Genetic Diversity Centre Zurich

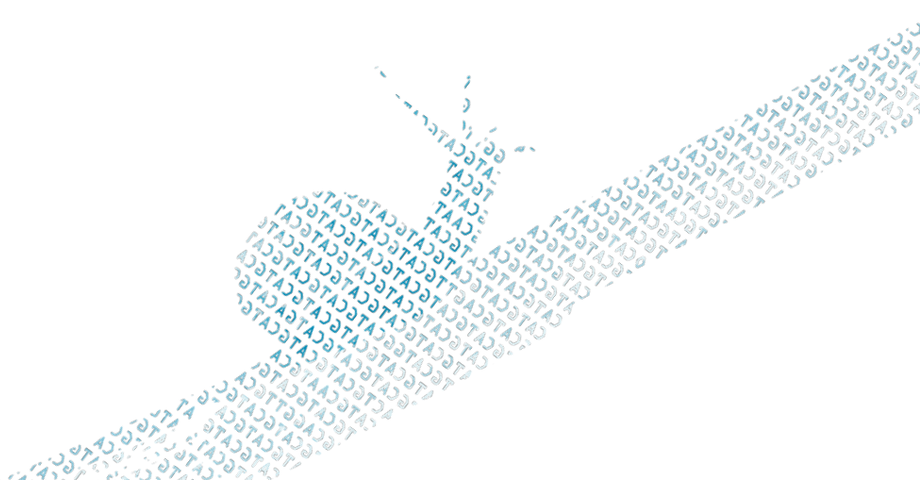
701-1425-00L - Genetic Diversity: Analysis

RNA-Seq

Wednesday, June 22, 2020

Jean-Claude Walser

jean-claude.walser@env.ethz.ch



GDC Genetic Diversity Centre Zurich

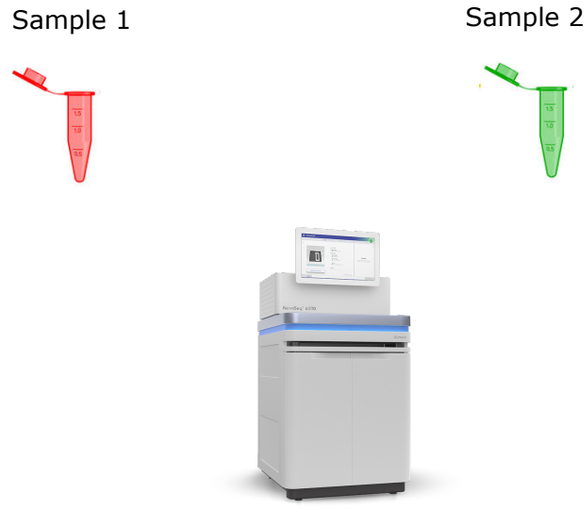
RNA-Seq is a comprehensive high-throughput sequencing approach for the **quantitative** (and **qualitative** analysis) of transcriptomes of model and **non-model organisms**.

GDC Genetic Diversity Centre Zurich

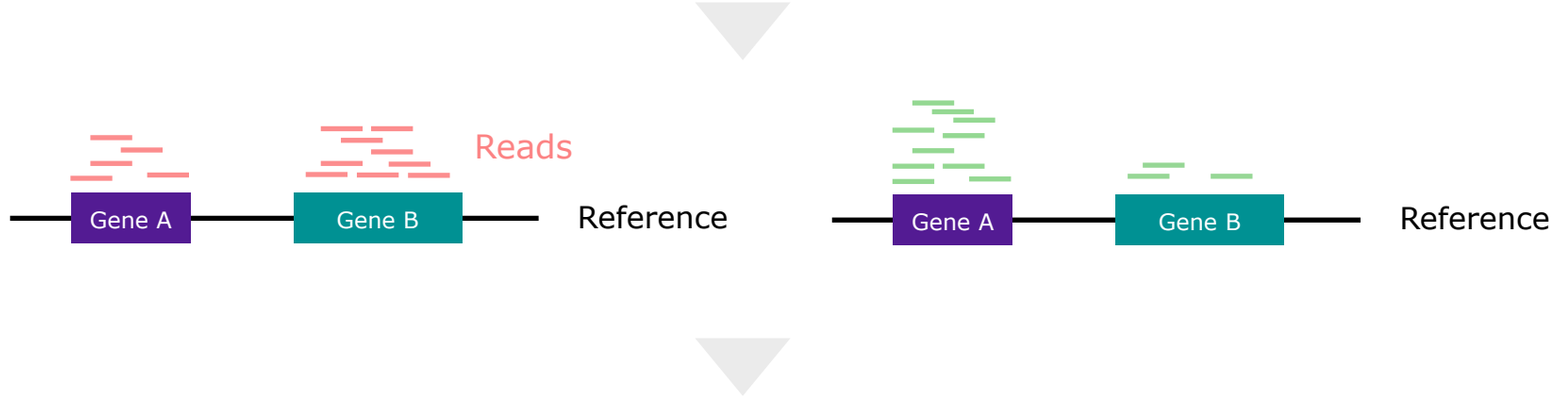
The Idea behind DEGs

- mRNA Isolation
- cDNA
- Library prep

- Sequencing



- Mapping



- Counts

Gene A: $\frac{5}{10} = 0.5$ fold change Gene B: $\frac{9}{3} = 3$ fold change

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

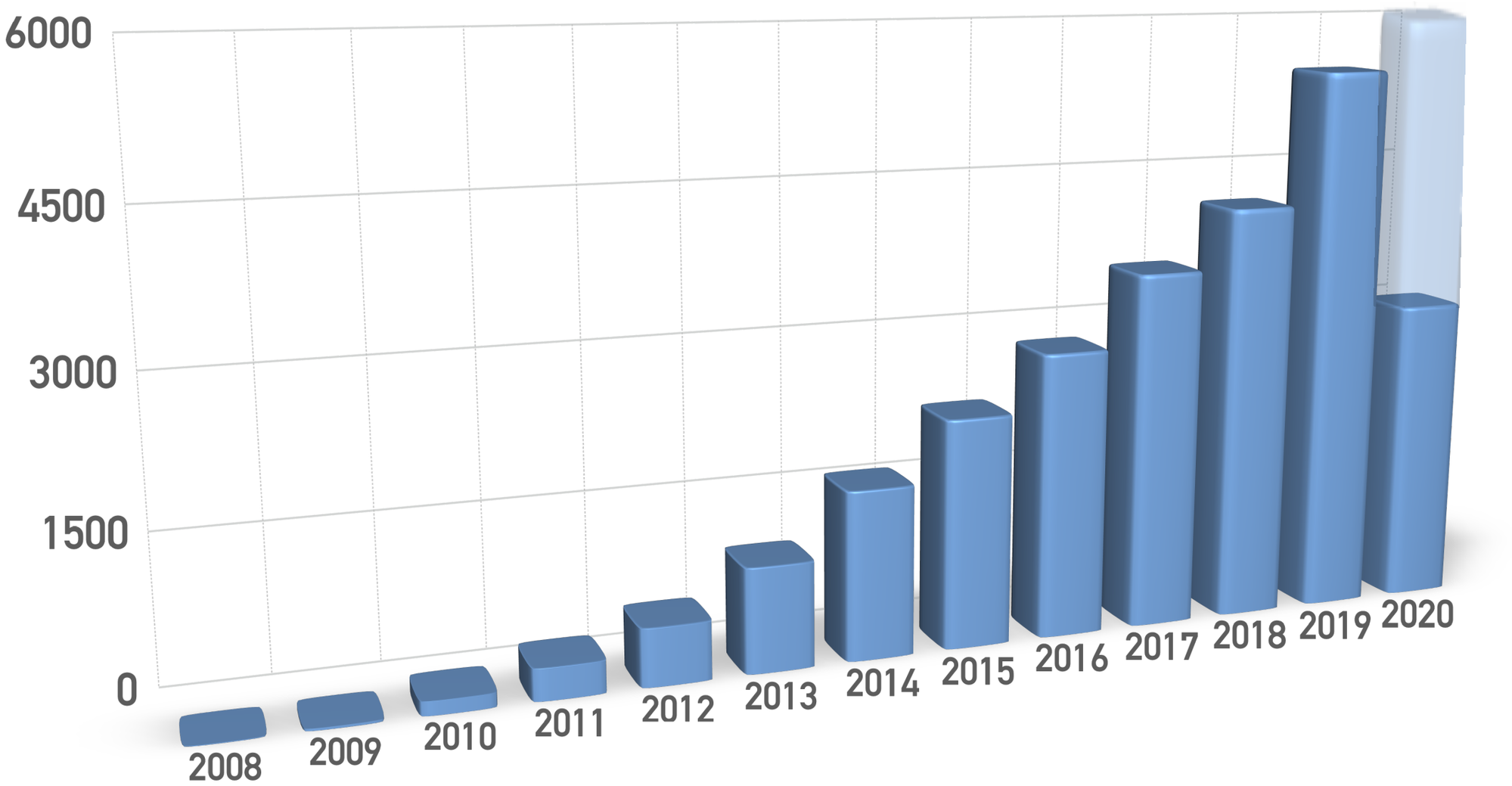
Zhong Wang, Mark Gerstein and Michael Snyder

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.

Genetic Diversity Centre Zurich

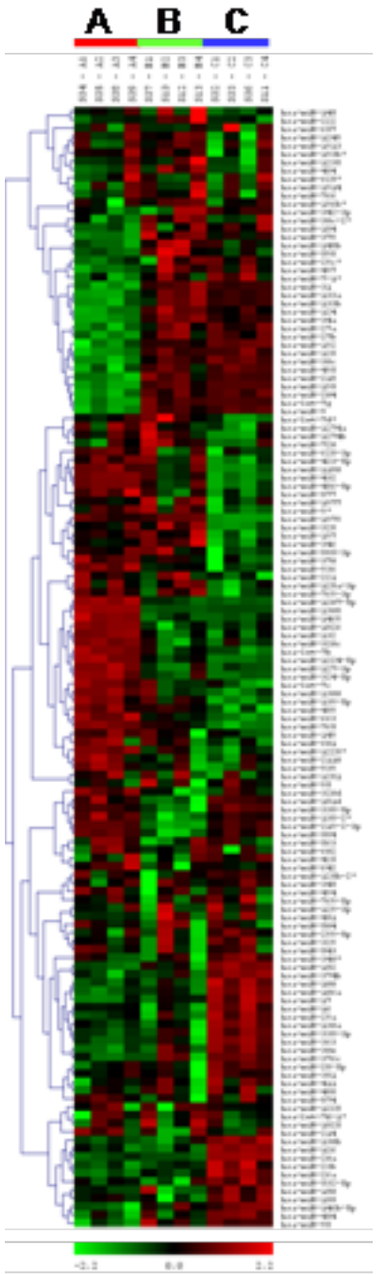
WEB OF SCIENCE™ THOMSON REUTERS™

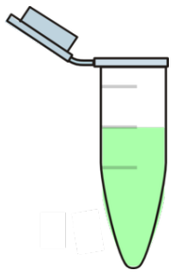


Database: MEDLINE | Search Term: "RNA-[Ss]eq" OR "RNA[Ss]eq" | Search Filed: Topic

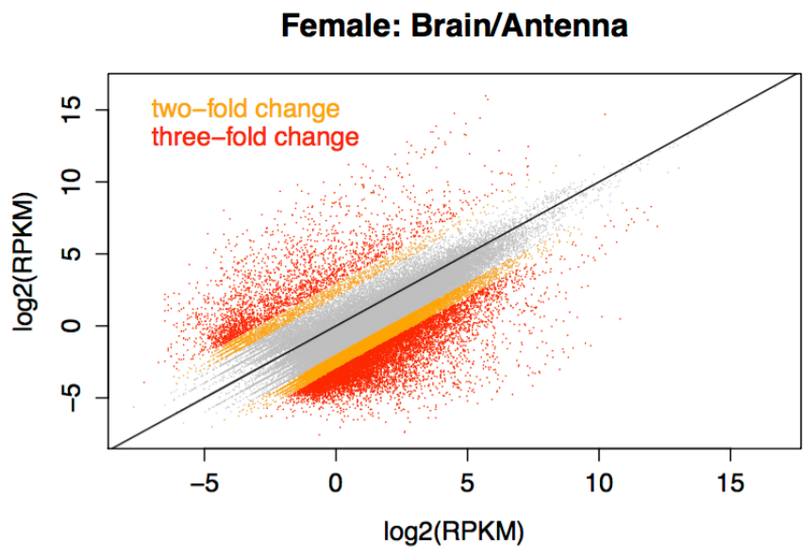
Genetic Diversity Centre Zurich

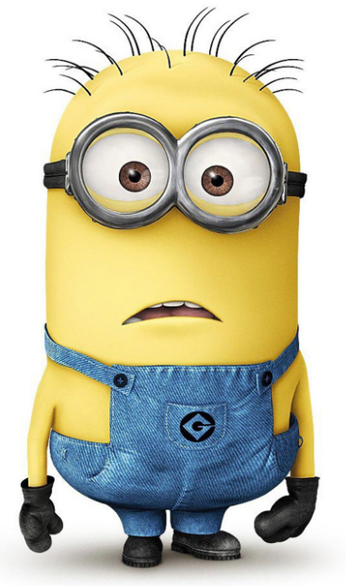
NovaSeq 6000
2 x 150 bp
800M - 1.6B rpr



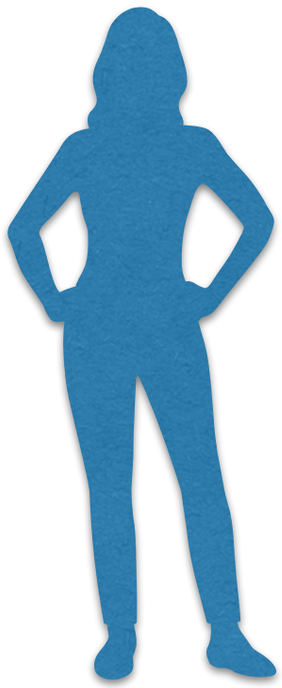



NovaSeq™ 6000


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



What do we know about **our** own transcriptome?

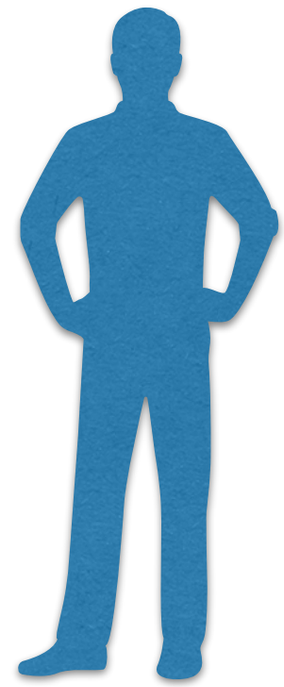


Number of (well-)validated genes: ?

Percentage of genes not encoding proteins: ?

Percentage of alternative splicing: ?

Average alternative transcribed forms: ?



RNA-Seq Blog

Transcriptome Sequencing Research & Industry News

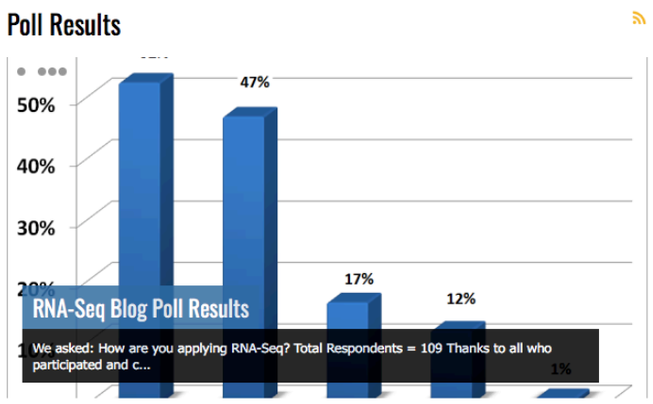
http://www.rna-seqblog.com

Researchers use AI, RNA-Seq to unlock the secrets of the genome

December 28, 2017 Leave a comment 2,576 Views

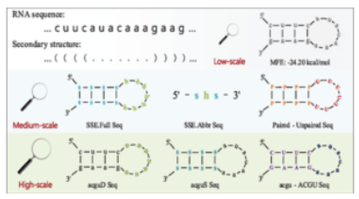
Every nine minutes, someone in the US dies from blood cancer which accounts for about 10 percent of all cancer deaths. And, every three minutes, one person in the US is diagnosed with a blood cancer — about 170,000 people ...

