新生物学丛书

新一代测序数据分析

NEXT-GENERATION SEQUENCING DATA ANALYSIS

〔美〕王忻琨 著
陈浩峰 主译

编辑推荐
Editor's Choice

科学出版社

新生物学丛书

新一代测序数据分析

Next-Generation Sequencing Data Analysis

〔美〕王忻琨 著

陈浩峰 主译

科学出版社

北京

图字：01-2017-6182

内 容 简 介

本书是一部介绍有关新一代测序（NGS）数据分析方法的著作。书中全面系统地介绍了新一代测序技术的生物学意义、测序原理、分析过程和应用领域等；详细介绍了新一代测序数据的分析方法，包括其在基因组从头测序和重测序、转录组测序、小 RNA 测序、ChIP 测序、表观基因组测序及宏基因组测序等应用中的具体分析方法，对读者学习新一代测序技术、促进该技术在生命科学各个领域中的应用有着重要意义。

本书可以作为具有生物学、遗传学、基因组学与生物信息学、农学等生命科学背景的高等学校师生和科研院所研究人员学习新一代测序技术原理及进行测序数据分析的参考书，也可供各大医院和医学院对于精准医学感兴趣的专业人士和其他生命科学从业者参考。

图书在版编目（CIP）数据

新一代测序数据分析 /（美）王忻琨著；陈浩峰主译. —北京：科学出版社，2018. 2
（新生物学丛书）
书名原文：Next-Generation Sequencing Data Analysis
ISBN 978-7-03-056469-6

Ⅰ. ①新… Ⅱ. ①王… ②陈… Ⅲ. ①生物信息论 Ⅳ. ① Q811. 4

中国版本图书馆 CIP 数据核字（2018）第 019094 号

责任编辑：罗 静 / 责任校对：彭 涛
责任印制：张 伟 / 封面设计：刘新新

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
*
2018 年 2 月第 一 版 开本：720 × 1000 1/16
2018 年 2 月第一次印刷 印张：13 1/4
字数：262 000

定价：98.00 元

（如有印装质量问题，我社负责调换）

译 者 名 单

主　　译　陈浩峰

翻译人员　陈浩峰　王　静　韩　瑶　齐　洺

译者单位　中国科学院遗传与发育生物学研究所

译者前言

随着现代科学技术的发展，生命科学研究已经进入了组学时代。新一代测序（NGS）技术的出现及其广泛应用，极大地推动了生命科学各个领域的发展，促生了包括“精准医学”在内的一大批新兴学科。人们运用NGS技术获得了海量的测序数据，如何从这些数据中分析总结出新知识、新发现，并将其应用于生命科学科研和生物产业及医疗产业，是摆在广大NGS应用者面前的一大难题。大多数NGS用户都仅具有生命科学或医学背景，对NGS数据分析过程了解不足，急需学习这方面的知识。目前，读者获取的NGS数据分析的相关知识，往往来自于散在的研究或综述论文，或者来自于互联网，不利于构成完整的知识体系。鉴于此，我们组织翻译了由美国西北大学王忻琨博士所著的《新一代测序数据分析》一书，以期填补国内中文书籍这方面的空白。

本书是一部全面介绍新一代测序及基因组学和生物信息学的著作，全面兼顾了生物学基本原理、测序技术、测序数据分析处理等方面的知识，为读者全方位详细介绍了新一代测序数据分析所涉及的方方面面的内容，可以作为有生物学、遗传学背景的教师及学生学习新一代测序技术原理和测序数据分析的参考。我们将其介绍给国内的读者，希望能够对促进新一代测序技术在我国生命科学研究中的应用，乃至精准医学的发展有所裨益。

本书译者曾编写过主要阐述NGS实验方法的《新一代基因组测序技术》一书，希望本书的出版可以与之相得益彰，共同构成有关新一代测序的完整知识体系。

感谢中国科学院遗传与发育生物学研究所基因组生物学研究中心和植物基因组学国家重点实验室对本书翻译工作的支持，感谢Illumina中国公司、北京英木和生物技术有限公司的赞助。

