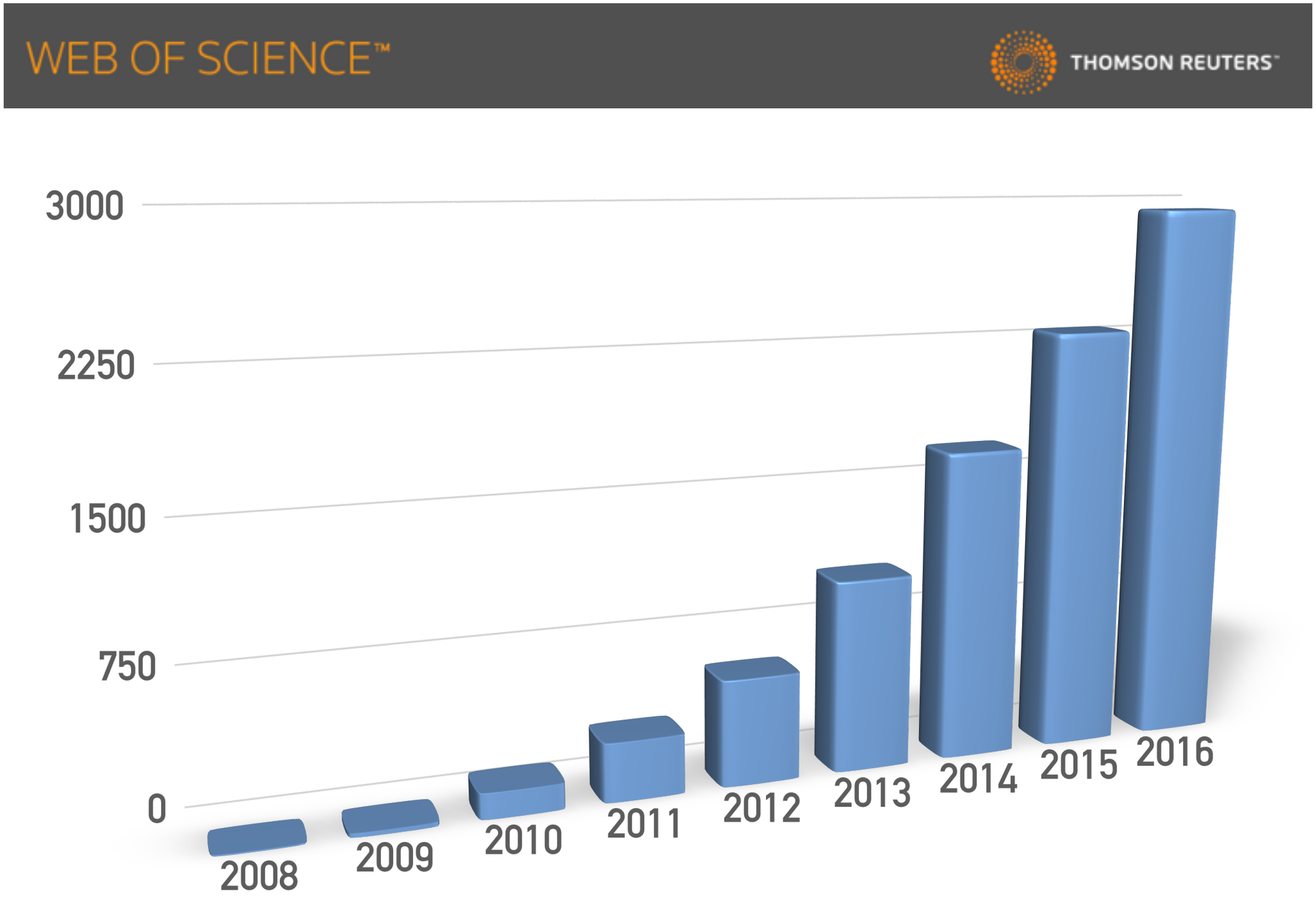
Non-Poisson technical variance - The observed imprecision between repeat measurements.

Biological variance - The natural variation in gene expression measurements.

What is RNA-SEQ?

http://www.rna-seqblog.com

Prospects of RNA-Seq



Database: MEDLINE | Search Term: “RNA-seq" OR “RNAseq” | Search Filed: Topic

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein and Michael Snyder

Abstract | RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10, 57–63.

December 2011

iCommunity NEWSLETTER

Transitioning from Microarrays to mRNA-Seq

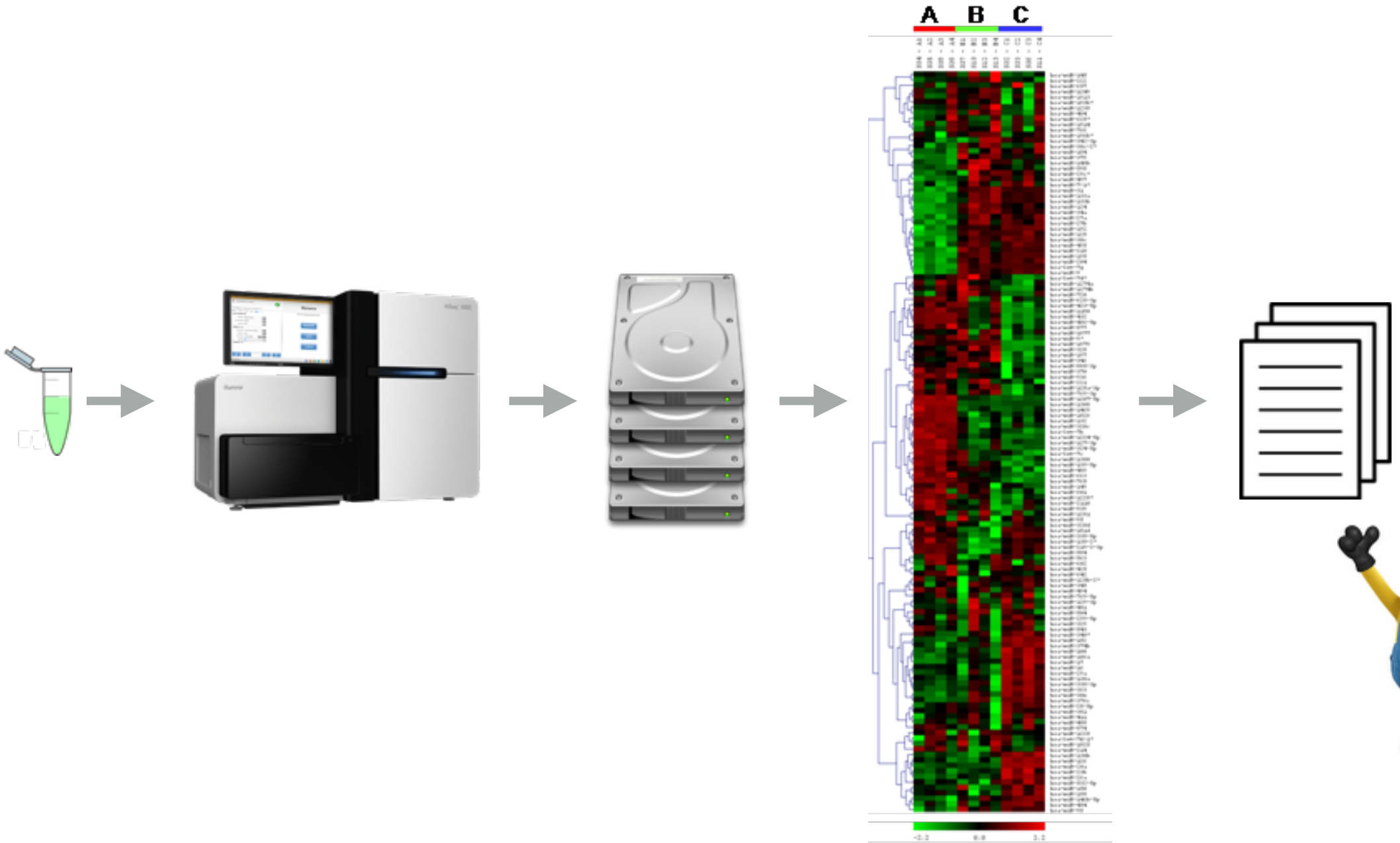
By providing HiSeq® 2000 mRNA-Seq data in two formats, Expression Analysis enables its customers to compare sequencing results with older array data, while gaining insight into the entire transcriptome.

