基因组学、转录组学
基因的结构分析-1

张百芳

主要内容

1. 基因组学
2. 转录组学
3. 基因调控序列的结构分析
4. 基因编码区的结构分析

基因组——一个细胞或生物体中，一套完整单倍体的遗传物质的总和。

- 不同生物基因组的大小和复杂程度各不相同；
- 基因组的功能：贮存和表达遗传信息。

不同生物基因组蕴含遗传信息量有差别。病毒<细菌<真核生物

- 基因组中不同的区域具有不同的功能：
  - 有些区域编码蛋白质的结构基因
  - 有些区域复制及转录的调控信号
  - 有些区域的功能尚不清楚

基因组学

基因组学是研究基因组的科学

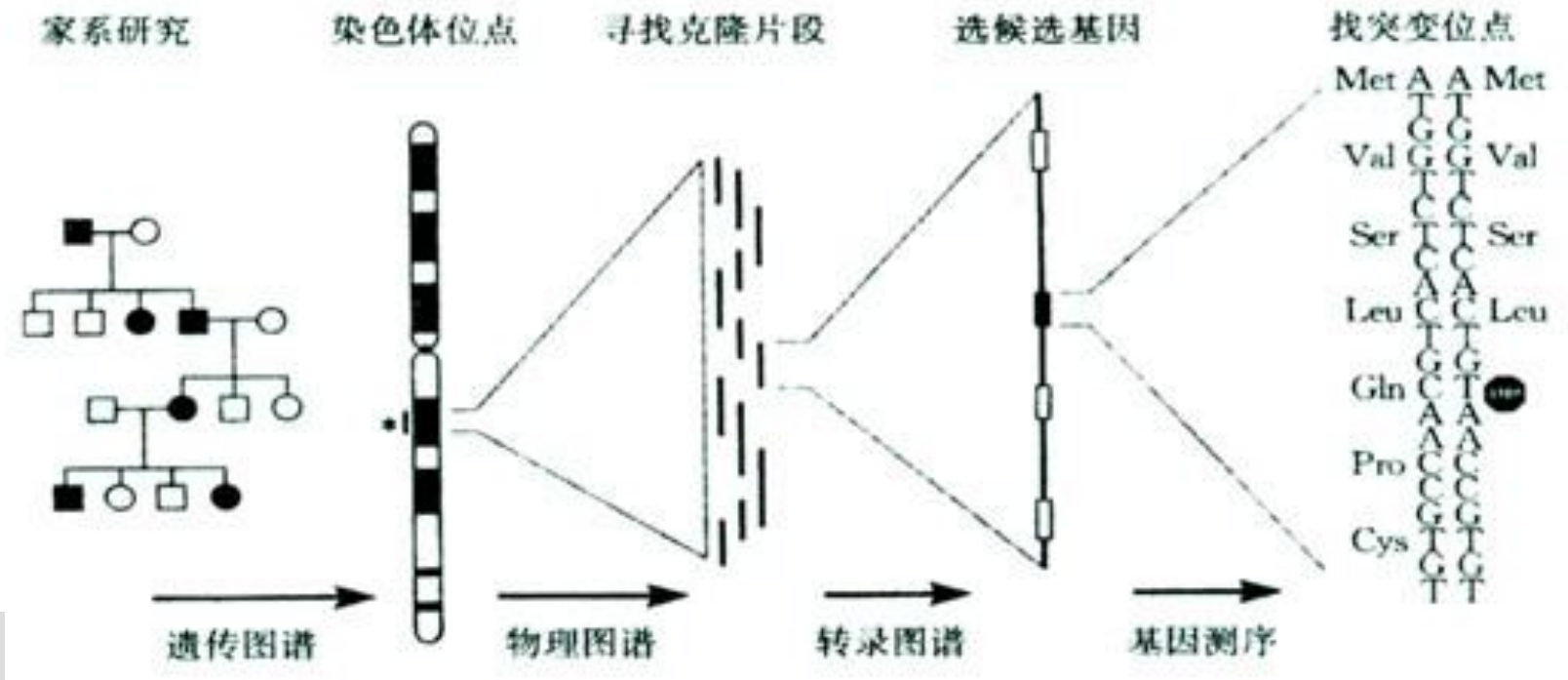
- 背景：1985年提出，1990年开始实施人类基因组计划(Human Genome Project, HGP)，产生基因组学。
- 概念：以分子生物学技术、计算机技术和信息网络技术为研究手段，以生物体内全部基因为研究对象，在全基因背景下和整体水平上探索生命活动的内在规律及其内外环境影响机制的科学。
- 目的：阐明基因组结构、结构与功能的关系以及基因与基因之间的相互作用。
- 分类：
  - 结构基因组学
  - 功能基因组学
  - 比较基因组学
    - 肿瘤基因组学
    - 植物基因组学
    - 药物基因组学
    - 环境基因组学

