Galaxy for linking bisulfite sequencing with RNA sequencing – Introduction to sequencing data analysis

Markus Wolfien, Andrea Bagnacani, Olaf Wolkenhauer

9th October 2019, Freiburg

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

deSTAIR
destair.bioinf.uni-leipzig.de

fli Leibniz Institute on Aging – Fritz Lipmann Institute

UNI FREIBURG

Universität Rostock Traditio et Innovatio

Outline

- Who are we?
- Gene expression
- Techniques to measure gene expression
- RNA Sequencing (RNA-Seq)
- Tools and materials

Who are we?

https://denbi.de

The *German Network for Bioinformatics Infrastructure* - **de.NBI** consists of eight service units that provide distinct bioinformatics services according to their areas of scientific expertise and their bioinformatics resources. The units complement each other in terms of thematic priorities, implementing user services in different areas of Life Sciences research and in industry. The services are complemented by comprehensive training activities.

Who are we?

https://destair.bioinf.uni-leipzig.de

The *Structured Analysis and Integration of RNA-Seq experiments* - **de.STAIR** provides comprehensive analyses of RNA-Seq experiments as a service. To bring ease of use, reproducibility, and accessibility for the developed approaches and services, we provide dedicated workshops, training programs and screen casts for bioinformaticians and other life scientists.

Who are we?

Who are you? :)

Gene expression

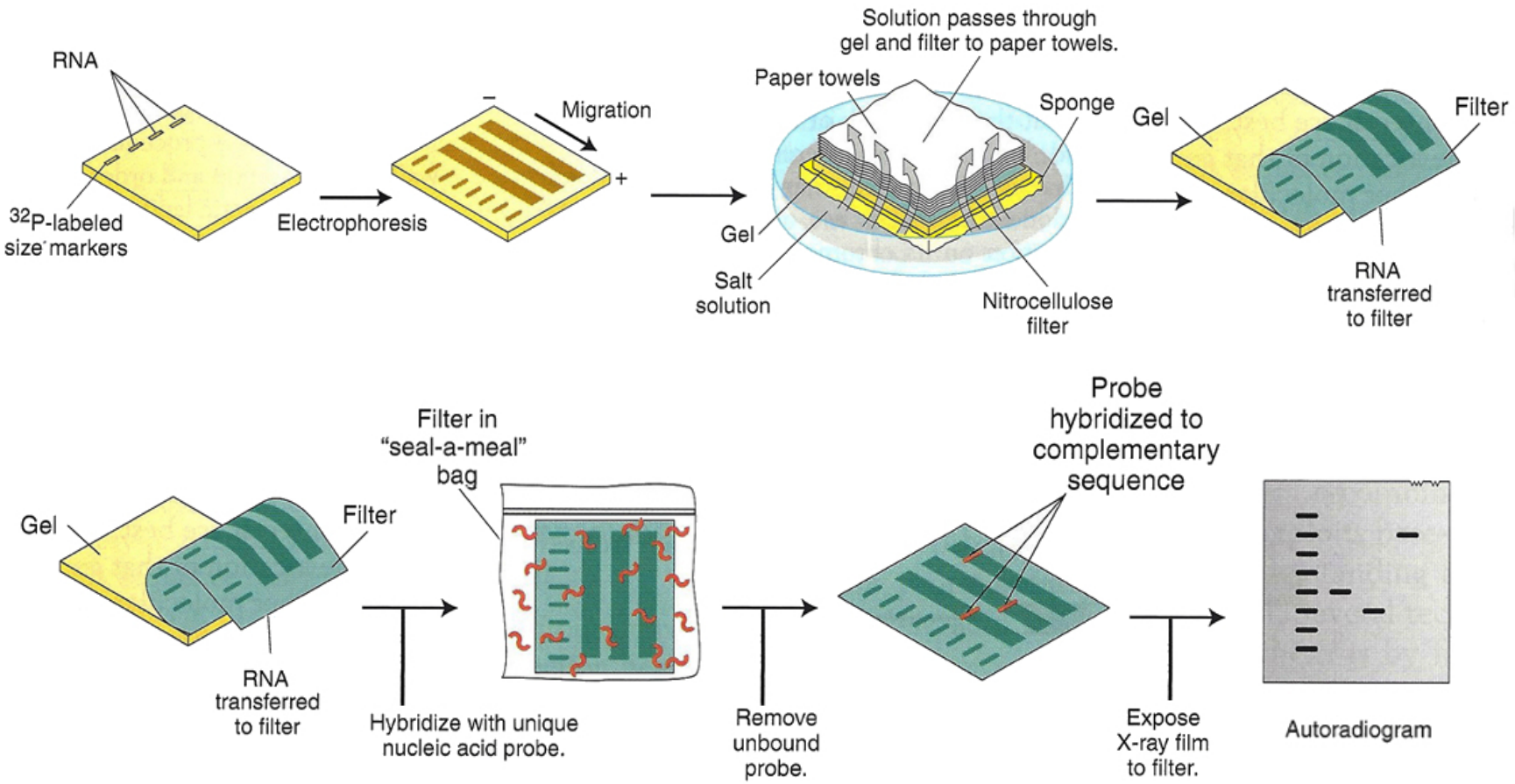
Gene expression is the process by which genes in a DNA strand synthesise *functional gene products*.

