

Genetics of Complex Human Diseases:
A Laboratory Manual



人类复杂疾病 遗传学实验指南

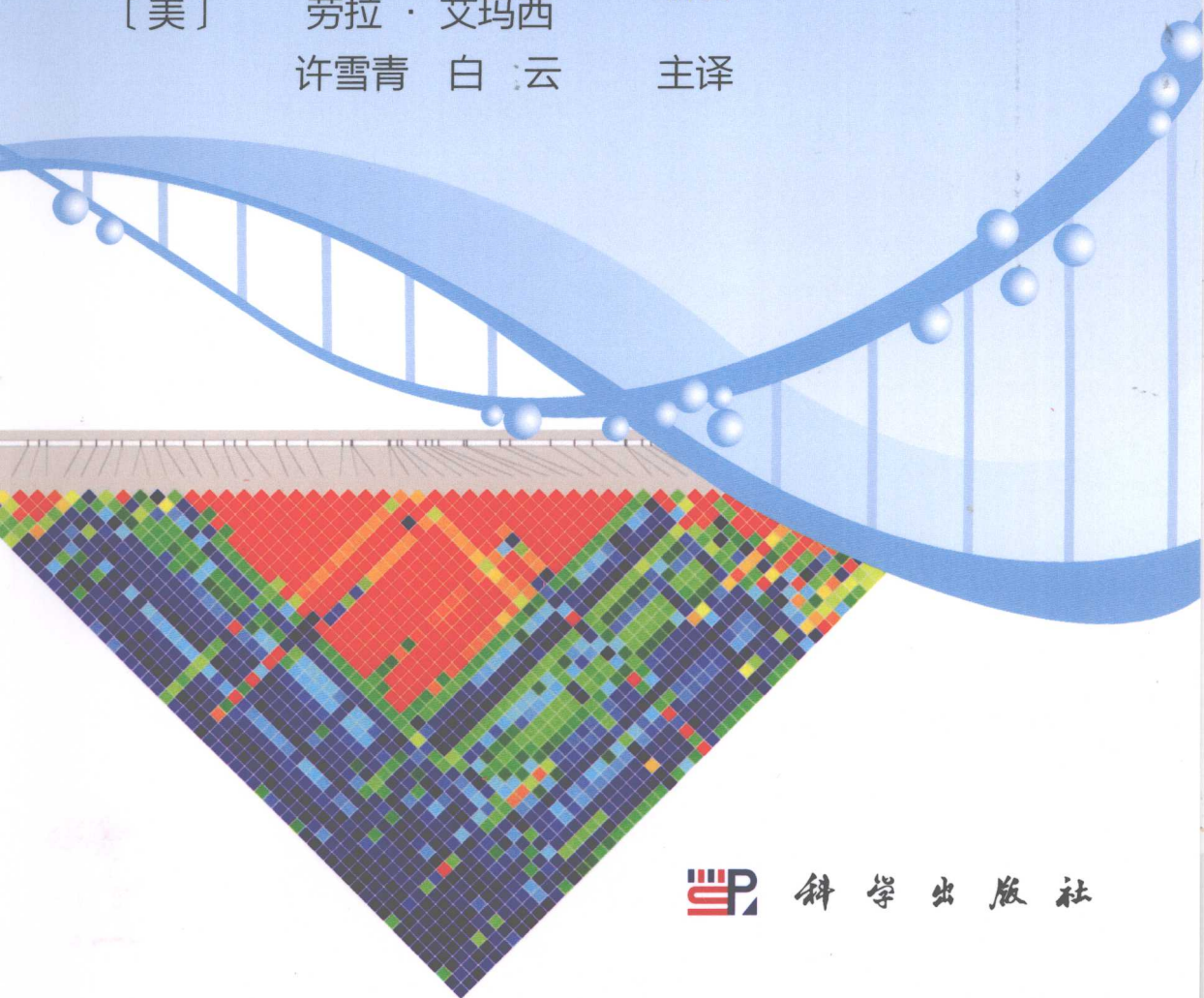
〔英〕 阿玛尔·阿尔沙拉比

编著

〔美〕 劳拉·艾玛西

许雪青 白云

主译



科学出版社

人类复杂疾病遗传学实验指南

Genetics of Complex Human Diseases:
A Laboratory Manual

[英] 阿玛尔·阿尔沙拉比 编著
[美] 劳拉·艾玛西
许雪青 白云 主译

科学出版社

北 京

图字: 01-2013-1881 号

内 容 简 介

本书的主要内容来自冷泉港实验室举办的人类复杂疾病遗传学课程,内容汇集了当前国际上遗传学工作者寻找致病基因的最新研究方法和这些方法背后的遗传学概念与统计学理论。既有后基因组时代以 GWAS 等为代表的热点研究领域的新方法,同时又以发展的观点介绍了连锁分析等经典的研究手段。具体内容包括基本遗传学与孟德尔遗传、统计方法、遗传流行病学、连锁研究、传递不平衡检验分析、变量成分分析、全基因组关联研究、拷贝数变异、高通量基因分型技术、RNA 编辑复杂性,以及遗传学计算机程序。

本书适合生物、医学等相关专业院校的师生阅读参考。

Originally published in English as Genetics of Complex Human Diseases: A Laboratory Manual, edited by Ammar Al-Chalabi, Laura Almasy © 2009 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA
© 2015 Science Press, Printed in China.

Authorized Simplified Chinese translation of the English edition © 2009 Cold Spring Harbor Laboratory Press. This translation is published and sold by permission of Cold Spring Harbor Laboratory Press, the owner of all rights to publish and sell the same.

图书在版编目(CIP)数据

人类复杂疾病遗传学实验指南/(英)阿尔沙拉比(Al-Chalabi, A.), (美)艾玛西(Almasy, L.) 编著;许雪青等译. —北京:科学出版社, 2015. 6