一、基因组学研究内容

（一）结构基因组学

以全基因组测序为目标的基因结构研究，
弄清基因组中全部基因的位置和结构，
为基因功能的研究奠定基础。

遗传图谱
物理图谱
转录图谱
序列图谱

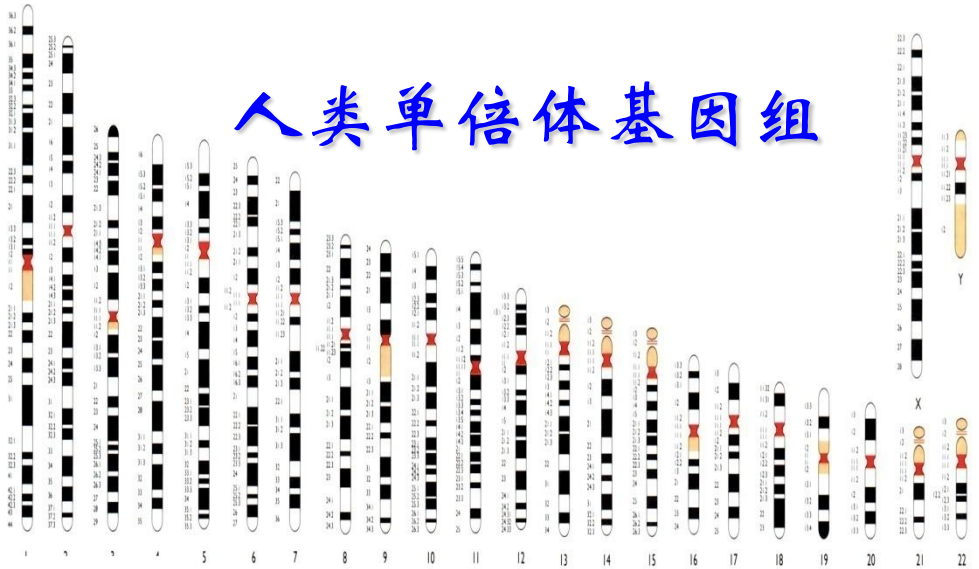


遗传图谱（genetic map）

——孟德尔的“新生”

指基因或DNA标记在染色体上的相对位置与遗传距离。是以DNA多态性遗传标记为界标，以遗传学距离（在减数分裂事件中两个位点之间进行交换、重组的百分率，1%的重组率称为1厘摩/cM）为图距的基因组图。

人类单倍体基因组



绘制遗传图谱的三个阶段

- 限制性片段长度多态性 (restriction fragment length polymorphism, RFLP)：

  较为粗略，分辨率2-5cM

- 短串联重复序列STR (short tandem repeat)：

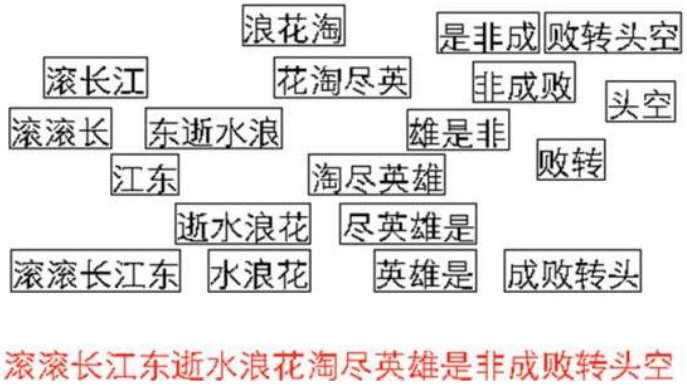
  分辨率1.6cM

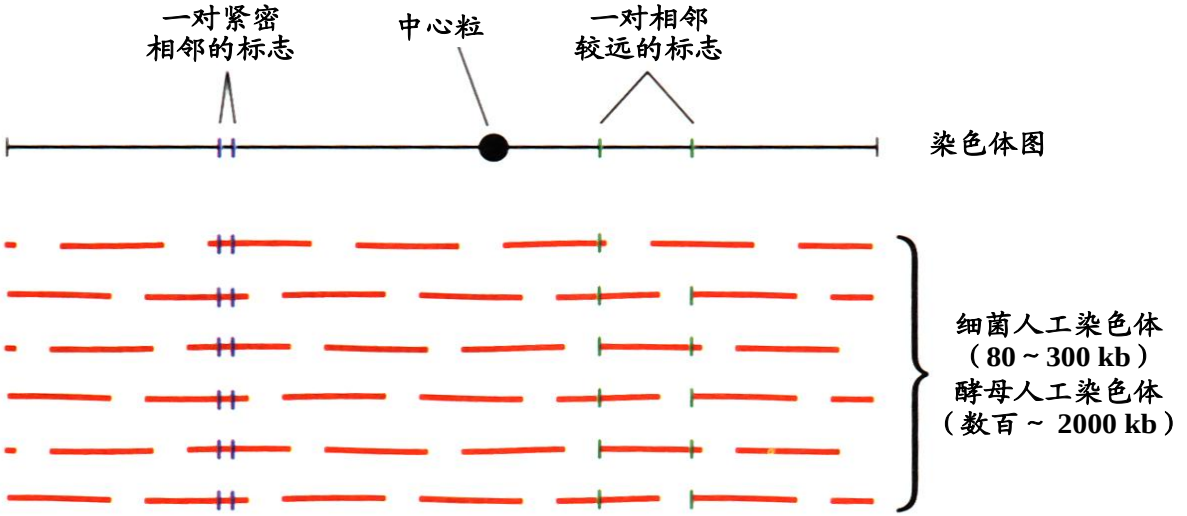
- 单核苷酸多态性 (single nucleotide polymorphism, SNP)

