

# METHODS IN MOLECULAR BIOLOGY

*Series Editor*  
**John M. Walker**  
**School of Life Sciences**  
**University of Hertfordshire**  
**Hatfield, Hertfordshire, AL10 9AB, UK**

For further volumes:  
<http://www.springer.com/series/7651>



# **Bacterial Pangenomics**

## **Methods and Protocols**

Edited by

**Alessio Mengoni**

*Department of Biology, University of Florence, Florence, Italy*

**Marco Galardini**

*EMBL-EBI, Cambridge, UK*

**Marco Fondi**

*Department of Biology, University of Florence, Florence, Italy*

*Editors*

Alessio Mengoni  
Department of Biology  
University of Florence  
Florence, Italy

Marco Galardini  
EMBL-EBI  
Cambridge, UK

Marco Fondi  
Department of Biology  
University of Florence  
Florence, Italy

ISSN 1064-3745                      ISSN 1940-6029 (electronic)  
ISBN 978-1-4939-1719-8        ISBN 978-1-4939-1720-4 (eBook)  
DOI 10.1007/978-1-4939-1720-4  
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014951682

© Springer Science+Business Media New York 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is a brand of Springer  
Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

---

## Preface

From a pioneering field a decade ago, now bacterial genomics is a mature research interdisciplinary field, which is approached by ecologists, geneticists, bacteriologists, molecular biologists, and evolutionary biologists working in medical, industrial, and basic science. The high diffusion of bacterial genomics in many different fields has been helped by the low costs of genome and transcriptome sequencing performed by the so-called Next Generation Sequencing (NGS) technologies. Now, the cost of a draft bacterial genome sequence is as low as few hundreds of Euro (or Dollars). This low cost is allowing many laboratories to perform genome sequencing of virtually every “interesting” bacterial strain they have in hand. In parallel, bioinformatic analysis of the data has grown and the specialized bioinformatician is an obliged professional figure in every laboratory that is interested in genome sequencing.

One of the most striking differences of bacterial genomics with respect to the genomics of eukaryotic multicellular organisms is the concept of pangenome, which was introduced in the late 2005 by researchers working on bacterial pathogenic species. The pangenome is defined as a genomic approximation to describe a species’ genome in terms of the sum of core (conserved in all strains) and dispensable (variable among strains) genes. For bacterial species, the pangenome concept is particularly relevant since closely related strains usually show large differences in gene content between them. Consequently, when speaking about bacterial genomics, often people are referring to comparative analysis of bacterial genomes and then to what we can call “bacterial pangenomics.” Understanding which genetic components of this large pangenomic variability are functionally, clinically, or evolutionary relevant is a challenging task; in fact, a large fraction of the dispensable genome is found to have a poor functional characterization. The availability of powerful and precise analysis tools is therefore of paramount importance.

Thanks to the large diffusion of bacterial genome analysis (or bacterial pangenomic studies), the present book is intended to provide the most recent methodologies about the study of bacterial pangenomes. Three major areas are covered, namely the experimental methods for approaching bacterial pangenomics (“Preparing the bacterial pangenome”), the bioinformatic pipelines for analysis and annotation of sequence data (“Defining the pangenome”), and finally the methods for inferring functional and evolutionary features from the pangenome (“Interpreting the pangenome”). In each of these sections, researchers from both academia and private leading companies of NGS and bioinformatic analysis (as Beijing Genome Institute, Life Technologies, Era7 Bioinformatics) are providing the most up-to-date protocols and procedures for bacterial genome analysis, from assessment of genome size and structure to the analysis of raw sequence data and their annotation and biological interpretation in terms of gene activity and metabarcoding diversity and genome evolution.

The aim of the present book is then to serve as a “field guide” both for qualified investigators on bacterial genomics who want to update their technical knowledge and for less-experienced researchers who want to start working with bacterial genomics and pangenomics.

Additionally, the book could serve to graduate students as a manual of methods used in bacterial pangenomics and as a supplemental textbook in classes of genomics and bioinformatics.

