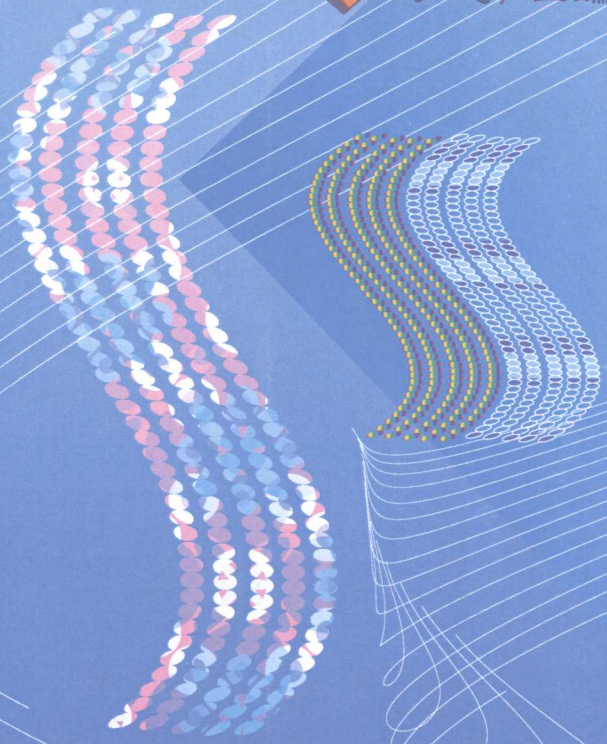
Introduction of Genomics

基因组学导论

（公共课）

刘 越／主编



中央民族大学出版社
PRESS OF THE CENTRAL UNIVERSITY FOR NATIONALITIES

基因组学导论

（公共课）

刘 越 主编

中央民族大学出版社

图书在版编目（CIP）数据

基因组学导论/刘越主编. —北京：中央民族大学出版社，2008.8
ISBN 978-7-81108-521-1
Ⅰ. 基… Ⅱ. 刘… Ⅲ. 基因组—研究 Ⅳ. Q343.1

中国版本图书馆 CIP 数据核字（2008）第 126546 号

基因组学导论

主　　编　刘　越
责任编辑　戴苏芽
封面设计　布拉格工作室
出 版 者　中央民族大学出版社
　　北京市海淀区中关村南大街 27 号　邮编：100081
　　电话：68472815（发行部）传真：68932751（发行部）
　　68932218（总编室）　68932447（办公室）
发 行 者　全国各地新华书店
印 刷 厂　北京宏伟双华印刷有限公司
开　　本　880×1230（毫米）　1/32　印张：12
字　　数　300 千字
版　　次　2008 年 8 月第 1 版　2008 年 8 月第 1 次印刷
书　　号　ISBN 978-7-81108-521-1
定　　价　30.00 元

前 言

基因组学来自于基因组，基因组指的是生物体内所有 DNA 的总和，基因组学是研究基因组的学科。随着人类基因组测序的完成和全球基因组计划的迅猛发展，超过 2000 种生物的基因测序工作接近或已经完成，其中包括拟南芥、水稻等植物；果蝇、小鼠等动物和大肠杆菌、流感病毒等微生物；特别值得提出的是，对于不可培养的微生物，人们已经开始着手宏基因组的研究。基因组及基因组学的研究对于破译物种的遗传密码、功能基因克隆、基因治疗等方面发挥重大的作用。

基因组学是现代生命科学前沿之一，为了让更多的学生了解基因组学的学科前沿发展动态和新的研究进展以及基因组学热点领域，拓展知识面，普及 21 世纪生命科学前沿知识，提高对基因组计划以及生命科学技术的认识，特编写这本教材。本教材是各位编者结合自己的教学和科研的相关研究实践并参考大量的国内外相关书籍和文献资料编写而成，可作为非生物类专业的公共课教材及其他专业的科普书籍。

全书的主要内容是介绍与人类生活密切相关的基因组学，主要包括基因组学的基本概念、研究内容和研究方法，以及耳熟能详的人类基因组计划、植物基因组计划、动物基因组计划、微生物基因组计划等研究领域的最新进展和基因在转移、利用过程中涉及的相关问题，内容广泛丰富。

全书由刘越主编，负责结构设计、章节安排、内容调整及全书的审核等工作。本书共分五章，第一章、第二章以及第四章的第一节和第四节由中央民族大学刘越编写；第三章由中国农业科学院张兰、张治国、于永涛和聊城大学的樊颖伦编写；第四章的第二节和第三节由中国农业科学院张俊成和北京市丰台区职工大学蒋黎明编写；第五章由中国农业科学院王海燕、中国协和医科

大学齐小强及李瑞芬编写。在编写过程中，中国农业科学院作物科学研究所贾继增研究员、孔秀英研究员、中央民族大学生命与环境科学学院的周宜君副教授及生物科学系的老师们给予了有益的指导，对本书提出许多合理的建议，在此一并表示衷心地感谢。

由于基因组学的发展迅速及编写时间有限，许多最新的研究进展、研究成果来不及编入本书，加之编者水平有限，书中难免出现错误和不足，敬请学术界同仁和广大读者批评指正，以便再版时改进。

编者

2008 年 6 月

目 录

第一章 绪论

第一节 基因组学的历史回顾

基因组（GENOME）一词是 1920 年德国汉堡大学植物学教授汉斯·温克勒（Hans Winkler）从 GENes 和 chromosOEs 铸成的，用于描述生物的全部基因和染色体组成的概念。作为基因组学研究的对象——基因，在基因组学的发展史上具有重要地位，对基因的发现、认识和关注是基因组学产生和发展的前提。如果我们从孟德尔（1866 年，分离和自由组合定律）和达尔文（1859 年，进化论）算起的话，基因组学最多只有 150 年的历史；如果我们从摩尔根（1908 年，果蝇遗传学实验）算起的话，基因组学只有 100 年的历史；如果我们从沃森和克里克发现 DNA 双螺旋结构（1953 年）算起的话，基因组学刚刚有 50 多年的历史；如果我们从 DNA 测序方法和第一个基因组全部测序（1977 年）算起的话，基因组学从梦想到现实的过程才用了短短的 30 年；如果我们从人类基因组计划的酝酿（1985 年）开始算起的话，基因组学所走过的路仅仅 20 多年而已。基因组学是一个相对年轻的学科，一个正在成长的学科。

1 对“基因”的认识过程

遗传学的奠基人奥地利人孟德尔（Gregor Johann Mendel 1822～1884），在布鲁诺（Brno，德 Brünn，现属捷克）的奥古斯丁教派修道院的菜园里，挥洒了 8 年的汗水，于 1865 年 2 月在奥地利自然科学学会会议上报告了自己植物杂交研究结果，第二年在奥地利自然科学学会年刊上发表了著名的《植物杂交试验》的

论文，发现了遗传学的两个基本规律——分离和自由组合规律。文中指出，生物每一个性状都是通过遗传因子来传递的，遗传因子是一些独立的遗传单位。这样把可观察的遗传性状和控制它的内在的遗传因子区分开来了，遗传因子作为基因的雏形名词诞生了。基因的存在最早是由孟德尔在 19 世纪推断出来的，并不是观察的结果。在达尔文发表进化论后不久，他试图通过对豌豆进行试验来解释该理论。但是直到 19 世纪末他的研究才被人们所重视。虽然孟德尔还不知道这种物质是以怎样的方式存在，也不知道它的结构是怎样，但孟德尔“遗传因子”的提出毕竟为现代基因概念的产生奠定了基础。孟德尔在他的豌豆杂交实验论文中，用大写字母 A、B 等代表显性性状如圆粒、子叶黄色等，用小写字母 a、b 等代表隐性性状如皱粒、子叶绿色等。他并没有严格地区分所观察到的性状和控制这些性状的遗传因子。但是从他用这些符号所表示的杂交结果来看，这些符号正是在形式上代表着基因，而且为了方便起见至今在遗传学的分析中仍沿用它们来代表基因。

