生物科学研究方法丛书

基因组学方法

中国生物技术发展中心
深圳华大基因研究院 编著

科学出版社

共回组学方法

生物科学研究方法丛书

基因组学方法

中国生物技术发展中心
深圳华大基因研究院 编著

主　　编　杨焕明
副 主 编　冯小黎
秘　　书　牛　力
编写人员　（按姓氏汉语拼音排序）

阿　叁　鲍　莉　陈　芳　陈　诠
程时锋　冯小黎　高　扬　耿春雨
胡学达　旷　苗　李　宁　李建文
李培培　栗东芳　林　哲　刘　慧
刘　晓　牛　力　潘胜凯　钱武斌
邱宏伟　田　埂　汪　杰　王　晶
王　莹　肖　斌　杨焕明　杨　琪
杨瑞珍　殷旭阳　余　玄　曾育章
赵山岑　赵饮虹　郑　杨

科学出版社
北　京

内 容 简 介

基因组学是从系统的角度研究生命体全部遗传信息的学科,自诞生以来只有20年的历程,但是发展非常迅猛,这与基因组学技术的飞速进步是分不开的。本书重点从方法学的角度阐述了基因组学的发展脉络和未来路线。首先介绍了基因组学的基本概念、历史发展。之后重点阐述了基因组学最根本的测序技术和生物信息分析技术的前沿进展,并着重讨论了几个对本领域研究具有重大意义的技术问题。接下来,本书综述了一些应用基因组前沿技术获得的科学研究成果,并进一步讨论了基因组学技术创新发展的策略和途径,特别是我国在这一学科的目标、方向和重点任务。

图书在版编目(CIP)数据

基因组学方法 / 中国生物技术发展中心,深圳华大基因研究院编著;杨焕明主编. —北京:科学出版社,2012.9
(生物科学研究方法丛书)
ISBN 978-7-03-035588-1

Ⅰ. 基… Ⅱ. ①中… ②深… ③杨… Ⅲ. 基因组-研究方法 Ⅳ. Q343.1-49

中国版本图书馆CIP数据核字(2012)第221343号

责任编辑:王　颖　邹梦娜　李国红 / 责任校对:包志虹
责任印制:肖　兴 / 封面设计:范璧合

科学出版社　出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
北京市安泰印刷厂印刷
科学出版社发行　各地新华书店经销
*
2012年9月第　一　版　　开本:787×1092　1/16
2012年9月第一次印刷　　印张:7
字数:161 000
定价:38.00元
(如有印装质量问题,我社负责调换)

前　言

——插上基因组学的翅膀

生命世界，纷繁多彩。是什么记录着生命的密码？是什么承担着遗传的使命？是什么指导着生命活动？这是古往今来困扰人们的问题。

19世纪，生物学家们终能一窥门径。第一是细胞学说的建立，生命有了统一的单位，但细胞的多样性又是源于什么呢？第二是孟德尔遗传学的创立，遗传因子作为独立的单位而代代相传，遗传因子能够自由组合，但这些遗传因子又是什么呢？第三是自然选择进化论的提出，物竞天择，适者生存，但这变异的来源和选择积累的载体又是什么呢？带着理论，更多的是带着疑问，生命科学走进了20世纪。

20世纪见证了生命科学的伟大进步。人们相继发现了DNA是遗传信息的载体，双螺旋是DNA的结构，遗传密码和中心法则是DNA指导生命活动的方式。但对于生命来讲，这些只是冰山一角，终于，生物学家决心完整地看待物种的遗传变异，从基因组的视角来研究生命，穷尽生命（首先是人类）全部的遗传和应对环境能力的可能。因为，只有基因组是所有生物体共有的部分且又能够解释它们为何如此不同，只有基因组蕴涵着生命最本质的奥秘。

提到基因组学，大家首先会想起测序。的确，测序是基因组学最基本的研究手段，没有序列，一切都无从谈起。所有基因组学研究都存在这样一个模式：生物学对象——序列——生物学问题。

从20世纪末起，测序技术实现了持续快速的进步，20年前获得某一单个基因的序列需要一个实验室一年的努力，而现在测定一个人的全基因组序列只需要在一台自动化测序仪上运行几天。其中不得不提的是毛细管自动测序仪的发明和广泛应用，它为“人类基因组计划”的顺利完成提供了最有力的技术保障，进而为基因组学的建立和成熟提供了重要支撑。

测序是基因组学的主要技术手段之一，但基因组学不仅是测序。测序前的样品准备和测序后的信息处理也一直在不断创新和发展中。一方面，基因组学研究领域的延伸要求我们研究的对象包含更多的处理方式和更多的生化层次；另一方面，规模日趋庞大的数据也要求我们在软件和硬件上都具备更强大的计算处理能力和更高的运算效率。现在计算机发展的速度基本符合摩尔定律，即每18个月翻一番；而测序通量的提高速度已经超过了这一数值。所以海量的数据处理能力日渐成为基因组学研究的瓶颈。迄今为止，我们对软件运行效率的持续改进以及硬件投入力度的增大成功地满足了基因组学不断发展的需求。生物学实验技术和信息技术相辅相成的创新使基因组学家追求的梦想正在实现：用基因组表述进化论，推演“生命大代数”，并将最终达到用数学语言描述、诠释和指导生命科学和生物产业的发展。

