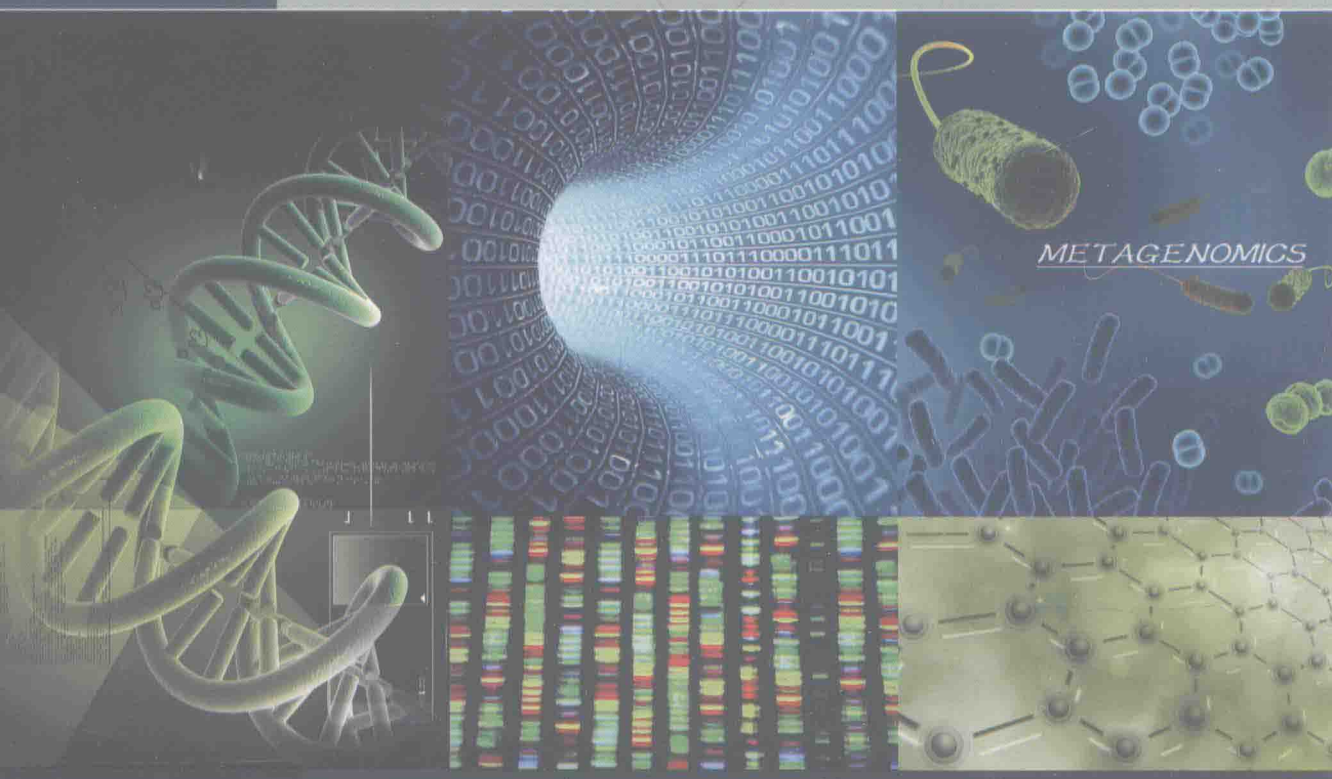
深圳市华大基因学院指定教材

English for Genomics and Its Application

基因组学及应用专业英语

主　编◎孙　颖
副主编◎黄　玉　石　琼



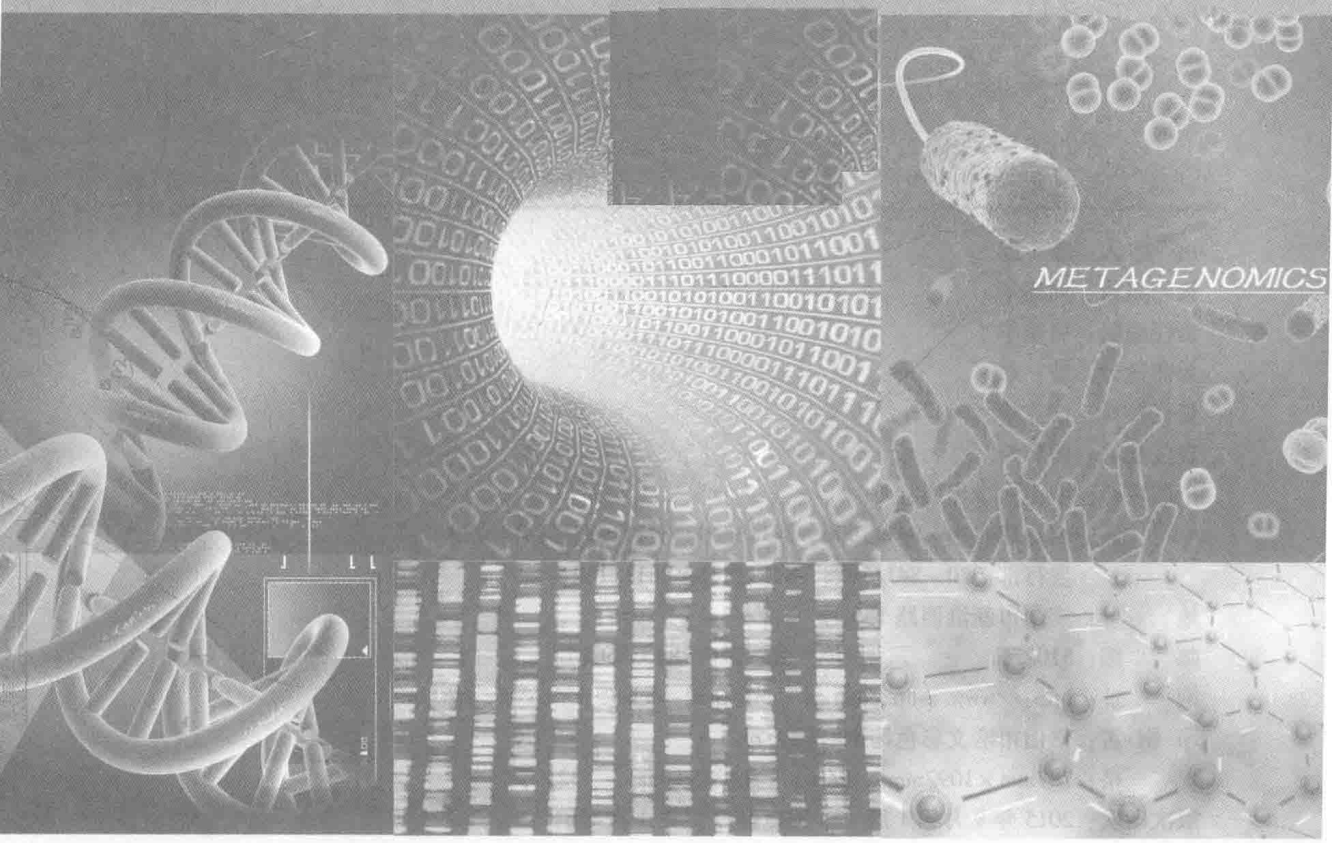
中山大學出版社
SUN YAT-SEN UNIVERSITY PRESS

深圳市华大基因学院指定教材

English for Genomics and Its Application

基因组学及应用专业英语

主　编◎孙　颖
副主编◎黄　玉　石　琼



中山大學出版社
SUN YAT-SEN UNIVERSITY PRESS
·广州·

图书在版编目（CIP）数据

基因组学及应用专业英语/孙颖主编；黄玉，石琼副主编．—广州：中山大学出版社，2015.9

ISBN 978-7-306-05394-7

Ⅰ.①基… Ⅱ.①孙… ②黄… ③石… Ⅲ.①基因组学—英语—高等学校—教材 Ⅳ.①H31

中国版本图书馆 CIP 数据核字（2015）第 185915 号

出 版 人：徐　劲
策划编辑：周建华
责任编辑：周建华
封面设计：曾　斌
责任校对：江克清
责任技编：何雅涛
出版发行：中山大学出版社
电　　话：编辑部 020-84111996，84113349，84111997，84110779
　　发行部 020-84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传　真：020-84036565
网　　址：http://www.zsup.com.cn　　E-mail:zdcbs@mail.sysu.edu.cn
印 刷 者：佛山市浩文彩色印刷有限公司
规　　格：787mm×1092mm　1/16　14.25 印张　321 千字
版次印次：2015 年 9 月第 1 版　2015 年 9 月第 1 次印刷
定　　价：38.00 元

