RNA biology I

Genomics Bio5488

Guoyan Zhao

Associate Professor of Genetics & Neurology

Class Information

- Lectures: Mon, Wed 10:00-11:30 am
  - February 16: RNA biology I
  - February 18: RNA biology I I

- Lab: February 20, 2026, Fri 10:00-11:30 am (Bring your laptop)
  - Yuchen (Alana) Cheng, c.yuchenalana@wustl.edu
  - Benjamin Van Court, b.vancourt@wustl.edu

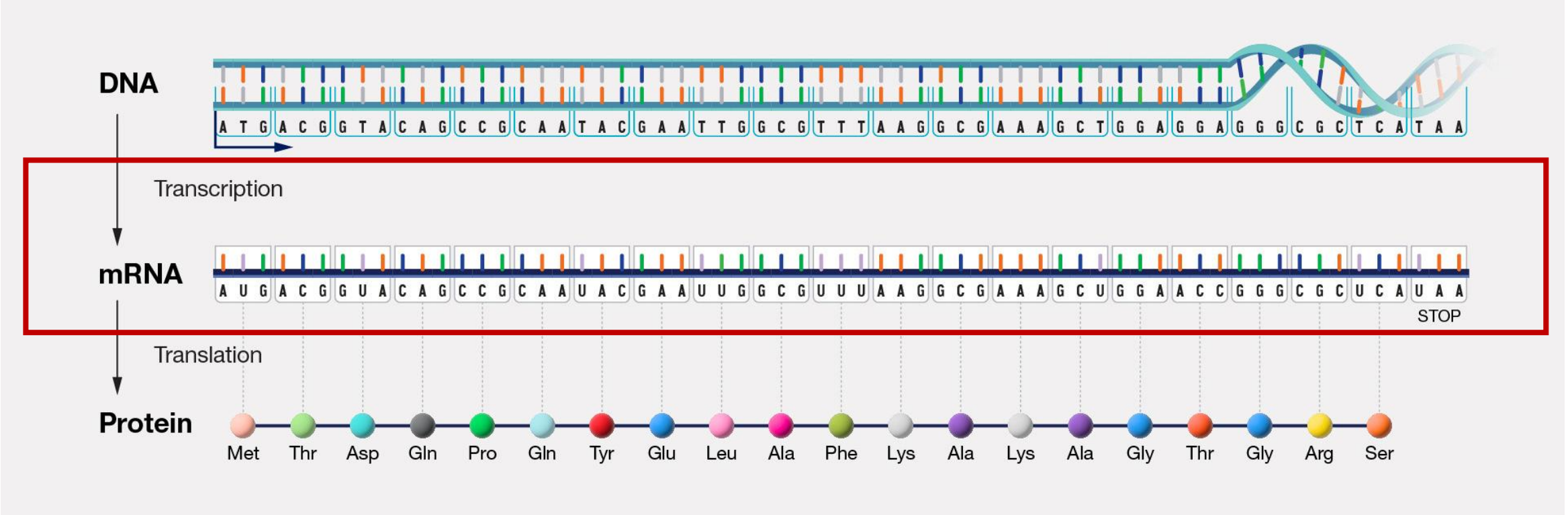
- Questions, suggestions and feedbacks: gzhao@wustl.edu

RNA biology

- RNA biology I
  - Introduction of RNA type and function
  - RNA quantification technology
    - Targeted
      - Northern blotting
      - RT-qPCR
      - *In situ* Hybridization
      - Reporter Assay
    - High-throughput
      - Microarray
      - RNA-seq
  - Experimental design principles
  - General workflow for RNA quantification
- RNA biology II
  - Transcriptome profiling by RNA-seq
    - Experimental design
    - RNA quantification
    - Quality control and normalization
    - Outlier detection
    - Differentially expressed gene detection
    - Result interpretation
- RNA biology Lab
  - RNA-seq data analysis

Goals: RNA biology, quantification, important concepts and considerations in experimental design and data analysis

Central Dogma



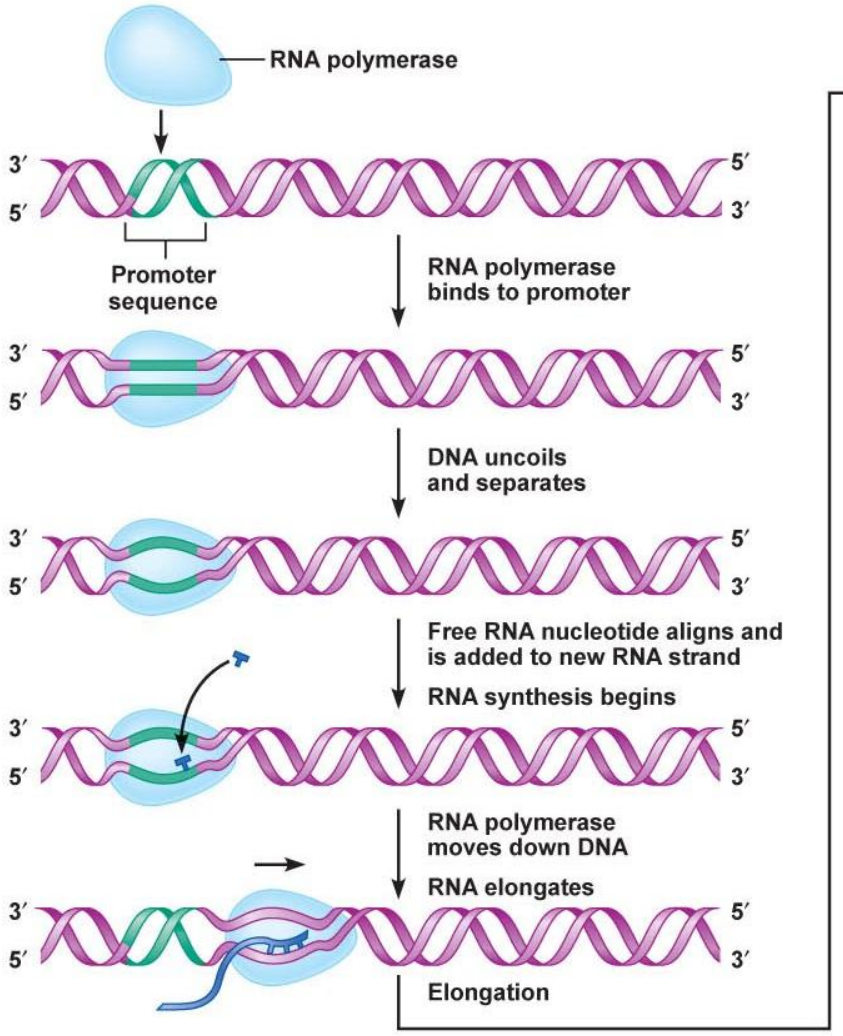
National Human Genome Research Institute

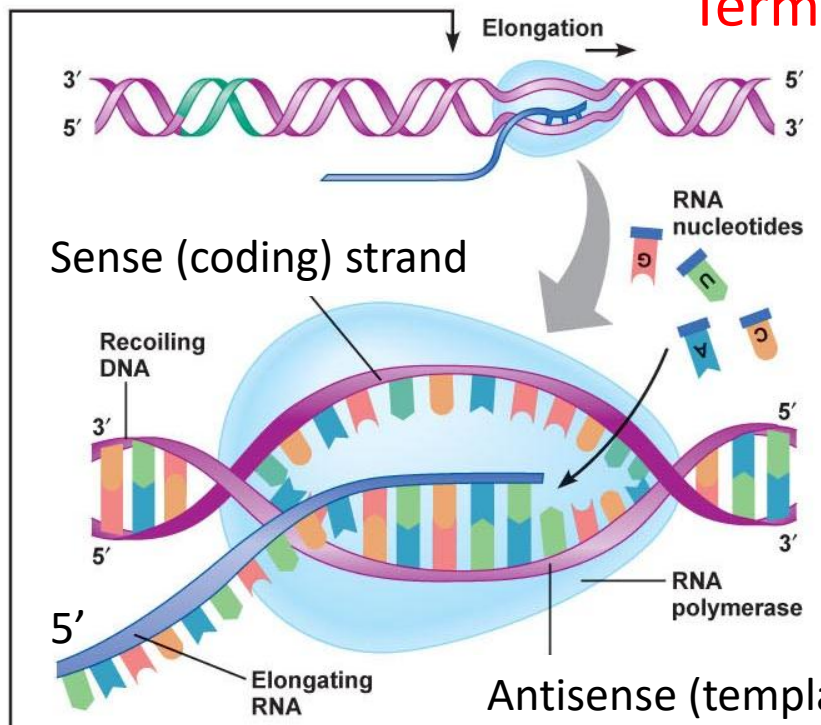
Transcription

Initiation

Elongation

Termination

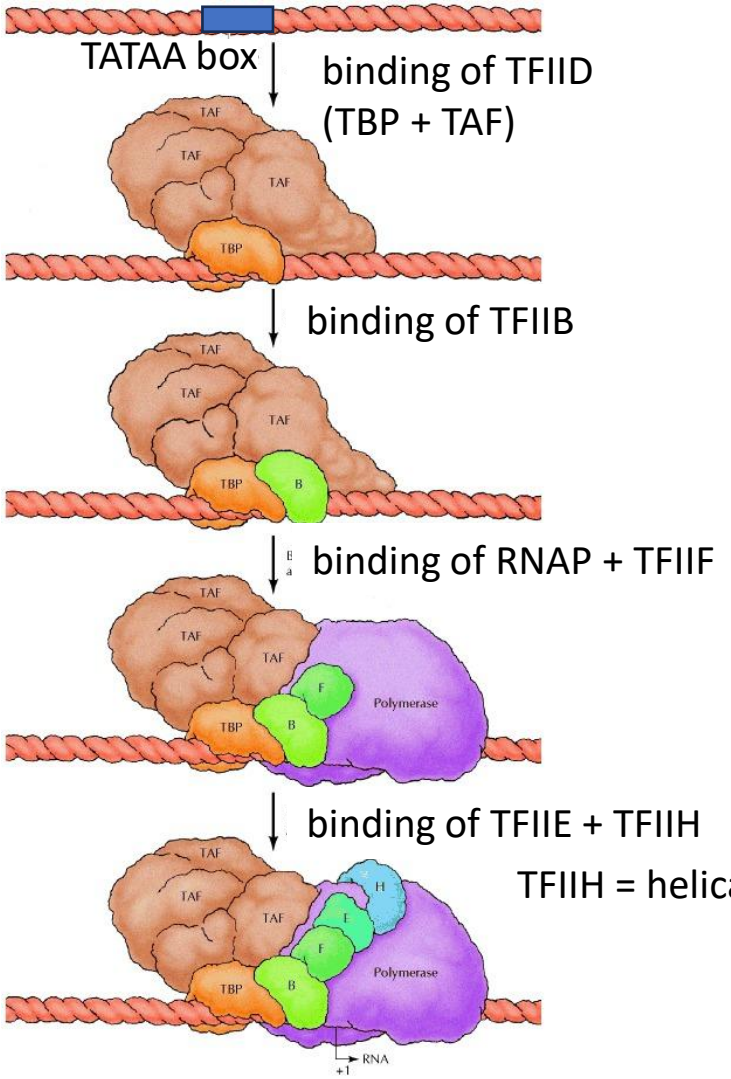




transcription bubble

Transcription Initiation

Formation of the initiation complex



RNA polymerase II is responsible for the synthesis of mRNA from protein-coding genes.

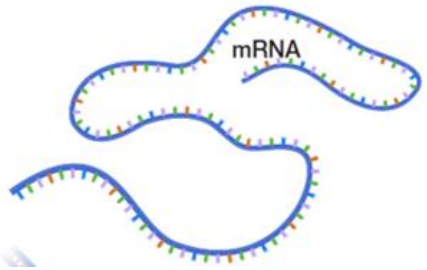
Two types of transcription factors:

- General transcription factors (basic transcription machinery): transcription from all polymerase II promoters
- Regulatory transcription factors

The Cell: A Molecular Approach. 2nd edition. Cooper GM. 2000. “Eukaryotic RNA Polymerases and General Transcription Factors”