Here, functional gene products can be Proteins or regulatory RNAs.

- ...on the production side, proteins determine a cell's function, and therefore a cell's differentiations
- ...on the regulatory side, regulatory RNAs shape the amount of proteins, providing the cell control over its structure and function

Techniques to measure gene expression

Northern blot detects transcribed mRNAs within a mixture of RNA molecules in a sample.



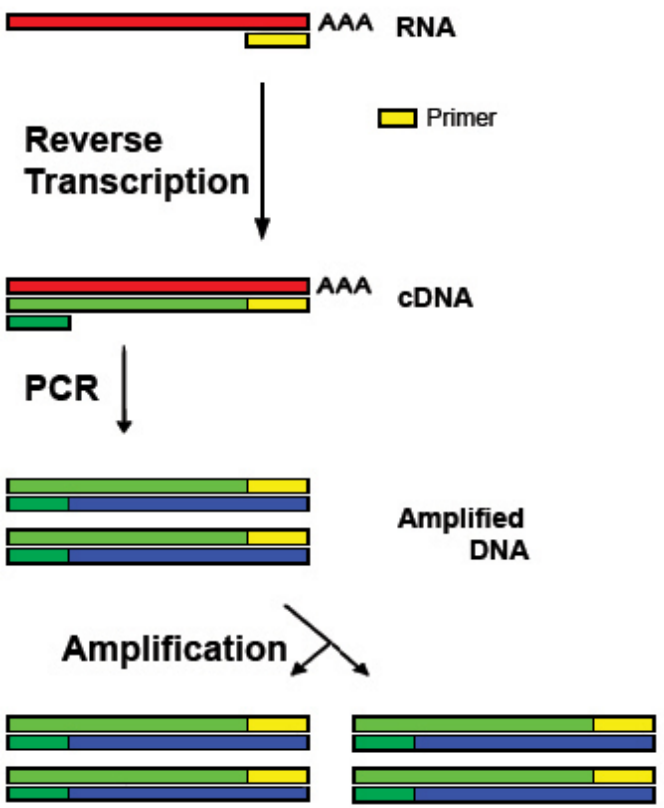
Techniques to measure gene expression

Northern blot procedure:

- RNA backbone has phosphates (negatively charged)
- By applying current to the gel, RNAs migrate towards the positive end, separating by size
- RNA molecules are transfered (blotted) from the gel to a membrane
- RNA molecules are denaturated to loose their 3D structures
- Probes (fluorescent or radioactive) are washed against the membrane
- Probes bind against the complementar RNA molecule
- Exceeding probes are washed off. The ones bound to RNAs, provide a visual display of the presence of the target RNA molecules

Techniques to measure gene expression

Northern blot is quantitatively inaccurate for small samples.



