Chapter 2. Genome sequencing and assemble technology

Lecture 2.4 RNA-seq and Big data

鲍坚东

PPT slides and Message @ http://jxpt.fafu.edu.cn/meol/homepage/common/

•Email: bajd@fafu.edu.cn

内容

1、基因组时代的困惑？

2、RNA种类与功能简介

3、RNA-seq原理与应用

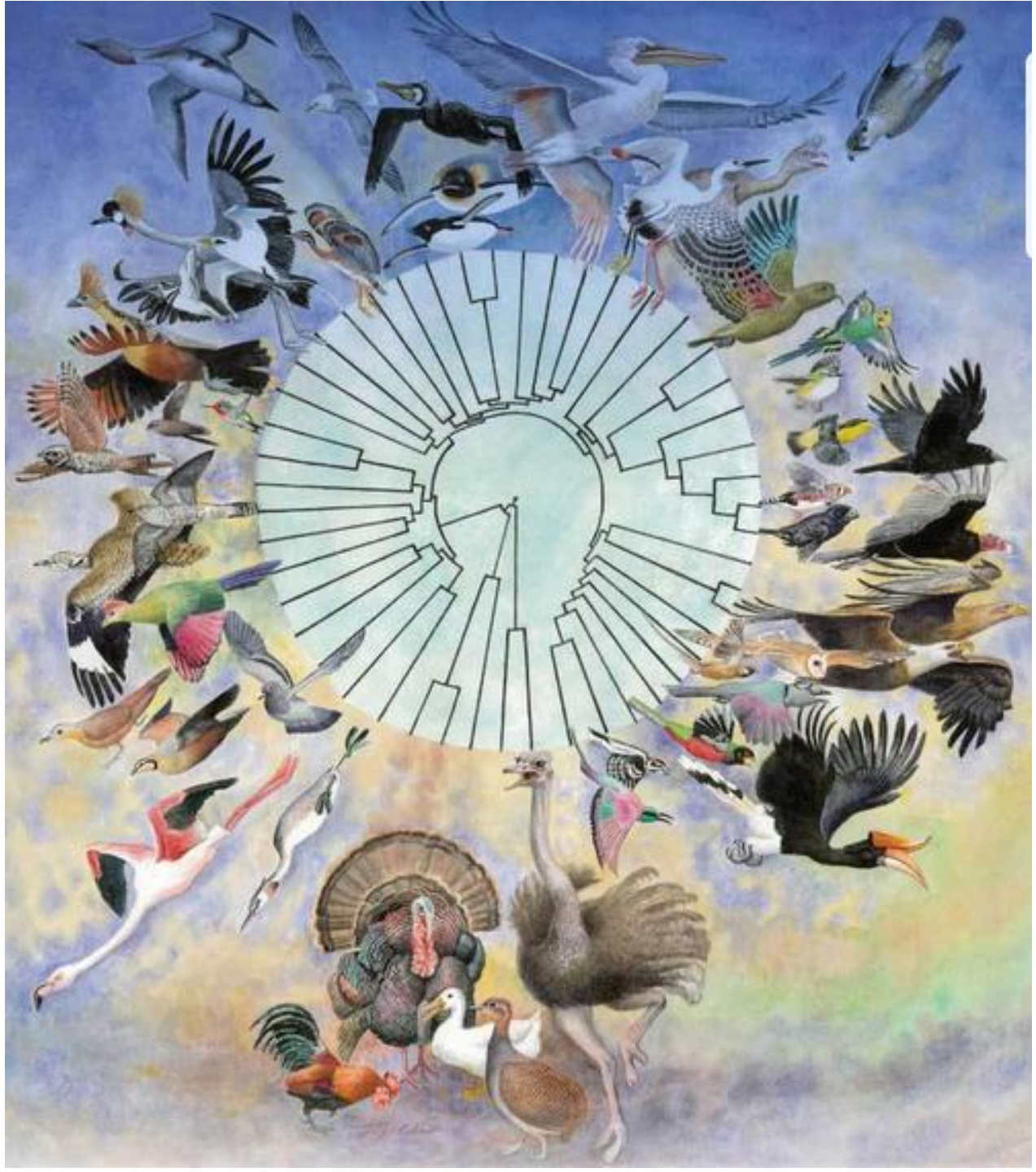
1、基因组时代的困惑？

- 基因组天书
- C值/N值悖论揭示了基因组“暗物质”存在
- 为什么人类基因数这么少
- 基因-RNA-蛋白，RNA=mRNA，简单信息传递？
- “暗物质”=垃圾？非编码RNA

大量物种基因组已经测序

NCBI-Genome
(2016-10-16)

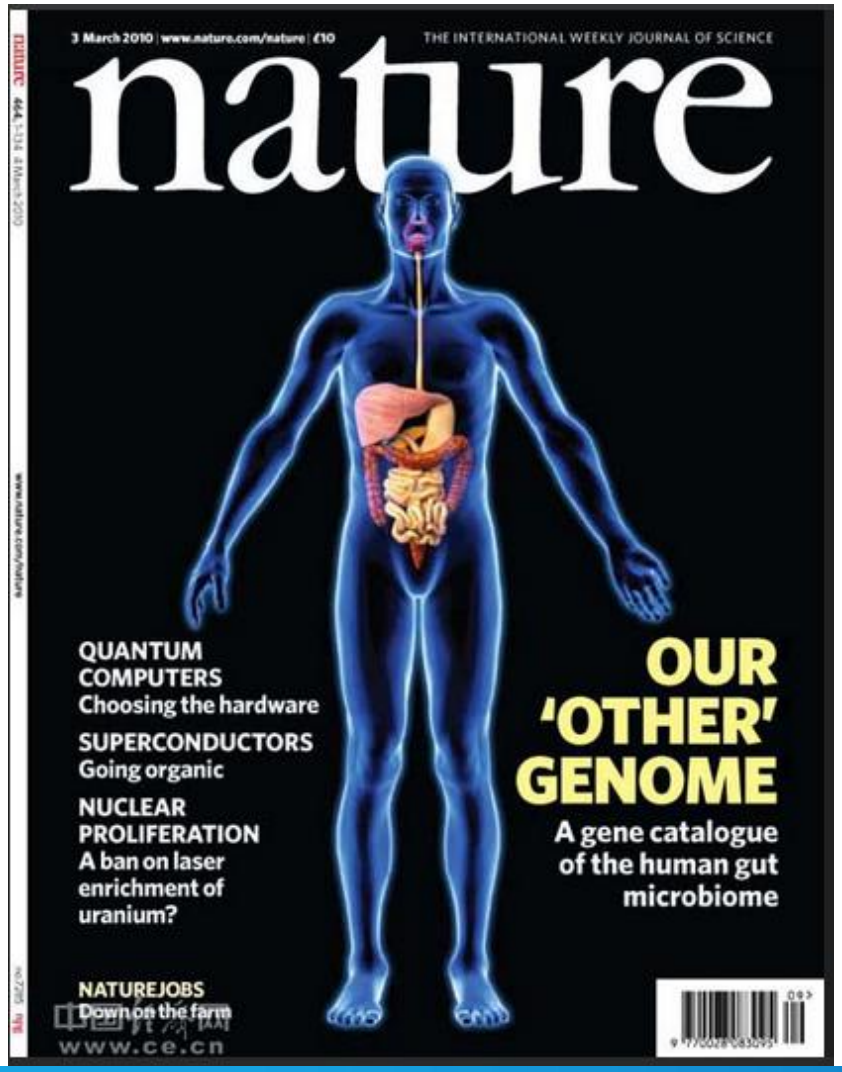
真核生物：3716
原核生物：75302
病毒：7799
细胞器：8748

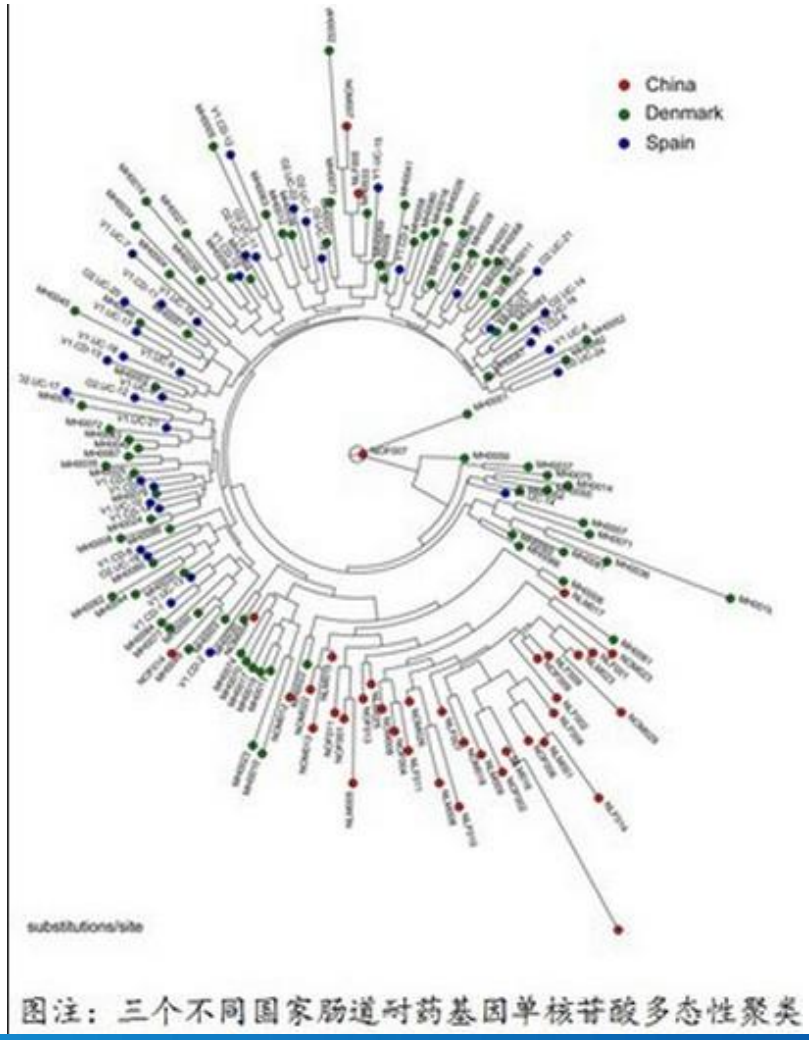


群体基因组测序揭示现代鸟类起源、分化与适应。

48种，全基因组测序，飞行、语言、牙齿丢失等机制

人类肠道宏基因组学





图注：三个不同国家肠道耐药基因单核苷酸多态性聚类

基因组测序进入个体测序时代

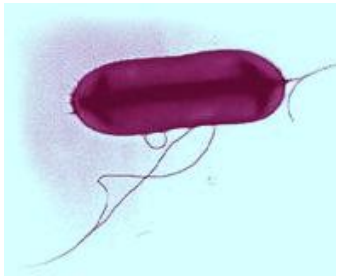
- 1990 年人来基因组计划　　10年 10亿美元
- 2008 年千人基因组计划
- 2012年 十万人基因组计划（英国）
- 2015年 百万人基因组计划