Read More »



LncFinder – an integrated platform for long non-coding RNA identification

August 9, 2018 Leave a comment 1,409 Views

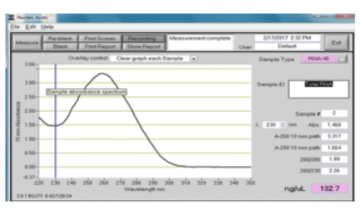


Discovering new long non-coding RNAs (lncRNAs) has been a fundamental step in lncRNA-related research. Nowadays, many machine learning-based tools have been developed for lncRNA identification. However, many methods predict lncRNAs using sequence-derived features alone, which tend to display unstable performances ...

Read More »

Optimized methodology for the generation of RNA-sequencing libraries from low-input starting material

February 12, 2018 Leave a comment 2,315 Views



RNA sequencing (RNA-seq) has become an important tool for examining the role of the transcriptome to biological processes. While RNA-seq has been widely adopted as a popular approach in many experimental designs, from gene discovery to mechanistic validation of targets, ...

Read More »

Bioinformatics Workshop – Introduction to RNA-seq Analysis Using High-Performance Computing and R

14 days ago Leave a comment 1,170 Views

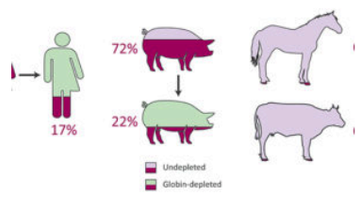
HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

In the Introduction to RNA-seq Analysis Using High-Performance Computing Workshop, participants will learn the basics of Unix/Linux and gain experience using the HMS compute cluster (O2). Participants...

Read More »

RNA Sequencing (RNA-Seq) Reveals Extremely Low Levels of Reticulocyte-Derived Globin Gene Transcripts in Peripheral Blood From Horses (Equus caballus) and Cattle (Bos taurus)

21 days ago Leave a comment 346 Views



RNA-seq has emerged as an important technology for measuring gene expression in peripheral blood samples collected from humans and other vertebrate species. In particular, transcriptomics analyses of whole blood can be used to study immunobiology and develop novel biomarkers of infectious disease. However, ...

Read More »

RNA-SEQ EXPERIMENTAL SETUP

Tomato - Flavor - Experiment

Flavor is a balance of acidity and sugar, plus the influence of elusive volatile compounds for aroma and flavor. Regardless of variety, grow conditions such as temperature can influence flavor.

Treatment #1 t_1=27°C / t_2=15°C

Treatment #2 t_1=29°C / t_2=18°C

ADH1A Gene - Experiment

Alcohol Dehydrogenase 1A (Class I), Alpha Polypeptide

This gene encodes class I alcohol dehydrogenase, alpha subunit, which is a member of the alcohol dehydrogenase family. Members of this enzyme family metabolize a wide variety of substrates, including ethanol.

Design a study to better understand metapopulation-based *ADH1A* gene expression in Marmoset?

Sample Design

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

Library Prep

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

Data Analysis

What is the purpose of your RNA-Seq experiment?

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

- to quantify transcription (DEGs 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

What resources are available and what is the quality of these resources?

References (e.g. genome, transcriptome)

Assembly Quality (e.g. draft, contamination)

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

Design	Preparation	Methode	Analysis	Extras

What factors influence the design?

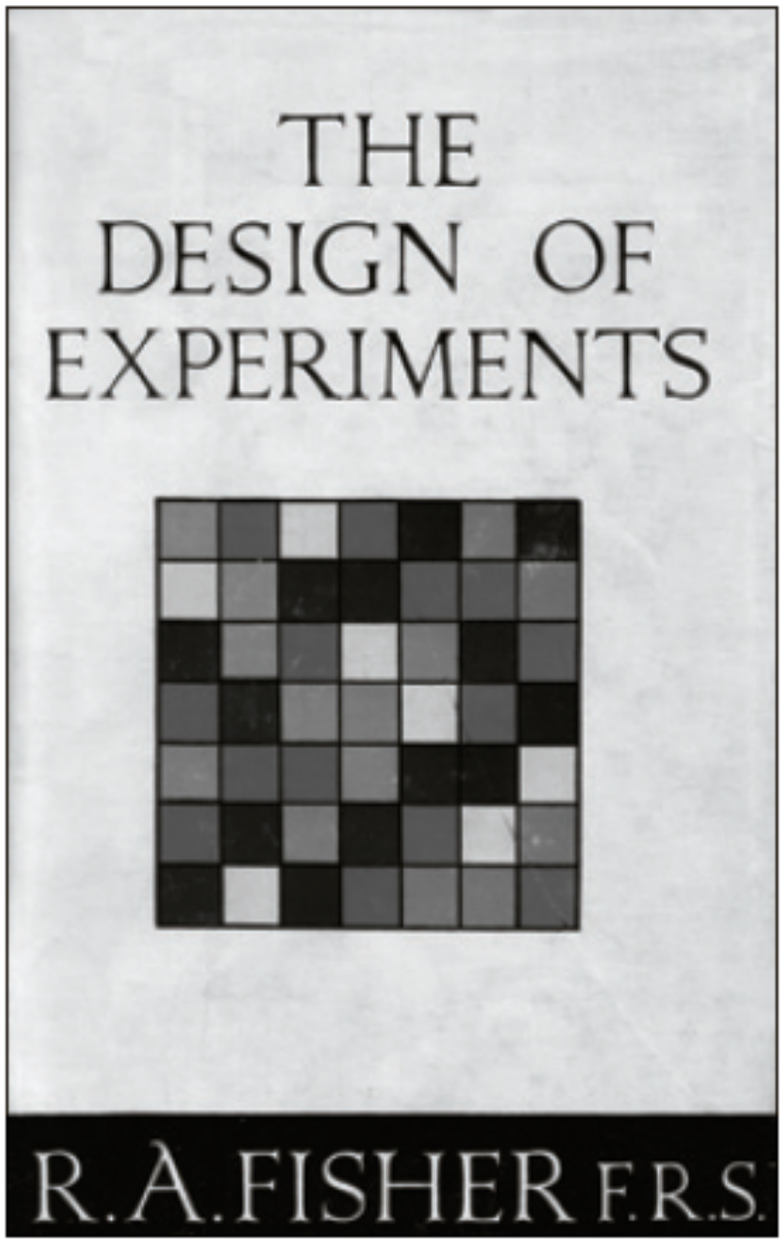
What is the **aim** of the experiment (e.g. expression, isoforms)?

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

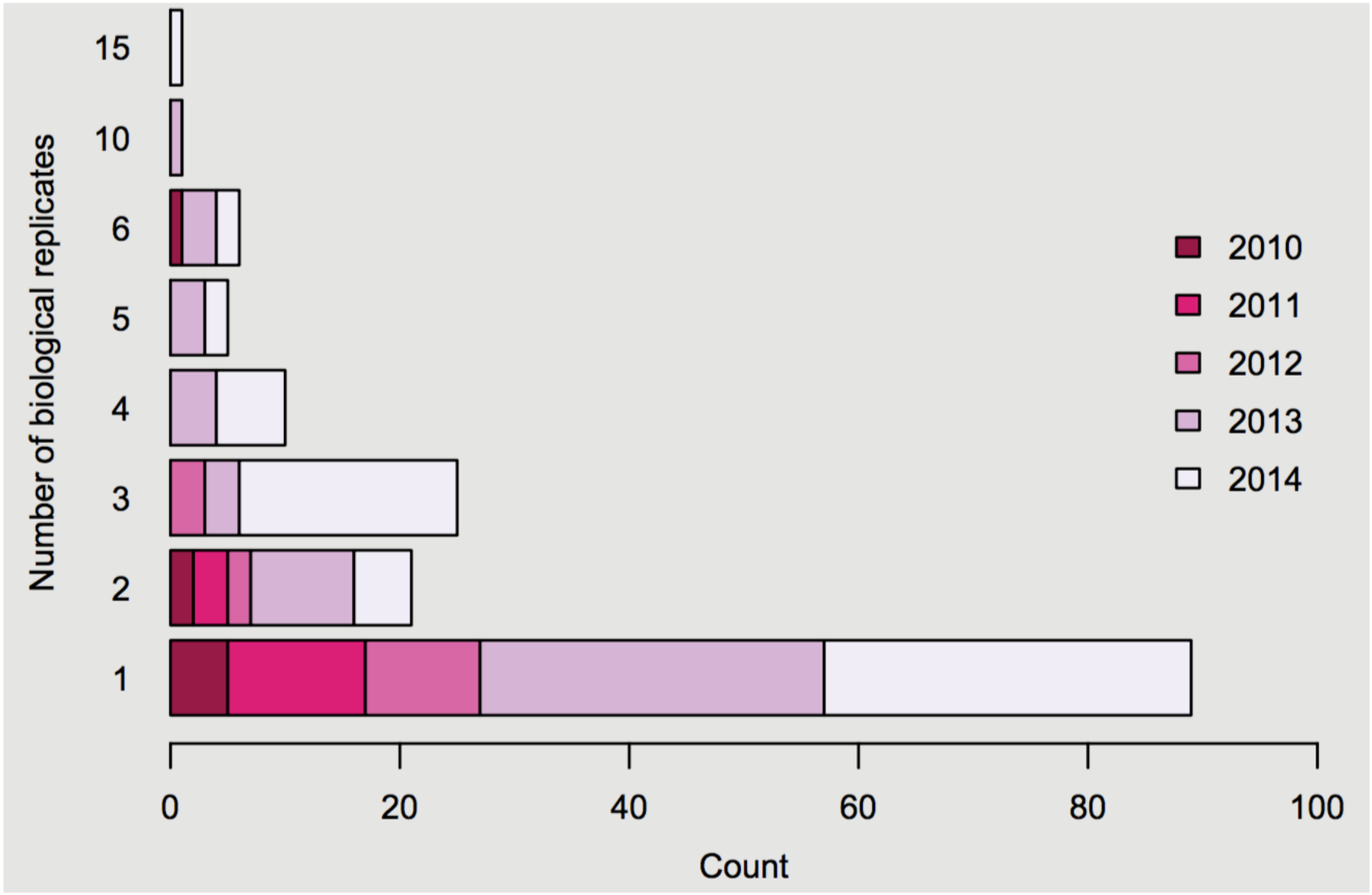
What (min) depth of sequencing **coverage** is required?

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

How much **money** do we have?



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



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

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?

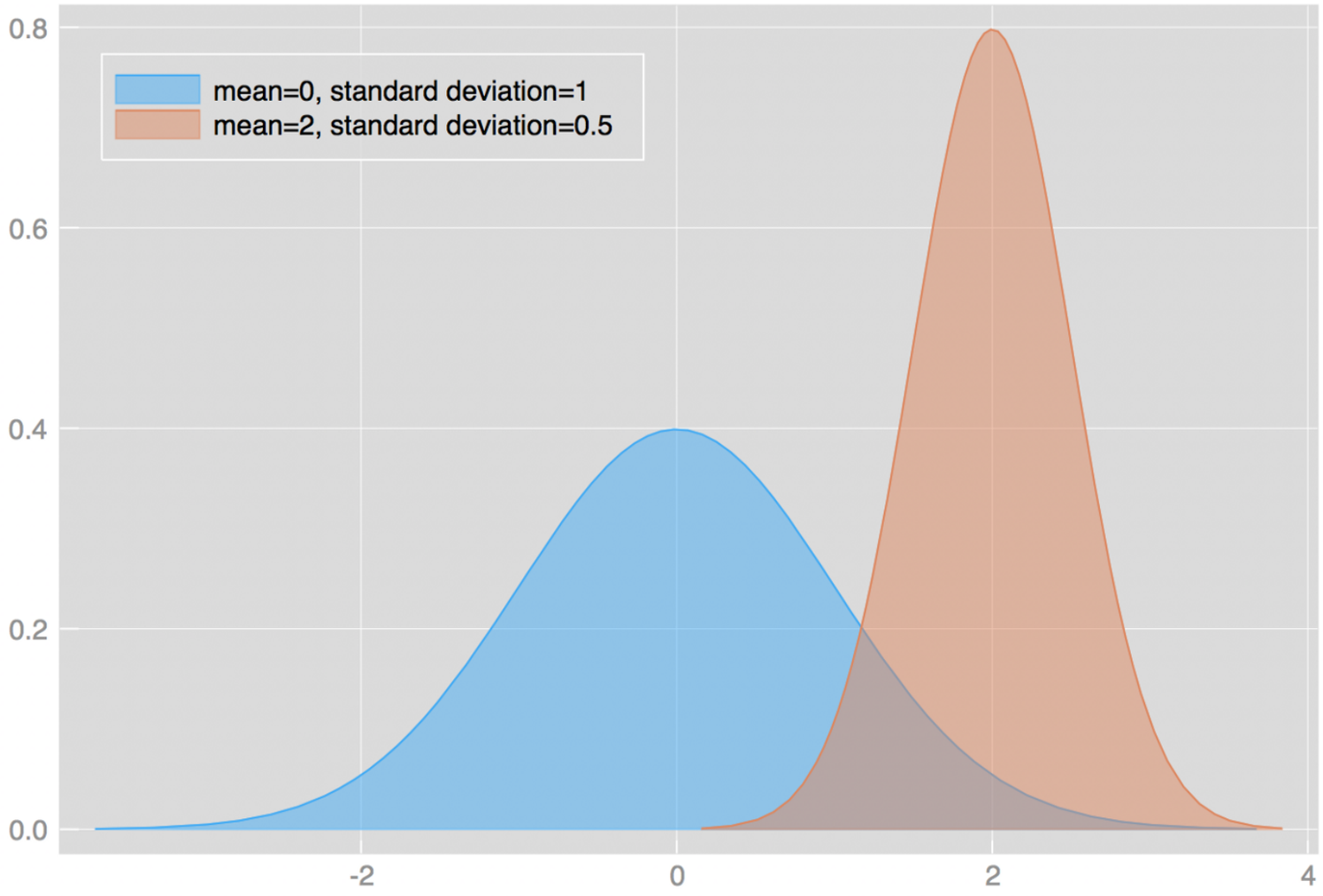
NICHOLAS J. SCHURCH,[1,6] PIETÀ SCHOFIELD,[1,2,6] MAREK GIERLIŃSKI,[1,2,6] CHRISTIAN COLE,[1,6] ALEXANDER SHERSTNEV,[1,6] VIJENDER SINGH,[2] NICOLA WROBEL,[3] KARIM GHARBI,[3] GORDON G. SIMPSON,[4] TOM OWEN-HUGHES,[2] MARK BLAXTER,[3] and GEOFFREY J. BARTON[1,2,5]

[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.

Probability Distributions



Probability Distributions

Binomial Distribution

Normal Distribution

Poisson Distribution

- named after the French mathematician Simeon Denis Poisson (1781-1840)
- probability model in biology and medicine
- count data
- the mean and the variance are equal

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

λ (lambda): average number of occurrences

e : constant 2.7183

x : number of occurrences

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.