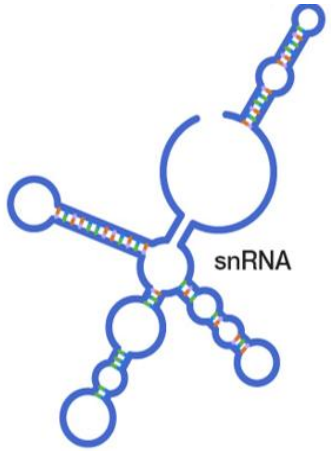
RNA types

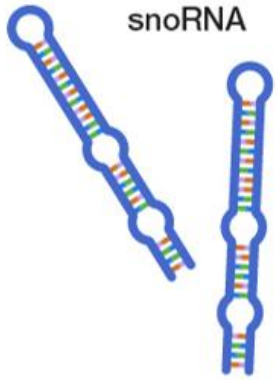
Coding RNA



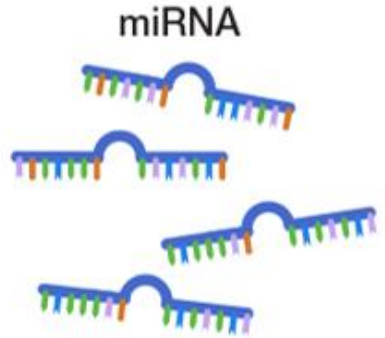
Non-coding RNA

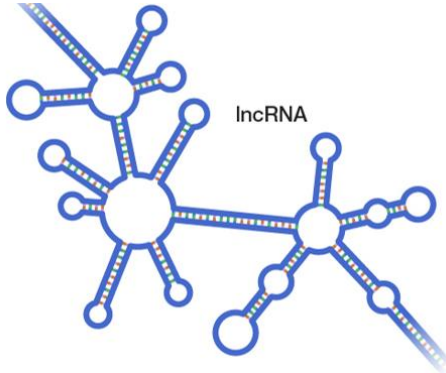
Infrastructural ncRNAs





Regulatory ncRNAs

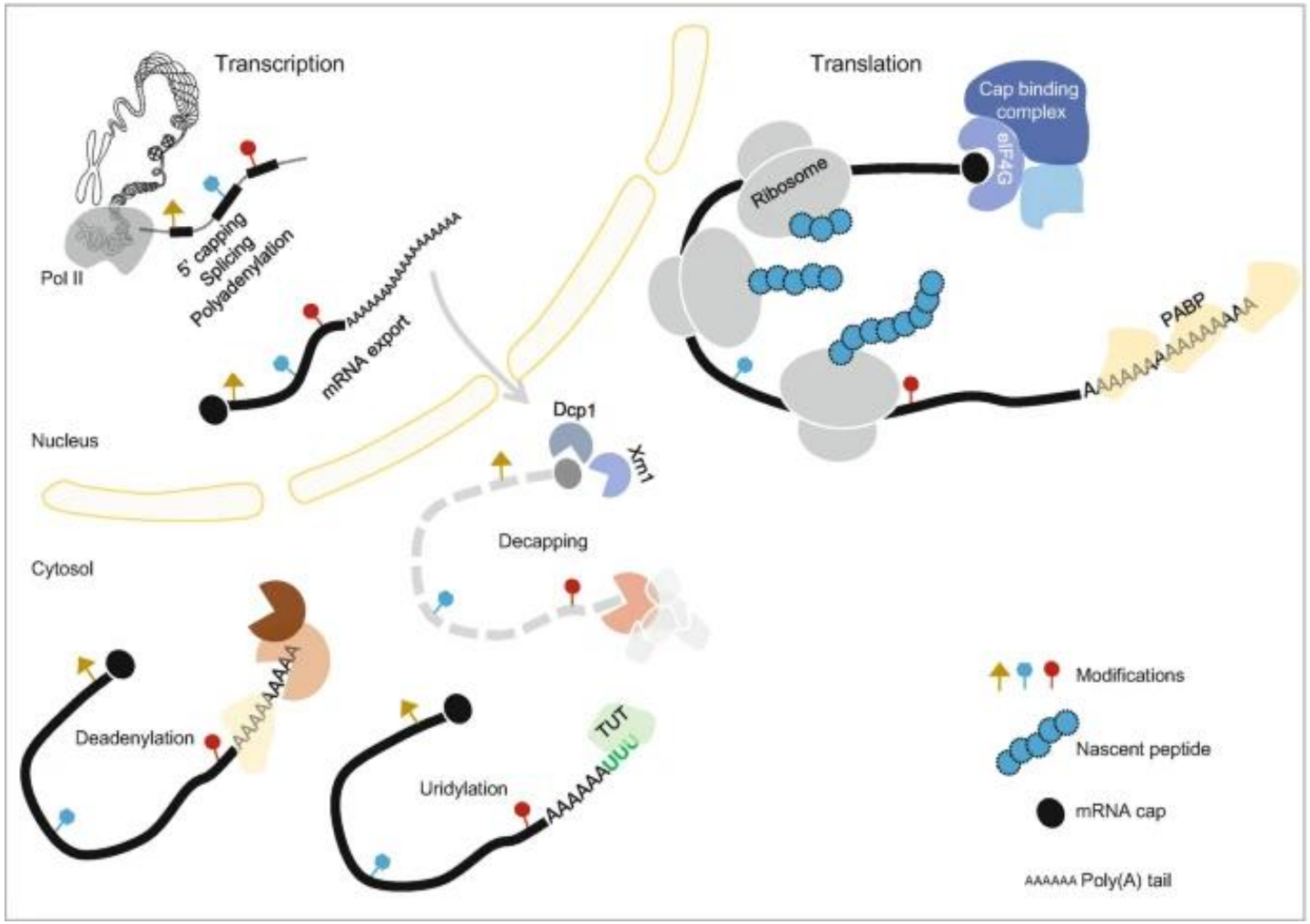




- small PIWI-interacting RNAs (piRNAs):
- promoter-associated RNAs (PARs)
- enhancer RNAs (eRNAs)

Coding RNA - mRNA

life cycle of an mRNA



mRNA modification

- in the coding sequence
  - alter translation fidelity
  - decoding
  - ribosome transit speeds and accuracy, often leading to pausing or stalling.
- in the untranslated regions
  - stability of the RNA structure
  - ability to form RNA–protein interactions
  - impact mRNA maturation, translation, and decay

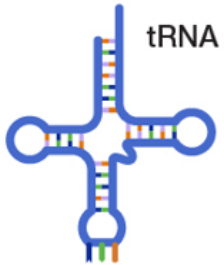
Effects of mRNA Modifications on Translation: An Overview, Bijoyita Roy, RNA modification, 2021

Non-coding RNAs (ncRNAs) - infrastructural ncRNAs

Non-coding RNAs (ncRNAs) are a heterogeneous group of transcripts that, by definition, are not translated into proteins

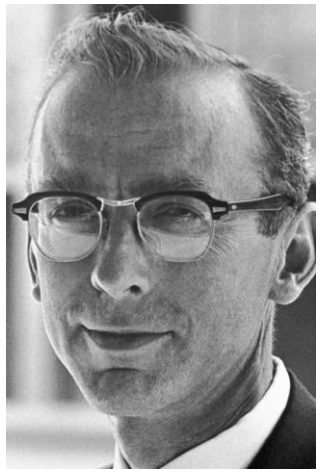
- Ribosomal RNA (rRNA):

1) structural elements in the assembly of ribosomes
2) transcribed from ribosomal DNA (rDNA). The rDNA sequence is arranged as tandem repeat units in the genome.
3) the most abundant form of RNA found in most cells (about 80% of cellular RNA).

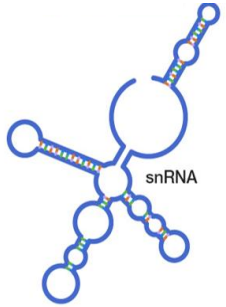


- Transfer RNA (tRNA):

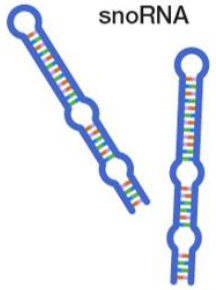
Robert W. Holley won the 1968 Nobel Prize in Physiology or Medicine for describing the structure and determining the nucleotide sequence of alanine transfer RNA, linking DNA and protein synthesis.



Non-coding RNAs (ncRNAs) - infrastructural ncRNAs cont.



- Small nuclear RNAs (snRNAs):
- critical components of the spliceosome that catalyze the splicing of pre-mRNA.



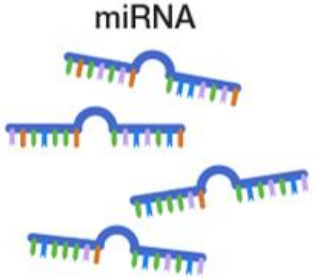
- Small nucleolar RNAs (snoRNAs):
- widely present in the nucleoli of eukaryotic cells
- important for RNA biogenesis and chemical modifications of rRNA, tRNA, and mRNA.

Spliceosomal snRNA Epitranscriptomics, Morais et al., 2021
snoRNAs: functions and mechanisms in biological processes, and roles in tumor pathophysiology, Huang et al., 2022

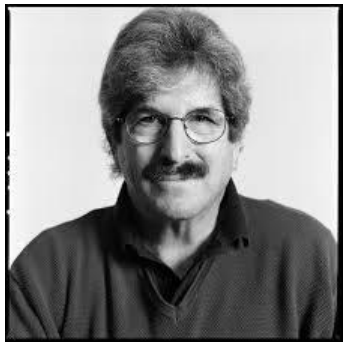
Non-coding RNAs (ncRNAs) – Regulatory ncRNAs

- small interfering RNAs (siRNAs):
- ~ 20 base pairs
- induces gene silencing by targeting complementary mRNA for degradation


Andrew Z. Fire and Craig C. Mello (2006)
siRNAs and RNA interference (RNAi)



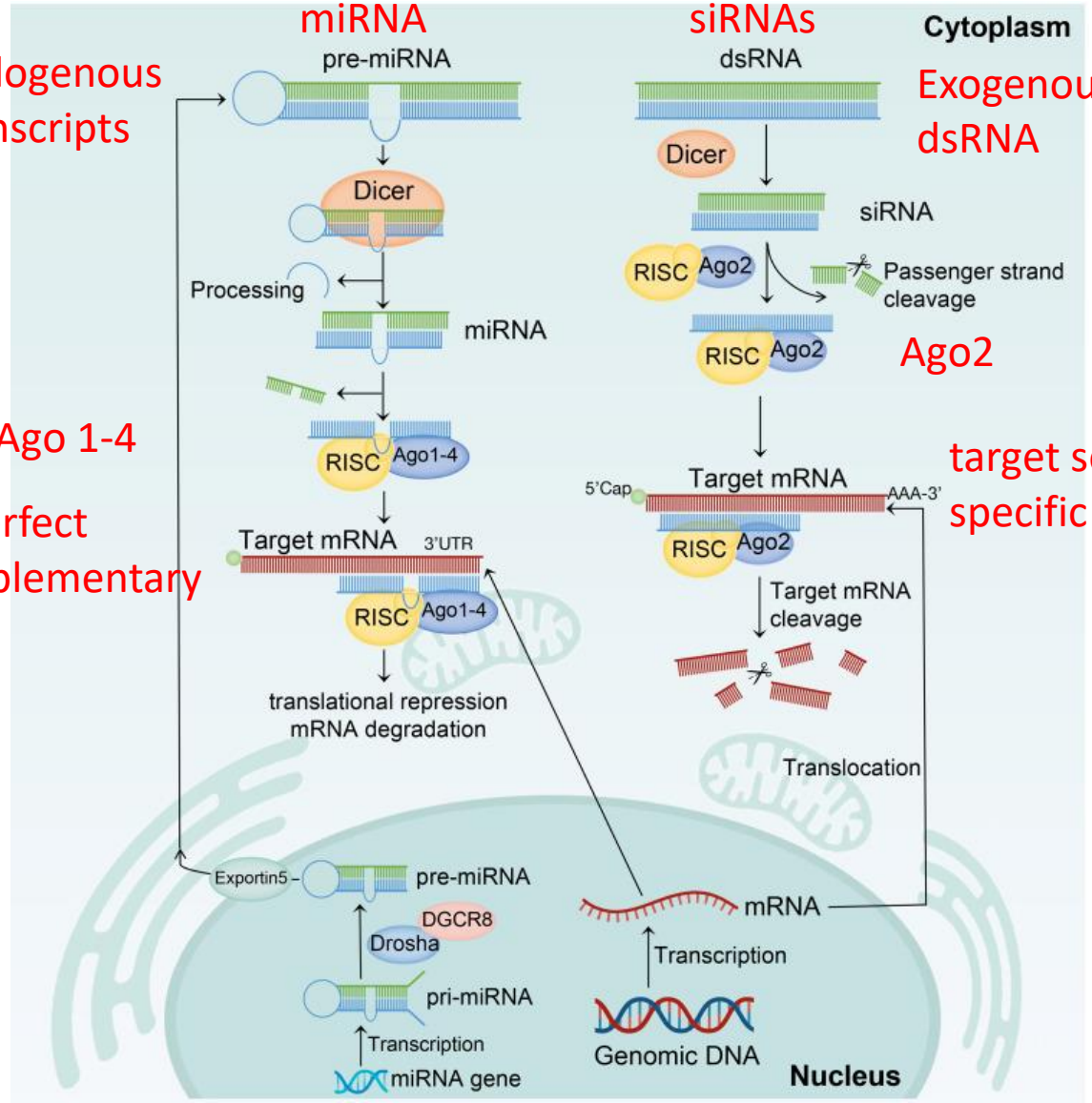
- microRNAs (miRNAs):
- ~22 nucleotides
- leads to mRNA degradation or inhibition of mRNA translation

Victor Ambros and Gary Ruvkun (2024)
miRNA and post-transcriptional gene regulation

Molecular Mechanisms and Biological Functions of siRNA, Dana et al., 2017
microRNAs in action: biogenesis, function and regulation, Shang et al., 2023