Transition

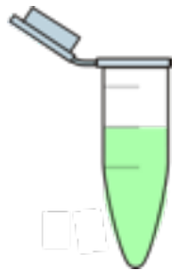
Intensity ▷ Count

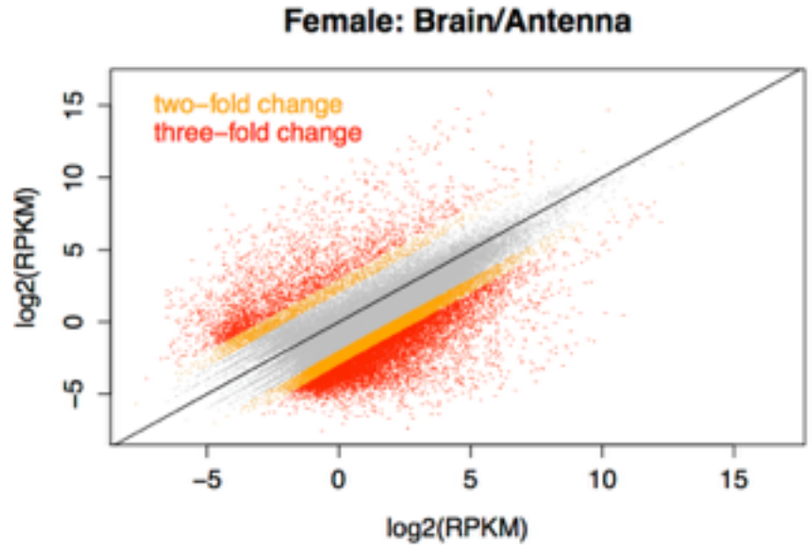
http://www.illumina.com/content/dam/illumina-marketing/documents/icommunity/article_2011_12_ea_rna-seq.pdf

Prospects of RNA-Seq



Prospects of RNA-Seq





Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

CrossMark click for updates

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

	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 at what quality?

Resources

Design | Preparation | Methode | Analysis | Extras

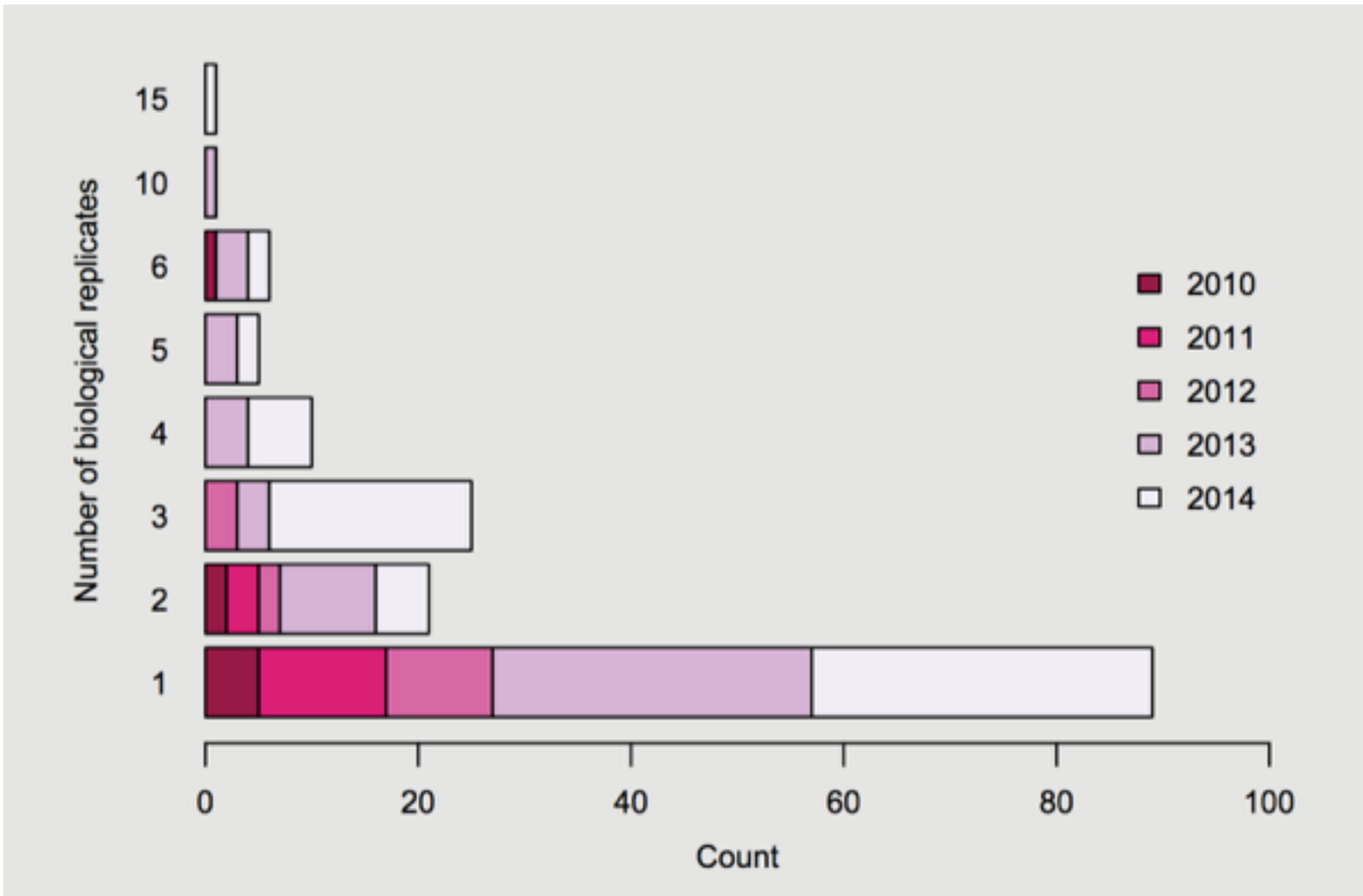
What factors influence the design?