本书编委会

主　编：孙　颖
副主编：黄　玉　石　琼
编　委：孙　颖　黄　玉　石　琼
　　赵晓萌　陈洁明　徐军民

前　言

基因组学在生命科学各个领域中发挥着越来越重要的作用，成为当今生命科学、医学等重要研究领域强有力的工具。基因组学及其研究涉及的相关方法和技术，如基因组学样本的规范化收集、测序技术、大数据的产生、存贮以及数据的分析和应用等，已成为目前生物医学领域关注的焦点。同时，近年来，随着第二代测序技术的迅速发展，数以亿计的原始测序数据正源源不断地从测序仪上产生，这些海量的数据给生物信息学、基因组学及应用等带来了全新的挑战。

为了帮助在校高年级大学生以及研究生学习、了解基因组和在生命科学各领域的研究进展及其应用，我们编写了《基因组学及应用专业英语》一书，系统地介绍在大数据背景下，新一代测序技术、基因组学及其在医学、农业、环境、生物多样性等方面的应用进展，为生命科学及医学等领域的学生、学者和研究人员提供专业英语教材及辅助阅读材料。本书既适用于研究生的专业阅读和参考材料，也可以作为教师和从事生物医学研究相关人员的专业英语学习资料。

本教材以内容精炼、知识新颖、难易适中作为选材的主要标准，以英文原版教材为参考，精选来自 *Nature*、*Science* 等顶级期刊中有关基因组及其应用的文献，确保使用最纯正的英文表达，讲述最核心的技术路线，涵盖最前沿的研究进展，使学生在学习过程中既能学到前沿的基因组学及应用专业知识，又能提高学术英语能力。

本教材内容覆盖广泛，包括基础的高通量测序技术（二代测序技术）的原理和应用，测序产出的大数据及其在个性化医疗、动物和植物基因组及环境基因组研究、生物多样性保护上的应用，以及基于基因组学的生物样本库建设和资源收集现状和展望。教材总共分为 8 章，每章分成 3 课，各围绕一个主题，选取 1～2 篇高质量综述或数篇文章的精华段落作为主要内容，由浅入深，涉及内容范围由大到小，让读者对基因组及其应用有一个广泛而深入的了解和把握。

教材每课课后均编排词汇表（Glossary），以帮助学生掌握基因组学核心词汇。此外，除常规的音标标注外，还加入了专业英语常用词根、词缀、构词法的标注和说明，常用短语和专有名词缩写也一一在列，使记单词变得事半功倍。同时，在每篇文章中以黑体突出显示专业词汇，使学生在学习时能更快找到它们在文中的位置，更好地理解其含义及使用方法。词汇表后设有难句解释（Notes to Difficult Sentences）模块，对难句进行翻译并对涉及的专业知识重点进行详细解释，从而进一步拓展专业知识，在提升学生的专业英语阅读能力的同时，帮助学生读懂、读通、读透，进而完全掌握课程内容，扩宽知识面。本教材每章节后编排了丰富的课后阅读材料（Reading Material），材料内容新颖有趣，简单易读，贴近生活，让学生充分消化吸收所学内容，进一步强化知识，做到活学活用。

本教材的撰写与出版得到深圳市科技计划深港创新圈项目“DNA 条形码系统支撑的深港水产食品安全体系建设”（SGLH20131010105856414）和江苏省镇江市高层次创新创业人才引进计划项目“长江流域经济鱼类的基因组学研究及育种应用”的资助，还得到深圳华大基因研究院与华大（镇江）水产科技产业有限公司的大力支持。此外，本书参考和使用了国内外一些专业文献资料和图表以及国外生物领域的杂志和网站内容，在此一并表示衷心的感谢！

本书已在华大基因研究院创新班和中国科学院大学华大教育研究中心研究生班试用，获得一致好评，并在此基础上，进一步修订。由于编者能力有限，错误、缺点在所难免。书中如有不妥之处，敬请同行与读者批评指正。