ISBN 978-7-03-044600-8

I. ①人… II. ①阿… ②艾… ③许… III. ①医学遗传学-实验-指南 IV. ①R394-33

中国版本图书馆 CIP 数据核字(2015)第 126191 号

责任编辑:李悦 孙青 / 责任校对:郑金红
责任印制:徐晓晨 / 封面设计:北京铭轩堂设计公司

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>

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2015 年 6 月第 一 版 开本: 720×1000 1/16

2015 年 6 月第一次印刷 印张: 16 1/4

字数: 330 000

定价: 98.00 元

(如有印装质量问题, 我社负责调换)

《人类复杂疾病遗传学实验指南》译校人员名单

(按姓氏音序排列)

白 云	陈雪丹	戴礼猛	符胜煜
郭 洪	黄 钢	雷佳凡	李发科
李俊霞	刘 丹	刘 波	涛孟慧
宋 敏	王 丰	王 凯	王 燕
王艳艳	徐小峰	许雪青	章 波
张 坤	张明杰		

译者序

该书原著是冷泉港实验室出版社 *A Laboratory Manual* 系列丛书中的一部,其主要内容也被编入 *Cold Spring Harbor Protocols* 系列丛书。能同时被冷泉港出版社纳入其久负盛名的两大研究方法系统丛书,足以表明本书在复杂疾病遗传学研究中的权威性和重要性。

研究人类疾病是医学遗传乃至整个生命科学研究的重要内容。人类常见疾病,如高血压、糖尿病、肿瘤、阿尔茨海默病等都因为太复杂而缺乏有效的分析手段。对这类复杂疾病的研究一直是遗传学和临床医学关注的重点和难点。

复杂疾病的传统经典研究方法是连锁分析和关联分析。在人类基因组计划之后,后基因组时代主要侧重从基因组的角度来研究生命现象,由此产生了以全基因组关联分析(genome-wide association studies, GWAS)为代表的一些新的研究方法。正是由于这些方法的出现,使得对复杂疾病的遗传分析已成为当前的研究热点。短短几年内,利用这些思路手段,人类已经在阿尔茨海默病、乳腺癌、糖尿病、冠心病、肺癌等一系列复杂疾病中找到疾病相关的易感基因。可以预见,对更多种类的复杂疾病,以及对复杂疾病在不同人群(尤其是人类世界最大的群体——汉族)中的遗传学分析将在相当长一段时间内都是医学遗传学和相关临床研究的重要内容。这些研究将为全面系统研究复杂疾病的遗传因素提供新的动力,也将为我们了解这些常见疾病的发病机制、早期诊断、开发有效防治措施等研究提供更多的线索。