人们一般将“人类基因组计划”的启动视为基因组学的起点，幸运的是，中国这一次没有输在起跑线上，我国科学家参与了这一国际大科学项目，圆满完成了其中1%的工作。而

后，我国又绘制了水稻、家蚕、家鸡等有重要农业价值的物种的全基因组图谱，为国家人口健康和绿色农业工作提供了原创性的支持。新一代测序技术的诞生使基因组学的研究能力得到极大解放，在此基础上，我国科学家又率先发表了第一个亚洲人基因组图谱，使中国在个人基因组时代再次走在国际前沿。此后，陆续发表了家蚕基因组甲基化谱、熊猫基因组、黄瓜基因组、人类泛基因组等一系列具有重大国际影响力的科学论文，保持了其在国际同行业中的领先地位。

国家提出要把生物技术作为未来高技术产业迎头赶上的重点，而生物科学的内在逻辑决定了基因组学将成为其最核心和最前沿的部分，因此，支撑中国经济的发展，引领未来产业走向，基因组学家当仁不让。为了担负起这份使命，我们必须认真总结过去的成功和失败，找准未来技术进步的方向，坚持走自主创新之路。

本书是一本基因组学方法的入门读物，我们希望借本书的编写和出版为大专院校学生、科研人员提供参考，为政府各部门在基因组学科学、技术和产业发展的战略决策和实施规划制订上提供参考，并希望借此书引导人们去思考中国基因组学及其产业体系的创新发展。

我们相信，插上基因组学的翅膀，中国生命科学和生物产业将迎来更加灿烂的明天。

目　录

第一章

基因组学的发展历程

第一节　基因组学的研究内涵

一、双螺旋模型对遗传信息储藏方式的启示

人们都知道生物具有代代相传的特性，这种遗传的物质基础就是基因组(genome)，每一种生物的基因组都包含着相应的遗传信息，这些信息决定了它们的个体建立和生物学特征。绝大多数基因组，包括人的基因组，都是由脱氧核糖核酸(DNA)组成的，但是也有一些病毒基因组由核糖核酸(RNA)组成。DNA 和 RNA 是由核苷酸(nucleotide)单体组成的多聚分子。组成 DNA 的核苷酸都由三部分组成，1 个戊糖基、1 个含氮的碱基和 1 个磷酸基团。含氮碱基包含 4 种，分别为胞嘧啶(cytosine，C)、胸腺嘧啶(thymine，T)、腺嘌呤(adenine，A)和鸟嘌呤(guanine，G)(图 1.1)。

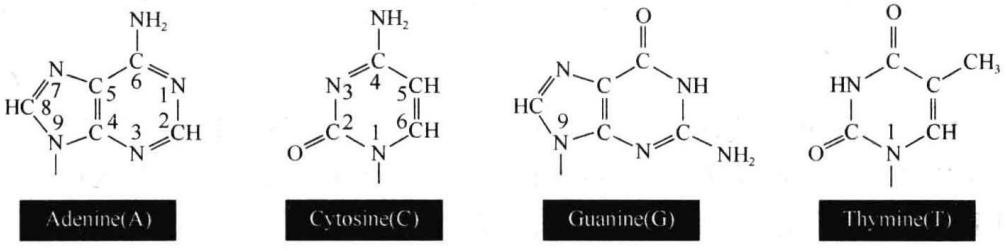

图 1.1　组成 DNA 基本结构单位的核苷酸的 4 种碱基

DNA 双螺旋结构被认为是 20 世纪最重要的科学发现之一。1953 年，Watson 和 Crick 在综合碱基比值和 X 射线衍射图谱的基础上，运用模型构建的方法推导出了 DNA 的双螺旋结构。根据 Watson 和 Crick 的双螺旋结构模型(图 1.2 左)，DNA 就像一个右手螺旋的楼梯，它由两条反向互补的多核苷酸单链相互缠绕而成。磷酸与核糖在外侧通过 3′，5′-磷酸二酯键相连接形成楼梯的骨架，而位于内侧的两条多聚核苷酸上的碱基通过氢键互补配对原则形成楼梯的阶梯。碱基配对形成的氢键和相邻碱基间疏水作用形成的碱基堆积力是维持这个楼梯稳定性的主要因素。碱基配对由于具有非常重要的生物学意义而显得更加重要。根据碱基互补配对原则(图 1.2 右)，即 A 仅与 T，G 仅与 C 互补配对，任何一个亲本 DNA 可以准确地产生出子代分子来，这也是细胞通用的 DNA 复制法则。此外，碱基间的配对在转录、翻译和调控等遗传信息流动过程中也至关重要。

二、动态变化中的染色体/染色质是遗传信息的载体

作为遗传物质的 DNA 在生物体中具有一定的组织形式，这就是染色质，包括 DNA 和蛋白质两部分。染色质(图 1.3)上的蛋白质包括组蛋白和非组蛋白。组蛋白可以看成染色

