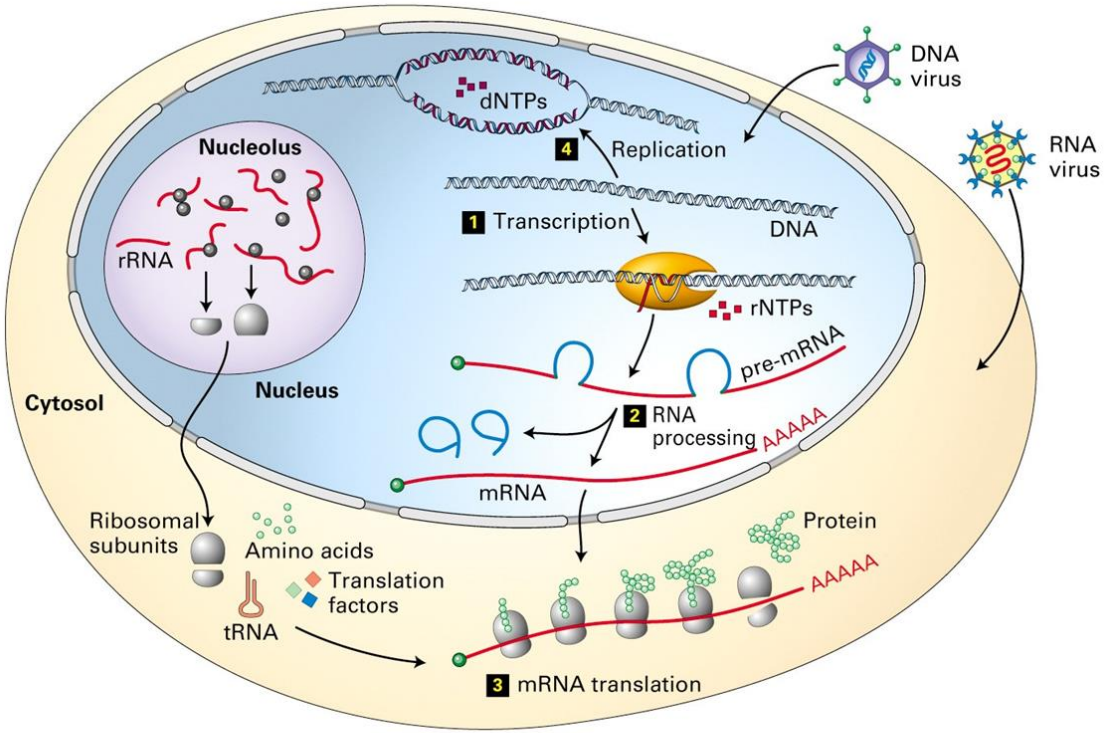
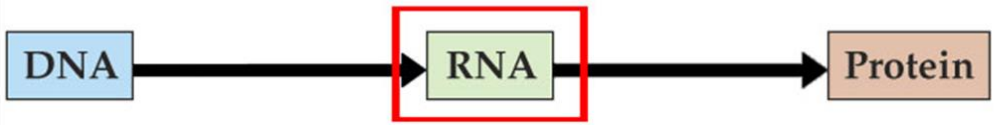
第七章　转录组学

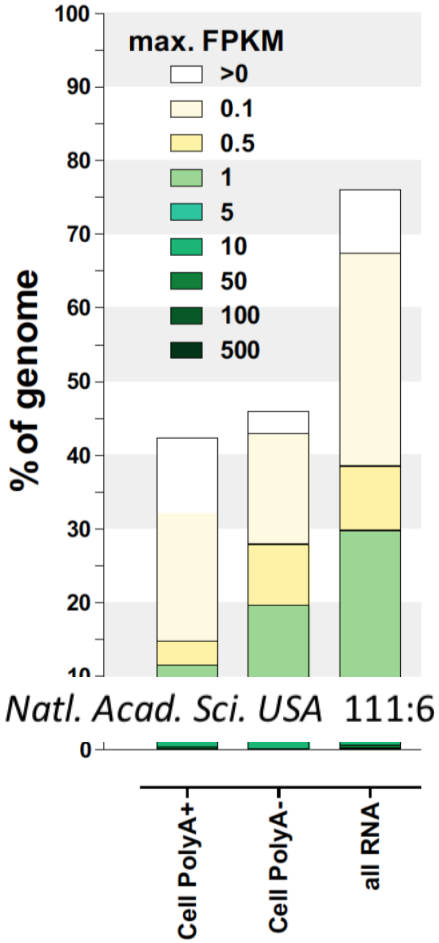
转录与转录组（Transcriptome）

RNA 生物学





ENCODE项目估计的基因组中可转录区域的比例



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

(Kellis M, et al. Proc Natl Acad Sci U S A, 2014)

第七章　转录组学

——第一节　转录组学概述

1 转录组学概述

1.1 转录组学的研究对象

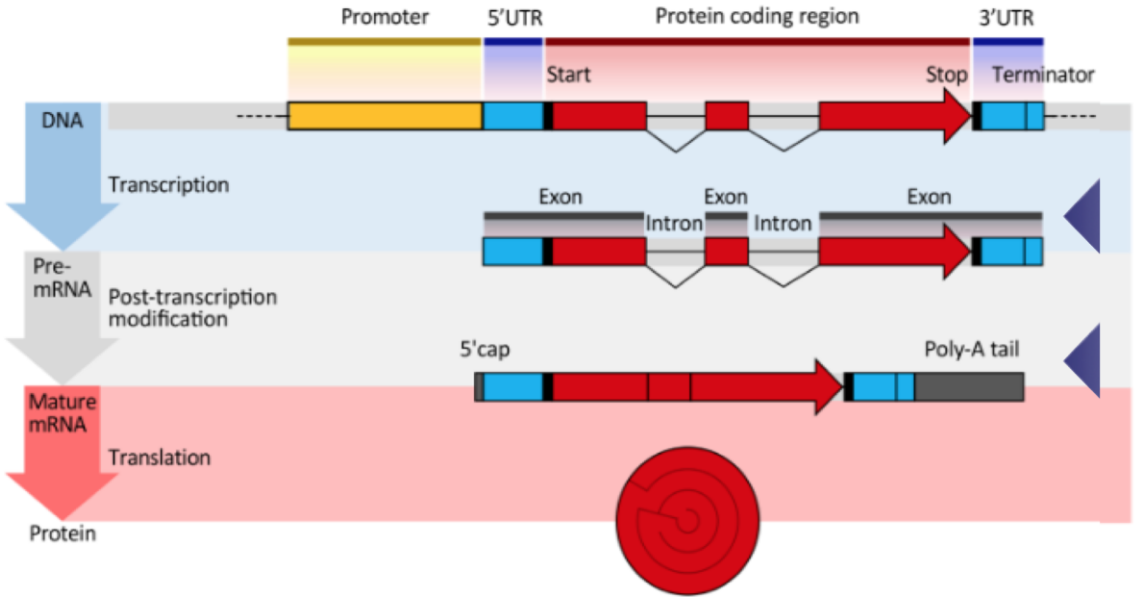
1.2 转录组学的研究方法

1.3 特殊测序建库方法

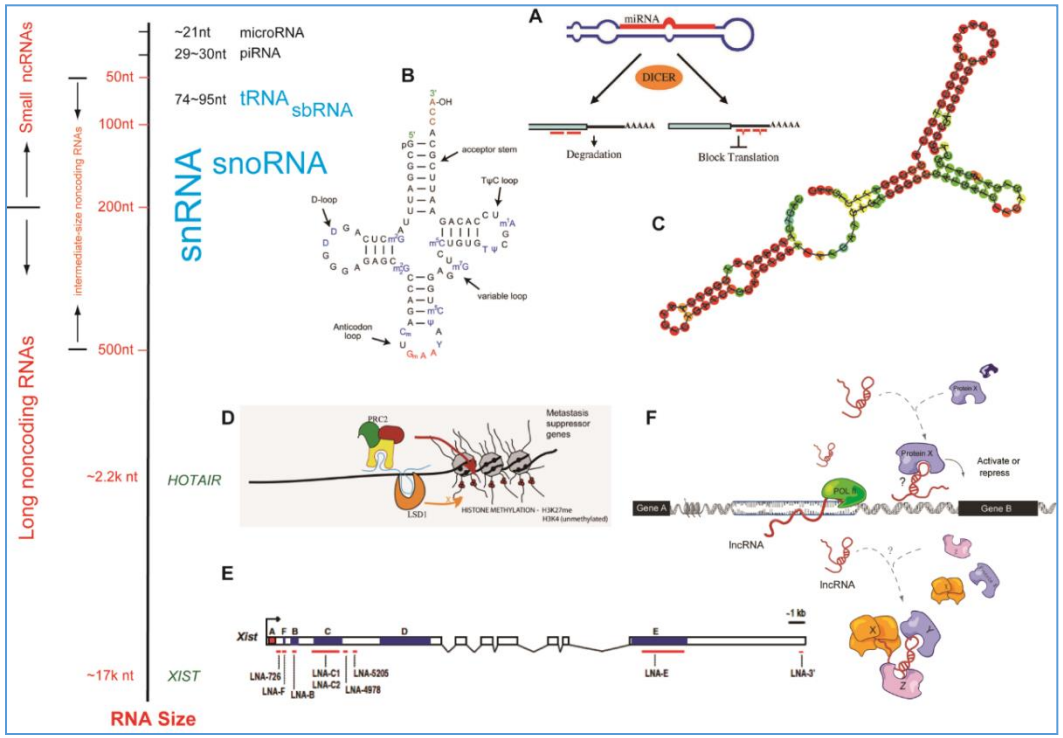
1.4 单细胞和空间转录组学技术

1.1 转录组学的研究对象

转录组学：在整体水平上研究细胞中基因转录的情况及转录调控规律的科学，包括解析mRNA、非编码 RNA（non-coding RNA，ncRNA）等各种转录产物的生物合成、转运、降解等生物学过程；注释基因的转录结构和剪接模式，构建与分析转录调控网络；发掘RNA剪接、修饰与调控规律等。



各个层级的编码基因转录本



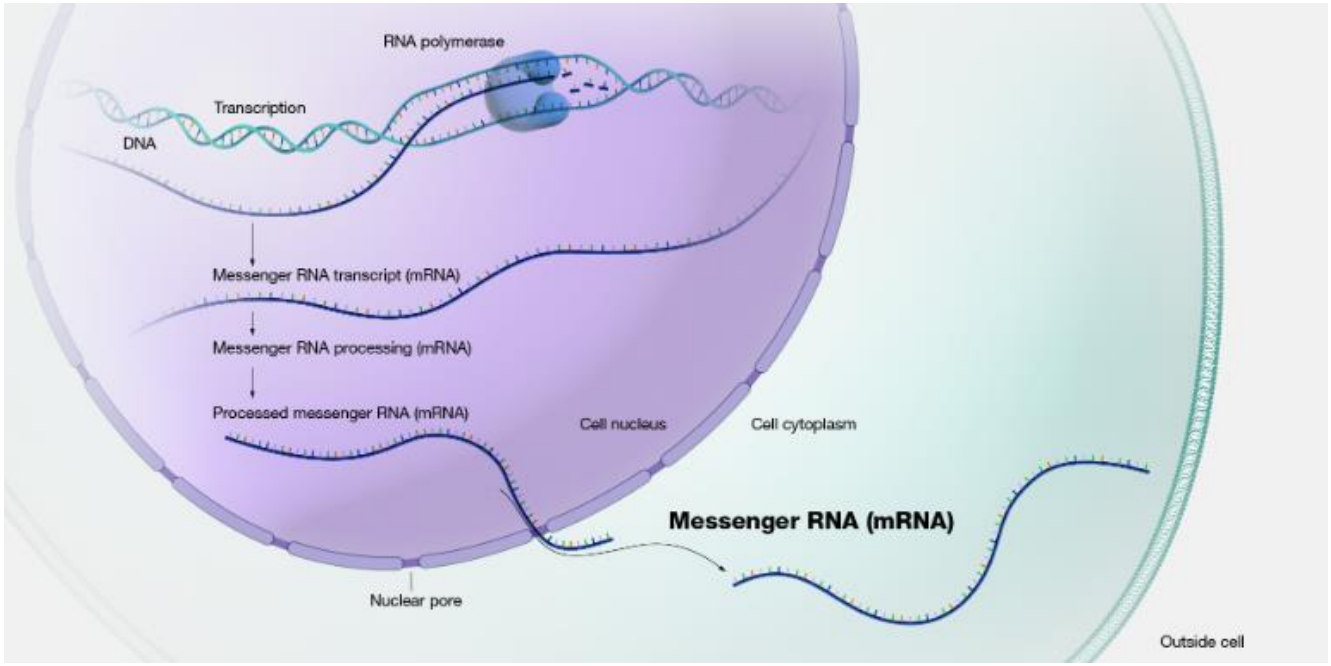
非编码RNA转录产物

1.1 转录组学的研究对象

编码mRNA及功能

mRNA是编码区的转录产物，由编码区(coding region)--从起始密码子AUG开始经一连串编码氨基酸的密码子直至终止密码子、5’端上游非编码区(5’UTR)、3’端下游非编码区(3’UTR)组成，负责将 DNA 中存储的遗传信息翻译成功能性蛋白质。

mRNA 于 1989 年首次被提出作为一种治疗方法，可用于治疗自身免疫、代谢和呼吸道炎症以及癌症等疾病。在新冠流行期间，也开发了有效的 COVID-19 mRNA 疫苗。

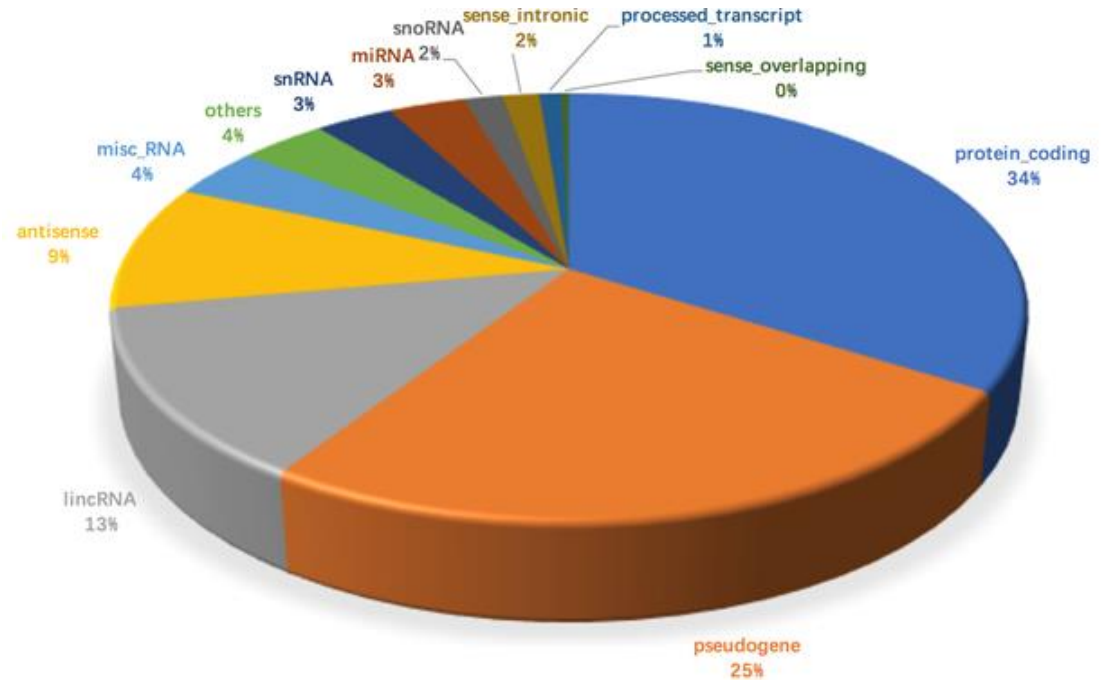


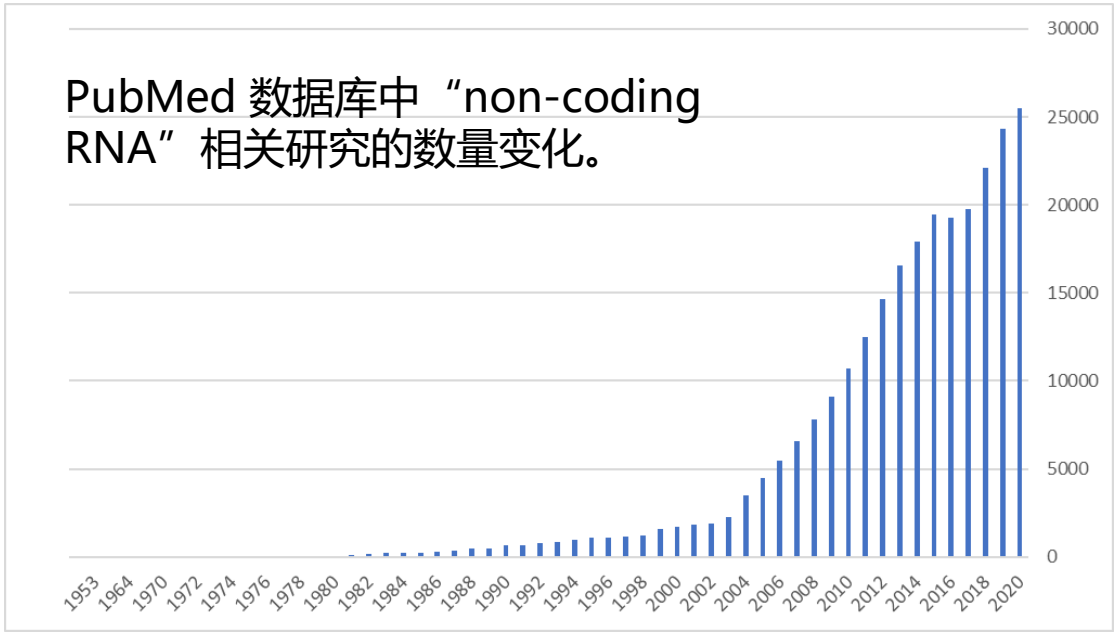
https://www.genome.gov/genetics-glossary/messenger-rna

1.1 转录组学的研究对象

非编码 RNA 及分类

- 非编码 RNA（ncRNA）是指任何可以不被翻译成蛋白质而发挥作用的 RNA 分子。
- 最终转录产物为非编码 RNA 的 DNA 序列通常被称为非编码基因或非编码 RNA 基因。





1.1 转录组学的研究对象

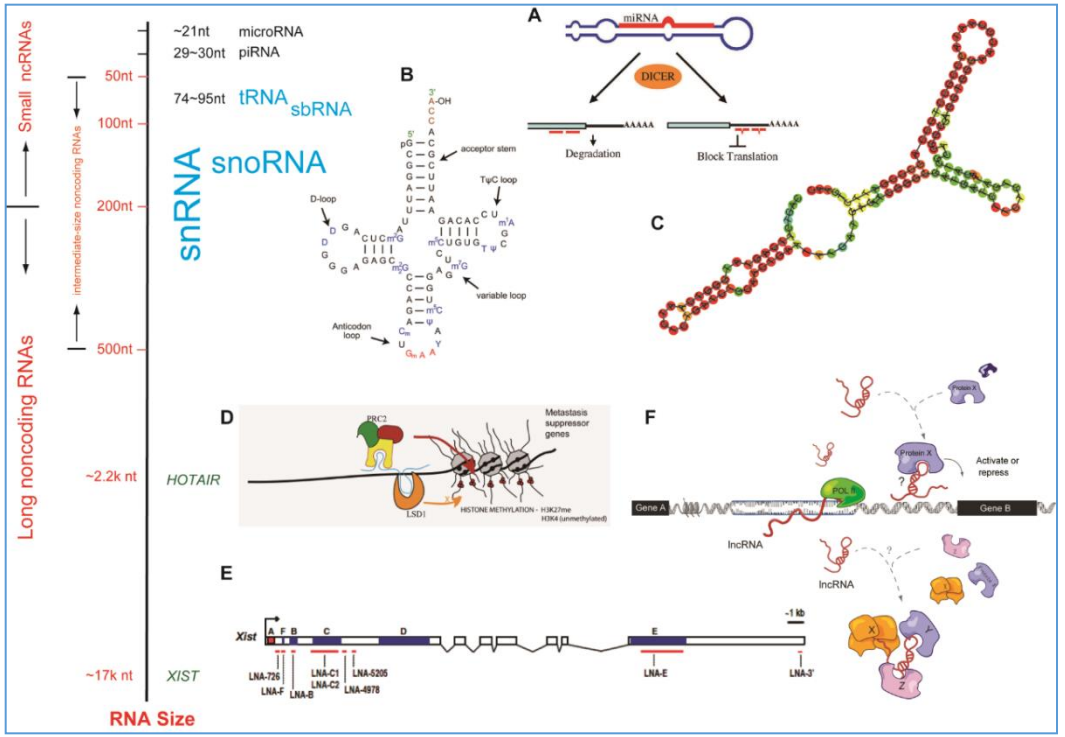
非编码 RNA 及分类

➢ 基础结构型ncRNA

- 核糖体 RNA（ribosomal RNA，rRNA）
- 转运 RNA （transfer RNA，tRNA）
- 小核 RNA （small nuclear RNA，snRNA）
- 小核仁RNA（small nucleolar RNA，snoRNA）
- 端粒相关RNA（telomere-associated RNA，TERC/TERRA）

➢ 调控型ncRNA

- 环状RNA （circular RNA，circRNA）
- 长非编码RNA（long non-coding RNA，lncRNA）
- 小非编码RNA
  - 微小RNA （microRNA，miRNA）
  - 小干扰RNA（samll interfering RNA，siRNA）
  - PIWI 相互作用 RNA （PIWI-interacting RNA，piRNA）

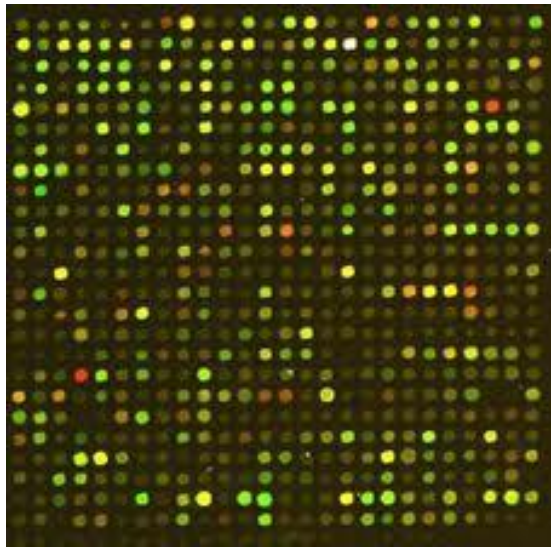


非编码RNA类型

1.2 转录组学的研究方法

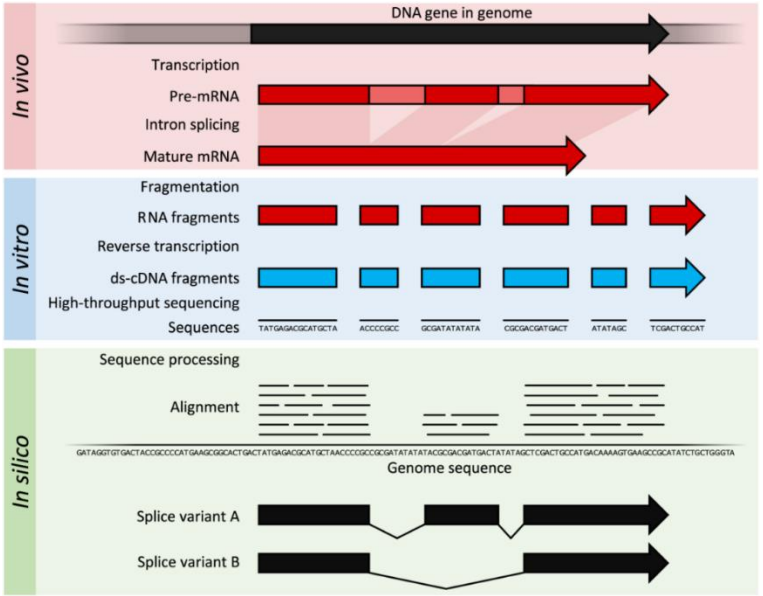
转录组学技术是用于研究生物体转录组（其所有 RNA 转录本的总和，包括非编码RNA）的技术。

➢ 基因芯片

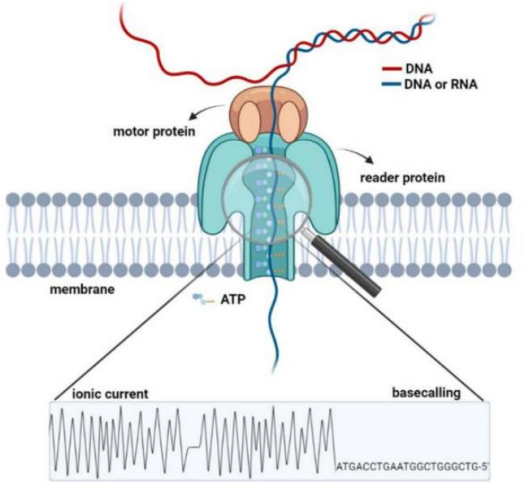


➢ RNA-seq

- 从生物体中提取 mRNA，将其分解并复制成稳定的双链 cDNA。使用高通量、二代短读测序方法对 ds-cDNA 进行测序。然后可以将这些序列与参考基因组序列比对，以重建正在转录的基因组区域。
- 第三代测序 (TGS)，长读长技术简化从头基因组组装的过程，无需组装或使用复杂的生物信息学工具即可识别全长转录本



二代测序技术



三代测序技术:纳米孔测序

(Athanasopoulou K, et al. Life (Basel), 2021)
(Lowe R, et al. PLoS Comput Bio, 2017)

1.2 转录组学的研究方法

（一）细胞解离

（二）RNA提取

（三）RNA富集或其他RNA去除

一般生物体中的RNA中，rRNA占绝大多数。在人体中，mRNA一般只占到2%。根据研究目的进行RNA类型的筛选。

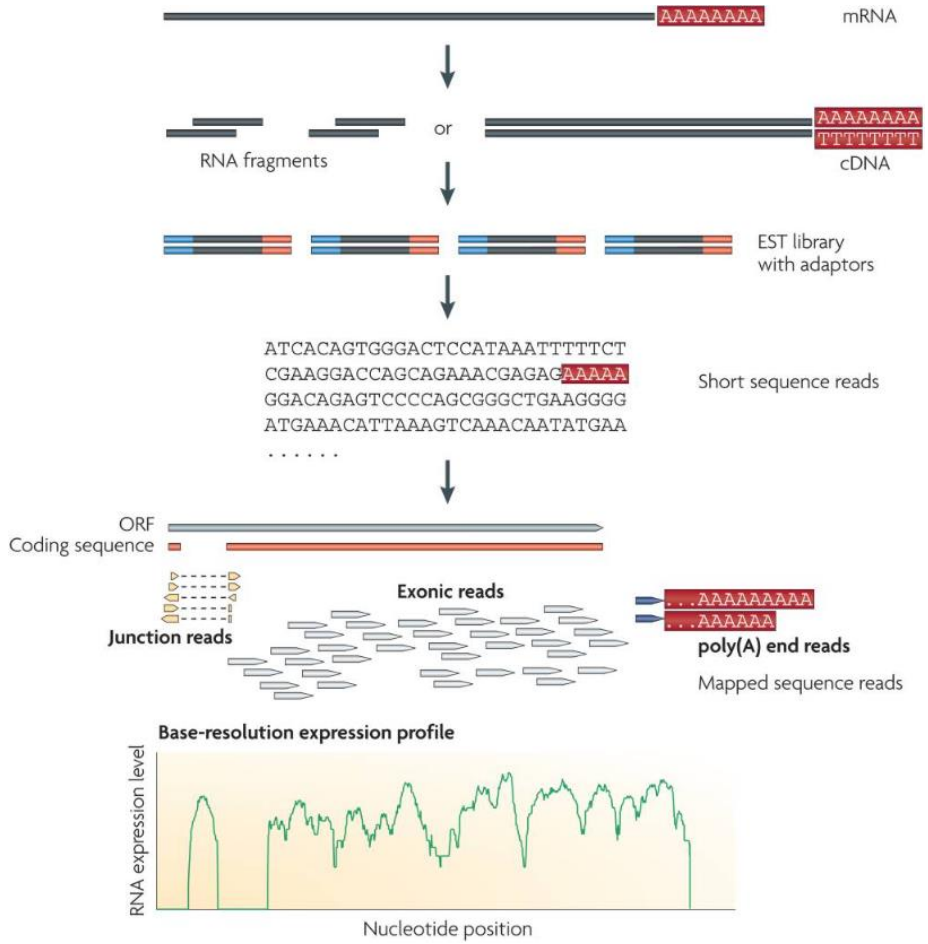
（四）片段化与cDNA合成

与miRNA、piRNA、siRNA和许多其他可以在接头连接后直接测序的小RNA不同，较大的RNA分子必须片段化为较小的片段(200-500 bp)才能与大多数深度测序技术兼容。对于不同测序技术，要处理成的片段长度不尽相同。

（五）建库与文库质检

一旦合成cDNA，就可以按照DNA建库流程去进行文库制备和质检。

（六）平台测序



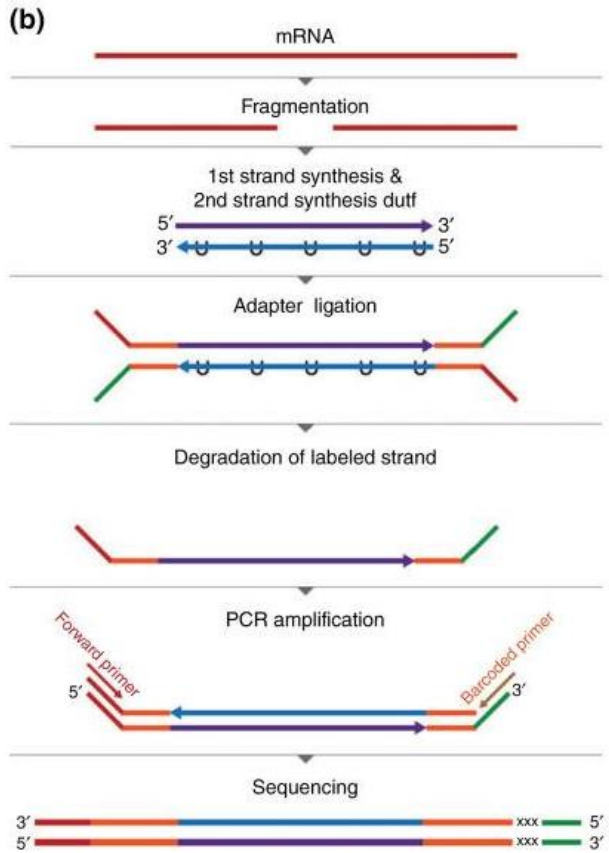
(Wang Z, et al. Nat Rev Genet, 2009)