精准医疗计划项目（美国，$2.15亿）

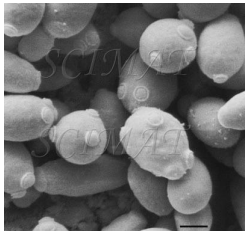
Illunima测序成本是￥1000元 3Gb+建库

Hiseq X Ten 18,000个人类基因组/年

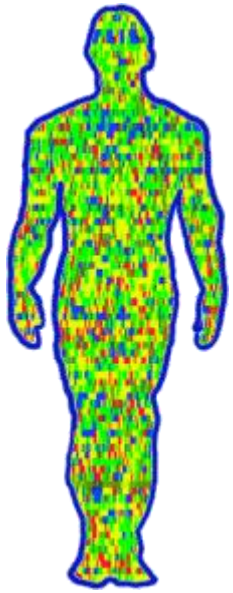
基因组大小(C值)和基因数目（N值）

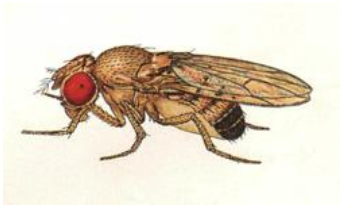


大肠杆菌
4.5Mb
4400基因



啤酒酵母
12Mb
6600基因



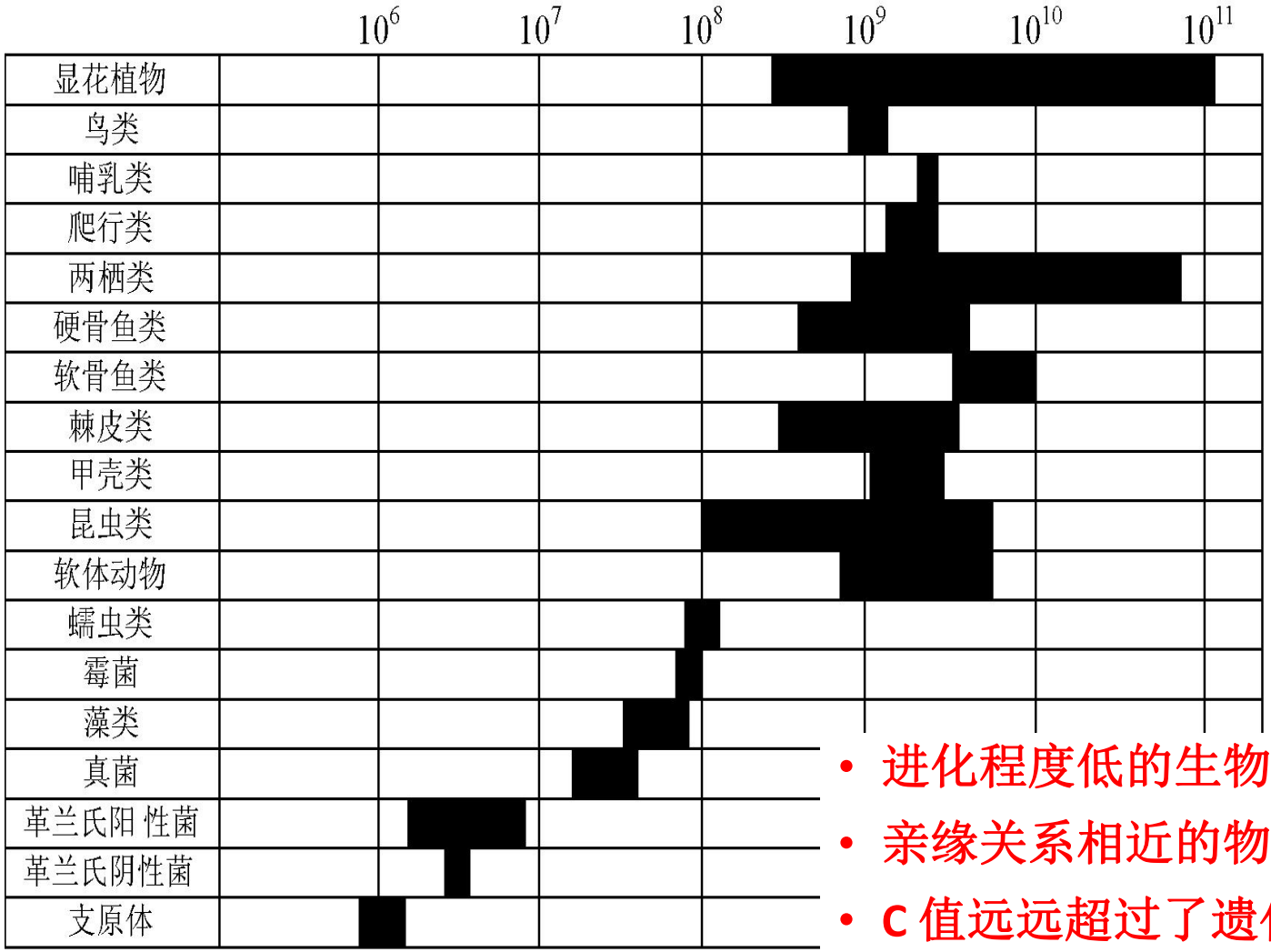


果蝇 120Mb 13000基因
人 3Gb 20000-25000基因

蚜虫
430Mb
34000基因

蝗虫
6.5Gb
17000基因

基因组大小/基因数量与
生物遗传信息量背离



C值悖论

- 进化程度低的生物 C 值反而更高。
- 亲缘关系相近的物种间 C 值差异很大。
- C 值远远超过了遗传信息量的需要。

疑问

Special Section

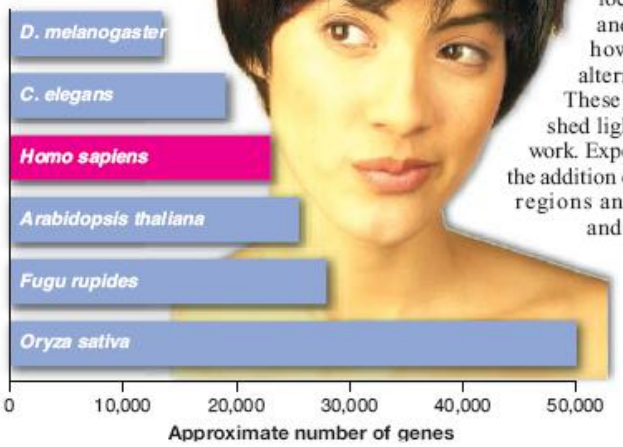
Why Do Humans Have So Few Genes

When leading biologists were unraveling the sequence of the human genome in the late 1990s, they ran a pool on the number of genes contained in the 3 billion base pairs that make up our DNA. Few bets came close. The conventional wisdom a decade or so ago was that we need about 100,000 genes to carry out the myriad cellular processes that keep us functioning. But it turns out that we have only about 25,000 genes—about the same number as a tiny flowering plant called *Arabidopsis* and barely more than the worm *Caenorhabditis elegans*.

That big surprise reinforced a growing realization among geneticists: Our genomes and those of other mammals are far more flexible and complicated than they once seemed. The old notion of one gene/one protein has gone by the board: It is now clear that many genes can make more than one protein. Regulatory proteins, RNA, noncoding bits of DNA, even chemical and structural alterations of the genome itself control how, where, and when genes are expressed. Figuring out how all these elements work together to choreograph gene expression is one of the central challenges facing biologists.

In the past few years, it has become clear that a phenomenon called alternative splicing is one reason human genomes can produce such complexity with so few genes. Human genes contain both coding DNA—exons—and noncoding DNA. In some genes, different combinations of exons can become active at different times, and each combination yields a different protein. Alternative splicing was long considered a rare hiccup during transcription, but researchers have concluded that it may occur in half—some say close to all—of our genes. That finding goes a long way toward explaining how so few genes can produce hundreds of thousands of different proteins. But how the transcription machinery decides which parts of a gene to read at any particular time is still largely a mystery.

The same could be said for the mechanisms that determine which genes or suites of genes are turned on or off at particular times and places. Researchers are discovering that each gene needs a supporting cast of hundreds to get its job done. They include proteins that shut down or activate a gene, for example by adding acetyl or methyl groups to the DNA. Other proteins, called transcription factors, interact with the genes more directly: They bind to landing sites situated near the gene under their control. As with alternative splicing, activation of different combinations of landing sites makes possible exquisite control of gene expression, but researchers have yet to figure out exactly how all these regulatory elements really work or how they fit in with alternative splicing.



In the past decade or so, researchers have also come to appreciate the key roles played by chromatin proteins and RNA in regulating gene expression. Chromatin proteins are essentially the packaging for DNA, holding chromosomes in well-defined spirals. By slightly changing shape, chromatin may expose different genes to the transcription machinery.

Genes also dance to the tune of RNA. Small RNA molecules, many less than 30 bases, now share the limelight with other gene regulators. Many researchers who once focused on messenger RNA and other relatively large RNA molecules have in the past 5 years turned their attention to these smaller cousins, including microRNA and small nuclear RNA. Surprisingly, RNAs in these various guises shut down and otherwise alter gene expression. They also are key to cell differentiation in developing organisms, but the mechanisms are not fully understood.

Researchers have made enormous strides in pinpointing these various mechanisms. By matching up genomes from organisms on different branches on the evolutionary tree, genomicists are locating regulatory regions and gaining insights into how mechanisms such as alternative splicing evolved. These studies, in turn, should shed light on how these regions work. Experiments in mice, such as the addition or deletion of regulatory regions and manipulating RNA, and computer models should also help. But the central question is likely to remain unsolved for a long time: How do all these features meld together to make us whole?

—ELIZABETH PENNISI

CREDIT: JUPITERIMAGES

What Genetic Changes Made Us Uniquely Human

mentary culture, parrots speak, and some rats seem to giggle when tickled.

What is beyond doubt is that humans, like every other species, have a unique genome shaped by our evolutionary history. Now, for the first time, scientists can address anthropology's fundamental question at a new level: What are the genetic changes that make us human?

With the human genome in hand and primate genome data beginning to pour in, we are entering an era in which it may become possible to pinpoint the genetic changes that help separate us from our closest relatives. A rough draft of the chimp sequence has already been released, and a more detailed version is expected soon. The genome of the macaque is nearly complete, the orangutan is under way, and the marmoset was recently approved. All

Half of the differences might define a chimp rather than a human. How can we sort them all out?

One way is to zero in on the genes that have been favored by natural selection in humans. Studies seeking subtle signs of selection in the DNA of humans and other primates have identified dozens of genes, in particular those involved in host-pathogen interactions, reproduction, sensory systems such as olfaction and taste, and more.

But not all of these genes helped set us apart from our ape cousins originally. Our genomes reveal that we have evolved in response to malaria, but malaria defense didn't make us human. So some researchers have started with clinical mutations that impair key traits, then traced the genes' evolution, an tide of genomic data with more phenotypic information on apes. Other researchers argue that clues to function can best be gleaned by mining natural human variability, matching mutations in living people to

subtle differences in biology and behavior. Both strategies face logistical and ethical problems, but some progress seems likely.

A complete understanding of uniquely human traits will, however, include more than DNA. Scientists may eventually circle back to those long-debated traits of sophisticated language, culture, and technology, in which nurture as well as nature plays a leading role. We're in the age of the genome, but we can still recognize that it takes much more than genes to make the human.