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

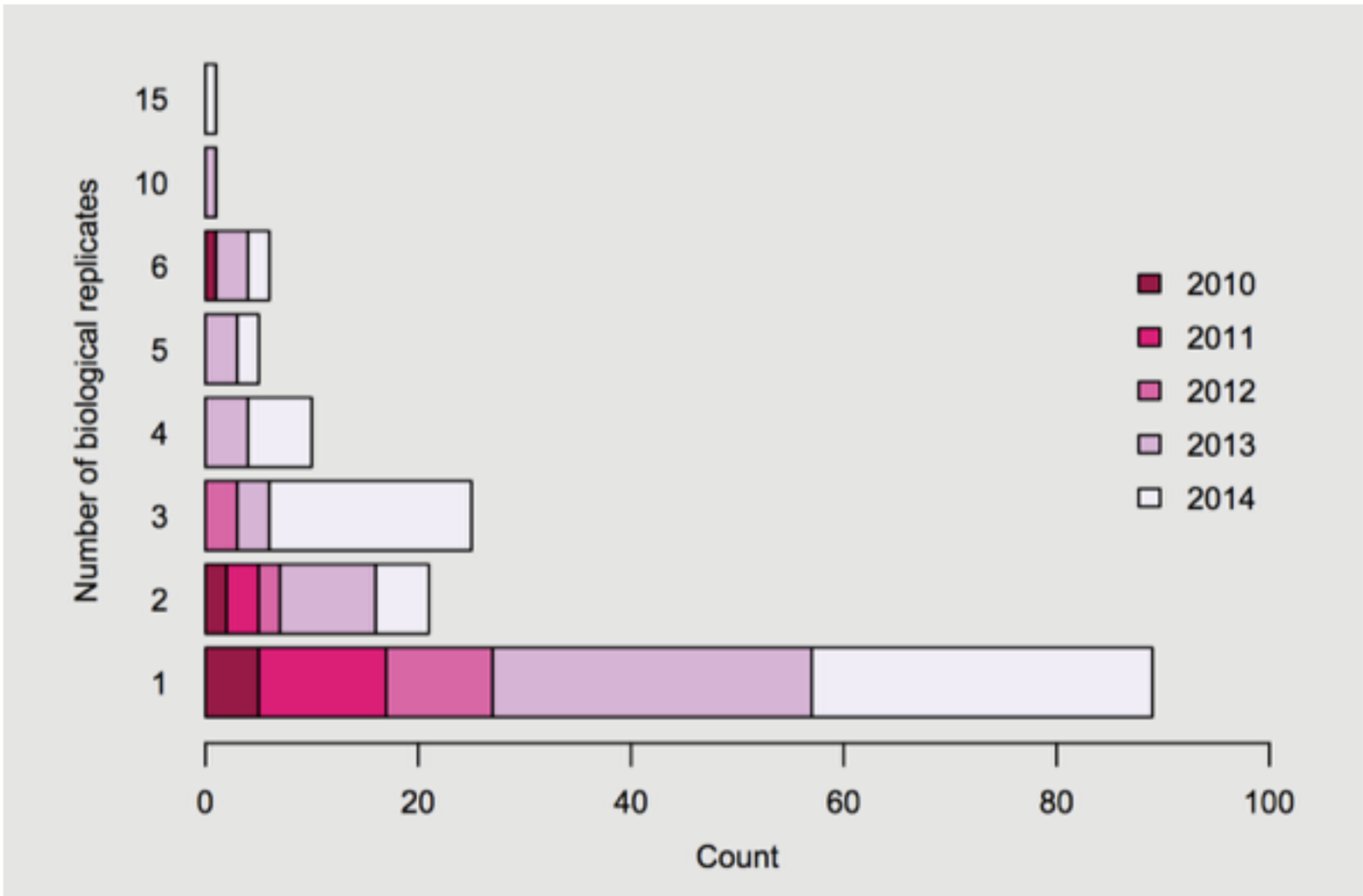
What (min) depth of sequencing **coverage** is required?

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

How much **money** do we have?



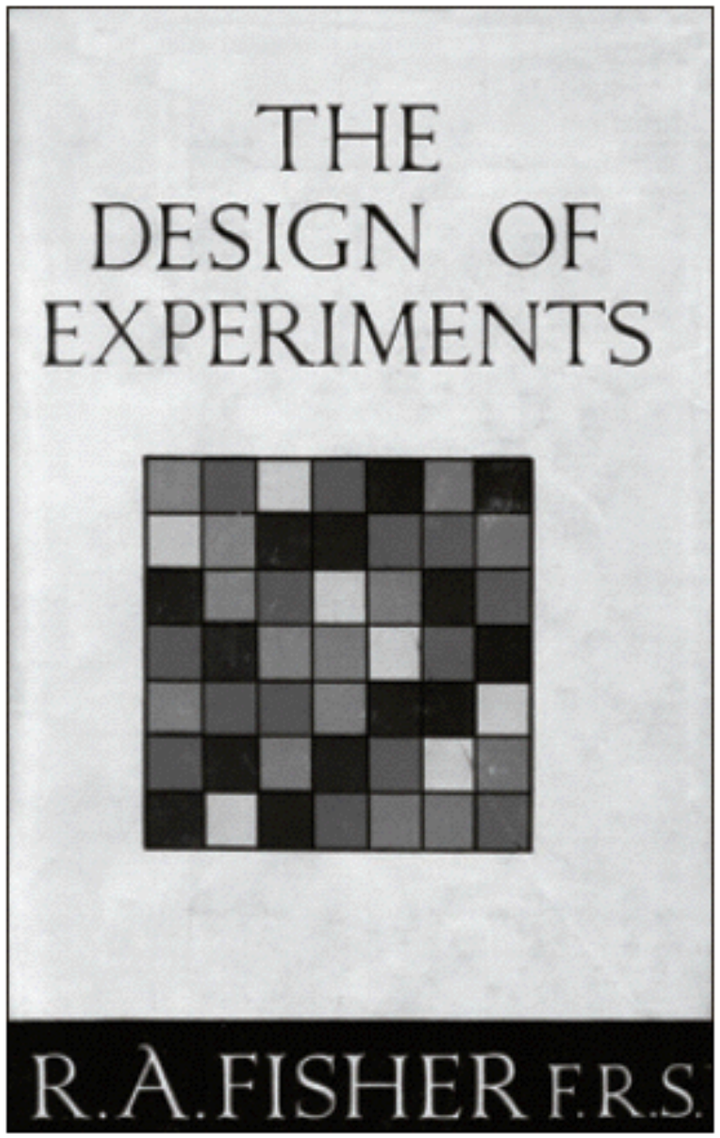
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.



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.

DOI: 10.1534/genetics.110.114983

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge[1]

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-Alcalde [1], Joaquín Dopazo[1], Alberto Ferrer[2], 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,*]

[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

"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
2x10M (20M) PE-reads > 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]

[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

"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.