该书汇集了当前国际上遗传学工作者寻找致病基因的最新研究方法以及这些方法背后的遗传学概念与统计学理论。既有后基因组时代以 GWAS 等为代表的热点研究领域的新方法,同时又以发展的观点介绍了连锁分析等经典的研究手段。具体内容包括基本遗传学与孟德尔遗传、统计方法、遗传流行病学、连锁研究、传递不平衡检验分析、变量成分分析、全基因组关联研究、拷贝数变异、高通量基因分型技术、RNA 编辑复杂性,以及遗传学计算机程序。书中各章节均由各领域著名研究者撰写,每个主题均包括实践信息及相关分析评述,力图阐明如何进行复杂疾病分析,以及具体进行分析的原因。

我们相信该书的翻译出版将有助于传播复杂疾病这一研究领域的新知识、新方法,提高相关科研工作者及广大研究生的研究分析能力,进一步促进我国复杂疾病遗传学的研究水平。此外,对提升相关研究生教学水平,更新研究生教学内容都会有很好的帮助。

目前我国还没有相关专著,即便该书原著也是在冷泉港学习班讲义基础上精选而来。翻译这本圭臬之作对译者而言是一个巨大的挑战,需要学习理解大量的前沿知识,众多新名词需要反复斟酌。好在参译人员都是学习工作在本领域第一线的青年科技工作者,其中多名具有海外学习经历,其科学研究方向与该书很多内容直接相关。最终在大家共同努力之下该书的翻译工作得以完成。但即便这样,不免有一些理解和翻译不尽如人意之处,还望读者不吝指出。

译 者

2015 年 3 月

前 言

这本书的主要内容来自冷泉港实验室举办的人类复杂疾病遗传学课程。该课程自 2004 年来每两年在冷泉港举办一次。聆听统计学和复杂疾病遗传学国际前沿大专家的授课,见证随着研究进展而完善的课程是一件让人感觉奇妙的事情。本书试图反映该课程的精华,而不仅仅是课程材料一模一样的拷贝。本书所包含的章节由分工良好的具有清晰教学风格的作者所写,我们期望它们同样具有冷泉港教程的特色:既为想知道结论后面的因果缘由的研究者提供细节,同时也为初入该领域的新人提供概况。

毫无疑问,如果没有众多人的鼎力相助,这本书是不可能完成的。感谢本书的作者们,他们将复杂的内容讲述得深入浅出,做得非常出色。也要感谢课程的演讲者,他们每两年奉献自己的时间将知识传授给他人。我们还要感谢在本书出版过程中给予指导的所有冷泉港实验室出版社员工:出版人 John Inglis,编辑主任 Alex Gann,项目协调人 Mary Cozza 和发展编辑 Judy Cuddihy,以及出版职员 Kathy Bubbeo、Rena Steuer、Susan Schaefer 和 Lauren Heller,没有他们的努力,这本书不可能面世。

Ammar Al-Chalabi

Laura Almasy

2009

目 录

译者序

前言

1 绪言	1
1.1 为什么遗传学重要	1
1.2 现代遗传学简明史	1
1.2.1 紧跟遗传学思想	1
1.2.2 孟德尔、达尔文以及遗传的波-粒辩论	2
1.2.3 分子生物学的中心法则	2
1.2.4 DNA 和遗传密码	2
1.2.5 基因组学时代的来临	2
1.2.6 后基因组时代	3
1.3 复杂疾病研究如何融入遗传学	3
1.4 最后感想	3
参考文献	4
2 “统计学 101”——人类复杂疾病遗传学的初级指南	5
引言	5
2.1 数集理论基础	5
2.2 遗传学分析中的概率理论	6
2.2.1 条件概率	6
2.2.2 独立性	6
2.3 遗传学分析中的变量、参数和分布	7
2.3.1 离散均匀分布	8
2.3.2 正态分布	8
2.3.3 卡方分布	9
2.3.4 二项分布	9
2.4 最大似然估计	10
2.5 假设检验为结果提供置信水平	11
2.5.1 p 值	11
2.5.2 似然比、似然比检验统计量, 以及优势对数计分法	12
2.5.3 效能	13
2.6 总结	14

