

致 谢

本书自 2006 年开始搜集资料，2014 年 9 月开始动笔，集中了很多人的辛劳和心血，特别向他（她）们致以衷心的感谢：

首先，要感谢助理主编夏志和王晓玲，感谢他们从资料收集到最后定稿的全过程中所显示的智慧 and 作出的贡献；

感谢参与本书有关章节初稿撰写的（以汉语拼音为序，下同）阿叁、董伟、方晓东、郭苗苗、胡学达、李波、倪培相、苏夜阳、沈玥、王晓玲、吴婷婷、夏志、赵宏翠、朱师达；

感谢参与本书最终校稿的董伟、刘韧、毛文静、王晓玲、夏志、张国捷、张丽、张秀清、朱燕楠；

感谢为本书构思、设计和手工描绘插图的李胜霆；

感谢参与本书资料收集的蔡锴晔、陈翠、陈姝婧、冯强、高长欣、栗东芳、李贵波、李京湘、李思淼、刘程章、刘汉奎、刘石平、聂超、田娟、王丽敏、吴静、叶辰、叶明芝、曾筱凡、周加利、张建国、张勇、赵倩、赵屹等；

感谢参与本书的组织、协调并提供资源的冯小黎、李生斌、汪建、王俊、徐讯、尹烨、杨碧澄、杨爽等；

感谢为本书提供部分素材与相关信息的华大基因相关部门和工作人员；

当然，还要衷心感谢拨冗阅读初稿，并给予鼓励、支持、指正的巴德年、强伯勤、陈润生、康乐、詹启敏和安锡培等诸位老师，特别是对《基因组学》（征求意见稿）提出宝贵修改意见的郝柏林、高翼之等老师；

衷心感谢对中国及世界的基因组学作出重大贡献的华大基因的几代工作人员及他们的家人，国家人类基因组研究中心（南方、北方）与所有其他中国团队。

此外，笔者特别感谢 Ewan Birney, Lars Bolund, Allen Bradley, Sydney Brenner, George Church, Brian Clark, Francis Collins, David Cox, Antoine Danchin, Richard Durbin, Hans Galjaard, Richard Gibbs, Thomas Hudson, Eric Lander, Michael Morgan, Arno Motulsky, Søren Nørby, Stephen O'Brien, Maynard Olson, Aristides Patrinos, Richard Roberts, Edward Rubin, Yoshiyuki Sakaki, John Sulston, Harold Varmus, Friedrich Vogel, James Watson, Jeans Weissenbach 等国际师友。他们的学术思想和科学贡献，以及对笔者 17 年来的鼓励和指教，是成就本书的重要基石。

最后，还要感谢早逝的父亲和 90 高龄的老母亲、兄弟姐妹和所有其他家人的关怀，特别感谢夫人刘韧博士对本书编著的支持和贡献。

杨焕明

2016 年 9 月 9 日

前言

15年前，当全球庆祝“人类基因组计划 (Human Genome Project, HGP)”完成人类基因组的序列草图之时，我便坚信基因组学一定会作为一门授课学科，登上高等院校的大雅之堂，也萌生了撰写一本基因组学教科书的念头。

15年过去了，大家认为理所当然堪当此重任者——基因组学的实践者和重要贡献者、国际人类基因组测序协作组 (International Human Genome Sequencing Consortium, IHGSC) 的主要成员居然无一人动笔，至少迄今没有一人完成这一工作，我自己也没有将此心愿付诸行动，尽管在很多国家基因组学已成为高校的选修或必修课。也许最可能的解释是：他们或陶醉于基因组学一个又一个的新突破；或穷于应付或乐于应答该领域一个又一个的新挑战；或者担心这样的教科书只能招惹非议；最为担心的恐怕还是——这样一个日新月异的研究领域，所有的教科书都难以逃脱“刚一出版就已过时”的命运。此类实例屡见不鲜。

15年来，我们矢志不移，使中国屹立于基因组学的世界民族之林。今天，我们既然要编写这样的教科书，就一定要写好：以我们的亲身经历，写出既反映基因组学的世界潮流和发展趋势，又体现中国特色和反映中国贡献的教科书。我们要共同努力、一起建立并完善基因组学的教学体系，正如合作建立我们的基因组学的研究体系那样。

我们与读者分享的首先是基因组学的理念，其次才是围绕这一理念和核心技术的知识和经验。

我们力图使本书成为一本“好读、好懂、好记、好学、好教”的入门教科书，特别注重科学性、系统性、可读性和趣味性相结合。我们明白，一本好的教科书的编写是一项长期的艰巨任务，本书打算“修订不止，再版不已”，计划将根据学科的发展，每年或数年修订再版，与时俱进地反映这一领域的进展、基本概念的更新；同时，也为了督促编者自己认真勘误，有错必纠。

“学而知之”乃求学之要旨——不管从师求学还是无师自通，一本好的教科书必不可少；“悟而知之”乃学习生命科学之优势——以“我是生命”来领悟基因组学；“习而知之”——“学而时习之，不亦乐乎？”乐在把书“读薄”。本书的正文只有全书的五分之一，“细目”则可一览全书。

写书难，写教科书更难，写一门新兴而又在快速发展的学科教科书更是难上加难。一本受欢迎的教科书的形成过程，正是编与读互动、学与教相长的过程。本书的命运如何，将取决于使用者与编写者现在开始的共同努力。

此为前言，言犹未尽，详见“编写说明”。

杨焕明

2016年9月9日，于中国深圳

编写说明

《基因组学》的编写是一次新的尝试。为使读者对我们的构思和本书的构架、体系及写作特点有更全面的了解，也便于对全书提出更有针对性的批评与建议，我们特参照享誉全球、久版不衰的《基因的分子生物学》（Watson 等著，中文第 7 版，科学出版社，2015）的相似方式，对本书的编写作较大篇幅的详细说明。

框架与构思

参照《基因的分子生物学》，本书将通常的目录拆分为“简目”与“细目”。“简目”仅有一页，可一目了然地通观全书的构架。“细目”共有 14 页，接近《基因的分子生物学》的“内容详注”，旨在展示我们对基因组学这一学科的科学体系与教学体系的构思，可作为全书的复习提纲。

本书共分五部分。

第一部分——基因组学概论分为两篇。

第一篇《基因组学》导读不同于一般的“导论”。导论是一书的入门之“导”，而导读则是全书的阅读之“导”。作为入门教科书，本书的导读共有九节，从不同的角度，概括了基因组学这一学科的定义、理念、源流、发展趋势及应用范例，以及基因组学相关技术在生命科学和生物产业中的地位及其对人类社会的影响，特别是必须重视的相关生命伦理学原则和应该了解的中国的贡献。其内容并没有与本书的框架一一对应，而其中发展趋势和新近进展的部分，将是再版的主要更新内容。

第二篇基因组学的发展史主要描述 HGP 这一基因组学的第一次实践，这是史实。尽管“基因组”一词问世于 1920 年，“基因组学”命名于 1987 年，但真正意义的基因组学是随着 HGP 的讨论才开始的。正因为如此，本书以 1984 年美国 HGP 的第一次讨论会开始，特别以笔者自己亲身参与的体会，列举了这一计划的学术源流、目标变迁、技术路线改进、模式生物选取，以及人类基因组序列草图和精细图的完成，特别是中国的参与和贡献。本书还将 HGP 的诸多后续计划，包括延续至今的“国际千人基因组计划 (The International 1000 Genomes Project, G1K Project)”和“国际癌症基因组计划 (The International Cancer Genome Project, ICGP)”，作为基因组学发展史的重要部分。

第二部分——基因组学的方法学也分为两篇，是本书最重要的部分。

这一部分的重要特点是开宗明义，将测序仪与其他的测序关键技术作了明确的切割。测序作为一项技术，测序仪只是其中的一部分，还包括诸多更为重要的关键技术。特别是对大学生和研究生来说，了解这些机器之外的关键技术，如上机前准备工作与所有相关技术，尤其是基因组序列的组装和分析，要比按照说明书操作机器更为重要。

第一篇 DNA 测序分为两章，第一章（测序仪）介绍测序仪的历史性技术突破、市场化的仪器和基本原理，以及“下一代”测序仪的发展趋势。第二章（关键技术）除了介绍策略与方案、测序材料外，还特别介绍了关键技术的发展方向。

第二篇序列的组装和分析的前两章（序列的组装、基因组概貌分析）内容包括下机序列 (reads) 的处理与质控、序列的组装，以及基因组概貌和变异等初步分析。本书主要列举分析的实例，以及部分开源软件工具的使用方法，并没有对所有的生物信息学技术作详尽的介绍。而第三章（基因组比较分析）只

