中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

国家基因组科学数据中心
National Genomics Data Center

第9章 RNA与转录组

章张 博士

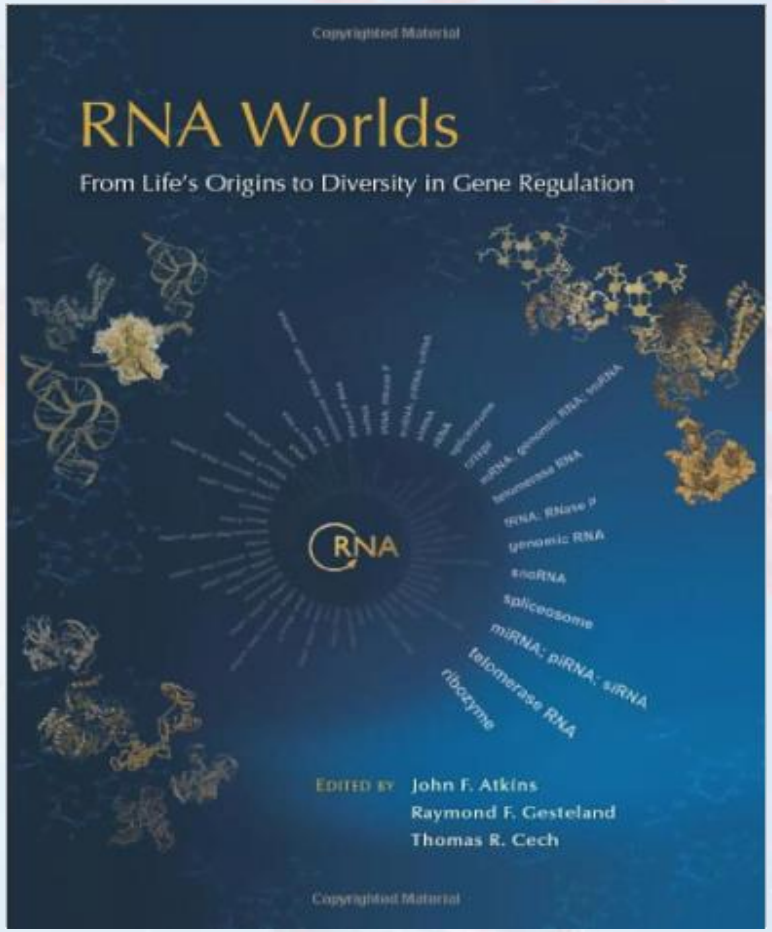
中国科学院大学
University of Chinese Academy of Sciences

基因组学 - Genomics

2024春季学期

内容提纲

- RNA基本概念与RNA世界假说
- RNA主要类型
- RNA生命动态
- 转录组
- 基因表达



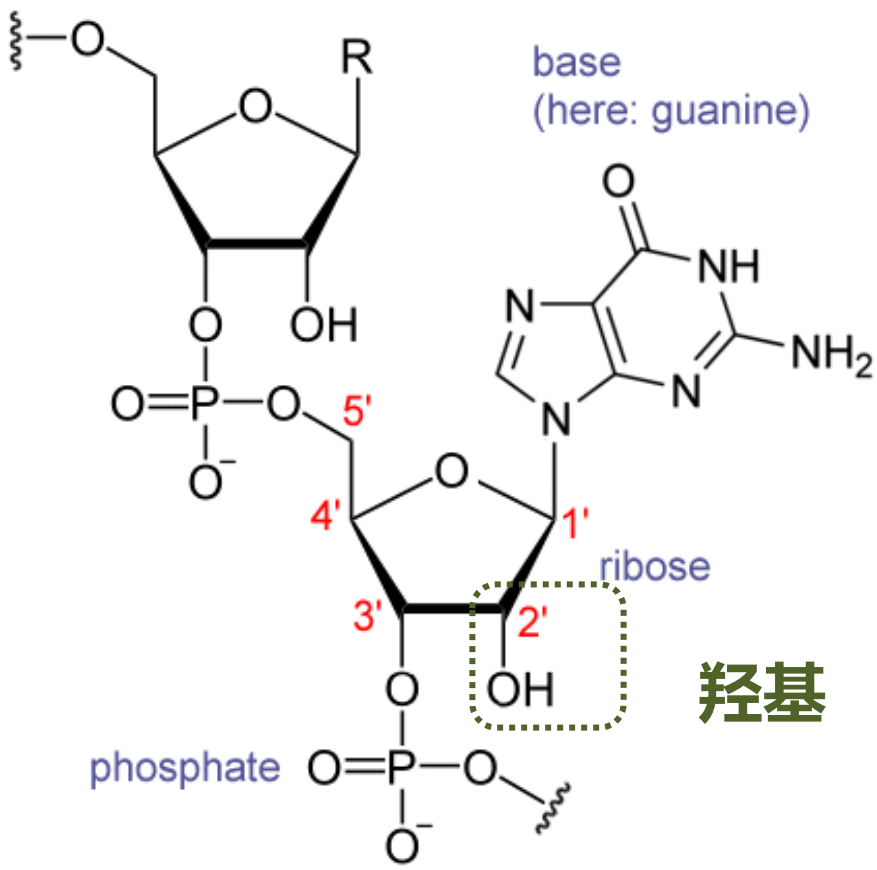
滚滚长江东逝水，浪花淘尽英雄。是非成败转头空。青山依旧在，几度夕阳红。——《临江仙》

RNA基本概念与RNA世界假说

RNA

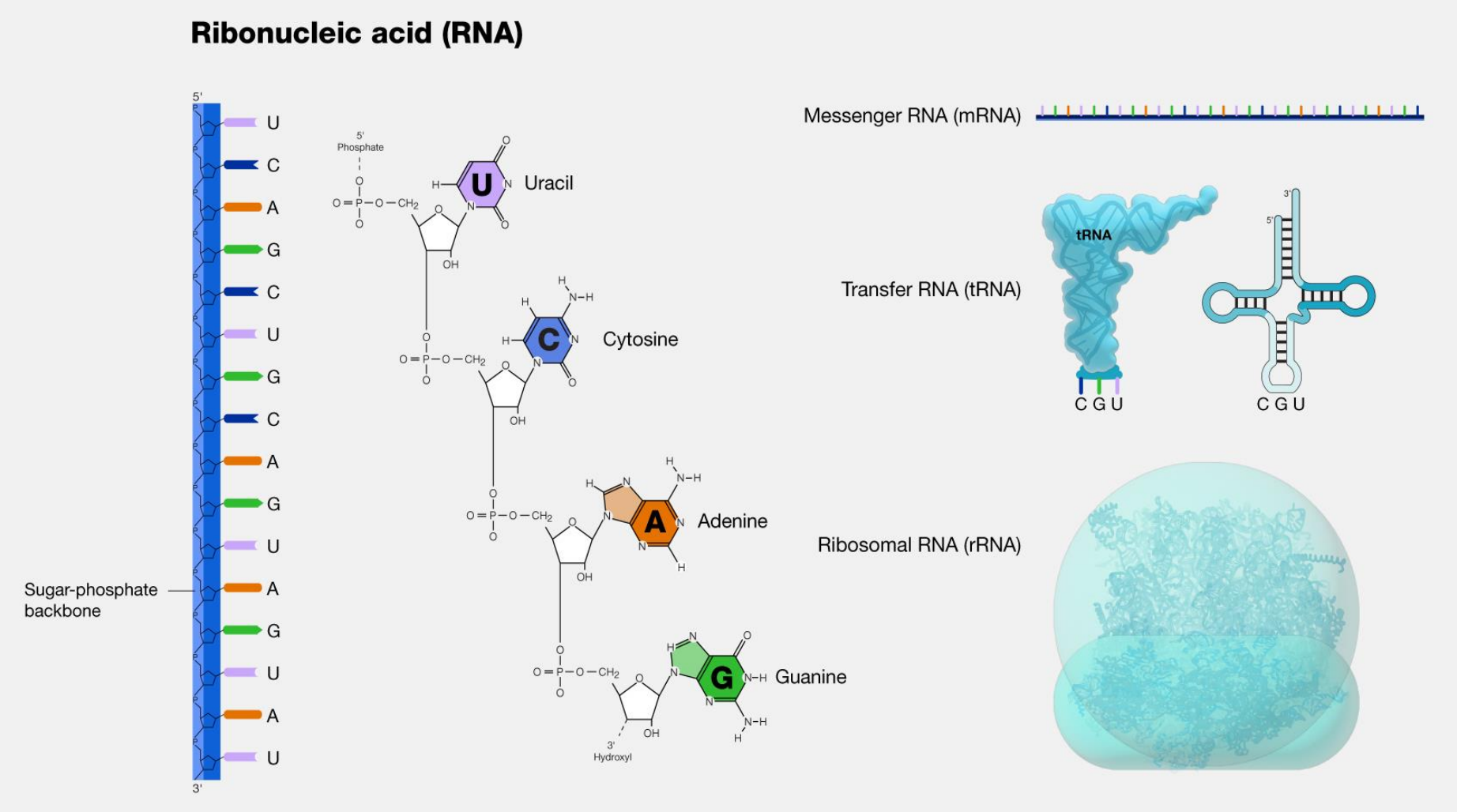
- 核糖核酸（缩写为RNA，即Ribonucleic Acid），存在于生物细胞以及部分病毒、类病毒中的遗传信息载体。
- RNA由核糖核苷酸经磷酸二酯键缩合而成长链状分子。
- RNA的碱基有四种，即A腺嘌呤、U尿嘧啶、G鸟嘌呤、C胞嘧啶，其中，U取代了DNA中的T。

RNA化学结构



https://en.wikipedia.org/wiki/RNA_world

RNA

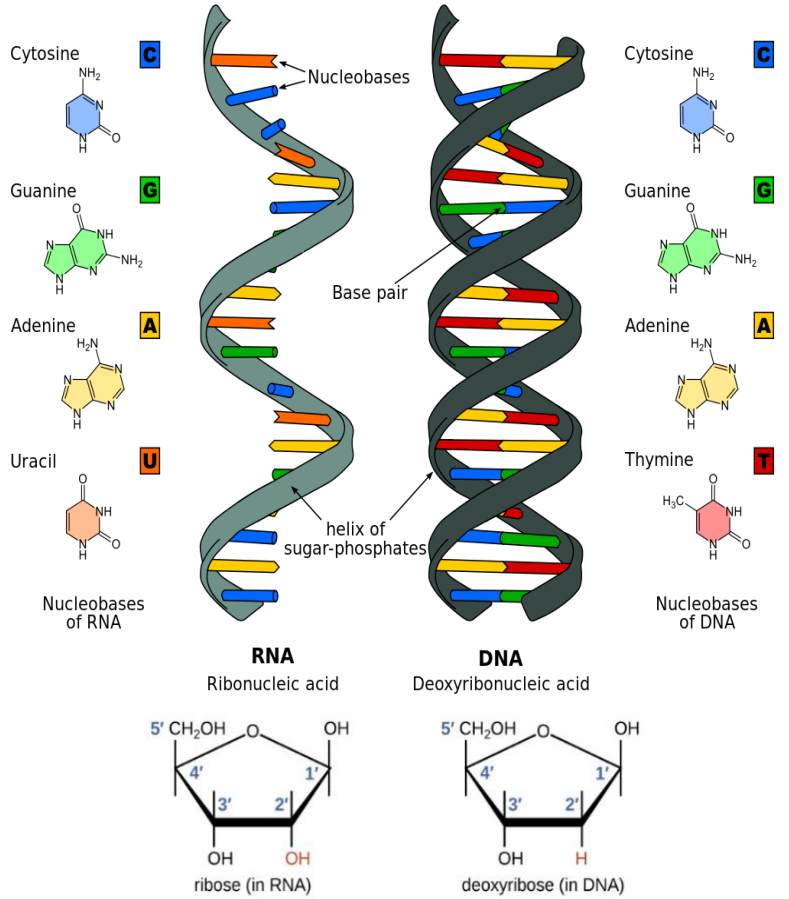


https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid

RNA与DNA区别

项目	DNA	RNA
名称	脱氧核糖核酸 DeoxyriboNucleic Acid	核糖核酸 Ribonucleic Acid
五碳糖种类	脱氧核糖	核糖
骨架结构	双螺旋结构	单链结构
碱基种类	腺嘌呤（A） 胸腺嘧啶（T） 胞嘧啶（C） 鸟嘌呤（G）	腺嘌呤（A） 尿嘧啶（U） 胞嘧啶（C） 鸟嘌呤（G）
五碳糖2位碳连接组分	氢原子	羟基
存在	细胞核（少量存在于线粒体、叶绿体）	细胞质

RNA与DNA结构、碱基组成



https://en.wikipedia.org/wiki/RNA_world

https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-rna/

RNA发现史

RNA领域重要发现

- 1868年，Friedrich Miescher首次在真核细胞核中发现
- 1939年，Torbjörn Caspersson 发现RNA在蛋白质合成中扮演重要角色
- 1967年，**Carl Woese提出RNA世界学说**，指出最早的生命形式可能依赖于RNA，**携带遗传信息**和**催化生化反应**
- 1976年，Walter Fiers确定**RNA病毒**完整基因组碱基序列

RNA研究相关诺贝尔奖

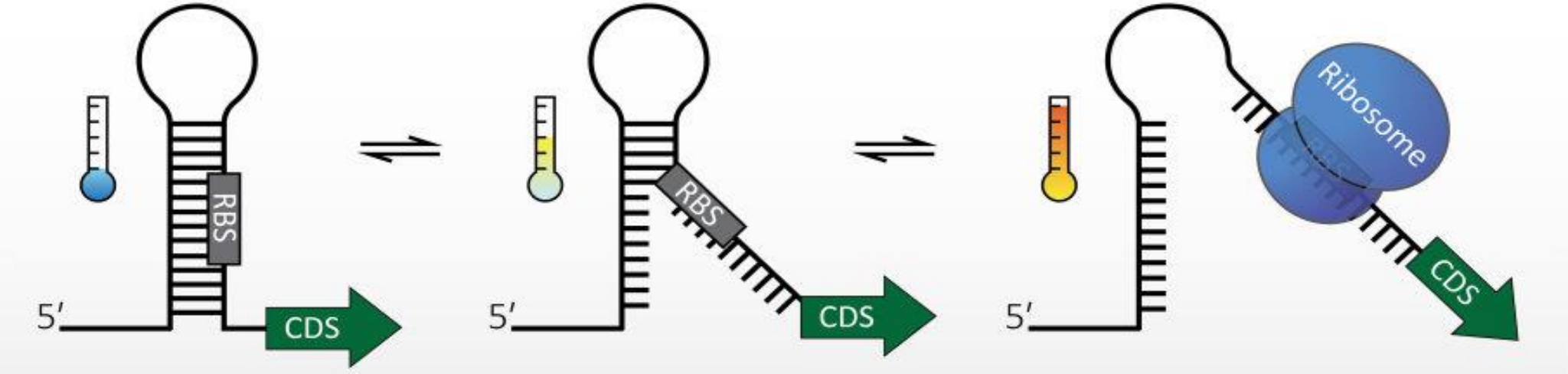
年份	获奖原因
1959	发现RNA合成酶
1968	发现tRNA
1975	发现逆转录病毒及逆转录酶
1989	发现RNA具有催化特性
1993	发现RNA剪接
2006	发现RNA干扰现象
2006	真核生物基因转录的分子基础
2023	mRNA新冠疫苗

RNA世界假说

- **RNA世界假说（RNA world hypothesis）**：认为RNA是地球早期生命分子，这些早期RNA分子同时拥有**DNA遗传信息储存功能**，及**蛋白质的催化能力**，从而支持早期细胞或前细胞生命的运作。
  - 1962年，Alexander Rich首次提出RNA世界的类似想法
  - 1968年，**卡尔·沃斯**在《遗传密码》建立RNA生命型态的概念
  - 1986年，诺奖得主**沃特·吉尔伯特**（Walter Gilbert）正式提出
- **RNA 作为酶**
  - 可自我复制
  - 可催化简单化学反应
  - 可催化肽键形成
- **RNA 作为信息存储介质**
  - 比较脆弱，易水解
  - 不稳定，变异可能性大
- **RNA作为调控物质**
  - 核糖开关（Riboswitch）：基因表达的调控物质之一，在细菌、植物和古菌中发现，改变其二级结构以响应所结合的代谢物，从而影响转录
  - RNA温度计（RNA thermometer）：受温度变化而调节基因表达

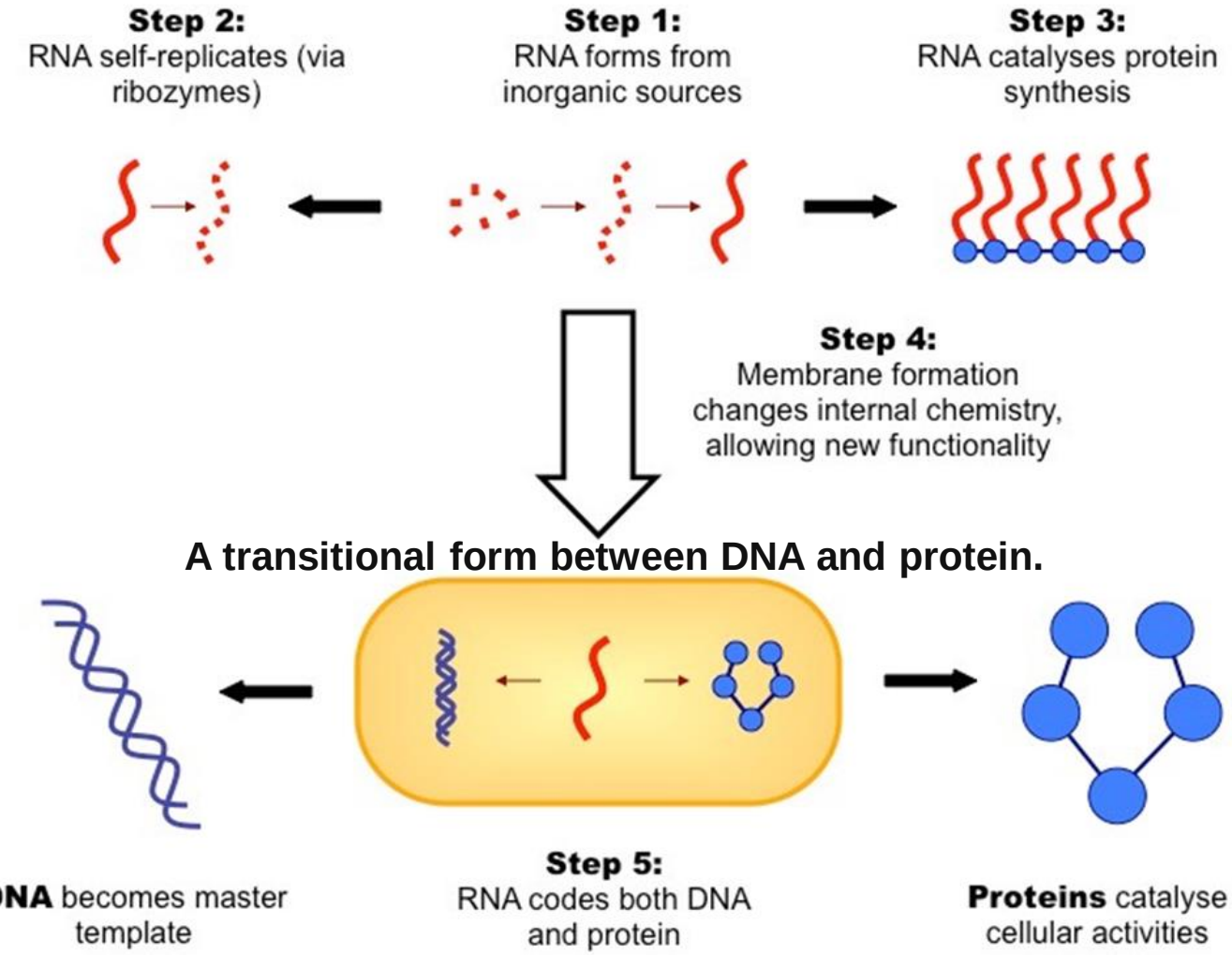
RNA thermometer

- RNA温度计（RNA thermometer）是一类**对温度敏感的非编码RNA**，能**随温度变化调控基因表达**。主要调控与**热休克和冷休克**反应有关的基因，与致病性、饥饿状态等过程相关的基因调控也有关系。
- **二级结构随温度变化而改变**，使RNA上核糖体结合位点（RBS）等重要区域暴露或遮蔽，进而**改变对应编码基因的翻译速率**。
- 一般认为，在早期的RNA世界中，RNA温度计的作用是对其他的RNA分子进行**温度依赖性的调控**。在现代生物中，RNA温度计可以说是一个RNA世界的良好**分子化石**。



Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology* 2012

RNA World



DNA, through its ***superior chemical stability*** (double helical structure) has taken over as the data storage form.

Protein, through its ***greater variability*** (20 amino acids as opposed to 4 nucleotides) has taken over as the catalytic form.

https://ib.bioninja.com.au/standard-level/topic-1-cell-biology/15-the-origin-of-cells/rna-world-hypothesis.html

其它假说：RNA-peptide world

Article | Open Access | Published: 11 May 2022

A prebiotically plausible scenario of an RNA–peptide world

Felix Müller, Luis Escobar, Felix Xu, Ewa Węgrzyn, Milda Nainytė, Tynchtyk Amatov, Chun-Yin Chan, Alexander Pichler & Thomas Carell

Nature **605**, 279–284 (2022) | Cite this article

76k Accesses | **39** Citations | **581** Altmetric | Metrics

Research Article | Open Access

Loading of Amino Acids onto RNA in a Putative RNA-Peptide World

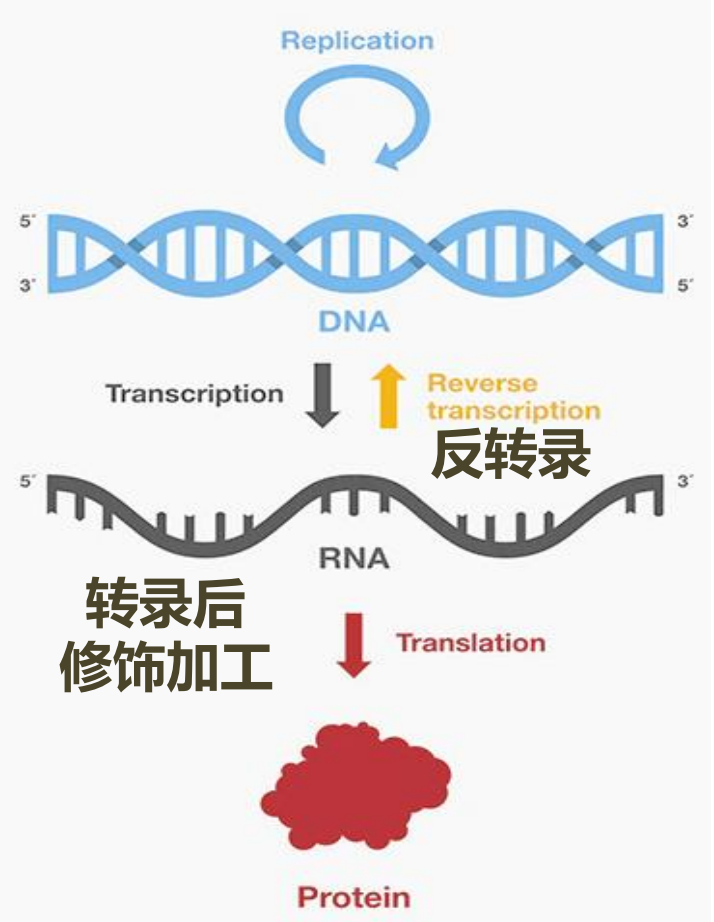
Johannes N. Singer, Felix M. Müller, Ewa Węgrzyn, Dr. Christina Hölzl, Hans Hurmiz, Chuyi Liu, Dr. Luis Escobar, Prof. Dr. Thomas Carell

First published: 07 March 2023 | https://doi.org/10.1002/anie.202302360

RNA复杂性

- 遗传信息传递与功能执行载体
- RNA生命动态
  - RNA合成
  - 调控
  - 加工修饰
  - 转运
  - 降解
  - 逆转录
- RNA类型与功能多样
- 与多种生物大分子存在交互作用

转录组（Transcriptome）：所有转录产物的集合，包括信使RNA、核糖体RNA、转运RNA及非编码RNA。

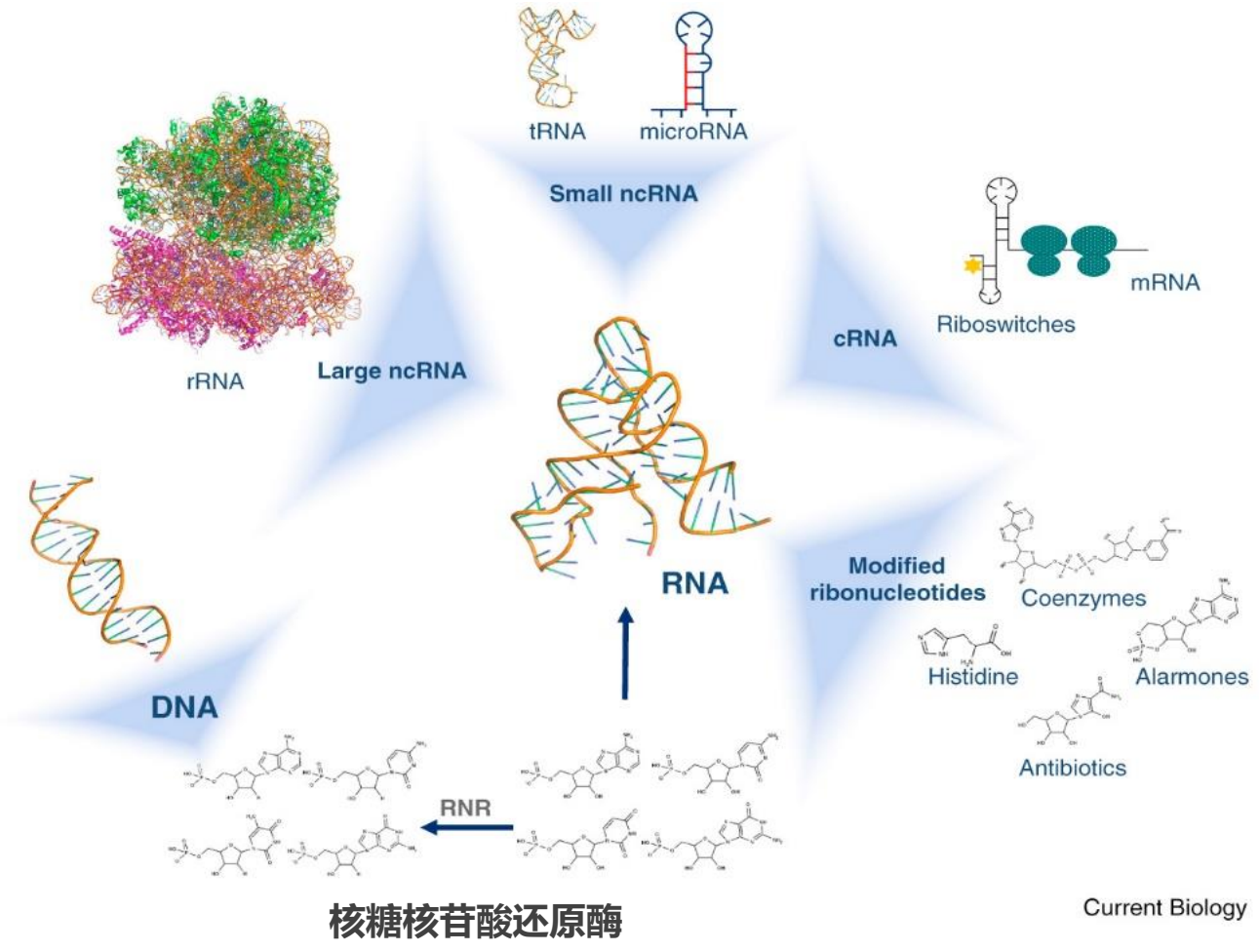


RNA组：全生命周期合成的RNA分子和共价结构集合

RNA主要类型与功能

RNA主要类型

- 在细胞生物中，mRNA为遗传讯息的传递者，指导蛋白质合成，被称为**编码RNA**（coding RNA）
- 没有编码蛋白质能力的RNA则被称为**非编码RNA**（non-coding RNA）
- 非编码RNA通过催化生化反应，或调控、参与基因表达过程而发挥相应的生理功能
- 功能：蛋白合成、调控、修饰、复制等



