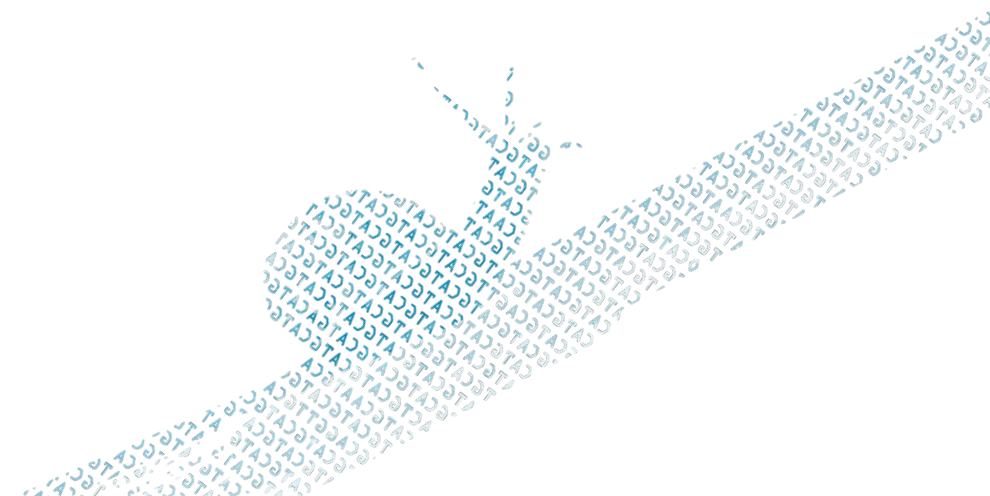




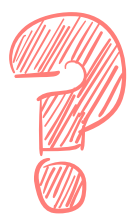
Genetic Diversity: Analysis

RNA-Seq

Thursday, 24. June 2025



Archaeogenetics has revolutionised the study of ancient diseases by providing direct genetic evidence of past infections such as **plague** (*Yersinia pestis*), **tuberculosis** (*Mycobacterium tuberculosis*) or **hepatitis** B virus (HBV), allowing researchers to trace the evolution and spread of pathogens through time. These discoveries not only improve our understanding of historical diseases, but also provide valuable insights into the co-evolution of humans and pathogens.



Why do we have historical evidence going back thousands of years for diseases like the **plague**, **tuberculosis**, or **hepatitis** but not **influenza**?

Stability and Preservation of Ancient DNA vs. RNA

DNA is chemically more stable than RNA, making it more likely to be preserved in ancient remains under favourable conditions.

Due to RNA's instability, direct evidence of ancient RNA viruses is rare. Influenza, being an RNA virus, is particularly challenging to study in ancient contexts because its genetic material degrades more quickly than DNA.

Nearly 5,000 years ago, a 20-year-old woman was buried in a tomb in Sweden - one of Europe's early farmers, dead in her prime. Now, researchers have discovered what killed her: *Yersinia pestis*, the bacterium that causes plague. The sample is one of the oldest ever found and belongs to a previously unknown branch of the *Y. pestis* evolutionary tree. This newly discovered strain may have contributed to the collapse of large Stone Age settlements across Europe, potentially triggering the world's first pandemic, according to the research team. However, other scientists argue that there isn't yet enough evidence to confirm this scenario.

Science News Archaeology 2018

Localisation and Preservation in Archaeological Remains

Preservation - Bacteria responsible for systemic infections, such as those that cause plague (*Yersinia pestis*) or tuberculosis (*Mycobacterium tuberculosis*), are often found in the bloodstream. This allows them to settle in hard tissues such as bones and teeth, which are better preserved over time.

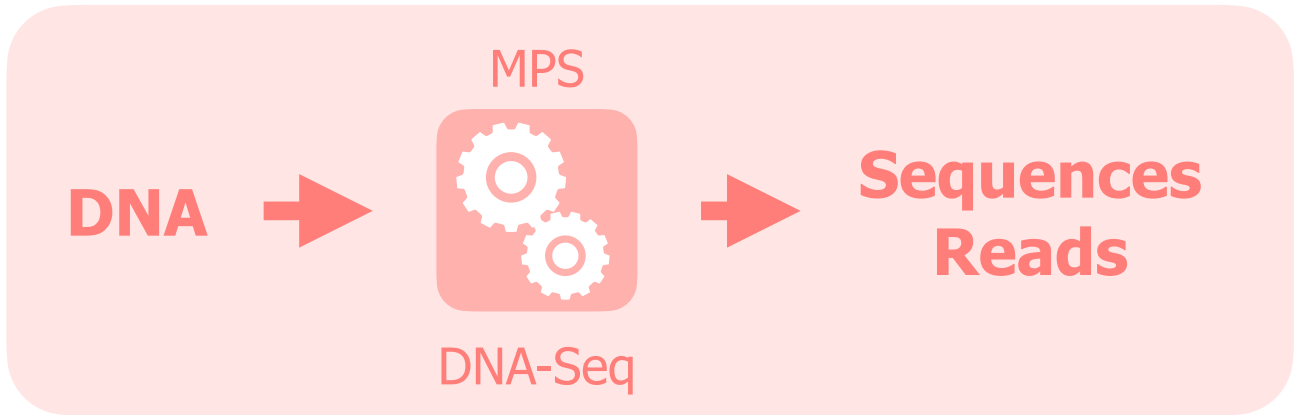
Soft tissue localisation - Many viruses cause localised infections, mainly in soft tissues, which decompose more rapidly after death. For example, influenza typically affects the respiratory tract, and hepatitis primarily affects the liver.

What is RNA?

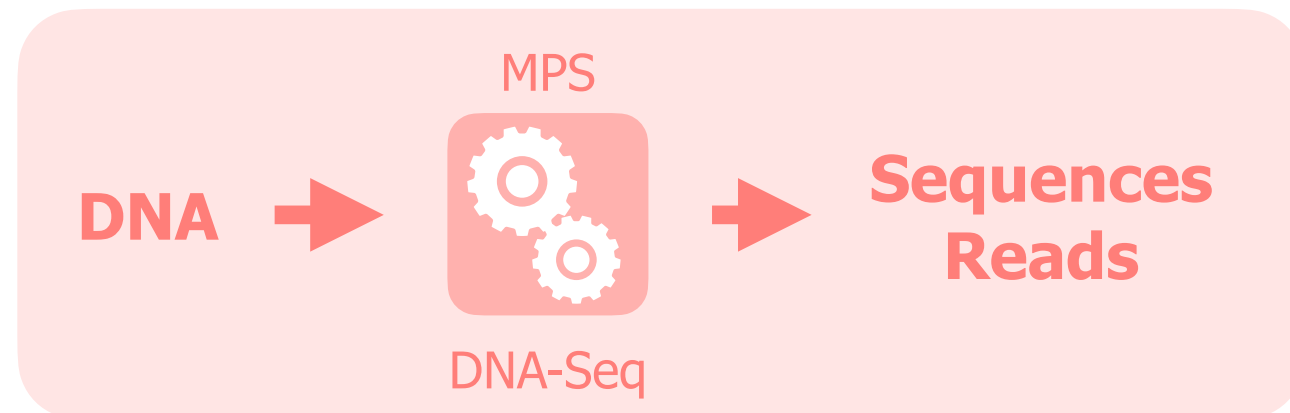
Why would I use RNA?

What is RNA-Seq?

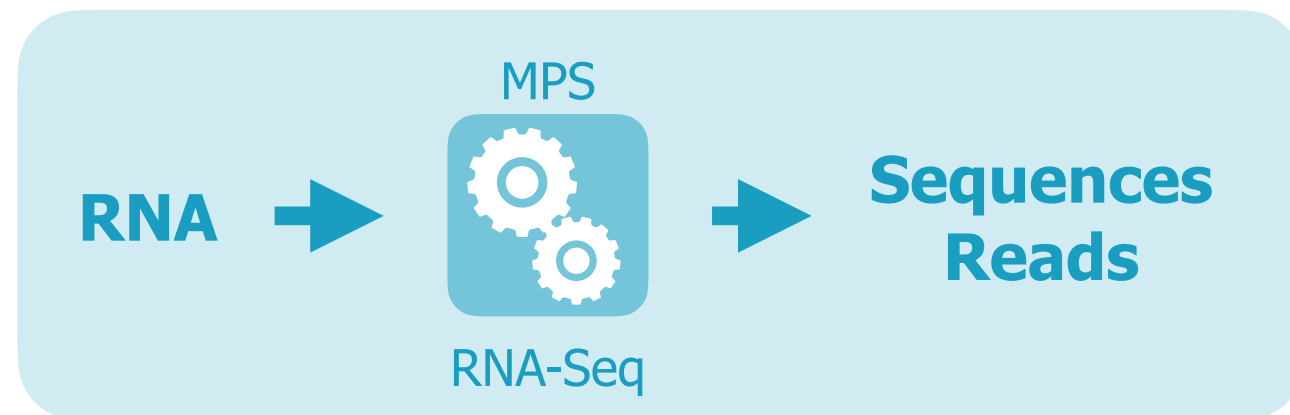
What can I do with RNA-Seq?



Genome Analysis

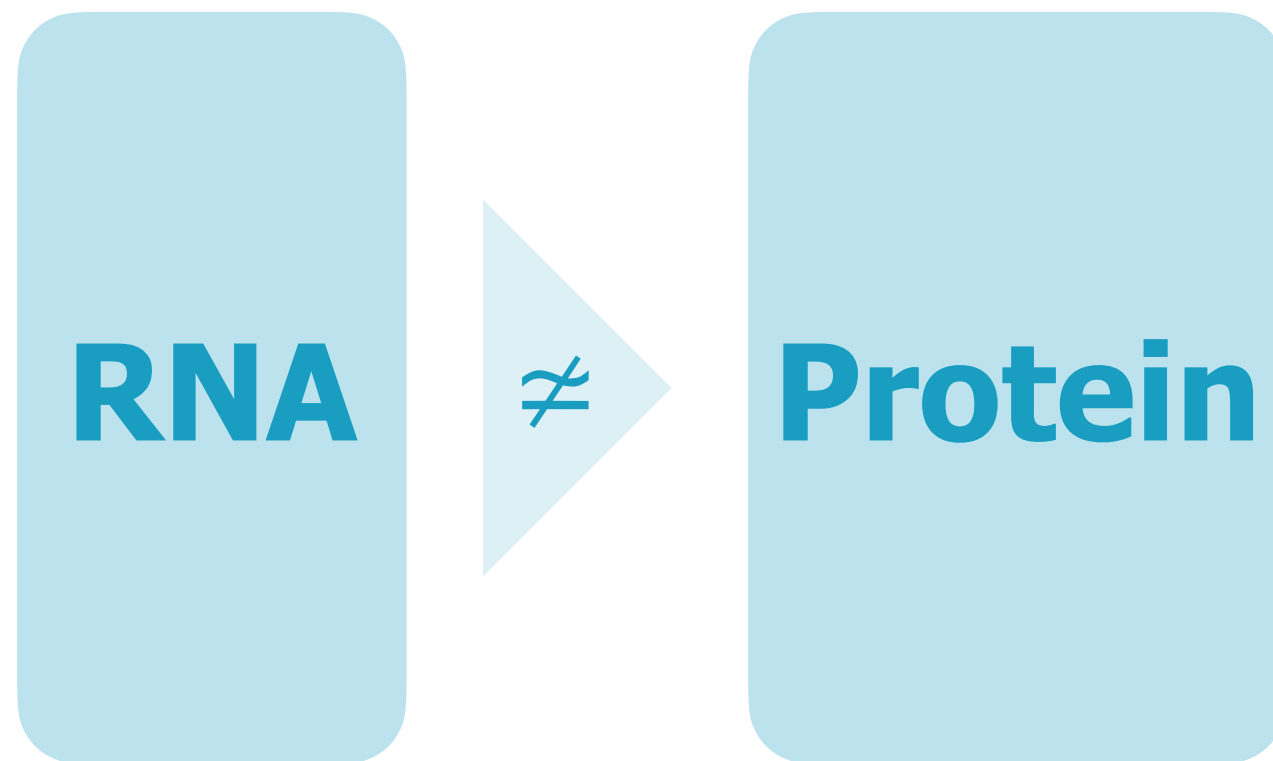


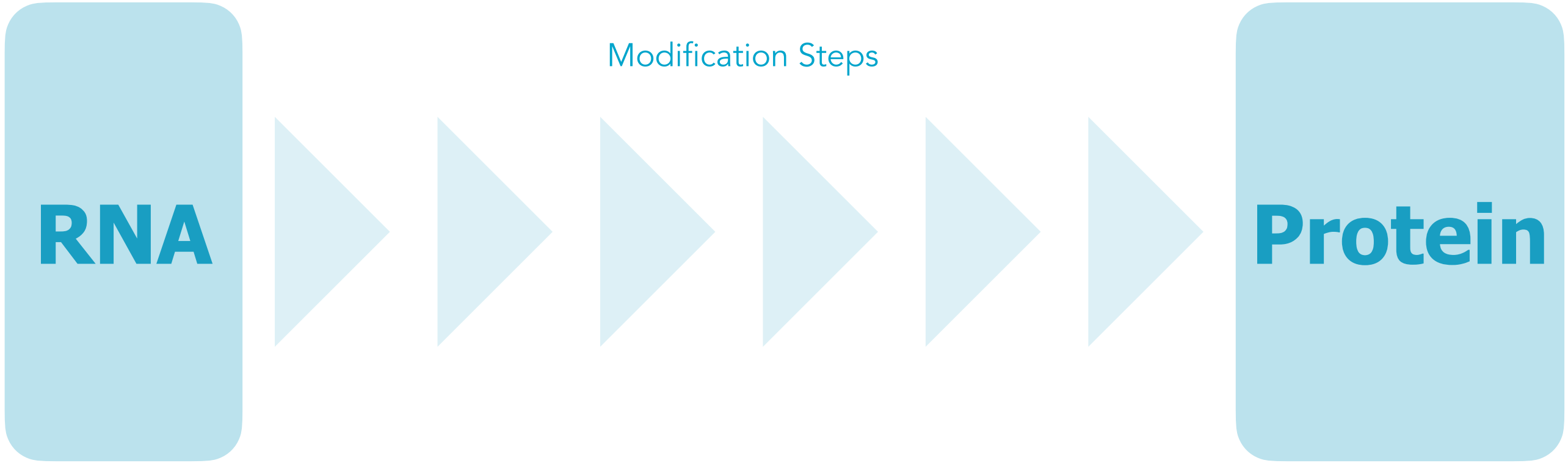
Genome Analysis

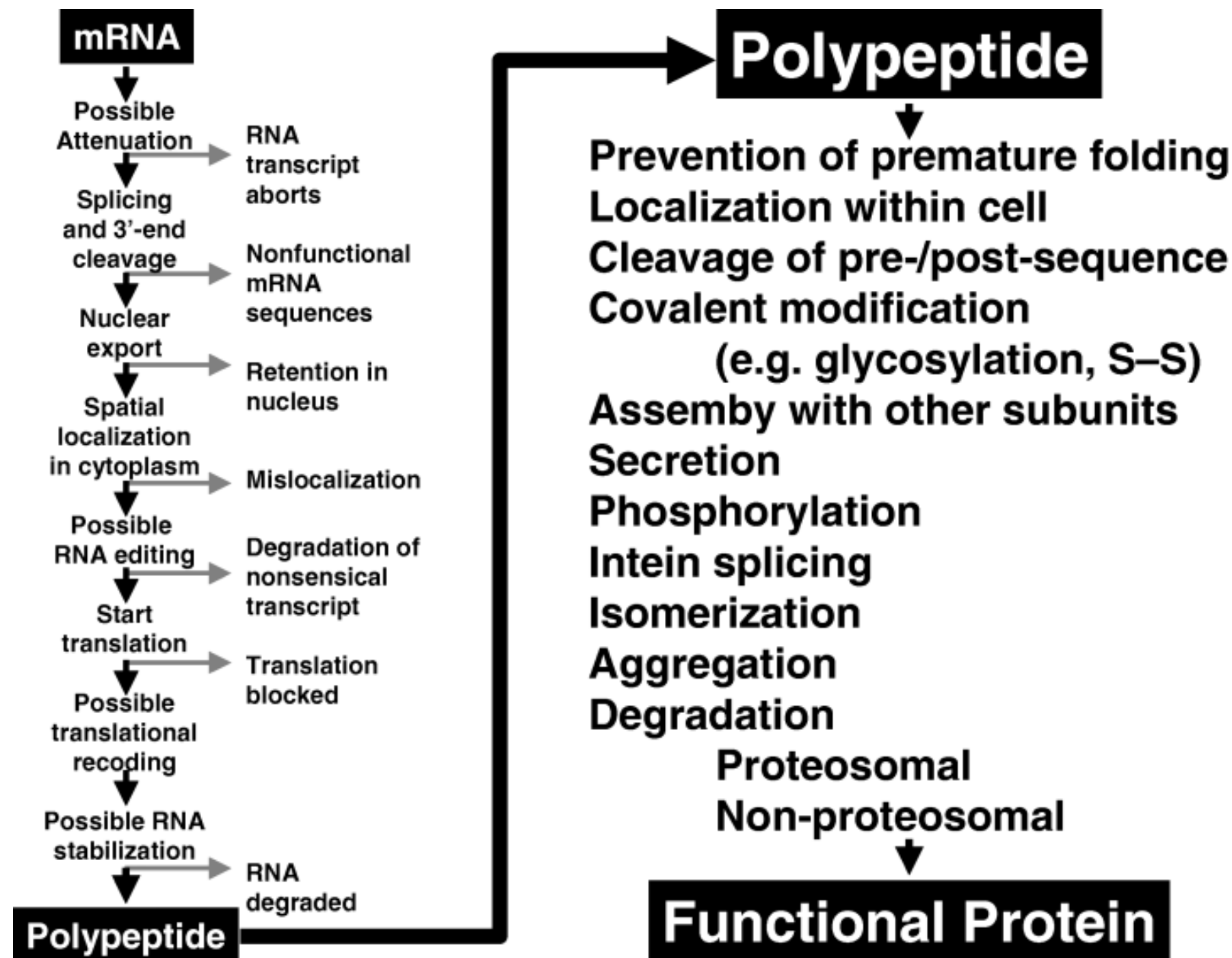


Transcriptome Analysis

MPS == Massive Parallel Sequencing







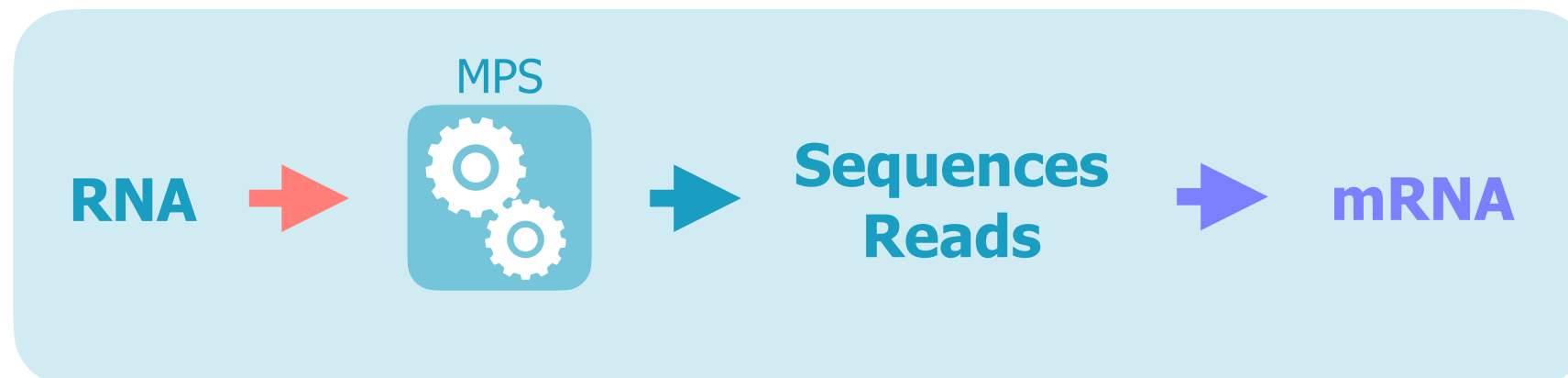
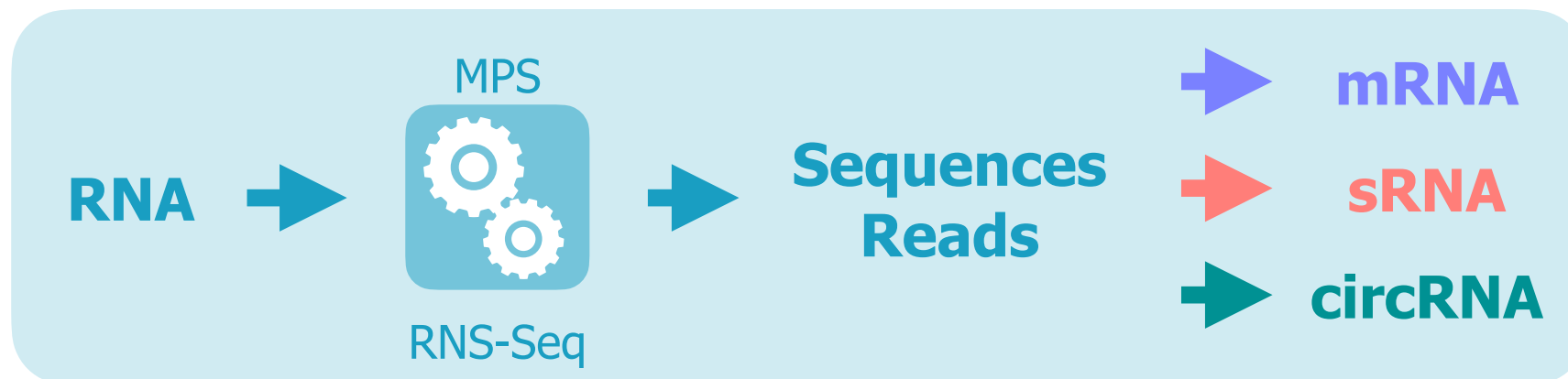
Feder & Walser (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology*.

total
RNA



mRNA
rRNA
tRNA

ncRNA, nmRNA, sRNA, smnRNA, tRNA, 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 – enriching RNA fragments within a specific size range
- Removal of non-target RNA (e.g., ribosomal RNA depletion)
- Targeted enrichment – capturing specific transcripts or regions of interest

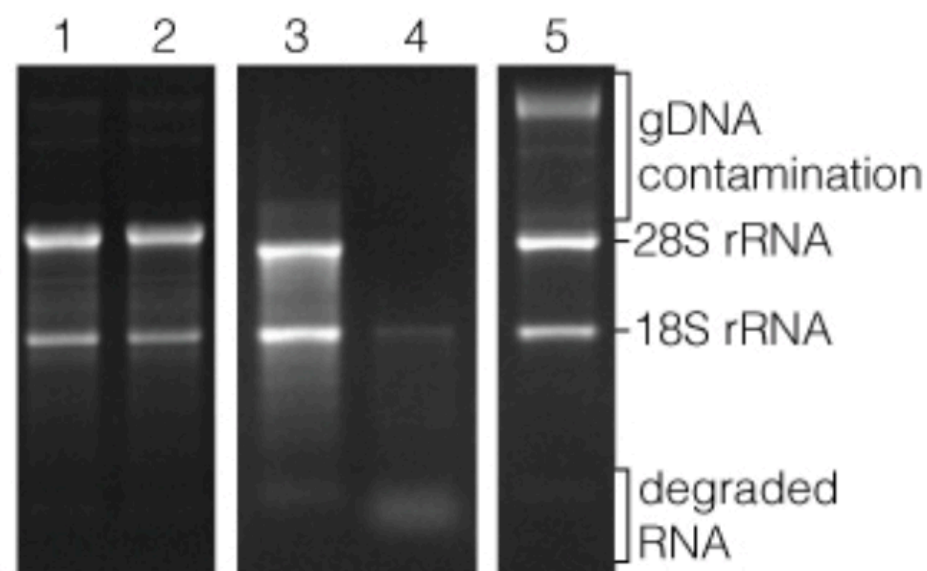
Every cleaning step improves RNA quality — but at a cost.