孙　颖

2015 年 6 月 8 日

Contents

Chapter 1 Next-generation Sequencing (NGS)

Over the past several years, next-generation DNA sequencing technologies have catapulted to prominence, with increasingly widespread adoption of several platforms that individually implement different flavors of massively parallel cyclic-array sequencing. Common characteristics extend beyond the technologies themselves, to the quantity and quality of data that are generated, such that all raise a similar set of new challenges for experimental design, data analysis and interpretation. The reduction in the costs of DNA sequencing by several orders of magnitude is democratizing the extent to which individual investigators can pursue projects at a scale previously accessible only to major genome centers. The dramatic increase in interest in this area is also evident in the number of groups that are now working on sequencing methods to supplant even the new technologies discussed here. Given the current state of flux, it is difficult to peer even a few years into the future, but we anticipate that next-generation sequencing technologies will become as widespread, commoditized and routine as microarray technology has become over the past decade. Also analogous to microarrays, we expect that the challenges will quickly shift from mastering of the technologies themselves to the question of how best to go about extracting biologically meaningful or clinically useful insights from a very large amount of data.

Lesson 1
From Convention to Second-generation Sequencing

DNA sequence represents a single format onto which a broad range of biological phenomena can be projected for high-throughput data collection. Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude, and democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators. These new technologies are rapidly evolving, and near-term challenges include the development of robust protocols for generating sequencing **libraries**, building effective new approaches to data-analysis, and often a rethinking of experimental design. Next-generation DNA sequencing has the potential to dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread, rather than requiring significant production-scale efforts.

The field of DNA sequencing technology development has a rich and diverse history. However, the overwhelming majority of DNA sequence production to date has relied on some version of the Sanger biochemistry. Over the past five years, the incentive for developing entirely new strategies for DNA sequencing has emerged on at least four levels, undeniably reinvigorating this field. First, in the wake of the Human Genome Project, there are few remaining avenues of optimization through which significant reductions in the cost of conventional DNA sequencing can be achieved. Second, the potential utility of short-read sequencing has been tremendously strengthened by the availability of whole genome assemblies for ***Homo sapiens*** and all major model organisms, as these effectively provide a reference against which short reads can be mapped. Third, a growing variety of molecular methods have been developed, whereby a broad range of biological phenomena can be assessed by high-throughput DNA sequencing (e. g. , genetic variation, RNA expression, protein-DNA interactions and **chromosome** conformation). And fourth, general progress in technology across disparate fields, including microscopy, surface chemistry, nucleotide biochemistry, **polymerase** engineering, computation, data storage and others, have made alternative strategies for DNA sequencing increasingly practical to realize.

1. Sanger sequencing

Since the early 1990s, DNA sequence production has almost exclusively been carried out with **capillary**-based, semi-automated implementations of the Sanger biochemistry (Fig. 1.1a). In high-throughput production pipelines, DNA to be sequenced is prepared by one of two approaches: first, for shotgun *de novo* sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *Escherichia coli*; or second, for targeted **resequencing**, PCR amplification is carried out with primers that flank the target. The output of both approaches is an amplified template, either as many "**clonal**" copies of a single plasmid insert present within a spatially isolated bacterial colony that can be picked, or as many PCR **amplicons** present within a single reaction volume. The sequencing biochemistry takes place in a "cycle sequencing" reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest. Each round of primer extension is stochastically terminated by the incorporation of **fluorescently labeled** dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of **emission spectra**, provides the readout that is represented in a Sanger sequencing "trace". Software translates these traces into DNA sequence, while also generating error probabilities for each **base-call**. The approach that is taken for subsequent analysis—for example, genome assembly or variant identification—depends on precisely what is being sequenced and why. Simultaneous electrophoresis in 96 or 384 independent capillaries provides a limited level of parallelization. After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to 1000 bp, and per-base "raw" accuracies as high as 99.999%. In the context of high throughput shotgun genomic sequencing, Sanger sequencing costs on the order of $0.50 per kilobase.