物理图谱（physical map） - *路标与路轨*

表示基因组中不同基因或DNA标记之间的实际距离。以**定位的序列标记位点STS**作为路标，以**DNA实际长度**即bp、kb、Mb（百万碱基对）为图距的基因组图谱。是进行DNA序列分析和基因组织结构研究的基础。

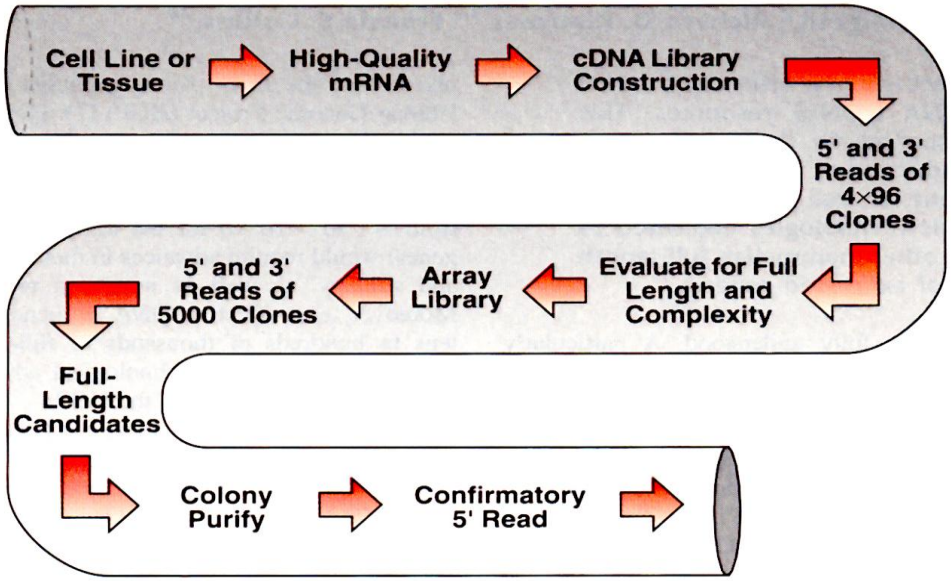
- 物理图谱首先是利用**限制性内切酶**将染色体切成片段，再根据重叠序列把片段连接成染色体，确定遗传标志之间的**物理距离**。
- 作图标志是已知序列的DNA片段，称为STS（sequence tagged site）。





❖转录图谱（表达图谱）——生命的乐谱

转录图谱是在识别基因组所包含的蛋白质编码序列的基础上绘制的结合有关基因序列、位置及表达模式等信息的图谱。以表达序列标签EST为位标，根据转录顺序的位置和距离绘制的图谱，它是染色体DNA某一区域内所有可转录序列的分布图（基因图谱）。



序列图谱(分子水平的物理图谱)——*重中之重*

以某一染色体上所含的全部碱基顺序绘制的图谱。是转录序列、调节序列和功能未知序列的总和。

DNA sequencing (DNA序列分析)是一个包括制备DNA片段化及碱基分析、DNA信息翻译的多阶段的过程。通过测序得到基因组的序列图谱。

2003.5.28—2003.6.2

冷泉港-人类基因组完成图发布会

（二）功能基因组学（后基因组学）

在基因组水平上阐明DNA序列的功能，从基因整体水平上对基因的活动规律进行系统研究。

1. 基因表达及调控
2. 人类基因信息的识别和鉴定
3. 基因功能信息的提取和鉴定
    1) 人类基因突变体的系统鉴定
    2) 基因表达图谱的绘制
    3) “基因改变----功能改变”的鉴定
    4) 蛋白质水平，修饰状态和相互作用的检测
4. 测序和基因多样性分析

（三）比较基因组学 comparative genomics

在基因组图谱和测序基础上，对已知的基因和基因组结构进行比较，以了解基因的功能、表达机理和物种进化的学科。

将人类基因组与模式生物基因组进行比较研究，根据同源性方法分析人类基因的功能、疾病分子机制，阐明物种进化关系和基因组内在结构。

疾病基因组学：比较基因组学的分支

（四）基因组学的研究技术和方法

结构基因组学

- 脉冲场凝胶电泳
- 毛细管电泳
- 基因芯片技术
- 全基因组随机测序

功能基因组学

- 基因转移技术
- 反义核苷酸技术
- 核酶技术
- 肽核酸(PNA)技术
- RNA干扰技术
- 基因敲除技术
- 基因表达系列分析(SAGE)
- cDNA芯片
- 反向遗传学
- 蛋白质组研究
- 生物信息学技术

二、基因组变异的生理病理意义

（一）基因组在进化过程中发生变异

基因组的序列变异：突变、插入、缺失、不同数目串联重复及SNP;
寡核苷酸微阵列分析(ROMA)检测基因组中“拷贝数多态性”