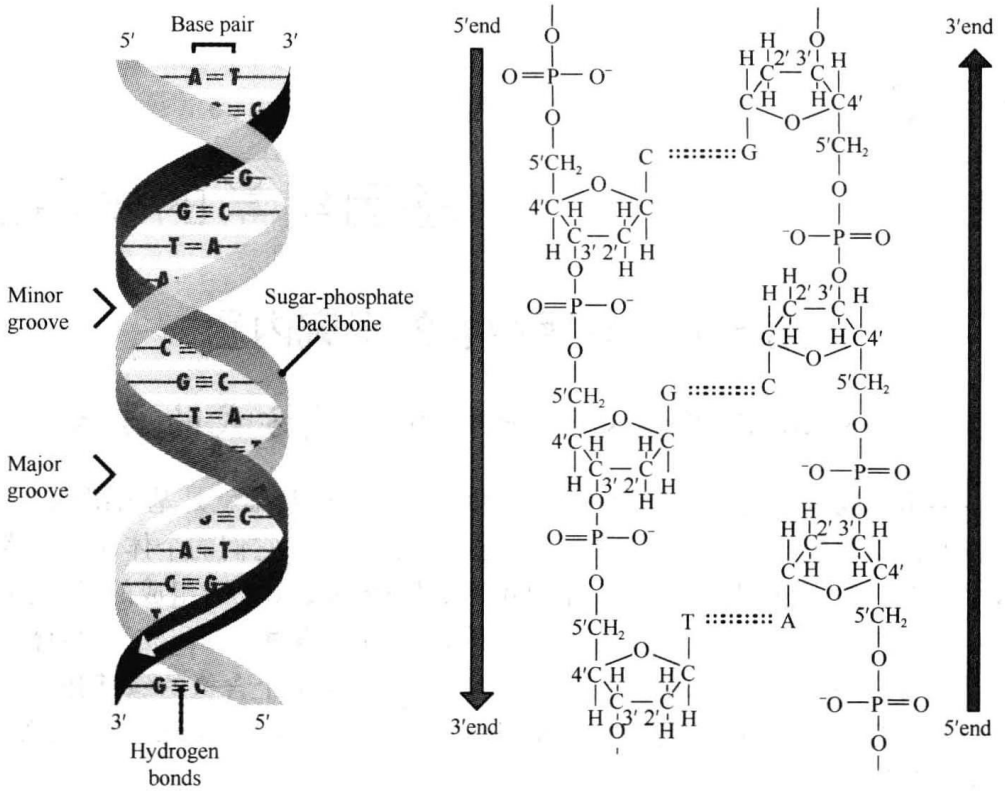

图 1.2　DNA 双螺旋结构的演示(左)和 DNA 双螺旋结构形成的化学基础:碱基互补配对(右)

质的结构蛋白。当细胞进行分裂时,DNA 紧密装配收缩成棒状,被称为染色体。细胞分裂完成后,染色体结构疏松呈线性排列,同时非组蛋白在一定的调控机制下与 DNA 结合,基因开始行使功能。染色质/染色体上 DNA 和蛋白质之间持续发生着互动,这对于基因的表达调控(即 DNA 的遗传信息)读取起着至关重要的作用。

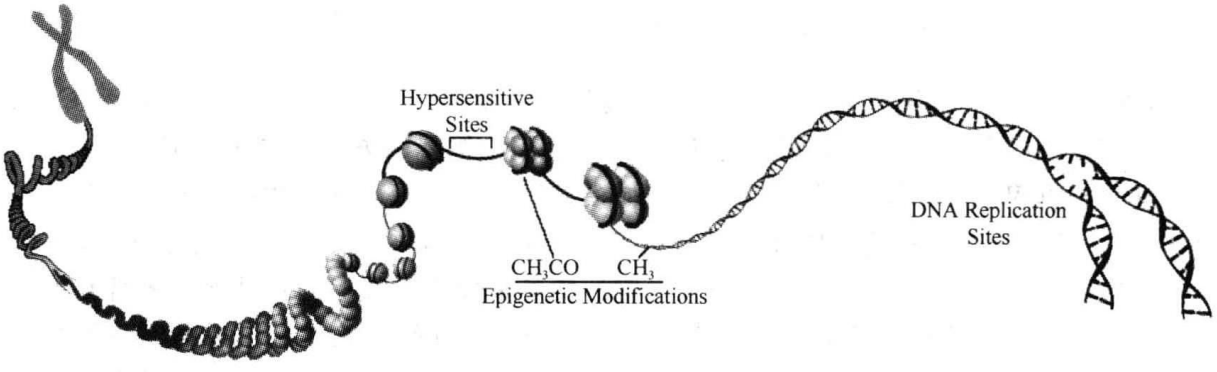

图 1.3　DNA 和组蛋白组装形成染色体

三、中心法则——遗传信息的使用

一个物种的遗传信息可以看做是一本书,书中的文字就是 DNA 序列。我们已经知道基因组(DNA)序列的每三个核苷酸对应一种氨基酸(终止密码子除外),氨基酸组成蛋白质,最后蛋白质行使生命的功能。总的来看,核苷酸序列的顺序决定了蛋白质的结构和功

能，其根本上又是由核苷酸和氨基酸的对应关系实现的，这种对应关系被称为遗传密码(图1.4)，每三种核苷酸被称为一个密码子。成千上万的密码子按照一定顺序排列在DNA上，还需要一些其他的功能序列来影响和决定其组织的方式以及何时何地被使用，这些功能序列和包含着密码子的编码序列一起决定了这个物种全部可能的生命活动。

第一个字母	第二个字母				第三个字母
	U	C	A	G	
U	苯丙氨酸	丝氨酸	酪氨酸	半胱氨酸	U
	苯丙氨酸	丝氨酸	酪氨酸	半胱氨酸	C
	亮氨酸	丝氨酸	终止	终止	A
	亮氨酸	丝氨酸	终止	色氨酸	G
C	亮氨酸	脯氨酸	组氨酸	精氨酸	U
	亮氨酸	脯氨酸	组氨酸	精氨酸	C
	亮氨酸	脯氨酸	谷氨酰胺	精氨酸	A
	亮氨酸	脯氨酸	谷氨酰胺	精氨酸	G
A	异亮氨酸	苏氨酸	天门冬酰胺	丝氨酸	U
	异亮氨酸	苏氨酸	天门冬酰胺	丝氨酸	C
	异亮氨酸	苏氨酸	赖氨酸	精氨酸	A
	甲硫氨酸(起始)	苏氨酸	赖氨酸	精氨酸	G
G	缬氨酸	丙氨酸	天门冬氨酸	甘氨酸	U
	缬氨酸	丙氨酸	天门冬氨酸	甘氨酸	C
	缬氨酸	丙氨酸	谷氨酸	甘氨酸	A
	缬氨酸(起始)	丙氨酸	谷氨酸	甘氨酸	G