Early Life: Embracing the RNA World. *Current Biology* 2018

RNA代表类型及功能

RNA类型	缩写	功能
Messenger RNA (信使RNA)	mRNA	蛋白编码
Ribosomal RNA (核糖体RNA)	rRNA	核糖体结构功能组成部分
Transfer RNA (转运RNA)	tRNA	翻译过程中接附氨基酸
MicroRNA (微RNA)	miRNA	mRNA降解与翻译抑制
Long non-coding RNA (长非编码RNA)	lncRNA	基因调控
Small interfering RNA (小分子干扰RNA)	siRNA	转录后基因沉默
Circular RNA (环形RNA)	circRNA	基因调控
……	……	……

https://en.wikipedia.org/wiki/List_of_RNAs

非编码RNA

	Type	Length (nt)	Function
Small ncRNAs (<200 nt)	MicroRNA (miRNA)	18–22	RNAi; Protein translation regulation
	Small interfering RNA (siRNA)	20–25	RNAi; Antiviral defense
	Piwi RNA (piRNA)	26–31	Regulation of transposable elements
	Trans-activating CRISPR (tracr) RNA	~65	CRISPR/Cas adaptive immunity in bacteria
	Small nuclear RNA (snRNA)	100–300	Intron splicing; RNA processing
	U-rich snRNA (snRNA)	100–300	snRNA subclass; intron splicing
	Small nucleolar RNA (snoRNA)	60–200	rRNA processing
	Signal recognition particle RNA (7SL)	300	RNA component of the signal recognition particle; protein synthesis
	Y RNAs	80–120	Components of ribonucleoproteins; DNA replication & RNA processing
	Small Cajal body-specific RNA (scaRNA)	200–300	Biogenesis of small nuclear ribonucleoproteins
Long ncRNAs (>200 nt)	Long intergenic ncRNA (lincRNA)	1 kb	Protein scaffolding
	Natural antisense transcript (NAT)	>200	RNAi, alternate splicing, genome imprinting
	Circular RNA (circRNA)	100–999	miRNA decoy, protein regulation
Housekeeping ncRNAs	Ribosomal RNA (rRNA)	>1,500	Protein synthesis
	Transfer RNA (tRNA)	76–90	Protein synthesis

https://www.bio-rad.com/en-uk/applications-technologies/coding-non-coding-rna?ID=Q1070M70KWE7

RNAs in human

Table 1. Human Genes

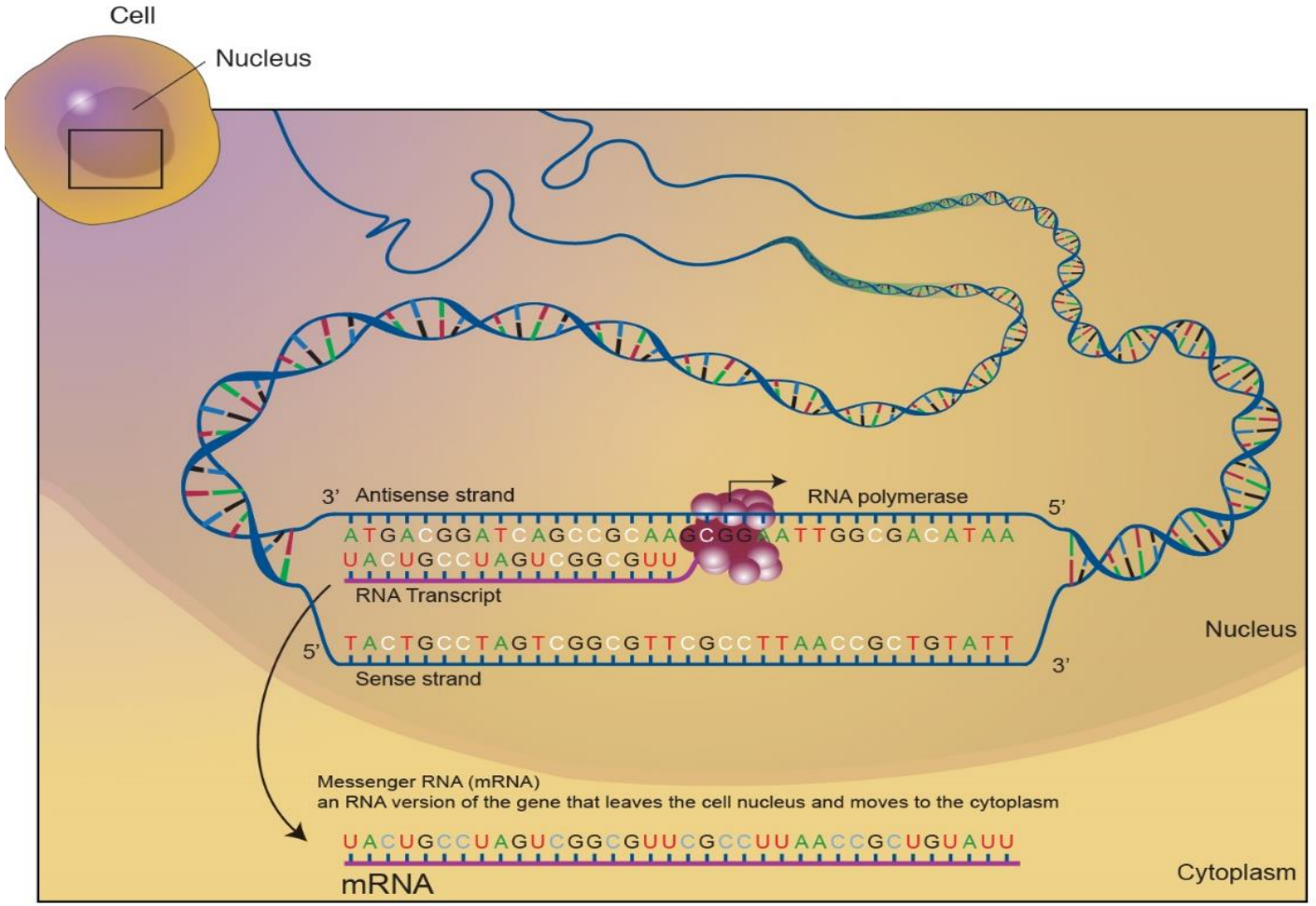
Class	Abbreviation	Function	Estimated Number
Messenger RNA	mRNA	protein coding	20,078[a]
Ribosomal RNA	rRNA	structural and functional component of ribosome	531[a]
Transfer RNA	tRNA	translational adaptor molecule	512[b]
Small nuclear RNA	snRNA	processing of pre-mRNA	1,923[a]
Small nucleolar RNA	snoRNA	processing of rRNA, tRNA, and snRNA	1,529[a]
Antisense RNA	aRNA	gene regulation	4,424[a]
Long noncoding RNA	lncRNA	gene regulation	12,933[a]
MicroRNA	miRNA	translational inhibition and mRNA degradation	>500[c]
Small interfering RNA	siRNA	posttranscriptional gene silencing	n/a[d]
Piwi-interacting RNA	piRNA	protect genome integrity	n/a[e]
Enhancer RNA	eRNA	unknown	variable[f]

[a]GENCODE (Harrow et al., 2012).
[b]HGNC database (Seal et al., 2011; http://www.genenames.org).
[c]2,000–3,000 putative miRNA have been annotated (Harrow et al., 2012), but the majority are not validated.
[d]Human endogenous siRNA are rare and have not been systematically identified.
[e]Piwi-interacting RNA (piRNA) have not been systematically identified, although estimates indicate that hundreds of piRNAs are derived from each of more than 100 loci (Aravin et al., 2006; Brennecke et al., 2007; Girard et al., 2006).
[f]Enhancer RNAs (eRNAs) are generated from active enhancers, thus the number of eRNAs depends on the set of active enhancers in a cell (Kim et al., 2010; Wang et al., 2011). Current estimates indicate that 25%–80% of active enhancers generate eRNAs.

Transcriptional Regulation and Its Misregulation in Disease. *Cell* 2013
https://doi.org/10.1016/j.cell.2013.02.014

信使RNA (Messenger RNA)

- **信使RNA**（mRNA）：单链RNA，携带遗传信息，利用**遗传密码**将每个碱基三联体（或**密码子**）翻译成相应氨基酸，将遗传信息从DNA传递到核糖体，指导**蛋白质合成**
- **mRNA**占细胞总RNA的2%~5%，**种类多，代谢活跃，半衰期短**，合成后数分钟至数小时即被分解



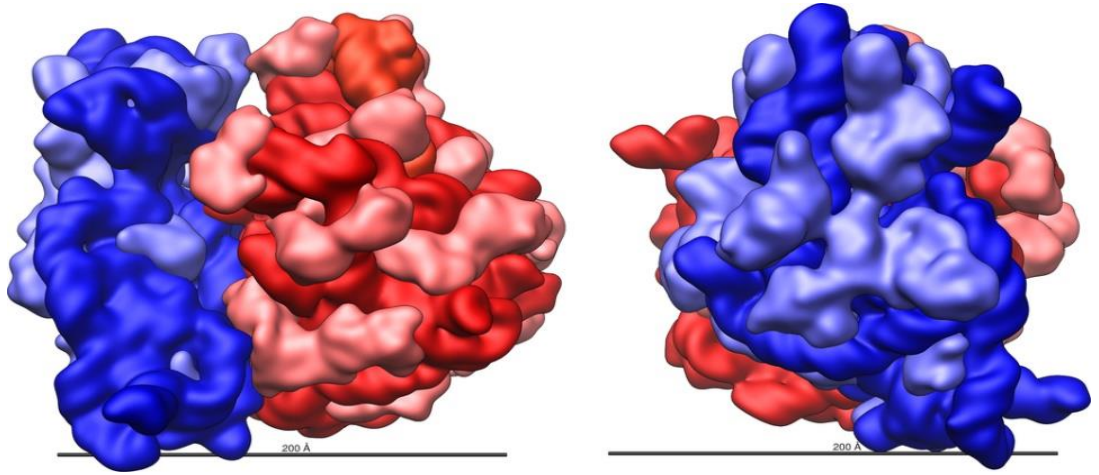
https://www.genome.gov/genetics-glossary/messenger-rna

核糖体RNA (Ribosomal RNA)

- **核糖体RNA**（rRNA），转录自核糖体DNA，分子质量大，结构复杂，是细胞内含量最多的一类RNA，占RNA总量的82%左右
- 功能为**在mRNA的指导下将氨基酸合成为肽链**，rRNA与多种蛋白质结合成**核糖体（Ribosome）**，作为**蛋白质生物合成的“装配机”**

生物种类	类型	大亚基	小亚基
原核生物	70S	50S (5S: 120nt, 23S: 2906nt)	30S (16S: 1542nt)
真核生物	80S	60S (5S: 121nt, 5.8S: 156nt, 28S: 5070nt)	40S (18S: 1869nt)

S：沉降系数（斯韦德伯格Svedberg单位），S为大分子物质在超速离心沉降中的一个物理学单位，可间接反映分子量的大小。



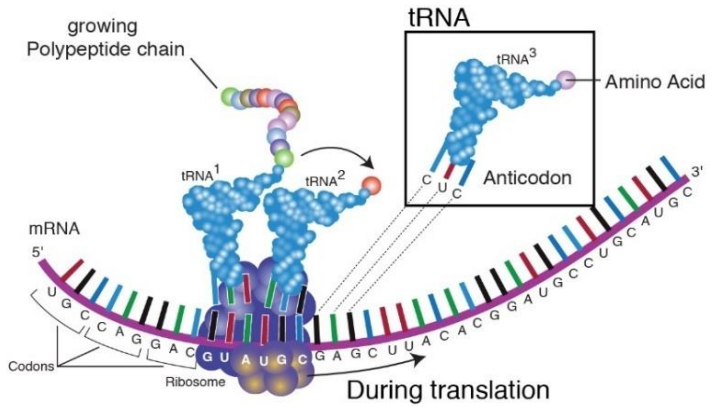
Three-dimensional views of the ribosome, showing rRNA in **dark blue (small subunit)** and **dark red (large subunit)**. Lighter colors represent **ribosomal proteins**.

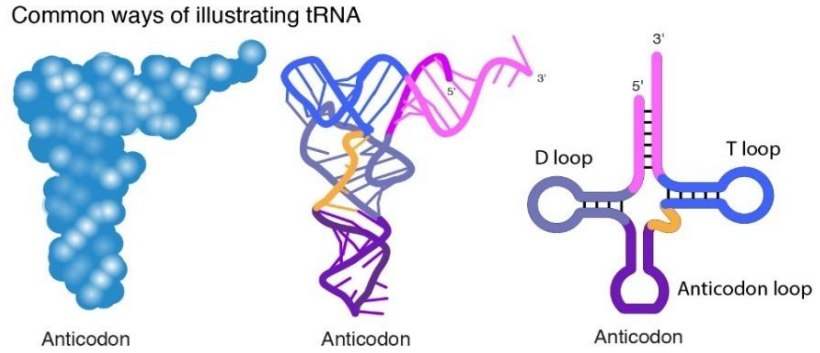
https://en.wikipedia.org/wiki/Ribosomal_RNA

转运RNA (Transfer RNA)

- **转运RNA**（tRNA），由70-90个核苷酸组成，折叠成三叶草形状
- **氨酰-tRNA合成酶**催化tRNA 3'端接附特定种类氨基酸
- 翻译过程中，tRNA**反密码子**识别mRNA的**密码子**，将密码子所对应**氨基酸转运至核糖体合成多肽链**
- 由于**遗传密码简并性**，存在多种tRNA与一种氨基酸接附现象
- 1968年，**罗伯特·W·霍利**因其首次分离tRNA，并阐明其序列与结构而获得诺贝尔生理学或医学奖

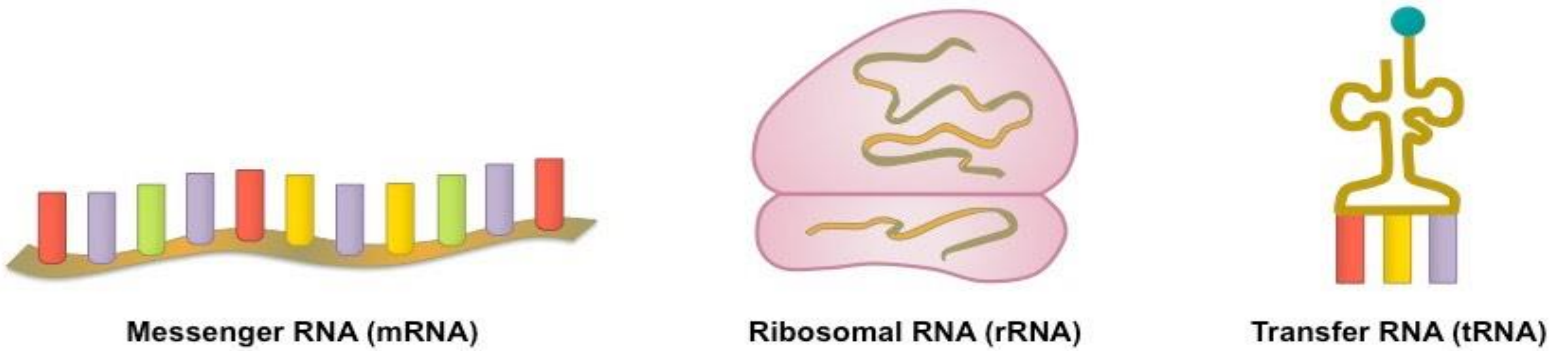
tRNA结构与功能

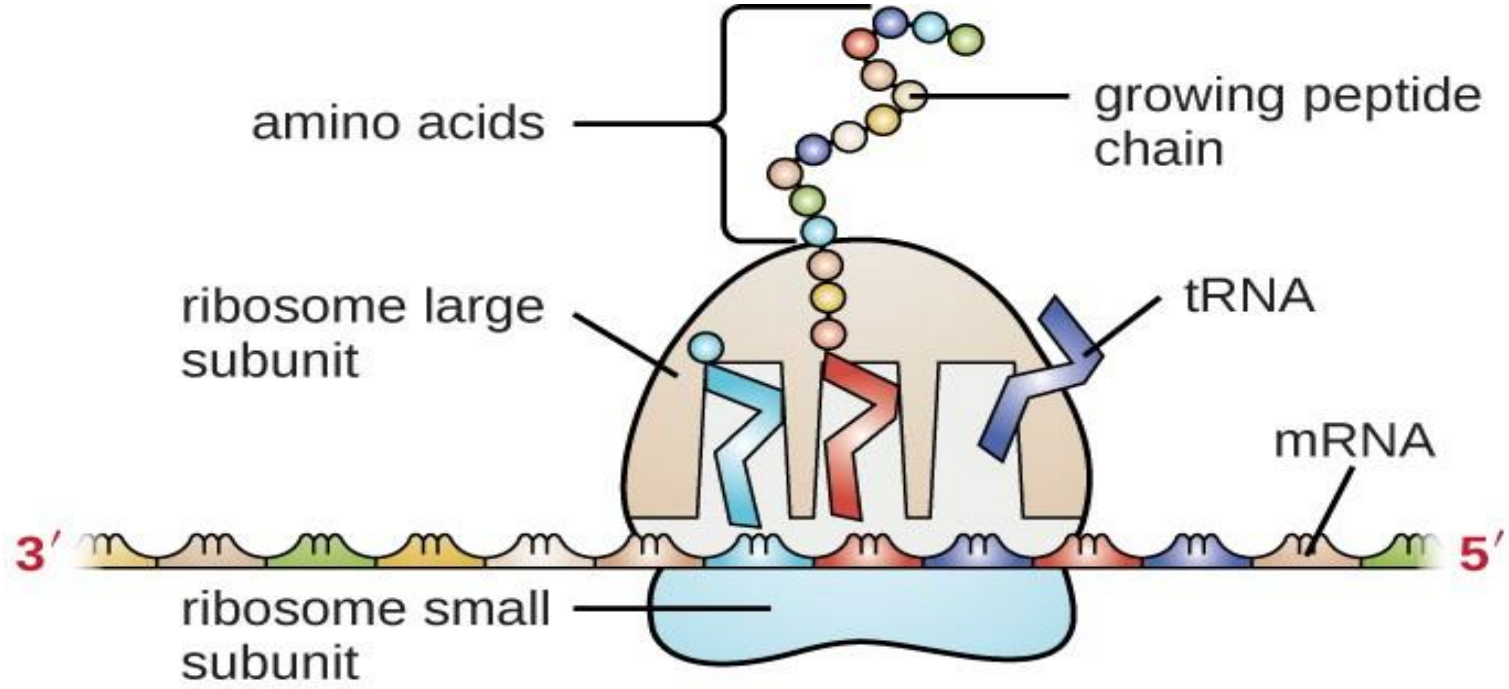




https://www.genome.gov/genetics-glossary/Transfer-RNA

蛋白合成：mRNA、rRNA、tRNA





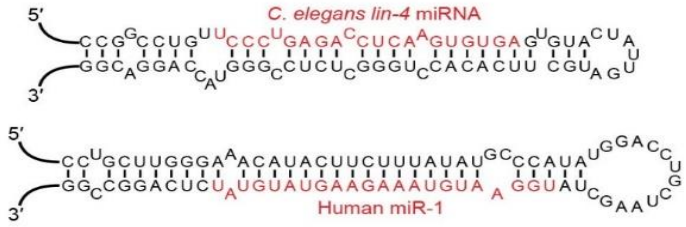
https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-rna/

微小RNA (miRNA)

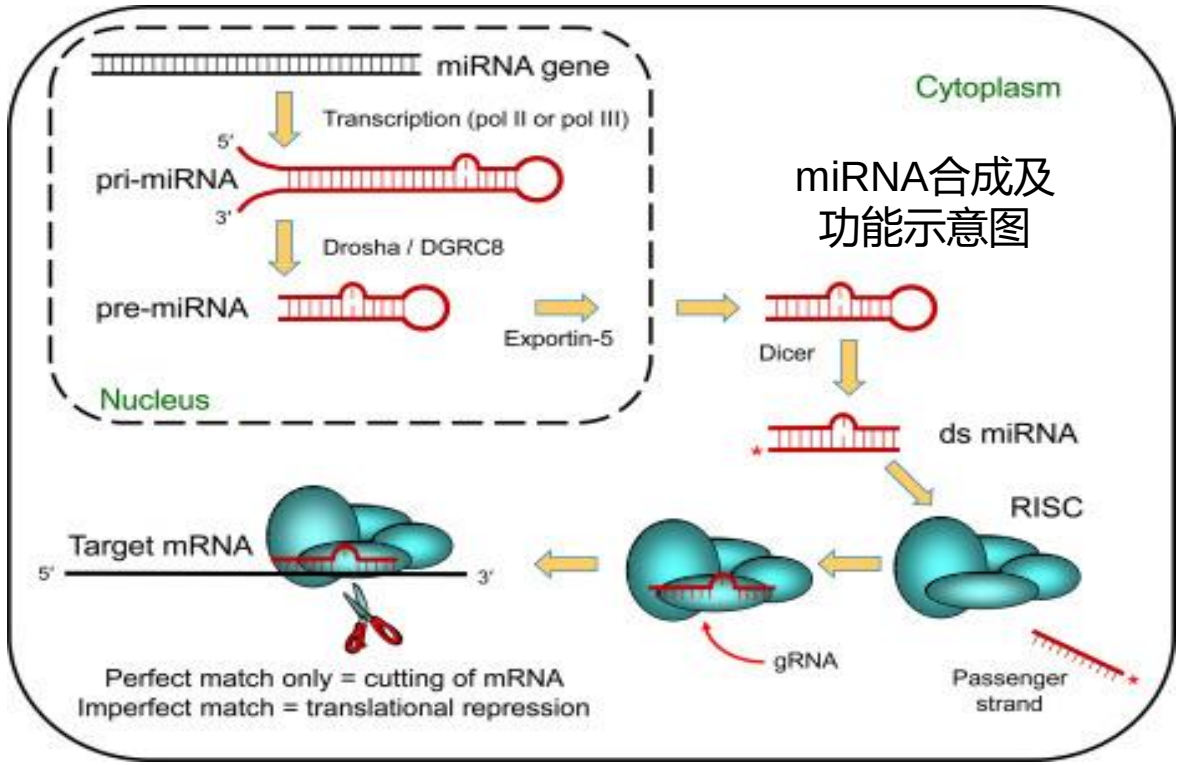
微小RNA（microRNA, miRNA）

内生小非编码RNA分子，~22 nt

- miRNA大多具有茎环结构，其加工过程包括三个阶段：
  - 原始pri-miRNA (300~1000nt)
  - 前体pre-miRNA (70~90nt)
  - 成熟miRNA (20~24nt)
- **miRNA在RNA沉默和基因表达转录后调控中发挥作用**，通过与mRNA互补配对，诱导mRNA链被切割，或促使mRNA缩短其poly(A)尾巴而降低稳性，或降低mRNA翻译成蛋白质的效率



动植物、病毒
红色部分显示成熟miRNA

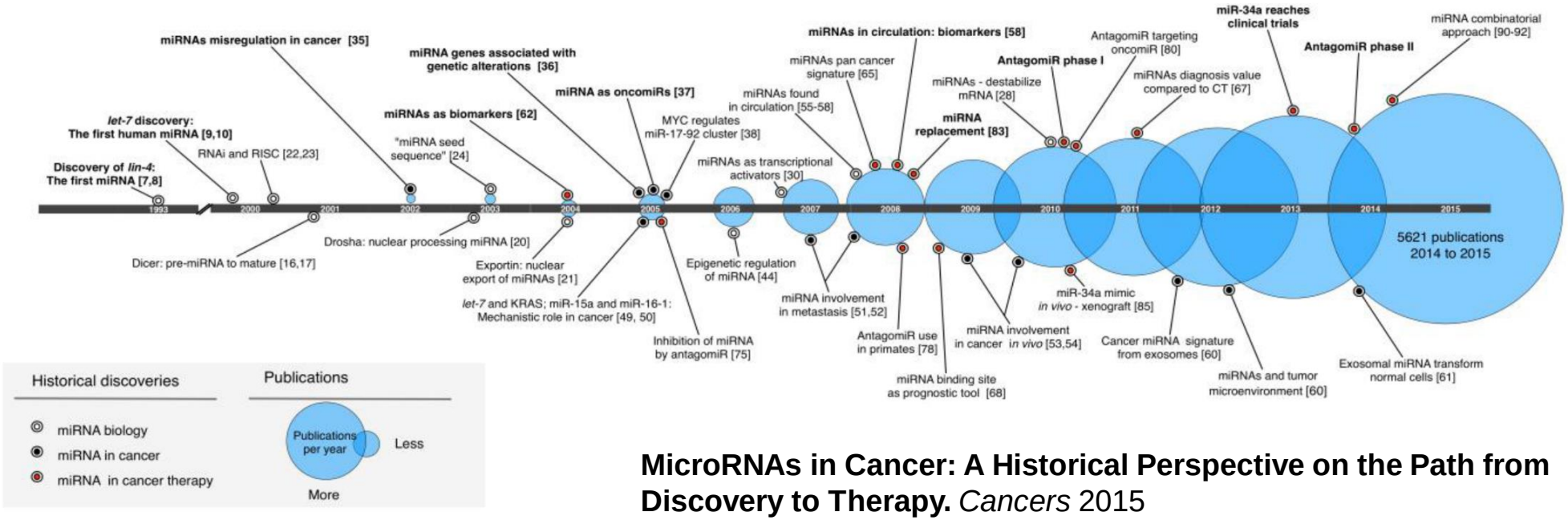


miRNA合成及功能示意图

https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/microrna

miRNA: from biology to cancer

- 在线虫中首次发现lin-4 和let-7（*Cell* 1993）
- 一个miRNA可以有多个靶基因，多个miRNA也可调控同一个基因
- miRNA 高度的保守性与其功能的重要性有着密切的关系
- miRNA在细胞分化、生物发育及疾病发生发展过程发挥重要作用