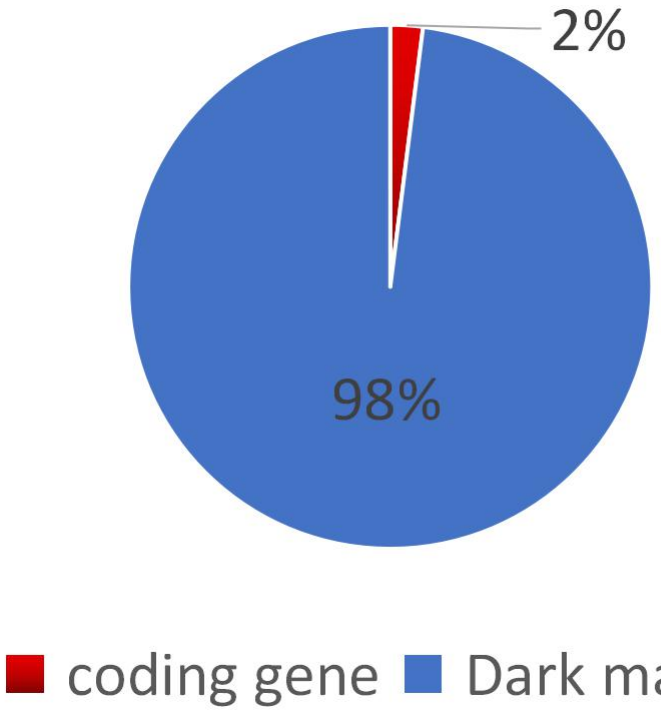
—Elizabeth Culotta

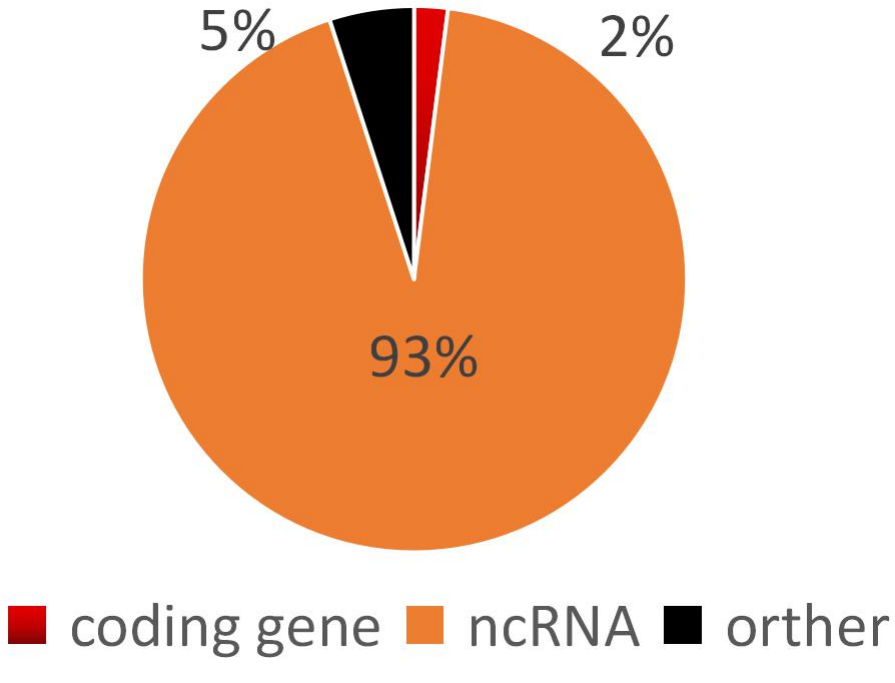
CREDITS (LEFT TO RIGHT): FRITZ POLKING/VISUALS UNLIMITED; TERRY HUSBEY/GETTY IMAGES

红毛猩猩：1400万，97%；
大猩猩：　1000万，98%；
黑猩猩：600万年前，99%

编码基因难以解释人与猩猩巨大分化
某种暗物质存在主导了这种分化？

人类基因组中的暗物质

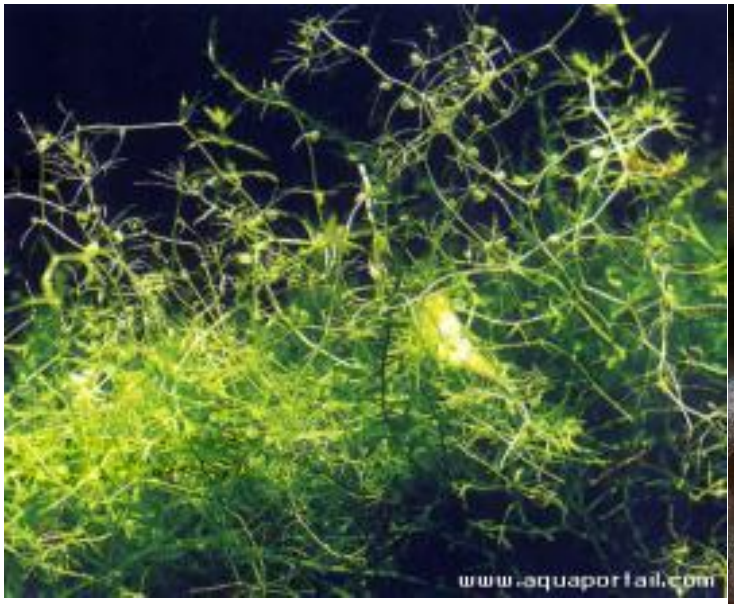


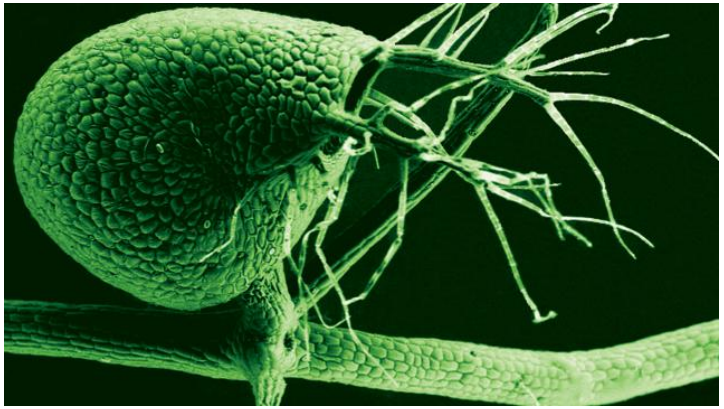


人类基因组95%可以转录成RNA，但只有2%区域是编码蛋白

非编码RNA有什么功能？

问题：高等生物基因组都是充满暗物质吗？





丝叶狸藻（*Utricularia gibba*）具有可活动囊状捕虫结构的小型食虫植物

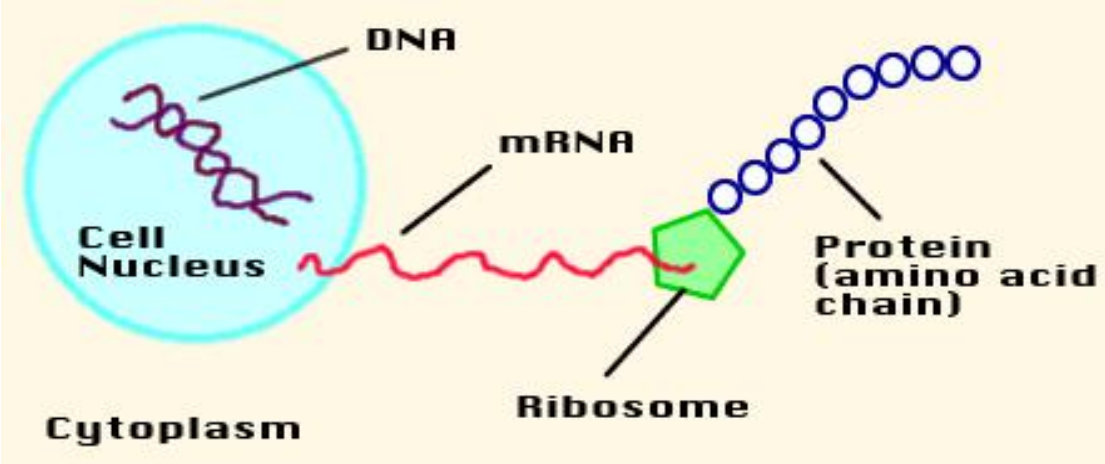
基因组：82Mb
28500个基因

只有3% ncRNA

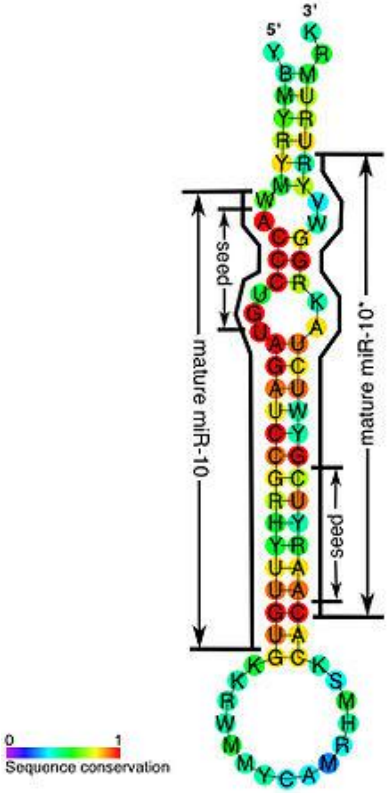
2、RNA种类与功能简介

三种主要RNA

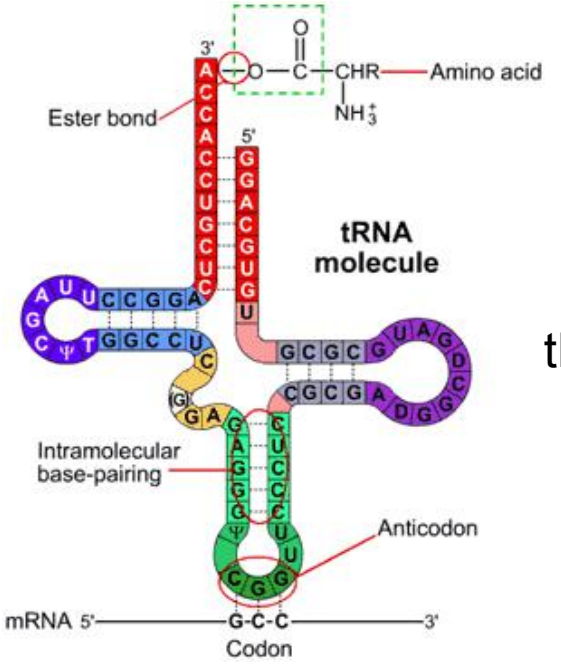
1) Messenger RNA(mRNA)： conding RNA
2) Transfer RNA (tRNA)：转运RNA
3) Ribosomal RNA (rRNA)：核糖体RNA