图1.4　遗传密码表：三核苷酸和氨基酸的对应关系

我们已经了解了生命之书是如何在生命的代代之间传递的，那么，作为生命的基本单位——细胞，又是如何使用这本生命之书的呢？这需要通过一个被称为“中心法则”的过程(图1.5)，首先在细胞核内以DNA的一条链为模板合成RNA，这个过程被称为转录，RNA的序列与模板链互补配对，在碱基组成上将T(胸腺嘧啶)换成了U(尿嘧啶)。RNA经过剪接和修饰后进入细胞质中，再于核糖体上通过密码子规则指导合成蛋白质，这个过程被称为“翻译”。“翻译”出的蛋白质还需经过一系列的折叠和修饰才能行使生命功能。后来科学家发现RNA也可以被反转录为DNA，同时RNA也可以自我复制，这就形成了“中心法则”理论现在的形式，细胞(生命体)正是通过它将“书”上的生命“剧本”转化实现为了丰富多彩的生命“舞剧”。

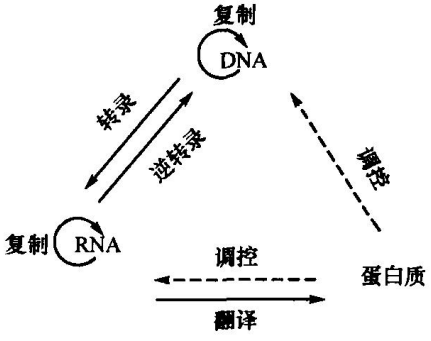

图1.5　中心法则

四、基因组和基因组学

基因组指单倍体细胞中的编码序列和非编码序列在内的全部DNA分子。基因组DNA序列编写着一切生命体活动最基本的生物信息，这些信息是生物体个体建立和维持其生物学特征所必需的。基因组学就是通过分析基因组DNA序列或其表达中间过程或产物等来解读这些信息的学科。在技术上，基因组学通过测序和解读两个相对独立的环节来

达到这一目标。

生命是序列的，如何获取序列成为基因组学的首要问题，测序技术也就成为基因组学最核心的技术。1977 年，Gilbert 等人报道了通过化学降解测定 DNA 序列的方法，同年，Sanger 建立了双脱氧链终止法。测序技术的发展给基因组学研究带来了革命性的改变。20 世纪 80 年代末，测序技术在分子实验室的日常化促使了“人类基因组计划”的诞生；到 20 世纪 90 年代末，Sanger 测序法高通量、自动化的实现促使了“人类基因组计划”的完成，并奠定了 21 世纪基因组学和医学发展的格局；2006 年以来，第二代测序技术的出现更使“万物基因组”和“个体基因组”推上议事日程；在未来几年内，我们将有幸看到下一代测序技术的通用和以基因组学为基础的生命科学时代的到来。

解读基因组序列中的遗传信息是基因组学研究的根本目标。定位、注释基因组序列中功能元件是解读基因组序列的重要内容。这是一个以生物信息为导向、与实验相结合的过程。对于多数功能元件来说，可以直接通过特征序列的寻找和同源分析进行定位和功能注释，也可以用基于转录的高通量实验分析达到目的。

基因组学研究的最终目的就是，通过测序和解读基因组为一切以生物学为基础的产业和应用提供基本的遗传信息。20 世纪 70 年代，基因重组技术的诞生使得分子生物学家借到了“上帝之手”，可以通过改造单个基因而获得相应的性状。基因组学的发展使遗传工程领域有了较快的发展，它为这种“上帝之手”提供了最基本的素材——物种所有基因序列。小鼠基因敲除计划就是一个例子。

然而，由于重组技术本身的缺陷，仅仅对单个基因改造，并不能与基因组学发展的规模和速度相称。一个全新的概念诞生了——“合成生物学”，即人为地从通路和基因组水平设计和制造新的生物部件、装置和系统；或重新设计已有的天然生物系统为人类的特殊目的服务。合成生物学需要两个基本的条件：一个是合成基因组序列的技术；一个是人为地设计能产出所需产物的代谢通路。而后一个完全依赖于大规模基因组测序所得的基因和代谢网络数据库。另外，最近也诞生了能快速改造整个代谢通路的技术。如果说“碱基序列”是基因组这本大书的基本字符，基因组的“解读”为我们提供了基本的语法和素材，那么，我们“书写”基因组的时代指日可待！

第二节　基因组学的发展历程

一、基因组学的催生婆：“人类基因组计划”

“人类基因组计划”（Human Genome Project，HGP）是一项规模宏大的科学计划，其旨在测定组成人类染色体（指单倍体）中所包含的 30 亿个核苷酸序列的碱基组成，从而绘制出人类基因组图谱，且辨识并呈现其上的所有基因及其他功能元件。“人类基因组计划”是人类为了解自身的奥秘所迈出的重要一步，是继曼哈顿计划和阿波罗登月计划之后，人类科学史上的又一个伟大工程。