2. Second-generation DNA sequencing

Alternative strategies for DNA sequencing can be grouped into several categories. These include (Ⅰ) microelectrophoretic methods; (Ⅱ) sequencing by **hybridization**; (Ⅲ) real-time observation of single molecules; (Ⅳ) cyclic-array sequencing. Here, we use "second generation" in reference to the various implementations of cyclic-array sequencing that have recently been realized in a commercial product (Fig. 1.1b). The concept of cyclic-array sequencing can be

summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection. Two reports in 2005 described the first integrated implementations of cyclic-array strategies that were both practical and cost-competitive with conventional sequencing, and other groups have quickly followed.

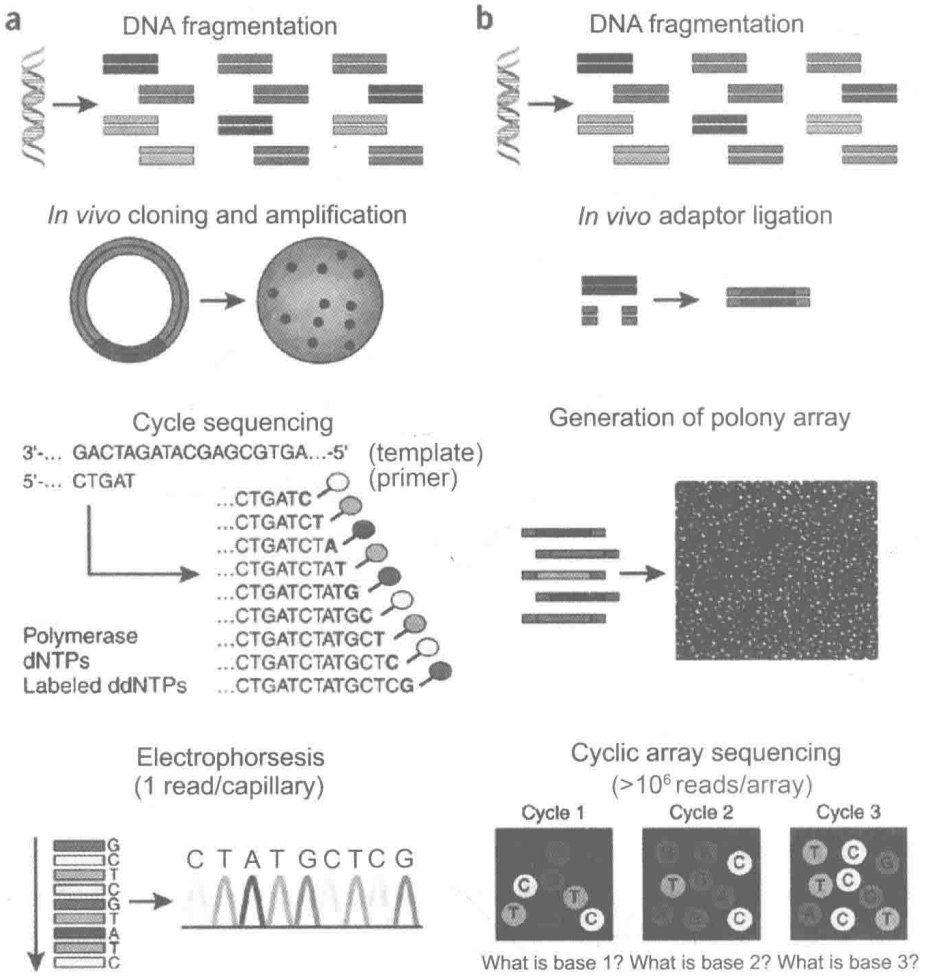


Figure 1.1 Work flow of conventional(a) versus second-generation(b) sequencing

Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their workflows are conceptually similar (Fig. 1.1b). Library preparation is accomplished by random fragmentation of DNA, followed by ***in vitro*** **ligation** of common **adaptor** sequences. Alternative protocols can be used to generate **jumping libraries** of mate-paired tags with controllable distance distributions. The generation of clonally clustered amplicons to serve as sequencing features can be achieved by several approaches, including ***in situ*** polonies, emulsion PCR or bridge PCR. What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (*in situ* polonies, bridge PCR), or to the surface of

micron-scale beads, which can be recovered and arrayed (emulsion PCR). The sequencing process itself consists of alternating cycles of **enzyme**-driven biochemistry and imaging-based data acquisition. The platforms that are discussed here all rely on sequencing by synthesis, that is, serial extension of primed templates, but the enzyme driving the synthesis can be either a polymerase or a **ligase**. Data are acquired by imaging of the full array at each cycle (e. g., of fluorescently labeled nucleotides incorporated by a polymerase). Global advantages of second-generation or cyclic-array strategies, relative to Sanger sequencing, include the following:

(1) *in vitro* construction of a sequencing library, followed by *in vitro* clonal amplification to generate sequencing features, circumvents several bottlenecks that restrict the parallelism of conventional sequencing (that is, transformation of *E. coli* and colony picking).
(2) Array-based sequencing enables a much higher degree of parallelism than conventional capillary-based sequencing. As the effective size of sequencing features can be on the order of 1 μm, hundreds of millions of sequencing reads can potentially be obtained in parallel by rastered imaging of a reasonably sized surface area.
(3) Because array features are immobilized to a planar surface, they can be **enzymatically** manipulated by a single reagent volume. Although **microliter**-scale reagent volumes are used in practice, these are essentially amortized over the full set of sequencing features on the array, dropping the effective reagent volume per feature to the scale of **picoliters** or **femtoliters**.

Collectively, these differences translate into dramatically lower costs for DNA sequence production.

Glossary

library /ˈlaɪbreri/
文库
A collection of DNA fragments that is stored and propagated in a population of micro-organisms through the process of molecular cloning.

Homo sapiens /ˌhoʊmoʊˈseɪpienz/
人；智人
Homo sapiens is the binomial nomenclature (also known as the scientific name) for the human species.

chromosome /ˈkroʊməsoʊm/
染色体
A chromosome is a packaged and organized structure containing most of the DNA of a living organism.

polymerase /ˈpɒləməˌreɪs/
聚合酶
An enzyme that synthesizes long chains or polymers of nucleic acids. DNA or RNA polymerases are used to assemble DNA or RNA molecules, respectively, by copying a DNA or RNA template strand using base-pairing interactions.

capillary /ˈkæpəleri/
毛细管；毛细管的
Capillaries are the smallest of a body's blood vessels that make up the microcirculation. These microvessels, measuring around 5 to 10 μm in diameter, connect

arterioles and venules, and they help to enable the exchange of water, oxygen, carbon dioxide, and many other nutrients and waste substances between the blood and the tissues surrounding them.

resequencing /reˈzekwənsɪŋ/

重测序

re + sequence + ing

The purpose of the sequencing of part of an individual (whose reference genome is already known) genome is to detect sequence differences between the individual and the standard genome of the species.

clonal /ˈkloʊnəl/

无性（繁殖）系的；无性（繁殖）系般的

clone—clonal, clonally, cloner, cloning, monoclone, polyclone

amplicon /æmplɪkɒn/

扩增子

An amplicon is a piece of DNA or RNA that is source and/or product of natural or artificial amplification or replication events.

fluorescently labeled

荧光标记的

fluorescent—fluorescently

emission spectra

发射光谱

base-call

碱基识别

Base calling is the process of assigning bases (nucleo-bases) to chromatogram peaks.

hybridization/ˌhaɪbrɪdaɪˈzeʃn/

杂交；配种

hybrid—hybridize; hybridization; hybridoma

(1) The process of combining different varieties of organisms to create a hybrid.

(2) The process of joining two complementary strands of nucleic acids (DNA, RNA or oligonucleotides).

in vitro /ɪnˈvɪtroʊ/

体外；离体

In vitro studies are performed with cells or biological molecules studied outside their normal biological context.

ligation /laɪˈgeɪʃn/

（核酸分子末端）连接

The covalent linking of two ends of DNA or RNA molecules, most commonly done using DNA ligase, RNA ligase (ATP) or other enzymes.

adaptor /əˈdæptə/

接头

An adapter is a short, chemically synthesized, double stranded DNA molecule which is used to link the ends of two other DNA molecules. It may be used to add sticky ends to cDNA allowing it to be ligated into the plasmid much more efficiently.

jumping library

跨步文库

A collection of genomic DNA fragments generated by chromosome jumping.