Bustin. Journal of Molecular Endocrinology 2000

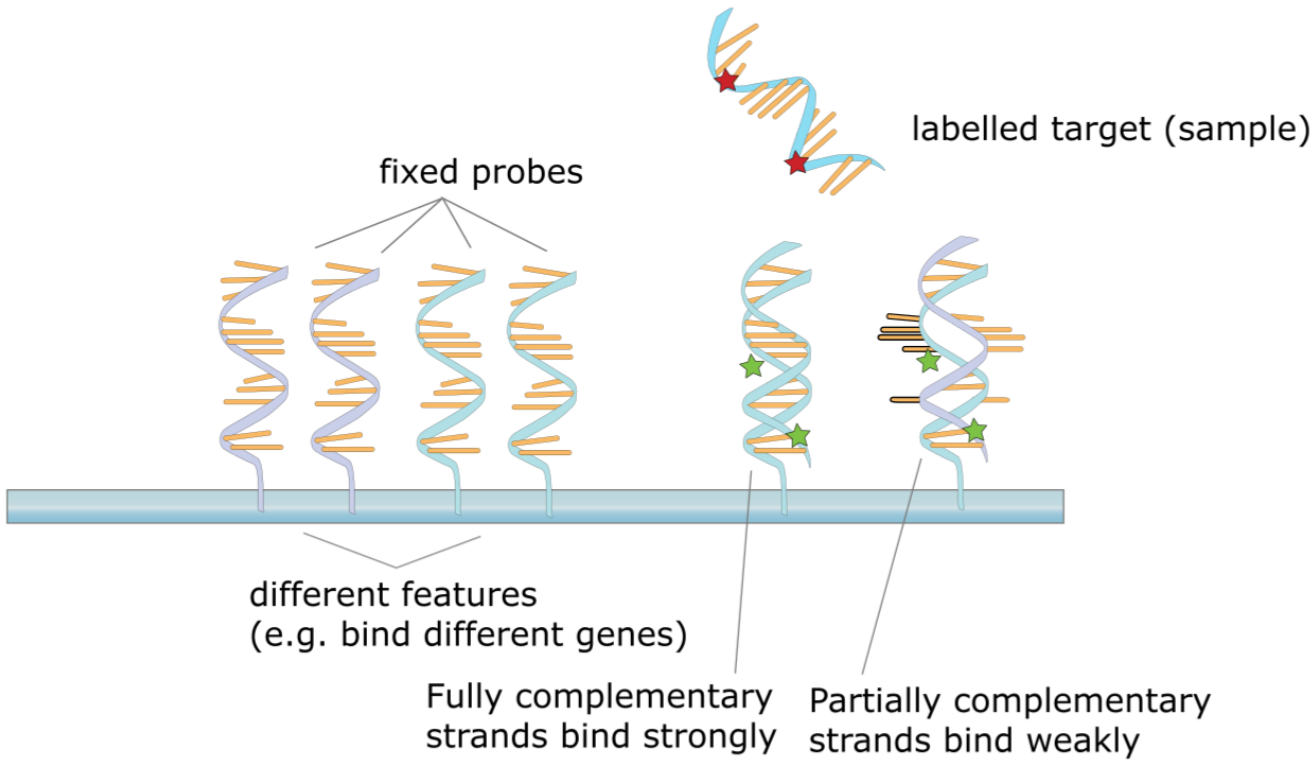
A solution to the inaccuracies coming from the availability of small RNA samples, has been provided by the **Reverse Transcription Polimerase Chain Reaction** (**RT-PCR**).

- RT-PCR amplifies expressed genes by reverse transcribing the RNA of interest into its DNA complement
- This is achieved using reverse transcriptase