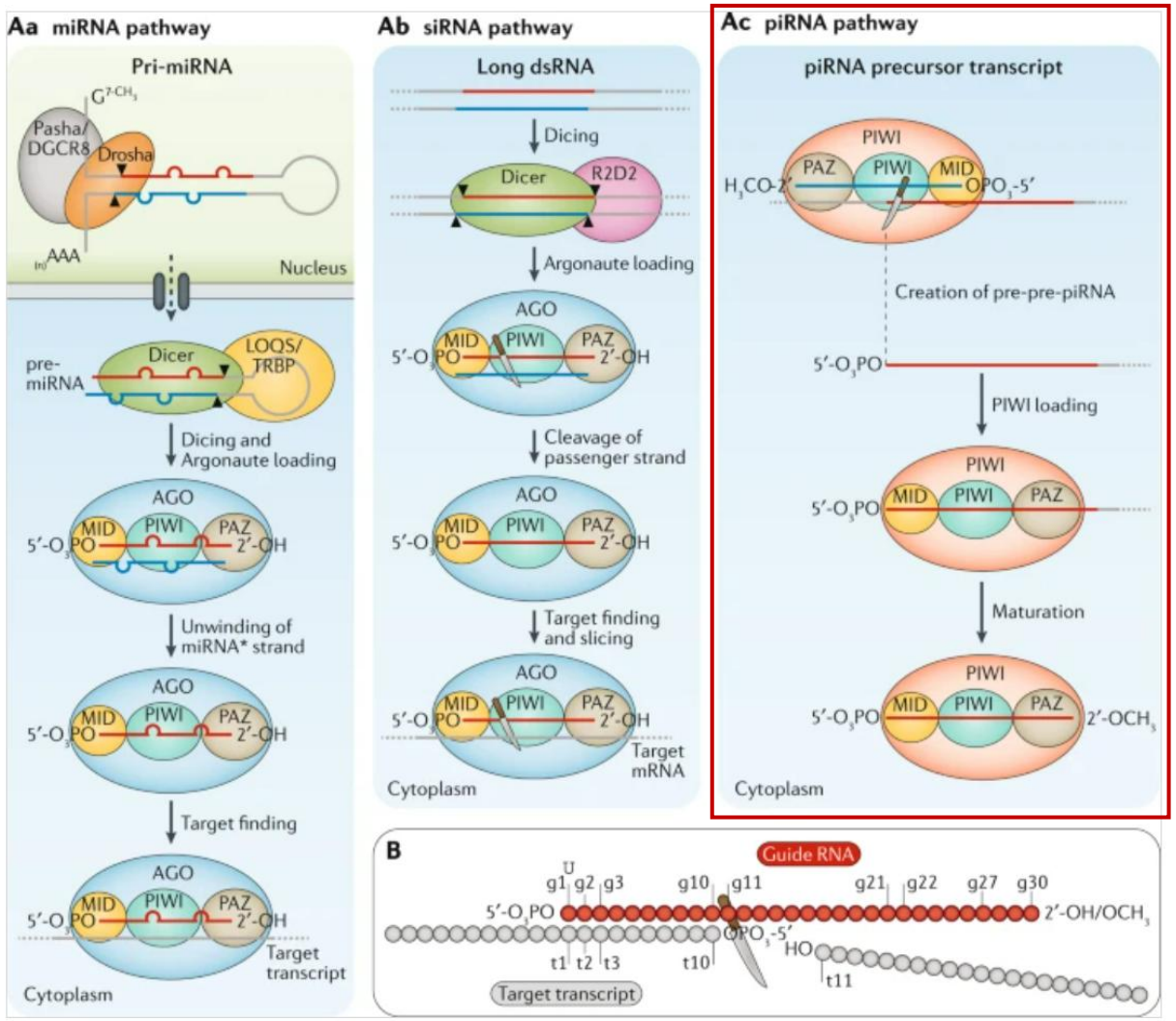
Mechanism comparison between siRNAs and miRNA



RISC:
RNA-induced silencing complex

Non-coding RNAs (ncRNAs) – Regulatory ncRNAs cont

Comparison of three RNA-silencing pathways
that all depend on the Argonaute family proteins



- small PIWI-interacting RNAs (piRNAs):
- 21/24/26–31 nucleotides
- piRISC: piRNA-induce silencing complex
- protect genome integrity

single- vs. double- stranded RNA precursors

Binding proteins: PIWI vs. AGO proteins

3’-end modification: 2’-O-methyl vs. 2’,3’ hydroxy

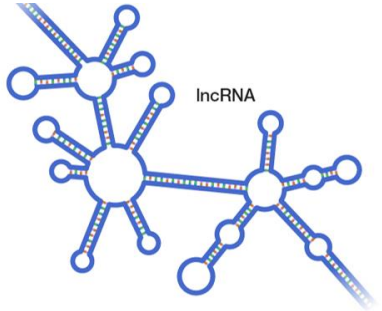
Different target genes:

- miRNAs: endogenous genes
- siRNA: exogenous genes
- piRNAs: transposon silencing

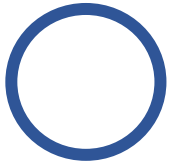
PIWI-Interacting RNAs (piRNAs): Promising Applications as Emerging Biomarkers for Digestive System Cancer, Cai et al, 2022

Non-coding RNAs (ncRNAs) – Regulatory ncRNAs cont

- Long non-coding RNAs (lncRNAs):
  - non-coding transcripts of more than 200 nucleotides
    - linear lncRNAs:



1. Most linear lncRNAs are transcribed by RNA polymerase II
2. Presumably are capped, polyadenylated, and contain exon-exon splice junctions like mRNAs
3. Based on their genomic locations relative to adjacent protein-coding genes, lncRNAs are classified as sense, antisense, bidirectional, intronic, and intergenic lncRNAs. They are also derived from 'pseudogenes',
4. Regulate many aspects of cell differentiation, development and other physiological processes.



- circular RNAs (circRNAs):

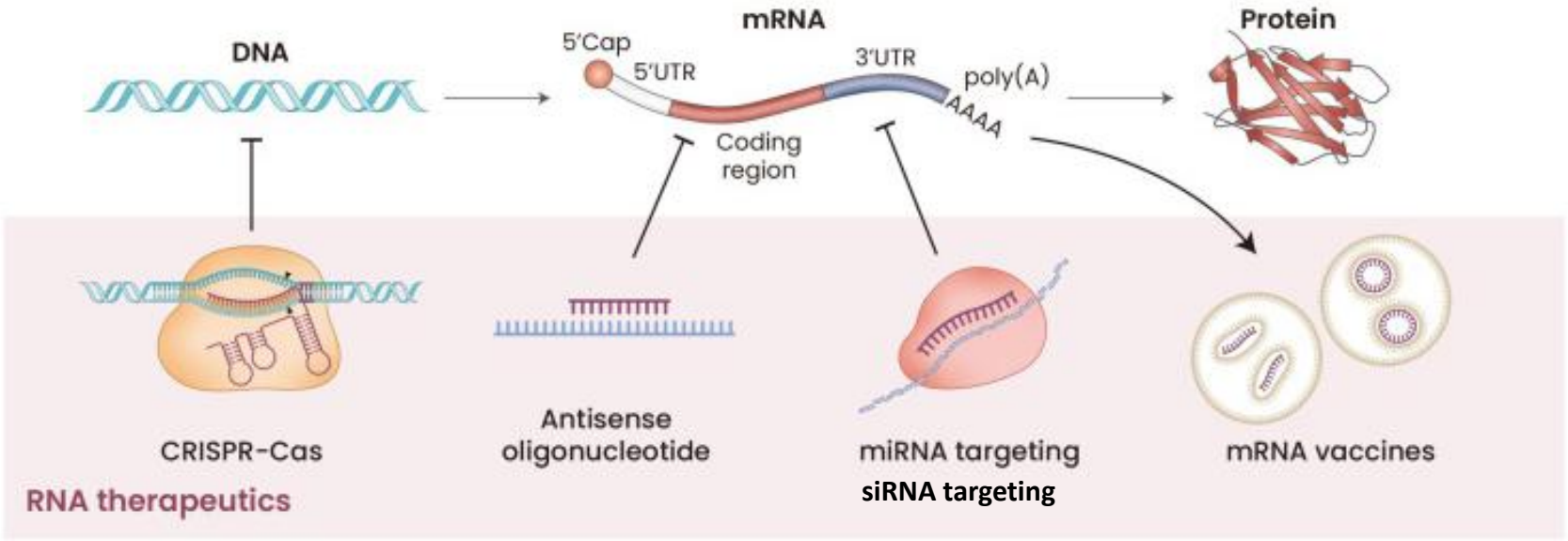
- generated by back-splicing of linear transcripts during the mRNAs splicing
- can be derived from exons, introns, exon–intron junctions or intergenic regions of the genome.
- tissue- and cell-specific expression patterns, abundant
- more resistant to degradation, have a longer half-life than mRNAs and linear lncRNAs
- carry out diverse regulatory functions

Long non-coding RNAs: definitions, functions, challenges and recommendations, Mattick et al., 2023
Structure, Regulation, and Function of Linear and Circular Long Non-Coding RNAs, Qin et al., Front Genet. 2020
Circular RNAs: the next level of gene regulation , Garrido and Flores , Am J Transl Res. 2023

Different Classes of Genes are Transcribed by Different Eukaryotic RNA Polymerases

Type of RNA synthesized		RNA polymerase
mRNA		II
tRNA		III
rRNA	5.8S, 18S, 28S	I
	5S	III
Some small RNA	e.g. snRNA	II or III
Mitochondrial genes		Mitochondrial RNA polymerases
Chloroplast genes		Chloroplast RNA polymerases

RNA in Therapeutics



Katalin Karikó and Drew Weissman was awarded the 2023 Nobel Prize for their discoveries concerning nucleoside base modifications that enabled the development of effective mRNA vaccines against COVID-19.

RNA in Biology and Therapeutics, Sunjoo Jeong, Mol Cells. 2023
Non-Coding RNA-Targeted Therapy: A State-of-the-Art Review, Nappi et al., 2024

Why Do We Care About Gene Expression?

- Gene expression profiles provide a snapshot of cellular/ tissue state at the molecular scale

- Gene expression profiles provide a snapshot of cumulative interactions of many regulatory relationships

- Gene expression is a ‘proxy’ measure for transcription/ translation/ functional events

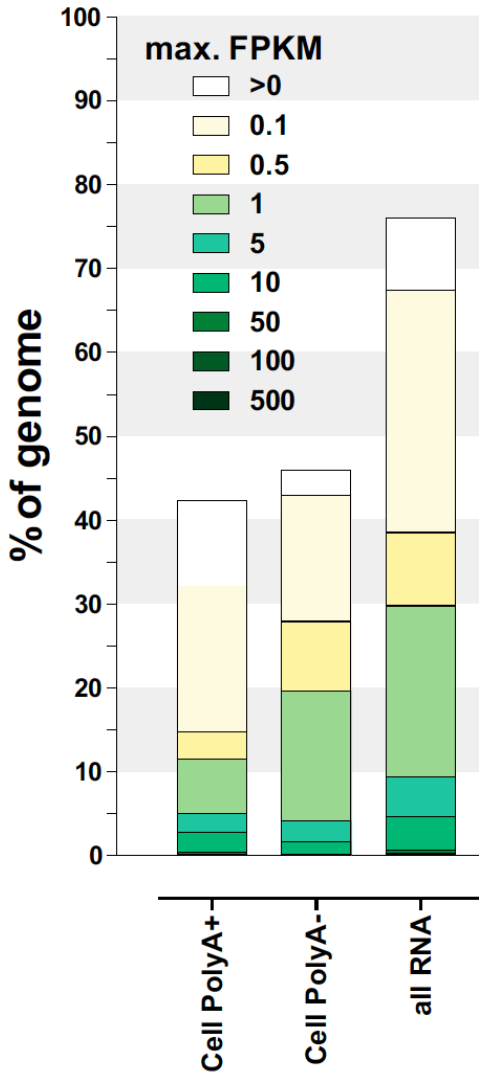
What can we learn by measuring gene expression?

- Cell function and behavior
- Developmental biology
- Environmental impacts
- Disease mechanisms
- Personalized medicine

The assumptions behind the conclusions

- Assume that gene expression levels correspond to functional protein levels
- Assume that a normal cell has a standard expression profile/ signature
- Assume that changes in expression profile indicate that some property or functional changes

How Much of the Genome is Transcribed?



Estimates from ENCODE

The ENCODE (ENCyclopedia Of DNA Elements) Project Consortium. Science. 2004

- Recent high-throughput transcriptomic analyses have revealed that eukaryotic genomes transcribe up to 90 % of the genomic DNA.
- mRNA: Only 1–2% of these transcripts encode for proteins
- the vast majority are transcribed as non-coding RNAs (ncRNAs).

Kellis *et al. Proc. Natl. Acad. Sci. USA* 111:6131 (2014)

Does mRNA Level == Protein Level?

