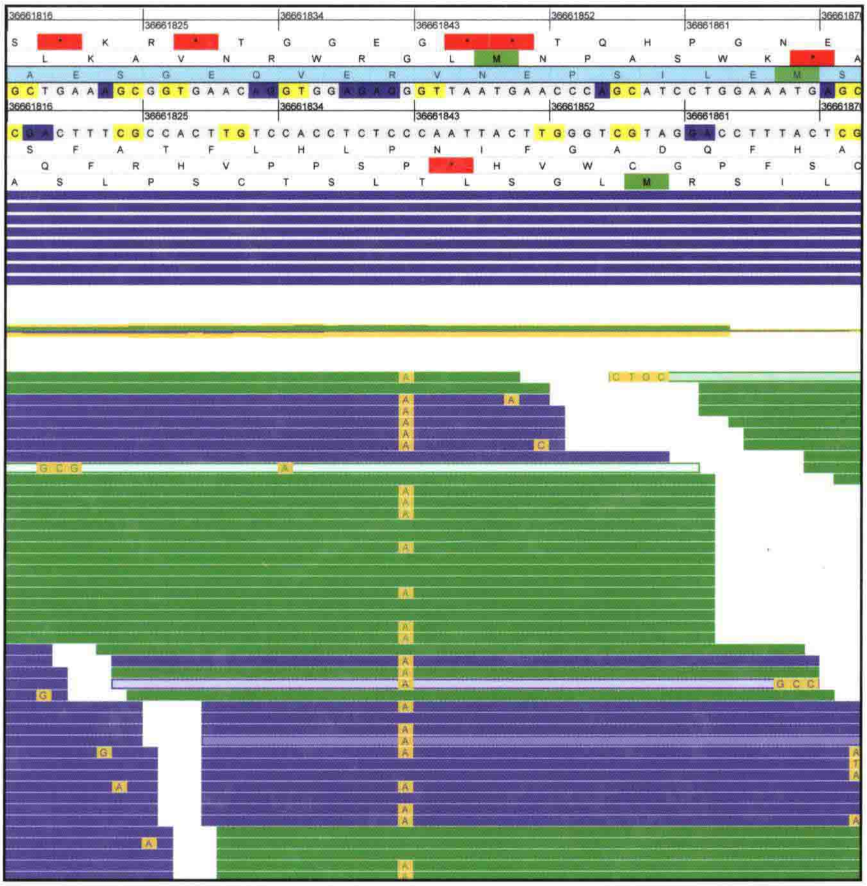
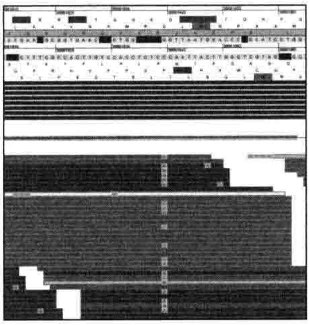
NEXT-GENERATION DNA SEQUENCING INFORMATICS


36661816
36661825
36661834
36661843
36661852
36661861
36661870
GCTGAAAGCGGTGAACAGGTGGAGAGGGTTAATGAACCCAGCATCCTGGAAATGAGC
CGACTTTCGCCACTTGTCCACCTCTCCCAATTACTTGGGTCGTAGGACCTTTACTCG

EDITED BY

STUART M. BROWN

NEXT-GENERATION DNA SEQUENCING INFORMATICS



EDITED BY

STUART M. BROWN

CSH PRESS
COLD SPRING HARBOR LABORATORY PRESS
Cold Spring Harbor, New York • www.cshlpress.org

NEXT-GENERATION DNA SEQUENCING INFORMATICS

Publisher and Acquisition Editor	John Inglis
Director of Editorial Development	Jan Argentine
Project Manager	Inez Sialiano
Permissions Coordinator	Carol Brown
Production Editor	Kathleen Bubbeo
Production Manager	Denise Weiss
Sales Account Manager	Elizabeth Powers
Cover Designer	Michael Albano

Front cover artwork: A heterozygous single-nucleotide G>A variant is verified by visualization in Genome View (genomeview.org) of short reads from a next-generation sequencing machine aligned to the reference genome. Forward reads are shown in blue, reverse reads are shown in green, and sequence variants are highlighted in yellow. Other visible sequence variants are probable sequencing errors.