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:

(1) sample size (e.g. number of replicates)

(2) effect size (e.g. fold-change)

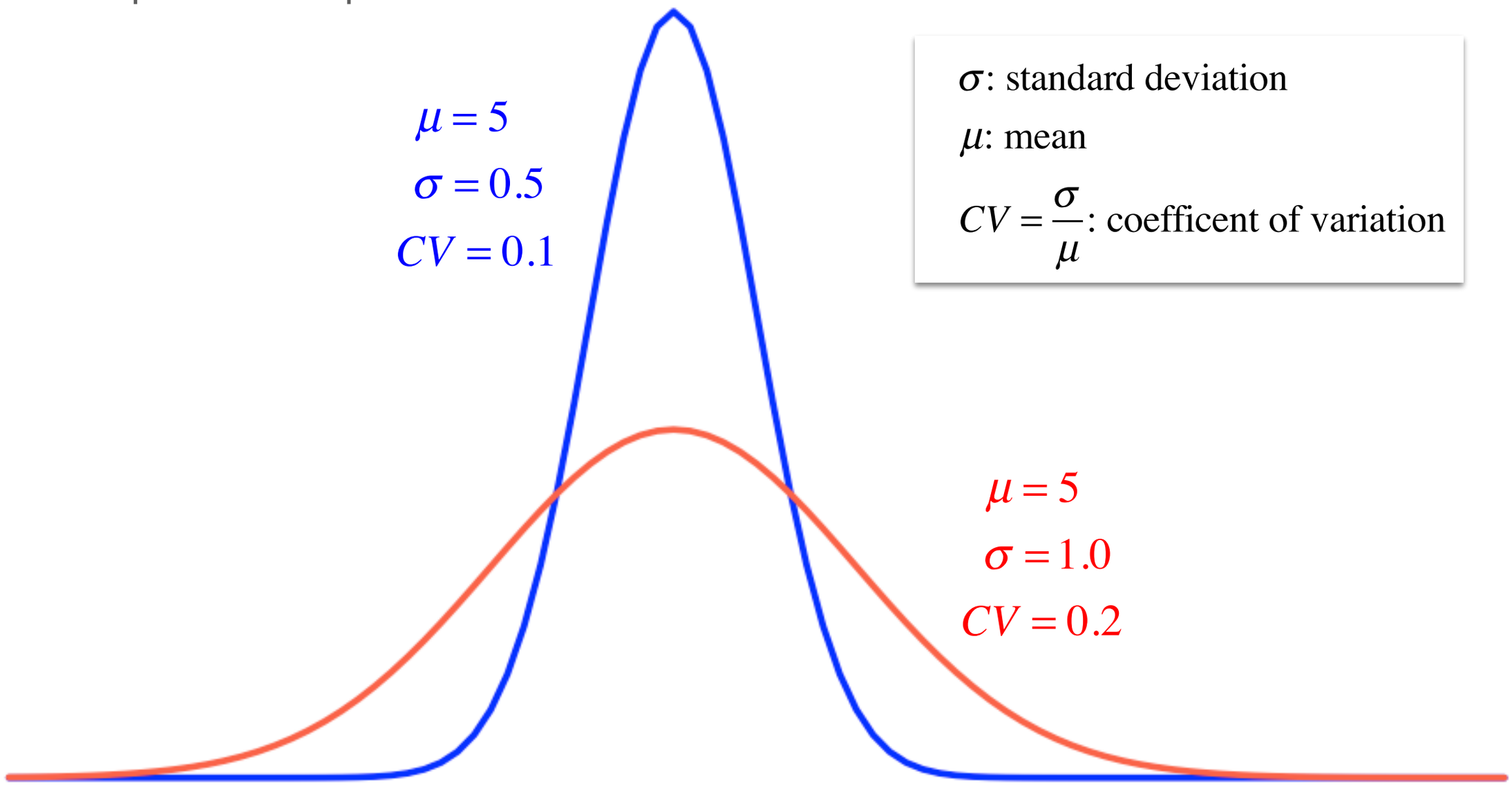
(3) significance level = P(Type I error) = probability of finding an effect that is not there

(4) power = 1 - P(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

A simple example:



σ: standard deviation

μ: mean

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

Inbred vs. **Wild**

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

Model Organsims vs. Non-Model Organsims

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

CV: coefficent of variation

σ: standard deviation

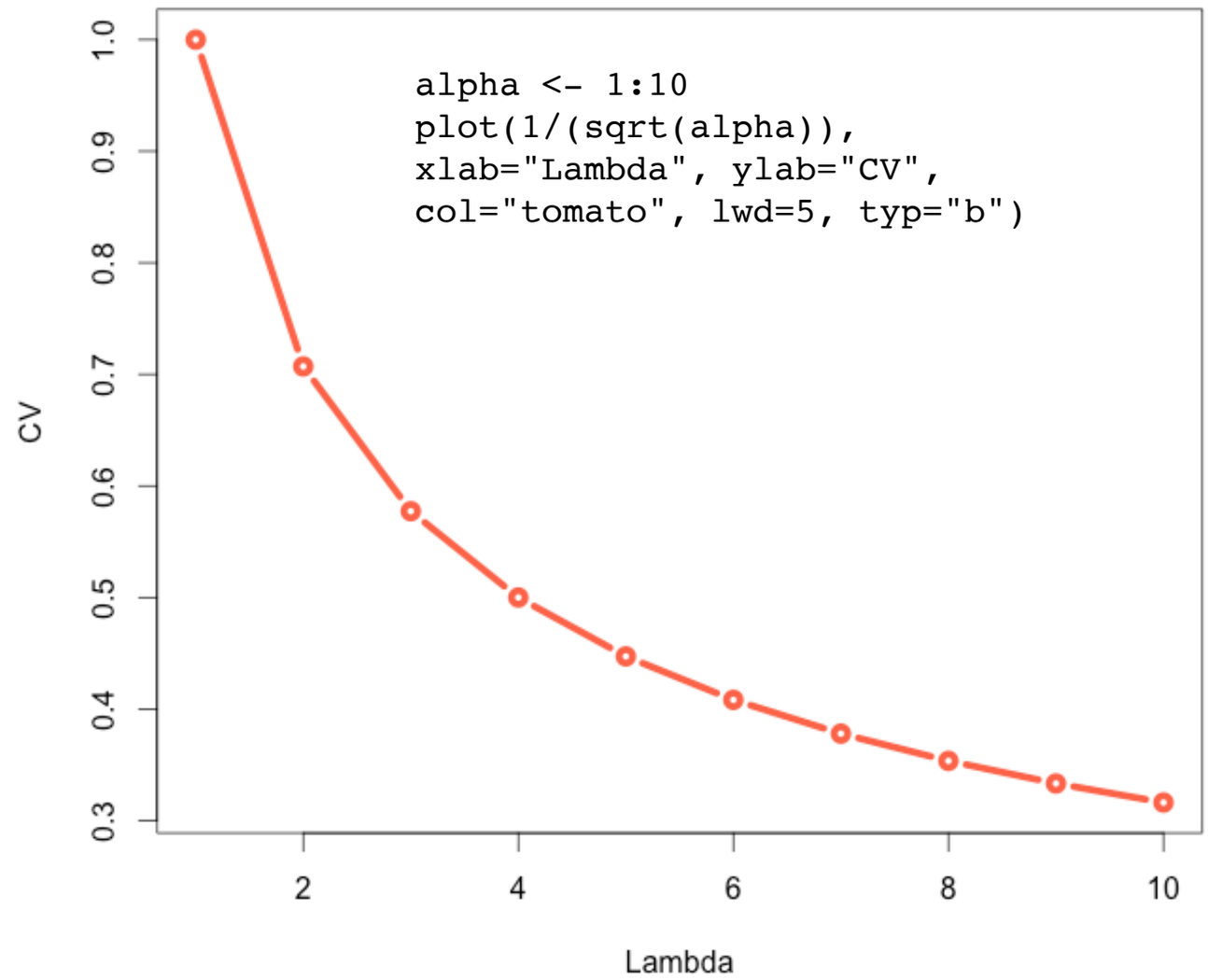
μ: mean

inbred animal strains: $CV \leq 0.2$

unrelated individuals: $CV > 0.3$

Coefficent of variance for a Poisson distribution

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



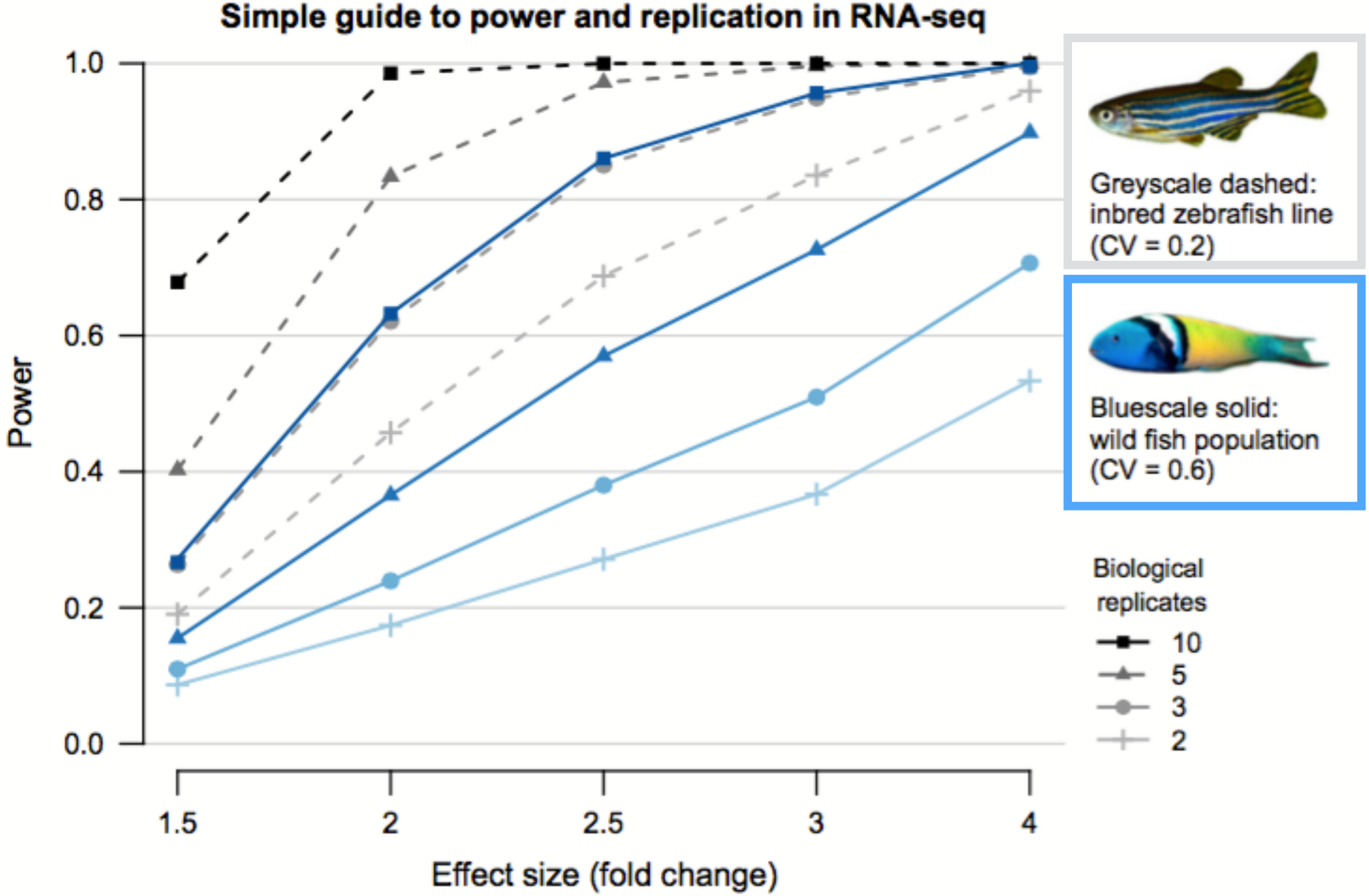
CV: coefficent of variation

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

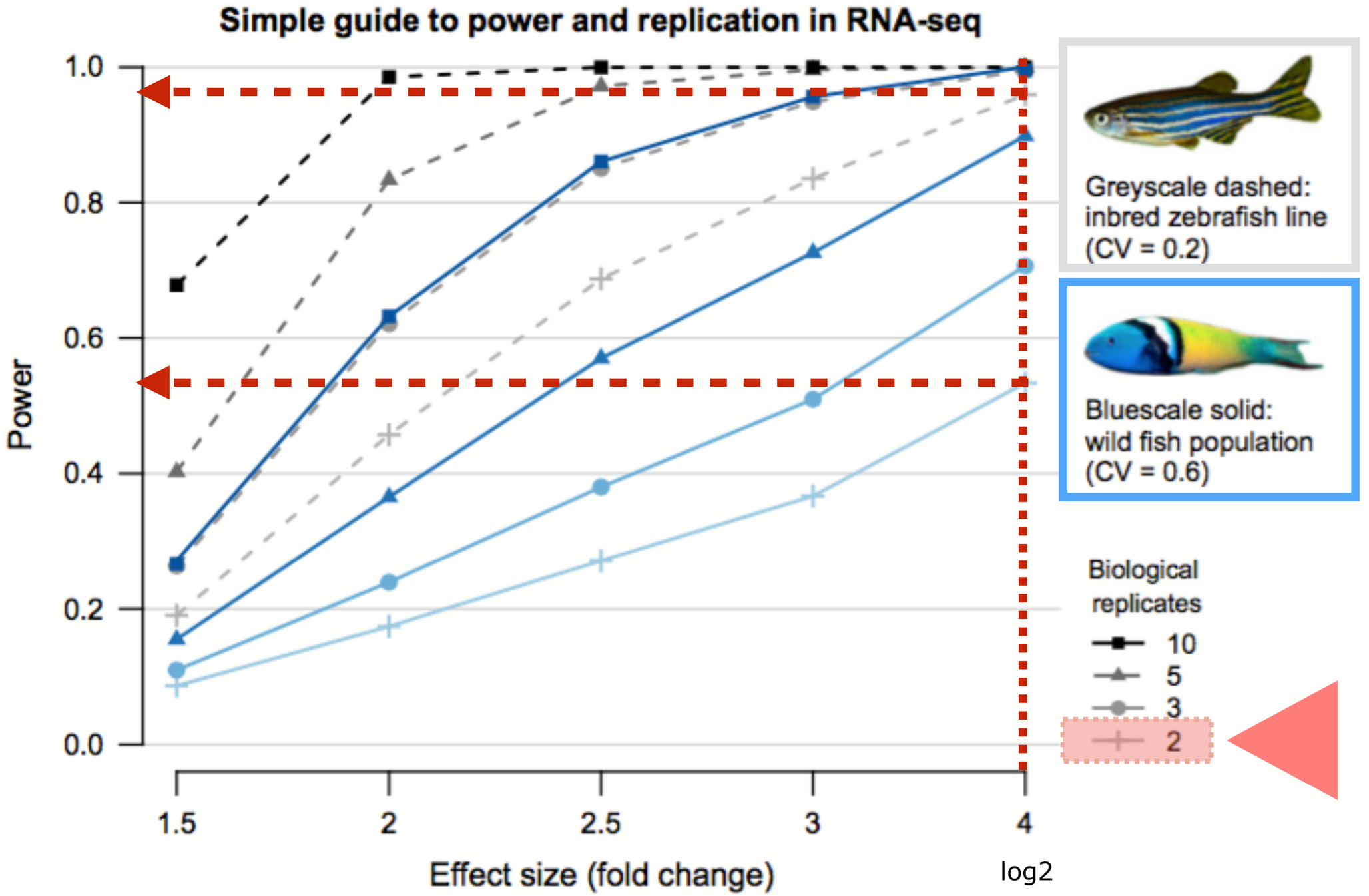
σ: standard deviation (= $\sqrt{Variance}$)

μ: mean

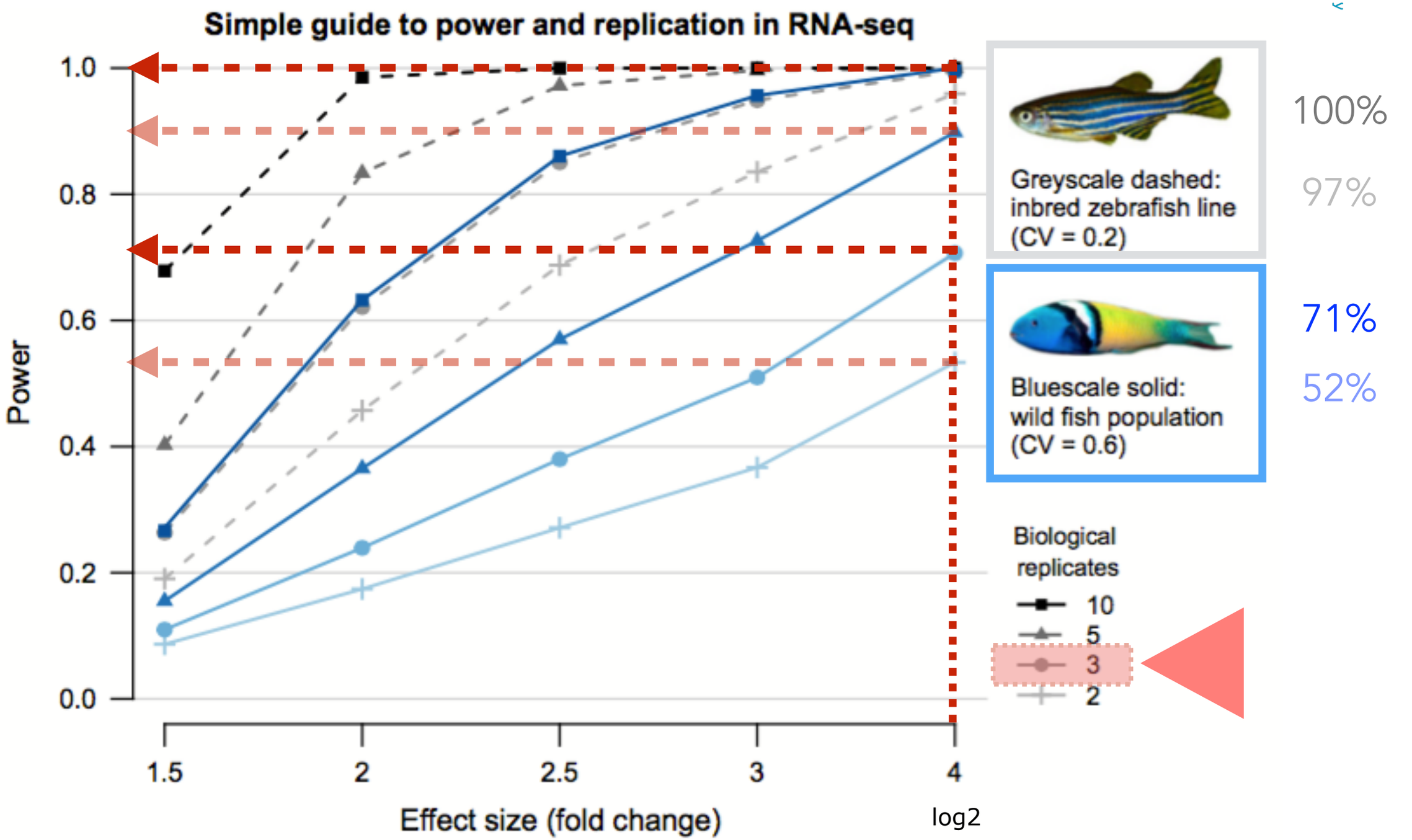
! The expected value and variance of a Poisson-distributed random variable are both equal to λ.



