Next-Generation Genome Sequencing

Towards Personalized Medicine

Edited by
Michal Janitz

WILEY-
BLACKWELL

WILEY-VCH Verlag GmbH & Co. KGaA

Next-Generation Genome Sequencing

Edited by
Michal Janitz

Related Titles

Dehmer, M., Emmert-Streib, F. (eds.)

Analysis of Microarray Data

2008
ISBN: 978-3-527-31822-3

Helms, V.

Principles of Computational Cell Biology

2008
ISBN: 978-3-527-31555-1

Knudsen, S.

Cancer Diagnostics with DNA Microarrays

2006
ISBN: 978-0-471-78407-4

Sensen, C. W. (ed.)

Handbook of Genome Research

Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues

2005
ISBN: 978-3-527-31348-8

Next-Generation Genome Sequencing

Towards Personalized Medicine

Edited by
Michal Janitz

WILEY-
BLACKWELL

WILEY-VCH Verlag GmbH & Co. KGaA

The Editor

Dr. Michal Janitz
Max Planck Institute for
Molecular Genetics
Fabeckstr. 60-62
14195 Berlin
Germany

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de.

Composition Thomson Digital, Noida, India
Printing Betz-Druck GmbH, Darmstadt
Bookbinding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN: 978-3-527-32090-5

Contents

Preface

The development of the rapid DNA sequencing method by Fred Sanger and co-workers 30 years ago initiated the process of deciphering genes and eventually entire genomes. The rapidly growing demand for throughput, with the ultimate goal of deciphering the human genome, led to substantial improvements in the technique and was exemplified in automated capillary electrophoresis. Until recently, genome sequencing was performed in large sequencing centers with high automation and many personnel. Even when DNA sequencing reached the industrial scale, it still cost $10 million and 10 years to generate a draft of the human genome. With the price so high, population-based phenotype–genotype linkage studies were small in scale, and it was hard to translate research into statistically robust conclusions. As a consequence, most presumed associations between diseases and particular genes have not stood up to scientific scrutiny. The commercialization of the first massive parallel pyrosequencing technique in 2004 created the first opportunity for the cost-effective and rapid deciphering of virtually any genome. Shortly thereafter, other vendors entered the market, bringing with them a vision of sequencing the human genome for only $1000.

This is the topic of this book. We hope to provide the reader with a comprehensive overview of next-generation sequencing (NGS) techniques and highlight their impact on genome research, human health, and the social perception of genetics.

There is no clear definition of next-generation sequencing. There are, however, several features that distinguish NGS platforms from conventional capillary-based sequencing. First, it has the ability to generate millions of sequence reads rather than only 96 at a time. This process allows the sequencing of an entire bacterial genome within hours or of the *Drosophila melanogaster* genome within days instead of months. Furthermore, conventional vector-based cloning, typical in capillary sequencing, became obsolete and was replaced by direct subjecting of fragmented, and usually, amplified DNA for sequencing. Another distinctive feature of NGS are the sequenced products themselves, which are short-length reads between 30 and 400 bp. The limited read length has substantial impact on certain NGS applications, for instance, *de novo* sequencing. The following chapters will present several innovative approaches, which will combine the obvious advantages of NGS, such as throughput and simplified template preparation, with novel challenging features in terms of short read assembly and large sequencing data storage and processing.

This book arose from the recognition of the need to understand next-generation sequencing techniques and their role in future genome research by the broad scientific community. The chapters have been written by the researchers and inventors who participated in the development and applications of NGS technologies. The first chapter of the book contains an excellent overview on Sanger DNA sequencing, which still remains the gold standard in life sciences. The second and fourth parts of the book describe the commercially available and emerging sequencing platforms, respectively. The third part consists of two chapters highlighting the bottlenecks in the current sequencing: data storage and processing. Once the NGS techniques became available, an unprecedented explosion of applications could be observed. The fifth part of this book provides the reader with the insight into the ever-increasing NGS applications in genome research. Some of these applications are enhancements of existing techniques. Many others are unique to next-generation sequencing marked by its robustness and cost effectiveness, with the prominent example of paleogenomics.

The versatility and robustness of the NGS techniques in studying genes in the context of the entire genome surprised many scientists, including myself. We know that the processes that cause most diseases are not the result of a single genetic failure. Instead, they involve the interaction of hundreds if not thousands of genes. In the past, geneticists have concentrated on genes that have large individual effects when they go wrong, because those effects are so easy to spot. However, combinations of genes that are not individually significant may also be important. It has become evident that next-generation sequencing techniques, together with systems biology approaches, could elucidate the complex dependences of regulatory networks not only on the level of a single cell or tissue but also on the level of the whole organism.