1.3 特殊测序建库方法

RNA-Seq 作为转录组学常用的研究方法，是研究基因表达和识别新的转录异构体首选方法。除了用于基因表达的一般分析外，已演绎出针对新生RNA、mRNA、环状RNA、非编码RNA、小RNA等不同类型RNA分子的高通量测序技术，还有几种特定应用的 RNA-Seq 方法，包括链特异性测序、靶向测序等。

链特异性测序建库

最常用的方法是在第二链cDNA合成中使用三磷酸脱氧尿 (dUTP) 代替dTTP，之后利用酶（如Uracil-DNA Glycosylase, UDG）消化含有U的DNA链，从而保留一条链的信息。



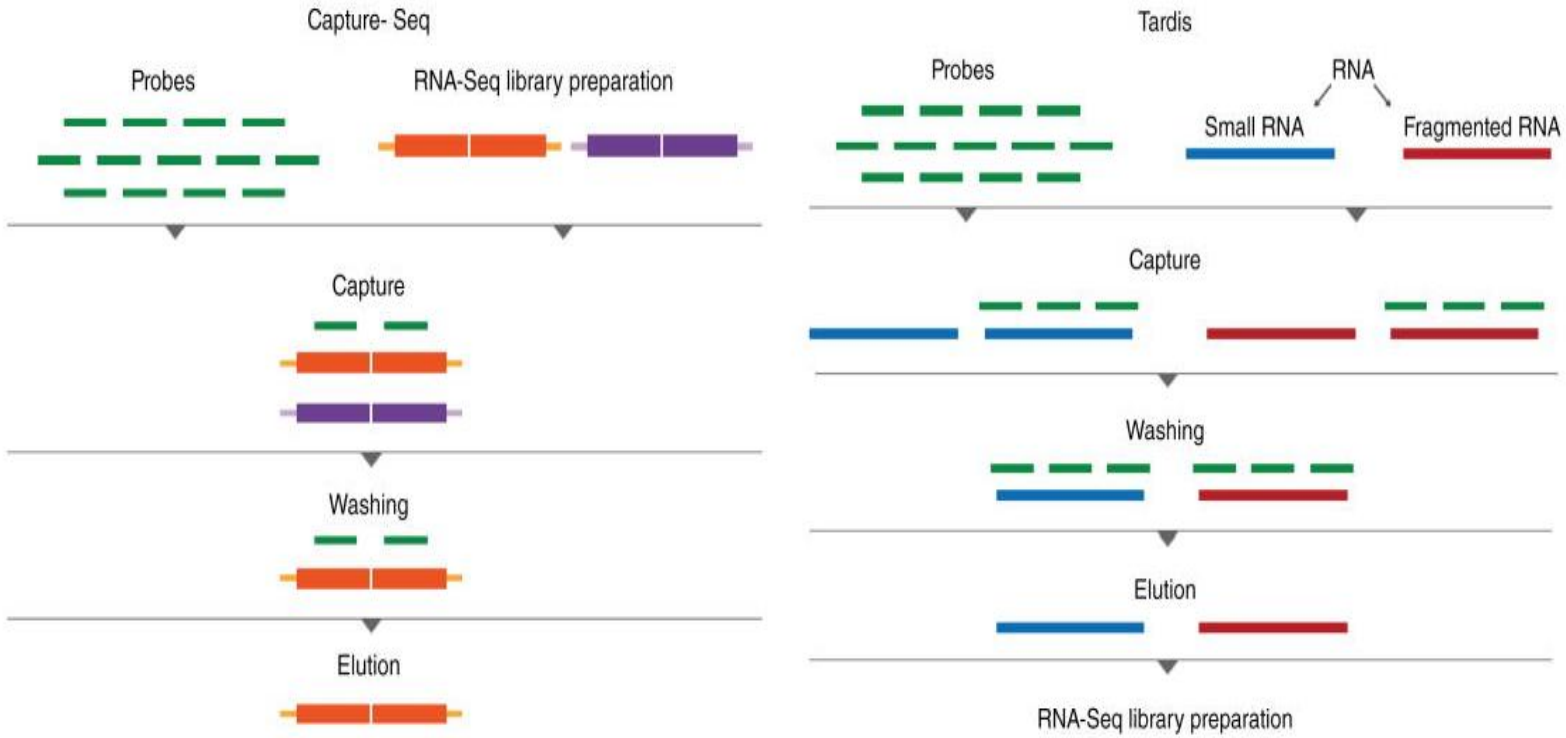
(Hrdlickova R, et al. Wiley Interdiscip Rev RNA, 2017)

1.3 特殊测序建库方法

靶向RNA测序建库

最常见的是使用专门设计的、与目标RNA互补的探针捕获特定的mRNA或IncRNA，这些探针可以是生物素标记的，使得目标RNA可以通过亲和柱（如磁珠）轻松分离。

另一种方法通常在目标区域较短时使用，通过设计特异性的引物直接对这些区域进行PCR扩增，这些引物将特异性地结合到目标cDNA的两端，用于扩增特定的基因或转录本区域。



两种靶向 RNA-Seq。Capture-Seq 方法基于通过将 RNA-Seq 文库与 DNA 寡核苷酸探针杂交来捕获感兴趣的区域。TARDIS 基于将输入 RNA 与 DNA 寡核苷酸探针杂交。

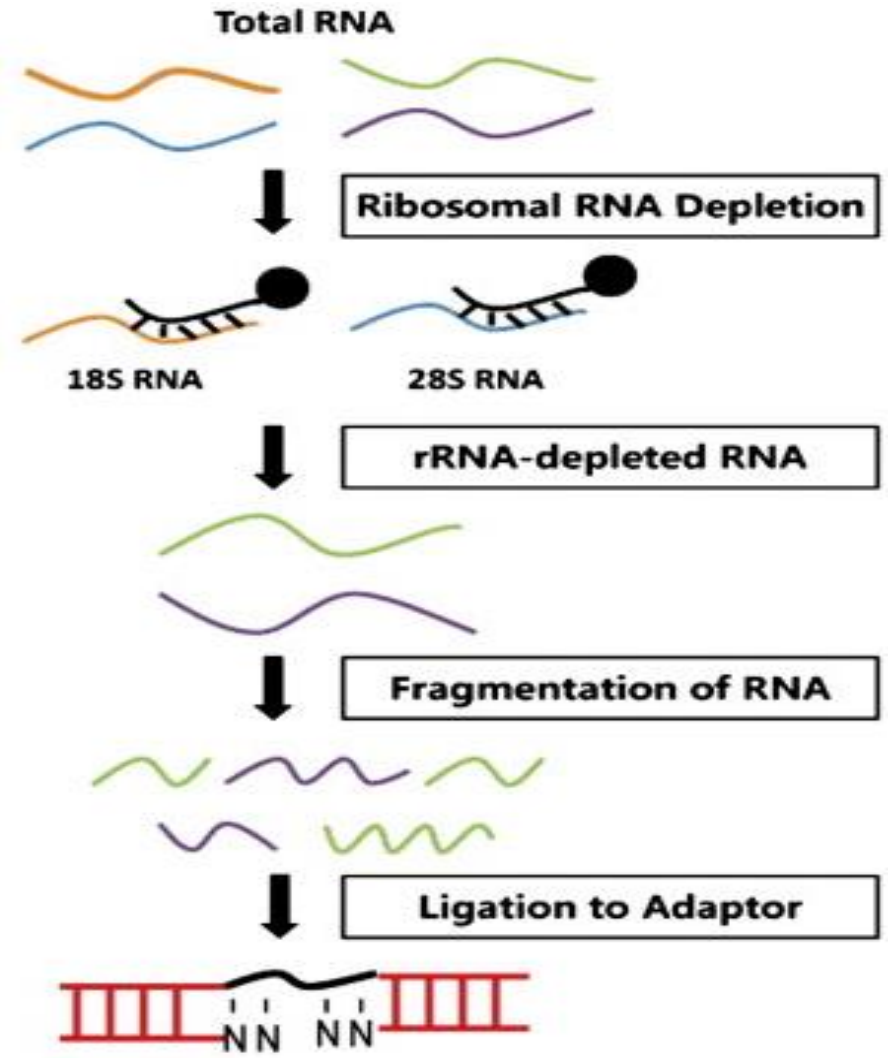
(Hrdlickova R, et al. Wiley Interdiscip Rev RNA, 2017)

1.3 特殊测序建库方法

非编码特异性建库测序—lncRNA

由于常规RNA-Seq测序以检测mRNA为主，只关注携带polyA的转录本，会丢失不含polyA的lncRNA；

为了便于lncRNA的研究，发展了专门的测序技术尽可能完整的检测各类lncRNA，如去除核糖体RNA（Ribo-minus）等提取手段。另外，链特异性的RNA-Seq能保留RNA的链方向信息，可以更准确的鉴定lncRNA，并且更好的与antisense重叠的编码基因进行区分。



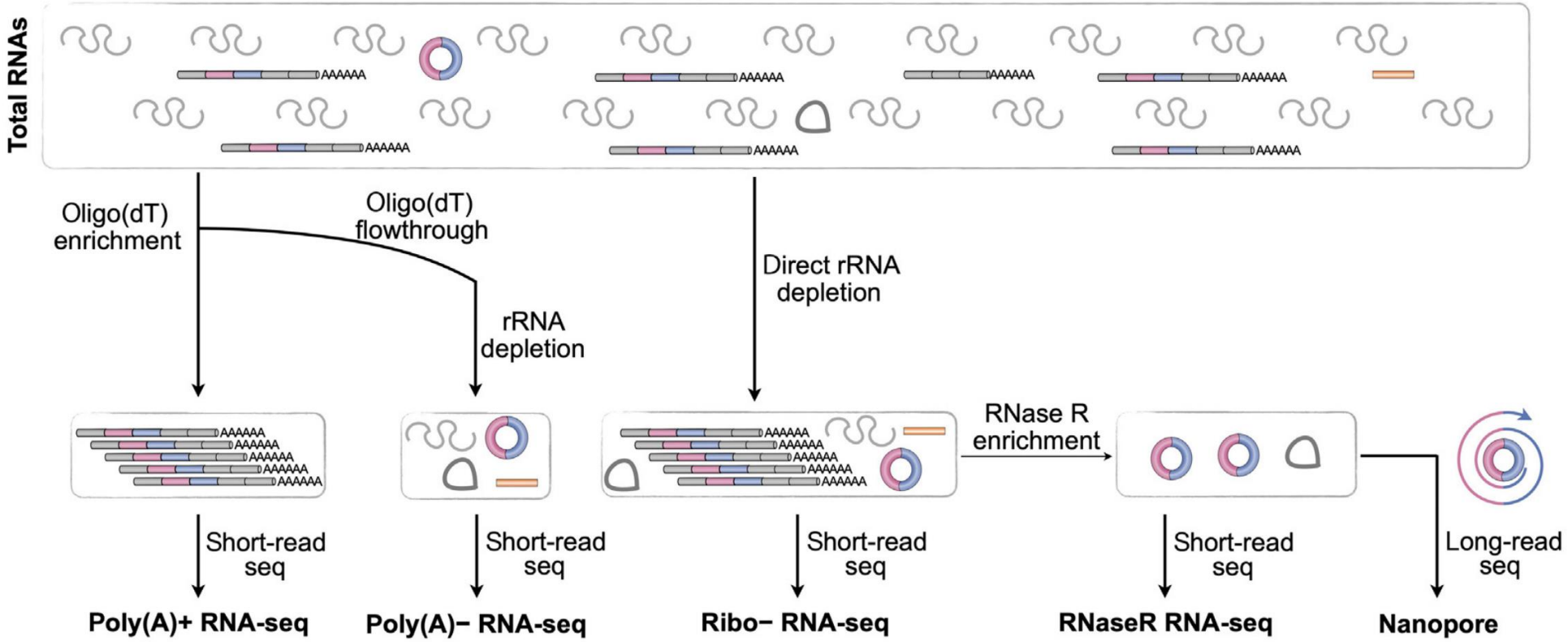
ribo-minus RNA-Seq流程图

(Cui P, et al. Genomics, 2010)

1.3 特殊测序建库方法

非编码特异性建库测序—环RNA

- 比如，使用探针结合并去除rRNA、使用RNase R消化linear RNA。
- RNase R：一种来源于大肠杆菌的核酸外切酶，沿RNA的3’-5’方向切割并降解RNA，能够消化几乎所有的linear RNA分子，但不易消化circRNA。



(Ma XK, et al. Trends Genet, 2010)

1.3 特殊测序建库方法

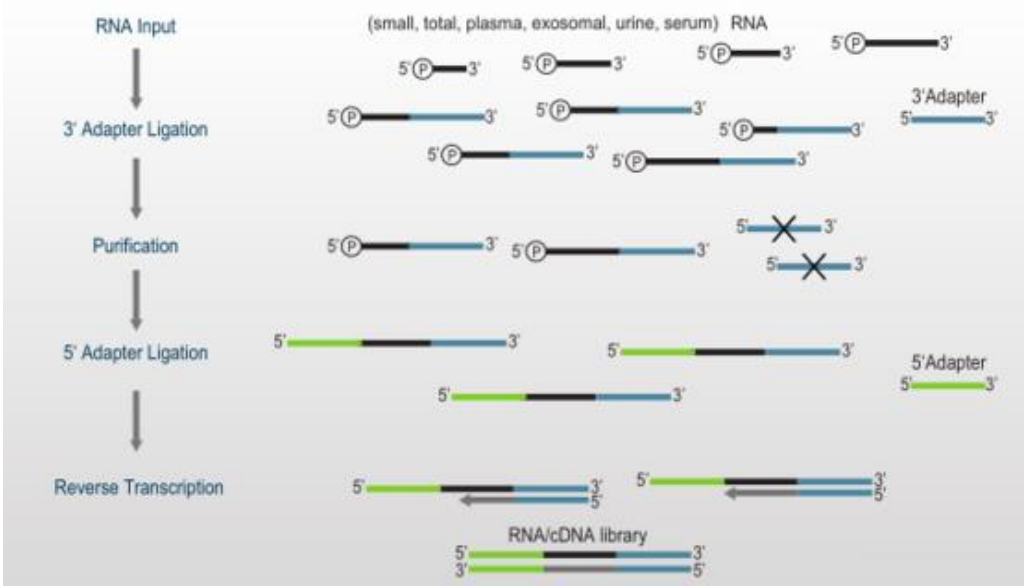
非编码特异性建库测序—小RNA

- 小RNA-seq

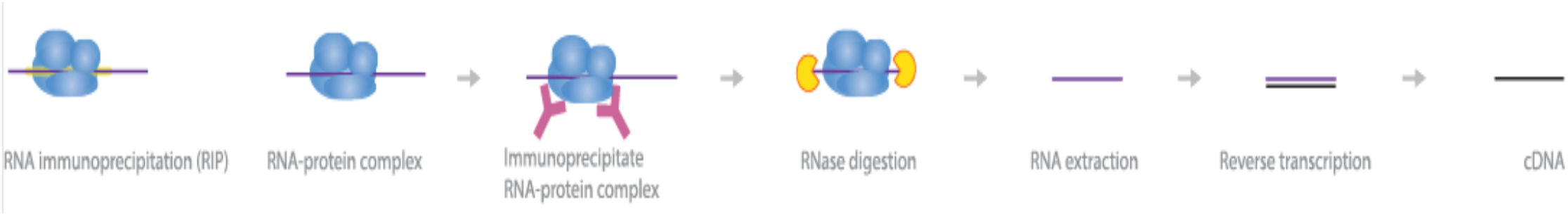
提取RNA后，选择长度在17-25nt的RNA进行测序

- 富集piRNA小RNA-seq

oxidized sRNA-seq，经NaIO4处理后再进行小RNA建库测序



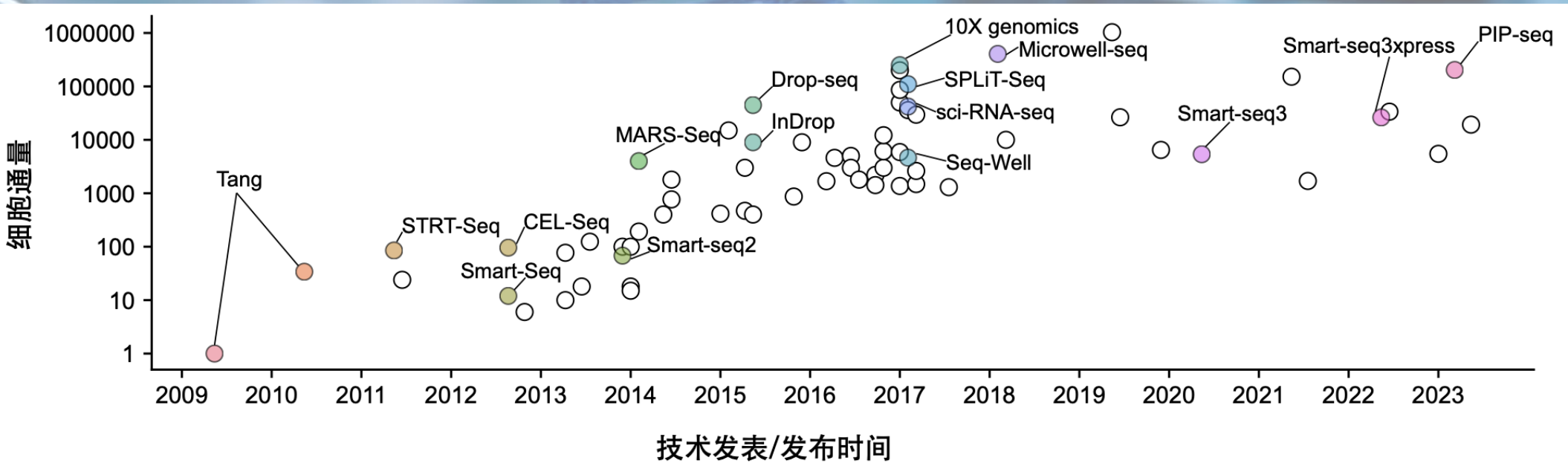
- RIP-seq(AGO2,Piwi等)



- CLIP-seq（AGO2,Piwi等）

(Wu X, et al. Mol Ther Nucleic Acids, 2020)

1.4.1 单细胞转录组

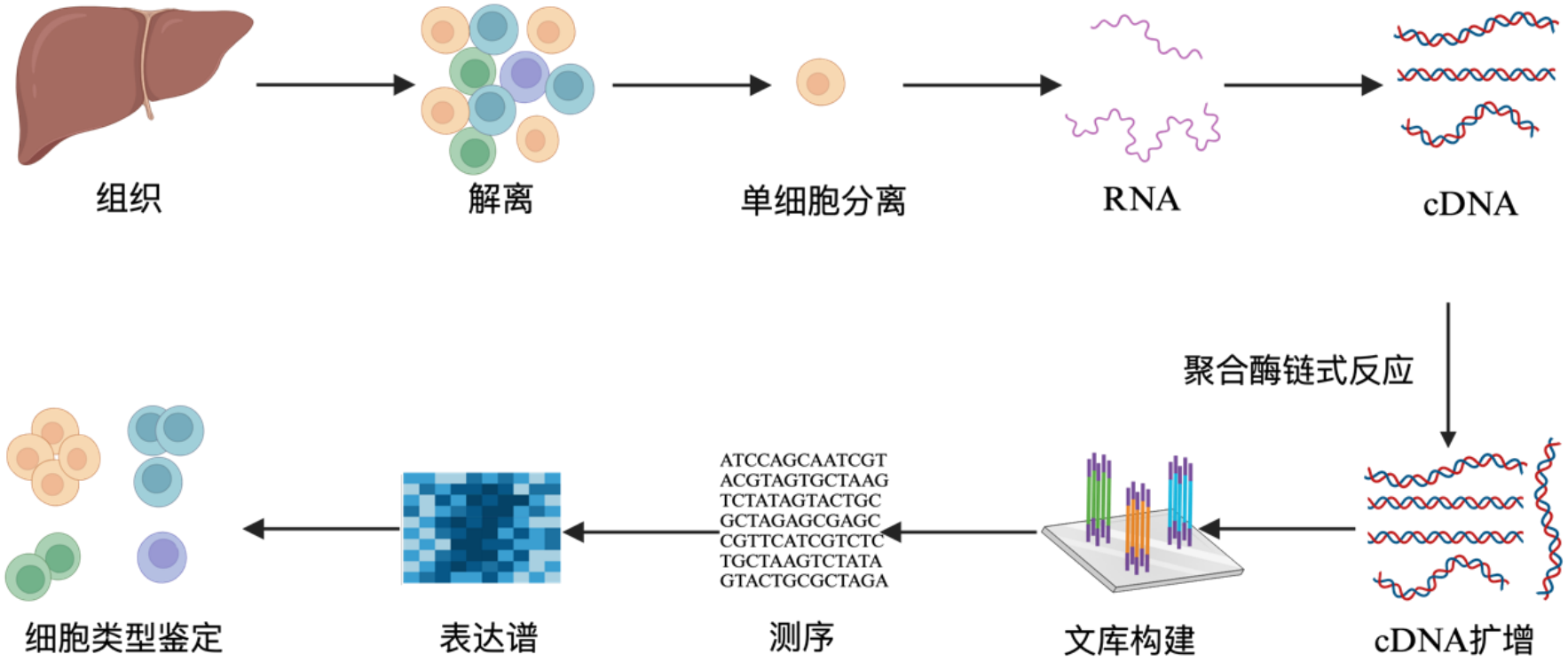


单细胞转录组学技术发展

single-cell RNA sequencing

2009年，研究者在Nature Methods杂志首次报道了单细胞转录组测序技术。此后，许多更为新颖、灵敏、准确的单细胞转录组测序技术相继被开发和应用，以更快的速度和更低的成本提供了丰富的生物学信息。

1.4.1 单细胞转录组

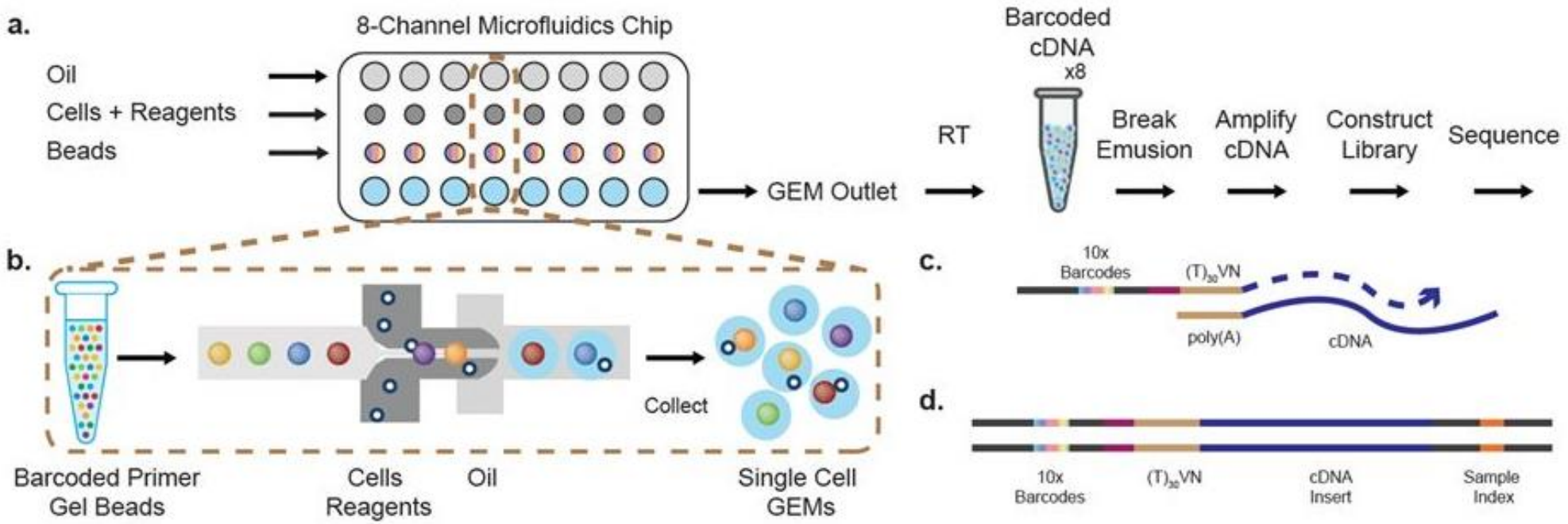


单细胞转录组测序流程

包括单细胞悬液制备、单细胞分选、cDNA扩增、文库构建和测序等步骤

1.4.1 单细胞转录组

scRNA-seq技术：10x Genomics

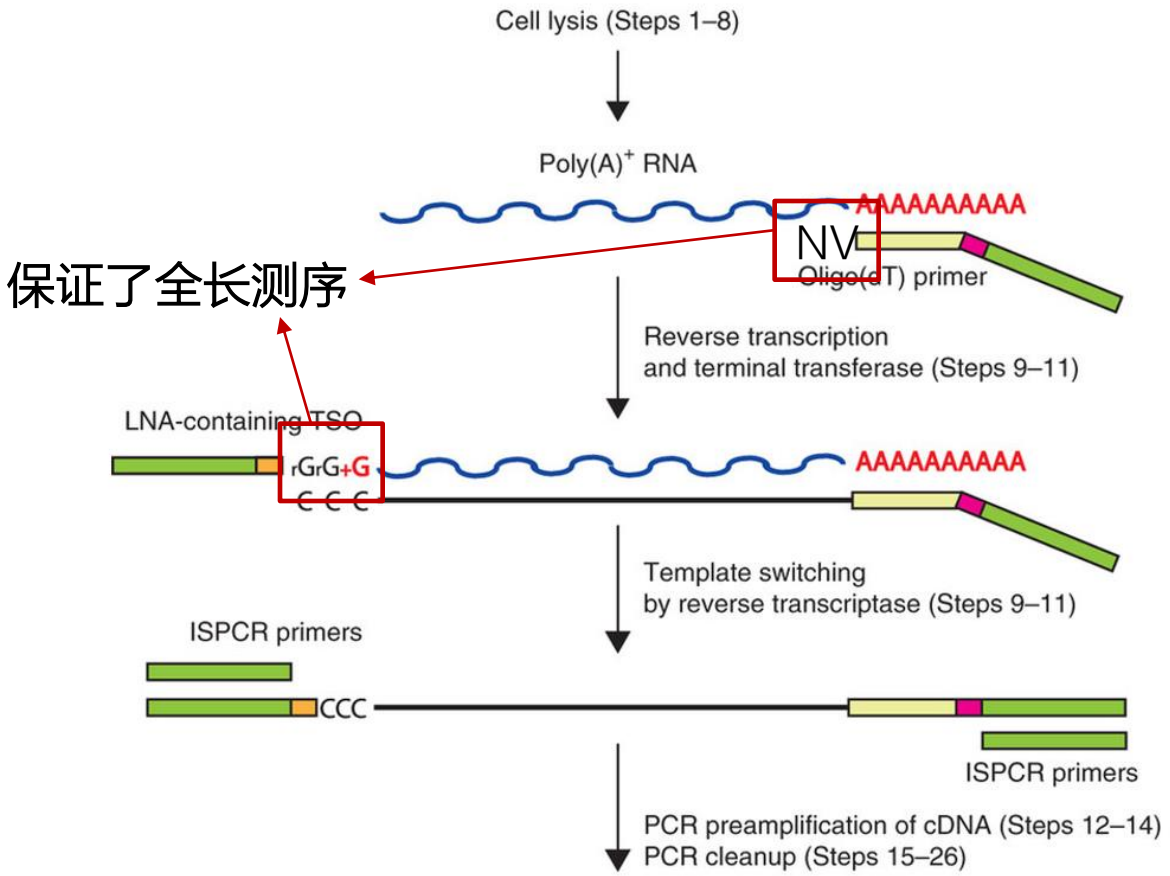


将样本解离成单细胞悬液，在微流控制芯片中通过液压的方式推进。在第一个进样孔进入酶（反转录的酶），结合凝胶微珠；第二个进样孔进入油滴，包裹细胞、酶、凝胶微珠，形成油包水的微体系，在这个微体系中细胞进行裂解、反转录，之后再进行后面的建库分析。

(Choi JR, et al. Cells, 2020)

1.4.1 单细胞转录组

SMART-seq：全长cDNA测序



- ✓ **设计特殊的3’端定位引物**
  引物NV末端结构保证cDNA的合成从mRNA的3’最末端开始; 合成的cDNA在下游连上通用PCR引物
- ✓ **设计特殊的5’端定位引物**
  利用MMLV逆转录酶合成碱基的特点（末端多加CCC），设计带通用PCR引物以及GGG的引物，保证mRNA 5’端的完整性
- ✓ **通用PCR引物的一致性，减少PCR扩增偏好**

1.4.1 单细胞转录组

SMART-seq2 vs. 10x Genomics

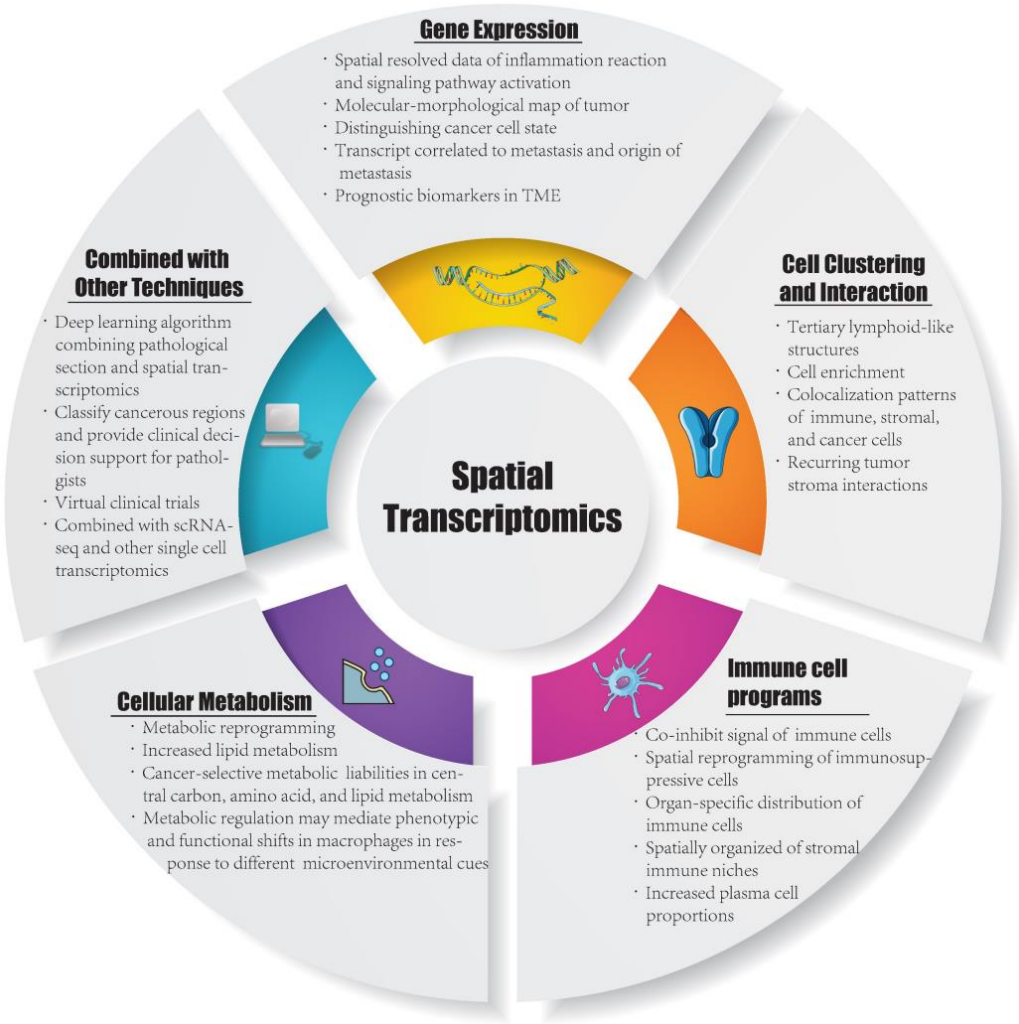
	SMART-seq2	10x Genomics
Technology	Microfluidic plate	Microfluidic droplet
Number of cells per sample	96/800 cells A limited number of cells depending on C1 IFC	500 -10,000 cells A large number of cells
Number of read per cell	100 -1,000 million reads Uniform among cells	5000 -10,000 reads Diverse among cells
Sequencing	Full-length (96 cells)	3'-end
Cell size	5-25 μm Depending on C1 IFC	<40 μm
Sequencing library	Separate Can resequencing the user's selected cells	Mixed
	For individual cells in detail	For individual cells in a population

- Cell suspension（细胞悬液）and nuclei suspensions（核悬液）samples can be studied.
- Recovers up to ~65% of cells; Low doublet rate (~0.9% per 1,000 cells).
- For the most common applications, 50,000 or more reads per cell should be sequenced.
- It is possible to work with cryo-preserved cells, enabling safe sample shipping and batching.