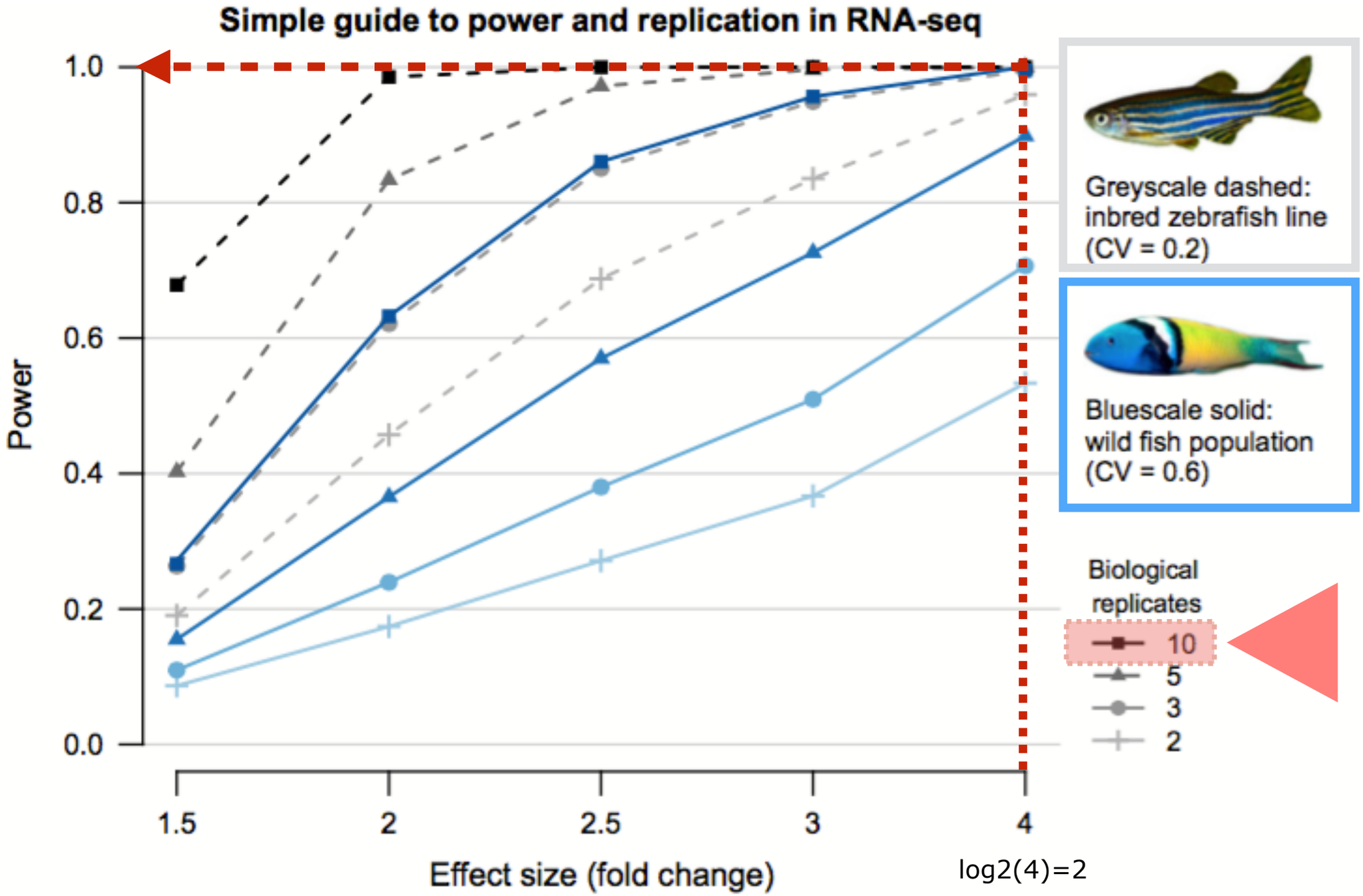
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.



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



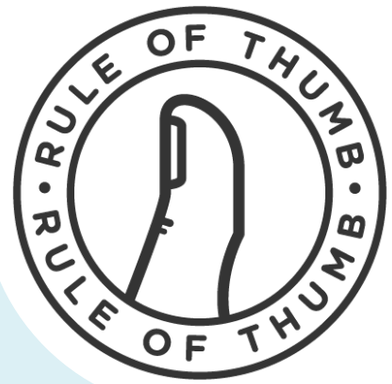
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.



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

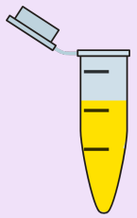
- Expression landscape?
- Library complexity?
- Read distribution?

☛ Pilot (Test) Sequencing



1. CLEAR SCIENTIFIC QUESTION
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

Sample Design



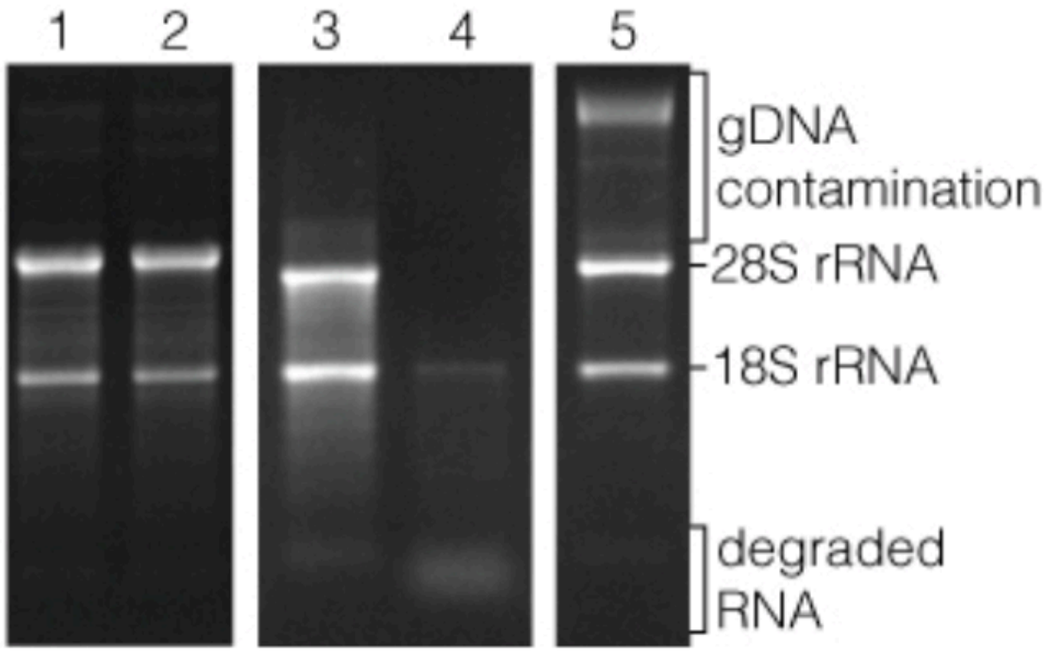
Sample preparation
RNA extraction
Cleaning (e.g. remove ribosomal RNA)

Library Prep

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

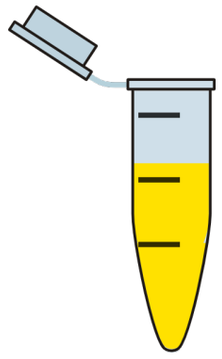
Data Analysis

Quantity and Quality of RNA



RNA analysis by agarose gel electrophoresis. Lanes 1 and 2 are examples of intact RNA with a 28S:18S rRNA ratio of approximately 2:1. Lane 3 is an example of degraded RNA with RNA smeaing below the 28S and 18S RNA bands. Lane 4 is an example of RNA degradation resulting in the loss of the 28S rRNA band and an accumulation of degraded RNA near the bottom of the gel. Lane 5 is an example of RNA with significant genomic DNA (gDNA) contamination.

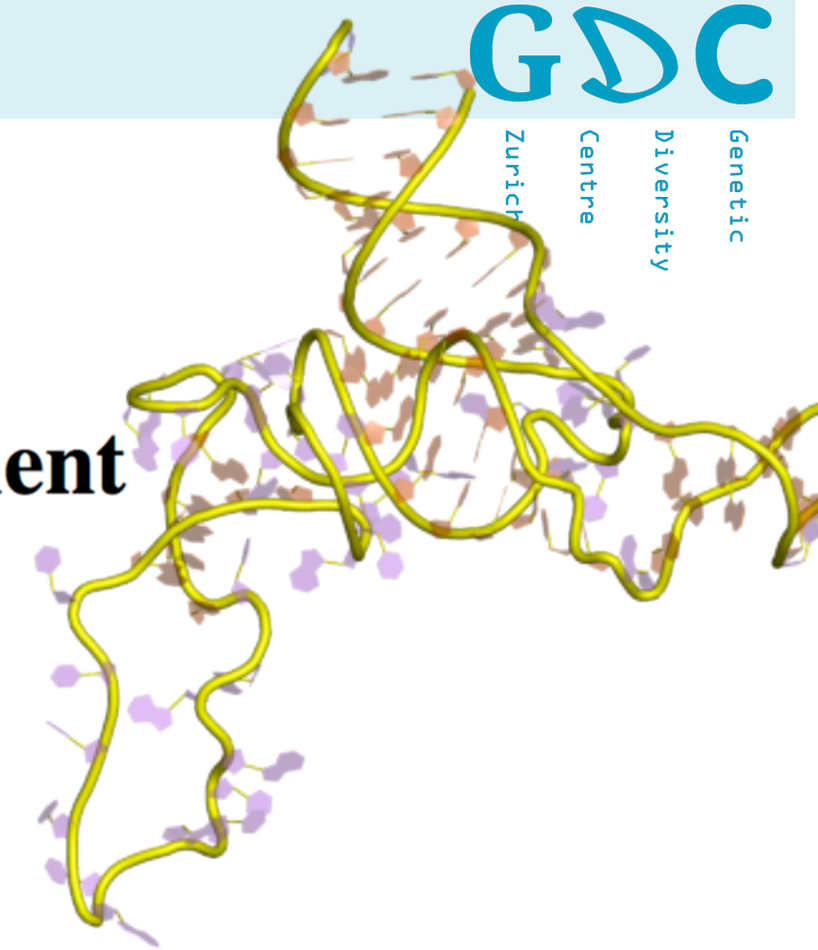
Source: Wieczorek *et al.* Promega Corporation



DNA

RNA ➜ **Total RNA**

mRNA, polyA RNA, polysomal RNA, tRNA, ribosomal RNA, lincRNA, miRNA, piRNA, siRNA, SRP RNA, tmRNA, snRNA, snoRNA, SmY RNA, scaRNA, gRNA, aRNA, crRNA, tasiRNA, rasiRNA, 7SK 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

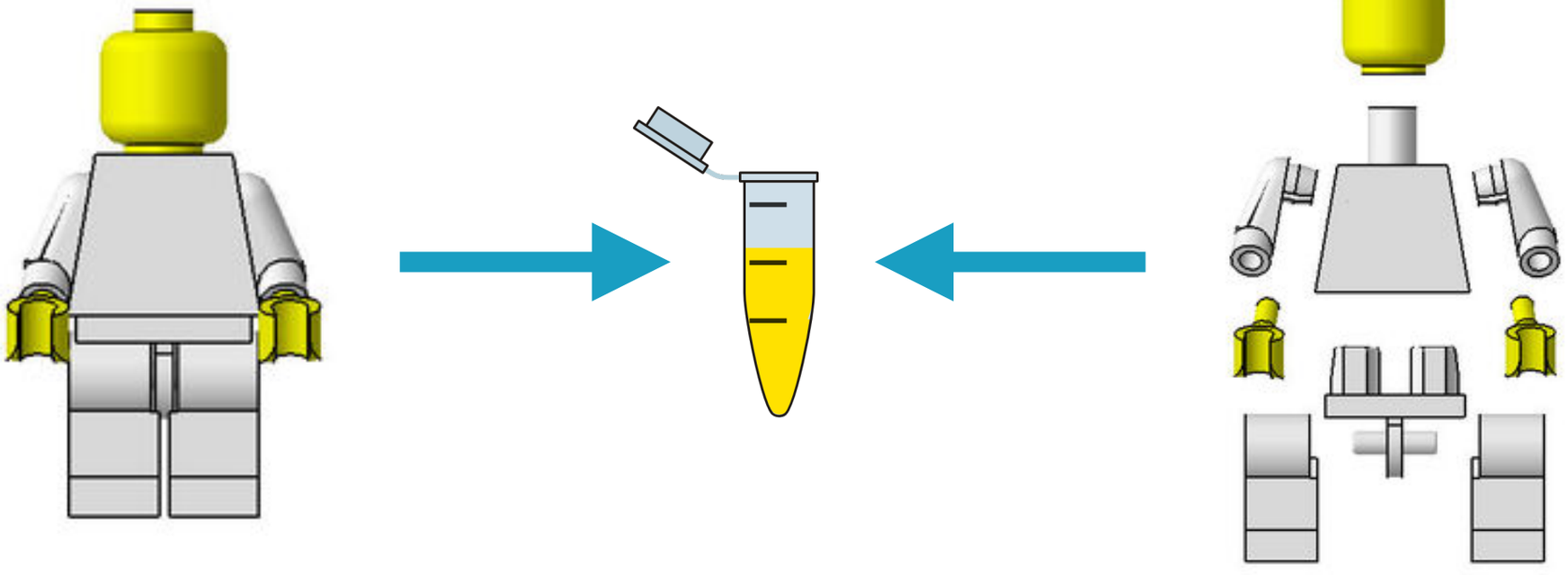
ABSTRACT

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.

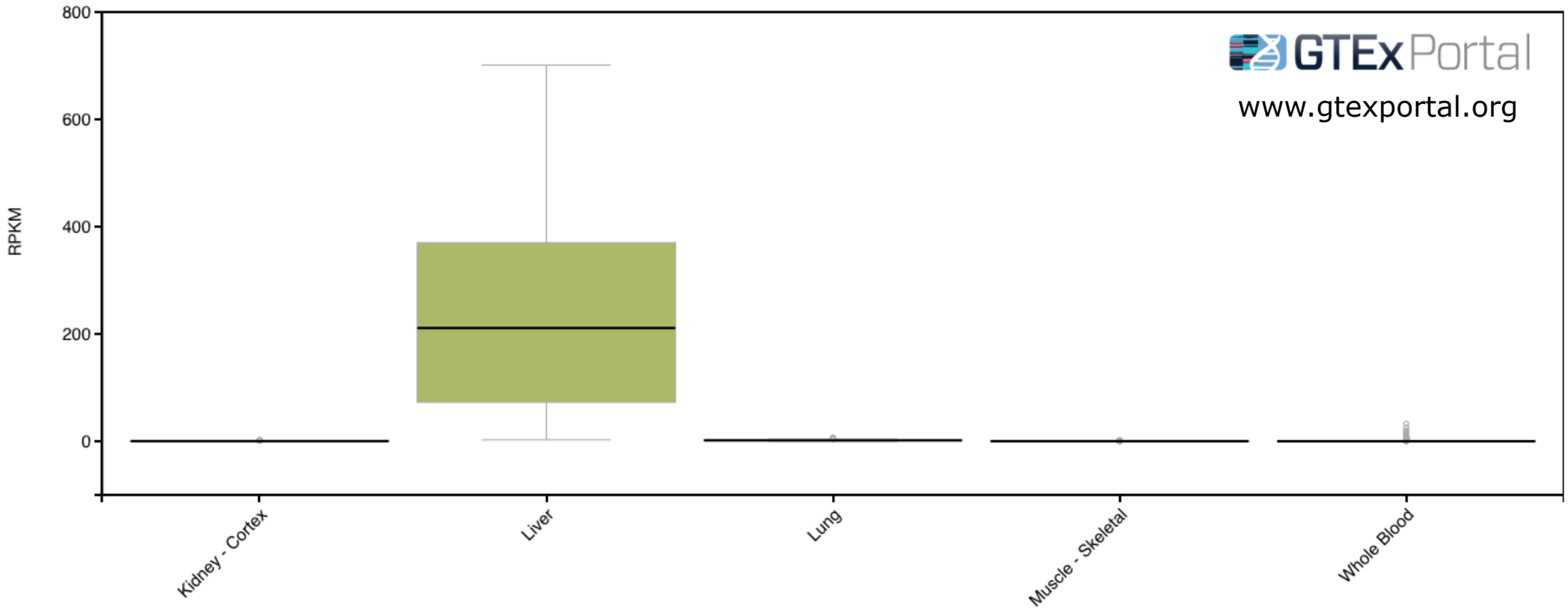
Keywords: rRNA depletion • sample preparation • RNA-seq • next generation sequencing • transcriptome

Genetic Diversity Centre Zurich

?