是例举了演化基因组与 META 基因组，显然不尽人意。

第三部分 —— 基因组的生物学是本书的另一重要部分和特色，即在基因组序列和基因组学研究的基础上，生命科学揭开的新篇章。

这一部分主要是以物种为基础的，着力体现一个物种的基因组序列所带来的该物种生物学各方面研究的革命性变化。物种是所有生物研究的起点，也是基因组学及其他“组学”研究的起点。本书试图从一直跟踪的已测序的诸多物种中选取一些有代表性的，如广为熟知的模式生物和具有重要科学和生态意义及经济价值的物种，来介绍基因组序列基础上的生物学新发现，也建议读者作进一步的选择。对照浩如烟海的新文献、日新月异的新知识，这部分内容显得不尽人意，也反映了笔者知识的局限性，这都将在再版时反思、再议、再补、再改。而所有跨物种、跨组学的分析，特别重要的是生物学的灵魂——演化，暂且穿插在各物种基因组的分析之中，也有待补充并将其在这一知识体系中更好定位。

第一篇人类基因组中的第三章（临床基因组学）就是这一构思的尝试。本章开门见山便是基于序列数据的检测技术。这是基因组学及测序技术给医学临床带来的重大影响。在将来的再版中还会作更多的调整与充实。本章中的法医基因组，则因其特殊性而单独阐述，而外饰基因组学和 META 基因组学等新兴学科，由于已写入技术部分，在这里没有占很大篇幅。

第二篇动物基因组将按动物界的系统分类选择脊椎动物门和无脊椎动物门以及为人熟知的其他门类的典型动物作简要介绍。

第三篇植物基因组将植物人为地分为模式植物、主粮类、蔬菜类、瓜果类、经济类、花卉类和单细胞植物类，并选择典型的物种作简要的介绍。

第四篇微生物基因组则分为真菌、细菌、病毒和环境微生物群，并选择代表物种简要介绍。物种甚多，描述甚简，仅供选择使用。

第四部分 —— 基因组的设计和合成之所以没有使用常见的合成生物学一词，是由于在现阶段尚未真正实现“从头（from scratch）”合成一个新的生物体的合成生物学。而“合成”一词易被误解成仅仅是化学与工程的范畴。为了突出全基因组“设计”在这一领域的核心地位，故使用了“设计和合成”这两个术语的结合，而所有其他方面则概括成“基因组组成”。合理与接受程度如何，有待实践的检验及笔者和读者的共同努力。

第五部分 —— 基因组伦理学是基于笔者十几年的工作经历和体会，深感理解生命伦理学双重使命——为生命科学和生物技术“鸣锣开道”（生命伦理学要呼唤科学技术的创新）和“保驾护航”（生命伦理学要保证生命科学和生物技术的正确方向）的极端重要性。主要内容是笔者长期担任联合国教科文组织（The United Nations Educational, Scientific and Cultural Organization, UNESCO）的国际生命伦理委员会（International Bioethics Committee, IBC）与政府间生物伦理委员会（International Governmental Bioethics Committee, IGBC），以及联合国人权委员会（Commission on Human Rights, UN）伦理小组成员的所得和体会。

特色与不足

本书的特色也正是本书的不足和风险。

1) 重理念而不只是讲授知识。还记得巴德年先生有一次对我们青年教师说：“三流教授讲书本，二流教授讲知识，一流教授讲理念”（尽管本书绝非一流）。在人手一机（手机或计算机）的信息时代，所有知识都是“现成的”，几乎都在网上可以找到，很多人都有“无师自通”的能力，但这并不影响一本好的教科书的存在意义，而更加突出了“理念”的重要性。集笔者十几年之经历，学习基因组学——掌

握理念尤为重要。有理念才能矢志不移，才能掌握知识，才能举一反三，游刃有余。本书所有知识部分的介绍，都是为了证明并分享基因组学的理念；也是为了激发读者的兴趣，挖掘读者的知识潜能，培养读者的创新意识；同时也为正在构思的网络版留下更多的空间。正因为此，即使在“基因组的生物学”这样的重要部分中，我们也是惜墨如金，点到为止。

作为一门新学科，基因组学博大精深，涉及生命科学的方方面面。本书主要是供刚学完普通遗传学的本科生及基因组学相关专业的研究生作为基因组学的综合性入门教科书。其主要目的是在掌握理念的基础上，了解基因组学的基本要点。正因为如此，本书的正文仅有易读、好记的要点（约占全书篇幅的五分之一），其他拓展性的知识约占五分之四的篇幅；而为了顾及正文内容叙述的完整性与连贯性，全部以小一号的字体呈现。当然，本书的风险或许在此。笔者的初衷是宁失为“过浅”，也不愿“过深”而有悖于入门教科书。

2) 讲源流而不重复相关基础。基因组学的基础几乎涉及生命科学所有学科，特别是遗传学和分子生物学。正因为这样，许多国内外的教科书或专著，尽管以基因组学为题，但究其内容，一半以上的篇幅实际是地地道道的遗传学和分子生物学。本书忍痛割爱，作了风险最大的尝试——将它们基本删除或作“入框”处理，以突出基因组学的源流和精髓。而“框”的设计和撰文，则侧重历史性、相关性和趣味性，不要求读者全部仔细阅读，仅供使用时选择。据笔者四十余年各类不同的教学经验，本书特别重视专用名词或术语的准确定义及其内涵和外延，分析这一术语与另一相近术语的异同，而不泛泛而论使学生不得要领。正是出于这样的考虑，本书特别重视“术语学（terminology）”与基本概念。在书后另附“常用术语与缩写简释”，以补正文中说明之不足。也同样待议、待改及完善。

3) 论原理而不详谈技术细节。基因组学也是一门理论与实验相结合的学科。除了其核心技术——测序技术以外，还涉及很多其他生物和理化技术。本书重在介绍这些技术的基本原理，而不是描述其操作细节。更多的“入框”处理也体现“授人以鱼不如授人以渔”的理念。这里的“渔”，不只是渔具与鱼饵，而是有关“渔”的理念、战略（包括环境变化、天时气候、地域水势，各种“渔”的手段，如垂钓、撒网、拖网、“竭泽而渔”，以至这些手段的选择、改进与创新）。这也为将来可能由笔者自己或其他同事编写本书的配套用书——《基因组学技术》与《基因组信息学技术》留下空间。正是出于这样的考虑，本书用较大的篇幅介绍常用技术包括生物信息学分析技术的基本原理和操作要点，特别是在第二部分之中。

4) 倡甚解而不推荐过多文献。这或许是本书的另一风险。作为教科书，理应鼓励学生广泛阅读原始文献。但作为过来之人，深知学生们所修科目甚多、学习任务繁重。列举过多的参考文献，要求学生阅读并真正读懂并不实际，也会流于形式。真正精选必读文献也绝非易事。因此本书期望读者把主要精力放在对本书正文的甚解上，而不在正文中直接关联参考文献。但为了一定程度的平衡，将所有必要的参考文献，特别是历史性的经典文献，加以汉语主题词的提示，分篇分章列于全书之后，并不要求学生必读。

5) 抒己见而又博采百家。这是笔者的又一纠结。教科书与在课堂上授课一样，一定有也应该有自己的学术观点，这又一定涉及国内外历史上的一些敏感而现在已无必要的学术争议。一些颇多争议的中译名，例如“META 基因组”在本书中暂且直接使用英文缩写。而 epigenomics 则摒弃易于误导的“表观基因组学”，而采用“外饰基因组学”。本书在博采百家的同时，力求培养学生的分析能力，形成有自己独立见解的能力。

6) 论贡献而不分中外老幼。本书首开先河，大量引用了中国科研团队的工作，特别是配用了很多年轻中国人的面孔。乍一看本书，一些读者或颇有微词，但如果仔细阅读有关章节对他们工作的描述，一定会觉得非常自然——中外并蓄、“老幼无欺”。当然，所有的引用仅为例举，而不是全面、平衡的评估，“文责自负”。本书使用的照片数量，还远远比不上《基因的分子生物学》。我们明白这一尝试的风险，乐意在再版时更换。

此外，本书对一些教材编写的具体问题也作了小试。对于诸多英语人名，第一次出现时皆为英文全名，而后仅用英文姓氏。地名、机构名称也作相似处理。对于常用的中译专业名词，仅在第一次出现时附上英文全名和中译名。而常用的 HGP 等缩写词，则在第一次出现时缩写词在先，然后在括号中用英文及中

