Edited by:
Jacques Izard and Maria C. Rivera

Metagenomics for Microbiology

Metagenomics for Microbiology

Edited by

Jacques Izard

The Forsyth Institute, Cambridge,
Massachusetts, USA

Maria C. Rivera

Virginia Commonwealth University,
Center for the Study of Biological Complexity,
Richmond, Virginia, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

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

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

ISBN: 978-0-12-410472-3

Metagenomics for Microbiology

LIST OF CONTRIBUTORS

Nadim J. Ajami

Department of Molecular Virology and Microbiology, The Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX, USA; Metanome Inc., Houston, TX, USA

Mathieu Almeida

Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA

Brett Bowman

Pacific Biosciences, Menlo Park, CA, USA

Erika del Castillo

Department of Microbiology, The Forsyth Institute, Cambridge, MA; Harvard School of Dental Medicine, Boston, MA, USA

Yong-Joon Cho

ChunLab, Inc. Seoul National University, Seoul, Korea

Georg K. Gerber

Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Shaomei He

Department of Bacteriology and Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

Jacques Izard

Department of Microbiology, The Forsyth Institute, Cambridge, MA; Harvard School of Dental Medicine, Boston, MA, USA

Mincheol Kim

School of Biological Sciences, Seoul National University, Seoul, Korea

Jonas Korlach

Pacific Biosciences, Menlo Park, CA, USA

Patricio S. La Rosa

Predictive Analytics, Monsanto, St. Louis, MI, USA

Joseph F. Petrosino

Department of Molecular Virology and Microbiology, The Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX, USA; Metanome Inc., Houston, TX, USA

Mihai Pop

Department of Computer Science; Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA

Maria C. Rivera

Department of Biology, Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, USA

Matthias Scholz

Centre for Integrative Biology, University of Trento, Trento, Italy

Nicola Segata

Centre for Integrative Biology, University of Trento, Trento, Italy

William D. Shannon

Department of Medicine, Washington University School of Medicine, St. Louis, MI, USA; BioRankings, LLC, St. Louis, MI, USA

Erica Sodergren

The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

Adrian Tett

Centre for Integrative Biology, University of Trento, Trento, Italy

George Weinstock

The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

Yanjiao Zhou

The Genome Institute, Washington University School of Medicine, St. Louis, MI, USA; Department of Pediatrics, Washington University School of Medicine, St. Louis, MI, USA

PREFACE

It is well known that only a small fraction of extant microbial life has been identified. Metagenomics, the direct sequencing and characterization of genes and genomes present in complex microbial ecosystems (e.g., metagenomes), has revolutionized the practice of microbiology by bypassing the hurdle of pure culture isolation. Metagenomics shows promise of advancing our understanding of the diversity, function, and evolution of the uncultivated majority.

Metagenomics as a field arose in the 1990s after the application of molecular biology techniques to genomic material directly extracted from microbial assemblages present in diverse habitats, including the human body. The application of metagenomic approaches allows for the acquisition of genetic/genomic information from the viruses, bacteria, archaea, fungi, and protists forming complex assemblages. The field of metagenomics addresses the fundamental questions of which microbes are present and what their genes are potentially doing.

In the mid-2000s, the availability of high-throughput or next-generation sequencing technologies propelled the field by lowering the monetary and time constraints imposed by traditional DNA sequencing technologies. These advances have allowed the scientific community to examine the microbiome of diverse environments/habitats, follow spatial and temporal changes in community structure, and study the response of the communities to treatment or environmental modifications.

In 2012, the publication of the large-scale characterization of the microbiome of healthy adults created high expectations about the influence of the microbiota in human health and disease. With the publication of the results of the Human Microbiome Project, metagenomics has emerged as a major research area in microbiology, particularly, when it comes to the characterization of the role of microbiota in complex disorders, such as obesity.

With contributions by leading researchers in the field, we provide a series of chapters describing best practices for the collection and analysis of metagenomic data, as well as the promises and challenges of the field. The chapters have been dedicated to different aspects of metagenomics. Chapter 1 provides an end-to-end overview of the metagenomic pipeline and its challenges. Chapter 2 showcases SMRT, one of the third-generation sequencing platforms, and its use in metagenomics. As high abundance of ribosomal RNA (rRNA) transcripts is a major hurtle for the application of transcriptomics to microbial communities, Chapter 3 describes methodology that can reduce the "noise" rRNA imposes on this type of studies. Chapters 4 and 5 showcase some of the computational approaches that are used to analyze the whole-community metagenome sequence data and available software, and highlight future research directions. The statistical challenges and solutions for cross-sectional and longitudinal data sets are explored in Chapters 6 and 7, respectively. Chapter 8 presents a historical perspective of the microbiome studies, the societal impact of microbial communities, and the challenges ahead for metagenomics, while advances in virome studies are explored in Chapter 9. A perspective on the current efforts, challenges, and the future of metagenomic is presented in Chapter 10.