In most RNA-seq workflows, cleanup steps are essential to remove contaminants that could interfere with accurate signal detection. The aim is not to balance purity and quantity, but to achieve the level of purity required for reliable downstream analysis — even if this results in some loss of material. To compensate, it is important to begin with a sufficiently high amount of input RNA.

Quantity and Quality of RNA

In RNA-seq experiments, the quality of your data depends heavily on the quality of your starting material. If you start with poor-quality RNA—degraded samples, contamination, or insufficient input—no amount of fancy sequencing or analysis can fix those problems. This is the principle of “garbage in, garbage out”: **bad input leads to unreliable results**. To get meaningful data, you must ensure good RNA quality and proper sample preparation from the very beginning.



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 smearing 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 Schlumpberger¹

¹Qiagen, Hilden, Germany

ABSTRACT

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

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

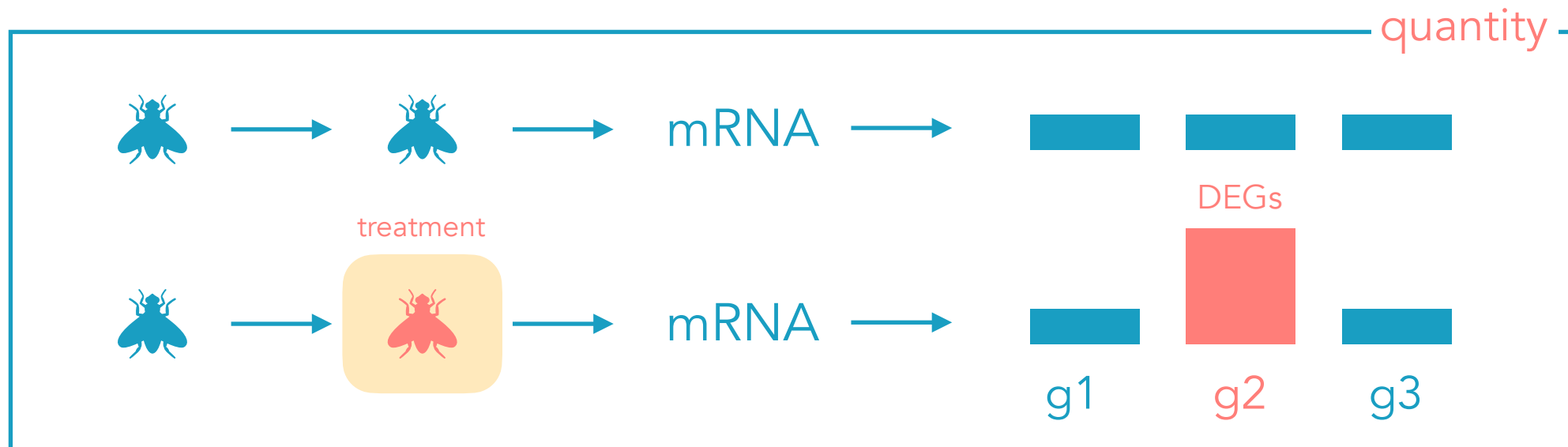


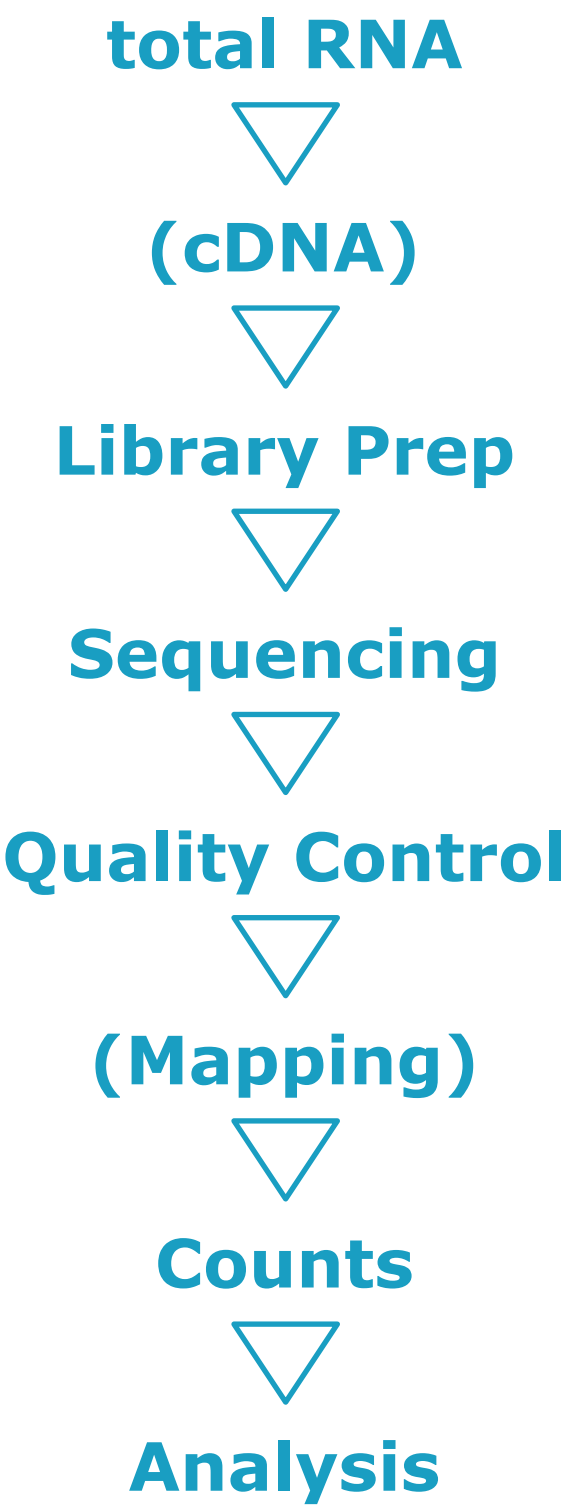
For transcriptome-based studies, RNA-seq libraries are generated by the synthesis of double stranded cDNA followed by the addition of sequencing adapters. This method however, does not retain any information about the DNA strand from which the RNA was transcribed. It is often desirable to create **libraries 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).

RNA-Seq ▷ Introduction

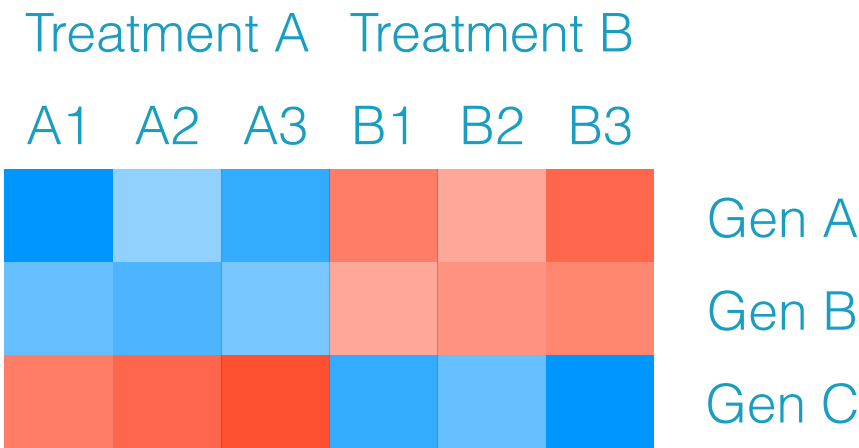
RNA-Seq is a comprehensive high-throughput sequencing approach for the **quantitative** and **qualitative** analysis of transcriptomes of model and **non-model organisms**.



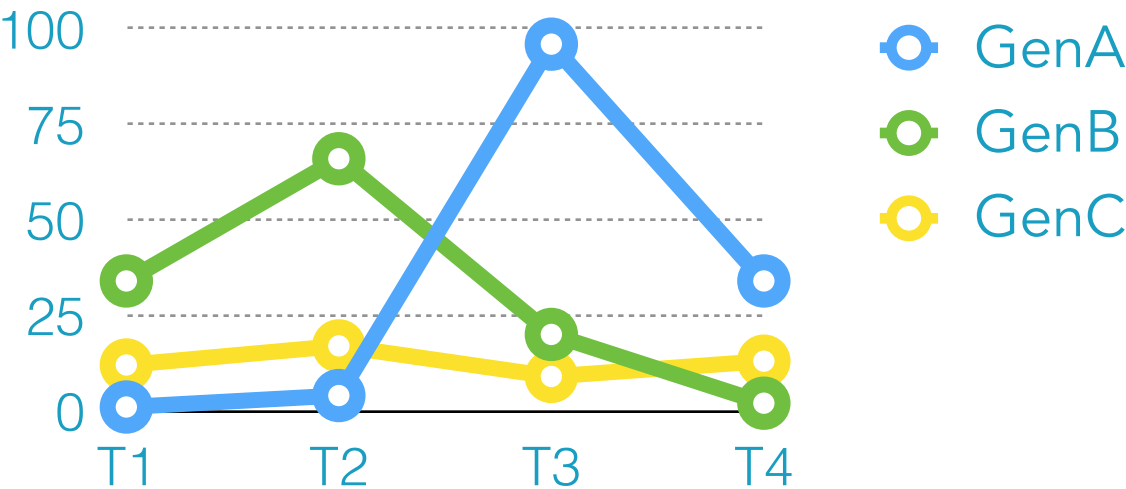


Quantitative RNA-Seq

DEG-Analysis



Time-Serie



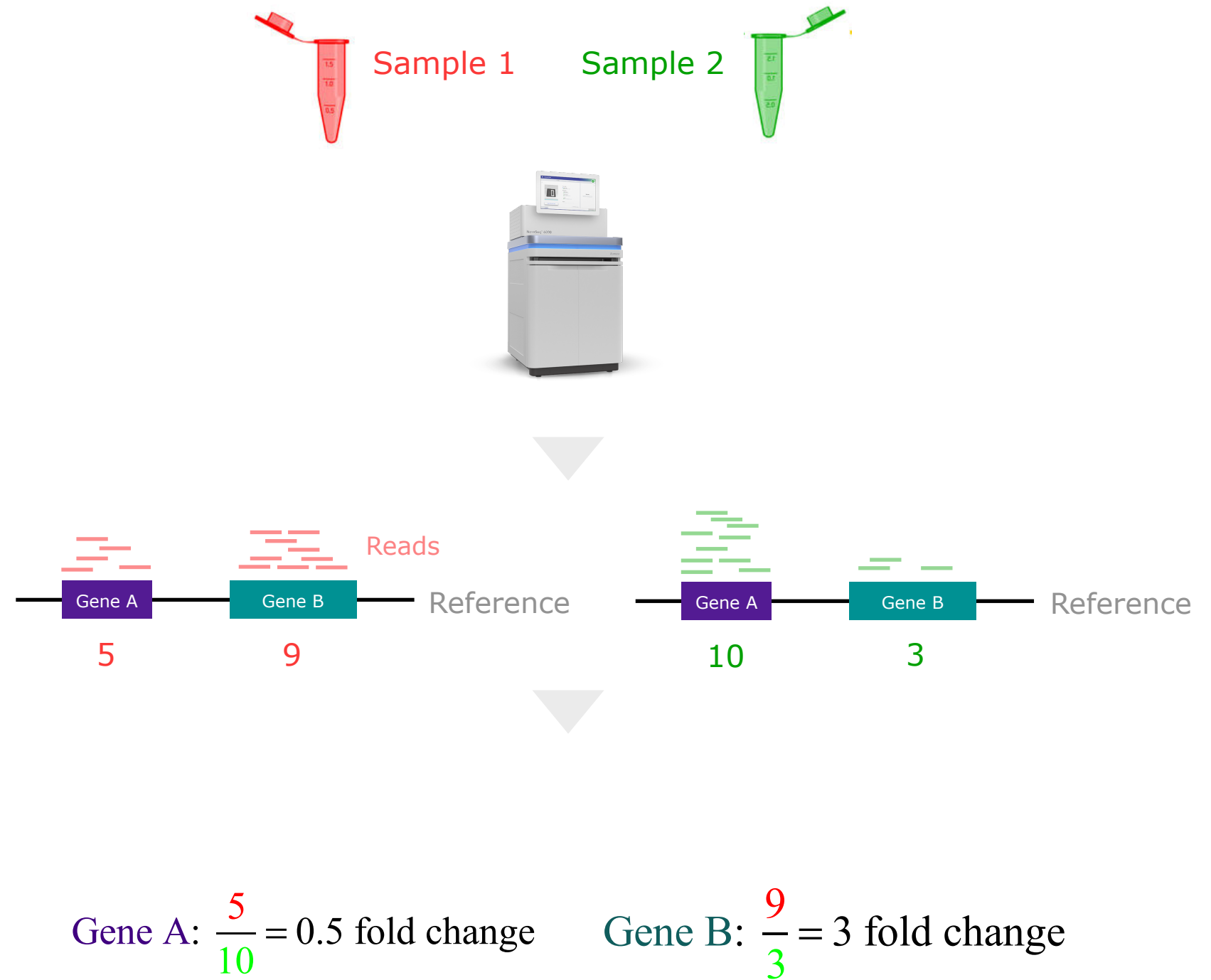
The Idea behind DEGs

- mRNA Isolation
- cDNA
- Library prep

- Sequencing

- Mapping

- Counts



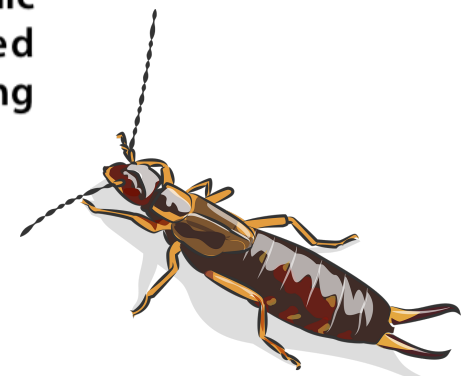
SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY BIOLOGY

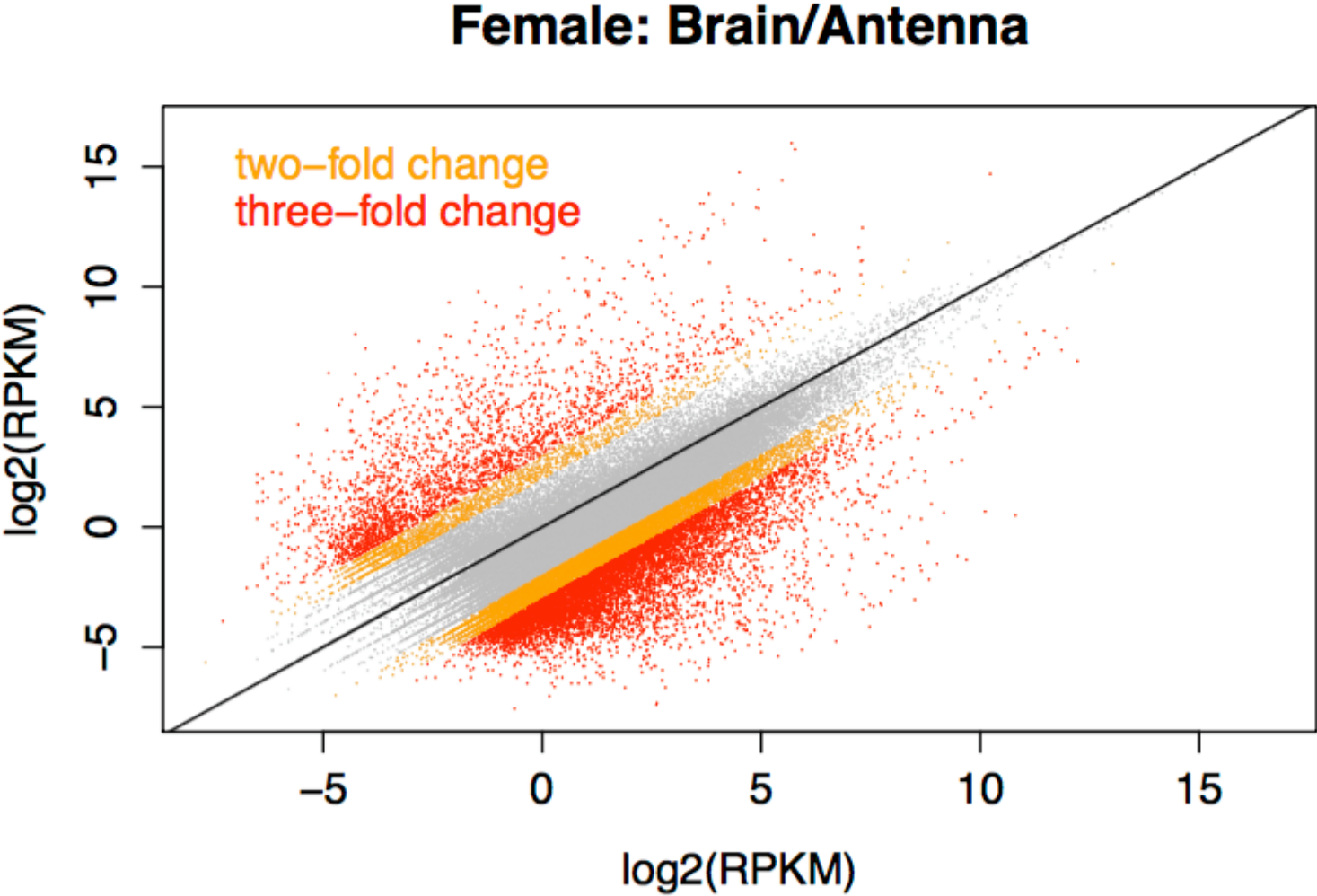
The genetic mechanism of selfishness and altruism in parent-offspring coadaptation