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)

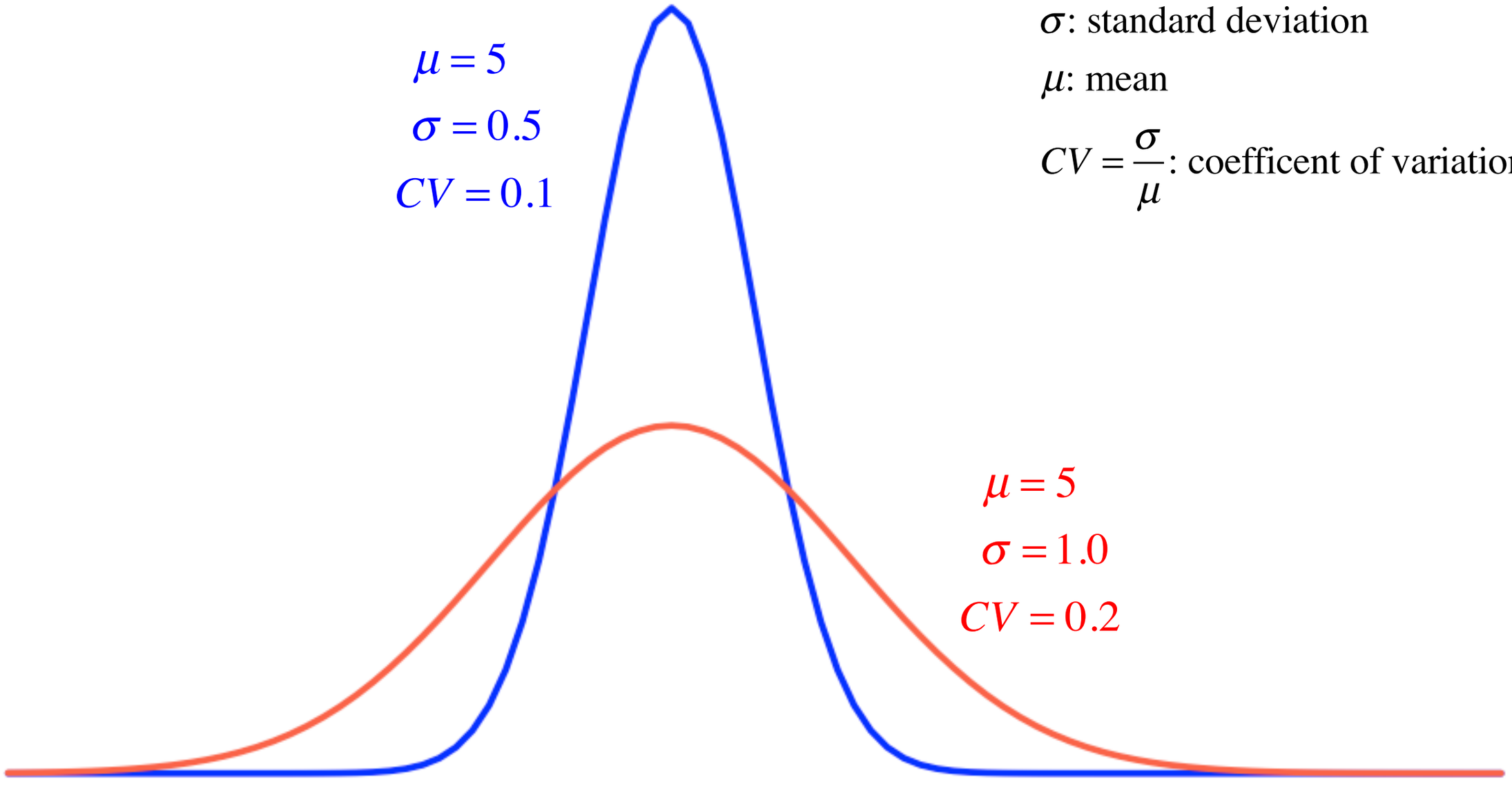
Given any three, we can determine the fourth.