文全称，再次出现时就使用英文缩写词。

仿照《基因的分子生物学》已受广大读者欢迎的编排，我们也在每页的外侧（双页左侧，单页右侧）留出空白。以供学生在听课时作简略的笔记，复习时作每页的归纳或眉批。

尽管笔者和参与编写本书的全体人员已尽了很大努力，力求完美，宁晚勿滥，而必须承认的事实仍是眼高手低。作为首版，本书的各类问题比比皆是，每次再阅，不禁汗颜。请诸位读者、老师和同学严格要求，多予指正。我们仍将征求对本书勘误的修改和评论意见，既包括对本书的框架设计及有关章节中要点的描述，也包括选字、遣词、造句、插图等的具体修改意见，以及概念的更新、补充和对新进展的及时反映。正出于这一目的，本书曾在 2015 年 9 月 9 日先出“征求意见稿（基因组学 2015）”仅供内部交流和师友试用，以期抛砖引玉。本版正是在整合诸位师长的勘误与雅正的基础上，也根据笔者自己授课的体会，作了较大程度的修改。

有关本书的图、表，如仍有涉及知识产权等问题，请与出版社联系，我们将在再版时定作相应的删改。

杨焕明

2016 年 9 月 9 日

简 目

基因组学

第一部分 基因组学概论

- 第一篇 《基因组学》导读
第二篇 基因组学的发展史

第二部分 基因组学的方法学

- 第一篇 DNA 测序
第二篇 序列的组装和分析

第三部分 基因组的生物学

- 第一篇 人类基因组
第二篇 动物基因组
第三篇 植物基因组
第四篇 微生物基因组

第四部分 基因组的设计和合成

第五部分 基因组伦理学

GENOMICS

Part I. An Introduction to Genomics

- Chapter 1. A Guide to *GENOMICS*
Chapter 2. History of Genomics

Part II. The Methodology of Genomics

- Chapter 1. DNA Sequencing
Chapter 2. Sequence Assembly and Analysis

Part III. The Biology of Genomes

- Chapter 1. Human Genomes
Chapter 2. Other Animal Genomes
Chapter 3. Plant Genomes
Chapter 4. Microbial Genomes

Part IV. Redesign and Synthesis of Genomes

Part V. Bioethics in Genomics



目 录

第一部分 基因组学概论

第一篇 《基因组学》导读	3
第一节 基因组学的一般定义	3
一、基因组学和基因组	3
二、基因组学与其他学科的关系及分科	5
第二节 基因组学的两个理念	5
一、生命是序列的	5
二、生命是数字的	5
第三节 基因组学的三个源流	7
一、演化学说	7
二、细胞学说	8
三、基因学说	10
第四节 测序技术的四个突破	14
一、直读	14
二、自动化	14
三、规模化	14
四、MPH	15
第五节 基因组学发展的五大趋势	15
一、重绘“生命之树”	15
二、群体基因组分析	16
三、个体基因组分析	16
四、“跨组学”分析	17
五、基因组的生物学	17
第六节 基因组学应用的六个方面	17
一、外显子和全外显子组测序——单基因性状与遗传病	17
二、全基因组测序——复杂性状与常见疾病	18
三、单细胞测序——基因组异质性	18
四、META 基因组测序——微生物及病原基因组	18
五、微(痕)量 DNA 测序——无创检测、法医和古 DNA	19
六、“数据化”育种与生物条码	19
第七节 正在改变世界的七项技术	19
一、合成基因组学	19
二、基因组编辑技术	20
三、干细胞与 iPS 技术	22
四、动物克隆	23

五、大数据与生物库	24
六、表型组分析	24
七、“组学”与相关技术	25
第八节 基因组伦理学的八个方面 (HELPCESS)	25
一、人文 (H, Humanity)	25
二、伦理 (E, Ethics)	25
三、法律 (L, Law)	25
四、公众关系和决策 (P, Public-relationship/Policy-making)	26
五、文化 (C, Culture)	26
六、经济和教育 (E, Economy/Education)	26
七、生物安全和防护 (S, Safety and Security)	26
八、社会 (S, Society)	26
第九节 中国对基因组学的九大贡献	27
一、1% 国际人类基因组计划 (HGP)	27
二、10% 国际单体型图计划 (HapMap 计划)	28
三、国际千人基因组计划 (G1K 计划)	28
四、国际癌症基因组计划 (ICGP)	29
五、水稻和家蚕基因组计划	30
六、META 基因组	31
七、“生命之树”与动植物基因组	31
八、“第二代酵母”全基因组设计和合成	31
九、倡导“HGP 精神”和“合作”的文化	32
第二篇 基因组学的发展史	33
第一节 HGP 的起始和学术源流	33
一、HGP 的讨论和启动	33
二、HGP 的国际化	35
三、HGP 的学术源流	40
第二节 HGP 的目标和技术路线	42
一、HGP 的技术目标	42
(一) 遗传图	43
(二) 物理图	43
(三) 转录图	45
(四) 序列图	45
二、HGP 的技术路线	45
(一) 定位克隆霰弹法	46
(二) 全基因组霰弹法	46
第三节 HGP 中的模式生物	46
一、大肠杆菌	47
二、酿酒酵母	47
三、秀丽线虫	48
四、黑腹果蝇	49
五、拟南芥	50
六、河豚鱼	51

七、小鼠	51
第四节 HGP 的完成与后续计划	52
一、HGP 的完成	52
二、HGP 的后续计划	54
(一) 国际 HapMap 计划	55
(二) ENCODE 计划	56
(三) 国际千人基因组计划 (G1K 计划)	56
(四) 国际癌症基因组计划 (ICGP)	57
(五) 百万人基因组计划与十亿人基因组建议 (G1M 与 G1B)	58
(六) 其他国际合作计划	58
第五节 HGP 的意义和影响	61
一、创造了一种新的文化：合作	61
二、催生了一门新的学科：组学	61
三、提供了一个新的技术：测序	61

第二部分 基因组学的方法学

第一篇 DNA 测序	67
------------------	----

第一章 测序仪	67
---------------	----

第一节 测序技术的四个突破	68
一、第一个突破：直读	68
(一) 前直读法	68
(二) 直读法	69
1. SBC 法	70
2. SBS 法	71
二、第二个突破：自动化	74
三、第三个突破：规模化	77
四、第四个突破：大规模并行高通量测序	80
(一) 焦磷酸测序	81
1. 454	81
2. Ion Torrent/Proton	83
(二) 新一代 SBS	85
1. Illumina	85
2. BGISEQ-500	88
3. Max-Seq	90
4. LaserGen	91
(三) 连接测序	92
第二节 下一代测序仪及其基本原理	94
一、单分子测序	94
(一) HeliScope	94
(二) PacBio	94
二、纳米孔测序	97
(一) 物理纳米孔	98

(二) 生物纳米孔	100
三、杂交测序	101
四、显微测序	102
第二章 关键技术与方法	104
第一节 Southern 测序	105
一、测序策略	105
(一) 点测序	105
1. PCR 扩增	106
2. 质粒克隆	108
(二) 靶区域测序	109
(三) 外显子组测序	111
(四) 全基因组测序	112
(五) 单细胞测序	113
(六) META 测序	114
(七) Indexing	114
二、基因组概貌评估	115
(一) 评估参数	115
1. 基因组大小	115
2. 测序深度	116
(二) 评估方法	116
1. 基因组速览	116
2. K-mer 分析	116
三、测序的技术路线	120
(一) 定位克隆霰弹法	120
(二) 全基因组霰弹法	121
四、测序材料	121
(一) 常用材料	122
1. 细胞培养物和外周血细胞	122
2. 其他组织细胞	122
(二) 特殊材料	122
1. 单细胞或痕量 DNA	122
2. 石蜡包埋材料	123
3. 古 DNA 样本	123
4. META 样品	124
(三) DNA 的制备、纯化和质检	124
1. 盐析法：血液样本	125
2. CTAB 法：植物组织、真菌类样本	125
五、测序文库	125
第二节 Northern 测序	126
一、测序材料及 RNA 模板制备	126
二、RNA 的测序和分析	129
(一) 数字化表达谱	129
(二) 全转录组	130

1. 已有基因组参考序列的转录组分析	131
2. 无基因组参考序列的转录组分析	132
(三) 单细胞的转录组或表达谱分析	132
第三节 Western 测序	132
一、染色质免疫沉淀测序	133
二、核糖体保护	134
第四节 Eastern 测序	134
一、化学法	135
二、亲和法	136
三、酶切法	137
第五节 其他重要技术	138
一、GWAS	138
二、基因组编辑	140
三、芯片技术	142
(一) 基因表达芯片	142
(二) SNP 芯片	142
(三) CGH 芯片	142
(四) 突变检测芯片	143
(五) 捕获芯片	143
(六) 测序芯片	143
第二篇 序列的组装和分析	143
第一章 序列的组装	143
第一节 测序数据的质控和预处理	143
一、碱基读取	143
二、载体和接头序列的去除	146
第二节 序列的组装	146
一、基于参考序列的组装	147
二、 <i>de novo</i> 组装	148
(一) OLC 法	148
(二) <i>de Bruijn</i> 图法	150
(三) “穷举”法	152
三、多策略结合组装	152
(一) 复杂基因组	152
(二) 基因组精细图	153
四、组装的质控	154
第二章 基因组概貌分析	156
第一节 基因组大小	156
第二节 GC 含量	157
一、GC 含量和染色体带型	158
二、GC 含量和基因密度	160
三、GC 含量和 CpG 岛	161
第三节 重复序列	161

一、重复序列的分类	161
二、重复序列的识别	162
第四节 编码基因	164
一、编码基因的注释	164
(一) <i>de novo</i> 预测	164
(二) 同源比对预测	166
二、基因的功能注释	168
(一) InterProScan	168
(二) GO	168
(三) COG	169
(四) KEGG	169
第五节 假基因	170
第六节 非编码基因	171
一、rRNA	171
二、tRNA	172
三、miRNA	173
(一) miRNA 比对预测	174
(二) miRNA 结构预测	175
(三) miRNA 靶基因注释	176
第七节 基因组变异	177
一、单核苷酸多态性	177
二、拷贝数变异	180
三、其他结构变异	181
(一) 倒位	181
(二) 易位	181
四、个体杂合度	182
第三章 基因组比较分析	183
第一节 演化基因组	183
一、基因家族分析	183
二、系统发生分析	186
三、同源与直系同源基因	190
四、基因的选择和演化	191
五、多基因家族	193
六、转位因子	195
七、保守性区域及共线性分析	196
八、全基因组加倍	198
第二节 META 基因组	202
一、16S rDNA 分析	203
二、环境基因标签和 K-mer 分析	204
三、参考基因组集和共有核心细菌	205
四、核心基因家族	207
五、整合基因集构建	207
六、META 微生物组群与疾病相关性分析	208

七、MWAS	209
--------------	-----

第三部分 基因组的生物学

第一篇 人类基因组	214
-----------------	-----

第一章 人类基因组概貌	214
-------------------	-----

一、人类基因组大小	216
二、编码基因	218
(一) 编码基因的估计数目	218
(二) 基因序列、外显子和编码序列	219
(三) 编码基因的分布和密度	220
三、假基因	222
四、GC 含量	222
五、重复序列	222
六、CpG 岛和甲基化	223
七、非编码 RNA	223
(一) rRNA	223
(二) tRNA	224
(三) miRNA	224
八、线粒体 DNA 基因组	224

第二章 人类基因组的生物学	226
---------------------	-----

一、人类基因组多样性与“泛基因组”	226
二、演化与群体基因组	227

第三章 临床基因组学	228
------------------	-----

第一节 基于序列数据的检测技术	229
-----------------------	-----

一、无创产前检测	229
二、植入前检测	233
三、症状前检测	235
四、精准医学和 21 世纪的医学	237

第二节 单基因病及基因定位	240
---------------------	-----

一、单基因病	240
二、基因定位	241
(一) 遗传分析定位	241
(二) 细胞遗传学定位	242
(三) 原位杂交	243
(四) 外显子组测序	244

第三节 癌症基因组	244
-----------------	-----

一、癌症与基因组学	244
二、癌症的组学研究	246
(一) 全基因组 MPH 测序与癌症基因组分析	246
1. 点突变与 InDel	247
2. 结构变异	247

3. 基因扩增	247
(二) 单细胞组学技术与癌症基因组异质性	248
(三) 外饰基因组变异与癌症发生	249
(四) 癌症病原组	251
1. 癌症病原组	251
2. 细胞内致癌病毒的检测	252
三、癌症基因组学与精准医学	254
第四节 人体内微生物组群与疾病	256
一、肥胖	256
二、冠心病	257
三、肝硬化	257
四、免疫系统病	257
五、糖尿病	258
六、神经类疾病	258
七、临床干预	259
(一) 饮食的调整	259
(二) 调节物的使用	259
(三) 肠内含物的移植	259
八、其他人体共生微生物	260
九、人体病原组	260
第五节 长寿与老年病	261
一、长寿与衰老	261
(一) 长寿基因与代谢途径	261
(二) 长寿与端粒	262
(三) 长寿与外饰基因组	263
(四) 长寿与线粒体基因组	263
(五) 其他动物的长寿“秘诀”	263
二、老年病	266
(一) 阿尔兹海默病	266
(二) 帕金森病	266
(三) 老年性黄斑变性	267
第六节 法医基因组	267
一、前 DNA 标记	268
二、DNA 标记	268
(一) 第一代 DNA 标记: RFLP 和 AFLP	268
(二) 第二代 DNA 标记: VNTR 及 STR	270
(三) 第三代 DNA 标记: SNP	274
(四) 第四代 DNA 标记: 全基因组序列	274
第二篇 动物基因组	279
第一章 脊椎动物门	281
第一节 哺乳纲	281
一、灵长目	281
(一) 黑猩猩 <i>Pan troglodytes</i>	281

(二) 中国恒河猴 <i>Macaca mulatta lasiota</i>	281
(三) 食蟹猴 <i>Macaca fascicularis</i>	282
二、食肉目	283
(一) 大熊猫 <i>Ailuropoda melanoleuca</i>	283
(二) 北极熊 <i>Ursus maritimus</i>	284
(三) 家犬 <i>Canis familiaris</i>	285
(四) 猫 <i>Felis catus</i>	286
(五) 东北虎 <i>Panthera tigris</i>	287
三、偶蹄目和奇蹄目	287
(一) 家猪 <i>Sus scrofa</i>	287
(二) 双峰驼 <i>Camelus bactrianus</i>	290
(三) 牦牛 <i>Bos grunniens</i>	291
(四) 牛 <i>Bos taurus</i>	292
(五) 藏羚羊 <i>Pantholops hodgsonii</i>	292
(六) 家山羊 <i>Capra hircus</i>	293
(七) 绵羊 <i>Ovis aries</i>	294
(八) 马 <i>Equus caballus</i>	294
(九) 驴 <i>Equus asinus</i>	295
四、啮齿目	296
(一) 小鼠 <i>Mus musculus</i>	296
(二) 大鼠 <i>Rattus norvegicus</i>	297
(三) 中国仓鼠 <i>Cricetulus griseus</i>	298
五、食虫目	298
(一) 裸鼯鼠 <i>Heterocephalus glaber</i>	298
(二) 盲鼯鼠 <i>Spalax galili</i>	299
(三) 达马拉鼯鼠 <i>Fukomys damarensis</i>	300
六、鲸目	301
(一) 小须鲸 <i>Balaenoptera acutorostrata</i>	301
(二) 白鲸豚 <i>Lipotes vexillifer</i>	301
七、后兽亚纲双门齿目	302
(一) 短尾负鼠 <i>Mondelphis domestica</i>	302
(二) 袋鼠 <i>Macropus eugenii</i>	303
八、原兽亚纲单孔目——鸭嘴兽	303
第二节 鸟纲	305
一、鸡 <i>Gallus gallus</i>	307
二、北京鸭 <i>Anas platyrhynchos</i>	307
三、朱鹮 <i>Nipponia nippon</i>	308
四、鸽子 <i>Columba livia</i>	309
五、游隼和猎隼 <i>Falco peregrinus & Falco cherrug</i>	310
六、流苏鹬 <i>Philomachus pugnax</i>	310
第三节 爬行纲	311
一、中华鳖和绿海龟 <i>Pelodiscus sinensis & Chelonia mydas</i>	311
二、眼镜王蛇和缅甸蟒 <i>Ophiophagus hannah & Python molurus</i>	312
三、壁虎 <i>Gekko japonicus</i>	312

第四节	两栖纲——爪蟾	313
第五节	鱼纲	313
一、	斑马鱼 <i>Danio rerio</i>	313
二、	姥鲨 <i>Callorhynchus milii</i>	314
三、	河豚鱼 <i>Fugu rubripes</i>	314
第二章	脊索动物门——文昌鱼	315
第三章	无脊椎动物门	315
第一节	昆虫纲	315
一、	家蚕 <i>Bombyx mori</i>	315
二、	飞蝗 <i>Locusta migratoria</i>	317
三、	黑腹果蝇 <i>Drosophila melanogaster</i>	317
四、	大头切叶蚁 <i>Atta cephalotes</i>	318
第二节	蜘蛛纲	318
一、	丝绒蜘蛛 <i>Stegodyphus mimosarum</i>	318
二、	马氏正钳蝎 <i>Mesobuthus martensii</i>	319
第三节	无脊椎动物	319
一、	线虫动物门——秀丽隐杆线虫 <i>Caenorhabditis elegans</i>	319
二、	软体动物门	320
(一)	牡蛎 <i>Crassostrea gigas</i>	320
(二)	章鱼 <i>Octopus bimaculoides</i>	321
三、	扁形动物门——血吸虫 <i>Schistosoma japonicum</i>	321
四、	腔肠动物门——水螅 <i>Hydra vulgaris</i>	322
五、	棘皮动物门——海胆 <i>Stongylocentrotus purpuratus</i>	322
六、	缓步动物门——水熊虫 <i>Hypsibius dujardini</i>	323
第三篇	植物基因组	332
第一章	模式植物	334
一、	拟南芥 <i>Arabidopsis thaliana</i>	334
二、	其他芥类植物	334
(一)	琴叶芥 <i>Arabidopsis lyrata</i>	335
(二)	条叶蓝芥 <i>Thellungiella parvula</i>	335
(三)	盐芥 <i>Thellungiella salsuginea</i>	335
第二章	主粮类	336
一、	水稻 <i>Oryza sativa</i>	336
二、	谷子 <i>Setaria italica</i>	338
三、	玉米 <i>Zea mays</i>	339
四、	高粱 <i>Sorghum bicolor</i>	340
五、	小麦 <i>Triticum aestivum</i>	341
六、	大麦 <i>Hordeum vulgare</i>	343
七、	马铃薯 <i>Solanum tuberosum</i>	343
第三章	蔬菜类	344
一、	黄瓜 <i>Cucumis sativus</i>	344