Library of Congress Cataloging-in-Publication Data

Next-generation DNA sequencing informatics / edited by Stuart M. Brown.

pages cm
Includes bibliographical references and index.
ISBN 978-1-936113-87-3 (hardcover : alk. paper)
1. Nucleotide sequence. 2. Bioinformatics. I. Brown, Stuart M., 1962-

QP625.N89N485 2013
572.8'633–dc23

2012034431

10 9 8 7 6 5 4 3 2 1

All World Wide Web addresses are accurate to the best of our knowledge at the time of printing.

For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our website at www.cshlpress.org.

NEXT-GENERATION DNA SEQUENCING INFORMATICS

OTHER RELATED TITLES FROM COLD SPRING HARBOR LABORATORY PRESS

A Short Guide to the Human Genome

Guide to the Human Genome

Molecular Cloning: A Laboratory Manual, Fourth Edition

HANDBOOKS

An A to Z of DNA Science: What Scientists Mean When They Talk about Genes and Genomes

At the Bench: A Laboratory Navigator, Updated Edition

At the Helm: Leading Your Laboratory, Second Edition

An Illustrated Chinese–English Guide for Biomedical Scientists

Binding and Kinetics for Molecular Biologists

Career Opportunities in Biotechnology and Drug Development

C. elegans *Atlas*

Experimental Design for Biologists

Fly Pushing: The Theory and Practice of Drosophila *Genetics*, Second Edition

Is It in Your Genes? The Influence of Genes on Common Disorders and Diseases That Affect You and Your Family

Lab Dynamics: Management and Leadership Skills for Scientists, Second Edition

Lab Math: A Handbook of Measurements, Calculations, and Other Quantitative Skills for Use at the Bench

Lab Ref, Volume 1: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

Lab Ref, Volume 2: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

Statistics at the Bench: A Step-by-Step Handbook for Biologists

Preface

Next-generation DNA sequencing (NGS) technology has been a huge stimulus for new and exciting ways to create and test new hypotheses in biology as well as to revisit old ones but with a novel and vastly enhanced perspective. It would be no exaggeration to state that many of the current dynamic advances in biomedical basic and translational science are being driven by this technology.

NGS is enabled by sophisticated and novel bioinformatics tools specifically created or adapted to make NGS possible. Not only has new software been developed for a wide range of novel applications and types of data analysis, but new algorithms have also been developed for old problems, such as sequence alignment and de novo assembly, to cope with the huge volume of data generated on new sequencing machines.

The cycle of software development has accelerated as vendors upgrade their machines and different groups compete to publish new methods and to meet investigator demands. As a result of the frenetic pace of development, new software tools for NGS data analysis are often released with bare bones command line user interfaces and minimal documentation. Making things even more complicated, many different software packages exist for each of the major NGS applications with few benchmarking studies available to guide users in the choice of the best solutions. In short, there is an urgent need for a scientifically rigorous, cutting-edge, and practical treatise to guide researchers about all major aspects of informatics needed to successfully operate and fully take advantage of NGS.

The authors of the present work have been very lucky that their home institution, NYU Langone Medical Center, has invested early and heavily in building both assay and informatics capacity and manpower in NGS. Specifically, in 2008 NYU Langone Medical Center built its Genome Technology Center to provide research and translational scientists access to the latest DNA sequencing, expanding upon previous technologies such as microarrays and real-time polymerase chain reaction (qPCR). In parallel, the Informatics Center at NYU Langone Medical Center has developed the Sequencing Informatics Group to provide research design, upstream data processing, data management, and data analysis consulting for all users of the sequencers within NYU Langone Medical Center and beyond.

As our group has grown in experience, we have evaluated many different software packages and built best practice workflows for many different types of NGS projects, including de novo sequencing (and genome annotation), amplicon sequencing for rare variant detection and for metagenomics, ChIP-seq, RNA-seq, and detection of somatic variants in cancer (including single-base substitutions, insertions, deletions, and translocations).

In this book, building on our own extensive experience that spans collaborations on more than 30 National Institutes of Health–funded projects, and by critically evaluating and synthesizing the literature in the field, we provide an overview of many core types of NGS projects, a discussion of methods embodied in popular software, and detailed descriptions of our own best practice workflows (including several tutorials). We have included advice designed to be helpful to both bioinformaticians implementing their own data analysis methods and to laboratory and clinical investigators planning to use NGS methods to address their own research questions.

The future of NGS and all the related informatics innovations is as bright as it is exciting, and we are gratified to be able to contribute to the field's development with the present volume.