参考文献	14
3 分离性状的连锁分析	16
引言	16
3.1 连锁	16
3.2 疾病表型的连锁分析	19
3.3 多位点连锁分析	21
3.4 遗传模式判断错误的后果?	22
3.5 血缘同一性的非参数连锁分析	23
3.6 LOD 值的意义	24
3.7 复杂疾病连锁分析中的遗传异质性	24
3.8 连锁分析在全基因组关联研究时代的价值	25
参考文献	26
互联网信息	27
4 复杂疾病遗传中的流行病学因素	28
引言	28
4.1 关联研究的实施	28
4.1.1 管理	28
4.1.2 设计选择	29
4.1.3 标记选择	29
4.1.4 样本选择	30
4.1.5 样本大小	31
4.1.6 随机误差	31
4.1.7 偏倚	31
4.1.8 暴露的变化	32
4.1.9 质控	32
4.2 关联研究的解释	32
4.2.1 理解因果	32
4.2.2 因果关系,修饰效应,混杂	33
4.3 大规模数据库	34
4.4 整合的“取代的”研究	36
4.5 结语	36
参考文献	37
互联网信息	38
5 复杂表型的方差组分分析方法	39
引言	39
5.1 什么是方差组分分析方法?	39

5.2	估计遗传率	40
5.3	环境共同效应的处理	41
5.4	采用测定的环境参数作为协变量	42
5.5	协变量选择对方差组分分析的影响	42
5.6	使用易感性阈值模型	43
5.7	连锁分析与方差组分的应用	44
5.8	对研究选择进行确认的重要性	45
5.9	非正态性的处理	45
5.10	多重分析与基因多效性	46
5.11	数量性状的关联分析	47
5.12	在方差组分分析框架下的基因与基因、基因与环境的相互作用	47
5.13	鉴定潜在的功能突变体	49
5.14	小结与结论	50
	参考文献	50
6	遗传关联研究中的多重检验和效能计算	52
	引言	52
6.1	多重检验导致 I 型错误的发生	52
6.2	三种主要的多重检验校正方法	53
6.2.1	控制总 I 型错误率	53
6.2.2	贝叶斯理论(Bayesian perspective)	54
6.2.3	错误发现率	55
6.3	成功有效的研究需要进行计算统计效能	59
6.3.1	效能计算举例	59
6.3.2	效能与样本量的关系	61
6.3.3	间接关联的效能统计	62
6.3.4	遗传研究中实际的效能统计	62
6.4	致谢	63
	参考文献	63
	互联网信息	64
7	遗传关联分析	65
	引言	65
7.1	具有显著效应的遗传相关性	66
7.2	寻找直接相关性	66
7.2.1	病例-对照研究	66
7.2.2	病例-对照研究的统计学分析	66
7.2.3	统计学分析范例	67