二、白菜 <i>Brassica rapa</i>	345
三、甘蓝 <i>Brassica oleracea</i>	345
四、西红柿 <i>Solanum lycopersicum</i>	346
五、辣椒 <i>Capsicum annuum</i>	346
六、萝卜 <i>Raphanus sativus</i>	347
第四章 瓜果类	347
一、西瓜 <i>Citrullus lanatus</i>	347
二、甜瓜 <i>Cucumis melo</i>	348
三、香蕉 <i>Musa acuminata</i>	348
四、葡萄 <i>Vitis vinifera</i>	349
五、梨 <i>Pyrus bretschneideri</i>	350
第五章 经济类	350
一、大豆 <i>Glycine max</i>	350
二、木豆 <i>Cajanus cajan</i>	351
三、雷蒙德氏棉 <i>Gossypium raimondii</i>	352
四、木本棉 <i>Gossypium arboreum</i>	352
五、杨树 <i>Populus trichocarpa</i>	353
六、胡杨 <i>Populus euphratica</i>	353
七、毛竹 <i>Phyllostachys heterocycla</i>	354
第六章 花卉类	354
一、梅花 <i>Prunus mume</i>	354
二、兰花 <i>Phalaenopsis equestris</i>	355
第七章 单细胞植物类——甲藻	355
第四篇 微生物基因组	361
第一章 真菌	361
一、酿酒酵母 <i>Saccharomyces cerevisiae</i>	361
二、黄曲霉 <i>Aspergillus flavus</i>	362
三、米曲霉 <i>Aspergillus oryzae</i>	362
四、白色念珠菌 <i>Candida albicans</i>	363
五、稻瘟病菌 <i>Magnaporthe oryzae</i>	363
六、灵芝 <i>Ganoderma lucidum</i>	363
七、兔脑炎原虫 <i>Encephalitozoon cuniculi</i>	364
第二章 细菌	364
一、医学细菌	366
(一) 大肠杆菌 <i>Escherichia coli</i>	366
1. 大肠杆菌 K12	368
2. EHEC	368
(二) 流感嗜血杆菌 <i>Haemophilus influenzae</i>	369
(三) 鼠疫耶尔森菌 <i>Yersinia pestis</i>	369
(四) 幽门螺旋杆菌 <i>Helicobacter pylori</i>	371

二、工业细菌	372
(一)嗜热细菌	372
1. 腾冲嗜热菌 <i>Thermoanaerobacter tengcongensis</i>	373
2. 嗜热栖热菌 <i>Thermus thermophilus</i>	374
3. 超嗜热棒菌 <i>Pyrobaculum aerophilum</i>	375
(二) 枯草芽孢杆菌 <i>Bacillus subtilis</i>	375
三、农业细菌	375
(一) 农杆菌 <i>Agrobacterium fabrum</i>	375
(二) 苏云金芽孢杆菌 <i>Bacillus thuringiensis</i>	376
第三章 病毒	377
一、DNA 病毒	378
(一) HBV	378
(二) HPV	379
二、RNA 病毒	382
(一) HIV	382
(二) 登革热病毒	383
(三) 冠状病毒	384
1. SARS	384
2. MERS	385
(四) 流感病毒	385
1. H1N1	386
2. H7N9	386
(五) 布尼亚病毒	386
(六) Ebola 病毒	387
(七) Zika 病毒	387
第四章 环境微生物组群	387
一、海洋环境	388
二、土壤环境	389
三、极端环境	390
四、大气环境	391
五、工业环境	391

第四部分 基因组的设计和合成

第一节 合成基因组学和合成生物学	395
第二节 合成生物学的发展史	398
一、人工合成“前基因”生物材料	399
二、人工合成寡核苷酸和基因	399
三、人工设计和合成遗传装置	401
四、人工设计和合成代谢途径	402
五、人工设计和合成全基因组	402
第三节 合成生物学方法学	406
一、全基因组的设计和合成	406

(一) Synthia	406
(二) 第二代酵母基因组计划	409
二、基因组组分的设计与合成	411
(一) 代谢途径的设计和改建	411
1. 青蒿素合成途径的重新设计和改造	411
2. 吗啡合成途径的重新设计和重建	413
(二) 系统和装置的设计与组装	415
(三) 基因元件的设计和“生物砖”	416
1. 基本元件	416
2. 标准化元件——“生物砖”	417
(四) 遗传密码的改写与遗传物质的改造	421
1. 非天然氨基酸的导入	421
2. 非天然核苷酸的合成	422
第四节 合成生物学的应用及展望	425
一、合成生物学的应用实例	425
(一) 环境及作物监测系统	425
1. 芳香烃生物传感器	425
2. 砷离子生物传感器	425
(二) 疾病监控系统	426
(三) 生物能源	426
1. 丁醇的合成生物学生产	426
2. 生物石油	427
(四) 生物材料	427
(五) DNA 计算	428
(六) DNA 信息存储	429
二、合成生物学和人类未来的思考	430
(一) 改变人类对生存方式的思维和生物产业的格局	430
(二) 合成生物学与人类的未来	432
1. 工程细胞的全基因组设计	432
2. 合成生物学的碳基原料	432
3. “一大四位”的未来农业	433

第五部分 基因组伦理学

第一节 生命伦理学是基因组学的题中应有之义	437
第二节 基因组伦理学的八个方面	437
一、H: 将“人”字写在天上	437
二、E: 生命伦理是生命科学的准则	438
三、L: 有法可依、“无”法则立、违法必究	439
四、P: 科学决策与“鱼水之情”	440
五、C: 科学也是美丽的	440
六、E: “无形之手”与科学的未来	440
七、S: 防患于未然	441
八、S: 为了人类福祉与社会和谐	442

参考书目	443
参考文献	445
常用术语与缩写简释	478

第一部分 基因组学概论

今天已经到了这样的时代，几乎所有的生物学现象都难以逃脱“组学化”。在接下来的 25 年里，街头巷尾谈论的主要话题都将是“组学”，即便不只是唯一的话题。

——史蒂芬·福仁德 (2011)

Today, we've gotten to the point where almost no biological phenomenon can escape "omicsization", and within the next 25 years, omics will be the biggest, if not the only, game in town.

——Stephen Friend (2011)

我有幸有这样一个机会，让我的科学生涯从双螺旋这一步直接跨入 30 亿步* 的人类基因组……不尽快将它完成将是非常不道德的。

——詹姆斯·沃森 (1999)

* 指人类基因组的 30 亿对核苷酸

I would only once have the opportunity to let my scientific career encompass a path from the double helix to the three billion steps of the human genome ... It's essentially immoral not to get it done as fast as possible.