*Florence, Italy*  
*Florence, Italy*  
*Cambridge, UK*

*Alessio Mengoni*  
*Marco Fondi*  
*Marco Galardini*

---

## Contents

|   |           |
|---|-----------|
| <i>Preface</i> . . . . .  | <i>v</i>  |
| <i>Contributors</i> . . . . .   | <i>ix</i> |
| 1 Pulsed Field Gel Electrophoresis and Genome Size Estimates . . . . .<br><i>Rosa Alduina and Annalisa Pisciotta</i>  | 1         |
| 2 Comparative Analyses of Extrachromosomal Bacterial Replicons,<br>Identification of Chromids, and Experimental Evaluation<br>of Their Indispensability . . . . .<br><i>Lukasz Dziewit and Dariusz Bartosik</i>   | 15        |
| 3 Choice of Next-Generation Sequencing Pipelines . . . . .<br><i>F. Del Chierico, M. Ancora, M. Marcacci, C. Cammà, L. Putignani,<br/>and Salvatore Conti</i>   | 31        |
| 4 The Pyrosequencing Protocol for Bacterial Genomes. . . . .<br><i>Ermanno Rizzi</i>  | 49        |
| 5 Bacterial Metabarcoding by 16S rRNA Gene Ion<br>Torrent Amplicon Sequencing. . . . .<br><i>Elio Fantini, Giulio Gianese, Giovanni Giuliano, and Alessia Fiore</i>   | 77        |
| 6 The Illumina-Solexa Sequencing Protocol for Bacterial Genomes . . . . .<br><i>Zhenfei Hu, Lei Cheng, and Hai Wang</i>   | 91        |
| 7 High-Throughput Phenomics . . . . .<br><i>Carlo Viti, Francesca Decorosi, Emmanuela Marchi, Marco Galardini,<br/>and Luciana Giovannetti</i>  | 99        |
| 8 Comparative Analysis of Gene Expression: Uncovering<br>Expression Conservation and Divergence Between <i>Salmonella</i><br><i>enterica</i> Serovar Typhimurium Strains LT2 and 14028S . . . . .<br><i>Paolo Sonogo, Pieter Meysman, Marco Moretto, Roberto Viola,<br/>Kris Laukens, Duccio Cavalieri, and Kristof Engelen</i> | 125       |
| 9 Raw Sequence Data and Quality Control . . . . .<br><i>Giovanni Bacci</i>  | 137       |
| 10 Methods for Assembling Reads and Producing Contigs. . . . .<br><i>Valerio Orlandini, Marco Fondi, and Renato Fani</i>  | 151       |
| 11 Mapping Contigs Using CONTIGuator . . . . .<br><i>Marco Galardini, Alessio Mengoni, and Marco Bazzicalupo</i>  | 163       |
| 12 Gene Calling and Bacterial Genome Annotation with BG7 . . . . .<br><i>Raquel Tobes, Pablo Pareja-Tobes, Marina Manrique,<br/>Eduardo Pareja-Tobes, Evdokim Kovach, Alexey Alekhin,<br/>and Eduardo Pareja</i>  | 177       |

|    |   |     |
|----|---|-----|
| 13 | Defining Orthologs and Pangenome Size Metrics . . . . .   | 191 |
|    | <i>Emanuele Bosi, Renato Fani, and Marco Fondi</i>  |     |
| 14 | Robust Identification of Orthologues and Paralogues<br>for Microbial Pan-Genomics Using GET_HOMOLOGUES:<br>A Case Study of pInCA/C Plasmids . . . . . | 203 |
|    | <i>Pablo Vinnuesa and Bruno Contreras-Moreira</i>   |     |
| 15 | Genome-Scale Metabolic Network Reconstruction . . . . .   | 233 |
|    | <i>Marco Fondi and Pietro Liò</i>   |     |
| 16 | From Pangenome to Panphenome and Back . . . . .   | 257 |
|    | <i>Marco Galardini, Alessio Mengoni, and Stefano Mocali</i>   |     |
| 17 | Genome-Wide Detection of Selection and Other Evolutionary Forces . . . . .  | 271 |
|    | <i>Zhuofei Xu and Rui Zhou</i>  |     |
| 18 | The Integrated Microbial Genome Resource of Analysis . . . . .  | 289 |
|    | <i>Alice Checcucci and Alessio Mengoni</i>  |     |
|    | Erratum to . . . . .  | E1  |
|    | <i>Index</i> . . . . .  | 297 |