Stuart M. Brown

Acknowledgments

The authors would like to acknowledge the senior leadership of the NYU Langone Medical Center, Dean and CEO Dr. Robert Grossman and the entire executive and scientific leadership team, for creating an exceptionally enabling environment in which we were given the means and the encouragement to pursue our scientific investigations of NGS informatics.

We are also immensely grateful to all our colleagues at NYU Langone Medical Center who have entrusted us with the success of their NGS projects and allowed to us to create, test, and deploy a plethora of innovative informatics solutions across a diverse spectrum of basic science and translational investigations of the highest quality.

We are deeply indebted to John Inglis at Cold Spring Harbor Laboratory Press for seeing the value of this book. We are thankful for the great patience of the CSHLP staff in working with our erratic schedules and the superb quality they bring to every phase of production. We especially thank Inez Sialiano for her editorial guidance in all phases of the writing and Kathleen Bubbeo for insisting on the highest quality graphics and for rooting out errors throughout the text.

The NYU Sequencing Informatics Group:

Alexander Alekseyenko
Constantin Aliferis
Silvia Argimón (affiliate)
Stuart M. Brown
Efstratios Efstathiadis
Frank Hsu (affiliate)
Kranti Konganti
Eric R. Peskin
Christina Schweigert (affiliate)
Steven Shen
Phillip Ross Smith
Alexander Statnikov
Zuojian Tang
Jinhua Wang

About the Authors

Alexander Alekseyenko is Assistant Professor in the Department of Medicine and Associate Operations Director of the Bioinformatics consulting group for the Center for Health Informatics and Bioinformatics, NYU School of Medicine. Dr. Alekseyenko received his Ph.D. in Biomathematics from the University of California at Los Angeles. He conducted postdoctoral training first at the European Bioinformatics Institute, Cambridge, United Kingdom and then at Stanford University. Dr. Alekseyenko is the primary informatics faculty member at NYU working in the area of metagenomics, understanding the diversity of microorganisms present in the human body through next-generation sequencing and the application of evolutionary and ecological statistical models.

Silvia Argimón is Associate Research Scientist in the Cariology and Comprehensive Care Department at NYU College of Dentistry. Her research interests include oral bacteria diversity and virulence. She received her Ph.D. in molecular biology from University of Aberdeen, Scotland.

Stuart M. Brown is Associate Professor in the Cell Biology Department and a senior faculty member in the Center for Health Informatics and Bioinformatics at NYU School of Medicine, where he serves as Operations Director for the Bioinformatics consulting group and leader of the Sequence Informatics group. He has taught graduate courses in Bioinformatics at NYU for 12 years and he is the author of textbooks on bioinformatics and medical genomics. He received his Ph.D. in molecular biology from Cornell University.

Efstratios Efstathiadis is Assistant Professor and the Technical Director of the High Performance Computing Facility of the NYU Langone Medical Center. Previously he served as Technology Architect of the Center for Computational Science at Brookhaven National Laboratory. Dr. Efstathiadis obtained his Ph.D. in nuclear physics in 1996 at the City University of New York.

Jeremy Goecks is a Postdoctoral Fellow in the Departments of Biology and Math & Computer Science at Emory University. He is a core member of the team developing Galaxy, a popular Web-based platform for computational biomedical research. Dr. Goecks earned his Ph.D. in computer science from the Georgia Institute of Technology.

D. Frank Hsu is Clavius Distinguished Professor of Science and Professor of Computer and Information Science at Fordham University. He is former chair of the Fordham Computer Science Department. Dr. Hsu is the former Editor-in-Chief of the *Journal of Interconnection Networks.* He received his Ph.D. from the University of Michigan.

Kranti Konganti is a programmer/bioinformatician in the Center for Health Informatics and Bioinformatics at NYU School of Medicine. He received his M.S. in Bioinformatics from Northeastern University. He has primary responsibility at NYU for data analysis of sequencing performed on the Roche 454 machine as well as genome data visualization in the GBrowse system.

Eric R. Peskin is Associate Technical Director of the High-Performance Computing Facility at the Center for Health Informatics and Bioinformatics, NYU School of Medicine. Dr. Peskin earned his Ph.D. in computer science from the University of Utah. Previously, he served at Intel as a Senior Software Engineer in logic technology development and as Assistant Professor of Electrical Engineering at the Rochester Institute of Technology.

Christina Schweikert is Assistant Professor in the Computer and Information Science Department at Fordham University. Dr. Schweikert obtained her Ph.D. in computer science from the City University of New York.