We hope that this book will enrich the understanding of the dramatic changes in genome exploration and its impact not only on research itself but also on many aspects of our life, including healthcare policy, medical diagnostics, and treatment. The best example comes from the field of consumer genomics. Consumer genomics promises to inform people of their risks of developing ailments such as heart disease or cancer; it can even advise its customers how much coffee they can safely drink. This information is retrieved from the correlation of the single nucleotide polymorphism (SNP) pattern of the individual with the SNP haplotype linked to a particular disease. Recent public discussions on the challenges posed by the availability of personal genome information have revealed a new perception of genomic information and its uses. For the first time, a desire to understand the genome has become important and relevant to people outside of the scientific community. In addition to the benefits of having access to genetic information, the ethical and legal risks of making this information available are emerging. The last part of the book introduces the reader to the debate, which will only intensify in the years to come.

In conclusion, I would like to express my sincere gratitude to all of the contributors for their extraordinary effort to present these fascinating technologies and their applications in genome exploration in such a clear and comprehensive way. I also extend my thanks to Professor Hans Lehrach for his constant support.

Berlin, July 2008 *Michal Janitz*

List of Contributors

Annelise E. Barron
Stanford University
Department of Bioengineering
W300B James H. Clark Center
318 Campus Drive
Stanford, CA 94305
USA

Eugene Berezikov
Hubrecht Institute
Uppsalalaan 8
3584 CT Utrecht
The Netherlands

Leonard N. Bloksberg
SLIM Search Ltd.
P.O. Box 106-367
Auckland 1143
New Zealand

Edwin Cuppen
Hubrecht Institute
Uppsalalaan 8
3584 CT Utrecht
The Netherlands

Lei Du
454 Life Sciences
20 Commercial Street
Branford, CT 06405
USA

Tim Durfee
DNASTAR, Inc.
3801 Regent Street
Madison, WI 53705
USA

Jeremy S. Edwards
University of New Mexico Health
Sciences Center
Cancer Research and Treatment Center
Department of Molecular Genetics and
Microbiology
Albuquerque, NM 87131
USA

University of New Mexico
Department of Chemical and Nuclear
Engineering
Albuquerque, NM 87131
USA

Michael Egholm
454 Life Sciences
20 Commercial Street
Branford, CT 06405
USA

Jeppe Emmersen
Aalborg University
Department of Health Science
and Technology
Fredrik Bajers Vej 3B
9000 Aalborg
Denmark

Anthony P. Fejes
Genome Sciences Centre
570 West 7th Avenue, Suite 100
Vancouver, BC
Canada V5Z 4S6

Susan Forrest
University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

Ryan E. Forster
Northwestern University
Materials Science and Engineering
Department
2220 Campus Drive
Evanston, IL 60208
USA

Christopher P. Fredlake
Northwestern University
Chemical and Biological Engineering
Department
2145 North Sheridan, Tech E136
Evanston, IL 60208
USA

M. Thomas P. Gilbert
University of Copenhagen
Biological Institute
Department of Evolutionary Biology
Universitetsparken 10
2100 Copenhagen
Denmark

William Glover
ZS Genetics
8 Hidden Pond Lane
North Reading, MA 01864
USA

Susan H. Hardin
VisiGen Biotechnologies, Inc.
2575 West Bellfort, Suite 250
Houston, TX 77054
USA

Steven J.M. Jones
Genome Sciences Centre
570 West 7th Avenue, Suite 100
Vancouver, BC
Canada V5Z 4S6

Pui-Yan Kwok
University of California, San Francisco
Cardiovascular Research Institute
San Francisco, CA 94143-0462
USA

University of California, San Francisco
Department of Dermatology
San Francisco, CA 94143-0462
USA

Abizar Lakdawalla
Illumina, Inc.
25861 Industrial Boulevard
Hayward, CA 94545
USA

Jeantine E. Lunshof
VU University Medical Center
EMGO Institute
Section Community Genetics
Van der Boechorststraat 7, MF D424
1007 MB Amsterdam
The Netherlands

Artem E. Men
University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

Kåre L. Nielsen
Aalborg University
Department of Biotechnology,
Chemistry and Environmental
Engineering
Sohngaards-Holms vej 49
9000 Aalborg
Denmark

Robert C. Nutter
Applied Biosystems
850 Lincoln Centre Drive
Foster City, CA 94404
USA

Vicki Pandey
Applied Biosystems
850 Lincoln Centre Drive
Foster City, CA 94404
USA