Min Wu^{1*}, Jean-Claude Walser², Lei Sun^{3†}, Mathias Kölliker^{1*‡}

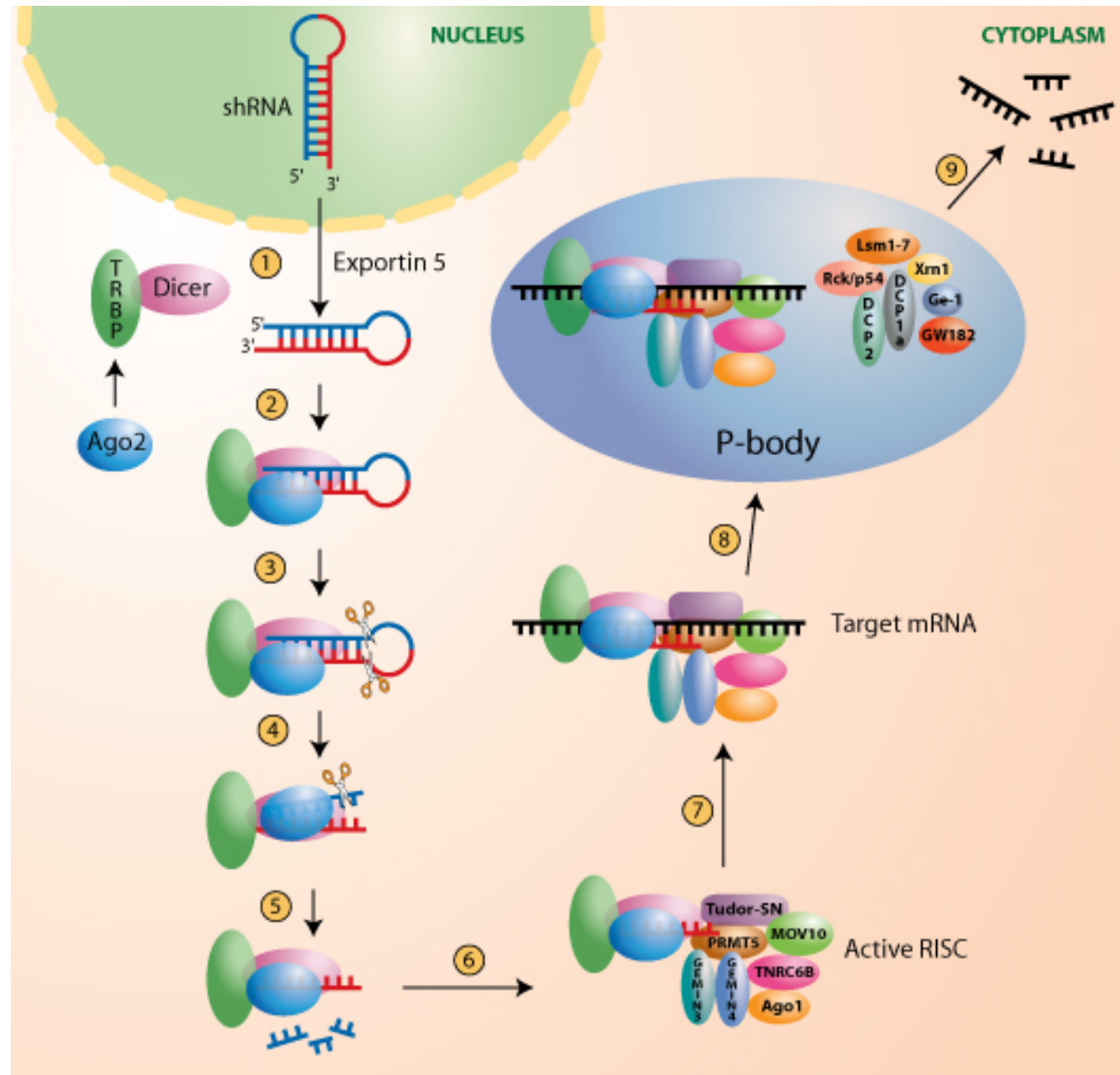
The social bond between parents and offspring is characterized by coadaptation and balance between altruistic and selfish tendencies. However, its underlying genetic mechanism remains poorly understood. Using transcriptomic screens in the subsocial European earwig, *Forficula auricularia*, we found the 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.







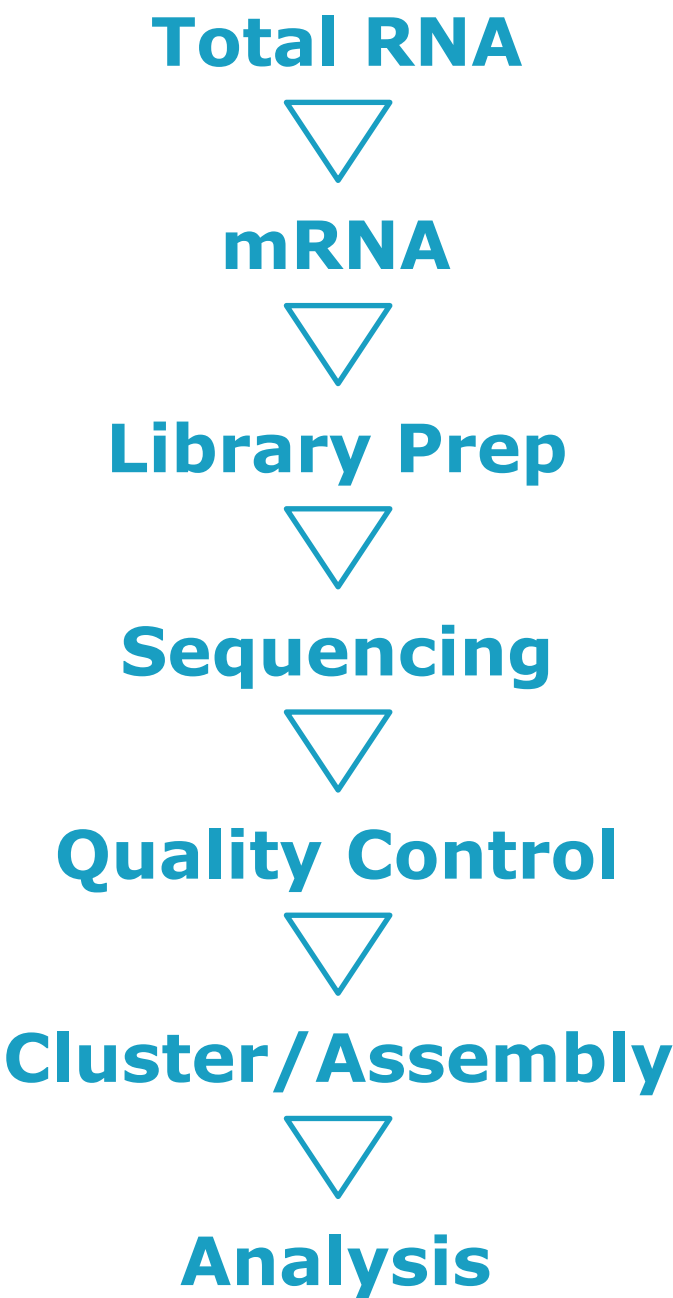
RNA interference (RNAi) is a post-transcriptional process triggered by the introduction of double-stranded RNA (dsRNA) which leads to gene silencing in a sequence-specific manner.



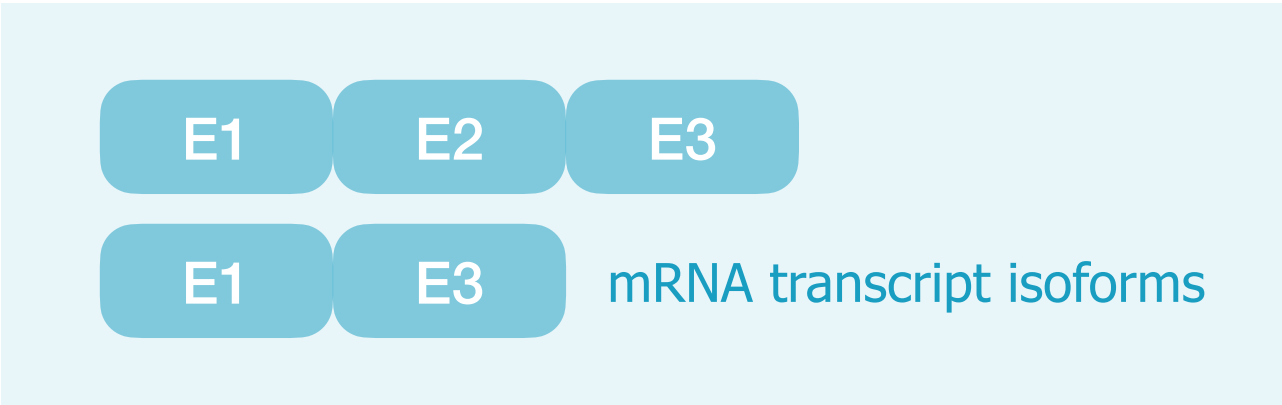
RNAi-mediated gene silencing in mammals using shRNAs

- ① Plasmid-expressed short hairpin RNA (shRNA) requires the activity of endogenous Exportin 5 for nuclear export.
- ② Ago2 (Argonaute 2) is recruited by TRBP, that forms a dimer with Dicer, and then receives the shRNA.
- ③ The shRNA is cleaved in one step by Dicer generating a 19-23 nt duplex siRNA with 2 nt 3' overhangs.
- ④ After identification of the "guide strand" in the siRNA duplex, the "passenger strand" is cleaved by Ago2.
- ⑤ The "passenger strand" is released.
- ⑥ The "guide strand" is integrated in the active RNA Interference Specificity Complex (RISC) that contains different argonautes and argonaute-associated proteins.
- ⑦ The siRNA guides RISC to the target mRNA.
- ⑧ RISC delivers the mRNA to cytoplasmic foci named processing bodies (P-bodies or GW-bodies) wherein mRNA decay factors are concentrated.
- ⑨ The target mRNA is cleaved by Ago2 and degraded.

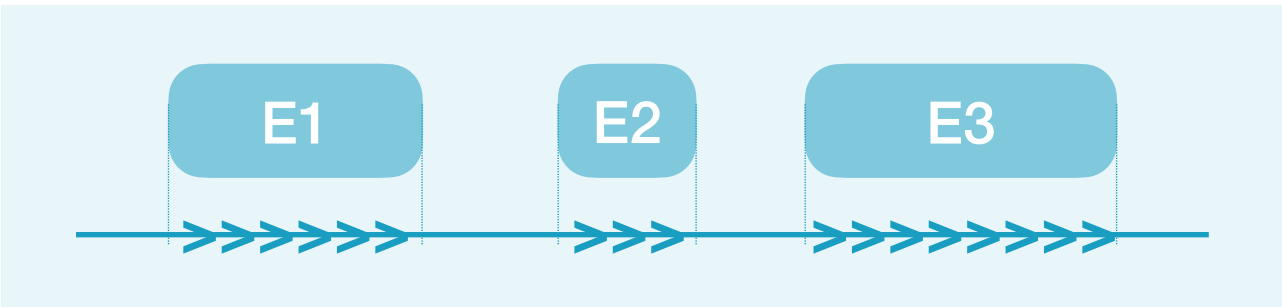
Qualitative RNA-Seq



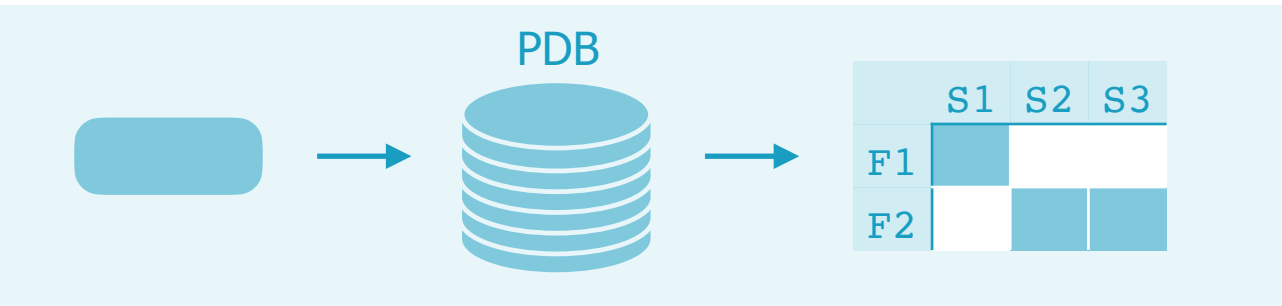
Alternative Splicing



Exon Mapping



Gene Annotation



DISCOVER FULL-LENGTH TRANSCRIPTS

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

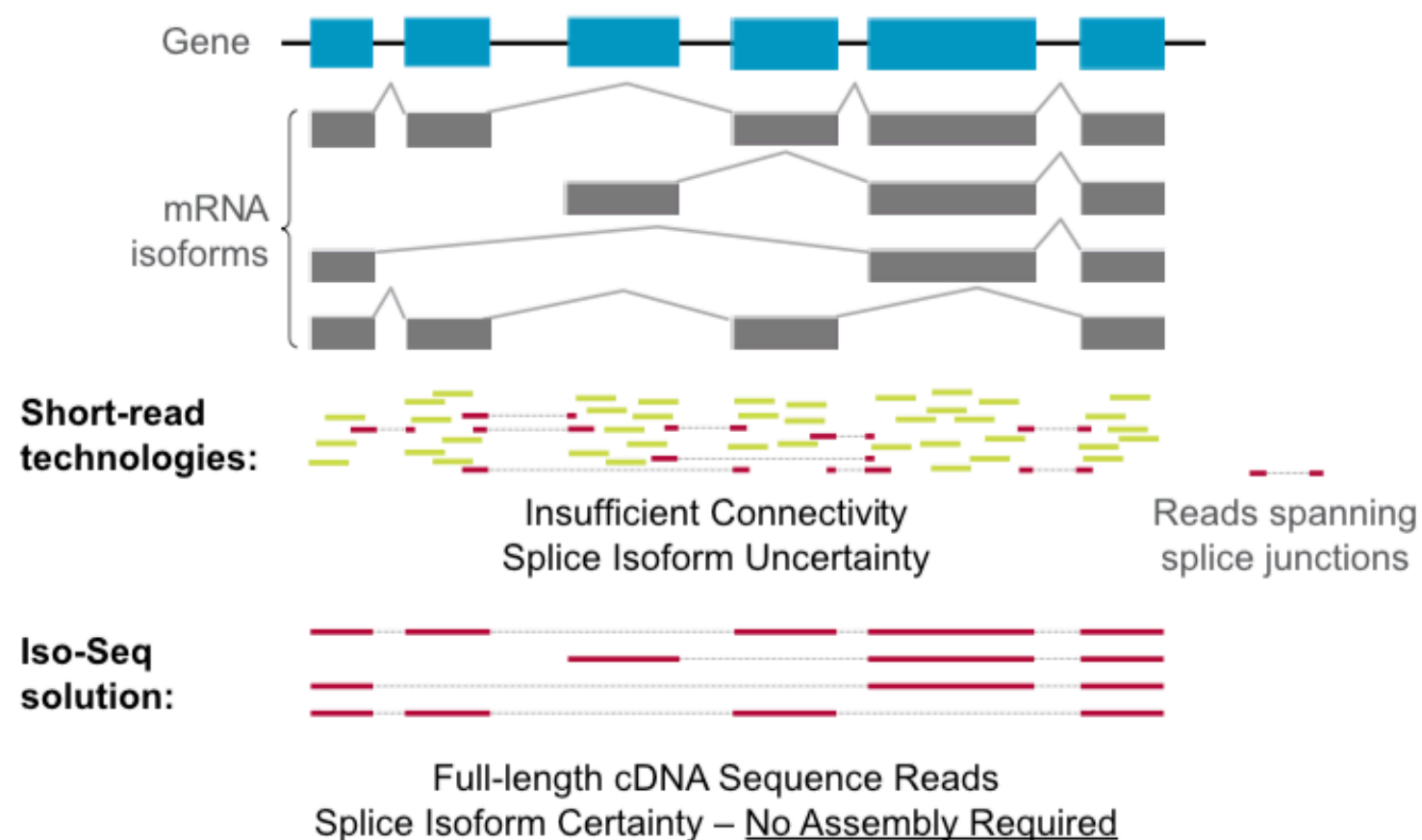
RNA Sequencing



[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.

Is RNA-seq still sexy?

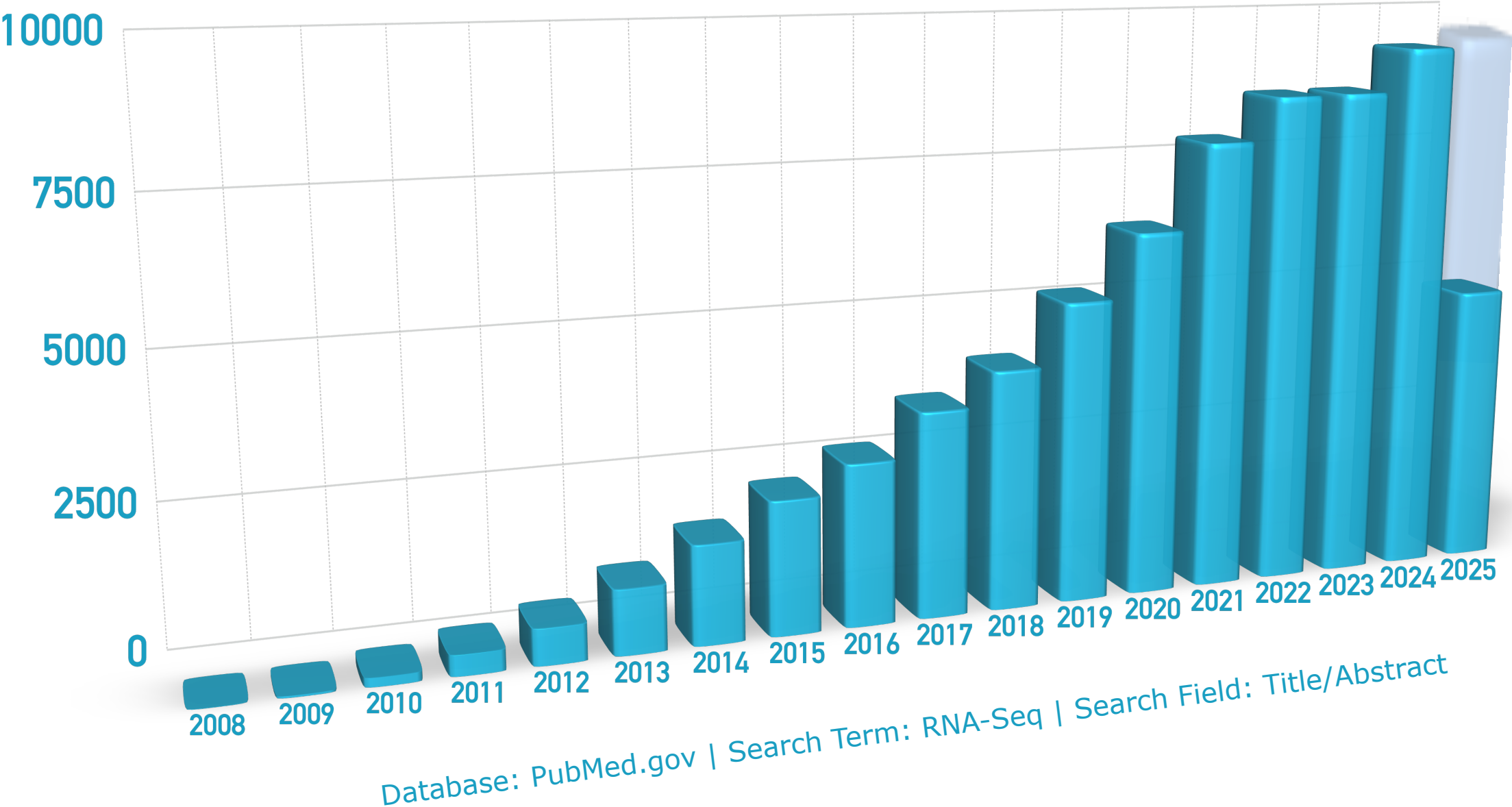
INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein and Michael Snyder

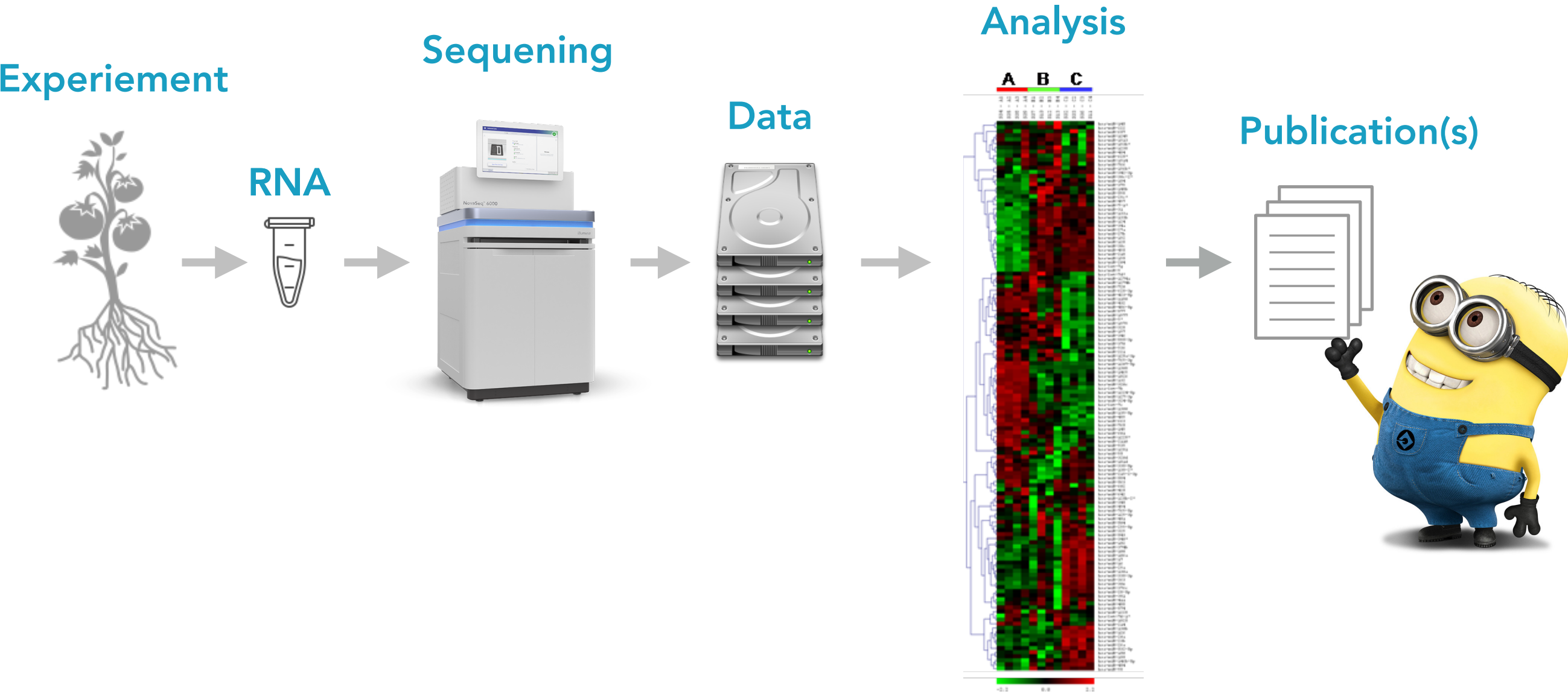
Abstract | RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts 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.