可以说，遗传因子实际上是孟德尔根据其实验结果所虚拟假想的某种东西，从那时起遗传学家踏上了寻找基因实体的道路。萨顿（W.S. Sutton 1877～1916）和鲍维里（T. Boveri 1862～1915）两人注意到在杂交试验中遗传因子的行为与减数分裂和受精中染色体的行为非常吻合，他们作出“遗传因子位于染色体上”的“萨顿—鲍维里假想”：他们根据各自的研究，认为孟德尔的“遗传因子”与配子形成和受精过程中的染色体传递行为具有平行性，并提出了遗传的染色体学说，认为孟德尔的遗传因子位于染色体上，即承认染色体是遗传物质的载体，第一次把遗传物质和染色体联系起来。这种假想可以很好地解释孟德尔的两大规律，在以后的科学实验中也得到了证实。1909 年丹麦遗传学家约翰逊（W. Johansen 1859～1927）在《精密遗传学原理》一书中提出“基因”

概念，以此来替代孟德尔假定的“遗传因子”。从此，“基因”一词一直伴随着遗传学发展至今。“基因”一词来自希腊语，意思为“生”。约翰逊还提出了“基因型”与“表现型”这两个含义不同的术语，前者是一个生物的基因成分，后者是这些基因所表现的性状，初步阐明了基因与性状的关系。不过此时的基因仍然是一个未经证实的，仅靠逻辑推理得出的概念。

美国实验胚胎学家、遗传学家摩尔根（Thomas Hunt Morgan 1866～1945）和他的学生们于 1908 年前后开始利用果蝇作了大量的潜心研究。他在 1910 年通过果蝇眼色突变性状的遗传实验发现了伴性遗传现象，第一次揭示出一种或多种遗传特性与某一特定染色体的明确联系；他和他的同事们进一步通过大量的果蝇杂交实验又发现了遗传学的第三个基本规律——连锁互换规律，从而继承和发展了孟德尔的遗传学说。他们为遗传染色体学说最终提供了更充分、直接、可靠的证据，并认为染色体是孟德尔式遗传性状传递机理的物质基础。1926 年他的巨著《基因论》出版，从而建立了著名的基因学说，他还绘制了著名的果蝇基因位置图，首次完成了当时最新的基因概念的描述，即基因以直线形式排列，它决定着一个特定的性状，而且能发生突变并随着染色体同源节段的互换而交换，它不仅是决定性状的功能单位，而且是一个突变单位和交换单位。

摩尔根等人还认为，基因是遗传的功能单位，它能产生特定的表型效应；基因又是一个独立的结构单位，在同源染色体之间可以发生基因的互换，但交换只能发生在基因之间而不是发生在基因之内；基因可以发生突变，由一个等位形式变为另一等位形式，因而基因又是突变单位。这就是 20 世纪 40 年代以前流行的所谓“功能、交换、突变”三位一体的基因概念。这种认识把基因与染色体联系起来，说明了基因的物质性，基因存在的场所及