限于译者水平，译文的不妥之处在所难免，望广大读者予以指正。

陈浩峰

2017年8月1日

目　录

第一部分　细胞与分子生物学概论

第二部分　新一代测序技术及数据分析概论

第三部分　新一代测序数据分析的具体应用

第四部分 发展中的新一代测序技术与数据分析

第一部分
细胞与分子生物学概论

1 细胞系统与生命密码

1.1 细胞面临的挑战

虽然细胞体积非常微小，直径只有不到50μm，但其运行的精巧程度可以和一切人类设计的系统相媲美，并且它可以通过以DNA形式储存的密码来实现自身繁殖。假如我们希望设计一个和细胞同样复杂的系统，将需要面临很多看起来无法逾越的挑战。细胞具有复杂的内部结构，含有多种生物大分子组分，它需要完成一系列任务来维持自身系统的稳定，其中最根本的任务是在复杂多变的环境条件下，维持它的内部秩序，防止其出现故障或者崩溃，并完成自身繁殖，甚至进一步提升自身系统的功能。

要维持细胞内部系统的秩序，就需要输入能量。根据热力学第二定律，一个系统如果没有持续的能量输入，其熵值将逐渐上升，最终将导致系统的解体。同时，因为细胞内部结构的存在状态是动态的，需要响应不断变化的外部环境，所以除了能量输入以外，细胞还需要不断补充“原材料”来更新它的内部部件，或者制造新的部件。因此，细胞要维持其内部平衡及与所处环境之间的互动，就需要不断地输入能量与原材料，并且排出废物。细胞的系统信息储存在它的DNA序列中，用于指导细胞获取生存所需的能量和原材料，以及实现细胞繁殖等生命过程。

进化使得大量的生物种类都不再像单细胞生物那样简单。例如，人体就是由数千亿个细胞组成的。在多细胞系统中，细胞发生分化以执行特定的功能，如人体胰腺中的β细胞执行合成与释放胰岛素的功能，大脑中的皮层神经元执行学习和记忆的神经生物学功能。尽管功能多样，但对这些细胞的某个个体来说，它所面对的挑战与单细胞生物是一样的。不同的是，其不直接面对外部环境，而是需要面对其周边微环境的变化。

1.2 细胞如何面对挑战

很多细胞，如藻类与植物的细胞，直接从阳光或其他来源汲取能量。另外一些细胞（或者生物体）则作为异养生物从环境中汲取能量。在原材料方面，细胞可以利用简单有机化合物中携带的能量来固定空气中的二氧化碳，或者从环境中直接获取有机分子转化为其所需物质。同时，细胞内现存的物质也可以被降解，作为新的原材料被重新加以利用。这个能量的捕获、利用、合成、转化，以及分

子物质的降解再利用过程，构成了细胞代谢过程。代谢是细胞最基础的特征，包含了大量的生物化学反应。

对环境中各种信号的接收与转导对细胞的生存至关重要。细胞依靠其表面分布的受体接收信号，有些信号则需要依靠细胞内结构接收。信号转导到细胞内以后，通常在胞内引起一系列反应，使原始信号得以放大与调节，与之对应，细胞的代谢过程也发生相应的变化。细胞的信号接收与转导网络是由不同的代谢途径环路组成的。如果这些代谢途径失去功能，将会损害细胞对环境的反应能力，最终导致细胞死亡。

细胞系统依靠 DNA 复制和细胞分裂来完成其繁殖和进化。DNA 的复制（详见第 2 章）是一个高度保真的过程，但并非完全不出错。在维持系统稳定的同时，这个过程也提供了细胞系统分化与进化的机制。细胞分裂在各个阶段都是被精密调控的，以保证复制后的 DNA 可以平均分配到子细胞中。对于大多数有性生殖的生物来说，在其生殖细胞的形成过程中，DNA 只复制一次，而细胞却分裂两次，其结果是每个配子中 DNA 量减半。而父本 DNA 和母本 DNA 发生的重组，又使得子代进一步发生分化。

1.3 细胞内的各种分子

细胞中的各种过程需要不同类型的生物分子去实现。在一个典型的细胞中大部分是水，约占细胞总重量的 70%。除了水之外，细胞中还含有大量不同大小的分子。小分子主要包括无机离子（如 Na^+、K^+、Ca^{2+}、Cl^-、Mg^{2+} 等）、单糖、脂肪酸、氨基酸及核苷酸。绝大多数的大分子是多糖、脂类、蛋白质和核酸（DNA 和 RNA）。在上述组分中，无机离子对信号转导过程（如 Ca^{2+} 的波动是重要的胞内信号）、细胞能量的储备（如 Na^+/K^+ 形成跨膜浓度梯度）、蛋白质的结构 / 功能（如 Mg^{2+} 是很多金属蛋白的关键辅助因子）至关重要。碳水化合物（包括单糖与多糖）、脂肪酸及脂类是细胞中提供能量的主要分子。其中脂类是细胞膜的主要成分。蛋白质由 20 种氨基酸以不同的顺序与长度组成，参与几乎所有的细胞活动，包括代谢、信号转导、DNA 复制和细胞分裂。它们还是很多胞内结构，如细胞骨架（详见 1.4 节）的组成部分。核酸以它几乎无穷的核苷酸排列顺序来储存生命的编码信息，它不仅指导细胞内蛋白质的合成组装，而且可以根据环境的变化对这些蛋白质如何组装进行调控。