(Kashima Y, et al. Exp Mol Med, 2020)

1.4.2 空间转录组

- 常规的单细胞转录组学技术在进行测序前将细胞解离成单细胞悬液，然后利用单细胞分离技术（微孔，微板，液滴）等方法进行单细胞建库，在这个过程中细胞失去了原本在组织的空间信息。
- 然而在一些研究，比如探究细胞命运机制及细胞谱系发生时，空间位置的信息尤为重要。所以需要在单细胞的基础上，联合空间转录组，二者相互印证，才能获取更真实的信息。



空间转录组学的应用

(Du J, et al. Int J Biol Sci, 2023)

1.4.2 空间转录组

空间转录组方法的分类

以靶标的最终读数以及如何获得空间内容为依据，可以分为以下几种：

1. 基于测序的方法

- 依赖于基于空间条形码DNA的全转录组/基因组测序。

2. 基于探针的方法

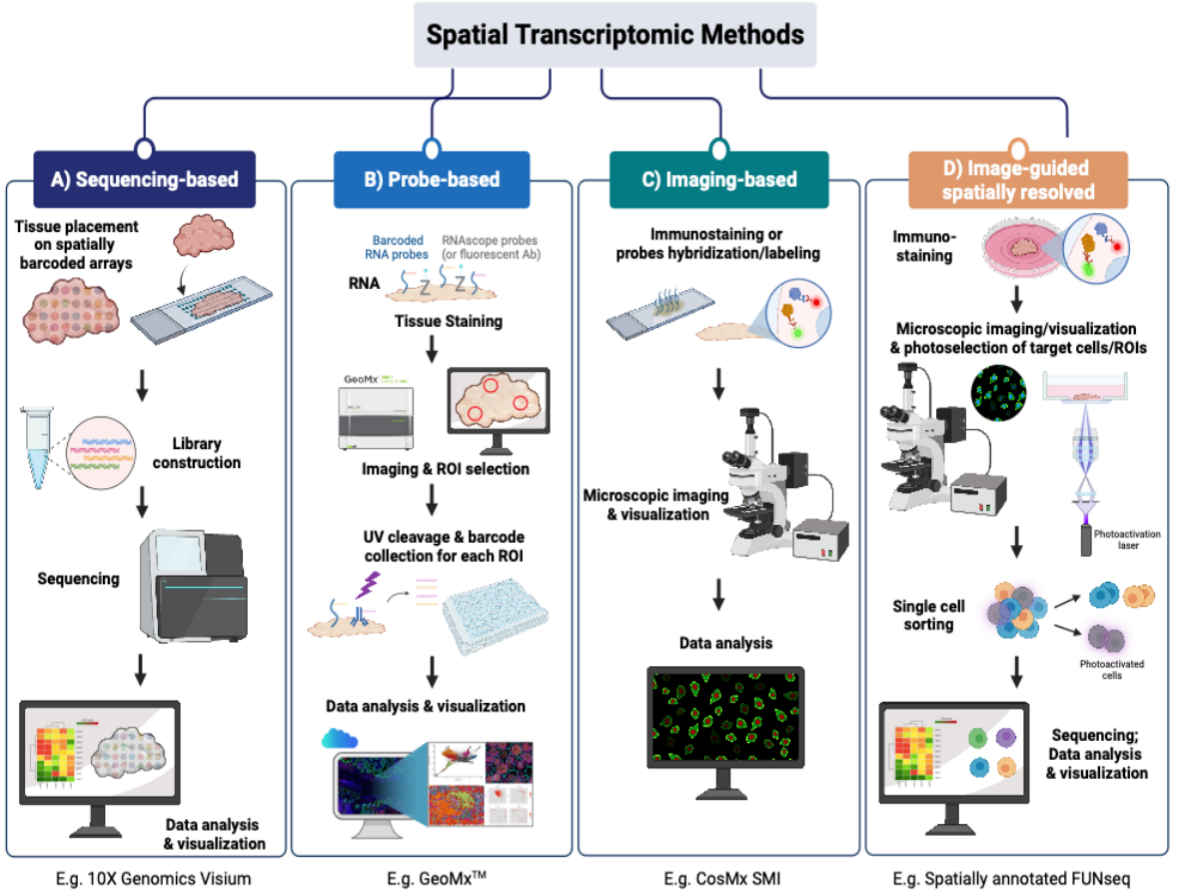
- 需要对已知目标的条形码探针进行计数。

3. 基于成像的方法

- 依赖于重复的成像周期来获得已知目标的最终读数。

4. Image-seq

- 原位和体内成像引导的空间分辨率单细胞测序方法。Image-seq基于显微图像选择和分离感兴趣区域 （ROI） 中的单细胞。



空间转录组学的分类

(Chen T, et al. Cells, 2023)

第七章　转录组学

——第二节　转录组数据的基础分析

2.1 转录组的组装

2.2 转录组的定量

2.3 转录组差异表达分析

2.4 转录组聚类分析

2.5 转录组降维分析

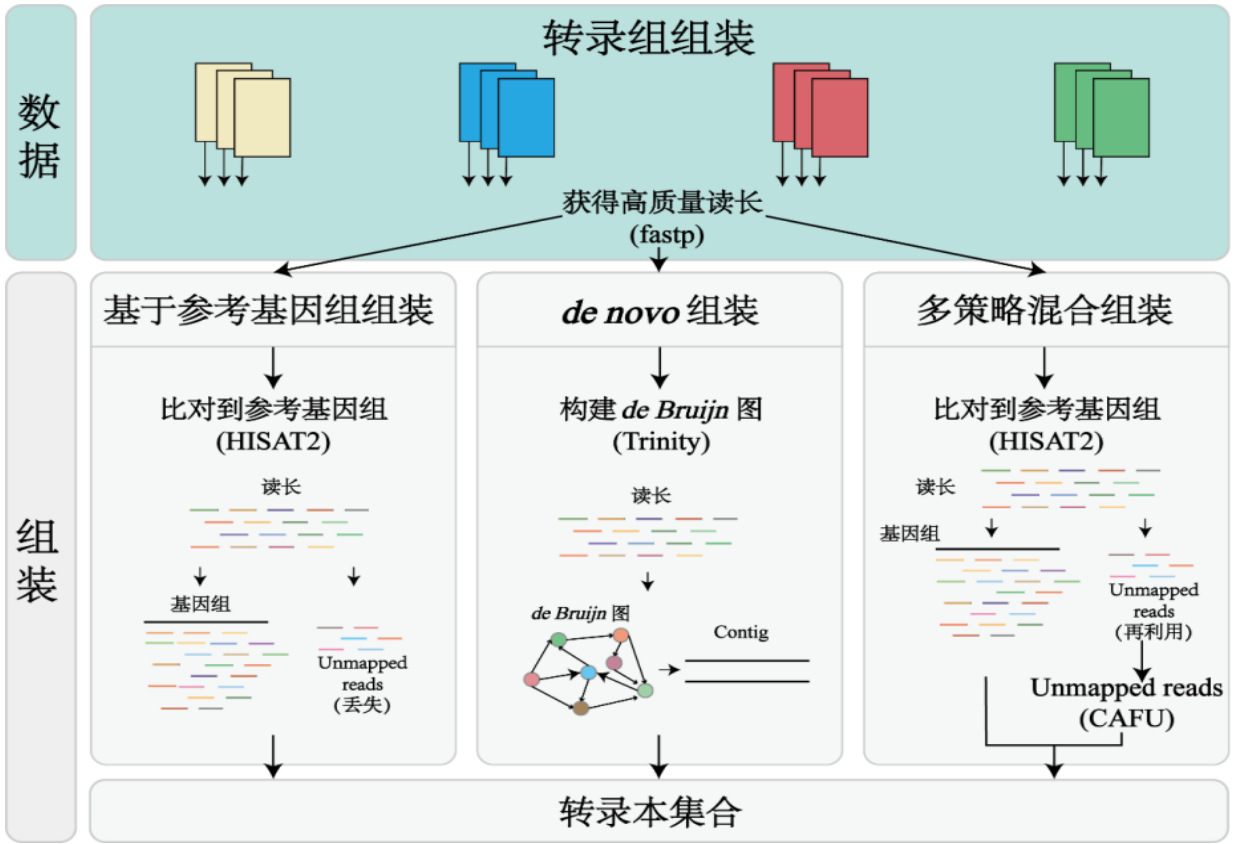
2.6 共表达网络分析

2.7 转录调控网络分析

2.8 基因的功能富集分析

2.1 转录组的组装

基于参考基因组, 从头组装，混合策略……



三种组装策略示意图

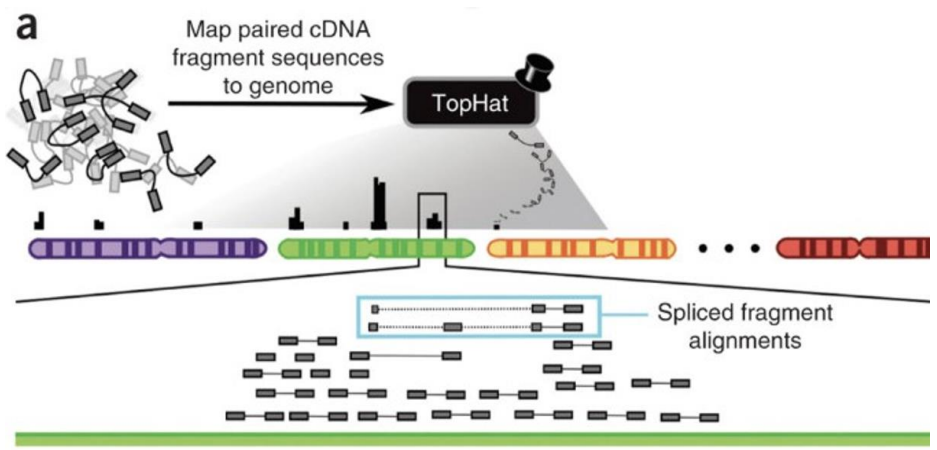
2.1 转录组的组装

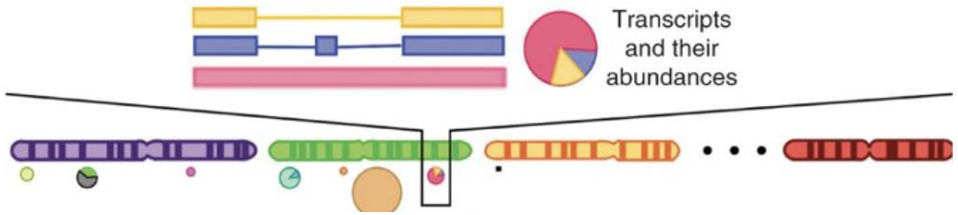
组装策略	优点	缺点
基于参考基因组组装	可靠度高；所需计算资源较少；能注释出新的转录本等	必须依赖于已知基因组信息；组装结果受基因组组装质量、序列比对软件的性能等多因素影响
从头组装	不需要提供参考基因组序列，直接从测序数据本身入手，可以避免低质量基因组组装等带来的误差和错误	受测序错误、测序偏差、同一基因不同转录本间序列高度相似等因素影响，组装结果存在片段化、错误率高等特征
多策略混合	更高的转录本覆盖度； 有助于发现新的、未知的基因	组装过程变得更加复杂，计算和存储需求高，结果的解释更加困难

2.2 转录组的定量

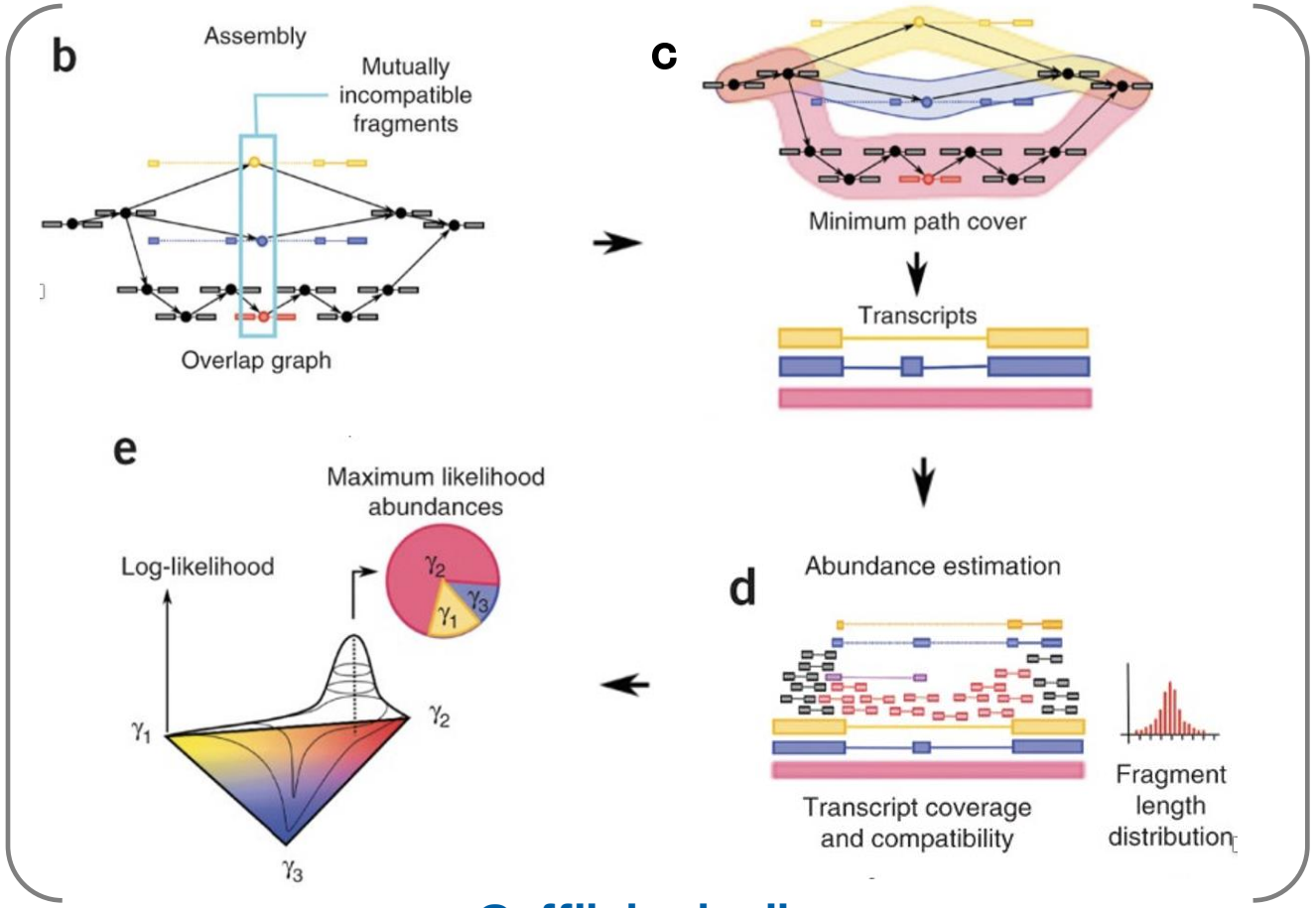
转录组定量分析常以测序数据与基因组的比对结果为基础，量化表达量。常见的软件有HTSeq，BEDTools，FeatureCounts，StringTie，Qualimap， Cufflinks等。

Mapping





Result



Cufflink pipeline

Cufflinks流程

(Trapnell C, et al. Nat Biotechnol, 2010)

2.2 转录组的定量

基因水平的表达定量

- RNA（或cDNA）分子在测序之前先进行片段化，较长的转录本会比较短的转录本被剪切成更多的片段。因此，转录本的reads数不仅与其表达水平成正比，而且与其长度成正比。
- 为了消除基因长度产生的固有技术误差，在过去十年中，已针对RNA-seq数据开发了许多归一化方法，目前，常采用TPM（Transcripts Per Million）、RPKM (Reads Per Kilobase Million)、 FPKM (Fragments Per Kilobase Million)估算基因水平的表达值.

$$\text{TPM of a gene: } \mathbf{TPM} = A \times \frac{1}{\sum(A)} \times 10^6 \text{ Where A} = \frac{\text{Total reads mapped to gene} \times 10^3}{\text{Gene length in bp}}$$

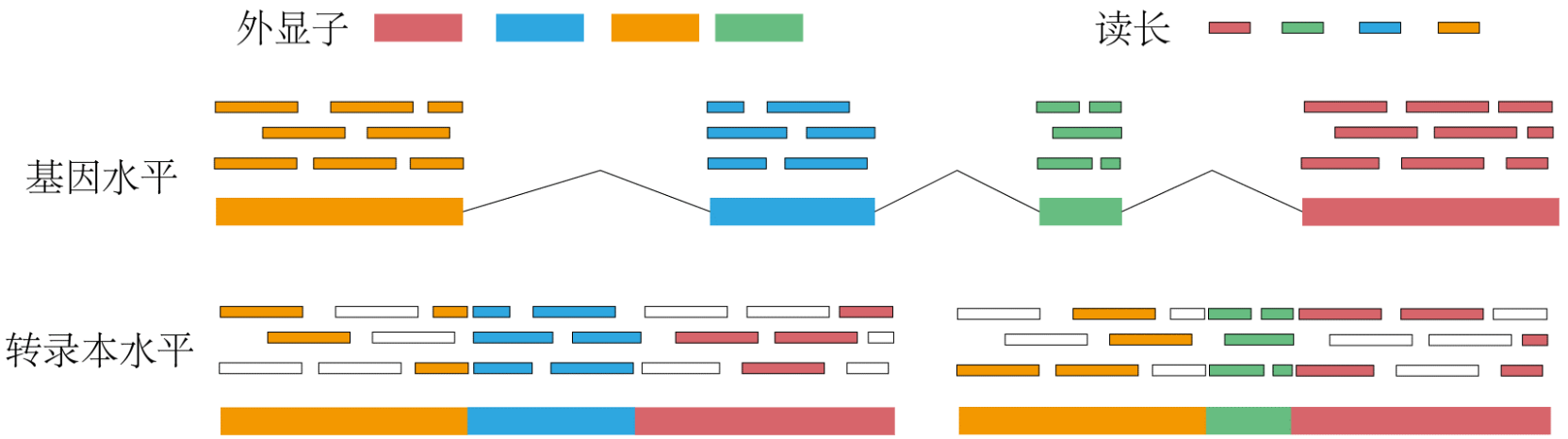
$$\text{RPKM of a gene: } \mathbf{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{Gene length in bp}}$$

$$\text{FPKM of a gene: } \mathbf{FPKM} = \frac{\text{Number of fragments mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{Gene length in bp}}$$

2.2 转录组的定量

转录本水平的表达定量

- 基因水平的表达分析常常忽略了同一基因不同转录本的表达差异。
- 以StringTie为例，其使用最优插入位置算法寻找剪接位点（junction），通过分析读长在junction上的分布模式来推断转录本是否存在和相对丰度。这种方法能够处理多比对和重叠比对的读长，更准确地识别转录本。



基因水平和转录本水平表达定量示意图

2.2 转录组的定量

转录组定量结果的质量控制

- 异常值（Outlier）。

  Outlier的处理通常基于统计学方法，如离群值检测算法，或根据特定实验设计和样本特征对异常值进行检测。然后通过删除、替换异常值或者转换数据（对数转换或归一化）对异常值进行处理，保证表达矩阵的可靠性。

- 批次效应（batch effect）

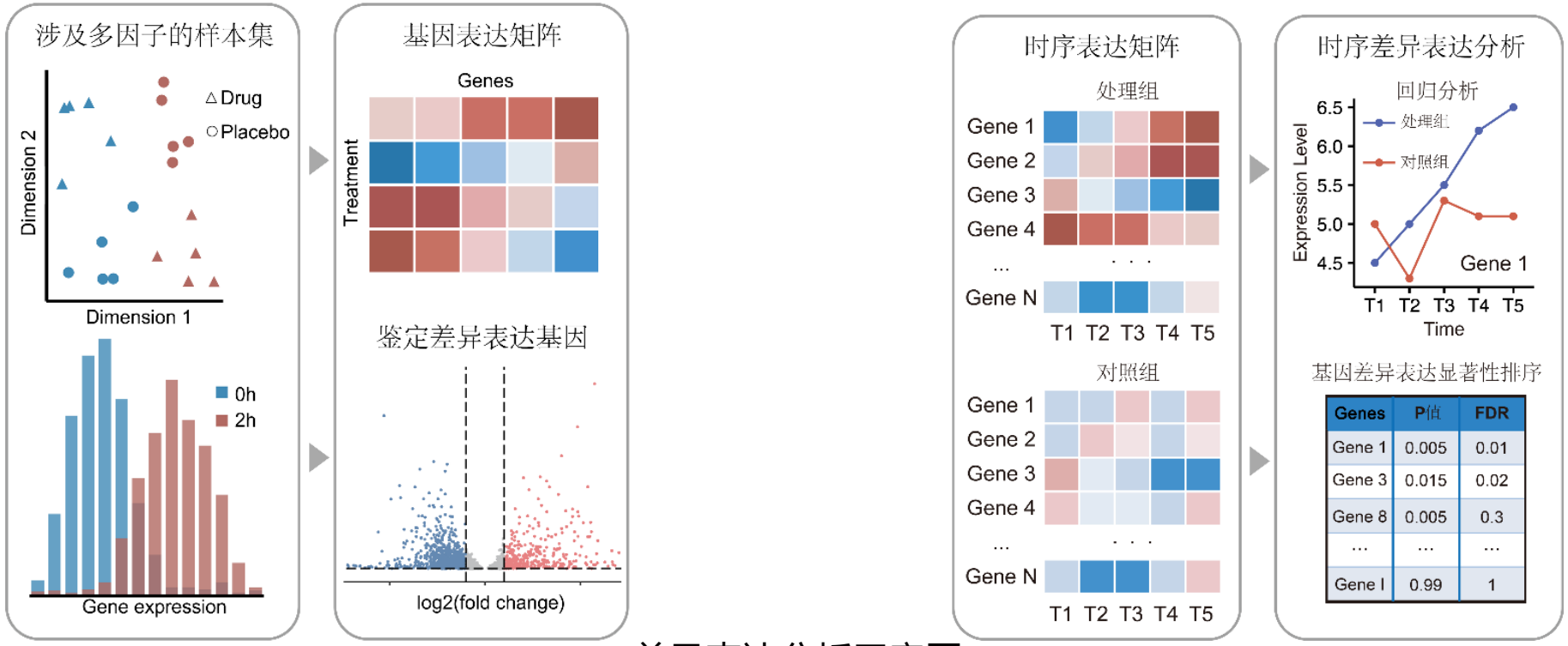
  通过sva包的combat、 limma R包removeBatchEffect函数， RUVseq和svaseq等可一定程度消除批次效应。在批次矫正之后，可以使用RLE（Relative Log Expression）等指标来评估不同批次的相同样本之间的相似性。消除批次效应后的数据，其 RLE 分布应接近于零，且在不同样本间相似。

2.3 转录组差异表达分析

差异表达（DE）基因分析：确定不同样本组的平均表达水平是否存在显著差异。

在进行差异表达基因/转录本鉴定时，通常采用两个指标进行筛选：（1）表达值的差异倍数（Fold Change，FC）；（2）错误校验率（False Discovery Rate，FDR），经统计学方法校正后的P值。

- 配对样本的转录组差异表达分析
- 时序样本的转录组差异表达分析

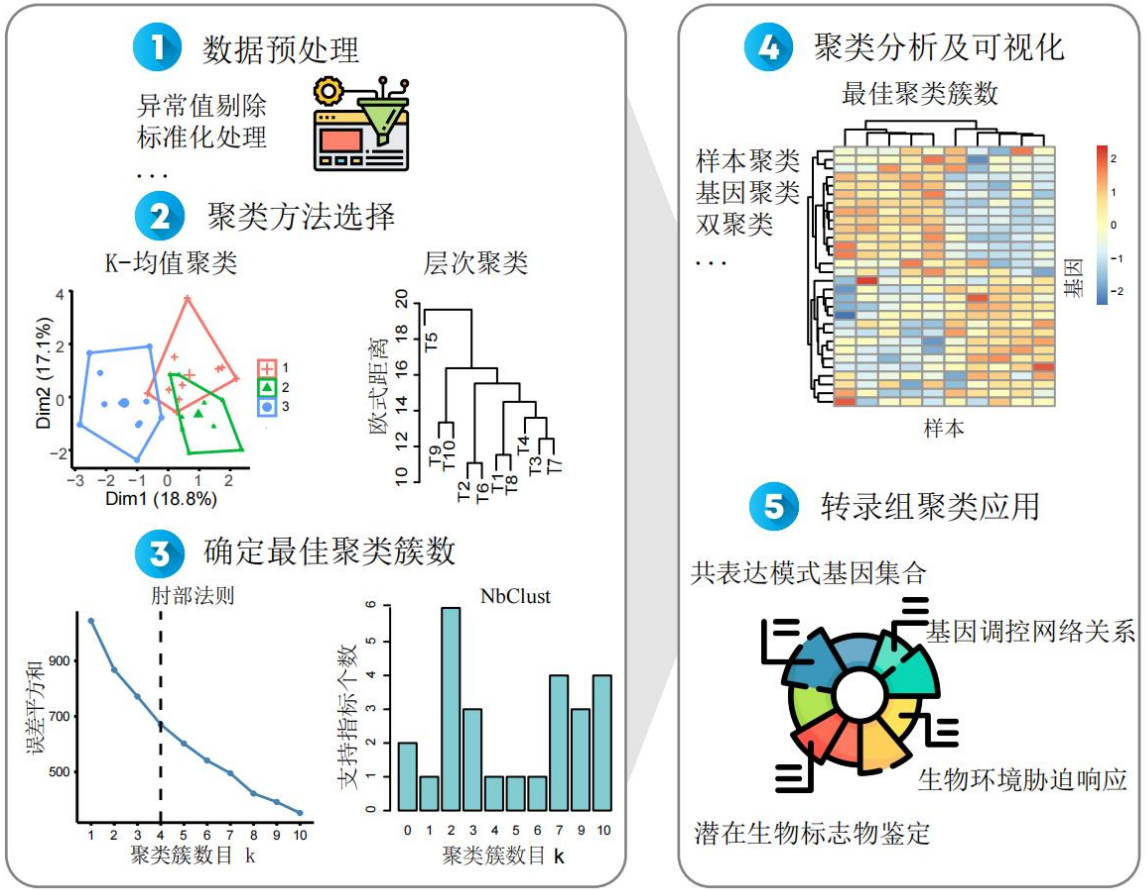


差异表达分析示意图

2.4 转录组聚类分析

对于一个给定的基因表达矩阵，可进行样本聚类、基因聚类、双聚类（Biclustering）、集成聚类（Integrative clustering）等不同角度的聚类分析。

- 样本聚类将不同实验条件或不同时间状态下的样本依据基因表达模式进行聚类，R stats软件包中hclust和kmeans函数可分别依据层次聚类（Hierarchical clustering）和K-均值聚类（K-means clustering）算法对样本进行聚类。
- 基因聚类可将表达模式高度相似的基因聚为一簇，有助于探究特定信号通路中的共同调节基因。
- 双聚类可同时对基因和样本进行聚类，以识别在特定生物条件下共同表达的基因集合和样本集合，此过程可用pheatmap包来实现。
- 集成聚类则是将多种聚类算法整合起来，用于将来源于不同平台、不同类型特点的组学数据信息进行聚类，以更全面、更系统的角度深入理解组学数据之间的关联与互作。