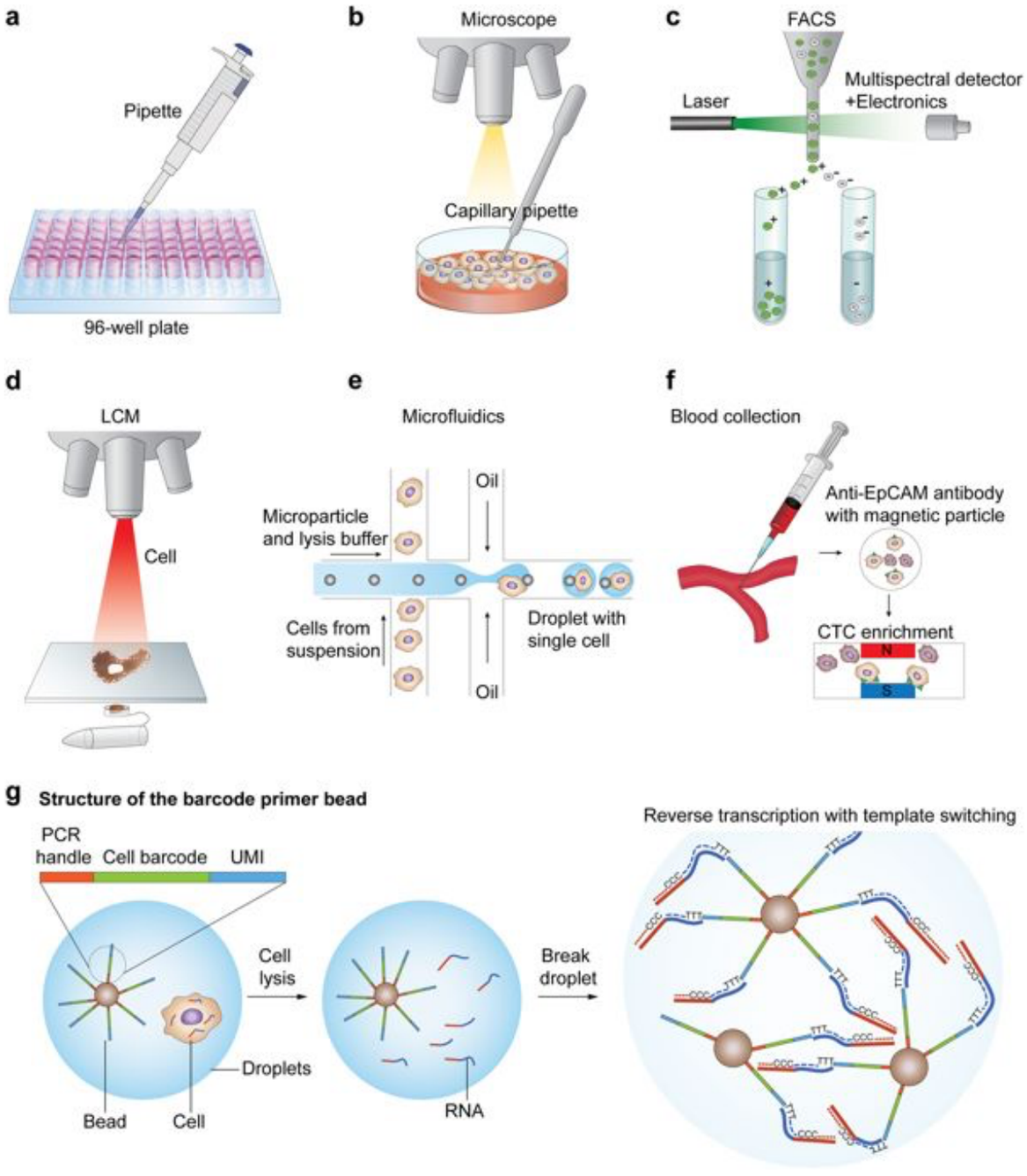
ADH1A Gene Expression



ADH1A encodes a member of the alcohol dehydrogenase family. The encoded protein is the alpha subunit of class I alcohol dehydrogenase, which consists of several homo- and heterodimers of alpha, beta and gamma subunits. **Alcohol dehydrogenases catalyze the oxidation of alcohols to aldehydes.** This gene is active in the **liver** in **early fetal life but only weakly active in adult liver.** This gene is found in a cluster with six additional alcohol dehydrogenase genes, including those encoding the beta and gamma subunits, on the long arm of chromosome 4. Mutations in this gene may contribute to variation in certain personality traits and substance dependence.

GeneCards
HUMAN GENE DATABASE

Single-cell RNA sequencing (scRNA-seq)



Single-cell isolation techniques:
a The limiting dilution method isolates individual cells, leveraging the statistical distribution of diluted cells. **b** Micromanipulation involves collecting single cells using microscope-guided capillary pipettes. **c** FACS isolates highly purified single cells by tagging cells with fluorescent marker proteins. **d** Laser capture microdissection (LCM) utilizes a laser system aided by a computer system to isolate cells from solid samples. **e** Microfluidic technology for single-cell isolation requires nanoliter-sized volumes. An example of in-house microdroplet-based microfluidics (e.g., Drop-Seq). **f** The CellSearch system enumerates CTCs from patient blood samples by using a magnet conjugated with CTC binding antibodies. **g** A schematic example of droplet-based library generation. Libraries for scRNA-seq are typically generated via cell lysis, reverse transcription into first-strand cDNA using uniquely barcoded beads, second-strand synthesis, and cDNA amplification.

Source: Lee and Bang (2019) Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine 50

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).

Sample Design

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

Library Prep

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

Data Analysis

Whole-Transcriptome Sequencing

NextSeq*†

HiSeq 4000*

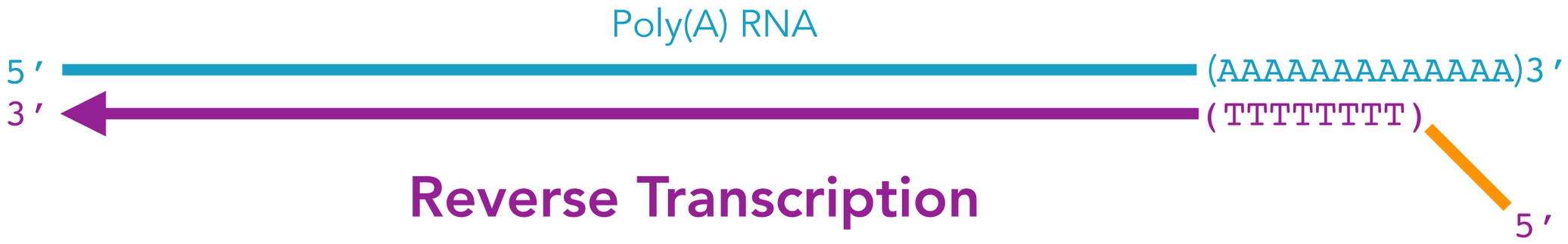
NovaSeq 6000*††

Output Range	20–120 Gb	125–1500 Gb	134–6000 Gb
Run Time	11–29 hr	< 1–3.5 days	13–44 hr
Reads per Run	130–400 million	2.5–5 billion	Up to 20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp
Samples per Run‡	2–8	50–100	26–400
Relative Price per Sample‡	Higher Cost	Mid Cost	Lower Cost
Relative Instrument Price‡	Lower Cost	Mid Cost	Higher Cost

mRNA

The **poly-A** tail is a long chain of adenine nucleotides that is added to a messenger RNA (mRNA) molecule during RNA processing to increase the **stability** of the molecule. Additionally, the poly-A tail allows the mature messenger RNA molecule to be **exported** from the nucleus and translated into a protein by ribosomes in the cytoplasm.

Source: Scitable by Nature Education



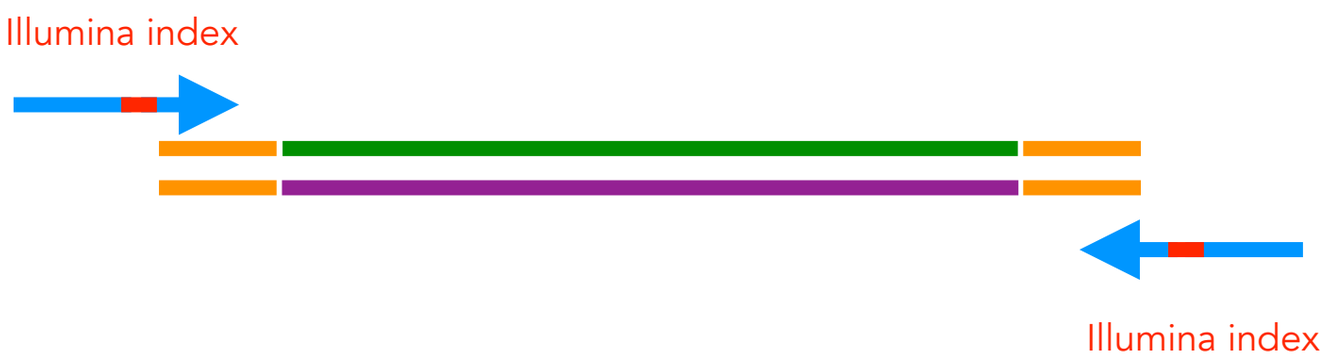
A **reverse transcriptase** is an enzyme used to generate complementary DNA from an RNA template.


removal of RNA
5′
3′
(AAAAAAAAAAAAA)3′
(TTTTTTTT)
5′

Reverse Transcription

Double-stranded cDNA library

Library Amplification for Multiplexing



Possible Bias:

- over-representation of transcript end
- non-random starting point
- short fragments are prefered

DISCOVER FULL-LENGTH TRANSCRIPTS

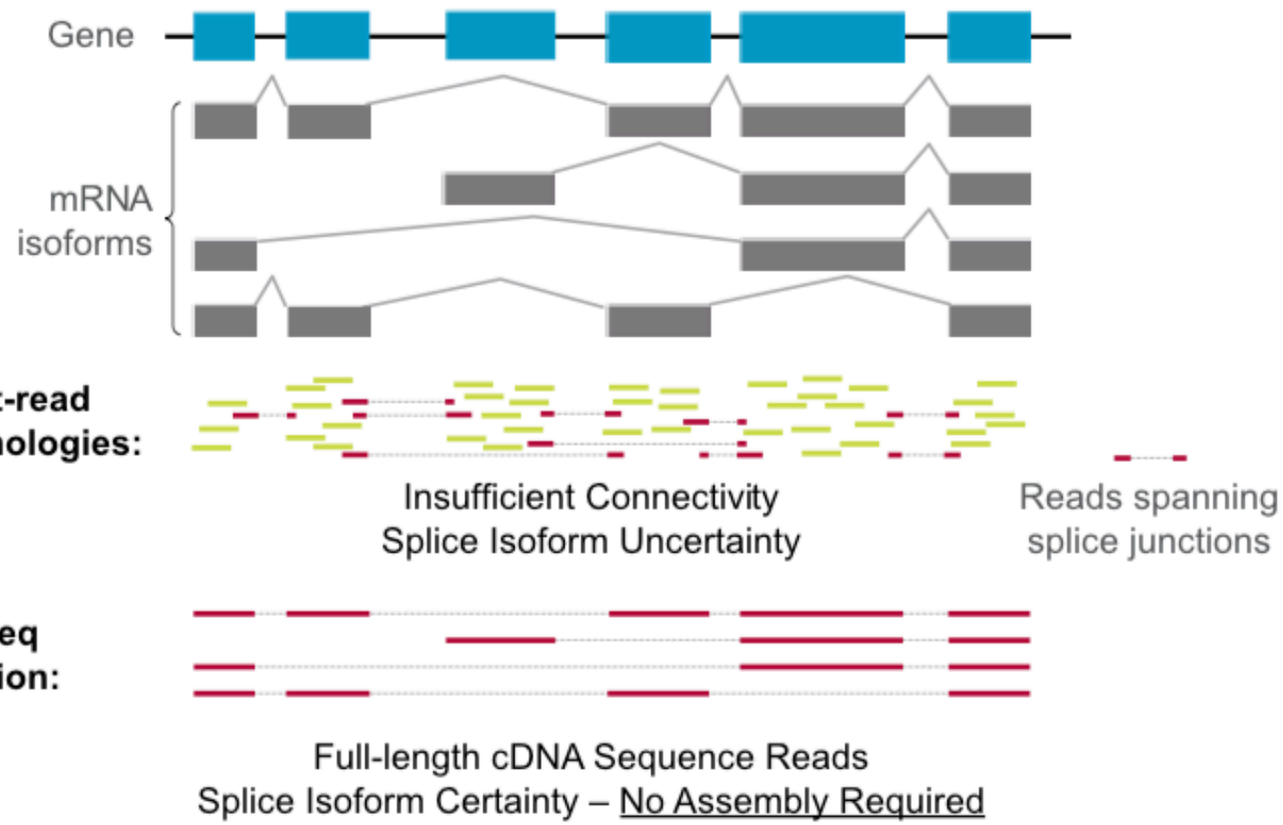
Get a complete view of transcript isoform diversity with PacBio long-read sequencing.

RNA Sequencing

Single Molecule, Real-Time (SMRT) Sequencing and Iso-Seq analysis allow you to generate full-length cDNA sequences — no assembly required — to characterize transcript isoforms within targeted genes or across an entire transcriptome so that you can easily and affordably:

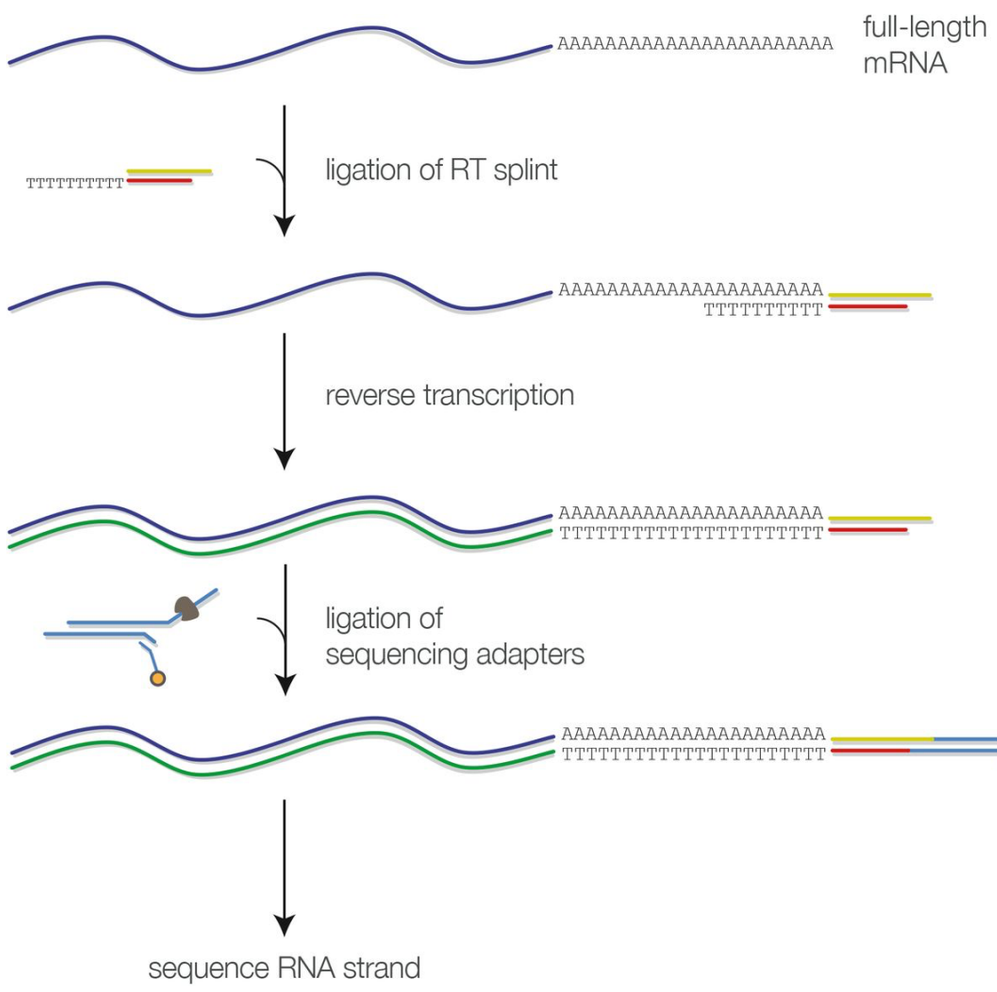
- Discover new genes, transcripts and alternative splicing events
- Improve genome annotation to identify gene structure, regulatory elements, and coding regions
- Increase the accuracy of RNA-seq quantification with isoform-level resolution

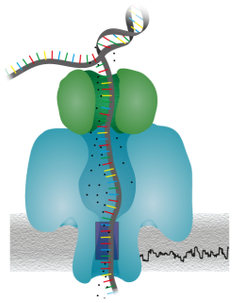
DETERMINATION OF TRANSCRIPT ISOFORMS



The Iso-Seq method allows you to make evidence-based genome annotations, discover novel genes and isoforms, identify promoters and splice sites to understand gene regulation, improve accuracy of RNA-seq quantification for gene expression studies, and distinguish important stress response, developmental, or tissue-specific isoforms.