Experimental Design

RNA-Seq Hype
Massive Parallel Sequencing Hype



RNA-Seq Reality

Massive Parallel Sequencing Reality

Setup



Quality
Amount



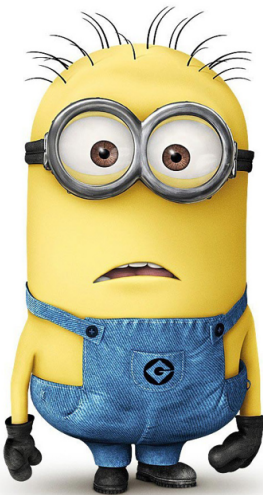
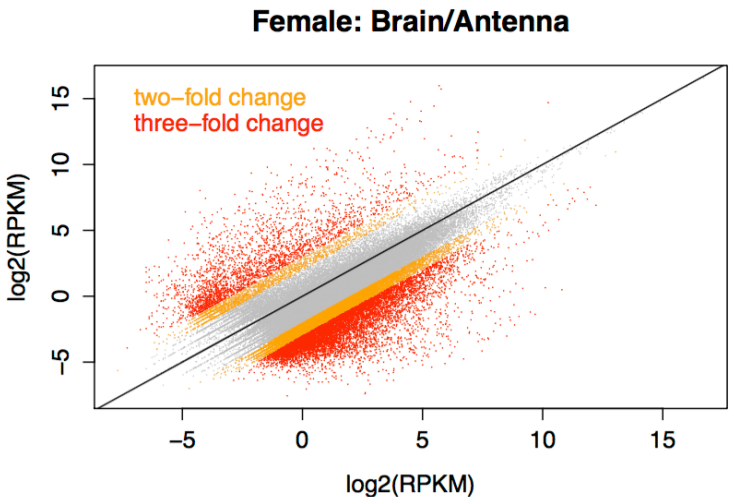
Method



Storage

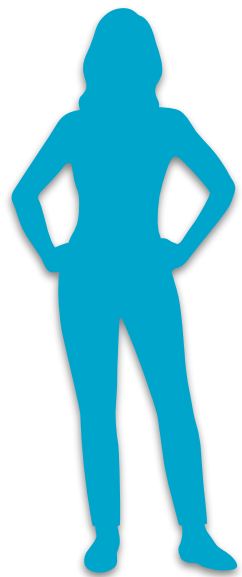


Interpretation





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

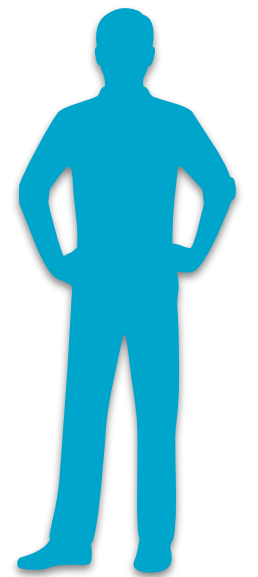


Number of well-validated genes: ?

Percentage of protein-coding genes: ?

Prevalence of alternative splicing: ?

Average number of transcript isoforms per gene: ?



**Molecular
BioSystems**



PAPER

[View Article Online](#)

[View Journal](#) | [View Issue](#)



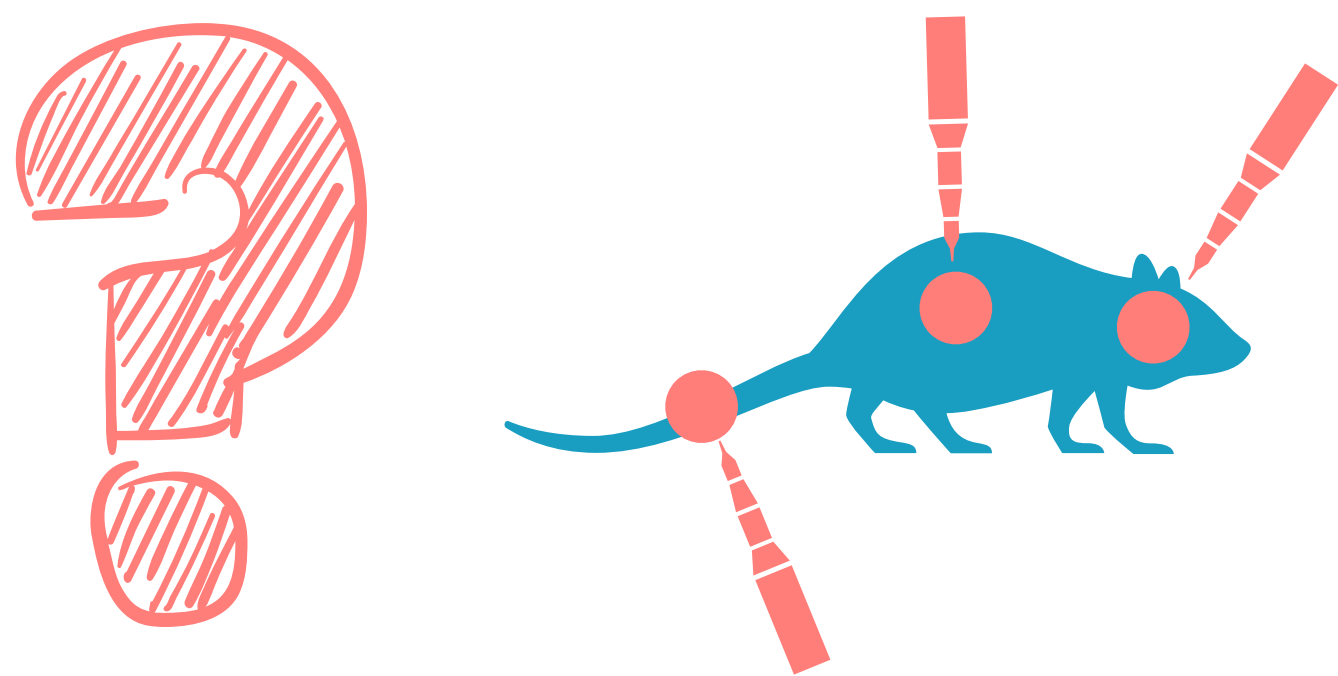
Cite this: *Mol. BioSyst.*, 2016,
12, 508

Strand-specific RNA-seq analysis of the *Lactobacillus delbrueckii* subsp. *bulgaricus* transcriptome†

Huajun Zheng,^{‡a} Enuo Liu,^{‡a} Tao Shi,^a Luyi Ye,^a Tomonobu Konno,^b
Munehiro Oda^c and Zai-Si Ji^{*ab}

Lactobacillus delbrueckii subsp. *bulgaricus* 2038 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





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

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)

Design

Preparation

Methode

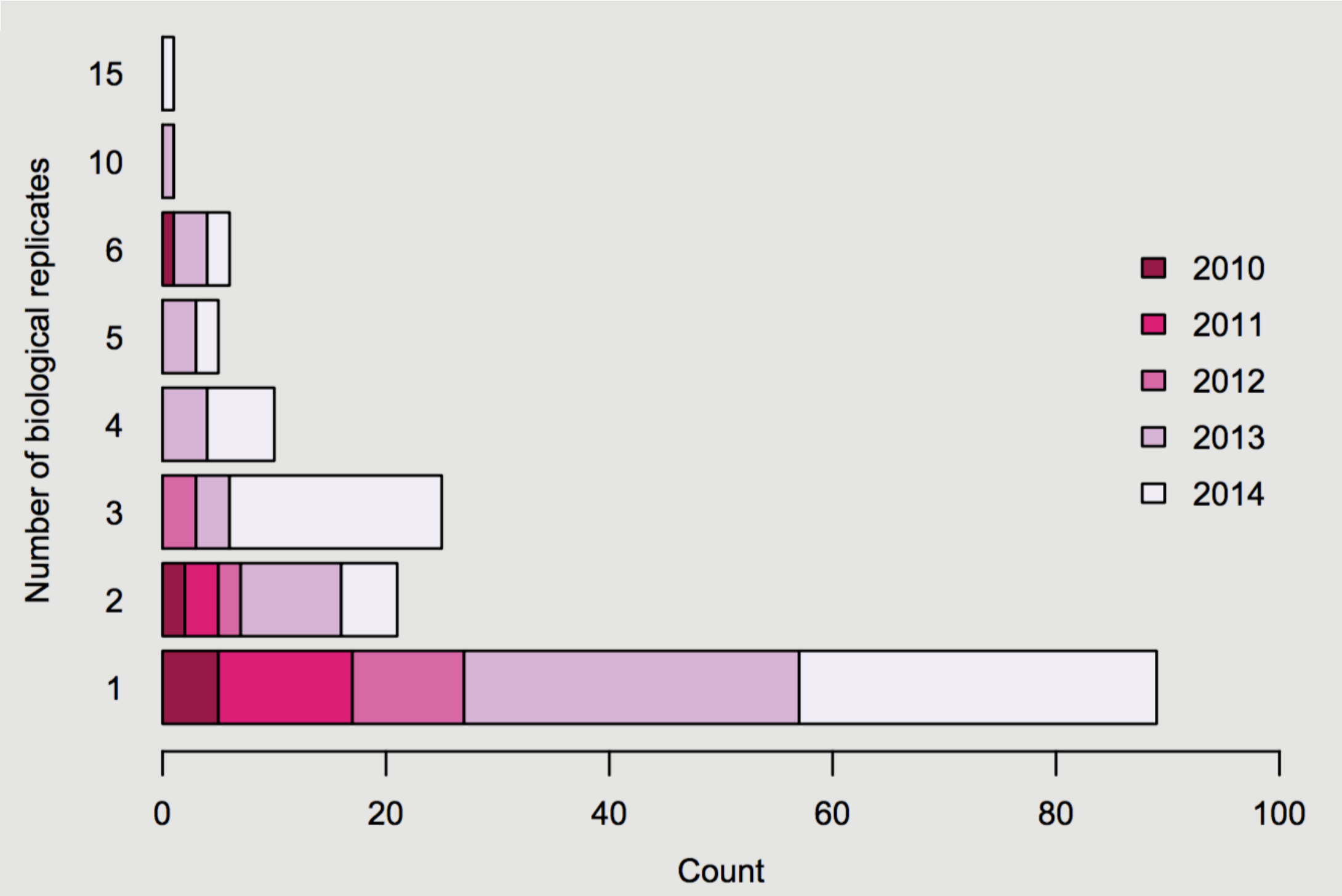
Analysis

Extras

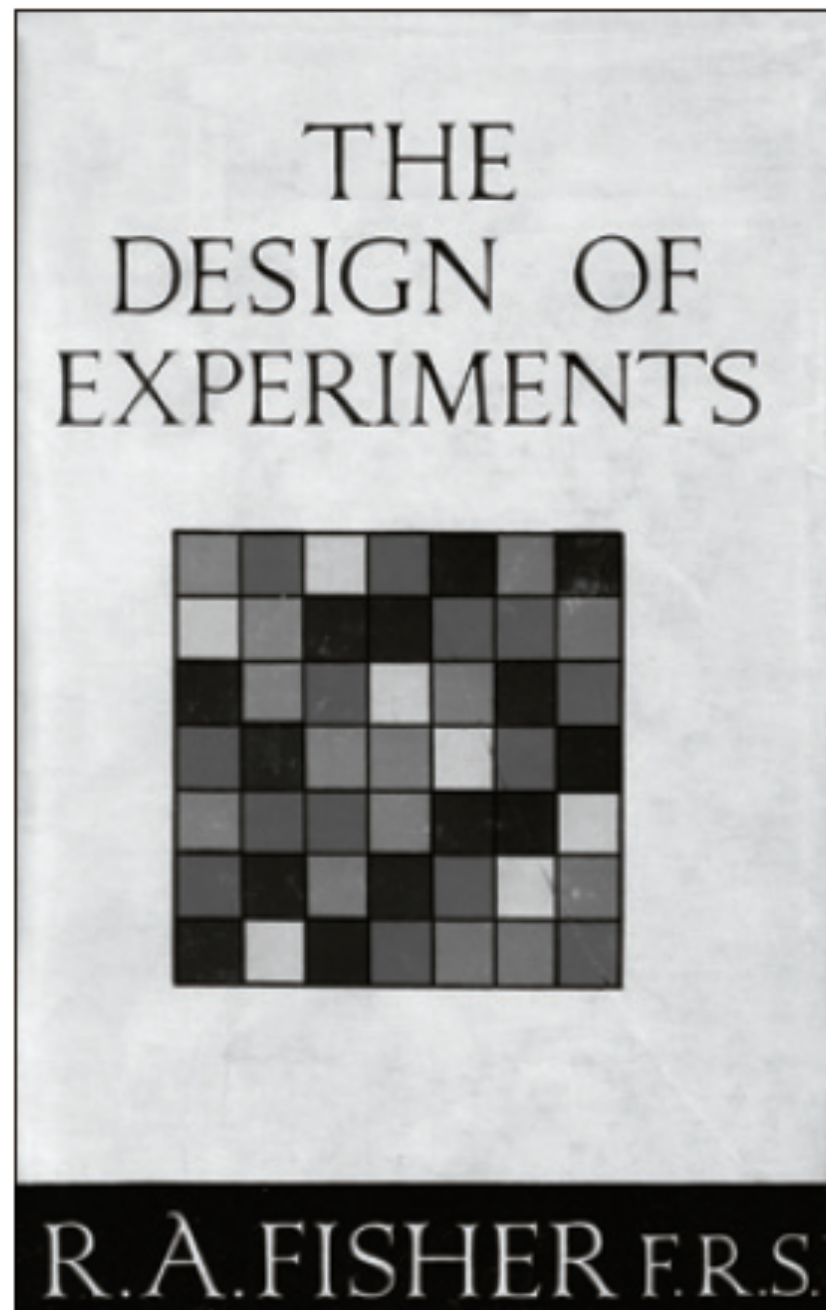


How much sequencing is needed?

- How many **biological replicates** are needed?
- What is the **minimum** required **sequencing depth**?
- What is the **trade-off** between sequencing depth and number of biological replicates?
- What is your budget?



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



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

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

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge¹

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

Manuscript received January 31, 2010
Accepted for publication March 15, 2010

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

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

Differential expression in RNA-seq: a matter of depth

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

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

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

** Corresponding author. Email: aconesa@cipf.es*

August 29, 2011

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

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

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

Yuwen Liu^{1,2}, Jie Zhou^{1,3} and Kevin P. White^{1,2,3,*}

¹Institute of Genomics and Systems Biology, ²Committee on Development, Regeneration, and Stem Cell Biology and

³Department 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 > 2x**15**M (30M) PE-reads => 6% increase

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

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

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

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

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

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

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

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

“With **three biological replicates**, nine of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates. This rises to >85% for the subset of SDE genes changing in expression by more than fourfold. To achieve >85% for all SDE genes regardless of fold change requires **more than 20 biological replicates.**”

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

A large benchmarking experiment with 48 replicates per condition (Schurch et al., 2016) found:

- **3 replicates** capture only ~20–40 % of differentially expressed genes.
- **6 replicates** recover >85 % of strong signals.
- **≥12 replicates** are needed to reliably detect all differential genes regardless of effect size

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 a critical step in **experimental design**. It helps determine the number of samples needed to detect an effect of a given size with a specified level of statistical confidence. Alternatively, when sample size is constrained, it allows us to **estimate the probability (statistical power) of detecting such effects**. If this probability is too low, it may be advisable to adjust the design, increase replication, or reconsider the feasibility of 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 = $P(\text{Type I error})$ = probability of finding an effect that is not there**
- (4) **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>

In **RNA-seq experiments**, gene expression is measured as **count data** — the number of reads mapping to each gene. These counts are not only affected by **biological differences**, but also by **technical variation**, resulting in **overdispersion** (variance > mean).

The **negative binomial (NB) distribution** is used to model this overdispersed count data. It introduces a **dispersion parameter** that accounts for variability beyond what the Poisson distribution can handle.

This makes the NB model ideal for **power analysis** in RNA-seq, as it allows researchers to:

- Estimate the probability of detecting a given fold change in expression,
- At a specified false discovery rate (FDR),
- Given a known or estimated dispersion and sequencing depth.

Using the NB model, tools like **RNASeqPower** help determine **how many biological replicates** are needed to reliably detect differential expression — balancing cost, statistical power, and biological variability.

Tools or approaches that assume a negative binomial distribution:

- **DESeq2** (`estimateSizeFactors` and `estimateDispersions`)
- **edgeR** (`RNAseqPower` and `glmLRT`)
- **RNASeqPower** R package



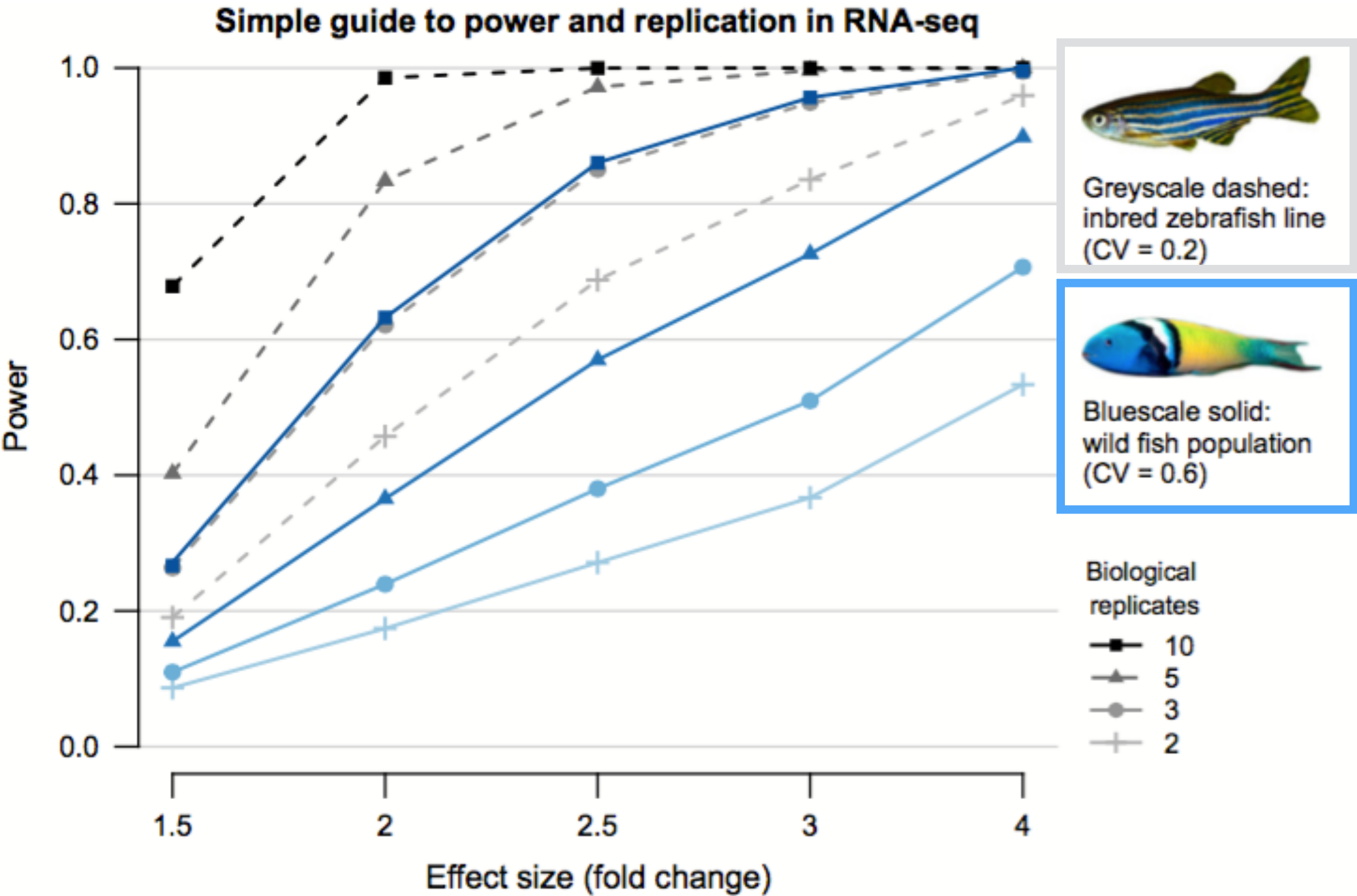
Greyscale dashed:
inbred zebrafish line
(CV = 0.2)