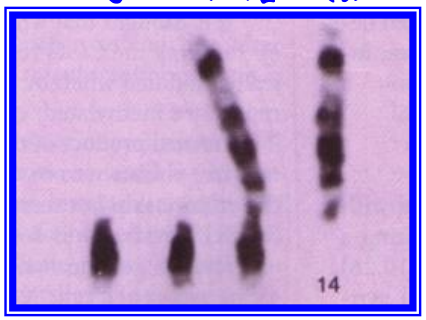
（二）染色体DNA变异可导致疾病发生

染色体数目的变异; 结构与排列的变异

（三）线粒体DNA突变可引起线粒体基因病

mtDNA无修复系统及His保护，突变频率高，如 Leber遗传性视神经病

家族性唐氏症



21染色体易位，接附在14号染色体

（四）易感基因与环境的相互作用

环境相关疾病易感基因的等位多态性。

（五）所有的疾病，都可以说是基因病

- 基因相关论
- 基因修饰论
- 变基因的表
- 基因的多态

单基因病

多基因疾病

线粒体基因

基因组与癌症

- 基因组某单个基因座上存在缺陷基因所引起的一类遗传病。（6000余种）红绿色盲、血友病等。
- 一种基因型可产生不同的表现型，或一种表现型可由若干基因型所影响，即基因型与表型非“一
- 突变基因在mtDNA上，其传递和表达完全不同于核基因突变引起的遗传病，而成为一组独特的遗传病。
- 遗传特点:母系遗传(突变的mtDNA通过卵子细胞质的线粒体传给子代)；数量特征：突变的mtDNA数量超过阈值时，会出现临床症状。突变mtDNA所占比例似与临床症状的表现程度相关。

基因组与癌症研究

- 癌症是最常见的基因病。
- 癌基因和抑癌基因
- 全部基因组序列对比
- 癌基因活化和抑癌基因失活：基因片段丢失，重排，碱基替换，小片段插入或缺失，扩增或甲基化。

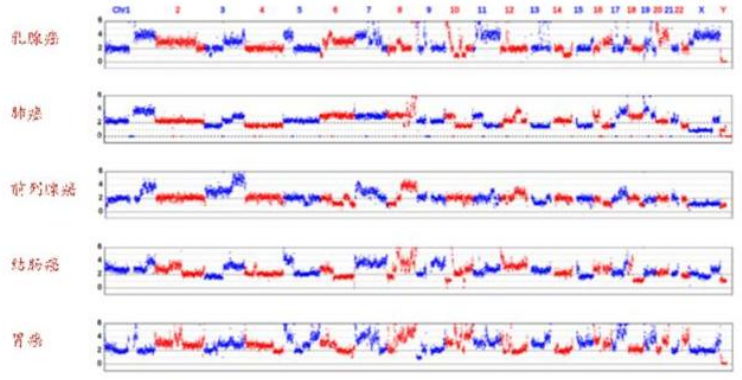
三、单细胞基因组学

- DNA是以单分子的形式存在于每个细胞中的。
- 在一个细胞中最常见的基因组改变包括点突变和基因拷贝数变化。这种变化是单分子的变化，所以是随机的，不同细胞是不同步的，不知道它什么时候发生，也不知道它在哪发生，因此每个细胞都拥有不同的基因组。这也使得单细胞和单分子水平上的检测成为必需。
- 单细胞基因组学是单分子技术与基因组学的交汇之处。在单细胞中测基因拷贝数以及单个点突变，现在不仅已成为可能，而且真正成为了一项日益重要的技术。

- 人类生殖细胞（精子、卵子）在分裂时发生随机重组，使得每个生殖细胞都不相同。
- 任何一个正常的男子都会有约5%的精子出现拷贝数不正常的现象。这种不正常是由于细胞分裂时染色体没有正常分裂。这种染色体不正常的精子会导致生殖障碍、流产、胚胎停育或者唐氏综合症等遗传疾病，尽管父母看起来完全健康，但就是有5%的出错几率。对男子而言，这5%的几率是不随年龄变化而变化的。
- 但对女士的卵子来讲，染色体不正常的几率在30岁之前是25%，此后很快随年龄的增长而上升，到40岁的时候是70%。这就导致发生生殖障碍的比率和流产的比率随年龄的增长而增加，生育成功率则随年龄的增长而递减。

- 癌症是由于基因组改变所引起的疾病，癌细胞中剧烈的基因组变化，使得原发肿瘤中的细胞之间存在高度不同。
- 针对癌症的很多重大课题都需要单细胞基因组学。首先是个体化治疗，即靶向治疗，通过个人的基因组测序，为预防、检测和治疗疾病提供个体化的解决方案。
- 癌症难以治愈和高死亡率的罪魁祸首是肿瘤的转移。其机理是癌症先出现在原发灶，然后通过血液循环扩散到身体的其他器官。然而，癌症病人血液中肿瘤细胞数量很少，一般只有几个，传统的研究手段往往基于大量细胞才能进行分析。单细胞测序技术可以用到循环肿瘤细胞的研究上。

