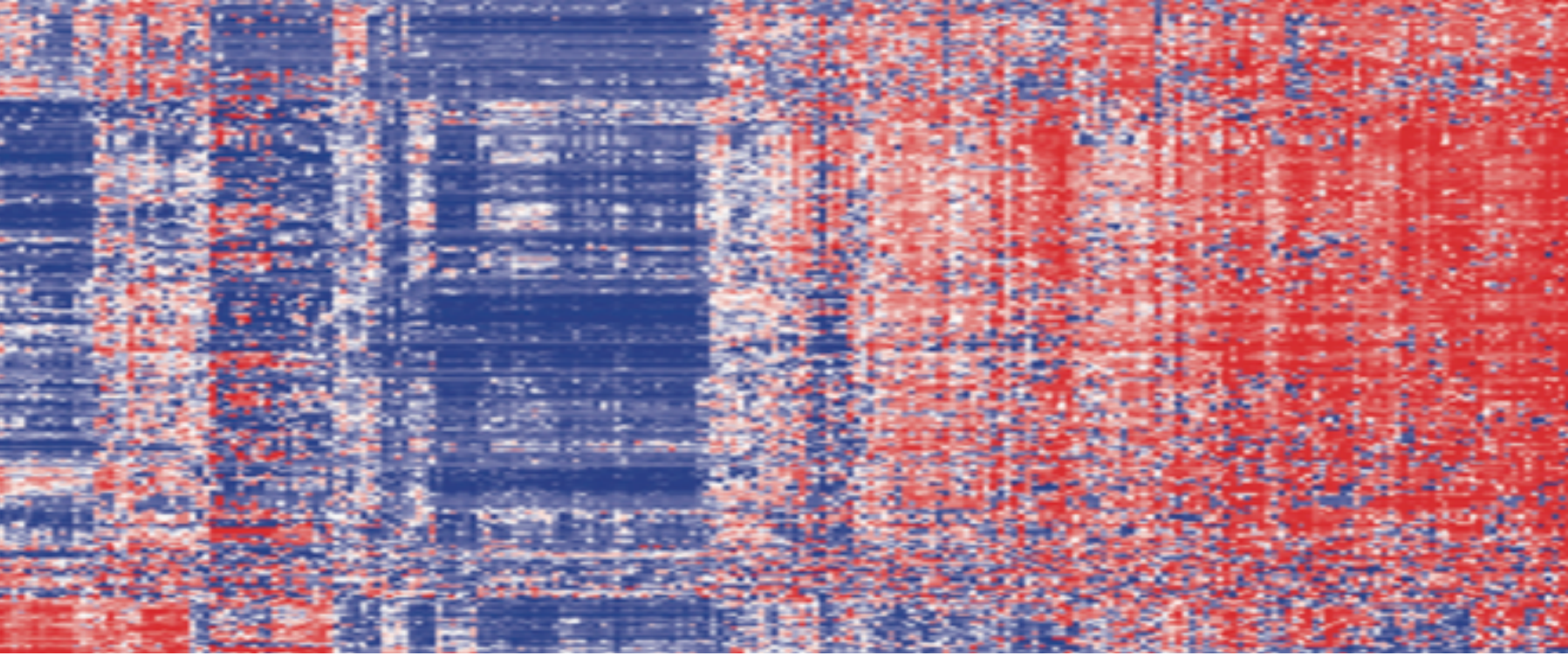


PROSPECTS OF RNA-SEQ

Jean-Claude Walser

GDC

Genetic Diversity Centre Zurich

http://www.gdc.ethz.ch

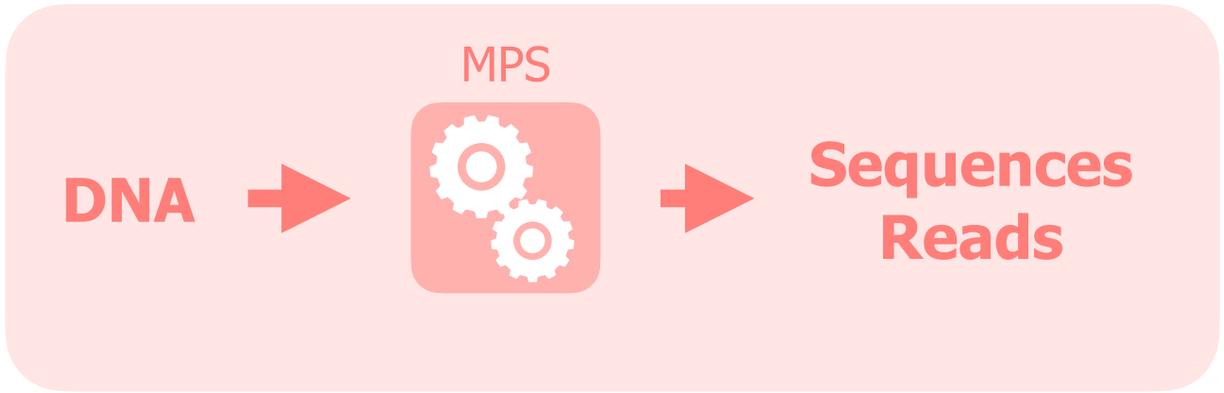
Genetic Diversity Centre Zurich

What is RNA?

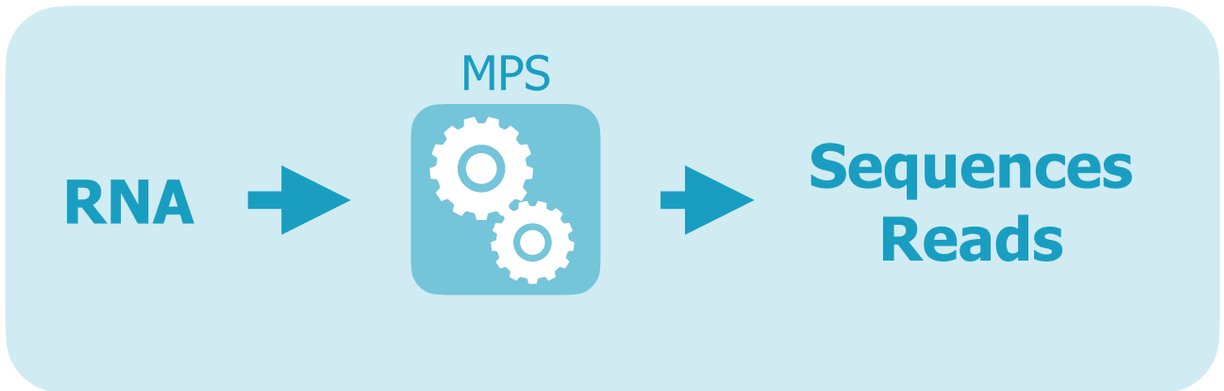
Why would I use RNA?

What is RNA-Seq?

What can I do with RNA-Seq?