Techniques to measure gene expression

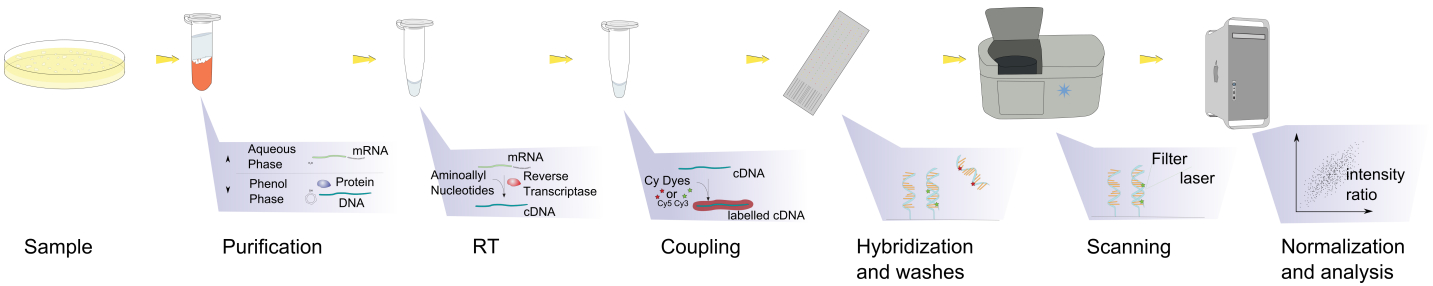
But **Northern blot and RT-PCR are very time consuming**!

Microarray chips represent a step forward in the automation and measurement of gene expression levels, and are able to assess it for large numbers of genes in parallel.



Techniques to measure gene expression

Microarray procedure:



- They consist of a surface organised into wells. Each well has *probes* i.e. known sequences of DNA
- Targets are labelled with a fluorophore that produces a chemoluminescence
- Targets *hybridise* to their corresponding probes, binding tightly or weakly
- The surface is washed to get rid of the weakly bound targets, and then scanned to analyse the chemoluminescence within each well

Techniques to measure gene expression

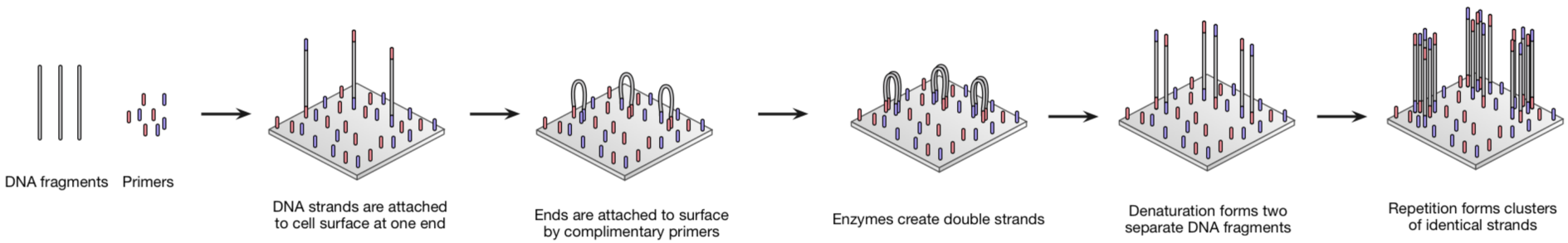
However, **microarray data is difficult to compare and reuse**:

- Lack of standardization in platform production
- Protocol diversity
- Multiple image analysis approaches