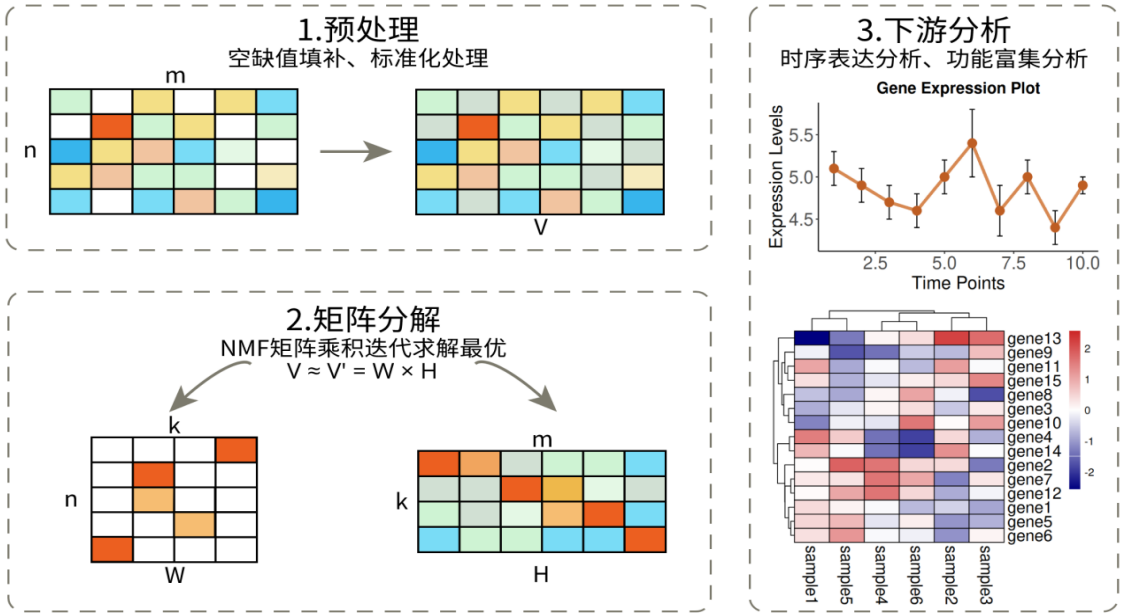
转录组聚类分析流程图

2.5 转录组降维分析

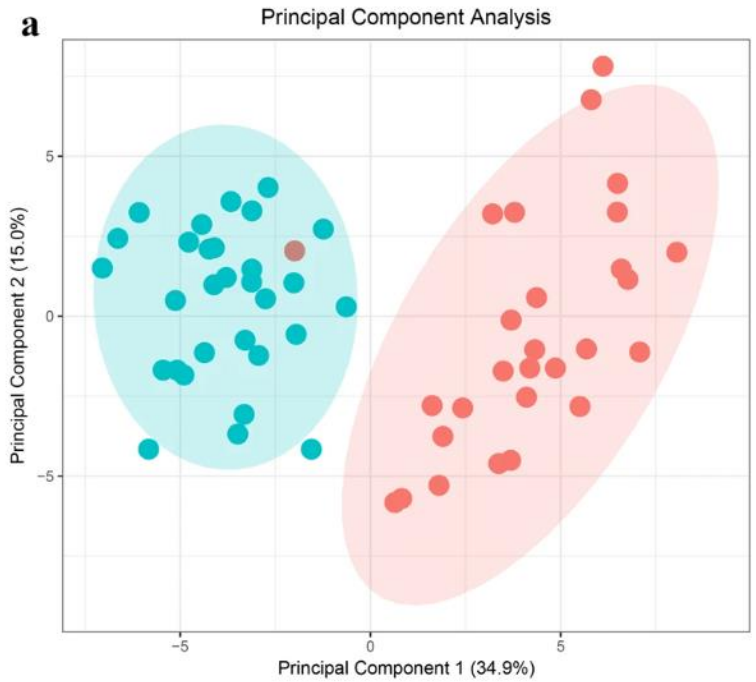
降维分析是通过线性或者非线性变换将高维数据映射到低维空间，在尽量保留原有数据的信息基础上，通过减少特征变量数目和降低数据复杂度，实现对数据的压缩处理和特征聚类。

常见的降维算法及工具:

- PCA（prcomp和princomp函数）
- NMF（CoGAPS包、easyMF）
- t-SNE（Rtsne包）…



NMF矩阵分解原理

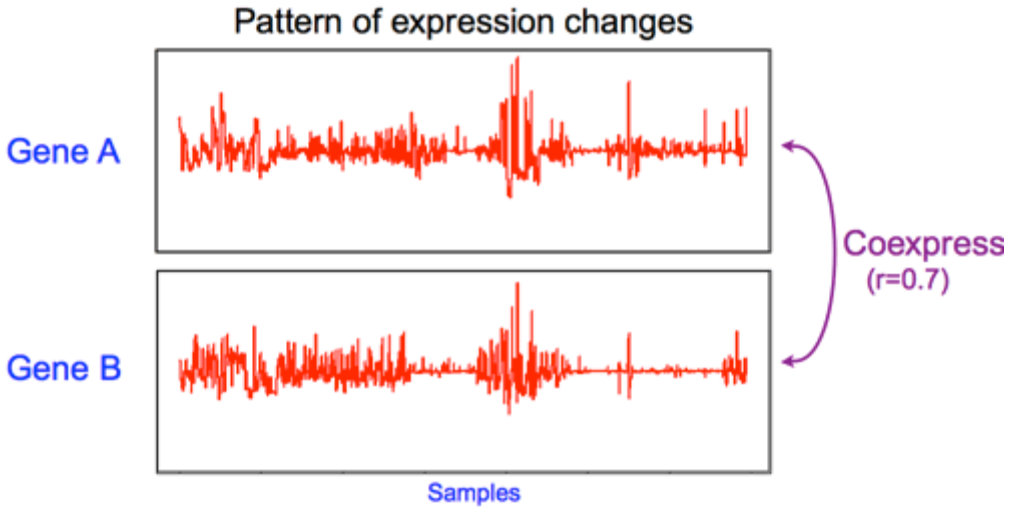


PCA降维数据的分组

(Xu W, et al. Sci Rep, 2018)

2.6 共表达网络分析

- 基因共表达网络分析（Gene Co-expression Network Analysis，GCNA）是一种用于解析基因表达关联模式的系统生物学分析方法。通过构建基因共表达网络，可以识别具有相似表达模式的基因模块，探索基因网络与表型的关联，挖掘网络中的核心基因。
- 具有相似表达模式（共表达）的基因在功能上相关，属于同一复合体，参与同一途径或调节机制，可能相互影响或可能受相同潜在机制的影响。



基因在样本中表现出协调的表达模式

https://coxpresdb.jp/static/overview.shtml

2.6 共表达网络分析

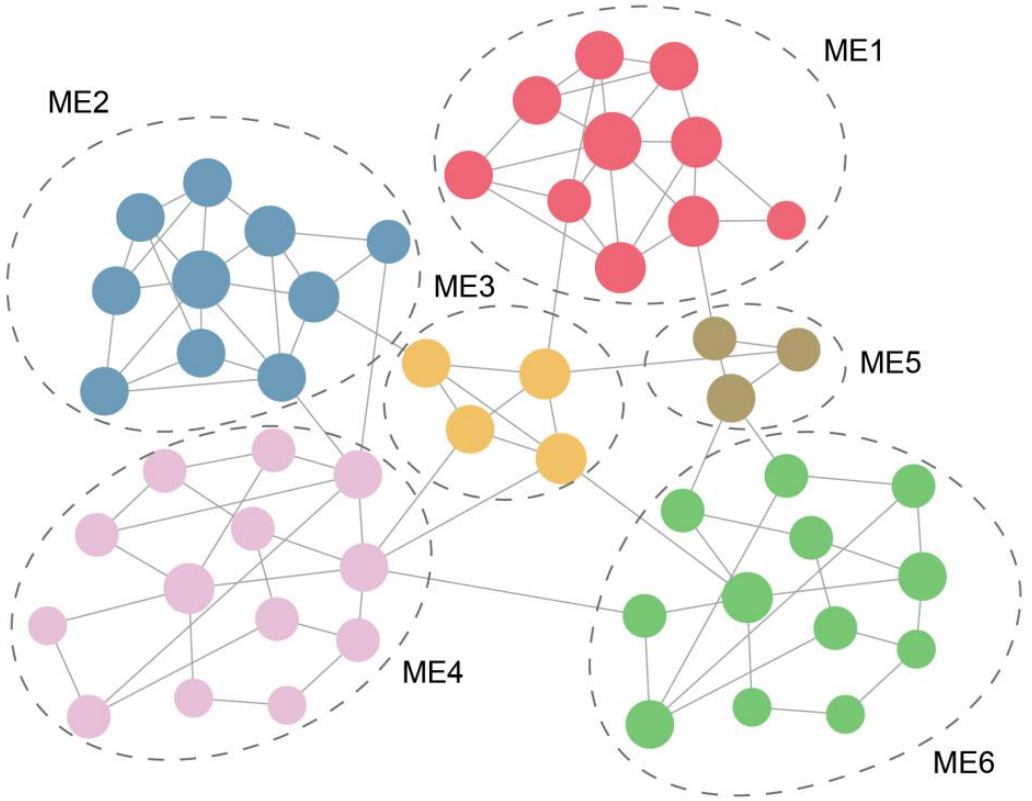
✓ 加权基因共表达网络分析（Weighted Gene Co-expression Network Analysis，WGCNA）是最常见的共表达网络分析方法。WGCNA主要包含四个主要步骤：网络构建、基因模块分类、基因模块分析、核心基因鉴定。

网络构建 -> 加权处理，构建无尺度网络

基因模块分类 ->层次聚类建树，划分具有相似表达模式的基因模块

基因模块分析 -> 关联分析、富集分析解析网络模块

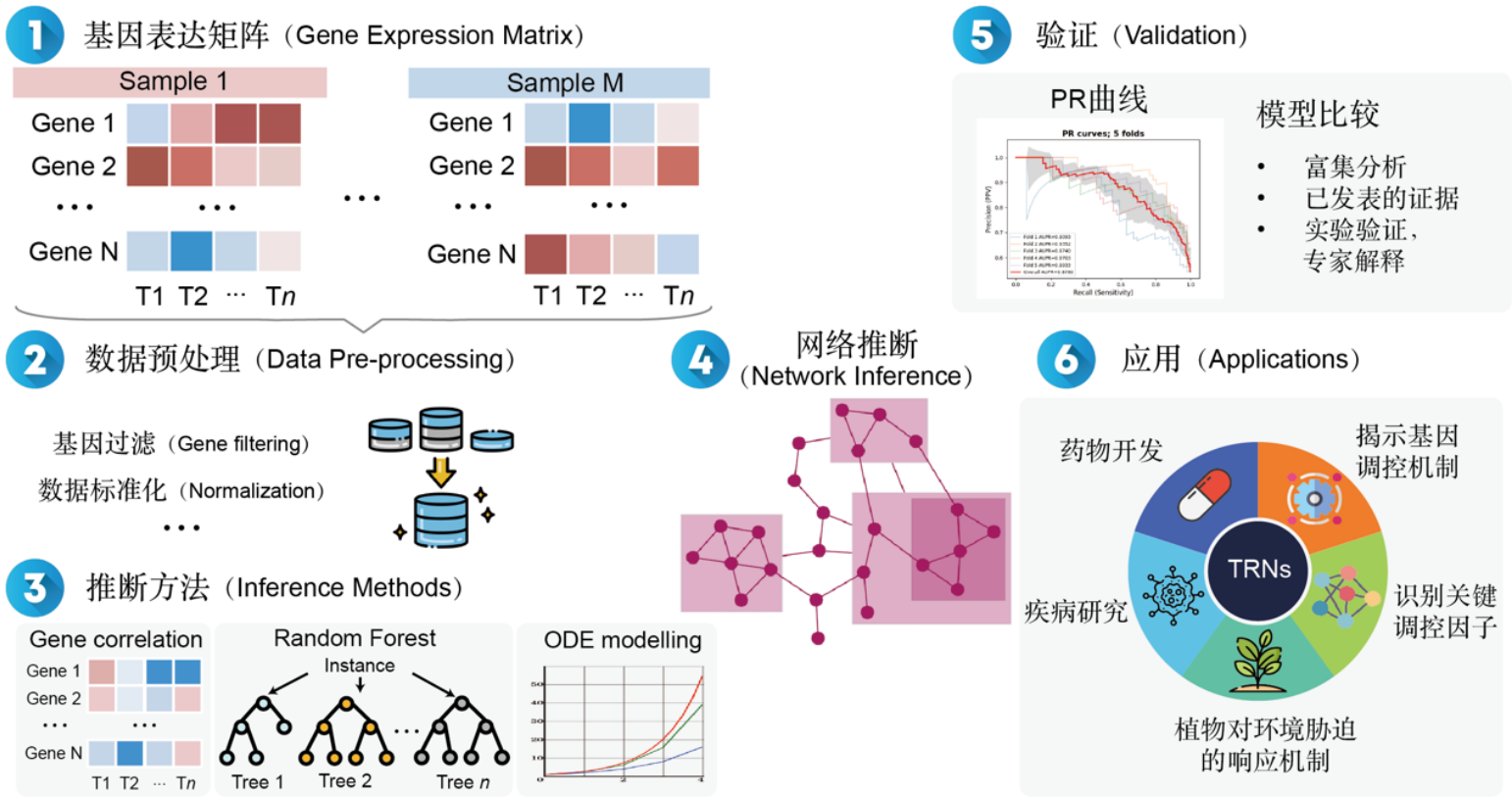
核心基因鉴定 -> 依据网络关系筛选目标模块中的hub基因



共表达网络分析模块示例图

2.7 转录调控网络分析

转录调控网络（Transcriptional Regulatory Network，TRN）是一类以转录因子及其靶基因为节点、基因间调控关系为边的生物学网络。全基因组层面的转录调控网络构建有助于获取新的基因交互作用，帮助研究人员理解特定实验条件下的基因调控规律。

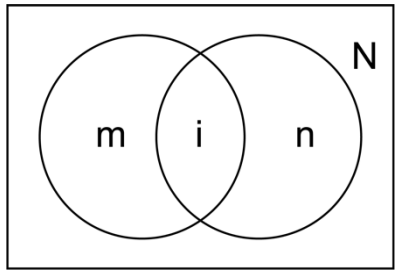


基于转录组学数据的基因调控网络分析

2.8 基因的功能富集分析

- 基因的功能信息主要包括结构域、功能分类以及所在的生物通路等信息。常用的基因功能注释数据库有：GO（Gene Ontology；https://geneontology.org/）、KEGG（Kyoto Encyclopedia of Gene and Genomes; https://www.genome.jp/kegg/pathway.html）、MSigDB（The Molecular Signatures Database; https://www.gsea-msigdb.org/gsea/msigdb）等。
- 统计学方面看，富集的目的是发现显著富集的基因功能集，假定N是所有表达的基因数目，n是读者感兴趣的基因（例如，差异表达基因）数目，m是某一GO term中包含的基因数目，k是该GO term中差异表达基因的数目，利用超几何分布（hypergeometric distribution），可以计算出这组差异表达基因是否显著富集于该GO term。富集概率计算如下：

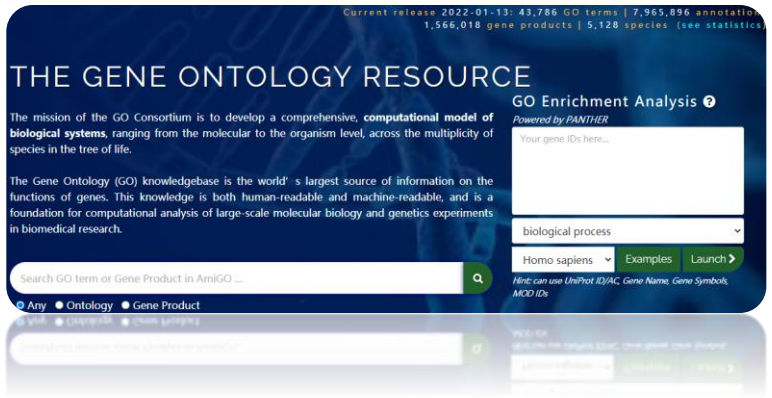
$$p = 1 - \sum_{i=0}^{k-1} \frac{C_{N-m}^{n-i} \times C_m^i}{C_N^n}$$

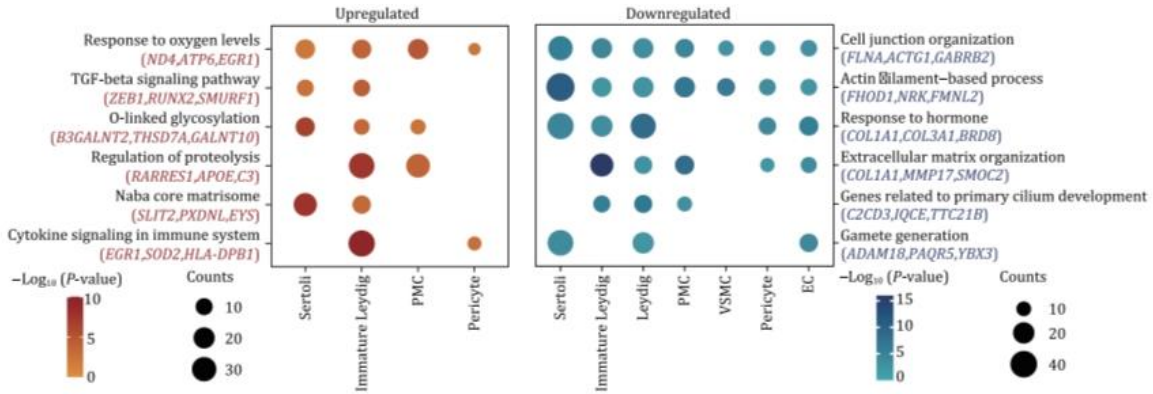


2.8 基因的功能富集分析

常用的富集方法还包括GSEA和GSVA等。

- GSEA是用一个预先定义的基因集中的基因来评估在与表型相关度排序的基因表中的分布趋势，从而判断其对表型的贡献。
- GSVA是一种非参数的无监督分析方法，通过将基因在不同样本间的表达量矩阵转化成基因集在样本间的表达量矩阵，从而来评估不同的功能通路在不同样本间是否富集。





GO富集结果示意图

(Huang DY, et al. Protein & Cell, 2023)

第七章　转录组学

——第三节　非编码RNA

3 非编码RNA

3.1 长非编码RNA

3.2 环状RNA

3.3 小非编码RNA

3.4 非编码RNA的结构

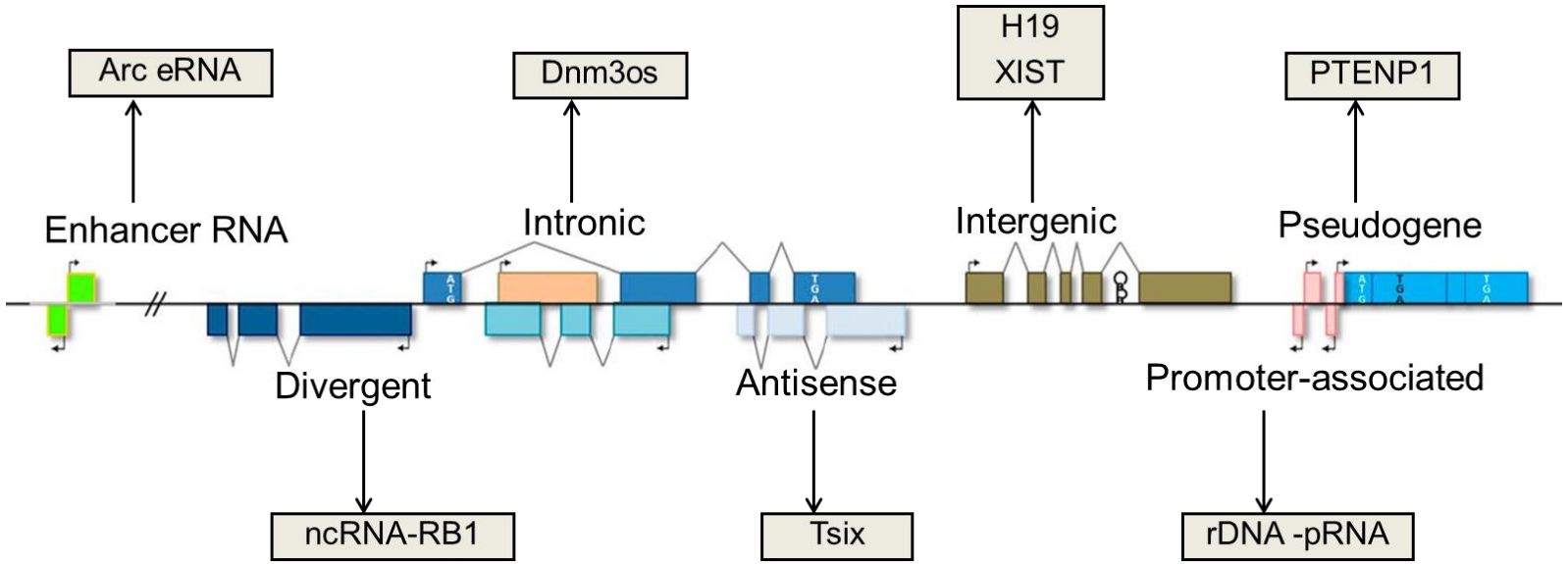
3.5 新非编码RNA的预测、发现和鉴定

3.6 非编码RNA研究展望

3.1 长非编码RNA

长链非编码 RNA(lncRNA)是一类转录本长度超过 200nt 的 RNA 分子，不编码蛋白，以 RNA 的形式在多种层面上（表观遗传调控、转录调控以及转录后调控等）调控基因的表达水平。

在胚胎发育、细胞分化、衰老等生物学过程以及多种复杂疾病中均发挥重要作用。



3.1 长非编码RNA

LncRNA的功能

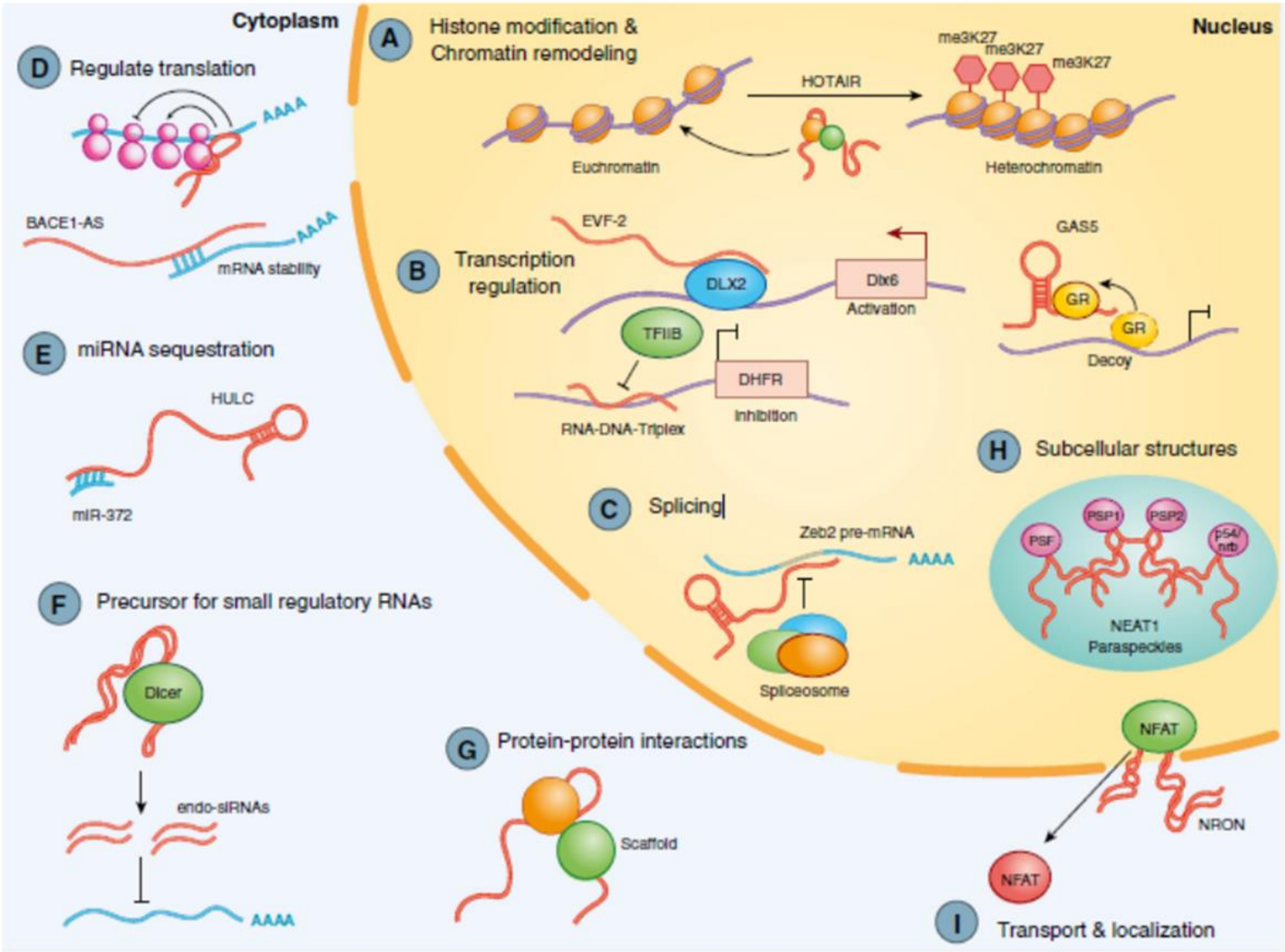
基因转录的激活与抑制

影响组蛋白修饰、染色质相互作用、与可变剪接

转录后/翻译调控

作为海绵吸附miRNA

作为内源小RNA前体，调控蛋白的功能活性、相互作用和细胞定位，参与形成核结构等



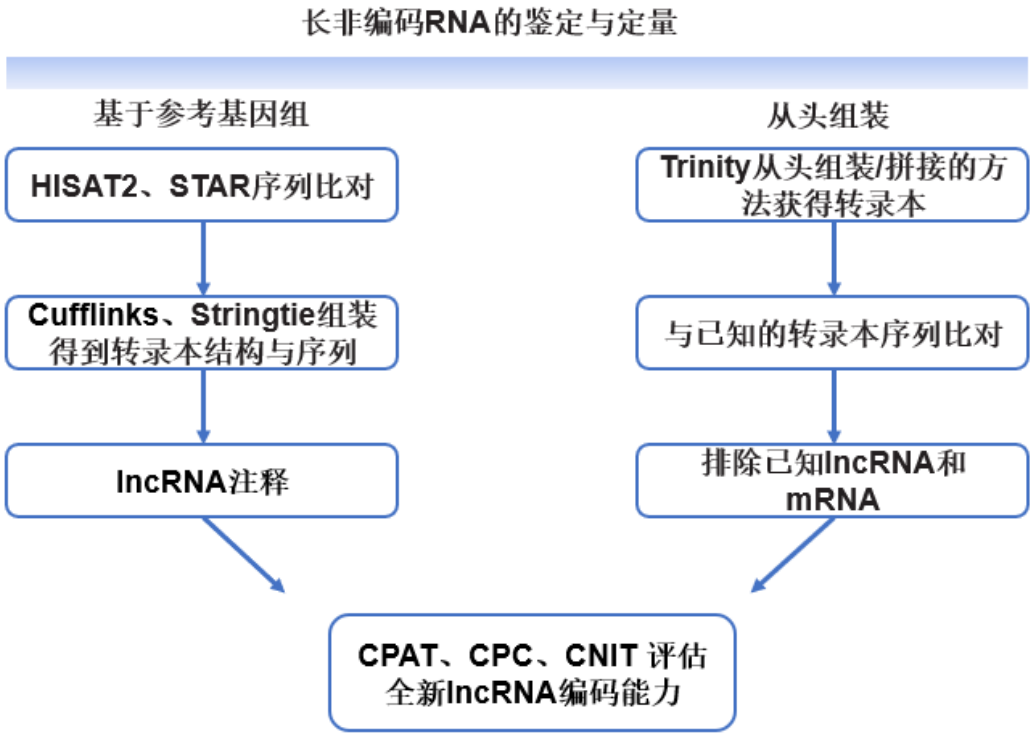
3.1 长非编码RNA

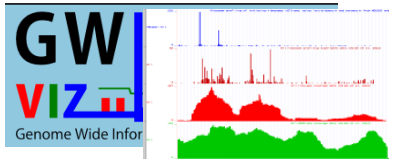
LncRNA 的鉴定与定量

测序： RNA-seq ， Ribo-minus，链特异性测序

鉴定：从 RNA-seq 中组装转录本，注释已知转录本；对于新转录本，判断其编码潜能。

定量：与mRNA转录组类似，以reads读数为基础，再进行归一化（主要包括RPKM、TPM、scale factor等）。

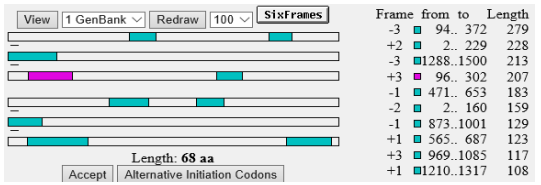


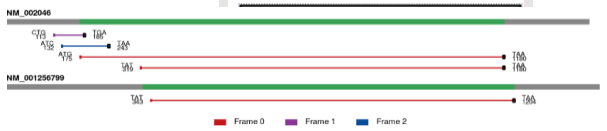


PhyloCSF browser tracks

"http://www.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub/hub.txt" into the "**My Hubs**" tab under "**track hubs**",

sORFs.org Biomart database

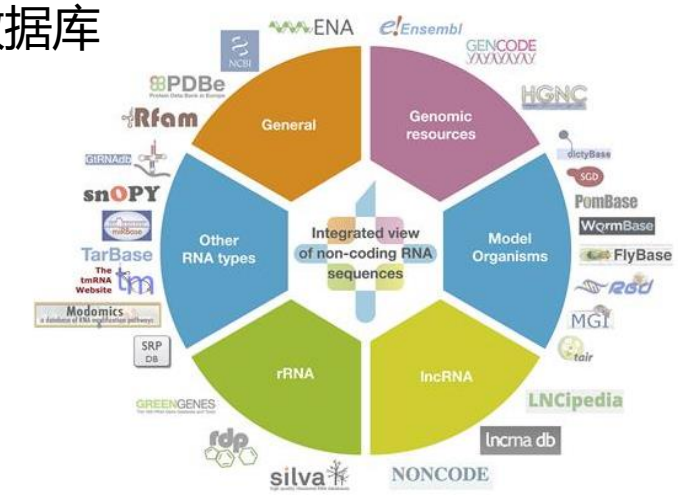




预测新转录本的编码潜能的工具

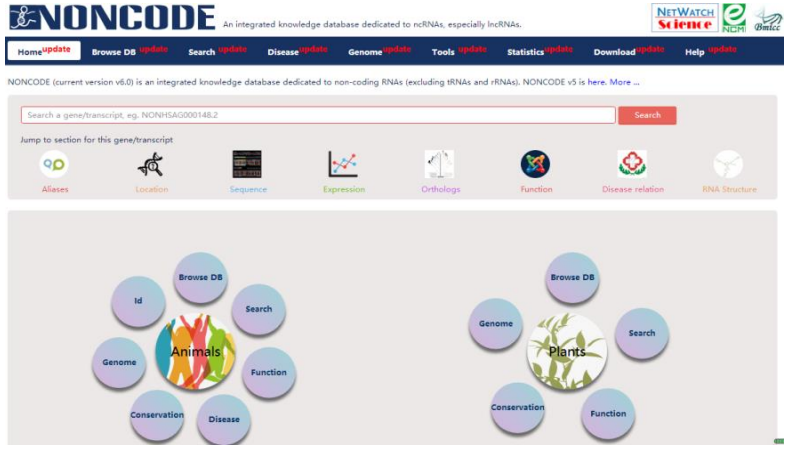
LncRNA 相关数据库

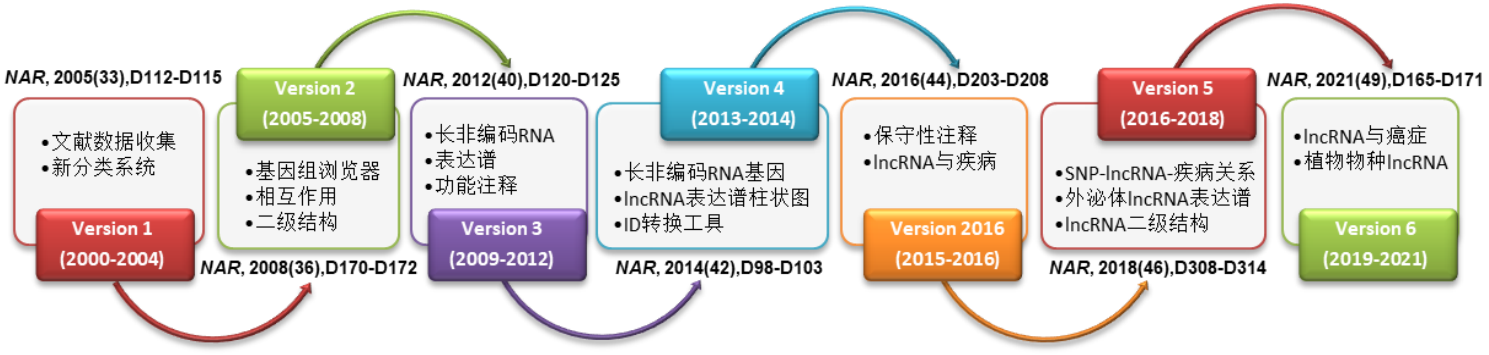
- RNAcentral :一个包含广泛物种 ncRNA 序列的综合数据库



- lncBook:整合人类lncRNA及多组学注释的数据库，并鉴定了疾病和不同环境中的特征lncRNA。

- NONCODE数据库：lncRNA的国际权威数据库



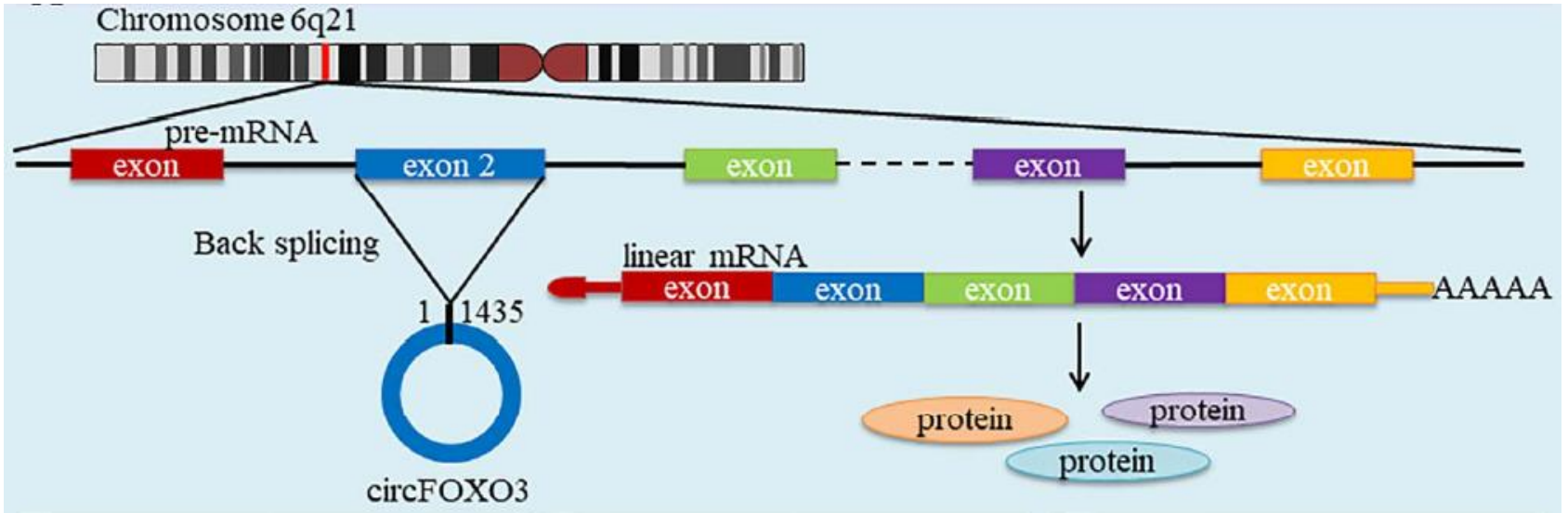


3.2 环状RNA

环形 RNA 结构特征

- 环状RNA（circular RNA，circRNA）是一种新型的内源性非编码RNA分子，与传统线性RNA不同，具有独特的闭环结构，没有 3‘poly A 尾巴 和 5’cap 结构。
- circRNA对核酸酶不敏感，比线性RNA更为稳定，半衰期长
- 很多circRNA由蛋白编码基因产生，比如circFoxO3，但转录形成的circFoxO3无编码能力。

circRNA结构



circFoxO3不编码蛋白 (Rao D, et al. Front Cell Dev Biol, 2021)

环形 RNA 的产生途径

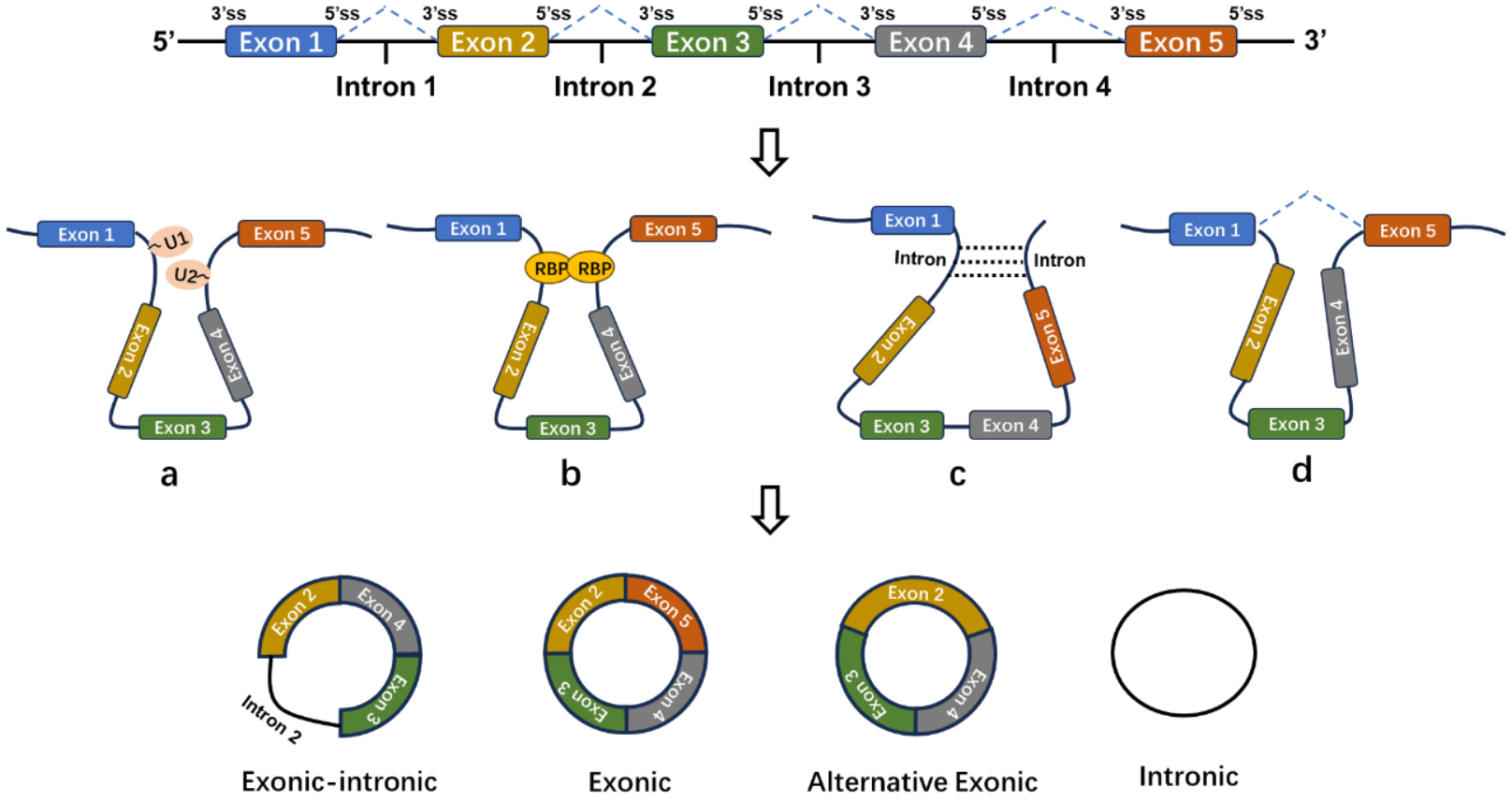
circRNA的四种环化途径

a. 剪接体依赖的环化途径

b. 蛋白因子结合环化途径

c. 内含子配对驱动的环化途径

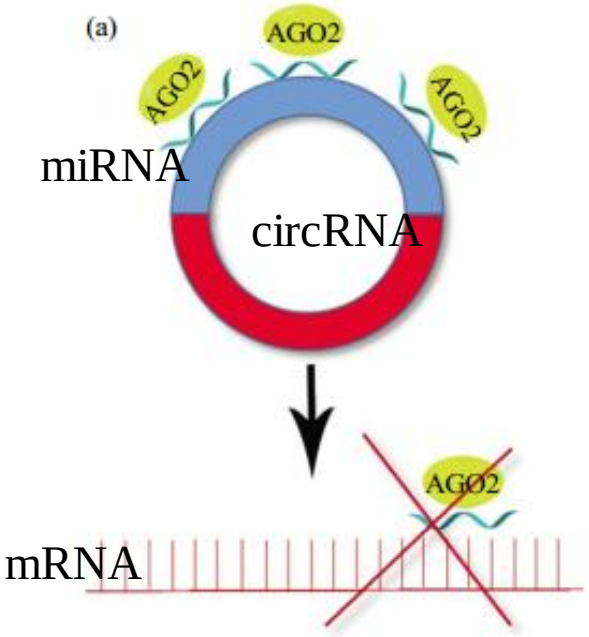
d. 套索驱动的环化途径



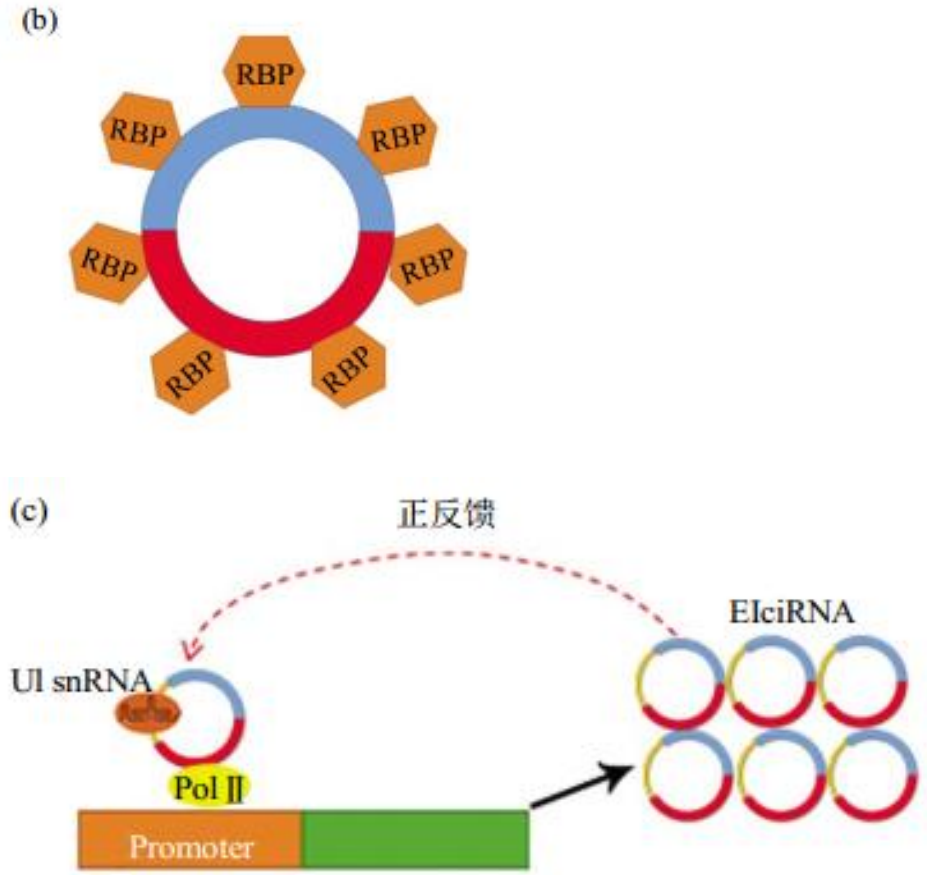
3.2 环状RNA

环形 RNA 的功能

- 充当蛋白支架，促进与之结合的蛋白相互作用。
- 充当蛋白质海绵，竞争性结合蛋白，调节蛋白作用。



作为miRNA海绵

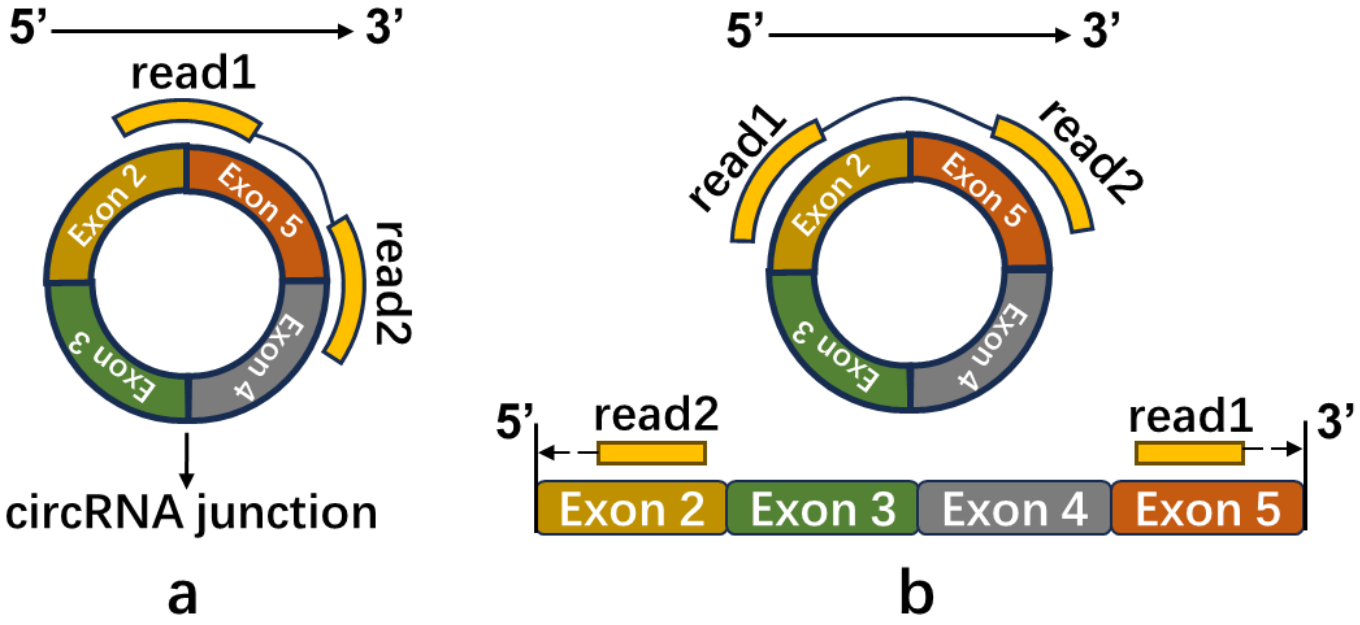


参与基因调控

(http://www.pibb.ac.cn/html/2018/6/20170443.htm)

环形 RNA 的鉴定

- 原理：在RNA-seq数据中，检测匹配到反向剪接位点的reads，识别circRNA的剪接点。
- 然后进行circRNA全长序列的重建，并识别可变剪接，确定circRNA内部的其他结构序列。



- 通过识别junction reads来预测circRNA。
- a. reads覆盖了反向剪接位点
- b. 成对的reads包含了反向剪接位点，且比对的位置是相反的。

1. **候选分子方法；**依赖于基因注释。
2. **子序列比对方法；**通过剪接读段（junction reads）来识别circRNA。
3. **机器学习类方法；**从转录本中提取保守性分数、序列组成等特征，然后使用不同的机器学习分类器和统计方法整合这些特征构建分类模型，来对circRNA进行预测。

3.2 环状RNA

环形 RNA 的检测工具

- CIRI
- FUCHS

通过比对到反向剪接点的reads 检测circRNA

- CIRI-AS
- circSplice

识别circRNA内部的可变剪切

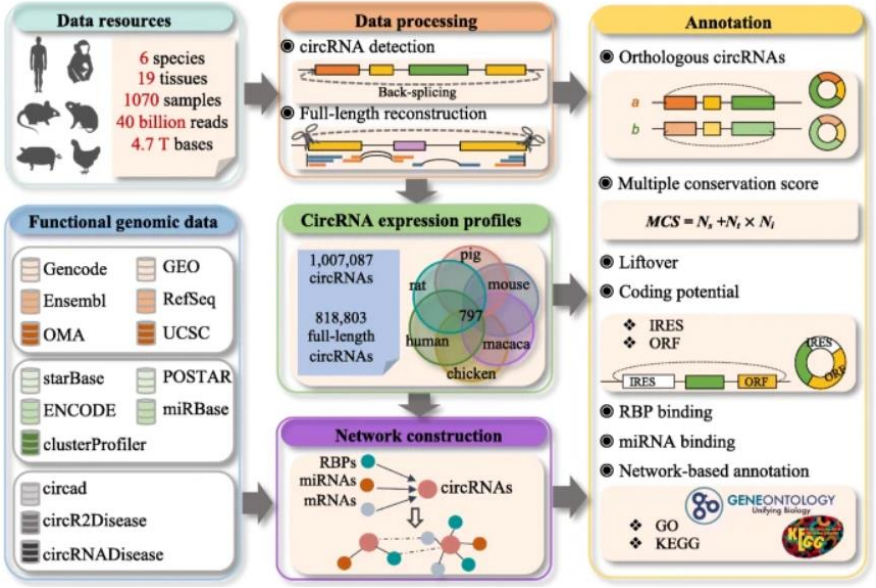
- DeepCirCode
- PredcircRNA

机器学习分类器和统计方法整合转录本特征预测circRNA

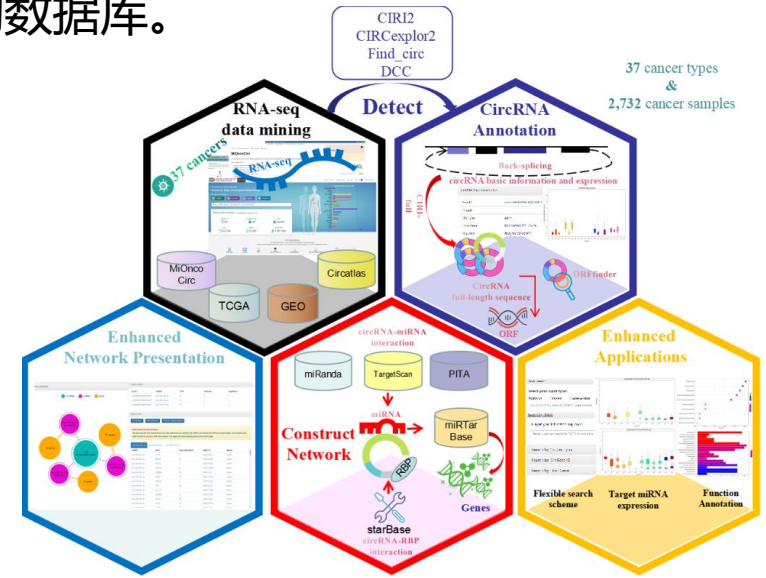
- CIRI-full: 全长 circRNA 的重建， isoform 水平的定量。更适合长片段，>250或300bp
- CIRI-vis: circRNA可视化

环形 RNA 相关数据库

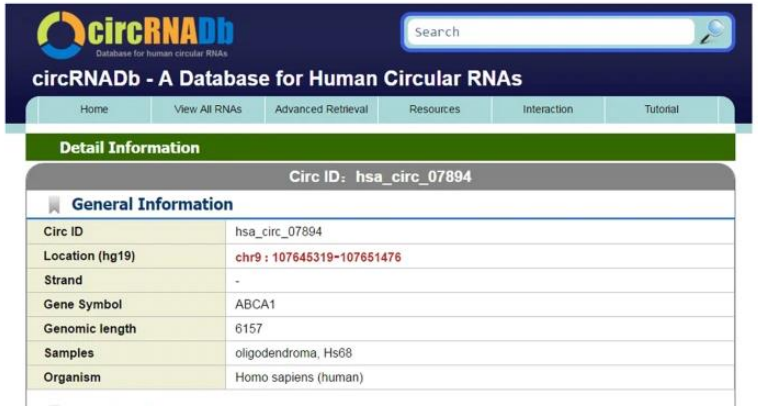
circAtlas:多物种的circRNA注释



CircNet:专注于癌症中环状RNA（circRNA）调控网络的数据库。



circRNADb:专注于编码蛋白的circRNA数据库，包含32，914个具有详细注释的外显子circRNA，并预测了circRNA中可能的ORF。



3.3 小非编码RNA

➢小非编码RNA长度在15-200， 通常在转录和/或转录后水平参与调控基因表达，其多样性随着物种生物复杂性的增加而进化增加。

- 三种被广泛研究的典型小RNA

  miRNA

  piRNA

  siRNA

- 两种非典型的小RNA

  tsRNA

  rsRNA

3.3 小非编码RNA

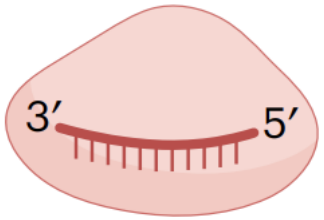
miRNA

定义：存在于大多数真核生物中，长度为20-25 nt，具有5’端单磷酸基团和3’端羟基的特征。

功能：miRNA通过碱基互补配对的方式识别靶基因，抑制靶mRNA的翻译或降解。

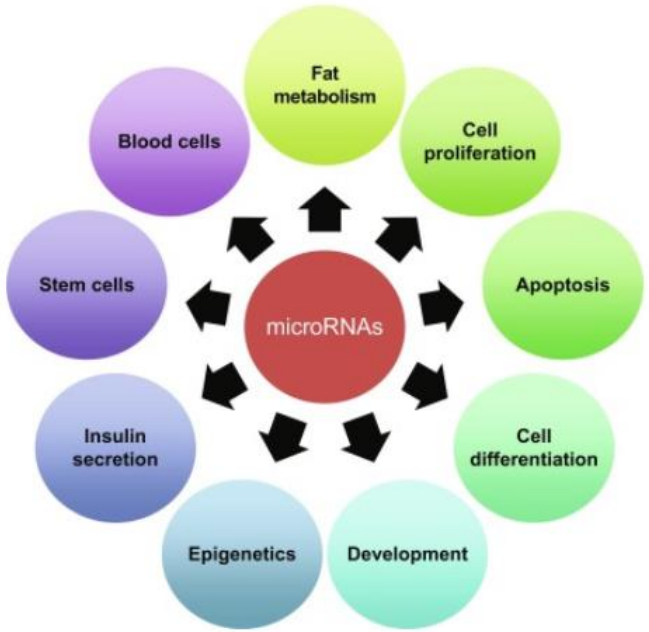
生物学特征：

高度保守性：不同生物发育过程中，miRNAs具有相同的调控机制。
时序表达特异性性：在不同组织、不同发育阶段，miRNA表达水平是动态变化的。
组织表达特异性：一些miRNA表达具有细胞和组织特异性。



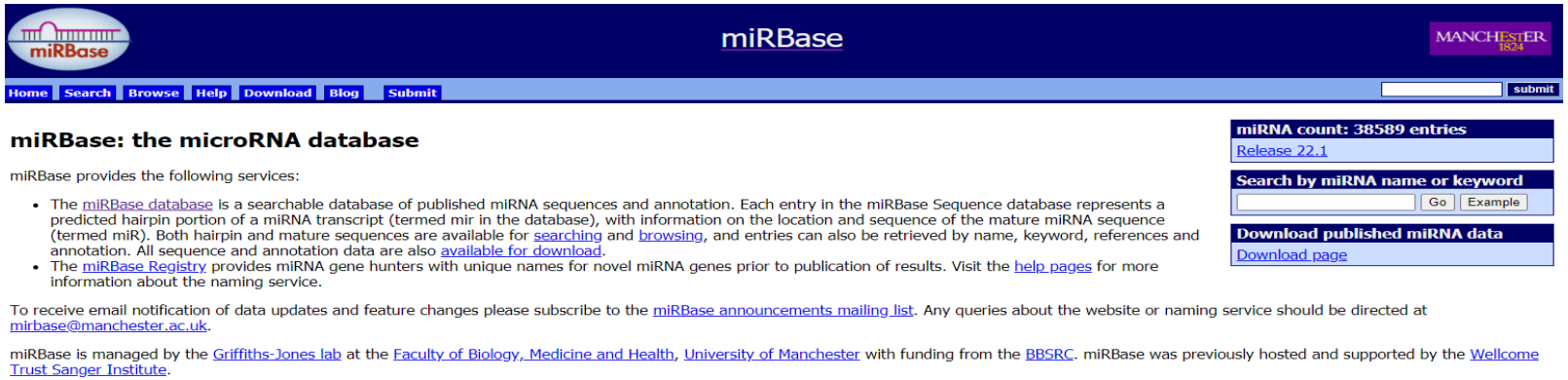
Ago2–miRNA

(Shang et.al, 2023)