Louise Pape
University of Wisconsin-Madison
Biotechnology Center
Departments of Genetics and Chemistry
Laboratory for Molecular and
Computational Genomics
Madison, WI 53706
USA

Annabeth H. Petersen
Aalborg University
Department of Biotechnology,
Chemistry and Environmental
Engineering
Sohngaards-Holms vej 49
9000 Aalborg
Denmark

Ellen Prediger
Applied Biosystems
850 Lincoln Centre Drive
Foster City, CA 94404
USA

Yijun Ruan
Genome Institute of Singapore
60 Biopolis Street
Singapore 138672
Singapore

David C. Schwartz
University of Wisconsin-Madison
Biotechnology Center
Departments of Genetics and Chemistry
Laboratory for Molecular and
Computational Genomics
Madison, WI 53706
USA

Thomas E. Schwei
DNASTAR, Inc.
3801 Regent Street
Madison, WI 53705
USA

Kirby Siemering
University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

William K. Thomas
Hubbard Center for Genome Studies
448 Gregg Hall, 35 Colovos Road
Durham, NH 03824
USA

Harper VanSteenhouse
Illumina, Inc.
25861 Industrial Boulevard
Hayward, CA 94545
USA

Chia-Lin Wei
Genome Institute of Singapore
60 Biopolis Street
Singapore 138672
Singapore

Peter Wilson
University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

Ming Xiao
University of California, San Francisco
Cardiovascular Research Institute
San Francisco, CA 94143-0462
USA

Shiguo Zhou
University of Wisconsin-Madison
Biotechnology Center
Departments of Genetics and Chemistry
Laboratory for Molecular and
Computational Genomics
Madison, WI 53706
USA

Part One
Sanger DNA Sequencing

1
Sanger DNA Sequencing

Artem E. Men, Peter Wilson, Kirby Siemering, and Susan Forrest

1.1
The Basics of Sanger Sequencing

From the first genomic landmark of deciphering the phiX174 bacteriophage genome achieved by F. Sanger's group in 1977 (just over a 5000 bases of contiguous DNA) to sequencing several bacterial megabase-sized genomes in the early 1990s by The Institute for Genomic Research (TIGR) team, from publishing by the European Consortium the first eukaryotic genome of budding yeast *Saccharomyces cerevisiae* in 1996 to producing several nearly finished gigabase-sized mammal genomes including our own, Sanger sequencing definitely has come a long and productive way in the past three decades. Sequencing technology has dramatically changed the face of modern biology, providing precise tools for the characterization of biological systems. The field has rapidly moved forward now with the ability to combine phenotypic data with computed DNA sequence and therefore unambiguously link even tiny DNA changes (e.g., single-nucleotide polymorphisms (SNPs)) to biological phenotypes. This allows the development of practical ways for monitoring fundamental life processes driven by nucleic acids in objects that vary from single cells to the most sophisticated multicellular organisms.

"Classical" Sanger sequencing, published in 1977 [1], relies on base-specific chain terminations in four separate reactions ("A", "G", "C", and "T") corresponding to the four different nucleotides in the DNA makeup (Figure 1.1a). In the presence of all four 2′- deoxynucleotide triphosphates (dNTPs), a specific 2′,3′-dideoxynucleotide triphosphate (ddNTP) is added to every reaction; for example, ddATP to the "A" reaction and so on. The use of ddNTPs in a sequencing reaction was a very novel approach at the time and gave far superior results compared to the 1975 prototype technique called "plus and minus" method developed by the same team. The extension of a newly synthesized DNA strand terminates every time the corresponding ddNTP is incorporated. As the ddNTP is present in minute amounts, the termination happens rarely and stochastically, resulting in a cocktail of extension

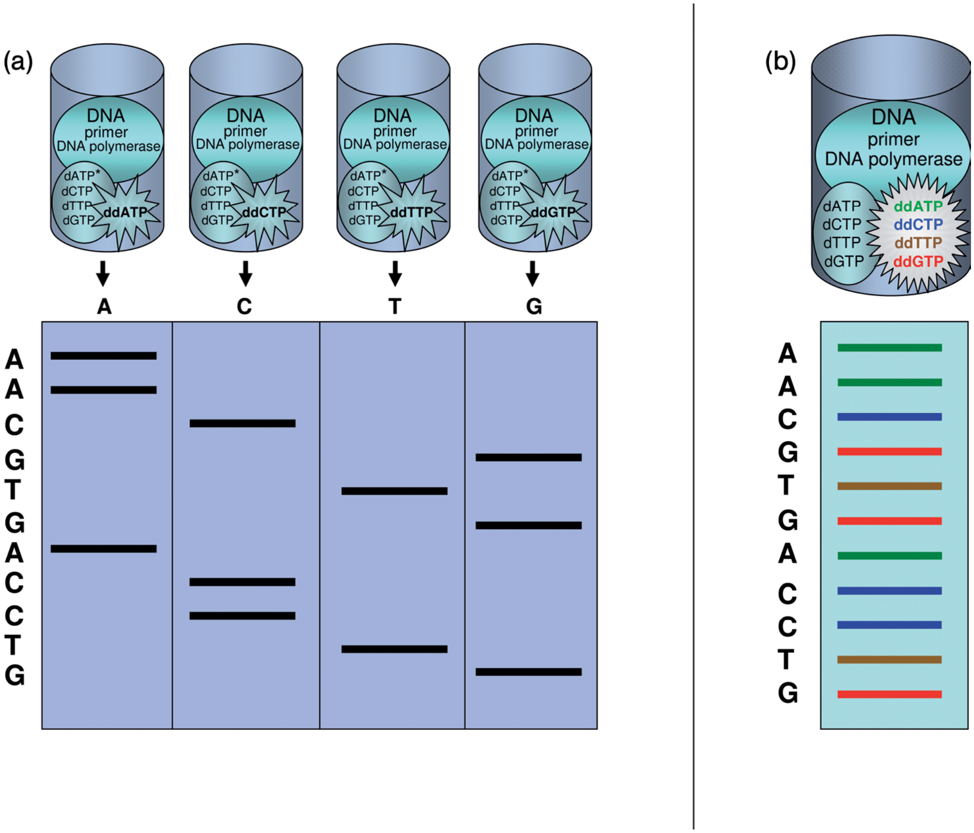

Figure 1.1 Schematic principle of the Sanger sequencing method. (a) Four separate DNA extension reactions are performed, each containing a single-stranded DNA template, primer, DNA polymerase, and all four dNTPs to synthesize new DNA strands. Each reaction is spiked with a corresponding dideoxynucleoside triphosphate (ddATP, ddCTP, ddTTP, or ddGTP). In the presence of dNTPs, one of which is radioactively labeled (in this case, dATP), the newly synthesized DNA strand would extend until the available ddNTP is incorporated, terminating further extension. Radioactive products are then separated through four lanes of a polyacrylamide gel and scored according to their molecular masses. Deduced DNA sequence is shown on the left. (b) In this case, instead of adding radioactive dATP, all four ddNTPs are labeled with different fluorescent dyes. The extension products are then electrophoretically separated in a single glass capillary filled with a polymer. Similar to the previous example, DNA bands move inside the capillary according to their masses. Fluorophores are excited by the laser at the end of the capillary. The DNA sequence can be interpreted by the color that corresponds to a particular nucleotide.

products where every position of an “N” base would result in a matching product terminated by incorporation of ddNTP at the 3′ end.

The second novel aspect of the method was the use of radioactive phosphorus or sulfur isotopes incorporated into the newly synthesized DNA strand through a labeled precursor (dNTP or the sequencing primer), therefore, making every product detectable by radiography. Finally, as each extension reaction results in a very complex mixture of large radioactive DNA products, probably the most crucial achievement was the development of ways to individually separate and detect these molecules. The innovative use of a polyacrylamide gel (PAG) allowed very precise sizing of termination products by electrophoresis followed by *in situ* autoradiography. Later, the autoradiography was partially replaced by less hazardous techniques such as silver staining of DNA in PAGs.

As innovative as they were 30 years ago, slab PAGs were very slow and laborious and could not be readily applied to interrogating large genomes. The next two major technological breakthroughs took place in (i) 1986 when a Caltech team (led by Leroy Hood) and ABI developed an automated platform using fluorescent detection of termination products [2] separating four-color-labeled termination reactions in a single PAG tube and in (ii) 1990 when the fluorescent detection was combined with electrophoresis through a miniaturized version of PAGs, namely, capillaries [3] (Figure 1.1b). Capillary electrophoresis (CE), by taking advantage of a physically compact DNA separation device coupled with laser-based fragment detection, eventually became compatible with 96- and 384-well DNA plate format making highly parallel automation a feasible reality. Finally, the combination of dideoxy-based termination chemistry, fluorescent labeling, capillary separation, and computer-driven laser detection of DNA fragments has established the four elegant "cornerstones" on which modern building of high-throughput Sanger sequencing stands today.