非编码RNA（ncRNA）

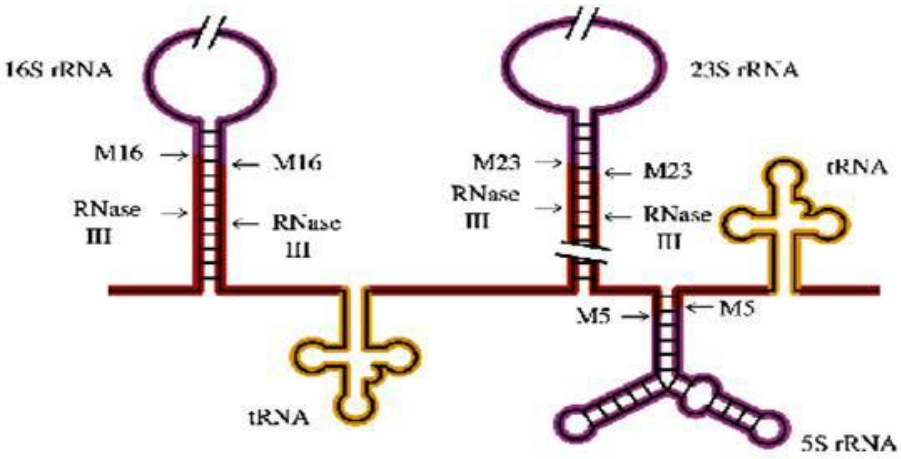


miRNA
20-24 nt



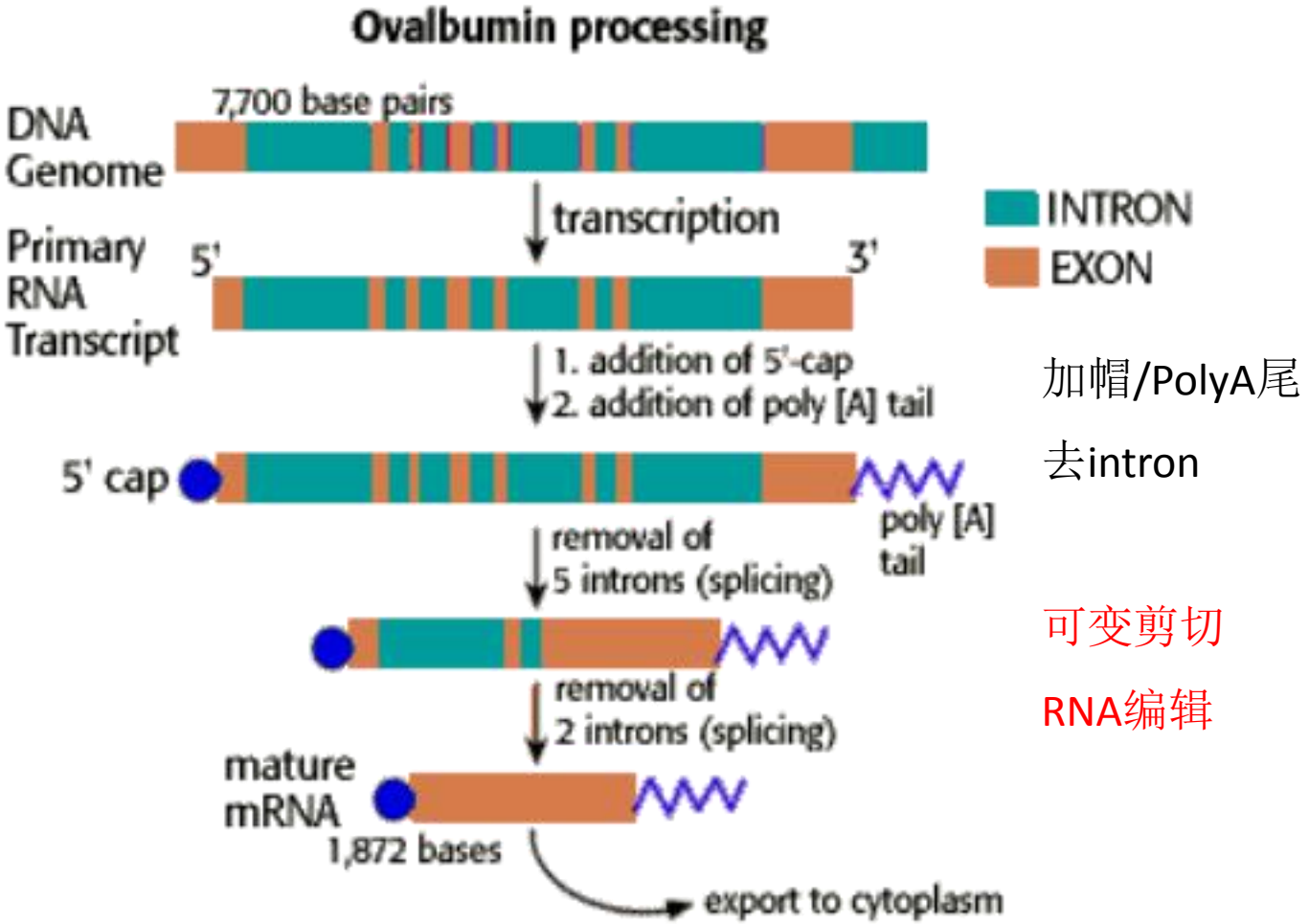
tRNA

snRNA
siRNA/tasiRNA
lncRNA
……



rRNA

mRNA成熟过程



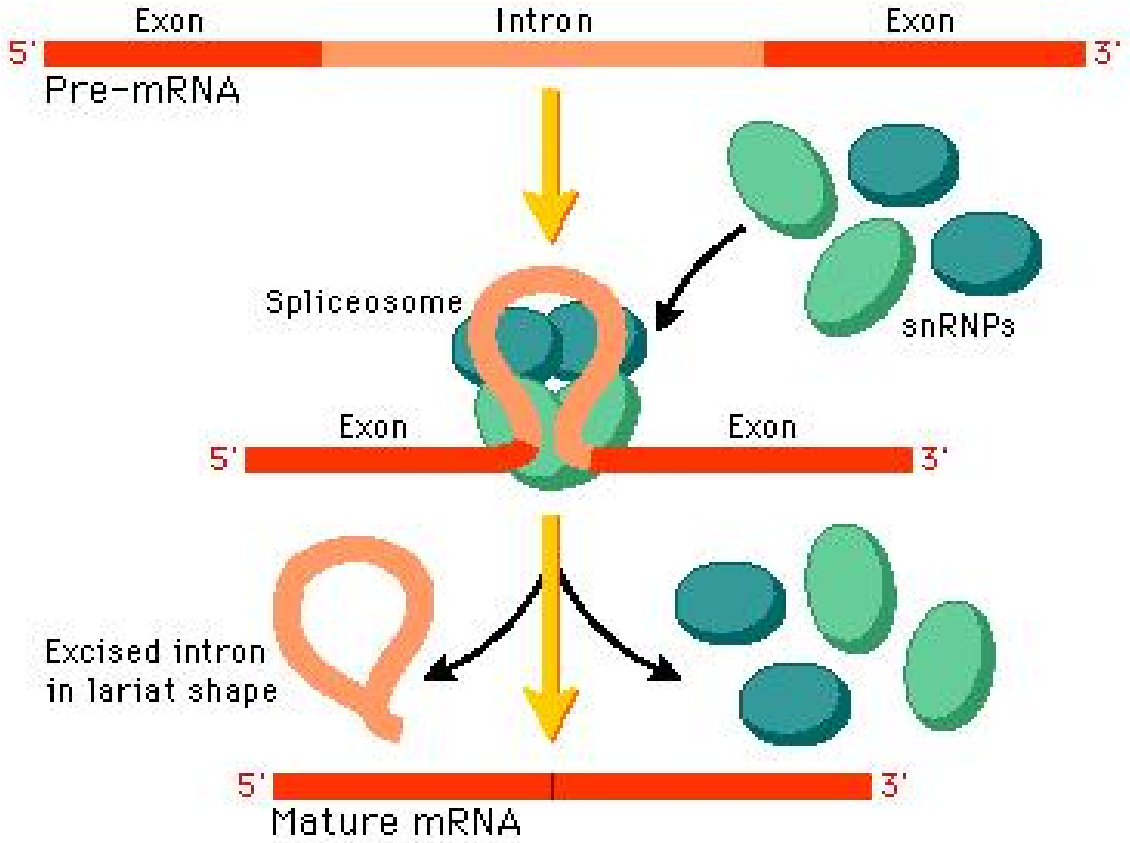
加帽/PolyA尾

去intron

可变剪切

RNA编辑

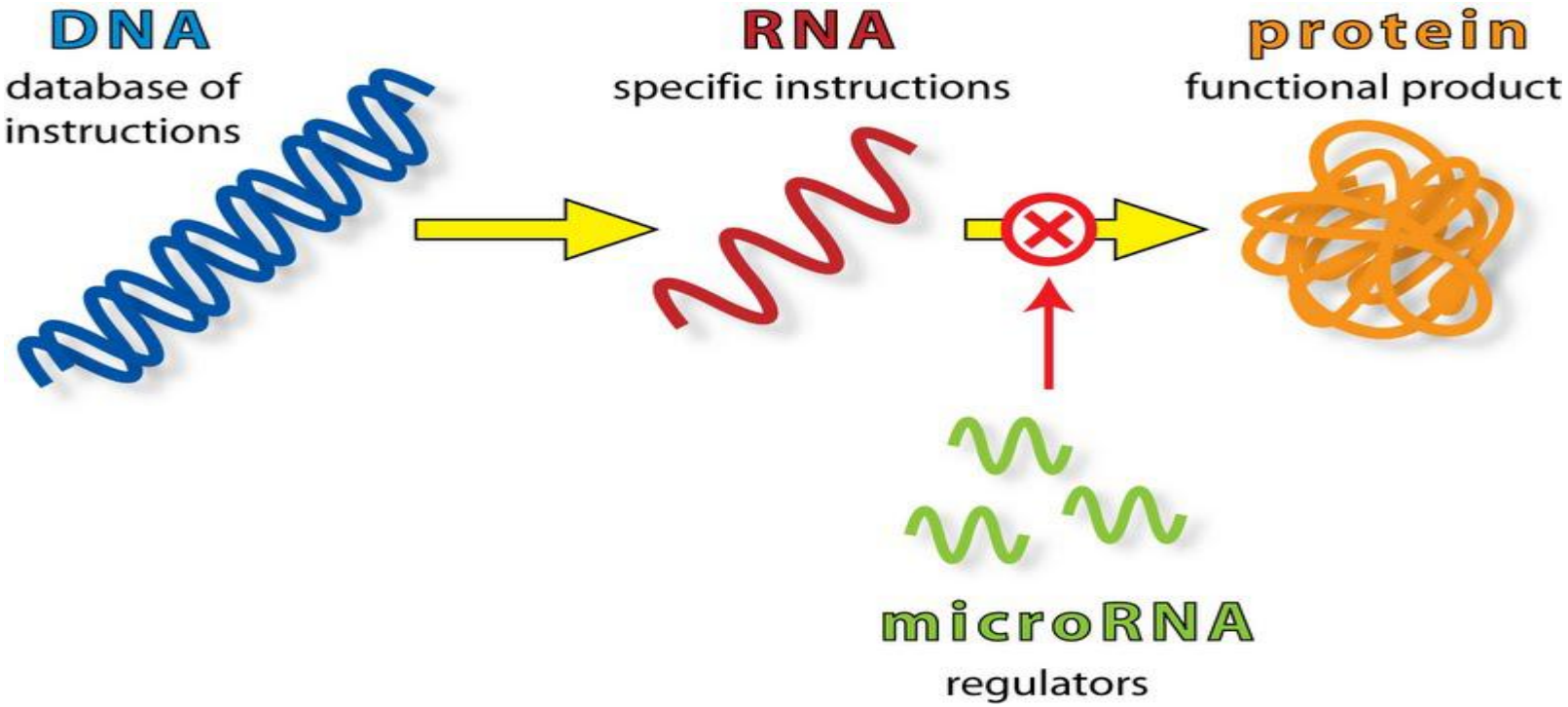
Small Nuclear RNAs (snRNAs)



