

Edited by Michal Janitz

 WILEY-
BLACKWELL

Next-Generation Genome Sequencing

Towards Personalized Medicine



Q78
N567

Next-Generation Genome Sequencing

Towards Personalized Medicine

Edited by
Michal Janitz



**WILEY-
BLACKWELL**



E2009000300

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>.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

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

**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

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
Uppsalaalan 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
Uppsalaalan 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, Gehrman 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

Jeanine 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

Contents

Preface XIII

List of Contributors XVII

Part One Sanger DNA Sequencing 1

1 Sanger DNA Sequencing 3

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

- 1.1 The Basics of Sanger Sequencing 3
- 1.2 Into the Human Genome Project (HGP) and Beyond 6
- 1.3 Limitations and Future Opportunities 7
- 1.4 Bioinformatics Holds the Key 8
- 1.5 Where to Next? 9
- References 10

Part Two Next-Generation Sequencing: Toward Personalized Medicine 13

2 Illumina Genome Analyzer II System 15

Abizar Lakdawalla and Harper VanSteenhouse

- 2.1 Library Preparation 15
- 2.2 Cluster Creation 17
- 2.3 Sequencing 19
- 2.4 Paired End Reads 19
- 2.5 Data Analysis 20
- 2.6 Applications 21
- 2.6.1 Genome Sequencing Applications 23
- 2.6.2 Epigenomics 23
- 2.6.3 Transcriptome Analysis 23
- 2.6.4 Protein–Nucleic Acid Interactions 26
- 2.6.5 Multiplexing 26

2.7	Conclusions	26
	References	27

3 Applied Biosystems SOLiD™ System: Ligation-Based Sequencing 29 *Vicki Pandey, Robert C. Nutter, and Ellen Prediger*

3.1	Introduction	29
3.2	Overview of the SOLiD™ System	29
3.2.1	The SOLiD Platform	30
3.2.1.1	Library Generation	30
3.2.1.2	Emulsion PCR	31
3.2.1.3	Bead Purification	31
3.2.1.4	Bead Deposition	33
3.2.1.5	Sequencing by Ligation	33
3.2.1.6	Color Space and Base Calling	35
3.3	SOLiD™ System Applications	35
3.3.1	Large-Scale Resequencing	35
3.3.2	<i>De novo</i> Sequencing	35
3.3.3	Tag-Based Gene Expression	36
3.3.4	Whole Transcriptome Analysis	37
3.3.5	Whole Genome Resequencing	38
3.3.6	Whole Genome Methylation Analysis	38
3.3.7	Chromatin Immunoprecipitation	39
3.3.8	MicroRNA Discovery	39
3.3.9	Other Tag-Based Applications	40
3.4	Conclusions	40
	References	41

4 The Next-Generation Genome Sequencing: 454/Roche GS FLX 43 *Lei Du and Michael Egholm*

4.1	Introduction	43
4.2	Technology Overview	44
4.3	Software and Bioinformatics	47
4.3.1	Whole Genome Assembly	47
4.3.2	Resequencing and Mutation Detection	47
4.3.3	Ultradeep Sequencing	47
4.4	Research Applications	49
	References	51

5 Polony Sequencing: History, Technology, and Applications 57 *Jeremy S. Edwards*

5.1	Introduction	57
5.2	History of Polony Sequencing	57
5.2.1	Introduction to Polonies	58
5.2.2	Evolution of Polonies	59
5.2.3	Current Applications of the Original Polonies Method	61

5.3	Polony Sequencing	62
5.3.1	Constructing a Sequencing Library	63
5.3.2	Loading the Library onto Beads Using BEAMing	64
5.3.3	Immobilizing the Beads in the Sequencing Flow Cell	65
5.3.4	Sequencing	66
5.3.5	Data Analysis	68
5.4	Applications	69
5.4.1	Human Genome Sequencing	69
5.4.1.1	Requirements of an Ultrahigh-Throughput Sequencing Technology	69
5.4.2	Challenges of Sequencing the Human Genome with Short Reads	70
5.4.2.1	Chromosome Sequencing	72
5.4.2.2	Exon Sequencing	72
5.4.2.3	Impact on Medicine	72
5.4.3	Transcript Profiling	73
5.4.3.1	Polony SAGE	73
5.4.3.2	Transcript Characterization with Polony SAGE	73
5.4.3.3	Digital Karyotyping	75
5.5	Conclusions	75
	References	76