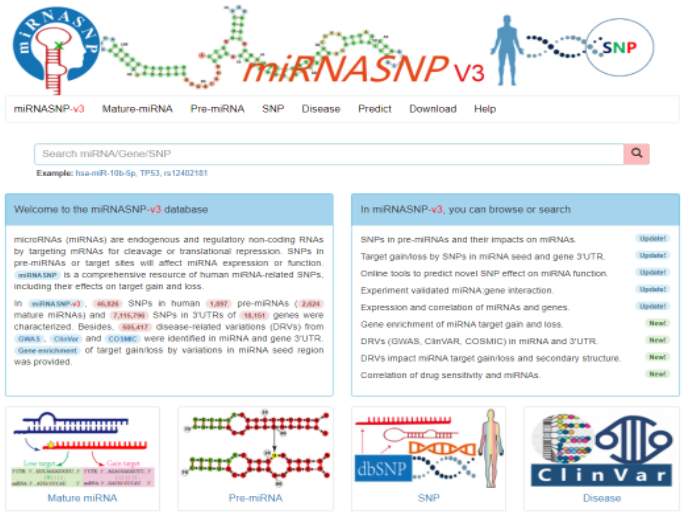


3.3 小非编码RNA

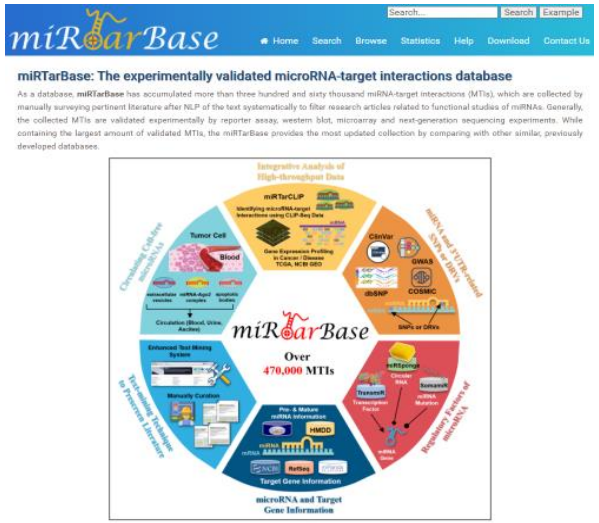
- miRBase:已发表的miRNA 序列数据、注释、预测基因靶标等。



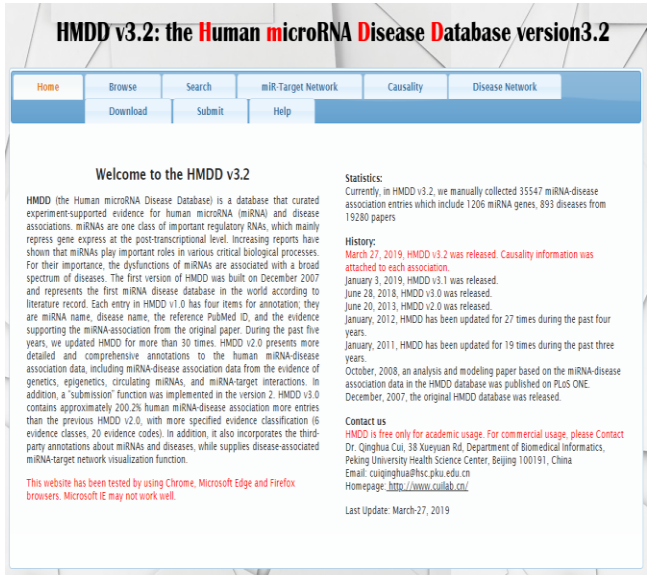
- miRNASNP:提供与miRNA相关的SNPs



- miRTarBase:收集被实验验证的靶标



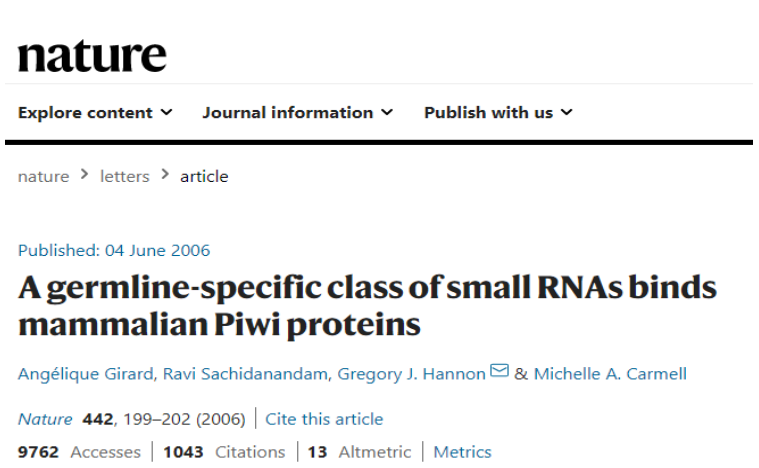
- HMDD:收集人类miRNA和疾病关联的实验支持证据



3.3 小非编码RNA

piRNA (PIWI-interacting RNAs) 是2006年报道发现的一类新的非编码小RNA，piRNA的发现被Science评为2006年十大科技进展之一。

piRNA主要在生殖细胞中表达，在生殖细胞及基因调控中发挥重要作用。


nature

Explore content · Journal information · Publish with us

nature › letters › article

Published: 04 June 2006

A germline-specific class of small RNAs binds mammalian Piwi proteins

Angélique Girard, Ravi Sachidanandam, Gregory J. Hannon & Michelle A. Carmell

Nature **442**, 199–202 (2006) | Cite this article

9762 Accesses | **1043** Citations | **13** Altmetric | Metrics


nature

Explore content · Journal information · Publish with us

nature › letters › article

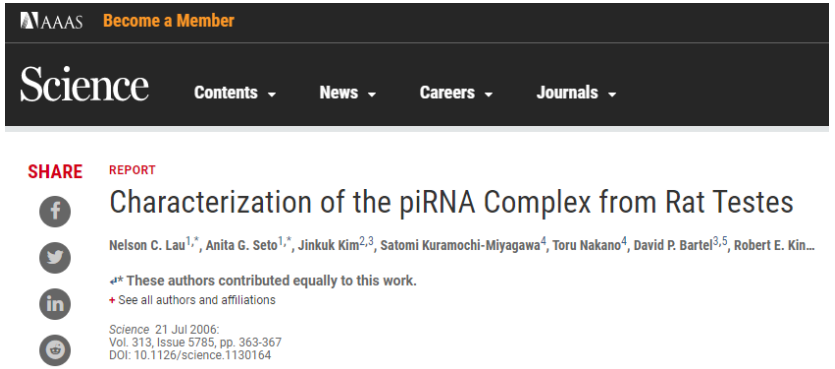
Published: 04 June 2006

A novel class of small RNAs bind to MILI protein in mouse testes

Alexei Aravin, Dimos Gaidatzis, Sébastien Pfeffer, Mariana Lagos-Quintana, Pablo Landgraf, Nicola Iovino, Patricia Morris, Michael J. Brownstein, Satomi Kuramochi-Miyagawa, Toru Nakano, Minchen Chien, James J. Russo, Jingyue Ju, Robert Sheridan, Chris Sander, Mihaela Zavolan & Thomas Tuschl

Nature **442**, 203–207 (2006) | Cite this article

6717 Accesses | **937** Citations | **6** Altmetric | Metrics


AAAS Become a Member

Science · Contents · News · Careers · Journals

SHARE

REPORT

Characterization of the piRNA Complex from Rat Testes

Nelson C. Lau[1,*], Anita G. Seto[1,*], Jinkuk Kim[2,3], Satomi Kuramochi-Miyagawa[4], Toru Nakano[4], David P. Bartel[3,5], Robert E. Kin…

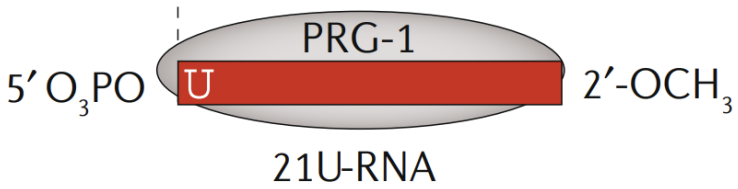
* These authors contributed equally to this work.

+ See all authors and affiliations

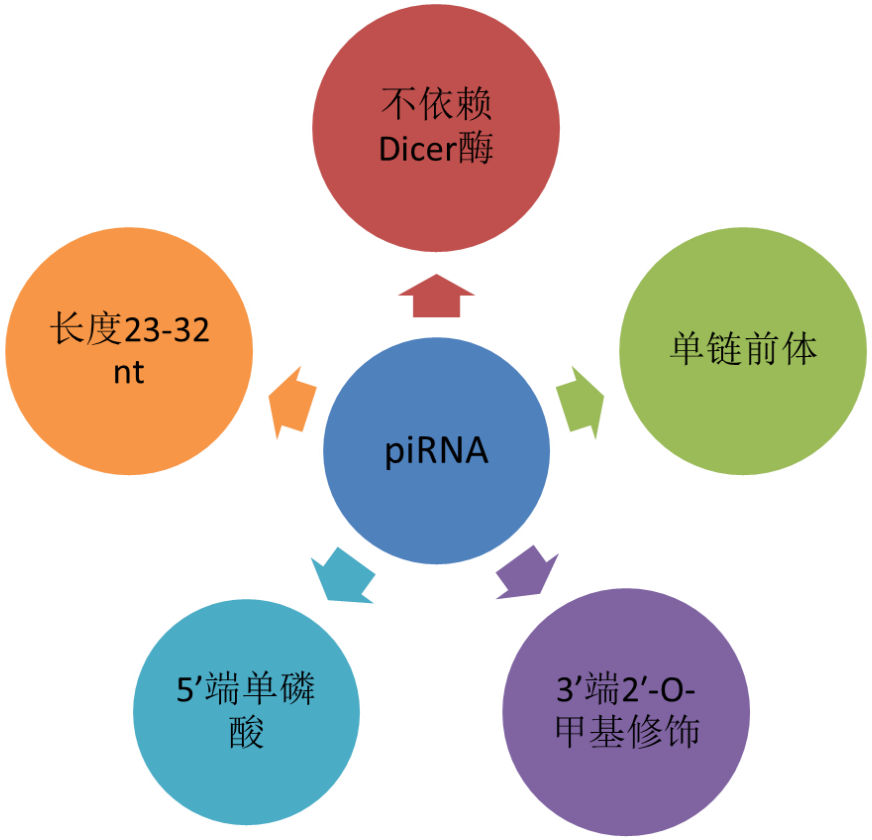
Science 21 Jul 2006:
Vol. 313, Issue 5785, pp. 363-367
DOI: 10.1126/science.1130164

piRNA的特征

- piRNA：长度约23-32nt，具有5’端1U偏好性和3’端2’-O-甲基化修饰的特征。



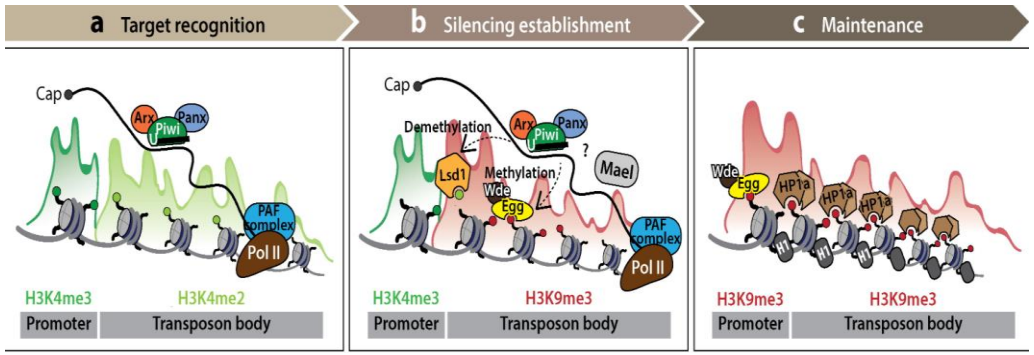
(Deniz M. Ozata et.al, 2019)



3.3 小非编码RNA

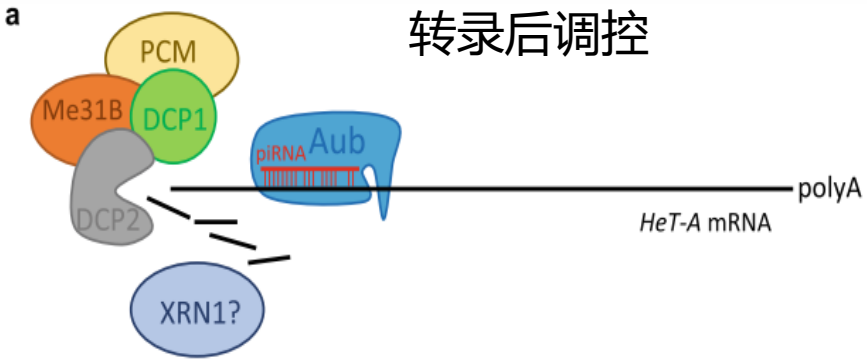
piRNA的功能

转录水平调控



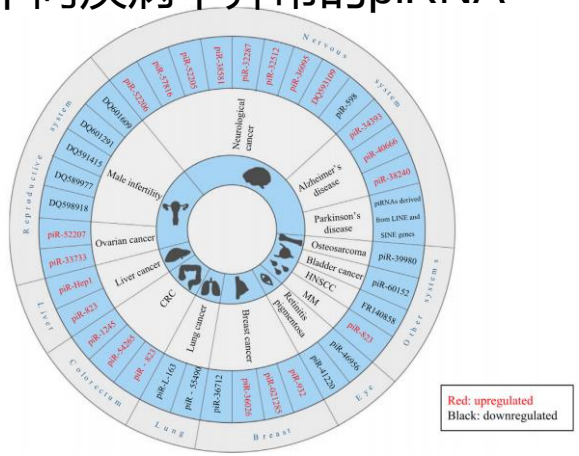
Piwi-piRNA通过序列互补识别新生转录本，建立沉默，并维持其抑制状态。

转录后调控



与癌症标志物相关的piRNA

不同疾病中异常的piRNA

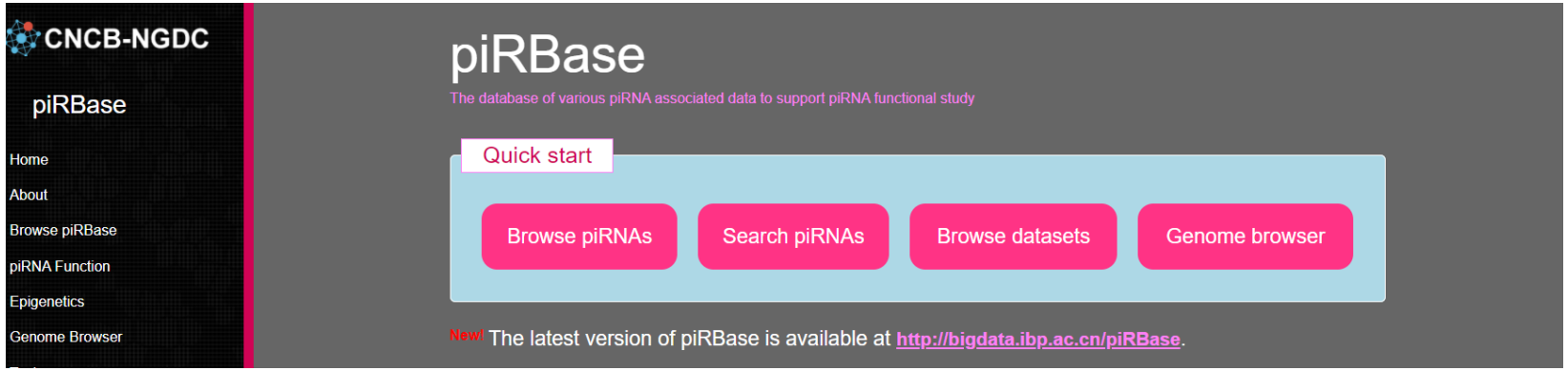


(Czech B, et al. Annu Rev Genet, 2018)

3.3 小非编码RNA

piRNA相关数据库

piRBase：辅助piRNA功能研究的数据库，是国际RNA联盟收录的唯一piRNA专业数据库。



piRNAClusterDB：piRNA cluster收集整理的数据库

3.3 小非编码RNA

小RNA的鉴定和定量

miRNA鉴定：

- 根据序列特征，碱基配对原则，最小自由能等特征模型，进行新miRNA及二级结构预测。
- 工具：miRDeep2，miRge，miRanalyzer等

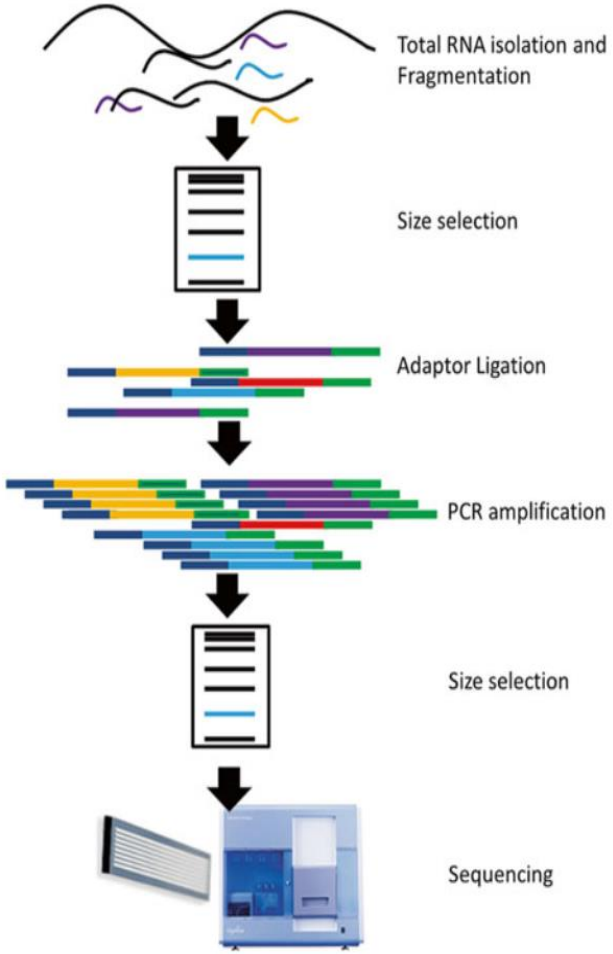
piRNA和piRNA cluster鉴定：

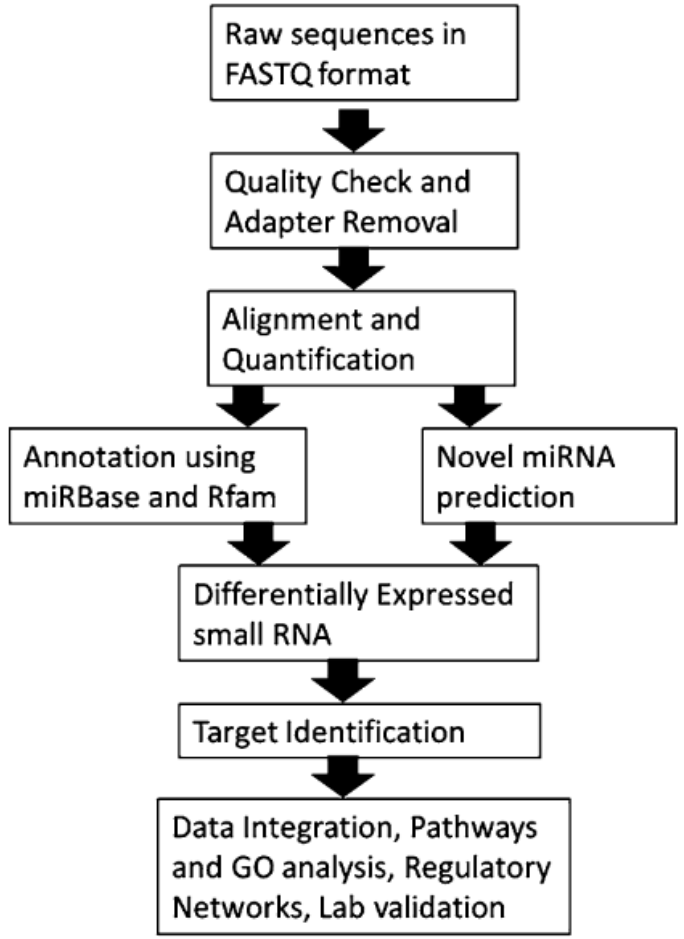
- piRNA缺乏保守的序列和结构特征，导致识别piRNA具有很大的挑战性。许多鉴定方法依赖于提取复杂的特征，或者主要针对转座子相关的piRNA进行预测。
- piRNApred：基于RNA序列、结构、热动力学和理化性质预测piRNA
- PiRPred：基于多核心和支持向量机进行机器学习预测piRNA
- Piano：基于转座子和支持向量机预测piRNA
- piRNN：应用卷积神经网络分类器进行piRNA的预测
- piClust和proTRAC：能够预测piRNA cluster

3.3 小非编码RNA

小RNA定量及差异表达分析

- RNA 提取
- 选择长度 17-25 nt 的 RNA
- 加接头（adapter）
- 扩增
- 测序
- 数据分析

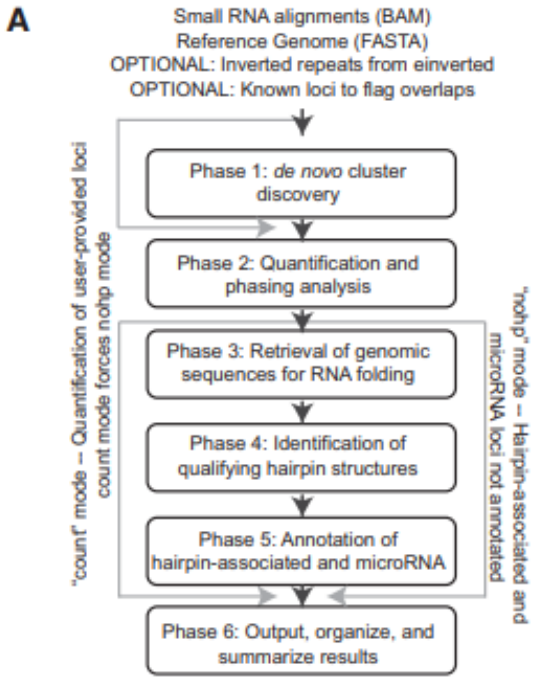




(Mehta, Methods Mol Biol, 2014)

3.3 小非编码RNA

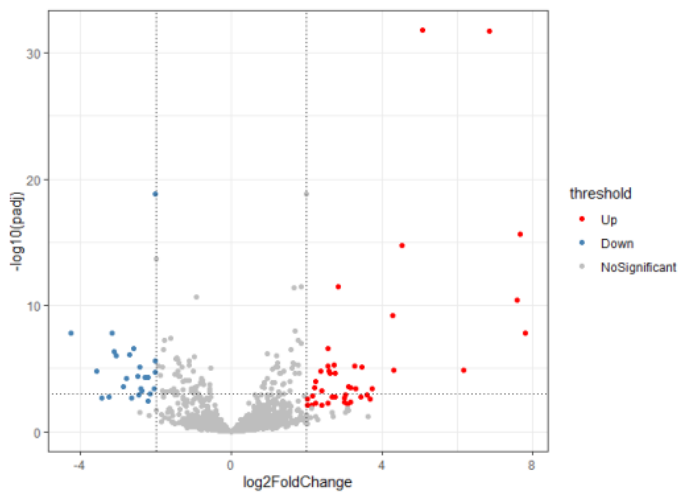
- 定量：ShortStack



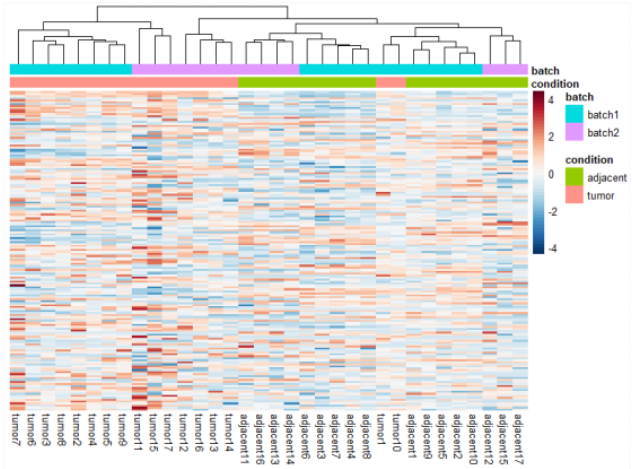
- 差异表达分析：

类型	软件	标准化方法	pvalue计算模型	FDR计算方法	差异基因筛选标准
有生物学重复	DESeq2(Anders et al, 2014)	DESeq	负二项分布	BH	|log2(FoldChange)| > 0 & padj < 0.05
无生物学重复	edgeR(Robinson et al, 2010)	TMM	负二项分布	BH	|log2(FoldChange)| > 1 & padj < 0.05

- 差异表达火山图



- 表达模式聚类图



3.3 小非编码RNA

miRNA靶基因分析

➢靶基因预测

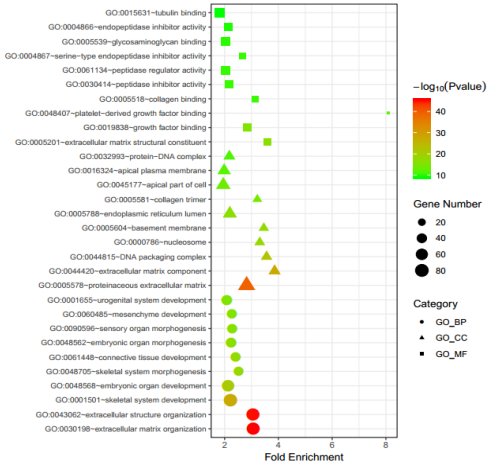
-miRanda：结合两个因素预测靶基因，一是miRNA和mRNA间的序列互补匹配程度，二是形成的复合结构的自由能。

-TargetScan：通过搜索和每条miRNA种子区域匹配的保守的8mer和7mer位点来预测靶基因。

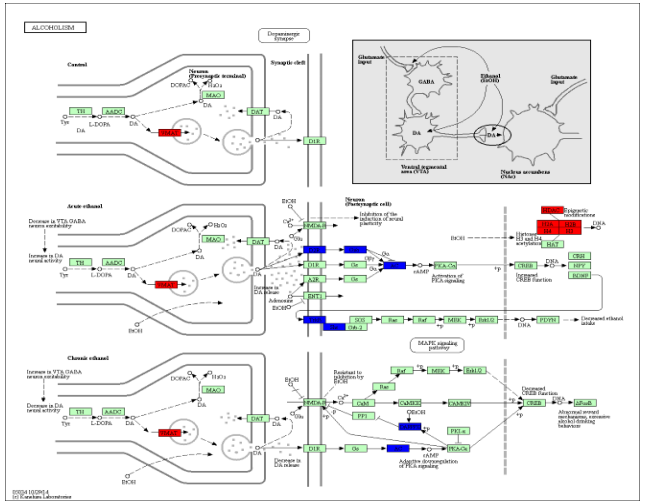
-RNAhybrid：基于分析miRNA和靶基因间形成双链的二级结构，从而预测miRNA靶基因。

-DMISO：基于深度学习方法捕获miRNA/isomiR(miRNA亚型)与mRNA相互作用的复杂特征。

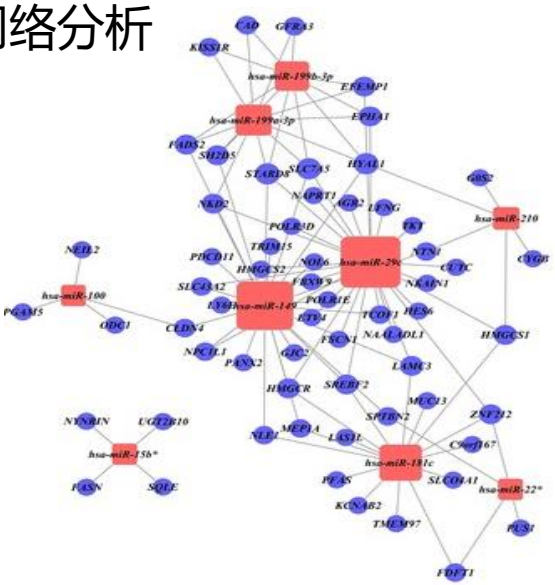
➢ GO富集分析



➢ KEGG富集分析

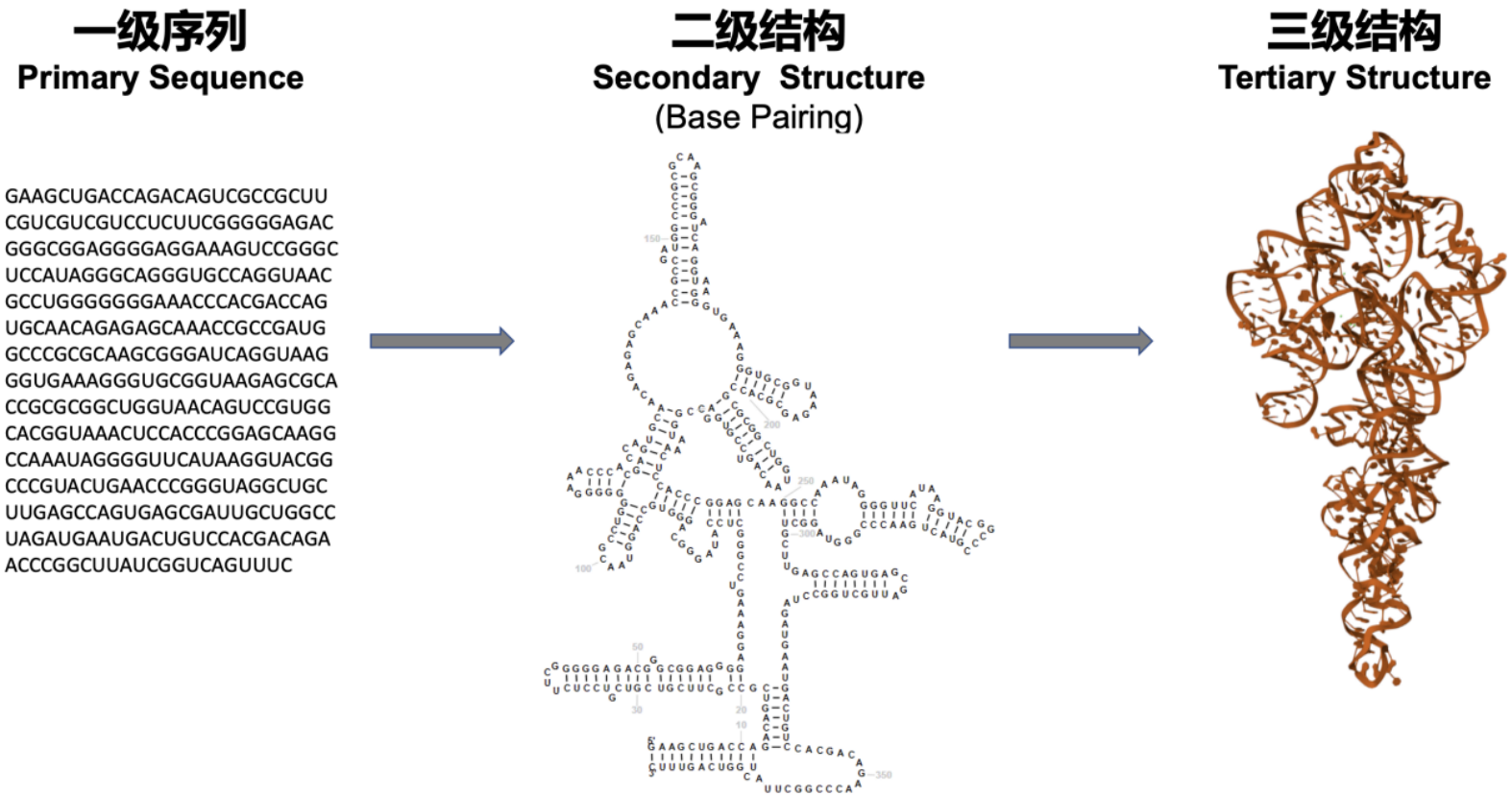


➢ 调控网络分析



3.4 非编码RNA的结构预测

- RNA二级结构：是指RNA可以通过碱基互补配对（base pairing）形成不同的单链区域（loop）和双链区域（duplex)
- RNA三级结构：在单链和双链区域组成的二级结构基础上，能进一步形成更高层次的复杂的三维结构。