SnRNA 参与了mRNA内含子剪切过程

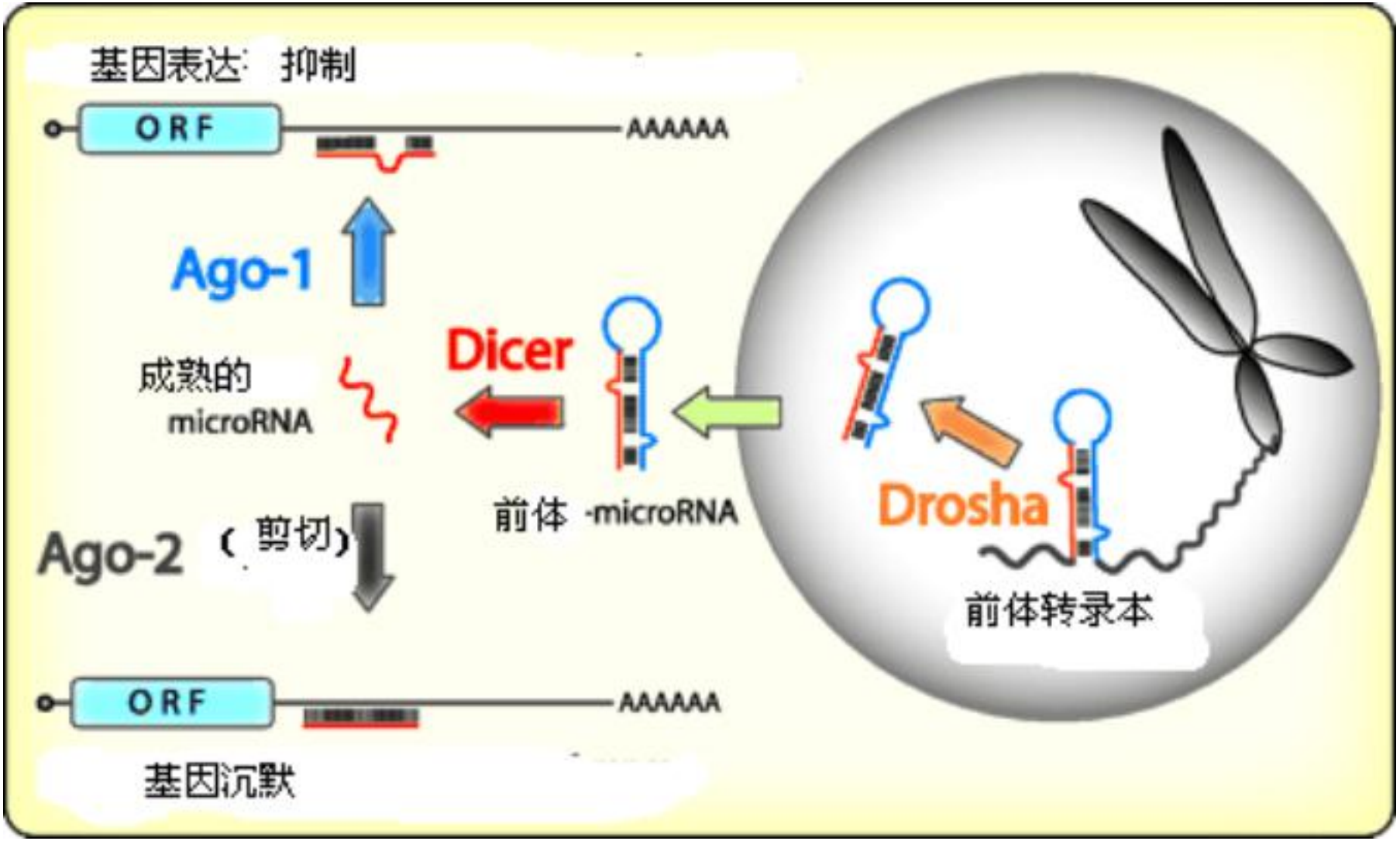
tRNA和rRNA参与mRNA翻译成蛋白质的过程

Micro RNAs (miRNAs)



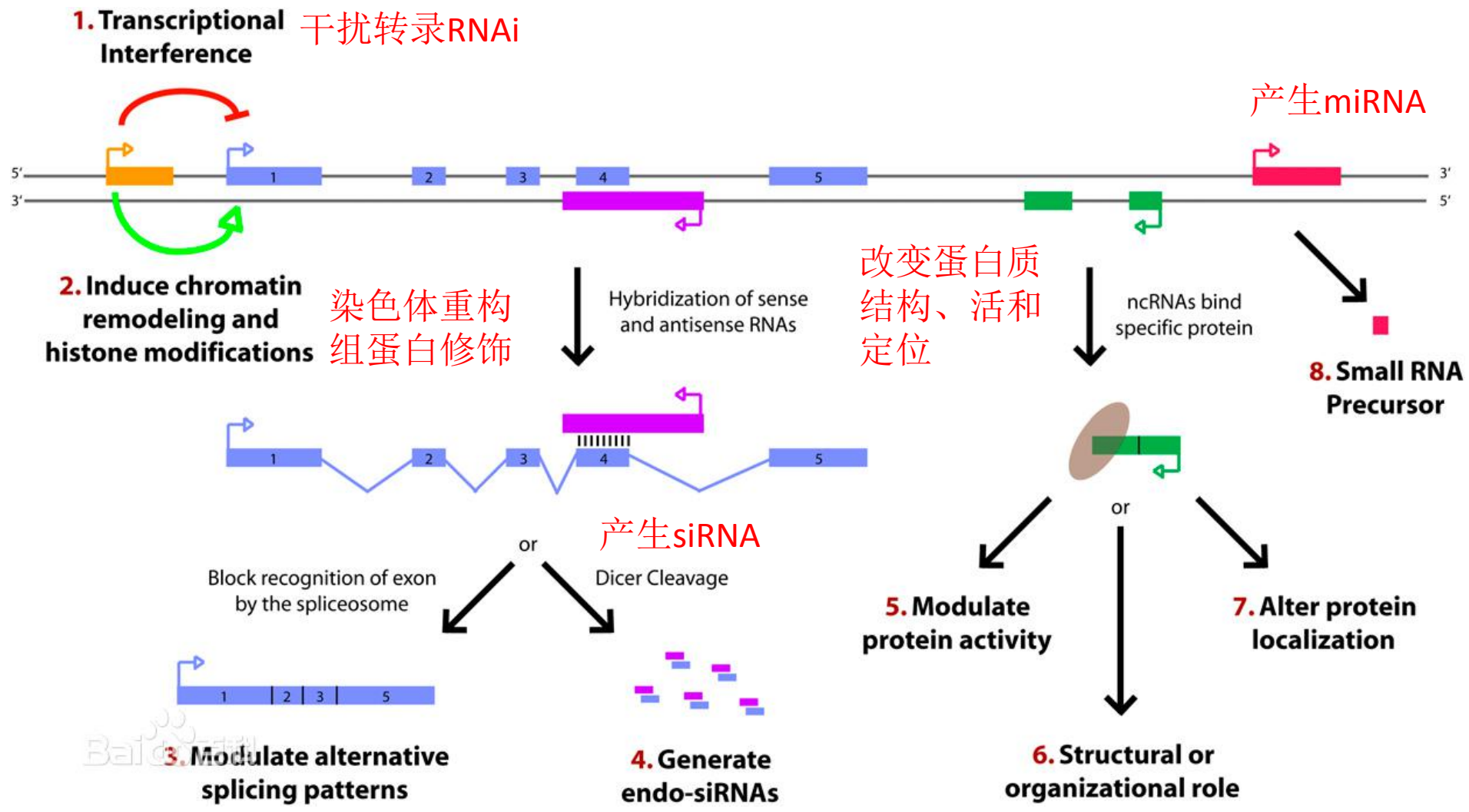
21-24bp,沉默或剪切mRNA导致降解

miRNA的产生以及功能



与siRNA有什么区别？外源/内源双链RNA剪切而来

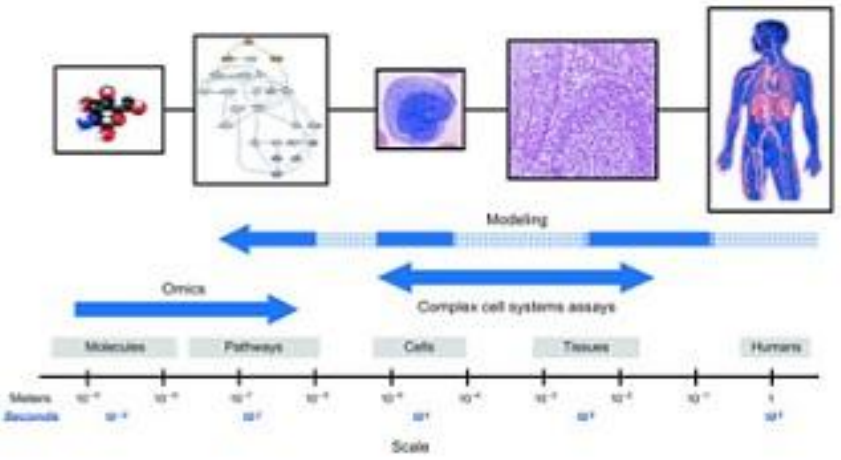
lncRNA （长非编码RNA）功能



生命在于运动：活动的RNA

存活不到“一秒”的人类个体有多大数据？

Data of a “one-second-life” Human Individual, How Big?



Picture by Kunisada Utagawa, “Rules of a Dietary Life”, ca. 1850.

个体仅在静态下：
$3X10^{10}$（G级别）的基因组；
10^{14}（T级别）细胞；
10^{15}（P级别）肠道菌群，而这些菌群
拥有人类100倍的基因集；

12

基因表达具有时空特异性：组织特异性/环境/时序

RNA的复杂程度超乎想象
如何解密？

RNA-seq助力揭开RNA世界神秘面纱

3、RNA-seq原理与应用

RNA-Seq(RNA sequencing)， 即RNA测序又称转录组测序，就是把mRNA，small RNA和 lncRNA）全部或者其中一些用高通量测序技术进行测序分析的技术