7.2.4 数量计量的使用	69
7.3 寻找间接相关性	70
7.3.1 连锁不平衡法	70
7.3.2 多标记及单体型分析	71
7.3.3 与疾病相关的基因间的交互作用	71
7.4 应对分析中的问题	72
7.4.1 质量控制	72
7.4.2 哈-温平衡	72
7.4.3 基因型缺失	73
7.4.4 群体分层	74
7.5 关联研究的缺点和问题	74
7.5.1 假阳性结果	75
7.5.2 重复实验效能的缺乏	75
7.5.3 研究之间的异质性	75
7.5.4 研究内的异质性	75
7.6 结论	76
参考文献	76
8 全基因组关联研究 GWAS	78
引言	78
8.1 GWAS 基于常见疾病-常见变异的假说	78
8.2 标签 SNP(tag SNP)与连锁不平衡(LD)是 GWAS 的基础	79
8.2.1 标签 SNP(tag SNP)	79
8.2.2 连锁不平衡(LD)	79
8.2.3 始祖突变与单倍型	81
8.3 GWAS 研究中使用的芯片平台	82
8.4 怎样做 GWAS 分析	82
8.4.1 数据处理	82
8.4.2 质控	83
8.4.3 群体分层	85
8.4.4 群体分层的校正	86
8.5 GWAS 中的数据分析方法	87
8.6 单倍型分析有助于定位功能变异	88
8.6.1 鉴定单倍型	88
8.6.2 E-M 算法	88
8.6.3 单倍型模块	90

8.7 GWAS产生的一些关键结果	90
8.7.1 老年性黄斑变性	90
8.7.2 Wellcome 病例对照研究信托基金会(WTCCC)	90
8.7.3 1型糖尿病	90
8.8 我的研究结果是阴性的——帮帮我.....	91
8.9 其他的策略也很重要.....	91
8.9.1 拷贝数变异	91
8.9.2 少见变异.....	91
8.10 结论	92
参考文献	92
互联网信息	95
9 连锁不平衡、HapMap 及插补介绍	96
引言	96
9.1 一些基本 LD 统计学知识.....	96
9.1.1 LD 统计量 D'	96
9.1.2 LD 统计量 r^2	97
9.2 利用国际单体型计划定位 LD 区域.....	97
9.3 从标签到填补.....	99
9.4 结论	100
参考文献.....	100
互联网信息.....	101
10 全基因组关联研究的 Meta 分析	102
引言.....	102
10.1 处理基因型数据缺失方面的输入软件.....	102
10.1.1 数据输入的准确性及质量.....	103
10.1.2 卡方校正	105
10.1.3 将不确定的数据整合到 Meta 分析中	107
10.2 Meta 分析准备工作	107
附录.....	109
参考文献.....	111
互联网信息.....	112
11 基因-环境相互作用与复杂疾病	113
引言.....	113
11.1 在遗传流行病学中基因-环境交互作用意味着什么	114
11.2 常见疾病中存在基因-环境交互作用的证据	115

11.3	从不同角度看基因-环境交互作用	115
11.3.1	遗传角度	117
11.3.2	公共卫生学角度	117
11.4	为什么研究基因-环境交互作用?	117
11.5	研究设计	120
11.5.1	横断面病例-对照研究	120
11.5.2	单纯病例研究设计	121
11.5.3	前瞻性队列研究设计	121
11.6	基因-环境交互作用研究中的统计效能和样本量	122
11.7	结果的重复性和 Meta 分析	124
11.8	候选基因与全基因组研究	124
11.9	概括与结论	125
	参考文献	126
	互联网信息	128
12	基于家系的遗传学关联分析	129
	引言	129
12.1	为什么要开展基于家系的关联研究?	129
12.2	早期基于家系的关联研究与病例-对照研究	131
12.3	传递不平衡检验是一种连锁分析	132
12.4	传递不平衡检验是一种关联和连锁分析	133
12.5	基于家系的关联分析具有广泛的应用范围	135
12.5.1	用于迟发型疾病的分析	135
12.5.2	用于一般核心家系的关联和连锁分析	135
12.5.3	用单体型进行分析	136
12.5.4	对连续数量性状的关联研究	137
12.5.5	X 连锁遗传的关联研究	137
12.5.6	对父母起源和母源效应的研究	138
12.5.7	基因间和基因与环境间相互作用的研究	138
12.6	结语	139
	致谢	139
	参考文献	140
	互联网信息	143
13	拷贝数变异与人类常见疾病	144
	引言	144
13.1	人类基因组中的拷贝数变异	144
13.2	结构变异的形成机制	145