Direct RNA Sequencing Kit





ONT

Input requirement: 500 ng RNA

Preparation time: 110 min

GDC Genetic Diversity Centre Zurich

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

RNA-SEQ DATA QUALITY

Sigurgeirsson et al. (2014) found than more than half of the genes were differentially expressed due to **in vitro RNA degradation**.

Wang L, Nie J, Sicotte H, et al. (2016) Measure transcript integrity using RNA-seq data. BMC Bioinformatics.

Sigurgeirsson B, Emanuelsson O, Lundeberg J. (2014) Sequencing degraded RNA addressed by 3′ tag counting. PLoS One.

Wang et al. BMC Bioinformatics (2016) 17:58
DOI 10.1186/s12859-016-0922-z

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Measure transcript integrity using RNA-seq data

Liguo Wang[1†], Jinfu Nie[1†], Hugues Sicotte[1], Ying Li[1], Jeanette E. Eckel-Passow[1], Surendra Dasari[1], Peter T. Vedell[1], Poulami Barman[1], Liewei Wang[3], Richard Weinshiboum[3], Jin Jen[4], Haojie Huang[5], Manish Kohli[2*] and Jean-Pierre A. Kocher[1*]

Abstract

Background: Stored biological samples with pathology information and medical records are invaluable resources for translational medical research. However, RNAs extracted from the archived clinical tissues are often substantially degraded. RNA degradation distorts the RNA-seq read coverage in a gene-specific manner, and has profound influences on whole-genome gene expression profiling.

Result: We developed the transcript integrity number (TIN) to measure RNA degradation. When applied to 3 independent RNA-seq datasets, we demonstrated TIN is a reliable and sensitive measure of the RNA degradation at both transcript and sample level. Through comparing 10 prostate cancer clinical samples with lower RNA integrity to 10 samples with higher RNA quality, we demonstrated that calibrating gene expression counts with TIN scores could effectively neutralize RNA degradation effects by reducing false positives and recovering biologically meaningful pathways. When further evaluating the performance of TIN correction using spike-in transcripts in RNA-seq data generated from the Sequencing Quality Control consortium, we found TIN adjustment had better control of false positives and false negatives (sensitivity = 0.89, specificity = 0.91, accuracy = 0.90), as compared to gene expression analysis results without TIN correction (sensitivity = 0.98, specificity = 0.50, accuracy = 0.86).

Conclusion: TIN is a reliable measurement of RNA integrity and a valuable approach used to neutralize in vitro RNA degradation effect and improve differential gene expression analysis.

Keywords: Transcript integrity number, TIN, RNA-seq quality control, Gene expression

RSeQC

tin.py

This program is designed to evaluate RNA integrity at **transcript** level. TIN (transcript integrity number) is named in analogous to RIN (RNA integrity number). RIN (RNA integrity number) is the most widely used metric to evaluate RNA integrity at **sample (or transcriptome)** level. It is a very useful preventive measure to ensure good RNA quality and robust, reproducible RNA sequencing. However, it has several weaknesses:

- RIN score (1 <= RIN <= 10) is not a direct measurement of **mRNA** quality. RIN score heavily relies on the amount of 18S and 28S ribosome RNAs, which was demonstrated by the four features used by the RIN algorithm: the "total RNA ratio" (i.e. the fraction of the area in the region of 18S and 28S compared to the total area under the curve), 28S-region height, 28S area ratio and the 18S:28S ratio24. To a large extent, RIN score was a measure of ribosome RNA integrity. However, in most RNA-seq experiments, ribosome RNAs were depleted from the library to enrich mRNA through either ribo-minus or polyA selection procedure.
- RIN only measures the overall RNA quality of an RNA sample. However, in real situation, the degradation rate may differs significantly among transcripts, depending on factors such as "AU-rich sequence", "transcript length", "GC content", "secondary structure" and the "RNA-protein complex". Therefore, RIN is practically not very useful in downstream analysis such as adjusting the gene expression count.
- RIN has very limited sensitivity to measure substantially degraded RNA samples such as preserved clinical tissues. (ref: http://www.illumina.com/documents/products/technotes/technote-truseq-rna-access.pdf).

To overcome these limitations, we developed TIN, an algorithm that is able to measure RNA integrity at transcript level. TIN calculates a score (0 <= TIN <= 100) for each expressed transcript, however, the medTIN (i.e. meidan TIN score across all the transcripts) can also be used to measure the RNA integrity at **sample** level. Below plots demonstrated TIN is a useful metric to measure RNA integrity in both transcriptome-wise and transcript-wise, as demonstrated by the high concordance with both RIN and RNA fragment size (estimated from RNA-seq read pairs).

Example output:

geneID	chrom	tx_start	tx_end	TIN
ABCC2	chr10	101542354	101611949	67.6446525761
IPMK	chr10	59951277	60027694	86.383618429
RUFY2	chr10	70100863	70167051	43.8967503948

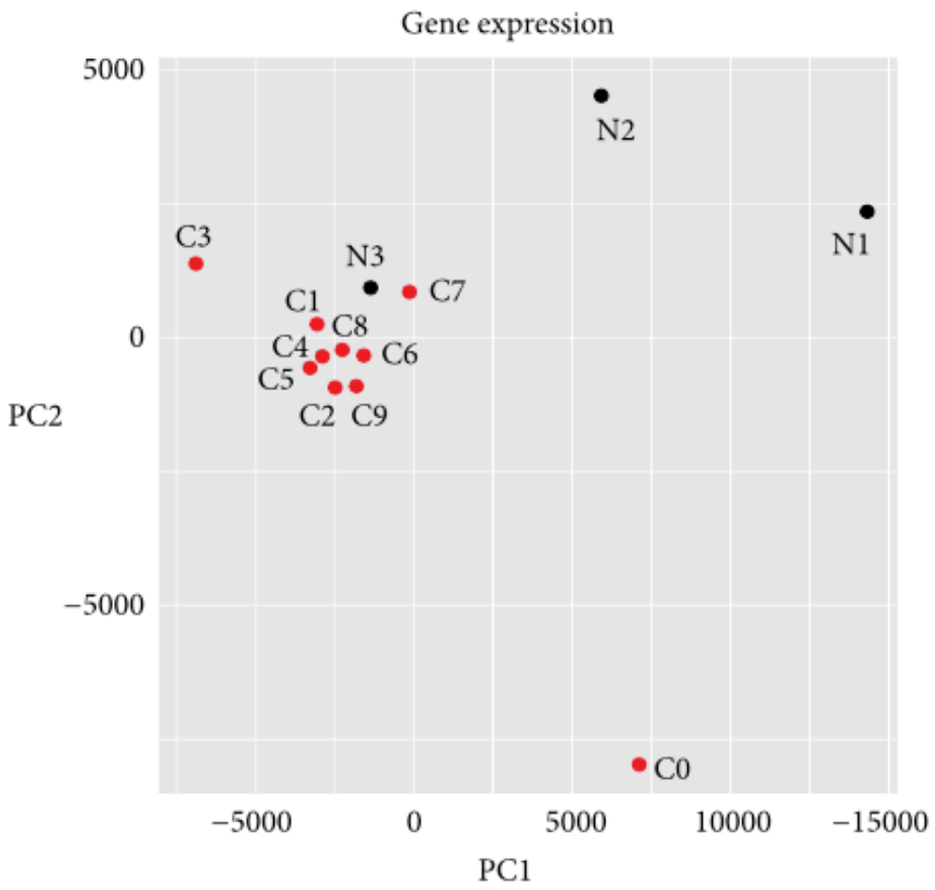
A Simple Guideline to Assess the Characteristics of RNA-Seq Data

Keunhong Son,[1] **Sungryul Yu**,[2] **Wonseok Shin**,[3] **Kyudong Han**,[3] **and Keunsoo Kang**[1]

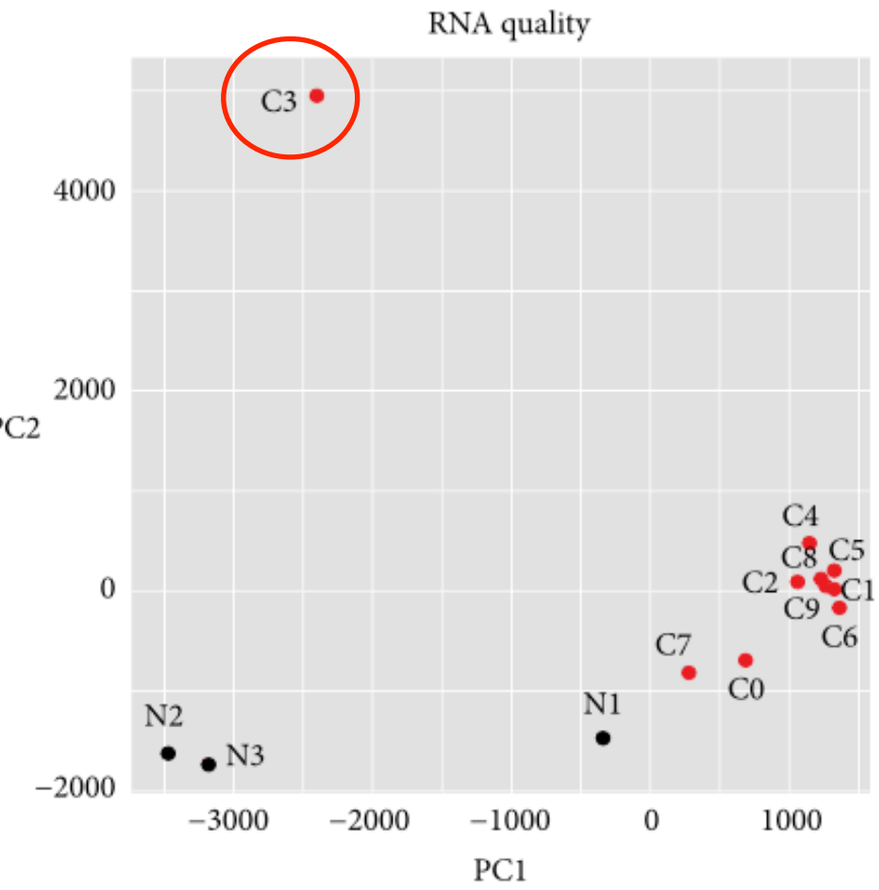
[1]*Department of Microbiology, College of Natural Sciences, Dankook University, Cheonan 31116, Republic of Korea*
[2]*Department of Clinical Laboratory Science, Semyung University, Jecheon 27136, Republic of Korea*
[3]*Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan 31116, Republic of Korea*

Next-generation sequencing (NGS) techniques have been used to generate various molecular maps including genomes, epigenomes, and transcriptomes. Transcriptomes from a given cell population can be profiled via RNA-seq. However, there is no simple way to assess the characteristics of RNA-seq data systematically. In this study, we provide a simple method that can intuitively evaluate RNA-seq data using two different principal component analysis (PCA) plots. The gene expression PCA plot provides insights into the association between samples, while the transcript integrity number (TIN) score plot provides a quality map of given RNA-seq data. With this approach, we found that RNA-seq datasets deposited in public repositories often contain a few low-quality RNA-seq data that can lead to misinterpretations. The effect of sampling errors for differentially expressed gene (DEG) analysis was evaluated with ten RNA-seq data from invasive ductal carcinoma tissues and three RNA-seq data from adjacent normal tissues taken from a Korean breast cancer patient. The evaluation demonstrated that sampling errors, which select samples that do not represent a given population, can lead to different interpretations when conducting the DEG analysis. Therefore, the proposed approach can be used to avoid sampling errors prior to RNA-seq data analysis.

PCA plots of RNA-seq data show the characteristics of samples according to gene expression (FPKM) levels (left) and RNA quality (TIN score).

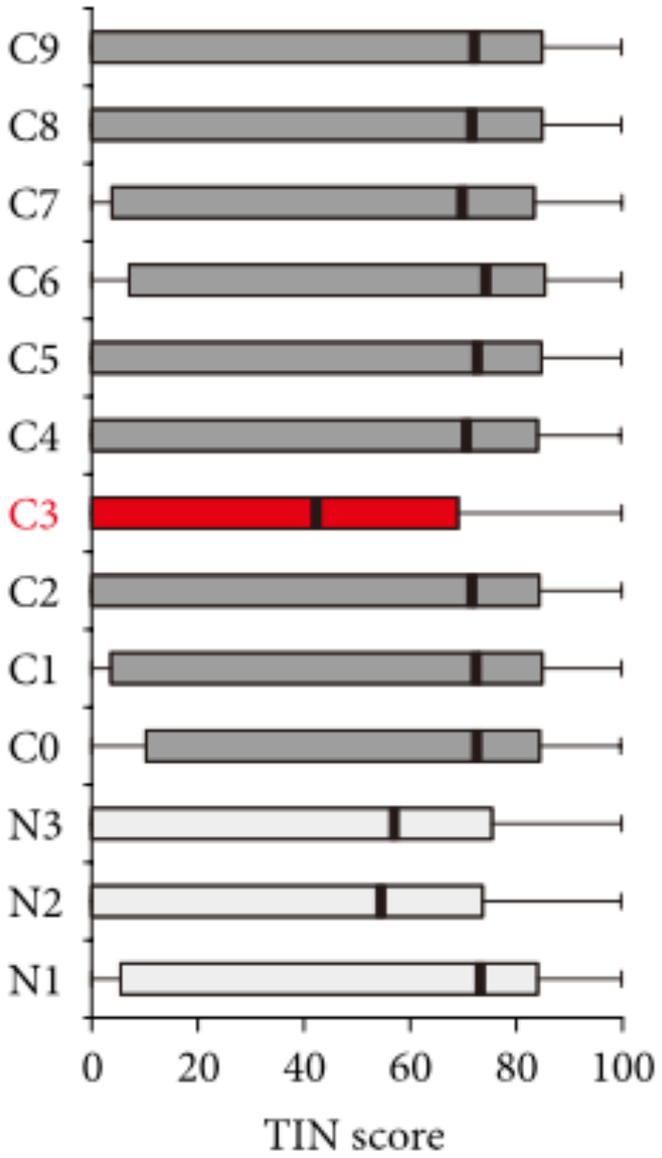


The gene expression PCA plot provides a map of the distances between samples from which the characteristics of RNA-seq data can be inferred.



The transcript integrity number (TIN) score PCA plot can infer the quality (not the sequencing quality) of RNA-seq data, which can effectively discriminate low-quality samples.