经历长达 10 年的酝酿，“人类基因组计划”于 1990 年正式启动，计划投资 30 亿美元，预期在 15 年内完成。该计划由美国能源部和国家卫生研究院率先启动，随后，英国、日本、法国、德国和中国先后加入。中国承担并完成“人类基因组计划”的 1%任务（简称“1%项目”）。

“人类基因组计划”的主要内容包括基因组的全序列测定，建立遗传图谱、物理图谱、序

列图谱、转录图谱；进行人类基因的鉴定；建立基因组研究技术、人类基因研究的模式生物；建立信息系统。此外，“人类基因组计划”还包括对社会、法律与伦理问题的研究，交叉学科的技术训练，技术的转让，研究计划的外延等九方面内容。自 1990 年正式启动后，“国际人类基因组协作组（International Human Genome Consortium）”先后完成了平均分辨率为 0.7cM 的遗传图谱（Murray，et al. 1994）和平均分辨率为 100kb 的物理图谱的绘制工作（Schuler，et al. 1996），并于 2001 年发表了人类基因组草图（HGP Consortium 2001），于 2004 年发表了常染色质完成序列（HGP Consortium 2004）。从 1999 年完成 22 号染色体序列分析到 2006 年完成 1 号染色体序列分析，全部 24 条染色体（22 条常染色体和 2 条性染色体 X、Y）的序列都已被全部解析。至此，基因组序列图整合了由 7000 个标记组成、分辨率为 0.7cM 的遗传图谱（Murray，et al. 1994；Dib，et al. 1996）和由 36 000 个标记组成、分辨率为 100kb 的物理图谱（Hudson，et al. 1995；Schuler，et al. 1996），序列全长 28.1 亿碱基对，覆盖 99%常染色质区，全基因组仅剩 341 个空洞，注释了 20 000～25 000 个蛋白质编码基因。至此，“人类基因组计划”终于完美谢幕了。

人类基因组学的启动标志着基因组学作为生物学的一个分支学科的诞生，而它的顺利完成标志着基因组学走向独立和成熟。可以说“人类基因组计划”就是基因组学的催生婆（图 1.6）。

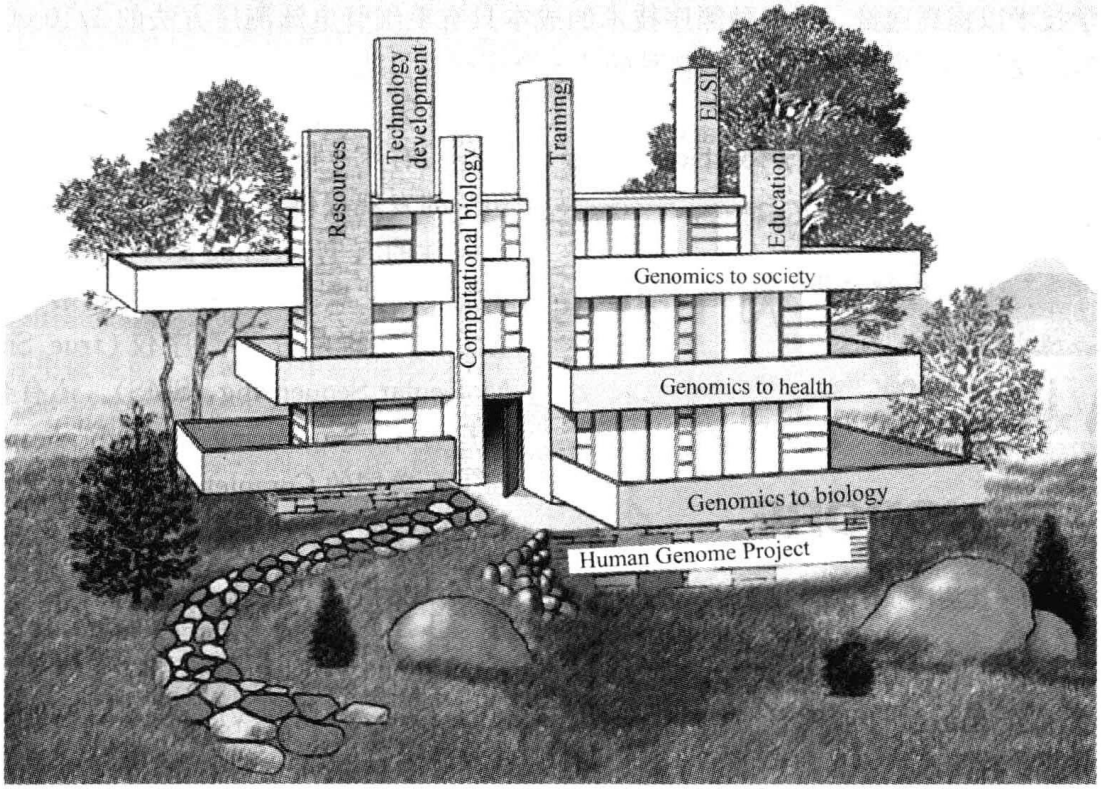

图 1.6 “人类基因计划”奠定了 21 世纪生物学和医学的基础并对社会发展产生深远影响

“人类基因组计划”最终勾画出人类基因组的共有的“参考性”图谱。但是，实际上每个人的基因组序列都是有着部分的差异，虽然比重不大，但从全部的 30 亿对碱基来看总的改变数目仍是不容忽视的。其中最常见的变异是单核苷酸多态性（single nucleotide polymorphism，SNP）。为了描述这些 SNP 在 DNA 上存在的位置、在同一群体内部和不同人群间的分布状况，“人类基因组单体型图计划（HapMap 计划）”于 2003 年启动。中国负责 10%的工作，所产生的全部数据与“人类基因组计划”一样免费向公众开放（IHGSC 2005；IHG-