1.4 细胞内结构与空间

细胞具有组织精巧的内部结构（图 1.1）。根据其内部结构的复杂程度，可将细胞分为两大类：原核细胞与真核细胞。它们之间的根本区别在于是否具有细胞

核。原核细胞是较为原始的形态，它没有细胞核，其 DNA 位于一个没有包被的区域，称为核区；原核细胞也没有细胞器。与之相比，真核细胞具有明显的细胞核，用于 DNA 的储存、维护与表达。真核细胞还具有各种特化的、可以将细胞各种功能分隔开的细胞器，如内质网（ER）、高尔基体、细胞骨架、线粒体及植物细胞特有的叶绿体等。这里我们对各种细胞内的结构与空间，包括细胞核、各种细胞器及其他亚细胞结构与空间，如细胞膜和细胞质等作一简单描述。

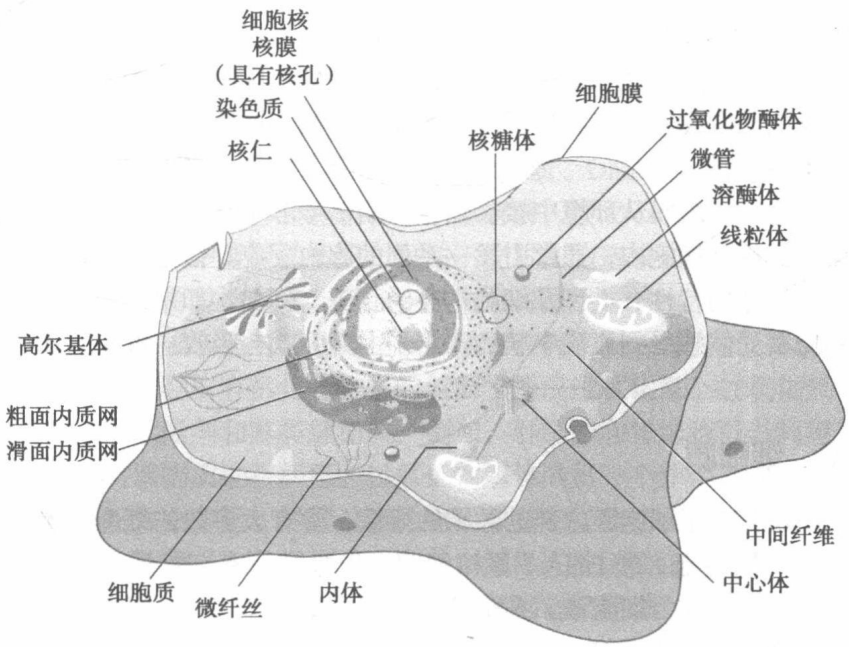

图 1.1　典型真核细胞的结构图，本图所示为一个动物细胞。

1.4.1　细胞核

DNA 是储存生命密码的载体，因此它需要妥善维护，以避免可能遭受的损伤，并且需要保证它的准确性与稳定性。又因为将储存在 DNA 中的遗传信息准确地表达出来是细胞正常发挥功能的关键，所以基因表达在任何时候都需要受到严格调控。在绝大多数真核细胞中，细胞核都位于中心位置，它可以为 DNA 的储存、维护和基因表达提供一个受到良好保护的环境。细胞核区被双层的核膜包被，核膜上分布有跨内膜与外膜的核孔，通过这些核孔可以进行蛋白质和各种 RNA 的跨膜运输，这对于基因表达过程非常关键。细胞核骨架是由核纤层蛋白组成的网络结构，为细胞核提供机械支持。在细胞核内，细长的 DNA 分子缠绕在一种称为组蛋白的蛋白质上，被紧密包裹，以便存储在有限的细胞核空间里。在原核细胞中，有一个细胞核形状的不规则区域，称为核区，外部没有核膜包被。核区可以为 DNA 提供一些类似真核细胞核膜的保护（但其功能不如核膜）。