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

$\mu = 5$

$\sigma = 0.5$

$CV = 0.1$



σ: standard deviation

μ: mean

$CV = \frac{\sigma}{\mu}$: coefficent of variation

Inbred vs. **Wild**

Model Organsims vs. **Non-Model Organsims**

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

CV: coefficent of variation

σ: standard deviation

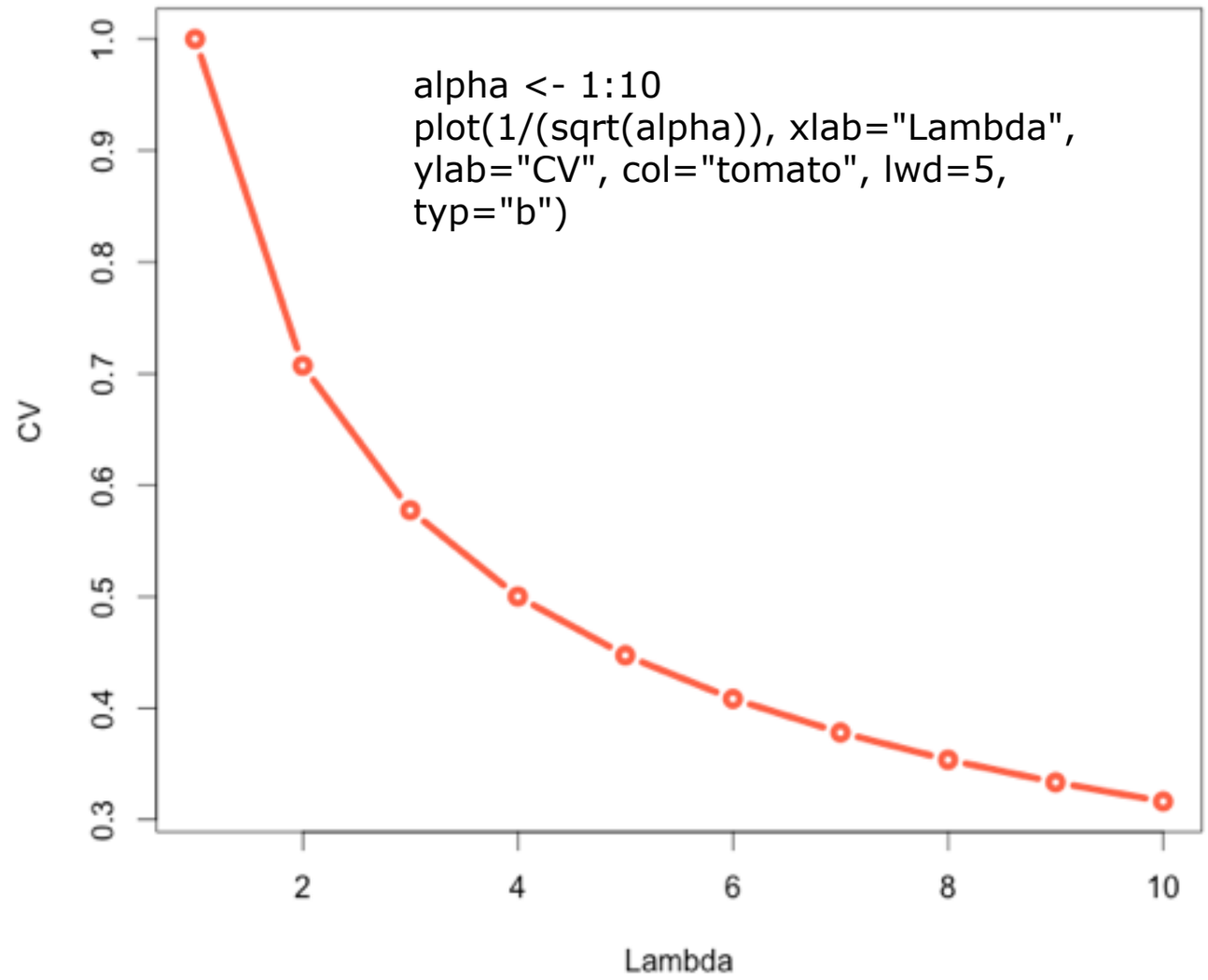
μ: mean

inbred animal strains: $CV \leq 0.2$

unrelated individuals: $CV > 0.3$

Poisson Distribution

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



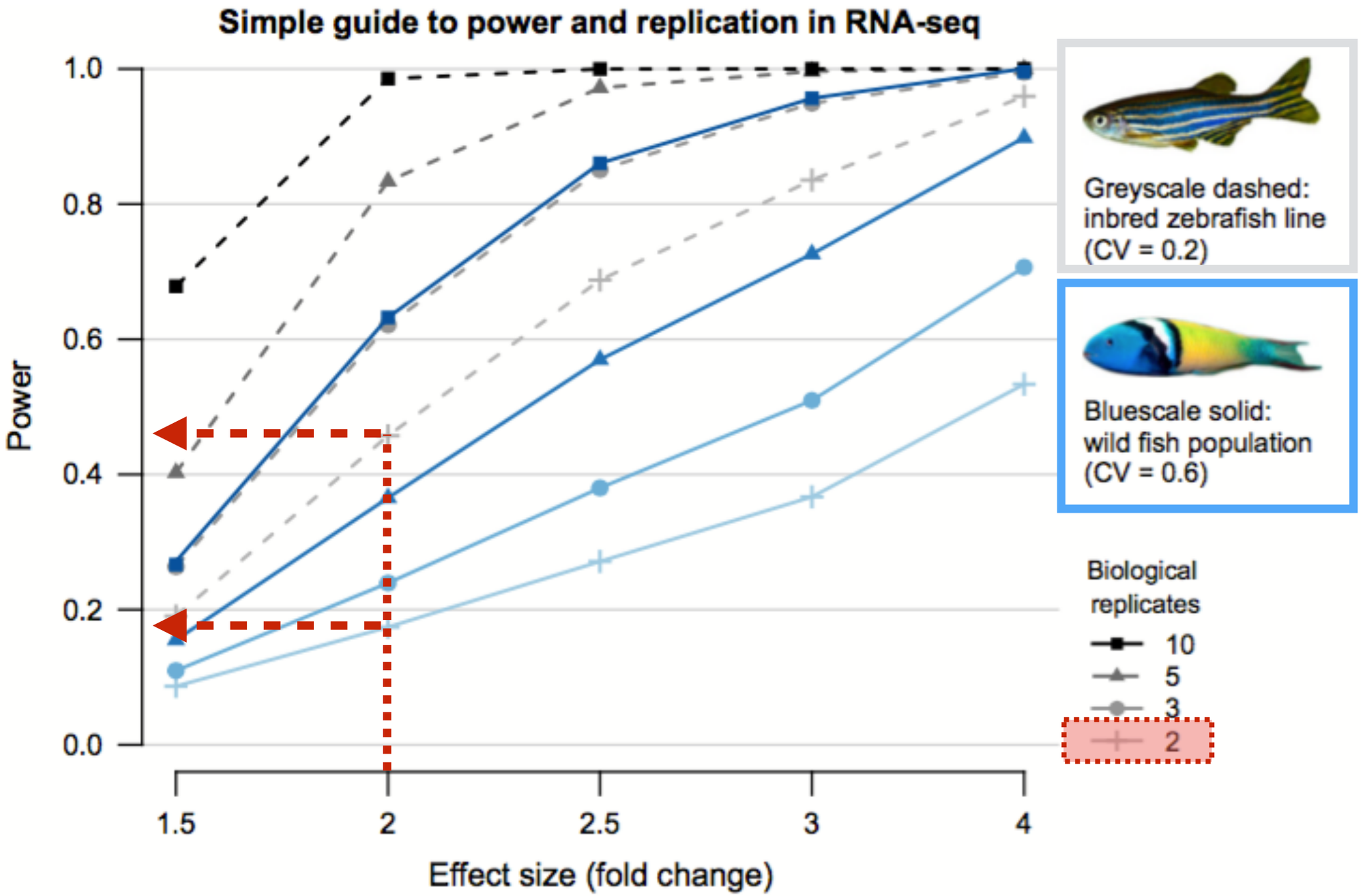
CV: coefficent of variation

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

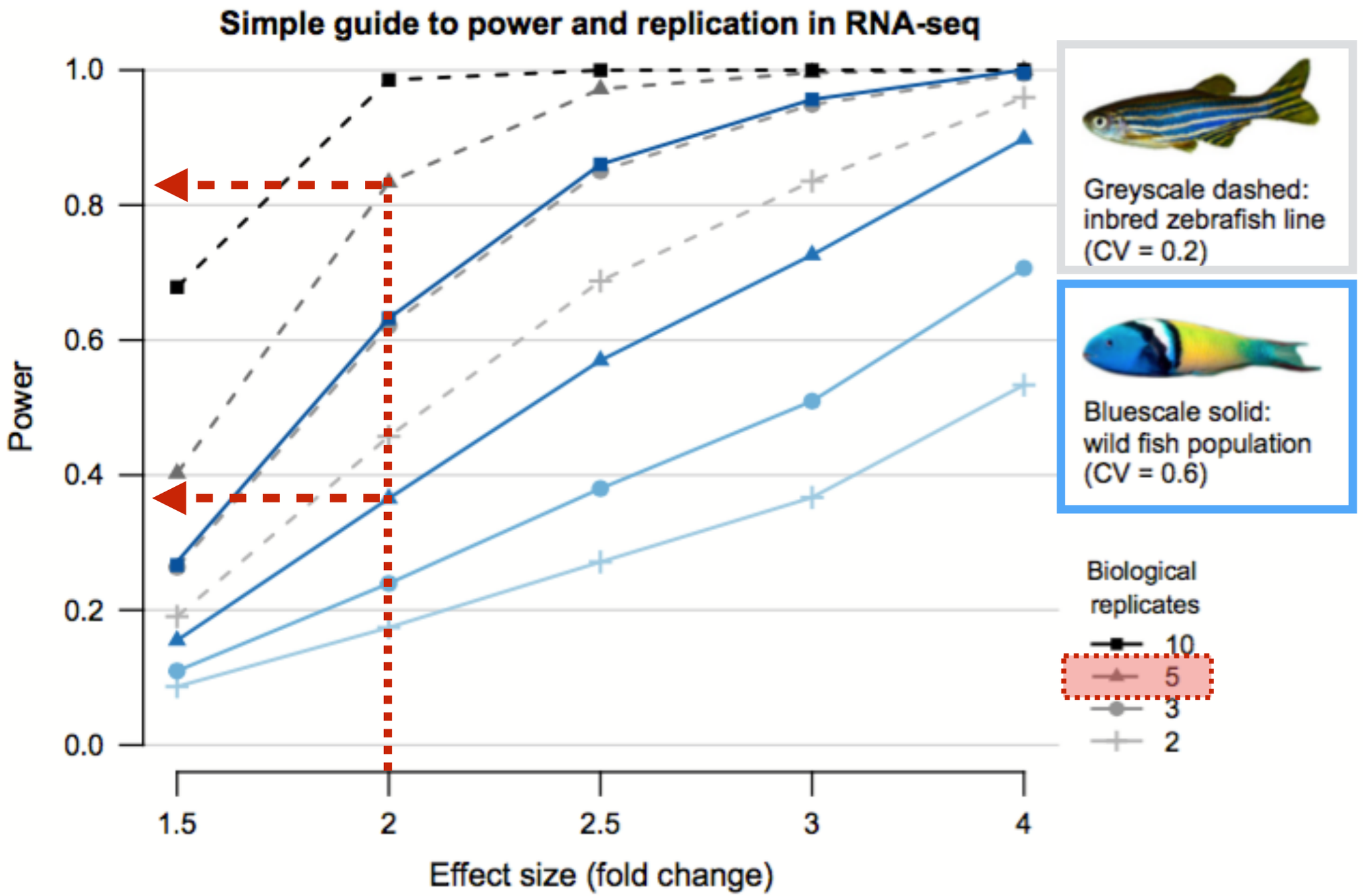
σ: standard deviation (= $\sqrt{Variance}$)