MicroRNAs in Cancer: A Historical Perspective on the Path from Discovery to Therapy. *Cancers* 2015

长非编码RNA (lncRNA)

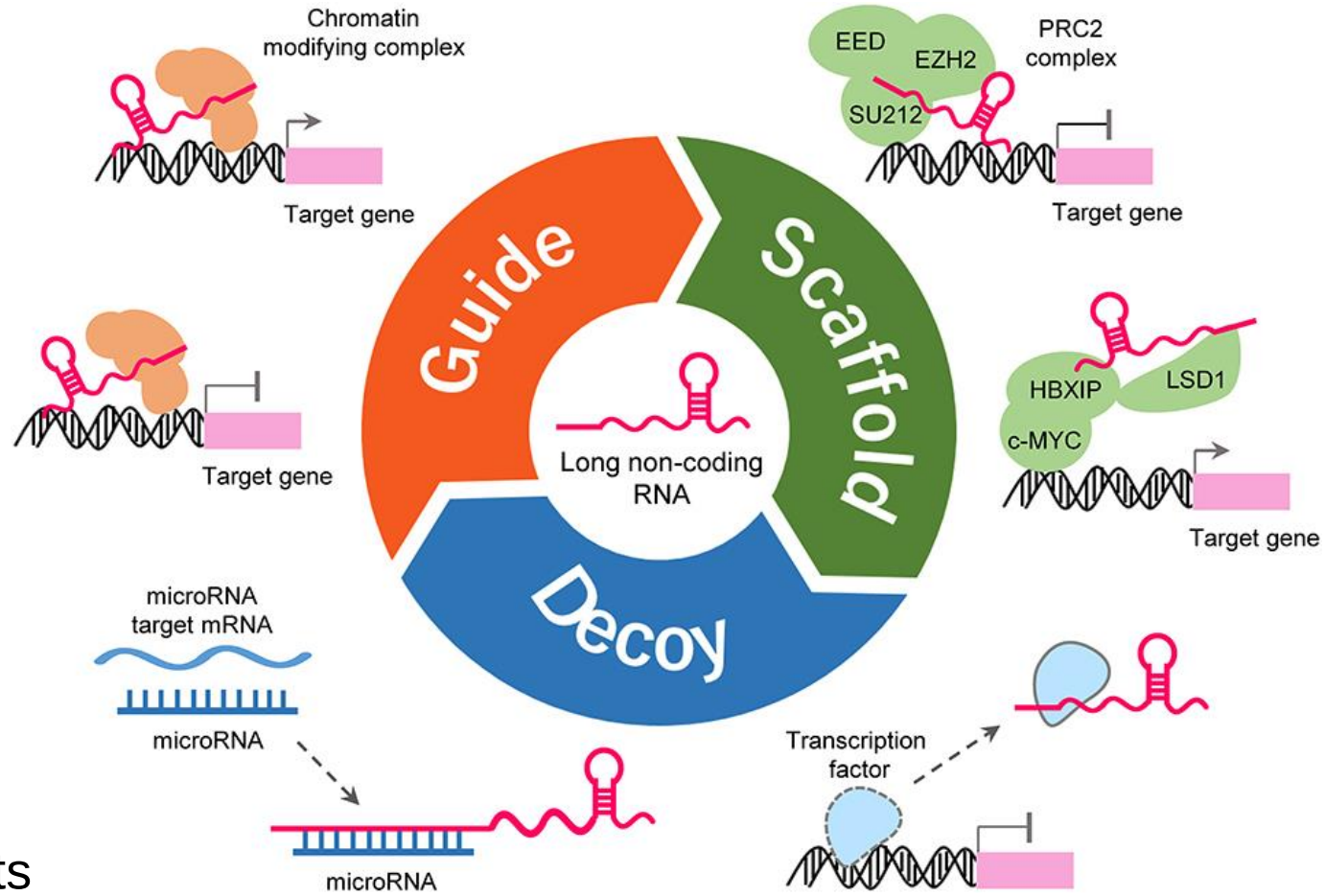
- 长链非编码RNA（Long non-coding RNA, lncRNA）是长度大于200个核苷酸的非编码RNA
- 在表观遗传调控、细胞周期调控和细胞分化调控等众多生命活动中发挥重要作用，是目前研究热点

lncRNA分类

- 反义长lncRNA
- 内含子lncRNA
- 基因间lncRNA
- 启动子相关lncRNA
- 非翻译区lncRNA

95243 lncRNA genes & **323950** transcripts in human (LncBook数据库)

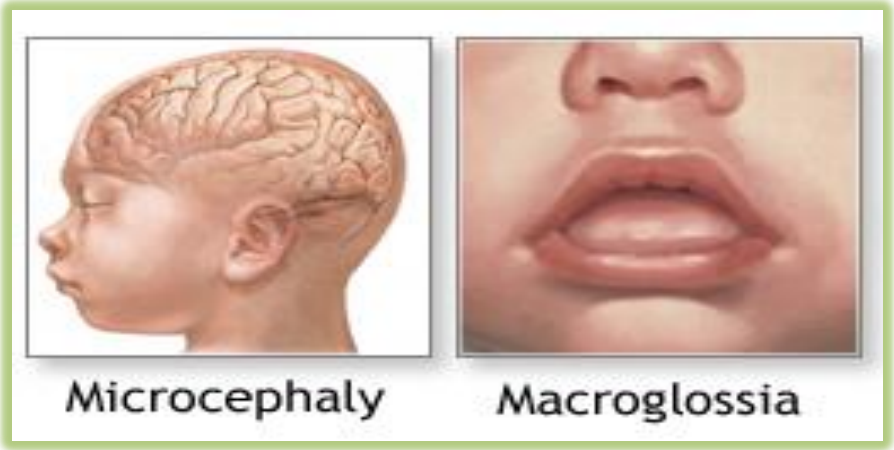
https://ngdc.cncb.ac.cn/lncbook



Front. Genet., 15 October 2018 | https://doi.org/10.3389/fgene.2018.00471

LncRNAs and X inactivation

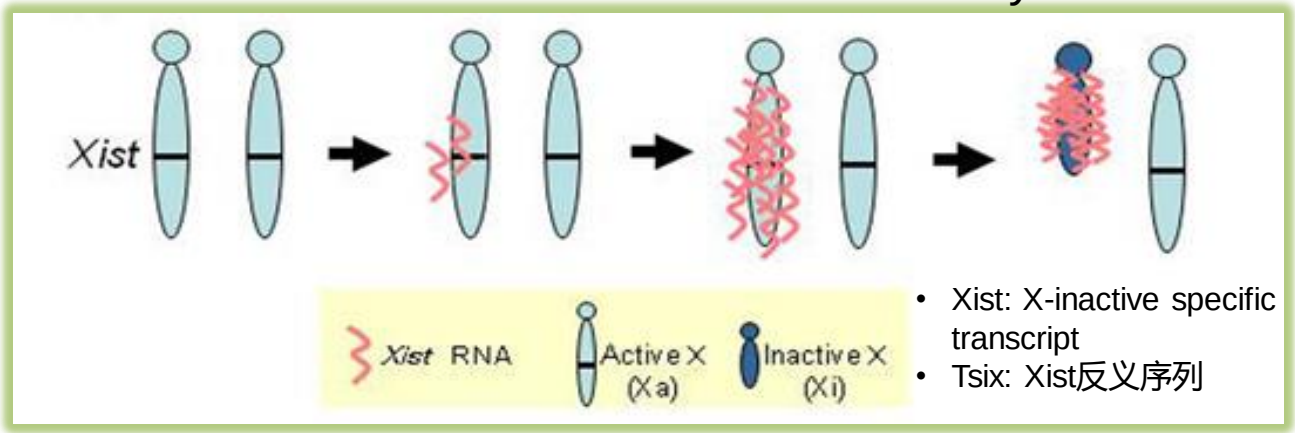
- Have intron/exon structures like protein-coding genes, but lack protein-coding potential
- *H19* (~2kb), only transcribed from the maternal allele, is linked to Beckwith-Wiedemann syndrome (BWS) and cancer (*Mol Cell Biol* 1990)
- *Xist* (~17kb), involved in X inactivation (*Nature* 1991)



小头畸形 巨舌

Beckwith-Wiedemann syndrome

雌性三色猫的斑驳毛色是X失活的表现

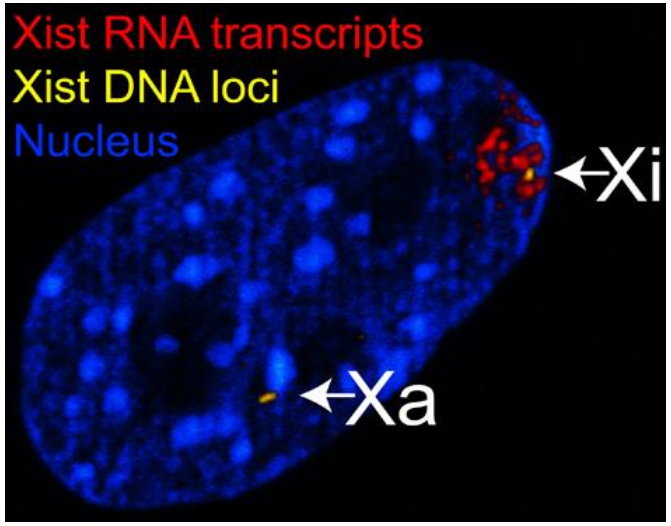


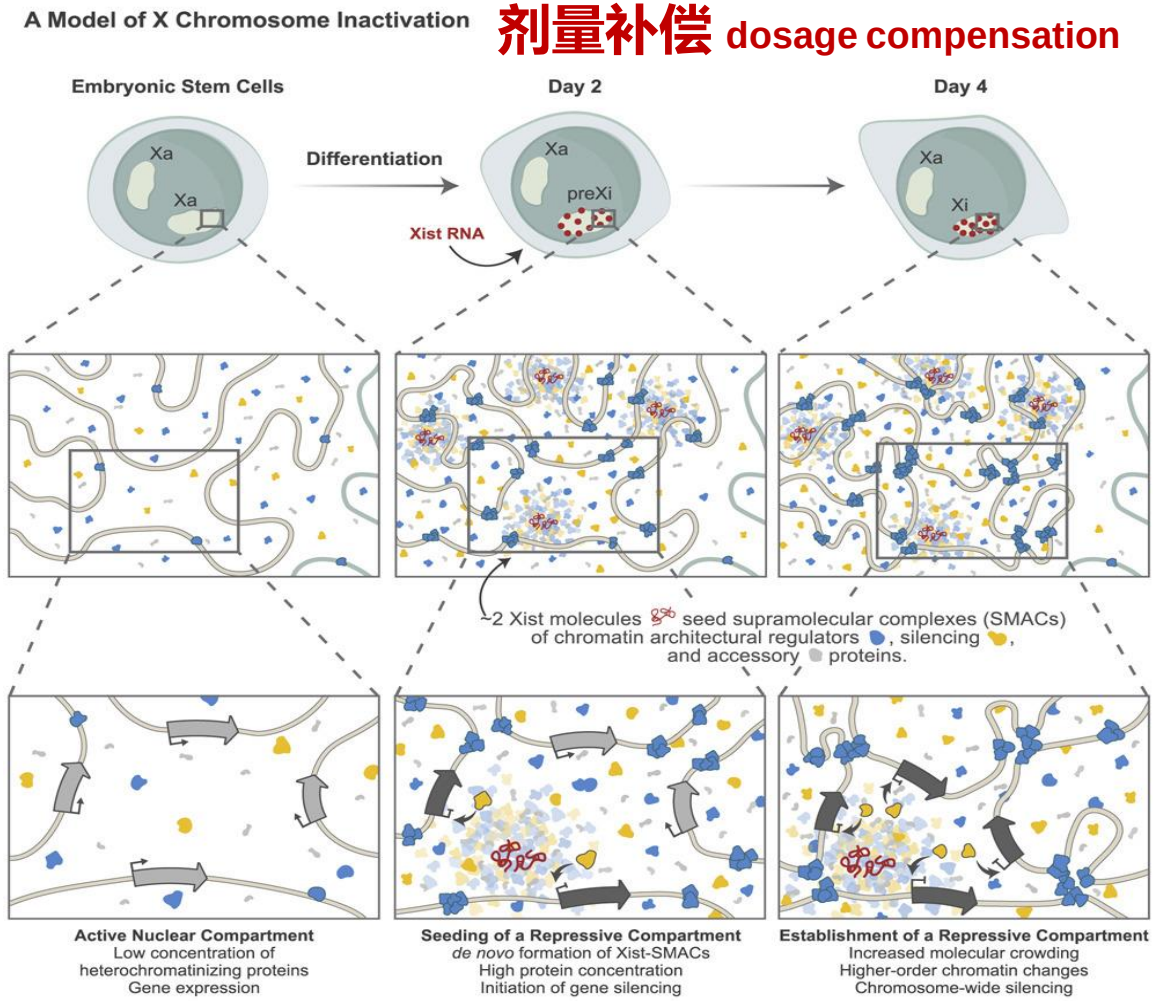
*Xist*包覆失活的X染色体，抑制基因表达使其成为异染色质

X染色体失活

X染色体失活（**X-inactivation**，简称**Xi**，或里昂化Lyonization），是指雌性哺乳类细胞中两条X染色体的其中一条失去活性的现象。X染色体会被包装成**异染色质（heterochromatin）**，进而因功能受抑制而沉默化。

1961年英国遗传学家Mary Lyon（1925—），提出了X染色体失活的概念

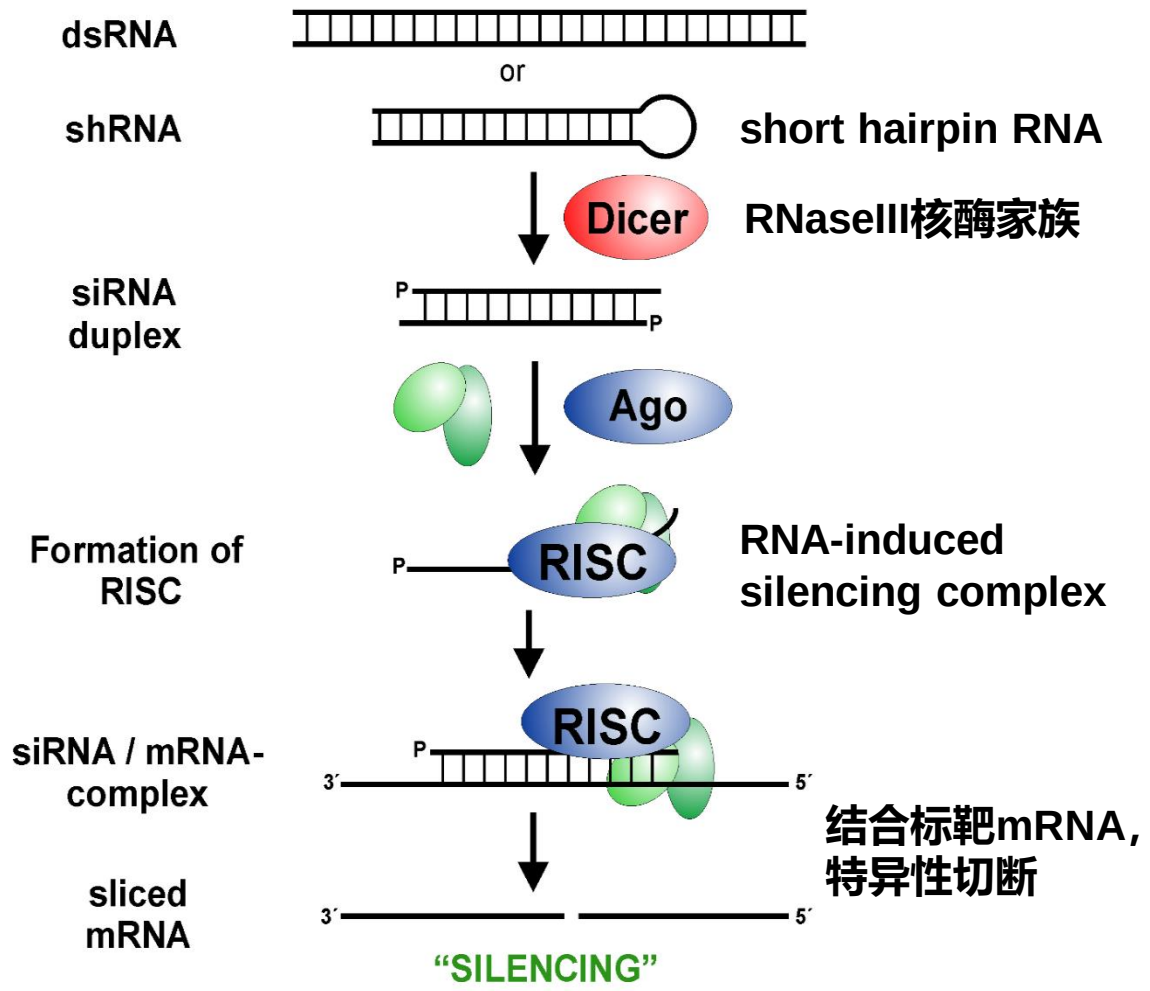




***Xist* nucleates local protein gradients to propagate silencing across the X chromosome**. *Cell* 2021

小干扰RNA (siRNA)

- **小干扰RNA** (Small interference RNA, siRNA)，也称短干扰RNA或沉默RNA，是一类双链非编码RNA分子，长度为20-25个核苷酸对
- 与miRNA相似， siRNA在**RNA干扰** (RNA interference, RNAi)通路中起作用
- **siRNA干扰**具有互补核苷酸序列的特定基因的表达，降解转录后mRNA，从而阻止蛋白翻译过程



https://en.wikipedia.org/wiki/Small_interfering_RNA

RNA干扰（RNAi）

- RNA干扰（RNA interference，RNAi）是指一种分子生物学上由双链RNA诱发的基因沉默现象，其机制是通过阻碍特定基因的转录或翻译来抑制基因表达。
- RNAi机制普遍存在于动植物，尤其是低等生物中，被认为是进化上相对保守的基因表达调控机制
- 1990年Jorgensen研究查尔酮合成酶对花青素合成速度的影响，得到了白色和白紫杂色的矮牵牛花。
- 1992年，Romano和Macino在粉色面包霉菌中发现外源导入基因可抑制具有同源序列的内源基因表达。
- 1998年，Fire和Mello在线虫中发现，双链RNA相比正义或反义RNA显示出了更强的抑制效果（*Nature* 1998）。

矮牵牛花中所观察到的RNA干扰现象

Andrew Fire
(1959-) Stanford

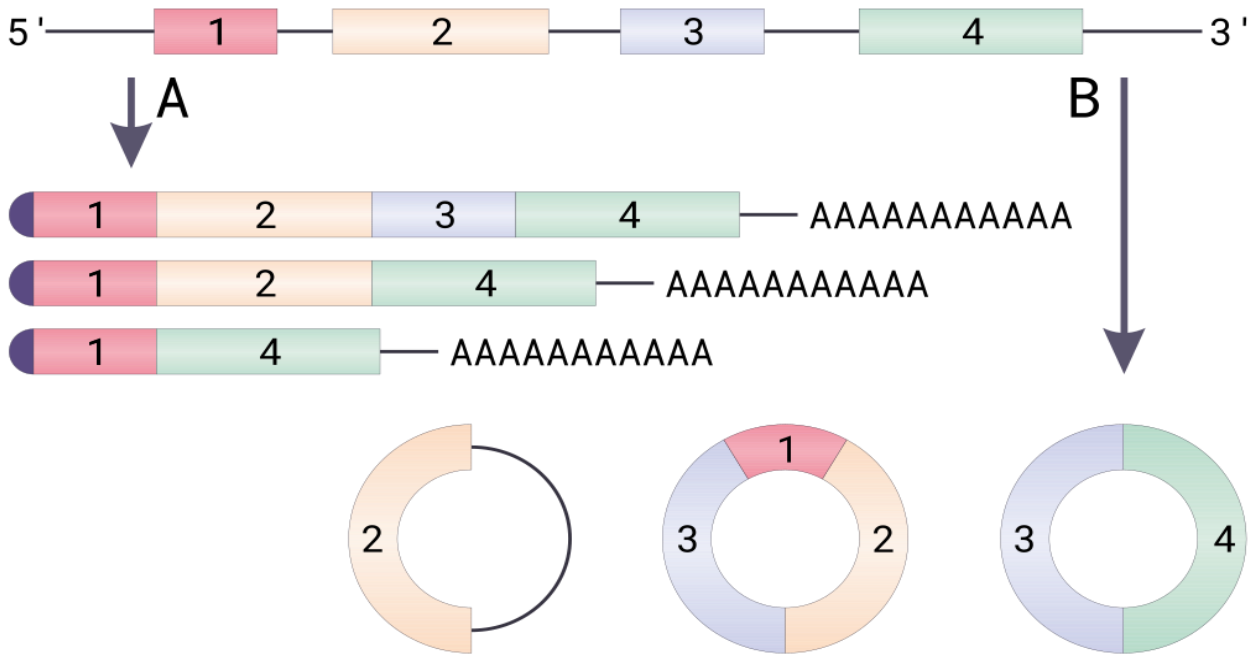
Craig Mello
(1960-) UMass

2006 Nobel Prize in Physiology or Medicine

环形RNA (circRNA)

- **环形RNA**（Circular RNA）是一种单链RNA，分子呈封闭环状结构，表达稳定，不易降解，是RNA领域最新研究热点
- 功能上，circRNA分子富含microRNA结合位点，在细胞中起到**miRNA海绵**（miRNA sponge）的作用，可解除miRNA对其靶基因的抑制作用，称为竞争性内源RNA（ceRNA）机制
- 人类：768986 circRNAs（circAtlas数据库）

https://ngdc.cncb.ac.cn/circatlas/

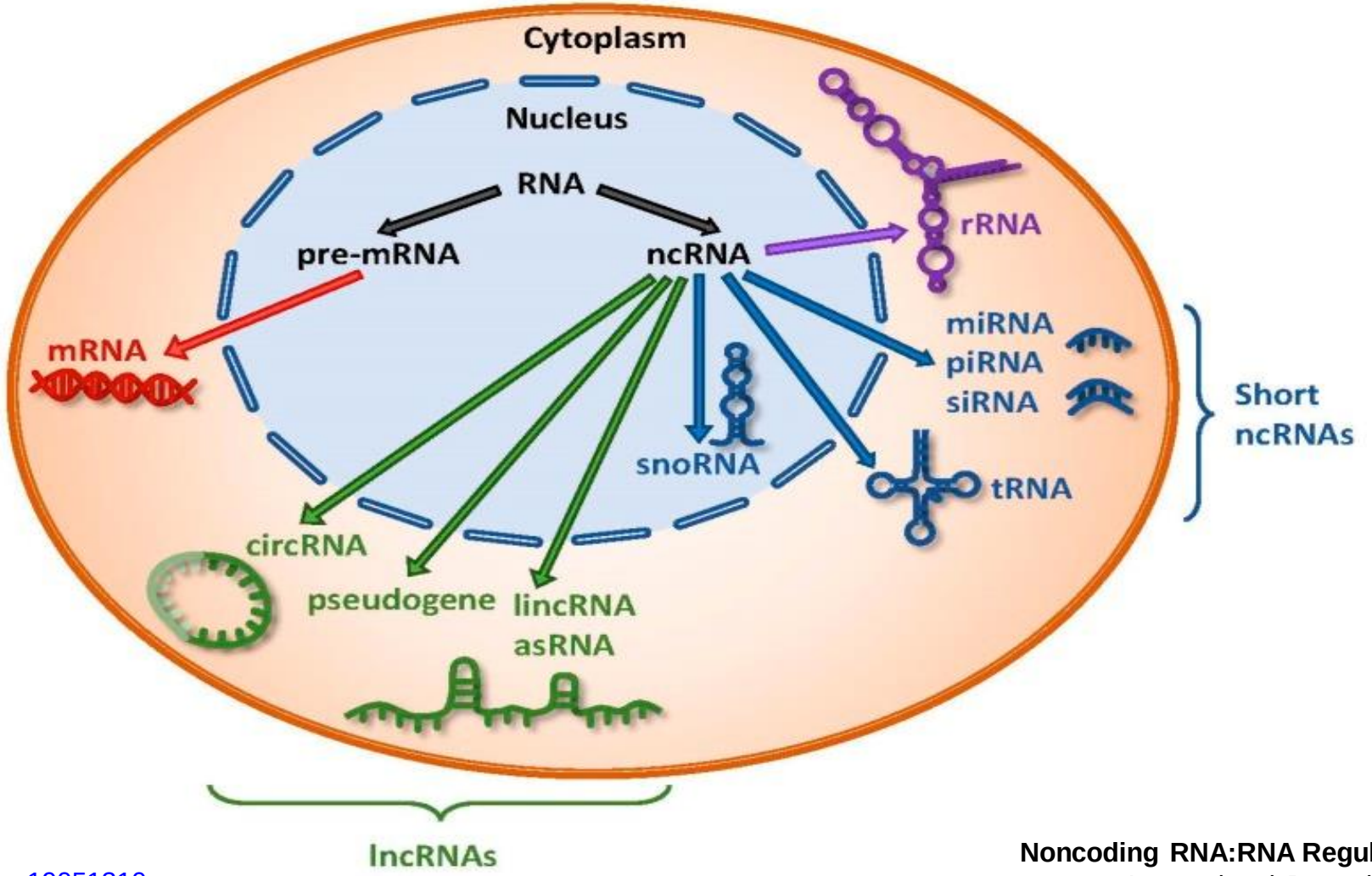


Circular RNA biogenesis

A. mRNA splicing, with alternative splice variants. All mRNAs have cap and polyA tail
B. CircRNA formation via backsplicing

https://en.wikipedia.org/wiki/Circular_RNA

Coding and noncoding RNAs



https://doi.org/10.3390/ijms19051310

Noncoding RNA:RNA Regulatory Networks in Cancer.
International Journal of Molecular Sciences 2018