- 循环肿瘤细胞的单碱基突变存在异质性——也就是说每个细胞都不一样，这对癌症检测意义不大。
- 正常体细胞的基因拷贝数是2，癌细胞要高于或低于2。实验发现同一个肺癌病人的8个循环肿瘤细胞表现出高度一致的拷贝数变异模式。更有趣的是，不同病人的同种癌的基因拷贝数图案也是一样的，这就展现了一个好的前景。我们可以通过分析循环肿瘤细胞拷贝数图案来推测癌症类型。乳腺癌、肺癌、前列腺癌、结肠癌和胃癌具有不同的基因拷贝图案。利用这些发现做不需活检的无创癌症检查，也许能够做到癌症的早发现。



转录组学

转录组与转录组学

➢转录组(transcriptome)

一种生物体或一个细胞在特定生理或病理状态下转录出的所有RNA。

➢转录组学(transcriptomics)(1997年提出转录组学)

以转录组为研究对象,在整体水平上研究基因转录及其调控规律,了解基因产生全部转录物的时空关系及其生物学意义。

狭义的转录组学通常特指针对mRNA的研究，是功能基因组学的重要分支，也是连接基因组结构和功能的桥梁和纽带。

转录组学研究内容

- 对特定细胞转录与转录后加工的研究
- 对转录物编制目录
- 绘制动态转录物图形
- 转录物的网络式调节

转录组学基本研究方法

◆基于测序的转录组学方法

表达序列标签（EST, expressed sequence tags）

全长cDNA文库

基因表达系列分析(serial analysis of gene expression, SAGE)

◆基于杂交的转录组学方法

Northern blot

原位杂交 (hybridization in site)

RNA酶保护试验 (RNase protection assay,RPA)

RT-PCR 及实时定量PCR（real-time PCR）

cDNA microarray （cDNA Chip）

◆基因表达聚类分析

- 表达序列标签（EST）是从已构建好的cDNA文库中随机抽取克隆，从5’末端到3’末端对插入的cDNA片段进行一轮单向自动测序，所获得的约60-500bp的一段cDNA序列。
- 主要应用

  在同一物种中搜寻基因家族的新成员

  在不同物种间搜寻功能相同的基因

  搜寻已知基因的不同剪切模式

  大规模分析基因表达水平

EST技术流程：

一、构建cDNA文库

- 非标准化的cDNA文库的构建。（可用于基因表达量的分析）
- 经标准化或扣除杂交处理的cDNA文库。（富集表达丰度较低的基因）
- Oligo d(T) cDNA文库。

（非翻译区由于不含有编码序列，与编码区保守序列相比所受到的选择压力比较小，因而其多态性程度比较高，便于多态性位点的选择以用于遗传图谱的构建。）

- 随机引物cDNA文库。

（所获得的EST在基因功能的鉴定时具有更多的信息含量，并且在构建EST数据库时更有优势，同时有利于利用EST数据库聚类完整的基因和阅读框的寻找，便于利用更敏感的蛋白质比较来寻找同源基因。）

技术流程：

二、序列测定及数据分析

EST软件平台

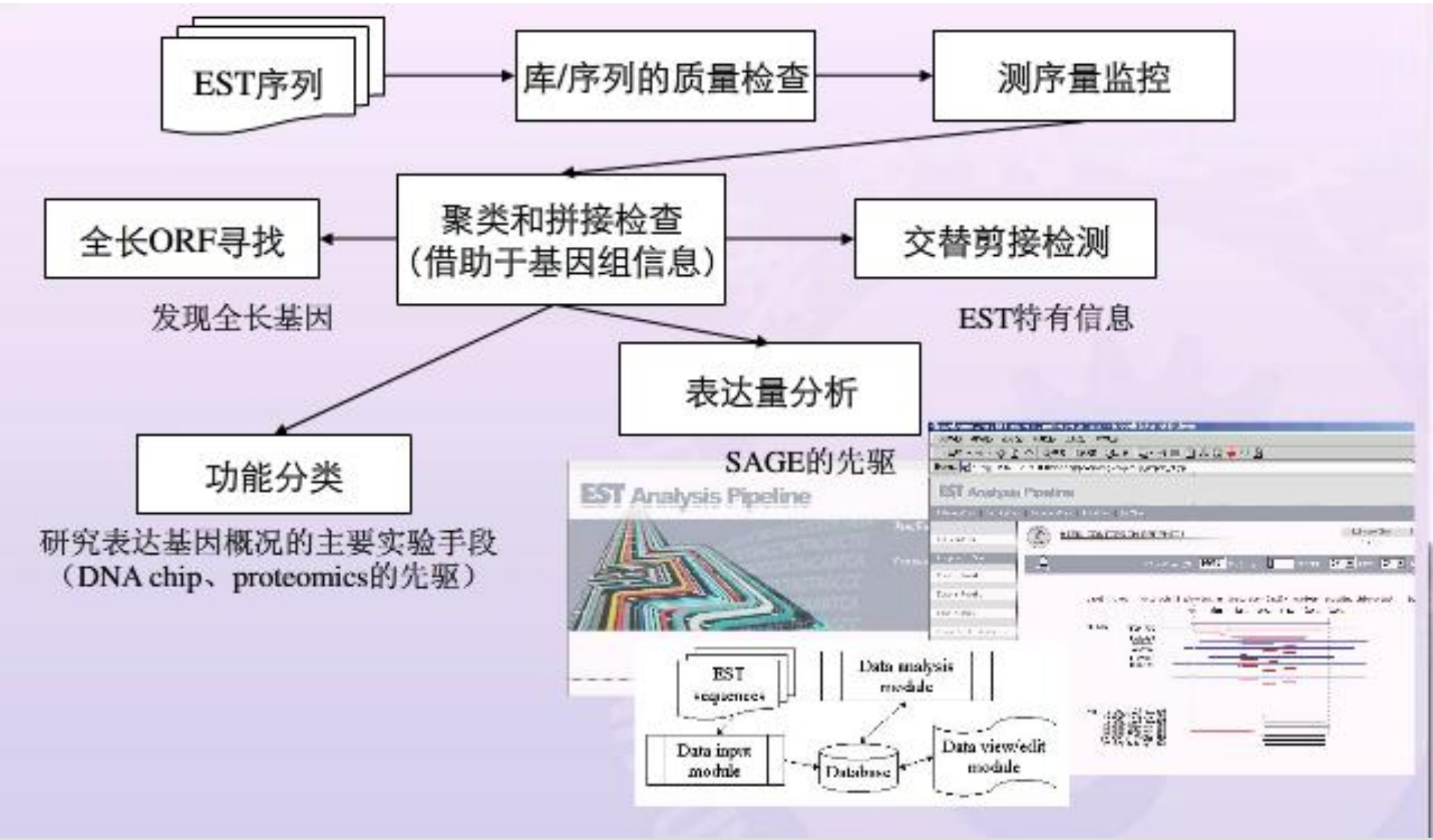

EST序列
库/序列的质量检查
测序量监控
聚类和拼接检查
（借助于基因组信息）
全长ORF寻找
发现全长基因
交替剪接检测
EST特有信息
表达量分析
SAGE的先驱
功能分类
研究表达基因概况的主要实验手段
（DNA chip、proteomics的先驱）
EST Analysis Pipeline
EST sequences
Data analysis module
Data input module
Database
Data view/edit module

表1：家猪脂肪组织的已知基因功能分类

	基因数目 Gene number	所占克隆数 Clone number	所占百分比（%） Percentage（%）
细胞分裂 Cell division	108	191	7.22
细胞信号传导 Cell signal/communication	280	682	14.33
细胞结构和运动 Cell structure and mobility	163	549	10.3
细胞机体防御 Cell and organism defense	372	573	12.1
基因和蛋白表达 Gene and protein expression	336	731	15.43
代谢 Metabolism	485	1135	23.96
未分类 Unclassified	708	789	16.66
总计（Total）	2458	4737	100

表2：猪脂肪组织与猪胚胎胸腺组织和猪甲状腺组织表达谱的比较

	胚胎胸腺（%） Embryonic thymus	脂肪（%） Fat tissure	甲状腺（%） Hypothyroid
细胞分裂 Cell division	5.4	7.22	3.88
细胞结构和运动 Cell structure and mobility	4.84	10.3	9.11
细胞机体防御 Cell and organ defense	6	12.1	10.2
基因和蛋白表达 Gene and protein expression	37.06	15.43	19.6
细胞信号传导 Cell signal /communication	7.88	14.33	14.8
代谢 Metabolism	8.12	23.96	18.1
未分类 Unclassified	30.72	16.66	19.8

参考文献：1、猪脂肪组织表达序列标签（ESTs）大规模测序及分析
邓亚军等，遗传学报，Vol.31, NO.11, 2004
2、两种家猪心脏组织基因表达谱的分析
曾燕舞等，遗传学报，Vol.31, No.6, 2004

基因表达系列分析(serial analysis of gene expression,SAGE)

cDNA的3’特定位置9-11bp序列可代表相应的转录本，称为SAGE标签。分离所有转录本中的SAGE标签，再串联插入到克隆载体中进行测序，从而接近完整地获得基因组的表达信息。

SAGE既可显示该标签所代表的基因在特定组织或细胞是否表达，又可根据各SAGE标签出现的频率来确定所代表基因的表达丰度。

基因表达聚类分析

- 转录组学方法尤其是DNA微阵列的应用导致基因表达数据爆炸性增长。如何对这些数据进行分析，从中提取有意义的生物学信息，已成为转录组学的研究热点和技术瓶颈。
- 聚类分析技术能将待处理的对象分配到相应的聚类中，使得同一聚类中的对象差别较小，不同聚类之间的对象差别较大。
- 聚类分析技术在转录组学研究中，非常适合大批量分析基因群的功能。

RNA组与RNA组学

- **RNA组(RNome)**

  一种生物体或一个细胞或组织中的全部RNA分子.

- **RNA组学(RNomics)**

  以RNA组为研究对象,研究细胞内所有RNA分子的结构和功能及其在不同生理条件下的动态变化规律的科学.