Steven Shen is Associate Professor in the Department of Biochemistry and the Center for Health Informatics and Bioinformatics at NYU School of Medicine. The primary focus of his work is to develop next-generation sequencing–related technology and computational methods for probing the epigenetic alteration in the genomes of ant species. Before coming to NYU School of Medicine, Dr. Shen was Assistant Professor at Boston University School of Medicine and Research Scientist at Massachusetts Institute of Technology. He also worked at Helicos Biosciences developing single-molecule sequencing technology.

Phillip Ross Smith is Associate Professor in the Cell Biology Department and a senior faculty member in the Center for Health Informatics and Bioinformatics at NYU School of Medicine. He is a former CIO of NYU School of Medicine and a former editor of the *Journal of Structural Biology.* Dr. Smith obtained his Ph.D. in high energy physics from the University of Cambridge, United Kingdom and his M.D. from NYU School of Medicine.

Zuojian Tang is Associate Research Scientist at the Center for Health Informatics and Bioinformatics at NYU School of Medicine. She manages computing support for Illumina Next-Generation Sequencing. She received her M.S. in computer science and bioinformatics from McGill University.

James Taylor is Assistant Professor in the Departments of Biology and Mathematics & Computer Science at Emory University. He is one of the original developers of Galaxy, a popular Web-based platform for computational biomedical research. Dr. Taylor received his Ph.D. in computer science from Pennsylvania State University, where he was involved in several vertebrate genome projects and the ENCODE project.

Jinhua Wang is Assistant Professor at NYU School of Medicine and a member of the NYU Cancer Institute. Dr. Wang completed his Ph.D. training in computational biology and genomics at the Chinese Academy of Sciences. He also served as bioinformatics research manager for the Chinese National Human Genome Center. He conducted postdoctoral research at Cold Spring Harbor Laboratory, where he focused on developing mathematical and statistical methods to identify functional elements in eukaryotic genomes, especially on sequence elements that regulate gene transcription and pre-mRNA splicing. He also served as bioinformatics scientist at St. Jude Children's Research Hospital.

Contents

Preface, vii

Acknowledgments, ix

About the Authors, xi

1 Introduction to DNA Sequencing, 1
Stuart M. Brown

2 History of Sequencing Informatics, 27
Stuart M. Brown

3 Visualization of Next-Generation Sequencing Data, 45
Phillip Ross Smith, Kranti Konganti, and Stuart M. Brown

4 DNA Sequence Alignment, 65
Efstratios Efstathiadis

5 Genome Assembly Using Generalized de Bruijn Digraphs, 83
D. Frank Hsu

6 De Novo Assembly of Bacterial Genomes from Short Sequence Reads, 97
Silvia Argimón and Stuart M. Brown

7 Genome Annotation, 111
Steven Shen

8 Using NGS to Detect Sequence Variants, 129
Jinhua Wang, Zuojian Tang, and Stuart M. Brown

9 ChIP-seq, 139
Zuojian Tang, Christina Schweikert, D. Frank Hsu, and Stuart M. Brown

10 RNA Sequencing with NGS, 171
Stuart M. Brown, Jeremy Goecks, and James Taylor

11 Metagenomics, 187
Alexander Alekseyenko and Stuart M. Brown

12 High-Performance Computing in DNA Sequencing Informatics, 195
Efstratios Efstathiadis and Eric R. Peskin

Glossary, 217
Index, 231

1

Introduction to DNA Sequencing

Stuart M. Brown

HISTORY OF DNA SEQUENCING

All of the DNA sequencing work for the **Human Genome Project** (1995–2003) was performed using modifications of the method invented by Frederick Sanger in 1975 (Sanger and Coulson 1975). Before Sanger's work, some nucleotide sequences were determined using ad hoc methods that involved RNA synthesis and enzymatic digestion. An interesting approximation of the Sanger method was published in 1971 by Ray Wu of Cornell University (Wu and Taylor 1971), where he was able to determine the 12-base single-stranded ends of bacteriophage λ DNA by the addition of complementary radiolabeled nucleotides to the single strand by DNA polymerase, followed by a complex scheme of nuclease digestion and chromatography. Walter Gilbert and Allan Maxam published a 24-bp sequence of the *lac* operator (a transcription repressor binding site) from the *Escherichia coli* genome in 1973 (Gilbert and Maxam 1973). Their method involved a complicated mixture of pyrimidine fingerprinting by partial nuclease digestion and chromatography as well as nuclease digestion of in vitro–transcribed RNA molecules. The entire sequence of the coat protein gene from bacteriophage MS2 was determined by Walter Fiers and coworkers at the University of Ghent, Belgium (Min Jou et al. 1972). This method relied on nuclease digestion of phage RNA, partial in vitro synthesis of RNA by RNA polymerase and incomplete nucleotide mixtures, and chemical characterization of the fragments. But Fiers also used information from the known protein sequence to limit the possible codons and to assemble overlapping fragments.