μ: mean

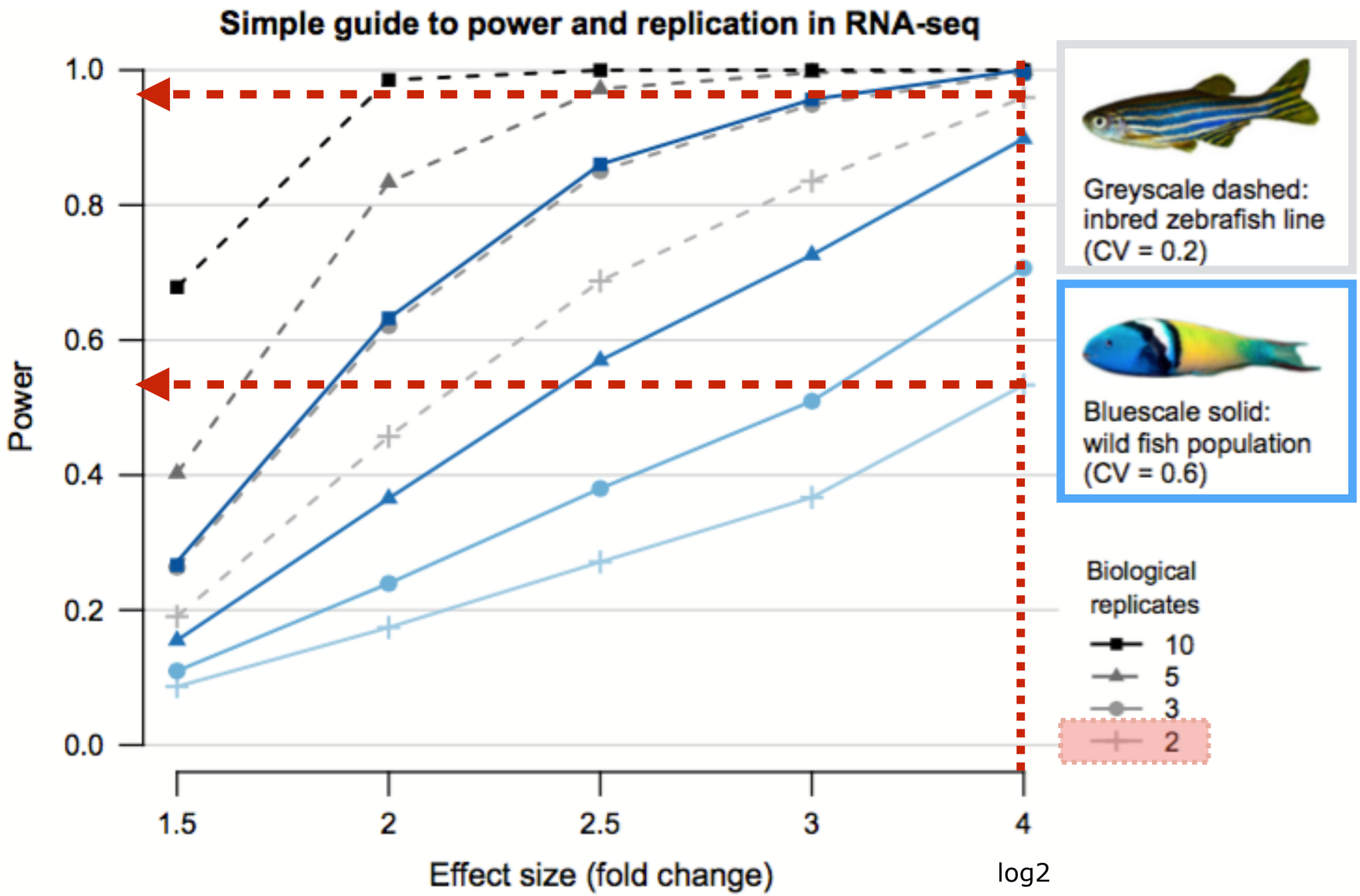
! The expected value and variance of a Poisson-distributed random variable are both equal to λ.



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



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



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 − β).
The number of samples **n** in each group.

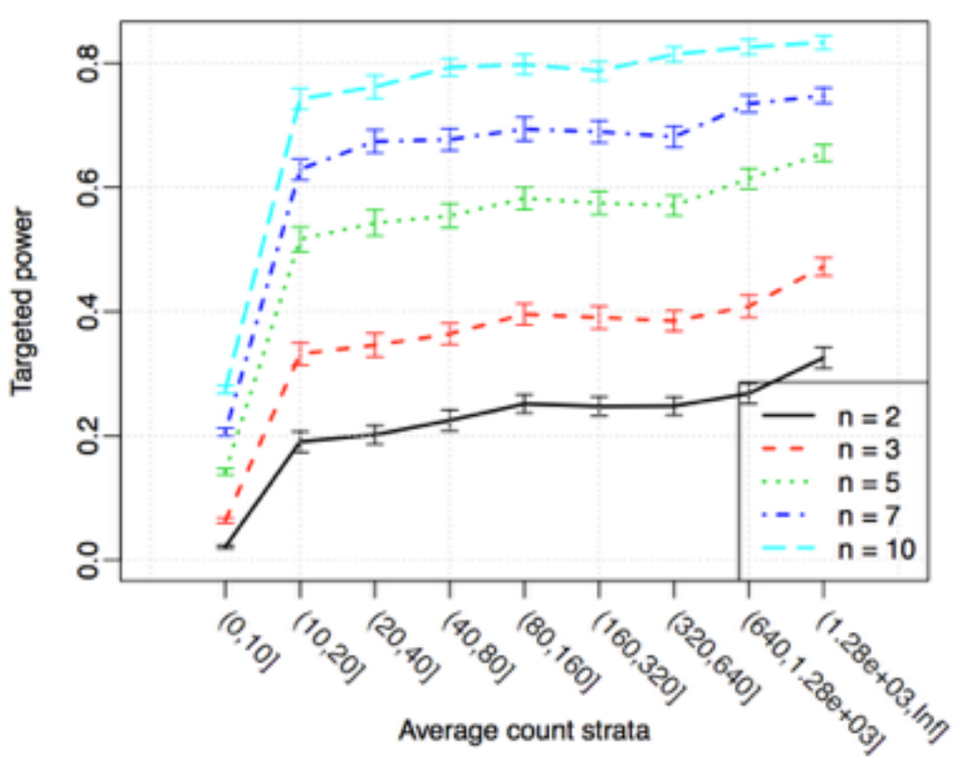
Steven N Hart, Terry M Therneau, Yuji Zhang and Jean-Pierre Kocher (2013) Calculating Sample Size Estimates for RNA Sequencing Data, J Comput Biol. 20(12):970-8

```
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

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

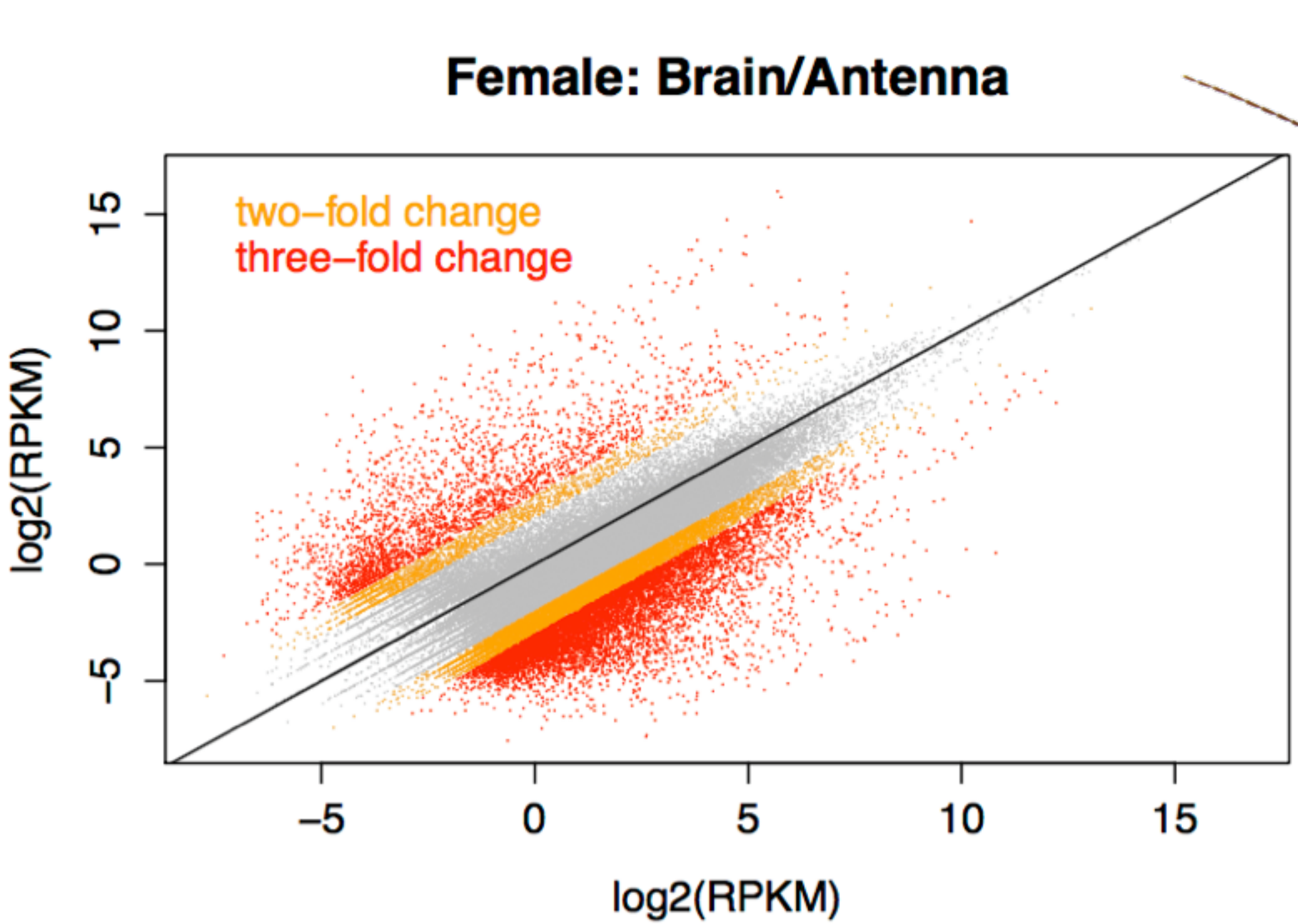
"Ideally, biological validation of gene function uses independent biological samples to confirm the up- or down-regulation of genes in response to a given treatment or condition of interest. Therefore, although we did not include studies that relied solely on qPCR in our survey of transcriptomics, the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

	Sample Design
	Sample preparation RNA extraction Cleaning (e.g. remove ribosomal RNA)
	Library Prep Illumina paired-end sequencing

Pooling Samples

Independent Samples

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

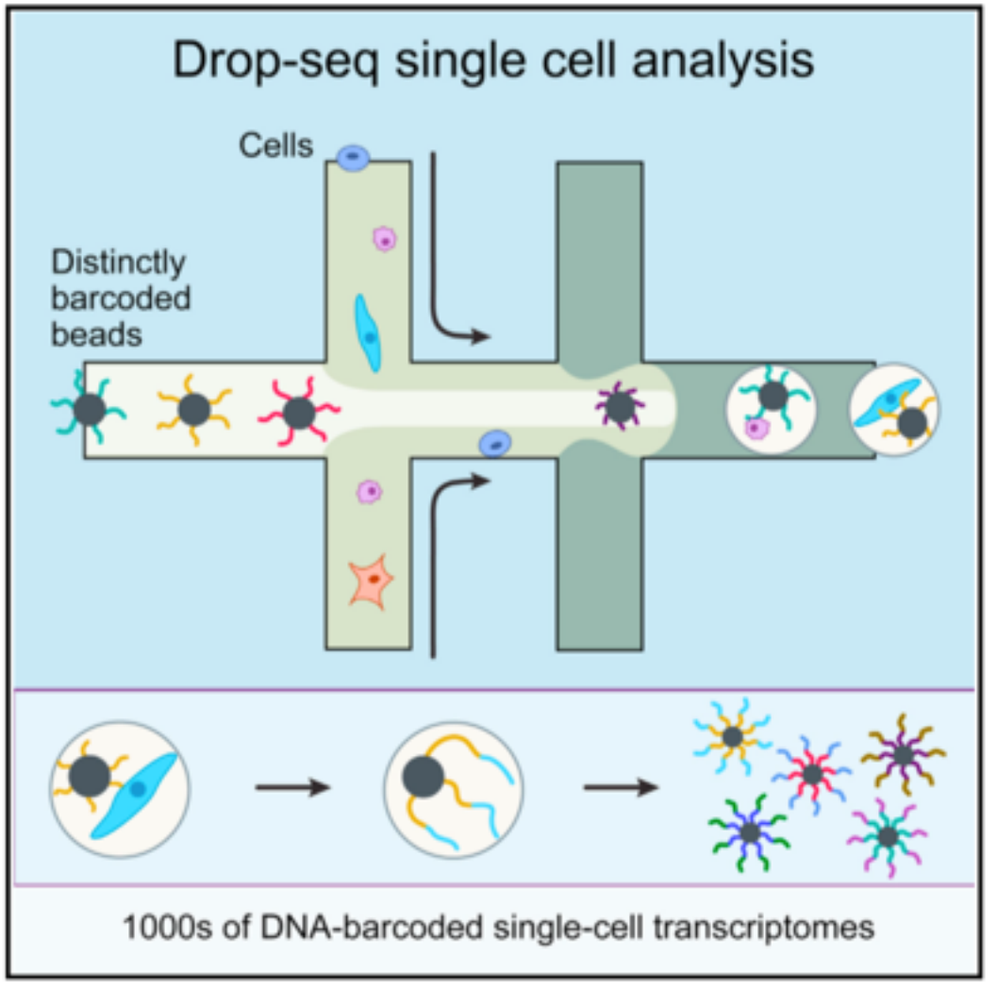

Female: Brain/Antenna
two–fold change
three–fold change
log2(RPKM)
15
10
5
0
−5
−5
0
5
10
15
log2(RPKM)

Resource

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract



Authors

Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence

emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief

Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

For transcriptome-based studies, RNA-seq libraries are generated by the synthesis of double stranded cDNA followed by the addition of sequencing adapters. This method however, does not retain any information about the DNA strand from which the RNA was transcribed. It is often desirable to create libraries that retain the strand orientation of the original RNA targets. For example, in some cases transcription creates anti-sense RNA constructs that may play a role in regulating gene expression.

Head et al. (2014). Library construction for next-generation sequencing: Overviews and challenges. BioTechniques, 56(2).

total RNA

target RNA

foreign RNA

degraded RNA

ribosomal RNA

contaminent RNA

...

Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity

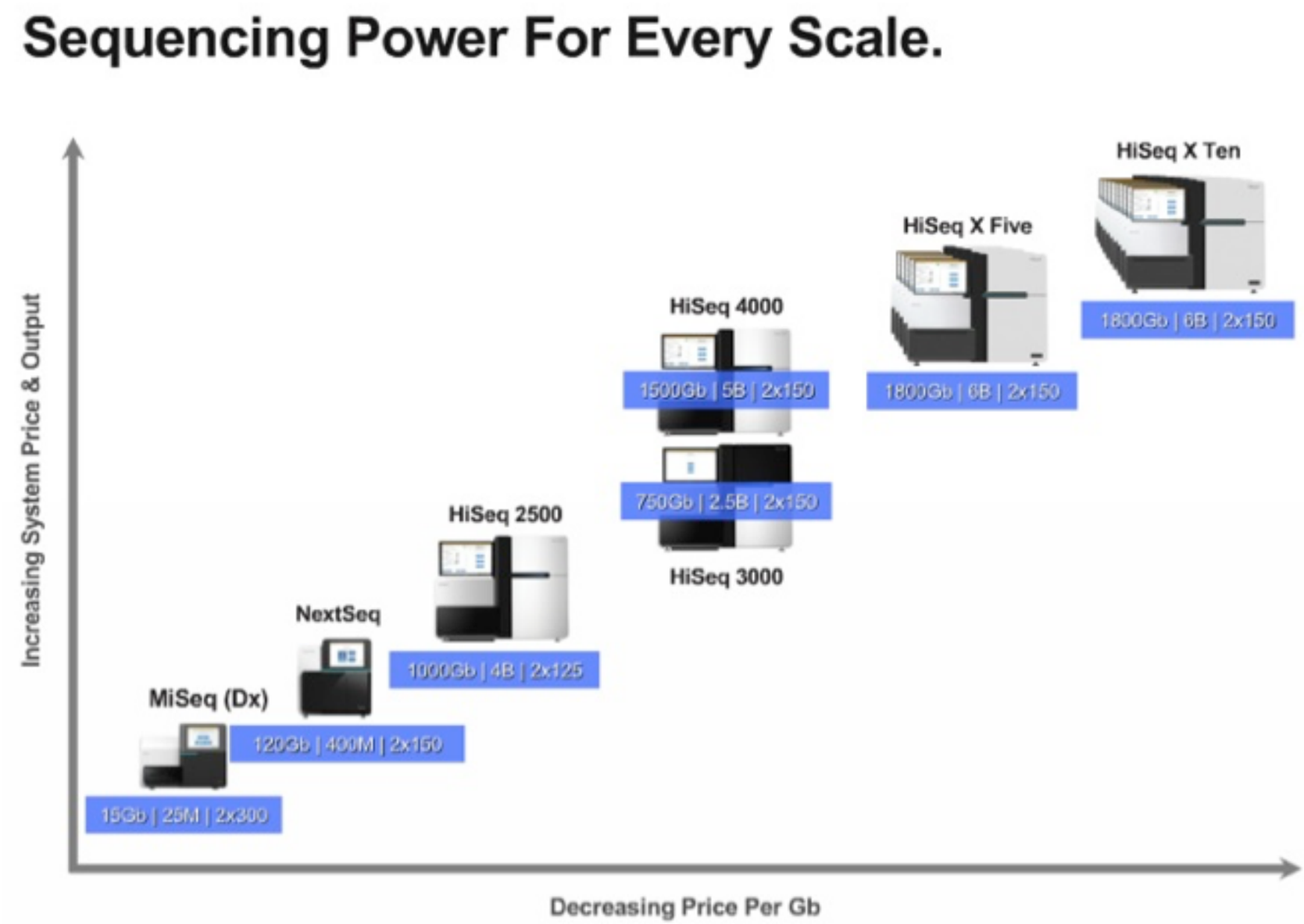
Dominic O'Neil,[1] Heike Glowatz,[1] and Martin Schlumpberger[1]