$\Longrightarrow$ Interoperability problem

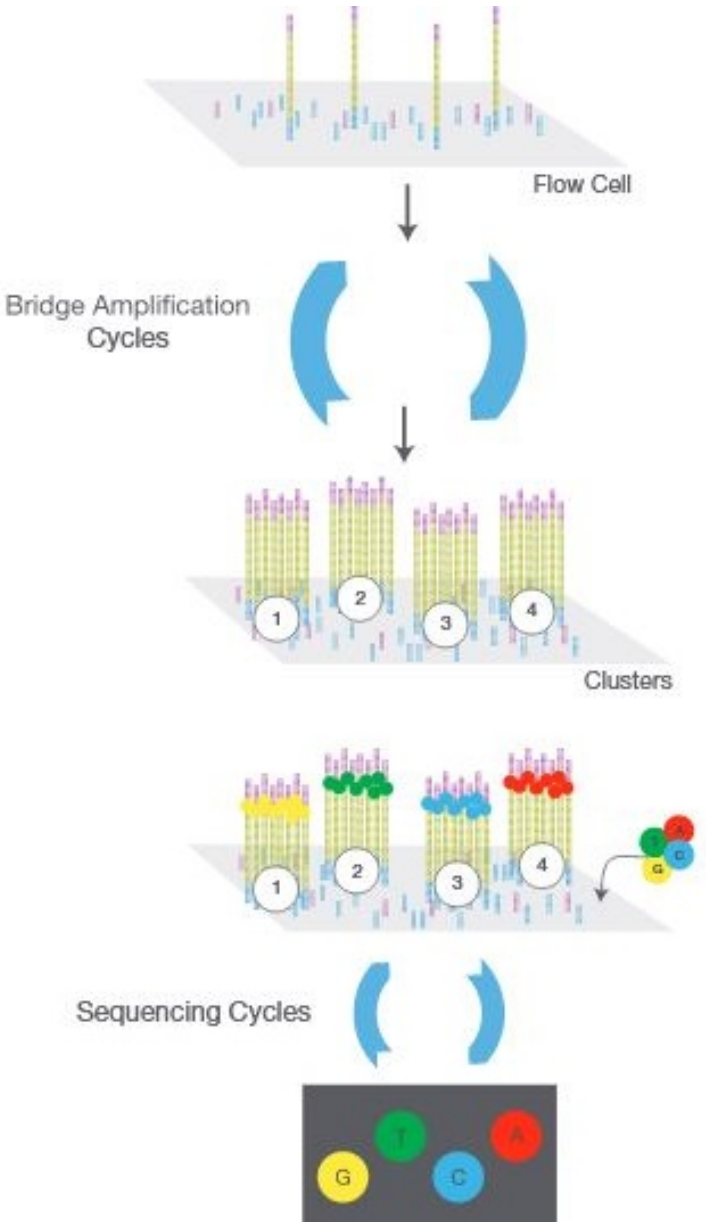
Techniques to measure gene expression

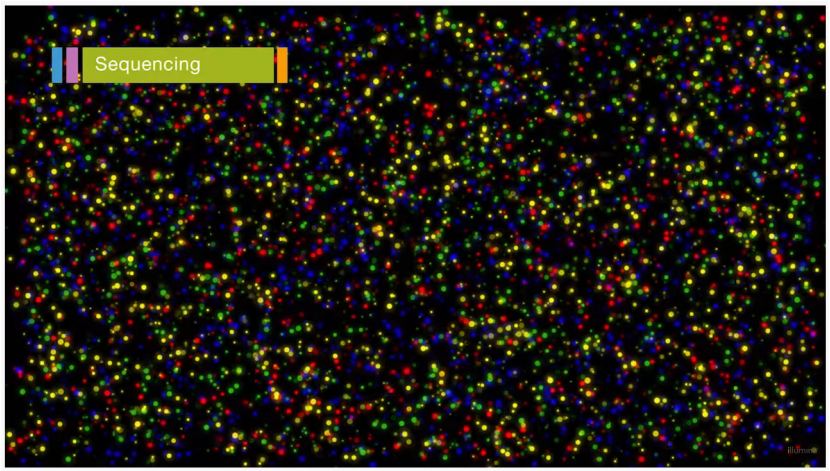
NGS enables massively parallel sequencing of DNA fragments, and can be used to analyse gene expression.



- DNA is isolated, fragmented, and primed on both ends
- Fragments are hybridised to the surface, where PCR amplifies them, creating clusters of DNA fragments
- Nucleotides are incorporated one after the other. They have a fluorophore and a terminator sequence to: 1) emit light, 2) prevent multiple binding
- The platform scans for the luminescence, then washes away fluorophores and terminator sequences