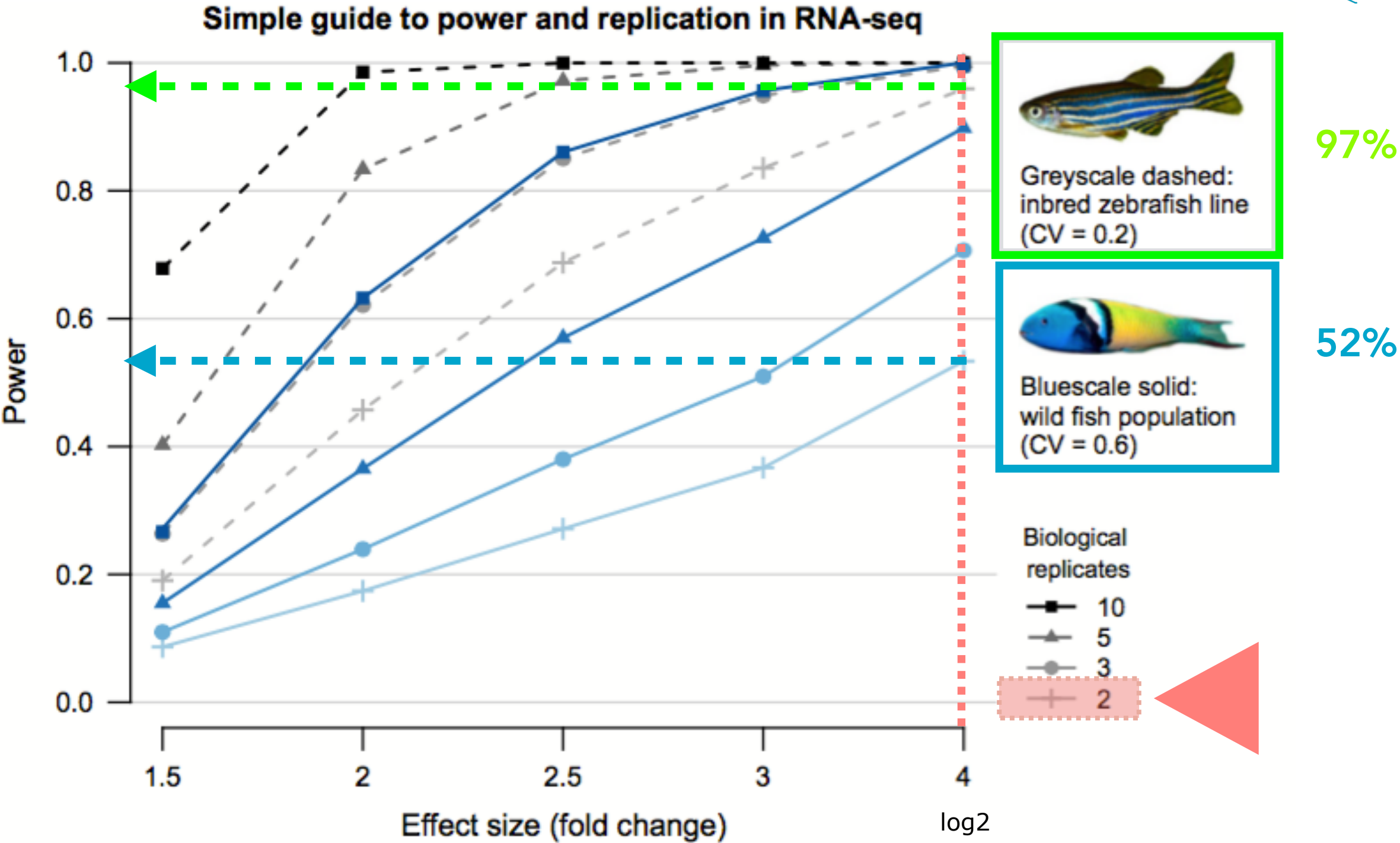
Bluescale solid:
wild fish population
(CV = 0.6)

Coefficient of Variation (CV)

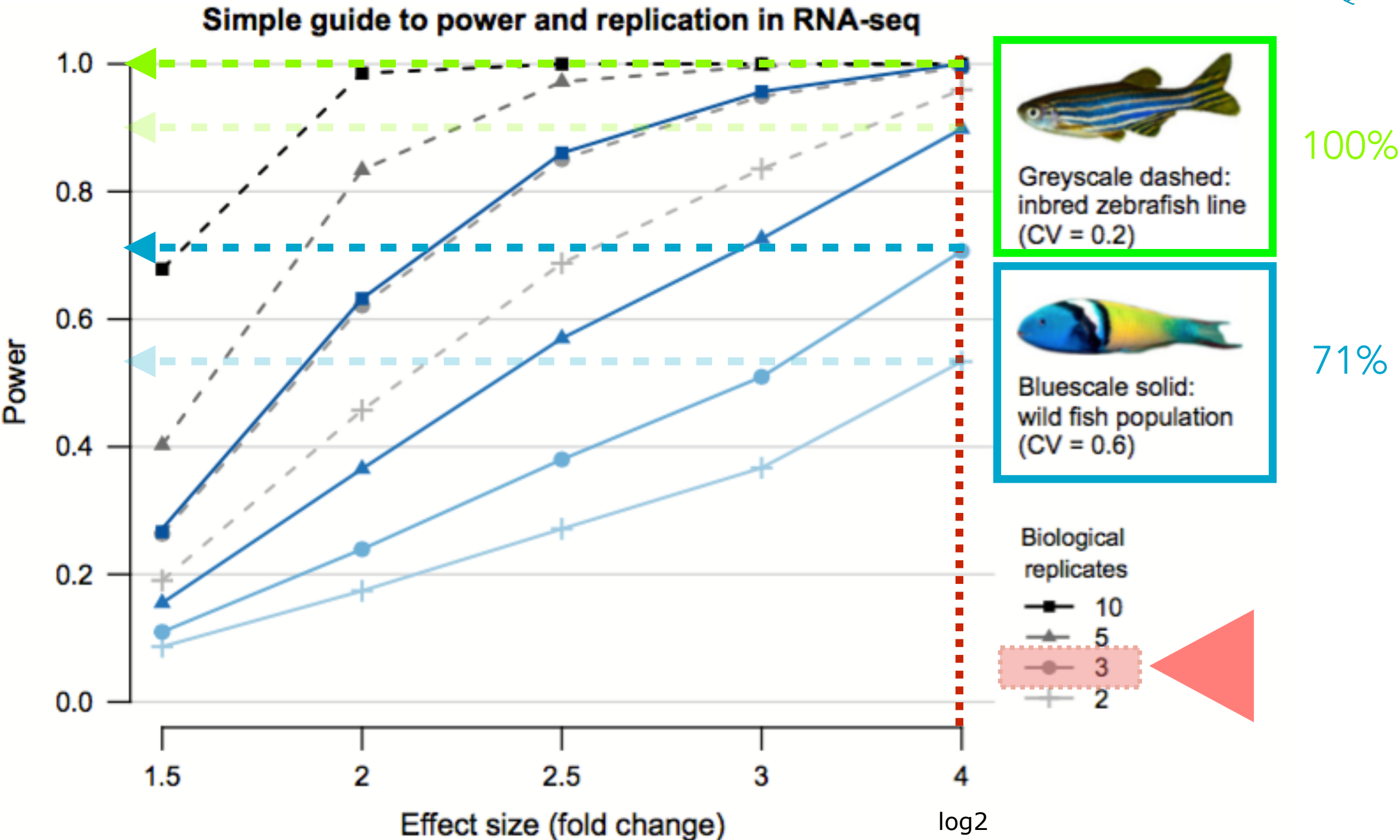
CV measures how much a gene's expression varies compared to its average level. It's calculated by dividing the standard deviation by the mean expression across samples. **A low CV means the gene's expression is consistent, while a high CV means it varies a lot.** Understanding CV helps us identify reliable genes and estimate how many samples are needed to detect real differences.



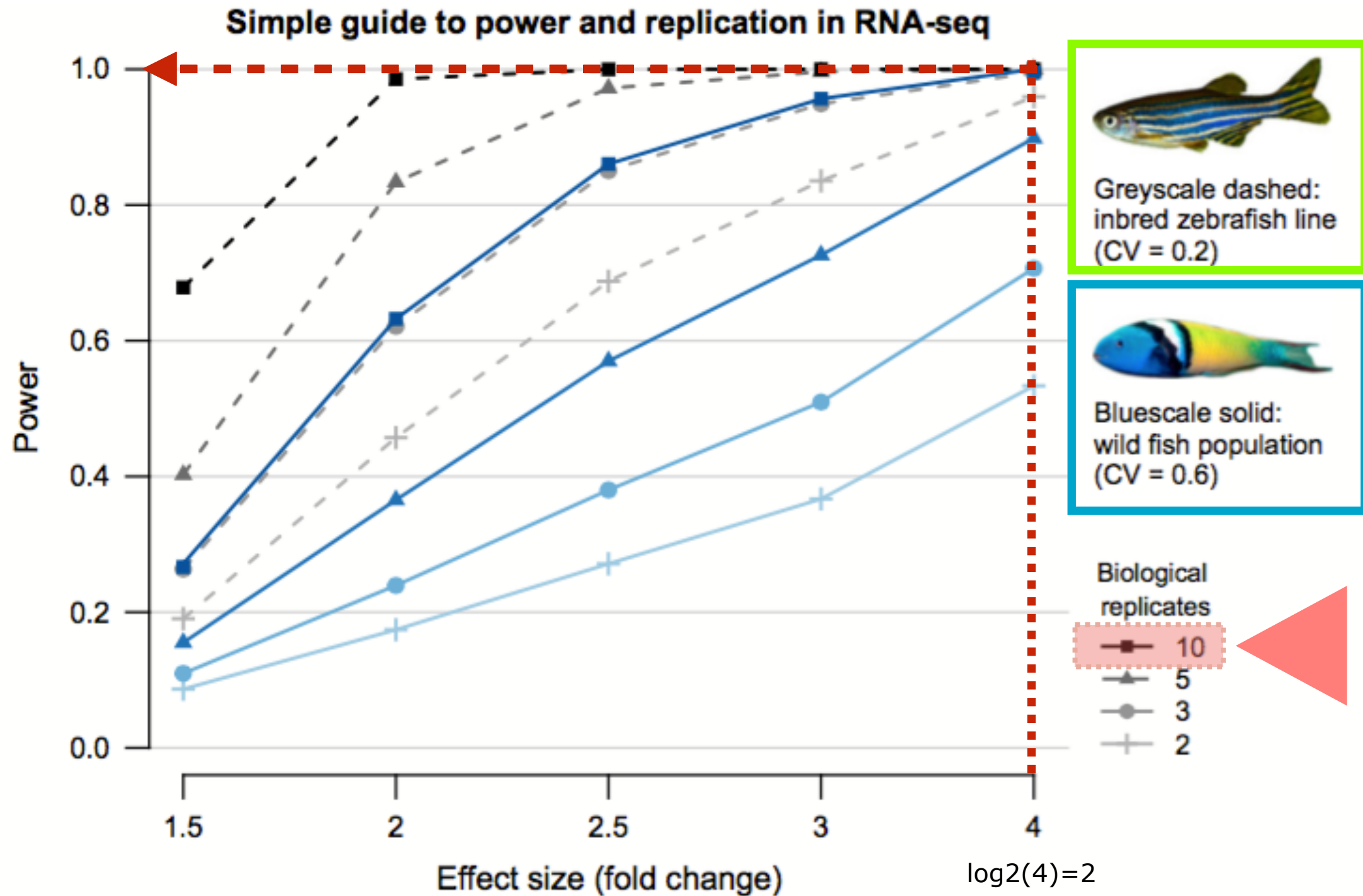
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, 25, 1224–1241.



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.



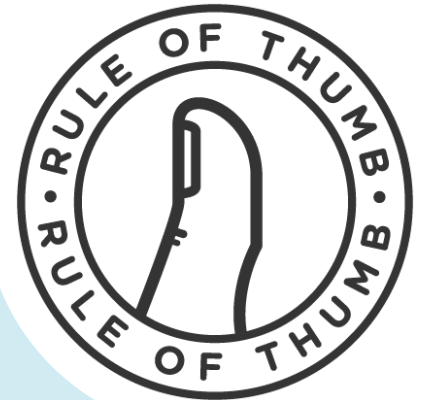
Todd et al. (2015) The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, 25, 1224–1241.



When planning an RNA-seq experiment, there are many uncertainties to consider:

- What does the expression landscape look like?
- How complex is the library?
- How will the reads be distributed across genes?

These factors can greatly affect the success of the study. Running a pilot experiment can provide valuable insights into these questions, helping to optimise the design and improve the chances of meaningful results.



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

Tomato - Flavour - Experiment



Flavour is a balance of acidity and sugar, along with the influence of elusive volatile compounds that contribute to aroma and taste. Regardless of variety, growing conditions - such as temperature - can also influence flavour.



Treatment #1 $t_1=27^{\circ}\text{C}$ / $t_2=15^{\circ}\text{C}$

Treatment #2 $t_1=29^{\circ}\text{C}$ / $t_2=18^{\circ}\text{C}$

ADH1A Gene - Experiment

Alcohol Dehydrogenase 1A (Class I), Alpha Polypeptide



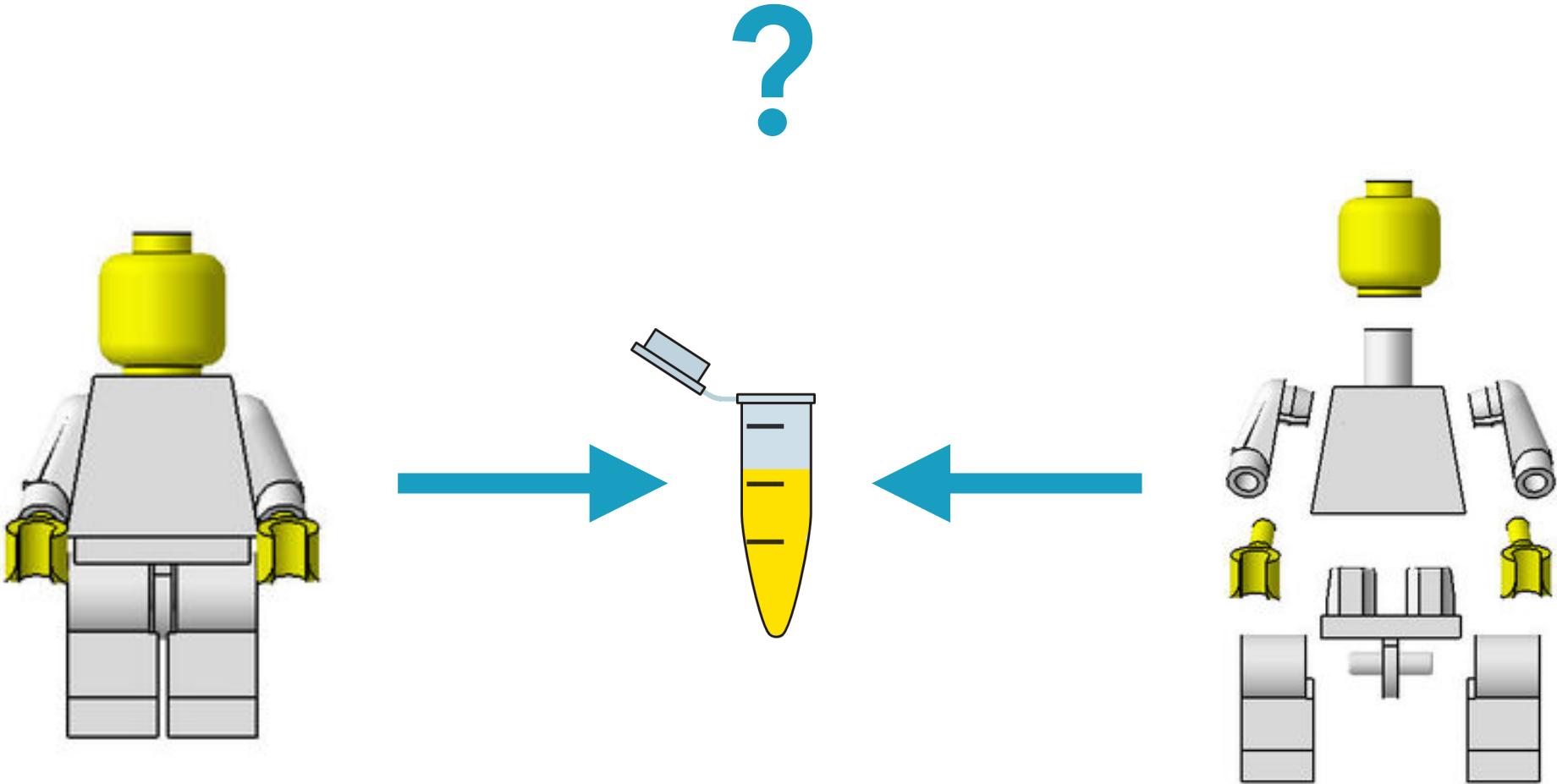
This gene encodes class I alcohol dehydrogenase, alpha subunit, which is a member of the alcohol dehydrogenase family. Members of this enzyme family metabolize a wide variety of substrates, including ethanol.



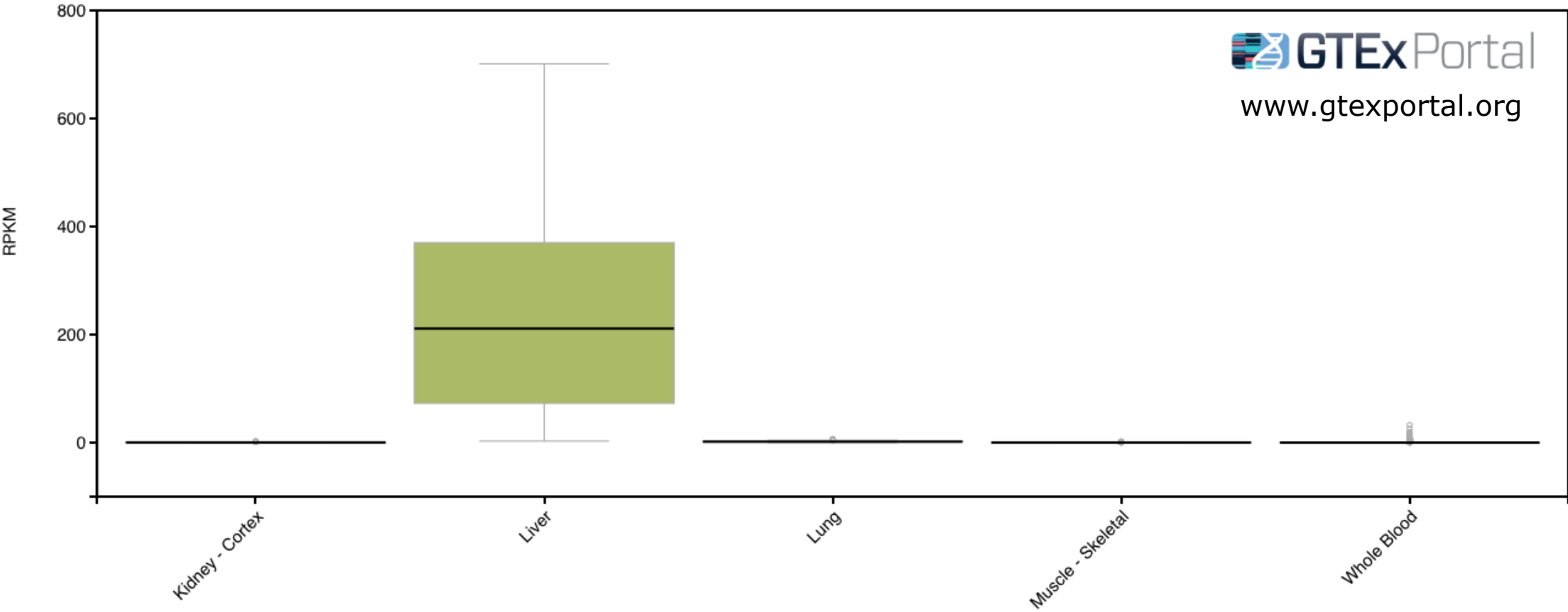
Design a study to investigate population-level variation in *ADH1A* gene expression across metapopulations of marmosets.

Sample considerations

When designing an RNA-seq experiment, it's important to consider which cells or tissues you are analyzing. Not all cell types respond the same way to a treatment or condition. Some cells may show strong changes in gene expression, while others remain largely unaffected. Choosing the right cell population is crucial to detecting meaningful biological signals and avoiding diluted or misleading results.



ADH1A Gene Expression

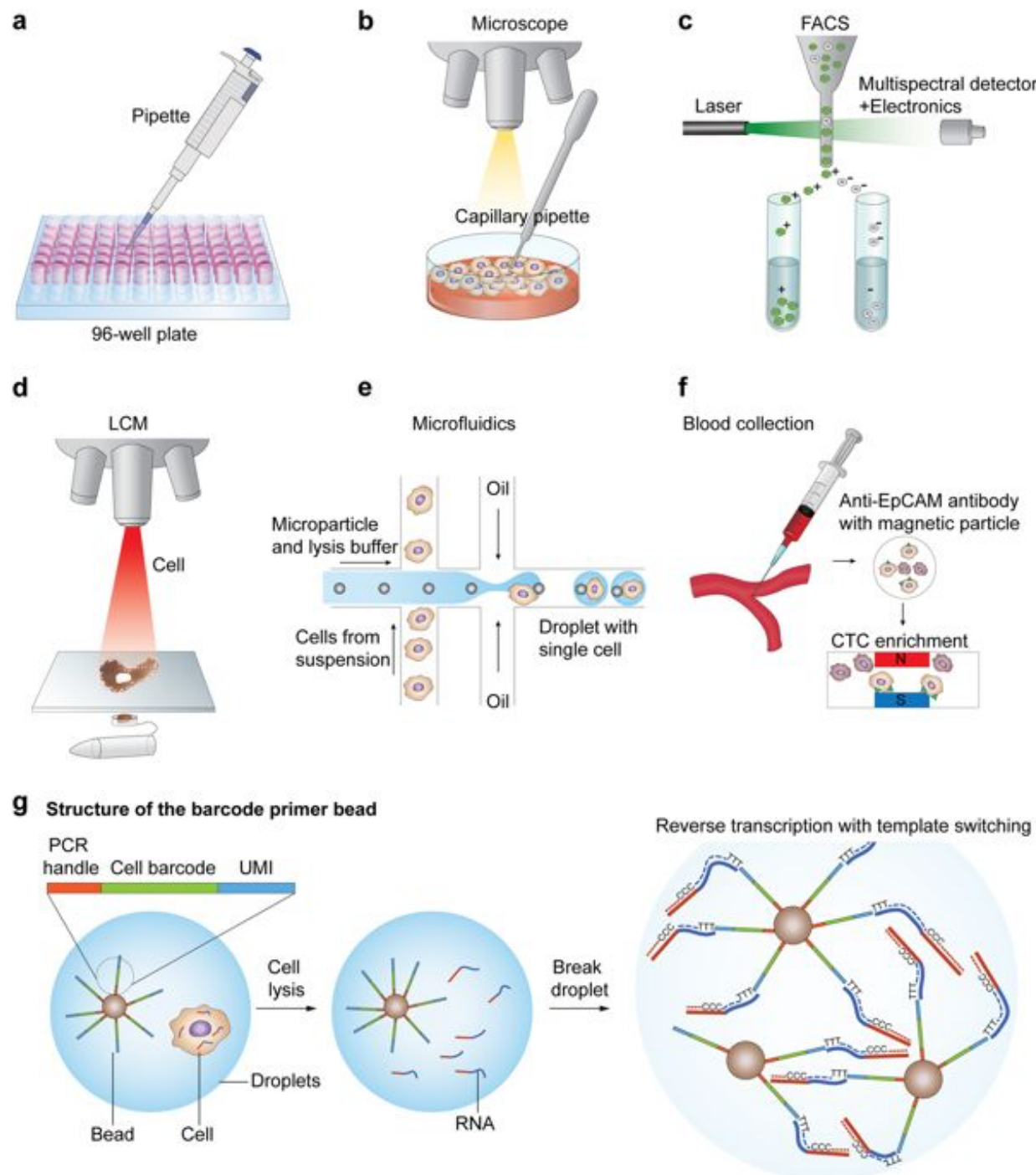


GTExPortal
www.gtexportal.org

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.



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

Library Preparation

Library preparation converts RNA into a form suitable for sequencing. Different sequencing platforms use distinct protocols: some require converting RNA into complementary DNA (cDNA) before sequencing, while others can directly sequence RNA molecules. Understanding these differences helps choose the right approach for your experiment.

Whole-Transcriptome Sequencing with Illumina

			
	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

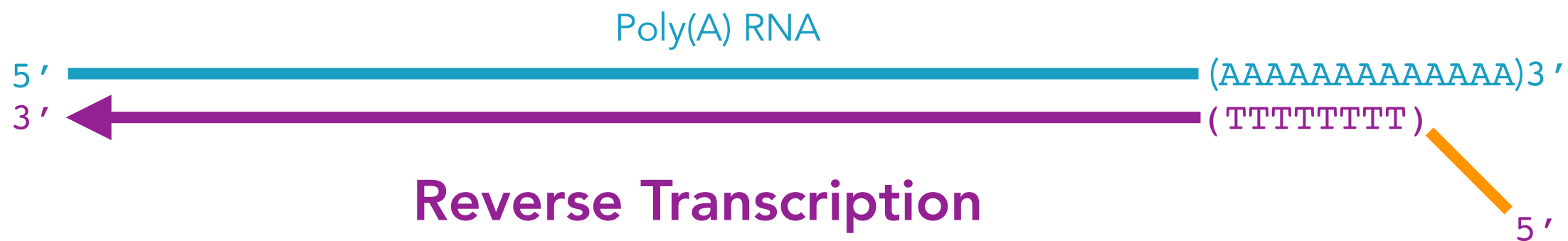
mRNA



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

3'-Method



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

removal of RNA



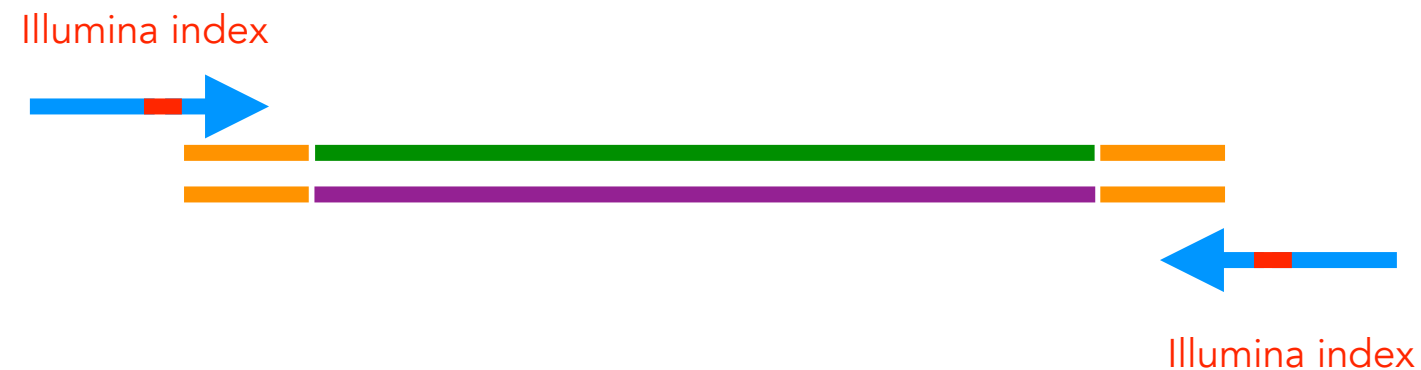
Reverse Transcription



Double-stranded cDNA library



Library Amplification for Multiplexing

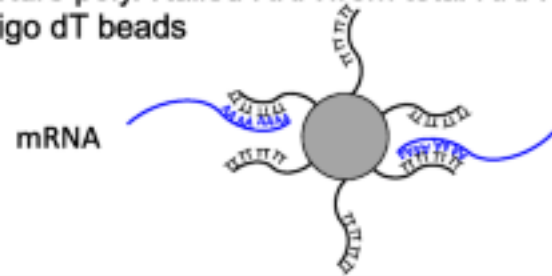


Possible Bias:

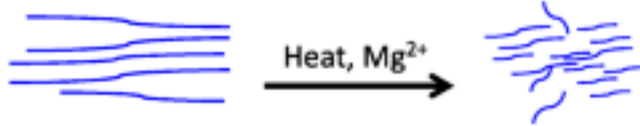
- over-representation of transcript end
- non-random starting point
- short fragments are preferred

Traditional method (KAPA)

Step 1: Capture polyA tailed RNA from total RNA using magnetic oligo dT beads



Step 2: mRNA fragmentation



Step 3: 1st strand synthesis with random primers



Step 4: 2nd strand synthesis with dUDP



Step 5: A-tailing and barcoded adapter ligation



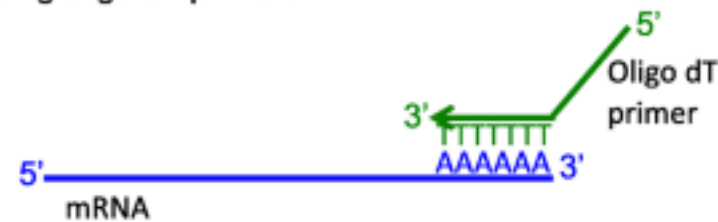
Step 6: Amplification (dUTP strand is not amplified)



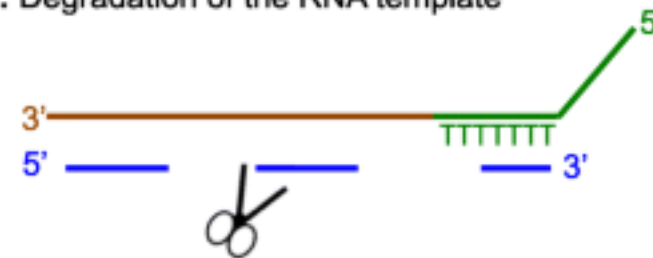
Step 7: Sequencing

3' method (LEXO)

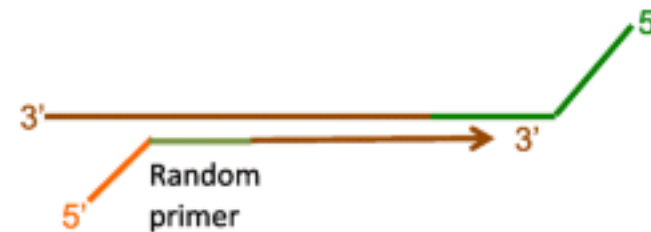
Step 1: 1st strand synthesis of polyA tailed RNA from total RNA using oligo dT primers



Step 2: Degradation of the RNA template



Step 3: 2nd strand synthesis with random primers containing 5' Illumina-compatible linker sequences



Step 4: Amplification using random primers that add barcodes and cluster generation sequences



Step 5: Sequencing

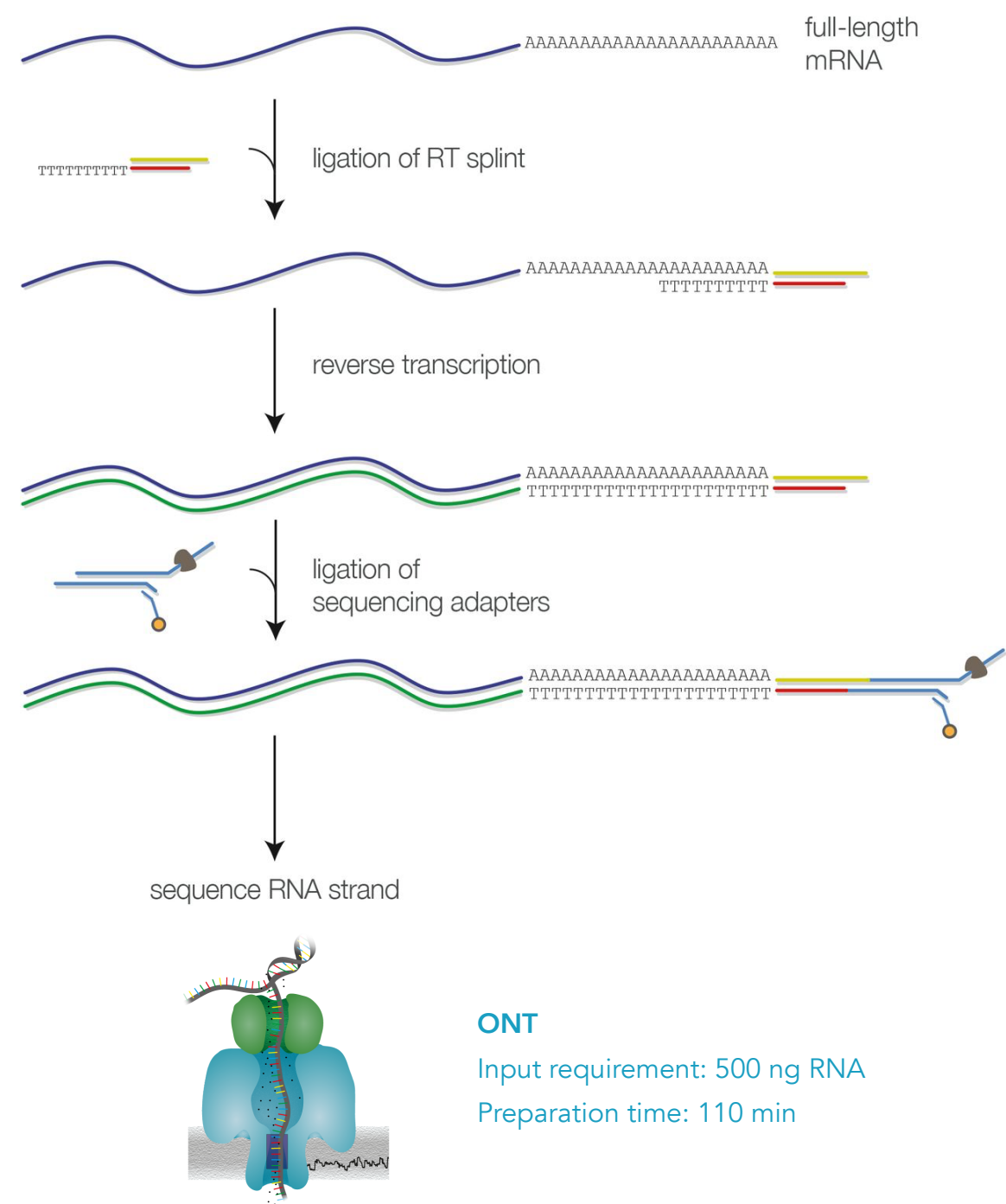
Source: Ma et al. (2019) BMC Genomics 20/9.

The number of differentially expressed transcripts detected by the Trad-KAPA and 3'-LEXO, before and after subsampling from 10 million reads

Sequencing Depth	Trad-KAPA	Intersection (with 10 m)	3'-LEXO	Intersection (with 10 m)	Intersection (Trad-KAPA and 3'-LEXO)
1 million	343	339 (98.8%)	257	249 (96.9%)	177
2.5 million	758	742 (97.9%)	474	460 (97.0%)	329
5 million	1234	1194 (96.8%)	777	740 (95.2%)	562
10 million	1982	1982	1157	1157	882

Source: Ma et al. (2019) BMC Genomics 20/9.

Direct RNA Sequencing Kit



Data Filtering

Filtering RNA-seq data is an important step to remove low-quality or uninformative reads and genes. However, the need for filtering depends heavily on the characteristics of each dataset - such as sequencing depth, sample quality, and experimental design. Because of this variability, general filtering recommendations don't always apply, and filtering strategies should be tailored to the specific data and research questions.



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

Transcript Integrity

RNA integrity is crucial for reliable RNA-seq results. Degraded RNA can lead to biased gene expression measurements and affect data quality.

Sigurgeirsson et al. (2014) found that 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, Lundberg J. (2014) Sequencing degraded RNA addressed by 3' tag counting. PLoS One.

METHODOLOGY ARTICLE

Open Access



Measure transcript integrity using RNA-seq data

Liguo Wang^{1†} , Jinfu Nie^{1†}, Hugues Sicotte¹, Ying Li¹, Jeanette E. Eckel-Passow¹, Surendra Dasari¹, Peter T. Vedell¹, Poulami Barman¹, Liewei Wang³, Richard Weinshiboum³, Jin Jen⁴, Haojie Huang⁵, 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



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 ($1 \leq \text{RIN} \leq 10$) 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 differ 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 ($0 \leq \text{TIN} \leq 100$) for each expressed transcript, however, the medTIN (i.e. median 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

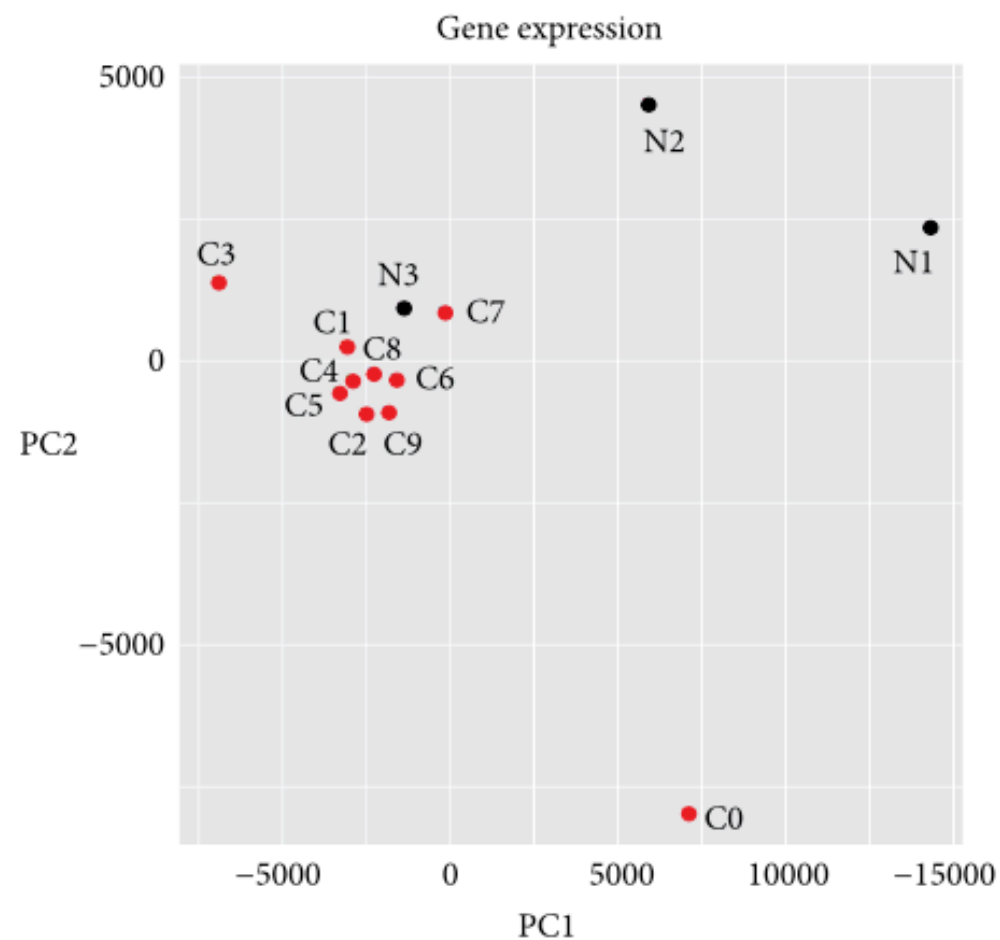
¹Department of Microbiology, College of Natural Sciences, Dankook University, Cheonan 31116, Republic of Korea

²Department of Clinical Laboratory Science, Semyung University, Jecheon 27136, Republic of Korea

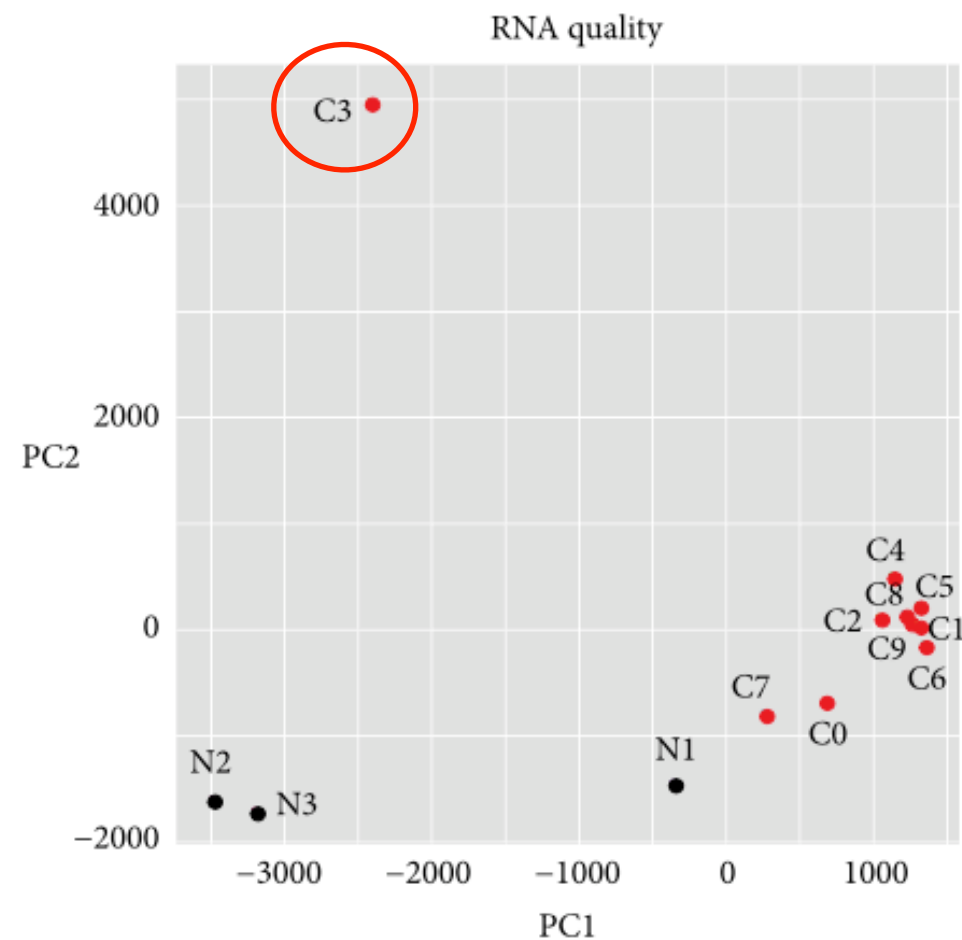
³Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan 31116, Republic of Korea

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).

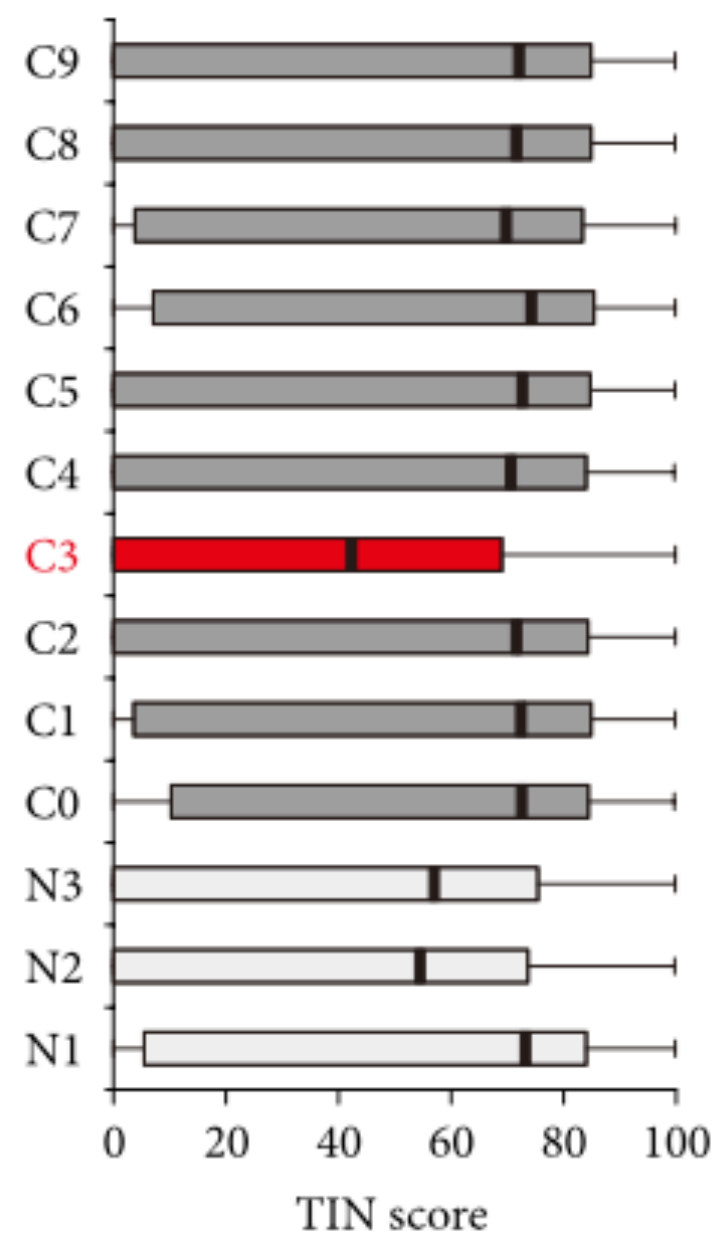


The gene expression PCA plot provides a map of the distances between samples from which the characteristics of RNA-seq data can be inferred.



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.

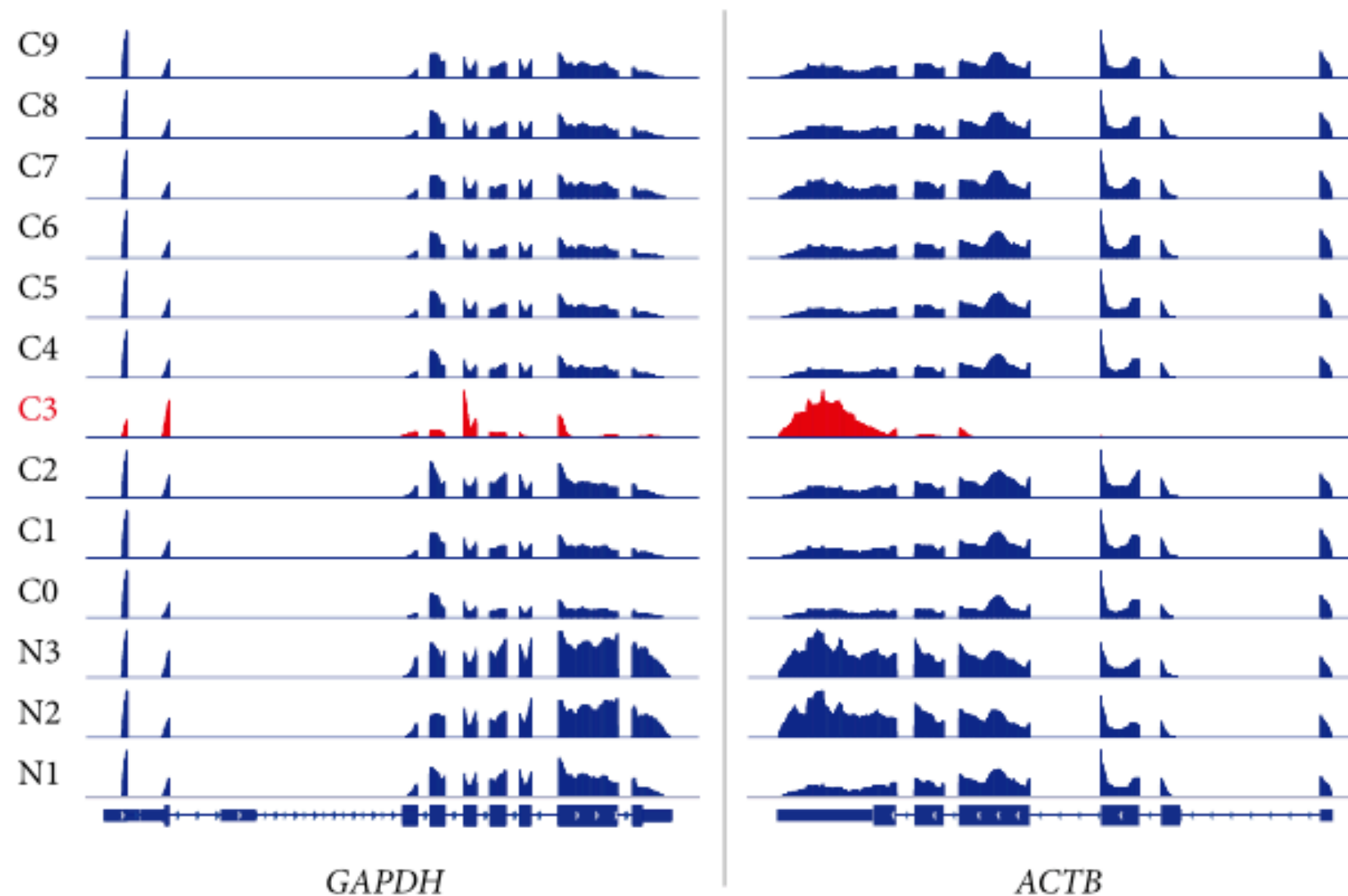
Source: Son et al. (2018). A Simple Guideline to Assess the Characteristics of RNA-Seq Data. BioMed research international.



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.

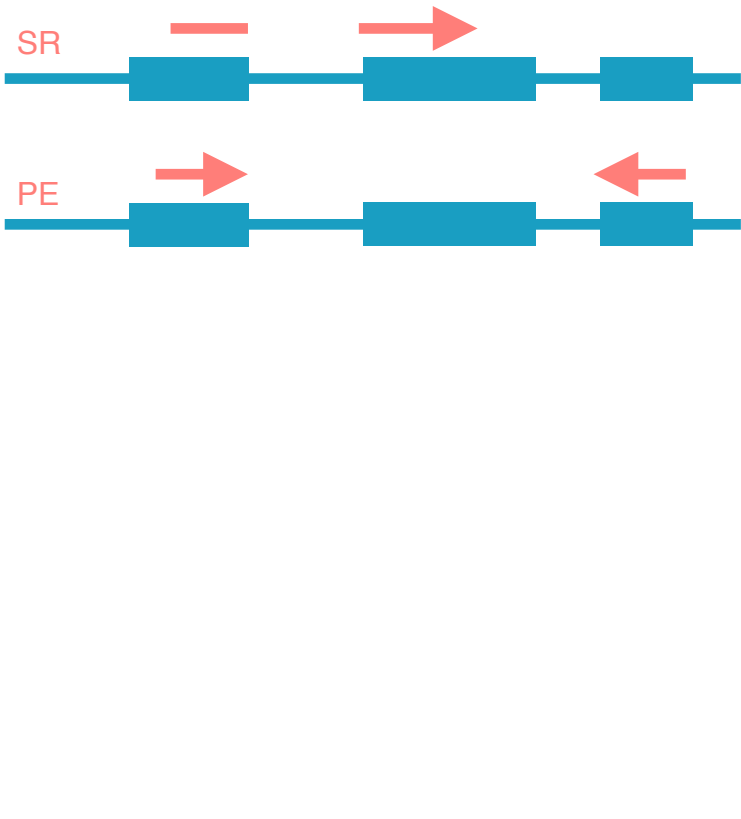
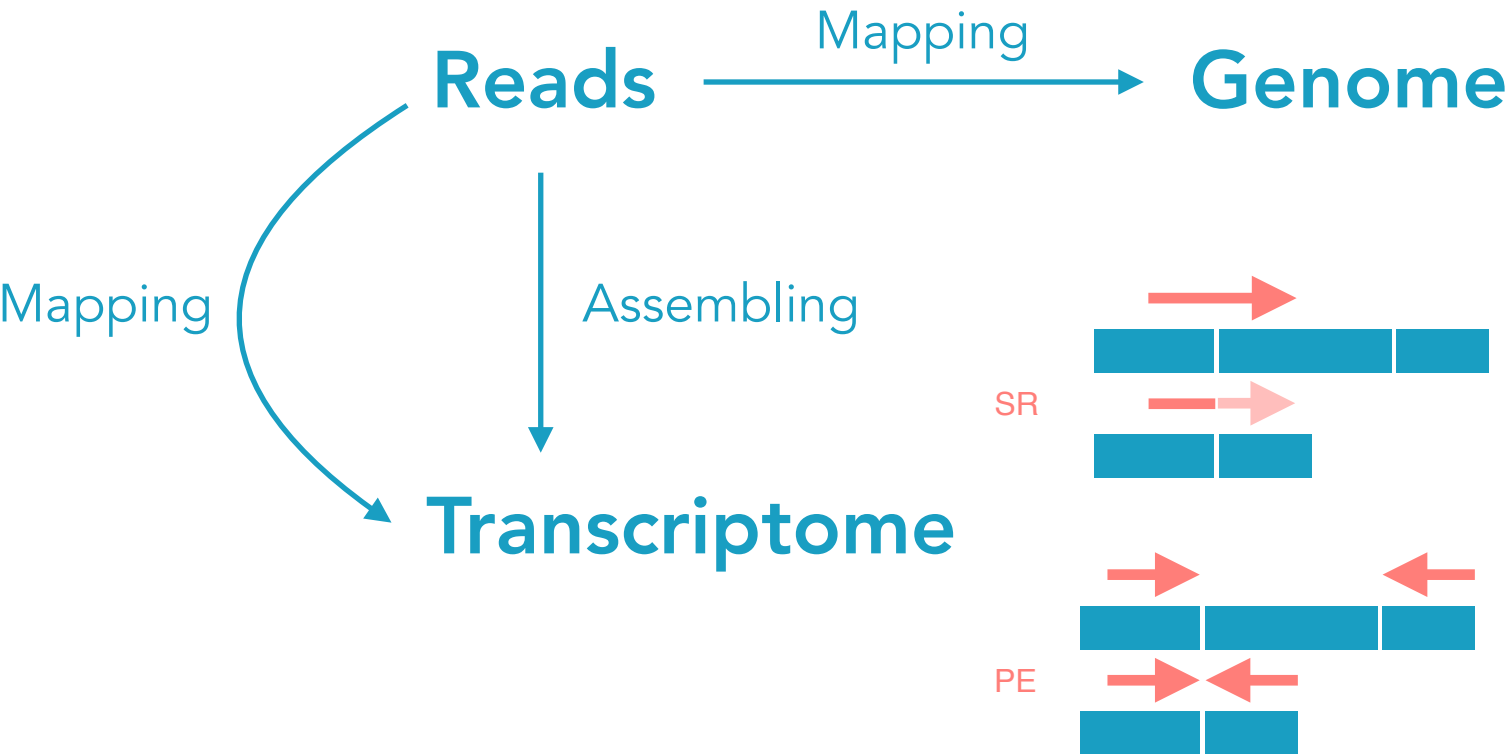
Genome browser snapshots of mapped read densities are shown using integrative genomics viewer (IGV). FPKM, fragments per kilobase of transcript per million mapped reads.



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

Read Mapping

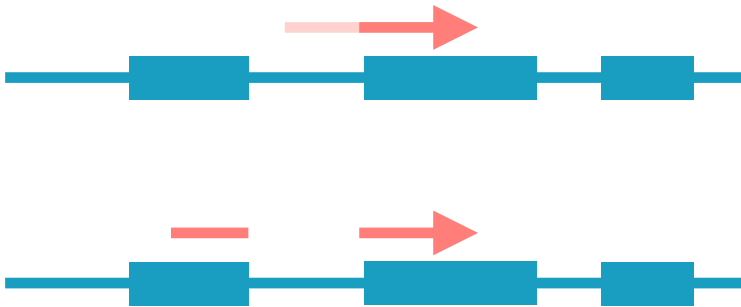
Read mapping is the process of assigning sequencing reads to their origin in the genome or transcriptome. Traditional methods align reads to a reference genome or transcriptome, while newer pseudo-mapping approaches use k-mer-based strategies to quickly assign reads without full alignment. The choice of method depends on the read type, experiment goals, and downstream analysis. Understanding these options is important before starting mapping and producing **count tables** to ensure accurate and meaningful results.



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

Minimap2	spliced (long-read)
GMAP	spliced (long-read)
GraphMap	spliced (long-read)



- **Minimap2** is widely used for mapping long noisy reads (e.g. from Oxford Nanopore or PacBio) and supports spliced alignment for RNA-seq.
- **GMAP** is specifically designed for aligning cDNA and RNA sequences.
- **GraphMap** also supports long-read alignment and handles RNA reads with splicing.

Traget Length



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



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

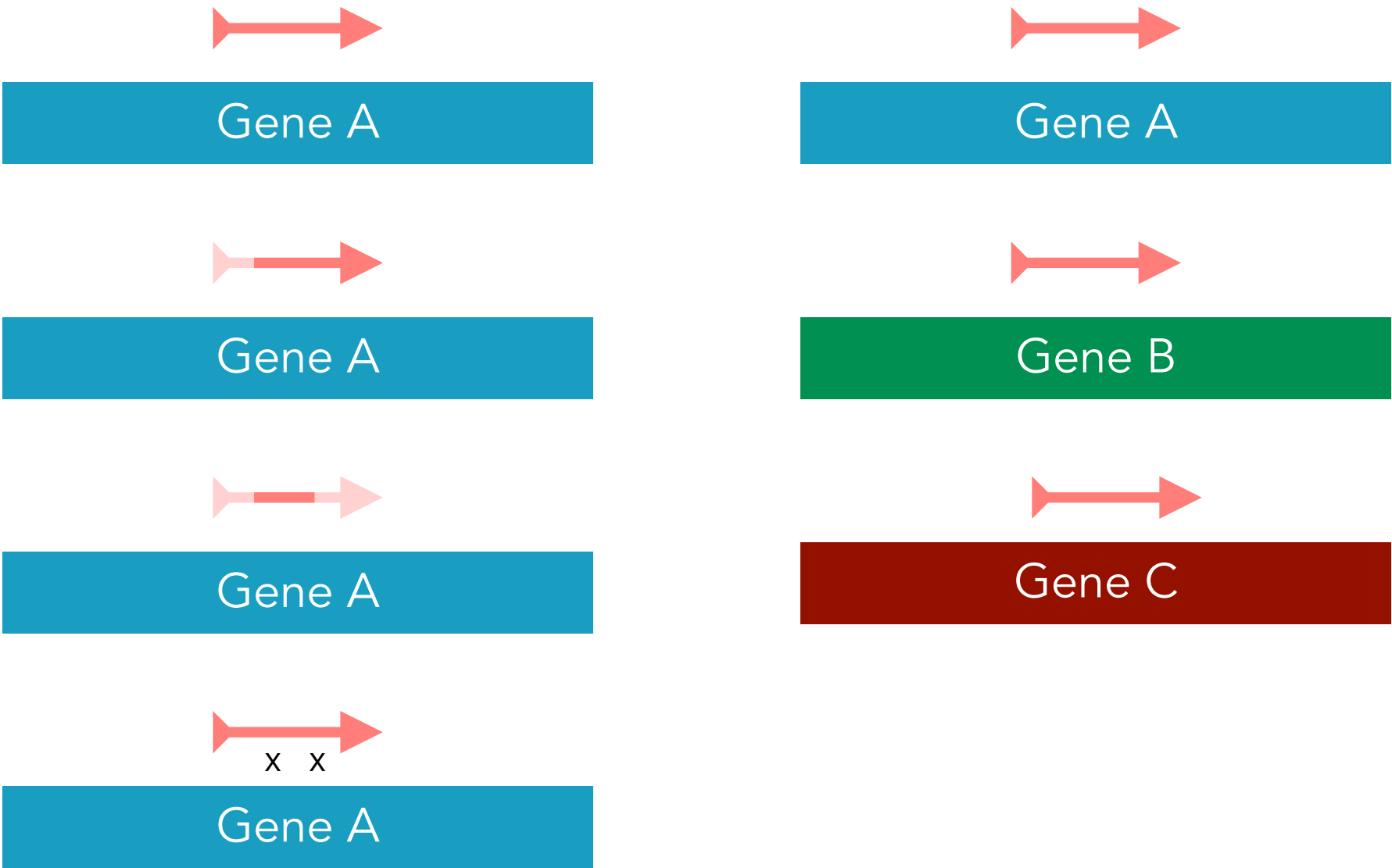


$$10 \times 1.5 = 15 \rightarrow \frac{15}{10} = 1.5$$



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

Mapping Quality



Traget Coverage



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



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



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

Mapping RNA-seq reads to a reference is complex and introduces several challenges that can affect downstream analyses:

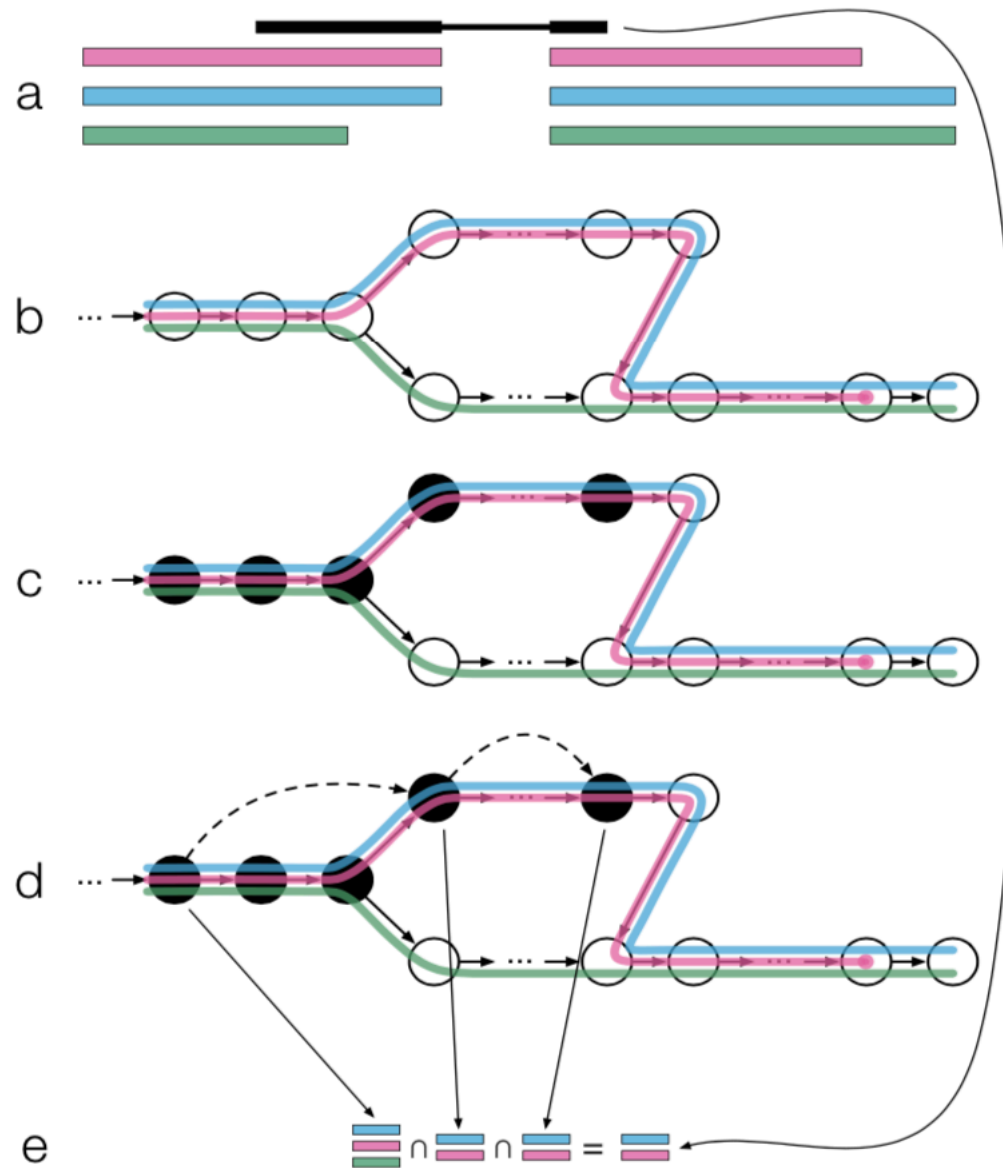
- **Coverage Bias:** Read coverage is rarely uniform across transcripts. Factors like GC content, fragmentation, or priming strategies can skew coverage, leading to under- or overrepresentation of transcript regions.
- **Even Coverage:** Ideally, reads should cover transcripts evenly, but biases in sequencing or library prep (e.g. 3' bias in poly(A) selection) can distort the true expression landscape.
- **Mapping Quality:** Low-complexity regions, repetitive sequences, or short reads can result in ambiguous mappings, reducing mapping quality and reliability of quantification.
- **Soft Clipping:** Mappers may "soft clip" ends of reads that don't align well, often due to sequencing errors or adapter contamination. This can affect alignment statistics and interpretation.
- **Splicing:** Eukaryotic RNA includes introns that must be accounted for. Splice-aware aligners (like HISAT2 or STAR) are needed to correctly map reads that span exon-exon junctions.
- **Multimapping Reads:** Reads that map equally well to multiple locations (e.g. in gene families or pseudogenes) pose a challenge. How these are handled can impact expression estimates.
- **Read Length and Type:** Short reads may be harder to map uniquely, especially in complex transcriptomes. Paired-end and longer reads improve mapping precision but increase computational demands.