——James Watson (1999)



第一篇 《基因组学》 导读

第一节 基因组学的一般定义

一、基因组学和基因组

基因组学 (genomics) 的一般定义是，基因组学是研究基因组 (genome) 的科学。

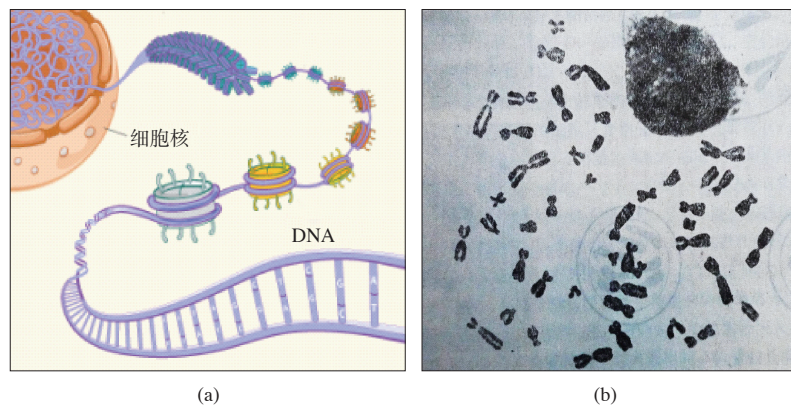


图 1.1 染色体 -DNA(a) 和中国医学科学院吴旻团队发表的人类外周血培养细胞的中期染色体 (b)

从术语学的角度，基因组学的这个一般定义，是基于“基因组”这一主体词派生的定义。而基因组本身的定义，则从不同学科的角度有不同的表述：从形式遗传学 (formal genetics, 即经典的孟德尔遗传学) 的角度，基因组是指一个生物体所有基因 (遗传和功能单位) 的总和；从细胞遗传学 (染色体遗传学) 的角度，基因组是指一个生物体 (单倍体) 所有染色体的总和 (如人类的 22 条常染色体和 X、Y 染色体)；从分子遗传学的角度，基因组是指一个生物体或一个细胞器所有 DNA 分子的总和，如真核生物的核基因组 DNA 分子和线粒体基因组 DNA 分子 (植物还另有叶绿体基因组 DNA 分子)，细菌的主基因组和数目不等的质粒 DNA 组分；现在有时还指某一特定生态环境样本中所有微生物 (microbiota) DNA 的总和；最重要的是，从现代信息学的角度，基因组是指一个生物体所有遗传信息的总和。

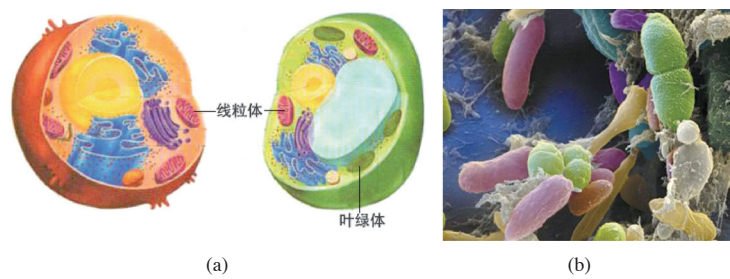


图 1.2 动植物细胞的线粒体和植物细胞的叶绿体 (模式图, a) 及人体肠道微生物组群 (电镜照片, b)

框 1.1

genome 和 genomics 的命名

Genome 是德国 Universität of Hamburg (汉堡大学) 的 Hans Winkles 于 1920 年将 gene 与 chromosome 两词组合而成的, 意为所有染色体上的全部基因; 而 genomics 则是美国 Johns Hopkins University (约翰·霍普金斯大学) 的 Victor McKusick 等在 1987 年为一个新杂志命名时提出的。

Gene 一词源于古希腊语 γένος (génos), 意为“种, 后代”。要注意的是, Gene (基因) 一词于 1909 年由丹麦植物学家 Wilhelm Johannsen 首次提出, 以取代孟德尔的“factor (因子)”等用语。而 genetics (基因学, 即遗传学) 一词首先是英国遗传学家 William Bateson 于 1905 年开始使用的。

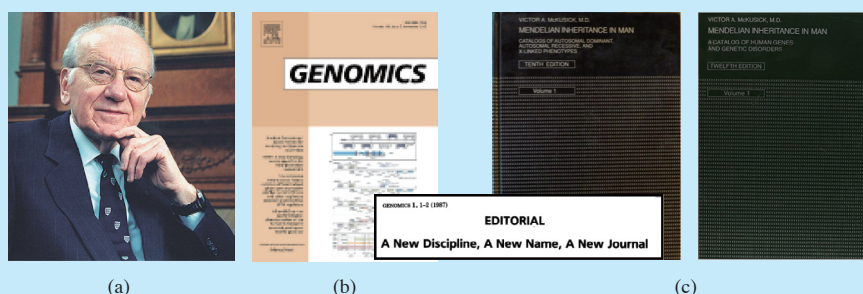


图 1.3 Victor McKusick (a) 和 1987 年创刊的 *Genomics* 杂志 (b) 及他的名著 *Mendelian Inheritance in Man* (c)

框 1.2

遗传学和基因组学的异同

遗传学是学习基因组学的基础。在一定意义上, 基因组学是遗传学的继续和发展, 是基因组层次和规模的遗传学。

基因组学和一般意义上的遗传学的相同之处, 是两者都以基因和其他遗传的功能因子为研究对象; 不同之处在于遗传学一般研究的是一个或少数几个基因, 而基因组学则是以一个生命体的所有基因和所有遗传的功能因子的自然存在单位为研究对象。也可以说, 基因组学是既有全基因组的规模和广度, 又有分子即核苷酸水平的深度的遗传学。

同遗传学从未分成“结构遗传学”和“功能遗传学”一样, 基因组学更加强调结构和功能密不可分的相互关系, 而不加以人为割裂与强化。也正是因为这样, 基于遗传学和基因组学的这一历史性和科学上的亲缘关系, 遗传学所有外延、派生的学科, 几乎毫无例外地已经开始并逐步升华为与基因组学对应的有关学科。

例如, 研究人类基因的遗传和变异的人类遗传学, 研究人类疾病发生和发展的遗传机制的医学遗传学, 着重研究遗传学在疾病诊断和治疗方面应用的临床遗传学, 研究基因和癌症相关性的癌症遗传学, 研究人类的演化与“生命之树”上其他生物亲缘关系的人类演化遗传学, 研究人群的基因组特点、基因变异、迁移和分布特点的人类群体遗传学等, 都已取其所需、与时俱进, 汲取了基因组学的精华, 部分或几乎全部升华为人类基因组学、医学基因组学、临床基因组学、癌症基因组学、人类演化基因组学、人类群体基因组学。作为学科主要特点而永远保留的只有遗传学的精髓, 那就是, 研究生命的遗传和变异 (稳定性和变异性)、连接基因型 (genotype) 和表现型 (phenotype, 简称表型)。

因此, 要学好基因组学这门科学, 需要掌握形式遗传学、细胞遗传学和分子遗传学的基础知识和基本技能, 以及现代信息学的基本概念和生物信息学的必要工具, 特别是要有从相关的重要网站获得信息并使用信息分析工具的基本技能。此外, 因为生物科学的特殊性, 学习基因组学, 还要学习生命伦理和生物安全相关问题的基本概念。

二、基因组学与其他学科的关系及分科

基因组学是生命科学所有学科的基础，又是 21 世纪生命科学的前沿和新的起点。基因组学是生命科学中最为年轻、最为活跃、进展最快的领域。

基因组学与遗传学一样，可依研究对象不同分为人类基因组学、动物基因组学、植物基因组学及微生物基因组学等亚学科，也可依研究策略、技术的不同分为比较基因组学和演化基因组学等。从研究的内容来说，主要分为基因组的概貌、基因组的生物学和基因组学的应用三个方面。

基因组学对生命科学其他相关学科最重要的影响和贡献是提供了“一组(-ome)”和“一组学(-omics)”的概念、策略和技术。

“组”与“组学”就是把生命科学几乎所有学科都“组化”和“组学化”，正像历史上的所有新生学科一样。

框 1.3 “一组(-ome)”和“一组学(-omics)”示例

“一组(-ome)”	“一组学(-omics)”
基因组 (genome)	基因组学 (genomics)
转录组 (transcriptome)	转录组学 (transcriptomics)
蛋白质组 (proteome)	蛋白质组学 (proteomics)
代谢组 (metabome)	代谢组学 (metabonomics)
调控组 (regulome)	调控组学 (regulomics)
表型组 (phenome)	表型组学 (phenomics)
甲基化组 (methylome)	甲基化组学 (methylomics)
组蛋白修饰组 (histone-modifiome)	组蛋白修饰组学 (histone-modificomics)
RNA 组 (RNAome)	RNA 组学 (RNAomics)
非编码 RNA 组 (non-coding RNAome, ncRNAome)	非编码 RNA 组学 (non-coding RNAomics, ncRNAomics)
外饰基因组 (epigenome)	外饰基因组学 (epigenomics)
病原组 (pathogenome)	病原组学 (pathogenomics)

还有一些几乎完全基于基因组学概念和技术的新学科，如综合研究生态、环境和群体微生物的 META 基因组 (metagenome) 和 META 基因组学 (metagenomics，类似的术语 hologenomes，意为宿主基因组与肠道微生物组群基因组之和)，研究脑科学的连接组 (connectome) 和连接组学 (connectomics)。也正是如此多样而又相互关联，本书的内容不限于基因组学，其内容涉及很多其他“组学”，特别是转录组学、蛋白质组学、表型组学、外饰基因组学和 META 基因组学等，即所谓“跨组学”或“贯穿组学 (trans-omics)”，简称“组学 (omics)”。

第二节 基因组学的两个理念

基因组学有两个最主要的理念：

一、生命是序列的！(Life is of sequence!)