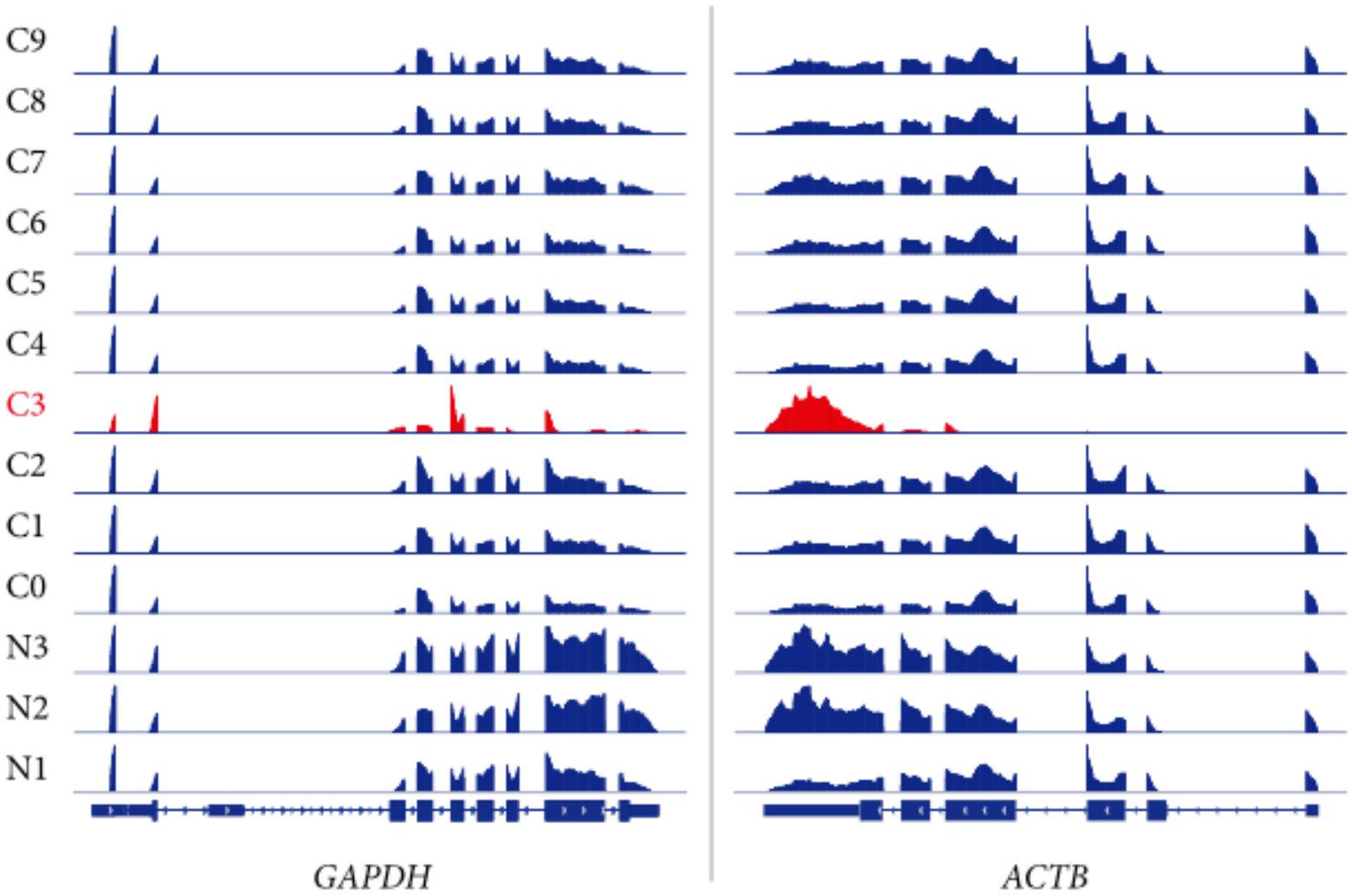
Source: Son et al. (2018). A Simple Guideline to Assess the Characteristics of RNA-Seq Data. BioMed research international.



Boxplot indicates the RNA quality of samples according to the TIN scores. A thick line (black) within the box marks the mean.

Source: Son et al. (2018). A Simple Guideline to Assess the Characteristics of RNA-Seq Data. BioMed research international.

Genome browser snapshots of mapped read densities are shown using integrative genomics viewer (IGV). FPKM, fragments per kilobase of transcript per million mapped reads.

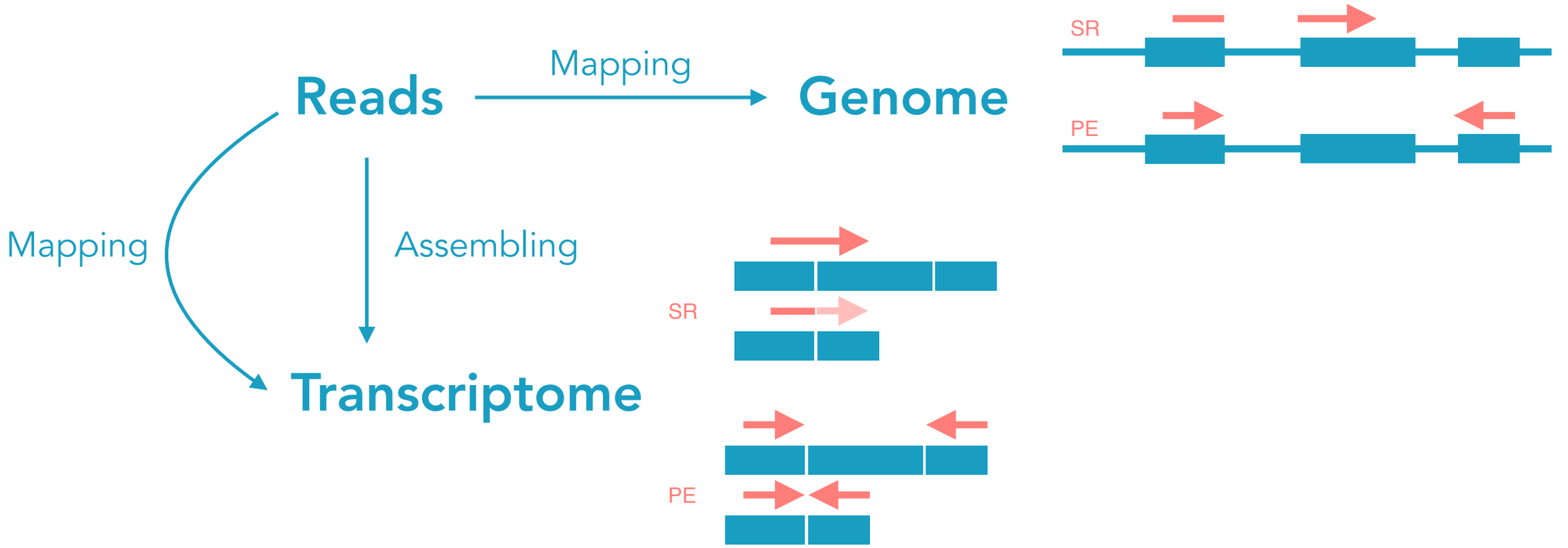


Source: Son et al. (2018). A Simple Guideline to Assess the Characteristics of RNA-Seq Data. BioMed research international.

RNA-SEQ DATA FILTERING

- Adaptor sequences (trim or remove)
- Non-mRNA (e.g. SSU rRNA)
- Low complexity sequences
- Contamination

RNA-SEQ READ MAPPING


Reads
Mapping
Genome
Mapping
Assembling
Transcriptome
SR
PE
SR
PE

Program	Mapping
BWA	unspliced
TopHat2	spliced
HISAT2	spliced
STAR	spliced
Kallisto	pseudo-alignment
Salmon	pseudo-alignment
Sailfish	pseudo-alignment

based on Costa-Silva et al. (2017) PLOS ONE

Traget Length

20 reads

$$20 \times 1.5 = 30 \rightarrow \frac{30}{10} = 3$$

15 reads

$$15 \times 1.5 = 22.5 \rightarrow \frac{22.7}{7} = 3.2$$

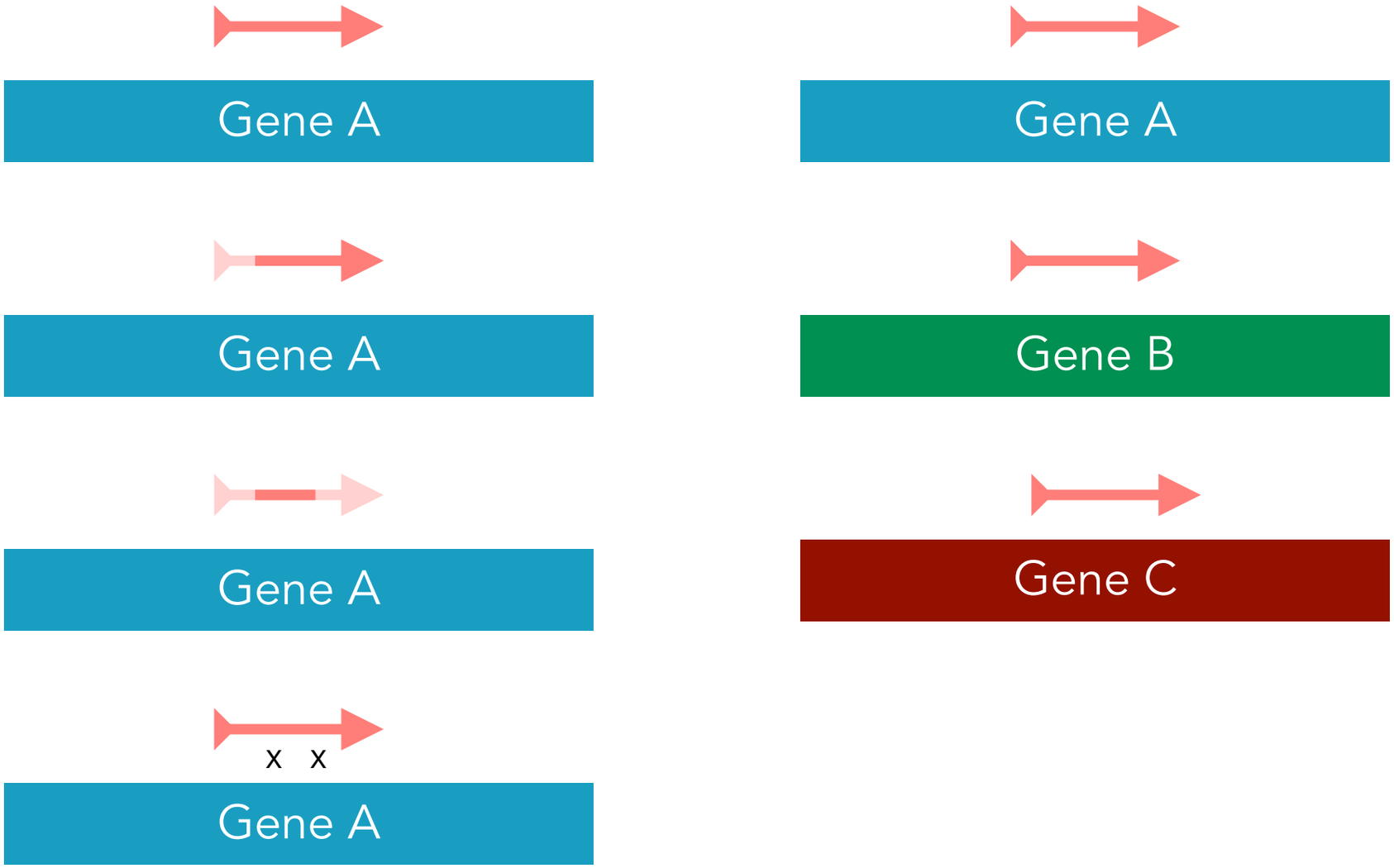
10 reads

$$10 \times 1.5 = 15 \rightarrow \frac{15}{10} = 1.5$$

10 reads

$$10 \times 1.5 = 15 \rightarrow \frac{15}{7} = 2.1$$

Mapping Quality



GDC Genetic Diversity Centre Zurich

Traget Coverage

20 reads

Gene A

$$20 \times 1.5 = 30 \rightarrow \frac{30}{10} = 3$$

20 reads

Gene A

$$20 \times 1.5 = 30 \rightarrow \frac{30}{10} = 3$$

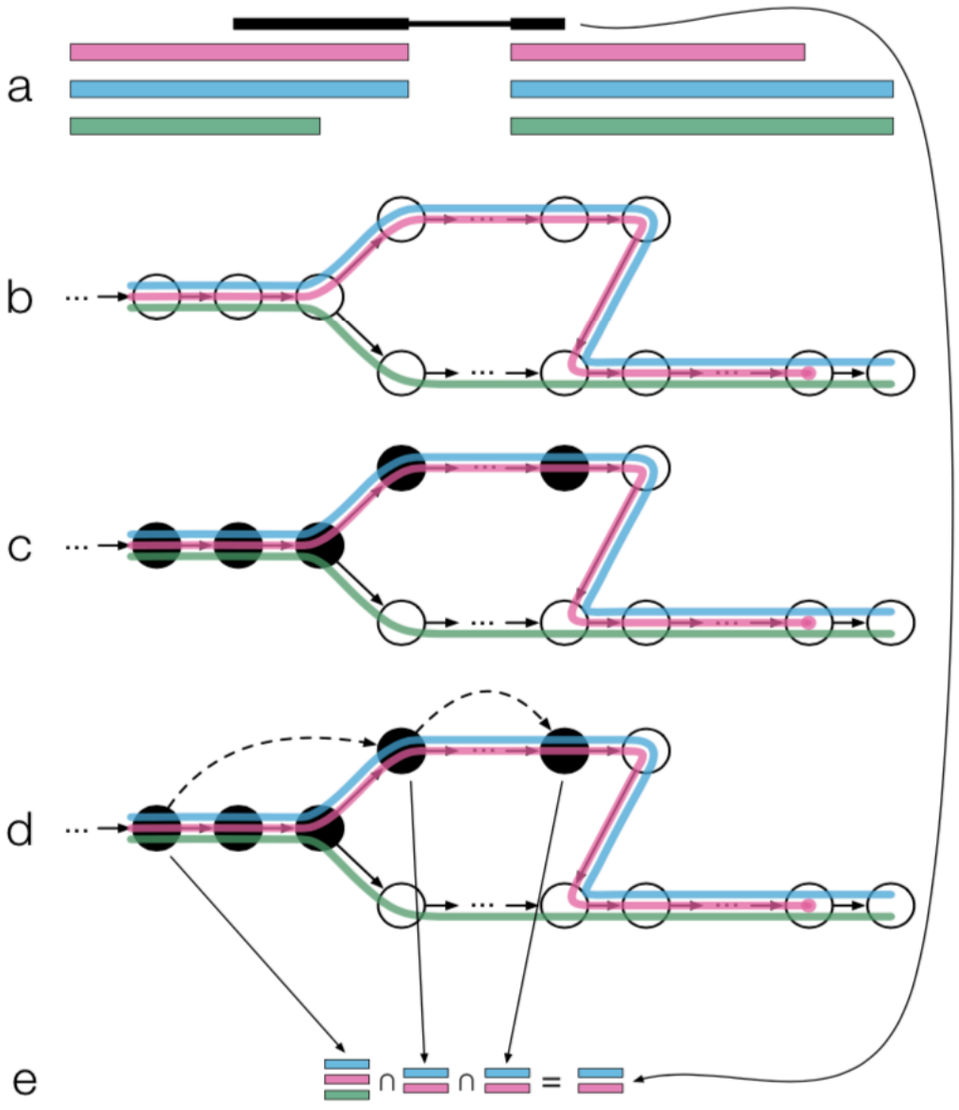
20 reads

Gene A

$$20 \times 1.5 = 30 \rightarrow \frac{30}{10} = 3$$

GDC Genetic Diversity Centre Zurich

Pseudo-Alignment



(**a**) An example of a read (in black) and three overlapping transcripts with exonic regions as shown.

(**b**) An **index** is constructed by creating the transcriptome **de Bruijn Graph** (T-DBG) where nodes (v_1, v_2, v_3, ...) are *k*-mers, each transcript corresponds to a colored path as shown and the path cover of the transcriptome induces a *k*-compatibility class for each *k*-mer.

(**c**) Conceptually, the *k*-mers of a read are hashed (black nodes) to find the *k*-compatibility class of a read.

(**d**) Skipping (black dashed lines) uses the information stored in the T-DBG to skip *k*-mers that are redundant because they have the same *k*-compatibility class.

(**e**) The *k*-compatibility class of the read is determined by taking the intersection of the *k*-compatibility classes of its constituent *k*-mers.

Source: Bray et al. (2016) Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology.

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

RNA-Seq data is …

(a) **compositional** (multiple parts of non-negative numbers).