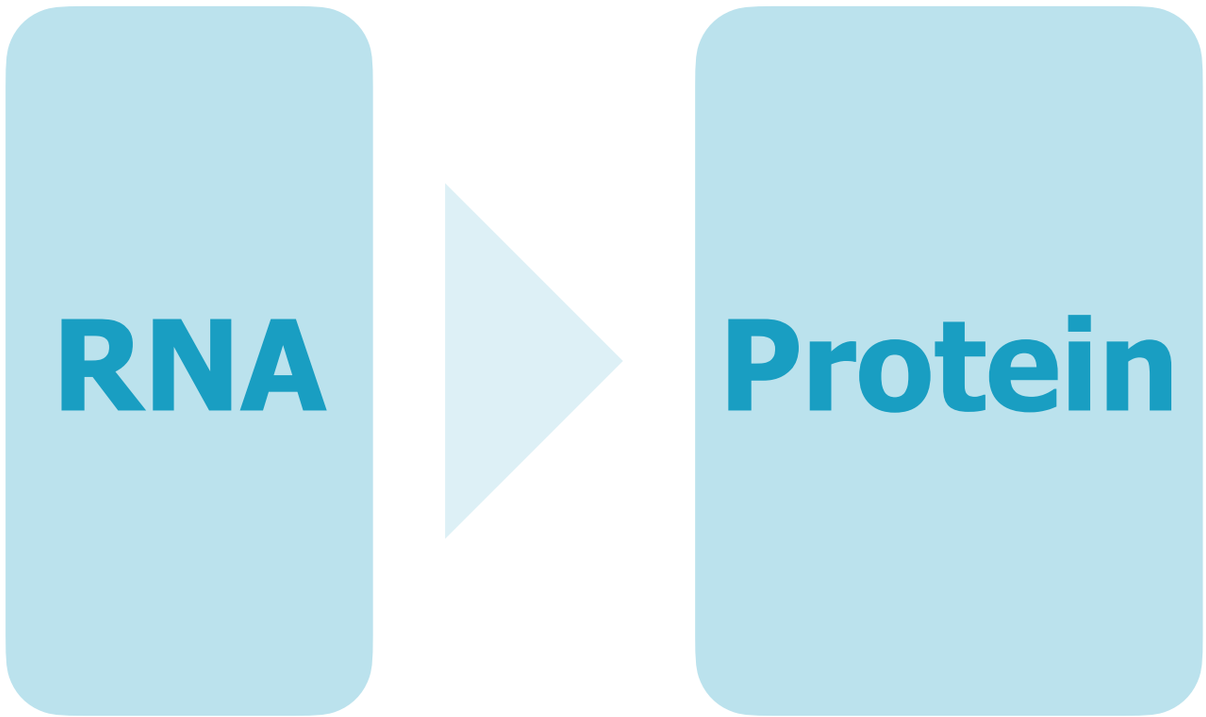


Genome Analysis

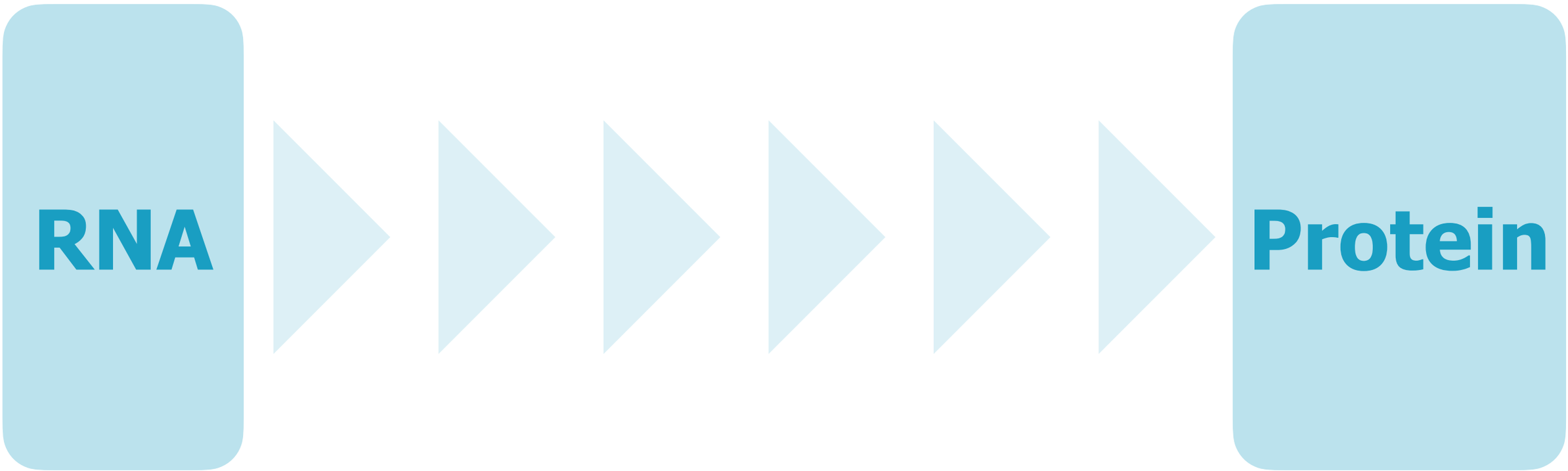


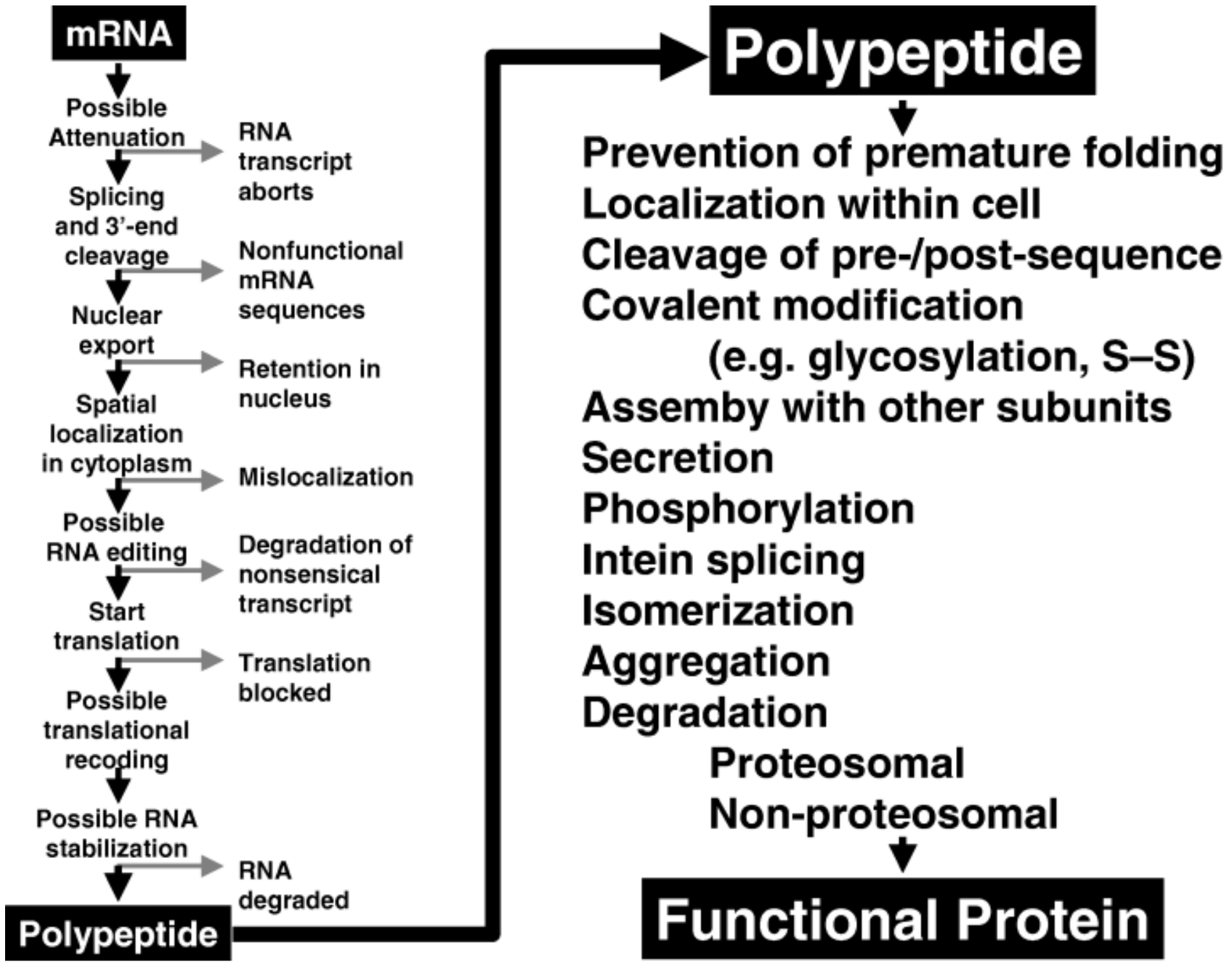
Transcriptome Analysis

MPS == Massive Parallel Sequencing


RNA
Protein

Genetic Diversity Centre Zurich





Feder & Walser (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. Journal of Evolutionary Biology.

GDC Genetic Diversity Centre Zurich

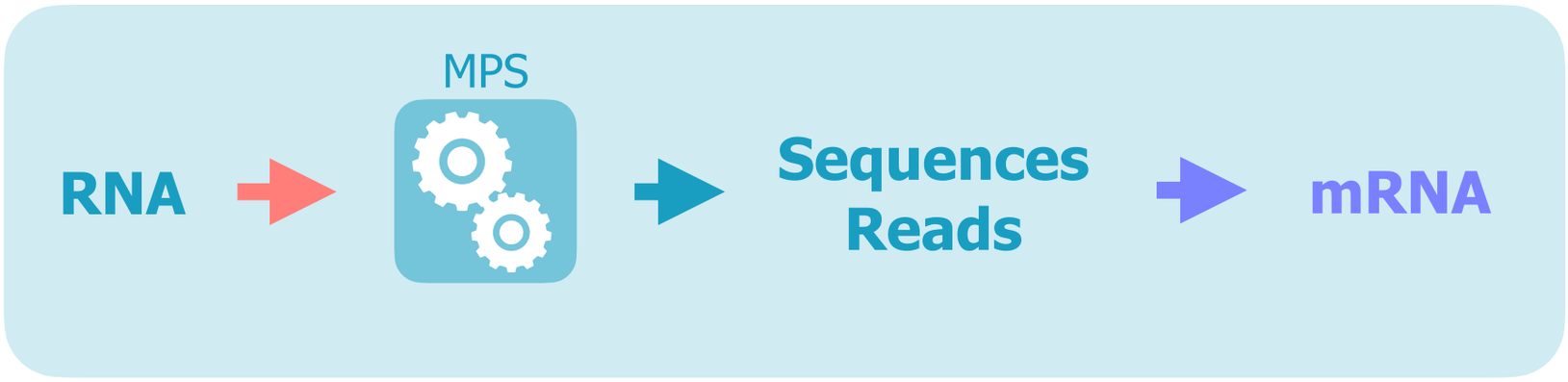
total **RNA**

mRNA

rRNA

tRNA

ncRNA, nmRNA, sRNA, smnRNA, tRNA, sRNA, mRNA, pcRNA, rRNA, 5S rRNA, 5.8S rRNA, SSU rRNA, LSU rRNA, NoRC RNA, pRNA, 6S RNA, SsrS RNA, aRNA, asRNA, asmiRNA, cis-NAT, crRNA, tracrRNA, CRISPR RNA, DD RNA, diRNA, dsRNA, endo-siRNA, exRNA, gRNA, hc-siRNA, hcsiRNA, hnRNA, RNAi, lincRNA, lncRNA, miRNA, mrpRNA, nat-siRNA, natsiRNA, OxyS RNA, piRNA, qiRNA, rasiRNA, RNase MRP, RNase P, scaRNA, scnRNA, scRNA, scRNA, SgrS RNA, shRNA, siRNA, SL RNA, SmY RNA, snoRNA, snRNA, snRNP, SRP RNA, ssRNA, stRNA, tasiRNA, tmRNA, uRNA, vRNA, vtRNA, Xist RNA, Y RNA, NATs, pre-mRNA, circRNA, msRNA, cfRNA, …



RNA library enrichment strategies

- size selection
- not target removal (e.g., ribosomal RNA)
- target enrichment


total RNA
cleaner RNA
cleanerer RNA
cleanest RNA
enriched RNA

GDC Genetic Diversity Centre Zurich

total RNA

▽

(cDNA)