排列方式，基因从此就不再是一个抽象的概念了。当然这时人们仍然不了解基因的化学本质以及基因是如何控制生物性状的。

从 20 世纪 40 年代起，人们开始注意基因与性状的关系，即开始研究基因如何控制性状的问题。1941 年，比得尔（G.W. Beadle 1903～1989）和塔特姆（E.L. Tatum 1909～1975）以红色链孢霉为材料进行生化遗传研究。他们通过诱变获得了多种氨基酸和维生素的大量营养缺陷突变体。这些突变基因不能产生某种酶，或只产生有缺陷的酶。例如，有一个突变体不能合成色氨酸是由于它不能产生色氨酸合成酶。于是，研究者提出了“一个基因一种酶”的假说，认为基因对性状的控制是通过基因控制酶的合成来实现的。这一假说在 20 世纪 50 年代得到充分验证，后来发现有些蛋白质不只由一种肽链组成，如血红蛋白和胰岛素，不同肽链由不同基因编码，因而又提出了“一个基因一条多肽链”的假设。“一个基因一种酶”和“一个基因一条多肽链”理论的提出，大大促进了分子遗传学的发展，人们急切期望能搞清楚基因的化学结构。1949 年鲍林（L.C. Pauling 1901～1994）与合作者在研究镰刀型细胞贫血症时推论：基因决定着多肽链的氨基酸顺序，这样 20 世纪 40 年代末至 20 世纪 50 年代初，基因是通过控制合成特定蛋白质以控制代谢决定性状的原理变得清晰起来。

虽然 DNA 在细胞核中很早就被发现，但证明其为遗传物质的决定性实验是 1944 年艾弗里（O.T. Avery 1877～1955）的肺炎双球菌的转化实验。他和麦卡蒂（M. McCarty 1911～2005）等人发表了关于“转化因子”的重要论文，首次用实验明确证实：DNA 是遗传信息的载体。1952 年赫尔希（A.D. Hershey 1908～1997）和蔡斯（M.M. Chase 1927～）进一步证明遗传物质是 DNA 而不是蛋白质。

这一实验不仅证明了 DNA 是遗传物质，揭示了遗传物质的化学本质，也大大推动了对核酸的研究。1953 年，美国分子生物

学家詹姆斯·沃森（J.D. Watson）和英国物理学家佛朗西斯·克里克（F.H.C. Crick）根据威尔金斯（M. Wilkins）和富兰克林（Rosalind Franklin 1920-1958）所进行的X射线衍射分析，提出了著名的DNA双螺旋结构模型，进一步说明基因成分就是DNA，它控制着蛋白质合成。进一步的研究证明，基因就是DNA分子的一个区段。每个基因由成百上千个脱氧核苷酸组成，一个DNA分子可以包含几个乃至几千个基因。基因的化学本质和分子结构的确定具有划时代的意义，它为基因的复制、转录、表达和调控等方面的研究奠定了基础，开创了分子遗传学的新纪元。

基因本质的确定为分子遗传学发展拉开了序幕。1955年，美国分子生物学家本泽（Benzer）对大肠杆菌T_4噬菌体作了深入研究，揭示了基因内部的精细结构，提出了基因的顺反子（cistron）概念。本泽把通过顺反实验而发现的遗传的功能单位称为顺反子，1个顺反子决定一条多肽链，顺反子即是基因。1个顺反子内存在着很多突变位点——突变子，突变子就是改变后可以产生突变型表型的最小单位。1个顺反子内部存在着很多重组子。重组子就是不能由重组分开的基本单位。理论上每一核苷酸对的改变，就可导致一个突变的产生，每两个核苷酸对之间都可发生交换。这样看来，一个基因有多少核苷酸对就有多少突变子，就有多少重组子，突变子就等于重组子。这个学说打破了过去关于基因是突变、重组、决定遗传性状的“三位一体”概念及基因是最小的不可分割的遗传单位的观点，从而认为基因是DNA分子上一段核苷酸顺序，负责着遗传信息传递，一个基因内部仍可划分若干个起作用的小单位，即可区分成顺反子、突变子和重组子。一个作用子通常决定一种多肽链合成，一个基因包含一个或几个作用子。突变子指基因内突变的最小单位，而重组子为最小的重组合单位，只包含一对核苷酸。所有这些均是基因概念的伟大突破。