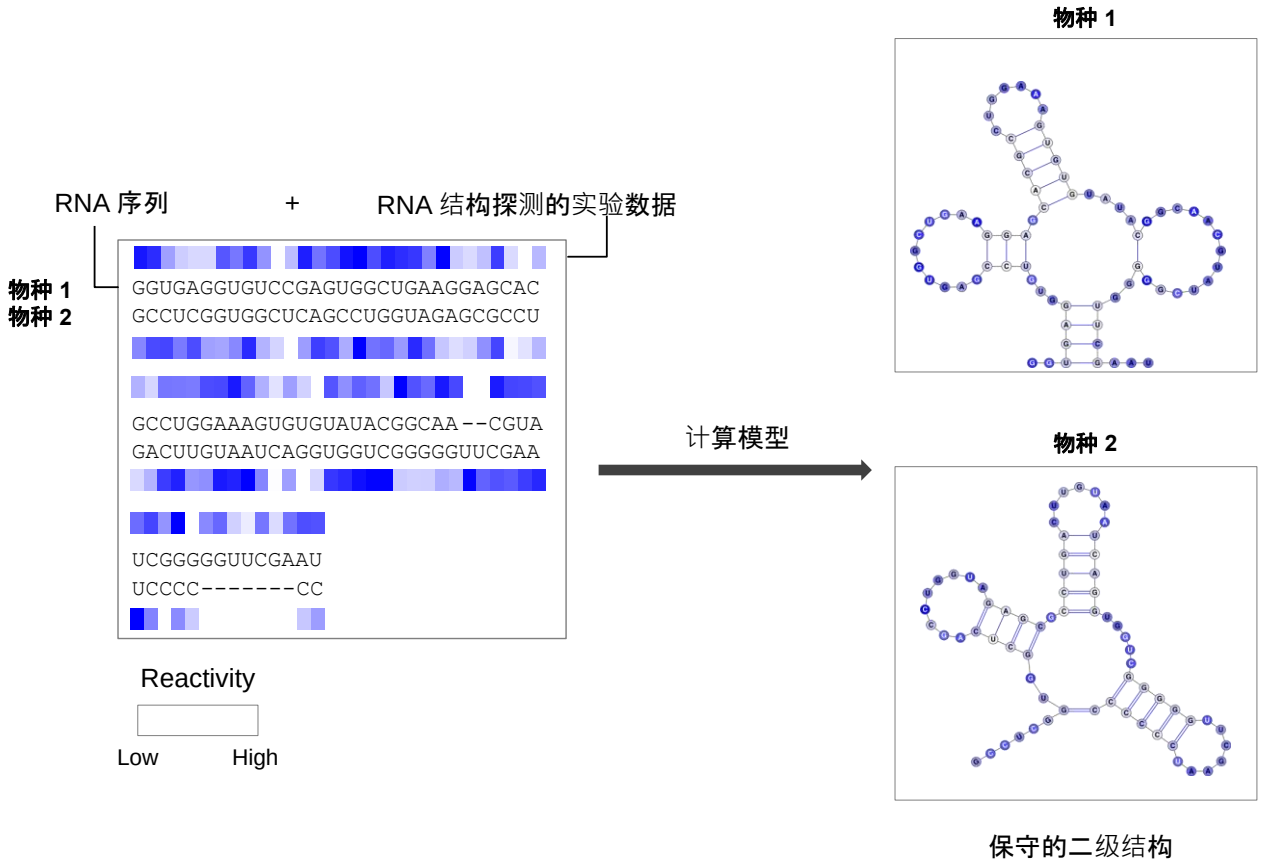


非编码RNA的二级结构预测

1. 基于能量最小化的预测（Free Energy Minimization）在一定的条件下，RNA的结构总是趋向于取其可能的最稳定状态，即自由能最低的状态。一些工具如 RNAfold 和RNAstructure就是基于此原理工作的。

2.比较序列的预测（Comparative Analysis）：这种方法基于的观点是在进化过程中，功能相关的RNA结构比序列更保守。Rfam 数据库就提供了这样的比对模式。

3.机器学习和AI方法（Machine Learning and AI）：近年来，随着深度学习、迁移学习等机器学习和AI技术的发展，也有一些新的RNA结构预测方法出现，这些方法往往需要大量已知的RNA序列和结构作为训练数据。



结合 RNA 结构探测实验数据和计算模型进行 RNA 二级结构预测

3.5 新非编码RNA的预测、发现和鉴定

1. **转录组测序技术:** RNA-seq通过测序全转录组,可以发现新的非编码RNA分子。
2. **生物信息学预测方法:** 基于序列特征的预测；保守性的预测；功能注释的预测等。
3. **实验验证方法:** Northern blot；Real-time PCR；原位杂交；功能实验。
4. **高通量实验技术:** Ribo-seq，通过测序核糖体结合的RNA片段,鉴定RNA转录本的编码潜能； SHAPE-seq，CLIP-seq等。

综合应用上述这些方法,可以有效地鉴定出非编码RNA分子,并探讨其生物学功能。
这些方法的基本原理和相关软件归纳如下:

1. **序列和结构特征分析**：核酸碱基组成特征，非编码RNA通常具有较高的GC含量,以及特殊的核酸组成模式等。
2. **比对和保守性分析:** 功能性非编码RNA在进化过程中往往表现出较高的序列保守性。涉及到的软件包括BLAST，Rfam 等。
3. **转录组分析**：表达特征分析；差异分析。涉及到的软件包括WGCNA，EdgeR，DEseq2等。
4. **综合预测模型**：整合多种特征，如序列、结构、表达、保守性等多个特征的组合预测，并通过机器学习方法利用监督或无监督学习算法构建预测模型。涉及到的软件包括COME，CPAT等。

3.6 非编码RNA研究展望

1. **新的非编码RNA类型的发现；**从更多物种、组织或细胞类型中发现新的非编码RNA。利用微生物宏转录组发现不同环境和条件下的新非编码 RNA。
2. **非编码RNA结构预测方法的研究；**进一步提高预测能力和效率，基于AI 技术、大模型的结构预测。
3. **非编码RNA的调控和功能研究；**非编码RNA与DNA、RNA、蛋白质等相互作用调控生物过程的机制。
4. **非编码RNA在疾病中的应用；**细胞外 RNA（cfRNA，cell free RNA）用于疾病的无创检测的分子标志物，直接通过抽血化验即可检测疾病状态。
5. **非编码RNA的疗法开发；**基于 siRNA 的RNA干扰（RNA interference， RNAi）系统用于基因治疗。circRNA用于RNA疫苗。

第七章　转录组学

——第四节　单细胞转录组学

4.1 单细胞转录组数据预处理和质控

4.2 数据标准化、校正及整合

4.3 降维、聚类和细胞注释

4.4 差异细胞比例和差异表达基因分析

4.5 细胞发育轨迹推断

4.6 基因调控网络

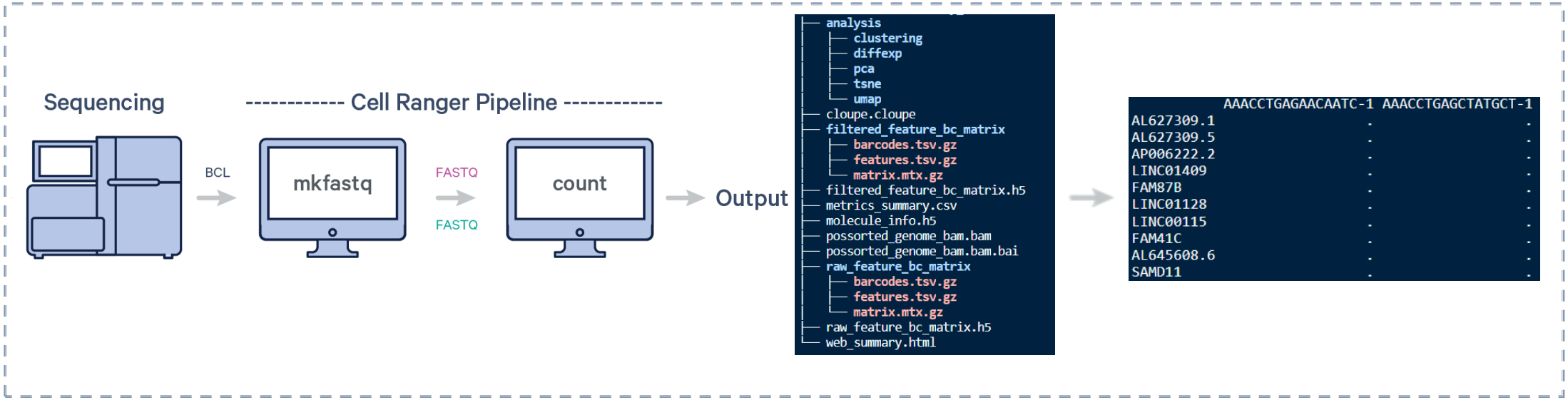
4.7 细胞通讯分析

4.8 单细胞转录组数据分析总结

4.9 单细胞转录组应用

4.1 单细胞转录组数据预处理和质控

原始数据预处理

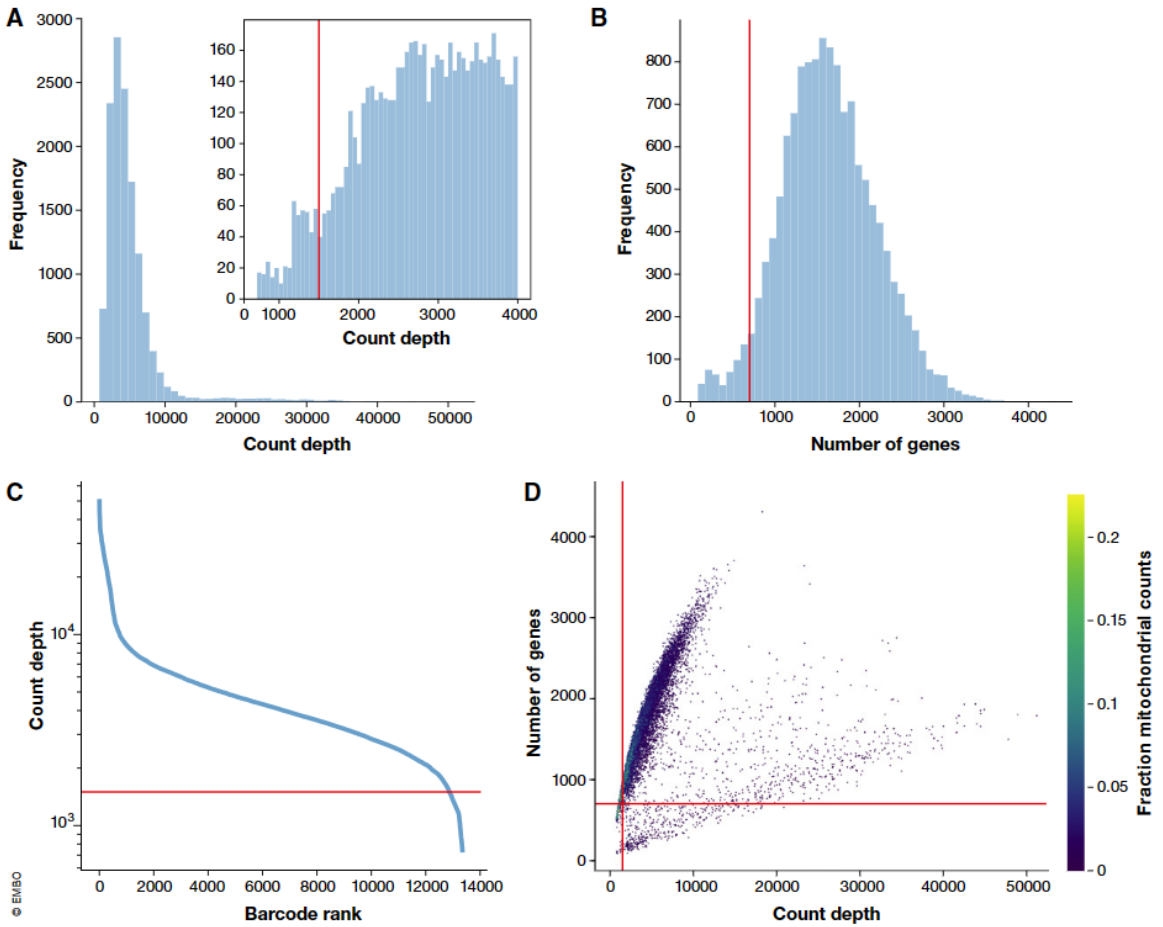


- 通过使用Cell Ranger等数据处理工具，可对原始数据进行预处理，从而得到单细胞表达矩阵。
- Cell Ranger是10x Genomics为其单细胞测序数据提供的上游分析软件，能够将序列与基因组比对并进行基因表达定量。

(www.10xgenomics.com)

4.1 单细胞转录组数据预处理和质控

质量控制



质控变量分布图

低质量细胞

检测基因数目少和线粒体UMI比例高的细胞代表膜破裂的濒死细胞。

双细胞/多细胞

基因数目过多的细胞可能为双细胞或多细胞。

DoubletFinder、Scrublet、scDblFinder等软件可用于识别双细胞。

为确保下游分析细胞的活性，通常基于三个指标进行细胞质量控制：

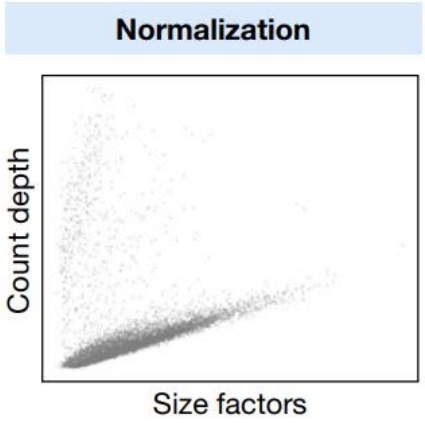
- 每个细胞的UMI数；
- 每个细胞的基因数；
- 每个细胞的线粒体基因来源UMI所占比例。

(Luecken MD, et al. Mol Syst Biol, 2019)

4.2 数据标准化、校正及整合

标准化

- 表达矩阵中的每个计数代表了细胞内mRNA成功捕获、逆转录和测序等多个步骤综合的结果。
- 每个步骤固有的可变性使得即使相同类型细胞的UMI深度也可能不同，进而影响细胞之间基因表达的比较。



标准化后矩阵

标准化

标准化旨在消除这些差异，使它们不干扰细胞之间表达谱的比较，确保在细胞群体中观察到的异质性或差异表达是由生物学而不是技术偏倚引起的。

标准化的处理

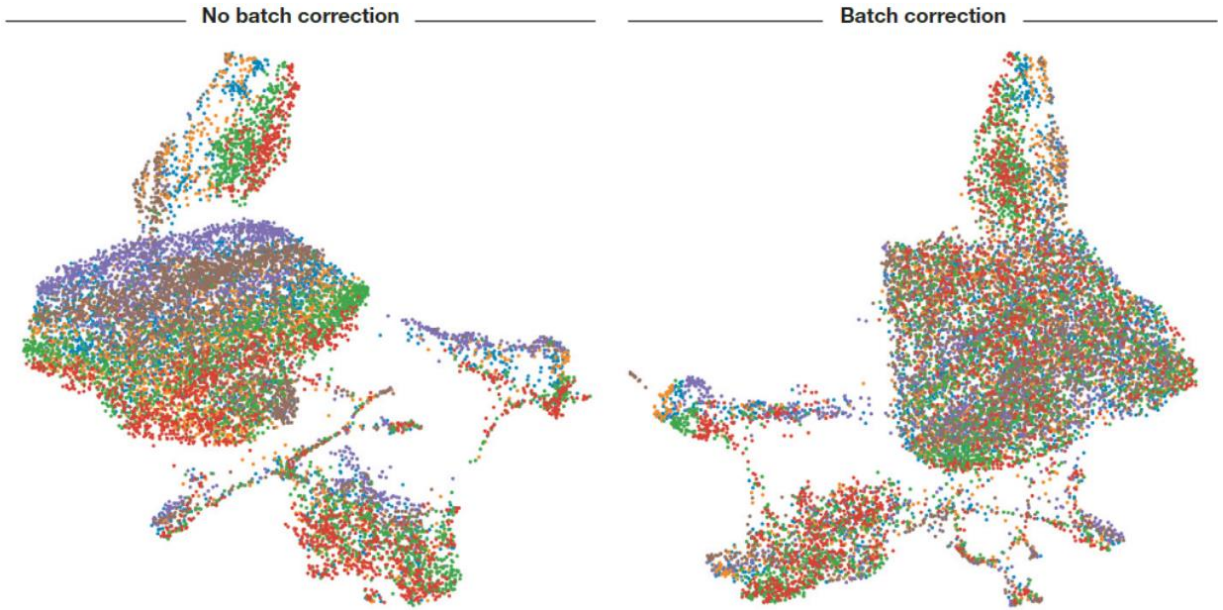
R包Seurat中NormalizeData函数可以实现标准化，即每个基因的UMI数除以该细胞总的UMI数，乘以scale.factor（默认10000），再使用log1p对这些值进行对数转换，从而实现数据的标准化处理。

(Luecken MD, et al. Mol Syst Biol, 2019)
(Heumos L, et al. Nat Rev Genet, 2023)

4.2 数据标准化、校正及整合

数据校正与整合

数据校正的对象包括技术和生物因素，例如不同批次、线粒体转录本、核糖体转录本、细胞周期等，这些都可能对后面的分析产生影响，因此需要消除这些差异来源。



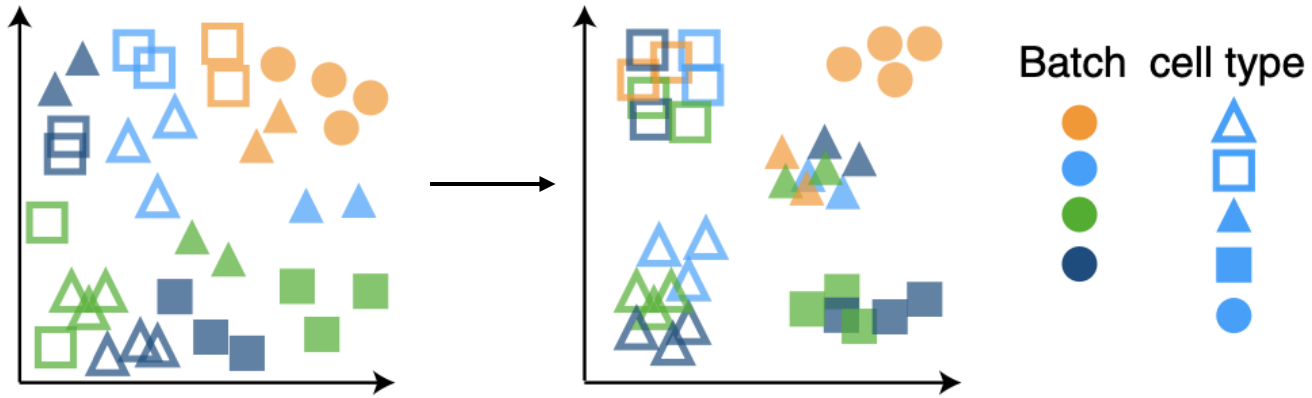
批次校正前后的UMAP图

- 常见生物因素细胞周期的校正可以通过对细胞周期评分进行简单线性回归来实现。
- 单细胞数据还有其他的噪音来源，其中一个就是dropout。一些工具可以用来推断dropout，用适当的表达量来替代0值，例如MAGIC、WEDGE、DCA、scVI等。
- 在多样本整合过程中，校正样本测序深度、批次、噪音等技术因素的方法包括线性校正和非线性校正。

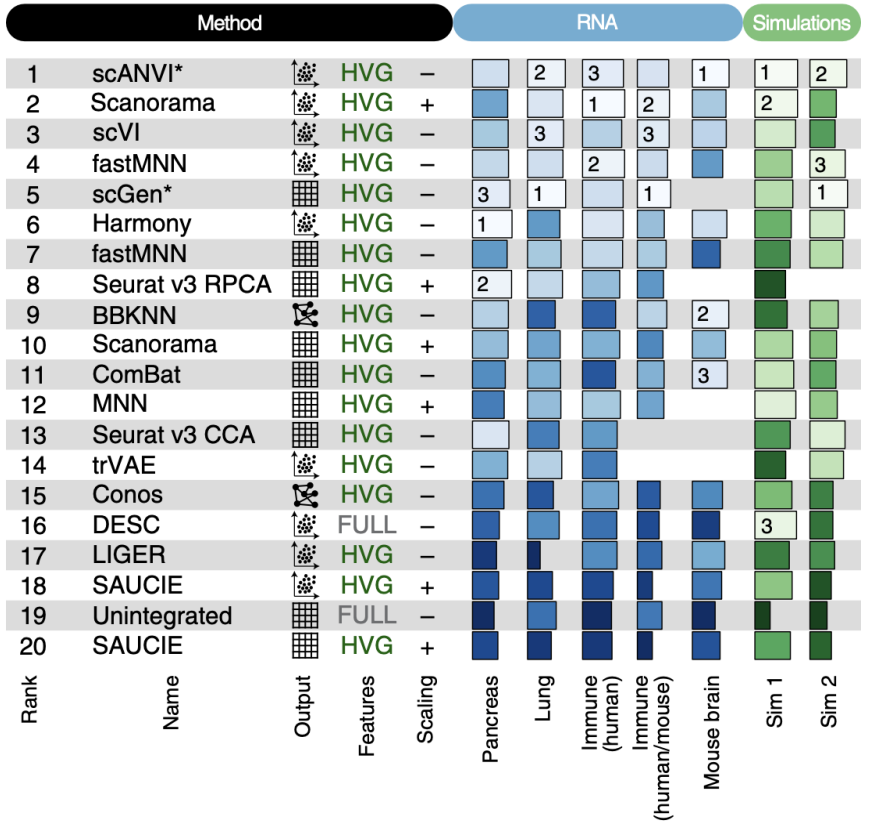
(Luecken MD, et al. Mol Syst Biol, 2019)

4.2 数据标准化、校正及整合

数据校正与整合



去除批次示意图



- 数据整合（去批次）的目的是将不同批次的数据整合在一起进行分析，包括不同的样本、不同的数据集、不同的测序技术等
- 去批次需要在去除样本批次效应的同时保留生物学意义
- 目前常用的算法包括Seurat软件的CCA、Harmony等。

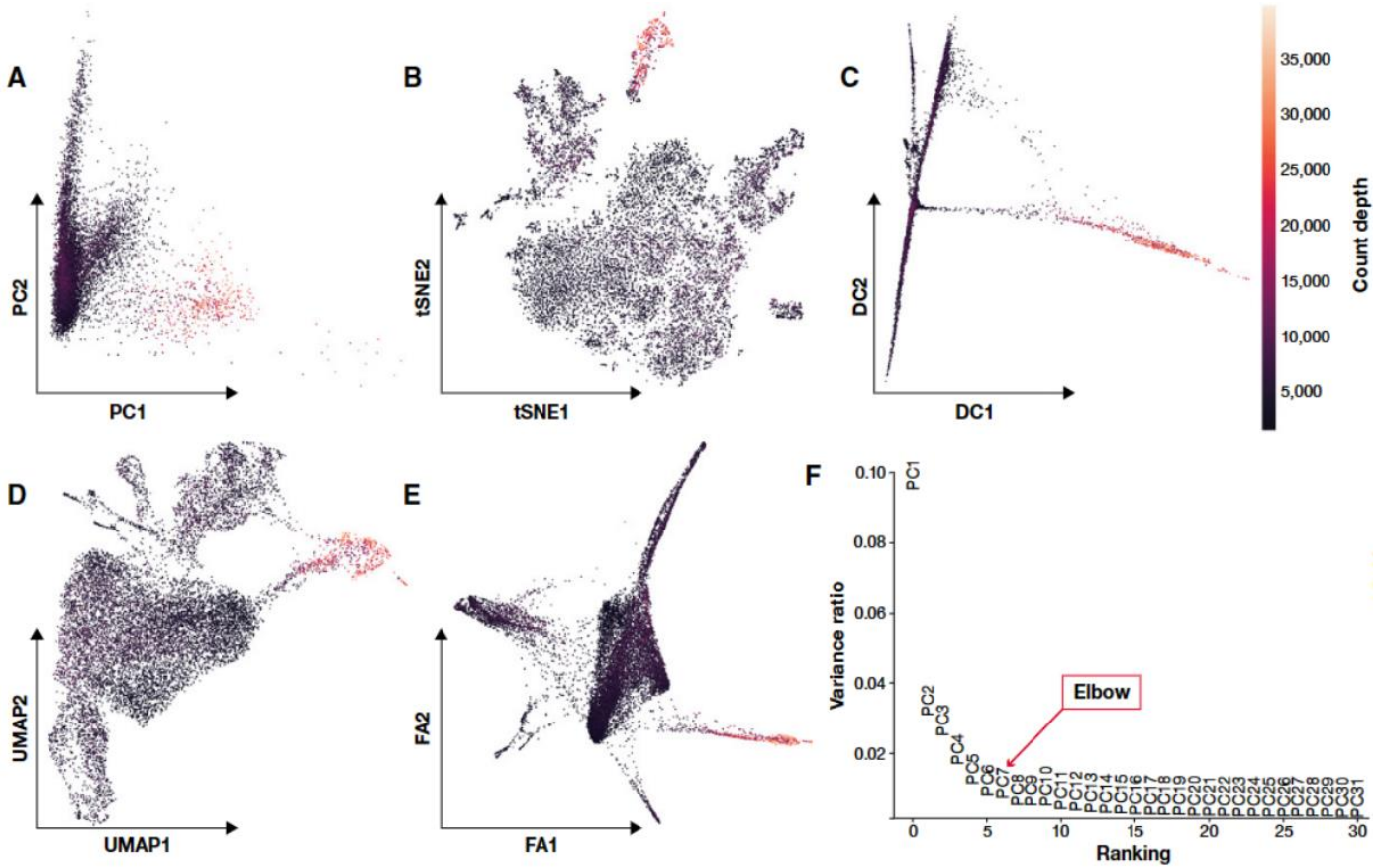
(Luecken MD, et al. Nat Methods, 2022)

4.3 降维、聚类和细胞注释

降维

单细胞测序数据作为一个高维的复杂数据，其包含上万个细胞样本和基因数量。

- 为了便于解读这些数据需要降低单细胞表达矩阵的维度，常用的降维方法包括PCA、t-SNE、UMAP、diffusion maps等。



scRNA测序数据的常见可视化方法

(Luecken MD, et al. Mol Syst Biol, 2019)

4.3 降维、聚类和细胞注释

聚类和注释

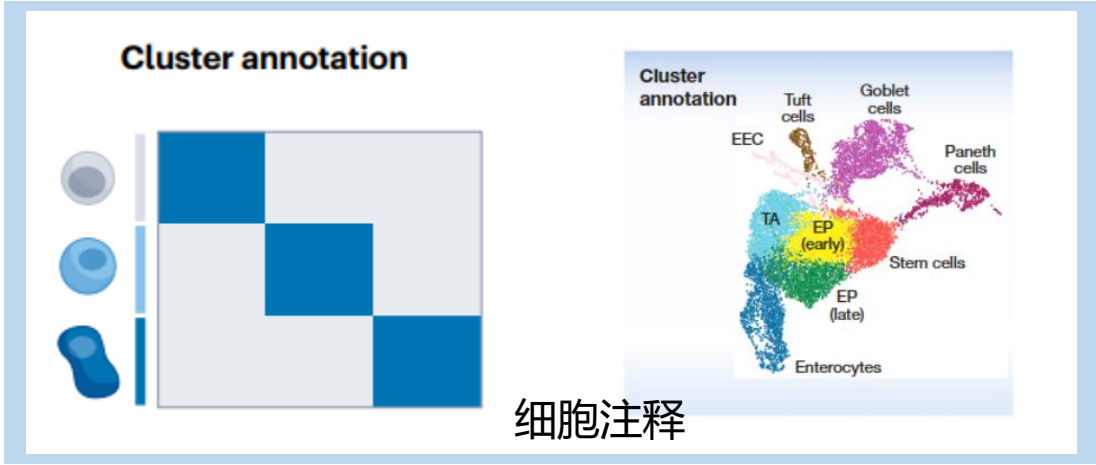
聚类

- 聚类是指根据细胞基因表达谱相似性对细胞进行分组而获得细胞簇。
- 表达谱相似性通常将降维结果作为输入，通过距离度量来确定。目前主要有两种方法根据相似性生成细胞簇：聚类算法（clustering algorithms）和社区检测算法（community detection methods）。

细胞聚类

细胞注释

- 理论上在一起的细胞属于同类细胞。但聚类结果并不能直观地呈现各细胞群所代表的细胞类型，因此需要通过多种方法来进行细胞注释以识别对应的细胞类型。
- 常用的方法包括手动注释、singleR自动注释、基于参考数据库注释等。

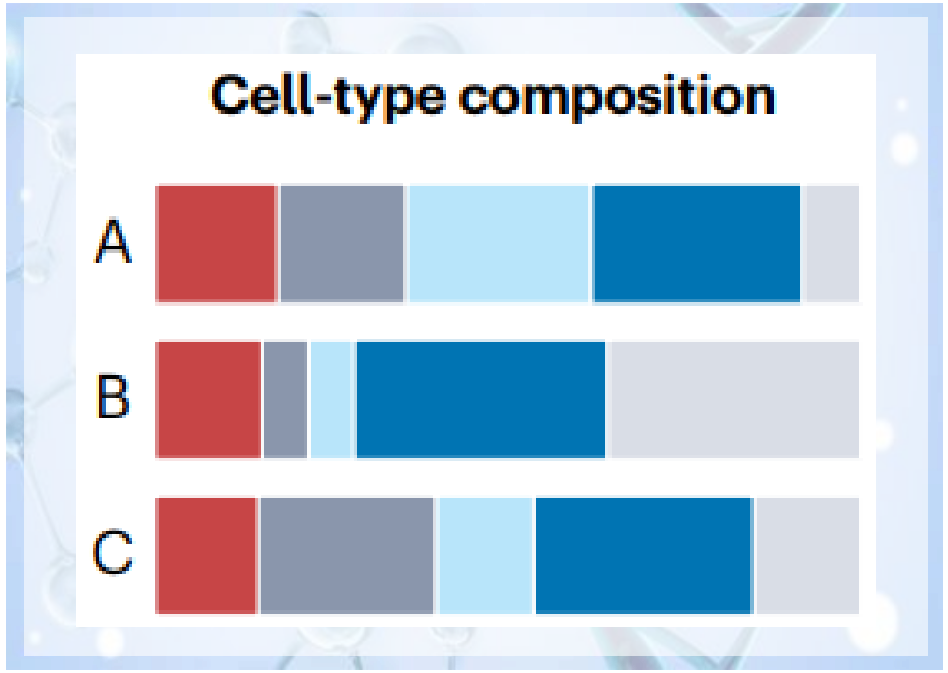


细胞注释

(Luecken MD, et al. Mol Syst Biol, 2019)
(Heumos L, et al. Nat Rev Genet, 2023)

4.4 差异细胞比例和差异表达基因分析

细胞比例



细胞组成

- 细胞比例研究主要关注不同细胞类型的相对丰度。在发育和疾病中经常观察到细胞比例变化或特异的细胞类型，但细胞比例分析方法缺乏独立的基准。

- 专门为单细胞数据中细胞组成分析而设计的方法包括scDC、scCODA和tascCODA等。这些方法可以纳入分层细胞类型信息。

(Heumos L, et al. Nat Rev Genet, 2023)