- Yeast: significant discrepancy between mRNA and protein levels
  - 156 genes in yeast (Gygi et al., 1999 Mol Cell Biol)
  - 245 genes in yeast on galactose/ ethanol medium (Griffen et al., 2002 Mol Cell Proteomics)

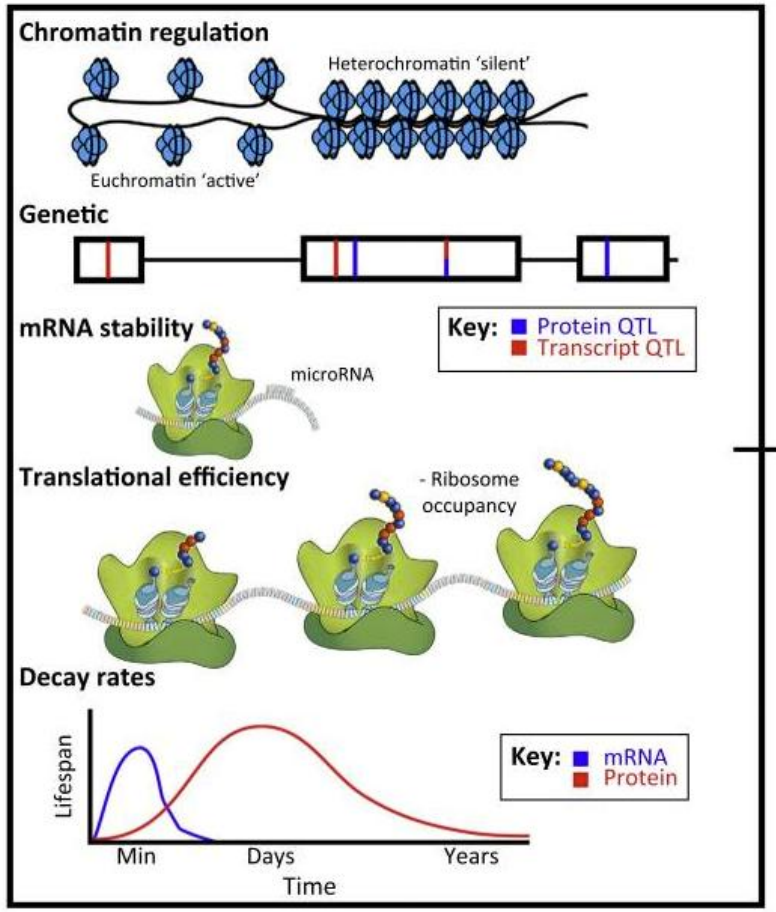
- mammalian cells: significant discrepancy between mRNA and protein levels
  - The mean correlations were ~0.25. Only 1/3 had significant correlations (> 0.445) between protein and RNA levels. (Gry et al., BMC Genomics, 2009)
  - 55 target proteins (Edfors et al., Mol Syst Biol. 2016)
  - ~30–40% of the variance in protein abundance is explained by mRNA abundance (Vogel & Marcotte, Nature Reviews Genetics, 2012

- Differentially expressed mRNAs correlate significantly better with their protein product than non-differentially expressed mRNAs. (Koussounadis et al. Scientific , 2015)

Regulation of protein abundance

Genome

Types of regulation



- Copy number variation
- Promoters, enhancers, silencers, insulators
- histone and DNA modifications
- 3D genome structures

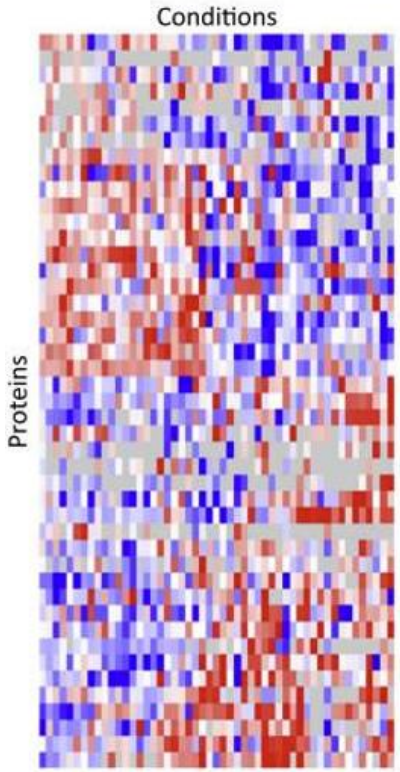
the loci controlling RNA expression (eQTLs) had only a 50% overlap with the loci controlling protein expression (pQTLs)

mRNA stability, modification

miRNAs
siRNA
piRNA
lncRNA

abundance of a transcript ≠ the use of a transcript

Protein abundance



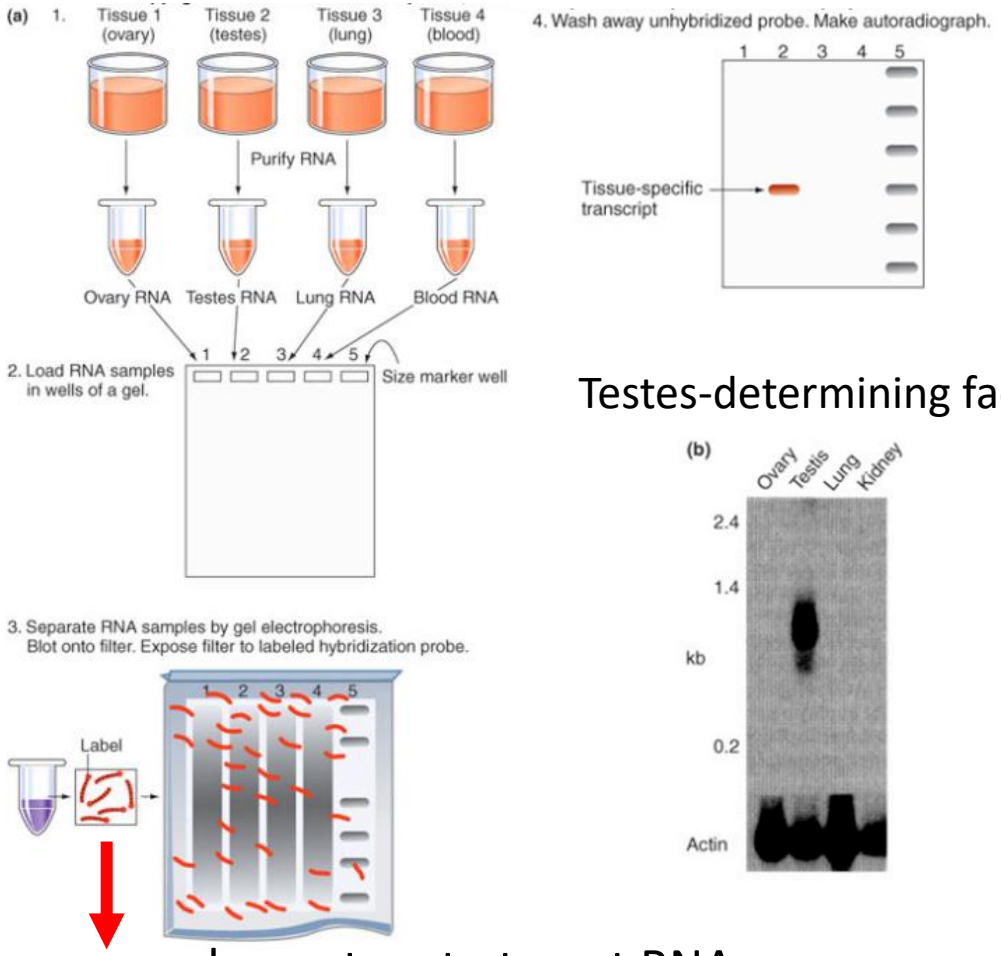
The utility of protein and mRNA correlation, Payne et al., Trends Biochem Sci , 2016

RNA quantification method

- Targeted
    1. Northern blotting
    2. *In situ* Hybridization
    3. Reverse transcription quantitative PCR (RT-qPCR)
    4. Reporter Assay
- High-throughput
    1. Microarray
    2. RNA-seq
    3. Single cell/nucleus RNA-seq

Mechanisms and Measurement of Changes in Gene Expression,
Singh et al., Biol Res Nurs. 2018

Measuring Gene Expression: 1. Northern blot



- To determine the size and quantity of specific RNA molecules among a mixture of RNA.

complementary to target RNA sequence
single strand
labeled with radioactive isotope or fluorescent dye

Northern Blotting Analysis, Josefsen & Nie

Northern blot pros & cons

Advantages

- the simplicity of the procedure and low cost
- Very sensitive due to use of radioactive probes
- Nearly infinite dynamic range

Disadvantages

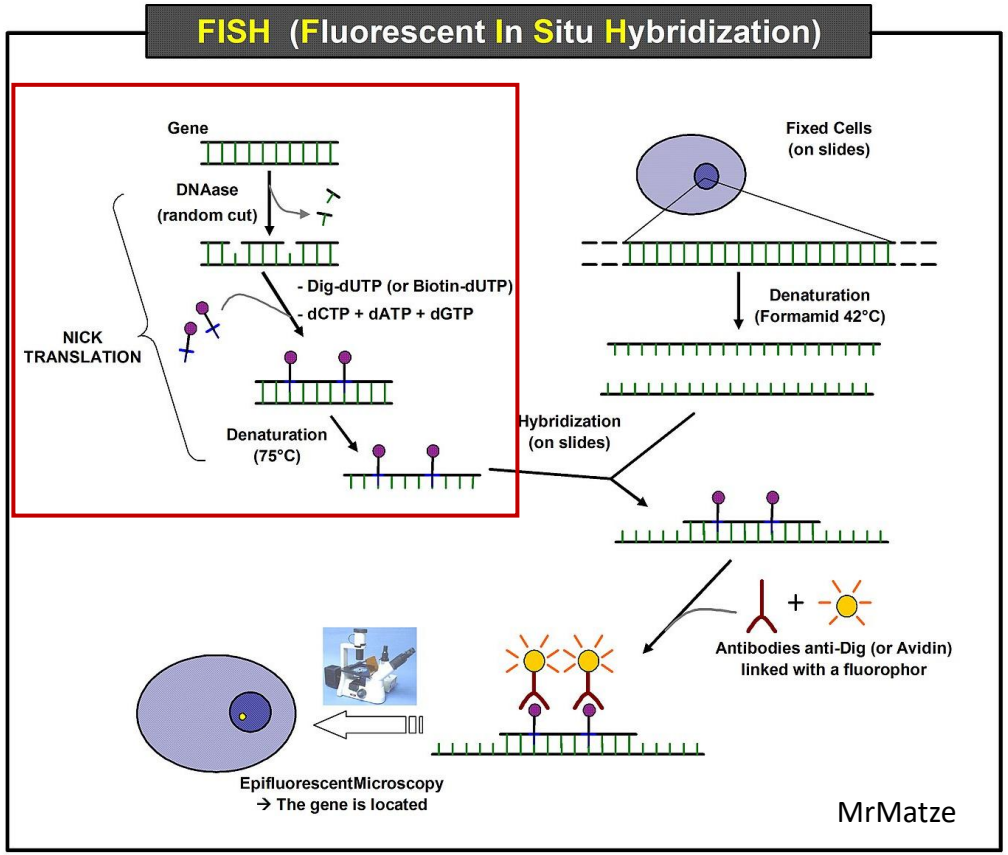
- time-consuming
- only a small number of samples can be analyzed at one time
- requires a large amount of starting material
- Quality control (non-specific hybridization)

Measuring Gene Expression: 2. *In situ* Hybridization

- To localize and detect DNA or RNA sequences in morphologically preserved cells, tissue sections, and even whole tissue. (*in situ* = in the original place)

Probe labeling	Visualization
radio-labeled bases	autoradiography
fluorescent-labeled bases	fluorescence microscopy
antigen-labeled bases	immunohistochemistry

zebrafish lyve1



Technical Review: In Situ Hybridization, THE ANATOMICAL RECORD (2014)

In situ Hybridization pros & cons

Advantages

- provide spatial information of cellular content
- Single-cell sensitivity
- Spatial and temporal analysis
- Can be used on archival tissues

Disadvantages

- Expensive
- Time consuming
- require experienced personnel
- probe- and sample-specific, have to be optimized for each set of conditions empirically.

Measuring Gene Expression: 3. RT-qPCR

Reverse transcription quantitative PCR (RT-qPCR)

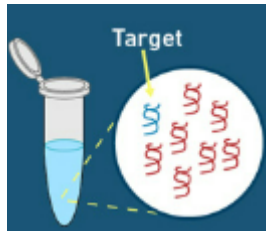
- the gold standard for accurate, sensitive and fast method of nucleic acid detection and quantification

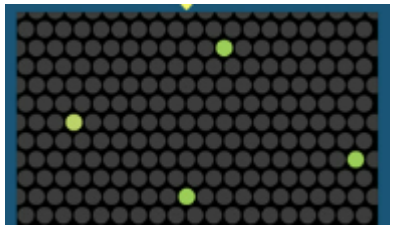
- relies on fluorescence to detect and quantify nucleic acid amplification products