RNA生命动态

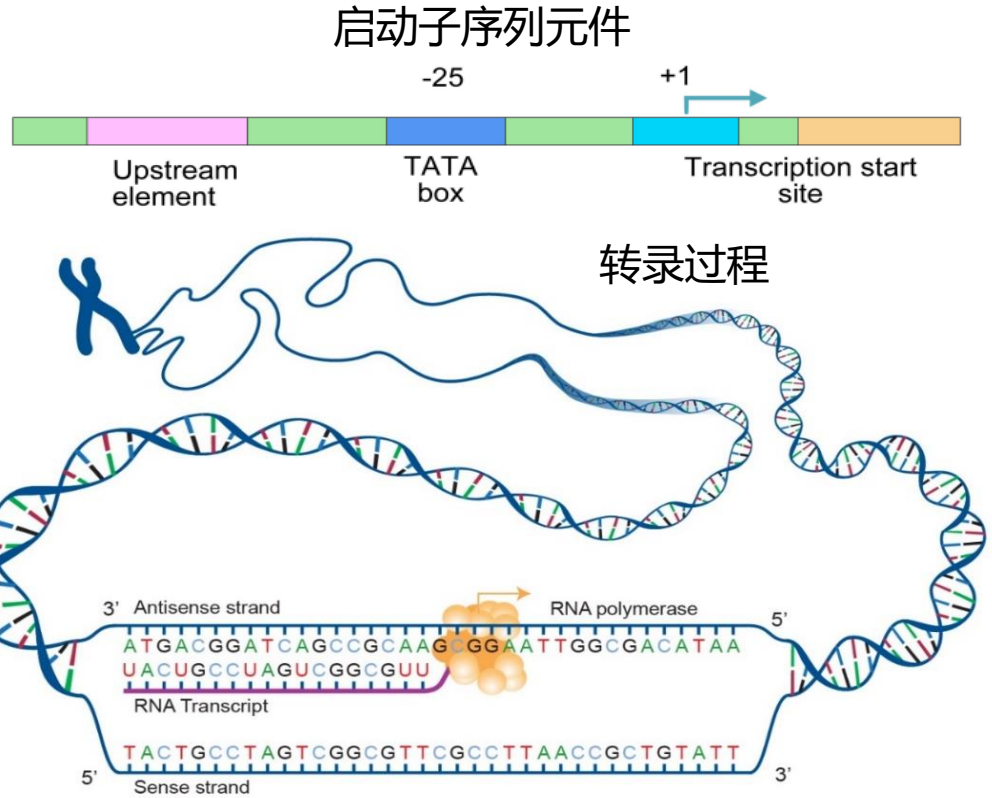
RNA生命动态

- 合成
- 转录调控
- 加工修饰
- 降解
- 干扰
- 交互作用

RNA合成（转录Transcription）

转录：在RNA聚合酶（RNA Polymerase）的催化作用下，遗传信息由DNA到RNA的过程，一般分为**启动、延伸、终止**三个阶段

- **启动：**转录起始于RNA聚合酶和通用转录因子与启动子的结合
- **延伸：**转录启动后，DNA双螺旋链解开，RNA聚合酶以其中一条DNA链作为模板链。RNA聚合酶以**3’-5’**端的方向读取模板链，并以**5’-3’端**的方向合成与之反向平行互补的RNA链，也称转录本（**Transcript**）
- **终止：**转录终止于DNA非编码序列末端附近的终止子处



https://en.wikipedia.org/wiki/Transcription_(biology)

https://www.genome.gov/genetics-glossary/Transcription

转录调控

- **转录调控（Transcriptional regulation）**：指通过改变转录效率从而调控RNA转录本表达水平的过程。转录调控可以控制基因的**时空动态表达**。原核与真核生物具有不同的转录调控机制。
- **顺式作用元件（*cis*-regulatory elements, CRE）**： 位于基因旁侧序列中能影响基因表达的DNA序列，可影响基因转录活性。
- **反式作用元件（*trans*-regulatory elements, TRE）**：转录模板上游基因编码的一类蛋白调节因子，又称转录因子(Transcription Factors)，通过与特异的顺式作用元件相互作用反式激活基因转录，分为两类：①**通用转录因子**：所有mRNA转录启动共有。②**特异转录因子**：个别基因转录所必需，决定基因的时空特异性表达。

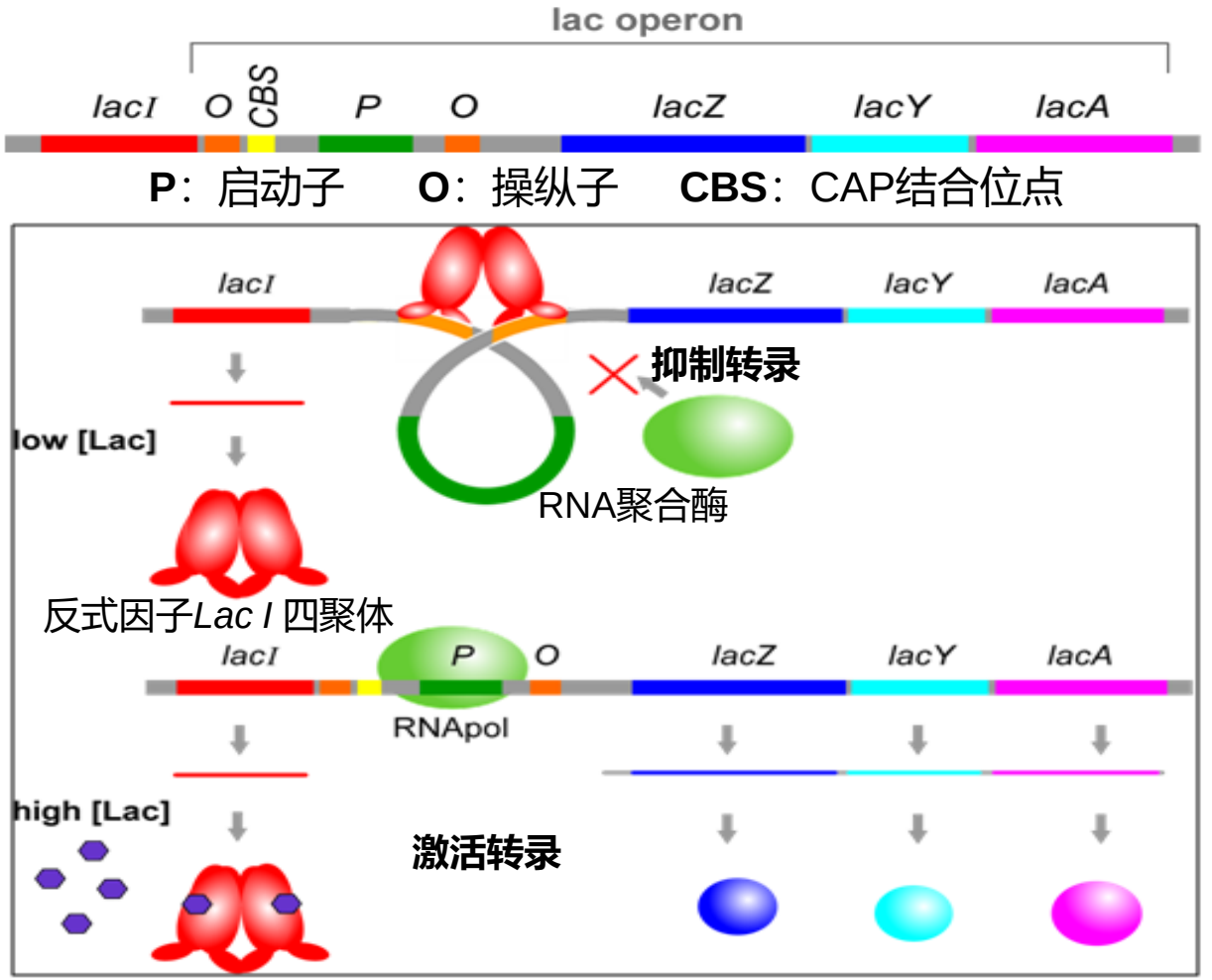
CRE are present on the same molecule of DNA as the gene they regulate, whereas TRE can regulate genes distant from the gene from which they were transcribed.

原核生物转录调控

- **原核生物转录调控规律**
  - 营养和环境因素影响转录调控
  - 多个基因可位于同一转录单元上，称为**操纵子（operon）**
  - 同一操纵子中基因受到相同的转录调控，翻译过程相互独立
- **原核生物转录调控机制类型**
  - 代谢产物对基因活性的调节
  - 弱化子对基因活性的调节
  - 降解物对基因活性的调节
  - 细菌的应激反应

https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Online_Open_Genetics_(Nickle_and_Barrette-Ng)/12%3A_Regulation_of_Gene_Expression/12.1%3A_The_lac_Operon

乳糖操纵子基本结构及调控过程

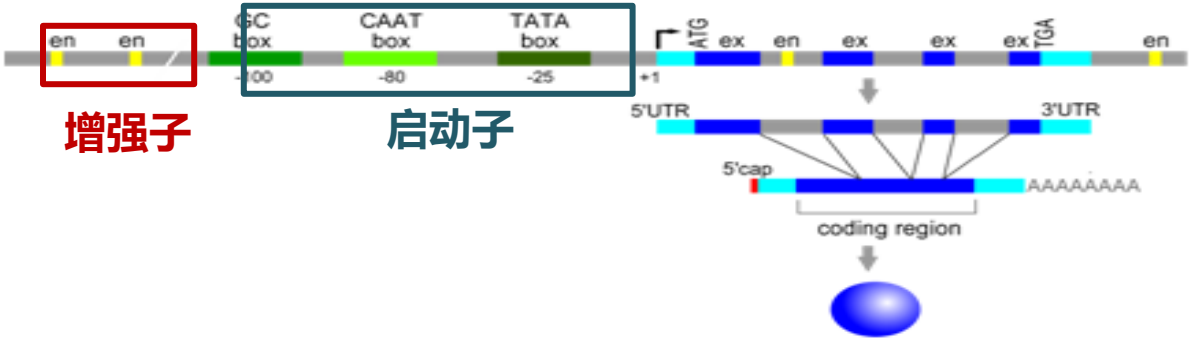


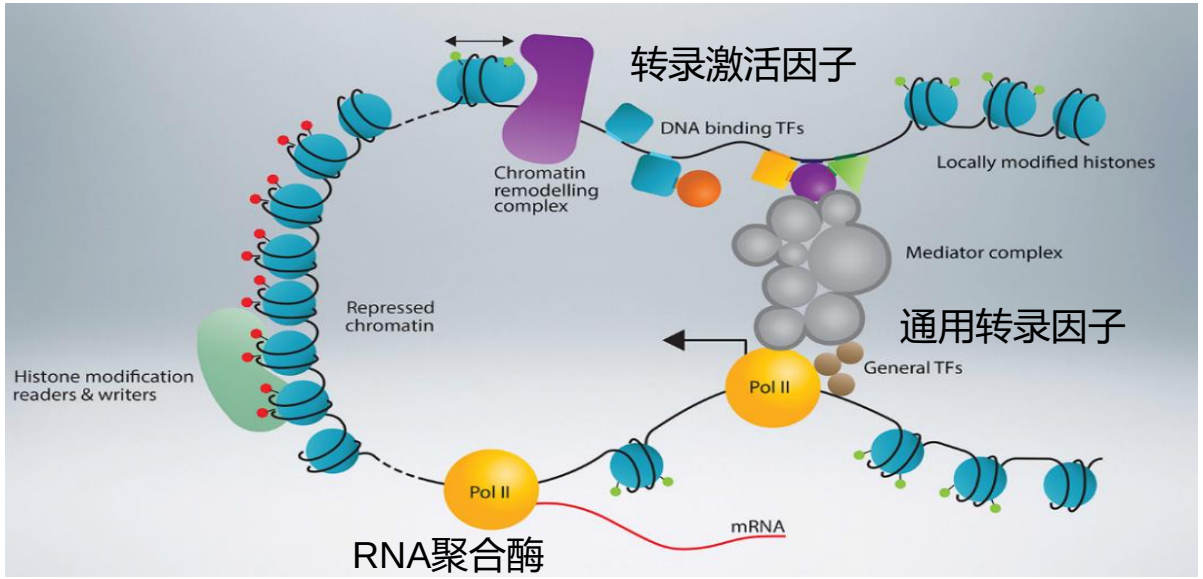
真核生物转录调控

真核生物转录调控特点

- 几种调控水平相互作用：**局部调控**特定细胞单个基因的打开或关闭，**全局调控**维持染色质水平的基因表达模式
- 转录调控元件包括**顺式元件**及**反式因子**，作用方式更为复杂
- **通用转录因子**和**RNA聚合酶**组成基础转录复合体，**转录激活因子**和**阻遏因子**与其作用激活或抑制转录
- **增强子**远离其所控制基因，与转录激活因子结合增强基因转录。**沉默子**作为负调控序列，与阻遏因子结合，抑制基因转录

真核转录调控示意图

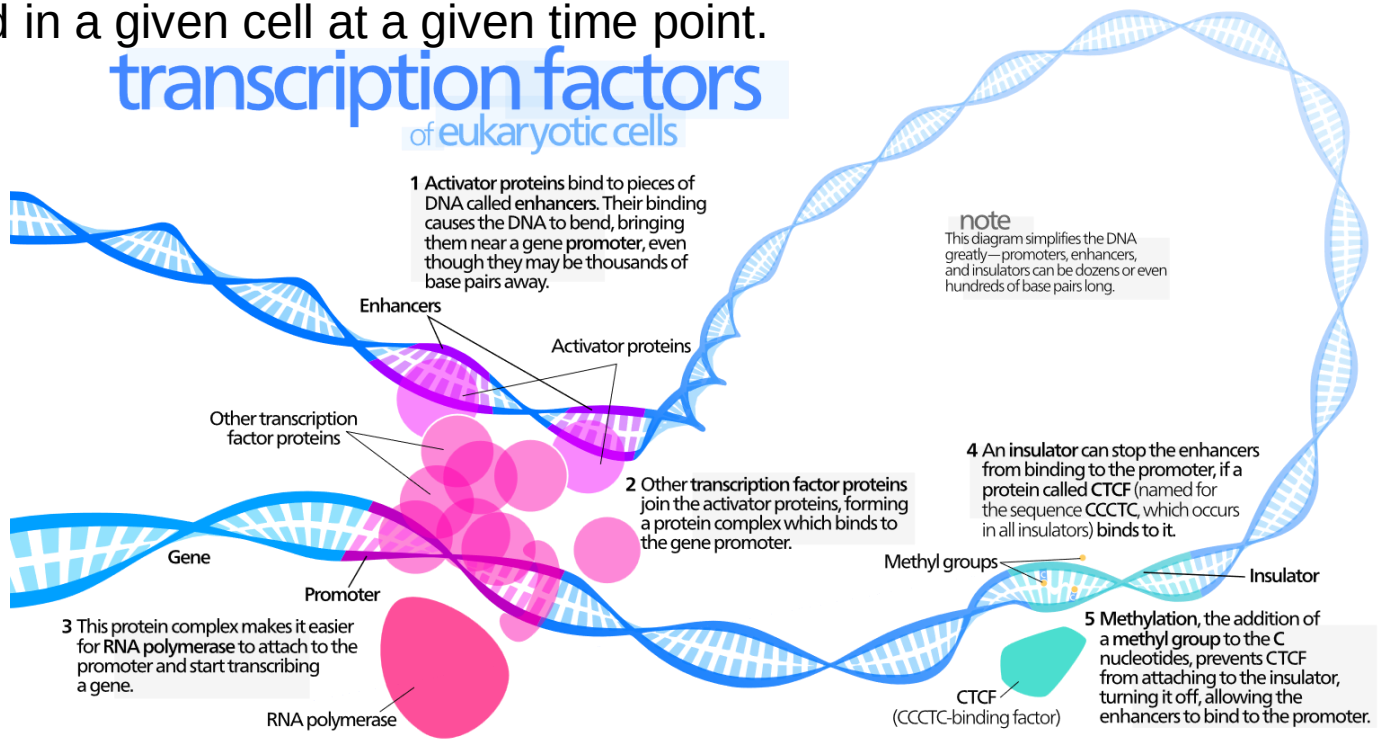




http://bruneaulab.gladstone.org/transcriptional-regulation

转录因子

- **Transcription Factor (TF)**: a type of protein that binds to a specific DNA sequence and regulates gene transcription—turn on & off—that gene are **expressed at the right time and in the right amount**.
  - There are **1659** TFs and **1024** TF cofactors in the human genome (AnimalTFDB as of Apr 2024).
  - Only a subset of TFs may be expressed in a given cell at a given time point.
  - TFs are essential for gene expression regulation in all living organisms.
  - TFs are of clinical significance because TF mutations can cause specific diseases, and medications can be potentially targeted toward them.
- **Transcription Factor Binding Sites (TFBS)**: 4-20bp short DNA sequences where TFs bind.
  - There are millions of TF binding sites in the human genome.



https://en.wikipedia.org/wiki/Transcription_factor

转录后加工与修饰

真核细胞中，将初级RNA转录本转化为成熟RNA的加工过程

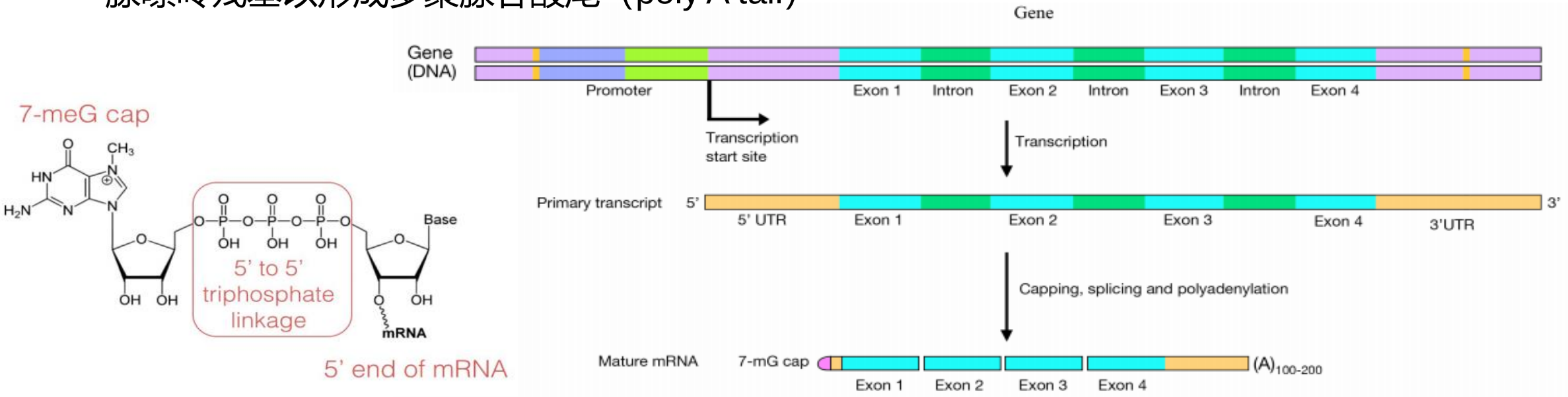
加工修饰类型

- 5'端加帽 (5' Capping)
- 3’端加尾 (Cleavage and polyadenylation)
- 剪接 (Splicing)
- 化学修饰 (Chemical modification)
- 编辑 (Editing)

mRNA加工

mRNA加工 (mRNA Processing)：将初级mRNA转化为成熟RNA的加工过程，发生在细胞核中mRNA翻译之前，包括**5'端加帽、3'端多聚腺苷酸化、RNA剪接**

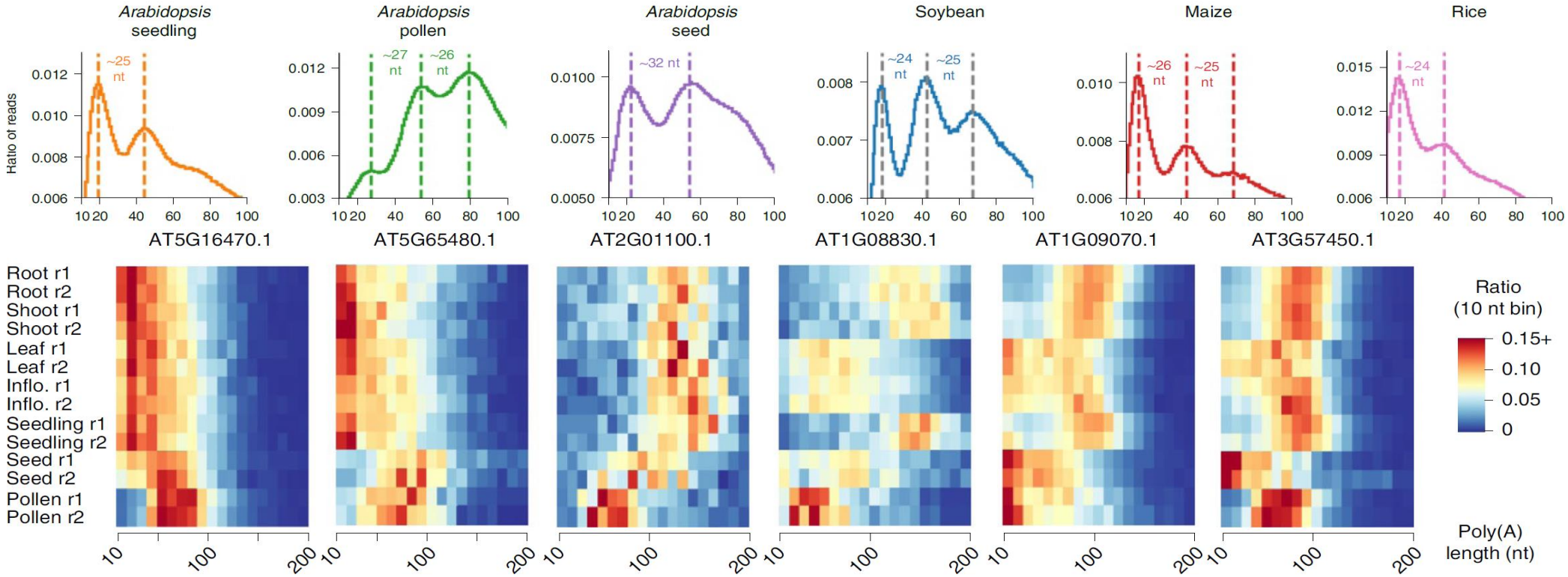
- **5'端加帽 (5' Capping)**：在mRNA前体5'端加上7-甲基鸟苷（m7G）
- **3'端加尾 (Cleavage and polyadenylation)**：对mRNA前体3’端进行切割，随后加上约200个腺嘌呤残基以形成多聚腺苷酸尾（poly A tail）



https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_and_Easy_(Ahern_and_Rajagopal)/05%3A_Flow_of_Genetic_Information/5.05%3A_RNA_Processing

Regulation of poly(A) tail length

- Poly(A) tail length plays an essential role in **regulating mRNA metabolism**.
- Poly(A) tail length is regulated in a **gene-specific manner—mRNAs with short half-lives** in general have **long poly(A) tails**, while **mRNAs with long half-lives** are featured with relatively **short poly(A) tails**.
- Across species, poly(A) tails in the **nucleus** are almost **twice** as long as in the **cytoplasm**.