---

## Contributors

- ROSA ALDUINA • *Department of Biological, Chemical and Pharmaceutical Sciences and Technologies, University of Palermo, Palermo, Italy*
- ALEXEY ALEKHIN • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- M. ANCORA • *Istituto Zooprofilattico Sperimentale dell’Abruzzo e Molise “G. Caporale”, National and OIE Reference Laboratory for Brucellosis, Teramo, Italy*
- GIOVANNI BACCI • *Department of Biology, University of Florence, Florence, Italy; Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Centro di Ricerca per lo Studio delle Relazioni tra Pianta e Suolo (CRA-RPS), Rome, Italy*
- DARIUSZ BARTOSIK • *Institute of Microbiology, Department of Bacterial Genetics, University of Warsaw, Warsaw, Poland*
- MARCO BAZZICALUPO • *Department of Biology, University of Florence, Florence, Italy*
- EMANUELE BOSI • *Department of Biology, University of Florence, Florence, Italy*
- C. CAMMÀ • *Istituto Zooprofilattico Sperimentale dell’Abruzzo e Molise “G. Caporale”, National and OIE Reference Laboratory for Brucellosis, Teramo, Italy*
- DUCCIO CAVALIERI • *Department of Computational Biology, Fondazione Edmund Mach, San Michele all’Adige, Trento, Italy*
- ALICE CHECCUCCI • *Department of Biology, University of Florence, Florence, Italy*
- LEI CHENG • *BGI, Shenzhen, China*
- F. DEL CHIERICO • *Unit of Parasitology and Unit of Metagenomics, Bambino Gesù Children’s Hospital, IRCCS, Rome, Italy*
- SALVATORE CONTI • *Thermo Fisher Scientific, Monza, Italy*
- BRUNO CONTRERAS-MOREIRA • *Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico; Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC), Zaragoza, Spain; Fundación ARAID, Zaragoza, Spain*
- FRANCESCA DECOROSI • *Dipartimento di Scienze delle Produzioni Agroalimentari e dell’Ambiente (DISPAA), University of Florence, Florence, Italy*
- LUKASZ DZIEWIT • *Department of Bacterial Genetics, Institute of Microbiology, University of Warsaw, Warsaw, Poland*
- KRISTOF ENGELEN • *Department of Computational Biology, Fondazione Edmund Mach, San Michele all’Adige, Trento, Italy*
- RENATO FANI • *Department of Biology, University of Florence, Florence, Italy*
- ELIO FANTINI • *Italian National Agency for New technologies, Energy and Sustainable development, Rome, Italy*
- ALESSIA FIORE • *Italian National Agency for New technologies, Energy and Sustainable development, Rome, Italy*
- MARCO FONDI • *Department of Biology, University of Florence, Florence, Italy*
- MARCO GALARDINI • *EMBL-EBI, Cambridge, UK*
- GIULIO GIANESE • *Italian National Agency for New technologies, Energy and Sustainable development, Rome, Italy*
- LUCIANA GIOVANNETTI • *Dipartimento di Scienze delle Produzioni Agroalimentari e dell’Ambiente (DISPAA), University of Florence, Florence, Italy*

- GIOVANNI GIULIANO • *Italian National Agency for New technologies, Energy and Sustainable development, Rome, Italy*
- ZHENFEI HU • *BGI, Shenzhen, China*
- EVDOKIM KOVACH • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- KRIS LAUKENS • *Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium; Biomedical Informatics Research Center Antwerp (biomina), Antwerp University Hospital, University of Antwerp, Edegem, Belgium*
- PIETRO LIÒ • *Computer Laboratory, University of Cambridge, Cambridge, UK*
- MARINA MANRIQUE • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- M. MARCACCI • *Istituto Zooprofilattico Sperimentale dell'Abruzzo e Molise "G. Caporale", National and OIE Reference Laboratory for Brucellosis, Teramo, Italy*
- EMMANUELA MARCHI • *Dipartimento di Scienze delle Produzioni Agroalimentari e dell'Ambiente (DISPAA), University of Florence, Florence, Italy*
- ALESSIO MENGONI • *Department of Biology, University of Florence, Florence, Italy*
- PIETER MEYSMAN • *Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium; Biomedical Informatics Research Center Antwerp (biomina), Antwerp University Hospital, University of Antwerp, Edegem, Belgium*
- STEFANO MOCALI • *Consiglio per la Ricerca e la sperimentazione in Agricoltura, Centro di Ricerca per l'Agrobiologia e la Pedologia (CRA-ABP), Florence, Italy*
- MARCO MORETTO • *Department of Computational Biology, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy*
- VALERIO ORLANDINI • *Department of Biology, University of Florence, Florence, Italy; Department of Protein Biochemistry, National Research Council, Napoli, Italy*
- EDUARDO PAREJA • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- PABLO PAREJA-TOBES • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- EDUARDO PAREJA-TOBES • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- ANNALISA PISCIOTTA • *Department of Biological, Chemical and Pharmaceutical Sciences and Technologies, University of Palermo, Palermo, Italy*
- L. PUTIGNANI • *Unit of Parasitology and Unit of Metagenomics, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy*
- ERMANNO RIZZI • *Institute for Biomedical Technologies (ITB), National Research Council (CNR), Segrate, MI, Italy*
- PAOLO SONEGO • *Department of Computational Biology, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy*
- RAQUEL TOBES • *Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain*
- PABLO VINUESA • *Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico*
- ROBERTO VIOLA • *Department of Computational Biology, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy*
- CARLO VIII • *Dipartimento di Scienze delle Produzioni Agroalimentari e dell'Ambiente (DISPAA), University of Florence, Florence, Italy*
- HAI WANG • *BGI, Shenzhen, China*
- ZHUOFEI XU • *Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark*
- RUI ZHOU • *Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark*