NEXT-GENERATION DNA SEQUENCING INFORMATICS

SECOND EDITION

OTHER TITLES FROM COLD SPRING HARBOR LABORATORY PRESS

A Bioinformatics Guide for Molecular Biologists

A Short Guide to the Human Genome

At the Bench: A Laboratory Navigator, Updated Edition

At the Helm: Leading Your Laboratory, Second Edition

Career Opportunities in Biotechnology and Drug Development

Experimental Design for Biologists, Second Edition

Guide to the Human Genome

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, Second Edition

Molecular Cloning: A Laboratory Manual, Fourth Edition

Navigating Metabolism

Quickstart Molecular Biology: An Introduction for Mathematicians, Physicists, and Computational Scientists

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

Using R at the Bench: Step-by-Step Data Analytics for Biologists

NEXT-GENERATION DNA SEQUENCING INFORMATICS

SECOND EDITION

EDITED BY

STUART M. BROWN

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

NEXT-GENERATION DNA SEQUENCING INFORMATICS, SECOND EDITION

Printed in the United States of America

Publisher and Acquisitions Editor	John Inglis
Director of Editorial Services	Jan Argentine
Project Manager	Inez Sialiano
Permissions Coordinator	Carol Brown
Director of Publication Services	Linda Sussman
Production Editor	Kathleen Bubbeo
Production Manager	Denise Weiss
Cover Designer	Lachina

Front cover artwork: Enormous volumes of DNA and RNA sequence data are produced on tiny surfaces (e.g., Illumina flow cell, *center*) using next-generation sequencing technologies. Making sense of these large and complex data sets requires the sophisticated informatics tools that are made accessible and understandable in this book.

Library of Congress Cataloging-in-Publication Data

Next-generation DNA sequencing informatics / edited by Stuart M. Brown. -- Second edition.
pages cm
Summary: "Next-generation DNA sequencing (NGS) technology has revolutionized biomedical research, making complete genome sequencing an affordable and frequently used tool for a wide variety of research applications. This book provides a thorough introduction to the necessary informatics methods and tools for operating NGS instruments and analyzing NGS data"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-62182-123-6 (hardback)
1. Nucleotide sequence. 2. Bioinformatics. I. Brown, Stuart M., 1962-

QP625.N89N485 2015
572.8'633--dc23

2015012098

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.

Contents

Preface

Next-generation DNA sequencing (NGS) is perhaps the most dynamic and rapidly growing area of modern biology. To keep pace with the rapid developments in sequencing technology, innovative biological applications, and software, it has been necessary to extensively revise the first edition of this book published in 2013. As the cost of sequencing continues to fall and the scope of sequencing projects expands to larger studies and advances into new areas (ecology, taxonomy, population genetics, and chromatin structure, to name a few), data processing and analysis remain key challenges. In addition to extensive revisions capturing new publications in every chapter, new chapters have been added on emerging NGS technologies, quality control, de novo transcriptome assembly, cloud-based informatics, and the nascent field of proteogenomics.

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 volumes of data generated on new sequencing machines. The cycle of software development is very rapid 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 practical guide for researchers about all major aspects of informatics needed to successfully operate and fully take advantage of NGS.

Many of the authors contributing to this book have been based at the New York University Langone Medical Center (NYULMC), which has invested early and heavily in building both assay and informatics capacity and manpower in NGS. Specifically, NYULMC built in 2008 its Genome Technology Center to provide basic 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, an NGS Informatics Group was created to provide research design, upstream data processing, data management, and data analysis consulting for all users of the sequencers within NYULMC and beyond. As the Informatics Group has grown in experience and expanded to include many collaborators, it has tested many different software packages and built best-practice workflows for many NGS projects, including de novo sequencing (and genome annotation), amplicon and exon sequencing for variant detection and for metagenomics, ChIP-seq, RNA-seq, and detection of somatic variants in cancer.

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 updated volume.

Stuart M. Brown

Acknowledgments

Many of the authors make their academic homes at the NYU Langone Medical Center, in its Center for Health Informatics and Bioinformatics (CHIBI). We are thankful for the steady support of 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 especially thankful for the constant support and guidance of Constantin Aliferis, founding Director of CHIBI, who originally suggested that this book be written.

We are also immensely grateful to all our basic and clinical science collaborators who have entrusted us with the analysis of their data for NGS projects and allowed us to create, test, and deploy 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 (CSHLP) 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 discovering and repairing errors throughout the text.

Alexander V. Alekseyenko
Silvia Argimón
Brian D. Athey
Miroslav Blumenberg
Stuart M. Brown
Lisa Cohen
Efstratios Efstathiadis
David Fenyö
Jeremy Goecks
Gerald A. Higgins
D. Frank Hsu
Kranti Konganti
Konstantinos Krampis
Guillermo I. Perez-Perez
Kelly V. Ruggles
Christina Schweikert
Steven Shen
Phillip Ross Smith
Zuojian Tang
Jinhua Wang

About the Authors

Alexander V. Alekseyenko is Associate Professor in the Department of Medicine, New York University School of Medicine and Department of Mathematics, Courant Institute of Mathematical Sciences, New York University. He received his Ph.D. in biomathematics from University of California, Los Angeles and received postdoctoral training at the European Bioinformatics Institute, Cambridge, United Kingdom and the Department of Statistics, Stanford University. Dr. Alekseyenko is Director of the Microbiomics Informatics Laboratory (MIL), which works in the area of understanding the contribution of human microbiome diversity to health and disease through utilization and development of evolutionary and ecological statistical models.

Silvia Argimón is an Associate Research Scientist in the Cariology and Comprehensive Care Department at New York University 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.

Brian D. Athey is the Michael A. Savageau Collegiate Professor and Chair of the Department of Computational Medicine and Bioinformatics at the University of Michigan Medical School. He is also a Professor of Psychiatry and of Internal Medicine. He is the founding Principal Investigator of the NIH National Center for Integrative Biomedical Informatics (NCIBI), one of eight NIH National Biomedical Computing Centers. Brian also serves as CSO of the tranSMART Foundation, an emerging U.S. and E.U. consortium to create and support an open data sharing platform and analytic software. Brian received his Ph.D. in cellular and molecular biology from the University of Michigan.

Miroslav Blumenberg is Associate Professor in the R.O. Perelman Department of Dermatology and the Department of Biochemistry and Molecular Pharmacology at New York University School of Medicine. Dr. Blumenberg received his Ph.D. in organic chemistry from the Massachusetts Institute of Technology. He conducted postdoctoral training first at the Department of Biochemistry with C. Yanofsky, and then at the Department of Genetics with L.L. Cavalli-Sforza, both at Stanford University. Dr. Blumenberg's research interests are focused on the molecular regulation in human epidermal keratinocytes. He cloned several keratin genes, determined their regulatory circuits, defined signaling in epidermal

differentiation as well as in inflammatory processes, and currently is working in "Skinomics" (i.e., bioinformatics approaches in dermatology and skin biology).

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 New York University School of Medicine, where he serves as Operations Director for the Bioinformatics consulting group and leader of the Sequence Informatics group. He is an adjunct professor of Computer Science at Fordham University. 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.

Lisa Cohen is a bioinformatics programmer in the Genome Technology Center at New York University Langone Medical Center, responsible for upstream processing of Illumina sequencing data in addition to microarray and RNAseq data analyses for the core facility. Her research interests are in comparative genomics and de novo genome/transcriptome assembly and annotation of nonmodel species. She received an M.S. in biology from the University of North Carolina at Wilmington.

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

David Fenyö's research focuses on providing a detailed understanding of the dynamics of cellular processes. He applies mathematical, statistical, and computational methods to the analysis of quantitative data and the modeling of biological systems. After receiving a Ph.D. in physics from Uppsala University in Sweden in 1991, he switched the emphasis of his research to bioinformatics, first as a postdoctoral fellow at the Rockefeller University, then as a cofounder of a bioinformatics start-up company, and subsequently as staff scientist and product manager at GE Healthcare. Dr. Fenyö joined the NYU School of Medicine in 2010 and he is currently Associate Professor of Biochemistry and Molecular Pharmacology, Interim Director for the Center for Health Informatics and Bioinformatics, and Graduate Director for the Ph.D. program in biomedical informatics.

Jeremy Goecks is an Assistant Professor of Computational Biology at George Washington University, where he leads a computational genomics research laboratory. He has more than 6 years of experience in bioinformatics, leading efforts to develop novel genome analysis methods and software for a wide variety of biological and biomedical projects. He earned his Ph.D. in computer science from the Georgia Institute of Technology.

Gerald A. Higgins is Vice President of Pharmacogenomic Science for AssureRx Health, Inc. and Adjunct Professor of Computational Medicine and Bioinformatics at the University of Michigan Medical School. He has previously served as Chief of Molecular Neurobiology at

the National Institutes of Health and Vice President of R&D for Hoffman-La Roche. He received doctorates in Neurobiology and Anatomy from the University of Vermont College of Medicine and has a M.D. in psychiatry, although he is not a practicing physician.

D. Frank Hsu is the Clavius Distinguished Professor of Science at Fordham University. He is former chair of the Computer and Information Science Department. Dr. Hsu is on the editorial board of the *Journal of Interconnection Networks, Brain Informatics,* and the *Health Information Science* book series (Springer). He received his Ph.D. degree from the University of Michigan.

Kranti Konganti is Senior Systems Analyst II at Texas A&M Institute for Genome Sciences and Society. He received his M.S. in bioinformatics from Northeastern University. He has primary responsibility for genome annotation, sequence alignment, comparative genomics, and creating and maintaining analysis pipelines. In his previous postion as bioinformatics programmer at NYU Medical Center, he implemented genome data visualization the GBrowse system.

Konstantinos Krampis is Associate Professor in the Department of Biological Sciences and Director of Bioinformatics at the Center for Translational and Basic Research (CTBR), Hunter College, City University of New York. He has implemented a bioinformatics facility at CTBR that will provide the required genomic sequencing and computational infrastructure for addressing health disparities and improving health outcomes through genomic and bioinformatics research. Previously, he was Assistant Professor at J. Craig Venter Institute. As PI of a NIH-NIAID grant he implemented scalable data analysis pipelines on cloud computing platforms for annotation, assembly, and visualization of newly sequenced genomes. He founded the Cloud BioLinux project, the first Virtual Machine with bioinformatics tools on the Amazon cloud.

Guillermo I. Perez-Perez is Associate Professor of Medicine and Microbiology at New York Langone Medical Center. Dr. Perez-Perez received his D.Sc. in medical bacteriology from the Escuela Nacional de Ciencias Biológicas in Mexico. His research interest is in microbiology with special emphasis in medical bacteriology. His most recent work is in the area of the gut and skin microbiome.

Kelly V. Ruggles is a Postdoctoral Fellow in the Center for Health Informatics and Bioinformatics at New York University Medical Center. She is a member of the Computational Proteomics team developing informatics methods to aid in proteomic quantitation and cancer proteogenomics. She received her Ph.D. in metabolic biology, M.S. degree in human nutrition from Columbia University, and B.S. from Cornell University in biological engineering.

Christina Schweikert is an Assistant Professor of Computer Science at St. John's University. She has been granted a prestigious Clare Boothe Luce professorship in computer science. Dr. Schweikert completed her Ph.D. degree in computer science from the City University of New York, Graduate Center, and she has previously taught at Fordham University and

the State University of New York. Dr. Schweikert's research interests include data mining, programming languages, and biomedical and health-care informatics.

Steven S. Shen is the Director for Genomics Bioinformatics at Genome Technology Center, Associate Professor in the Department of Biochemistry and the Center for Health Informatics and Bioinformatics at New York University 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. Before coming to NYUMC, 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 an Associate Professor in the Department of Cell Biology and a senior faculty member in the Center for Health Informatics and Bioinformatics at New York University School of Medicine. He is a former Interim CIO of NYU School of Medicine and a former member of the editorial board 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 New York University.

Zuojian Tang is Research Scientist at the Center for Health Informatics and Bioinformatics at New York University School of Medicine. Her research work has focused on complete downstream data analysis including algorithm development of large amount of biological data. She received her M.S. in computer science and bioinformatics from McGill University.

Jinhua Wang is an 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.