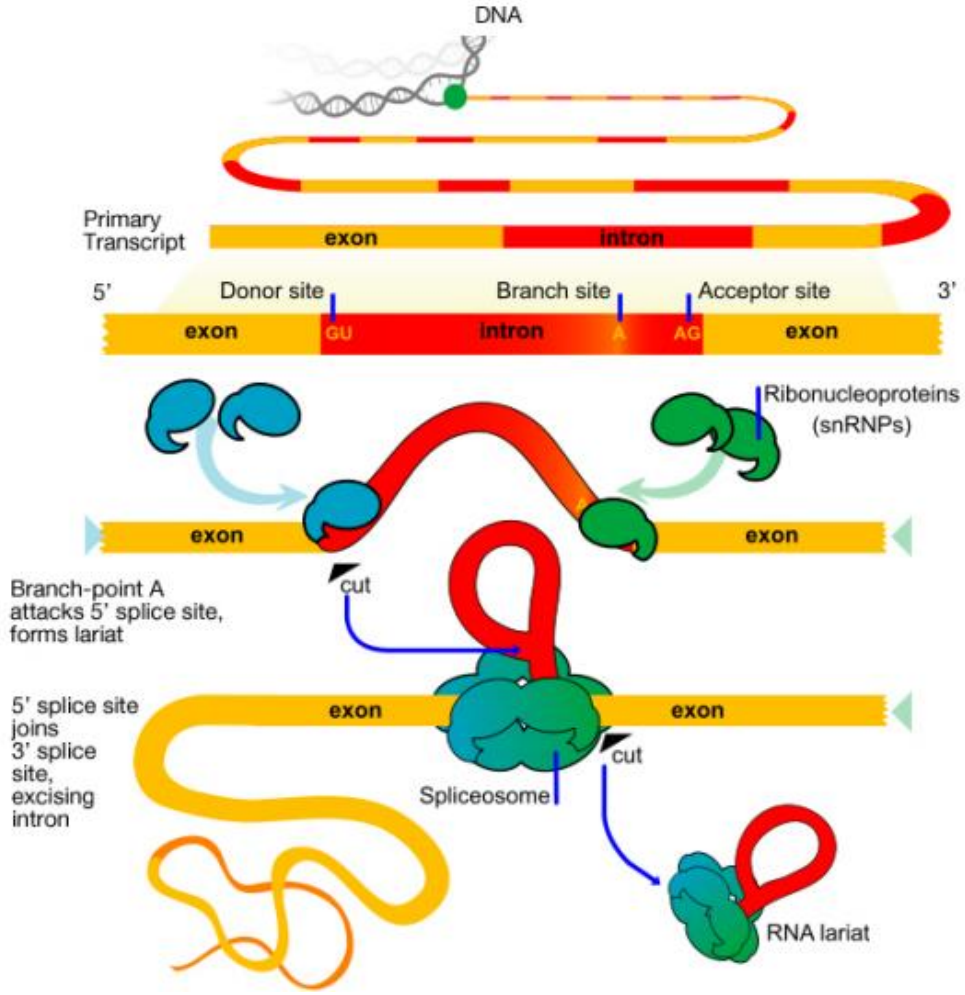


An atlas of plant full-length RNA reveals tissue-specific and monocots–dicots conserved regulation of poly(A) tail length. *Nat. Plants* 2022

RNA剪接

RNA剪接（RNA Splicing）：将非编码内含子从RNA前体上除去，将外显子拼接在一起的加工过程

- **剪接体（Spliceosome）**：由识别位于mRNA前体序列剪接位点的蛋白质和小核RNA(snRNA)聚合而成
- **RNA剪接反应：**由剪接体蛋白复合物催化，可发生在RNA前体合成加帽后及转录过程中
- **可变剪接（Alternative Splicing）**：mRNA前体以多种方式被剪接，产生编码不同蛋白质序列的成熟mRNA

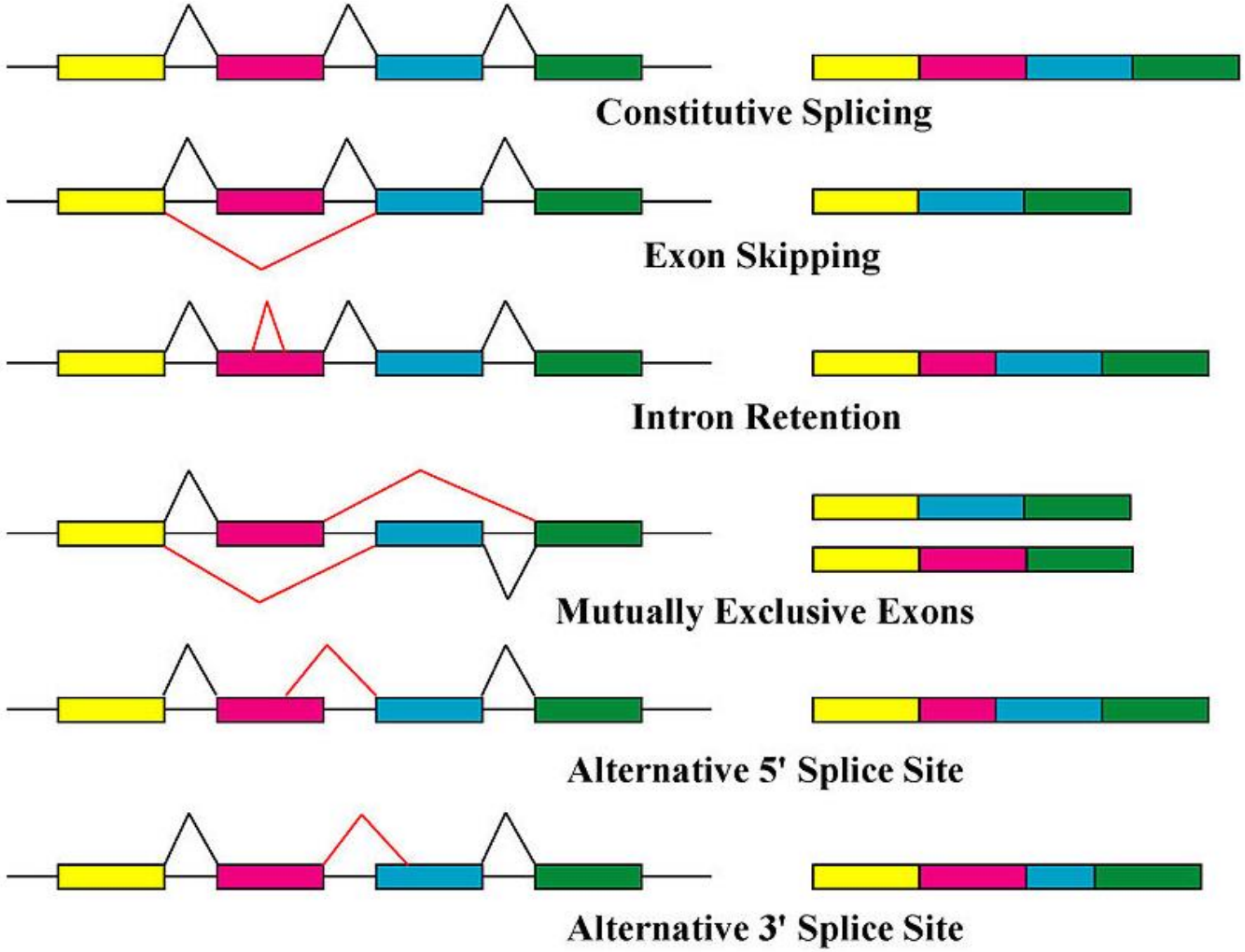


https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_and_Easy_(Ahern_and_Rajagopal)/05%3A_Flow_of_Genetic_Information/5.05%3A_RNA_Processing

单基因转录模式的多样性

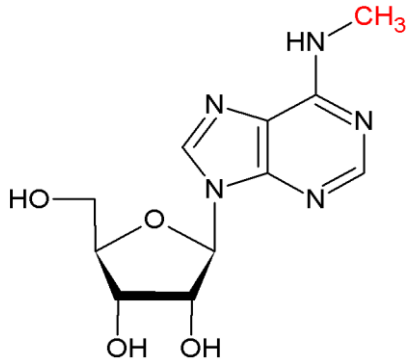
可变剪接（alternative splicing，AS，选择性剪接）：在 mRNA 前体到成熟 mRNA 的过程当中，不同的剪切方式使得**同一个基因可以产生多个不同的成熟 mRNA**，最终产生不同的蛋白质。

在真核生物体内广泛存在，人类基因组中95%的基因都存在可变剪接现象，导致**转录本和蛋白质结构与功能的多态性**，是一种**重要的转录调控机制。**

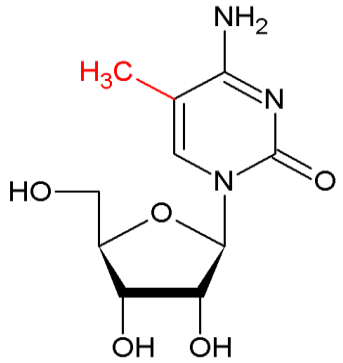


RNA化学修饰

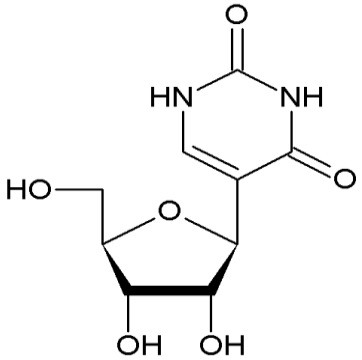
- RNA化学修饰（**Chemical Modification**）：RNA分子合成后发生的化学组分变化，具有**改变RNA结构、功能和稳定性等**作用
- 目前已知的RNA化学修饰类型包含**100多种**，代表示例： m^6A, m^5C, Ψ, m^7G



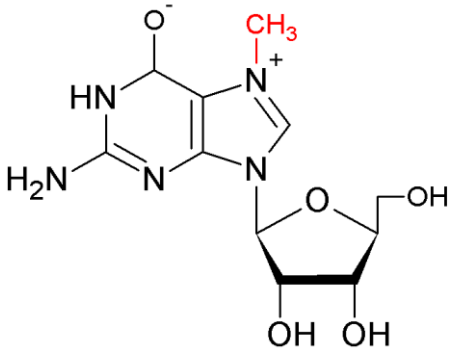
m^6A



m^5C

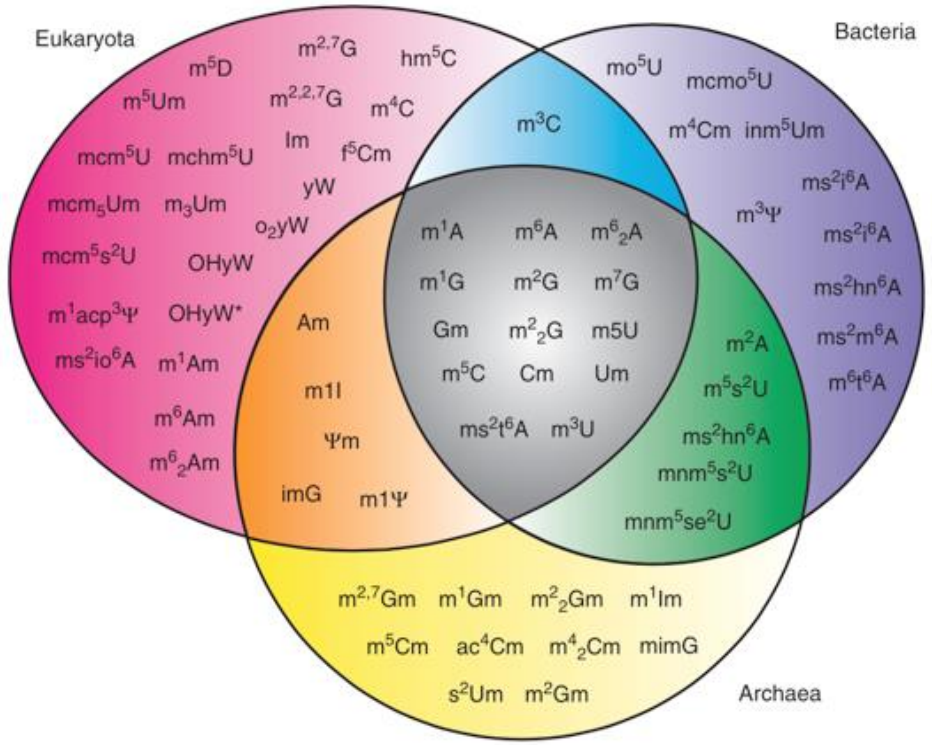


ψ (pseudouridine)



m^7G

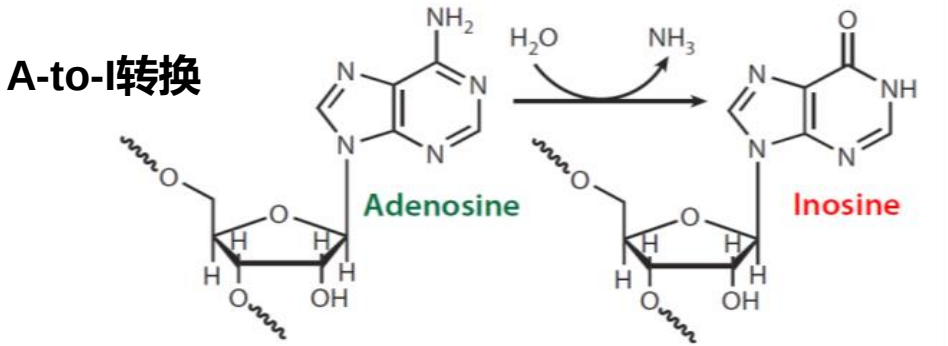
https://mods.rna.albany.edu

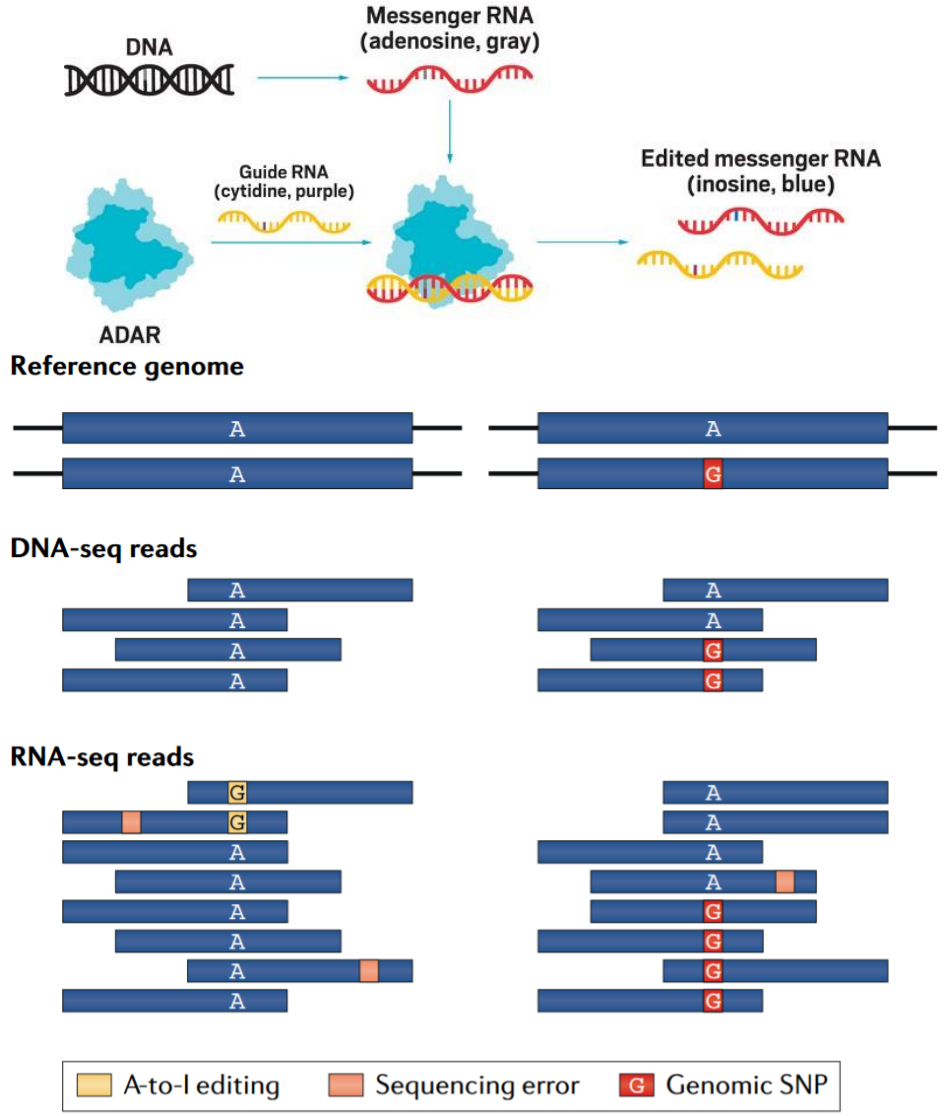


Motorin Y. and Helm M.. Wiley Interdiscip Rev *RNA*, 2011

RNA编辑

- RNA编辑（RNA editing）对RNA分子内特定核苷酸序列进行修饰的分子过程，包括核苷酸的插入、删除和碱基替换
- 存在于真核生物、原核生物、病毒的一些tRNA、rRNA、mRNA或miRNA分子中
- 发生在细胞核和胞浆内，也发生在线粒体和质体内
- 由对应编辑体（Editosome）介导

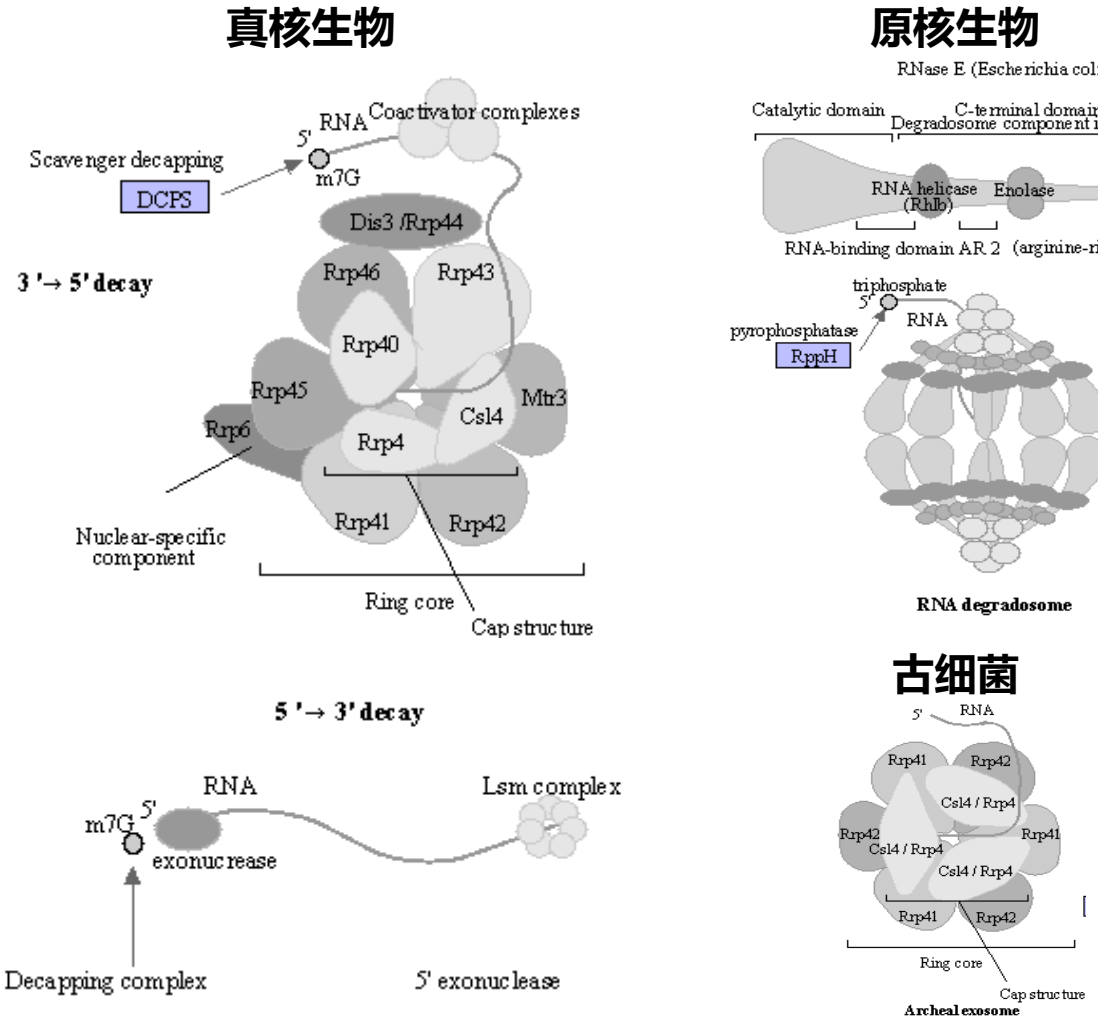




A-to-I RNA editing — immune protector and transcriptome diversifier. *Nat Rev Genet* 2018
https://doi.org/10.1038/s41576-018-0006-1

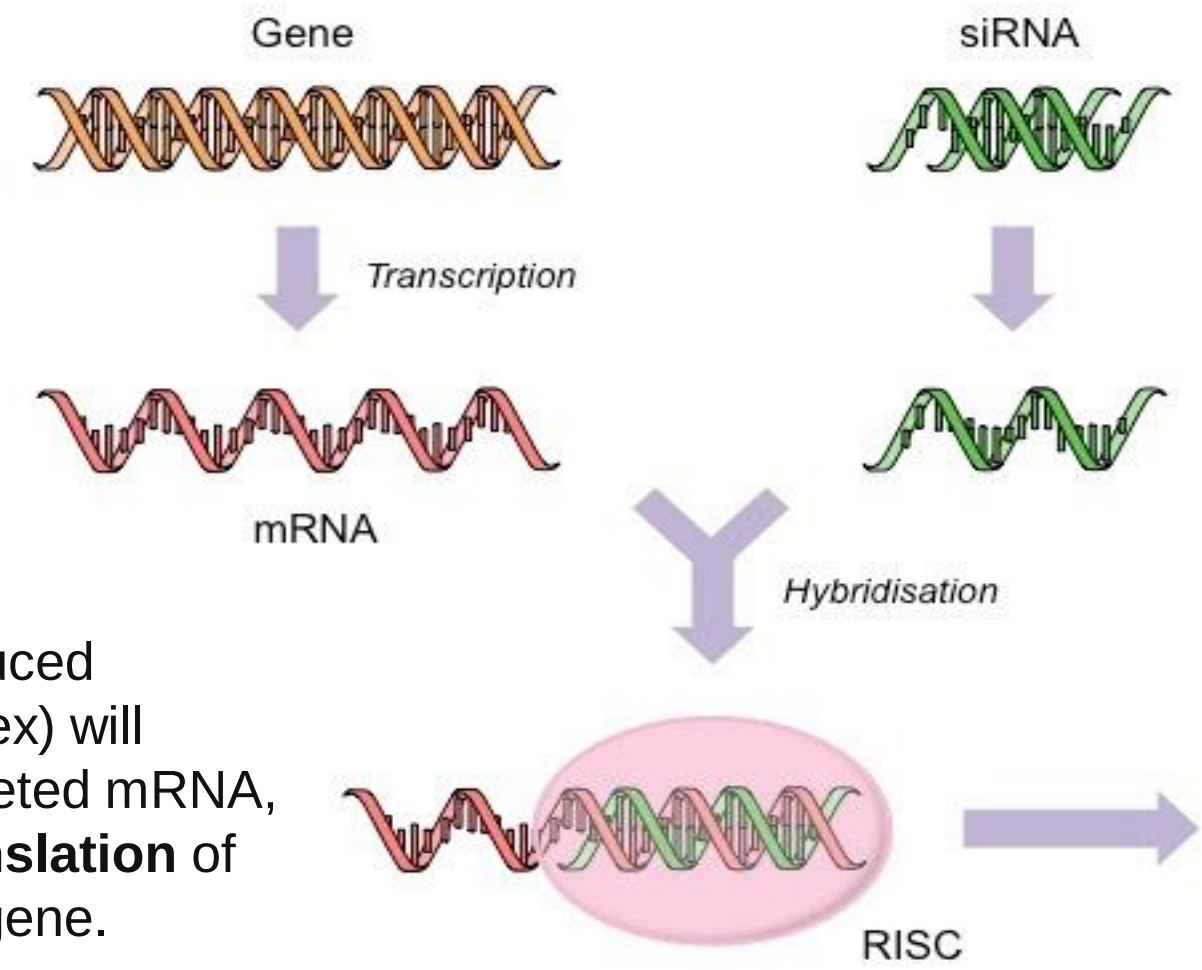
RNA降解

- RNA分子的质量控制和**降解（RNA Decay）**对遗传信息的传递起着至关重要的作用
- mRNA具有不同的稳定性，其有限寿命使细胞快速改变蛋白质合成以响应变化需求
- 多种机制可导致mRNA降解。真核生物包括两条主要降解途径，皆由mRNA的**poly(A)尾缩短**引起。**3’至5’途径**，由外泌体多亚基复合体起关键作用。**5’至3’途径**，由5’至3’的外切酶降解转录本
- 细菌**核糖核酸内切酶E**为RNA降解关键酶
- 古细菌由**外泌体**介导RNA降解



https://www.genome.jp/keg-bin/show_pathway?map=ko03018&show_description=show

RNA Interference with siRNA



siRNA is a **double-stranded RNA molecule** that is roughly 20 – 25 base pairs in length.

siRNA **interferes with the expression of genes** by causing the mRNA transcripts to be broken prior to translation.

RISC (RNA-induced silencing complex) will destroy the targeted mRNA, **preventing translation** of the associated gene.

RNA Interference (RNAi)-Based Therapeutics: Delivering on the Promise?
Annual Review of Pharmacology and Toxicology 2016

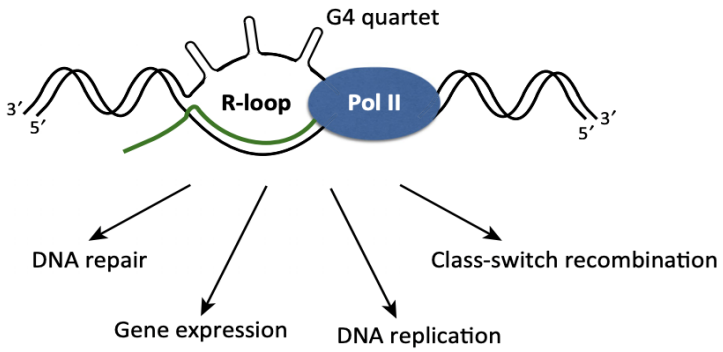
应用

功能鉴定
药物标靶
疾病治疗

RNA交互网络

RNA molecules play crucial roles in various biological processes. Their regulation and function are mediated by **interacting with other molecules**.

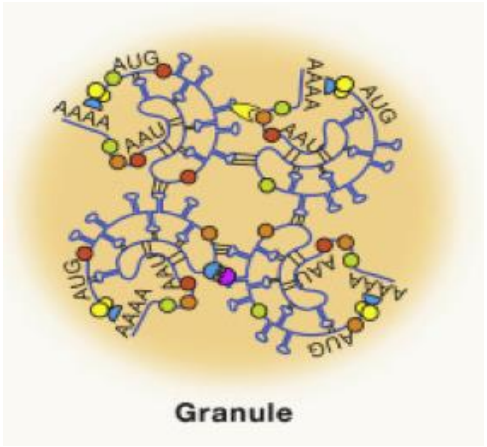
RNA-DNA



An R-loop is a **three-stranded** nucleic acid structure, composed of a **DNA:RNA hybrid** and the associated non-template **single-stranded DNA**.

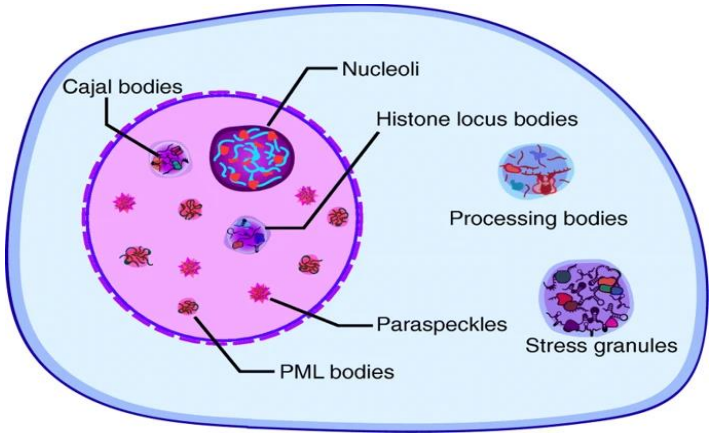
Breaking bad: R-loops and genome integrity.
Trends in Cell Biology 2015

RNA-RNA



RNA-RNA interactions are important in **many basic cellular activities** including transcription, RNA processing, localization, and translation.