Nowadays, the CE coupled with the development of appropriate liquid-handling platforms allows Sanger sequencing to achieve a highly automatable stage whereby a stand-alone 96-capillary machine can produce about half a million nucleotides (0.5 Mb) of DNA sequence per day. During the late 1980s, a concept of "highly parallel sequencing" was proposed by the TIGR team led by C. Venter and later successfully applied in human and other large genome projects. Hundreds of capillary machines were placed in especially designed labs fed with plasmid DNA clones around the clock to produce draft Sanger reads (Figure 1.2). The need for large volumes of sequence data resulted in the design of "sequencing factories" that had large arrays of automated machines running in parallel together with automated sample preparation pipelines and producing several million reads a month (Figure 1.3). This enabled larger and larger genome projects to be undertaken, culminating with the human and other billion base-sized genome projects.

Along the way, numerous methods were developed that effectively supported template production for feeding high-throughput sequencing pipelines, such as the whole genome shotgun (WGS) approach of TIGR and Celera, or strategies of subgenome sample pooling of YAC, BAC, and cosmid clones based on physical maps of individual loci and entire chromosomes (this strategy was mainly used by the International Human Genome Project team). Not only did the latter methods help to perform sequencing cheaper and faster but also facilitated immensely the genome assembly stage, where the daunting task of putting together hundreds of thousands of short DNA pieces needed to be performed. Some sophisticated algorithms based on paired end sequencing or using large-mapped DNA constructs, such as fingerprinted BACs from physical maps, were developed. Less than 20 years ago,

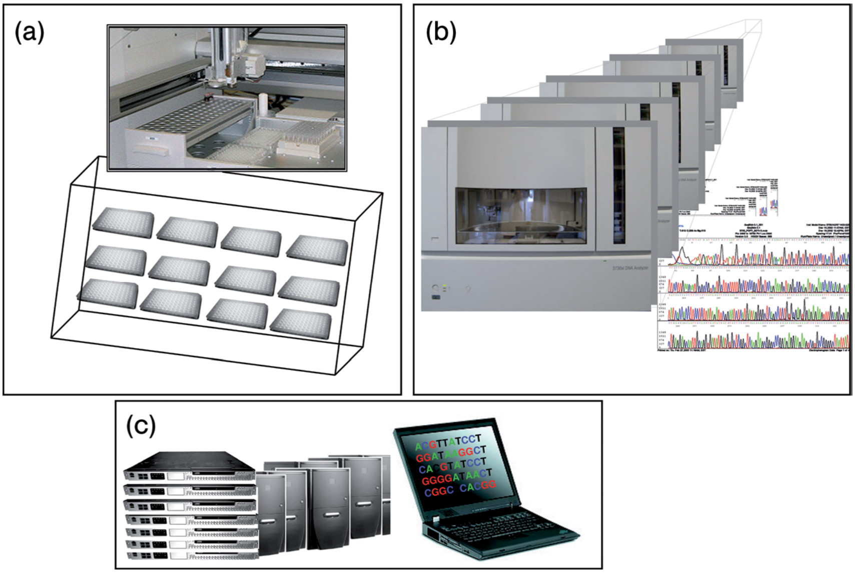

Figure 1.2 Sanger sequencing pipeline. (a) DNA clone preparation usually starts with the isolation of total DNA (e.g., whole genomic DNA from an organism or already fragmented DNA, cDNA, etc.), followed by further fragmentation and cloning into a vector for DNA amplification in bacterial cells. As a result, millions of individual bacterial colonies are produced and individually picked into multiwell plates by liquid-handling robots for isolation of amplified DNA clones. This DNA then goes through a sequencing reaction described in Figure 1.1. (b) Processed sequenced DNA undergoes capillary electrophoresis where labeled nucleotides (bases) are collected and scanned by the laser producing raw sequencing traces. (c) Raw sequencing information is converted into computer files showing the final sequence and quality of every scanned base. The resultant information is stored on dedicated servers and also is usually submitted into free public databases, such as the GeneBank and Trace Archive.

assembling a 1.8 Mb genome of *Haemophilus influenzae* sequenced by the WGS approach [4] was viewed as a computational nightmare, as it required putting together about 25 000 DNA pieces. Today, a typical next-generation sequencing machine (a plethora of which will be described in the following chapters of this book) can produce 100 Mb in just a few hours with data being swiftly analyzed (at least to a draft stage) by a stand-alone computer.

1.2 Into the Human Genome Project (HGP) and Beyond

The HGP, which commenced in 1990, is a true landmark of the capability of Sanger sequencing. This multinational task that produced a draft sequence published in 2001 [5] was arguably the largest biological project ever undertaken. Now, 7 years later, to fully capitalize on and leverage the data from the Human Genome Project, sequencing technologies need to be taken to much higher levels of output to study