Successfully navigating these issues requires choosing appropriate tools and carefully evaluating mapping outputs before proceeding to quantification.

Traditional alignment methods attempt to match each read to its exact position in the genome or transcriptome, accounting for splicing, mismatches, and other complexities. While accurate, this process can be computationally intensive and sometimes unnecessarily detailed for quantification purposes. This leads us to an alternative approach — **pseudo-alignment**.

Pseudo-alignment is a lightweight mapping strategy used primarily for transcript quantification. Instead of aligning each read base-by-base, pseudo-aligners (like *kallisto*, *Salmon*, or *Sailfish*) determine which transcripts a read could originate from based on shared k-mers. This makes the process much faster while still providing accurate expression estimates. Pseudo-alignment does not produce traditional alignment files (like BAM), but outputs transcript-level abundance estimates directly — ideal for differential expression analysis pipelines.

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, \dots) 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.

Data Analysis

RNA-seq **count tables** are the backbone of expression analysis, but they come with challenges. The data are **high-dimensional** (many genes, few samples), often **sparse**, and typically **overdispersed**, requiring specialized models like the negative binomial. Counts are also affected by technical noise, sequencing depth, and batch effects. Careful normalization, filtering, and statistical modeling are essential to extract reliable biological insights.



Online version

Fast-track your RNA-seq insights in hours – not days

Is RNA-seq data analysis often a bottleneck for you, negatively impacting your downstream experiments? What if a bioinformatics solution allows you to seamlessly build gene expression workflows to scale up your RNA-seq projects without requiring extensive bioinformatics expertise?

Our RNA-seq Analysis Portal is an easy online tool that is:

- **Easy and fast:** Seamless transition from raw sequencing data (FASTQ files) to differential gene expression, pathway analysis and downstream validation tools in just one day
- **Compatible and flexible:** Supports RNA-seq data generated by all commonly used RNA library preparation kits
- **Economical:** Forget about committing to an annual license agreement. Purchase GeneGlobe Analyze Credits to customize what you spend as per your need to access the portal

Deepen, improve and accelerate your RNA-seq analyses anytime from the comfort of your lab, home or even a coffee shop.

RNA-Seq data is ...

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

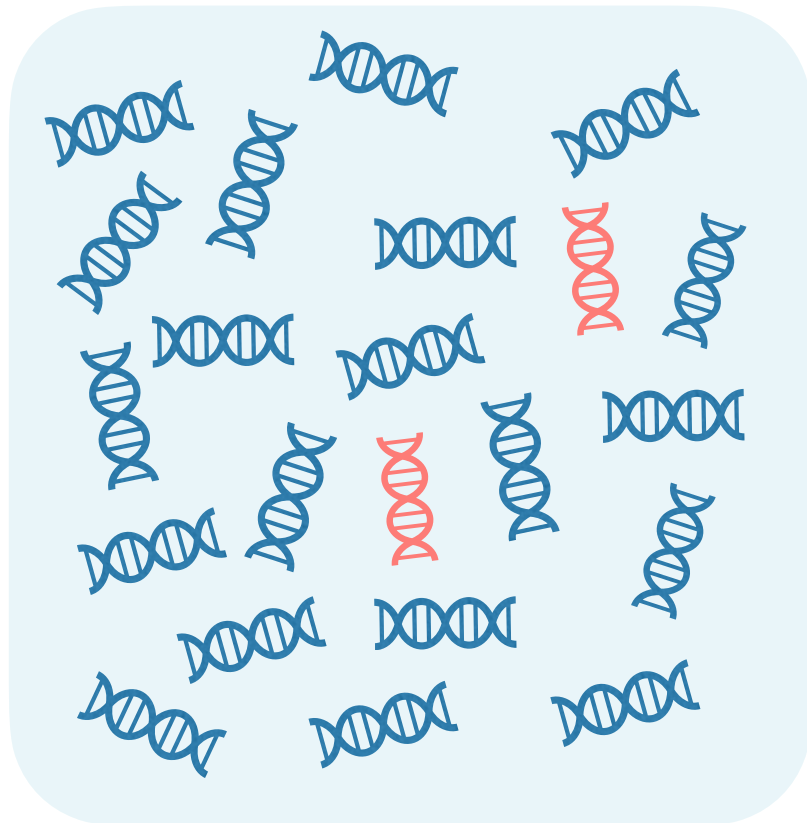
(b) **high dimensional** (many variables/genes)

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

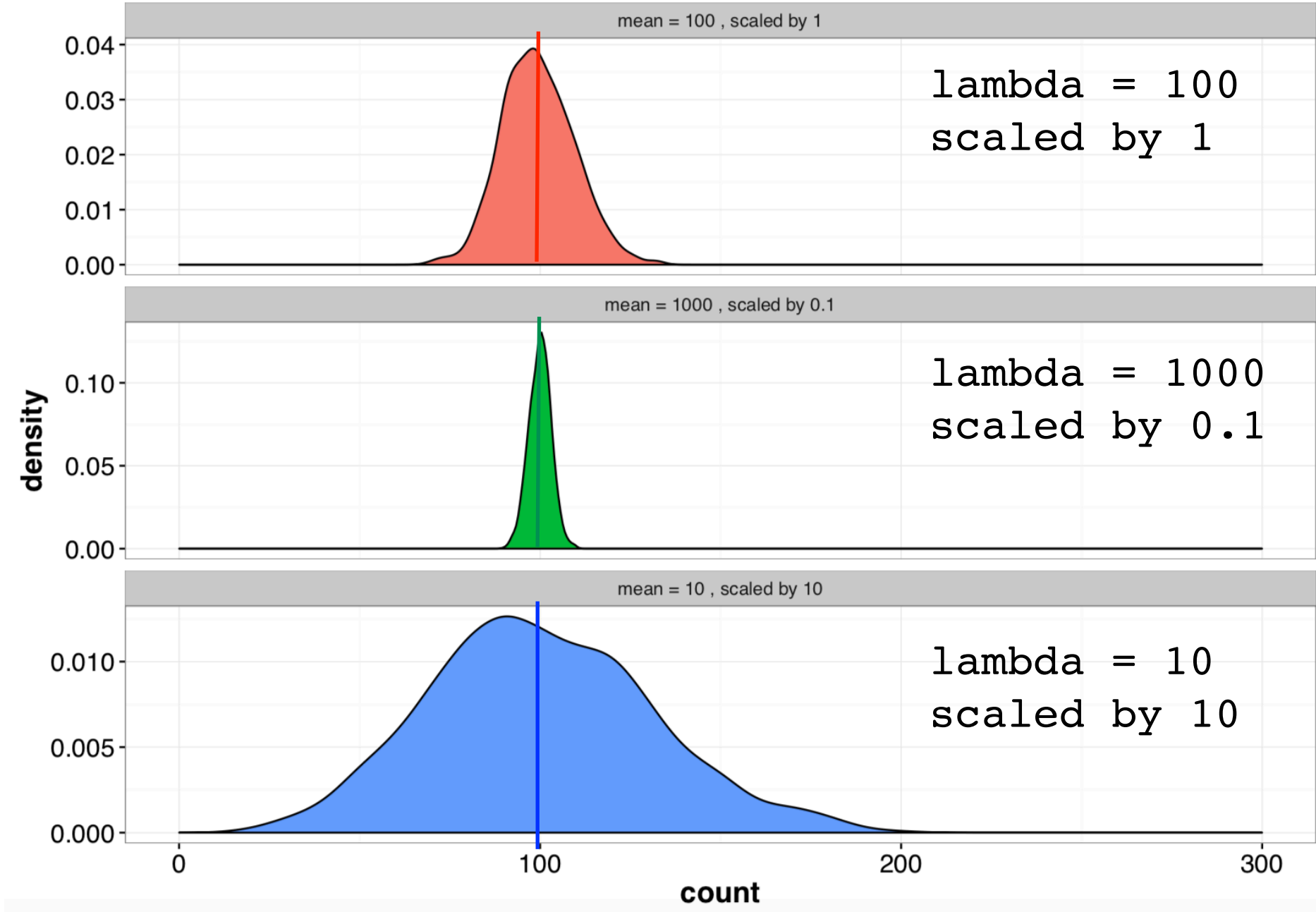


Subsamples:

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 & Compositionality - Technical variation during sequencing results in varying sequencing depths. To reduce/remove sequencing depth variation, counts should be normalised. 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 (expression table) are zero rich. While log-ratios (network inference in general) can be used to address compositionality, it is sensitive to zeros (i.e. negative infinities). Pseudo-counts could solve the problem, but could affect the results by changing the covariance structure of the data. Alternative treatments of zeros have been proposed but are problematic as zeros could indicate missing or under-sampling.

```
set.seed(240617)
x2 <- sample(1:100, 10, replace = TRUE)
y2 <- sample(1:100, 10, replace = TRUE)
cor(x2,y2)
# 0.466
x1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
y1 <- sort(sample(1:100, 10, replace = TRUE), TRUE)
cor(x1,y1)
# 0.883
x3 <- c(x2, rep(0,20))
y3 <- c(y2, rep(0,20))
cor(x3,y3)
# 0.790
```

easyRNASeq

DEGseq

DESeq / DESeq2

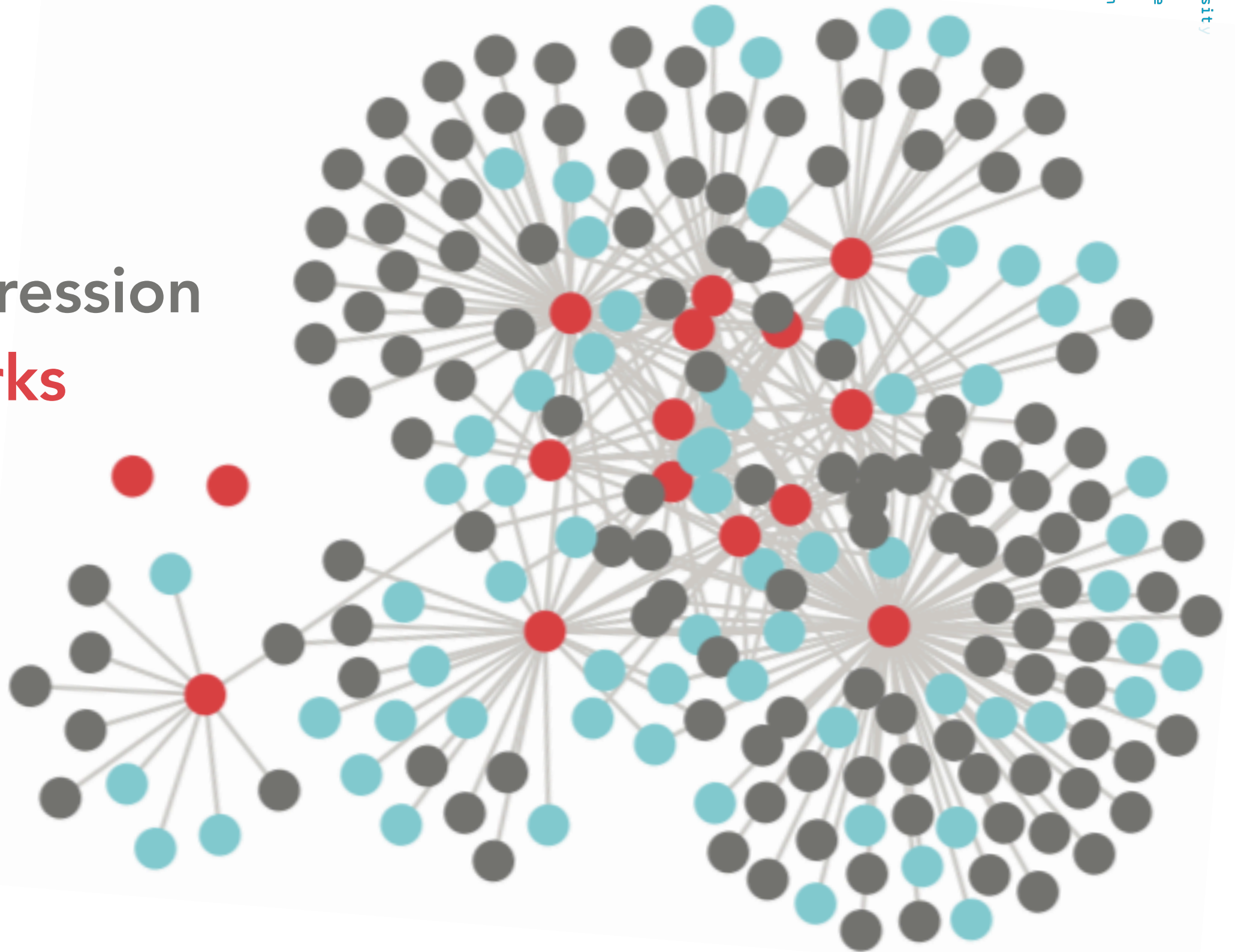
NOISeq

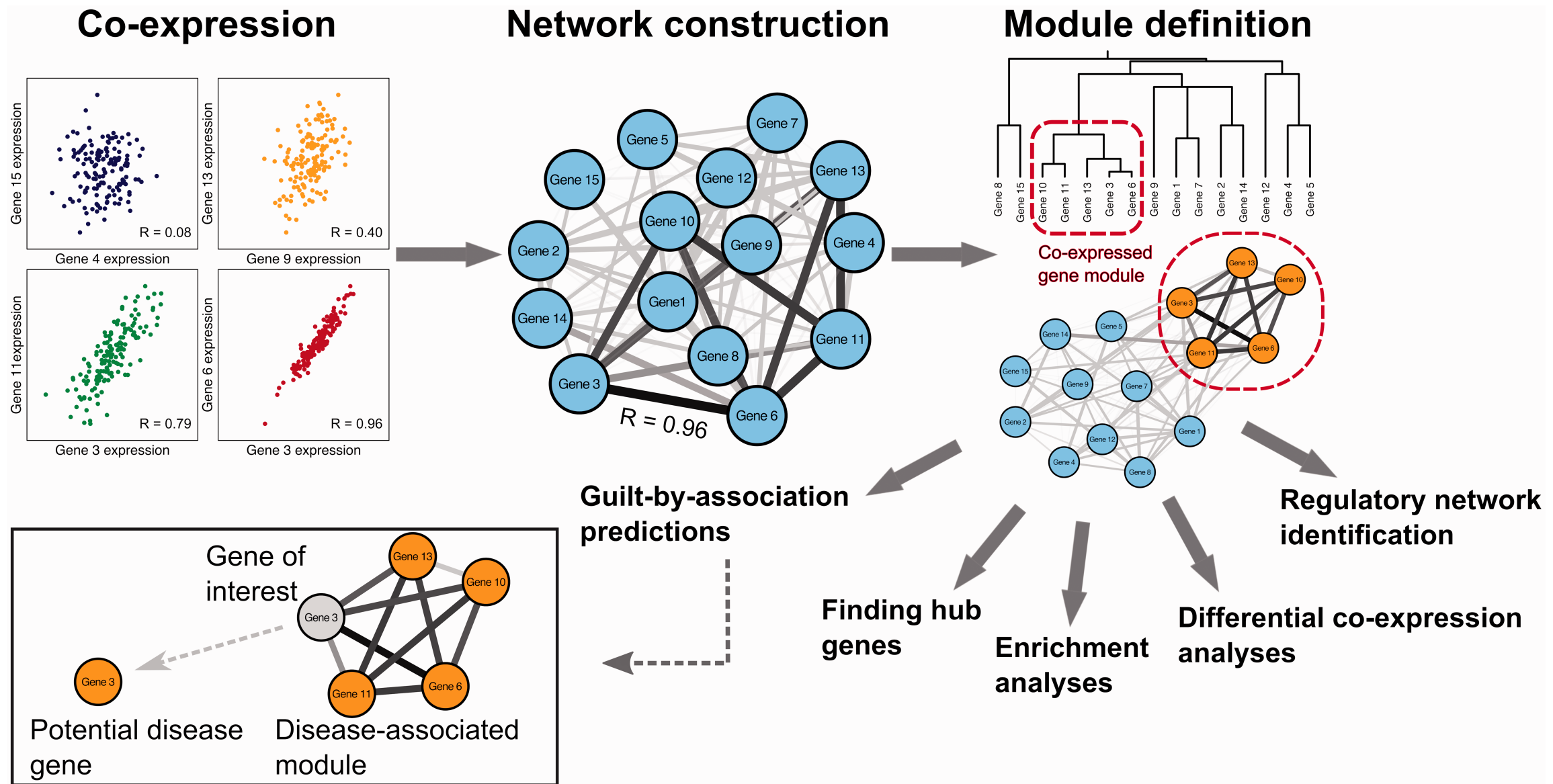
edgeR

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.

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 Brinza³, Camille Marchet⁴, Janice Kielbassa⁵,
Sylvère Bastien^{1,2}, Mathilde Boutigny^{1,2}, David Monnin¹, Adil El Filali¹, Claudia
Marcia Carareto⁶, Cristina Vieira^{1,2}, Franck Picard¹, Natacha Kremer¹, Fabrice Vavre^{1,2},
Marie-France Sagot^{1,2} and Vincent Lacroix^{1,2,*}

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

RNA-seq: Power and Limits Across Systems

RNA-seq is a powerful and reliable tool, particularly for well-annotated model organisms, for which genomes and gene functions are well characterised. However, challenges arise more quickly in wild or non-model systems: incomplete or low-quality genomes, sparse functional annotation, and unpredictable transcript diversity make analysis and interpretation more difficult. Many differentially expressed genes lack a known function, and attributing roles based on model organisms can be risky, since gene function may not be conserved and genes often fulfil multiple functions.

Furthermore, these systems tend to exhibit higher levels of biological variability, resulting in a higher coefficient of variation (CV). Consequently, more biological replicates are required to reliably detect real effects. In such contexts, careful experimental design is as important as sequencing depth or the analysis pipeline.

A Quick Recap

1

Question

Start with a precise scientific question.

Gather Knowledge

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

2

3

Design

Think carefully about the design 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.

**MORE
THINGS
CONSIDERED**

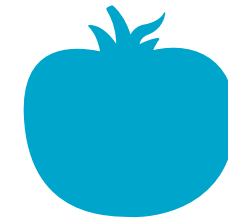
1. Taste Is Polygenic and Complex

"Perfect taste" is subjective and influenced by many genes—those controlling sugar accumulation, acid metabolism, volatile organic compounds (for aroma), cell wall breakdown (texture), etc. There's no single "taste gene."

2. RNA-seq Is a Snapshot

RNA-seq tells you what's being expressed at a given time in a specific tissue. But taste develops over time, and expression changes depending on:

- Developmental stage
- Tissue type (skin vs. pulp)
- Environmental factors (light, temperature, water stress)
- Post-harvest ripening



So, your study would need very careful experimental design to capture relevant differences.

3. Expression ≠ Function

Differential expression doesn't prove causation. A gene might be highly expressed in tasty fruit, but:

- It could be a downstream effect.
- The key change might be in a regulatory region or involve epigenetics, not transcript abundance.