The sequencing method developed by Sanger in 1975 relies on the synthesis of new **DNA fragments** using DNA polymerase to extend a short synthetic oligonucleotide primer hybridized to a single-stranded DNA template. The first version of Sanger's sequencing method used a two-phase DNA synthesis reaction. In the first phase, the **sequencing primer** was partially extended using a mixture of all four deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP), generating a set of newly synthesized DNA fragments, all starting at the primer but extending for "random" lengths. In the second phase, the partially extended templates were split into four parallel DNA synthesis reactions, each one including only three of the four deoxyribonucleotide triphosphates. "Synthesis then proceeds as far as it can on each chain: thus, if dATP is the missing triphosphate, each chain will terminate at its 3′ end at a position before an A residue" (Sanger and Coulson 1975). The newly synthesized DNA fragments were then denatured from the template and separated by size by electrophoresis in adjacent lanes of an acrylamide gel. "Ideally, the sequence of the DNA is read off from the radioautograph" (Sanger and Coulson 1975) (see Fig. 1).

The **Sanger sequencing method** was revolutionary in several ways. Most importantly, it could be applied to any DNA molecule, and it could be used to determine long DNA sequences. However, this system, as first presented, had two critical limitations that prevented its immediate widespread adoption. First, the requirement for an oligonucleotide primer means that some DNA sequence must be known at a

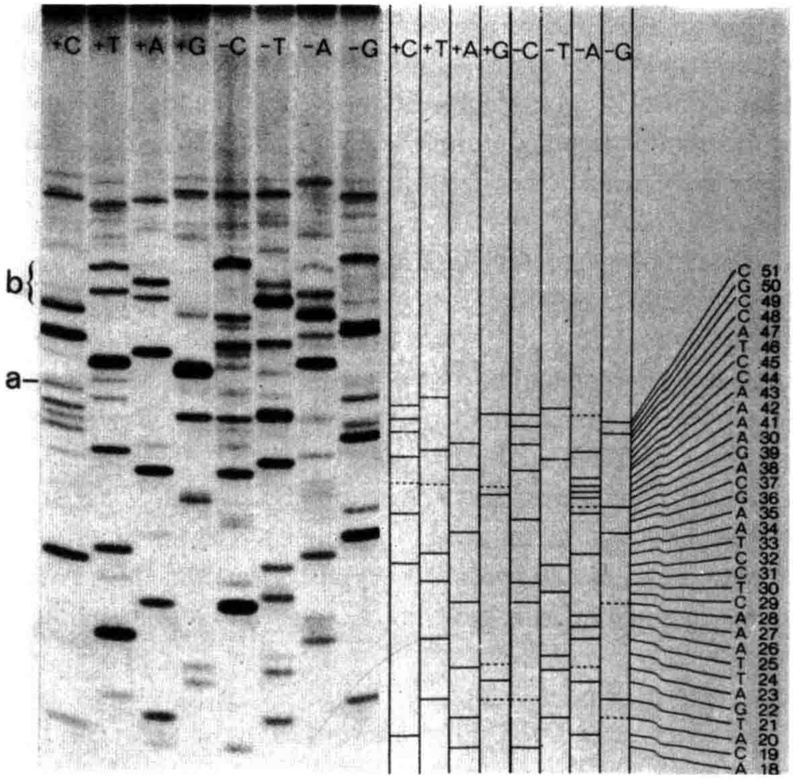


FIGURE 1. The autoradiograph produced by Sanger and Coulson in the 1975 *Journal of Molecular Biology* (**94:** 441–448) paper to document their method for DNA sequencing "by Primed Synthesis with DNA Polymerase."

location directly adjacent to the region of DNA where the sequence is to be determined. Second, the "random extension" of the primer does not necessarily generate an even distribution of fragments of all desired lengths.