Techniques to measure gene expression

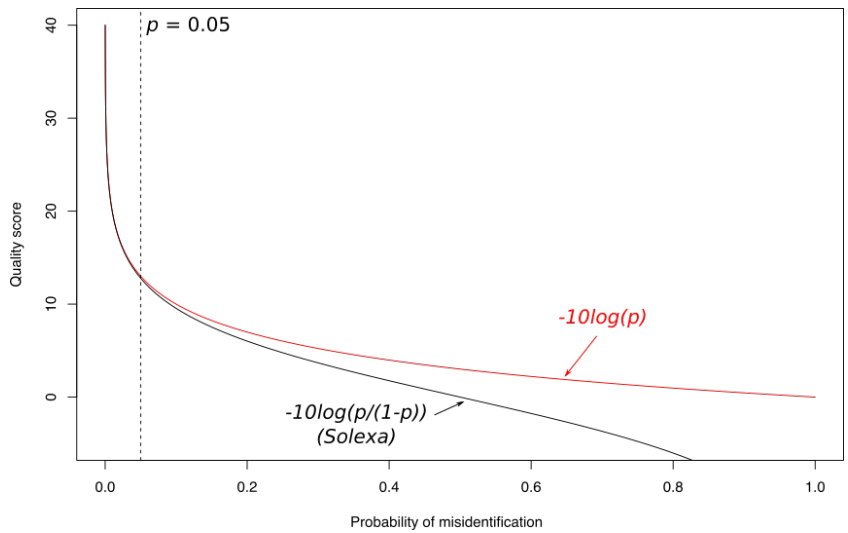


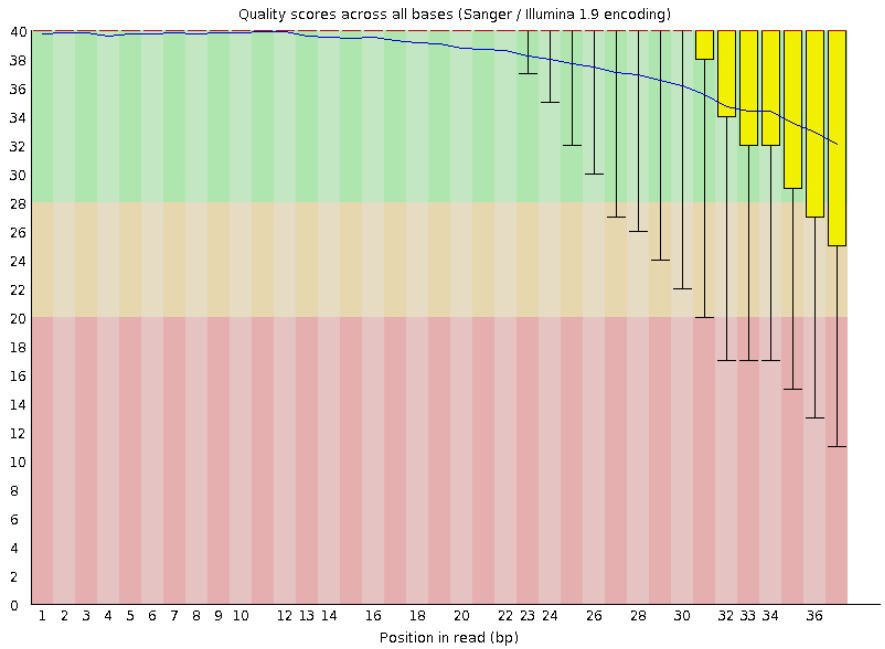


https://www.youtube.com/watch?v=fCd6B5HRaZ8

Oxford Genomics Centre, 2017
Illumina, 2015

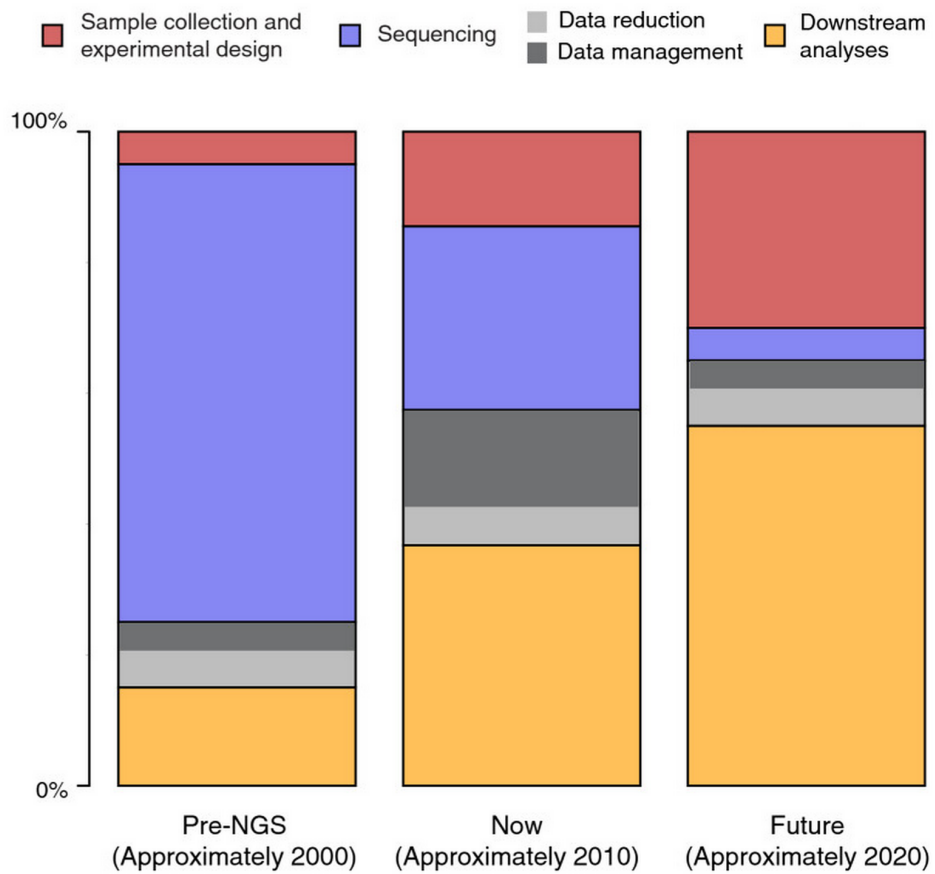
The more the readout process progresses, the less confident the reads become. This is due to the removal of the fluorescent dye, which becomes progressively difficult, leaving background noise to the further readouts.





Techniques to measure gene expression

The decrease of sequencing costs is met with an increased effort in data processing $\Longrightarrow$ Increased need of Bioinformatics expertise.