SC 2007)。HapMap 集合了频率高于 5%的 SNPS 和包括插入、缺失、拷贝数变异、结构变化等其他形式的人类遗传变异。随后中国深圳华大基因研究院、英国 Sanger 研究所(Wellcome Trust Sanger Institute)和美国国家人类基因组研究所(National Human Genome Research Institute,NHGRI)于 2008 年 1 月启动了“千人基因组计划”,旨在提供最详尽的人类遗传变异图谱,鉴别出所有在人群中出现频率高于 1%的突变,以支持疾病的研究。

随着测序技术的发展,“ENCODE 计划”、“癌症基因组计划”、“千种动植物基因组计划”等一系列重大计划相继启动,开启了人类解读生命密码的新征程。

二、测序技术的发展

测序技术是基因组学的核心技术。正是近年来测序技术突飞猛进的发展带来了基因组学今天的繁荣。没有毛细管电泳自动测序仪(Sanger 测序法)的应用,“人类基因组计划”不可能于本世纪初完成。在此之后,测序技术的发展更是日新月异。首先公布的是焦磷酸测序技术,由 Roche 公司推出了相应的测序仪器 454。之后,Illumina 公司推出基于“边合成边测序(Sequencing by Synthesis,SBS)”的 Solexa 测序技术,ABI 公司推出的“边连接边测序(Sequencing by Ligation,SBL)”的 SOLiD 测序技术已经发展成熟,并在不断改进化学和光学技术以提高通量。这三种测序技术的成本只有毛细管电泳测序方法的 1/10 000～1/100,这使得大规模重测序和更多物种的从头测序成为可能。

图 1.7 新一代测序技术:提供新一代测序仪产品的部分公司

今天,新的测序技术仍然不断涌现,虽然还没有商业化的推广,但已为今后基因组学的发展提供了有力的信心(图 1.7)。代表性的有 HeliScope 测序技术,上机前不需要对文库进行任何扩增,是第一台真正意义上的单分子测序仪(true Single Molecular Sequencing,tSMS)。还有杂交测序技术(Sequencing By Hybridization,SBH),美国的 Complete Genomics 公司使用高通量的芯片(单张芯片达到上亿通量)在芯片表面纳米球上扩增 DNA 片段,使用“组合探针锚定连接(Combinational Probe-anchored Ligation)”技术对片段两端的 35 个碱基进行双向测序。除此以外,Pacific Bioscience 公司和 Visigen 公司分别开发的单分子实时测序技术也有广阔的前景。当然,最有可能真正实现 1000 美元/人基因组测序的是纳米孔(nanopore)测序技术,是近年来发展最快、最热门的领域,是纳米技术和生物技术的完美结合。

三、基因组学研究领域的拓展

(一) 基因表达与转录组学

随着越来越多的基因组被测序,接下来的问题是这些基因的功能是什么、不同的基因

参与了哪些细胞内不同的生命过程、基因表达的调控、基因与基因产物之间的相互作用以及相同的基因在不同的细胞内或者疾病和治疗状态下的表达水平等。因此，在“人类基因组计划”后，转录组的研究迅速受到科学家的青睐。转录组学(transcriptomics)是基因组学的新兴学科，即研究细胞在某一状态下所含 mRNA 的序列、类型、拷贝数及转录过程。

转录组学可以提供各种条件下各种基因表达的信息，并据此推断相应未知基因的功能，揭示特定调节基因的作用机制。通过这种基于基因表达谱的分子标签，不仅可以辨别细胞的表型归属，还可以用于疾病的诊断。转录组学最初的技术是表达序列标签(Expression Sequence Tag，EST)，将 mRNA 反转录成 cDNA，再随机地从 cDNA 库中挑选克隆进行测序，由每个克隆获得 100～500bp 的一段序列。ESTs 已经被广泛地应用于基因识别。之后开发的用于转录组数据获得和分析的方法主要有 cDNA 芯片检测技术。随着高通量测序技术的出现，RNA 测序(RNA-Seq)和数字化表达谱(Digital Gene Expression，DGE)已经成为现在转录组学研究的主流技术。

(二) 染色体的修饰与表观基因组学

染色体上各种各样的修饰对基因的表达有很大影响，而这些修饰并不改变核苷酸本身，如甲基化和组蛋白修饰等(图 1.8)，这些被称为表观遗传的改变，在全基因组范围研究表观遗传的变异即是表观基因组学(epigenomics)。

甲基化因其与人类发育和肿瘤的密切关系，已经成为表观遗传学和表观基因组学的重要研究内容。甲基化是指从活性甲基化合物(如 S-腺苷基甲硫氨酸)上将甲基催化转移到其他化合物的过程。DNA 甲基化是指生物体在 DNA 甲基转移酶(DNA methyltransferase，DMT) 的催化下，以 s-腺苷甲硫氨酸(SAM)为甲基供体，将甲基转移到特定的碱基上的过程。DNA 甲基化可以发生在腺嘌呤的 N-6 位、胞嘧啶的 N-4 位、鸟嘌呤的 N-7 位或胞嘧啶的 C-5 位等。但在哺乳动物中 DNA 甲基化主要发生在 5′-CpG-3′的 C 上生成 5-甲基胞嘧啶(5mC) 。人类基因组序列中的 CpG 二核苷酸主要以两种形式存在：一种分散在重复序列中，并且总是处于甲基化状态，另一种则以大小为 300～3000bp 且富含 CpG 二核苷酸岛的形式存在，由于这些 CpG 岛通常位于基因的转录起始位点(启动子或第一外显子)附近并可能参与了基因的表达调控，因而受到人们的广泛关注，特别是 CpG 岛异常高甲基化所致抑癌基因转录失活以及异常低甲基化所致原癌基因的激活已经成为肿瘤研究中的热点问题。

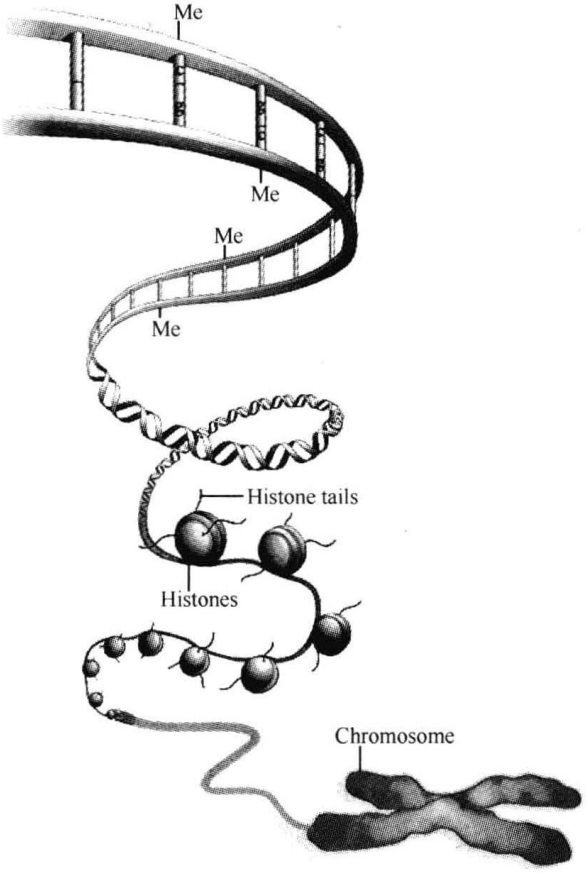

图 1.8 染色体上主要的两种修饰方式：DNA 甲基化和组蛋白的修饰

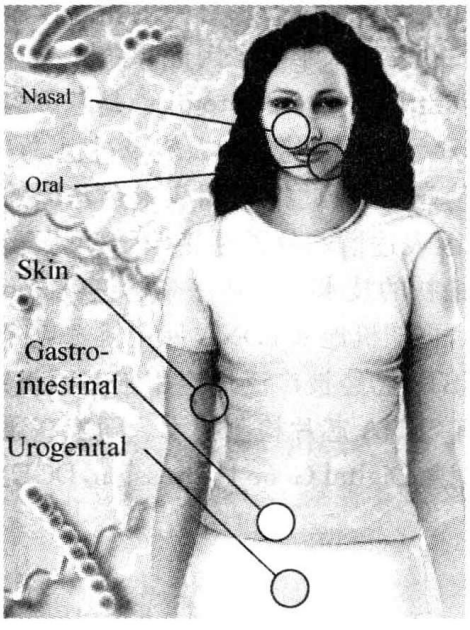

图 1.9　人体的多个器官均与环境微生物形成复杂的作用关系

（三）宏基因组学：从基因组的视角看环境微生物

人类生活在一个到处都有微生物存在的世界里，成人体表和体内含有的微生物数量高达 10^{15}，约为人体总细胞数的 10 倍。人体微生物种类繁多，数量巨大，通过多种方式影响人体，同时又受到多种内外因素的影响，形成一个复杂的系统（图 1.9）。“人体微生物基因组”指的是人体内共生的菌群，包括肠道、口腔、呼吸道、生殖道等处菌群基因组总和，人体微生物基因组计划的目标就是测定人体共生菌群的基因组序列信息，研究与人体发育、健康有关的基因功能。这种通过直接从环境样品中提取全部微生物的 DNA 构建基因组文库，利用基因组学的研究策略研究环境样品所包含的全部微生物的遗传组成及其群落功能被称作宏基因组学（metagenomics）。我们现在能够通过深度测序来鉴定复杂群落中包含的微生物，跳过了传统研究方法中微生物培养这一步。这些工作一旦完成，人类对自身的认识将会达到一个空前的水平。

（四）翻译组学：以测序来研究蛋白质组学

mRNA 翻译成蛋白质是基因表达的重要阶段，研究翻译过程要比研究转录过程困难得多。在转录水平只需研究转录出的 mRNA 的量，就能了解各个基因的表达情况，而翻译水平牵涉 RNA 和蛋白质的复杂作用机制。核糖体作为蛋白质翻译的场所，自然成为蛋白质翻译水平研究的对象。翻译起始，核糖体结合到 mRNA 5′非翻译区，并不断向下游移动，当遇到起始密码子，核糖体开始招募氨基酸合成多肽链，直到遇到终止密码子，核糖体从 mRNA上脱落，肽链合成结束（图 1.10）。真核生物在翻译过程中，一个核糖体在 mRNA 上能够结合约 30 个核苷酸。通过 RNA 酶消化，mRNA 中不与核糖体结合的部分被降解，而那些受核糖体保护的 mRNA 片段留下，我们将这些 mRNA 进行测序就能得到翻译的序列，并可推测蛋白质的氨基酸组成，这被称 Ribosomal Profiling 技术，也称作 Translational Profiling 技术，用以分析细胞内所有因突变或环境变化而改变的 mRNA 的翻译状态。由此，我们可以用测序替代蛋白质组的一系列实验，充分发挥测序成本低和数字化信号的优势，为蛋白质组学的发展提供有力的支持。比较研究细胞内蛋白质丰度和 mRNA 表达水平的方法被称作翻译组学（translatomics）。