Shortly after Sanger's "primer extension" sequencing method was published, Allan Maxam and Walter Gilbert invented a sequencing method based on chemical cleavage of DNA (Maxam and Gilbert 1977). Like the Sanger method, Maxam–Gilbert sequencing splits the DNA template into four reactions. In each reaction, the template is radioactively labeled at the 5′ end, then subjected to chemicals that specifically cleave DNA at one of the four bases. The reactions are conducted under conditions that produce, on average, just one cleavage per DNA molecule, at a random location. Then, like the Sanger method, the four reactions are loaded into adjacent lanes of an acrylamide gel, and fragments are separated by size by electrophoresis. The DNA sequence can then be read from an autoradiograph of the acrylamide gel. Maxam and Gilbert sequencing was initially more popular than Sanger sequencing because it can be conducted directly on purified DNA fragments, with no requirement for a single-stranded template and a complementary oligonucleotide primer.

Sanger rapidly improved his method by using dideoxynucleotides as "chain terminators" in the primer extension reaction in place of the clumsy two-phase procedure described in 1975 (Sanger et al. 1977). The improved method again starts with a single-stranded DNA template hybridized with a short complementary oligonucleotide primer. The primed template is split into four reaction mixtures, each containing DNA polymerase, the four normal deoxyribonucleotide triphosphates (with one radiolabeled nucleotide), and one dideoxynucleotide. As the polymerase extends the primer, whenever a dideoxynucleotide is incorporated, the reaction stops, producing a mixture of truncated fragments of varying lengths, all starting at the same primer and ending with the same base. Once again, the four reactions are loaded onto an acrylamide gel, the fragments are separated by electrophoresis, and the DNA sequence is read off the autoradiograph. Sanger reported reading sequences up to 300 bases long on a single gel.

Sanger sequencing and Maxam–Gilbert sequencing remained in competition for many years. The Sanger method became more popular, possibly because of the complexity of the steps and the toxicity of the reagents used in the Maxam–Gilbert method. Many refinements have been developed to improve the Sanger method, including a variety of cloning methods for the preparation of single-stranded templates that span a gene (or an entire genome) of interest as well as commercial kits to streamline the preparation of reagents. One very significant improvement in the Sanger technology was the development of fluorescent dyes to replace radioactive labels on the newly synthesized DNA fragments (Smith et al. 1986). This led to the development of semiautomated DNA sequencers by Leroy Hood, Michael Hunkapiller, and others that were commercialized by Applied Biosystems

Inc. (ABI). Crucial innovations in the ABI sequencers included attaching the four different colors of dye labels to the four dideoxynucleotide chain terminators, so that fragments terminated at all four bases could be generated in a single reaction tube and assayed on a single lane of an acrylamide gel, and using a computer to monitor a real-time fluorescent detector, so that the sequence data could be collected automatically as the gel electrophoresis was run. These ABI sequencers provided nearly all of the data for the Human Genome Project. Another incremental improvement in the ABI automated fluorescent sequencers was the use of acrylamide gel in capillary tubes rather than in a large thin slab between two glass plates. This saved setup work for technicians in the sequencing laboratory, allowed for more consistent results from electrophoresis, allowed for increased speed of electrophoresis, and allowed the machines to be scaled up to run more samples simultaneously (see Fig. 2).

Cloning for Sequencing

The Sanger sequencing reaction uses a single-stranded DNA template, a short single-stranded oligonucleotide primer that is complementary to the template, DNA polymerase enzyme, and a mixture of chain-extension and chain-terminating nucleotides. The usual strategy to prepare DNA for sequencing involves **cloning** a target fragment of DNA into a plasmid vector that provides a cloning site between binding sites for standard sequencing primers that can be used by single-strand DNA polymerase II. This allows any DNA target to be sequenced in both directions using standard oligonucleotide primers, so that the sequence of the target molecule does not need to be known in advance (see Fig. 3).

DNA sequencing using the Sanger method is capable of reading ~500–800 bases in a single read. This is a limitation of both the Sanger primer extension/chain-terminator chemistry and the ability to separate DNA fragments by electrophoresis with accurate single-base resolution. Because most interesting biological nucleic acid molecules (such as genes, mRNA transcripts, plasmids, and genomes) are much longer than 800 bp, DNA sequencing projects usually involve some strategy of breaking DNA molecules into shorter pieces, sequencing them, and then using bioinformatics tools to assemble the data into complete sequences of the target molecules.

For small sequencing targets, a strategy based on restriction digest fragments may be effective, although it may be difficult to keep track of the size and orientation of all of the fragments for assembly of the sequences. Another strategy developed by Henikoff (1984) involves the generation of progressively smaller fragments of DNA by directional digestion with exonuclease III. The nested sequences are then assembled by overlapping the reads to build a contiguous sequence (**contig**). Because all DNA sequencing methods generate some errors, it became standard procedure to combine several overlapping **sequence reads** over the entire extent of the target