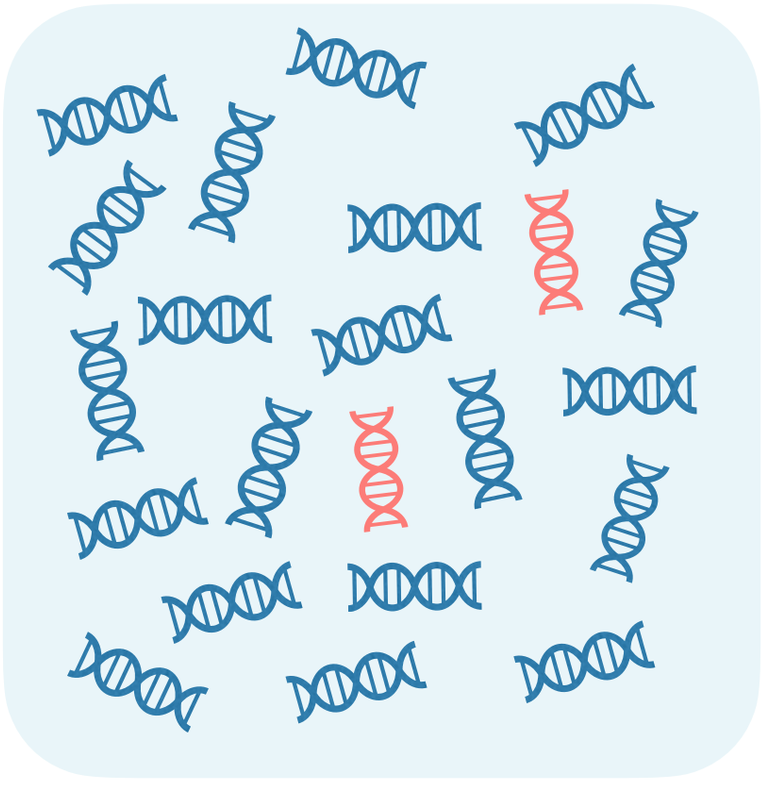
(b) **high dimensional** (many variables/genes)

and **underdetermined** (the number of genes is much greater than the number of samples).

(c) **overdispersed** (variance of the counts of read is larger than expected).

(d) often spares with **many zeros** (zero-inflated).

Sequencing Depths

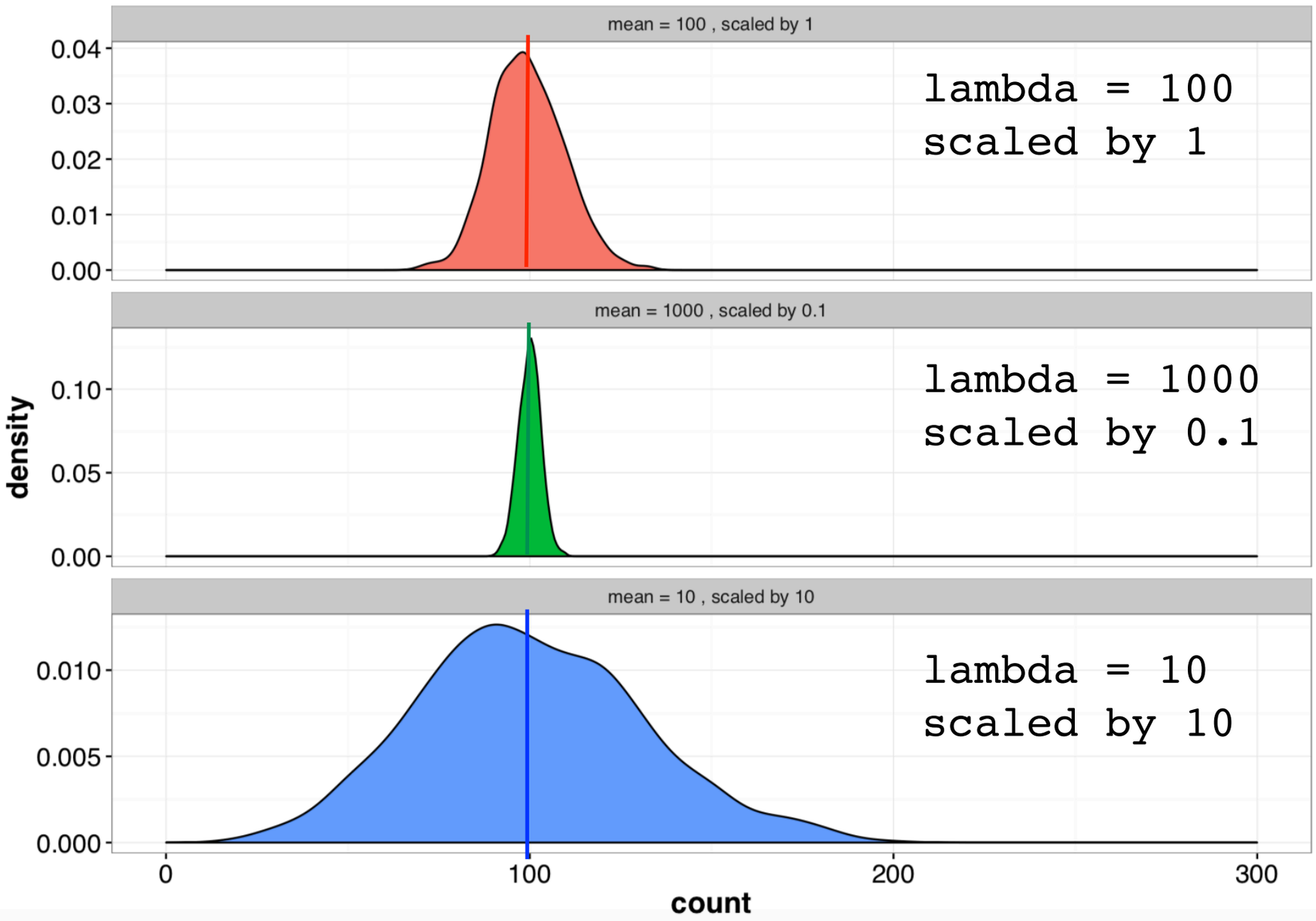


Subsamples:

N=1	➜ n=1 n=0	N=1	➜ n=0 n=1
N=5	➜ n=5 n=0	N=5	➜ n=2 n=1
N=10	➜ n=8 n=2	N=10	➜ n=8 n=2

Sequencing depth & Compsitionality - Technical variation during sequencing results in varying sequencing depths. To reduce/remove sequencing depth variation, counts should be normalized. As a result, we are dealing with compositional rather than absolute data.

Poisson distributed variables with different means, scaled to mean = 100



Sparsity - RNA-Seq data is zero-rich. While log-ratios (network inference in gneral) can be used to tackle compositionality it is sensitive to zeros (i.e. negative infinities). Pseudocounts could resolve the issue but might impact the results as they alter the covariance structure of data. Alternive treatments of zeros have been proposed but are problematic since zeros could indicate absence or undersampling.

```
set.seed(200617)
x1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
y1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
cor(x1,y1)
# 0.883
x2 <- sample(1:100, 10, replace = TRUE)
y2 <- sample(1:100, 10, replace = TRUE)
cor(x2,y2)
# 0.466
x3 <- c(x2, rep(0,20))
y3 <- c(y2, rep(0,20))
cor(x3,y3)
# 0.790
```

GDC Genetic Diversity Centre Zurich

easyRNASeq

DEGseq

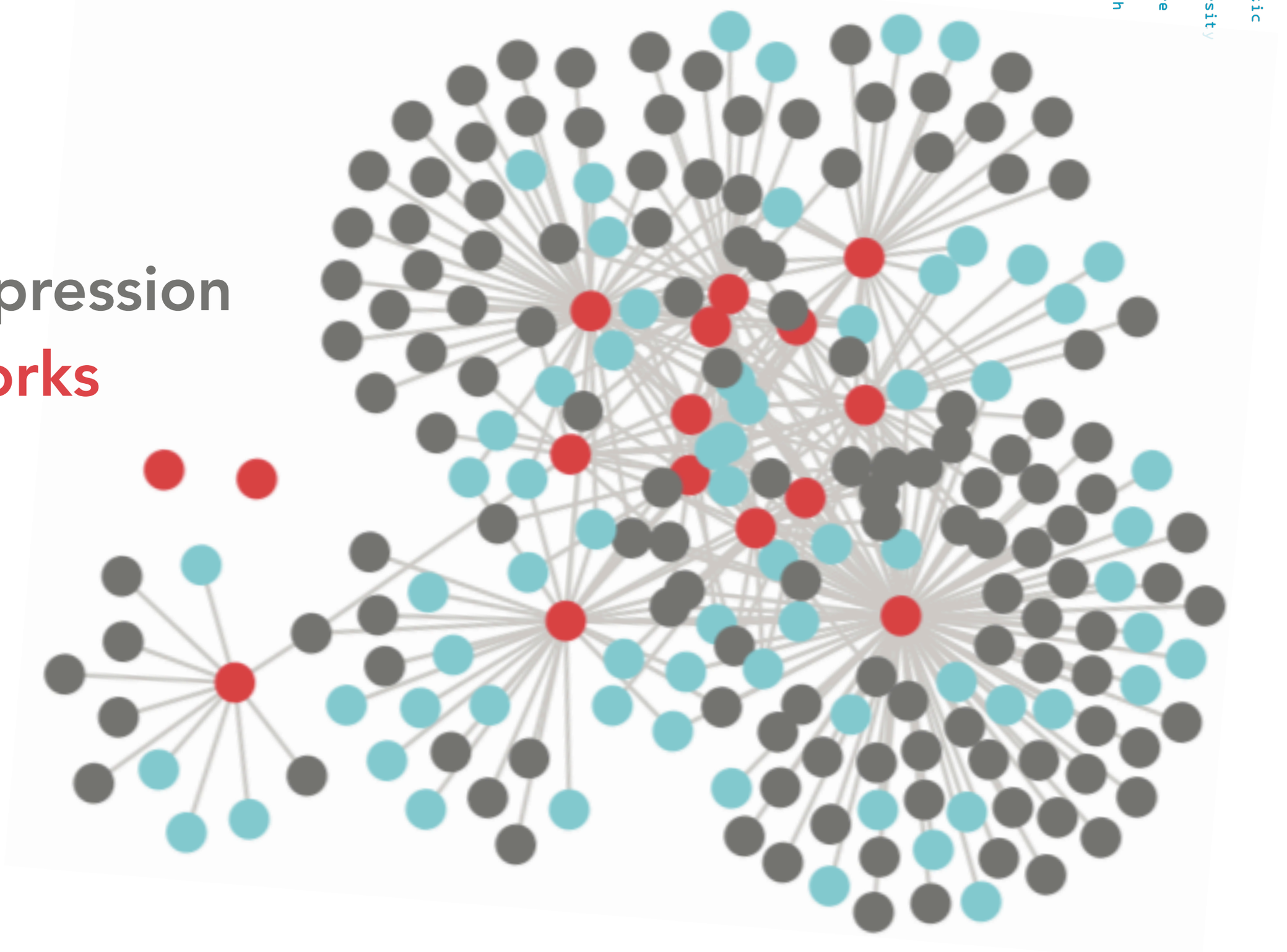
DESeq / DESeq2

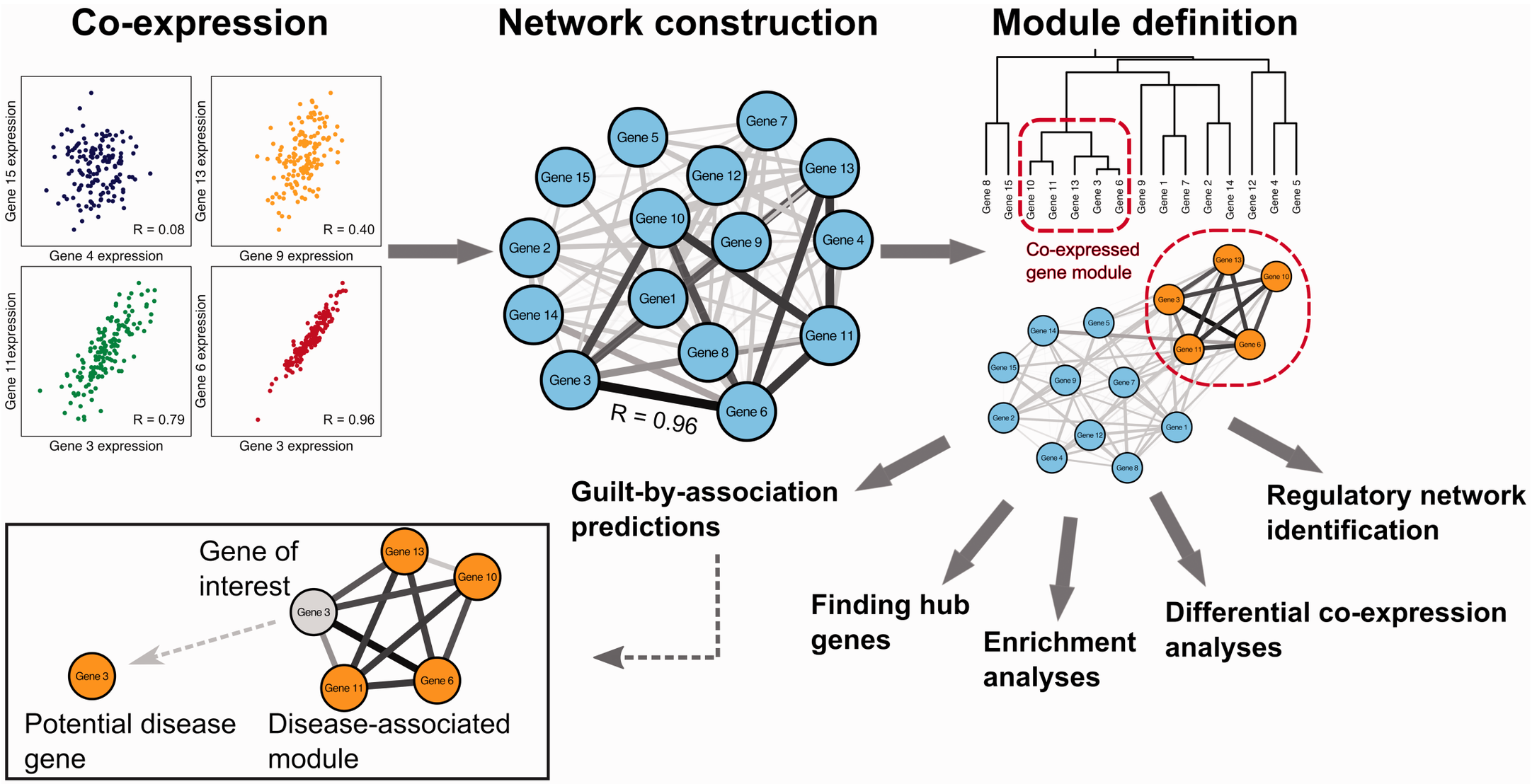
NOISeq

edgeR

baySeq

Gene Co-Expression Networks





Example of a co-expression network analysis. First, pairwise correlation is determined for each possible gene pair in the expression data. These pairwise correlations can then be represented as a network. Modules within these networks are defined using clustering analysis. The network and modules can be interrogated to identify regulators, functional enrichment and hub genes. Differential co-expression analysis can be used to identify modules that behave differently under different conditions. Potential disease genes can be identified using a guilt-by-association (GBA) approach that highlights genes that are co-expressed with multiple disease genes.

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**."

SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY BIOLOGY

The genetic mechanism of selfishness and altruism in parent-offspring coadaptation

Min Wu[1]*, Jean-Claude Walser[2], Lei Sun[3†], Mathias Kölliker[1*‡]

The social bond between parents and offspring is characterized by coadaptation and balance between altruistic and selfish tendencies. However, its underlying genetic mechanism remains poorly understood. Using transcriptomic screens in the subsocial European earwig, *Forficula auricularia*, we found the expression of more than 1600 genes associated with experimentally manipulated parenting. We identified two genes, *Th* and *PebIII*, each showing evidence of differential coexpression between treatments in mothers and their offspring. In vivo RNAi experiments confirmed direct and indirect genetic effects of *Th* and *PebIII* on behavior and fitness, including maternal food provisioning and reproduction, and offspring development and survival. The direction of the effects consistently indicated a reciprocally altruistic function for *Th* and a reciprocally selfish function for *PebIII*. Further metabolic pathway analyses suggested roles for *Th*-restricted endogenous dopaminergic reward, *PebIII*-mediated chemical communication and a link to insulin signaling, juvenile hormone, and vitellogenin in parent-offspring coadaptation and social evolution.

A Quick Recap

1

Question

Start with a precises scientific question.

Gather Knowledge

What do you know, what do you have and what would you still need?

2

3

Design

Think carfully about the desing and do not just use the newest technology or cheapest solution.

Pilots

A few well designed tests might be a good investment.

4

5

Replicates

Always use biological replicates.