- broad applications

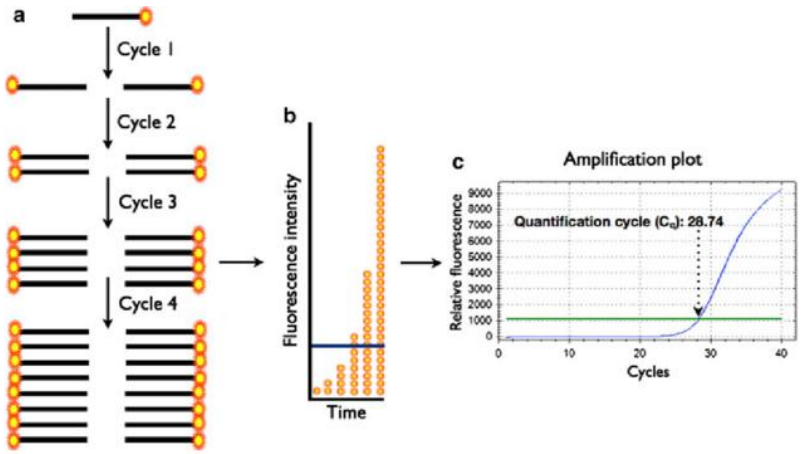
1. Convert total RNA to cDNA

2. Add cDNA to RT-qPCR master mix and aliquot mixture across PCR array



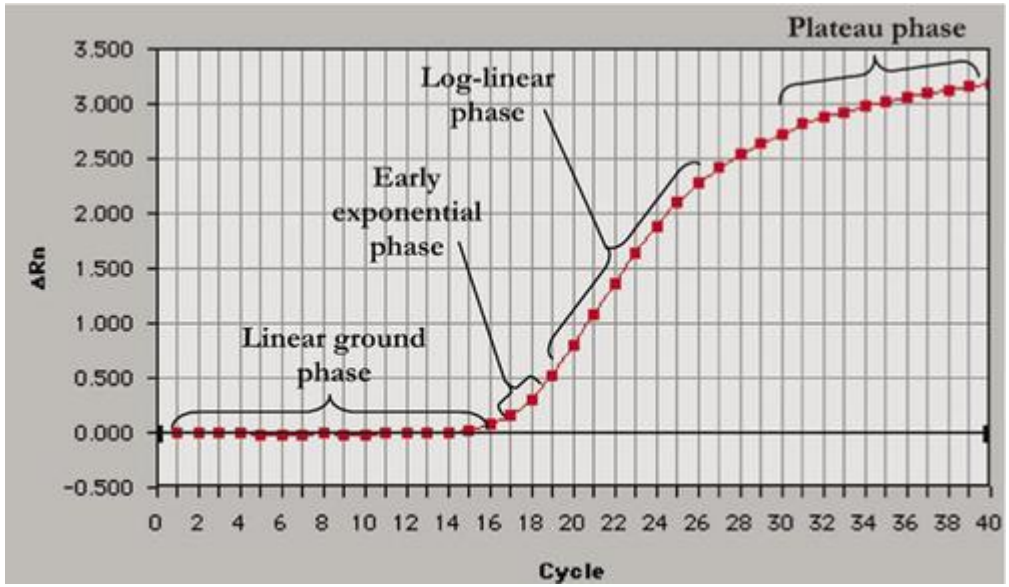


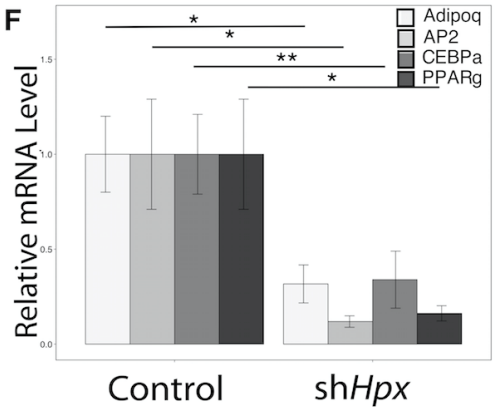
3. Run in your RT-qPCR instrument



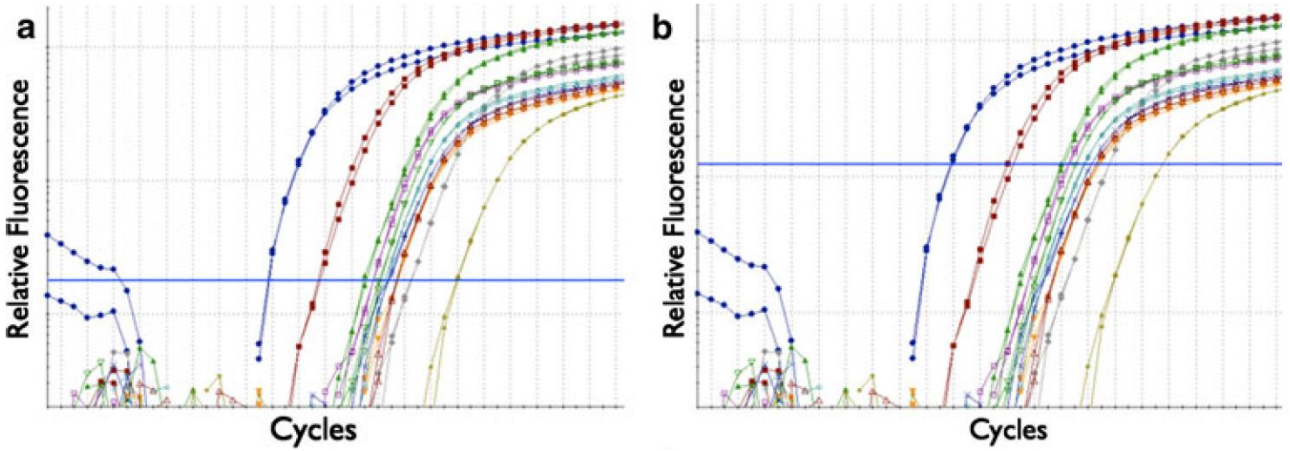
Good vs. bad RT-qPCR design

Phases of the PCR amplification curve

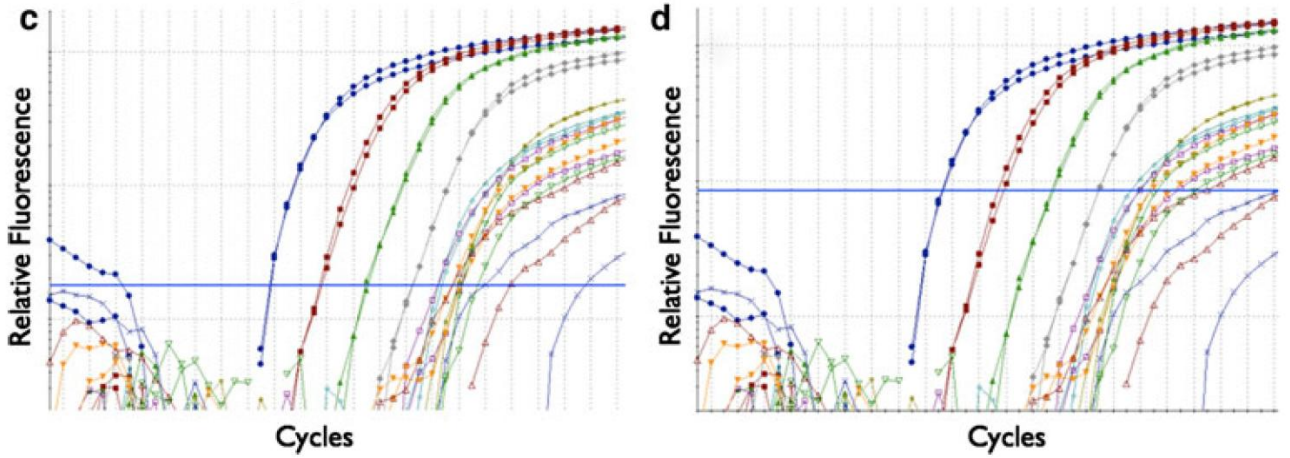




well-designed experiment



poorly-designed experiment



RT-qPCR pros & cons

Advantages

- Highly sensitive, quantitative and reproducible
- ‘Gold-standard’
- Excellent dynamic range
- Fast results

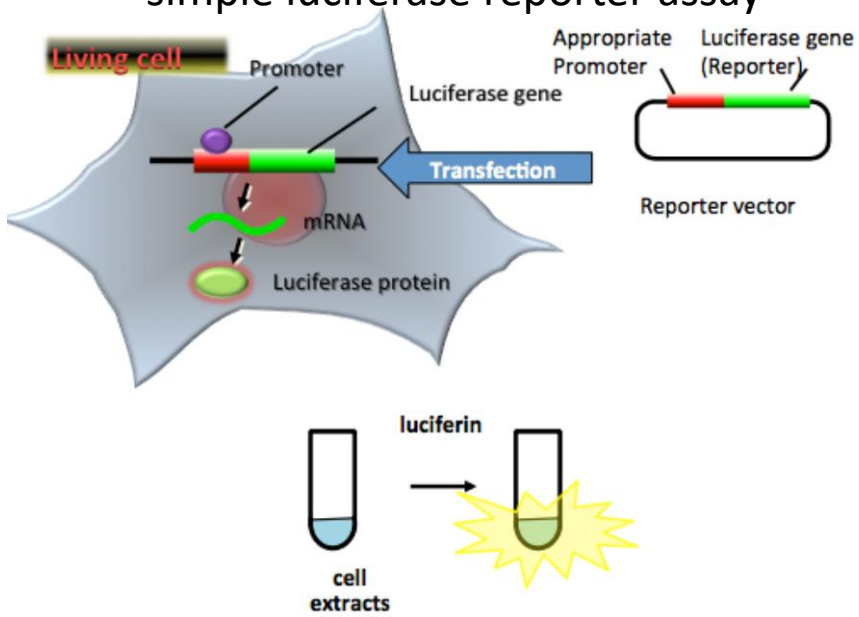
Disadvantages

- Expensive
- Not high-throughput
- Non-specific amplification can lead to false positives
- Always have controls – positive and negative controls

How to do successful gene expression analysis using real-time PCR, Derveaux et al., 2009

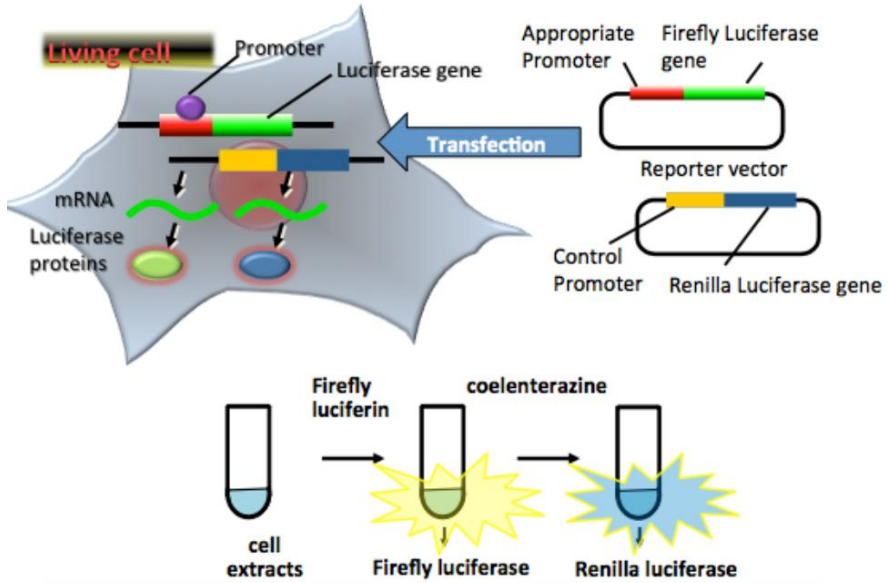
Measuring Gene Expression: 4. Reporter Gene Assay

simple luciferase reporter assay



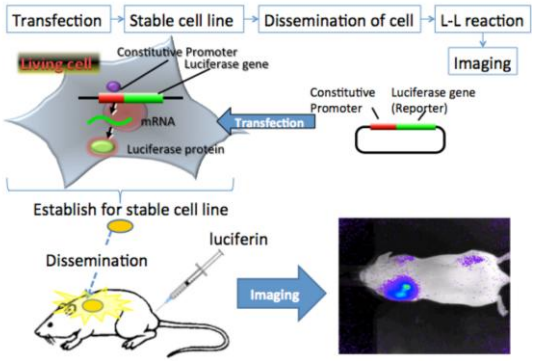
Promoter activity = Luciferase activity / Cell number or cellular enzyme activity

dual nonsecreted luciferase reporter assay



Promoter activity = Firefly luciferase activity / *Renilla* luciferase activity

In vivo bioluminescence imaging



Application of Reporter Gene Assay

- Application:
  - Determine promoter or enhance strength
  - Interactions between promoters and transcription factors
  - Protein-protein interactions
  - Protein trafficking
  - Signal transduction
  - Drug screening both in vitro and in vivo

- Commonly used reporter genes:
  - luciferases, green fluorescent protein (GFP), β-galactosidase, chloramphenicol acetyltransferase, β-glucuronidase

Reporter Assay pos & cons

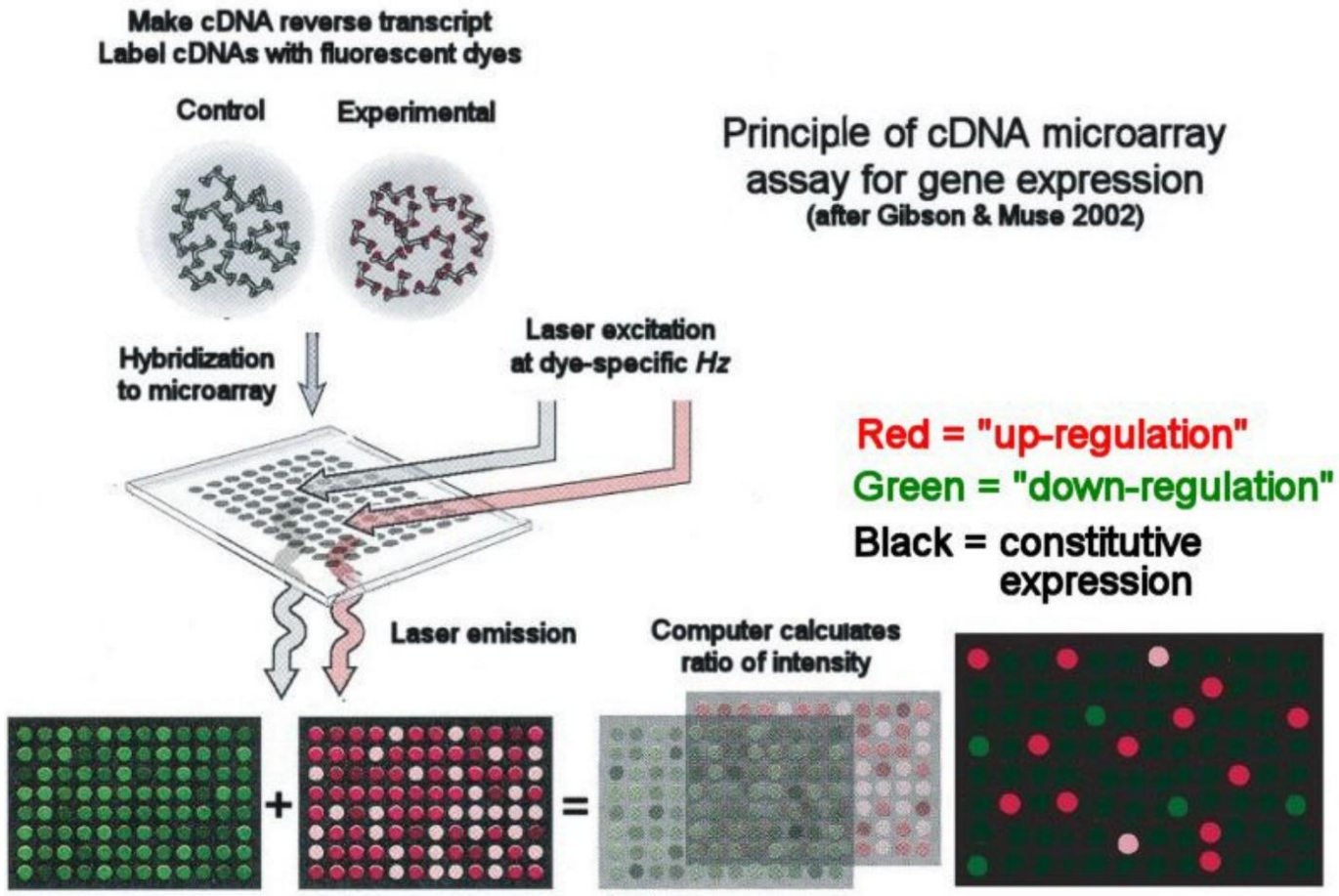
Advantages

- *In vivo* applications
- Highly sensitive
- New technology enables longitudinal studies

Disadvantages

- Stability issues
- Not high-throughput
- Quantification *in vivo* is affected by many variables

High-throughput technology: 1. Microarray



- Principle: hybridization between two DNA strands
- Quantifies RNA through template hybridization and dye intensity

Microarray pros & cons

Advantages

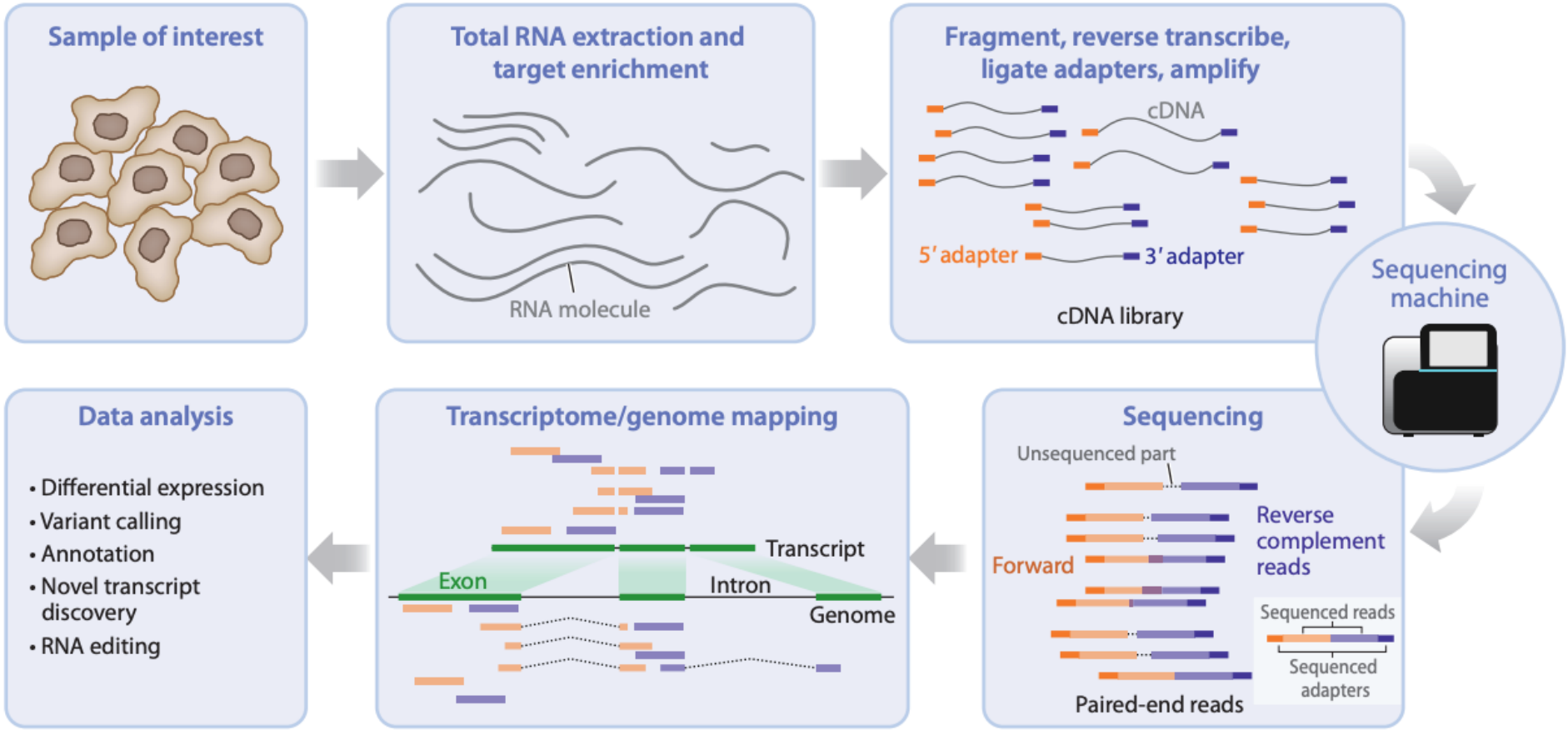
- High-throughput
- Reliable and more cost effective than RNA-Seq for gene expression profiling in model organisms
- Kit systems (easy/ accessible)

Disadvantages

- Need target transcripts information
- Quality and quality control highly variable
- Cross-hybridization
- Multiple tissue samples cannot be tested in one assay; a control and a test tissue sample need to be prepared separately

High-throughput technology: 2. RNA-seq

- RNA sequencing (RNA-seq) is a genomic technique that uses next-generation sequencing to analyze the quantity and presence of RNA molecules in a biological sample.



RNA-seq pros & cons

Advantages

- Transcript identification and quantification in a single assay
- Very direct and quantitative
- No prior knowledge of genome required
- A greater dynamic range to quantify transcripts allows more differentially expressed gene detection
- single-nucleotide resolution allows the detection of genetic variants, transcript isoforms and splice variants

Disadvantages

- Amplification steps can offset balance between high/ low abundance transcripts
- High cost than microarray
- Analysis is non-trivial – no optimal pipeline for the variety of different applications and analysis scenarios

Experimental Design and General Workflow

"Seventy percent of whether your experiment will work is determined before you touch the first test tube" - Sun, *Nature Reviews Mol Cell Biol* (2004)

How much risk can I afford to take?

↓

How much can I trust various components in an experiment?

↓

How thoroughly should I plan my experiment?

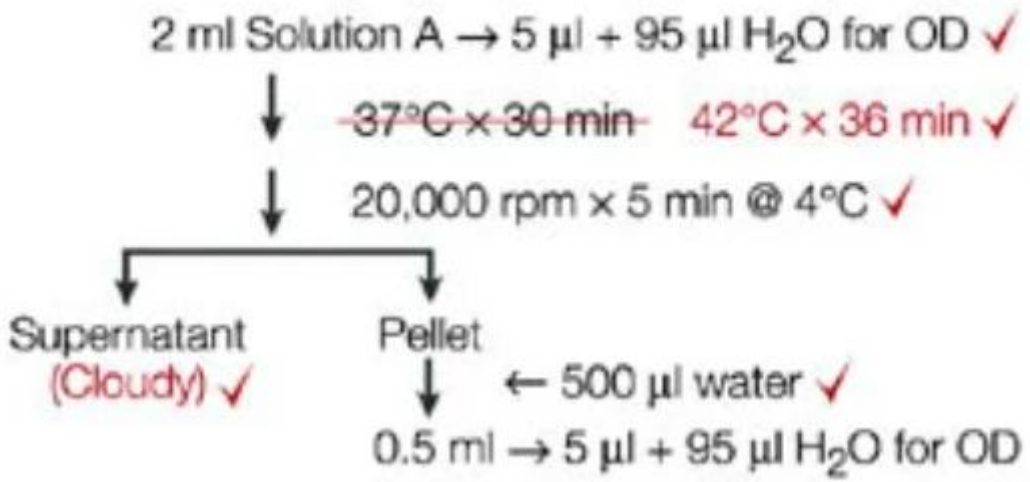
↓

Carrying out the experiment

Excessive trust in authorities and its influence on experimental design, Sun, *Nature Reviews Mol Cell Biol* (2004)

Maximizing experimental success

- To pay attention to even minute details, to the point that you become a perfectionist, will save you invaluable time in the long run.
- ***Thorough planning and understanding of the protocol.***
  - write a step-by-step flowchart that details the entire procedure from the beginning to the end thoroughly understand every step
  - compare the flowchart you generate with the protocols from several sources to see whether they are consistent.
  - check your detailed design of a particularly crucial or time-consuming experiment with your mentor or experienced colleague
- ***Importance of positive and negative co***
- ***A well-kept notebook.***
  - date, the hypotheses and/or questions, key protocols (in the form of flowcharts), reage



Gene Expression Experimental Design

Which technology should I use? – what's your research question?

- **Northern blotting:**
  - Best for determining the size of a specific RNA transcript.
- **RT-PCR:**
  - The most sensitive method for detecting and quantification of gene expression
  - Ideal for analyzing a few genes with high accuracy
- **In situ hybridization:**
  - visualize the location of a specific RNA within a tissue or cell, spatial information
- **Reporter Assay:**
  - Used to study the regulatory elements of a gene by monitoring the activity of a reporter gene under different conditions.
- **Microarray:**
  - Allows for simultaneous analysis of thousands of genes, useful for exploring global gene expression patterns of model organisms.
- **RNA-seq:**
  - Provides a comprehensive view of the transcriptome, including novel transcripts and isoforms, offering the most detailed information about gene expression

Comparison of gene expression between conditions

Hypothesis Statement:
Formulate your research question into two competing hypotheses:

Case Study: Detecting disease-associated APOE expression changes in Alzheimer's disease (AD) brain tissues.

Scientific Question: Do APOE expression differ between brains from individuals with AD and cognitively normal controls?

Null hypothesis (H_o): Mean expression of APOE is the same in AD and control brains

$$\mu_{AD} = \mu_{Control}$$

Alternative hypothesis (H_a): APOE expression differs (or is higher) in AD

two-tailed test: $\mu_{AD} \neq \mu_{Control}$

one-tailed test: $\mu_{AD} > \mu_{Control}$

Hypothesis Testing

State Hypothesis

H_0: Mean expression of APOE is the same in AD and control brains: $\mu_{AD} = \mu_{Control}$
H_a: APOE expression differs (or is higher) in AD: $\mu_{AD} \neq \mu_{Control}$

Set significance level (α)

It is important to set α before the experiment (a priori). typical value of $\alpha = 0.05$

Collect and prepare data

Ensure the data is representative of the population.
Selecting an appropriate sampling method.
Determining the sample size (power calculation).

Choose appropriate statistical test

Data type (continuous, categorical, etc.), Distribution of the data (normal, non-normal), Sample size, Number of groups being compared: t-tests, chi-square tests, ANOVA

Calculate test statistic and p-value

quantifies how much the sample data deviates from the null hypothesis.
p-value: probability of observing results as extreme as the sample data, assuming the null hypothesis is true.

Make a decision

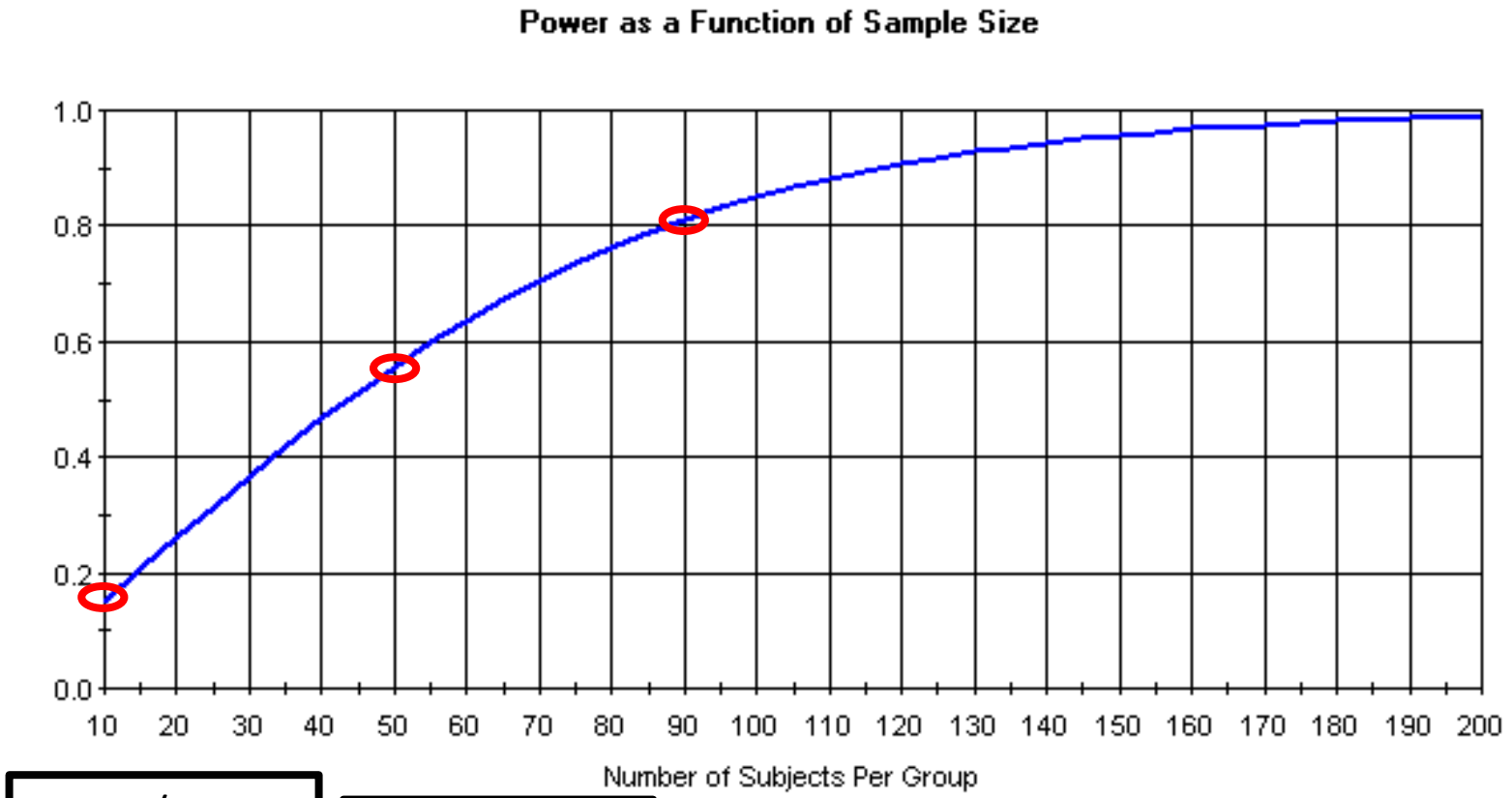
If p-value ≤ α:
Reject the null hypothesis, evidence supports the alternative hypothesis.
If p-value > α:
Fail to reject the null hypothesis, insufficient evidence to support the alternative hypothesis.

Errors in Decision Making

Conclusion based on data	Null Hypothesis (H_0) is True	Null Hypothesis (H_0) is False
Reject Null Hypothesis	Type-I Error (α) Common values: 0.05	Correct conclusion (Power = 1-β) Common values: 0.8
Fail to Reject Null Hypothesis	Confidence interval (1-α) Common values: 0.95	Type II Error (β)

- Alpha: The maximum probability of making Type I error. $\alpha = 0.05$
- Beta: When the alternative hypothesis is true, the probability of rejecting it (the probability of making a type II error)
- Power: The ability of the test to detect a true effect when it's there

Power and sample size



N = 10/group
Power = 0.15

N = 50/group
Power = 0.55

N = 90/group
Power = 0.8

- Increase sample size leads to increase of power
- diminishing returns for adding more and more subjects

Choosing an Effect Size

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Cohen's d = the difference between two means divided by the pooled standard deviation

Effect Size	d
Very Small	0.01
Small	0.2
Medium	0.5
Large	0.8
Very Large	1.2
Huge	2.0

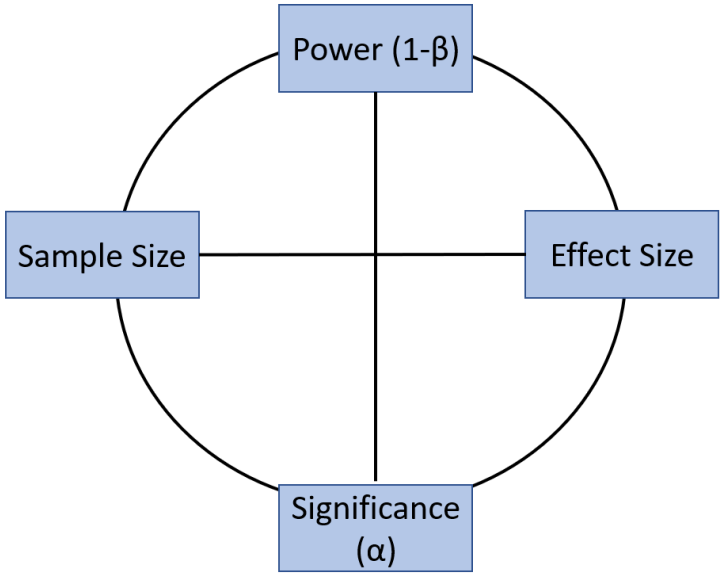
Common Effect Size Indices

Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's *d*[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (*M*); *s* is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR)	$\frac{\text{Group 1 odds of outcome}}{\text{Group 2 odds of outcome}}$ If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Measures of association			
Pearson's *r* correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables

Using Effect Size—or Why the P Value Is Not Enough, Sullivan and Feinn, J Grad Med Educ. 2012

Power Analysis

- Four inter-related concepts:
  - power
  - effect size
  - sample size
  - significance level (α)



- Each is a function of the other three. If three of these values are fixed, the fourth is completely determined (Cohen, 1988)

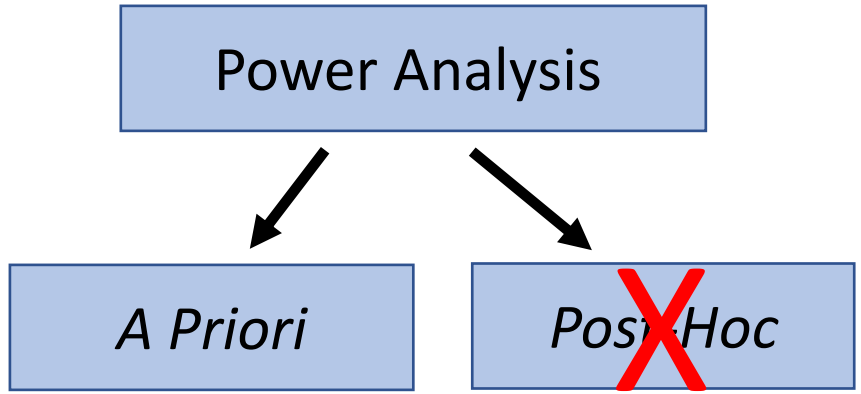
Software

- G*Power 3: Faul et al., Behavior Research Methods, 2007
- Power Analysis & Sample Size (PASS)
- R: pwr package

Why and when do we need power analysis?

- To determine the necessary number of subjects needed to detect an effect of a given size.
- To determine power, given an effect size and the number of subjects available.
- No point to conducting a study that is seriously underpowered.

Power analysis is conducted before the study begins.



The Journal of Rheumatology 2022;xx:xxxx
doi:10.3899/jrheum.211115
First Release May 15 2022

Expert Review

Post Hoc Power Calculations: An Inappropriate Method for Interpreting the Findings of a Research Study

Michael G. Heckman[1], John M. Davis III[2], and Cynthia S. Crowson[3]

The Journal of Rheumatology

Post-hoc power estimates (power calculated for hypothesis tests after performing them) are sometimes requested by reviewers in an attempt to promote more rigorous designs. However, they should never be requested or reported because they have been shown to be logically invalid and practically misleading. We review the problems associated with post-hoc power, particularly the fact that the resulting calculated power is a monotone function of the *p*-value and therefore contains no additional helpful information. We then discuss some

The Interpretation of Statistical Power after the Data have been Gathered, Dziak et al., Curr Psychol. 2018

Power analysis is only meaningful for future experimental design.

A priori vs. Post-Hoc power analysis

- Power analysis is essential before data collection:
  - Determine required sample size (n) for detecting a biologically meaningful effect
  - Avoid underpowered studies (false negatives)
  - Prevent overpowered studies (wasted resources, inflated trivial significance)
- A design-stage (a priori) power analysis is the gold standard use.

- Use of power for completed studies (Post-Hoc Power analysis)
  - Interpreting negative results: no true effect vs. lacked power to detect a realistic effect size
  - A high-profile claim from an under-powered study should be viewed with extreme skepticism.

Smallest Effect Size of Interest (SESOI): An effect size that is biologically or clinically meaningful, not just statistically significant.

General Workflow: Multiple Testing Problem

- Measure the expression of thousands of genes at a time
  - For each gene we test the null hypothesis that there is no differential expression.
  - Many thousands of statistical tests are performed

- Suppose *g* null hypotheses are being tested
  - g*α = The expected number of false positives if ALL null hypotheses were true.
  - 10,000 X 0.05 = 500 false positive DEGs

- Family-wise error rate (FWER): the probability of rejecting at least one null hypothesis given they are all true (1 false positive): Bonferroni Correction: $\frac{\alpha}{g}$

- False Discovery Rate: $p(k) \leq (k / g) * \alpha$

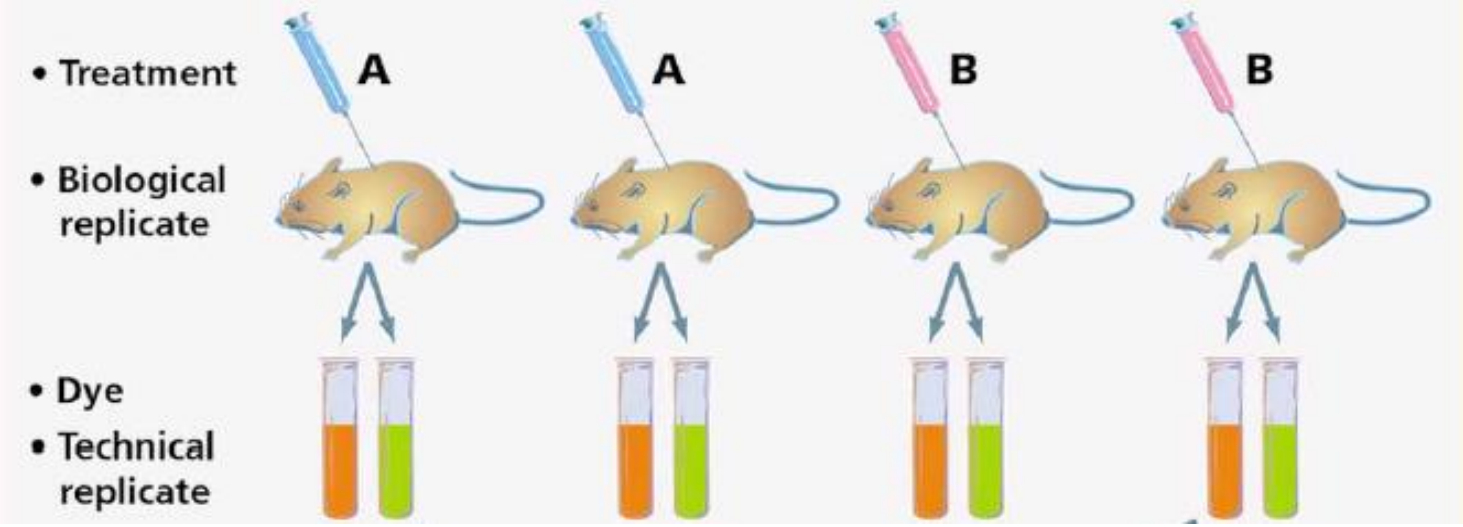
Gene Expression General Workflow

- Sample acquisition and handling
- RNA extraction
- RNA concentration and quality assessment
- Assay optimization and data collection
- Data analysis

General Workflow: Sample acquisition and handling

- Type of replicates:
  - Biological replicates
  - Technical replicates

- As many Biological replicates as possible



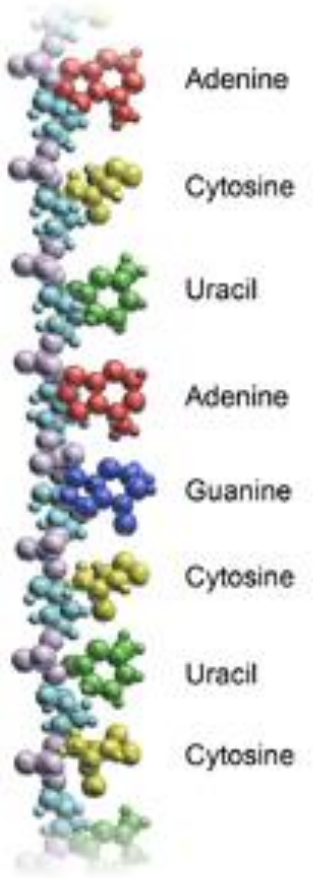
Churchill *Nature Genetics Supplement 32: 490-495* (2002)

General Workflow: RNA Extraction

- RNA extracted from tissue is very heterogeneous
  - Many cells and difference cell types

- Total RNA contains different types of RNA
  - Only 1-2% is mRNA
  - Remainder is rRNA, tRNA, etc

- extremely susceptible to degradation



General Workflow: RNA Quality Control

It is important to ensure you have an intact RNA preparation, at the correct concentration and purity, before jumping into your experiments.

1. Measuring RNA concentration

- Measure UV absorption of the sample using a spectrophotometer at 260nm
- the Beer Lambert Law: an A260 of 1.0 = 40 µg/mL of RNA.

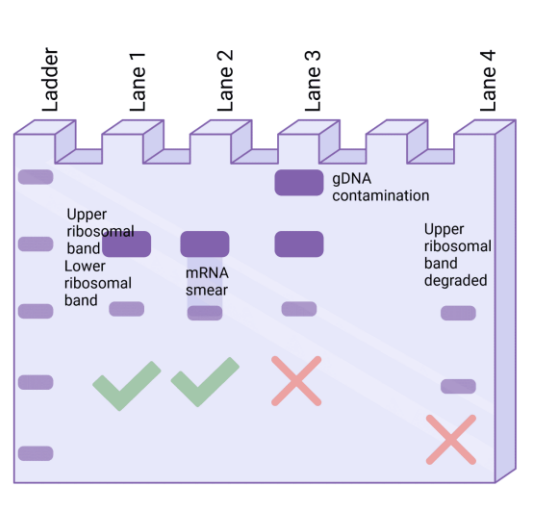
2. Measuring RNA Purity

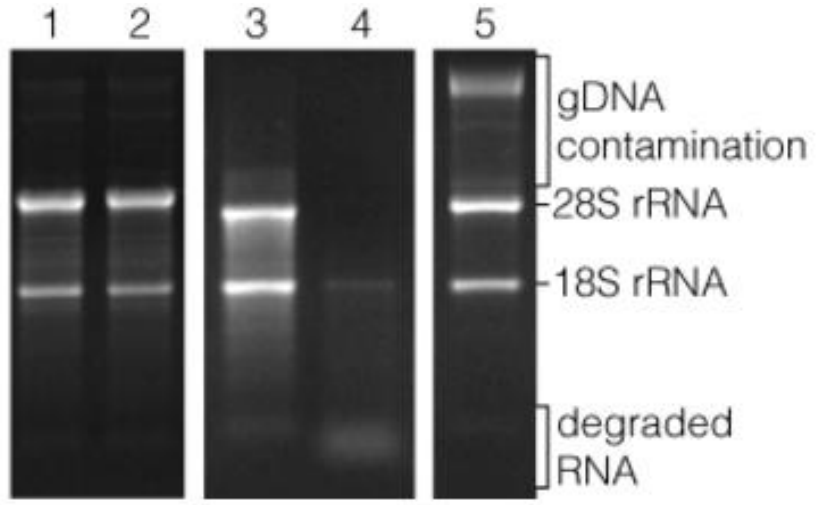
- Protein Contamination: A260/A280 ratio
  - Pure RNA A260/A280 = 2.1
  - Values between 1.8–2.0 are considered acceptable for many protocols.

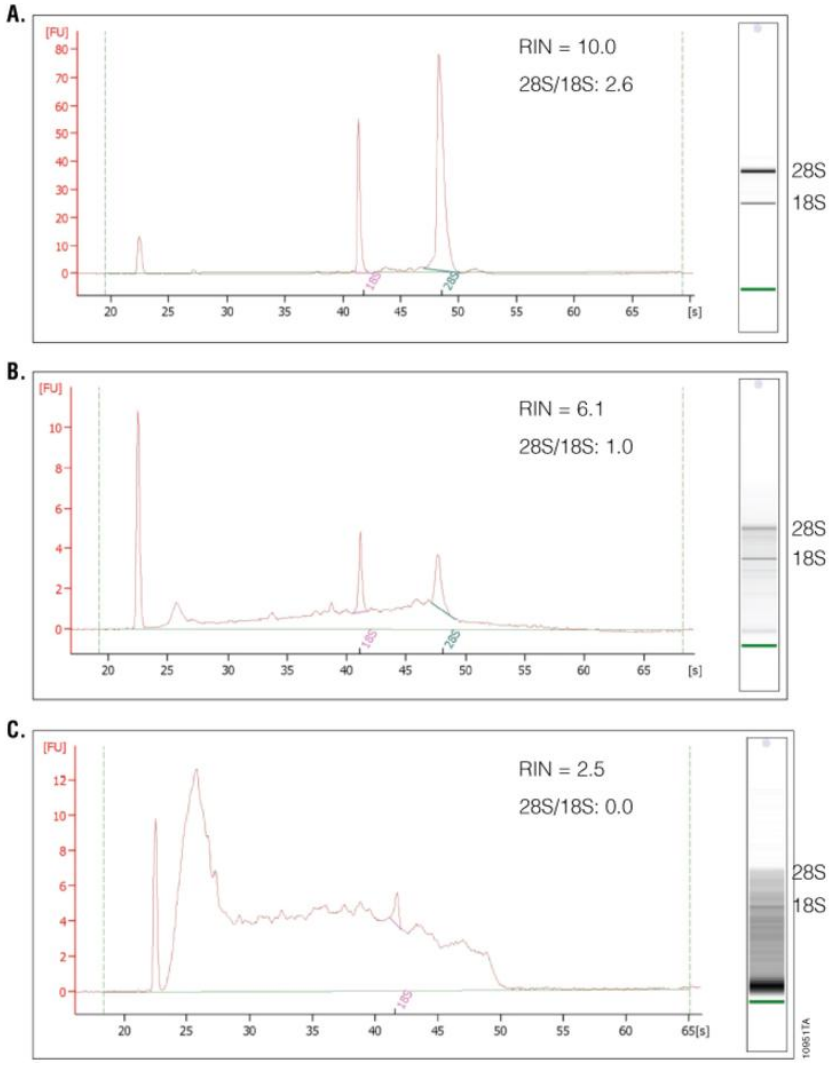
- Other contaminants: A230
  - A high peak at A230 indicates contamination with guanidine salts or phenol.
  - The ideal A260/A230 ratio is greater than 1.5

Measuring RNA Integrity

- Agarose Gel



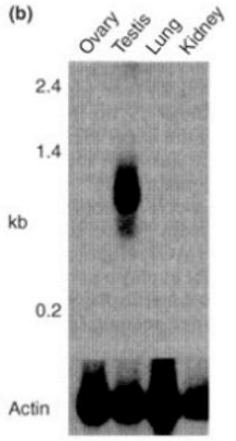


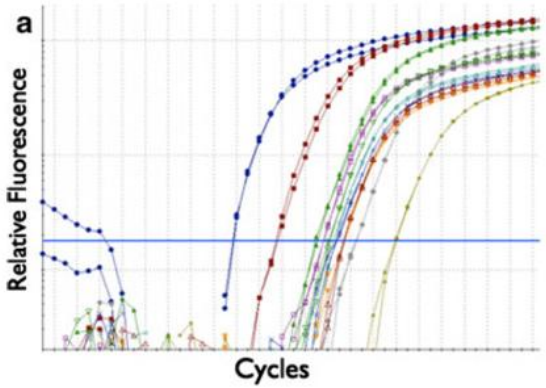


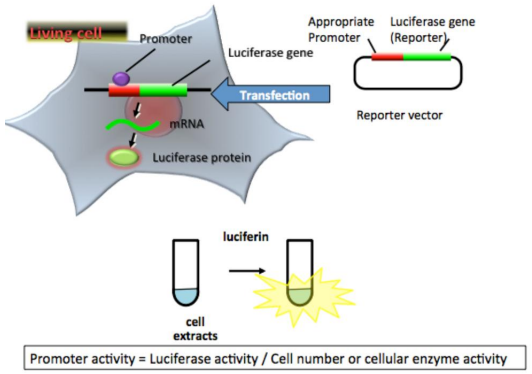
- Bioanalyzers
  - more expensive
  - use small amounts of RNA (1–2µL) and microfluidics
  - determine the quantity and quality of RNA samples.
  - RNA Integrity Number (RIN)

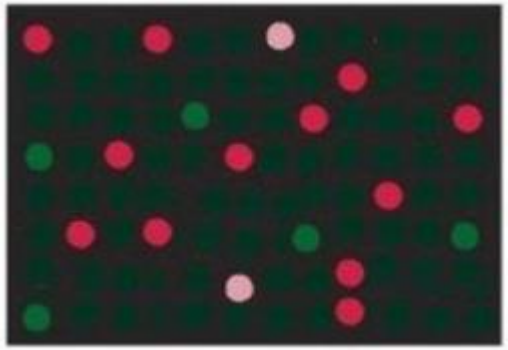
General Workflow: Assay Optimization and Data Collection

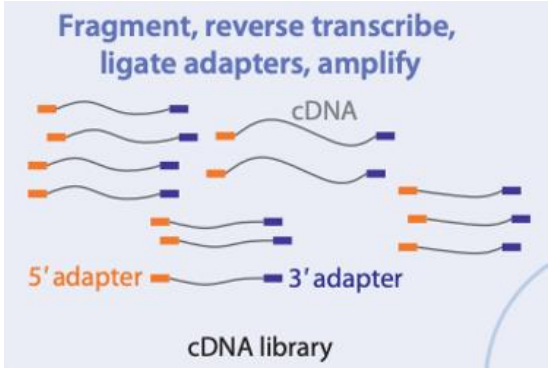
- Assay-specific











General Workflow: Data Analysis

- Reference sample, genome, gene,
  - Controls could be heterogenous
  - Reference genome versions
  - Reference gene expression could change in treatment group (GAPDH, ACTB)

- Data normalization
  - Different loading quantity
  - Different input number of cells
  - Different transfection efficiency

- Outliers: biological vs. technical

- Statistical test

Choosing the appropriate statistical test

- The decision for a statistical test is based on the scientific question to be answered, the data structure and the study design.

- The question to be answered and the null hypothesis must be formulated before the data are recorded and the statistical test is selected.

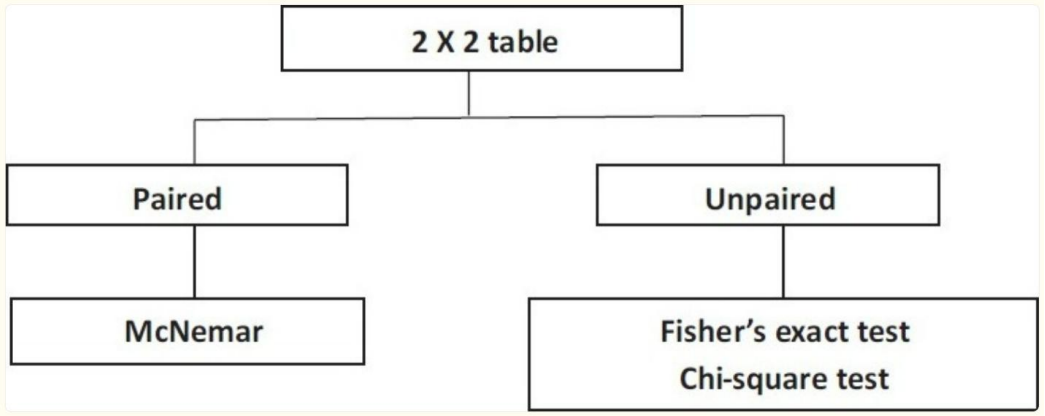
- Select statistical test and the level of significance before the study begins to ensure that the study results do not influence the test selection.

- It must be decided whether the test should be one-tailed or two-tailed.
  - two-tailed means that no particular direction of expected difference is assumed.
  - one-tailed test should only be performed when there is clear evidence that the intervention should only act in one direction.

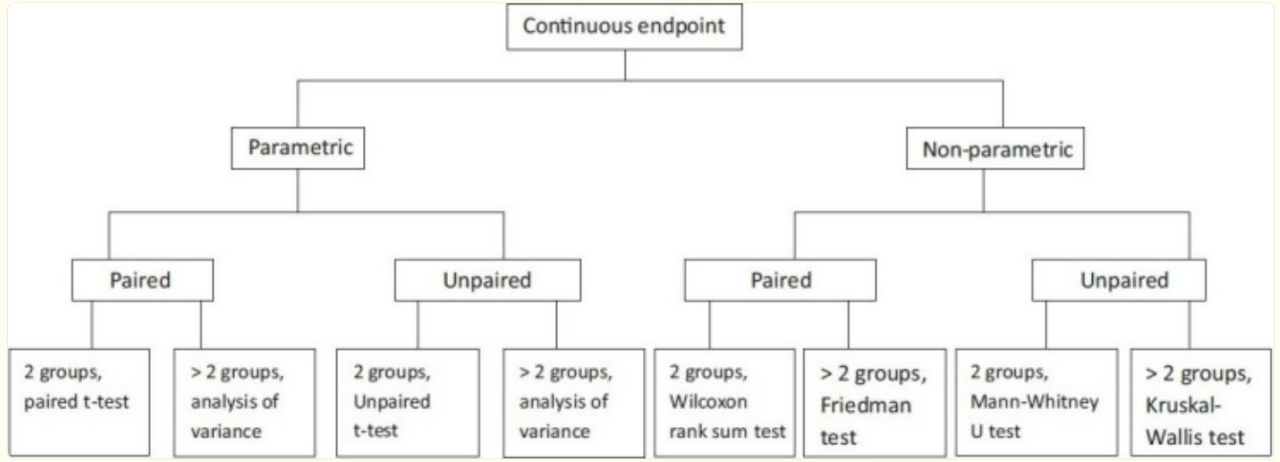
Choosing statistical test, Parab and Bhalerao, Int J Ayurveda Res. 2010

Choosing the appropriate statistical test

group comparison with two categorical endpoints



group comparison of a continuous endpoint



Choosing statistical test, Parab and Bhalerao, Int J Ayurveda Res. 2010

Summary

- RNA biology I
  - Introduction of RNA type and function
  - RNA quantification technology
    - Targeted
      - Northern blotting
      - *In situ* Hybridization
      - RT-qPCR
      - Reporter Assay
    - High-throughput
      - Microarray
      - RNA-seq
  - Experimental design principles
  - General workflow
- RNA biology II
  - Transcriptome profiling by RNA-seq
    - Experimental design
    - RNA quantification
    - Quality control and normalization
    - Outlier detection
    - Differentially expressed gene detection
    - Result interpretation

Common Effect Size Indices

Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's d[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (M); s is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR)	$\frac{\text{Group 1 odds of outcome}}{\text{Group 2 odds of outcome}}$ If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR)	Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups	Small 2 Medium 3 Large 4	Compares probabilities of outcome occurring from one intervention to another
Measures of association			
Pearson's r correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables
r^2 coefficient of determination	Range, 0 to 1; Usually expressed as percent	Small 0.04 Medium 0.25 Large 0.64	Proportion of variance in one variable explained by the other

Using Effect Size—or Why the P Value Is Not Enough, Sullivan and Feinn, J Grad Med Educ. 2012