▽

Library Prep

▽

Sequencing

▽

Quality Control

▽

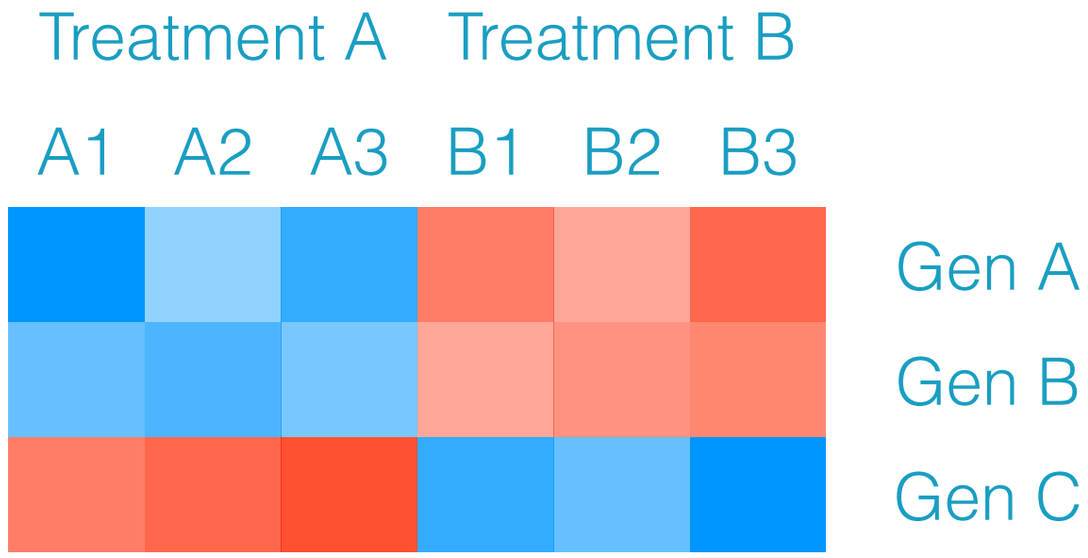
(Mapping)

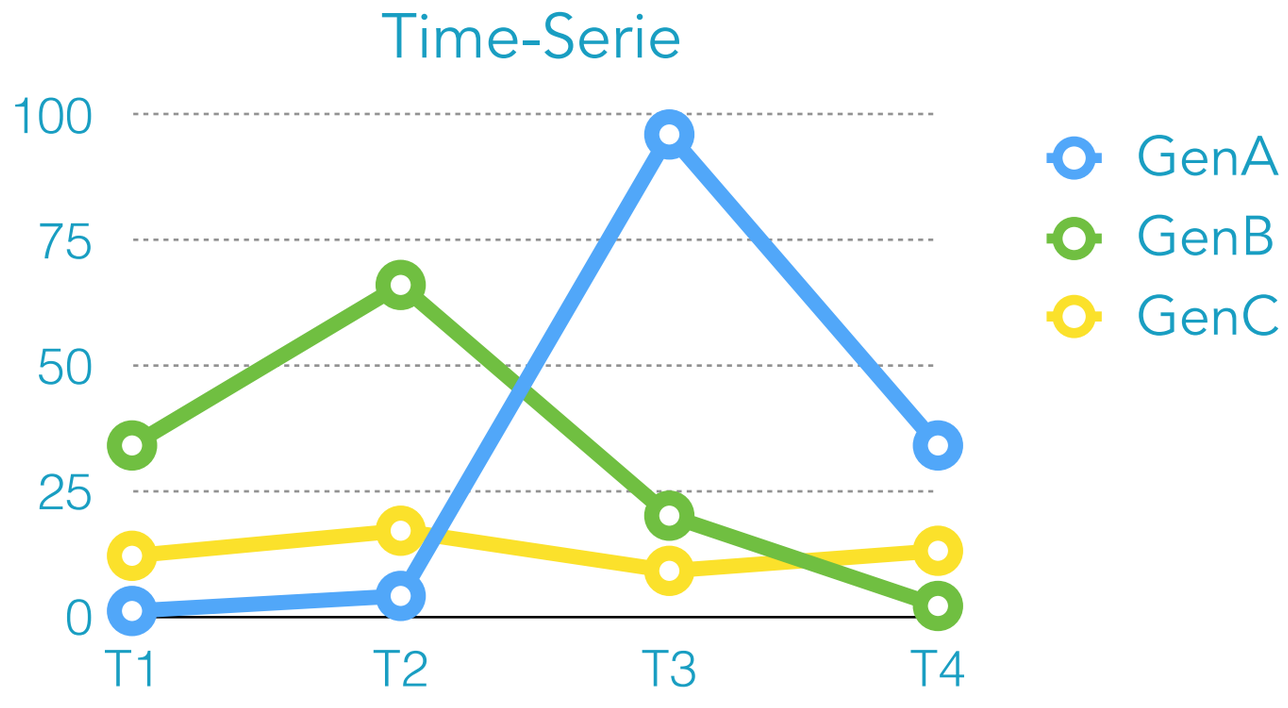
▽

Counts

▽

Analysis





Total RNA

▽

mRNA

▽

Library Prep

▽

Sequencing

▽

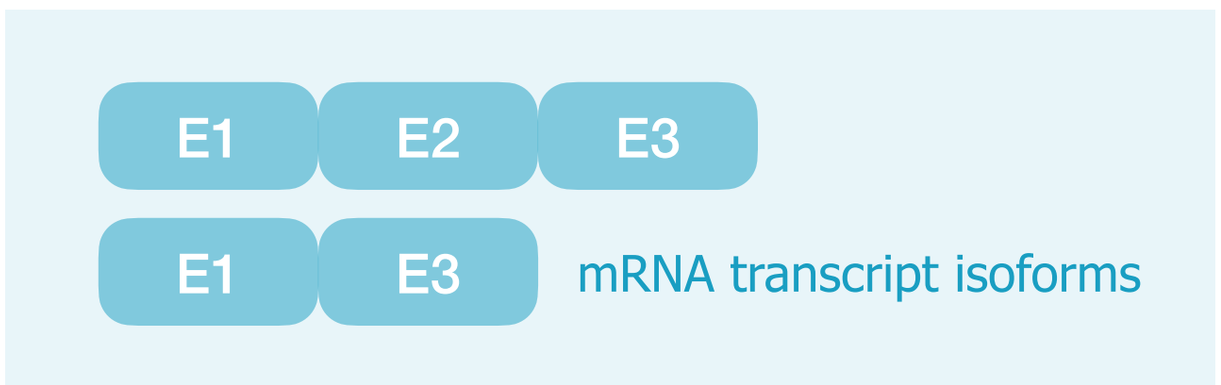
Quality Control

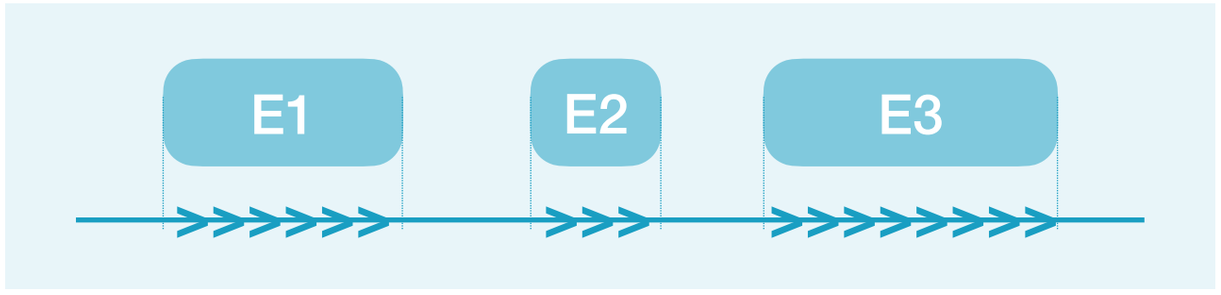
▽

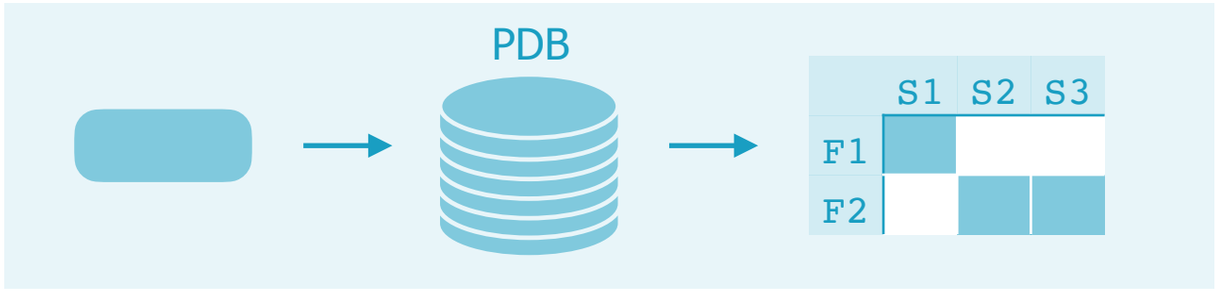
Cluster/Assembly

▽

Analysis







GDC Genetic Diversity Centre Zurich

RNA-Seq Blog

Transcriptome Sequencing Research & Industry News

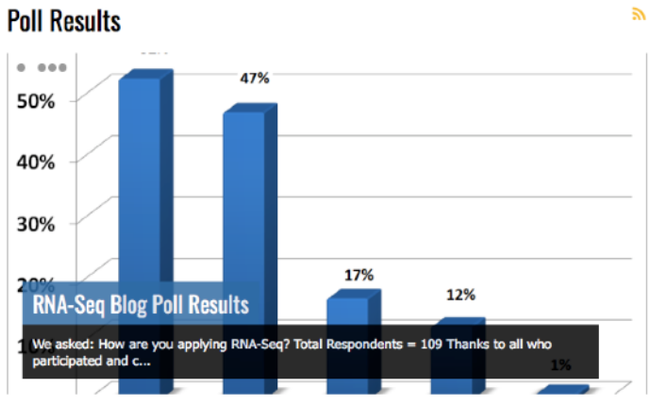
http://www.rna-seqblog.com

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

December 28, 2017 Leave a comment 2,576 Views

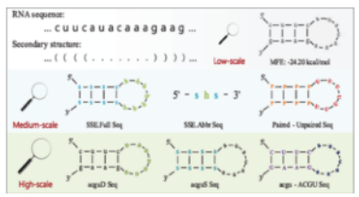
Every nine minutes, someone in the US dies from blood cancer which accounts for about 10 percent of all cancer deaths. And, every three minutes, one person in the US is diagnosed with a blood cancer — about 170,000 people ...