RNA-Protein



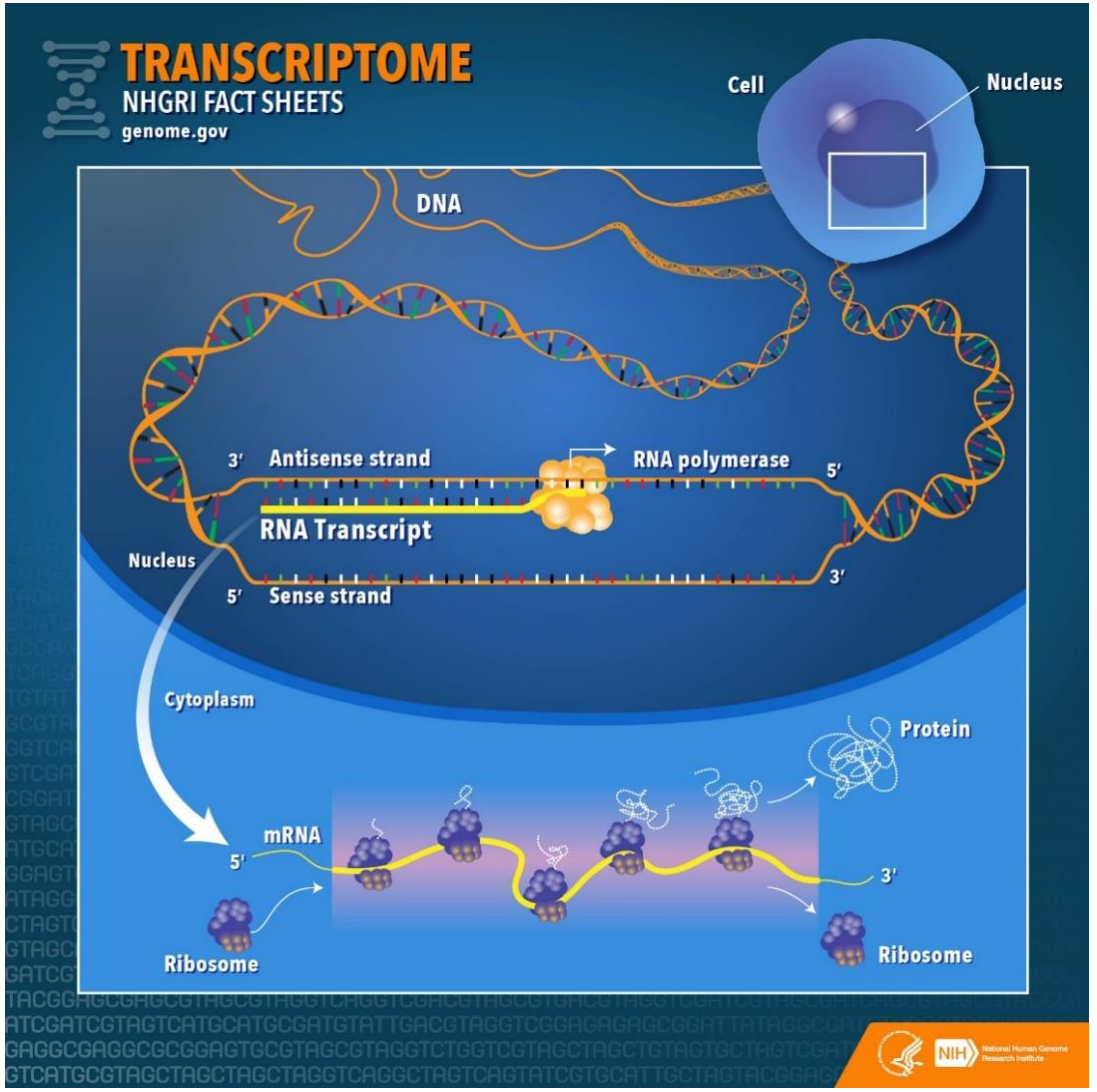
RNA granules contain various ribosomal subunits, translation factors, decay enzymes, helicases, scaffold proteins, and RNA-binding proteins.

RNA granules: post-transcriptional and epigenetic modulators of gene expression.
Nature Reviews Molecular Cell Biology 2009

转录组及研究方法

转录组（Transcriptome）

- 转录组：在某一特定条件下，**所能转录出的所有RNA的总和**，包括信使RNA（mRNA）、核糖体RNA（rRNA）、转运RNA（tRNA）及非编码RNA；狭义上指细胞所能转录出的mRNA。
- 转录组可以被视为**蛋白质组的前体**，即由基因组表达的整组蛋白质。
- 转录组学：研究在单个细胞或特定类型细胞、组织、器官或发育阶段的细胞群内所产生的**各类RNA分子的类型和数量。**



Transcriptome is the set of all RNA transcripts, including coding & non-coding.

https://en.wikipedia.org/wiki/Transcriptome

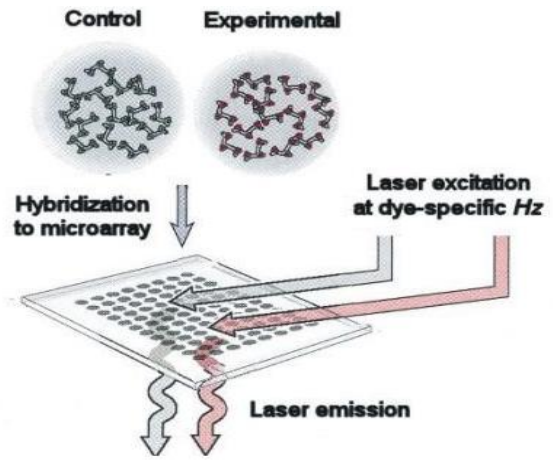
https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet

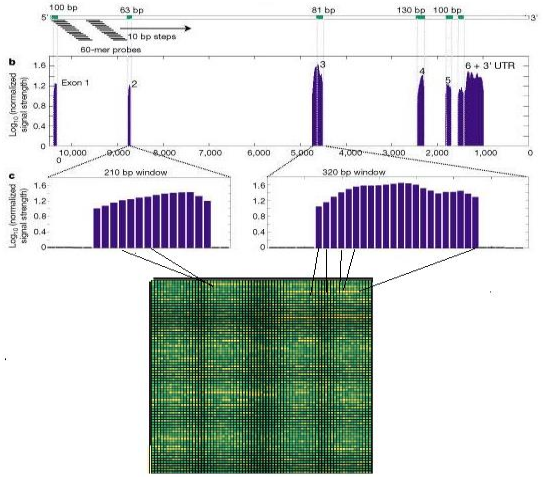
转录组的特点

- Transcriptome is highly **dynamic and complex** in comparison to the relatively stable genome.
- Transcriptome reflects the genes that are being **actively expressed at any given time—gene expression**.
- Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease.
- The key aims of transcriptomics are:
  - to **catalogue all transcripts**, including mRNAs, non-coding RNAs and small RNAs;
  - to **determine the transcriptional structure of genes**, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications;
  - to **quantify the changing expression levels** of each transcript during development and under different conditions.

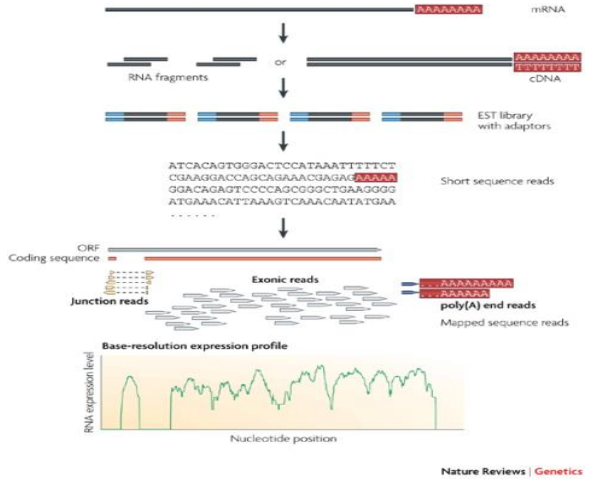
转录组研究方法

Microarray: Hybridization-based





RNA-seq: sequencing-based



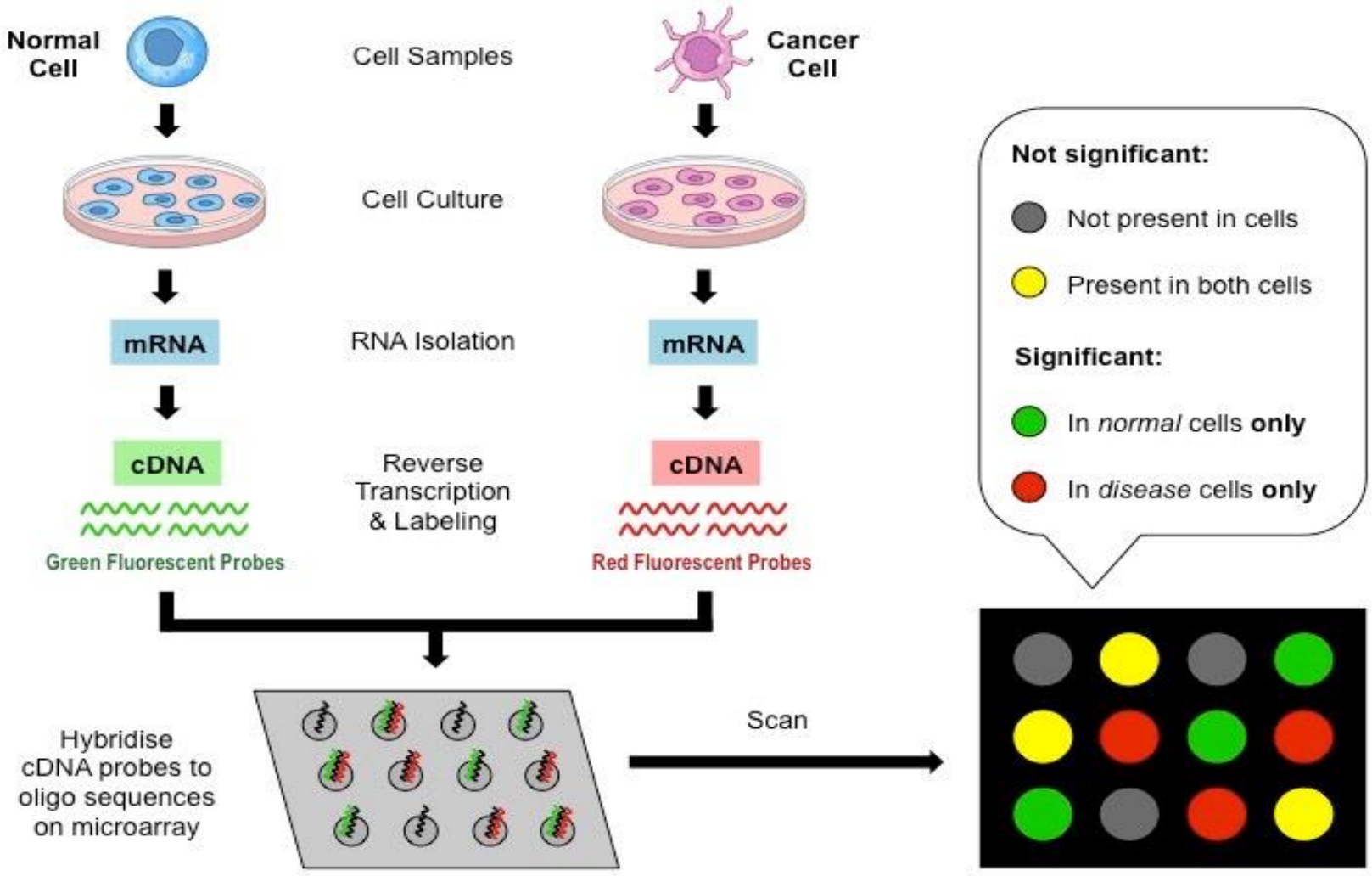
Short-read cDNA
Long-read cDNA
Long-read Direct

1995 Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

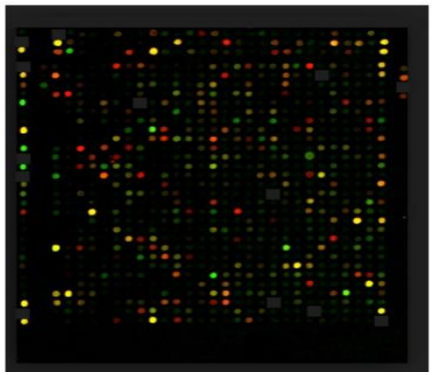
2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

Microarray



双色寡核苷酸基因芯片

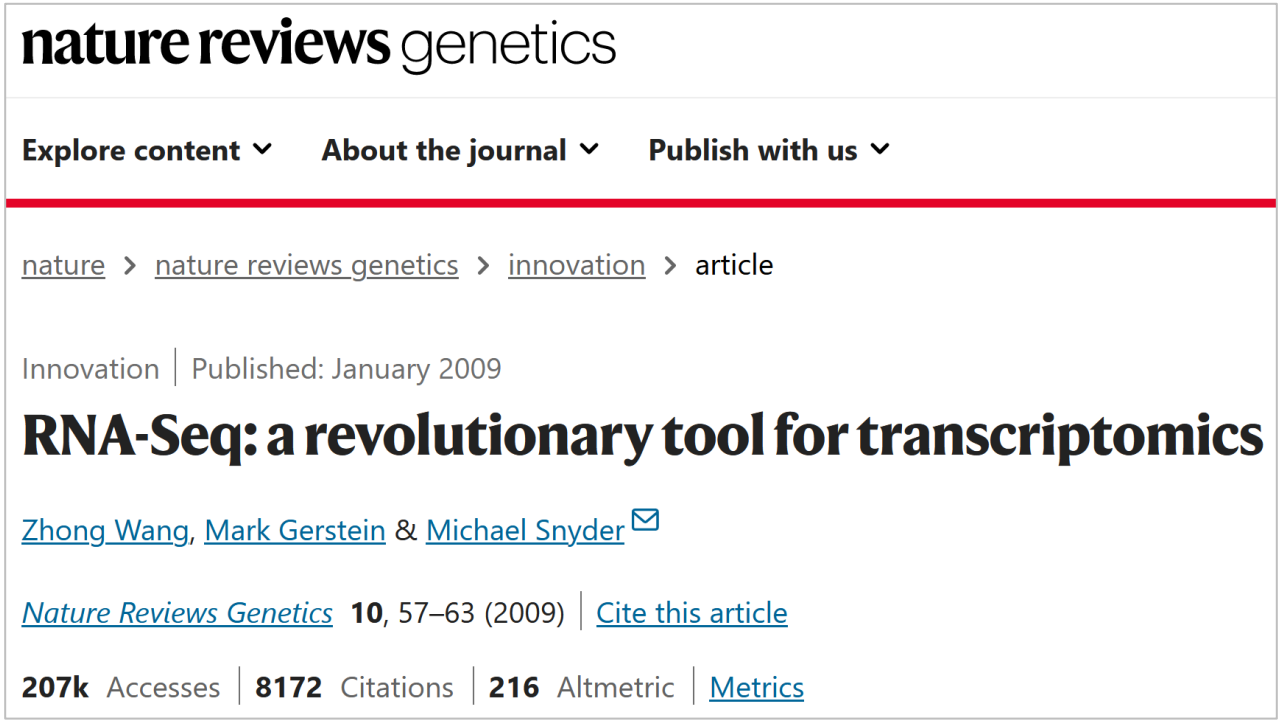
DNA microarray used to detect gene expression in a human (left) and a mouse (right)



https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/microarrays.html

What is RNA-seq?

RNA-seq利用高通量测序技术来检测组织或细胞在特定状态中所有的转录产物

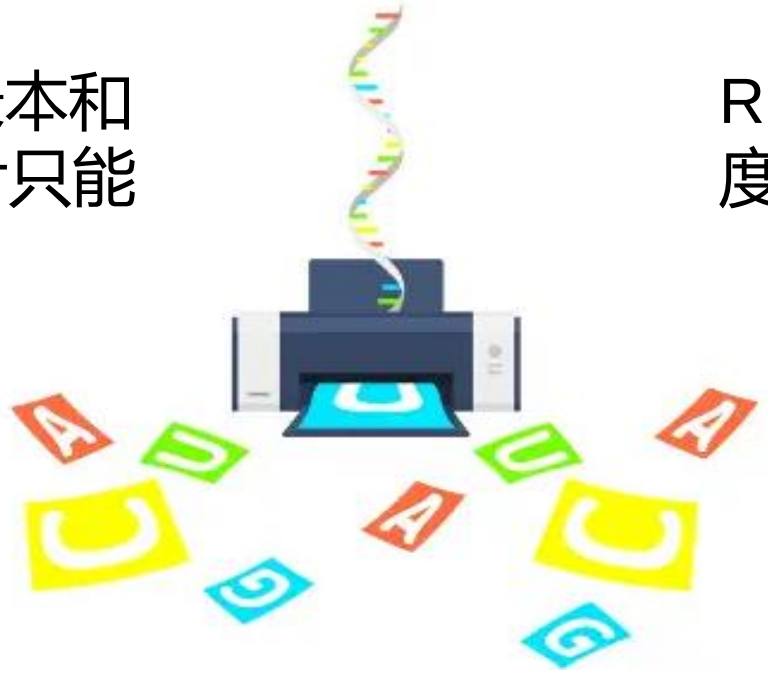


- also called "Whole Transcriptome Shotgun Sequencing" ("WTSS")
- refers to the use of high-throughput sequencing technologies
- sequence cDNA to get information about a sample's RNA content
- a powerful tool to detect the whole transcriptome in cell and tissue

Wang Z, *et al.* ***Nat Rev Genet*** (2009) https://doi.org/10.1038/nrg2484

转录组测序（RNA-sequencing）

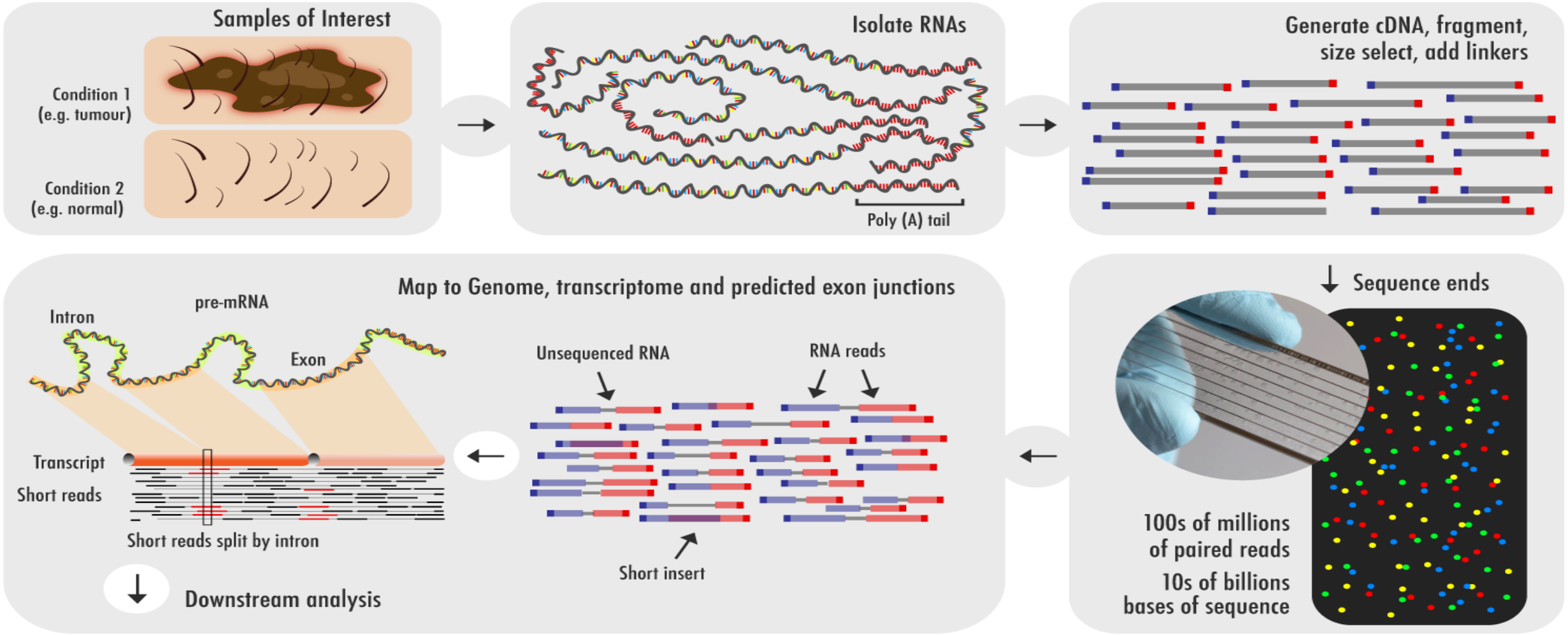
RNA-seq可以发现新的转录本和可变剪切异构体，DNA芯片只能检测已知转录本的表达



RNA-seq比芯片的解析精度更高：单个碱基

文库制备 **测序** **数据分析**

转录组测序基本流程

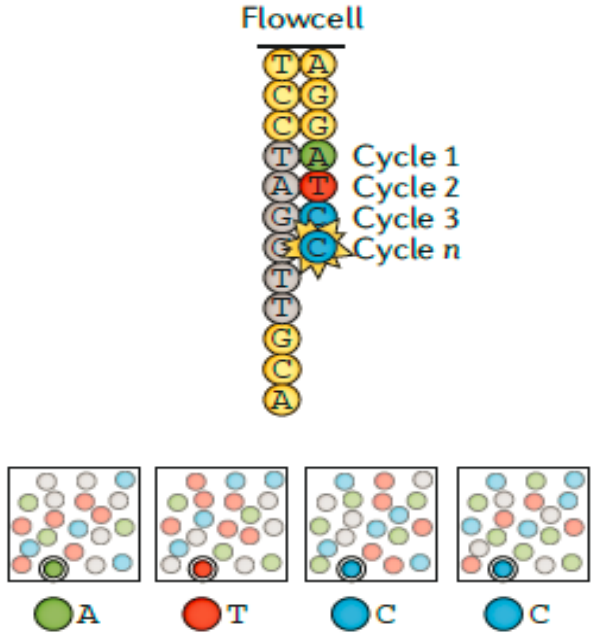


https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

RNA测序主要代表

二代测序

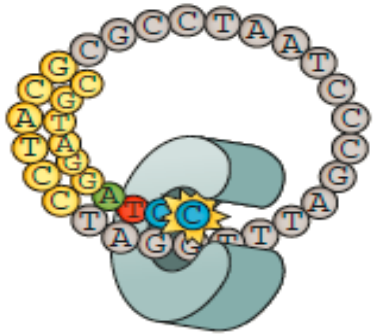
Illumina

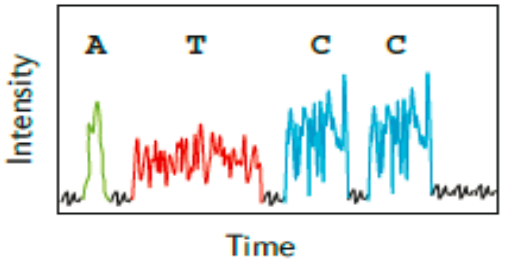


Short read Sequencing

三代测序

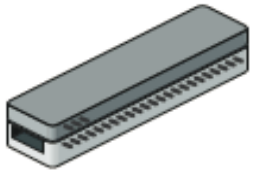
Pacific Biosciences



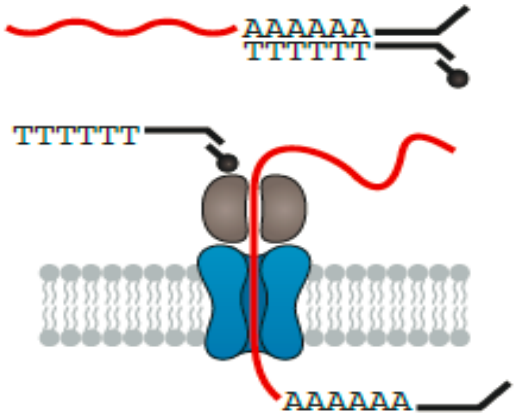


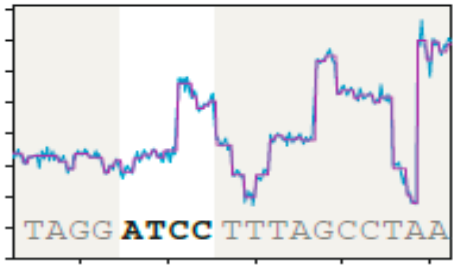
Long read Sequencing

RNA直测



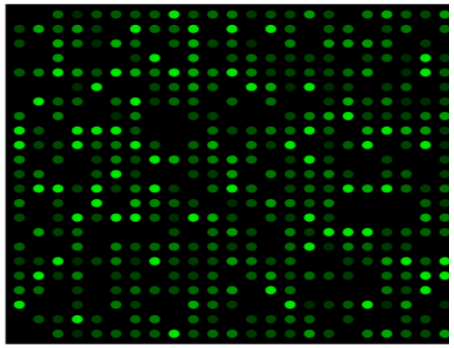
Oxford Nanopore





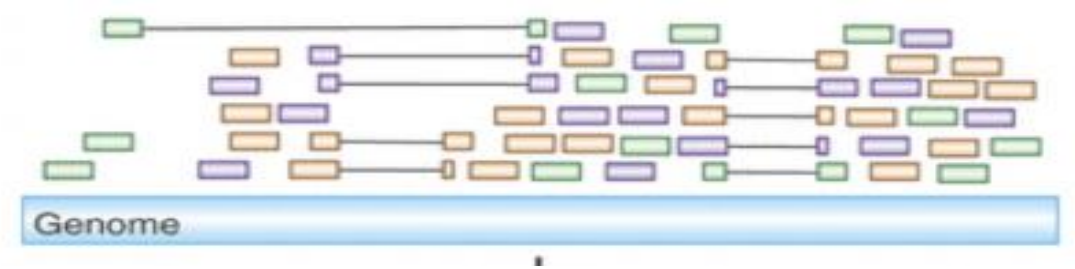
Direct Sequencing

Microarray vs. RNA-seq



Microarray

- Explore changes in expression of known transcripts
- Dynamic range: 44-fold
- Simple data analysis, < 1G



RNA-seq

- Explore expression changes in known transcripts and find novel ones
- Dynamic range: 9,560-fold
- Complex data analysis, big data > 5G

Note: Dynamic range is the ratio between the maximum and minimum expression level.