This book is intended for researchers, teachers, students, and the citizen scientists contemplating performing microbial metagenomics studies. For microbiologists generating metagenomic next-generation sequencing data, the book will provide an introduction and support to the computational and statistical specifics of the data. For the statisticians and computational scientist contemplating working with metagenomic data, it will provide some of the initial background needed. For the community, in general, it will provide the basis for further investigation of this transformative and fascinating field.

We would like to thank all authors for their contributions. We need to acknowledge the public and private funding entities that made this technological and conceptual advance a possibility, as well as the researchers and consortia that broke the grounds for those innovations to flourish. Last, we would like to thank Elsevier for the short book format and allowing a more focused and didactic approach.

Jacques Izard
Maria C. Rivera

CONTENTS

CHAPTER 1

Steps in Metagenomics: Let's Avoid Garbage in and Garbage Out

Jacques Izard

WHY METAGENOMICS?

Is metagenomics a revolution or a new fad? Metagenomics is tightly associated with the availability of next-generation sequencing in all its implementations. The key feature of these new technologies, moving beyond the Sanger-based DNA sequencing approach, is the depth of nucleotide sequencing per sample.[1] Knowing much more about a sample changes the traditional paradigms of "What is the most abundant?" or "What is the most significant?" to "What is present and potentially significant that might influence the situation and outcome?"

Let's take the case of identifying proper biomarkers of disease state in the context of chronic disease prevention. Prevention has been deemed as a viable option to avert human chronic diseases and to curb healthcare management costs.[2] The actual implementation of any effective preventive measures has proven to be rather difficult. In addition to the typically poor compliance of the general public, the vagueness of the successful validation of habit modification on the long-term risk, points to the need of defining new biomarkers of disease state.

Scientists and the public are accepting the fact that humans are superorganisms, harboring both a human genome and a microbial genome, the latter being much bigger in size and diversity, and key for the health of individuals.[3,4] It is time to investigate the intricate relationship between humans and their associated microbiota and how this relationship modulates or affects both partners.[5] These remarks can be expanded to the animal and plant kingdoms, and holistically to the Earth's biome. By its nature, the evolution and function of all the Earth's biomes are influenced by a myriad of interactions between and among microbes (planktonic, in biofilms or host associated) and the surrounding physical environment.

Metagenomics for Microbiology. http://dx.doi.org/10.1016/B978-0-12-410472-3.00001-4

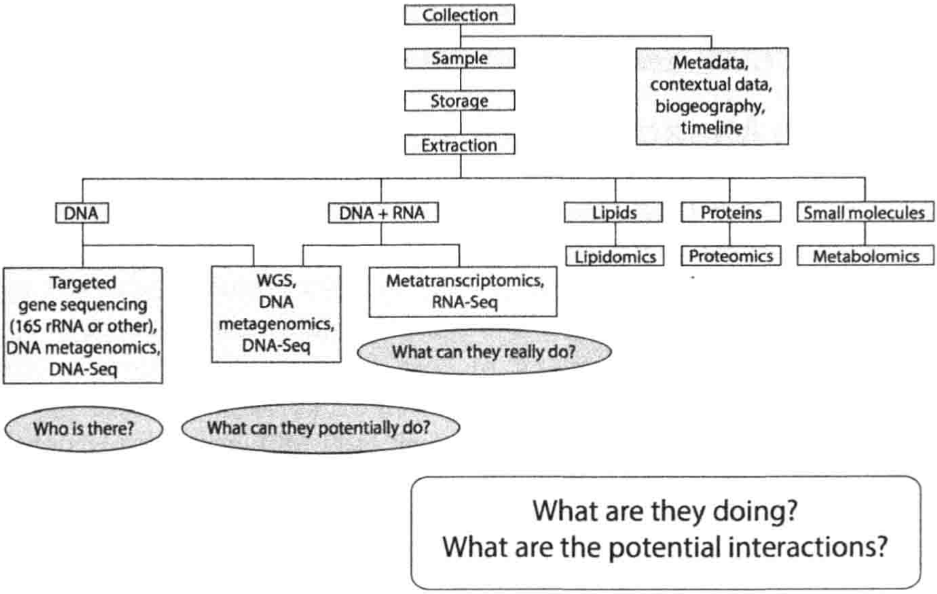
Collection
Sample
Storage
Extraction
Metadata, contextual data, biogeography, timeline
DNA
DNA + RNA
Lipids
Proteins
Small molecules
Lipidomics
Proteomics
Metabolomics
Targeted gene sequencing (16S rRNA or other), DNA metagenomics, DNA-Seq
WGS, DNA metagenomics, DNA-Seq
Metatranscriptomics, RNA-Seq
What can they really do?
Who is there?
What can they potentially do?
What are they doing?
What are the potential interactions?

Fig. 1.1. Metagenomic analysis process and some of the overarching questions that can be answered by the different methodologies.

The general definition of metagenomics is the cultivation-independent analysis of the genetic information of the collective genomes of the microbes within a given environment based on its sampling. It focuses on the collection of genetic information through sequencing that can target DNA, RNA, or both. The subsequent analyses can be solely focused on sequence conservation, phylogenetic, phylogenomic, function, or genetic diversity representation including yet-to-be annotated genes. The diversity of hypotheses, questions, and goals to be accomplished is endless. The primary design is based on the nature of the material to be analyzed and its primary function (Figure 1.1).

IT ALL STARTS WITH THE STUDY DESIGN

The goal is not to tell you how to do your science but to emphasize some aspects of study design that need careful attention because of the characteristics of the methodologies used in metagenomic studies. It begins by identifying the primary objective of the metagenomics project. What is the main scientific question you are trying to answer? More than one hypothesis can be tested depending on the scope of the experiment and the amount of associated data, or metadata, that you collect and use for your subsequent analyses.

The high-dimensionality characteristic of the metagenomics datasets is challenging and is revolutionizing microbiology analytical methodology. What is meant by high-dimensional dataset? Let's take as an example the Human Microbiome Project (HMP) 16S ribosomal RNA (rRNA)-based characterization of 10 sites from the digestive tract of 200 individuals. Such analysis required the collection of over 2000 samples, generating approximately 23 million high-quality sequence reads that were assigned to 674 taxonomic clades with their respective relative abundance per taxonomic level (e.g., from phylum to genus). For example, for the genus *Pyramidobacter*, the database stores the relative abundance at each taxonomic level, from the phylum (e.g., "Bacteria|Synergistetes"), the most inclusive taxonomic level, to the genus (e.g., "Bacteria|Synergistetes|Synergistia|Synergistales|Synergistaceae|*Pyramidobacter*"), the least inclusive taxonomic level, and all the taxonomic levels between the two.[6] From the same study, four body sites were further analyzed using whole metagenome shotgun (WMS) sequencing from approximately 100 individuals, generating a trillion nucleotides.[6] Another example can be extracted from the work of Giannoukos et al.[7] while developing rRNA depletion methodology for fecal samples. They obtained over 100,000 reads per sample.[7] In each example, each sample has a tremendous amount of genotypic and phenotypic information in addition to the metadata (e.g., age, sex, race, and others). In addition to the nucleotide data, information about other molecules (e.g., lipids, proteins, and metabolites) can be collected; increasing the complexity and multidimensionality of the dataset. The type of data collected will determine the type of analyses performed. These analyses can help answer questions such as: "What are the organisms present?", "What can these organisms potentially do?",' "What is their metabolic capability?", and "How do they influence the host?" (Figure 1.1). Planning the structure of samples and metadata acquisition as well as the analysis pipeline to be used, prior to the start of the experiment, will avoid bottlenecks and optimize utilization of funds.

During the study design phase, investigators need to take into consideration the ethical and legal issues related to metagenomics data collection and analysis. Some of the constrains of metagenomics studies utilizing human subjects include Institutional Review Boards, informed consent, and other issues related to the protection of the identifiable health information of the human subjects (e.g., HIPAA Privacy Rule in the United States). For examples of consent documentation and standard operating procedures, the National Institutes of Health HMP has made those document public and available online (http://www.hmpdacc.org).[8] It is essential for the consent procedures to accurately state what data will be gathered, how it will be used, and how it will be stored. All efforts should be made to secure information and confidentiality of the genetic material and associated data over time. This includes both the physical storage of the information, data deposition and data sharing, even when the samples are de-identified. For environmental samples, having the right of access and sampling permits is critical as geolocation is now required with the sample data submission to repository.

It is important to point out that any samples collected from a host will contain a significant amount of the host genetic material. The potential contamination of samples with the host genetic material adds to the complexity of the metagenomics studies, and sophisticated computational pipelines for the removal of the contaminating reads are essential to generate meaningful conclusions and, in the case of human subjects, to protect the privacy and confidentiality of the sample donor. Figure 1.2 shows the impact of human "contamination" on the amount and quality of the data collected using shotgun sequencing of human samples from 16 different body sites.[8] When working with different models, it should be noted that the genome of a brown rat is not that much smaller than that of a human (over 3 billion base pairs), and that the corn genome is over 2 billion base pairs. Although protists and fungi are much smaller, their genomes are still composed of few million base pairs. The knowledge of your biological system of interest will be critical to optimize the study design.

HAVE A STATISTICAL ANALYSIS PLAN IN PLACE BEFORE STARTING

Planning for statistical analysis should be an integral part of the study design. Although many experimental designs can be performed in metagenomics project, there is no single path to a successful strategy.

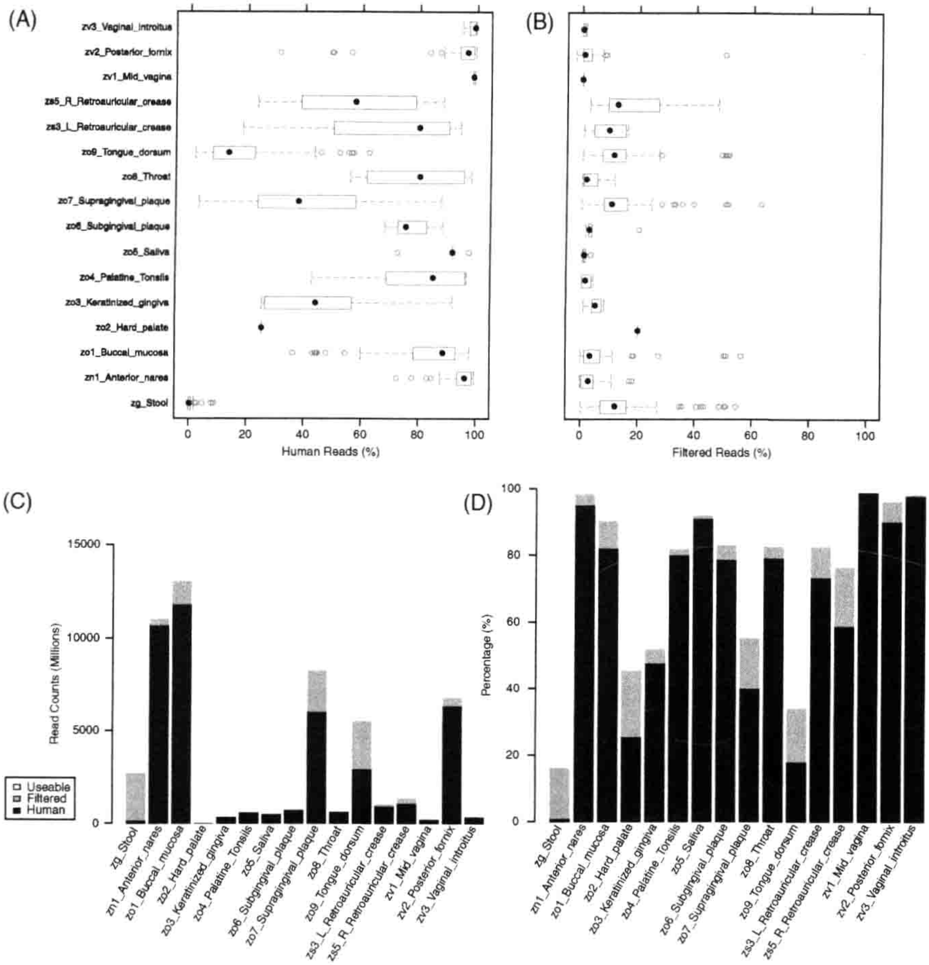

Fig. 1.2. Impact of quality and human filtering on shotgun metagenomic dataset. Thorough quality filtering and removal of reads resulting from human DNA contamination was performed on all shotgun metagenomic data of the Human Microbiome Project (average of 13 Gb/sample). The variation in fraction of reads per sample removed across the 18 body sites is shown by (A) boxplots for % of human and of (B) quality filtered reads. (C) Total amount of usable data (white) per site significantly varied because of (i) the different number of samples per site, (ii) the differential impact of human contamination (dark gray), and (iii) the differential impact of quality filtering (light gray). (D) Summary view of the usable fractions versus human and quality filtered data, per body site. (Reprinted by permission from Macmillan Publishers Ltd.[8])

While using metagenomic or metatranscriptomic approaches, it is essential to refer to the specific needs of each experiment.

The statistical analysis plan should take into account the characteristics of the experiment (in human studies, this would be the inclusion and exclusions criteria), the rate of sample acquisition (this would include the rate of human subject recruitment that will determine if you are