Read More »



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

August 9, 2018 Leave a comment 1,409 Views

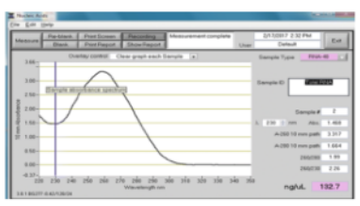


Discovering new long non-coding RNAs (lncRNAs) has been a fundamental step in lncRNA-related research. Nowadays, many machine learning-based tools have been developed for lncRNA identification. However, many methods predict lncRNAs using sequence-derived features alone, which tend to display unstable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

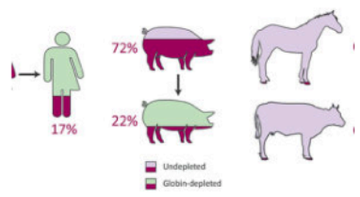
HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

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 »

Experimental Design

GDC Genetic Diversity Centre Zurich

RNA-Seq Hype

Massive Parallel Sequencing Hype

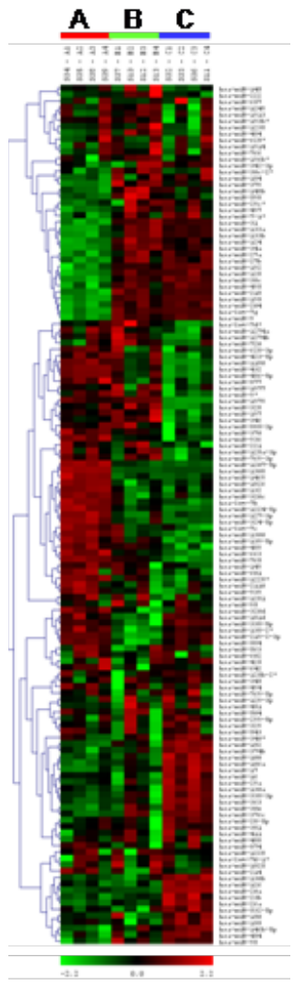
Experiement

RNA

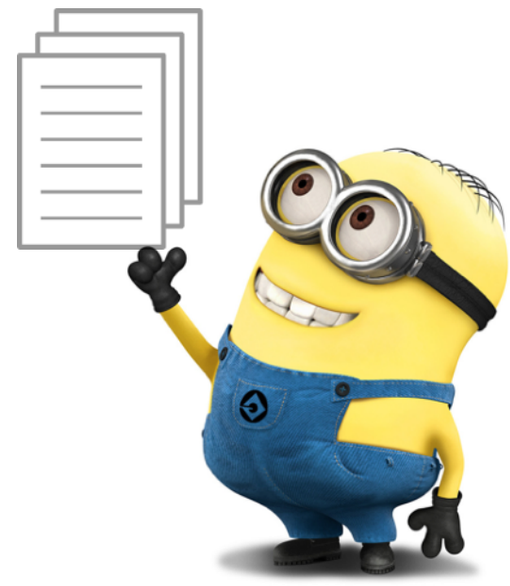
Sequening

Data

Analysis



Publication(s)



GDC Genetic Diversity Centre Zurich

RNA-Seq Reality

Massive Parallel Sequencing Reality

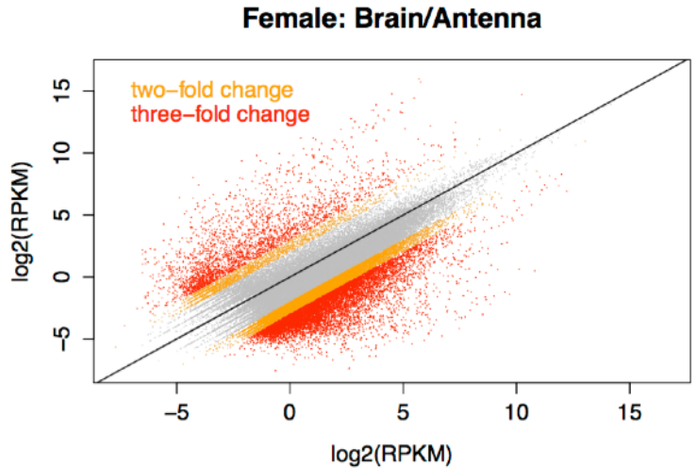
Setup

Quality Amount

Method

Storage

Interpretation



What do you know about our transcriptome?

- Number of (well-)validated genes?
- What is the percentage of genes not encoding proteins?
- How frequent is alternative splicing?
- Average alternative transcribed forms?

Number of (well-)validated genes: **>55,000**

Percentage of genes not encoding proteins: **majority**

Percentage of alternative splicing: **almost all**

Average alternative transcribed forms: **>9**

Flavor is a balance of **acidity** and **sugar**, plus the influence of elusive **volatile compounds.** Regardless of which variety you grow, how you grow a tomato and external factors (e.g., temperature) impact flavor.

Treatment #1 t_1=27°C / t_2=15°C

Treatment #2 t_1=29°C / t_2=18°C

GDC Genetic Diversity Centre Zurich

Molecular BioSystems

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

CrossMark click for updates

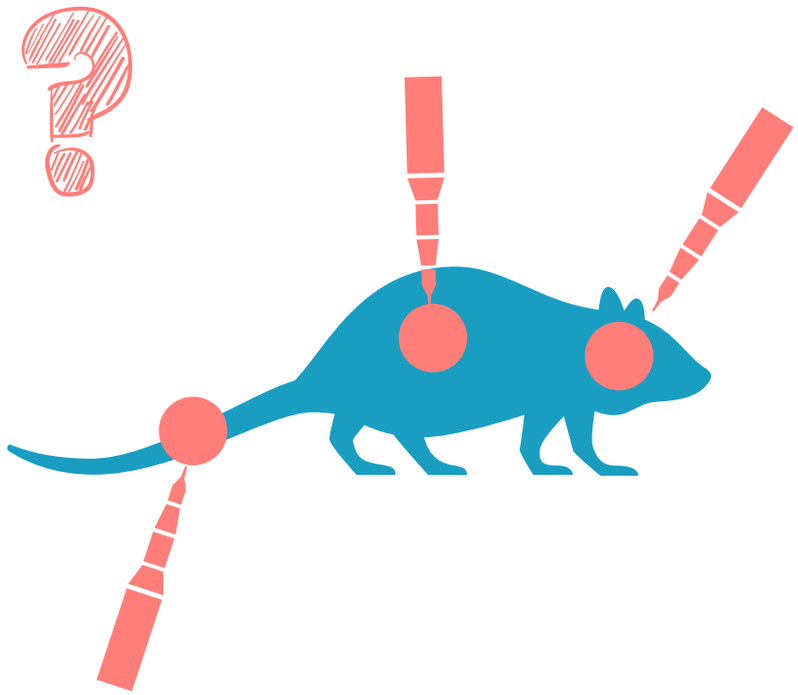
Cite this: *Mol. BioSyst.*, 2016, 12, 508

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]

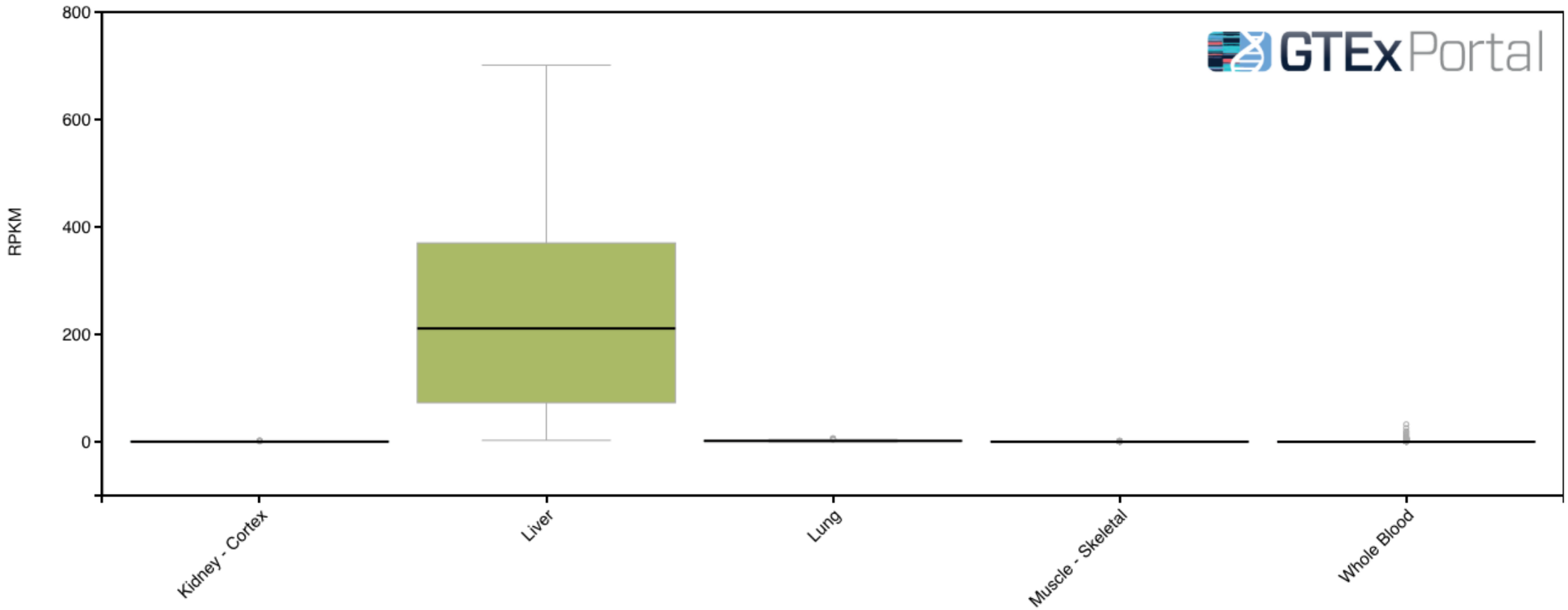
Lactobacillus delbrueckii subsp. bulgaricus 2038 is an industrial bacterium that is used as a starter for dairy products. … Here, we utilized RNA-seq to explore the transcriptome of *Lb. bulgaricus 2038* from four different growth phases under whey conditions. The most abundantly expressed genes in the 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



ADH1A Gene Expression

ADH: **A**lcohol **D**e**h**ydrogenase



http://www.gtexportal.org/home/

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 / orientation

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)

Design | Preparation | Methode | Analysis | Extras

How much sequencing is needed?

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 you have?