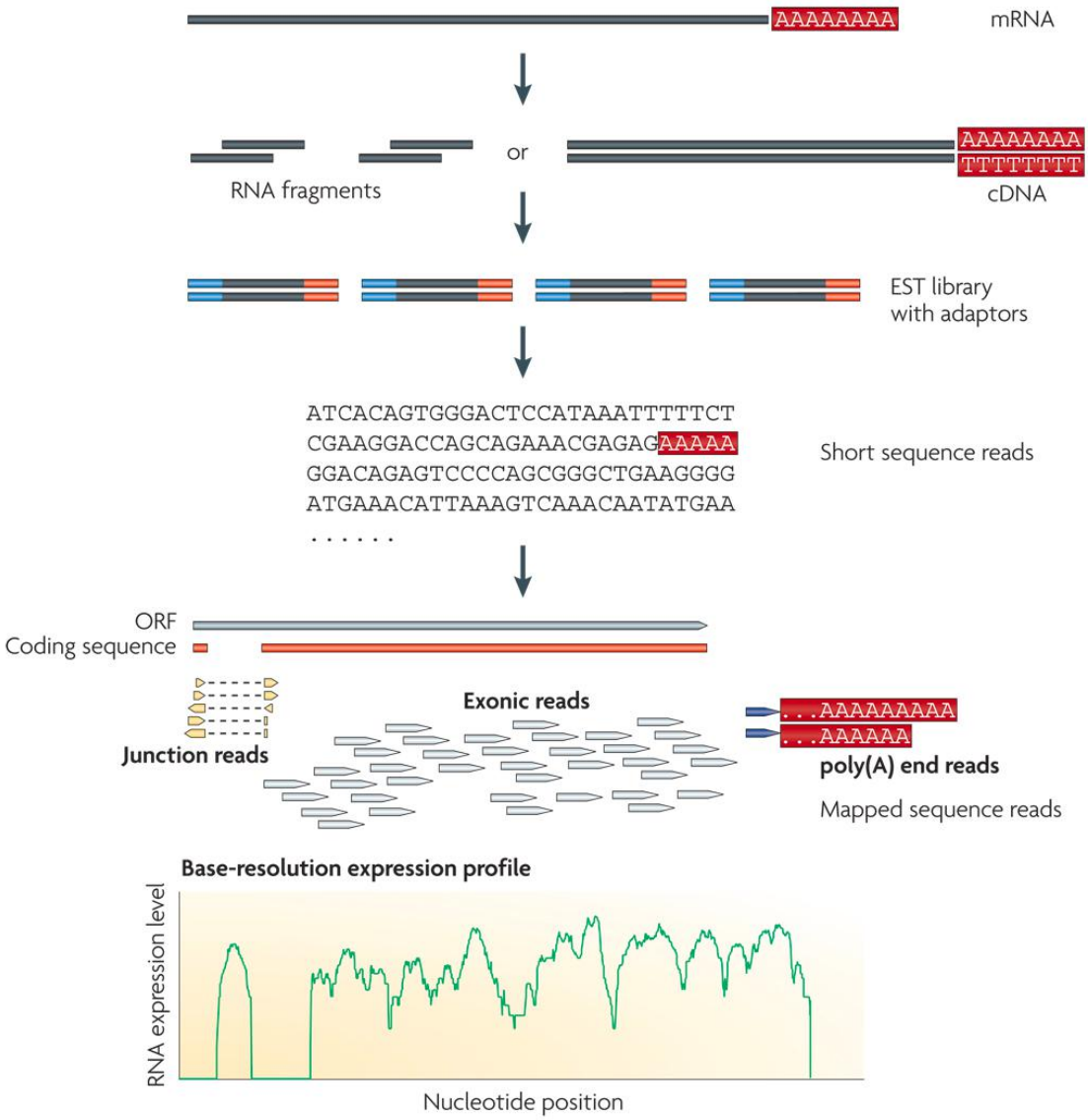
转录组是指特定组织或细胞在某一发育阶段或功能状态下转录出来的所有RNA的总和，主要包括mRNA和非编码RNA(non-coding RNA，ncRNA)。

转录本结构/可变剪切/RNA编辑/miRNA与目标基因

转录组研究技术比较

技术	cDNA library-seq	Microarray	RNA-Seq
原理	一代测序	核酸杂交	高通量测序
分辨率	单碱基	几十到一百碱基	单碱基
通量	低	高	高
是否依赖基因组信息	否	是	否
背景噪音	低	高	低
成本/基因	非常高	低	低
应用			
同时分析所有表达基因	否	是	是
基因表达检测范围	无	±100倍	±10,000倍
基因结构研究	是	否	是