四、从解读生命到书写生命

当今基因组学的研究主要通过序列的分析来解读生命，但我们的最终目的是通过基因组序列的设计来“书写”生命。以往的遗传工程，通过改造个别基因来获得相应的性状，如 RNA 干涉和转基因技术。而随着第一代“人造细菌”的问世，“人造生命（合成生物学）”的时代已悄然到来。合成生物学（synthetic biology）的基础是合成核苷酸序列的技术，包括直接合成、化学合成和大片段转移技术等；而其核心则是基于诸多基因组序列和“三大系统”

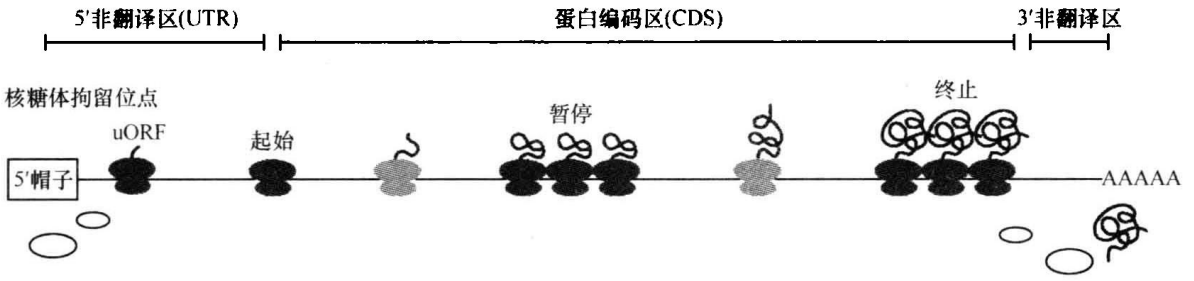

图 1.10　mRNA 翻译的过程

(代谢途径、信号传导通路、基因表达调控网络)的"人造基因组"序列的设计和组合,这完全依赖于大规模基因组测序所得的基因和代谢网络数据库,因此我们得到的信息越全面,创造的手段就越自由。

合成生物学的研究首先要明确维持一个生命正常活动的最小基因组有多小?哪些基因对于生命体是必不可少的?对线虫进行 RNA 干扰研究发现大多数的基因对生命活动并不起主要作用(图 1.11)。*M. genitalium* 基因组的测序得出大约 470 个编码区是必需的,其中包括与 DNA 修复、能量代谢和其他重要生命途径等有关的成分(Peterson, et al. 1993)。随后的研究成果使这一数字缩小到了 386 个(Fraser, et al. 1995)。

同时,科学家们都在试图构建出遗传物质不同于核酸分子的新生命。例如,肽核酸(Peptide Nucleic Acids, PNA)是 20 世纪 90 年代丹麦科学家发明的一类全新的以多肽骨架取代糖磷酸主链的 DNA 类似物,近十年来,人们为其在许多高科技领域找到了用途。

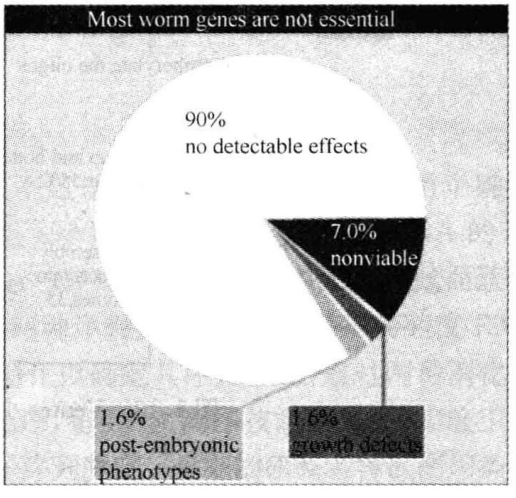

图 1.11　线虫的多数基因对其生命活动不是必需的

最小基因组指一定条件下生物保证存活所必需的基因构成的集合,这些基因被称为必需基因(essential gene)。最小基因组和人工合成基因组研究一方面是解析生命存活所必需基因和人工合成生命体的首要步骤,另一方面也可用于研究难以获取的基因或人工设计的核酸序列的生物学特性,使研究者们能从各个层次上更深入地理解生命现象的本质,为下一步创建人工生命体奠定基础。2002 年,德国埃科德·威默(Eckard Wimmer)的研究团队合成了有生物活性的脊髓灰质炎病毒基因组(Cello, et al. 2002)。2003 年,克雷格·文特(Criag Venter)研究小组合成了噬菌体 ΦX174 基因组(5386 bp)(Smith, et al. 2003),2008 年合成了生殖道支原体基因组(Gibson, et al. 2008)。2010 年该研究团队剔除了山羊蕈状支原体基因组中的部分基因并加入"水印"标记基因,制造了世界上第一个人工生命体"Synthia",并使其成功自我复制(Gibson, et al. 2010)。2011 年 7 月,Church 教授课题小组利用 MAGE 技术实现了大肠杆菌中 314 个终止密码子替换(Isaacs, et al. 2011)。2011 年 9 月,杰夫·布克(Jef Boeke)领导的科研团队人工合成出两个染色体片断并将其放入一个活酵母菌体内,酵母菌仍能正常存活,未出现明显异常(Dymond, et al. 2011)。该研究是世界上首次成功合成真核生物的部分基因组,标志人工合成生物基因组的研究又迈出了重要一步(表 1.1)。