GDC Genetic Diversity Centre Zurich

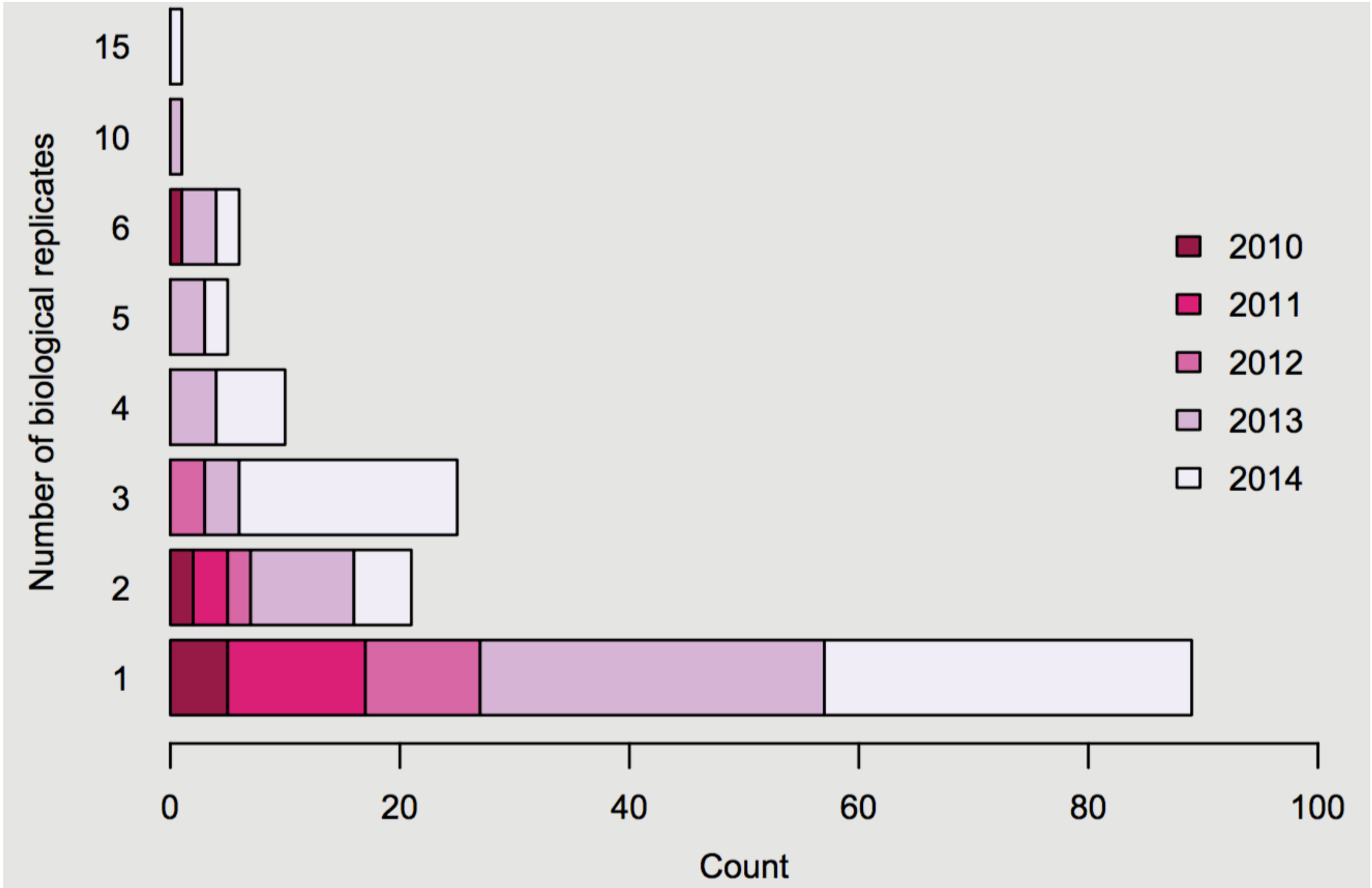
Signal-to-noise ratio

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

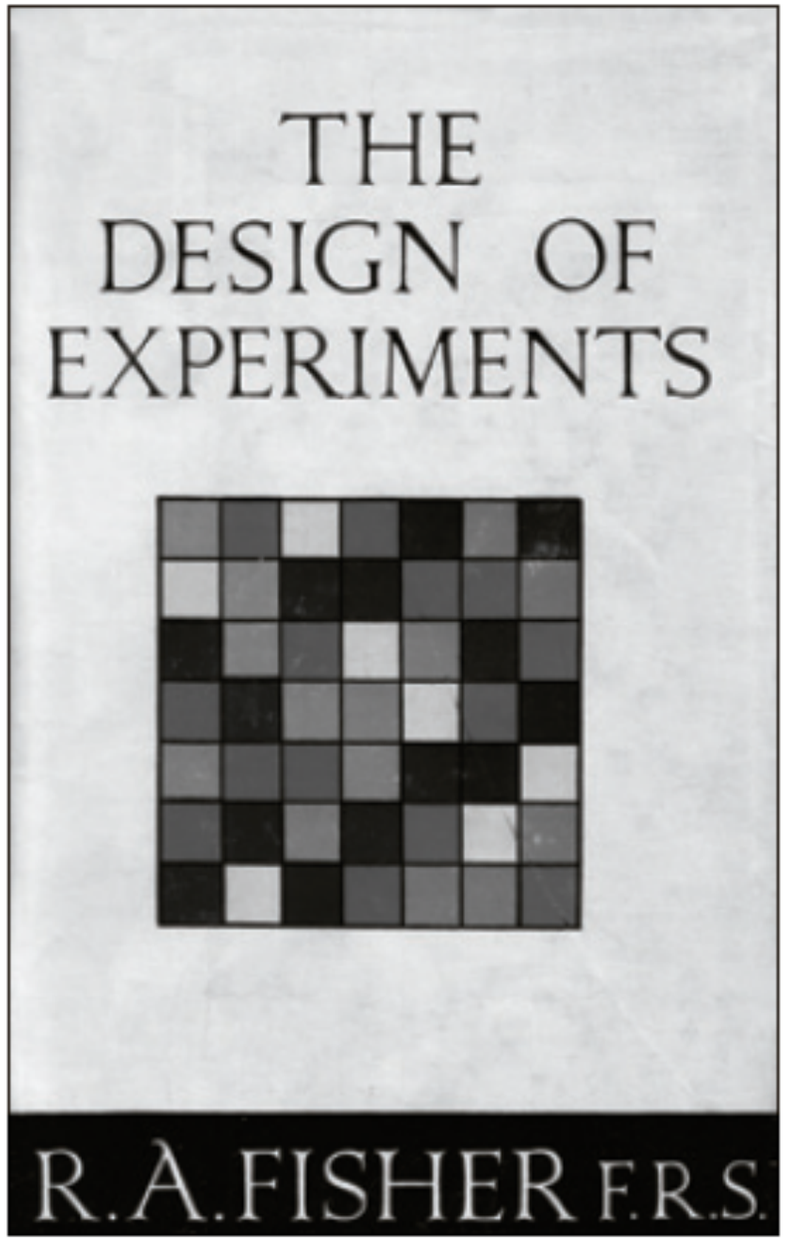
Poisson counting errors - The uncertainty inherited in any count-based measurements.

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

Biological variance - The natural variation in gene expression measurements.



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

GDC Genetic Diversity Centre Zurich

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:

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

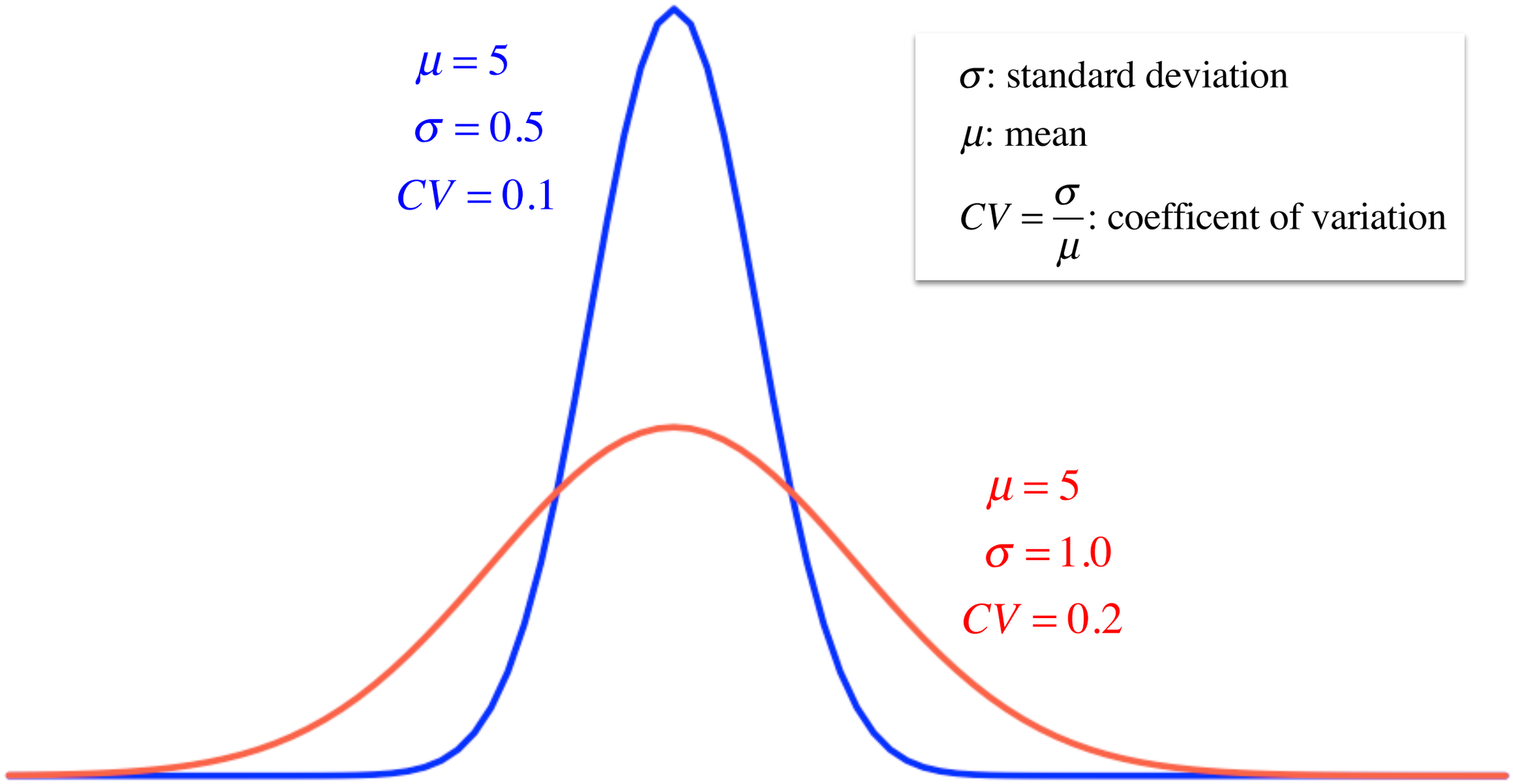
(2) effect size (e.g. fold-change)

(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



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

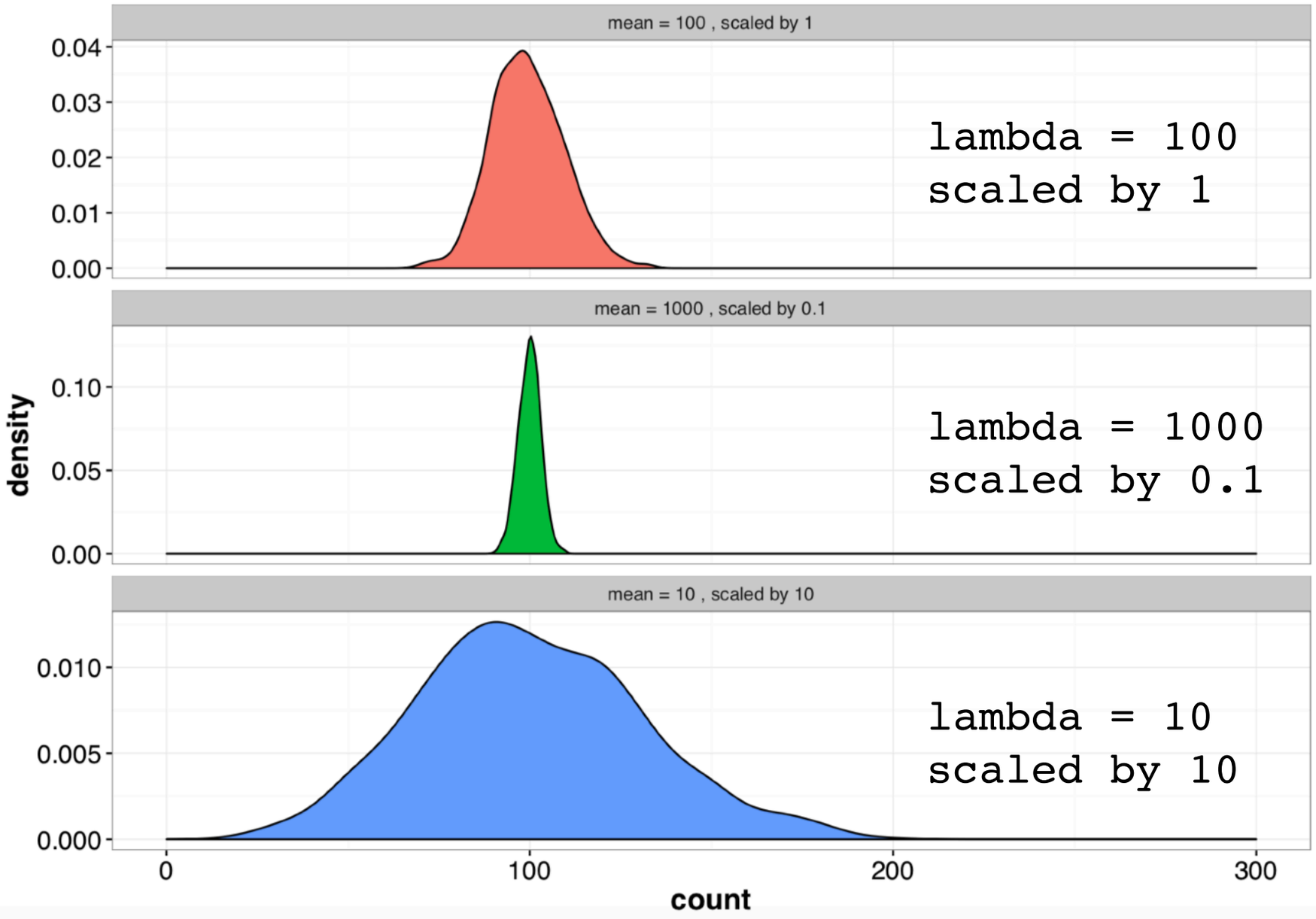
λ: average number of event per interval (= mean)

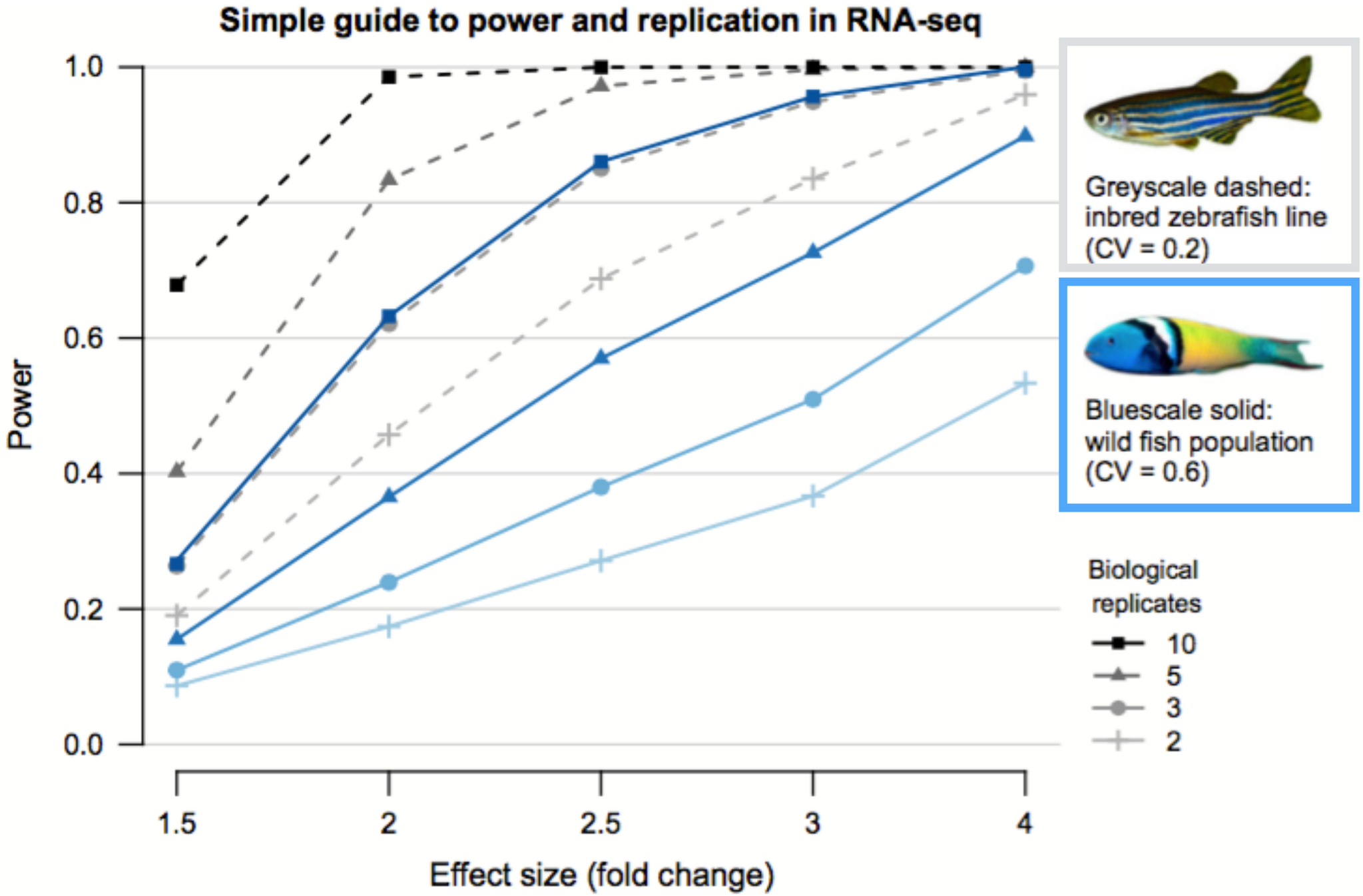
σ: standard deviation (= $\sqrt{Variance}$)

μ: mean

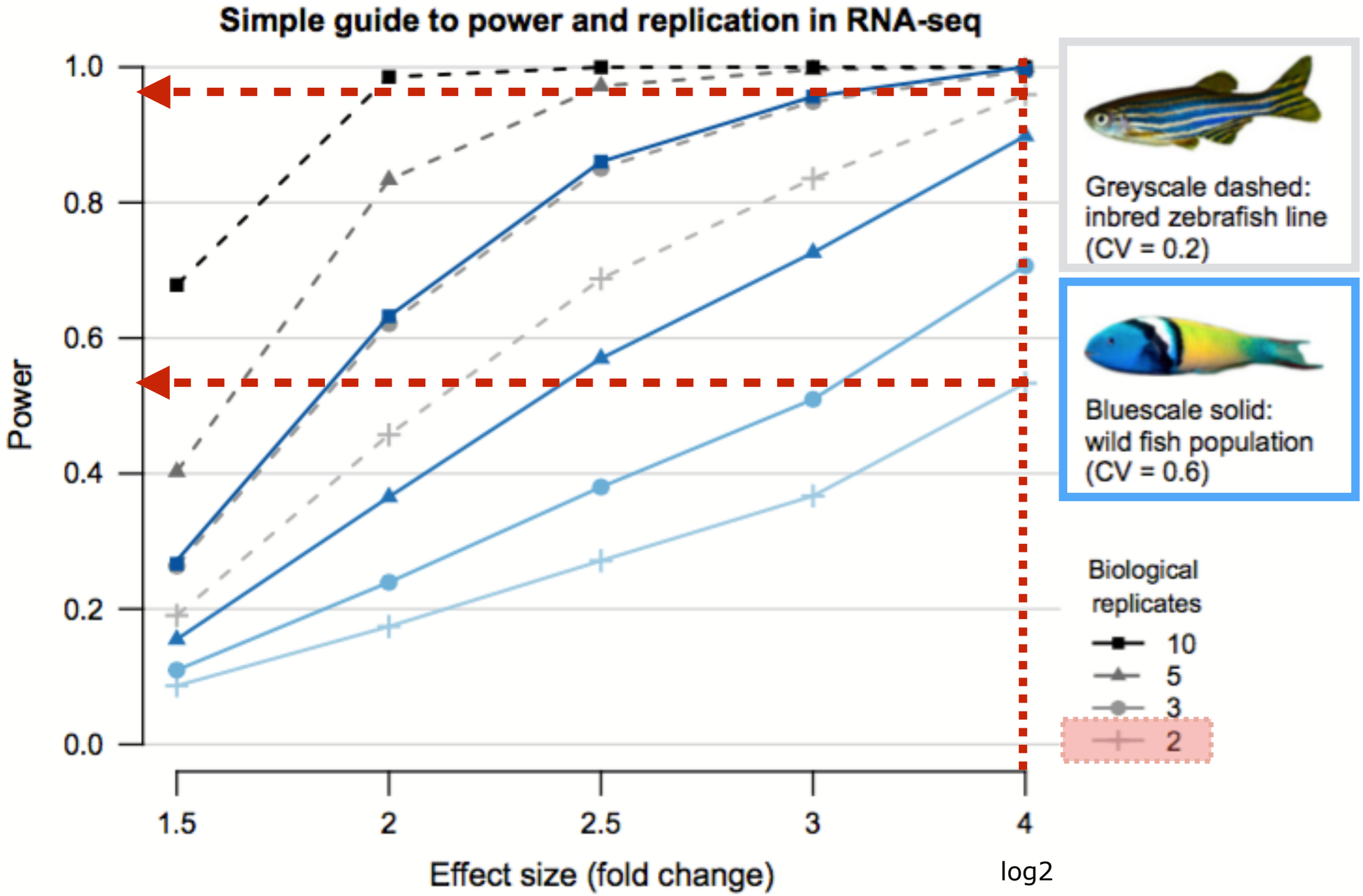
The expected value and variance of a Poisson-distributed random variable are both equal to λ.

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

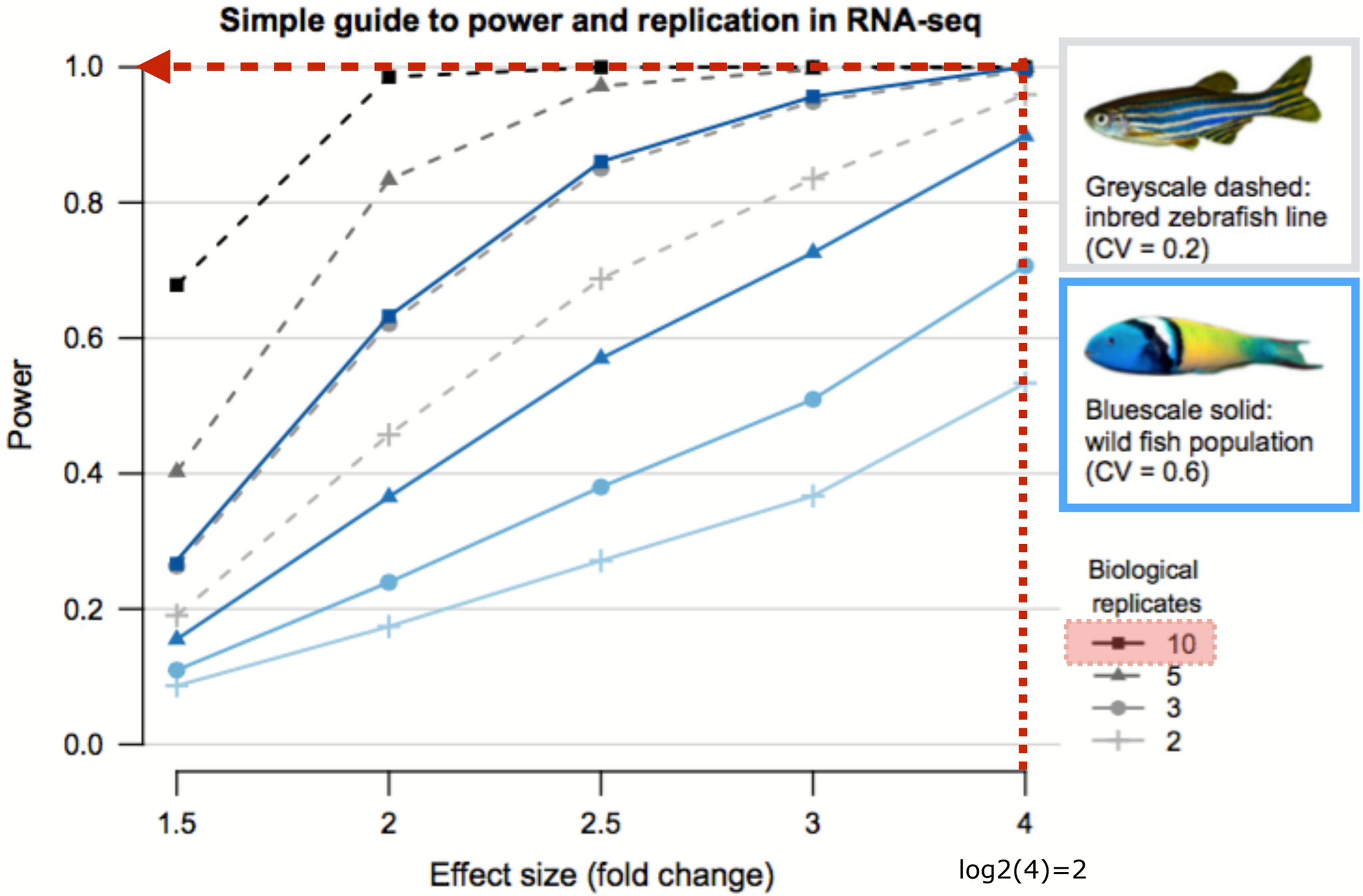




Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.



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



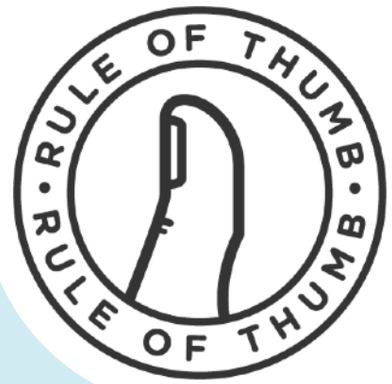
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. Molecular Ecology, 25, 1224–1241.

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

☛ available data set

☛ pilot sequencing

GDC Genetic Diversity Centre Zurich



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* TRANSCRIPTOME ASSEMBLY

Is RNA-Seq still sexy?

INNOVATION

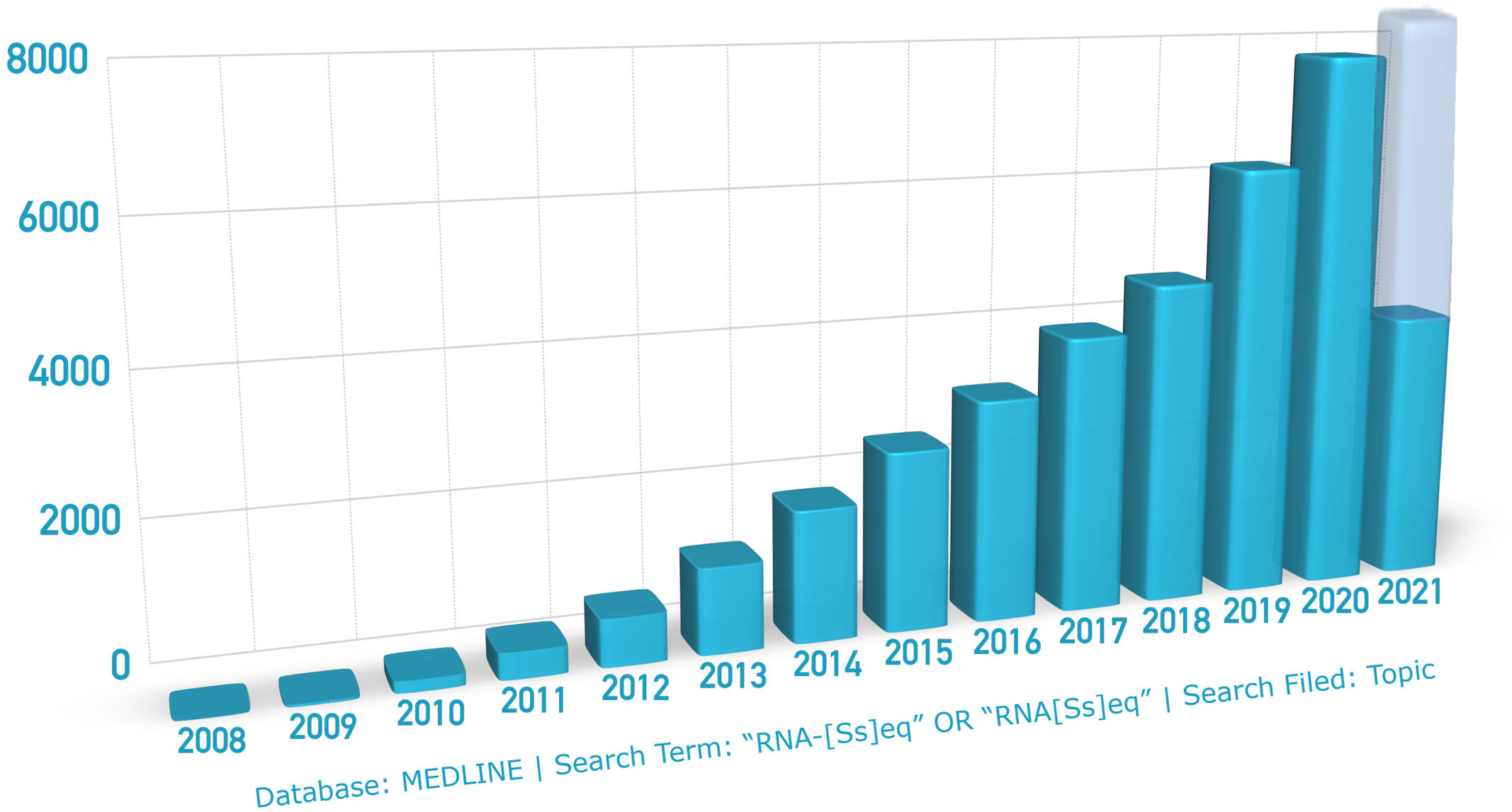
RNA-Seq: a revolutionary tool for transcriptomics

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.

RNA-Seq ▷ Antiquated or not?



GDC Genetic Diversity Centre Zurich

DISCOVER FULL-LENGTH TRANSCRIPTS

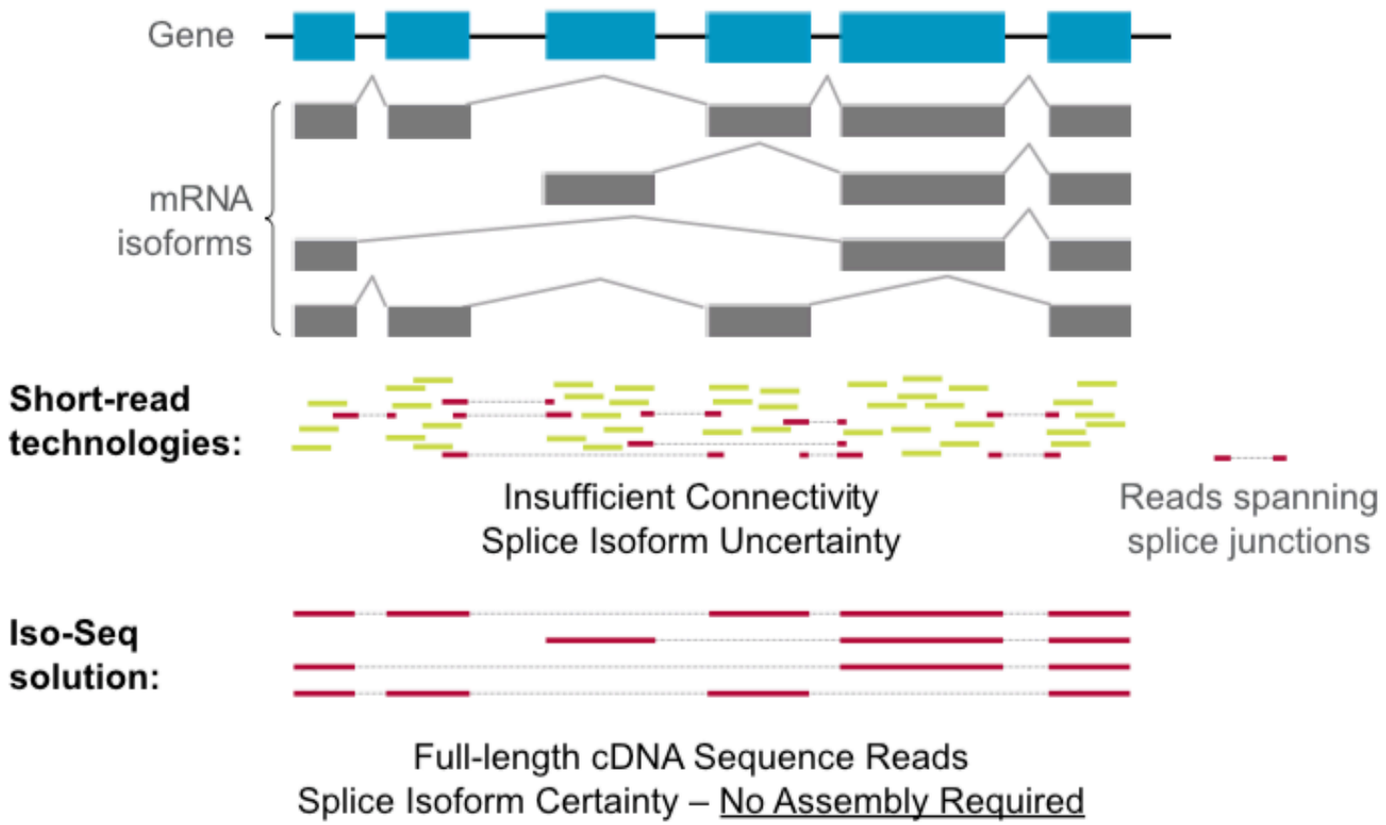
Get a complete view of transcript isoform diversity with PacBio long-read sequencing.

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

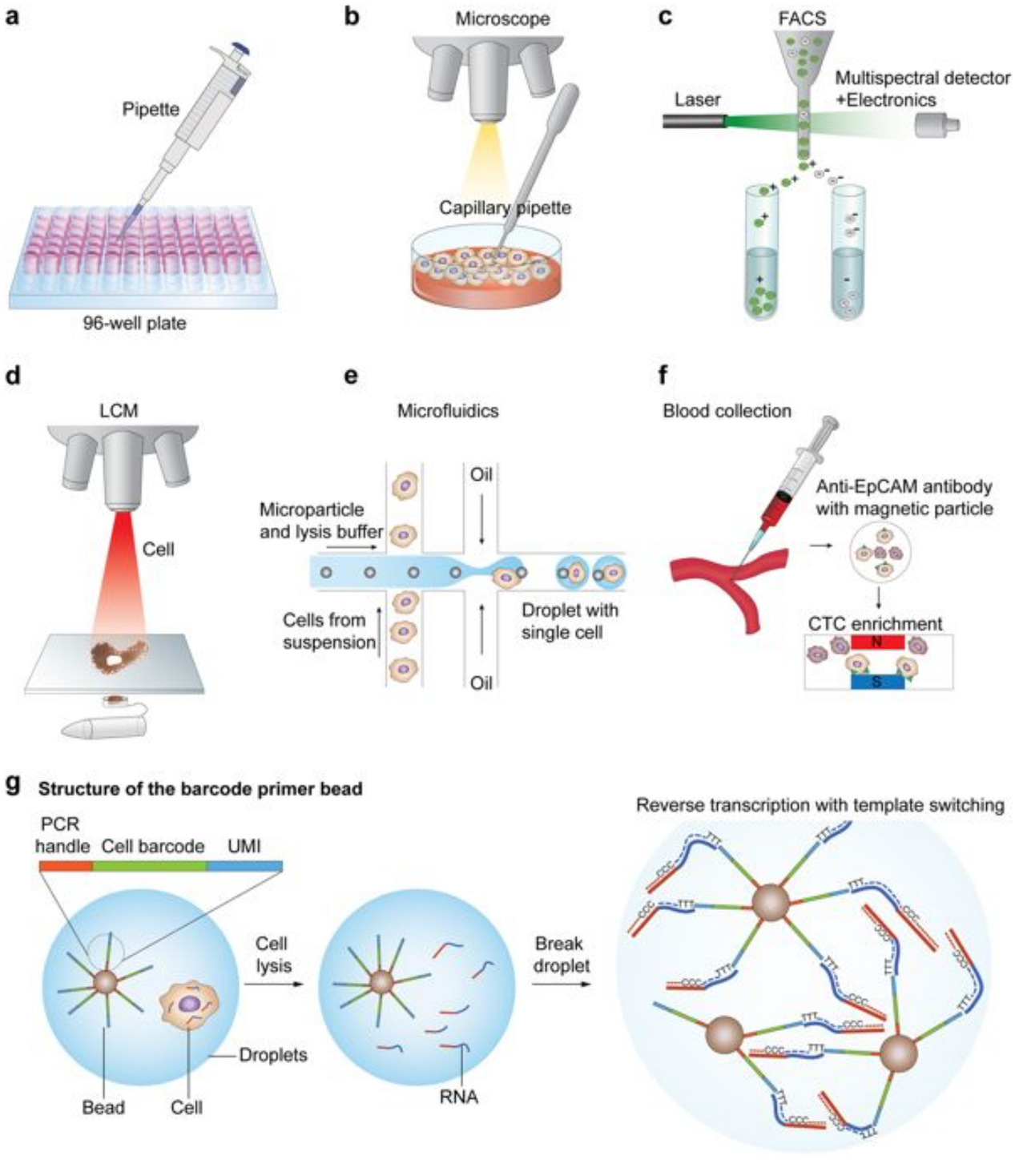
DETERMINATION OF TRANSCRIPT ISOFORMS



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.

GDC Genetic Diversity Centre Zurich

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

Published online 25 July 2016

Nucleic Acids Research, 2016, Vol. 44, No. 19 **e148**
doi: 10.1093/nar/gkw655

SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence

Hélène Lopez-Maestre[1,2], Lilia Brinza[3], Camille Marchet[4], Janice Kielbassa[5], Sylvère Bastien[1,2], Mathilde Boutigny[1,2], David Monnin[1], Adil El Filali[1], Claudia Marcia Carareto[6], Cristina Vieira[1,2], Franck Picard[1], Natacha Kremer[1], Fabrice Vavre[1,2], Marie-France Sagot[1,2] and Vincent Lacroix[1,2,*]

[1]Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France, [2]EPI ERABLE - Inria Grenoble, Rhône-Alpes, [3]PT Génomique et Transcriptomique, BIOASTER, Lyon, France, [4]Université de Rennes, F-35000 Rennes; équipe GenScale, IRISA, Rennes, [5]Synergie-Lyon-Cancer, Universite Lyon 1, Centre Leon Berard, Lyon, France and [6]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil

Zurich Centre Genetic Diversity GDC

A Quick Recap

Question

Start with a precises scientific question.

Gather Knowledge

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

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.

Replicates

Always use biological replicates.

THE END