RNA-seq具有荧光定量PCR的表达量检测灵敏性，芯片的高通量，还可以转录本测序，且价格便宜



RNA-seq流程

RNA提取

cDNA

建库-加接头

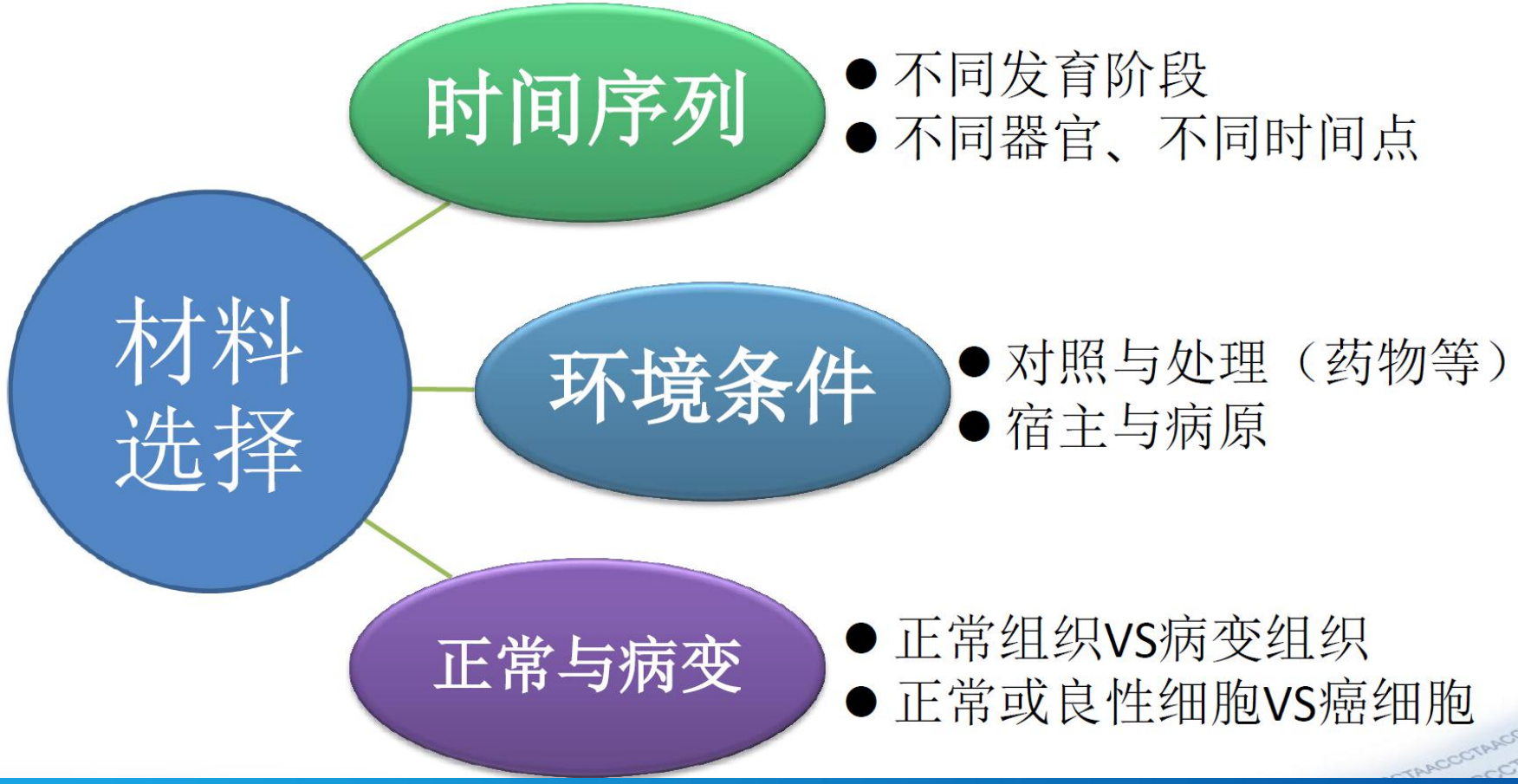
测序

分析：mapping/

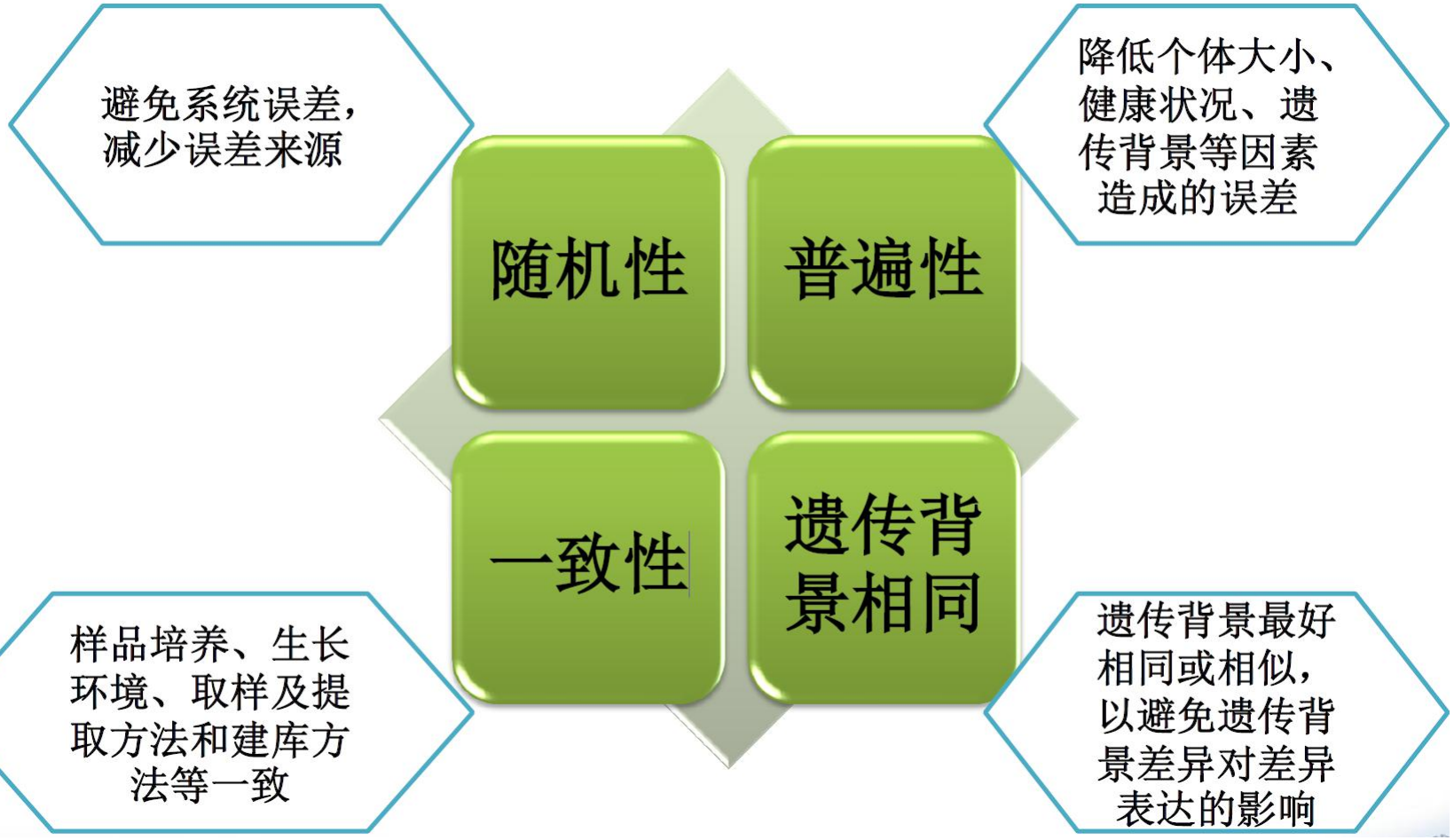
转录本/表达量

样本间比较

RNA-seq实验方案设计



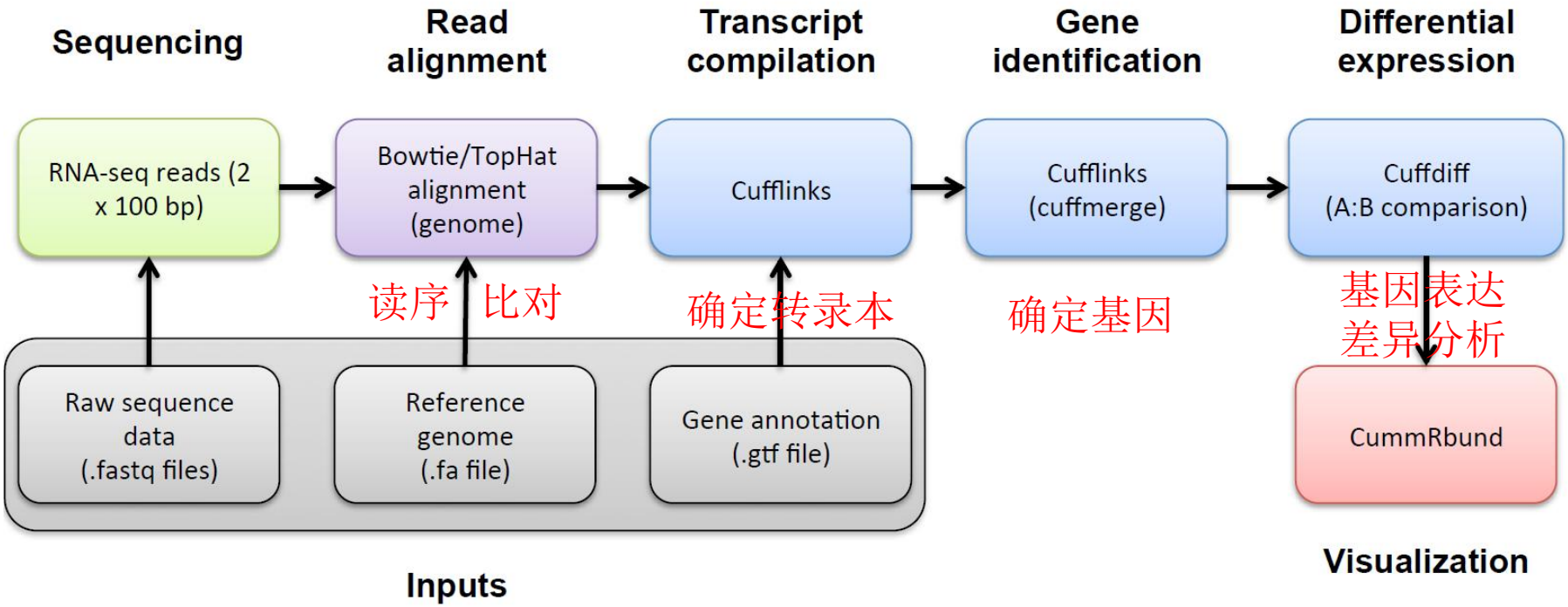
RNA-seq取样关键点



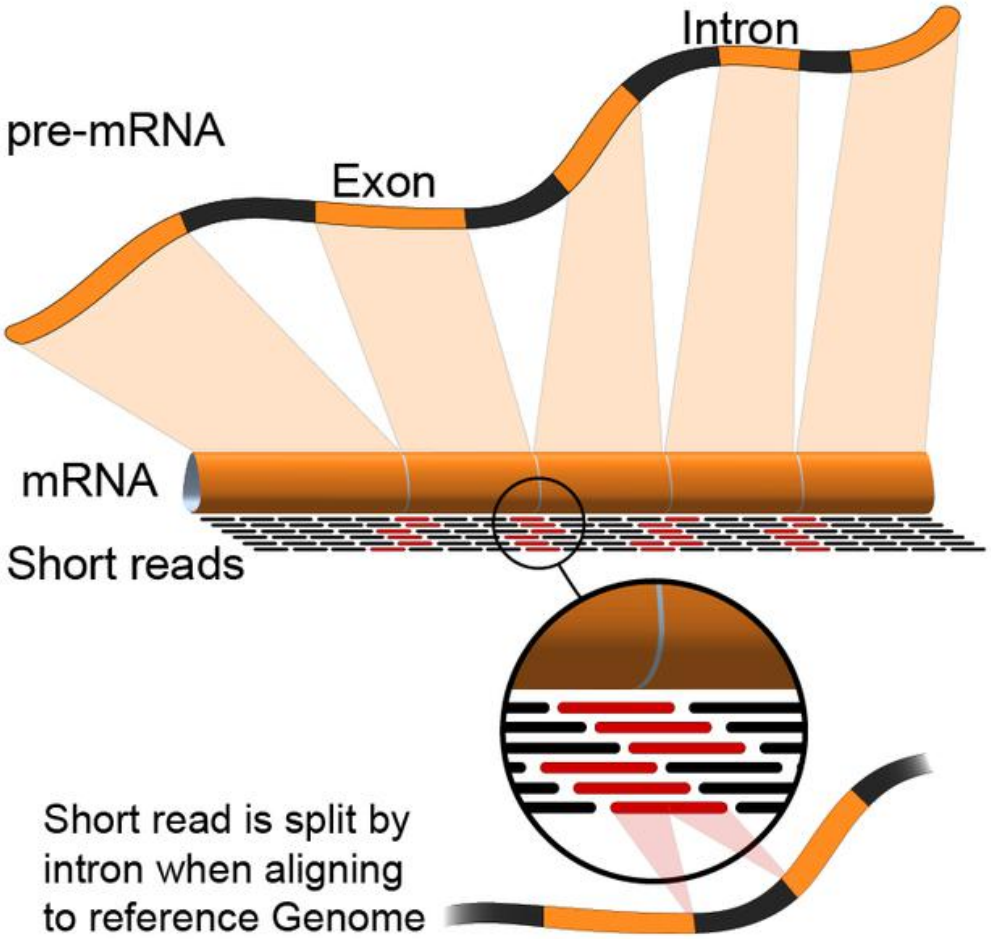
RNA-seq研究目标与文库选择

普通转录组文库	• 适用真核生物转录组和表达谱 • 应用广泛，技术成熟
链特异性文库	• 适用真核生物转录组和表达谱 • 区分方向信息，信息分析优势
DSN均一化文库	• 适用表达低丰度基因检测 • 适用于lncRNA、全转录本建库
全转录本文库	• 适用转录组、表达谱、lncRNA • 保留除rRNA外全部RNA信息

RNA-seq分析流程Tophat/Cufflinks/



RNA-seq比对算法：Spliced Alignment



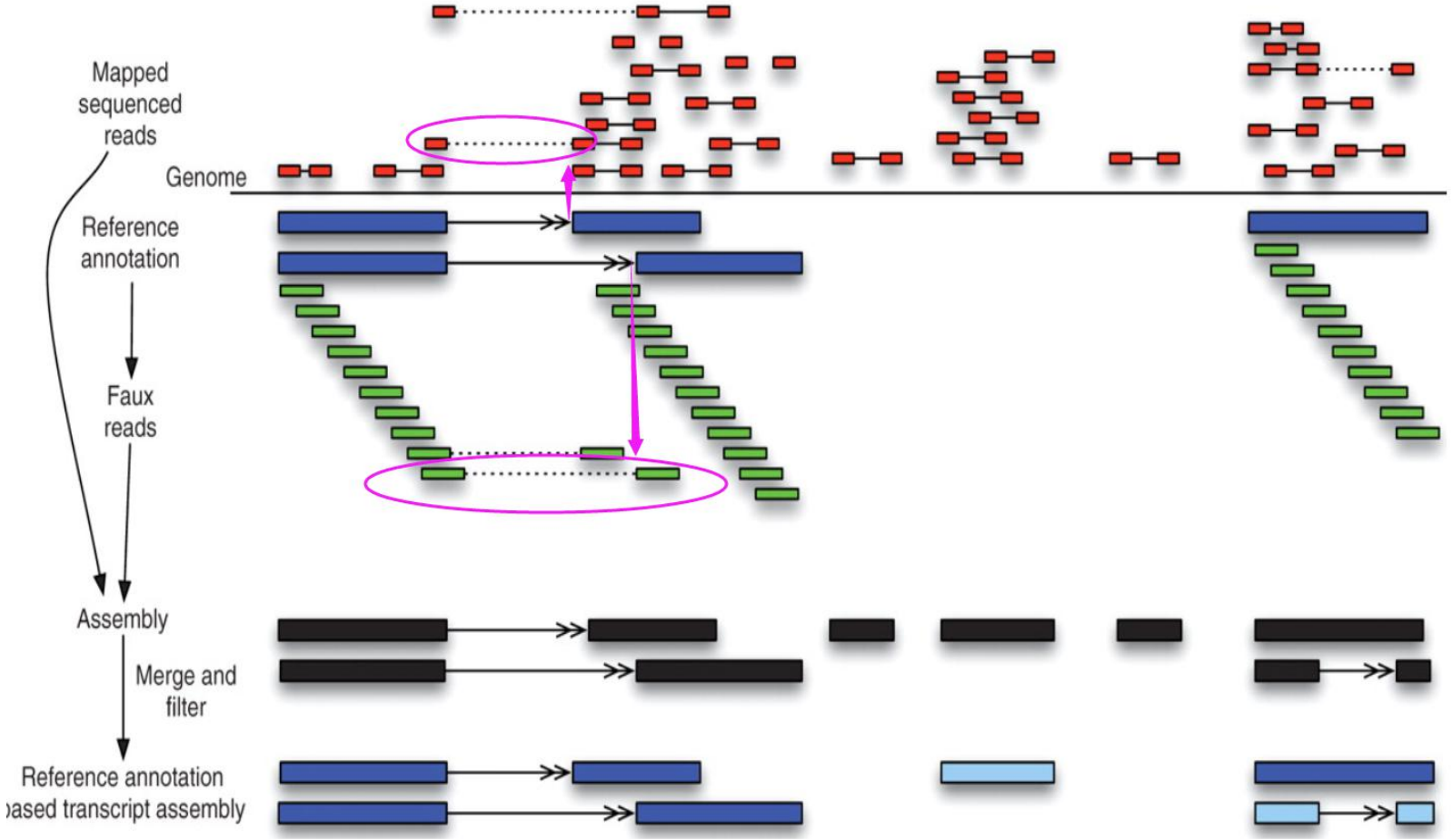
读序不连续比对，确定内含子

比对算法挑战
2千万读序比对时间

blastn：1秒/1序列，
200天

Bowtie：1.5小时
BWA：1小时

RNA-seq可变剪切分析



可变剪切由跨内含子读序确定

基因表达量RPKM值

RPKM:Reads Per Kilobase per Million of mapped reads

基因每Kb长度没百万读序比对到读序数据量

$$\text{RPKM for transcript } t = 10^6 \times 10^3 \times \frac{X_t}{l_t N}$$

PRKM值排除了基因大小与测序深度的影响，理论上可以直接拿来做基因表达量比较

人类基因总数可能是永远解不开的迷?

已报道的人类蛋白质基因总数的版本:

1) Celara: 27 894 HGR: 29 304 (Esemble)(2000)

Celara与HGR的注释基因有7000个不同, 相同的为20000左右, 加上不同的注释约34 000个.

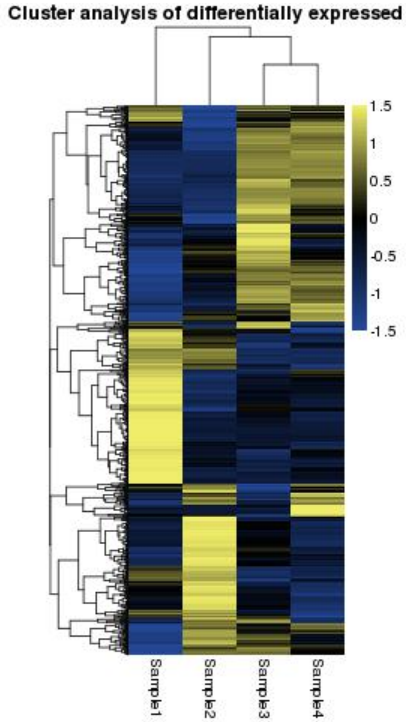
2) Esemble注释: 21 561 (http://www.ensembl.org, 2008).