[1]Qiagen, 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.

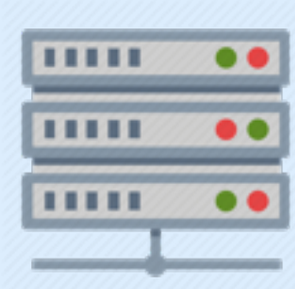
Keywords: rRNA depletion • sample preparation • RNA-seq • next generation sequencing • transcriptome


Sequencing Power For Every Scale.
Increasing System Price & Output
Decreasing Price Per Gb
HiSeq X Ten
1800Gb | 6B | 2x150
HiSeq X Five
1800Gb | 6B | 2x150
HiSeq 4000
1500Gb | 5B | 2x150
750Gb | 2.5B | 2x150
HiSeq 3000
HiSeq 2500
1000Gb | 4B | 2x125
NextSeq
120Gb | 400M | 2x150
MiSeq (Dx)
15Gb | 25M | 2x300

Prospects of RNA-Seq

200-250 CHF per Library

2200 CHF per Lane PE100 HiSeq 2500 (200 Mio Reads)

	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

Briefings in Bioinformatics Advance Access published December 2, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 12 doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

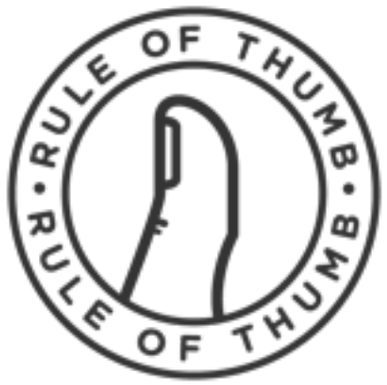
Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Submitted: 20th August 2013; Received (in revised form): 9th October 2013

Abstract

RNA-sequencing (RNA-seq) has rapidly become a popular tool to characterize transcriptomes. A fundamental research problem in many RNA-seq studies is the identification of reliable molecular markers that show differential expression between distinct sample groups. Together with the growing popularity of RNA-seq, a number of data analysis methods and pipelines have already been developed for this task. Currently, however, there is no clear consensus about the best practices yet, which makes the choice of an appropriate method a daunting task especially for a basic user without a strong statistical or computational background. To assist the choice, we perform here a systematic comparison of eight widely used software packages and pipelines for detecting differential expression between sample groups in a practical research setting and provide general guidelines for choosing a robust pipeline. In general, our results demonstrate how the data analysis tool utilized can markedly affect the outcome of the data analysis, highlighting the importance of this choice.

Keywords: *RNA-seq; gene expression; differential expression*



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