13.3	拷贝数变异对基因表达的影响	146
13.4	CNV 的群体特征	146
13.5	人类常见疾病的遗传结构:基于 SNP-和 CNV-的 GWAS	147
13.6	CNV 的研究是怎样改变我们对复杂疾病遗传结构认识的?	148
13.7	CNV 检测技术进展及在复杂疾病研究中的应用	150
13.8	结语	151
	参考文献	151
	互联网信息	158
14	肿瘤基因组学	159
	引言	159
14.1	肿瘤的遗传基础	159
14.2	肿瘤中基因组学以及基因改变研究	162
14.2.1	肿瘤中拷贝数变异分析	162
14.2.2	基因组重排的发现	162
14.2.3	肿瘤中的体细胞突变	164
14.2.4	肿瘤中的常见变异	166
14.2.5	肿瘤中的表观遗传改变	167
14.3	肿瘤转录组改变	171
14.3.1	研究转录基因组改变的方法	171
14.4	候选癌基因的优选	172
14.4.1	癌基因筛选	172
14.4.2	癌基因的计算优选	173
14.5	不同类型肿瘤基因组学数据的整合	174
14.5.1	研究方法	174
14.5.2	肿瘤基因组学整合研究计划及资料库	175
14.6	小结	177
	致谢	177
	参考文献	177
	互联网信息	188
15	序列变异影响信使 RNA 前体剪接并导致(复杂)疾病的机制:不局限于遗传密码	190
	引言	190
15.1	剪接的发生过程	191
15.1.1	为定位剪接位点,剪接增强子和沉默子是必需的	193
15.2	剪接失调导致疾病	196

15.3	剪接突变不直接参与剪接位点的构成	199
15.3.1	MCAD 中 SRE 突变:单倍型的重要性	202
15.3.2	CFTR 9 号外显子的 SRE 突变和错误剪接	202
15.3.3	SMN1/2,剪接和脊髓肌萎缩	203
15.3.4	外显子跳跃和脆弱性	204
15.4	剪接调节蛋白和剪接体成分中的常见遗传变异能导致亚最佳剪接	204
15.5	环境影响剪接	205
15.6	In Silico 评估工具用于评估剪接效应	207
15.7	结论	208
	参考文献	209
	互联网信息	214
16	高通量基因分型的实验室研究方法	215
	引言	215
16.1	为什么要选用 SNP?	216
16.2	候选基因有用吗?	217
16.3	连锁区域可以通过高通量 SNP 基因分型进行精确定位	219
16.4	GWAS 及其相关基因分型	220
16.5	SNP 数据需要做好质量控制	222
16.6	下一代测序技术将带来基因分型的革命性变化	225
16.7	小结和结论	226
	致谢	226
	参考文献	227
	互联网信息	229
17	全基因组关联研究的基因集分析和网络分析	230
	引言	230
17.1	基因集分析和基因网络分析	232
17.1.1	将遗传关联定位到基因	234
17.1.2	GWAS 的 GSA 和网络分析:到更深处捕捞更小的鱼	235
17.1.3	GSA 的应用	237
17.1.4	网络 and 上位分析的应用	238
17.1.5	GWAS 联合和网络与 eQTL 分析	238
17.2	挑战与前景	239
	参考文献	240
	互联网信息	244

1 绪 言

Ammar Al-Chalabi¹ and Laura Almasy²

¹MRC Centre for Neurodegeneration Research, King's College London, London SE5 8AF, United Kingdom; ²Southwest Foundation for Biomedical Research, San Antonio, Texas 78227

1.1 为什么遗传学重要

现代遗传学研究的核心在于:通过了解何种遗传变异导致疾病表型,我们能搞清楚疾病发病机制并因此可能干预或预防疾病。我们距离全面理解人类基因组以及其与疾病或其他表型的关系这一努力目标还非常遥远,但我们已经取得了显著进步,某些疾病已经开始向我们暴露它们的秘密了。