关于基因的本质确定后，人们又把研究视线转移到基因传递遗传信息的过程上。在20世纪50年代初人们已懂得基因与蛋白质间似乎存在着相应的联系，但基因中信息怎样传递到蛋白质上这一基因功能的关键课题在20世纪60年代至70年代才得以解决。从1961年开始，尼伦伯格（M.W. Nirenberg）和科拉纳（H.G. Khorana）等人逐步搞清了基因以核苷酸三联体为一组编码氨基酸，并在1967年破译了全部64个遗传密码，这样把核酸密码和蛋白质合成联系起来。然后，沃森和克里克等人提出的“中心法则”更加明确地揭示了生命活动的基本过程。1970年泰明（H.M. Temin）以在劳斯肉瘤病毒内发现逆转录酶这一成就进一步发展和完善了“中心法则”，至此，遗传信息传递的过程已较清晰地展示在人们的眼前。过去人们对基因的功能理解是单一的，即作为蛋白质合成的模板。但是1961年法国雅各布（F. Jacob）和莫诺（J.L. Monod）提出了“操纵子”学说，又大大扩大了人们关于基因功能的视野。他们在研究大肠杆菌乳糖代谢的调节机制中发现了有些基因不起合成蛋白质模板作用，只起调节或操纵作用。从此根据基因功能把基因分为结构基因、调节基因和操纵基因。基因的概念随着多学科渗透和实验手段日新月异又有突飞猛进的发展，主要有以下几个方面：

结构基因和调控基因 根据操纵子学说，并不是所有的基因都能为肽链进行编码。于是便把能为多肽链编码的基因称为结构基因，包括编码结构蛋白和酶蛋白的基因，也包括编码阻遏蛋白或激活蛋白的调节基因。有些基因只能转录而不能翻译，如tRNA基因和rRNA基因。还有些DNA区段，其本身并不进行转录，但对其邻近的结构基因的转录起控制作用，被称为启动基因和操纵基因。启动基因、操纵基因与其控制下的一系列结构基因组成一个功能单位叫做操纵子（operon）。就其功能而言，调节基因、

操纵基因和启动基因都属于调控基因。这些基因的发现，大大拓宽了人们对基因功能及相互关系的认识。

断裂基因 20 世纪 70 年代中期，法国生物化学家查姆帮（Chamobon）发现，细胞内的结构基因并非全部由编码序列组成，而是在编码序列中间插入无编码作用的碱基序列，这类基因被称为间隔或断裂基因。这一发现于 1977 年被英国的查弗里斯和荷兰的弗兰威尔在研究兔 β-球蛋白结构时所证实。1928 年，生化学家吉尔伯特（Gilbert）提出基因是一个转录单位的设想，他认为基因是一个嵌合体，包含两个区段：一个区段由遗传密码组成，将被表达，称为“外显子”；一个区段由非遗传密码组成，将在 mRNA 中被删除，称为“内含子”。近年来的研究发现，原核生物的基因一般是连续的，在一个基因的内部没有非遗传密码的序列（即不含“内含子”），而真核生物的绝大多数基因都是不连续的断裂基因。断裂基因的表达程序是：整个基因先转录成一条长 RNA 前体，其中的非编码序列被一种称为“剪接”的酶切除，两端再相互连接成一条连续的密码顺序，以形成成熟的 mRNA。DNA 分子断裂基因的存在为基因功能的发展赋予了更大的潜力。

重叠基因 长期以来，人们一直认为在同一段 DNA 序列内是不可能存在重叠的读码结构的。但是，1977 年，维纳（Weiner）在研究 Q0 病毒的基因结构时，首先发现了基因的重叠现象。1978 年，费尔（Feir）和桑格（Sanger）在研究分析 ΦX174 噬菌体的核苷酸序列时，也发现由 5375 个核苷酸组成的单链 DNA 所包含的 10 个基因中有几个基因具有不同程度的重叠，但是这些重叠的基因具有不同的读码框架。以后在噬菌体 G4、MS2 和 SV40 中都发现了重叠基因。基因的重叠性使有限的 DNA 序列包含了更多的遗传信息，是生物对它的遗传物质经济而合理的利用。

假基因 1977 年，G. Jacp 在对非洲爪蟾 5S rRNA 基因簇的研究后提出了假基因的概念，这是一种核苷酸序列同其相应的正常