Advantages of RNA-seq compared with other transcriptomics methods

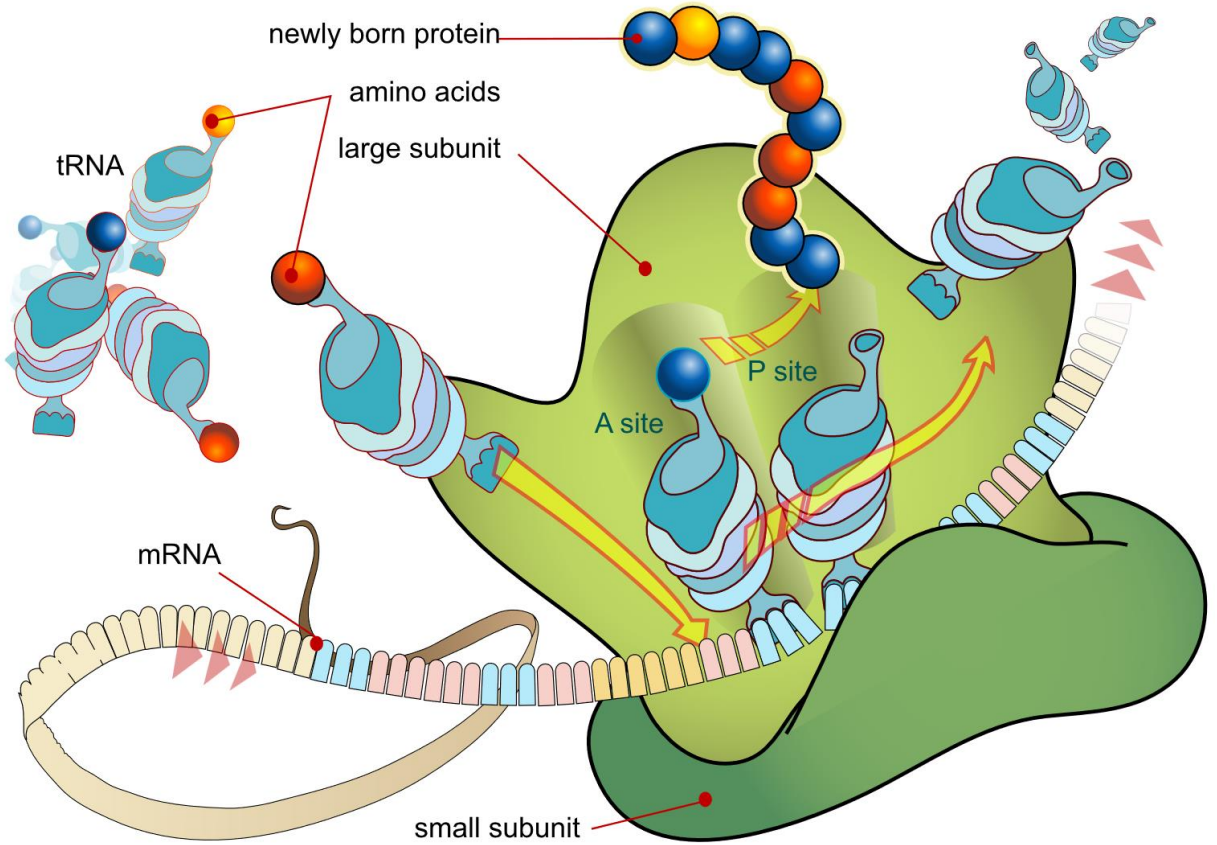
Technology	Tiling microarray	cDNA or EST sequencing	RNA-seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 2009

基因表达

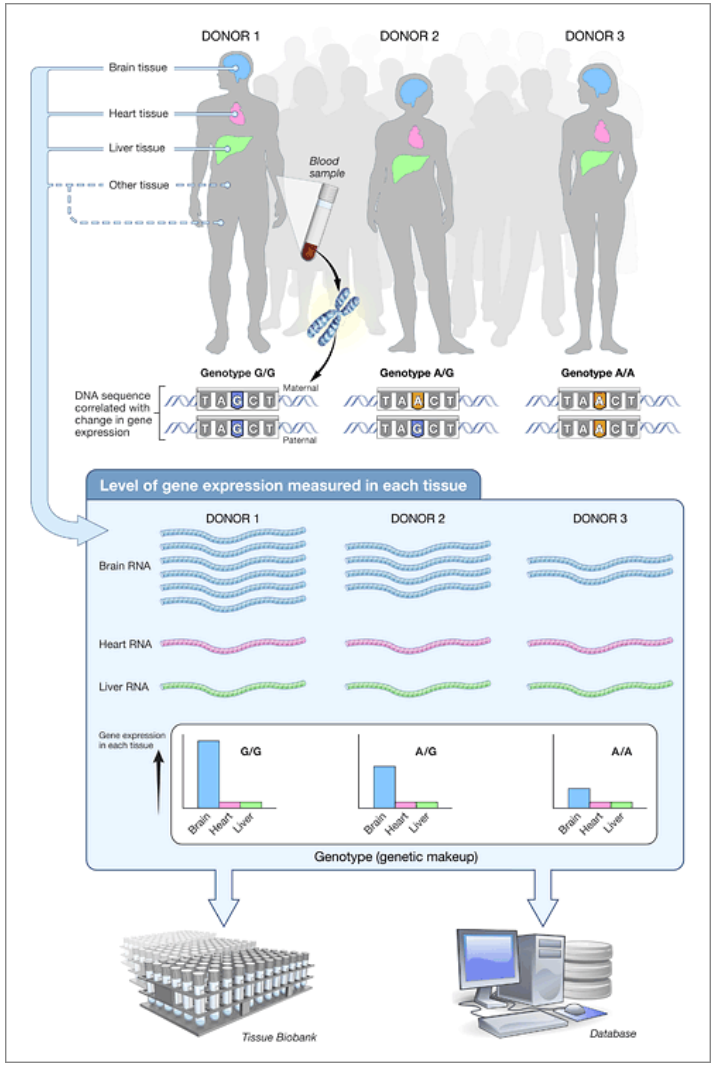
What is gene expression?

- Gene expression is the process by which information from a gene is used in the **synthesis of a functional gene product**.
- **Regulation of gene expression** refers to the **control of the amount and timing** of appearance of the functional product of a gene.
- Gene expression is the most fundamental level at **which the genotype gives rise to the phenotype**, i.e. *observable trait*.



转录组学也被称为“基因表达谱”，检测在一个特定条件下的RNA表达水平

GTEx, Genotype-Tissue Expression



GTEx (基因型-组织表达研究联盟)，为研究遗传差异通过分子表型影响人类健康提供资源。收集并研究了来自948名生前健康的人类捐献者的17382多份尸检样本，涵盖54个组织，包括31个实体器官组织、13个脑分区、全血、两个来自捐献者血液和皮肤的细胞系。利用这些样本研究基因表达的组织和个体差异。

GTEx Portal

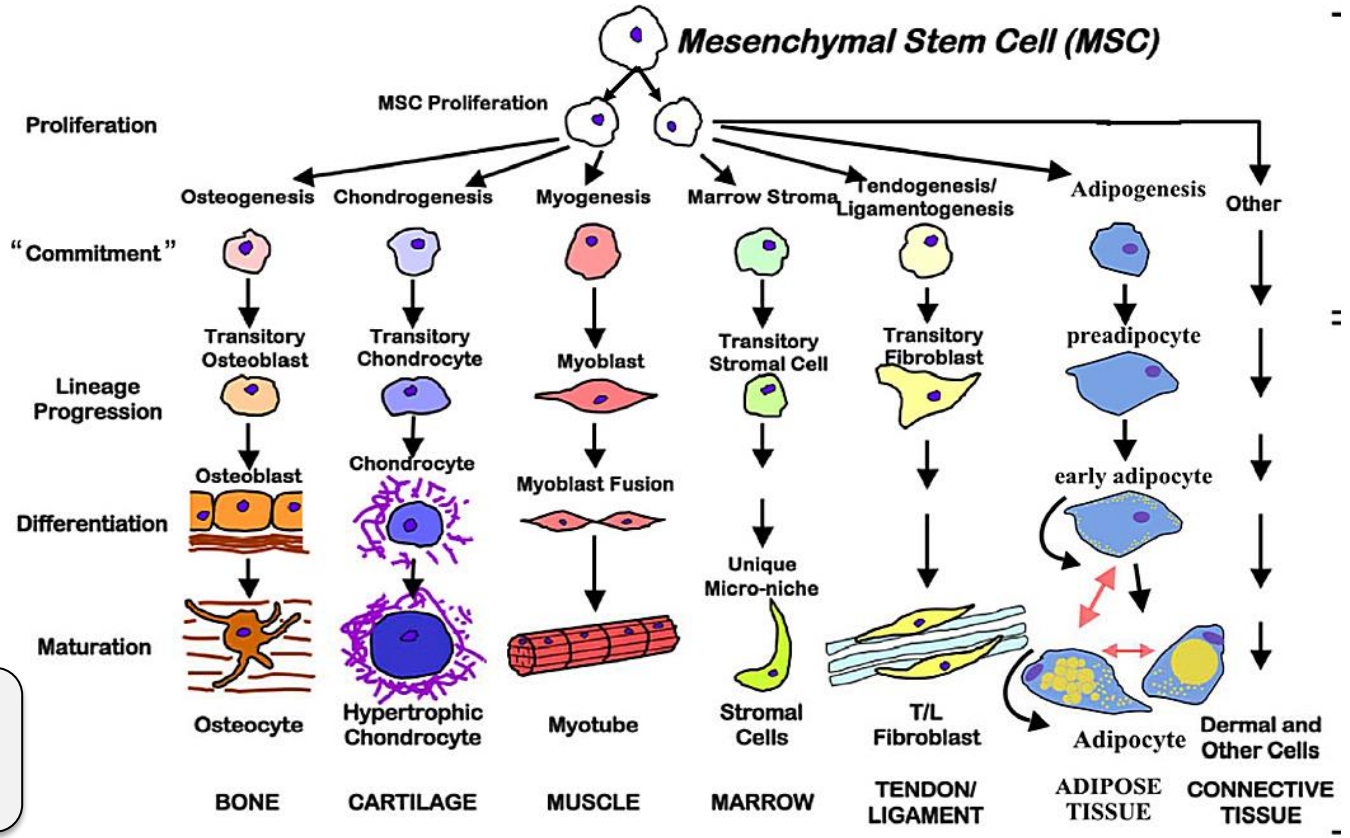
Science 348: 618, 640, 648, 660, and 666 **(2015)**
Nature 550: 190, 204, 239, 244, 249 **(2017)**

https://gtexportal.org

基因表达调控

- Regulation of gene expression includes **a wide range of mechanisms** that are used by cells to **increase or decrease** the production of gene products.
- Regulation of gene expression gives **control over the timing, location, and amount** of a given gene product present in a cell.
- Regulation of gene expression is the **basis** for development, morphogenesis, cellular differentiation, and the versatility and adaptability of any organism.
- Gene expression is regulated through **changes in the number and type of interactions between molecules** that collectively influence transcription of DNA and translation of RNA.

Regulate the **right genes**, at the **right time**, in the **right cells**, and with the **right amount**.

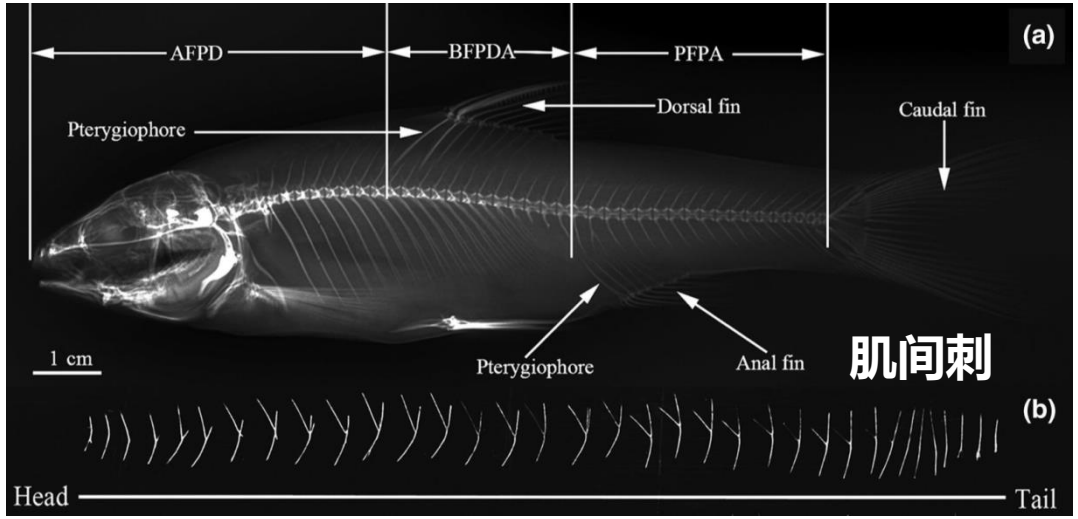


基因表达影响生物性状

2021.12，华中农大高泽霞组提出无刺鱼的专利—斑马鱼

2022.10，高泽霞组实验无刺武昌鱼，敲除基因*Runx2b*

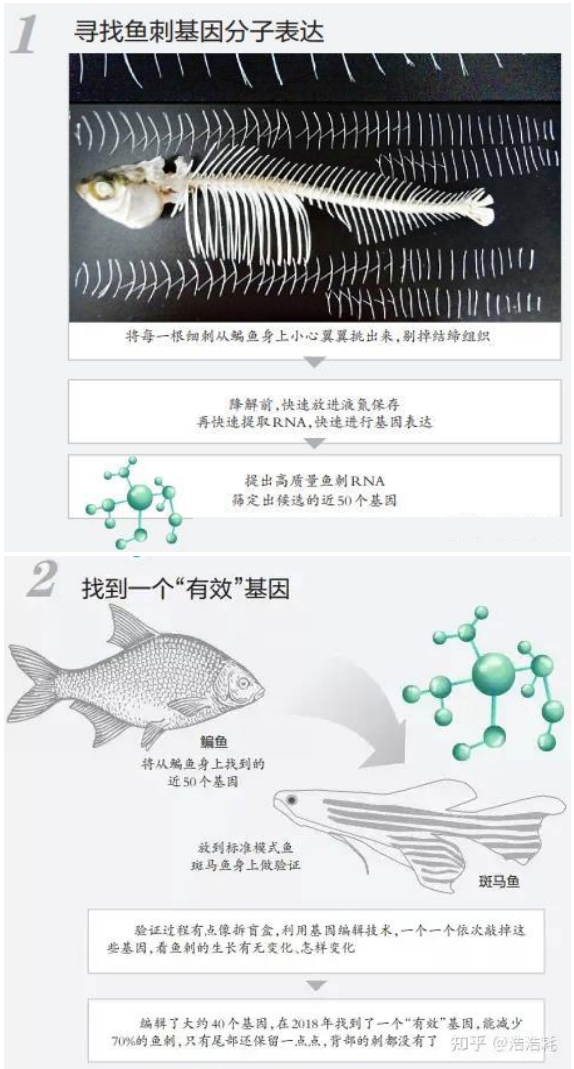
2023.01，水生所桂建芳组和华农高泽霞—无刺鲫鱼



Creation of intermuscular bone-free mutants in amphitriploid gibel carp by editing two duplicated runx2b homeologs[J]. Aquaculture, 2023: 739300.

Human *RUNX2* – RUNX family transcription factor 2: a member of the RUNX family of transcription factors and encodes a nuclear protein with an Runt DNA-binding domain. This protein is essential for **osteoblastic differentiation and skeletal morphogenesis** and acts as a scaffold for nucleic acids and regulatory factors involved in **skeletal gene expression**.

https://www.zhihu.com/question/507977981



基于RNA-seq的基因表达定量计算

$$\mathrm{RPKM} = \frac{10^6 \times n_r}{L \times N}$$

$$\mathrm{FPKM} = \frac{10^6 \times n_f}{L \times N}$$

$$\mathrm{TPM} = \frac{10^6 \times n_r}{L \times T}$$

$$T = \sum_{g \in G} \frac{n_{rg}}{L_g}$$

- **RPKM (Reads Per Kilobase per Million)**：n_r：比对至目标基因的read数量；L：目标基因的外显子长度之和除以1000，单位是kb；N：总有效比对至基因组的read数量
- **FPKM (Fragments Per Kilobase per Million)**：FPKM与RPKM的区别：F是fragments，R是reads，对于双端测序，每个fragments有两个reads，FPKM只计算两个reads比对到同一个转录本的fragments数量，RPKM计算的是可以比对到转录本的reads数量
- **TPM (Transcripts Per Kilobase per Million)**：n_r：比对至目标基因的read数量；L：目标基因的外显子长度之和除以1000，单位是kb；分母T为比对至各基因的read数量n_{rg}与该基因长度L_g的比值的总和。

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012 https://doi.org/10.1007/s12064-012-0162-3

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

TPM vs RPKM

基因	长度 (kb)	样本A (10^6)	样本B (10^6)	样本C (10^6)
Gene 1	100	300	400	500
Gene 2	500	700	750	800
Gene 3	1000	1000	1300	1800
Read depth		2000	2450	3100

Step 1: Normalize for gene length

基因	样本A	样本B	样本C
Gene 1	3	4	5
Gene 2	1.4	1.5	1.6
Gene 3	1	1.3	1.8
Seq. depth	5.4	6.8	8.4

TPM Step 2: Normalize for sequencing depth

基因	样本A	样本B	样本C
Gene 1	0.56	0.59	0.60
Gene 2	0.26	0.22	0.19
Gene 3	0.19	0.19	0.21
	1.00	1.00	1.00

RPKM Step 2: Normalize for read depth

基因	样本A	样本B	样本C
Gene 1	1.50	1.63	1.61
Gene 2	0.70	0.61	0.52
Gene 3	0.50	0.53	0.58
	2.70	2.78	2.71

A sample RNA-Seq pipeline

Library preparation
Illumina HiSeq-2000 sequencing

Data collection

Tuxedo analysis pipeline

Step 1. Read Mapping (TopHat2)

Step 2. Transcript assembly (Cufflinks2)

Step 3. Transcript merging (Cuffmerge2)

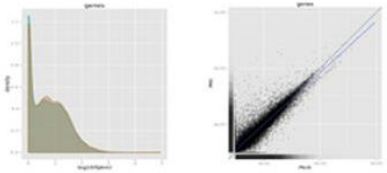
Pre-processing

Step 4. Differential expression analyses (Cuffdiff2)

Transcript ID	Control FPKM	Treatment FPKM	Fold-change
	2.0	10.0	+5
	6.0	3.0	-2
	12.0	12.0	0

Differential expression analysis

Step 5. Data inspection and global plots (CummeRbund)



Genome-Wide Analysis of Alternative Splicing Landscapes Modulated during Plant-Virus Interactions in *Brachypodium distachyon*. *Plant Cell* 2015

第一步：转换为log(ratio)数据形式

聚类分析是为了鉴定表达模式上发生共同变化趋势的基因，这些基因有可能是功能相同，共同调控的基因。聚类分析是在鉴定完毕差异表达基因的基础上做的。

数据形式为 $\log_2$(sample1/sample2)

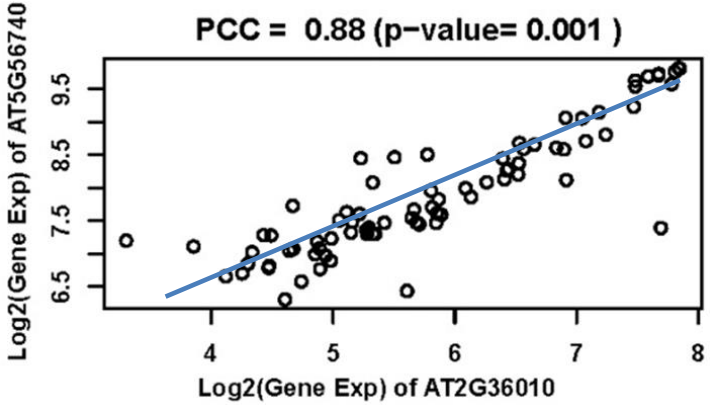
Gene X in sample (stress) 表达水平是 800
Gene X in sample (control) 表达水平是 200
Log2(800/200) = log2(4) = 2，变化倍数为4，**正调控基因**，用红色表示“red”

Gene Y in sample (stress) 表达水平是 is 200
Gene Y in sample (control) 表达水平是 is 800
$\log_2$ (200/800) = $\log_2$(1/4) = -2，变化倍数为4，**负调控基因**，用绿色表示“green”

$\log_2$ (400/400) = $\log_2$ (1) = 0，表达无变化，用黑色表示 “black”

ID	A/control	B/control	C/control	D/control
Gene 1	-2.3	-3.4	0	0.02
Gene 2	-10.2	3.9	0.01	1.03
Gene 3	2.4	2.01	3.1	1.2

第二步：计算基因对间相关系数并聚类



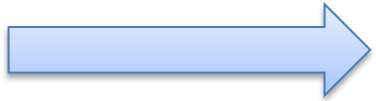
表达变化趋势相同的一对基因具有高相关性，Pearson correlation，线性相关

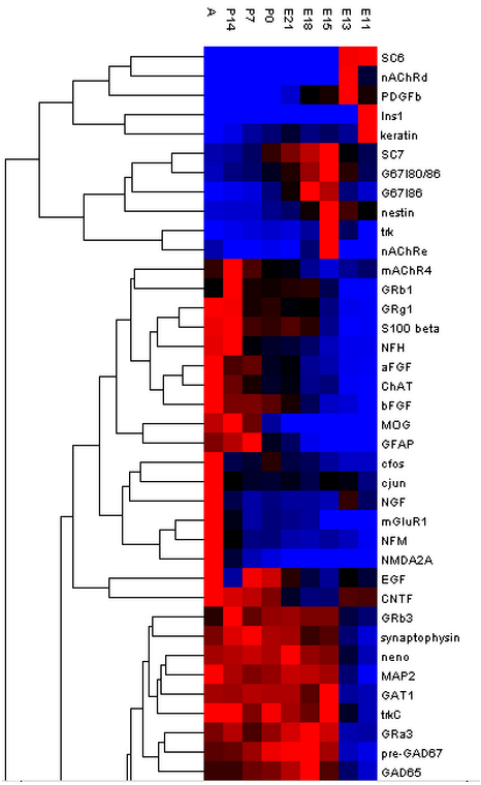
Correlation matrix for any pair of genes

	1	2	3	4	5	6	7	8	9	10
1	1	0.00000	0.80178	0.61237	0.61237	0.91218	0.25000	0.40825	0.40825	0.37500
2	0.00000	1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.0000	0.00000
3	0.80178	0.00000	1	0.76376	0.21822	0.97590	0.53452	0.87287	0.32733	0.80178
4	0.61237	0.00000	0.76376	1	0.16667	0.74536	0.40825	0.66667	–0.16670	0.61237
5	0.61237	0.00000	0.21822	0.16667	1	0.37268	–0.61230	–0.16670	–0.16670	–0.61237
6	0.91287	0.00000	0.97590	0.74536	0.37268	1	0.45644	0.74536	0.37268	0.68465
7	0.25000	0.00000	0.53452	0.40825	–0.61240	0.45644	1	0.61237	0.61237	0.25000
8	0.40825	0.00000	0.87287	0.66667	–0.16670	0.74536	0.61237	1	0.16667	0.91856
9	0.40825	0.00000	0.32733	–0.16670	–0.16670	0.37268	0.61237	0.16667	1	–0.10210
10	0.37500	0.00000	0.80178	0.61237	0.10206	0.68465	0.25000	0.91856	–0.10210	1