“生命是序列的”源于 Watson 和 Crick 于 1953 年提出的一个论点：“……碱基的精确序列是携带遗传信息的密码。”（“…the precise sequence of the bases is the code which carries the genetical information.”）[Watson J D & Crick F H. *Genetic implications of the structure of deoxyribonucleic acids*. *Nature* 171: 964, 1953.]

二、生命是数字的! (Life is digital!)

“生命是数字的”来自 Sulston 在 2002 年所讲的一段话“……代代相传的生命指令是数据的。而不是模拟的……”(“…the instructions for making a life from one generation to the next is digital, not analogue…”)[Sulston J & Ferry G. *The common thread: A story of science, politics, ethics, and the human genome*. Joseph Henry Press, 2002.]

这两个理念是从基因组学的角度对生命和生命世界的理解,是基因组学的基石和支柱。

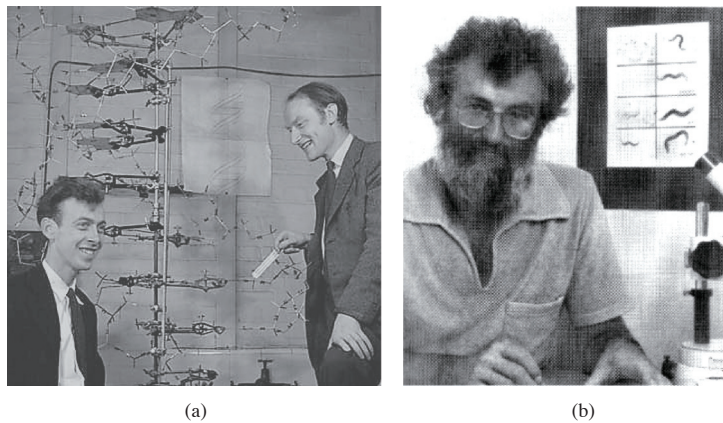


图 1.4 奠定基因组学理论基础的 James Watson (左) 和 Francis Crick (右) (a) 与 John Sulston (b)

基于这两个理念,现阶段的基因组有两项主要的核心技术:序列测定(sequencing, 测序)与基因组信息学(genome bioinformatics)。可以说,基因组学就是把生命科学“序列化”和“数字化”。测序(包括 DNA、RNA、甲基化组等测序)旨在“拿到”生命的这本“天书”,基因组信息学分析就是要借助计算机和相关软件结合表型分析来“解读”这本“天书”。

框 1.4

基因组学的两个理念和支柱

“生命是序列的”,准确的理解是:生命之所以为生命(遗传,生命的连续性),生命的遗传性改变(变异,生命的多样性)的所有信息,都蕴藏在 DNA 的核苷酸序列之中。要注意的是,外饰基因组学作为基因组学的一个分支,丰富了这一理解:遗传信息蕴藏在 DNA 的核苷酸序列以及不同方式修饰的核苷酸之中。

“生命是序列的”,并不排除任何新的重要发现或突破,更不能把 DNA 序列误解为只是基因组学的“结构”,DNA 序列的解读是基因组功能研究的重要部分。

“生命是数字的”是基因组学的支柱,连接了 21 世纪影响最大的两门科学——生命科学与信息科学,特别是当今世界的重要趋势——数字化。如果把生命的基本语言——A/T/C/G 与信息科学的基本语言——0/1 并排放在一起,就会豁然开朗:生命在本质上就是数字的。

ATCTACAACGTTATCGTCACAGCCCATG	1001000100101000101010001011
CATTGTAAATATCTTCTCATAGTAAT	0001000001000100010001001001
ACCATTTCACTATCATATTTCATCGGCGT	0001001000010010101000101010
AAATCTAACTTTCTTCCCACACACATTT	1000001110000011111111000010
CTCTGAAGCTTCACCGGCGAGTCATTC	1001010101010000100010100101
TCATAGTCGCCACGGACTTACATCCTC	0101010101010001000100101010
ATTACTAACAGACCGCAACCTCAACAC	1110001111000010101010101010
CACCTTCTTCGACCCGCGCGAGGAGGA	1010010101010101010101000000
GACTGCAAAACCCCACTCTGCATCAACT	0000001010100100111111000001
GAACGCAATCAGCCACTTTAATTAAGC	1010101000010000001111111111
TCATTAGCAGGAATACCTTTCTCACAG	1111100000100101010101010010
GTTTCTACTCCAAGACCACATCATCGA	1000001000111110000100010101

生命的现象太复杂了，导致生物学自成为科学以来，描述性的发现远远多于规律性的发现。“生命是数字的”，是通向新的规律性发现的重要方向与必经之路，是对生命本质和规律的新的发现。

框 1.5 数字化 (digitalization) 和全球化 (globalization)

20 世纪末至 21 世纪初，生命科学的发展在很大程度上受惠于当今世界的两大潮流——数字化和全球化。

首先，数字化使实验试剂和其他经典实验材料跨越了地理与物理的时空隔离，为 HGP 的全球化提供了可能性。早期最典型的例子便是遗传学最重要的试剂——DNA 探针，通过互联网远距离输送序列信息而取代了物质（如含“DNA 探针”的细菌克隆或 DNA）的转运。当今的实例之一是，一个病原尚未“入境”，它的基因组信息早已到达，足以启动疫苗及抗体制备或合成模拟，而严阵以待了。

其次，数字化和全球化使读取、储存、传输、分析庞大的基因组序列信息成为可能，使基因组学的全球性合作和数据分享成为可能，也使基因组学有幸成为人类科学史上合作最为“全球化”的范例。

更重要的是，人类的复杂表型（如所有的临床数据，包括影像）的数字化分型 (phenotyping，即表型分析)，使生物库 (BioBank) 和基因分型 (genotyping) 成为可能，也使生命科学与其他学科一样进入了“大数据 (Big Data)”的新纪元。

第三节 基因组学的三个源流

基因组学的发展，是生命科学发展的必然趋势与自然产物。

生命科学史上三大重要的规律性发现，即演化学说、细胞学说与基因学说，是现代生物学的基础，也是基因组学的源流。

一、演化学说

演化 (evolution) 是生命最重要的特点，演化学说是现代“生命科学的灵魂和基石”。

框 1.6 evolution 之汉译

本书之所以将 evolution 翻译为演化，一是承继严复先生的初译，二是听取了很多同行的建议。最直接的考虑是“进化”易被误解成生物的演化过程只能向一个方向“进步”。

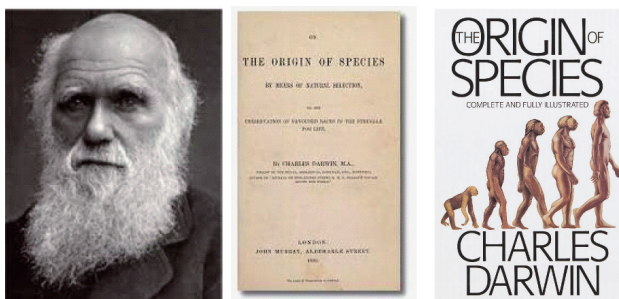
现代演化学说至少包括以下几个方面：

(1) “同祖同宗”——所有的生物都由演化而来，而基因组序列则是所有生命体演化的“活化石”，古 DNA 基因组学则为生命演化提供了最直接、最有说服力的证据。

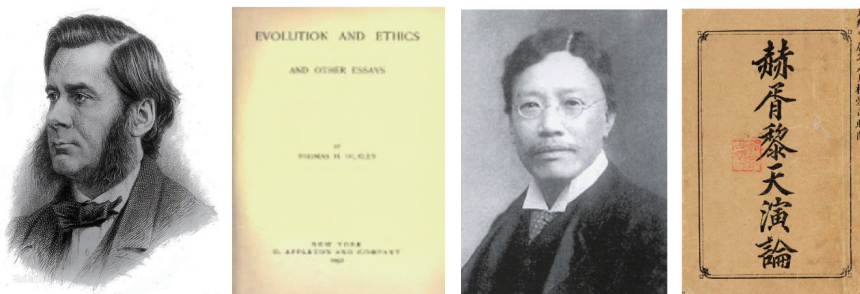
(2) “变异为源”——演化的原始动力是基因组 DNA 的自发变异，基因组序列的分析就是重建数字化的“生命之树”。由于多数变异是中性的，因而基因组变异事件及其发生的时间，可以根据基因组序列变异估算。

(3) “选择定向”——自然选择是有方向的，与表型相关的基因组区段或基因的被选择（正选择或负选择）的信息，是可以通过基因组序列的分析得到。

现代演化生物学还有一个十分重要的任务，就是以前所未有的序列大数据和新技术来解释物种起源、环境选择以及新表型的产生，阐明生命的连续性和多样性及其成因，使演化学说真正成为“生命科学的灵魂和基石”。