Part Three The Bottleneck: Sequence Data Analysis 77

6	Next-Generation Sequence Data Analysis	79
	<i>Leonard N. Bloksberg</i>	
6.1	Why Next-Generation Sequence Analysis is Different?	79
6.2	Strategies for Sequence Searching	80
6.3	What is a “Hit,” and Why it Matters for NGS?	82
6.3.1	Word Hit	82
6.3.2	Segment Hit	82
6.3.3	SeqID Hit or Gene Hit	82
6.3.4	Region Hit	82
6.3.5	Mapped Hit	83
6.3.6	Synteny Hit	83
6.4	Scoring: Why it is Different for NGS?	83
6.5	Strategies for NGS Sequence Analysis	84
6.6	Subsequent Data Analysis	86
	References	87
7	DNASTAR’s Next-Generation Software	89
	<i>Tim Durfee and Thomas E. Schwei</i>	
7.1	Personalized Genomics and Personalized Medicine	89
7.2	Next-Generation DNA Sequencing as the Means to Personalized Genomics	89

7.3	Strengths of Various Platforms	90
7.4	The Computational Challenge	90
7.5	DNASTAR's Next-Generation Software Solution	91
7.6	Conclusions	94
	References	94

Part Four Emerging Sequencing Technologies 95

8	Real-Time DNA Sequencing	97
	<i>Susan H. Hardin</i>	
8.1	Whole Genome Analysis	97
8.2	Personalized Medicine and Pharmacogenomics	97
8.3	Biodefense, Forensics, DNA Testing, and Basic Research	98
8.4	Simple and Elegant: Real-Time DNA Sequencing	98
	References	101
9	Direct Sequencing by TEM of Z-Substituted DNA Molecules	103
	<i>William K. Thomas and William Glover</i>	
9.1	Introduction	103
9.2	Logic of Approach	104
9.3	Identification of Optimal Modified Nucleotides for TEM Visual Resolution of DNA Sequences Independent of Polymerization	106
9.4	TEM Substrates and Visualization	107
9.5	Incorporation of Z-Tagged Nucleotides by Polymerases	108
9.6	Current and New Sequencing Technology	109
9.7	Accuracy	111
9.8	Advantages of ZSG's Proposed DNA Sequencing Technology	111
9.9	Advantages of Significantly Longer Read Lengths	112
9.9.1	<i>De novo</i> Genome Sequencing	112
9.9.2	Transcriptome Analysis	113
9.9.3	Haplotype Analysis	114
	References	115
10	A Single DNA Molecule Barcoding Method with Applications in DNA Mapping and Molecular Haplotyping	117
	<i>Ming Xiao and Pui-Yan Kwok</i>	
10.1	Introduction	117
10.2	Critical Techniques in the Single DNA Molecule Barcoding Method	118
10.3	Single DNA Molecule Mapping	120
10.3.1	Sequence Motif Maps of Lambda DNA	121
10.3.2	Identification of Several Viral Genomes	123

10.4	Molecular Haplotyping	124
10.4.1	Localization of Polymorphic Alleles Tagged by Single Fluorescent Dye Molecules Along DNA Backbones	125
10.4.2	Direct Haplotype Determination of a Human DNA Sample	127
10.5	Discussion	129
	References	131
11	Optical Sequencing: Acquisition from Mapped Single-Molecule Templates	133
	<i>Shiguo Zhou, Louise Pape, and David C. Schwartz</i>	
11.1	Introduction	133
11.2	The Optical Sequencing Cycle	135
11.2.1	Optical Sequencing Microscope and Reaction Chamber Setup	137
11.2.1.1	Microscope Setup	137
11.2.1.2	Optical Sequencing Reaction Chamber Setup	137
11.2.2	Surface Preparation	137
11.2.3	Genomic DNA Mounting/Overlay	139
11.2.4	Nicking Large Double-Stranded Template DNA Molecules	139
11.2.4.1	Nicking Mounted DNA Template Molecules	139
11.2.4.2	Gapping Nick Sites	139
11.2.5	Optical Sequencing Reactions	140
11.2.5.1	Basic Process	140
11.2.5.2	Choices of DNA Polymerases	140
11.2.5.3	Polymerase-Mediated Incorporations of Multiple Fluorochrome-Labeled Nucleotides	140
11.2.5.4	Washes to Remove Unincorporated Labeled Free Nucleotides and Reduce Background	141
11.2.6	Imaging Fluorescent Nucleotide Additions and Counting Incorporated Fluorochromes	141
11.2.7	Photobleaching	147
11.2.8	Demonstration of Optical Sequencing Cycles	147
11.3	Future of Optical Sequencing	148
	References	149
12	Microchip-Based Sanger Sequencing of DNA	153
	<i>Ryan E. Forster, Christopher P. Fredlake, and Annelise E. Barron</i>	
12.1	Integrated Microfluidic Devices for Genomic Analysis	154
12.2	Improved Polymer Networks for Sanger Sequencing on Microfluidic Devices	156
12.2.1	Poly(<i>N,N</i> -dimethylacrylamide) Networks for DNA Sequencing	156
12.2.2	Hydrophobically Modified Polyacrylamides for DNA Sequencing	159
12.3	Conclusions	160
	References	160