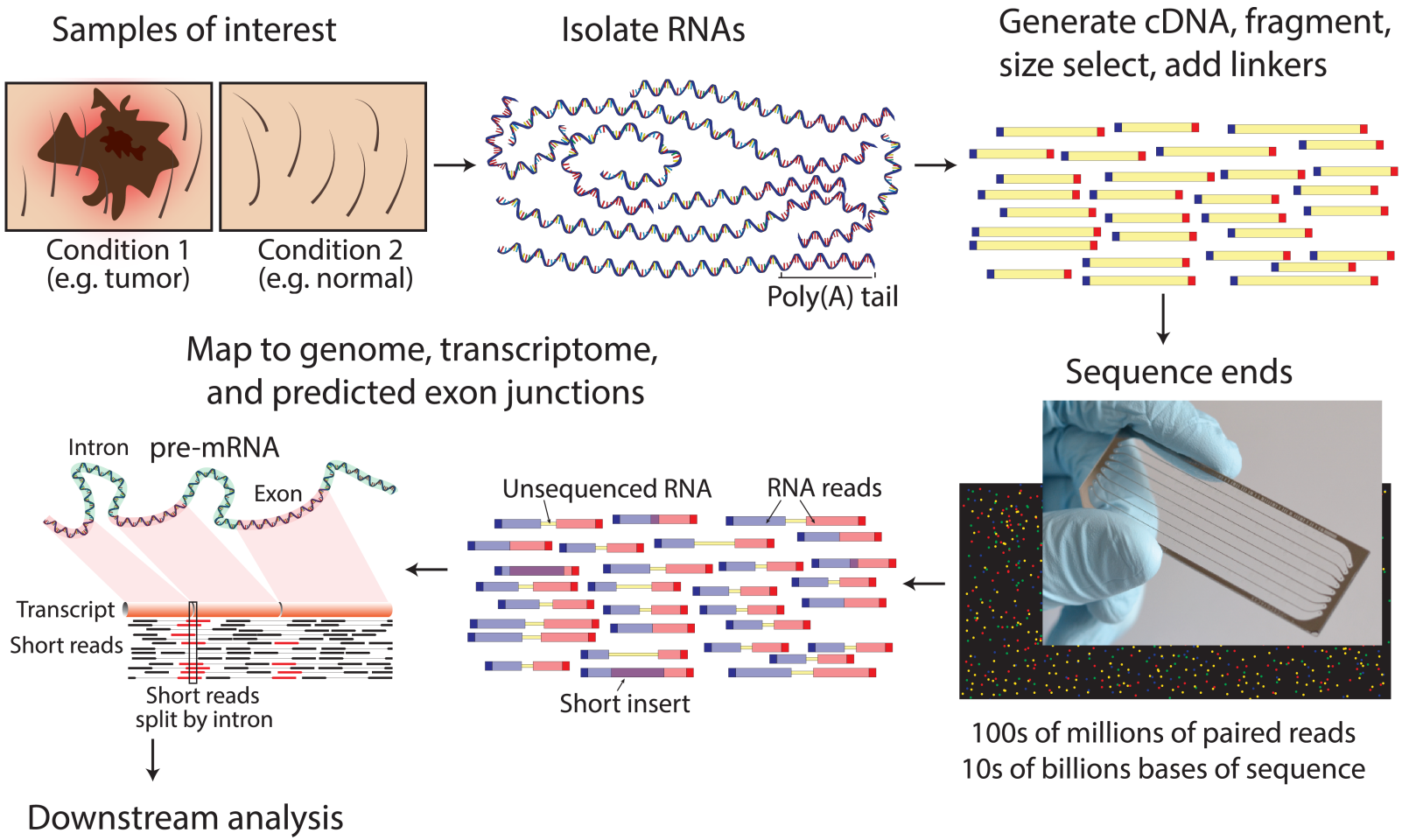
Sboner et al., Genome Biology 2011

RNA Sequencing (RNA-Seq)

RNA-Seq is able to identify thousands of differentially expressed genes, tens of thousands of differentially expressed gene isoforms and can detect mutations and germline variations for hundreds to thousands of expressed genetic variants, as well as detecting chimeric gene fusions, transcript isoforms and splice variants.

Wang, Nature 2009

RNA Sequencing (RNA-Seq)



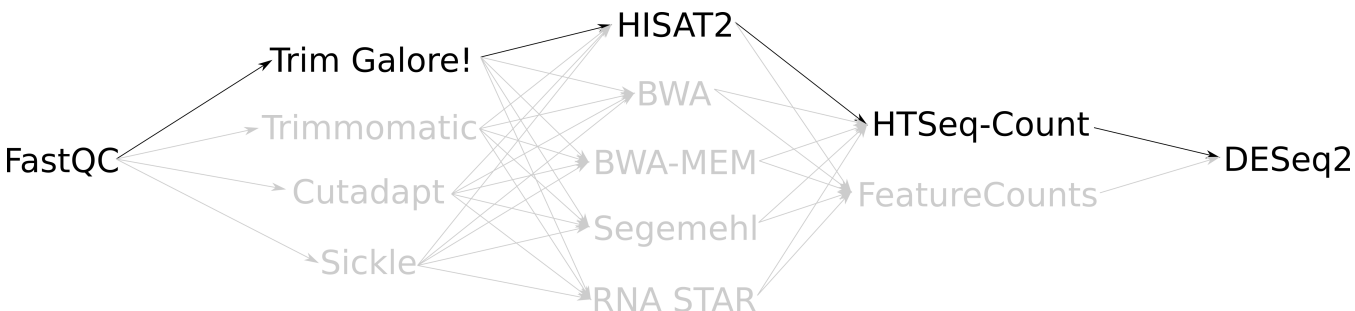
Griffith et al., PLOS 2015

RNA Sequencing (RNA-Seq)

No unique way of carrying out any data analysis workflow.

RNA-Seq workflow:

Tools:



Each tool has its own software dependencies!

Tools and materials

https://github.com/destairdenbi/trainings

https://usegalaxy.eu

https://galaxyproject.github.io/training-material/