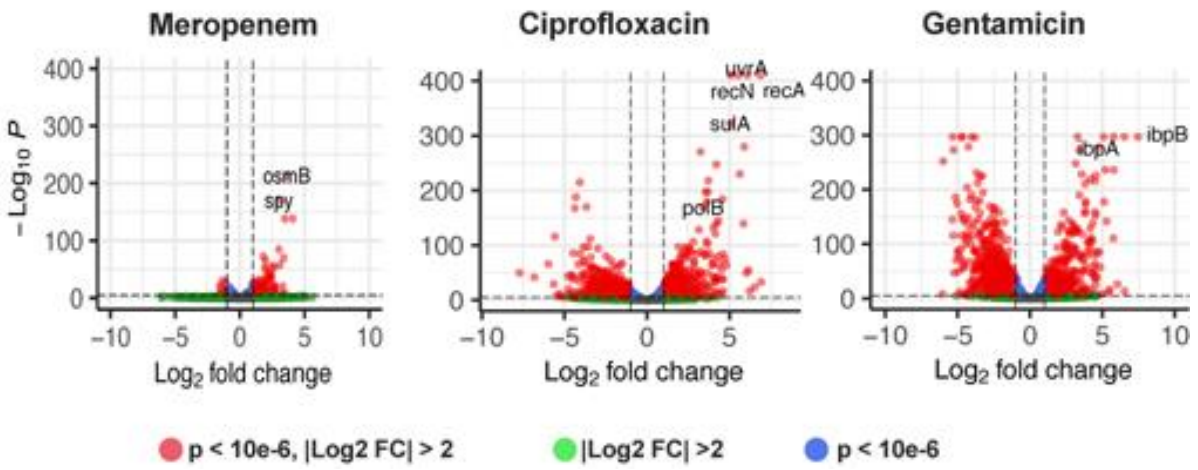
4.4 差异细胞比例和差异表达基因分析

差异表达基因分析

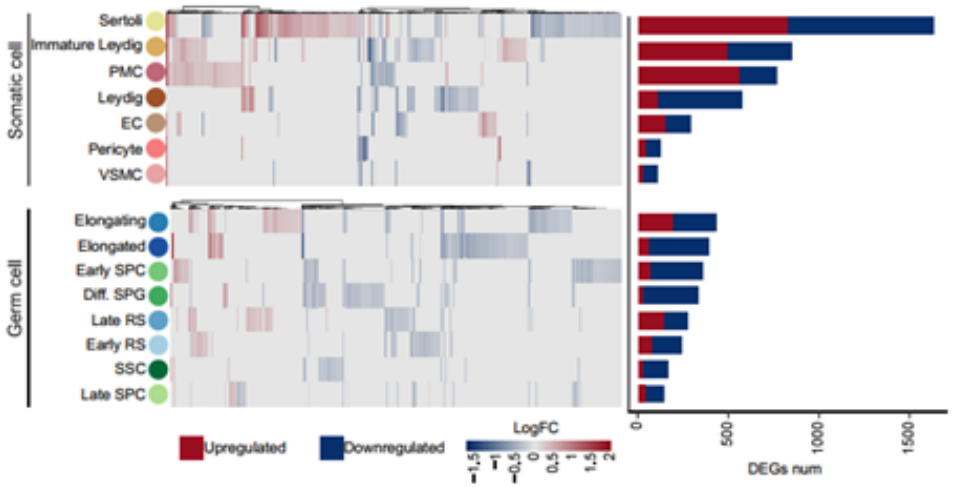
差异表达基因分析可以确定聚类之间的差异，识别细胞类型及其标记，研究细胞分化、疾病发展或暴露于外源性和/或内源性因素过程中的转录动力学

单细胞转录组数据的差异基因表达分析目前主要从两个角度进行:

- 在样本层面，可以通过计算聚合（pseudo-bulk）表达，对每个样本细胞中的基因表达求和或平均，从而得到与每个样本对应的pseudo-bulk表达量，并使用软件包DESeq2、edgeR或limma进行差异表达分析.
- 在细胞群层面，Seurat 软件包提供了多种计算差异表达基因的方法，可以通过FindMarkers 函数中的 test.use 参数设置检验差异表达基因的方法。



火山图 Bulk RNA-seq 展示药物处理组和对照的差异



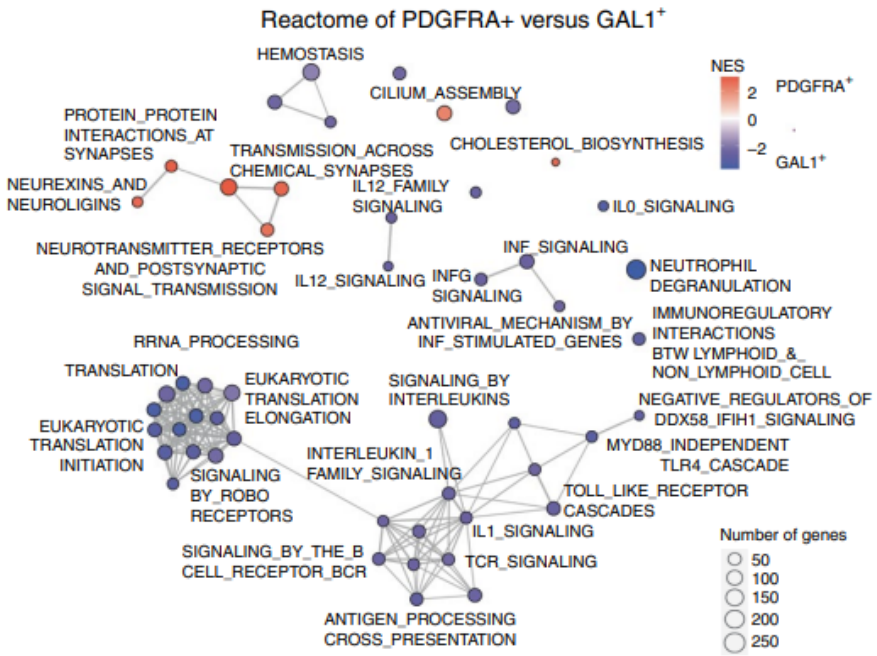
差异表达基因的热图，统计数量图

(Ma PJ, et al. Cell, 2023)
(Huang DY, et al. Protein & Cell, 2023)

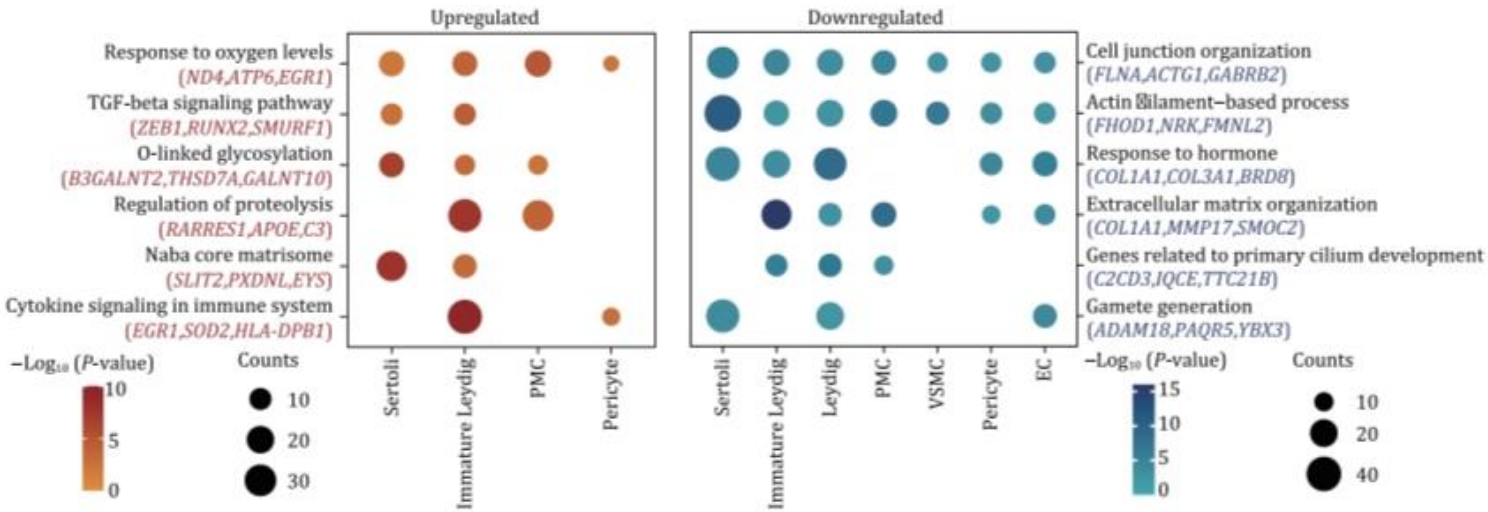
4.4 差异细胞比例和差异表达基因分析

基因集富集分析

- 对目标差异表达基因集进行基因集富集分析，可以将差异表达基因集总结为可解释的术语，例如通路。常见的数据库包括 MSigDB、Gene Ontology（GO）、KEGG 或 Reactome。



Reactome通路富集



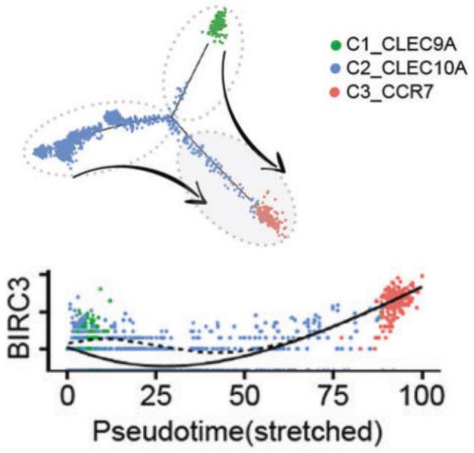
GO通路富集

(Yeo AT, et al. Nat Immunol, 2022)
(Huang DY, et al. Protein & Cell, 2023)

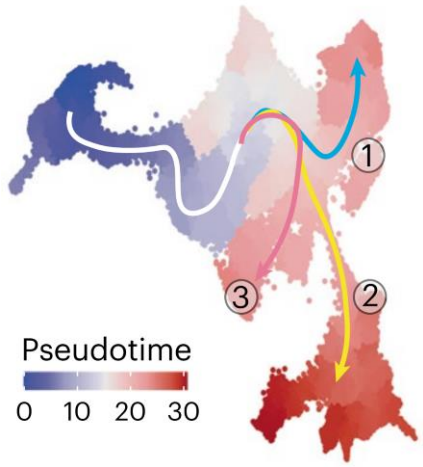
4.5 细胞发育轨迹推断

细胞发育轨迹推断

- 基于基因表达模式相似性排序细胞的伪时间方法。
- 最早用于构建发育轨迹且应用广泛的伪时间分析工具之一是Monocle，它借鉴显式主图（如t-SNE和UMAP）来展示细胞转录特征相似性关系，并通过嵌入反向图来重建单细胞轨迹，有较高的稳健性和准确性。

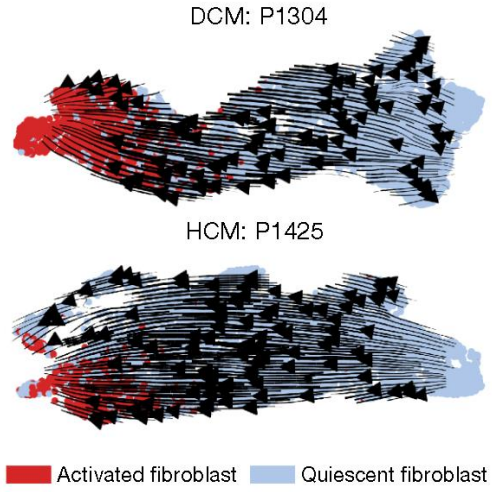


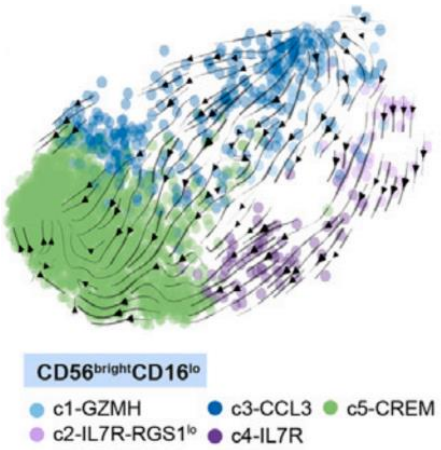
monocle2



monocle3

- 基于RNA剪接动力学的RNA速率（RNA velocity）方法。
- RNA速率是一种可以预测单细胞未来状态的高维矢量，通过估计基因的未剪接RNA（前体RNA，pre-mRNA）与成熟RNA（已剪接RNA）的比例来推断基因的转录状态和预测细胞的发育轨迹。



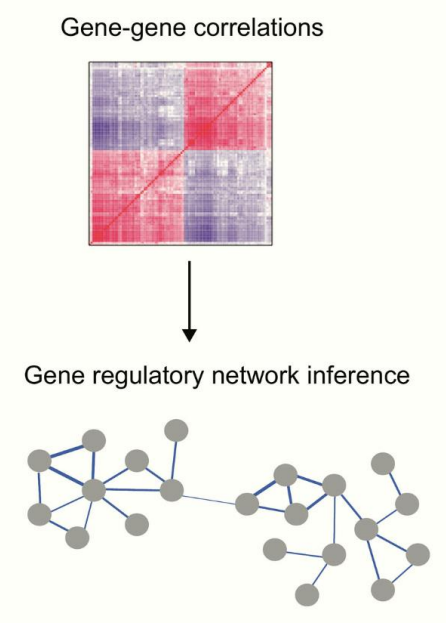


RNA velocity

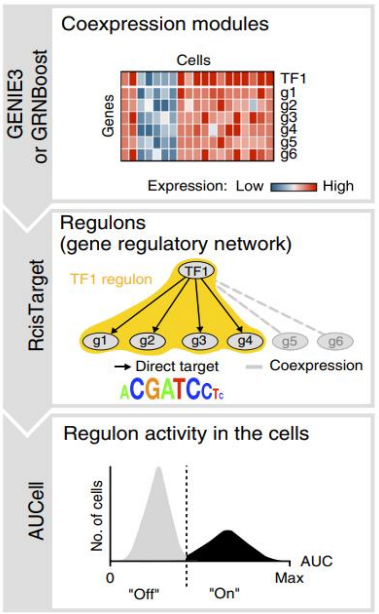
(Qian JB, et al. Cell Res, 2020)
(Chu YS, et al. Nat Med,2023)
(Chaffin M, et al. Nature, 2022)
(Tang F, et al. Cell, 2023)

4.6 基因调控网络

基因调控网络分析



SCENIC workflow



- 基因表达受到转录以及转录后调控多个层面的调节。转录因子及其下游靶基因形成的基因调控网络（Gene Regulatory Network, GRN）在基因表达调控中起着重要作用。
- 基因调控网络参与细胞稳态的维持和细胞异质性的形成。
- 单细胞转录组层面的基因调控网络分析方法主要有SCENIC（single-cell regulatory network inference and clustering）和hdWGCNA（high dimensional weighted gene co-expression network analysis）。
- SCENIC基于共表达和DNA调控基序（motif）分析来推断单细胞转录组基因调控网络。
- hdWGCNA建立在WGCNA基础上，扩展了其在单细胞和空间转录组学领域的应用，可用于分析高维转录组学（单细胞和空间转录组学）数据中的共表达网络。

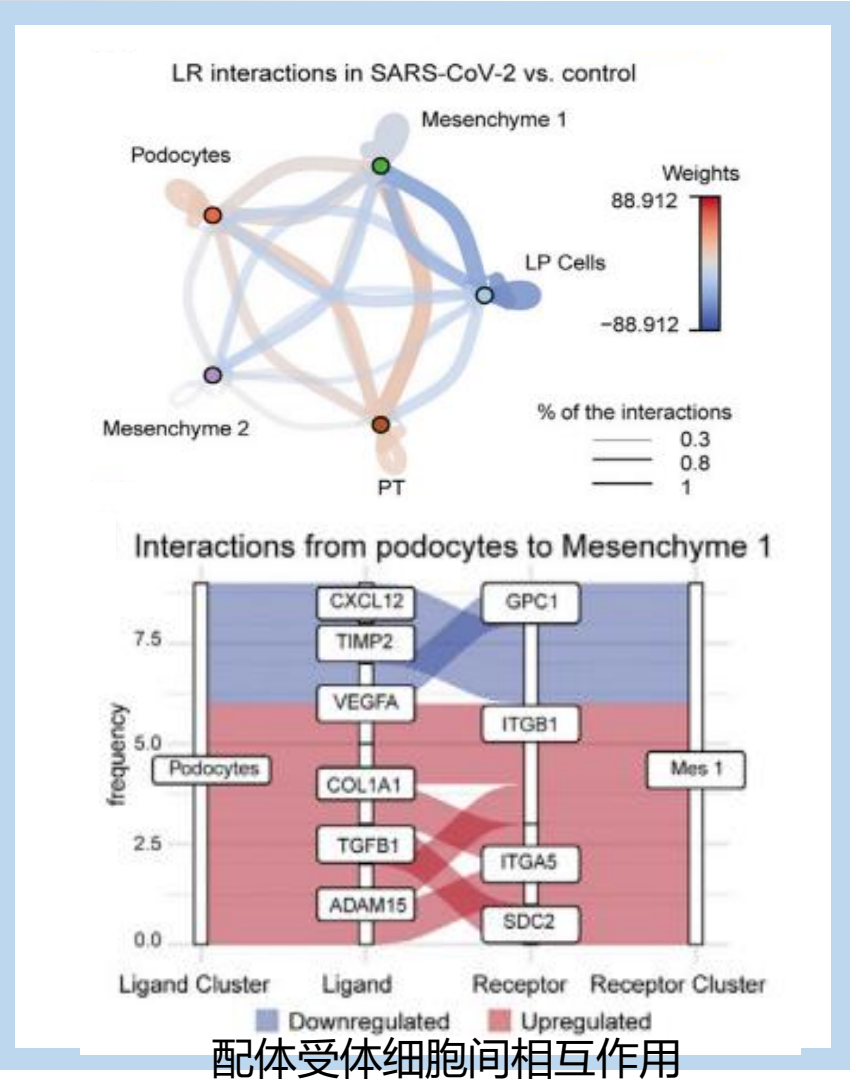
(Aibar S, et al. Nat Methods, 2017)
(Morabito S, et al. Cell Rep Methods, 2023)

4.7 细胞通讯分析

细胞通讯

细胞间通讯的推断方法通常使用配体、受体及其相互作用库来预测不同细胞类型之间的相互作用。

- 常见的分析工具包括CellPhoneDB、CellChat、NicheNet 、 SingleCellSignalR 、 Cytotalk 和 iTalk 等 。



配体受体细胞间相互作用

(Jansen J, et al. Cell Stem Cell, 2022)

4.8 单细胞转录组数据分析总结

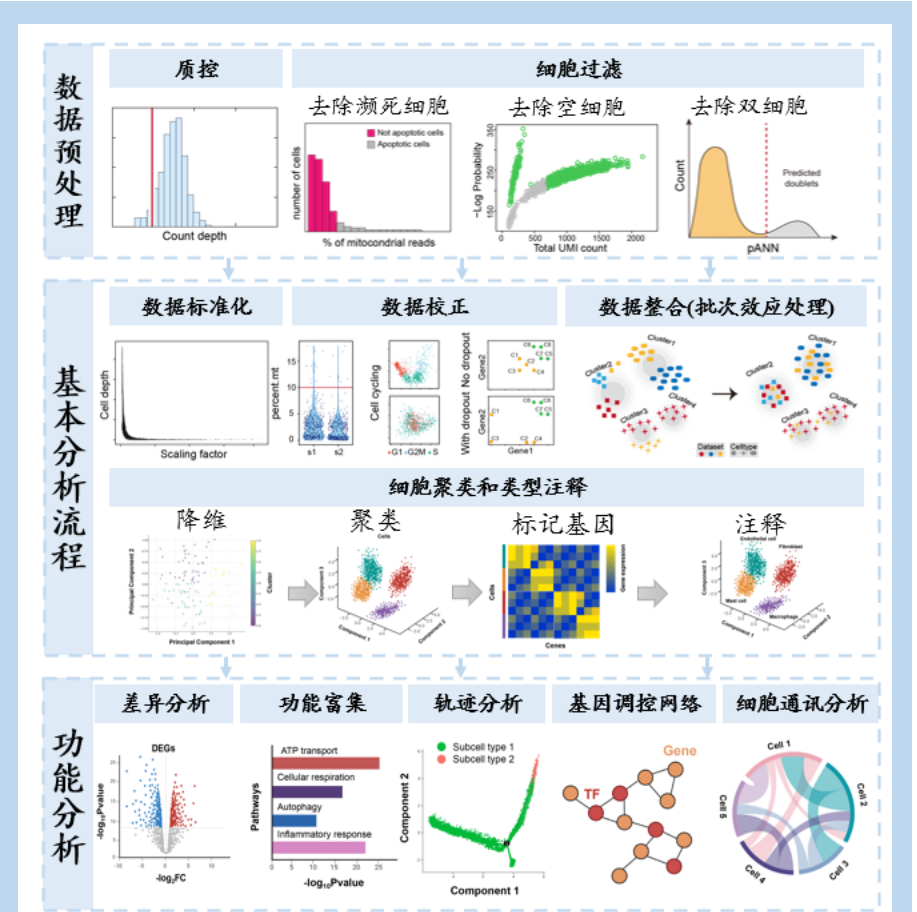
总结

单细胞转录组数据分析包括多个步骤，它主要涉及到以下几个环节。

1、数据预处理：通过细胞质控，在此基础上对数据进行标准化，校正及整合，消除各种技术和生物因素对数据分析造成的偏差。

2：降维、聚类和细胞注释：降维聚类之后，通过手动和自动注释方法完成细胞群及细胞亚群注释。

3：细胞及基因层面多维度分析：在细胞注释后，可以开展细胞群或亚群细胞比例和差异表达基因分析，并对差异表达基因开展功能富集分析等；对于不同细胞亚群，通过拟时序或RNA速率方法对它们进行细胞发育轨迹推断；在细胞群或亚群内，通过SCENIC和WGCNA等解析细胞内基因调控网络；在细胞群或亚群间，通过CellPhoneDB、CellChat、NicheNet等方法推断细胞间通讯。



单细胞转录组数据分析

4.8 单细胞转录组数据分析总结

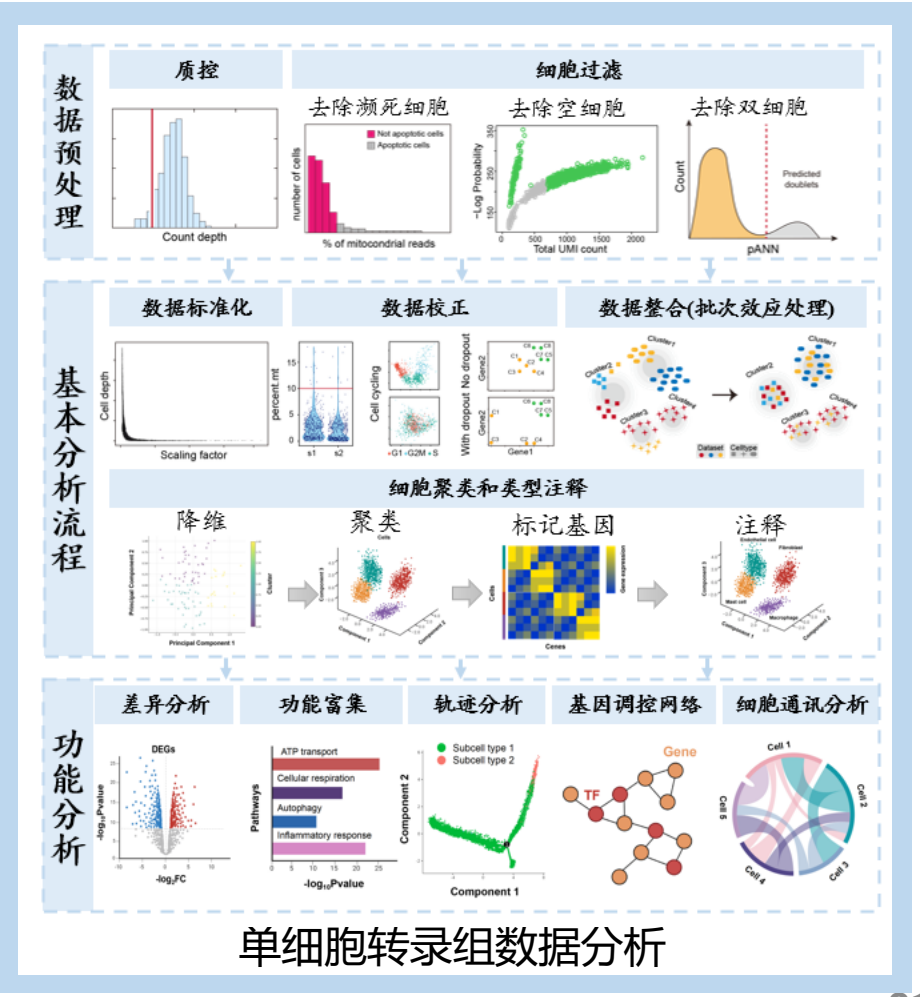
总结

单细胞转录组数据分析包括多个步骤，它主要涉及到以下几个环节。

1、数据预处理：通过细胞质控，在此基础上对数据进行标准化，校正及整合，消除各种技术和生物因素对数据分析造成的偏差。

2：降维、聚类和细胞注释：降维聚类之后，通过手动和自动注释方法完成细胞群及细胞亚群注释。

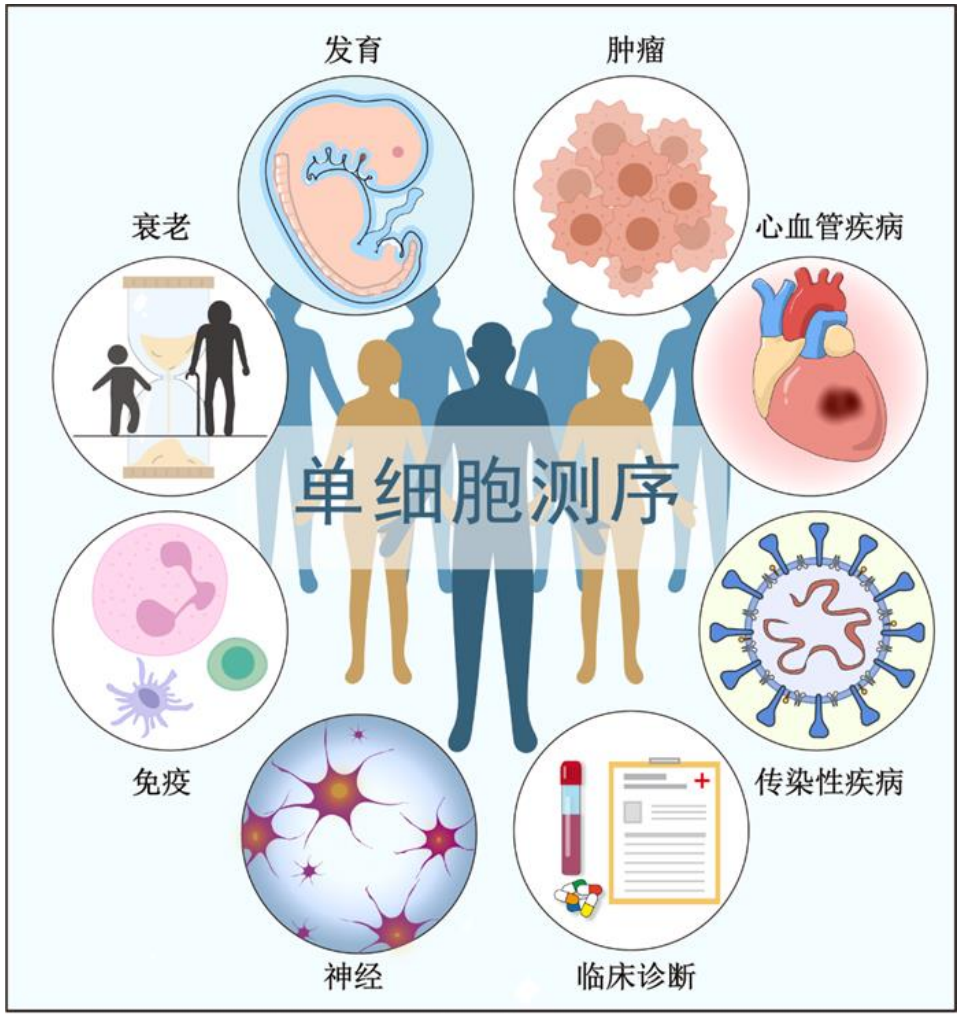
3：细胞及基因层面多维度分析：在细胞注释后，可以开展细胞群或亚群细胞比例和差异表达基因分析，并对差异表达基因开展功能富集分析等；对于不同细胞亚群，通过拟时序或RNA速率方法对它们进行细胞发育轨迹推断；在细胞群或亚群内，通过SCENIC和WGCNA等解析细胞内基因调控网络；在细胞群或亚群间，通过CellPhoneDB、CellChat、NicheNet等方法推断细胞间通讯。



单细胞转录组数据分析

4.9 单细胞转录组应用

- 单细胞转录组测序技术能够以前所未有的高分辨率准确且快速地鉴定组织中的稀有和新型细胞，能够揭示单个细胞内RNA转录本的异质性和复杂性及组织、器官和生物体中各类型细胞功能。
- 单细胞转录组测序技术的发展使得该技术可应用于生命科学几乎所有的领域。



单细胞转录组测序技术的应用

推荐阅读

- Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Review Genetics, 2019, 20:631-656
- Chen Q, Meng X, Liao Q, Chen M. Versatile interactions and bioinformatics analysis of noncoding RNAs. Brief Bioinform. 2019;20(5):1781-1794.
- Lotfollahi M, Hao Y, Theis FJ, et al. The future of rapid and automated single-cell data analysis using reference mapping[J]. Cell, 2024, 187(10): 2343-2358.

思考题

1. 请说明三种经典转录组组装策略（基于参考基因组序列的转录组组装、从头组装以及多策略混合的转录组组装）的异同，并指出各方法的主要特点。
2. 转录组测序数据可用于基因和转录本水平的定量分析。基因和表达水平的准确估计受到哪些因素的影响？如何克服这些因素的影响？
3. 举例阐明各类非编码RNA之间的关系？
4. scRNA-seq数据预处理过程去除了哪些细胞？
5. scRNA-seq数据和bulk RNA-seq数据的差异分析方法有何异同？