3) EBI(European Bioinformatics Institute): 27 462 (2003, nature 423:576)

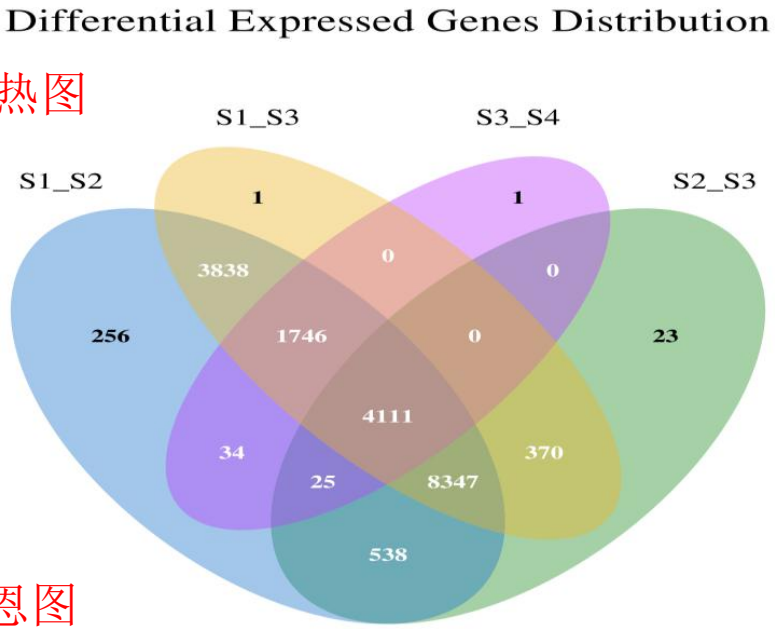
4) Genscan: 65 452

基因的组织特异/时序/环境表达的特征，无法穷尽有多少表达基因/转录本

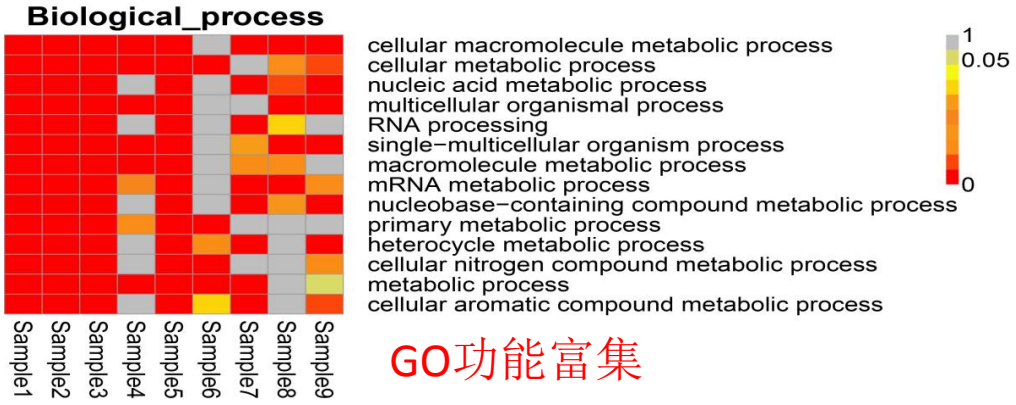
RNA-seq 差异表达分析



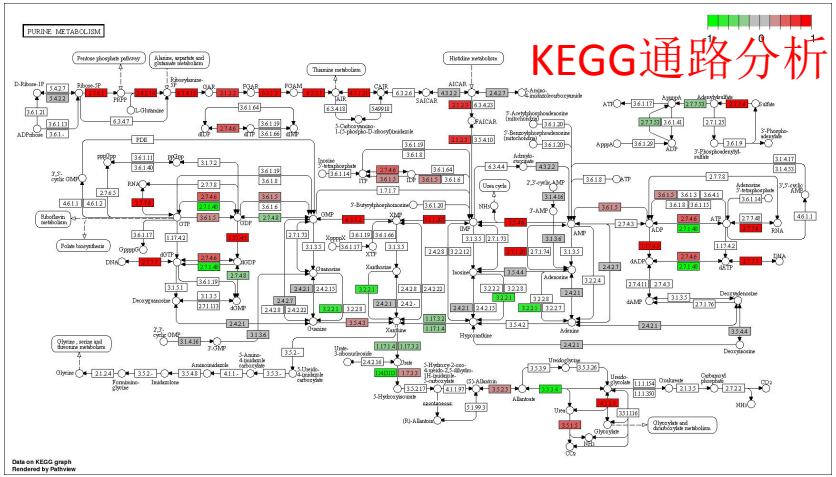
差异基因的聚类热图



差异基因的维恩图



GO功能富集



KEGG通路分析

miRNA的Cell封面故事：昼夜节律调控新机制

miR-279通过JAK/STAT信号调控了果蝇的运动行为节律。

生物钟，用以协调各种不同组织与器官的昼夜节律

miRNA测序与降解组联合测序分析

PNAS: 美科学家发现促进癌症的非编码RNA

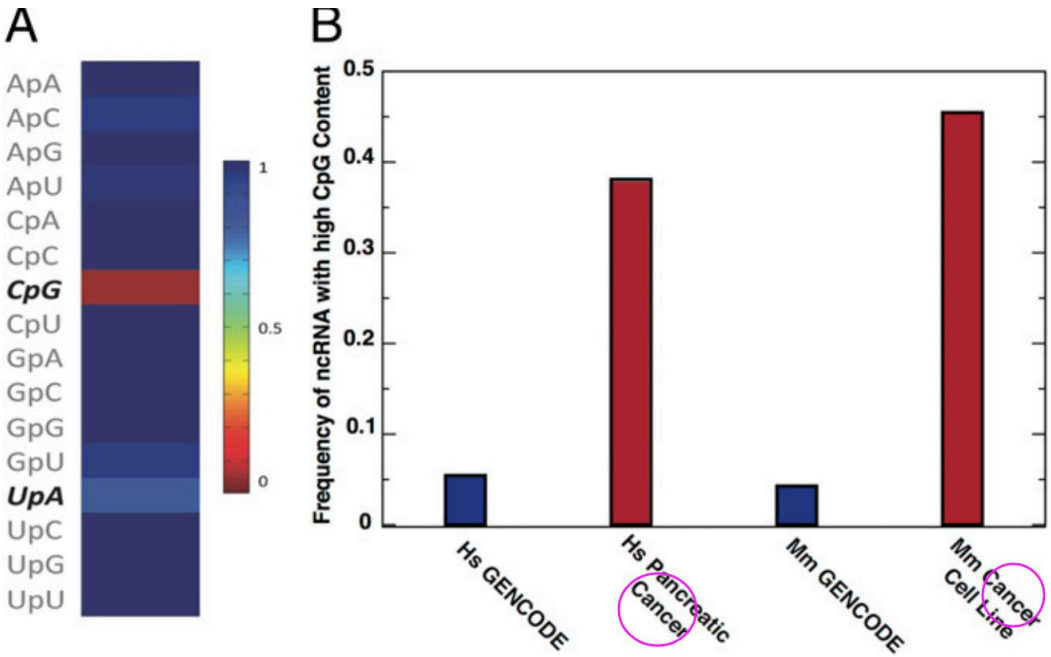
PNAS

CrossMark click for updates

Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells

Antoine Tanne[a], Luciana R. Muniz[a], Anna Puzio-Kuter[b], Katerina I. Leonova[c], Andrei V. Gudkov[c], David T. Ting[d], Rémi Monasson[e], Simona Cocco[f], Arnold J. Levine[b,g,1], Nina Bhardwaj[a,2], and Benjamin D. Greenbaum[a,g,h,1,2]

[a]Tisch Cancer Institute, Department of Medicine, Hematology, and Medical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [b]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903; [c]Roswell Park Cancer Institute, Buffalo, NY 14263; [d]Massachusetts General Hospital, Charlestown, MA 02129; [e]Laboratoire de Physique Théorique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [f]Laboratoire de Physique Statistique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [g]The Simons Center for Systems Biology, School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [h]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY 10029



RNA-seq展望

- 单细胞测序技术
- 第三代RNA测序技术
- 鉴定更多类型RNA，新功能
- 基因结构，调控网络，人类健康与疾病

Nature发表单细胞RNA-seq成果

nature International weekly journal of science

ARTICLE PREVIEW

view full access options

NATURE | ARTICLE Accelerated Article Preview

Single-cell RNA-seq identifies a PD-1^{hi} ILC progenitor and defines its developmental pathway

Yong Yu, Jason C.H. Tsang, Cui Wang, Simon Clare, Juexuan Wang, Xi Chen, Cordelia Brandt, Leanne Kane, Lia S. Campos, Liming Lu, Gabrielle T. Belz, Andrew N. J. McKenzie, Sarah A. Teichmann, Gordon Dougan & Pentao Liu

Affiliations | Corresponding author

Nature (2016) | doi:10.1038/nature20105

Received 23 April 2016 | Accepted 22 September 2016 | Published online 29 September 2016

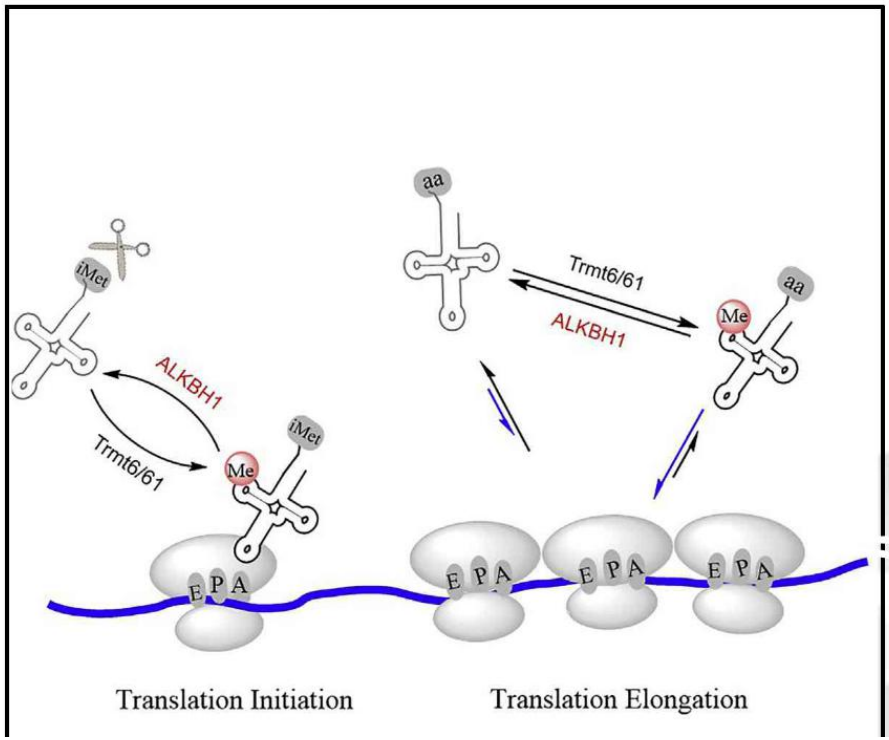
Cell里程碑成果：tRNA居然还有另外的作用

Article

Cell

ALKBH1-Mediated tRNA Demethylation Regulates Translation

Graphical Abstract



Authors

Fange Liu, Wesley Clark, Guanzheng Luo, ..., Arne Klungland, Tao Pan, Chuan He

Correspondence

taopan@uchicago.edu (T.P.), chuanhe@uchicago.edu (C.H.)

In Brief

Reversible tRNA methylation facilitates translation response to nutrient availability.

Liu et al., 2016, Cell *167*, 1–13
October 20, 2016 © 2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2016.09.038

CellPress