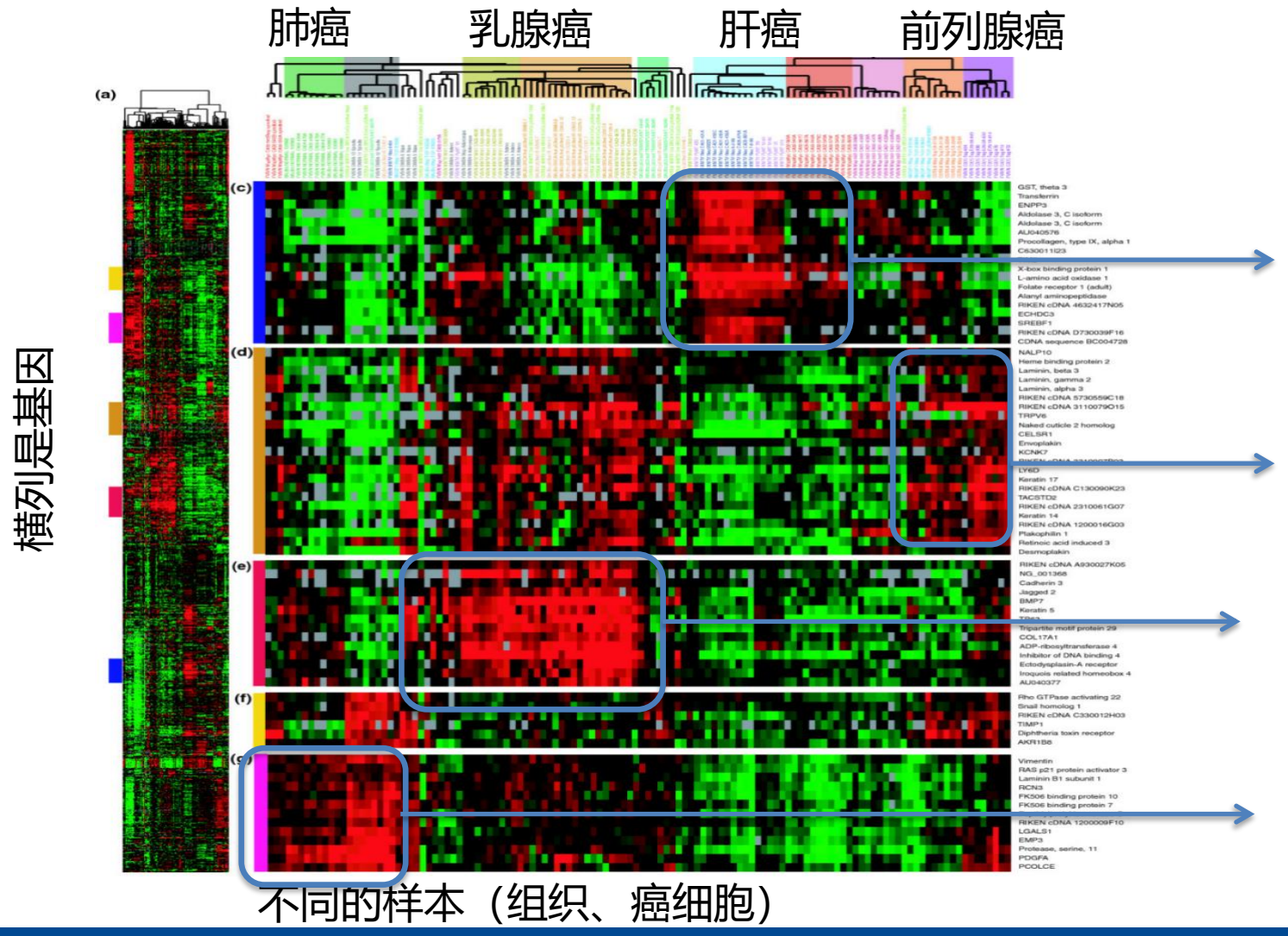
具有相关系数较高的基因会聚类在一起

Hierarchical Clustering Algorithm





基因表达模式



癌症标志物

肝癌里特异表达的基因

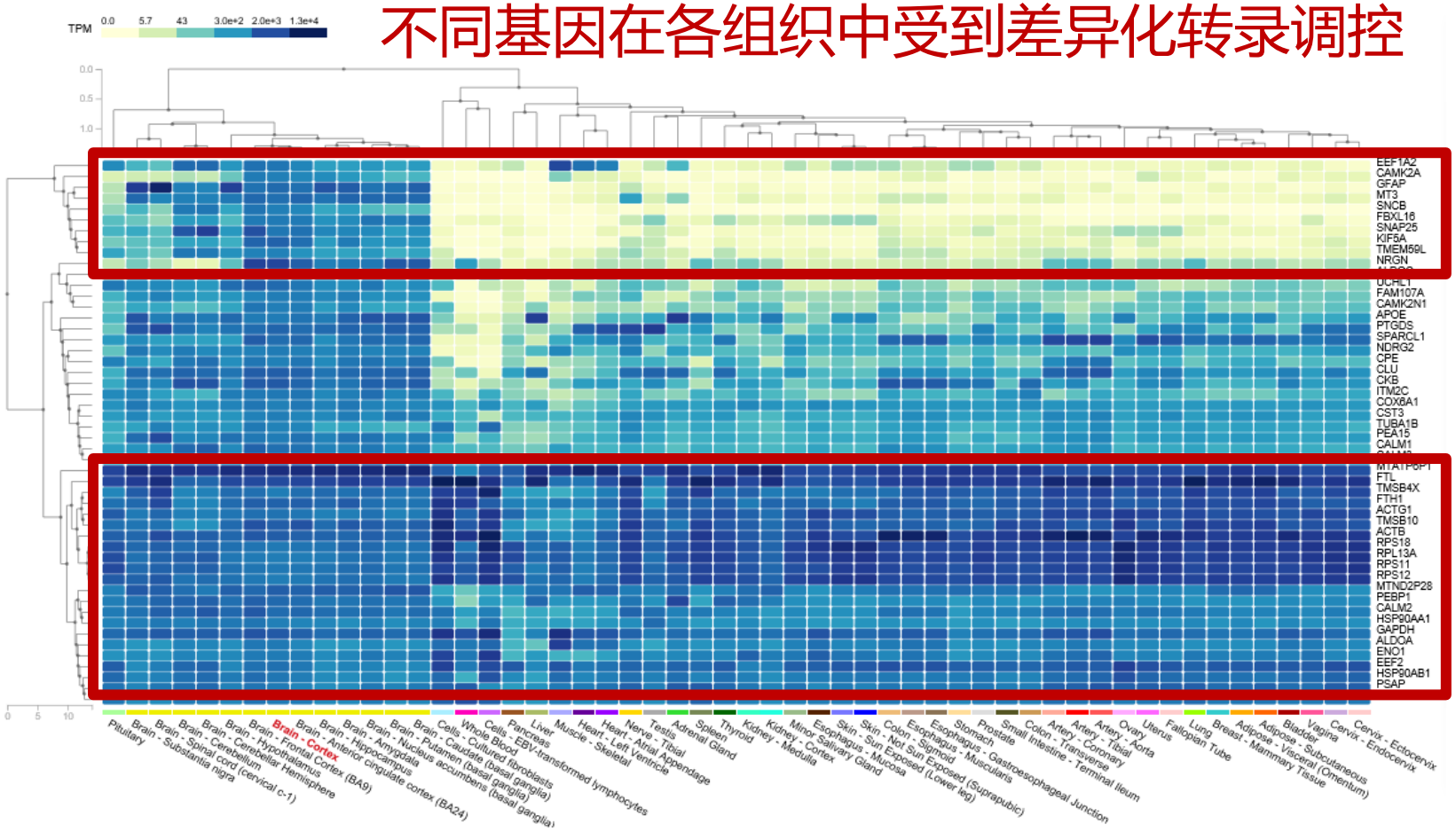
前列腺癌里特异表达的基因

乳腺癌里特异表达的基因

肺癌里特异表达的基因

转录调控影响基因表达模式

不同基因在各组织中受到差异化转录调控



特定组织
表达基因

多组织
表达基因

GTEx Portal

https://gtexportal.org/home/topExpressedGenePage

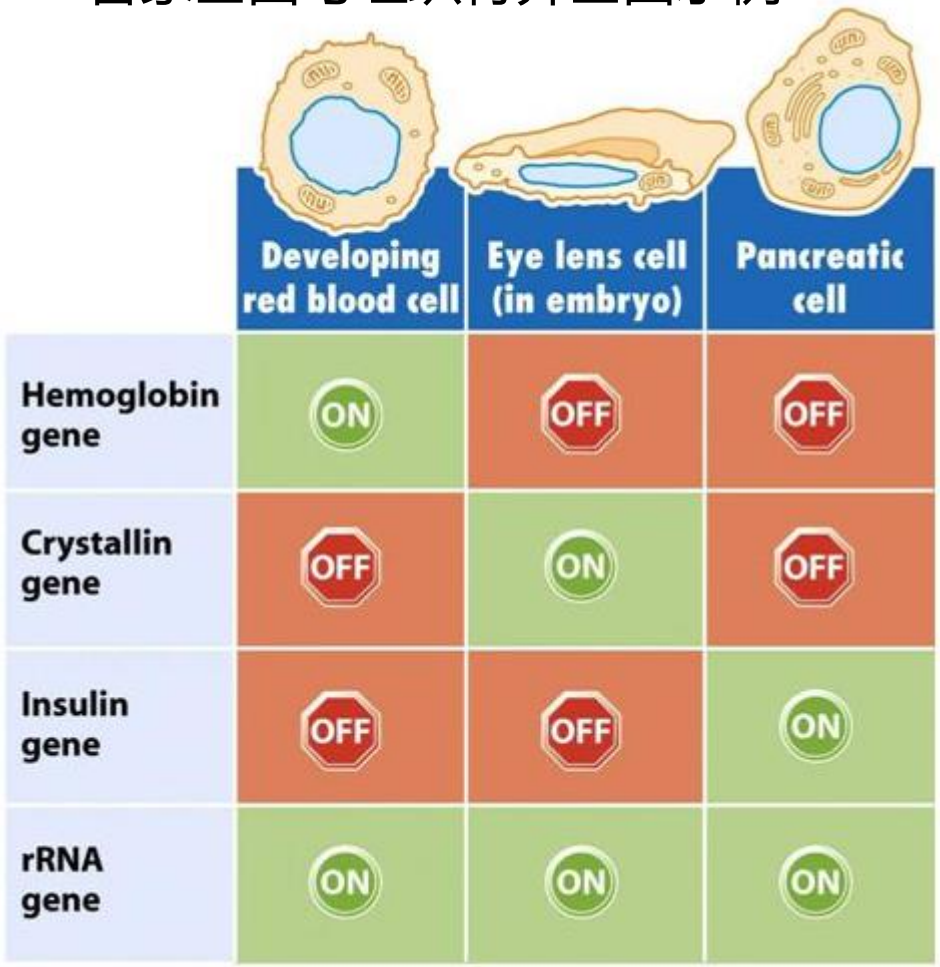
管家基因与组织特异基因

管家基因 House-Keeping Genes	组织特异基因Tissue-Specific Genes
Functional in all cell types	Functional in only certain cell types
Expressed constitutively	Expressed only when needed
Expression is not regulated	Expression is regulated by regulatory proteins, hormones, etc, as required under certain conditions
Needed for essential basic cellular activities	Needed for specific cellular activities
Genes coded for enzymes of glycolysis are active in all cell types throughout the life	Genes expressed only in specific conditions

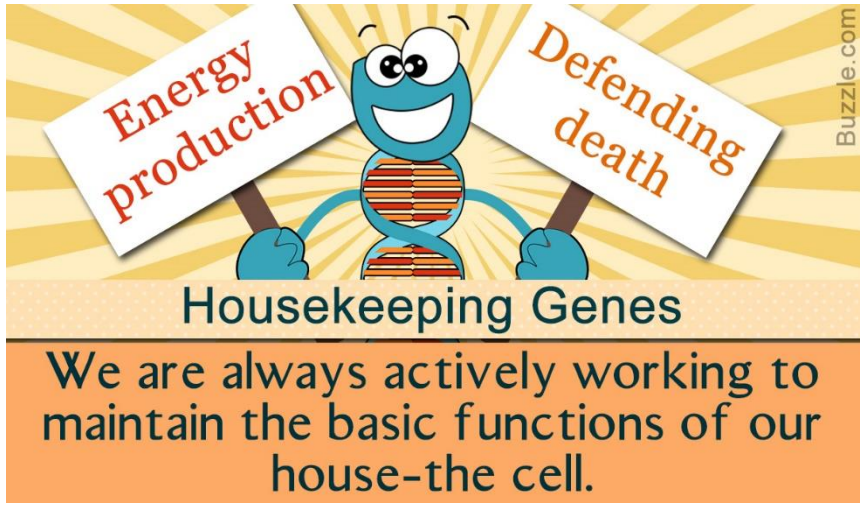
https://www.majordifferences.com/2013/10/difference-between-house-keeping-and_4.html

管家基因与组织特异基因

管家基因与组织特异基因示例



https://slideplayer.com/slide/5337297/



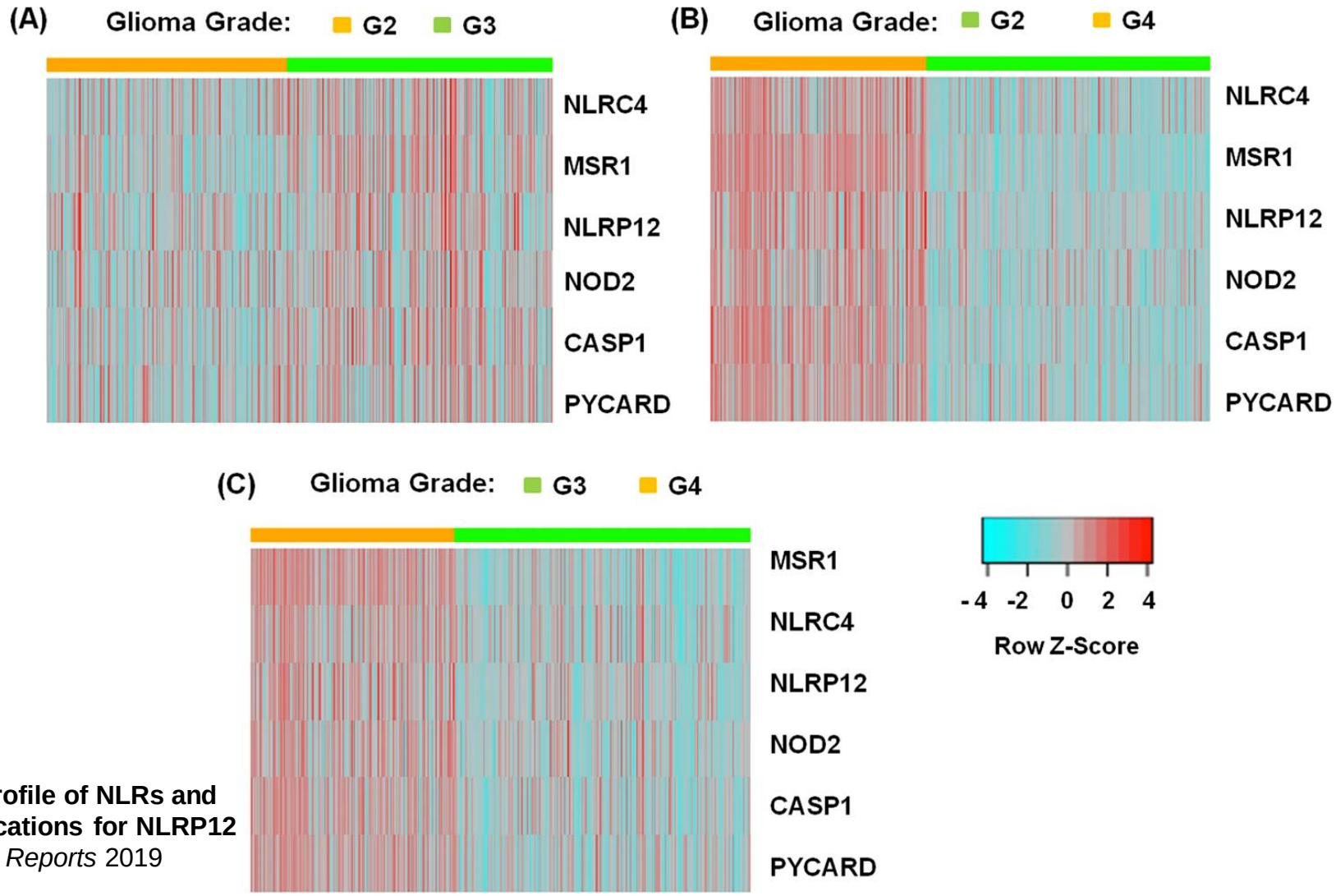
https://biologywise.com/the-importance-of-housekeeping-genes

GO terms specific for HK & TS genes

GO terms	GO ID	Description	% of HK genes	% of TS genes
HK genes	GO:0006412	protein biosynthesis	29.1	-
	GO:0003735	structural constituent of ribosome	23.3	-
	GO:0005840	ribosome	18.4	-
	GO:0005842	cytosolic large ribos. subunit (sensu Euk.)	10.7	-
	GO:0030529	ribonucleoprotein complex	4.9	-
	GO:0006414	translational elongation	3.9	-
TS genes	GO:0005615	extracellular space	-	14.8
	GO:0004295	trypsin activity	-	8.4
	GO:0004263	chymotrypsin activity	-	8.4
	GO:0005576	extracellular region	-	6.7
	GO:0006810	transport	-	6.6
	GO:0008233	peptidase activity	-	6.6

Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics* 2006

Differential gene expression profiles

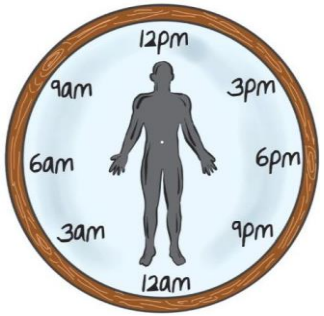


Differential Expression Profile of NLRs and AIM2 in Glioma and Implications for NLRP12 in Glioblastoma. *Scientific Reports* 2019

Circadian gene

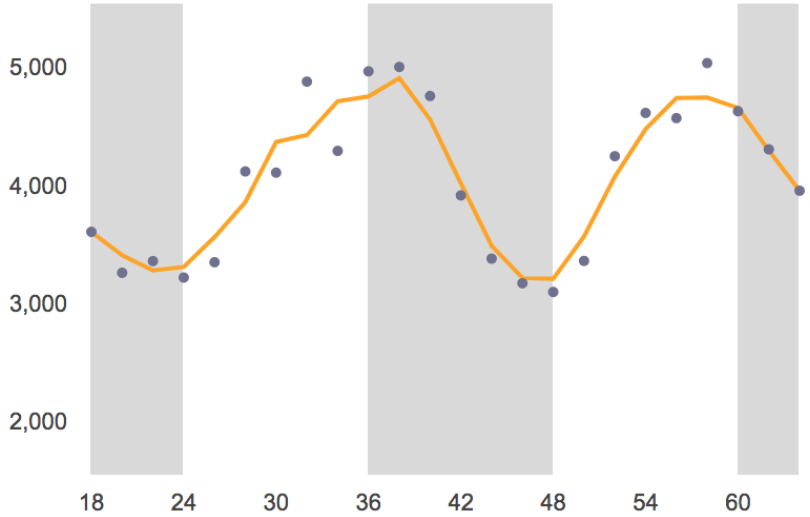
Circadian genes are defined as genes whose protein products are necessary components for the **generation and regulation of circadian rhythms**.

Wall Clock

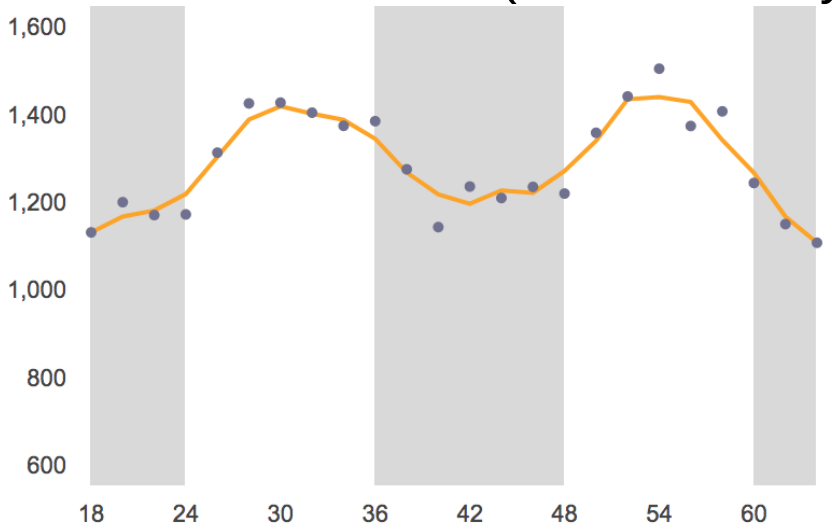


Body Clock

SLC2A2 (mouse liver)

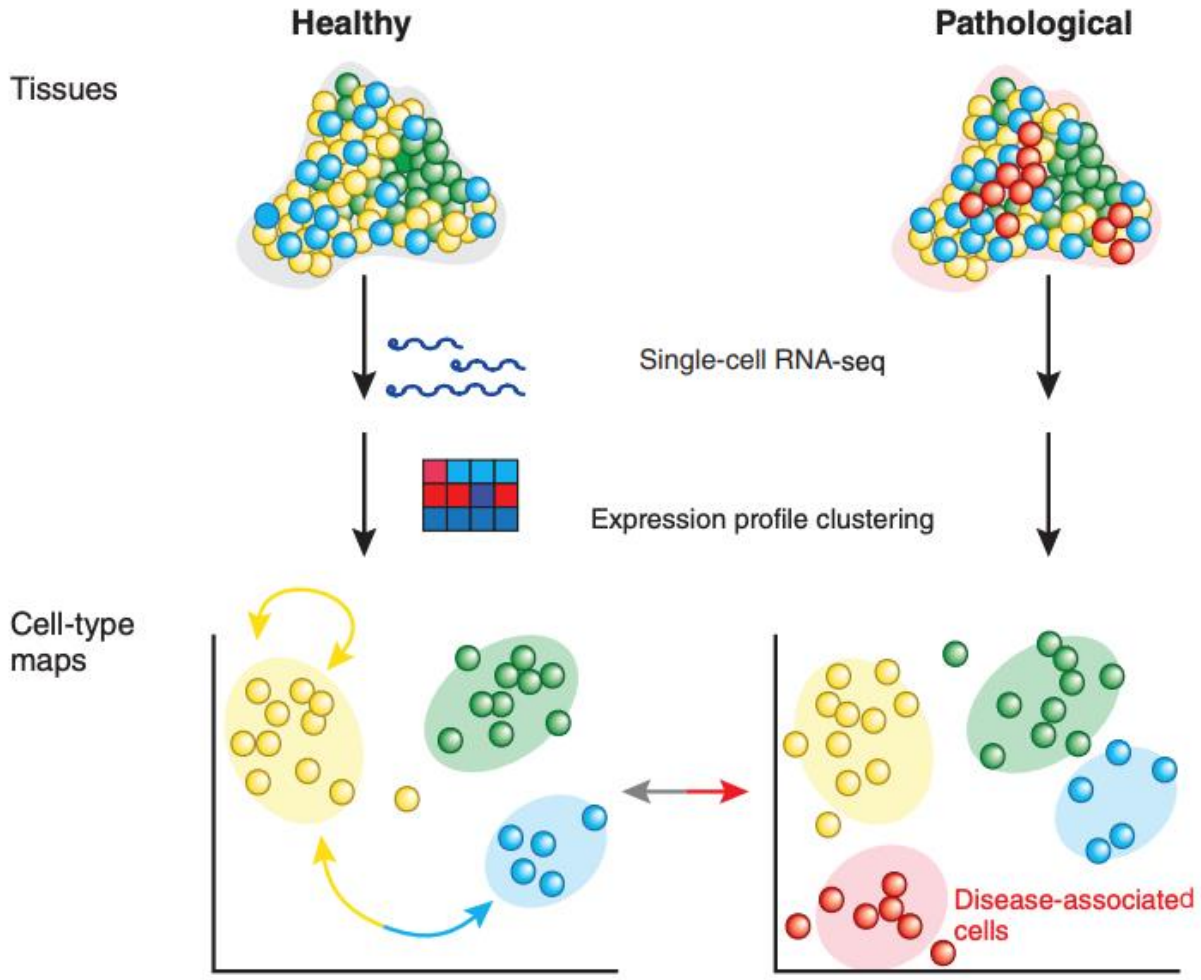


SLC2A2 (mouse kidney)



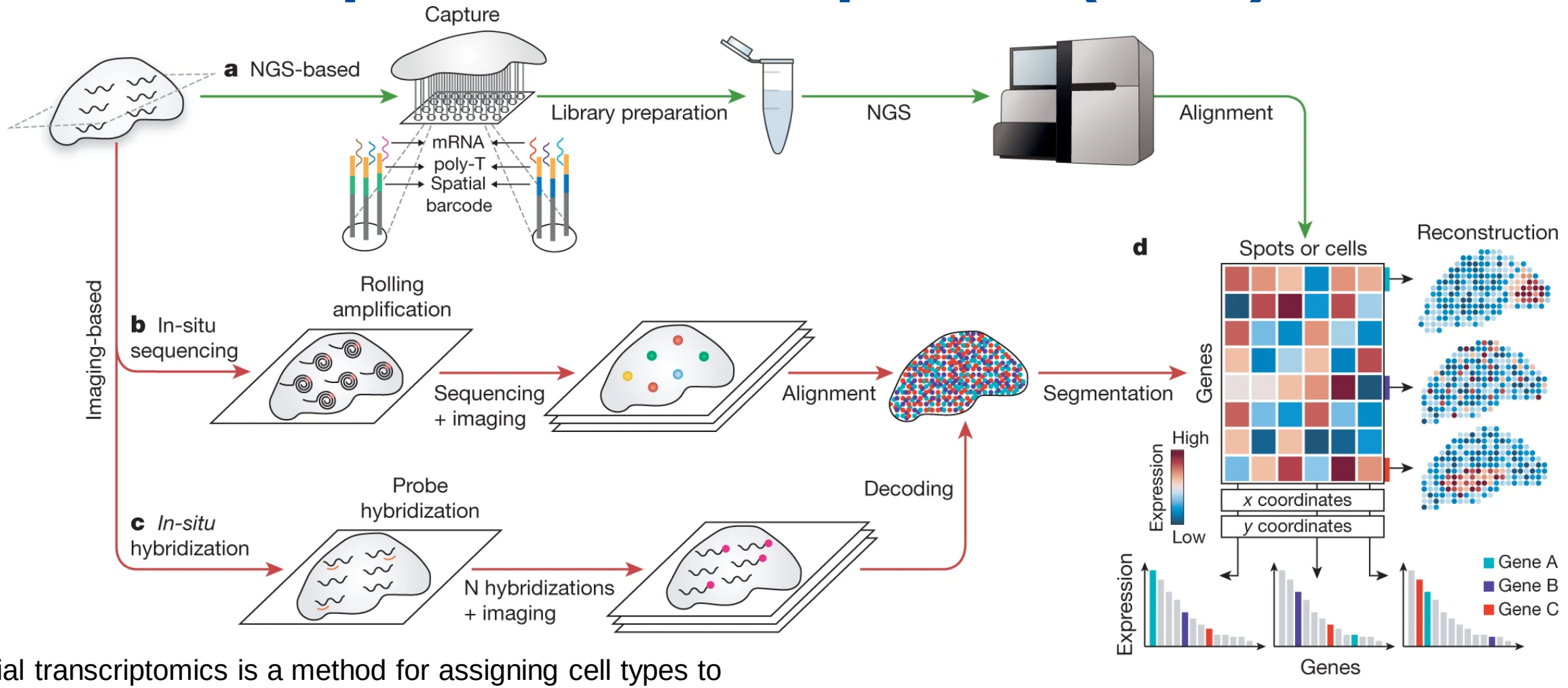
Single-cell transcriptomics

- **Single-cell transcriptomics** is a powerful way for **high-throughput, high-resolution transcriptomic analysis of cell states and dynamics**.
- **Single-cell RNA-Seq** (scRNA-Seq) captures **transcriptional profiles in each cell type** to describe the genetic basis of their identity and function.



Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* 2014

Spatial Transcriptome (STM)



Spatial transcriptomics is a method for assigning cell types to their locations in the histological sections and can also be used to determine subcellular localization of mRNA molecules.

Exploring tissue architecture using spatial transcriptomics. *Nature* 2021
https://www.nature.com/articles/s41592-022-01409-2

Summary

- RNA基本概念、RNA世界假说
- RNA类型及其特点：mRNA、ncRNA、miRNA、lncRNA、circRNA
- RNA生命动态：转录、加工、修饰、降解、交互网络
- 转录组与转录组学
- 基因表达芯片与RNA-seq的优缺点
- 基因表达
  - 定量：TPM、FPKM、RPKM；聚类分析；差异分析
  - 模式：管家基因、组织特异基因、节律基因
- 单细胞转录组学，空间转录组学

Further Reading

- **Origin of life: the RNA world.** *Nature* 1986
- **Bacterial RNA thermometers: molecular zippers and switches**. *Nature Reviews Microbiology* 2012
- **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature Reviews Genetics* 2009
- **Human housekeeping genes, revisited.** *Trends in Genetics* 2013
- **MicroRNAs in Cancer.** *Annual Review of Pathology: Mechanisms of Disease* 2014
- **The roles of RNA processing in translating genotype to phenotype**. *Nature Reviews Molecular Cell Biology* 2016
- **Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale**. *Quantitative Biology* 2018
- **The Role of Non-coding RNAs in Oncology**. *Cell* 2019
- **RNA sequencing: the teenage years**. *Nature Reviews Genetics* 2019
- **The RNA Atlas expands the catalog of human non-coding RNAs**. *Nat Biotechnol* 2021
- **Exploring tissue architecture using spatial transcriptomics**. *Nature* 2021
- **Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression**. *Nature Reviews Molecular Cell Biology* 2022
- **The Expanding Vistas of Spatial Transcriptomics**. *Nature Biotechnology* 2022