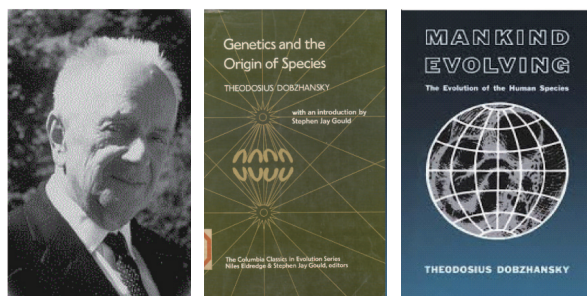


Charles Darwin 和他的 *The Origin of Species*



Thomas Huxley 和他的 *Evolution and Ethics*

严复和他的汉译《赫胥黎天演论》



Theodosius Dobzhansky 与他的两篇名著

图 1.5 演化学说的奠基人及著作

框 1.7

Theodosius Dobzhansky

Dobzhansky, 美国生物学家（出生于俄国），现代演化论的奠基人之一。他在科学上的最重要贡献是，在一版再版、经久不衰的《遗传学与物种起源》(*Genetics and the Origin of Species*, 1937、1951、1953) 一书中实现了遗传学与自然选择学说的完美结合，继承和发展了演化学说。此外，他还提出了“人类的演化不能理解作为一种纯生物学的过程，也不能完全描写成一部文化史。它是生物学和文化史的相互作用”。他的一段话：“离开了演化的观点，生物学的一切都毫无意义。(Nothing in biology makes sense except in the light of evolution.)”已成为生物学中流传最广的名言之一。

二、细胞学说

细胞学说的核心是，细胞是一切生物（除了前细胞生物）基本的结构单位与功能单位。

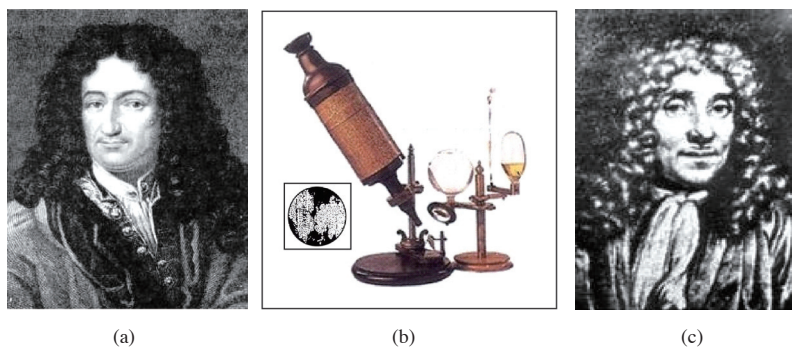


图 1.6 Robert Hooke (a) 和他所用的显微镜 (b) 及 Antony Leeuwenhoek (c)

1665 年, 英国的 Robert Hooke 用自己发明的显微镜发现了“细胞”(软木片的“小室”)。1675 年, 荷兰的 Antony Leeuwenhoek 也用自己设计的显微镜观察到了完整的动物细胞。



图 1.7 光镜与电镜示例

所有生物体都是由细胞组成的。所有生物细胞都来源于始祖细胞, 细胞分裂、分化和死亡(老化和凋亡)是所有生物体发生、生长与发育、衰老和死亡的基础。细胞学说是 1838~1839 年由德国植物学家 Matthias Schleiden 和动物学家 Theodor Schwann 最早提出, 到 1858 年德国病理学家 Rudolf Virchow 以“一切细胞来自细胞”的著名论断而完善的。

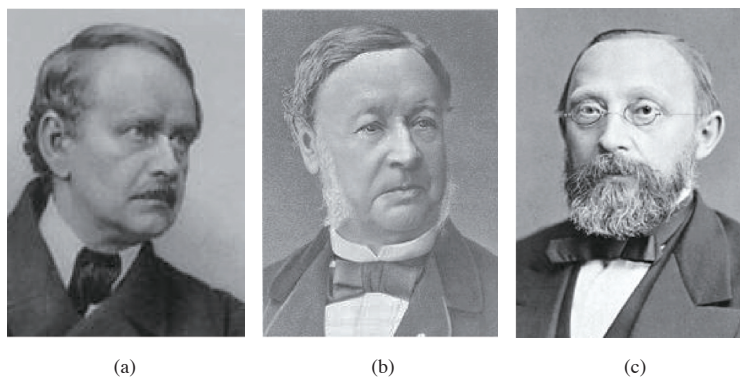


图 1.8 细胞学说创始人

Matthias Schleiden (a)、Theodor Schwann (b) 和 Rudolf Virchow (c)

框 1.8

基因组学与细胞学

细胞学有助于理解基因组的另一定义: 基因组(严格来说, 核基因组)是遗传物质/信息的自然存在单位。一个基因从来都不能单独遗传和行使功能, 只是所在基因组的一个组分。指导细胞活动的所有指令以及细胞对环境的反应, 都来自基因组。细胞分裂的过程就是基因组 DNA 分子复制的过程; 细胞分化的过程就是基因组指导相关基因在特定时空和特定环境条件下的表达和互作。基因组技术的近期重要进展之一就是单细胞 DNA 和 RNA 分析技术。

三、基因学说

基因学说开始于 20 世纪初的 Gregor Mendel 遗传学定律的再发现和 30 年代 Thomas Morgan 的基因论，发展自经典遗传学和分子遗传学的“联姻”。

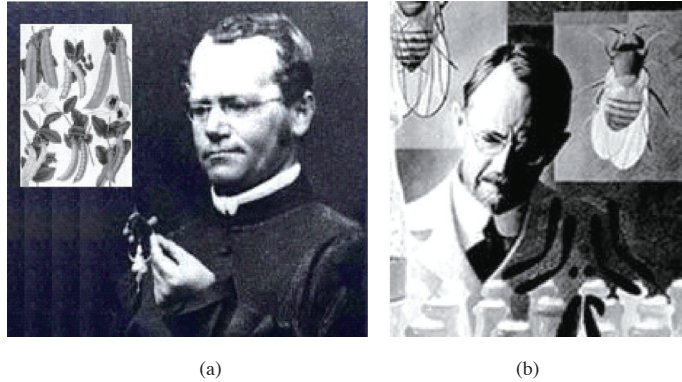


图 1.9 基因学说创史人 Gregor Mendel (a) 和 Thomas Morgan (b)

框 1.9

Mendel、Morgan 和遗传学三大定律

Mendel 出生于奥地利的 Brunn (现捷克)，从 1856 年到 1863 年，他通过 8 年的豌豆杂交实验，发现了遗传的基本规律而奠定了现代遗传学的基础。

Mendel 曾把他的研究成果寄给提出“颗粒说”的 Carl Nägeli。但是 Nägeli 对 Mendel 的发现不予重视，认为这些发现是“依靠经验而不是依靠理智的”。

Mendel 于 1865 年在奥地利自然科学研究协会上报告了研究结果，1866 年又在该会会刊上发表了《植物杂交实验》(*Experiments in Plant Hybridization*) 的论文。他在这篇论文中提出了“遗传因子”(现称 gene) 及显性性状、隐性性状等重要概念，并阐明其遗传规律。

直到 1900 年，这些规律才由 3 位植物学家：荷兰的 Hugo de Vries、德国的 Carl Correns 和奥地利的 Erich von Tschermak 通过各自的工作分别予以证实，而 Mendel 也被公认为经典遗传学的奠基人。有趣的是，由于 Mendel 有关“3:1 分离定律”的实验数据太准确了，以致于有“先有理论，后有数据”之嫌。最后，著名的统计学家 Fisher 证明 Mendel 的实验数据千真万确！

Morgan 及其团队在 1925 年发现果蝇有 4 对染色体，鉴定了约 100 个不同的基因，并以交配实验测量染色体位点之间的距离。1911 年提出了“染色体遗传理论”，后又提出了连锁和交换定律，被称为遗传学的“第三定律”。1926 年出版《基因论》(*The Theory of the Gene*) 一书，1933 年获诺贝尔生理学或医学奖。

20 世纪 40 年代的细菌转化实验和噬菌体转导实验，为“DNA 是遗传物质”提供了最重要的确凿证明，而真正的革命性突破是 1953 年发表的 Watson-Crick 模型即 DNA 双螺旋结构模型。60 年代中期解读的遗传密码，是人类解读的第一种“自然语言”，随后建立的阐明了表明遗传信息流向的“中心法则”，奠定了遗传学和整个生物学的分子和信息学基础。70 年代 DNA 克隆等遗传工程和分子生物学技术的发展，迎来了生物产业的第一个春天。