in situ /ɪnˈsaɪtuː/

原位；在原地

To examine the phenomenon exactly in place where it occurs (e. g. without moving it to special medium).

enzyme /ˈenzaɪm/

酶

ligase /ˈlaɪˌgeɪs/

连接酶

Ligase is a specific type of enzyme that facilitates the joining of DNA/RNA strands together by catalyzing the formation of a phosphodiester bond.

enzymatically /enzaɪˈmætɪklɪ/

酶促地

Enzymes are molecules that accelerate, or catalyze the chemical reactions.

microliter /ˈmaɪkroʊlɪtə/

微升（μL）$1\mu L = 1mm^3$

picoliter /pɪkəˈlaɪtə/

皮升（pL）$1pL = 10^3 \mu m^3$

femtoliter /ˈfemtoʊlɪtə/

飞升（fL）$1fL = 1\mu m^3$

Notes to Difficult Sentences

(1) Second, the potential utility of short-read sequencing has been tremendously strengthened by the availability of whole genome assemblies for *Homo sapiens* and all major model organisms, as these effectively provide a reference against which short reads can be mapped.

第二，人类及其他主要模式生物的全基因组组装项目有效地为短序列的匹配提供参考序列，从而极大提高了短序列测序的潜在效用。

拓展：自 1985 年美国科学家率先发起人类基因组计划以来，随着测序技术的不断发展，越来越多的基因组计划不断被推出和实施，例如，经济作物如水稻、棉花、番茄、花生的基因组计划，重要动物如大熊猫、鲤鱼、牛、蜜蜂的基因组计划，等等。现在，涵盖数以千万个体或物种的大规模基因组计划不断被提出，如 i5k（5000 种昆虫基因组计划）、G10K（万种脊椎动物基因组计划）、100 000 Genomes Project（英国 10 万基因组计划）等，生命科学领域全面迈入基因组和后基因组时代。

(2) What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (*in situ* polonies, bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR).

这些技术的共同点是，通过任何给定的单个文库得到的 PCR 扩增子最后都会在一定空间上聚集成群，要么聚集到平面上的某一点（原位克隆阵列，桥式 PCR），要么聚集到微米级的小磁珠上，可再次重新获取和排序（乳液 PCR）。

拓展：乳液 PCR 主要用于基因组测序。把用作 DNA 扩增模板的 DNA 片段分散在油液的小水滴内，逐一对小水滴内的 DNA 模板进行 PCR 扩增。理想状态下，一个小水滴只含有一个 DNA 模板和一个磁珠，由于小水滴内的 DNA 模板数目极少，可以减少当大量长短不一的模板一起扩增时所产生的短片段易被扩增的偏差，也可降低模板之间通过同源重组而扩增产生非天然连接的 DNA 片段的机会。

Lesson 2
Application of Next-generation Sequencing

The past several years have seen an accelerating flurry of publications in which next-generation sequencing is applied for a variety of goals. Important applications include: (Ⅰ)full-genome resequencing or more targeted discovery of **mutations** or **polymorphisms**; (Ⅱ) mapping of structural rearrangements, which may include copy number variation, balanced translocation breakpoints and chromosomal inversions; (Ⅲ) "RNA-Seq", analogous to **expressed sequence tags** (**EST**) or **serial analysis of gene expression** (**SAGE**), where shotgun libraries derived from mRNA or small RNAs are deeply sequenced; the counts corresponding to individual species can be used for quantification over a broad dynamic range, and the sequences themselves can be used for **annotation** (e. g., splice junctions and transcript boundaries); (Ⅳ) large-scale analysis of DNA **methylation**, by deep sequencing of bisulfite-treated DNA; (Ⅴ) "ChIP-Seq", or genome-wide mapping of DNA-protein interactions, by deep sequencing of DNA fragments pulled down by chromatin **immunoprecipitation**. Over the next few years, the list of applications will undoubtedly grow, as will the sophistication weth which existing applications are carried out.

The applications of NGS seem almost endless, allowing for rapid advances in many fields related to the biological sciences. Resequencing of the human genomes is being performed to identify genes and regulatory elements involved in pathological processes. NGS has also provided a wealth of knowledge for comparative biology studies through whole-genome sequencing of a wide variety of organisms. NGS is applied in the fields of public health and **epidemiology** through the sequencing of bacterial and viral species to facilitate the identification of novel virulence factors.