在本书中,我们着眼于将遗传学家用来鉴定疾病基因的工具和概念以及支撑这些遗传学概念的统计学理论结合起来。对人类遗传学感兴趣的或者在研究中需要用到遗传学技术的研究者会发现这些内容很有用,尤其对那些研究复杂遗传疾病的研究者而言。本书涉及一些统计学和数学知识,但通过一个章节对统计学进行简要介绍以及每个特定章节进行详细的用法解释,理解这些内容并不需要特别的统计学能力。书中覆盖的主题内容广泛,但主要强调关联研究,这是由于关联研究是目前众多复杂疾病研究设计的基础。另外,本书也涵盖了经典的方法,如连锁分析,这是由于这些方法对研究各种表型非常有用,也是后续很多方法的基础,研究者需要理解这些方法才能合理评价已有的研究结果。这些章节对从事遗传研究的人来说既是一个操作指南合集,也是各个领域的内容综述,这使得其成为那些既想知道如何做又想知道为什么的研究人员的非常宝贵的资源。

1.2 现代遗传学简明史

1.2.1 紧跟遗传学思想

遗传学正以非常快的速度发展,一个很好的体现是里程碑式的重要进展被飞速超越。尽管遗传学发展到当前认识水平的过程经历了很多重要的阶段,但知识的主要跨越性发展要么取决于新的方法(数学、概念或技术),要么来自现有的观念被推翻或者融合。

1.2.2 孟德尔、达尔文以及遗传的波-粒辩论

1859 年达尔文发表了其依赖于遗传观点的进化理论(Darwin 1859)。之后的 1865 年出现了第一个真正的针对遗传现象的现代科学分析:奥地利布隆城圣詹姆斯修道院的牧师格列高尔·孟德尔实施了他精细的豌豆杂交与计数实验(Mendel 1866)。1905 年,William Bateson 提出了遗传学(genetics)这个词,成为第一个遗传学教授,那是在英国剑桥大学。在当时,遗传机制到底是基于粒子还是基于波存在很大的争议。Bateson 认为基因是波或者是震动,而 Karl Pearson(卡方检验、回归和相关等概念的提出者)认为基因是一个个的颗粒(历史上对光的本质也有类似的甚至更长时间的争论)。孟德尔定律只能用遗传的粒子理论来解释,而似乎带有渐变过程的进化理论则难以用粒子理论来解释,相反用波-动理论就能很好地解释(可以容许亲本性状混合)。另外,波-动理论则无法解释进化理论中的多样性问题:任何混合方式最终都会随着时间带来多样性的丢失。这个矛盾在 1918 年被 Ronald Fisher(他提出了“统计变量”这一概念)解决了,他和 J. B. S. Haldane 以及 Sewall Wright 一起证实了基于粒子的遗传多基因学说可以同时符合孟德尔定律以及进化理论(Fisher 1918)。1903~1910 年,Walter Sutton 和 Thomas Hunt Morgan 发现位于染色体上的基因是遗传单元。因此,19 世纪末和 20 世纪初产生了现代遗传学和统计学的基础理论,这也是本书各种概念的直接源头。

1.2.3 分子生物学的中心法则

20 世纪 30 年代至 40 年代产生了遗传学的中心法则(1941 年提出,1958 年正式形成),该法则认为遗传信息从 DNA(1933 年发现位于染色体内)到 RNA,再到蛋白质,而不是相反方向(Crick 1970)。由于逆转录病毒和朊病毒的发现,中心法则在 1964 年和 1982 年两次被修订。

1.2.4 DNA 和遗传密码

1953 年,James D. Watson 和 Francis H. C. Crick 利用 Rosalind Franklin (Maurice Wilkins 实验室的工作人员)的晶体数据确定了 DNA 的碱基配对双螺旋结构。然后在 1967 年,我们现在称为遗传密码的碱基-蛋白质对应关系被众多科学家一一破解。现在,我们知道只有不到 2% 的人类基因组 DNA 是编码蛋白质的,显然遗传密码远没有被彻底破解,我们仍然未完全理解基因组内剩余的 98% 的信息。

1.2.5 基因组学时代的来临

20 世纪 70 年代见证了基因组学时代的到来。1972 年 Walter Fiers 小组发表