2000年底提出**RNA组学新概念**：研究全部非mRNA的小RNA(snmRNA)，在特定状态下表达差异、功能及其与蛋白质的相互作用。

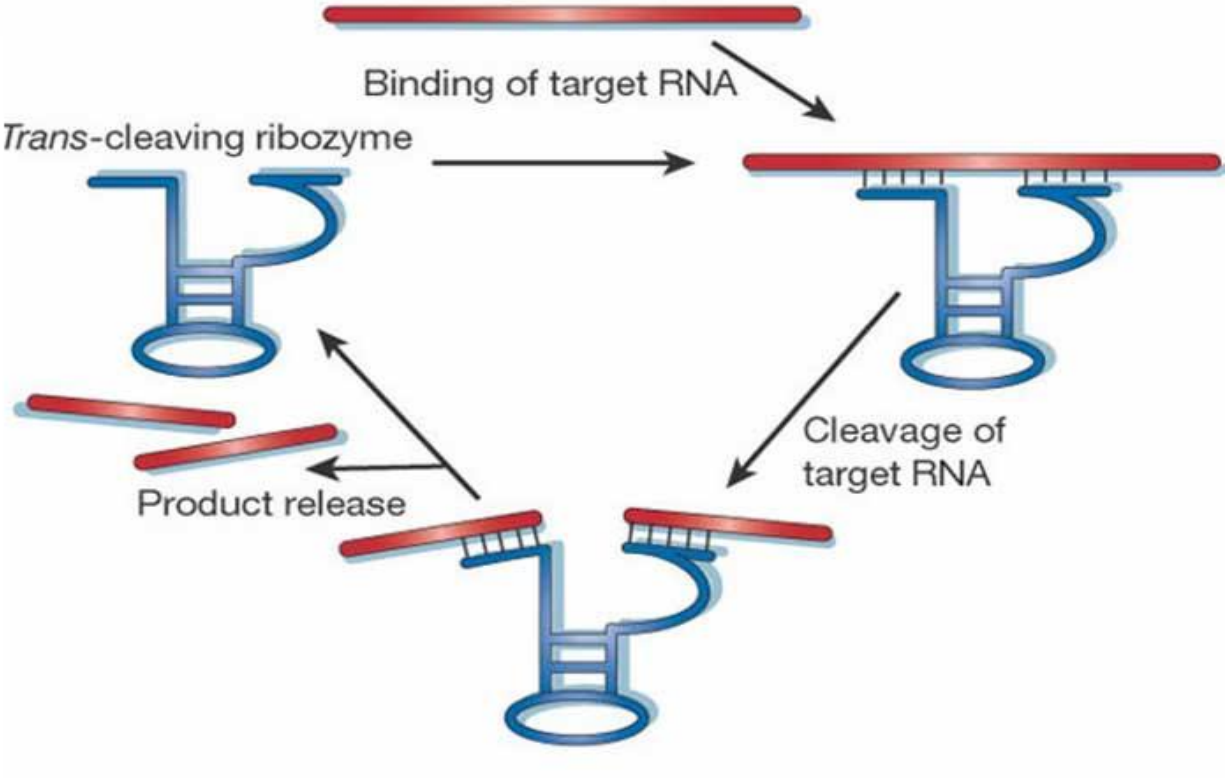
RNA组学研究的主要内容

研究snmRNAs的重点

- 核酶（ribozyme）
- 反义RNA（antisense RNA）
- RNA干扰 （RNA interference；RNAi）
- 微小RNA (MicroRNA；miRNA)

(一) 利用核酶抑制基因表达

- 基因治疗
- 杀肿瘤细胞
- 防病毒复制
- 研究基因功能



Applications of trans-cleaving ribozymes for gene inhibition
(Sullenger BA, Gilboa E. Emerging clinical applications of RNA. Nature. 2002)

（二）利用反义RNA抑制基因表达

1. 反义RNA（antisence RNA）是一类能与特异mRNA 顺序互补配对的小RNA分子。能阻断mRNA翻译为蛋白质。除调节翻译过程外，还能选择性关闭基因、阻止RNA合成。
2. 对内、外源基因均有调节作用。
3. 可利用克隆表达获得长链反义RNA或直接合成反义寡脱氧核苷酸进行反义技术（Antisence technology）

(三) RNA干扰(RNAi)

RNAi是指小的外源性双链RNA与mRNA形成双链RNA，导致mRNA降解，特异性地抑制靶基因的转录后表达的现象。

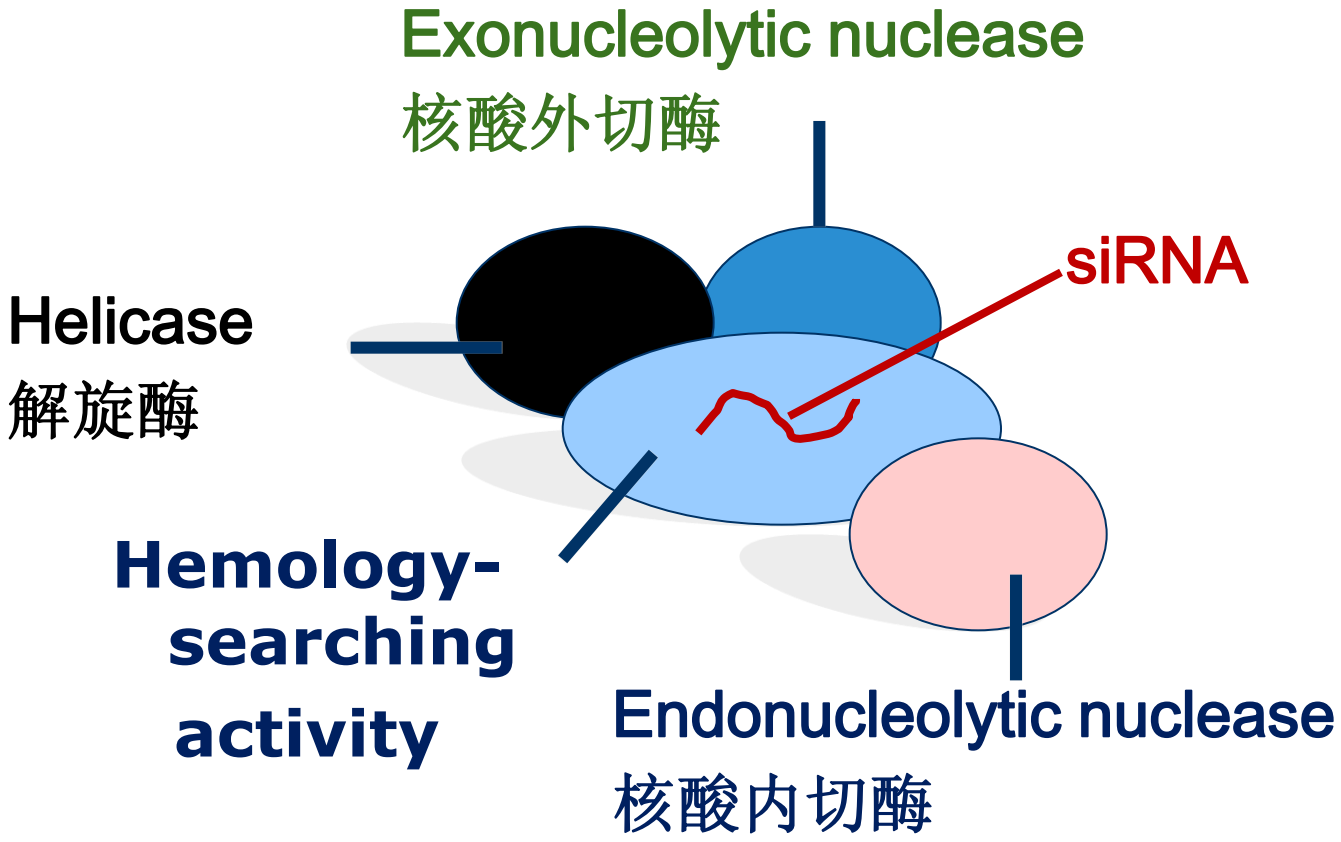
引起靶标mRNA的降解

抑制靶标mRNA的翻译

引起靶标启动子的转录沉默

RNA诱导的沉默复合物

(RNA-inducing silencing complex, RISC)



❖siRNA介导的转录后基因沉默

siRNA与RNA诱导的沉默复合物(RISC)结合，将RISC激活进而使双链siRNA解链，其中的一条链将去识别与其序列互补的特异性mRNA，然后RISC对特异性mRNA进行酶切，阻断相应蛋白的翻译，抑制相关基因的表达。

因为RNA诱导的RISC可以循环使用，能够继续酶切其他的靶mRNA，所以siRNA对靶mRNA的抑制作用可以被放大，特异性、高效性抑制基因表达。

(四) miRNA

- 在真核细胞中广泛存在，可以稳定存在于体液中，包括唾液、尿液、母乳和血液
- 没有开放阅读框，是一类非编码RNA
- 普遍长度约为20～23 nt，在3′端有1～2 nt的变化，成熟的miRNA 5’端有磷酸基团，3’端有羟基
- 序列有高度保守性，表达水平具有时空特异性

- 生物信息学预测发现，每个miRNA都有众多的靶基因，而每个基因的mRNA又有可能受到多个miRNA的调控，由此构成了复杂的调控网络。
- 在哺乳动物基因组中，30%以上基因的mRNA都受到miRNA的调节。
- 此外，一些miRNA隶属于同一个家族，它们具有相同的种子区域——与目标mRNA结合的关键部位，即从5’端起2～7位具有相同的序列。

miRNA与siRNA的区别

- miRNA是内源性的；siRNA是外源性或内源性的
- miRNA合成起始于胞核，再被转运至胞质；siRNA一般不在胞核合成
- miRNA前体是不完整的发夹结构RNA；siRNA前体是完全互补的长双链RNA
- 成熟的miRNA是单链；成熟的siRNA是双链结构
- miRNA与靶mRNA并非严格互补； siRNA通常与靶mRNA完全互补
- miRNA参与正常机体生长发育调节；而siRNA通常不参与

miRNA 与siRNA使靶mRNA沉默机制