Additionally, gene expression studies using RNA-Seq (NGS of RNA) have begun to replace the use of microarray analysis, providing researchers and clinicians with the ability to visualize RNA expression in sequence form. These are simply some of the broad applications that begin to skim the surface of what NGS can offer the researcher and the clinician. As NGS continues to grow in popularity, it is inevitable that there will be additional innovative applications.

1. Whole-exome sequencing

Mutation events that occur in gene-coding or control regions can give rise to indistinguishable clinical presentations, leaving the **diagnosing** clinician with many possible causes for a given condition or disease. With NGS, clinicians are provided a fast, affordable, and thorough way to determine the genetic cause of a disease. Although high-throughput sequencing of the entire human genome is possible, researchers and clinicians are typically interested in only the protein-coding regions of the genome, referred to as the exome. The exome comprises just over 1% of the genome and is therefore much more cost-effective to sequence than the entire genome, while providing sequence information for protein-coding regions (Fig. 1.2).

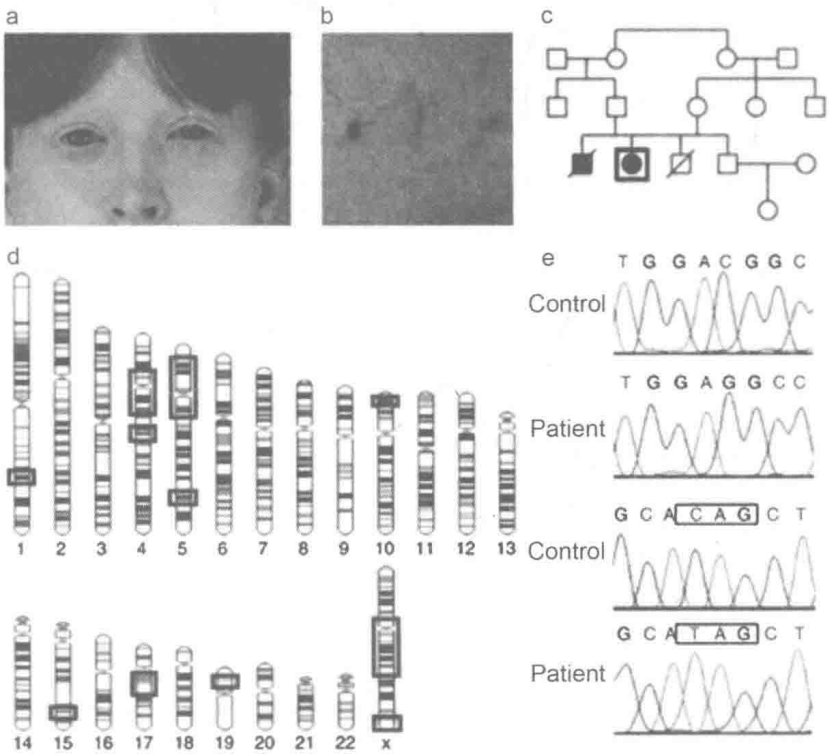


Figure 1.2 Clinical application of whole-exome sequencing in the detection of two disease-causing mutations

Figure 1.2 demonstrates the direct effect that NGS can have on the correct diagnoses of a patient. It summarizes the use of homozygosity mapping followed by whole-exome sequencing to identify two disease-causing mutations in a patient with **oculocutaneous albinism** and congenital neutropenia. **a** and **b** display the phenotypic traits common to oculocutaneous albinism type 4 and neutropenia observed in this patient. **c** is a pedigree of the patient's family, both the affected and unaffected individuals. The idiogram (graphic chromosome map) in **d** highlights the areas of genetic homozygosity. These regions were identified by single-nucleotide-polymorphism array analysis and were considered possible locations for the disease-causing mutation (s). **e** and **f** display chromatograms for the two disease-causing mutations identified by whole-exome sequencing. **e** depicts the mutation in SLC45A2, and **f** depicts the mutation in G6PC3. This case portrays the valuable role that NGS can play in the correct diagnosis of an individual patient who displays disparate symptoms with an unidentified genetic cause.