

Methods in
Molecular Biology 1399

Springer Protocols

A detailed microscopic image of wood grain, showing various layers and textures in shades of blue, white, and brown. The image is positioned in the upper middle section of the cover, partially overlapping a dark blue vertical bar on the left.

Francis Martin
Stéphane Uroz *Editors*

Microbial Environmental Genomics (MEG)

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:

<http://www.springer.com/series/7651>

Microbial Environmental Genomics (MEG)

Edited by

Francis Martin

*UMR1136 INRA/University of Lorraine "Tree-Microbe Interactions" (IAM),
Labex ARBRE, Champenoux, France*

Stéphane Uroz

*UMR1136 INRA/University of Lorraine "Tree-Microbe Interactions" (IAM),
Labex ARBRE, Champenoux, France*

UMR1138 INRA "Biogeochemistry of Forest Ecosystems" (BEF), Labex ARBRE, Champenoux, France

Editors

Francis Martin
UMR1136 INRA/University of Lorraine
“Tree-Microbe Interactions” (IAM)
Labex ARBRE
Champenoux, France

Stéphane Uroz
UMR1136 INRA/University of Lorraine
“Tree-Microbe Interactions” (IAM)
Labex ARBRE
Champenoux, France

UMR1138 INRA “Biogeochemistry
of Forest Ecosystems” (BEF)
Labex ARBRE
Champenoux, France

ISSN 1064-3745

ISSN 1940-6029 (electronic)

Methods in Molecular Biology

ISBN 978-1-4939-3367-9

ISBN 978-1-4939-3369-3 (eBook)

DOI 10.1007/978-1-4939-3369-3

Library of Congress Control Number: 2015957951

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2016

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.

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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Humana Press is a brand of Springer
Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

Although microorganisms (archaea, bacteria), micro-eukaryotes (fungi), and macro- and mesofauna represent major components of the environment, we are far from appreciating their identity, diversity, functions, the interactions established between them, and lastly their relative impact on the ecosystem functioning [1, 2]. In both terrestrial and aquatic ecosystems, they represent a considerable fraction of the living biomass [3] and several studies have now highlighted their key role in processes such as nitrogen and methane cycles, organic matter degradation, soil quality, and plant health and nutrition [4]. Most of the current knowledge was generated using monospecific or reductionist approaches, balancing between cultivation-dependent (sampling of organisms, morpho/phenotyping, physiological and biochemical characterization) and -independent approaches mostly based on low-throughput sequencing technologies (e.g., fingerprinting or cloning/sequencing). Such approaches were, and remain, very important as they enroot the current physiological and biochemical knowledge of the microorganisms, macro- and mesofauna, and give the relevance to the content of gene or protein sequences of the international databases. However, the recent revolution in sequencing technologies with the advent of the high-throughput methods (454 pyrosequencing, Illumina, Ion Torrent, PacBio, etc.), associated with a real decrease in the sequencing cost, is now opening the way to really appreciate the tremendous distribution and diversity of our micro- and macroorganisms neighbors [5]. Aside from the sequence-based approach, more and more analysis based on high-throughput chemical screening of environmental libraries (genomic DNA and cDNA cloned in expression vectors) are developed, revealing the common effort of the biologists to decipher the diversity and function of these organisms, especially the nonculturable and rare ones. At last, statistical analysis, modeling, and bioinformatics are rapidly becoming more accessible to single investigator laboratories [6]. All these aspects have really revolutionized microbial ecology giving emergence to a new research field entitled “Microbial Environmental Genomics.” Microbial environmental genomics seeks to understand how organisms and gene functions are influenced by environmental (biotic and abiotic) cues while accounting for variation that takes place within and among environmental populations and communities. By combining multiscale and multidisciplinary methods, we are now able to depict the complex assembly of organisms of the environment and to decipher their functional role (Fig. 1). Such developments should permit to improve our ability to develop predictive models to better integrate the relative role of these organisms in the biogeochemical cycles and the ecosystem functioning [7].

In this context, this book presents a series of 17 chapters to guide research into the identification of still unknown organisms, of novel functional genes, and how environmental conditions drive gene responses and the fitness of the complex guilds of organisms inhabiting our environment. Methods to analyze the diversity of different organism types are presented in Chapters 1–8, covering the archaea, bacteria, fungi, protists, and soil fauna. Chapter 9 presents a method to decipher the interactions between fungi and trees using RNA stable isotope probing (RNA-SIP). Notably, methods to identify and characterize functions and functional diversity of both pro- and eukaryotes are presented in Chapters 10–16. Those include protocols for gene hybridization (gene capture, geochips), DNA stable isotope probing, construction and screening of metagenomic and metatranscriptomic libraries,

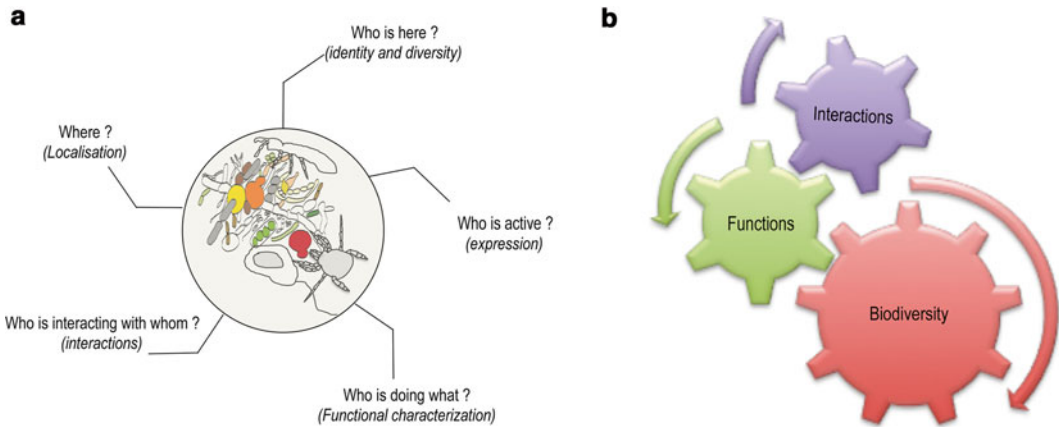


Fig. 1 Conceptual questioning in microbial environmental genomics. **(a)** The different challenging questioning in microbial environmental genomics. **(b)** The conceptual framework that needs to be integrated in models

and for bioinformatics analyses (MG-RAST). Chapter 17 presents a method to analyze both taxonomic and functional diversity using ancient DNA. We envision that this book will serve as a primary research reference for researchers and research managers in environmental microbiology working in the expanding field of molecular ecology and environmental genomics. The level of presentation is technically advanced with a strong emphasis on describing cutting-edge protocols in light of the possible future directions for research.

Champenoux, France

*Francis Martin
Stéphane Uroz*

Acknowledgments

S. Uroz and F. Martin are supported by the French National Agency of Research (ANR) through the Laboratory of Excellence Arbre (ANR-11-LABX-0002-01).

References

1. Averill C, Turner BL, Finzi AC (2014) Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage. *Nature* 505:543–545
2. Bardgett RD, van der Putten WH (2014) Belowground biodiversity and ecosystem functioning. *Nature* 515:505–511
3. Reid A, Greene SE (2012) How microbes can help feed the world. Report on an American Academy of Microbiology Colloquium, Washington, DC
4. Chaparro JM, Sheflin AM, Manter DK, Vivanco JM (2012) Manipulating the soil microbiome to increase soil health and plant fertility. *Biol Fertil Soils* 48:489–499
5. Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L (2015) High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *mBio* 6:e02288-14
6. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C (2013) Computational meta'omics for microbial community studies. *Mol Syst Biol* 9(1)
7. Treseder KK, Balsler TC, Bradford MA, Brodie EL, Dubinsky EA, Eviner VT et al (2012) Integrating microbial ecology into ecosystem models: challenges and priorities. *Biogeochemistry* 109:7–18

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>
1 “Deciphering Archaeal Communities” Omics Tools in the Study of Archaeal Communities <i>Lejla Pašić, Ana-Belen Martin-Cuadrado, and Purificación López-García</i>	1
2 Investigating the Endobacteria Which Thrive in Arbuscular Mycorrhizal Fungi <i>Alessandro Desirò, Alessandra Salvioli, and Paola Bonfante</i>	29
3 GenoSol Platform: A Logistic and Technical Platform for Conserving and Exploring Soil Microbial Diversity <i>Samuel Dequiedt, Pierre-Alain Maron, and Lionel Ranjard</i>	55
4 Sample Preparation for Fungal Community Analysis by High-Throughput Sequencing of Barcode Amplicons. <i>Karina Engelbrecht Clemmensen, Katarina Ihrmark, Mikael Brandström Durling, and Björn D. Lindahl</i>	61
5 Fungal Communities in Soils: Soil Organic Matter Degradation <i>Tomáš Větrovský, Martina Štursová, and Petr Baldrian</i>	89
6 DNA-Based Characterization and Identification of Arbuscular Mycorrhizal Fungi Species. <i>Carolina Senés-Guerrero and Arthur Schüßler</i>	101
7 Molecular Identification of Soil Eukaryotes and Focused Approaches Targeting Protist and Faunal Groups Using High-Throughput Metabarcoding <i>G. Arjen de Groot, Ivo Laros, and Stefan Geisen</i>	125
8 Identification and In Situ Distribution of a Fungal Gene Marker: The Mating Type Genes of the Black Truffle. <i>Herminia De la Varga and Claude Murat</i>	141
9 Stable-Isotope Probing RNA to Study Plant/Fungus Interactions. <i>Amandine Lé Van, Marie Duhamel, Achim Quaiser, and Philippe Vandenkoornbuysse</i>	151
10 Targeted Gene Capture by Hybridization to Illuminate Ecosystem Functioning. <i>Céline Ribière, Réjane Beugnot, Nicolas Parisot, Cyrielle Gasc, Clémence Defois, Jérémie Denonfoux, Delphine Boucher, Eric Peyretailade, and Pierre Peyret</i>	167
11 Hybridization of Environmental Microbial Community Nucleic Acids by GeoChip <i>Joy D. Van Nostrand, Huaqin Yin, Liyou Wu, Tong Yuan, and Jizhong Zhou</i>	183

12 Reconstruction of Transformation Processes Catalyzed
by the Soil Microbiome Using Metagenomic Approaches 197
Anne Schöler, Maria de Vries, Gisle Vestergaard, and Michael Schloter

13 MG-RAST, a Metagenomics Service for Analysis of Microbial
Community Structure and Function 207
Kevin P. Keegan, Elizabeth M. Glass, and Folker Meyer

14 Analysis of Active Methylo trophic Communities: When DNA-SIP
Meets High-Throughput Technologies 235
*Martin Taubert, Carolina Grob, Alexandra M. Howat,
Oliver J. Burns, Yin Chen, Josh D. Neufeld, and J. Colin Murrell*

15 Functional Metagenomics: Construction and High-Throughput
Screening of Fosmid Libraries for Discovery of Novel
Carbohydrate-Active Enzymes. 257
*Lisa Ufarté, Sophie Bozonnet, Elisabeth Laville, Davide A. Cecchini,
Sandra Pizzut-Serin, Samuel Jacquiod, Sandrine Demanèche,
Pascal Simonet, Laure Franqueville, and Gabrielle Potocki-Veronese*

16 Metatranscriptomics of Soil Eukaryotic Communities 273
*Rajiv K. Yadav, Claudia Bragalini, Laurence Fraissinet-Tachet,
Roland Marmeisse, and Patricia Luis*

17 Analysis of Ancient DNA in Microbial Ecology 289
*Olivier Gorgé, E. Andrew Bennett, Diyendo Massilani, Julien Daligault,
Melanie Pruvost, Eva-Maria Geigl, and Thierry Grange*

Index 317

Contributors

- E. ANDREW BENNETT • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- G. ARIEN DE GROOT • *ALTERRA-Wageningen UR, Wageningen, The Netherlands*
- PETR BALDRIAN • *Laboratory of Environmental Microbiology, Institute of Microbiology of the CAS, Praha 4, Czech Republic*
- RÉJANE BEUGNOT • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- PAOLA BONFANTE • *Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy*
- DELPHINE BOUCHER • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- SOPHIE BOZONNET • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- CLAUDIA BRAGALINI • *Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Université de Lyon, Villeurbanne Cedex, France; Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy*
- MIKAEL BRANDSTRÖM DURLING • *Department of Forest Mycology and Plant Pathology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden*
- OLIVER J. BURNS • *School of Biological Sciences, University of East Anglia, Norwich, UK*
- DAVIDE A. CECCHINI • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- YIN CHEN • *School of Life Sciences, University of Warwick, Coventry, UK*
- JULIEN DALIGAULT • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- CLÉMENTE DEFOIS • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- SANDRINE DEMANÈCHE • *Laboratoire Ampère, CNRS UMR5005, Ecole Centrale de Lyon, Université de Lyon, Ecully, France*
- JÉRÉMIE DENONFOUX • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France; Genoscreen, Campus de l'Institut Pasteur de Lille, Lille, France*
- SAMUEL DEQUIEDT • *INRA, UMR Agroécologie, Dijon Cedex, France*
- ALESSANDRO DESIRÒ • *Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy; Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA*
- MARIE DUHAMEL • *UMR 6553 Ecobio, CNRS, Université de Rennes 1, Rennes, France; Department of Biology, Stanford University, Stanford, CA, USA*
- KARINA ENGELBRECHT CLEMMENSEN • *Department of Forest Mycology and Plant Pathology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden*
- LAURENCE FRAISSINET-TACHET • *Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Université de Lyon, Villeurbanne Cedex, France*

- LAURE FRANQUEVILLE • *Laboratoire Ampère, CNRS UMR5005, Ecole Centrale de Lyon, Université de Lyon, Ecully, France*
- CYRIELLE GASC • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- EVA-MARIA GEIGL • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- STEFAN GEISEN • *Netherlands Institute for Ecology (NIOO), Wageningen, The Netherlands; Department of Terrestrial Ecology, University of Cologne, Cologne, Germany*
- ELIZABETH M. GLASS • *Argonne National Laboratory, Argonne, IL, USA*
- OLIVIER GORGÉ • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- THIERRY GRANGE • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- CAROLINA GROB • *School of Environmental Sciences, University of East Anglia, Norwich, UK*
- ALEXANDRA M. HOWAT • *School of Environmental Sciences, University of East Anglia, Norwich, UK*
- KATARINA IHRMARK • *Department of Forest Mycology and Plant Pathology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden*
- SAMUEL JACQUIOD • *Laboratoire Ampère, CNRS UMR5005, Ecole Centrale de Lyon, Université de Lyon, Ecully, France*
- KEVIN P. KEEGAN • *Argonne National Laboratory, Argonne, IL, USA; University of Chicago, Chicago, IL, USA*
- IVO LAROS • *ALTERRA-Wageningen UR, Wageningen, The Netherlands*
- ELISABETH LAVILLE • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- AMANDINE LÈ VAN • *UMR 6553 Ecobio, CNRS, Université de Rennes 1, Rennes, France*
- BJÖRN D. LINDAHL • *Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden*
- PURIFICACIÓN LÓPEZ-GARCÍA • *Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, Orsay, France*
- PATRICIA LUIS • *Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Université de Lyon, Villeurbanne Cedex, France*
- ROLAND MARMEISSE • *Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Université de Lyon, Villeurbanne Cedex, France*
- PIERRE-ALAIN MARON • *INRA, UMR Agroécologie, Dijon Cedex, France*
- FRANCIS MARTIN • *UMR1136 INRA/University of Lorraine "Tree-Microbe Interactions" (IAM), Labex ARBRE, Champenoux, France*
- ANA-BELEN MARTIN-CUADRADO • *Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Alicante, Spain*
- DIYENDO MASSILANI • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- FOLKER MEYER • *Argonne National Laboratory, Argonne, IL, USA; University of Chicago, Chicago, IL, USA*
- CLAUDE MURAT • *UMR1136 INRA/University of Lorraine "Tree-Microbe Interactions" (IAM), Labex ARBRE, Champenoux, France*
- J. COLIN MURRELL • *School of Environmental Sciences, University of East Anglia, Norwich, UK*

- JOSH D. NEUFELD • *Department of Biology, University of Waterloo, Waterloo, ON, Canada*
- NICOLAS PARISOT • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- LEJLA PAŠIĆ • *Department of Biology, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia*
- PIERRE PEYRET • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- ERIC PEYRETAILLADE • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- SANDRA PIZZUT-SERIN • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- GABRIELLE POTOCKI-VERONESE • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- MELANIE PRUVOST • *Institut Jacques Monod, UMR 7592, CNRS, Université Paris Diderot, Paris, France*
- ACHIM QUAISER • *UMR 6553 Ecobio, CNRS, Université de Rennes 1, Rennes, France*
- LIONEL RANJARD • *INRA, UMR Agroécologie, Dijon Cedex, France*
- CÉLINE RIBIÈRE • *EA 4678, CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France*
- ALESSANDRA SALVIOLI • *Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy*
- MICHAEL SCHLOTTER • *Research Unit for Environmental Genomics, Helmholtz Zentrum München, Neuherberg, Germany*
- ANNE SCHÖLER • *Research Unit for Environmental Genomics, Helmholtz Zentrum München, Neuherberg, Germany*
- ARTHUR SCHÜßLER • *Department of Biology, Genetics, SYMPLANTA GmbH & Co. KG, Munich, Germany*
- CAROLINA SENÉS-GUERRERO • *Department of Biology, Genetics, Ludwig-Maximilians University, Planegg-Martinsried, Germany*
- PASCAL SIMONET • *Laboratoire Ampère, CNRS UMR5005, Ecole Centrale de Lyon, Université de Lyon, Ecully, France*
- MARTINA ŠTURSOVÁ • *Laboratory of Environmental Microbiology, Institute of Microbiology of the CAS, Praha 4, Czech Republic*
- MARTIN TAUBERT • *School of Environmental Sciences, University of East Anglia, Norwich, UK; Institute for Ecology, Friedrich Schiller University Jena, Jena, Germany*
- LISA UFARTÉ • *INSA, UPS, INP; LISBP, Université de Toulouse, Toulouse, France; UMR792 Ingénierie des Systèmes Biologiques et des Procédés, INRA, Toulouse, France; UMR5504, CNRS, Toulouse, France*
- STÉPHANE UROZ • *UMR1136 INRA/University of Lorraine "Tree-Microbe Interactions" (IAM), Labex ARBRE, Champenoux, France; UMR1138 INRA "Biogeochemistry of Forest Ecosystems" (BEF), Labex ARBRE, Champenoux, France*
- PHILIPPE VANDENKOOORNHUYSE • *UMR 6553 Ecobio, CNRS, Université de Rennes 1, Rennes, France*
- JOY D. VAN NOSTRAND • *Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA*

- HERMINIA DE LA VARGA • *UMR1136 INRA/University of Lorraine “Tree-Microbe Interactions” (IAM); Labex ARBRE, Champenoux, France*
- GISLE VESTERGAARD • *Research Unit for Environmental Genomics, Helmholtz Zentrum München, Neuherberg, Germany*
- TOMÁŠ VĚTROVSKÝ • *Laboratory of Environmental Microbiology, Institute of Microbiology of the CAS, Praha 4, Czech Republic*
- MARIA DE VRIES • *Research Unit for Environmental Genomics, Helmholtz Zentrum München, Neuherberg, Germany*
- LIYOU WU • *Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA*
- RAJIV K. YADAV • *Ecologie Microbienne, UMR CNRS 5557, USC INRA 1364, Université Lyon 1, Université de Lyon, Villeurbanne Cedex, France*
- HUAQIN YIN • *Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA*
- TONG YUAN • *Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA*
- JIZHONG ZHOU • *Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA; Stephenson Research & Technology Center, University of Oklahoma, Norman, OK, USA*

Chapter 1

“Deciphering Archaeal Communities” Omics Tools in the Study of Archaeal Communities

Lejla Pašić, Ana-Belen Martin-Cuadrado,
and Purificación López-García

Abstract

Archaea constitute one of the three recognized phylogenetic groups of organisms living on the planet, and the latest to be discovered. Most Archaea resist cultivation and are studied using molecular methods. High-throughput amplicon sequencing and metagenomic approaches have been key in uncovering hitherto unknown archaeal diversity, their metabolic potential, and have even provided an insight into genomes of a number of uncultivated members of this group. Here, we summarize protocols describing sampling, molecular, metagenomic, and metatranscriptomic analyses as well as bioinformatics approaches that have proved useful for the study of archaea in natural samples.

Key words Archaeal communities, Metagenomic DNA, Small insert-size library, Large insert-size library, Single-cell genomics, Bioinformatic analysis

1 Introduction

Archaea constitute one of the three recognized phylogenetic groups of organisms living on the planet, and the latest to be discovered [1]. Exploration of microbial diversity in diverse ecosystems made manifest that archaea hold records of extremophily [2, 3] and seem particularly well adapted to limiting energy conditions [4]. However, they also thrive in non-extreme environments [5, 6].

Because archaea are difficult to get in culture, the use of molecular methods has been a key in uncovering a hitherto unknown archaeal diversity, including novel, highly divergent lineages (Fig. 1). For instance, the molecular exploration of archaeal diversity in the environment revealed the occurrence of several new groups of archaea (groups I–IV) in marine plankton [7–10]. Whereas groups II–IV defined new mesophilic lineages within the Euryarchaeota, group I archaea turned out to be an independent archaeal phylum, the Thaumarchaeote [11]. The discovery

of other potential new phyla, Aigarchaeota and Korarchaeota, forming a monophyletic group with Thaumarchaeota and with the Crenarchaeota has led to the proposal of the “TACK” superphylum [12]. It was thanks to metagenomic analyses showing the presence of *amo* genes, encoding different subunits of ammonium monooxygenase, that the ability to oxidize ammonia aerobically was first proposed as potential energy metabolism for the so-far uncultured members of Thaumarchaeota [13, 14]. This led to the isolation of the first axenic culture for a thaumarchaeote, the aerobic ammonia-oxidizing chemolithoautotroph *Nitrosopumilus maritimus*, and subsequently led to the discovery that Thaumarchaeota play a major ecological role as nitrifiers in the global nitrogen cycle [15]. Other metagenomic studies of archaea have also provided clues on the metabolic potential of uncultured archaea, for instance, Group II Euryarchaeota [16–19]. Moreover, the reconstruction of archaeal genomes or pangenomes from metagenomes has also started to provide interesting evolutionary information, for instance on the genome plasticity of extreme halophilic archaea in saturated brines [20] or on the large impact of interdomain gene transfer affecting Thaumarchaeota and marine Groups II and III Euryarchaeota [18]. Single-cell genomics appears as a complementary approach, not exempt of difficulties (contamination, biased genome amplification) to metagenomics, and a more direct approach to get functional and evolutionary information from specific organisms in complex communities [21]. In addition to the above approaches, metatranscriptomics and other RNA-based analysis (e.g., [22]) as well as metaproteomics and metabolomics are promising tools to better understand archaeal function in nature.

With increasingly improving and cheaper high-throughput sequencing, the bottleneck no longer lies on sequencing techniques, but on access and appropriate preparation of environmental samples and on the subsequent bioinformatic analyses. Archaea present specific challenges on two grounds. From an experimental point of view, environmental studies of archaea are challenging because many of them are extremophiles thriving in difficult-to-sample habitats, such as deep-sea vents or the deep subsurface. Special equipment and conditions are therefore required. Also, the physicochemical nature of their environment imposes constraints on sample preparation. For instance, low or high pH or high salt may alter the efficacy of cell lysis and nucleic acid purification of most commercial kits. In addition, because archaea have specific molecular adaptations, especially resistant lipids and S-layers, cell lysis may be challenging and part of the archaeal population may escape lysis protocols that work well for their bacterial or eukaryotic counterparts. Therefore, samples must be treated accordingly to overcome such difficulties. For the bioinformatic analyses,

archaea may represent a challenge because of the lack of reference genomes for many archaeal lineages that remain uncultured. Some archaea have also extremely low GC content [23], making more difficult some *in silico* analyses. In the present chapter, we briefly describe protocols describing sampling, molecular, metagenomic, metatranscriptomic, and single-cell genomics strategies that specifically deal with these problems and have proved useful for the study of archaea in natural samples (Fig. 2).

2 Materials

1. Discrete sampler (CTD incorporated in an array of Niskin bottles) or continuous sampling equipment. The latter may be a hose reinforced with fibers or steel cord (15 mm diameter hose LT 362 25 031318, NOVA Agricola (<http://www.novaagricoladebraga.com/>)) connected to a water pump (Sterwins Jet-2 900 W, H_{\max} 40 m, Q_{\max} 3600 L/h, ADEO services (<http://www.adeo.com/>)).
2. Sterile water/soil/sediment containers. Use amber containers when sampling aphotic environments, and heat-resistant thermo-containers when sampling hot environments.
3. Grab sampler or core box. Alternatively, sediment can be sampled using a sterile forceps or a syringe.
4. Dry ice/liquid nitrogen containers and transport containers.
 1. 0.22 μm Sterivex™ filter units and/or mixed cellulose ester hydrophilic filters with radius 90 mm/142 mm/293 mm and pore sizes 5.0 μm and 0.22 μm (MF-Millipore®, EMD Millipore, <http://www.emdmillipore.com/>).
 2. Peristaltic pump Millipore Easy-load Master-Flex (MF-Millipore®, EMD Millipore (<http://www.emdmillipore.com/>)) (*see Note 1*).
 3. Standing Stainless Steel Filter Holders (90 mm, 142 mm, and 293 mm) (MF-Millipore®, EMD Millipore (<http://www.emdmillipore.com/>)).
 4. Sucrose lysis buffer (SLB): 40 mM EDTA, 50 mM Tris-base, 0.75 M sucrose, pH 8.3. Filter or sterilize by autoclaving for 20 min at 121 °C.
 5. Low temperature freezer (−80 °C) (Heto, ThermoFisher Scientific (<http://www.thermoscientific.com/>)).
 6. RNAlater (Ambion Inc. (www.lifetechnologies.com/)). Cryovials V5132 Sigma (Nalgene, (<http://www.thermoscientific.com/>)).
 7. Acid-washed beads 710–1180 μm (Sigma-Aldrich (<http://www.sigmaaldrich.com/>)).

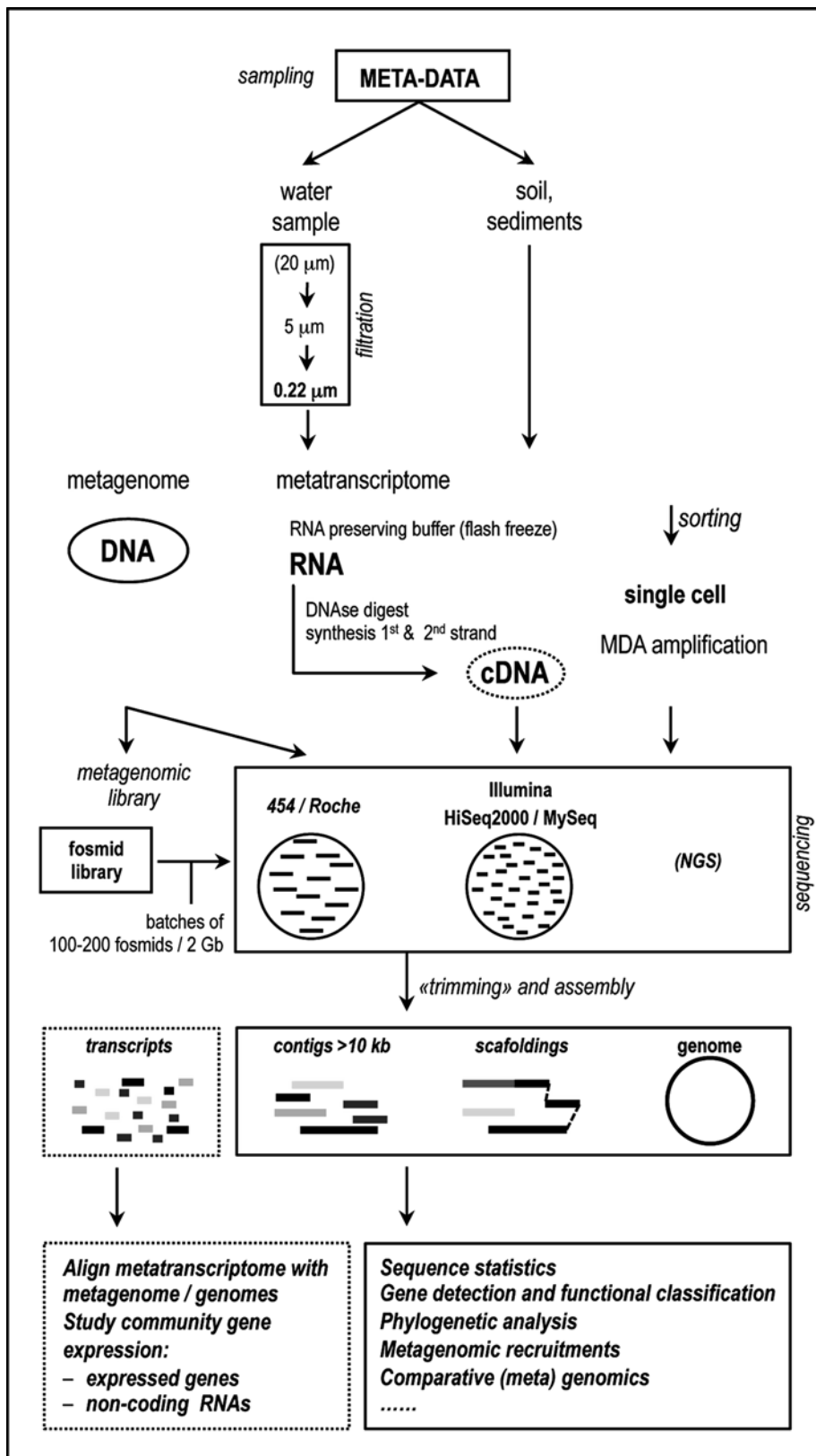


Fig. 2 Workflow of methods currently used in omics studies of archaeal communities

8. TE buffer: 10 mM Tris, 1 mM EDTA, pH 8.0 (all Merck & Co., Inc. (<http://www.merck.com>)). Autoclave for 20 min at 121 °C and store at room temperature.
9. 50 % (v/v) glycerol stock solution (50 %): molecular-grade 87 % glycerol aqueous solution (Sigma-Aldrich (<http://www.sigmaaldrich.com>)) in TE buffer (Sigma (<http://www.sigmaaldrich.com>)). Pass it through a 0.1 or 0.22 µm syringe filter. Store at -20 °C for up to 1 year.
10. 50 % betaine stock solution: anhydrous betaine (Sigma-Aldrich (<http://www.sigmaaldrich.com>)) in distilled water. Pass it through a 0.1 or 0.22 µm syringe filter. This stock can be stored at 4 °C for up to 1 year. Re-filter it before every use.
1. Acid-washed beads 150–212 µm (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
2. 15 mL Conical Centrifuge Tubes (Fisher Scientific (<http://www.fishersci.com/>)).
3. Heating cabinet (www.genlab.co.uk/drying-warming-cabinets).
4. Digester (Waldner Laboratory Systems).
5. Amicon-15 Centrifugal Filter Units with NMWL 30 kDa (EMD Millipore (<http://www.emdmillipore.com/>)).
6. 10 mg/mL lysozyme (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
7. 20.2 mg/mL proteinase K (Fermentas (www.thermoscientificbio.com/fermentas/)).
8. 10 % sodium dodecyl sulfate (SDS) (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
9. Phenol:chloroform:isoamyl alcohol (25:24:1; pH 8.0) (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
10. Chloroform:isoamyl alcohol (24:1) (Sigma-Aldrich (<http://www.sigmaaldrich.com>)). Absolute ethanol (Merck & Co., Inc. (<http://www.merck.com>)).
11. Mo Bio PowerSoil™ DNA extraction kit (Mo Bio (www.mobio.com/)).
12. Resuspension buffer: TE buffer (pH 8.0), 4 mg/mL lysozyme, 0.2 mg/mL proteinase K. (The reagents were purchased from Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
13. RNeasy Mini Kit (Qiagen (<http://www.qiagen.com>)).
14. 2000 U/mL DNase I with 10× buffer (New England Biolabs Inc (<http://www.neb.com>)).
15. 50 µM Random Primers (Invitrogen (<http://www.lifetechnologies.com>)).
16. SuperScript® III Reverse Transcriptase Kit (Invitrogen (<http://www.lifetechnologies.com>)).

17. RNaseOUT[®] Recombinant Ribonuclease Inhibitor (Invitrogen (<http://www.lifetechnologies.com>)).
18. 10 mM dNTP Mix (Invitrogen (<http://www.lifetechnologies.com>)).
19. 10,000 U/mL DNA polymerase I with 10× NEBuffer 2 (New England Biolabs Inc (<http://www.neb.com>)).
20. 10,000 U/mL *E. coli* DNA ligase (New England Biolabs Inc (<http://www.neb.com>)).
21. 2 U/μL ribonuclease H (Invitrogen (<http://www.lifetechnologies.com>)).
22. PowerSoil[™] Total RNA Isolation Kit (Mo Bio (www.mobio.com/)).
23. 5 U/μL *Taq* polymerase (Invitrogen (<http://www.lifetechnologies.com>)).
24. RNase ZAP RNase Decontamination Solution (Life Technologies (<http://www.lifetechnologies.com>)).
25. End-It[™] DNA End-Repair Kit (Epicentre (www.epibio.com)).
26. QIAquick GEL extraction kit (Qiagen (www.qiagen.com/)).
27. QIAquick PCR purification kit (Qiagen (www.qiagen.com/)).
28. pGEM[®]-T Easy Vector System (Promega (www.promega.com/)).
29. Electrocompetent *E. coli* JM109 cells (Promega (www.promega.com/)).
30. CHEF-DR[®] III Pulsed Field Electrophoresis Systems (BioRad (<http://www.bio-rad.com/>)).
31. CopyControl[™] Fosmid Library Production Kit with pCC1FOS vector (Epicentre (www.epibio.com)).
32. Nunc[®]96 DeepWell[™] plates (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
33. 2 mL Eppendorf Safe-Lock Tubes[™] (Eppendorf (<http://www.eppendorf.com/>)).
34. 0.22 μm filter (Millex[®]-GS filters, and EMD Millipore (<http://www.emdmillipore.com/>)).
35. 40 μm BD falcon nylon cell strainer (BD Biosciences (<http://www.bdbiosciences.com>)).
36. Sheath fluid: dissolve combusted (2 h at 45 °C) NaCl in DNA-free deionized water to a final concentration of 1 %.
37. 10,000× SYBR Green fluorescent nucleic acid dye (Invitrogen (<http://www.lifetechnologies.com/>)).
38. Lysis buffer D2 and stop solution. It contains potassium hydroxide and it is corrosive and harmful. Avoid skin contact,

- eye contact, and ingestion (REPLI-g Single Cell Kit (Qiagen, www.qiagen.com/)).
39. RepliPHI Phi29 reagents kit (Epicentre (www.epibio.com)).
 40. 50 μ M random hexamers (Invitrogen (<http://www.lifetechnologies.com/>)). Order with “standard desalting” and “hand-mix randomization” parameters. Hexamers should have phosphorothioate bonds between the last two nucleotides at the 3' end (5'-NNNN*N*N-3'). Store aliquots at -20 °C for 1 year.
 41. Dimethylsulfoxide (DMSO) (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
 42. 50 mL Conical Centrifuge Tubes (Fisher Scientific (<http://www.fishersci.com/>)).
 43. Cell sorter (e.g., Influx (BD Biosciences) or MoFlo (Beckman Coulter)).
 44. Corning® 96 well plates, UV-transparent (Sigma-Aldrich (www.sigmaaldrich.com)).
 45. SYTO 13 nucleic acid stain (Invitrogen <http://www.lifetechnologies.com/>).
 46. Nextera® XT DNA Sample Preparation Kit (<http://www.illumina.com/>).
 47. GS FLX Titanium Rapid Library Preparation Kit (<http://454.com/products/gs-FLX-system/>).
 48. 3 M sodium acetate (pH 5.5): 24.61 g of sodium acetate in 80 mL of double distilled water. pH should be adjusted to 5.5 with glacial acetic acid and the solution should be made to 100 mL with double distilled water (The reagents were purchased from Sigma-Aldrich (<https://www.sigmaaldrich.com/>)). Sterilize by filtration and store at room temperature.
 49. Absolute ethanol (Merck & Co., Inc. (<http://www.merck.com>)).
 50. Isopropanol (Merck & Co., Inc. (<http://www.merck.com>)).
 51. 98 % dithiothreitol (DTT) (Sigma-Aldrich (<https://www.sigmaaldrich.com/>)).
 52. 100 mM dATP (Thermoscientific (<http://www.thermoscientific.com/>)).
 53. Phosphate buffered saline (PBS): 8 g NaCl, 0.2 g KCl, 1.44 g Na_2HPO_4 , 0.24 g KH_2PO_4 per 1 L, pH 7.4 (The reagents were purchased from Merck & Co., Inc. (<http://www.merck.com>)). Autoclave for 20 min at 121 °C or filter-sterilize and store at room temperature.
 54. LB medium: 10 g tryptone, 5 g yeast extract, 10 g NaCl, dissolve in 1 L of distilled water. Autoclave for 20 min at 121 °C and store at room temperature.

55. Agarose Low Melting (Roth, <https://www.carlroth.com/>).
56. 10× TBE buffer (Thermoscientific (<http://www.thermoscientific.com/>)).
57. 34 mg/mL chloramphenicol in 100 % ethanol (Sigma-Aldrich (<http://www.sigmaaldrich.com>)): Store at -20°C .
58. 100 mg/mL ampicillin in water (Sigma-Aldrich (<http://www.sigmaaldrich.com>)). Store at -20°C .
59. Ultrapure nuclease-free water (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).
60. Sodium hypochlorite solution (Sigma-Aldrich (<http://www.sigmaaldrich.com>)).

3 Methods

3.1 Sampling Methods

Sampling is perhaps the most important step in an environmental study and should be carefully planned (*see Note 2*).

3.1.1 Water Samples

1. Collect between 50 and 300 L of water using the CTD incorporated in an array of Niskin bottles or using a pump and appropriate pre-cleaned containers (*see Note 3*).

3.1.2 Soil and Sediment Samples

1. Sample the soil (~500 g) under sterile conditions. Collect triplicate samples from soil surface and sieve them through a 2 mm mesh to remove large particles and plant material. Note that at least 0.25 g will be required for DNA extraction. Place the samples into individual sterile pre-cleaned containers (*see Note 3*) and keep them on ice until arrival in the laboratory.
2. Use grab sampler or other sediment sampling tool to collect sediment samples in triplicate.
3. Upon collection, subsample the sediment samples, homogenize them, and transfer them to the appropriate pre-cleaned sample containers. Transport the samples on dry ice. Supplement the hot spring sediment samples with equal volume of sucrose lysis buffer (SLB) immediately upon sampling. For SLB preparation *see* Subheading 2.

3.2 Sample Processing and Preservation

3.2.1 Processing and Preservation of Water Samples for Metagenomic Studies

1. Sequentially filter the water samples through 5 μm and 0.22- μm pore size filters using filter holders and peristaltic pumping system until clogging. At least triplicate filters should be produced (*see Note 4*).
2. Conserve Sterivex™ filters in Lysis buffer (*see* Subheading 2) at -20°C until DNA extraction. If water is filtered through mixed cellulose ester filters, store the filters in 50 mL conical centrifugation tubes at -80°C until DNA extraction.

3.2.2 Processing and Preservation of Soil and Sediment Samples for Metagenomic Studies

1. Process the soil samples within the 24 h from sampling—sampling itself disturbs the soil and can alter the composition of microbial community.
2. Upon arrival to the laboratory store the sediment samples at $-80\text{ }^{\circ}\text{C}$ until DNA extraction.

3.2.3 Processing and Preservation of Water Samples for Metatranscriptomic Studies

1. Allow the filtration (*see* Subheading 3.2.1, **step 1**) to proceed for 10 min then fill the Sterivex™ filter with 2 mL of RNAlater and freeze in liquid nitrogen (*see* **Note 5**).
2. Produce at least triplicate filters and store at $-80\text{ }^{\circ}\text{C}$ until RNA extraction.

3.2.4 Processing and Preservation of Soil and Sediment Samples for Metatranscriptomic Studies

1. Ground the soil samples in liquid nitrogen using a mortar and pestle until a fine powder is obtained. Suspend this powder in equal volume of RNAlater. Keep at $-80\text{ }^{\circ}\text{C}$ until RNA extraction (preferably within 24 h).
2. Supplement the sediment samples with equal volume of RNAlater and keep at $-80\text{ }^{\circ}\text{C}$ until RNA extraction.

3.2.5 Processing and Preservation of Samples for Single-Cell Genomics

1. Transfer aliquots of water samples (1.7 mL) to cryovials in triplicate.
2. Resuspend soil and sediment samples (~5 g) in 10–30 mL of $1\times$ PBS (*see* Subheading 2).
3. To disrupt cell aggregates, add 0.1 g acid-washed beads (diameter 710–1180 μm) and vortex the sample for 30 s at the highest setting. Alternatively, expose the sample to sonication in an ultrasonic water bath for 10 min at room temperature.
4. To confirm the absence of aggregates, examine the sample under microscope.
5. Remove acid-washed beads by passing the sample through a 40 μm filter.
6. To remove the remaining large particles centrifuge the sample for 30 s at $2500\times g$.
7. Transfer the supernatant (~1.7 mL) to a new cryovial.
8. Add 240 μL (final concentration 6 %) of 50 % glycerol or 50 % betaine stock solution (*see* Subheading 2) and store at $-80\text{ }^{\circ}\text{C}$.

3.3 Nucleic Acid Purification

Environmental nucleic acids of sufficient yield, purity, and integrity are the crucial starting material in metagenomic studies. The below protocols aim to extract high yields of DNA and RNA while minimizing shearing DNA using mechanical lysis which is presumed to introduce minimal bias.

Working with RNA requires an RNase-free working environment. To achieve this, dedicate a separate laboratory area, pipettors, and materials. Use only RNase-free reagents and plastic tubes.

Wear gloves at all times and treat the gloves, the utensils, and working surfaces with RNase ZAP. Pipet at a 45° angle with open tubes facing away from you and use PCR hood. When working with low biomass samples, scale up the volume of sample used for isolation (e.g., from ~0.5 g to 25 g). To ensure the absence of aerosolized contaminants include extraction blanks and confirm the absence of DNA and RNA contaminants by no visible amplification of 16S rRNA from extraction blanks after 35 cycles of PCR (*see Note 6*).

3.3.1 DNA Purification from Water Samples

The protocol assumes that the sample was filtered through Sterivex™ filters, and the quantities should be adjusted if cut mixed cellulose ester filters are used for the filtration of the sample.

1. To the filters add 1.8 mL of SLB lysis buffer (*see Subheading 2*) and 50 µL of lysozyme (final concentration 1 mg/mL). Add 200 µL of acid-washed beads (150–212 µm) and incubate at 37 °C for 45 min in slight movement.
2. Add 50 µL of proteinase K (final concentration 0.2 mg/mL) and 210 µL of 10 % SDS (final concentration 1 % (w/v)). Incubate at 55 °C for 1 h in slight movement. Allow the glass beads to settle on the bottom of the tube.
3. Recoverate lysate from the filter and transfer it to a 15 mL conical centrifugal tube. To remove remaining material, add 1 mL of Lysis buffer to the Sterivex™ filter, incubate at 55 °C for 15 min, and transfer the lysate to a fresh 15 mL conical centrifugal tube.
4. Extract twice with equal volume (3 mL) of phenol:chloroform:isoamyl alcohol (25:24:1; pH 8.0). Add 3 mL of phenol:chloroform:isoamyl alcohol, vortex 1 min, and centrifuge 5 min at 5000 ×g. Transfer the aqueous phase to a new tube and repeat the extraction and the centrifugation. Extract once with equal volume of chloroform:isoamyl alcohol (24:1). Transfer the aqueous phase to a new tube. Make sure no organic phase is transferred.
5. Concentrate DNA on a 30 kDa Amicon Ultra filter, by spinning it down to a volume of 200 µL.
6. Add 1 mL of TE buffer and spin down to a volume of 200 µL. Repeat this three times before collecting the DNA. The DNA can be stored at –20 °C.
7. Alternatively, precipitate the DNA by adding to the aqueous phase 0.1 V (300 µL) of 3 M sodium acetate pH 5.5 (*see Subheading 2*) and 2 V (6 mL) of 100 % ethanol. Allow the precipitation to proceed for at least 2 h at –20 °C. To collect the DNA, centrifuge at 20,800 ×g for 20 min at 4 °C.
8. Wash the pellet twice with 1 mL of 80 % ethanol, air-dry, and resuspend in 100 µL of TE buffer.

9. Determine the concentration of isolated DNA (*see* **Notes 7 and 8**).

3.3.2 DNA Purification from Soil and Sediment Samples

1. Wash the sediment sample (0.25 g to 1 g) at room temperature for 1 h with gentle shaking in 2 mL of 3 % NaCl. This will remove extracellular DNA. If working with soil samples use 0.25 g to 1 g of sample and start with **step 5**.
2. Centrifuge the sample for 10 min at $3000 \times g$ and room temperature and remove the supernatant.
3. To 1 g of washed sample add phosphate buffered saline (PBS) (*see* Subheading 2) to a final volume of 0.5 mL. This will minimize the effect of sample pH on DNA yields obtained.
4. Subject the sample to six to six freeze/thaw cycles in liquid nitrogen to facilitate cell lysis of archaeal cells and increase DNA yield.
5. Perform DNA purification using Mo Bio PowerSoil™ DNA extraction Kit [24].
6. Determine the concentration of isolated DNA (*see* **Notes 7 and 8**).

3.3.3 RNA Purification from Water Samples

1. To isolate RNA from the RNA preservation buffer in which the filter was stored:

Transfer the buffer into a 15 mL conical centrifuge tube.

Add 1/10 V of 3 M potassium acetate (pH 5.5) (*see* Subheading 2) and 1 V of isopropanol. Vigorously vortex the tube for 2 min. Incubate the tube at room temperature for 2 h with slight movement.

Centrifuge the sample for 30 min at 4 °C and $12,000 \times g$ and remove the supernatant.

To the pellet, add 1 mL of ice-cold 70 % ethanol.

Centrifuge the tube for 10 min at 4 °C and $12,000 \times g$ and remove the supernatant.

Repeat **steps 4 and 5**.

Air-dry the sample and resuspend the dried pellet in 600 µL of Resuspension buffer (*see* Subheading 2). Transfer 200 µL aliquots into three separate tubes.

Isolate RNA using RNeasy Mini Kit according to the manufacturer's instructions.

2. To isolate RNA from the filter:

Add 600 µL of Resuspension buffer to the filter (Sterivex™ or mixed cellulose filter that was cut in small pieces).

Incubate 10 min at room temperature. Vortex the sample 10 s every 2 min. Transfer 200 µL aliquots into separate tubes.

Isolate RNA using RNeasy Mini Kit [25].

3. Combine the RNAs obtained in **steps 1** and **2** and quantify the amount of RNA obtained (*see Notes 7–10*).
4. To the isolated RNA add 1/10 V of 10× DNase I reaction buffer. Add 1 U of DNase I per each µg of RNA. Mix by pipetting and incubate 10 min at room temperature.
5. Add 1/10 V of DNase I Stop solution. Mix by pipetting and incubate 10 min at 65 °C.
6. Concentrate RNA as described in Subheading 3.3.1, **step 7**. Resuspend the pellet in 10 µL of RNase-free water and quantify the amount of RNA obtained.
7. Confirm the absence of DNA and RNA contaminants by no visible amplification of 16S rRNA from extraction blanks after 35 cycles of PCR (*see Note 6*).
8. To synthesize the first strand of cDNA combine 10 µL of 0.5 µg/mL RNA with 1 µL of 3 µg/µL Random Primers, incubate 10 min at 70 °C, and chill on ice.
9. To the tube on ice add 4 µL 5× first-strand buffer, 2 µL 100 mM DTT, 1 µL of SuperScript® III Reverse Transcriptase (all included in SuperScript® III Reverse Transcriptase Kit), 1 µL of RNaseOUT™ Recombinant Ribonuclease Inhibitor, and 1 µL 10 mM dNTP Mix. Incubate the reaction mixture 10 min at 25 °C, 6 min at 55 °C, and 15 min at 70 °C. Chill the tube on ice.
10. To synthesize the second strand of cDNA add the following to the tube from previous step: 15 µL 10× NEBuffer 2 (supplied with DNA polymerase I), 4 µL DNA polymerase I, 1 µL *E. coli* DNA ligase, 2.5 µL Ribonuclease H, 3 µL 10 mM dNTP Mix, and 104.5 µL RNase-free water.
11. Incubate the reaction mixture 2 h at 16 °C. Stop the reaction by adding 10 µL 0.5 M EDTA (pH 8.0).
12. Purify cDNA by QIAquick PCR Purification Kit [26]. Quantify the amount of obtained cDNA (*see Note 9*).

3.3.4 RNA Purification from Soil and Sediment Samples

1. Extract the RNA from up to 2 g of soil (or up to 5 g of sediment) using PowerSoil™ Total RNA Isolation Kit [27] (*see Notes 8–11*). Include the extraction blanks.
2. Remove the DNA from RNA samples using DNase I treatment (*see Subheading 3.3.3, steps 12 and 13*).
3. Purify RNA by using RNeasy Mini Kit [25].
4. Confirm the absence of DNA and RNA contaminants by no visible amplification of 16S rRNA from extraction blanks after 35 cycles of PCR (*see Note 6*).
5. Proceed with **steps 15–19** of Subheading 3.3.3.

3.4 Cloning

In order to successfully clone the environmental DNA, a number of precautions should be undertaken. Avoid the prolonged storage of isolated environmental DNA at either 4 °C (days) or at -20 °C (months) as well as repeated freeze-thaw cycles. Preferably, proceed with cloning immediately upon the isolation of environmental DNA.

3.4.1 Small (≤ 10 kb) Insert-Size Shotgun Libraries

1. Determine the size and the quantity of isolated environmental DNA (*see Note 12*) by running an aliquot (1–2 μ L) of it on a 1 % agarose gel in 1 \times TBE buffer on a pulse field gel electrophoresis under following conditions: temperature 14 °C, voltage 6 V/cm, initial switch time 0.1 s, final switch time 2 s, angle 120°, length of a run 11 h.
2. If more than 50 % of isolated environmental DNA fragments is of desired insert size, proceed with Subheading 3.4.1, **step 5**.
3. Shear the DNA by passing it through 200 μ L pipette tips. Place 2–10 μ g of DNA diluted to 100 μ L with TE buffer into a clean microcentrifuge tube. Aspire and expel the DNA up to 200 times (*see Note 13*).
4. Examine 1–2 μ L of sheared insert DNA on an agarose gel as described in **step 1**.
5. To generate end-repaired insert DNA add the following reagents on ice to a final volume of 80 μ L (included in End-It™ DNA End-Repair Kit as well as in CopyControl™ Fosmid Library Production Kit with pCC1Fos Vector): 8 μ L 10 \times End-Repair Buffer, 8 μ L 2.5 mM dNTP Mix, 8 μ L 10 mM ATP, sheared insert DNA (up to 20 μ g), 4 μ L End-Repair Enzyme Mix, sterile water.
6. Incubate at room temperature for 1 h.
7. Inactivate the enzyme mix at 70 °C for 10 min.
8. Fractionate the blunt-ended DNA in the absence of any DNA stain using pulse field gel electrophoresis on a 1 % Agarose Low melt gel prepared with 1 \times TBE buffer. Prepare the gel with wide combs and use DNA size markers at both outside lanes of the gel.
9. Upon electrophoresis, cut off the outer lanes of the gel that contain the DNA size marker and stain them.
10. Visualize the DNA size markers with UV light and mark the position of the desired fragment size on both ladders with sterile scalpel or pipette tip.
11. Assemble the gel and cut out the gel slice containing insert DNA of desired size. Transfer the gel slice to a tared tube.
12. Extract the DNA from the slice by using QIAquick Gel Extraction Kit [26].

13. Generate 3' A-overhangs to the insert DNA from **step 12** by adding the following reagents: 1 μ L dATP (final concentration 0.2 mM), 5 μ L *Taq* polymerase buffer with MgCl₂ (final MgCl₂ concentration 1.5 mM), 0.2 μ L (1 U) *Taq* DNA polymerase. Add sterile water to 50 μ L.
14. Incubate at 72 °C for 30 min.
15. Purify the resulting insert DNA by QIAquick PCR Purification Kit [26] (*see Note 14*).
16. Clone the insert DNA into pGEM®-T Easy Vector System [28] and introduce the recombinant vectors into electrocompetent *E. coli* JM109 cells.
17. Randomly pick several insert-positive clones, grow each overnight in 10 mL LB medium with 100 mg/mL ampicillin. Use standard techniques to isolate, digest, and analyze plasmid DNA.
18. Store the insert-positive clones in 96-well plates as cultures in LB medium with 100 mg/mL ampicillin and 15 % glycerol. Keep at -80 °C for prolonged storage.

3.4.2 Large (10–40 kb) Insert-Size Shotgun Libraries

1. Prepare the large insert-size shotgun library by using CopyControl™ Fosmid Library Production Kit [29] (*see Note 15*).
2. To evaluate the quality of the library, select a subsample of insert-positive clones and grow them in LB medium with 12.5 mg/mL chloramphenicol overnight at 37 °C and 250 rpm.
3. To induce high copy number of fosmids in the cells, inoculate 500 μ L of overnight cultures from **step 2** into individual flasks that contain 5 mL of LB medium with 12.5 mg/mL chloramphenicol and 5 μ L of the 1000 \times CopyControl™ Induction Solution (included in CopyControl™ Fosmid Library Production Kit). Incubate for 5 h at 37 °C and 250 rpm.
4. Use standard techniques to extract, digest, and visualize fosmid DNA.
5. Store positive clones by picking them with sterile toothpicks and transferring them to separate wells of 96- or 384-well plates as a culture in LB medium with 12.5 mg/mL chloramphenicol supplemented with 15 % glycerol at -80 °C.

3.5 Single-Cell Genome Amplification

Single-cell genomics consists of a series of integrated processes which imply the physical separation of the cells from a population, their lysis, and the whole genome amplification of the individual cells (*see Note 16*). Although new, this approach already yielded genomic sequence of several noncultured archaeons [e.g., 30].

1. To decontaminate cell sorting instrument, install fresh sheath fluid (*see* Subheading 2) and clean the sample lines by a succession of warm water, 5 % sodium hypochlorite solution, and an overnight flush with DNA-free deionized water. Adjust the instrument following the manufacturer's instructions. Perform cell sorting and genome amplification in a HEPA-filtered environment.
2. To stain the cells, add 10,000× SYBR Green fluorescent nucleic acid dye to a final concentration of 1×. Incubate 15 min at 4 °C in the dark.
3. Sort the targeted microbial population in UV-treated microtiter plates containing 1 μL of TE buffer (pH 8.0) per well. Centrifuge the plate for 1 min at room temperature and 1000×*g* to spin down the reagents when necessary. Sorted cells can be stored at -80 °C for several months.
4. Add 1 μL of Lysis buffer D2 (included in REPLI-g Single Cell Kit) to each individual well and incubate 5 min at room temperature (*see* Note 17).
5. Add 1 μL of stop solution (included in REPLI-g Single Cell Kit). Centrifuge the plate for 1 min at room temperature and 1000×*g*. Do not leave the centrifuged plate for more than 1 h at 4 °C.
6. Prepare the multiple displacement amplification master mix: mix 1.5 μL 10× Phi29 DNA polymerase buffer, 0.24 μL 25 mM dNTP solution, 1.5 μL 0.5 mM Random hexamers, 0.15 μL 1 M DTT, 0.75 μL DMSO, 7.5 μL water, and 0.4 μL of Phi29 DNA polymerase. UV-irradiate the master mix for 90 min in a reflective container on ice.
7. To the master mix add 0.0015 μL of SYTO13 per reaction (final concentration 0.5 μM). Add 12 μL of master mix to 3 μL of lysed cell sample.
8. Allow the amplification to proceed for 30 min at 30 °C. Inactivate the polymerase by 15 min incubation at 65 °C.
9. Keep the amplified genomic DNA (up to 40 μg) at -80 °C until sequencing. Then proceed to either targeted *loci* or whole genome sequencing. To identify archaea genomes of interest, amplify and sequence the 16S rRNA (*see* Note 6). Other genes can be amplified, such as the *amo* gen.

3.6 Sequencing

3.6.1 Sanger Sequencing

Prior to sending the samples to sequencing facilities, carefully consider the guidelines for sample preparation. Shotgun library clones are usually sent on 96-well plates as bacterial stab cultures or as cultures in glycerol. The choice of 96-well plate and plate seals plays an important role as it must be compatible with

the sequencing platform and able to withstand pressure and temperature changes during transportation. If samples are transported to the sequencing facility by airmail, we recommend sending them early in the week to avoid prolonged storage during labor-free days.

3.6.2 High-Throughput Technologies

Next-generation sequencing (NGS) is commonly used in metagenomic studies of complex microbial communities. Roche/454 and Illumina platforms have been shown to provide very comparable results for abundances of genes or genomes [31].

The DNA that is to be sequenced should be free of all impurities and should not contain any biological macromolecules, chelating agents, divalent metal cations, denaturants, or detergents as these will interfere with the construction of the library. $A_{260}/_{280}$ ratio of DNA should be 1.8–2.0 and an $A_{260\text{ nm}}/_{230\text{ nm}} \geq 1.9$ and determined according to **Note 7**. The DNA should be dissolved in molecular-grade water (RNase and DNase free) or in TE buffer. While as little as 50–100 ng can be used as a starting material to construct sequencing libraries, the sequencing companies frequently ask for 1 μg of DNA.

When sending the samples, place sealed individual microcentrifuge tubes in a 50 mL disposable screw up tube for additional insulation during the shipment. Prior to sealing, pack any remaining space in the tube with clean tissue paper. Send the samples preferably with ice pack or on dry ice.

The 454 pyrosequencing platform was the first NGS introduced in the market. The GS FLX instrument generates ~400,000 reads (per instrument run) of length up to 1 kb (~800 bp). The greatest advantage of this platform is the read length, making the system well suited for larger genomic projects. The major disadvantage of this method is the misinterpretation of homopolymers (consecutive nucleotides, e.g., AAA or CCC) and, regarding the cost, the low amount of DNA sequenced in comparison with other NGS technologies.

1. To construct the 454/Roche library use GS FLX Titanium kit [32].

Illumina sequencing platform is one of the most popular in metagenomics. Here, the sequencing process takes place on a flow cell with eight channels, each of which can contain a different sample (or many samples if multiplexing, up to 96 with the Nextera system). In early 2011, Illumina released HiSeq 2000 v.3 kits, which can produce >187.5 million reads (or >37.5 Gb per lane) of 2×100 bp. MiSeq instrument offers the possibility of longer reads, 2×300 bp, but produces less quantity of data: >25 million reads (or 15 Gb) per run. There is also the option of using micro and nano flow cells which produce up to four million and one million reads per run (1.2 Gb and 500 Mb).

3.7 Bioinformatic Analyses

3.7.1 Quality Control of Sequences

1. To construct the Illumina library use Nextera[®] XT DNA Sample Preparation Kit [33].

To get reliable results in downstream analysis, it is essential to remove sequences of low quality. Phred quality scores have become widely accepted to characterize the quality of DNA sequences and are assigned to each nucleotide base call in automated sequencer traces. The most commonly used method is to count the bases with a quality score of a minimum of 30 (base call accuracy 99.9 %).

1. To trim the sequences using SolexaQA DynamicTrim (http://es.sourceforge.jp/projects/sfnet_solexaqa/downloads/src/DynamicTrim.pl.zip/) type:


```
> perl DynamicTrim.pl -h 30 trimseq input_file.fastq
```

 This command will individually crop each read to its longest contiguous segment for which quality scores are greater than a user-supplied quality cutoff (defined here as flag *-h 30*).
2. To remove duplicate sequences from 454 pyrosequencing data using CD-HIT submit your data to web server at <http://weizhongli-lab.org/cd-hit/servers.php>.

3.7.2 Introduction to Metagenomics Sequence Assembly

De novo assembly of metagenomic sequences has been successfully used to reconstruct genomes of a number of uncultivated archaea [17, 19]. Keep in mind that de novo assembly is computationally demanding in terms of memory and CPU resources.

1. To assemble the sequences using Velvet assembler type:


```
> velvetg assembly_71 71 -fastq file.fastq
```

 where *file.fastq* is input file in fastq format; *assembly_71* is the output file folder where the results are stored, and 71 is the size of k-mer used.
2. In the next step type:


```
> velvetg assembly_71 -cov_cutoff auto
```

 This command will generate assembly, which will be stored in folder *assembly_71* along with some statistical descriptors, which can be used to evaluate the quality of the assembly, including N50 (defined as the largest length L such that 50 % of all nucleotides are contained in contigs of size at least L).
3. To assemble the sequences using meta-IDBA assembler type:


```
> idba_odb -o output_file -r file.fastq -mink 70 -maxk100 -step 10 -pre_correction
```
4. To assemble larger contigs use overlap-based assembly tool Geneious (<http://www.geneious.com>). To import the data use Import command.
5. Generate contig assembly using Generate assembly command. Set strict overlapping parameters: at least 98 % of identity in 100 bp, do not allow gaps or ambiguities.

3.7.3 *Sequence
Statistics, Gene Detection,
and Functional
Classification*

6. Focus your further studies on contigs larger than 10 kb.
1. Calculate GC content for each nucleotide sequence using “geecee” from EMBOSS package (<http://emboss.sourceforge.net/download/>). Use this data to generate a plot of distribution of GC content per sequence. These plots will be either uni- or bimodal indicating the predominating GC content of distinct population subgroups.
2. GC content is used to bin the data as it carries a phylogenetic signal. For each sequence calculate tetranucleotide frequencies using the TETRA package [34] (<http://www.megx.net/tetra/index.html>). Use z scores data values derived from the frequency matrix to conduct principal component analysis (PCA) using the MeV program [35] (<http://www.tm4.org/mev.html>) or FactoMineR [36] (<http://cran.r-project.org/web/packages/FactoMineR/index.html>).
3. Calculate codon usage using “cusp” from EMBOSS package. Use this data to generate a plot of codon usage.
4. Use Glimmer [37] (<https://ccb.jhu.edu/software/glimmer/>) with the bacterial, archaeal, and plant plastid code (transl_table=11) to identify all open-reading frames (ORFs) \geq 30 amino acids.
5. Define candidate coding DNA sequences (CDS) using MetaProdigal [38] (<http://prodigal.ornl.gov/>). Match the two generated sets. Validate as genes the predicted CDS that have matches in the RefSeq database with an e value $\leq 1e-10$. Make a subset of CDS that match orphan RefSeq genes (i.e., hypothetical proteins).
6. Submit all ORFs to similarity search using BLASTP [39] (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download) against the latest version of RefSeq_protein non-redundant database (GenBank), SWISSPROT database, the clusters of orthologous group (COG) databases (COG+KOG, arCOG), and KEGG pathways database (Kanehisa Laboratories (<ftp://ftp.bioinformatics.jp/>)). Record first hits to these databases and consider the respective e values informative if they remain below the $1e-05$ threshold.
7. To search for motifs, submit all ORFs to the latest versions of conserved domain databases (CDD), Pfam, SMART, COG, arCOG, KOG, TIGR, and PRK using RPS-BLAST [40] (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download). Record first hits to these databases and consider the respective e values informative if they remain below the $1e-05$ threshold.

8. Retain the CDS that match orphan RefSeq genes if they (1) match a COG functional category; (2) contain any known motif in CDD databases provided their BLASTP and RPS-BLAST e values remain below the $1e-05$ threshold. Switch the accepted annotation to that of the relevant match.
9. Remove small (<100 aa) CDS that match orphan Refseq genes and look for significant matches in RefSeq, COG, and CDD databases (with similar e value thresholds as above). Validate these as genes if they do not overlap any other gene having high similarity in searched databases.
10. Identify tRNAs using tRNA-scanSE [41] (<http://lowelab.ucsc.edu/tRNAscan-SE/>).
11. Identify ribosomal RNA genes with rRNA_hmm_fs/hmmsearch 3.0 [42] (<http://hmmer.janelia.org/software>).

3.8 Phylogenetic and Phylogenomic Analysis

3.8.1 Phylogenetic Analysis of Metagenomic Sequences

1. Filter the metagenomic dataset for sequences of interest (e.g., using BLAST) and transfer them into a local database. Name this database “query database.”
2. Search publicly available bacterial and archaeal genomes (use one genome sequence per species) for sequences that are homologous to those of interest and gather them into a separate local database. Name this database “reference database.”
3. Align the sequences in “reference database” using MUSCLE [43] (<http://www.drive5.com/muscle/downloads.html>) or ClustalOmega [44] (<http://www.clustal.org/omega/>).
4. Detect the conserved positions in the “reference database” alignment using BMGE [45] (<https://wiki.gacrc.uga.edu/wiki/BMGE>) with default parameters and the BLOSUM62 substitution matrix.
5. Manually verify the trimmed “reference database” alignment using the program NET of the MUST package [46] (<http://megasun.bch.umontreal.ca/Software/HPLab/must/must.html>).
6. Use Prottest [47] (<https://code.google.com/p/prottest3/>) to select the best-fit models of amino acid replacement to be used in phylogenetic reconstruction.
7. Reconstruct maximum likelihood phylogenetic trees with RaxML v.7.2.4 [48] (<http://sco.h-its.org/exelixis/software.html>) using trimmed “reference database” alignment and the selected model. Estimate tree robustness using the Rapid Bootstrapping method as implemented in RaxML.
8. Separately, align the “query database” using ClustalOmega.
9. Place the aligned “query database” sequences onto the obtained reference tree using RaxML Evolutionary

3.8.2 Phylogenomic Analysis

Placement Algorithm [49] (<http://sco.h-its.org/exelixis/web/software/epa/index.html>).

1. Gather the protein sequences of publicly available bacterial and archaeal genomes (use one genomic sequence per species) into a local database.
2. Perform BLASTP searches against Refseq database using protein sequences in genomes that you are interested in as queries. Retain only those protein sequences that give matches to least four phylum-level sequences with an ϵ value threshold of $1e-05$.
3. Subject the retained protein to preliminary phylogenetic analysis. Manually inspect the phylogenetic trees and retain only sequences where the species belonging to different classes are monophyletic, irrespective of the relative order of emergence of the different classes.
4. Remove the sequence datasets that show evidence of horizontal gene transfer (HGT) (i.e., branch with members of other class/phylum).
5. Remove the sequence datasets that produced trees that suggest the presence of paralogs (i.e., contain multiple sequences for some species, or species of the same class that branch in different parts of the trees).
6. Use the remaining sequences to reconstruct final phylogenetic trees.
7. Select species that should be used as an outgroup.
8. Align the homologous protein sequences using MAFFT [50] (<http://mafft.cbrc.jp/alignment/software/>) with default parameters.
9. Perform **steps 4–7** from Subheading **3.8.1**.
10. Reconstruct Bayesian inference trees using MrBayes 3.2.1. [51] (<http://mrbayes.sourceforge.net/download.php>) with selected model. Run the four independent chains for 2,000,000 generation and sample trees every 100 trees. To construct a majority rule consensus tree, eliminate the first 5000 trees as burn-in.

3.9 Metagenomic Recruitment

In a typical metagenomics recruitment experiment, a bacterial genome is compared to the metagenome of an environment it inhabits. As a result, similarities of metagenomic sequences to genomic sequence are plotted along the length of bacterial genome. Recruitment plots are commonly used to indicate the presence of a microbe in an environment (shown by a large number of reads that are highly similar ($\geq 95\%$) to bacterial genome) and to delineate variable genomic regions (visualized as genomic regions with little or no homologous sequences in the metagenome).

1. Download the genome of interest in .fna format from <ftp://ftp.ncbi.nlm.nih.gov/>
2. Create a custom database from the genomic sequence using the following blast command:

```
>formatdb -i RefGen.fna -n CustomDatabase -p F
```

 where RefGen.fna is genome of interest in .fna format, CustomDatabase the name of database
3. Blast metagenome in fasta format against CustomDatabase as follows:

```
>blastall -p blastn -i Metagenome.fasta -o Result.blastn -d CustomDatabase -m 8 -e 1e-5 -b 1 -v 1
```

 where *blastn* specifies comparison of two nucleotide sequences, *-i* indicates the input fasta file, *-o* specifies the name of the output file, *-d* specifies the name of the custom database. *-m 8* specifies that the results will be in tabular format, *-e* indicates that only hits with e values equal or smaller than indicated will be retained, *-b 1* and *-v 1* specify that only the best hit will be retained.
4. From the resulting file, extract the column that contains the percent identity between metagenomics and genomic fragment (column 3 entitled “percent identity”) and the coordinates of metagenomics fragment in reference genome (columns 9 and 10 entitled “subject start” and “subject end”). To do this in linux environment type:

```
>cat Result.blastn | cut -f3,9,10 >ResultFinal.csv
```

 where *cat* opens the file Result.blastn, *cut -f* cuts the desired columns, *>* writes the results in novel file.
5. Plot these data using a graphic software with the complete length of the genomic sequence on the *x* axis.

3.10 Comparative (Meta)genomic Tools

Comparative metagenomics provides an insight into processes and microbial groups which confer specific characteristics to a given environment. Metagenomes typically differ in sequence composition, taxonomic diversity, population structure, and diversity of functional genes. Currently available web servers which use a number of bioinformatic tools to provide comparative metagenomics data are listed below:

Camera <http://camera.calit2.net/#>

Galaxy <https://main.g2.bx.psu.edu/u/aun1/w/metagenomic-analysis>

IMG/M <http://img.jgi.doe.gov/cgi-bin/m/main.cgi>

MEGAN <http://metagenomics.anl.gov/>

MetaLook <http://www.megx.net/metalook/index.php>

MetaMine <http://www.megx.net/metamine/>

MetaStats <http://metastats.cbc.umd.edu/detection.html>

MG-RAST <http://metagenomics.anl.gov/>

ShotgunFunctionalizeR <http://shotgun.math.chalmers.se/>

STAMP <http://kiwi.cs.dal.ca/Software/STAMP>

UniFrac <http://bmf.colorado.edu/unifrac/>

4 Notes

1. Filtration can also be performed using a positive pressure system using a dispensing pressure vessel (Millipore, 20 L 10 PSI, XX6700D20) in combination with an air compressor (Compressor DX25, DEXTER power, 24 L, 2.5 HP, 115 PSI/8 bar). We recommend not to overcome the 4 mbar to prevent the breakage of the filters.
2. When sampling for metagenomic studies, special care should be taken to not introduce contaminant DNA. It is highly recommended to collect replica samples, as these will allow to obtain sufficient quantity of nucleic acids and to evaluate the amount of technical variability between samples which arises from “noise-prone” steps in downstream analysis. When sampling for metatranscriptomic studies, special care should be taken to minimize sampling time. To this aim, it is imperative to transfer the samples to an RNA preservation buffer immediately upon sampling and to freeze the samples in liquid nitrogen.
3. The containers should be made from inert material such as glass, high density polyethylene (HDPE), or polytetrafluoroethylene (PTFE). Pre-clean the containers prior to sampling by soaking them overnight in 1:1 concentrated acid and rinsing them thoroughly with double distilled water. When sampling hot springs, keep the samples in thermal flasks until filtration.
4. Filtration can also be performed using a positive pressure system using a dispensing pressure vessel (Millipore, 20 L 10 PSI, XX6700D20) in combination with an air compressor (Compressor DX25, DEXTER power, 24 L, 2.5 HP, 115 PSI/8 bar). To prevent breaking of the filters do not overcome pressure of 4 mbar.
5. The transcription profiles of deep-sea (below 500 m depth) marine microbial communities are known to change due to the technical inability to rapidly process and preserve such samples. Although little can be done to minimize the transportation time of Niskin bottle along the water column, the processing time can be minimized by filtering small volume (1 L) of water sample immediately upon shipboard retrieval of the CTD.

6. Working with RNA requires an RNase-free working environment. To achieve this, dedicate a separate laboratory area, pipettors, and materials. Use only RNase-free reagents and plastic tubes. Wear gloves at all times and treat the gloves, the utensils, and working surfaces with RNase ZAP. Pipet at a 45° angle with open tubes facing away from you and use PCR hood. When working with low biomass samples, scale up the volume of sample used for isolation (e.g., from ~0.5 g to 25 g). To ensure the absence of aerosolized contaminants, include extraction blanks and confirm the absence of DNA and RNA contaminants by no visible amplification of 16S rRNA from extraction blanks after 35 cycles of PCR. Perform amplifications of bacterial 16S rRNA using respective environmental DNA templates, *Taq* DNA polymerase, 100 μM primers 27 F (5-AGA GTT TGA TCC TGG CTC AG-3), 1492R (5-GGT TAC CTT GTT ACG ACTT-3) and the following program: 94 °C (5 min), followed by 25 cycles of 94 °C (1 min), 45 °C (45 s), 72 °C (1 min), and a 20-min extension step at 72 °C. Perform amplifications of archaeal 16S rRNA gene using 100 μM primers Arch21F (5-TTC CGG TTG ATC CYG CCG GA) and Arch958R (5-YCC GGC GTT GAM TCC AAT T) and the following program: 94 °C (5 min), followed by 25 cycles of 94 °C (1 min), 58 °C (1 min), 72 °C (2 min), and a 20-min extension step at 72 °C.
7. For accurate quantification of DNA for next-generation sequencing, we recommend using a fluorescent dye based method (e.g., PicoGreen with Agilent 2100 Bioanalyzer System (www.genomics.agilent.com)) rather than an absorbance based method.
8. To remove contaminating polysaccharides, add 1 V of 20 % PEG 8000 and centrifuge 10 min at 20,800 × *g* and 4 °C.
9. If the quantity of cDNA is lower than 300 ng, perform an extra step of total RNA amplification (e.g., use Illumina® TotalPrep™ RNA Amplification Kit (Illumina (www.illumina.com))) to obtain sufficient cDNA for downstream sequencing.
10. Additional steps may be required to deplete the amount of ribosomal RNA (rRNA) transcripts in a sample in order to maximize mRNA recovery. However, keep in mind that a number of kits for rRNA depletion are not suitable for archaea. If quantification of transcript abundance is required following sequencing, internal RNA standards may be added to biomass samples.
11. If humic contaminants are still present in the sample upon extraction, include an additional precipitation step upon cellular lysis (addition of solutions SR1, SR2, phenol:chloroform,

and bead beating). Precipitate organic material by adding 1/10 V of 2 M sodium acetate (pH 4.0). Mix vigorously and incubate 5 min at room temperature. Centrifuge for 10 min at 4 °C and 20,800 × *g* and transfer the upper phase to a new 1.5 mL tube. Precipitate nucleic acids by adding 0.7 V of isopropanol. Incubate for 1 h at -20 °C, centrifuge for 20 min at 4 °C and 20,800 × *g*. Remove the supernatant, wash the pellet with 70 % ethanol, air-dry, and resuspend in 0.2 mL of RNase-free water. Continue with the RNA extraction protocol.

12. Cloning of low quantity DNA generally results in poor quality of a shotgun library, even if pooled from several isolations. We recommend that you proceed further only if satisfying concentrations (~500 ng/mL) of DNA are obtained.
13. To find best parameters, the number of repeats should be optimized.
14. If concentration of obtained DNA is too low, concentrate the DNA using freeze-drying. We use Speed Vac Model Savant SC 110 (Global Medical Instrumentation, Inc. Ramsey, MN, USA).
15. When size-selecting metagenomic DNA minimize the weight of cut gel slice as the amount of GELase enzyme is limited. Special care is needed not to introduce bubbles into transduction mixtures. Avoid shaking the phage extracts or transduction mixtures. Phage extracts expire by the end date. To determine the optimal ratio of packaged phage particles/cell, add 10, 20, 30, 40, and 50 μL of packaged phage particles to 100 μL of EPI300-T1® cells. Incubate the transduction mixtures 1 h at 37 °C and spread the transduction mixtures on individual LB plates supplemented with 12.5 mg/mL chloramphenicol. Incubate overnight at 37 °C. Count the colonies on each plate. Use the packaged phage/cell ratio that has yielded the highest number of clones and prepare the remaining transduction mixtures.
16. We perform single-cell genome amplification as developed by Bigelow laboratories using REPLI-g Single Cell Kit [52] designed to uniformly amplify genomic DNA from single cells without introducing bias [53].
17. To lyse archaeal cell walls, other protocols have been successfully used. Alkaline lysis has been used to obtain single amplified genomes of marine Thaumarchaeote [54] while DNA extraction under hot (60–70 °C) and alkaline conditions efficiently disrupted archaeal cells in marine sediment samples while minimizing fragmentation of DNA [55].

Acknowledgements

L.P. is supported by Slovenian Research Agency project J1-6741 and programme P1-0198. A.B.M.C. is supported by Spanish Ministerio de Economía y Competitividad project MEDIMAX BFP2013-48007-P. P.L.G. acknowledges the support of CNRS. The authors thank Céline Petitjean and David Moreira for providing rooted Bayesian phylogenetic tree of Archaea.

References

1. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* 87:4576–4579
2. Rothschild LJ, Mancinelli RL (2001) Life in extreme environments. *Nature* 409:1092–1101
3. López-García P (2005) Extremophiles. In: Gargaud M, Barbier B, Martin H, Reisse J (eds) *Lectures in astrobiology*. Springer-Verlag, Heidelberg, pp 657–679
4. Valentine DL (2007) Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat Rev Microbiol* 5:316–323
5. Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* 276:734–740
6. DeLong EF (1998) Everything in moderation: archaea as ‘non-extremophiles’. *Curr Opin Genet Dev* 8:649–654
7. DeLong EF (1992) Archaea in coastal marine environments. *Proc Natl Acad Sci U S A* 89:5685–5689
8. Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaeobacterial group from marine plankton. *Nature* 356:148–149
9. Fuhrman JA, Davis AA (1997) Widespread Archaea and novel Bacteria from the deep sea as shown by 16S rRNA gene sequences. *Mar Ecol Prog Ser* 150:275–285
10. López-García P, Moreira D, López-López A, Rodríguez-Valera F (2001) A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environ Microbiol* 3:72–78
11. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6:245–252
12. Guy L, Ettema TJ (2011) The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol* 19:580–587
13. Quaiser A, Ochsenreiter T, Klenk HP, Kletzin A, Treusch AH, Meurer G, Eck J, Sensen CW, Schleper C (2002) First insight into the genome of an uncultivated crenarchaeote from soil. *Environ Microbiol* 4:603–611
14. Schleper C, Jurgens G, Jonuscheit M (2005) Genomic studies of uncultivated archaea. *Nat Rev Microbiol* 3:479–488
15. Nicol GW, Schleper C (2006) Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? *Trends Microbiol* 14:207–212
16. Pester M, Schleper C, Wagner M (2011) The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* 14:300–306
17. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335:587–590
18. Deschamps P, Zivanovic Y, Moreira D, Rodríguez-Valera F, Lopez-García P (2014) Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome Biol Evol* 6:1549–1563
19. Martin-Cuadrado AB, Garcia-Heredia I, Molto AG, Lopez-Ubeda R, Kimes N, Lopez-García P, Moreira D, Rodríguez-Valera F (2015) A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. *ISME J* 9(7):1619–1634. doi:[10.1038/ismej.2014.249](https://doi.org/10.1038/ismej.2014.249)
20. Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D’Auria G, Zhaxybayeva O, Papke RT, Rodríguez-Valera F (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J* 1:235–245
21. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA,

- Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woynke T (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437
22. Hugoni M, Taib N, Debroas D, Domaizon I, Jouan Dufournel I, Bronner G, Salter I, Agogue H, Mary I, Galand PE (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proc Natl Acad Sci U S A* 110:6004–6009
 23. Martin-Cuadrado AB, Rodriguez-Valera F, Moreira D, Alba JC, Ivars-Martinez E, Henn MR, Talla E, Lopez-Garcia P (2008) Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J* 2:865–886
 24. PowerSoil® DNA isolation kit instruction manual. www.mobio.com/images/custom/file/protocol/12888.pdf
 25. RNeasy mini handbook. <http://www.qiagen.com/si/resources/resourcedetail?id=14e7cf6e-521a-4cf7-8cbc-bf9f6fa33e24&lang=en>
 26. QIAquick spin handbook. <http://www.qiagen.com/si/products/catalog/sample-technologies/dna-sample-technologies/dna-cleanup/qiaquick-pcr-purification-kit/#resources>
 27. RNA PowerSoil® Total RNA isolation kit instruction manual. <http://www.mobio.com/images/custom/file/protocol/12866-25.pdf>
 28. pGEM®-T easy vector system technical manual. <http://www.promega.com/~media/files/resources/protocols/technical%20manuals/0/pgem-t%20and%20pgem-t%20easy%20vector%20systems%20protocol.pdf>
 29. Protocol for CopyControl™ Fosmid library production kit with pCCI1Fos vector. <http://www.epibio.com/docs/default-source/protocols/copycontrol-fosmid-library-production-kit-with-pcc1fos-vector.pdf?sfvrsn=6>
 30. Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR (2011) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS One* 6, e16626
 31. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 7, e30087
 32. GS FLX Titanium Rapid Library preparation kit. <http://lifescience.roche.com/shop/products/gs-flx-titanium-rapid-library-preparation-kit>
 33. Nextera® XT DNA sample preparation kit. http://support.illumina.com/sequencing/sequencing_kits/nextera_xt_dna_kit.html
 34. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5:163
 35. Saced AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisnt J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovuch E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34:374–378
 36. Lê S, Josse J, Husson F (2008) FactoMineR: an R package for multivariate analysis. *J Stat Softw* 25:1
 37. Delcher A, Harmon D, Kasif S, White O, Salzberg S (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27:4636–4641
 38. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28:2223–2230
 39. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
 40. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokadov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) CDD: a conserved domain database for protein classification. *Nucleic Acids Res* 33:D192–D196
 41. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964
 42. Huang Y, Gilna P, Li W (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 25:1338–1340
 43. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113

44. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins D (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539
45. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210
46. Philippe H (1993) MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res* 21:5264–5272
47. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164–1165
48. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690
49. Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol* 60:291–302
50. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780
51. Ronquist F, Teslenko M, van der Mark P, Azres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBazes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61:539–542
52. Repli-G single cell kit. <http://www.qiagen.com/si/resources/resourcedetail?id=38fac1c-64b0-4281-aab3-aa8324bbd181&lang=en>
53. Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* 9:1038–1048
54. Luo H, Tolar BB, Swan BK, Zhang CL, Stepanauskas R, Moran MA, Hollibaugh JT (2014) Single-cell genomics shedding light on marine Thaumarchaeota diversification. *ISME J* 8:732–736
55. Morono Y, Terada T, Hoshino T, Inagaki F (2014) Hot-alkaline DNA extraction method for deep-subseafloor archaeal communities. *Appl Environ Microbiol* 80:1985–1994
56. Petitjean C, Deschamps P, López-García P, Moreira D, Brochier-Armanet C (2015) Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of life. *Mol Biol Evol* 32(5):1242–1254. doi:10.1093/molbev/msv015 (advanced access publication)

Investigating the Endobacteria Which Thrive in Arbuscular Mycorrhizal Fungi

Alessandro Desirò, Alessandra Salvioli, and Paola Bonfante

Abstract

The study of the so-called unculturable bacteria is still considered a challenging task. However, given recent improvements in the sensitivity of culture-free approaches, the identification and characterization of such microbes in complex biological samples is now possible. In this chapter we report how endobacteria thriving inside arbuscular mycorrhizal fungi (AMF), which are themselves obligate biotrophs of plants, can be studied using a combination of in vitro culture, molecular biology, and microscopy techniques.

Key words Endobacteria, Arbuscular mycorrhizal fungi, *Candidatus* Glomeribacter gigasporarum, Mollicutes-related endobacteria, Transmission electron microscopy, Fluorescent in situ hybridization, Real-time quantitative PCR, Bacterial enrichment

1 Introduction

Thousands of microbes are commonly associated with plant roots, forming the so-called root microbiota which plays a pivotal role in plant life. Among them, a group of soil fungi that colonize the roots of most land plants, the arbuscular mycorrhizal fungi (AMF), has a key role in improving mineral nutrition and protection of their host plant. AMF, which belong to the phylum Glomeromycota [1], have been considered as the oldest group of fungi capable of positively interacting with plants: they have been hypothesized to be crucial for the terrestrialization of first land plants c. 450 Mya [2]. In addition to some distinctive features (AMF are asexual, multinucleated and obligated biotrophs), Glomeromycota may harbor endobacteria in their cytoplasm [3]. Two types of endobacteria have been so far described and identified in AMF: a rod-shaped Gram-negative β -proteobacterium called *Candidatus* Glomeribacter gigasporarum (*CaGg*) [4] and a coccoid bacterium which represents a still enigmatic taxon of Mollicutes-related endobacteria (*Mre*) [5]. Differently from *Mre*, which show a wide distribution across the Glomeromycota,

the presence of *CaGg* is limited to the Gigasporaceae family. *CaGg* has been deeply investigated: its genome sequence has revealed that the endobacterium is nutritionally dependent on the fungus and has a potential role in providing its host with essential factors like vitamin B12 [6]. The fungus, on the contrary, is not obligately dependent on the endobacterium, even if the removal of *CaGg* causes some morphological changes in *Gigaspora* spores and a reduced proliferation of the presymbiotic fungal hyphae [7]. A recent step forward was made in the understanding of this specific fungal-endobacterial association: it has been demonstrated that the *CaGg* endobacterium improves the fungal fitness by increasing the sporification success, priming the mitochondrial activity and rising the detoxification of reactive oxygen species. Furthermore, the bacteria seem to enhance the fungal responsiveness to plant strigolactones, which are perceived by AMF as branching factors [8]. By contrast, information on *Mre* is much more limited: based on 16S rRNA gene sequences, this novel bacterial taxon represents the sister clade of the Entomoplasmatales and Mycoplasmatales, within the Mollicutes and shows high level of sequence variability [5, 9]. Only recently the genome of *Mre* has been sequenced showing i) a highly reduced gene content and metabolic capacity, ii) metabolic dependence on the fungal host, and iii) extensive chromosomal rearrangements and trans-kingdom gene transfer between the two partners [10, 11]. Interestingly, *CaGg* and *Mre* have been simultaneously detected in some *Gigaspora margarita* isolates (fam. Gigasporaceae) hosting what has been described as a new fungal microbiota [9]. Hence, single or multiple bacterial populations can thrive inside AMF that, being themselves obligate symbionts, need a host plant to complete their life cycle. Consequently, this complex “Russian dolls-like” organization is difficult to dissect with traditional culture-based techniques. Furthermore, the endobacteria so far detected in AMF are considered unculturable microbes; therefore, they cannot be obtained in pure culture.

In this chapter, we present several techniques that can be applied to investigate the endobacteria thriving in AMF. First, we describe three methods which have been set up to obtain the fungal material and, thereby, their associated endobacteria. The in pot culture is the recommended method for *G. margarita* routine propagation and the obtainment of large amounts of spores. The Millipore sandwich allows producing clean and intact extraradical mycelium, while the root organ cultures (ROCs) can be used to obtain in vitro spores and mycelium. As for bacteria living inside insect [12], the demonstration of the intracellular localization is the first criterion to speak about endobacteria. Therefore, as a second step, we illustrate the protocols to process the biological material for transmission electron and confocal microscopy, in order to detect the endobacteria and describe their morphology. Third, we present some molecular techniques that make possible a direct investigation of the endobacteria

from their fungal hosts. Having as a prerequisite the availability of specific primers and probes, PCR allows to identify the endobacteria, real-time quantitative PCR (qPCR) to evaluate their abundance and fluorescent in situ hybridization (FISH) to reveal and simultaneously localize their presence. Last, we describe a filtering-based protocol first described by Ghignone and colleagues [6] that allows to obtain a spore lysate enriched in the endobacterial component with a limited carryover of fungal nuclei. All the techniques described in this chapter have been applied with success to the study of the endobacteria within AMF [5–9, 13–15].

The presence of endobacteria thriving inside the cytoplasm of fungi has been reported few times so far. In addition to Glomeromycota, endobacteria have been described inside other groups of fungi, such as Mucoromycotina. Among them, *Rhizopus microsporus*, a rice pathogenic fungus whose pathogenicity is related to the presence of a strain of *Burkholderia rhizoxinica* [16]; *Mortierella elongata*, a filamentous fungus which hosts a *Burkholderia*-related endobacterium [17]; *Endogone*, one of the oldest plant-associated fungi which host Mre in fruiting body-forming spores [18]. All these findings strongly suggest that the presence of endobacteria in the cytoplasm of fungi is more widespread than expected. Novel bacterial populations still wait to be discovered and characterized. As a consequence, the application of protocols which allow to unambiguously detect and identify these bacterial dwellers of fungi will drastically improve the knowledge of such complex symbiotic systems.

2 Materials

Prepare all solutions using analytical reagents and deionized water (unless indicated otherwise). Use sterile consumables and reagents, or, prior to use, autoclave them at 120 °C for 20 min. Wear personal protective equipment, carefully handle dangerous reagents, and follow all waste disposal current regulations when disposing of waste materials.

2.1 Biological Materials and Growing Media

The fungal material employed in the experiments described in the *Biological Materials and Growing Media* sections (see Subheadings 2.1 and 3.1) consists of monoxenic inocula of the AMFG. *margarita* BEG34, purchased from specialized companies and/or in-house propagated following the in pot culture technique (see Subheading 3.1.1).

2.1.1 Pot Cultures

1. Plastic pots (0.9 L volume).
2. Sterilized quartz sand (oven-sterilized at 180 °C for 3 h).
3. *Trifolium repens* (clover) seeds (see Note 1).
4. AMF spores.

5. Long Ashton fertilization solution [19], modified (low phosphate): 0.75 mM MgSO₄ 7H₂O, 1 mM NaNO₃, 1 mM K₂SO₄, 2 mM CaCl₂ 2H₂O, 32 μM Na₂HPO₄, 0.025 mM Fe Na EDTA, 0.005 mM MnSO₄ 12H₂O, 0.00025 mM CuSO₄ 5H₂O, 0.0005 mM ZnSO₄ 7H₂O, 0.025 mM H₃BO₃, 0.0001 mM Na₂MoO₄ 2H₂O.
6. Sieves (aperture 100 μm).

2.1.2 Spore Sterilization

1. Ultrapure water.
2. Chloramine T.
3. Streptomycin sulfate.

2.1.3 *Lotus japonicus* Seeds Sterilization and Germination

1. Sulfuric acid (99.99 %).
2. Sterile water.
3. Water agar (0.6 % agar) plates, diameter 9 cm.

2.1.4 Millipore Sandwich Method

1. Water agar (0.6 % agar) plates, diameter 9 cm.
2. Magenta boxes.
3. Sterile cellulose nitrate membranes (pore size 45 μm).
4. Sterilized quartz sand (oven-sterilized at 180 °C for 3 h).
5. 1:2 diluted modified Long Ashton solution (see Subheading 2.1.1).
6. Surface-sterilized AMF spores.
7. 3–4 days germinating *L. japonicus* seedlings.

2.1.5 In Vitro Fungal Propagation under Root Organ Culture (ROC) Conditions

1. *Cichorium intybus* (chicory) transformed root cultures (see Note 2).
2. Surface-sterilized AMF spores.
3. Minimal (M) Medium [21]: 3 mM MgSO₄ 7H₂O, 0.79 mM KNO₃, 0.87 mM KCl, 1.22 mM Ca(NO₃)₂ 4H₂O, 35.0 μM KH₂PO₄, 21.7 μM Na Fe EDTA, 4.5 μM KI, 30.3 μM MnCl₂ 4H₂O, 9.2 μM ZnSO₄ 7H₂O, 24.0 μM H₃BO₃, 0.5 μM CuSO₄ 5H₂O, 0.01 μM Na₂MoO₄ 2H₂O, 40 μM glycine, 0.3 μM Thiamin HCl, 0.5 μM Pyridoxin HCl, 4 μM Nicotinic Acid, 277 μM Myo-inositol, 10 g/L Sucrose, Phytigel 4 g/L, pH 5.5.

2.2 Morphological Analyses

2.2.1 Transmission Electron Microscopy

1. Ultramicrotome with glass and diamond knives.
2. 1 % toluidine blue (w/v).
3. Polypropylene spot plate (well Ø 2 cm).
4. Hot plate.
5. Uranyl acetate solution: prepare a saturated solution by dissolving uranyl acetate in ddH₂O.
6. NaOH (pellets).

7. Lead citrate solution.
 - (a) Solution A (Lead nitrate stock solution): add 31.25 g in 500 mL of doubled-distilled water (ddH₂O). Add ten drops of HNO₃ 10 N (HNO₃ 10 N: 630 g in 1 L of ddH₂O).
 - (b) Solution B (Sodium citrate stock solution): add 41.50 g in 500 mL of ddH₂O. Add five drops of solution A.
 - (c) Solution C (NaOH 1 N): add 0.2 g in 5 mL of ddH₂O.
Prepare the solution as follows: 2.1 mL solution A+2.1 mL solution B. Mix. Add 0.8 mL solution C. Mix.
Solution A and B can be stored at room temperature. Solution C must be freshly prepared every time.
8. Philips CM10 transmission electron microscope (FEI, Hillsboro, OR, USA).

2.2.2 Confocal Microscopy

1. SYTO 9[®] Green-Fluorescent Nucleic Acid Stain (Life Technologies, Carlsbad, CA, USA), 5 mM solution in DMSO. Store at -20 °C and protect from light. Freshly prepared working solution for bacterial (50 nM–20 μM) and eukaryotic (10 nM–5 μM) cell visualization: make a 1:1000 dilution in ultrapure water. Vortex to mix. Store on ice and protect from light.
2. Leica TCS-SP2 confocal microscope (Leica Microsystems, Wetzlar, Germany).

2.3 Molecular Analyses

2.3.1 DNA Extraction

Rapid DNA Extraction

CTAB-Based DNA Extraction

1. 2× CTAB (Cetyltrimethylammonium bromide) extraction buffer: 100 mM Tris-HCl, pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2 % CTAB (w/v). Adjust the volume of the solution with ultrapure water. Autoclave for 20 min at 120 °C.
2. Polyvinylpyrrolidone (PVP).
3. Chloroform:isoamyl alcohol (24:1).
4. Chloroform.
5. Cold 2-propanol.
6. Cold 70 % ethanol.

2.3.2 PCR

1. PCR reagents: 10 μM of suitable primers (*see* Table 1), 2.5 mM of each dNTP, 5× Phusion[®] HF Buffer, Phusion[®] DNA Polymerase (2 U/μL) (Thermo Fisher Scientific, Waltham, MA, USA), ultrapure water.

Table 1
List of the primers used in PCR and qPCR experiments

	Target gene	Primer pair	Primer sequence (5'–3')	Expected amplicon size	References
Bacterial target (<i>Ca</i> Gg)	16S rRNA gene	CaGgADf	AGATTGAACCGCTGGGGCAT	1460 bp	Desirò et al. [9]
	16S rRNA gene	CaGgADr	ATGCGTCTACCGTGGCCATC		
	16S rRNA gene	CaGgAD7f	CACCTAAGGAGACTGCCAGTGAC	121 bp	
	23S rRNA gene	CaGgAD6r	AGGTGGCATCCCTCTGTACAG	587 bp	Bianciotto et al. [13]
Bacterial target (<i>Mre</i>)	16S rRNA gene	GlomGIGf	GGGTCCATTGCGGATTACTTC		
	23S rRNA gene	GlomGIGr	GGGACCAGGACTTCCATCCCCC	106 bp	Bianciotto et al. [13]
Fungal target	16S rRNA gene	GlomGIGf	GGGTCCATTGCGGATTACTTC		
	16S rRNA gene	GlGrA	GTGTGGCCCTCTTGACACC		Lumini et al. [7]
	109F	ACGGGTGAGTAAATRCTTATCT (109F-1)		1040–1090 bp	Naumann et al. [5]
	2:1 mixture of 1184R	ACGAGTGAGTAATGCTTATCT (109F-2)			
Bacterial target (<i>Mre</i>)	16S rRNA gene	2:1 mixture of 1184R	GACGACCAGACGTGATCCTY (1184R-1)		
	16S rRNA gene	2:1:1 mixture of CMsAD1f	GACGACCAAACTTGATCCTC (1184R-2)		
	16S rRNA gene	CMsAD1f	GATGATCAGACGTGATCCTC (1184R-3)	81–107 bp	Desirò et al. [9]
	16S rRNA gene	CMsAD2r	GAKGAAGGTCTAYGGATTGTAAACTTCTGGCACRTAGTAGTCGTG		
Fungal target	18S rRNA gene	AML1	ATCAAACITTCGATGGTAGGATAGA	800 bp	Lee et al. [23]
	18S rRNA gene	AML2	GAACCCAAACACTTTGGTTTC		
	ITS region	ITS1F	CTTGGTCATTTAGAGGAAGTAA	520–610 bp	Gardes and Bruns [24]
	ITS region	ITS4	TCCCTCCGTTAATTGATATGC		White et al. [25]
	Ef1- α	Ef1g1f	CGTTCCAATATCTGGTTGGCATGGTG	289 bp	Salvioli et al. [14]
	Ef1- α	Ef1g2r	GGTAAGACCAACTGGGGCGAATG		
	Ef1- α	Ef1g2f	TGAAACCTCCAAACAGACCAACTG	130 bp	
	Ef1- α	Ef1g1r	GGTTTCAACACGACCTACAGGGAC		
	rpoB	rpoBf	TCCGAGCTGTCCGAGTTTCAT	536 bp	
	rpoB	rpoBr	CGCTGCATGTTGAGCCCCAT		
rpoB	RpoBRTf	CGCGGCAAAGTCAACGGATAC	109 bp		
rpoB	RpoBRTr	ATCGGTGAGTGGCCCATCCTC			

2. PCR purification and clean-up: Wizard® SV Gel and PCR Clean-Up System (Promega, Fitchburg, Wisconsin, USA).

2.3.3 Cloning

1. Cloning vector: pGEM®-T Easy Vector Systems (Promega, Fitchburg, WI, USA).
2. Chemically competent bacterial cells: One Shot® TOP10 Chemically Competent *Escherichia coli* (Life Technologies, Carlsbad, CA, USA).
3. Fresh LB plates (9 cm diameter) with ampicillin (final concentration 100 mg/mL).

2.3.4 Real-Time qPCR

1. 48-well StepOne™ Real-time PCR system and StepOne™ software (Life Technologies, Carlsbad, CA, USA) or similar Real-time PCR equipment.
2. Qubit® 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA).
3. PCR reagents: 3 µM of each suitable primers (*see* Table 1), 2× Power SYBR® Green PCR Master Mix (Life Technologies, Carlsbad, CA, USA), ultrapure water.

2.4 Fluorescent In Situ Hybridization (FISH)

1. Sterile 1× and 10× phosphate buffered saline (PBS), pH 7.2.
2. Fixative solution: 4 % paraformaldehyde [26]. Heat up 45 mL of ultrapure water at 55–58 °C (avoid exceeding 60 °C). Add 2 g of paraformaldehyde and stir with a magnet. If necessary, add a few drops of NaOH 10 N while stirring continuously until powder dissolves. Add 5 mL of 10× PBS. Cool on ice. Adjust pH to 7.2–7.4 with HCl. Filter the solution with 0.45 µm filter. Store the solution a few days at 4 °C, or 2–3 weeks at –20 °C. Avoid repeated freeze-thaw cycles.
3. Microscope slides with eight individual wells (well Ø 6 mm) and large cover slides.
4. Low melting point agarose.
5. 50 %, 75 %, and 100 % ethanol.
6. Coplin jars.
7. Hybridization oven.
8. Proteinase K: prepare a 1 mg/mL stock solution in ultrapure water. Aliquot and store at –20 °C. Prepare a 10 µg/mL working solution in ultrapure water.
9. RNase A (40 µg/µL).
10. Tween 20.
11. Suitable labeled oligonucleotide probes (*see* Table 2). Probes are 5'-end labeled with fluorochromes like fluorescein isothiocyanate (FITC) or cyanine dyes (Cy3 or Cy5). Resuspend in ultrapure water the probes to obtain a 500 ng/µL probe

Table 2
List of the oligonucleotide probes used in FISH experiments

Target gene	Organism	Probe	Probe sequence (5'–3')	Fluorochrome	References
16S rRNA gene	<i>Ca Gg</i>	CaGgADf1	CTATCCCCCT CTACAGGAYAC	Cy5	Desirò et al. [9]
	Mre	BLOsADf2	ATCCRTAGACC TTCMTCCTTC	Cy3	Desirò et al. [15]
	Bacteria	EUB338	GCTGCCTCCC GTAGGAGT	Fluorescein	Amann et al. [26]
	<i>Buchmera</i>	Apis2Pa	CCTCTTTGGG TAGATCC	Fluorescein	Koga et al. [27]
None	None	non-CaGgADf1	GTRTCCTGTAG AGGGGGATAG	Cy5	Desirò et al. [9]
		non-BLOsADf2	GAAGGAKGAAGGT CTAYGGAT	Cy3	Desirò et al. [15]

stocks. Aliquot and store the stocks in the dark at -20°C . Dilute probe stocks at the working concentration of $50\text{--}70\text{ ng}/\mu\text{L}$ with ultrapure water and aliquot in individual tubes ($50\text{ }\mu\text{L}$ per tube). Avoid repeated freeze-thaw cycles. Store in the dark at -20°C .

12. Sterile $20\times$ saline-sodium citrate (SSC) buffer (3 M NaCl and 0.3 M sodium citrate, pH 7). Aliquot and store at -20°C .
13. 100% formamide. Aliquot and store at -20°C (*see Note 3*).
14. $50\times$ Denhardt's solution: dissolve the Denhardt's powder in ultrapure water according to the manufacture's instruction. Store at -20°C for up to 2 years. Prepare $25\times$ working solution. Aliquot and store at -20°C .
15. Antifade mounting medium: 1,4-Diazabicyclo[2.2.2]octane (DABCO) solution ($25\text{ mg}/\text{mL}$). Dissolve 250 mg DABCO in 10 mL $1\times$ PBS. Add 90 mL Glycerol. Adjust pH to 8.6 with HCl or NaOH. Store the solution at 4°C .
16. Leica TCS-SP2 confocal microscope (Leica Microsystems, Wetzlar, Germany).

2.5 Bacterial Enrichment

1. 0.9% NaCl in ultrapure water (w/v). Autoclave for 20 min at 120°C .
2. Sterile plastic pestles from 1.5 mL tubes.
3. Sterile $3\text{ }\mu\text{m}$ cellulose nitrate filters.
4. Sterile syringes and syringe filter holders.
5. RQ1 RNase-Free DNase ($1\text{ U}/\mu\text{L}$) (Promega, Fitchburg, WI, USA).

3 Methods

3.1 Biological Materials and Growing Media

3.1.1 Pot Cultures

The in pot culture is the recommended method for *G. margarita* routine propagation and for the obtainment of large amounts of spores, since it represents a “nearly natural” method to obtain AMF material under controlled conditions.

1. Place the oven-sterilized quartz sand in 0.9 L pots.
2. Soak the sand with the modified low phosphate Long Ashton solution and let it drain. Inoculate 100 *G. margarita* spores under the sand surface by pipetting.
3. Spread 80–100 *T. repens* seeds on the sand surface and cover them with a thin sand layer.
4. Put the pots in a climatic chamber with a photoperiod of 16 h and a temperature of 23 °C during the day and 21 °C during the night. Keep the pots in culture for at least 3 months.
5. During the entire culturing period, fertilize the pots once a week with the modified low phosphate Long Ashton solution. Irrigate the pots with water whenever needed.
6. After 3 months of cultivation, collect the newly formed spores from the sand soil by putting a 100 mL aliquot of substrate in a beaker and adding water. Shake the beaker so that spores are temporarily kept in suspension and immediately pour the water in a 100 µm aperture sieve. Recover the sieve content in a large glass Petri dish by washing it with water. Manually collect the spores under a stereomicroscope by pipetting with a P1000 pipette or by individually collecting them with laboratory tweezers (*see Note 4*).

3.1.2 Spore Sterilization

The entire procedure should be performed under a biological hood.

1. Prepare a sterilization solution containing 3 % chloramine T and 0.03 % Streptomycin sulfate in ultrapure water. Shake well until the powders are completely dissolved. Typically, 50 mL of solution is prepared to sterilize up to 2000 AMF spores.
2. Place the collected spores in a tube and remove the residual water. Add the sterilization solution and shake the tube. Typically, 1.5 mL of sterilization solution is used for 100 spores (*see Note 5*).
3. Place the tube horizontally, so that spores are not pelleted at the bottom, and wait 10 min.
4. Remove the sterilization solution by pipetting and add the same volume of ultrapure water; shake well and wait 5 min.
5. Remove the water by pipetting and add the same volume of sterilization solution. Shake well, place the tube horizontally, and incubate 10 min.

6. Remove the sterilization solution by pipetting and add the same volume of ultrapure water; shake well and wait 5 min.
7. Remove the water by pipetting and add the same volume of ultrapure water; shake well and wait 10 min.
8. Repeat **step 7**. After having waited 10 min, remove by pipetting all traces of water.

Spores are now ready for subsequent treatments. If their intended use is DNA/RNA extraction, place tubes in liquid nitrogen and store the frozen spores at $-80\text{ }^{\circ}\text{C}$ until use. If they will be employed for further vital manipulation, store them at $4\text{ }^{\circ}\text{C}$ for at most 1 week.

3.1.3 *Lotus japonicus* Seeds Sterilization and Germination

The procedure should be performed under a chemical hood and wearing suitable gloves until **step 4**, since sulfuric acid is toxic and corrosive.

1. Extract *L. japonicus* seeds from pods and place them in a plastic tube (*see Note 6*).
2. Add sulfuric acid to the tube so that seeds are completely soaked.
3. Mix well by vortexing and leave the seeds soaked for 3 min.
4. Eliminate the sulfuric acid by pipetting and rinse the seeds with sterile water for 10 min.
5. Repeat **step 4** twice.
6. Under a biological hood, take individual seeds with flame-sterilized tweezers and place them on water agar plates. Put approx. 5 seeds per plate.
7. Incubate plates containing the surface-sterilized seeds in the dark at $22\text{ }^{\circ}\text{C}$ for 4 days, then put them in the light at the same temperature until the cotyledons become green (*see Note 7*).

3.1.4 Millipore Sandwich Method

The Millipore sandwich method is the technique of choice to obtain clean and intact *G. margarita* extraradical mycelium, allowing at the same time the collection of colonized root portions. The entire procedure needs to be performed under a biological hood.

1. Fill the Magenta boxes with quartz sand for 1/3 of their volume. Autoclave for 20 min at $120\text{ }^{\circ}\text{C}$.
2. Pick an autoclaved cellulose nitrate membrane with flame-sterilized tweezers and place it on an water agar plate to let it moisten.
3. Take a germinated *L. japonicus* seedling (*see Subheading 3.1.3*) and place it on the membrane, with the shoot apex going beyond the edge.
4. Collect with a pipette 20–25 surface-sterilized *G. margarita* spores and place them below the seedling rootlet; eliminate the excess liquid by pipetting.

5. Pick a second cellulose nitrate membrane with flame-sterilized tweezers and superpose it to the first one to close the sandwich.
6. Open the Magenta box and soak the sand with the 1:2 diluted modified Long Ashton solution.
7. With long, thick laboratory tweezers dig a groove in the sand, at the center of the Magenta box.
8. With long, thick laboratory tweezers take the previously prepared sandwich from the water agar plate and place it vertically in the groove, with the plantlet apex upside. Gently close the groove and soak again with the 1:2 diluted modified Long Ashton solution (*see Note 8*).
9. Eliminate the excess of 1:2 diluted modified Long Ashton solution by pipetting and close the Magenta box.
10. Incubate the Magenta boxes in a climatic chamber with a photoperiod of 16 h and a temperature of 23 °C during the day and 21 °C during the night.
11. After 30 days, disassemble the sandwiches and collect the material under a stereomicroscope.
12. For the external mycelium, peel the root surface and place the collected hyphal bundle in tubes; for mycorrhizal roots, cut the root fragments in small pieces with a sharp scalpel and collect the material in tubes. Freeze the samples in liquid nitrogen and store them at -80 °C until use.

3.1.5 *In Vitro Fungal Propagation under Root Organ Culture (ROC) Conditions*

Root organ cultures can be used to monoxenically produce *in vitro* *G. margarita* BEG34 spores and mycelium (*see Note 9*).

1. Propagate clones of root-inducing T-DNA-transformed roots previously established by subculturing them every 4 weeks on minimal (M) medium in round Petri dishes (9 cm diameter). Keep the dishes in the dark at 26 °C.
2. Using flame-sterilized tweezers, transfer a 4–5 cm long T-DNA transformed root fragment in the center of a round Petri dish (9 cm diameter) containing M medium. Gently sink the root explant below the medium surface with tweezers to avoid desiccation.
3. Transfer about ten surface-sterilized *G. margarita* spores in the plate, all around the transformed root explant.
4. Carefully seal the dishes with Parafilm and incubate them in the dark at 26 °C.
5. A new spore generation is produced within 1.5–2 months as a result of mycorrhizal colonization. To keep the spores sterile, collect them under a biological hood using flame-sterilized tweezers.

3.2 Morphological Analyses

3.2.1 Transmission Electron Microscopy

In order to preserve fungal structures and organelles, as well as the small endobacteria, single spores were processed by using cryo-methods, that is, high-pressure and freeze-substitution preparation. Subsequently, spore samples were infiltrated with Epon/Araldite resin and then embedded in resin blocks. For details on the cryo-preparation and the subsequent resin infiltration and polymerization refer to [9]. In this section, we provide details on the processing of the samples for transmission electron microscopy starting from the sectioning of the resin blocks.

1. Pre-warm the hot plate to about 50 °C.
2. After embedding, the resin blocks are sectioned by using an ultramicrotome. Cut the blocks into semithin sections (1 µm) with a glass knife.
3. Place the sections on a microscope slide. Stain the sections with 1 % toluidine blue (w/v). Let the microscope slide dry on the hot plate (about 50 °C). Observe the stained sections under a light microscope for the orientation of the sample within the block. Select a small area of the section for ultrathin sections.
4. Cut the selected area of the block into ultrathin sections (70 nm) with a diamond knife. Treat ultrathin sections as floating sections until the end of the counterstaining step.
5. Prepare a humid chamber to prevent the sections from drying out. Place a wet paper towel inside a covered Petri dish (Ø 9 cm). Place the spot plate on the paper towel.
6. Fill the spot plate well with 500 µL of uranyl acetate solution. Lay down the sections on the solution for 20 min. Cover the Petri dish and put it on the hot plate (about 50 °C). Place a piece of aluminum foil on the cover of the dish to keep the samples in the dark during the staining.
7. Rinse twice the sections in water for 10 min.
8. Fill the spot plate well with 500 µL of lead citrate solution. Lay down the sections on the solution for 2–3 min. In order to prevent excessive lead citrate precipitation by exposure to CO₂, add NaOH pellets (~10 g) near the spot plate. Cover the Petri dish during the staining.
9. Rinse twice the sections in water for 10 min.
10. Mount the sections on a 200- or 300-mesh copper grids.
11. Observe under a transmission electron microscope (*see* Fig. 1a, b).

3.2.2 Confocal Microscopy

Prepare a fresh aliquot of the SYTO 9® Green-Fluorescent Nucleic Acid Stain working solution (1:1000 dilution in ultrapure water). Store on ice and protect from light.

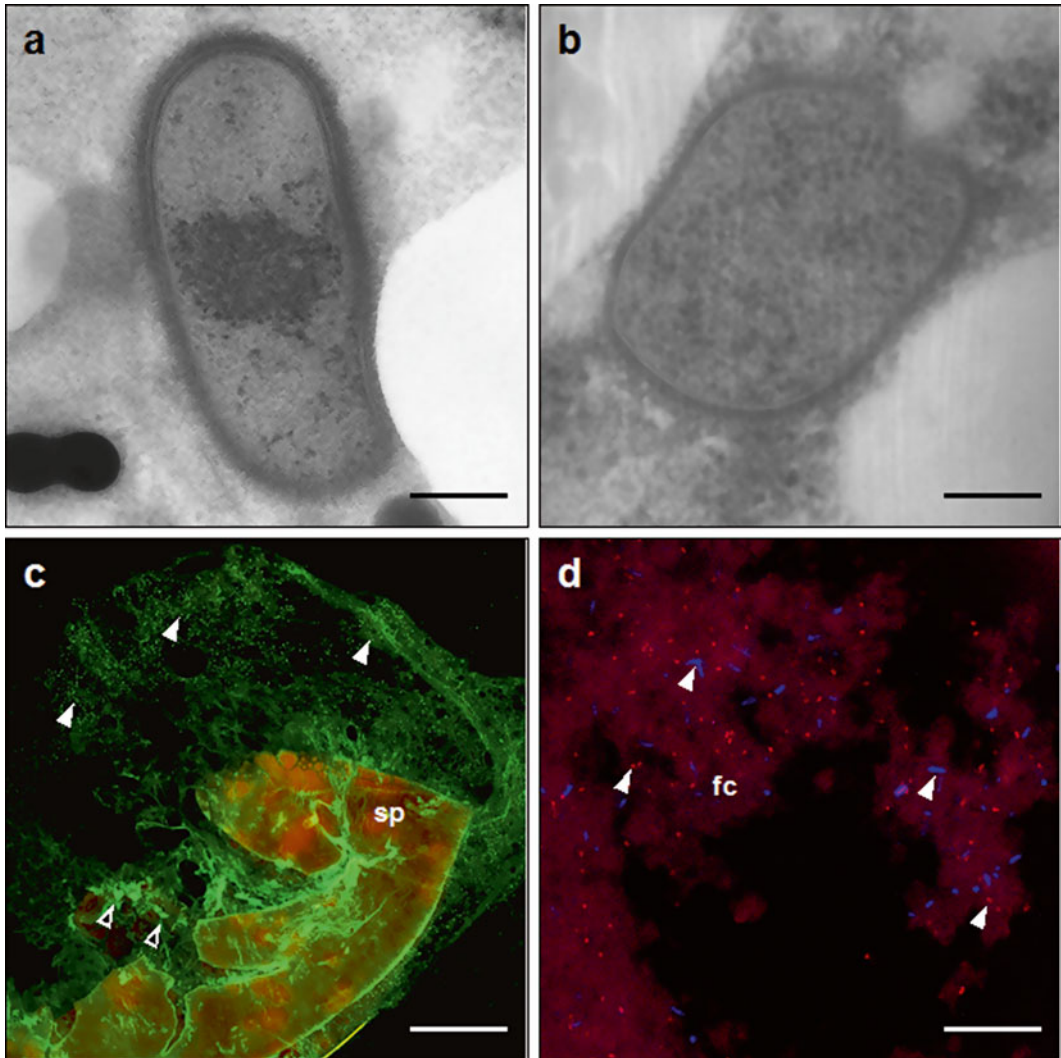


Fig. 1 Transmission electron and confocal microscopy of *Gigaspora margarita* (a, c, d) and *Rhizophagus clarus* (b) spores. Ultrastructure of (a) the rod-shaped *Candidatus Glomeribacter gigasporarum* and (b) the coccoid Mollicutes-related endobacterium as seen under a transmission electron microscope. (c) Crushed *G. margarita* spore (sp) after staining with SYTO 9[®]: the cytoplasm spreads over the slide forming a halo rich in endobacteria (arrowheads). Fungal nuclei (empty arrowhead) are trapped inside the cytoplasm. (d) FISH on a crushed spore of *G. margarita*: the double labeling with the *CaGg*-specific probe *CaGgAD1f* (blue) and the *Mre*-specific probe *BLOsADf2* (red) confirms the simultaneous presence of the two endobacterial types in the same AMF spore; bacteria are seen as rod-shaped or coccoid fluorescent spots (arrowheads). Fungal cytoplasm (fc). Scale bars, (a) 0.13 μm ; (b) 0.12 μm ; (c) 150 μm ; (d) 13 μm

Spore Processing

1. Transfer surface-sterilized spores (one to three spores per slide) on a microscope slide (*see Note 10*).
2. Add a 40–100 μL drop (depending on the size and number of spores) of SYTO 9[®] directly on the spores.
3. Add a large cover slide covering the entire microscope slide. Slightly press the cover slide down until you crash the spores.
4. Incubate the microscope slide for 5 min in the dark.
5. Observe under a confocal microscope (*see Note 11*) (*see Fig. 1c*).

Bacterial Suspension Processing

1. Transfer 10 μL of filtered bacterial suspension (*see Subheading 3.5*) on a microscope slide.
2. Add 10 μL drop of SYTO 9[®] directly on the bacterial suspension and mix gently by pipetting.
3. Add a large cover slide covering the entire microscope slide.
4. Incubate the microscope slide for 5 min in the dark.
5. Observe under a confocal microscope (*see Note 11*).

3.3 Molecular Analyses

In order to avoid contaminations, carry out all steps under a biological hood (unless indicated otherwise). Prior to use, clean all the instruments (*i.e.*, micropipettes, tube racks, centrifuge, etc.). Use sterile filter tips.

3.3.1 DNA Extraction

Rapid DNA Extraction

1. Prepare a fresh aliquot of extraction buffer with 10 \times PCR buffer:ultrapure water (1:1 dilution).
2. Place 1–10 surface-sterilized spores in 1.5 mL tube.
3. Crush the spore in a volume of 30 μL (single spore), 50 μL (pool of five spores), or 70 μL (pool of ten spores) of freshly prepared extraction buffer.
4. Incubate crashed spores at 95 °C for 15 min.
5. Centrifuge the crude extract at 16,000 $\times g$ for 10 min.
6. Collect the supernatant and store at –20 °C.

CTAB-Based DNA Extraction

The procedure should be performed under a fume hood.

1. Pre-warm water bath at 65 °C.
2. Prior to starting extraction, add 1 % (w/v) PVP to the 2 \times CTAB. Prepare 700 μL per sample of extraction buffer and pre-warm to dissolve PVP at 65 °C.
3. Place 1–10 surface-sterilized spores or add up to 100 μL of bacterial suspension (*see Subheading 3.5*) in a 1.5 mL tube.
4. Add 600 μL of extraction buffer. Crush the spore(s) (if extracting from spores) with a plastic pestle.
5. Incubate in the water bath at 65 °C for 1 h.

6. Centrifuge at $9,500 \times g$ for 10 min.
7. Transfer the supernatant to a new 2 mL tube without disturbing the pellet (if present).
8. Add 1 volume of chloroform:isoamyl alcohol (24:1). Mix by inverting the tube until the solution becomes lactescent and homogeneous.
9. Centrifuge at $6,000 \times g$ for 10 min.
10. Transfer the supernatant to a new 2 mL tube without disturbing the interface.
11. Add 1 volume of chloroform. Mix by inverting the tube 10–15 times.
12. Centrifuge at $6,000 \times g$ for 10 min.
13. Transfer the supernatant to a new 1.5 mL tube without disturbing the interface.
14. Add $2/3$ of the volume of cold 2-propanol. Mix by inverting the tube 10–15 times.
15. Incubate the tube at $-80\text{ }^{\circ}\text{C}$ for 5–10 min.
16. Centrifuge at $9,500 \times g$ for 10 min.
17. Carefully discard the supernatant to avoid dislodging the pellet.
18. Add 1 mL of cold 70 % ethanol and shake gently.
19. Centrifuge at $9,500 \times g$ for 1 min.
20. Remove the alcohol supernatant without disturbing the pellet. Air-dry the pellet.
21. Resuspend the pellet in 30–50 μL of ultrapure water.

3.3.2 PCR

1. Carry out individual PCR reactions in a final volume of 20 μL containing Phusion[®] DNA Polymerase, Phusion[®] HF Buffer, 375 μM of each dNTP, 750 nM of each primer (*see* Table 1), 1–4 μL (from rapid DNA extraction) or 40–50 ng (from CTAB extraction) DNA template. Bring the mix to the final volume with ultrapure water.
2. Use the following cycling conditions:
 - (a) Initial step of $98\text{ }^{\circ}\text{C}$ for 4 min
 - (b) Cycles (cycle conditions vary according to the primer pair used):
 - CaGgADf-CaGgADr: 35 cycles at $98\text{ }^{\circ}\text{C}$ for 13 s, $69\text{ }^{\circ}\text{C}$ for 30 s, $72\text{ }^{\circ}\text{C}$ for 55 s
 - GlomGIGf-GlomGIGr: 30 cycles at $98\text{ }^{\circ}\text{C}$ for 10 s, $58\text{ }^{\circ}\text{C}$ for 25 s, $72\text{ }^{\circ}\text{C}$ for 30 s

- 109F-1184R: 30 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 45 s (*see Note 12*)
- AML1-AML2: 35 cycles at 98 °C for 10 s, 58 °C for 30 s, 72 °C for 35 s
- ITS1f-ITS4: 35 cycles at 98 °C for 10 s, 57 °C for 30 s, 72 °C for 30 s
- Efg1g1f-Efg1g2r: 30 cycles at 98 °C for 10 s, 60 °C for 25 s, 72 °C for 20 s
- rpoBf-rpoBr: 35 cycles at 98 °C for 10 s, 60 °C for 30 s, 72 °C for 35 s

(c) Final extension step of 72 °C for 7 min

3. Purify PCR products directly from an amplification reaction or extract DNA fragments from agarose gel by using Wizard® SV Gel and PCR Clean-Up System following the manufacturer's instruction.
4. Clone purified PCR products using the pGEM®-T Easy Vector System following the manufacturer's instruction (*see Note 13*).
5. Insert the cloned vector into One Shot® TOP10 Chemically Competent *E. coli* following the manufacturer's instruction.
6. Plate and grow transformed *E. coli* cells in the selective medium containing ampicillin.
7. Screen clones for insert length by PCR. Select positive clones.
8. Sequence the cloned inserts.

3.3.3 Real-Time qPCR

Real-time qPCR is widely used for cultivation-independent detection and quantification of microorganisms. The estimation of the starting target quantities based on the amplification threshold cycle in each sample allows microbes (or nuclei for multinucleate organisms) to be quantified when single-copy genes are considered. The present application of qPCR can be used to quantify the abundance of endobacteria in a fungal sample, to determine the bacterial–fungal ratio (number of endobacteria vs. number of fungal nuclei detected), and to relatively quantify different endobacterial populations when simultaneously present in a fungal sample (*see Note 14*). In order to avoid contaminations carry out all steps under a biological hood. Use sterile filter tips.

1. Obtain plasmids containing the target DNA sequences. Quantify plasmids with the Qubit® 2.0 fluorometer and estimate the copy number/μL based upon the molecular weight of the template. Generate serial plasmid dilutions so that a 10⁶ to 10¹ plasmid copies are present in 1 μL of sample solution (*see Note 15*).

2. Serially dilute by tenfold the fungal sample(s) to be tested in qPCR (*see Note 16*).
3. Carry out individual real-time qPCR reactions in a final volume of 20 μL containing 2 \times real-time mix, 150 nM of each primer (*see Table 1*), and 1 μL of appropriate DNA dilution. Bring the mix to the final volume with ultrapure water. Prepare three technical replicates for each sample.
4. Use the following cycling conditions: initial step of 95 $^{\circ}\text{C}$ for 3 min, 40 cycles of 95 $^{\circ}\text{C}$ for 15 s followed by 60 $^{\circ}\text{C}$ for 40 s, with fluorescence measurement during the 60 $^{\circ}\text{C}$ step (*see Note 17*). At the end of the amplification add a melting curve analysis as follows: 55–95 $^{\circ}\text{C}$ with a heating rate of 0.5 $^{\circ}\text{C}$ per 10 s, with continuous fluorescence measurement (*see Note 18*).
5. As a first qPCR run, perform a standard curve using serial dilutions of plasmids to calculate the PCR efficiency. Do the same with the serially diluted fungal sample(s) to check that PCR efficiency is comparable to that obtained from plasmid standards (*see Note 19*). Those fungal sample dilutions falling into the standard curve dynamic range can be selected for the quantification of the endobacteria/fungal nuclei (*see Note 20*).
6. Perform the quantification assay. Set up the qPCR plate so that both fungal sample(s) and serial dilutions of plasmids are amplified in the same run. The number of target DNA sequences present in each PCR mixture is calculated by comparing the crossing points of the sample TCs with those of the standard plasmids using the StepOne™ software. If diluted fungal sample(s) were used for the quantification run, multiply the figure obtained by the dilution factor to retrieve the actual number of target DNA sequences present in the starting sample.

3.4 Fluorescent In Situ Hybridization (FISH)

1. Prior to hybridization, fix surface-sterilized spores in fresh and cold 4 % paraformaldehyde buffered with PBS. Incubate the spores at 4 $^{\circ}\text{C}$ for 3–6 h or, alternatively, at room temperature for 1–2 h.
2. Remove the fixative and wash the spores three times in 1 \times PBS. Process the spores immediately for the next hybridization steps, or suspend them in 1:1 1 \times PBS/100% ethanol and store at –20 $^{\circ}\text{C}$ until use.
3. Prepare a 2 % agarose solution in ultrapure water (w/v) in a small autoclaved flask.
4. Transfer fixed surface-sterilized spores (*see Note 10*) on an 8-well microscope slide and immobilize them with a 20–30 μL drop of 2 % agarose (one to three spores per well).

5. Dehydrate immobilized spores: plunge the entire microscope slide in an ethanol series (use Coplin jars): 3 min each, first in 50 % ethanol, then 75 % and 100 %. Let the ethanol evaporate but avoiding desiccation of the agarose drop.
6. Crush the spores to allow the penetration of the probes into the cytoplasm during the hybridization: crush the spores by adding a cover slide on the spore-embedded agarose drop, and pressing slightly. Gently remove the cover slide.
7. Carry out a pre-hybridization treatment with proteinase K for 10 min: add 50–70 μL of proteinase K on each well (agarose drop) (*see Note 21*).
8. Remove proteinase K and carry out the following steps: rinse with 1 \times PBS for 5 min; wash with 1 % Tween20 in 1 \times PBS (freshly prepared) for 5 min; rinse twice with 1 \times PBS for 5 min.
9. Pre-warm the hybridization buffer to 46 $^{\circ}\text{C}$.
 - (a) Carry out the next steps in the dark and avoid exposing probes to the light.
10. During the pre-hybridization steps freshly prepare the hybridization buffer (at 35 % formamide stringency) in a 2 mL tube, one tube per microscope slide, as follows:
 - (a) 700 μL of 100 % formamide (final concentration 35 %) (*see Note 3*)
 - (b) 200 μL of 20 \times SSC (final concentration 2 \times)
 - (c) 100 μL of Denhardt's solution (final concentration 1.25 \times)
 - (d) 1000 μL of ultrapure water (according to the volume of formamide)
 - (e) Store on ice.
11. Deposit a 60 μL drop of the hybridization buffer on each well. Add 3 μL of each probe (working concentration of 50–70 ng/ μl) (*see Table 2*) directly on the surface of each drop. Avoid using more than three probes (labeled with different fluorochromes) at the same time (*see Note 22*). Gently mix with a pipet tip without disturbing the agarose drop.
12. Prepare the humid chamber to prevent probe and buffer evaporation during the hybridization: fold a paper towel and place it into a 50 mL tube. Pour the towel with the remaining hybridization buffer.
13. Place the microscope slide horizontally inside the 50 mL tube, over the moist towel, and close it tightly (the humid chamber must not dry out). Incubate in the hybridization oven at 46 $^{\circ}\text{C}$ for 1 h and 30 min.
14. After hybridization, remove the hybridization buffer and rinse the samples twice with 2 \times SSC for 10 min and once with 0.1 \times SSC for 10 min. Let the microscope slide dry vertically.

15. Mount the microscope slide with a 20–30 μL drop of DABCO per well. Add a large cover slide to cover all wells (*see Note 23*). Remove excess DABCO.
16. Observe the microscope slide under a confocal microscope (*see Note 11*) (*see Fig. 1d*). Store the microscope slide at $-20\text{ }^{\circ}\text{C}$ in the dark for several months.

3.5 Bacterial Enrichment for Genome Sequencing

1. Distribute 1000–1200 surface-sterilized spores in 1.5 mL tubes (about 100 spores per tube) (*see Note 24*).
2. Resuspend the spores in 400 μL of 0.9 % NaCl.
3. Crush the spores using a plastic pestle until the spore walls are well smashed, and the 0.9 % NaCl solution becomes opaque and with a slightly pasty consistency.
4. Bring final volume to 1 mL, adding 600 μL of 0.9 % NaCl.
5. Using flame-sterilized forceps, place the 3 μm filter in the filter holder, and connect the syringe to the filter holder. Use a new 1.5 mL tube to collect the filtrate.
6. Transfer the suspension into the syringe and filter the crushed spores. Push thoroughly twice in order to allow the bacteria to pass through the filter pores. Remove the filter and place a new one in the filter holder. Repeat the passage with the other 100-spore batches. Maintain the tubes separated.
7. Centrifuge at $9,500\times g$ for 10 min.
8. Gently remove the supernatant to avoid losing the pellet.
9. Resuspend the pellet in 30 μL of 0.9 % NaCl.
10. Stain with SYTO 9[®] 10 μL of filtered bacterial suspension and observe under a confocal microscope (*see Subheading 3.2.2*).
11. Treat bacterial suspension with RQ1 RNase-Free DNase A according to the manufacturer's instruction: incubation at $37\text{ }^{\circ}\text{C}$ for 30 min followed by 10 min at $65\text{ }^{\circ}\text{C}$ to inactivate the enzyme. The 100-spore batch tubes are still separated. Store at $-20\text{ }^{\circ}\text{C}$.
12. Pool the bacterial suspension and extract genomic DNA with a CTAB-based method (*see Subheading 3.3.1*).
13. Check DNA extraction for fungal contamination using specific primers for AMF and bacteria (*CaGg* and/or *Mre*) (*see Subheading 3.3.2*). Fungal primers should not provide any PCR amplification (*see Fig. 2*).
14. Check quantity and quality of extracted genomic DNA.
15. Sequencing.

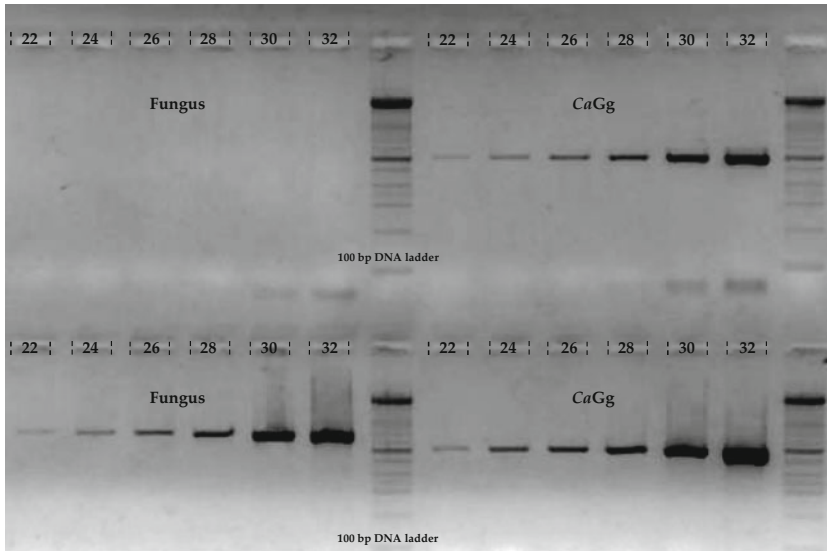


Fig. 2 Agarose gel electrophoresis patterns of the PCR amplification targeting the 18S rRNA gene of *G. margarita* (left) and the 16S rRNA gene of *CaGg* (right). The fungal and bacterial detection was carried out using the primer pairs AML1-AML2 [23] and GlomGIGf-GlomGIGr [13], respectively (see Subheading 3.3.2). After a rapid DNA extraction (see Subheading 3.3.1), the samples in the upper part of the gel were subjected to bacterial enrichment prior to amplification (see Subheading 3.5), whereas the ones in the lower part of the gel were directly amplified. Six subsamples for each assay were prepared. Each subsample was amplified with a different number of cycles (22, 24, 26, 28, 30, 32, respectively). As expected, the greater was the cycle number, the higher was the amount of amplicons obtained. Interestingly, any fungal amplification was observed from the samples enriched in endobacteria, suggesting a limited or absent carryover of fungal nuclei. These results confirmed the efficiency of the bacterial enrichment procedure here described

4 Notes

1. *Trifolium repens* is used here as host plant for *G. margarita* propagation since it has a small size, suitable for climatic chamber cultivation, and provides a good yield in terms of AMF spores in a relatively short time. However, other host plants, such as *Sorghum bicolor* and *Allium porrum*, can be used with success.
2. *Cichorium intybus* T-DNA-transformed roots are used in this protocol, and they have been obtained as described by Fontaine and colleagues [20]. However, both *Daucus carota* and *Medicago truncatula* transformed roots can be used to set up in vitro cultures of AMF, and the protocol to obtain the ROCs of such plant species is described by Bécard and Fortin [21] and Boisson-Dernier and colleagues [22], respectively.
3. Avoid subjecting formamide to freeze-thaw cycles. After thawing an aliquot, it should be stored at 4 °C and used shortly. Adjust the concentration of the formamide depending on the stringency necessary for the used probes.

4. If the spores still look very dirty (*i.e.*, several residues are attached to the spore surface), a mild sonication can be added prior to performing the sterilization (for not more than 30–40 s).
5. Depending on the spore amount being surface-sterilized, the suitable tube should be chosen. As an example, use 1.5 mL tubes to sterilize individual batches of 100 spores each, and 50 mL tubes when groups of 1000 spores are treated together.
6. Depending on the seed amount being surface-sterilized, the suitable plastic tube should be chosen. As an example, use 2 mL tubes to sterilize individual batches of 20 seeds each, and 15 mL tubes when up to 100 seeds are treated together.
7. After 4 days of germination in the dark, check whether the germination occurred and, if so, place the Petri dishes in the light, otherwise wait 2–3 days more. The seedlings are ready to be transplanted when cotyledons become green. From that moment, they can be kept in the Petri dish prior to use for at most 1 week.
8. The sandwich, composed by two cellulose nitrate membrane containing the *L. japonicus* seedling and the *G. margarita* spores, should be handled with extreme care; it should be placed in the Magenta box so that approximately 2/3 of its height is embedded in the sand.
9. The mycorrhizal efficiency is fast reducing in subsequent ROC cycles for this specific AMF isolate. Furthermore, the population of *Candidatus Glomeribacter gigasporarum* is dramatically reduced in successive spore generations obtained with this cultivation method, and this effect is further amplified when single spore inocula are employed [7]. Thus, it is recommended not to perform more than one ROC cycle to monoxenically produce *G. margarita* BEG34, unless the aim of the experiment is to obtain a cured line of the fungus, which is devoid of endobacteria [7].
10. When transferring the spores on a microscope slide, remove with a micropipette or absorb with blotting paper the remaining liquid (*i.e.*, PBS, ethanol-PBS mixture, water).
11. FITC and SYTO 9[®] fluorescence is excited at 488 nm and imaged with an emission window at 500–540 nm. Cy3 fluorescence is excited at 546 nm and imaged at 550–600 nm. Cy5 fluorescence is excited at 633 nm and imaged at 640–700 nm.
12. If the primer pair 109F-1184R [5] does not succeed, use a semi-nested PCR approach. Carry out a first PCR with 109F and 1387R [28]: cycling conditions were the same mentioned for 109F-1184R but with 55 s of extension in the cycles. Then apply a semi-nested PCR using the reverse primer 1184R:

cycling conditions were the same mentioned for 109F-1184R but with 25 cycles. Use semi-nested approach for particularly difficult templates (*i.e.*, templates from scarce or poor quality starting material).

13. Certain DNA polymerases add a single adenine to the 3' ends of amplified DNA fragments. The pGEM[®]-T Easy linearized Vector contains a single 3' terminal thymidine at each end which binds to the A overhang added by DNA polymerase. However, the DNA polymerase used in this protocol, as the other DNA polymerases that have a proofreading function, produce greater than 95 % blunt-end fragments. Thus, PCR fragments generated with such proofreading enzymes should be tailed at 72 °C for 15 min with dATP (200 μM final concentration) prior to cloning into the pGEM[®]-T Easy Vector.
14. The primer pairs GlomGIGf-GIGrA [7, 13] and RpoBRTf-RpoBRTTr [14] were used to detect and quantify *CaGg* inside *G. margarita* BEG34, whereas the primer pair Efgig2f-Efgigr [14] was used to target the fungal host DNA. The same qPCR approach was used to relative quantify the two bacterial populations (*Mre* and *CaGg*) hosted inside *G. margarita* CM23, using the primer pairs CMsAD1f-CMsAD2r [9] and CaGgAD7f-CaGgAD6r [9] designed to specifically target the 16S rRNA gene of *Mre* and *CaGg*, respectively.
15. Since the endobacteria of AMF are considered unculturable microbes, like their fungal hosts, plasmids carrying the target DNA inserts were used for the construction of the standard curve for each target gene.
16. The maximum dilution to be used in qPCR assays depends on the concentration of the starting material. For the quantification to be reliable, the fungal samples to be quantified must generate threshold cycles (TCs) that fall in the dynamic range established with the standard curve generated by plasmid amplification (*see* **Notes 18–20**).
17. qPCR primers here described were designed so that their melting temperature is between 65 and 70 °C and the amplified fragment for each primer pair is comprised between 80 and 150 bp. If these parameters are not respected, change annealing temperature and time accordingly.
18. Melting curve analysis is necessary to assess the absence of primer-dimer formation and that only a target-specific amplification occurs.
19. The amplification efficiency (*E*) can be obtained from the slope of the generated standard curve using the following formula: $E = (10^{(-1/\text{slope})} - 1) \times 100$. This value should be between 90 % and 105 % for each primer pair and on each tested template (both whether coming from plasmid or from total DNA

extraction). Since the precision of microbial quantification using qPCR relies on the assumption that the unknown sample and standard solutions share a comparable PCR efficiency, this should be verified prior to performing qPCR quantification.

20. The qPCR output recorded for each sample is represented by threshold cycle (TC), which is the intersection between an amplification curve and the threshold line in the qPCR graph. The dynamic range represents the TC interval in which the linearity of the target quantity with the TCs has been verified for those specific reaction conditions, and within which the absence of an inhibition effect and the sensitivity of the amplification are assessed. Thus, the fact that the sample TCs fall in the dynamic range established with the standard curve assures the reliability of the quantification.
21. Prior to hybridization, prepare negative controls treating the samples with RNase A.
22. In addition to *CaGg*- and/or *Mre*-specific probes, use a non-specific probe, such as the universal bacterial probe EUB338 [26], as positive control. Use also a negative control probe which specifically targets other bacterial taxa, such as the *Buchnera*-specific probe ApisP2a [27]. As further negative control, use nonsense probes, such as the probe non-BLOsADf2, which is the reverse complement of the probe BLOsADf2 [15]. Nonsense probes have no known rRNA target, thus ensuring that nonspecific probe incorporation into the samples does not occur: they must not provide any fluorescent signal.
23. DABCO drops should spread when the cover slide is placed. If not, slightly press the cover slide down until DABCO is homogeneously distributed. Due to the thickness of the agarose drops, it could happen that the cover slide does not adhere well to the microscope slide. If so, to avoid DABCO and samples from drying out, use nail polish to seal the gap created by the agarose drop.
24. The number of spores to be used as starting material to carry out the bacterial enrichment can vary depending on the goal of the experiment (*i.e.*, bacterial genome sequencing, FISH on bacterial suspension, DNA extraction, etc.), the abundance of the endobacteria within the spores, and the size of the spores. This protocol describes the steps necessary to prepare the material for the genome/metagenome sequencing of the endobacteria from *Gigaspora margarita* (BEG34 and MR104) and *Racocetra verrucosa* (VA105B) isolates. 1000 (for BEG34)—1200 (for MR104 and VA105B) spores used as starting material allow the obtainment of about 1 μ g

of enriched bacterial DNA. *G. margarita* and *R. verrucosa* produce relatively big spores (mean spore Ø 321 and 308 µm, respectively), but different AMF isolates or species could require a lower/higher spore number to be used as starting material for a bacterial enrichment and following genome/metagenome sequencing.

Acknowledgements

The authors wish to thank Mara Novero and Maria Teresa Della Beffa for having provided details on fungal culture conditions. Research in PB laboratory has been funded by the University of Turin (Local project 60 %).

References

- Schüßler A, Schwarzott D, Walker C (2001) A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycol Res* 105:1413–1421
- Bonfante P, Genre A (2008) Plants and arbuscular mycorrhizal fungi: an evolutionary-developmental perspective. *Trends Plant Sci* 13:492–498
- Bonfante P, Anca IA (2009) Plants, mycorrhizal fungi, and bacteria: a network of interactions. *Annu Rev Microbiol* 63:363–383
- Bianciotto V, Lumini E, Bonfante P, Vandamme P (2003) ‘*Candidatus Glomeribacter gigasporarum*’ gen. nov., sp. nov., an endosymbiont of arbuscular mycorrhizal fungi. *Int J Syst Evol Microbiol* 53:121–124
- Naumann M, Schüßler A, Bonfante P (2010) The obligate endobacteria of arbuscular mycorrhizal fungi are ancient heritable components related to the Mollicutes. *ISME J* 4:862–871
- Ghignone S, Salvioli A, Anca I, Lumini E, Ortu G, Petiti L, Cruveiller S, Bianciotto V, Piffanelli P, Lanfranco L, Bonfante P (2012) The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. *ISME J* 6:136–145
- Lumini E, Bianciotto V, Jargeat P, Novero M, Salvioli A, Faccio A, Becard G, Bonfante P (2007) Presymbiotic growth and sporal morphology are affected in the arbuscular mycorrhizal fungus *Gigaspora margarita* cured of its endobacteria. *Cell Microbiol* 9:1716–1729
- Salvioli A, Ghignone S, Novero M, Navazio L, Venice F, Bagnaresi P, Bonfante P (2015) Symbiosis with an endobacterium increases the fitness of a mycorrhizal fungus, raising its bioenergetic potential. *ISME J*: doi: 10.1038/ismej.2015.91
- Desirò A, Salvioli A, Ngonkeu EL, Mondo SJ, Epis S, Faccio A, Kaech A, Pawlowska TE, Bonfante P (2014) Detection of a novel intracellular microbiome hosted in arbuscular mycorrhizal fungi. *ISME J* 8:257–270
- Torres-Cortés G, Ghignone S, Bonfante P, Schüßler A (2015). Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: Transkingdom gene transfer in an ancient mycoplasma-fungus association. *Proc Natl Acad Sci USA* 112:7785–7790
- Naito M, Morton JB, Pawlowska TE (2015) Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. *Proc Natl Acad Sci USA* 112:7791–7796
- Engel P, Moran NA (2013) The gut microbiota of insects—diversity in structure and function. *FEMS Microbiol Rev* 37:699–735
- Bianciotto V, Genre A, Jargeat P, Lumini E, Becard G, Bonfante P (2004) Vertical transmission of endobacteria in the arbuscular mycorrhizal fungus *Gigaspora margarita* through generation of vegetative spores. *Appl Environ Microbiol* 70:3600–3608
- Salvioli A, Lumini E, Anca IA, Bianciotto V, Bonfante P (2008) Simultaneous detection and quantification of the unculturable microbe *Candidatus Glomeribacter gigasporarum* inside its fungal host *Gigaspora margarita*. *New Phytol* 180:248–257
- Desirò A, Naumann M, Epis S, Novero M, Bandi C, Genre A, Bonfante P (2013)

- Mollicutes-related endobacteria thrive inside liverwort-associated arbuscular mycorrhizal fungi. *Environ Microbiol* 15:822–836
16. Partida-Martinez LP, Hertweck C (2005) Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature* 437:884–888
 17. Sato Y, Narisawa K, Tsuruta K, Umezu M, Nishizawa T, Tanaka K, Yamaguchi K, Komatsuzaki M, Ohta H (2010) Detection of betaproteobacteria inside the mycelium of the fungus *Mortierella elongata*. *Microbes Environ* 25:321–324
 18. Desirò A, Faccio A, Kaech A, Bidartondo MI, Bonfante P (2015) *Endogone*, one of the oldest plant-associated fungi, host unique Mollicutes-related endobacteria. *New Phytol* 205:1464–1472
 19. Hewitt EJ (1966) Sand and water culture methods used in the study of plant nutrition, 2nd edn. Commonwealth Agricultural Bureau: The Eastern Press, London
 20. Fontaine J, Grandgougin-Ferjani A, Glorian V, Durand R (2004) 24-Methyl/methylene sterols increase in monoxenic roots after colonization by arbuscular mycorrhizal fungi. *New Phytol* 163:159–167
 21. Bécard G, Fortin JA (1988) Early events of vesicular-arbuscular mycorrhiza formation on Ri T-DNA transformed roots. *New Phytol* 108:211–218
 22. Boisson-Dernier A, Chabaud M, Garcia F, Bécard G, Rosenberg C, Barker DG (2001) *Agrobacterium rhizogenes*-transformed roots of *Medicago truncatula* for the study of nitrogen-fixing and endomycorrhizal symbiotic associations. *Mol Plant Microbe Interact* 14:695–700
 23. Lee J, Lee S, Young JPW (2008) Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiol Ecol* 65:339–349
 24. Gardes M, Bruns TD (1993) ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Mol Ecol* 2:113–11
 25. White TJ, Bruns T, Lee S, Taylor JW (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols. Academic Press, San Diego, pp. 315–322
 26. Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* 56:1919–1925
 27. Koga R, Tsuchida T, Fukatsu T (2003) Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. *Proc R Soc B* 270:2543–2550
 28. Marchesi JR, Sato T, Weightman AJ, Martin TA, Fry JC, Hiom SJ, Wade WG (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl Environ Microbiol* 64:2333

GenoSol Platform: A Logistic and Technical Platform for Conserving and Exploring Soil Microbial Diversity

Samuel Dequiedt, Pierre-Alain Maron, and Lionel Ranjard

Abstract

In 2008, the platform “GenoSol” (http://www.dijon.inra.fr/plateforme_genosol) was created at the INRA (French National Institute for Agronomic Research) of Dijon. This platform was launched by several soil microbial ecologist senior scientists to provide a logistics and technical structure dedicated to the acquisition, conservation, characterization, and supply of genetic resources (DNA) of soils from very large-scale samplings (several hundred to several thousand corresponding to large spatial and/or temporal scales). Thanks to this structure metagenomic analysis of soil microbial communities has been standardized as well as a reliable reference system for analysis of the microbial genetic resources of the collected soils (more than 10,000 soil samples to date). This platform also illustrates the usefulness of existing soil archives in providing a readily available source of ecological information that is relevant to microbial ecology, probably more than we can currently fathom.

Key words Soil, Microorganisms, Biodiversity, Molecular tools, Soil conservatory

1 Introduction

Soils are the principal reservoirs of microbial diversity and represent a core component of terrestrial ecosystems. There is an increasing demand for assessing the impact of human activities on this environmental matrix, which provides at small and large scales various ecosystem services [1]. To address this demand, taxonomic and functional diversity of soil microbial communities and their stability over time need to be characterized for predicting soil quality upon agricultural and industrial activities, the evolution of this quality being expected to affect environment quality and public health. Recent methodological progresses have led to the development and automation of molecular biological tools (based on the extraction and characterization of nucleic acids), which can be applied, with moderate throughput, to characterize soil microbial genetic resources (taxonomic diversity and functional potential) [2]. These tools should now be applied systematically to large-scale

samples so as to extend their general usefulness and produce a reliable reference system for the characterization and interpretation of the soil microbial diversity.

In this context, the platform “GenoSol” (http://www.dijon.inra.fr/plateforme_genosol) was created in 2008 by several senior soil microbial ecologist from INRA (French National Institute for Agronomic Research). This initiative aimed to fill the gap in technical and logistical standardization of soil conservatory and molecular tools to assess soil microbial diversity on large-scale sampling. The aim of this platform is to provide an appropriate logistic structure for the acquisition, storage, and characterization of soil genetic resources obtained by extensive sampling (several hundred to several thousand soils), on very large space and/or time scales (network of national soil survey, long-term experimental sites, etc.), and to make these resources readily available for the whole scientific community and policy makers. The ultimate goal is to produce a reliable reference system based on molecular characterization (taxonomic and functional features) of the soil microbial communities that provide scientific interpretations of the analyses from large scales of time and space sampling. The platform also aims at building up and storing for long-term periods a library of soil genetic resources that is made available to national and international scientific communities.

2 GenoSol Facilities

2.1 Soil Conservatory

The platform GenoSol is the first “microbiological” soil conservatory designed to manage, store, and make available soil microbial genetic resources obtained from large-scale sampling such as soil survey or network of experimental sites. Once received at the platform, soils are freeze-dried at $-40\text{ }^{\circ}\text{C}$ and stored. The impact of processing on subsequent variability of the molecular analyses has been tested and optimized (S. Dequiedt, pers. comm.). A protocol producing high-quality nucleic acids, compatible with the molecular characterization tools, is systematically applied to extract, purify, and quantify the DNA from the stored soil samples and is currently undergoing normalization (Association Française de Normalization, International Standard Organization). The purified DNAs are stored at $-30\text{ }^{\circ}\text{C}$ (INRA quality control). An automated, computerized procedure ensures sample management and traceability. So far, more than 10,000 soils and corresponding DNAs have been referenced by the platform GenoSol. It is estimated that subsequently more than 1500 new soils will be processed each year. These soils come from French national soil survey and experimental sites in majority but also a non-negligible part from international collaborations in the five earth continents. Figure 1 presents

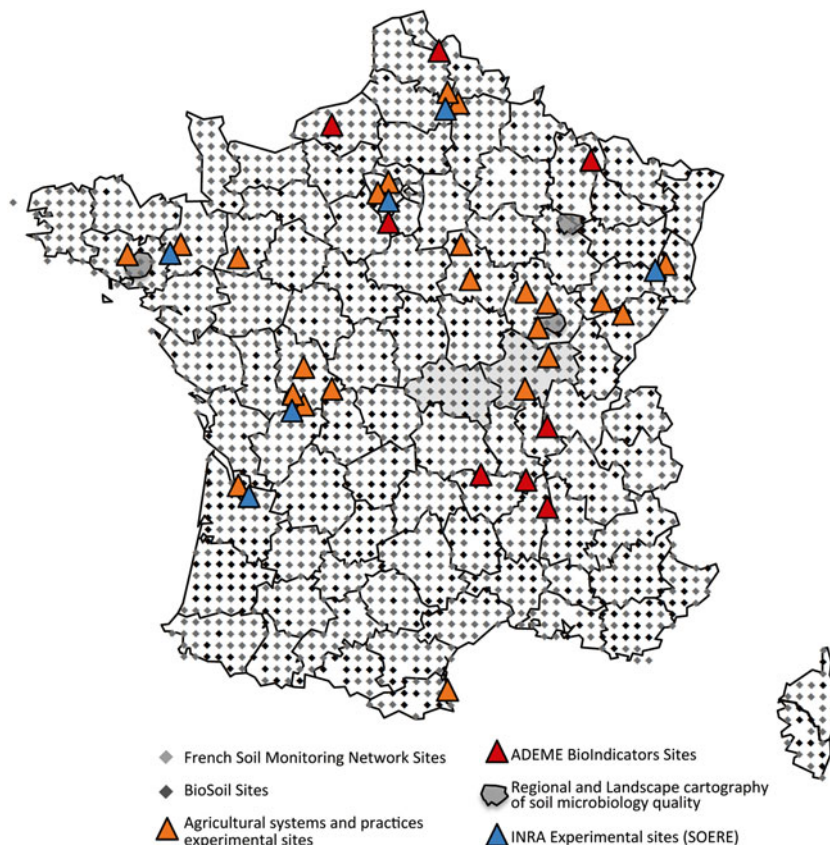


Fig. 1 Mapping of French soils stored in Genosol conservatory originating from the French Soil Monitoring Network (réseau de mesures de la qualité des sols: RMQS) and other experimental field sites managed by research or technical institutes

the distribution on the French territory of the 10,000 soils stored in genosol conservatory and originating from either the French Soil Monitoring Network (réseau de mesures de la qualité des sols: RMQS) or other experimental field sites managed by research or technical institute.

2.2 Molecular Tools Development

Microbial genetic resources of natural ecosystems are very difficult to characterize, which can be explained by the different degrees of accessibility of populations within a heterogeneous and structured matrix but also by the difficulty of resolving an information representing 100,000–1,000,000 different species per gram/mL of material. However, during the past 20 years, major advances in molecular biology have allowed the development of “molecular ecology approaches” to investigate the diversity of natural microbial communities in situ. In this context, the GenoSol platform will provide facilities and dedicated services to all researchers’ community (Fig. 2). GenoSol platform will develop metagenomics approach directly on the DNA extracted from environmental

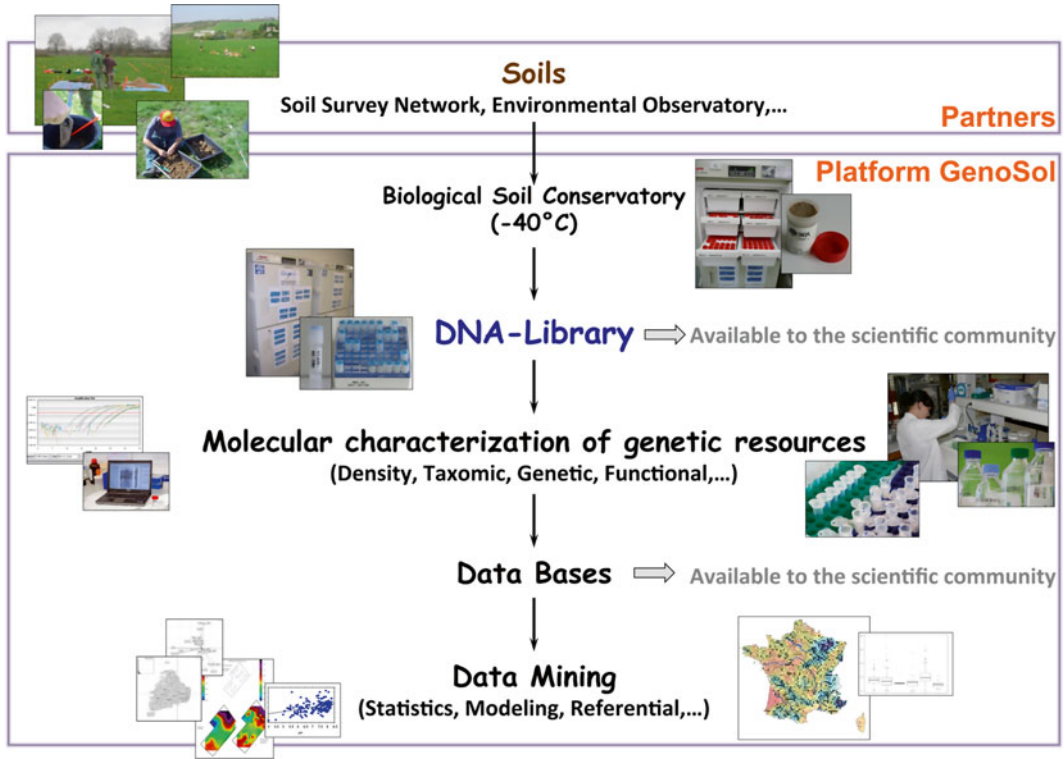


Fig. 2 Genosol activities

matrices to assess the density, taxonomic and functional diversity of indigenous microbial. More precisely, this technical platform is involved in the development and technological surveillance of methods for extracting nucleic acids from soils and tools for characterizing microbial genetic resources (genotyping, pyrosequencing, metagenomics, metaproteomics, measurement of activity, etc). This platform is involved in improving the standardization of procedures and molecular tools, with a view to normalization (AFNOR and ISO). A partnership with Genoscope was obtained in 2010 to develop high-throughput sequencing techniques for investigating soil microbial diversity.

Various significant technical outputs have been obtained in the platform. Below are presented some examples:

- Improvement of soil DNA extraction procedure to assess abundance and diversity of soil microbial community [3]
- Evaluation of the ISO standard soil DNA extraction procedure for assessing soil microbial abundance and diversity [4]
- Development of a robust procedure for applying pyrosequencing approach to assess soil microbial taxonomic inventory [3, 5]
- Development of a bioinformatics pipeline tool for analyzing pyrosequencing data for soil microbial taxonomic inventory [6]

- Optimization and validation of real-time PCR of bacterial and fungal communities in soils [7].

In addition, the development of robust strategies for long-term storage and archiving of soil samples from large systematic surveys and long-term field experiments and of molecular characterization of indigenous microbial communities allowed answering research questions unforeseen at the time of sampling. For example, GenoSol platform by making available the RMQs soil sampling for microbial ecologist allowed elucidating the turnover of bacterial diversity and the processes associated at the scale of France [8].

2.3 Reference System and Database

To develop a reliable reference system for analysis of the microbial genetic resources of the collected soils, the platform GenoSol has set up a database called “MicroSol” [9]. This database is designed to allow interactions with other databases managed by the GenoSol partners via computerized links. These external databases provide information on soil physicochemical characteristics, climatic data, plant cover, history of agricultural practices, and land use. This network of databases is only accessible to partners and users of the platform, and within the technical and legal frameworks laid down in the GenoSol charter. The platform GenoSol also promotes the selection and standardization of national and European procedures and tools for assessing soil quality.

In an agroecological context which requires the development of bioindicators for evaluating soil quality and the impact of agricultural practices, the MicroSol database developed is an operational tool to develop and promote microbial indicators, but also the associated standards essential for their interpretation. Indeed, this database contains the results of molecular analyses of the abundance and diversity of microbial communities acquired in a standardized methodological framework from samples covering the entire France territory. One of the main outputs is the development of statistical polynomial model allowing the prediction of optimum soil microbial biomass and biodiversity according to environmental parameters. This model is an innovative tool providing optimal value of microbial biomass for a given pedoclimatic condition, which must be compared with the corresponding measured data to allow a robust diagnostic of soil quality and of the impact of land use (Horrigue et al., in revision Ecological Indicators).

More recently, the Genosol platform was integrated in the AnaEE-France (Analysis and Experimentation on Ecosystems) research infrastructure services for experimental studies on soil biodiversity and associated ecological functions. AnaEE-France is the French node of a European research infrastructure dedicated to experimental research on continental ecosystems. It gathers a set of platforms selected for their originality, their quality, and their access to the scientific community. The infrastructure offers clear

access rules for a large set of services including in vitro, in natura experimental facilities, front hedge equipment, modeling platforms, and data bases. In particular, a wide panel of complementary services is relevant for soil ecology and biodiversity.

References

- Balvanera P, Pfisterer AB, Buchmann N, He JS, Nakashizuka T, Raffaelli D, Schmid B (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol Lett* 9:1146–1156
- Maron PA, Mougel C, Ranjard L (2011) Soil microbial diversity: spatial overview, driving factors and functional interest. *C R Biol* 334: 403–411
- Terrat S, Christen R, Dequiedt S, Lelievre M, Nowak V, Bachar D, Plassart P, Wincker P, Jolivet C, Bispo A, Lemanceau P, Maron PA, Mougel C, Ranjard L (2012) Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. *J Microbial Biotechnol* 5:135–141
- Plassart P, Terrat S, Griffiths R, Thomson B, Dequiedt S, Lelievre M, Regnier T, Nowak V, Bailey M, Lemanceau P, Bispo A, Chabbi A, Maron P-A, Mougel C, Ranjard L (2012) Evaluation of the ISO standard 11063 DNA extraction procedure for assessing soil microbial abundance and community structure. *PLoS One* 7, e44279
- Terrat S, Plassart P, Bourgeois E, Ferreira S, Dequiedt S, Adele-Dit-De-Renseville N, Lemanceau P, Bispo A, Chabbi A, Maron PA, Ranjard L (2015) Meta-barcoded evaluation of the ISO Standard 11063 DNA extraction procedure to characterize soil bacterial and fungal community diversity and composition. *Microb Biotechnol* 8:131–42
- Terrat S, Dequiedt S, Horigue W, Lelievre M, Cruaud C, Saby N, Jolivet C, Arrouays D, Maron PA, Ranjard L, Chemidlin Prévost-Bouré N (2015) Improving soil bacterial taxon-area relationships assessment using DNA meta-barcoding. *Heredity* 114(5):468–75
- Chemidlin Prévost-Bouré NC, Christen R, Dequiedt S, Mougel C, Lelievre M, Jolivet C, Ranjard L (2011) Validation and application of a PCR primer set to quantify fungal communities in the soil environment by real-time quantitative PCR. *PLoS One* 6(9), e24166
- Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, Thioulouse J, Saby NPA, Lelievre M, Maron PA, Morin FER, Bispo A, Jolivet C, Arrouays D, Lemanceau P (2013) Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nat Commun* 4:134. doi:10.1038/ncomms2431
- Morin FER, Dequiedt S, Koyao-Darinet V, Toutain B, Terrat S, Lelièvre M, Nowak V, Faivre-Primot C, Lemanceau P, Maron PA, Ranjard L (2013) MicroSol database, le Premier Système d'Information Environnemental sur la Microbiologie des Sols. *Etud Gest Sols* 20:27–38

Chapter 4

Sample Preparation for Fungal Community Analysis by High-Throughput Sequencing of Barcode Amplicons

Karina Engelbrecht Clemmensen, Katarina Ihrmark, Mikael Brandström Durling, and Björn D. Lindahl

Abstract

Fungal species participate in vast numbers of processes in the landscape around us. However, their often cryptic growth, inside various substrates and in highly diverse species assemblages, has been a major obstacle to thorough analysis of fungal communities, hampering exhaustive description of the fungal kingdom. Recent technological developments allowing rapid, high-throughput sequencing of mixed communities from many samples at once are currently having a tremendous impact in fungal community ecology. Universal DNA extraction followed by amplification and sequencing of fungal species-level barcodes such as the nuclear internal transcribed spacer (ITS) region now enable identification and relative quantification of fungal community members across well-replicated experimental settings.

Here, we present the sample preparation procedure presently used in our laboratory for fungal community analysis by high-throughput sequencing of amplified ITS2 markers. We focus on the procedure optimized for studies of total fungal communities in humus-rich soils, wood, and litter. However, this procedure can be applied to other sample types and markers. We focus on the laboratory-based part of sample preparation, that is, the procedure from the point where samples enter the laboratory until amplicons are submitted for sequencing. Our procedure comprises four main parts: (1) universal DNA extraction, (2) optimization of PCR conditions, (3) production of tagged ITS amplicons, and (4) preparation of the multiplexed amplicon mix to be sequenced. The presented procedure is independent of the specific high-throughput sequencing technology used, which makes it highly versatile.

Key words Meta-barcoding, High-throughput sequencing, ITS (internal transcribed spacer), DNA extraction, Multiplexing

1 Introduction

The biogeography and autecology of fungal species producing macroscopic sporocarps have long been studied within the disciplines of botany and vegetation ecology. Classical fungal taxonomy is also based primarily on sporocarp morphology, traditionally being close to the discipline of botany. The microscopic nature of the vegetative mycelium and the lack of sporocarps in many species, however, have rooted studies of fungal ecophysiology in the field

of microbiology, with a parallel taxonomy based on anamorphic stages. Furthermore, the often cryptic growth of fungi within various substrates—including other living organisms—and their presence in highly diverse species assemblages have been major obstacles to thorough analysis of fungal communities, hampering exhaustive description of the fungal kingdom. Recent technological developments allowing rapid, high-throughput sequencing of mixed communities simultaneously from many samples are currently having a tremendous impact in mycology, yielding new insights into ecological constraints on fungal niches (e.g. [1]), roles of fungal communities in ecosystem-level processes (e.g. [2]), and enabling the discovery and description of novel major fungal lineages (e.g. [3]).

Universal DNA extraction followed by amplification and sequencing of the fungal species-level barcode—for Dikarya primarily the nuclear internal transcribed spacer (ITS) region of the ribosomal RNA genes [4]—is now common practice when studying fungal communities in various substrates. A key to the applicability of large-scale sequencing techniques for community studies is the use of sample-tagged primers [5] to generate DNA amplicons from multiple samples that can subsequently be sequenced in one run at one of the high-throughput sequencing platforms. Based on the sample-specific tags, each DNA sequence can later be traced back to the original sample, and occurrences and relative abundances of barcode sequences can be analyzed, as a representation of community composition.

Particularly within the fungal kingdom, in which sexual reproduction is a widespread feature, species-level taxonomic resolution is useful and highly informative, similar to the situation in plants and animals. The species identity determines a certain set of traits that have been unified during adaptation and speciation. For example, traits related to both adaptations to the environment and resource acquisition should be unified within a species. Analysis of fungal species composition across various experimental and natural conditions followed by organization of available data in public databases, such as the UNITE database [6], enable us to link sequence data to complementary data on functional and ecological characteristics in a taxonomic setting. Furthermore, barcoding across many substrates and biomes is presently generating fungal sequence data that enable a continuous verification of new fungal species hypotheses and re-evaluation of total fungal diversity (<https://unite.ut.ee>).

Here, we present the sample preparation procedure (Fig. 1) presently used in our laboratory for fungal community analysis by high-throughput sequencing of amplified ITS2 markers. We focus on the procedure optimized for studies of total fungal communities in humus-rich soils, wood, and litter, although the procedure can be adjusted to other sample types and markers as well. We are not considering experimental design and field sampling, although these issues represent major challenges that in many

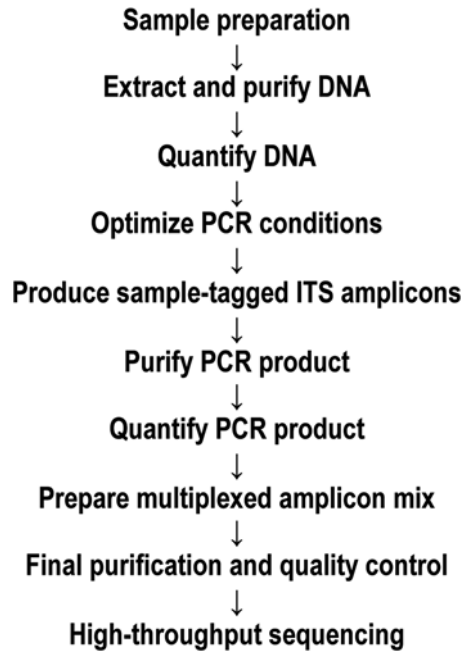


Fig. 1 Flow chart of fungal barcode amplification and multiplexing

cases may be the main restriction to which conclusions can be drawn from fungal community studies. Particularly, scale of interest, independent replication, pooling of samples, and subsampling are factors which should be carefully considered [7–9]. Once samples are collected, laboratory-based sample preparation consists of four main parts: (1) extracting pure DNA, (2) optimizing PCR conditions, (3) producing tagged ITS amplicons from each sample, and (4) preparing the multiplexed sample to be sequenced.

The development of new extraction protocols and kits is fast, but the fundamentals of DNA extraction remain the same: DNA must be purified from cellular and substrate material in a manner that prevents degradation and enables amplification. Optimal DNA extraction protocols, thus, vary depending on the organisms and substrates of interest. Although various “ready-to-use” kits are used in our lab, we here present our much-used CTAB-based protocol, as it is cheap, flexible toward different sample types and extraction volumes, albeit labor-intensive in comparison to many commercial kits. Unlike animals, both bacteria and fungi have sturdy cell walls, and initial freeze-drying and grinding acts to homogenize samples and breaks down cell wall material, while harmful cellular enzymes remain inactivated. We further macerate subsamples in a beadbeater—first in dry condition, then suspended in extraction buffer. By combining dry and wet bead beating with heating of the samples, DNA extraction from organisms with thick

cell walls or within tight aggregates is optimized. Insoluble particles are removed through centrifugation, while soluble proteins and other organic substances are removed through extraction with chloroform and centrifugation. Thereafter, DNA is precipitated with 2-propanol from the aqueous phase and washed thoroughly with ethanol to remove contaminating salts. Finally, the purified DNA is resuspended and stored in water. This method has been shown to give sufficiently intact genomic DNA from various sample types for the PCR reaction to work, although further DNA purification may be needed for some sample types.

It is crucial that the extracted DNA is representative of the total DNA pool of the entire sample, which is usually much larger than the small amounts used in traditional DNA extraction protocols. In the presented protocol, the proportion of sample subjected to the different homogenization and extraction steps may be altered according to the size and complexity of the original samples. In particular, this is important when samples are pooled into fewer composite samples to, e.g. represent a larger geographical area, and extraction of larger subsamples may be required to properly reflect the combined diversity.

Although non-targeted approaches, based on sequencing of entire meta-genomes from mixed communities, are rapidly gaining in feasibility, targeted sequencing after PCR amplification of specific genetic markers in most cases remains more attractive [10]. With read lengths currently limited to a maximum of 400–600 bp as available for the 454-pyrosequencing (Roche), Illumina MiSeq and Ion torrent technologies, an important step in the development of our sample preparation protocol has been to design new fungal-specific primers that enable amplification of the ITS2 region only [11]. The fungal ITS2 region has equally good species resolution as the full ITS region [12], and the amplicons targeted by our ITS7–ITS4 primer combination rarely extend beyond 500 bp in length (Fig. 2), allowing sequencing throughout their entire length with the available technologies. The shorter ITS2 amplicons with

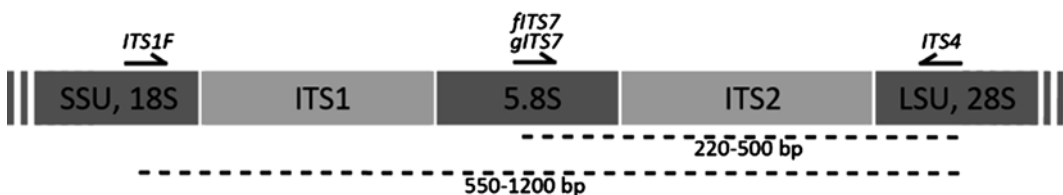


Fig. 2 Diagram of the fungal rDNA gene cluster. Genes encoding 18S, 5.8S, and 28S ribosomal RNA subunits (SSU=small subunit, LSU=large subunit) are separated by the internal transcribed spacer regions 1 (ITS1) and 2 (ITS2) that are useful as a species-level barcode for most fungi. Primer sites used for obtaining amplicons covering the majority of fungi are indicated by *single-headed arrows* above the diagram. The approximate size variation in amplicons of ITS2 (*upper broken line*) and the full ITS (*lower broken line*) are indicated below the diagram

much less size variation than full ITS amplicons have proved to give much better translation of DNA template relative abundances into sequence read relative abundances, probably due to fewer biases during both amplification and sequencing [11]. However, it should be mentioned that most primers have mismatches for some fungal groups (including the used primers), and that competition for primers in amplifications of complex communities may be strongly biased even by single mismatches between primer and template [11, 13].

In order for the composition of sequences in the amplified sample to resemble the template community as much as possible, it is pivotal to optimize both the template concentration and the number of PCR amplification cycles, preferably for each sample individually. The DNA template should be diluted enough to overcome any inhibition of the PCR reaction caused by too high concentrations of inhibitors in the extracts. However, the template should not be diluted more than necessary, as larger amounts of template decrease the impact of stochastic processes and give more predictable and less biased PCR amplification [14]. The number of PCR cycles should allow the reaction to reach (the middle of) the phase of exponential increase of product, but not to enter the saturated phase in which the community could be altered due to, e.g. primer or dNTP limitation. Generally, PCR biases due to both random drift and selection bias are minimized if the number of PCR cycles is reduced. Optimal dilutions and cycle numbers can be tested either with normal PCR or with quantitative real-time PCR (Fig. 3). Here, we describe the procedure based on normal PCR and visualization of PCR products by agarose gel electrophoresis. Once the PCR conditions are settled, final amplicons are produced with tagged primer combinations that are specific for each sample. We have designed [15] 104 primer pairs in which both primers are extended by a unique 10 bp tag (Fig. 4, Table 1). Tagging at both ends allows us to later filter out sequences with unexpected tag combinations caused by chimera formation or tag switching [16]. All tagged primer pairs were tested for amplification efficiency, using qPCR, and primer combinations that clearly deviated from the average were discarded.

Different sequencing platforms require different adaptors to be added to templates in a sample before sequencing. These adaptor sequences can, potentially, be included in the ITS primers, directly upstream the identification tag, which would give control of the sequencing direction (sequencing starts at one of the adaptors). However, we have chosen to prepare our amplicons by primers fused with only the identification tags. This precludes the use of excessively long fusion primers, which usually results in low amplification efficiency, higher risk of primer dimers (and multimers), and often requires complex nested PCR approaches, leading to a high potential for biases. The adaptors are instead added to our

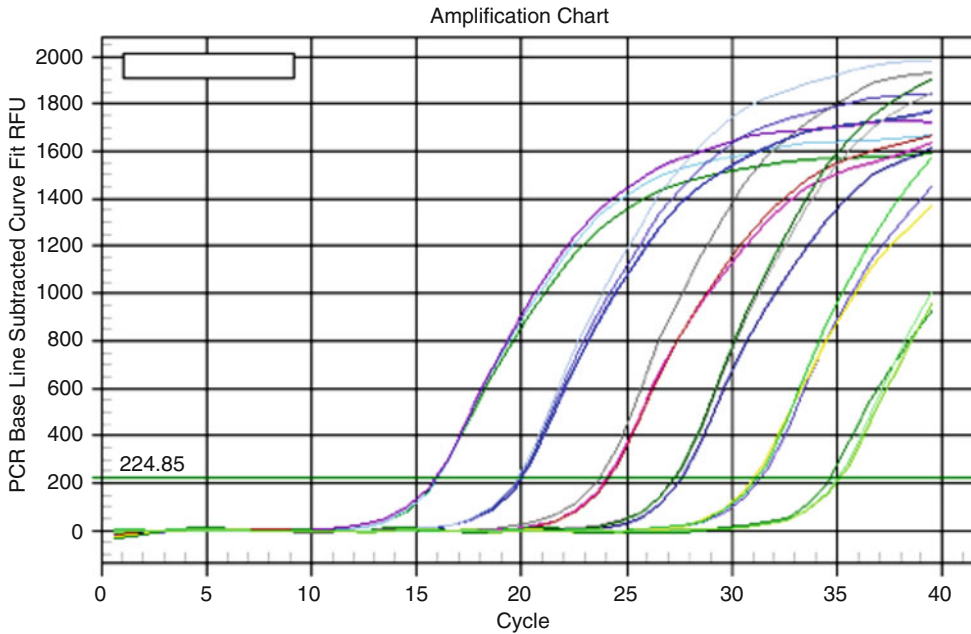


Fig. 3 Standards with different ITS copy numbers run in a quantitative real-time PCR (qPCR). Colored lines show the increase in PCR product for each cycle (x-axis) for a series of triplicate standard samples differing in ITS copy concentrations by a factor 10. The optimal number of cycles for producing amplicons for sequencing can be determined by qPCR. The total number of ITS copies in a sample can also be determined by relating the sample to a standard curve like this

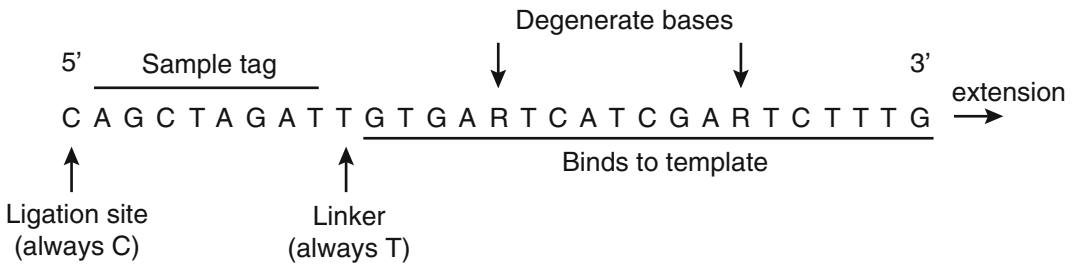


Fig. 4 Example of a sample tagged gITS7 primer [10]. The 19 bp long part that is complementary to the template contains two degenerated positions (R = equal mix between G and A) that increases the generality of the primer to cover a large fraction of all fungi (and also some other eukaryotes). The linker ensures that the binding part is restricted to 19 bp for all templates. We have developed a set of 104 primers that all differ in their sample specific tags on at least three positions. The conserved C at the 5' end facilitate non-biased ligation of sequencing adaptors

amplicons in random orientation by ligation according to the demands of the sequencing platform in question (adapter ligation is regularly performed by commercial sequencing facilities). However, since all our amplicons are short enough to be sequenced all through, obtained reads are easily reversed bioinformatically later. Addition of adaptors by ligation also means that our

Table 1
Sample identification tag sequences used in combination on ITS7 and ITS4 primers to be able to filter out sequences with unexpected tag combinations from the final data. Sequences include the ligation and linker sites as specified in Fig. 4

tag	ITS7	ITS4	tag	ITS7	ITS4	tag	ITS7	ITS4
tag_1	CACACGATCT	CACACGGTGT	tag_40	CCACAGTAIT	CCATCTACAT	tag_79	CTACTGATCT	CTAGCATGAT
tag_4	CACATAGTCT	CACATGTCTG	tag_41	CCACGTCACT	CCGACTGTCT	tag_80	CTAGAGCACT	CTAGTGATCT
tag_5	CACATGACIT	CACGCAGCAT	tag_42	CCACTATCCG	CCGATACTGT	tag_81	CTAGCTATCT	CTATACAGCT
tag_6	CACGATCAGT	CAC TAGCGGT	tag_44	CCAGATACTT	CCGCTATACT	tag_82	CTAGTCATGT	CTATAGCTCT
tag_7	CACGTGCTCT	CACTATGCAT	tag_45	CCAGCGTAGT	CCGTACATCT	tag_83	CTATAGCTGT	CTATATGCGT
tag_8	CACTATAGCT	CACTCACACT	tag_46	CCAGTAIGTT	CCGTAGCTGT	tag_84	CTATATCGCT	CTATCTCTGT
tag_9	CAC TATGTGT	CAC TCTGAGT	tag_47	CCATGTAGTT	CCGTACACAGT	tag_85	CTATCACTCT	CTATCTGACT
tag_10	CAC TCAGAGT	CAC TGATCAT	tag_48	CCGACTGATT	CCGTTCATGAT	tag_86	CTATCTCAGT	CTATGACACT
tag_11	CAC TCTCACT	CAC TGTATGT	tag_49	CCGAGCACTT	CCGTGTGTGT	tag_87	CTATCTGCTT	CTATGAGTGT
tag_12	CAC TGCTACT	CAG ACATAGT	tag_50	CCGAGTGTCT	CCTAGATACT	tag_88	CTATGCTAGT	CTATGGCCAT
tag_13	CAG ACAGTGT	CAG ACTATGT	tag_51	CCGATAGACT	CCTAGTCTGT	tag_89	CTATGTGTGT	CTCAGCACGT
tag_14	CAG ACATCTT	CAG AGCTCAT	tag_52	CCGCAITCGTT	CCTATCTCAT	tag_90	CTCACTTAGCT	CTCAGTATCT
tag_15	CAG AGACGGT	CAG ATACACT	tag_53	CCGTAGCAIT	CCTCGACTCT	tag_91	CTCACTCAIT	CTCATACTCT
tag_16	CAG AGCTCGT	CAG ATGCTAT	tag_54	CCGTATATGT	CCTCGAGCAT	tag_92	CTCATAGAGT	CTCATAGAGT
tag_17	CAG AGTATGT	CAG CACGACT	tag_55	CCGTGTCTCAGT	CCTCTCTGCT	tag_93	CTCATCAGTT	CTCATCGCAT
tag_18	CAG ATACAGT	CAG CACTCTG	tag_56	CCTACATCCG	CCTGACAGAT	tag_94	CTCATCTCGT	CTCATGCGAT
tag_20	CAG CACTAIT	CAG CTGACGT	tag_57	CCTACGGCTCT	CCTGCTCTAT	tag_95	CTCGCACTGT	CTCATGTACT
tag_21	CAG CGATACT	CAG TATCTCT	tag_58	CCTFAGACTGT	CCTGTACACT	tag_96	CTCGCAGAIT	CTCGATGAGT
tag_22	CAG CTAGAIT	CAG TCAGTCT	tag_60	CCTCGAGTCT	CGACTCTCAT	tag_97	CTCTAGATCT	CTCGCAGACT

(continued)

Table 1
(continued)

tag	ITS7	ITS4	tag	ITS7	ITS4	tag	ITS7	ITS4
tag_23	CAGCTCACC	CAGTCTGCT	tag_61	CCTCTCTGTT	CGATAGTCGT	tag_98	CTCTATGACT	CTCGTATGCT
tag_24	CAGTATCTCT	CAGTGAGCGT	tag_63	CCTGTATAGT	CGCTAGACAT	tag_99	CTCTCTATGT	CTCTACATGT
tag_25	CAGTGCAGTT	CAGTGTGAGT	tag_65	CGACTAGCCT	CGTAGATCAT	tag_100	CTCTGACACT	CTCTAGCAGT
tag_26	CAGTGTGATT	CATACAGTGT	tag_66	CGAGTAGAGT	CGTAGTCCGT	tag_101	CTCTGCCGATT	CTCTATGTCT
tag_28	CATAGTCTCT	CATAGCGAGT	tag_67	CGCGTCTATT	CGTACACACAT	tag_104	CTGAGAGATT	CTGACTGAGT
tag_29	CATAGTGAGT	CATCACTGAT	tag_68	CGCTATACGT	CGTACACGTGT	tag_105	CTGAGCTGTT	CTGAGATCCGT
tag_30	CATATGTTCGT	CATCAGCTAT	tag_69	CGTACAGATT	CGTCCGCTACT	tag_106	CTGAGTCACT	CTGAGTCACT
tag_31	CATCGACAGT	CATCATACGT	tag_70	CGTACTCACT	CGTCTGAGAT	tag_107	CTGATACTCT	CTGATCTAGT
tag_32	CATCGCTCTT	CATCGTCGCT	tag_71	CGTAGCGTCT	CGTGACTCGT	tag_108	CTGATCCGTGT	CTGCAGTACT
tag_33	CATCTACGCT	CATGAGACAT	tag_72	CGTATCGCGT	CGTGTCCGACT	tag_109	CTGATCTACT	CTGCCGTAGCT
tag_34	CATCTCCGTGT	CATGTACTGT	tag_73	CGTCACTCTT	CGTGTGTCAT	tag_112	CTGCTATGTT	CTGCTCACAT
tag_35	CATCTCTACT	CATGTCACGT	tag_74	CGTCCGCTAGT	CTACATCACT	tag_113	CTGCTGACGT	CTGCTCATACT
tag_36	CATGATACGT	CCACATGTAT	tag_75	CGTCCGTGTGT	CTACGATGCT	tag_114	CTGTCCACATT	CTGTCTATCT
tag_37	CATGCGTCTT	CCAGAGTACT	tag_76	CGTGTAGTCT	CTACTCATGT	tag_116	CTGTCTGTCT	CTGTGCCAGAT
tag_38	CATGTACTGT	CCAGTCCGAGT	tag_77	CGTGTGTACT	CTACTGCCAGT	tag_117	CTGTGATCTT	CTGTGCCGTCT
tag_39	CATGTCGCTT	CCAGTGTGAT	tag_78	CTACATGAGT	CTAGCACACAGT			

procedure is independent of sequencing platform. Current sequencing platforms give excessive numbers of reads for most community studies, and our use of tag-encoded primers integrated in our laboratory procedure gives the possibility to further multiplex samples (to get fewer reads per sample, but more samples in one run) using the multiplex identifier tags (MIDs) available from the sequencing platforms.

Initial evaluation suggests that both the Illumina and IonTorrent platforms are much more sensitive to variation in amplicon length than 454-sequencing, and preferably return sequences from shorter amplicons. These biases are probably the result of size fractionation during diffusion of DNA fragments toward the sequencing substrates (beads and flow cell for Ion Torrent and Illumina, respectively) as well as efficiency differences during emulsion PCR combined with relatively low detection sensitivity (Ion Torrent). Length biases may be overcome by fractionation of DNA pools based on amplicon lengths, followed by individual sequencing or MID tagging of the resulting pools. An alternative technology is the Pacific Bioscience (SMRT sequencing), which preliminarily showed much less length biases than the other platforms.

In-house, we have developed a publicly available bioinformatics pipeline adapted to fungal community studies based on ITS sequencing, SCATA (Sequence Clustering and Analysis of Tagged Amplicons, <http://scata.mykopat.slu.se>). This pipeline has been well-tested both by in-house projects and external users, and to date more than 2500 analyses have been run. No detailed accounts of the pipeline are given here, but settings optimized for fungal community studies can be found at the web page or in recent publications (e.g. [2]). In short, the bioinformatics procedure can be divided into three main parts: sequence quality filtering, sequence clustering into OTUs (operational taxonomic units), and OTU verification and identification. The initial quality filtering remove sequences that are incomplete (i.e. miss one or both primers) or are of low quality. Sequences are also scrutinized after reverse complementing. The sequences are then de-multiplexed based on the tag sequences, to recover the sample identities, which are kept in a metafile. Once this is done, amplicons are grouped by sequence similarity (i.e. clustered) into OTUs at the desired similarity threshold. For the ITS7-ITS4 primer combination, clustering is based on 41 bp of the conserved 5.8S region, about 105–330 bp of the ITS2 region and 38 bp of the LSU (Fig. 2). Homopolymers can be collapsed to, e.g. 2 bp before clustering (and raw reads kept in the metafile), which is desirable for sequencing technologies with high error rates in such regions. The number of sequences in each OTU will then represent the relative abundance of that OTU. Several program packages and web services are available for analysis of tagged sequence amplicons, e.g. QIIME ([17]; <http://qiime>).

org/) and MOTHUR ([18]; <http://www.mothur.org/>). However, most of these have initially been developed to handle amplicon data from hypervariable prokaryote 16S regions. This can be problematic to the analysis of fungal ITS sequences, since the evolutionary patterns of the bacterial 16S sequences and fungal ITS sequences are different and, thus, incur different assumptions to the algorithms. The most commonly used clustering method for bacterial data is full-linkage clustering with closed reference OTU picking (e.g. QIIME). However, given the unconstrained evolution of the ITS region and the lack of dense reference databases, other approaches are more appropriate for fungal ITS data [13]. In our pipeline, sequence similarity is established using USEARCH (<http://www.drive5.com/usearch/>) as a search engine, and sequences are assembled into OTUs by single-linkage clustering. The major advantage of this approach is that OTU delimitation is based on the “barcoding gap” rather than the intraspecific variation and, thus, harmonizes better with taxonomic species delimitation [4]. Clustering settings to approximate species level are validated by including known reference sequences in clustering runs at different similarity thresholds and scoring settings. Settings most closely reflecting species taxonomy in fungal clades with many well-known references are then implemented for all sequences. To approximate species-level OTUs across kingdom-wide fungal community studies, we have mostly used a 98.5 % sequence similarity threshold required for sequences to enter an OTU in the single-linkage clustering process. The entire UNITE database ([19]; <http://unite.ut.ee>) and a curated selection of sequences from the NCBI nr database (<https://blast.ncbi.nlm.nih.gov>) are optionally included in the clustering procedure, providing validation of clustering stringency in a taxonomic context and identification of OTUs based on the same criteria as the clustering.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents.

Avoid cross-contamination at all stages.

2.1 DNA Preparation Components

1. Freeze-drier.
2. Mortar and pestle, alternatively ball mill.
3. Screw cap tubes (2 mL) and microcentrifuge tubes (1.5 mL).
4. 2-mm-diameter glass beads.
5. Pipettes and filter tips for volumes of 1 mL and 200 μ L.

6. CTAB extraction buffer: 3 % CTAB, 0.15 M Tris-HCl, 2.6 M NaCl, 2 mM EDTA, pH 8. To prepare 50 mL CTAB buffer, dissolve 1.5 g CTAB (Hexadecyltrimethylammonium bromid) in 16.3 mL water by gently heating in microwave oven. Mix with 7.5 mL 1 M Tris-HCl (*see Note 1*), 26 mL 5 M NaCl (*see Note 2*), and 0.2 mL 0.5 M EDTA (*see Note 3*). Make a new bottle of 3 % CTAB buffer every day to obtain best extraction efficiency and to avoid cross-contamination.
7. Microwave oven.
8. Beadbeater machine for 1.5–2 mL tubes.
9. Heating block (65 °C) for 1.5–2 mL tubes.
10. Vortex.
11. Microcentrifuge for 1.5–2 mL tubes.
12. Fume hood.
13. Chloroform.
14. Isopropanol (2-propanol).
15. Ice.
16. Ethanol: 70 % dilution in water.
17. NanoDrop machine
Optional; to be used for purification with the Wizard DNA clean-up kit (Promega):
18. Wizard minicolumns, one per sample.
19. Wizard DNA clean-up resin.
20. Disposable 3 mL Luer-Lock syringes, one per sample (*see Note 4*).
21. Isopropanol (2-propanol): 80 % dilution in water.

2.2 Amplicon Preparation Components

1. PCR thermal cycler.
2. PCR-strips or plates.
3. Pre- and post-PCR pipettes and tips (1000 µL, 200 µL, and 10 µL filter tips).
4. Ice.
5. The *Taq* polymerase enzyme DreamTaq (stock at 5 units/µL) (*see Note 5*).
6. Reaction Buffer supplied with the *Taq* polymerase (stock at ×10 of required final concentration).
7. Nucleotides, dNTPs (stock at 2 mM of each of dATP, dCTP, dGTP, and dTTP in a mixture).
8. MgCl₂ (stock at 25 mM).
9. The forward and reverse ITS primers, gITS7 [11] and ITS4 [20], both extended at the 5'-end with a ten base pair identification

tag. In our laboratory, we have 104 uniquely tagged primer mixtures that are kept in mixtures (with unique tags on both primers) at stock concentration of 5 μ M (gITS7; CXXXXXXXXT-GTGARTCATCGAATCTTTG) and 3 μ M (ITS4; CXXXXXXXXT-TCCTCCGCTTATTGATATGC) (Fig. 4, Table 1) (*see Note 6*).

10. Centrifuge(s) for tubes and plates.
11. Gel tray system, combs and a submerged horizontal electrophoresis cell.
12. 500 mL glass bottle.
13. Agarose.
14. 1xSB buffer: 5 mM sodium tetraborate (stock solution 50 mM).
15. Microwave oven.
16. Nancy-520 dye (Sigma-Aldrich).
17. DNA size standard GeneRuler DNA ladder mix (ThermoScientific).
18. Gel documentation system with UV or Blue light.
19. AMPure (Beckman Coulter) magnetic bead solution.
20. AMPure magnetic plate.
21. 96-well PCR microplates with raised wells (chimney wells).
22. Ethanol: 70 % dilution in water.
23. Drying oven (37 °C).
24. Thin-walled, clear 0.5 mL PCR tubes.
25. Qubit instrument (Life Technologies).
26. Qubit dsDNA HS Assay Kit components A, B, C, D.
27. E.Z.N.A. Cycle-Pure Kit (Omega).
28. Timer.
29. Agilent 2100 Bioanalyzer (Agilent Tech) instrument.
30. Agilent chip priming station and vortex.
31. Agilent DNA 7500 Assay Kit, including gel-dye mix, marker, DNA ladder, chip.

3 Methods

Take precautions to avoid cross-contamination at all stages. Be careful to not use the same equipment and work surfaces for pre- and post-PCR work.

3.1 Sample Preparation

1. Freeze environmental samples at -20 °C as fast as possible to avoid unintended growth of opportunistic fungi after sampling (*see Note 7*).

2. Freeze-dry as large a sample volume as possible, typically 0.5–3 dL of soil/litter.
3. Grind the sample to a very fine powder manually by mortar and pestle or by using a ball mill for more sturdy material.

3.2 DNA Extraction and Purification

3.2.1 DNA Extraction Using the CTAB Method

1. Weigh 50–500 mg powdered substrate into 2 mL screw cap tubes; less material for organic samples, more for mineral soils (*see Note 8*). Note down the exact amount extracted for each sample (*see Note 9*). For every 23 samples, include an extraction blank (empty tube), which is treated as the samples in all following steps. Add three 2-mm-diameter glass beads to each tube (*see Note 10*) and close the lids.
2. Homogenize samples in the beadbeater machine set at low speed for 10 s.
3. Add 1000 μ L CTAB extraction buffer (*see Note 11*).
4. Run samples in the beadbeater machine once again. The CTAB will froth like detergent.
5. Incubate for 60 min at 65 °C in a heating block and vortex every 15 min.
6. Spin down particles in a table top centrifuge at 9600 $\times g$ for 5 min.
7. Transfer 500–800 μ L of the upper phase to a new, marked microcentrifuge tube using a pipette and filter tips (*see Note 12*).
8. Add 500–800 μ L chloroform (1 \times volume) and shake the samples vigorously by hand. All work with chloroform must be carried out in a fume hood (*see Note 13*).
9. Centrifuge the mixture at 9600 $\times g$ for 5 min.
10. Transfer the upper phase (typically 400–600 μ L, record amount) to a new marked 1.5 mL tube. Take care not to include any chloroform (lower phase) or interphase with the transferred phase.
11. Repeat the chloroform extraction **steps 8–10**.
12. Mix the supernatant with 600–900 μ L 2-propanol (1.5 \times volume) and leave on ice for 30 min (*see Note 14*). At this stage you may stop and store the samples at –20 °C.
13. Centrifuge at 16,500 $\times g$ for 10 min.
14. Discard the supernatant by gently decanting it into a beaker. The DNA should now be in the pellet.
15. Wash the pellet with 70 % ethanol (500 μ L). Centrifuge at 4100 $\times g$ for 5 min. Discard ethanol by decantation.
16. Optional: carefully remove remaining liquid with a pipette.
17. Let the pellet air dry by resting tubes upside down on a paper towel for 30 min.

18. Resuspend the pellet in 50 μL water. Gently tap tube to dissolve pellet. The DNA template may be stored in the tube at 4 $^{\circ}\text{C}$ for short term or at -20°C for longer term. To avoid contamination of the DNA template, subsample with care and as little as possible.

3.2.2 Optional:
*Purification with the Wizard
DNA Clean-Up System
(Promega)*

19. Use one Wizard minicolumn for each sample. Remove and set aside the plunger from a 3 mL disposable syringe. Attach the syringe barrel to the Luer-lock extension of each minicolumn (*see Note 4*).
20. Add 1 mL of Wizard DNA clean-up resin (*see Note 15*) to a 1.5 mL microcentrifuge tube. Add the DNA extract (*see Note 16*) to the clean-up resin and mix by gently inverting several times.
21. Pipet the Wizard DNA clean-up resin containing the bound DNA into the syringe barrel. Insert the syringe plunger slowly and gently push the slurry into the minicolumn with the syringe plunger. Discard the flow-through.
22. Detach the syringe from the minicolumn and remove the plunger from the syringe. Reattach the syringe barrel to the minicolumn. To wash the column, pipet 2 mL of 80 % 2-propanol into the syringe. Insert the plunger into the syringe and gently push the solution through the minicolumn. Discard the flow-through.
23. Remove the syringe barrel and transfer the minicolumn to a 1.5 mL microcentrifuge tube. Centrifuge the minicolumn for 2 min at $10,000\times g$ to dry the resin.
24. Transfer the minicolumn to a new microcentrifuge tube. Apply 50 μL of pre-warmed (65–70 $^{\circ}\text{C}$) water to the minicolumn and wait for 1 min. The DNA will remain intact on the minicolumn for up to 30 min. Centrifuge the minicolumn for 20 s at $10,000\times g$ to elute the bound DNA fragment.
25. Remove and discard the minicolumn. The purified DNA may be stored in the tube at 4 $^{\circ}\text{C}$ for short term or at -20°C for longer term. To avoid contamination of the DNA template, subsample with care and as little as possible.

**3.2.3 DNA Quantification
and Purity Check
by NanoDrop**

26. Blank measurement: Load your blank sample (1.5 μL water, same as template DNA eluate) onto the lower pedestal, close the sample chamber, and press “Blank” on the screen. Confirm that the blank has yielded a reproducible zero, by analyzing a blank as though it was a sample (*see Note 17*).
27. Clean the instrument between each measurement by wiping the sample from both the upper and lower pedestals using paper wipes.
28. Sample measurements: Load your sample (1.5 μL), and select “Measure” on the measurement screen. DNA concentration and purity can be assessed (*see Note 18*).

3.3 PCR Optimization

3.3.1 Test of Template Concentrations and Cycle Numbers

1. Choose DNA extracts representing different sample types, typically 4–6 DNA extracts at a time (*see Note 19*).
2. Dilute the samples to 5, 0.5, and 0.05 ng/μL with water based on the NanoDrop measurement (*see Note 20*).

$$\text{Dilution formula : } C_1 \times V_1 = C_2 \times V_2$$

C : concentration (ng/μL), V : volume (μL), 1: initial, 2: final

3. Prepare a master mix with sufficient material for one 50 μL PCR reaction per sample plus three–five extra samples to allow for pipetting losses (*see Note 21*). Run negative controls (blanks) with water added instead of DNA template; run at least one blank per 15 samples. Optionally, run the extraction blanks at the lowest dilution (highest concentration). Table 2 can be used to establish how much of each ingredient is needed.
4. Take out the reagents from the freezer, defrost them and put them on ice.
5. Pipet everything except your template DNA into a microcentrifuge tube. Vortex gently and quickly spin down the mixture using a centrifuge.
6. Aliquot 25 μL into each PCR tube (*see Note 21*).

Table 2
Overview table of ingredients in PCR mix

	Stock	Final	×1 reaction	×1 reaction	×__ reactions
	μM	μM	μL	μL	μL
<i>Master mix:</i>					
Water			0.165	8.25	
Buffer	×10	×1	0.1	5	
dNTPs	2,000	200	0.1	5	
MgCl ₂	25,000	750	0.03	1.5	
DreamTaq polymerase	5 units/μL	0.025 units/μL	0.005	0.25	
<i>Total volume:</i>			0.4	20	
<i>Per reaction:</i>					
Master mix			0.4	20	20
Tagged gITS7/ITS4 primer mix	5/3	0.5/0.3	0.1	5	5
Template DNA			0.5	25	25
<i>Reaction volume:</i>			1	50	50

7. Add the templates (25 μL) using new pipette tips for each sample.
8. Cap the tubes properly and spin down in a centrifuge.
9. Place tubes in the thermal cycler (PCR machine) and run the samples at the following cycle conditions: 5 min at 94 °C, typically 20–35 cycles of 30 s at 94 °C, 30 s at 56 °C and 30 s at 72 °C, and a final 7 min at 72 °C (*see Note 22*). Enter information on reaction volume and desired maximum number of cycles to run.
10. Pause the PCR machine when the desired test cycle numbers are reached (for example 22, 25, 28, 31, and 35) and take out an aliquot of 10 μL from each PCR reaction into a new tube for each cycle number.

3.3.2 PCR Products

Visualized by Gel

Electrophoresis

11. Set up electrophoresis cell, gel tray and combs to run all test PCR products including blanks. Fill the electrophoresis cell with 1xSB buffer (*see Note 23*).
12. To prepare 220 mL 1 % agarose gel (*see Note 24*), weigh 2.2 g of agarose into a 500 mL glass flask, and add 220 mL 1xSB buffer to the flask. Add 4 μL Nancy-520 dye (*see Note 25*).
13. Melt the agarose in a microwave oven. Make sure there are no bubbles or agarose crystals left.
14. Allow the melted agarose to cool to about 60 °C or cool enough to handle with your hands. Seal the ends of the gel tray and insert the combs, then pour the cooled agarose into the tray and allow it to solidify for about 30 min. When the agarose has solidified, carefully remove the combs and seals.
15. Load 3 μL of standard ladder in the wells closest to both sides of each sample row of the gel. Then load 5 μL of your samples into the wells. We usually load the gel in dry condition.
16. Place the tray with the gel in an electrophoresis cell and make sure that it is covered by buffer. Run the gel at about 10 V/cm for 20–30 min.
17. Stop the power supply, remove your gel and take a photo under UV or blue light in a gel documentation system.
18. Evaluate test PCR results.

Test PCRs: If any of the blanks have obvious PCR products, consider re-running the PCR, or even the DNA extraction. For each sample, evaluate which dilution worked the best, i.e. which gave strongest bands on the gel. Observe that if inhibitors are present in the DNA template, more diluted (less concentrated) samples may work better. Evaluate which cycle number worked the best; chose as few cycles as possible still giving sufficient PCR product, i.e. ideally weak to intermediate, but not very strong, bands on the

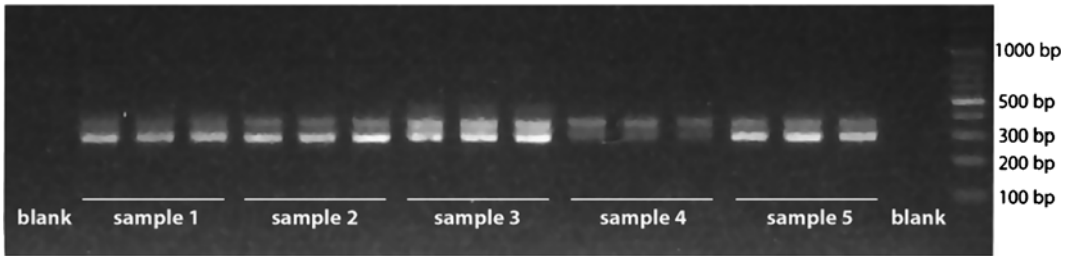


Fig. 5 The PCR products obtained with the gITS7–ITS4 primer combination separated on a 1 % agarose gel, stained with Nancy dye and visualized under UV light. Each sample is run in technical triplicates

gel (Fig. 5) (*see Note 26*). Note that a complex community is amplified, so PCR will not yield a single, sharp band, but commonly a pair of broad, diffuse bands (representing ascomycetes—shorter—and basidiomycetes—longer) or a smear located between the 300 bp and 500 bp markers.

3.4 Production of Tagged ITS Amplicons

3.4.1 PCRs with Sample-Tagged ITS Primers

1. Make new dilutions from the template as decided upon for each sample type. At least enough diluted sample to run three PCR reactions, each with 25 μL diluted sample, is needed.
2. Group samples according to the number of PCR cycles they should be subjected to. Assign a tagged primer mix to each sample. Calculate how many PCR reactions to run, starting with the PCR with fewest cycles. Run three technical PCR replicates of each sample.
3. Prepare a master mix with sufficient mix for all samples plus a few extra samples to allow for pipetting losses. Also include extraction blanks and PCR negative controls (blanks) with water added instead of DNA template; run at least one PCR blank per 48 samples. Table 2 can be used to establish how much of each ingredient is needed.
4. Take out the reagents from the freezer, defrost them and put them on ice.
5. Pipette everything but the tagged primer mix and your template DNA into a microcentrifuge tube. Vortex gently and quickly spin down the mixture using a table top centrifuge.
6. Aliquot 20 μL master mix into each PCR tube.
7. Add 5 μL of the tagged primer mix to each PCR tube using new pipette tips for each primer mix. Use different ITS tags for each sample as well as for extraction and PCR blanks, but same for the three PCR replicates of each sample. Make sure to record which tagged primer mix is used for each sample.
8. Add your template (25 μL) using new pipette tips for each sample. Prepare three technical PCR replicates of each.
9. Cap the tubes properly and spin down in a centrifuge.

10. Place tubes in the thermal cycler (PCR machine) and run the samples at the following cycle conditions: 5 min at 94 °C, typically 20–30 cycles of 30 s at 94 °C, 30 s at 56 °C and 30 s at 72 °C, and a final 7 min at 72 °C (*see Note 22*). Add information on reaction volume and desired number of cycles to run. Observe that when producing the final amplicons, all PCRs should be run to the end of the program to ensure the highest possible quality of the product.
11. Run 5 µL of the PCR products on an agarose gel (*see Subheading 3.3.2*). Choose only good products for further processing, ideally with weak to intermediate, but not very strong, bands on the gel (Fig. 5). In cases where samples gave no or very weak PCR products, these samples can be re-run at a higher cycle number, or other template concentrations considered.

3.4.2 Clean PCR Products with the AMPure Kit

12. Add 81 µL AMPure magnetic bead solution (1.8 × sample volume) to each PCR product of 45 µL (*see Notes 27, 28*).
13. Transfer the PCR-product/bead mix to a PCR-plate with raised wells.
14. Incubate at room temperature for 3–5 min.
15. Place the plate on the magnetic plate. Incubate for 5–10 min.
16. Keeping the plate on the magnet, turn the plate upside-down and try to get rid of the liquid. The plate can be gently hit against a table with kitchen tissue paper to absorb the liquid.
17. Add 200 µL 70 % ethanol to each well and incubate for 30 s at room temperature. Get rid of the liquid as in 16.
18. Repeat **step 17**. This time it is important to get rid of as much liquid as possible. Hit the plate hard several times against the table, until no drops appear on the kitchen-roll paper. Keep the plate in the magnetic plate at all times.
19. Let the plate dry at 37 °C for about 60 min; it is important to get rid of all ethanol. The magnetic plate is not necessary at this stage.
20. Remove the plate from the magnet. Add 60 µL elution buffer to each well, cover with plastic foil, vortex and spin down.
21. Place the plate on the magnet before pipetting the supernatant. Alternatively, the magnetic beads can be pelleted by centrifugation at 1900 × *g* for 10 s.

3.5 Amplicon Mixing and Sequencing

3.5.1 Quantification of Double-Stranded DNA with the Qubit dsDNA HS Assay Kit

1. All reagents should be allowed to adjust to room temperature (*see Notes 25, 27, 29*).
2. Set up the number of 0.5 mL tubes you will need for standards and samples. The assay requires 2 standards.
3. Prepare the working solution by diluting the dsDNA HS reagent (component A) 1:200 in dsDNA HS buffer (component B) in a tube. Prepare 200 µL per sample/standard, plus

for one extra sample to account for pipetting losses. That is, mix 199 μL of component B (\times number of samples) with 1 μL of component A (\times number of samples).

4. Load 190 μL of the working solution into each of the tubes used for standards.
5. Add 10 μL of each standard (C and D) to the appropriate tube and mix by vortexing for 2–3 s. Careful pipetting is critical to ensure that exactly 10 μL of each standard is added to 190 μL of working solution.
6. Load 197 μL working solution into individual sample assay tubes.
7. Add 3 μL of your samples to assay tubes containing the working solution and mix by vortexing for 2–3 s. The final volume in each tube should be 200 μL .
8. Centrifuge the samples at $1000 \times g$ for 10 s to get rid of air bubbles.
9. Allow all tubes to incubate at room temperature for 2 min.
10. On the Home Screen of the Qubit Fluorometer, press “DNA,” and then select “dsDNA High Sensitivity” as the assay type. The Standards Screen is automatically displayed.
11. On the Standards Screen, press “Yes” to run a new calibration.
12. Running a New Calibration: Insert the tube containing Standard 1 in the Qubit Fluorometer, close the lid and press “GO.” Also press “START” on the computer screen if using complementary software to store recordings. The reading will take approximately 3 s. Remove Standard 1. Repeat for Standard 2.
13. Insert a sample tube into the Qubit Fluorometer, close the lid, and press “GO.”
14. Upon the completion of the measurement, the result will be displayed on the screen. The Qubit machine can do calculations to account for dilution in the assay directly after each measurement. Alternatively, see calculation below.
15. Repeat sample readings until all samples have been measured.
16. The Qubit Fluorometer (QF) gives values for the Qubit dsDNA HS assay in ng/mL . This value corresponds to the concentration after samples were diluted into the assay tube. To calculate the dsDNA concentration in your sample, use the following equation:

*3.5.2 Calculating
the Concentration
of dsDNA in Each Sample*

$$\text{Sample dsDNA concentration (ng / } \mu\text{L)} = \text{QF} \times (200 \mu\text{L} / X \mu\text{L}) / 1000$$

where QF = the concentration given by Qubit in ng/mL
 X = the number of μL of sample you added to the assay tube

3.5.3 *Make an Equal Mix of all Amplicons*

17. Decide how much DNA (ng) you wish to pool from each sample, aiming for the same amount from each sample (*see Note 29*). Typically, this would correspond to the total amount of DNA in the sample with the lowest concentration. The DNA amount required per DNA pool that is sent for sequencing depends on the sequencing platform. Typically, at least 1 μg of total dsDNA is needed for amplicon samples. Also account for losses of DNA (typically one-third to half) in the final cleaning of the DNA mix.
18. Calculate the volume needed from each sample to obtain the decided amount of DNA in the mix. Typically, mix 50–100 ng DNA from each of the PCR reactions. If some samples have very little PCR product you may have to scale amounts of these samples down relative to the rest of the samples, i.e. take fewer ng of these samples, well-knowing that they may be under-represented in the sequence data. Calculate the amount to be pooled from each sample:

$$\mu\text{L sample to be pooled} = (\text{DNA amount to be pooled, ng}) / (\text{sample DNA concentration, ng} / \mu\text{L})$$

3.5.4 *Clean Amplicon Mix with Cycle-Pure Kit (Omega)*

19. The final amplicon mix is purified to remove leftovers from the PCR mix and smaller (unwanted) DNA fragments, such as primer dimers. Determine the volume of the amplicon mix to be purified. The volume can be reduced by speedvac or freeze-drying, but avoid drying out the sample completely. Pellet any remaining magnetic beads (from AMPure step) by centrifugation. Place the tube against a magnet and transfer the sample into a clean tube; avoid transferring magnetic beads as much as possible. Add 6 volumes of Cycle-Pure buffer CP; e.g., if your PCR mix is 100 μL , add 600 μL of buffer CP.
20. Vortex thoroughly to mix. Briefly centrifuge the tube to collect any drops from the inside of the lid.
21. Place a Cycle-Pure minicolumn into the 2 mL collection tube.
22. Add the mixed sample from **step 2** to the minicolumn and centrifuge at $13,000 \times g$ for 1 min at room temperature. Discard the flow-through liquid and place the minicolumn back into the same collection tube. The column can hold up to 700 μL at a time. If sample volume is larger than this, it can be loaded onto the same column multiple times and the centrifugation repeated.

23. Add 700 μL of Cycle-Pure wash buffer (ethanol diluted) and centrifuge at $13,000\times g$ for 1 min. Discard the flow-through liquid and place the minicolumn back into the same collection tube.
24. Add 500 μL of wash buffer and centrifuge at $13,000\times g$ for 1 min. Discard the flow-through liquid and place the minicolumn back into the same collection tube.
25. Centrifuge the empty minicolumn for 2 min at maximal speed ($\geq 13,000\times g$) to dry the column matrix. Do not skip this step; it is critical for the removal of ethanol from the minicolumn.
26. Place the minicolumn into a clean 1.5 mL microcentrifuge tube. Depending on the desired concentration of the final product, add 30–50 μL of water directly onto the centre of the column matrix. Incubate at room temperature for 2 min. Centrifuge for 1 min at $13,000\times g$ to elute the DNA. This eluates approximately 80–90 % of bound DNA. Repeat the eluation step once and pool the two eluates.
27. Store samples at 4 °C for short term or at –20 °C for longer term.

3.5.5 PCR Product
Quantification and Quality
Control by Bioanalyzer
(Agilent Tech)

28. Allow the Agilent gel-dye mix to equilibrate to room temperature for 30 min before use (*see* **Notes 25, 30**).
29. Put a new Agilent DNA chip on the chip priming station.
30. Pipette 9.0 μL gel-dye mix into the well marked with **G**.
31. Make sure that the plunger is positioned at 1 mL and then close the chip priming station.
32. Press the plunger until it is held by the clip.
33. Wait for exactly 30 s, then release the clip.
34. Wait for 5 s and then slowly pull back the plunger to the 1 mL position.
35. Open the chip priming station and pipette 9.0 μL gel-dye mix into the wells marked G.
36. Pipette 5 μL of marker (green) into all 12 sample wells and the ladder well. Do not leave any wells empty.
37. Pipette 1 μL of DNA ladder (yellow) into the well marked with a ladder.
38. In each of the 12 sample wells, pipette 1 μL of sample (used wells) or 1 μL of de-ionized water (unused wells).
39. Put the chip horizontally in the adapter and vortex for 1 min at the indicated setting (2400 rpm).
40. Run the chip in the Bioanalyzer within 5 min. The result will appear in the screen (Fig. 6).

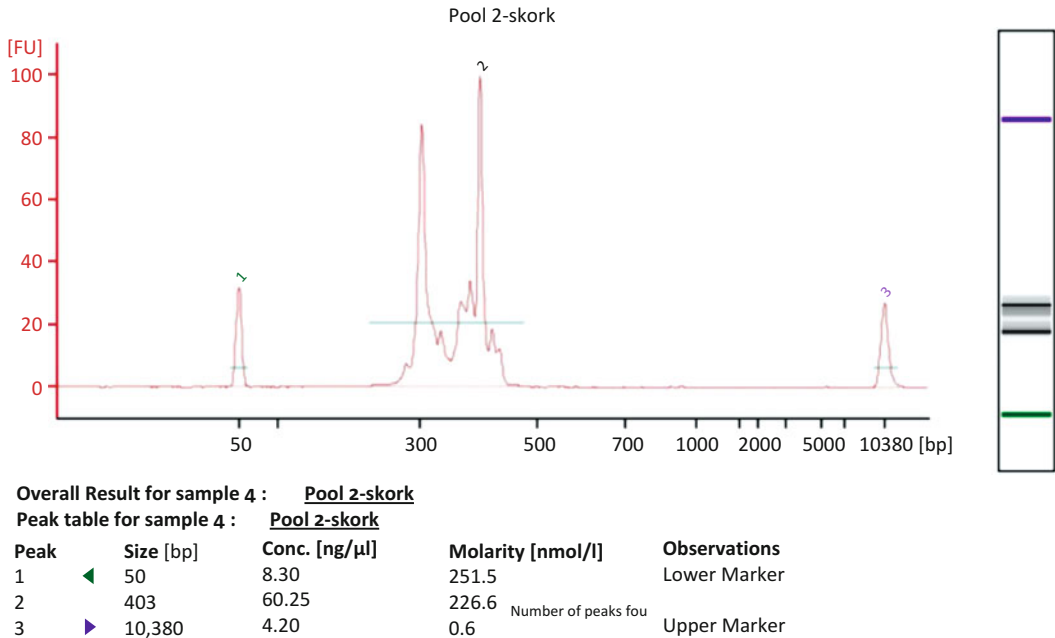


Fig. 6 DNA size distribution profiles in a multiplexed ITS2 amplicon mix from 100 samples as analyzed by Bioanalyzer. The amplicon mix is represented by a smear of sizes between 200 and 500 bp. The main peaks around 300 and 400 bp probably represent ascomycetes and basidiomycetes, respectively. Two markers are labeled by *green* (50 bp) and *violet* (10,380 bp) numbers or bands in the electropherogram (*left*) and in the virtual gel image (*right*), respectively

41. An amplicon mix with sufficient DNA and of the expected size distribution can be sent for sequencing, in either frozen or freeze-dried condition.

4 Notes

1. To prepare 1 L stock of 1 M Tris-HCl [pH 8], dissolve 121.1 g Trisbas in 800 mL water on a magnetic stir. Adjust pH to 8 by adding about 42 mL concentrated HCl (12 N). Adjust the volume to 1 L with water. Sterilize by autoclaving and check pH again.
2. To prepare 1 L stock of 5 M NaCl, dissolve 292.2 g NaCl in 800 mL water on a magnetic stir. Adjust the volume to 1 L with water. Sterilize by autoclaving.
3. To prepare 1 L stock of 0.5 M EDTA [pH 8], dissolve 186.1 g disodium ethylenediamine tetraacetate $\times 2$ H₂O (EDTA) in 800 mL water on magnetic stir. Adjust pH to 8 by adding about 20 g NaOH. Adjust volume to 1 L with water. Sterilize by autoclaving and check pH again.

4. Alternatively, a vacuum manifold speeds up sample handling significantly when working with many samples.
5. Depending on sample type and DNA concentration another polymerase may work better.
6. The used fungal-specific forward primer, gITS7, paired with the general, eukaryote ITS4 primer targets some plant species as well. Usually, for soils including roots, at the most 15 % of sequences represent plants. Alternative fungal-specific forward primers, e.g. fITS7 or fITS9, may be used to target the ITS2 region (Fig. 2). See the paper by Ihrmark et al. [11], for specificities of these primers.
7. If subsampling is needed before freeze-drying, mix the samples thoroughly in fresh condition, e.g. by cutting sample with scissors. Split the sample into evenly sized subsamples for different purposes to avoid subsampling biases.
8. This protocol, based on 3 % CTAB extraction buffer, has been used for a range of different substrates in our lab, although originally developed for DNA extractions from fungal tissues. DNA extracts from some sample types, such as humus-rich soils, need to be further purified. An alternative commercial DNA extraction kit that we have found efficient for difficult substrates is the NucleoSpin soil kit (Macherey-Nagel).
9. Multiple extractions from each sample may be a way to obtain more correct representation of DNA present in the sample in the extracted DNA. *See* also **Note 11**.
10. Alternatively, add two nuts to enhance fragmentation of more sturdy materials such as needle litters and wood.
11. The first step of the extraction can be done in a larger volume of CTAB if a larger amount of sample is extracted. This could be desirable if lowly abundant organisms or highly heterogeneous substrates are studied. For example, homogenization and extraction of 1 g of organic matter in 10 mL of extraction buffer followed by centrifugation and subsampling of extraction buffer (e.g. 500 μ L) before the next step in the protocol may increase representativity of extracted DNA significantly.
12. Depending on how the phase separation works for different sample types, the volumes transferred from the supernatants can be adjusted. Keep track of the dilution factor of the DNA throughout the procedure, to enable later calculations of amount of DNA per original sample mass.
13. Chloroform is a strong organic solvent. All work with chloroform must be carried out in a fume hood, wearing gloves and a lab coat. Waste chloroform must be kept in the fume hood and should not be poured out in the sink.

14. This step can also be done at room temperature, which gives less but cleaner DNA, or in the freezer, which gives more but less clean DNA.
15. Thoroughly mix the Wizard DNA clean-up resin before removing an aliquot. If crystals or aggregates are present, dissolve by warming the resin to 37 °C for 10 min. The resin itself is insoluble. Cool to 25–30 °C before use.
16. The sample volume must be between 50 and 500 µL. If the sample volume is less than 50 µL, bring the volume up to at least 50 µL with water. If the sample volume is >500 µL, split the sample into multiple purifications. The binding capacity of 1 mL of resin is approximately 20 µg of DNA.
17. The DNA concentration and purity in each extract is measured spectrophotometrically by the NanoDrop machine using the “DNA-50” software. DNA concentration in the final eluate can be calculated from its absorption maximum at 260 nm (A_{260}) as an absorbance of $A_{260}=1$ Absorbance Units (AU) corresponds to 50 ng/µL double-stranded DNA. This calculation assumes the absence of any other compound that absorbs UV light at 260 nm. Any contamination with, for example, RNA, protein, or especially humic substances significantly contributes to the total absorption at 260 nm and therefore leads to an overestimation of the DNA concentration. This method is therefore not recommended for exact concentration measurements at <5 ng/µL. Confirm that the blanks yield a reproducible zero by analyzing blanks as though they were samples; the result should vary no more than 0.03 AU (± 1.5 ng/µL) from the stored blank value.
18. The ratio of absorbance at 260/280 and 260/230 nm is used to assess the purity of DNA; 260/280 ratios of about 1.7–1.8 are accepted as pure for DNA. If the ratios are appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 230 or 280 nm. Also note that extracts purified with the Wizard DNA clean-up kit have substances that interfere with DNA absorbance at 260 nm, and therefore spectrophotometric methods cannot be used for DNA quantification in Wizard cleaned samples. Instead, PCR tests are run without prior DNA quantification (*see Note 20*).
19. In order to minimize amplification bias—i.e. aiming to conserve the original relative abundances of ITS types, it is important to optimize both the template concentration and the number of PCR cycles for each sample (or sample type) individually. The samples should be diluted to overcome any inhibition of the PCR reaction, but not more, enabling the desired amount of PCR product to be reached after as few PCR cycles

as possible. The PCR reaction should be interrupted during (the middle of) the phase of exponential increase in product, and should not be allowed to enter the “saturated” phase, in which the community could be altered due to, e.g. primer or dNTP limitation. Optimal dilutions and cycle numbers can be tested either with normal PCR or with quantitative real-time PCR (qPCR, Fig. 3). Here, we describe a procedure using normal PCR. When working with many samples, it may be practical to do the optimizations for a subset of samples representing all sample types, and optimizations can subsequently be generalized to other samples of the same type.

20. In samples, such as humus-rich soils, where inhibition is often a larger problem than small or varying amounts of DNA, it is often more time-efficient to leave out the NanoDrop measurements and instead dilute DNA extracts directly, e.g. by factors $\times 10$, $\times 100$, and $\times 1000$ with water (giving final PCR mix dilutions of $\times 20$, $\times 200$, and $\times 2000$). Note that suggested dilutions may not necessarily be optimal for all sample types and more levels could be tested.
21. In the test PCRs there is no need to use differently tagged primer pairs for each sample, and a single primer mix can be added directly to the master mix, of which 25 μL should then be aliquoted to each reaction well.
22. These PCR cycling conditions are adjusted to the gITS7–ITS4 primer pair, and should be adjusted if other primers are used.
23. PCR products can be visualized by gel electrophoresis in order to inspect amount and size of amplified product. In electrophoresis, DNA molecules are separated according to their size. When an electrical field is applied to the agarose gel, DNA molecules, with their negatively charged phosphate groups, will migrate from the negatively charged starting point toward the positively charged end. Smaller DNA molecules move more quickly than larger molecules, and after some time DNA molecules will be spatially separated according to their sizes in the gel.
24. This amount (220 mL) of gel fits into an approximately 15 by 25 cm tray, with enough space for four combs, each with 26 teeth. Observe that the gels do not need to run for very long to validate the products.
25. Nancy-520 dye, Qubit and Bioanalyser kit components contain DMSO, an organic solvent that can facilitate the entry of organic molecules into tissues. Because the dye binds to nucleic acids, it should be treated as a potential mutagen and used with appropriate care. Wear hand and eye protection and follow good laboratory practices when preparing and handling reagents and samples. Handle the DMSO stock solutions with particular caution.

26. Not every PCR is successful. The quality of the DNA may be poor, further purification of the DNA may be necessary, the primers may not fit, the concentration of starting template or any of the PCR ingredients may be non-optimal or cycle conditions may be imperfect. Further, if the PCR product is of a different size than expected, at least one of the primers may be unspecific enough to amplify a different DNA region—eventually a different organism—than the targeted region/organism. This may be caused by unspecific primers or by PCR settings, e.g. to low annealing temperature.
27. To save time and resources, the technical PCR replicates may be pooled either before the AMPure cleaning or the Qubit analysis. However, such pooling makes the procedure more sensitive to deviating PCR products, as the three technical replicates will be pooled in equal proportion of PCR product volume (μL), but not necessarily in DNA amount (ng). Therefore, it is important to inspect that all three technical replicates look similar on the gel picture before such pooling. Also, the total volume of pooled PCR product to be cleaned with AMPure will be too large for one well in the PCR plate, with the bead solution also added. The maximum volume to clean with AMPure in one well is 70 μL giving a volume of 200 μL after adding the bead solution. The volume of pooled technical PCR replicates may be decreased by, e.g. freeze-drier or speedvac.
28. The PCR products must be purified to get rid of salts, unincorporated dNTPs, and unused primers. The AMPure kit consists of small magnetic beads that bind DNA. By using a magnet to retain the beads (with the DNA), it is possible to wash the DNA. Once the beads have been washed and dried, the DNA can be eluted from the beads with water. A multi-pipette is useful in this protocol. AMPure contains sodium azide, which is toxic. Be careful and use gloves.
29. To use high-throughput sequencing technologies optimally to cover amplicons from many samples simultaneously, multiplexing of amplicons using sample-specific tags is done by pooling of all the amplicons, to enable sequencing a single composite sample. To obtain the most even pooling among samples, exact quantification of DNA concentrations is essential. Since amplicons consist of ITS sequences of different lengths, a true equimolar mix with same number of sequences pooled from all samples would require analyses of size distributions and molarity of all samples (e.g. by Bioanalyzer). However pooling same amount (ng) of dsDNA from each sample, for mixed communities, normally gives reasonably similar coverage of all samples. For exact quantification of the PCR products, we use the Qubit HS (high sensitivity) assay. This method is based on a

fluorophore that binds to the double-stranded DNA (dsDNA), is highly specific for dsDNA over RNA, and is accurate for initial DNA concentrations between 10 pg/ μ L and 100 ng/ μ L. The assay is performed at room temperature, and the signal is stable for 3 h, but it takes only a couple of min to measure a sample. The assay is insensitive to common contaminants, such as salts, free nucleotides, solvents, detergents, or protein.

30. The final DNA mix is run on a “gel on a chip” on a Bioanalyzer, providing information on the size distribution, concentration, and the molarity of the DNA. *See* Fig. 6 for an example of a Bioanalyzer output from an ITS2 amplicon mix. Observe that it may be necessary (depending on sequencing platform) to size-fractionate amplicons with wide size range before sequencing in order to secure good coverage of all amplicon sizes.

References

1. Tedersoo L, Bahram M, Pöhlme S, Kõljalg U, Yorou NS, Wijesundera R et al (2014) Global diversity and geography of soil fungi. *Science* 346:1078
2. Clemmensen KE, Finlay RD, Dahlberg A, Stenlid J, Wardle DA, Lindahl BD (2015) Carbon sequestration is related to mycorrhizal fungal community shifts during long-term succession in boreal forests. *New Phytol* 205:1525–1536
3. Rosling A, Cox F, Cruz-Martinez K, Ihrmark K, Grelet GA, Lindahl BD et al (2011) Archaeorhizomycetes: unearthing an ancient class of ubiquitous soil fungi. *Science* 333:876–879
4. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A* 109:6241–6246
5. Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R et al (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* 2:e197
6. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M et al (2013) Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol* 22:5271–5277
7. Petersen L, Minkinen P, Esbensen KH (2005) Representative sampling for reliable data analysis: theory of sampling. *Chemometr Intell Lab* 77:261–277
8. Lennon JT (2011) Replication, lies and lesser-known truths regarding experimental design in environmental microbiology. *Environ Microbiol* 13:1383–1386
9. Prosser JI (2010) Replicate or lie. *Environ Microbiol* 12:1806–1810
10. Lindahl BD, Kuske CR (2013) Metagenomics for study of fungal ecology. In: Martin F (ed) *Ecological genomics of the fungi*. Wiley-Blackwell, Hoboken, NJ, USA
11. Ihrmark K, Bodeker I, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J et al (2012) New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol Ecol* 82:666–677
12. Balaud R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kausrud H (2013) ITS1 versus ITS2 as DNA metabarcodes for fungi. *Mol Ecol Resour* 13:218–224
13. Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjoller R et al (2013) Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide. *New Phytol* 199:288–299
14. Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD (2014) Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Appl Environ Microbiol* 80:5717–5722
15. Frank DN (2009) BARCRAWL and BARTAB: software for design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics* 10:362
16. Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kausrud H (2012) Don't make a mistake: is tag switching an overlooked

source of error in amplicon pyrosequencing studies? *Fungal Ecol* 5:747–749

17. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK et al (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336
18. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541
19. Abarenkov K, Nilsson RH, Larsson KH, Alexander IJ, Eberhardt U, Erland S et al (2010) The UNITE database for molecular identification of fungi - recent updates and future perspectives. *New Phytol* 186:281–285
20. White TJ, Bruns TD, Lee S, Taylor JW, Innis M, Gelfand D et al (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Shinsky JJ, White TJ (eds) *PCR protocols. A guide to methods and applications*. Academic, Orlando, USA, pp 315–322

Fungal Communities in Soils: Soil Organic Matter Degradation

Tomáš Větrovský, Martina Štursová, and Petr Baldrian

Abstract

Stable isotope probing (SIP) provides the opportunity to label decomposer microorganisms that build their biomass on a specific substrate. In combination with high-throughput sequencing, SIP allows for the identification of fungal community members involved in a particular decomposition process. Further information can be gained through gene-targeted metagenomics and metatranscriptomics, opening the possibility to describe the pool of genes catalyzing specific decomposition reactions in situ and to identify the diversity of genes that are expressed. When combined with gene descriptions of fungal isolates from the same environment, specific biochemical reactions involved in decomposition can be linked to individual fungal taxa. Here we describe the use of these methods to explore the cellulolytic fungal community in forest litter and soil.

Key words Stable isotope probing, Microbial communities, Soil ecology, Organic matter decomposition, Metagenomics, Metatranscriptomics, Cellulose

1 Introduction

Soils contain one of the largest pools of organic carbon on the Earth, and soil processes therefore play a major role in the global C cycle. Understanding organic matter decomposition and the involvement of microorganisms in this process are essential for current and future carbon balance predictions. Most of the organic matter in soils is of plant origin and is composed of the polymers of the plant cell wall—cellulose, hemicelluloses, and lignin or of the cell wall of fungi—chitin. Cellulose is the most abundant of these biopolymers in litter where it typically constitutes 20–30 % of its mass while chitin is an important component of fungal mycelia abundant in soils [1, 2]. Decomposition of soil organic matter, especially cellulose, was the subject of intensive research for decades. Cellulolytic capabilities are relatively common in saprotrophic fungi and while it is these fungi that dominate cellulose decomposition in soils [3–5], it is currently known that certain bacterial

groups also harbor these functions. Knowledge of microbial decomposers has been largely derived from laboratory studies on a small number of isolated strains. Nowadays, novel molecular approaches, such as stable isotope probing (SIP) and next-generation sequencing, make it possible to analyze substrate utilization by all members of microbial communities concurrently and at sufficient resolution [6]. Here, we demonstrate the use of these methods to explore the fungal community actively involved in cellulose decomposition in forest litter and soil and to characterize the potential and active producers of the *ccbI* gene encoding for GH7 glycosyl hydrolase with exocellulase activity [7–9]. The results reveal that several fungal taxa contain and actively transcribe exocellulases and that many of them are able to accumulate cellulose-derived C in their biomass in high quantities.

2 Materials

2.1 Stable Isotope Probing

1. Sterile 160-mL serum bottles with airtight rubber stoppers.
2. ^{13}C -labeled *Zea mays* cellulose (97 atom% ^{13}C ; IsoLife, Wageningen, the Netherlands) or other substrate with high level of ^{13}C enrichment (>90 %) (*see Note 1*).
3. N_2 purged syringes, tubes with airtight rubber septa, and mass spectrometer such as IsoPrime (GV Instruments, Manchester, UK) for the analysis of respired ^{13}C in CO_2 .
4. Fast DNA Spin Kit for Soil (MP Biomedicals, Solon, OH) and a spectrophotometer or fluorimeter for DNA quantification such as ND1000 (NanoDrop, Wilmington, DE) and Qubit (Life Technologies, Carlsbad, CA).
5. ND1000 (NanoDrop, Wilmington, DE) and quantify dsDNA using Qubit (Life Technologies, Carlsbad, CA).
6. CsTFA (Cesium Trifluoroacetate) solution: Mix 3.17 mL of CsTFA with 1.93 mL nuclease-free water in 5.1 mL tube (*see Note 2*). Unlabeled DNA will band around buoyant density (BD) of 1.60 g/mL. Adjust BD of master mix to 1.61 (because addition of sample will reduce the final BD).
7. For centrifugation, Beckman polyallomer quick-seal tubes (13 × 51 mm, 5.1 mL) and L-100XP Optima Ultracentrifuge with near vertical rotor such as the NVT 100 rotor (Beckman Coulter, Brea, CA).
8. For fractionation of gradients after centrifugation: Fractionator—Fraction Recovery System, Puncturing, for Thin-Walled Tubes (Beckman Coulter, Brea, CA), ethanol, nuclease-free water, and syringe pump such as NE-1000 (New Era Pump Systems, Farmingdale, NY) and respective syringes.

9. For processing of DNA fractions: vacuum concentrator such as SpeedVac (Thermo Fisher Scientific, Waltham, MA).
10. For RT-PCR screening in DNA fractions: 96 well optical PCR plates (Life Technologies, part no. 4306737), optical adhesive covers for PCR plates (part no. 4360954), SYBR green PCR master mix (part no. 4334973), 10 mg/mL Bovine Serum Albumin solution, PCR-grade water.
11. For the quantification of fungal rDNA: RT-PCR 5 pmol primers, such as ITS1 (5'-TCC GTA GGT GAA CCT GCG G-3') and ITS2* (5'-TTY GCT GYG TTC TTC ATC G-3') [10] and rDNA standard such as the cloned rDNA region of *Saccharomyces cerevisiae*.

2.2 Gene-Targeted Metagenomics and Metatranscriptomics

1. For DNA and RNA co-extraction and purification: RNA PowerSoil Total RNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA), RNA PowerSoil DNA Elution Accessory Kit (MoBio Laboratories), OneStep PCR Inhibitor Removal Kit (Zymo Research, Irvine, CA, USA).
2. For reverse transcription of RNA: 200 U/ μ L Superscript III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA), DNase I (Sigma), Random hexamer primers (Sigma).
3. For specific amplification of partial sequence of fungal *cbhI* exocellulase (GH 7 glycosyl hydrolase): 2.5 U/ μ L Pfu DNA Polymerase (Fermentas, Thermo Fisher Scientific, Waltham, MA), 2 U/ μ L DyNAZyme II DNA Polymerase (Finnzymes, Thermo Fisher Scientific, Waltham, MA), PCR Nucleotide Mix 10 mM (Finnzymes), 10 mg/mL Bovine Serum Albumin solution, PCR-grade water.
4. 10 pmol PCR primers *cbhIF* (5'-ACC AAY TGC TAY ACI RGY AA-3') and *cbhIR* (5'-GCY TCC CAI ATR TCC ATC-3') [8] with sample-specific barcodes (short oligonucleotides that extend the primers at the 5'-end).
5. For purification of PCR products: MinElute Kit (Qiagen, Valencia, CA) or Agencourt AMPure XP beads (Beckman Coulter, Brea, CA).

3 Methods

3.1 SIP to Identify Cellulose-Utilizing Fungi

1. Collect soil or litter sample (*see Note 3*). Sieve soil through a 5-mm screen or cut litter it into small pieces. Incubate at 4 °C for 1 month to stabilize substrates. Nutrients liberated as a result of sample collection and homogenization are consumed during this time.

2. Transfer samples to the desired experiment incubation temperature (in our case 11 °C an average annual temperature at site in depth of 3 cm) and allow them to stabilize for 36–48 h.
3. Mix 5.0 g wet mass sample with 100 mg of ¹³C-labeled cellulose in a serum bottle sterilized by autoclaving. Seal bottles with rubber stoppers and aluminum crimp caps to keep them airtight. This allows for later sampling and quantification of the respired ¹³C-CO₂.
4. Incubate microcosms in the dark at the desired temperature (*see Note 4*). Harvest samples immediately (day 0) and after 8, 15, and 22 days (*see Note 5*). Before collecting soil from the microcosms, collect 1 mL of the headspace gas using an N₂-purged syringe to determine the carbon isotopic composition of CO₂ and store it in 12-mL airtight serum bottles with rubber septa at room temperature until analysis (*see Note 6*).
5. After harvest, proceed immediately with DNA extraction or freeze microcosm materials at –80 °C for later extraction.
6. Extract DNA from 0.5 g aliquots of microcosm material using the Fast DNA Spin Kit for Soil (MP Biomedicals, Solon, OH) according to the manufacturer's instructions, check DNA yield and ND1000 (NanoDrop, Wilmington, DE), and quantify dsDNA using Qubit (Life Technologies, Carlsbad, CA). The minimum amount of DNA required for subsequent steps is 1 µg.
7. Before the preparation of density gradients to be used for DNA separation, allow CsTFA solution to warm to room temperature (*see Note 7*).
8. Prepare DNA for separation. One microgram or more of DNA is needed per gradient (*see Note 8*). Before application, standardize all DNA extracts to the same concentration (e.g., dilute to 100 ng/µL) in order to use same volume of all samples to be combined with CsTFA. Place the DNA samples on ice.
9. Label Beckman polyallomer quick-seal tubes with sample numbers. With a 10-mL syringe, transfer 5.1 mL CsTFA solution to each tube. Leave a small airspace below the neck of the tube. Take the weight of all tubes and ensure that their weights do not differ by more than 10 mg.
10. Using a micropipette, add DNA sample to the tube, ensuring that it enters the CsTFA solution. Add the same volume for all samples. All centrifugation runs should contain blank tubes used to determine buoyant density of each fraction. To blanks add water instead of sample.
11. Heat-seal the tubes and place them into the NVT 100 rotor. Tubes should be placed in opposite positions and distributed

as equally as possible within the rotor when all positions are not used—check the centrifuge manual for recommended positions.

12. Use torque wrench with plug adaptor to tighten down plugs to 11 N.m, pushing down on the plug adaptors as you turn it. Do not use plugs in empty rotor spaces. Carefully place rotor onto the drive hub of the ultracentrifuge.
13. Perform centrifugation under the following conditions: speed: 141,400 $\times g$, temperature: 25 °C, acceleration: maximum, deceleration: 9. Run the centrifugation for 36–48 h.
14. Prepare the fractionator by washing its needle with a 1-mL syringe full of ethanol followed by two washes of nuclease-free water. Prime the syringe pump by filling 10-mL syringe with nuclease-free water and allow running until no bubbles remain in lines going to the fractionator. Be sure that the syringe pump is set to the proper speed (0.5 mL/min).
15. Stop the ultracentrifuge, remove two opposite samples for fractionation, and allow others to continue spinning. Fractionate blanks first.
16. For each tube, use a razor blade to slice off the top of the sealed tube. Using a 1-mL syringe, carefully fill airspace at the top of the tube with water. Cover the top with two layers of parafilm and puncture 8–10 times with a syringe needle.
17. Place the tube into the fractionator. While holding 1.5 mL tube beneath outflow, quickly turn on syringe pump and timer and puncture the bottom of the tube with needle assembly. Collect fractions every 30 s (250 μ L each).
18. For the blanks, collect all 20 fractions and check the BD of each fraction by weighing 100 μ L (2 replicates per fraction). Select your target range for sample fractions: unlabeled DNA will band around BD 1.60, and ^{13}C DNA will be around 1.65 or slightly less. As a rule of thumb, include about four to six fractions above 1.60 and six to eight fractions below it.
19. Fractionate remaining samples. Fractions can be stored at room temperature on the bench until all samples are processed.
20. After fractionation, add 1 mL of isopropanol (molecular biology grade) to each DNA fraction and mix. Incubate DNA fractions at room temperature for 1–2 h or at -20 °C overnight.
21. Centrifuge at maximum speed in a microcentrifuge for 30 min at room temperature and pour off the isopropanol. Wash the pellet with 0.5 mL isopropanol, spin again for 30 min, and pour off isopropanol again. Repeat pellet wash and carefully pipet off isopropanol.

22. Dry the pellets (DNA) in a vacuum concentrator or on bench (*see Note 9*).
23. Resuspend pellets in 20 μL of nuclease-free water or in a PCR-compatible elution buffer. Mix well by flicking. Store DNA at $-20\text{ }^{\circ}\text{C}$.
24. Quantify the amount of DNA in the fractions using Real-time PCR (*see Note 10*). For one 96-well plate prepare 1500 μL of PCR premix by mixing 7.5 μL of SYBR green master mix, 0.6 μL of BSA (10 mg/mL), 0.9 μL of forward and reverse primers, and 4.1 μL of nuclease-free sterile water per one reaction.
25. Use three replicates of each standard including negative control and two to three replicates of each fraction.
26. Use 96-well optical PCR plates. In each well mix 1 μL of a DNA sample or DNA standard solution with 14 μL of PCR premix to make a total volume of 15 μL . Spin the plate down to mix all reagents and remove air bubbles.
27. Run PCR with the following cycling parameters: 2 min at $56\text{ }^{\circ}\text{C}$, 10 min at $95\text{ }^{\circ}\text{C}$ followed by 40 cycles of $95\text{ }^{\circ}\text{C}$ for 15 s and $60\text{ }^{\circ}\text{C}$ for 1 min. Use heated cover option ($103\text{--}105\text{ }^{\circ}\text{C}$) and include the melting curve option (if available) to check the size of fragments.
28. After RT-PCR make a plot with fraction numbers on the x axis and relative amount of DNA on the y axis. For each sample, plot the DNA amounts of the sample and the control (unlabeled DNA).
29. Those fractions that contain DNA after separation of the control DNA should be collected and pooled as “light” ^{12}C -DNA; those fractions that contain DNA after ultracentrifugation of the labeled sample but no DNA in the fractions containing the control DNA should be collected as the “heavy” ^{13}C -DNA (*see Note 11*, Fig. 1).
30. Use the “light” and “heavy” DNA pools of each sample to analyze microbial community composition by PCR amplification of a fungal marker gene of choice and high-throughput sequencing and sequence analysis as described in other chapters of this book or [11, 12].
31. Identify the microorganisms utilizing cellulose as those that occur in the “heavy” fraction from the ^{13}C -enriched microcosm but exclude those where the ratio of their relative abundance in the “heavy”/“light” fraction from the ^{13}C -enriched microcosm is smaller or similar to the same ratio in the unlabeled microcosms because their appearance in the “heavy” fraction may not necessarily be the result of ^{13}C -enrichment (*see Note 11*).

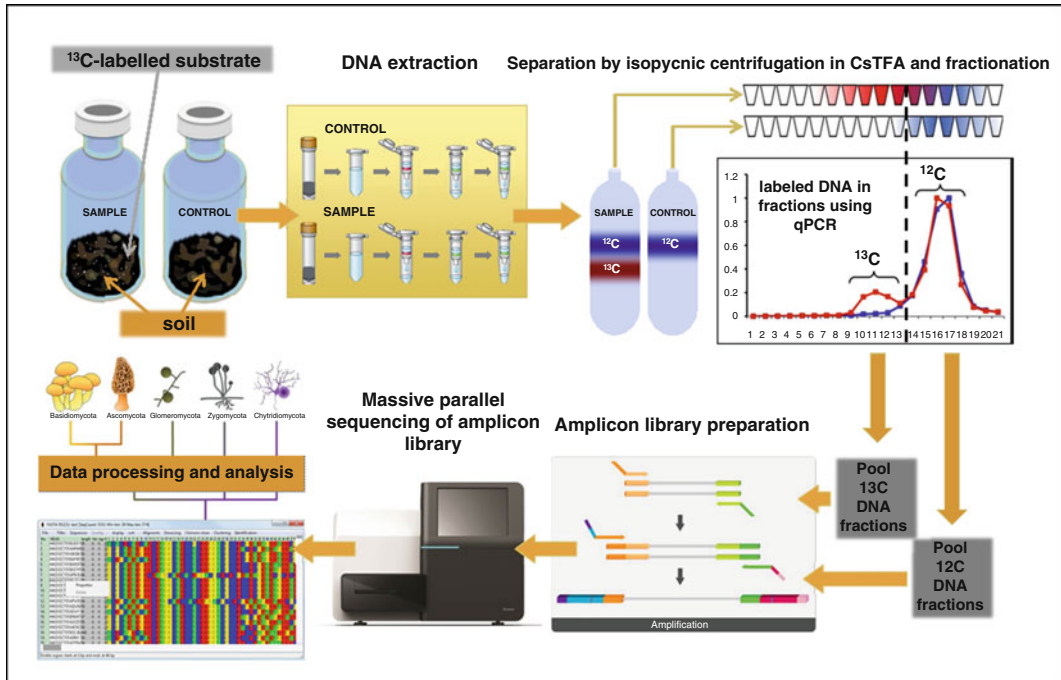


Fig. 1 Analysis of the members of fungal community involved in the decomposition of biopolymers using stable isotope probing

3.2 Gene-Targeted Metagenomics and Meta-transcriptomics to Identify Exocellulase Producers

1. In the area of study, select several study sites to represent the ecosystem. To ensure a representative sample at each sampling point, collect eight 5-cm-diameter soil cores around the circumference of a 4-m-diameter circle. In the field, separate the litter horizon material and soil material up to the required depth, and treat them separately. Remove larger roots from soil and sieve it through a 5-mm sterile mesh. Combine the eight subsamples and mix well. Cut the litter material into 0.5 cm pieces with sterile scissors. Combine the eight subsamples and mix well.
2. Prepare at least four aliquots of sample material for DNA/RNA co-extraction (0.5–3.0 g) in cryogenic vials. Freeze the aliquots in liquid nitrogen immediately and store them on dry ice. Upon arrival to the laboratory, store the samples frozen at $-80\text{ }^{\circ}\text{C}$ for no more than 6 months.
3. Co-extract RNA and DNA from each aliquot of each sample independently using the RNA PowerSoil Total RNA Isolation Kit and the DNA Elution Accessory Kit combined with the OneStep PCR Inhibitor Removal Kit (*see Note 12*, Fig. 2).
4. For RNA extraction, follow steps 1–8 of the RNA PowerSoil Total RNA Isolation Kit instructions.

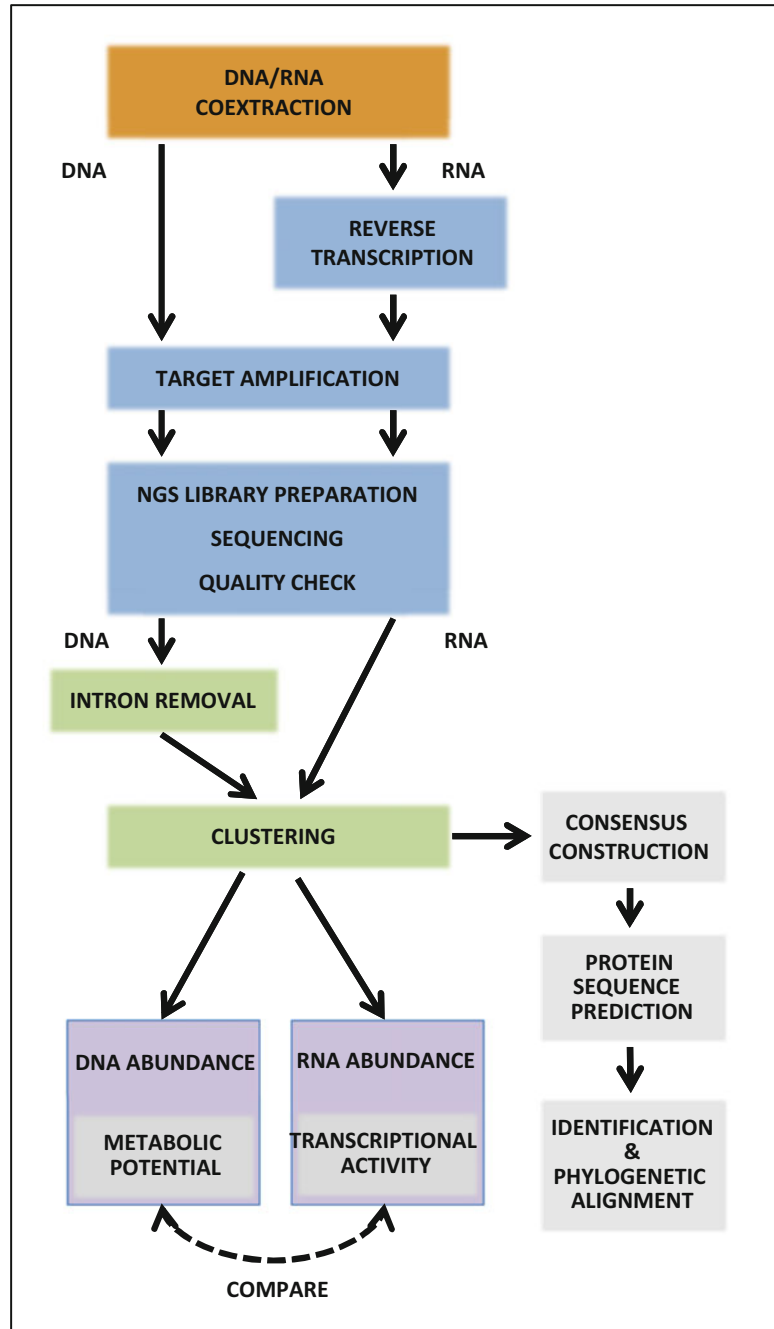


Fig. 2 Combination of gene-targeted metagenomics and metatranscriptomics to explore the potential and active decomposers of organic matter

5. Before proceeding to step 9 (the addition of the Solution SR4) use the OneStep PCR Inhibitor Removal Kit columns to clean the supernatant following the manufacturer's instructions. Use as many columns as needed to clean the whole volume of supernatant.

6. Collect the cleaned supernatant in a new 15 mL Collection Tube of the RNA PowerSoil Total RNA Isolation Kit. Follow the steps 9–20 of the RNA PowerSoil Total RNA Isolation Kit instructions.
7. To co-extract DNA, keep the columns from step 16 and use the DNA Elution Accessory Kit to elute DNA.
8. Store the recovered DNA at $-20\text{ }^{\circ}\text{C}$ and RNA at $-80\text{ }^{\circ}\text{C}$.
9. Treat 50 ng of each RNA sample with DNase I and perform reverse transcription using Superscript III Reverse Transcriptase and random hexamer primers according to the manufacturer's protocol to obtain single-stranded cDNA. Store the cDNA at $-20\text{ }^{\circ}\text{C}$.
10. Pool the aliquots of DNA or cDNA originating from the same soil sample before PCR amplification.
11. Perform PCR amplification of each DNA or cDNA samples with the primers cbhIF and cbhIR containing sample-specific barcodes. Set up the PCR reactions in at least three independent 50 μL reactions per sample containing 5 μL of 10 \times buffer for DyNAzyme DNA Polymerase, 3 μL of BSA (10 mg/mL), 2 μL of each primer (0.01 mM), 1 μL of PCR Nucleotide Mix (10 mM each), 1.5 μL polymerase (2 U/ μL , Pfu DNA Polymerase : DyNAzyme II DNA Polymerase mixed in a ratio 1:24), and 2 μL of template DNA or cDNA (*see Note 13*).
12. Run the PCR with the following cycling parameters: 94 $^{\circ}\text{C}$ for 3 min; 35 cycles of 94 $^{\circ}\text{C}$ for 30 s, 51 $^{\circ}\text{C}$ for 45 s, 72 $^{\circ}\text{C}$ for 1 min 30 s, followed by 72 $^{\circ}\text{C}$ for 15 min. Use heated cover option (103–105 $^{\circ}\text{C}$).
13. Pool the replicate PCR products and clean them using the MinElute Kit, elute DNA with 20 μL of sterile water. Measure the dsDNA concentrations in samples using Qubit and combine the barcoded samples so that the same amount of DNA is included from each sample.
14. To remove any short fragments that might interfere with high-throughput sequencing, clean the combined sample using the AMPure XP beads according to the manufacturer's protocol. Measure the dsDNA concentrations in composite samples using Qubit and use them to prepare a library for high-throughput sequencing.
15. Perform the high-throughput sequencing and sequence analysis as described in other chapters of this book or [11, 12].
16. When processing sequences after cleaning, separate the sequences from the DNA samples. For each sequence identify the positions of introns and remove introns from the sequences (Fig. 2).

17. Combine DNA sequences originating from RNA and those originating from DNA after intron removal and perform clustering at desired similarity level. For each OTU, record the abundance of sequences in the DNA and RNA-derived sequence pools. The sequences contained in the DNA-derived pool represent the *cbhI* genes present in the environment (in the genomic DNA of fungi) while those derived from the RNA-derived pool represent those sequences being transcribed by the members of the fungal community at the time of sampling [7].

4 Notes

1. The ideal setup would be to use ^{13}C litter from the dominant plant species of the considered experimental site/plot; however the *Zea* may cellulose is the most accessible substrate.
2. The amount of water to be added is theoretical. In practice, it is better to add less than suggested, check the buoyant density (BD), and add more water until the desired BD is reached. Check BD by weighing 100 μL on an analytical balance with a 0.1 mg resolution or using refractometer.
3. The amount of soil or litter to be collected depends on the number of intended replicates and sample moisture, e.g.: 500 g of soil for three replicates.
4. Incubation temperature should preferably reflect the temperature in the study area. When other temperature is used, there is a danger that selective growth of microorganisms with specific temperature optima occurs.
5. The times of sampling indicated here are optimal for the study of cellulose utilization in C-rich acidic forest soils. When other soils are studied or other substrates are used, it is advisable to perform a pilot test to determine optimal sample collection times. For the optimal results of SIP, the time of incubation should be sufficient for the ^{13}C -enrichment of microbial DNA but not longer than necessary as cross-labeling with ^{13}C may occur if the DNA of those microbes that originally fed on cellulose are used as food for other organisms (cross-feeders). Please note that the relative rate of utilization is substrate-dependent with the time of biomass labeling increasing with the decreasing decomposability of the substrate.
6. Analyze the CO_2 concentration and $^{12}\text{C}/^{13}\text{C}$ ratio in the stored bottles within 7 days using suitable equipment such as the Trace Gas system interfaced to an IsoPrime mass spectrophotometer.

7. Buoyant density of the CsTFA solution depends on temperature. The CsTFA should be stored at 4 °C.
8. One microgram of DNA is the minimum for separation. Three micrograms for DNA still separates well and the separated fractions after ultracentrifugation are easier to analyze by qPCR. Higher amounts of DNA should be avoided because these may fail to resolve during ultracentrifugation.
9. Please note that the pellet is often invisible due to low amount of DNA.
10. RT-PCR is highly sensitive, keep all solutions sterile and nuclease-free. Always set up reactions in a PCR hood; wipe down pipettes with ethanol before use, use new or aliquoted PCR water, etc.
11. To make sure that “light” and “heavy” DNA is identified properly, it is possible to plot DNA concentrations in fractions against fraction BD. When pooling fractions, avoid those where ^{12}C and ^{13}C -DNA may overlap. If no “heavy” DNA is observable, it is probable that ^{13}C -labeling was insufficient. In such a case, use longer incubation time before DNA extraction. The “heavy” DNA fraction contains DNA with sufficient ^{13}C -enrichment (at least 20 %) but it may also contain DNA with a specific sequence, such as the DNA with a high GC content. Please be sure to pool also the fractions of the control DNA that correspond to those observed as ^{13}C -enriched in the ^{13}C -supplemented sample. For more details, check [8].
12. To increase the yield of DNA and RNA, grinding of samples in liquid N_2 with a mortar and pestle should be applied. RNA extraction is extremely sensitive to contamination and nucleases. Please be sure to use nuclease-free water and plasticware throughout.
13. The use of the Pfu DNA Polymerase/DyNAZyme II DNA Polymerase mixture decreases the error level during the PCR reaction. As an alternative, another DNA polymerase with a proofreading activity can be used.

Acknowledgements

This work was supported by the Czech Science Foundation grant 13-06763S and by the Institute of Microbiology of the Czech Academy of Sciences (RVO61388971).

References

1. Berg B, Laskowski R (2006) Litter decomposition: a guide to carbon and nutrient turnover. Academic, Amsterdam
2. Clemmensen KE, Bahr A, Ovaskainen O, Dahlberg A, Ekblad A, Wallander H, Stenlid J, Finlay RD, Wardle DA, Lindahl BD (2013) Roots and associated fungi drive long-term carbon sequestration in boreal forest. *Science* 339:1615–1618
3. Baldrian P, Valášková V (2008) Degradation of cellulose by basidiomycetous fungi. *FEMS Microbiol Rev* 32:501–521
4. Berlemont R, Martiny AC (2013) Phylogenetic distribution of potential cellulases in bacteria. *Appl Environ Microbiol* 79:1545–1554
5. de Boer W, Folman LB, Summerbell RC, Boddy L (2005) Living in a fungal world: impact of fungi on soil bacterial niche development. *FEMS Microbiol Rev* 29:795–811
6. Baldrian P, López-Mondéjar R (2014) Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. *Appl Microbiol Biotechnol* 98:1531–1537
7. Baldrian P, Kolařík M, Štursová M, Kopecký J, Valášková V, Větrovský T, Žifčáková L, Šnajdr J, Rídl J, Vlček Č, Voříšková J (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *ISME J* 6:248–258
8. Edwards IP, Upchurch RA, Zak DR (2008) Isolation of fungal cellobiohydrolase I genes from sporocarps and forest soils by PCR. *Appl Environ Microbiol* 74:3481–3489
9. Štursová M, Žifčáková L, Leigh MB, Burgess R, Baldrian P (2012) Cellulose utilization in forest litter and soil: identification of bacterial and fungal decomposers. *FEMS Microbiol Ecol* 80:735–746
10. Šnajdr J, Dobiášová P, Větrovský T, Valášková V, Alawi A, Boddy L, Baldrian P (2011) Saprotrophic basidiomycete mycelia and their interspecific interactions affect the spatial distribution of extracellular enzymes in soil. *FEMS Microbiol Ecol* 78:80–90
11. Větrovský T, Baldrian P (2013) Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED. *Biol Fertil Soils* 49:1027–1037
12. Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjølter R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J, Kauserud H (2013) Fungal community analysis by high-throughput sequencing of amplified markers—a user's guide. *New Phytol* 199:288–299

DNA-Based Characterization and Identification of Arbuscular Mycorrhizal Fungi Species

Carolina Senés-Guerrero and Arthur Schüßler

Abstract

Arbuscular mycorrhizal fungi (AMF) are obligate symbionts of most land plants. They have great ecological and economic importance as they can improve plant nutrition, plant water supply, soil structure, and plant resistance to pathogens. We describe two approaches for the DNA-based characterization and identification of AMF, which both can be used for single fungal spores, soil, or roots samples and resolve closely related AMF species: (a) Sanger sequencing of a 1.5 kb extended rDNA-barcode from clone libraries, e.g., to characterize AMF isolates, and (b) high throughput 454 GS-FLX+ pyrosequencing of a 0.8 kb rDNA fragment, e.g., for in-field monitoring.

Key words 454 GS-FLX+ pyrosequencing, Arbuscular mycorrhizal fungi (AMF), DNA-based species identification, Evolutionary placement algorithm (EPA), Extended DNA barcoding, Nuclear rDNA, Phylogenetics

1 Introduction

Arbuscular mycorrhizal fungi (AMF) are asexual, clonal organisms that currently cannot be defined by a biological species concept [1]. Thus, AMF species were historically characterized by spore morphology. However, this can be misleading, e.g., because some species form several spore morphs, and the occurrence of spores, as resting stages, does not represent the active AMF community [2]. Moreover, from intraradical or extraradical hyphae AMF species can only be identified by molecular methods, a prerequisite being that the species to be identified are already characterized for the used molecular markers.

For molecular systematics and molecular ecological studies with a wide taxonomic coverage, the nuclear rRNA gene regions are the most frequently used markers targeting the small subunit (SSU) [3] and/or internal transcribed spacer (ITS) [4] and/or large subunit (LSU) [5] rDNA regions. Yet, due to the low variability in the SSU,

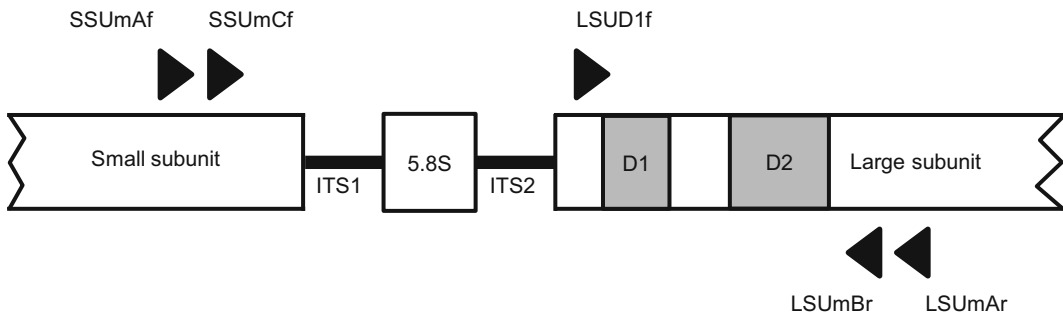


Fig. 1 Schematic representation of the nuclear ribosomal DNA regions studied (not to scale). *Triangles* indicate the position of the priming sites; primer names are shown

an extremely high intraspecific variability in the ITS region of up to >23 % [6, 7], or the use of relatively short LSU fragments, most analyses led to a phylogenetic resolution at an undefined taxonomic level above species [8]. An extended DNA barcoding region with resolution power also for closely related AMF species was suggested [6]; it comprises a part of the SSU, the complete ITS region (including the 5.8S rRNA gene), and approx. 0.8 kb of the LSU rRNA gene. It can be (nested) PCR amplified as a single ~1.5 kb fragment from spores, root fragments, and/or soil with AMF-specific primers [9]. These primers have the widest taxon coverage when compared to other commonly used ones targeting a single nuclear rDNA marker [10] and they can be used to amplify DNA from field samples [11]. A 0.8 kb fragment of this extended DNA barcode region can also be used for 454 GS-FLX+ pyrosequencing, also to analyze field samples (*see* Fig. 1). This fragment provides resolution of closely related species when using advanced methods based on an evolutionary placement algorithm (EPA) for data analyses and a solid and comprehensive reference sequence database. The presented method appears to be much more precise than other commonly used methods for monitoring AMF in the field by 454 GS-FLX+ sequencing [12].

2 Materials

The water and any other reagents used should be of molecular biological grade, unless stated otherwise.

2.1 Root Fragments

1. Entire root system.
2. Scalpel, tweezers, and a flame source.
3. 100 % ethanol.
4. Clean and sterile 2 mL microcentrifuge tubes.

**2.2 DNA Extraction:
FastDNA® SPIN Kit
for Soil or CTAB
Method (See Note 1)**

2.2.1 *FastDNA® SPIN Kit*
(MP Biomedicals,
Heidelberg, Germany)
for Soil

2.2.2 *Cetyl*
Trimethylammonium
Bromide (CTAB) Method

1. FastPrep® Instrument (MP Biomedicals, Heidelberg, Germany).
2. Lysing Matrix A tubes with an extra big bead instead of Lysing Matrix E.
3. Add the indicated volume of 100 % ethanol to the SEWS-M wash solution, mix, and store at room temperature.
4. Clean and sterile 2 mL microcentrifuge tubes and 15 mL Falcon tubes.

1. Tissue lyser.
2. 2× CTAB buffer: weigh 1 g of CTAB, 4.09 g of NaCl, and 0.5 g of polyvinylpyrrolidone (PVP). Add 5 mL of 1 M Tris-HCl, pH 8.0, and 4 mL of 0.25 M Na₂EDTA. Make up to 50 mL with water.
3. Autoclave the 2× CTAB buffer and add 1 mL of 99 % β-mercaptoethanol (*see Note 2*).
4. Prepare a 24:1 (v/v) chloroform/isoamylalcohol solution and store at room temperature.
5. Prepare 10 mg/mL RNaseA and store at -20 °C.
6. Prepare 50 mL of Tris-EDTA (TE) buffer: add 0.5 mL of 1 M Tris and 0.1 mL of 0.5 M EDTA. Make up to 50 mL with deionized water. Autoclave and adjust the pH to 8.0.
7. Clean and sterile 2 mL microcentrifuge tubes.

**2.3 PCR
Amplification**

2.3.1 *454 Sequencing*
PCR Amplification

1. Prepare primers and the AMF-specific mixture of primers (*see Table 1*). First, prepare primer stocks by diluting your primers to 100 μM. Prepare working solutions with a concentration of 10 μM for each primer. E.g. for the forward primer SSUMAf, mix 10 μL of each primer SSUMAf1-2 with 80 μL of water. For the reverse primer LSUMAr, mix 10 μL of each primer LSUMAr1-4 with 60 μL of water.
 2. Store primer stocks and working solutions at -20 °C.
 3. 2× Phusion® High-Fidelity DNA polymerase Master Mix HF (NEB, Frankfurt, Germany).
 4. 20 mg/mL Bovine Serum Albumin (BSA) (NEB, Frankfurt, Germany).
 5. Thermal cycler.
1. Prepare the LSU primers. We use the forward primer LSUD1f and the reverse primer mixture LSUMBr (*see Table 1*). The forward (fusion-) primer has to be synthesized together with the 454 adaptor A and different MIDs (5'-adaptorA-MID-LSUD1f-3'). The reverse primer is synthesized with the 454 adaptor B (5'-adaptorB-LSUMBr-3'). For details regarding

Table 1
Primers and primer mixtures

Step	Primer	Nucleotide sequence (5'–3')	Primer mixtures
Subheading 3.5, First PCR (1.8 kb)	SSUmAf1	TGGGTAATCTTTTGAAACTTYA	SSUmAf: mix SSUmAf1-2 (equimolar)
	SSUmAf2	TGGGTAATCTTRTGAAACTTCA	
	LSumAr1	GCTCACACTCAAATCTATCAAA	LSumAr: mix LSumAr1-4 (equimolar)
	LSumAr2	GCTCTAACTCAATTCTATCGAT	
	LSumAr3	TGCTCTTACTCAAATCTATCAAA	
LSumAr4	GCTCTTACTCAAACCTATCGA		
Subheading 3.6, Nested PCR (1.5 kb)	SSUmCf1	TCGCTCTTCAACGAGGAATC	SSUmCf: mix SSUmCf1-3 (equimolar)
	SSUmCf2	TATTGTTCTTCAACGAGGAATC	
	SSUmCf3	TATTGCTCTTNAACGAGGAATC	
	LSumBr1	DAACACTCGCATATATGTTAGA	LSumBr: mix LSumBr1-5 (equimolar)
	LSumBr2	AACACTCGCACACATGTTAGA	
	LSumBr3	AACACTCGCATAACATGTTAGA	
	LSumBr4	AAACACTCGCACATATGTTAGA	
LSumBr5	AACACTCGCATATATGCTAGA		
Subheading 3.9, Analyzing clones by PCR	M13F (-20)	GTAAAACGACGGCCAG	
	M13R	CAGGAAACAGCTATGAC	
	R377mod	CTCTCTTTTCAAAGTNCTTTTCATCT	
Subheading 3.15.1, 454 sequencing PCR (0.8 kb)	LSUD1f	TAAGCGGAGGAAAAGAAAMTAAC	
	LSumBr	Mixture of LSumBr1-5	

the fusion-primer adaptors and multiplex identifiers (MIDs), refer to the data provided by the supplier. Prepare stocks and working solutions of each primer as previously described.

2. Store primer stocks and working solutions at -20°C .

2.4 Cloning and Clone Analyses

1. Zero Blunt® TOPO® PCR Cloning Kit (Invitrogen, Darmstadt, Germany) using TOP10 chemically competent cells, according to the instructions of the supplier. However, we use only one third of the volumes and amounts of salt and plasmid vector and half of the chemically competent cells, to reduce costs.
2. Go Taq® DNA Polymerase (Promega, Mannheim, Germany) or the Go Taq® Green Master Mix (*see Note 3*).
3. Prepare M13F (-20) and M13R primer stocks (*see Table 1*) of 100 μM and working solutions of 10 μM , each. Store at -20°C .

4. Prepare LB agar: weigh 10 g of peptone, 5 g of yeast extract, 5 g of NaCl, and 12 g of agar. Add 1 L of deionized water and autoclave. Cool the LB agar until it reaches an approx. temperature of 55 °C and add 1 mL of 50 mg/mL kanamycin (Sigma-Aldrich, Munich, Germany) to obtain a final concentration of 50 µg/mL.
5. Pour the LB agar with kanamycin in Petri dishes (approx. 90 mm diameter). For some plates, draw a grid in the backside and write numbers in ascending order in each square of the grid. Store the plates at 4 °C.
6. Prepare LB medium with kanamycin as described before. Do not add agar.
7. For restriction fragment length polymorphism (RFLP), use three restriction enzymes: *Rsa I*, *Hinf I*, and *Mbo I* (NEB, Frankfurt, Germany).
8. For electrophoresis prepare 1 L of 50× Tris-acetate-EDTA (TAE) buffer: weigh 242 g of Tris base, 18.6 g of EDTA, and add 750 mL of deionized water to dissolve. Add 57.1 mL of glacial acetic acid. Make up to 1 L with deionized water and adjust the pH to 8.2. Dilute to 1× TAE buffer to prepare agarose gels and electrophoresis buffer.
9. Clean and sterile 200 µL tubes.
10. SOC media (20 g/L tryptone, 5 g/L yeast extract, 4.8 g/L MgSO₄, 3.6 g/L dextrose, 0.5 g/L NaCl, 0.19 g/L KCl; provided with the Zero Blunt® TOPO® PCR Cloning Kit).

2.5 Plasmid Preparation (“miniprep”)

1. NucleoSpin® Plasmid Kit (Macherey-Nagel, Düren, Germany) for single reactions or the NucleoSpin® 8 Plasmid Kit, requiring the NucleoVac 96 Vacuum Manifold and the Starter Kit A, for multiple reactions.
2. Dissolve RNase A with 1 mL of Buffer A1 (both provided in the kit) and vortex.
3. Add the indicated volume of 100 % ethanol to Buffer A4. Ethanol is not provided in the kit.

2.6 Sample Processing for 454 Sequencing (See Note 4)

1. Agencourt AMPure XP beads using solid phase reversible immobilization (SPRI) paramagnetic bead based technology (Beckman Coulter, Krefeld, Germany) to purify the amplicons.
2. PicoGreen dsDNA Assay Kit (Invitrogen, Darmstadt, Germany) to quantify the purified amplicons.
3. The GS FLX+ Titanium Sequencing Kit (Roche, Basel, Switzerland) applying the LongAmplicon3 processing pipeline which allows for 3' end trimming, recommended for processing long amplicon reads.

2.7 Basic Bioinformatic Analyses

1. Create an account at the CIPRES Science Gateway Portal (<https://www.phylo.org/portal2/>).
2. Install a program to proofread and edit sequences. We use SEQASSEM (www.sequentix.de).
3. Install a program to manually edit sequence alignments. We use ALIGN (www.sequentix.de).
4. Familiarize with the Basic Local Alignment Search Tool (BLAST) at the National Center for Biotechnology Information (NCBI) webpage (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).
5. Familiarize with MAFFT [13] (<http://mafft.cbrc.jp/alignment/software/>) to make multiple sequence alignments. It can be either downloaded or used through a web interface.
6. Install a program to visualize phylogenetic trees. We use FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

2.8 454 Sequencing Bioinformatic Analyses

1. Install the software Quantitative Insights Into Microbial Ecology (QIIME) [14] (<http://qiime.org/>).
2. Familiarize with QIIME. There are tutorials and explanatory documents found at the webpage.
3. Install the Graphical User Interface (GUI) of the RAxML Workbench in Linux (<http://sco.h-its.org/exelixis/web/software/epa/index.html>).
4. Install Archaeopteryx to visualize phylogenetic trees (<https://sites.google.com/site/cmzmasek/home/software/archaeopteryx>).

3 Methods

When working with small amounts of DNA (0.2–2 μ L) and using nested PCR protocols, it is necessary to work in a contamination-free environment. For example, use separate rooms for pre- and post-PCR steps and never expose the samples to an environment where, e.g., target DNA carrying plasmids were extracted or PCR products handled; use only clean molecular biology grade water and make sure that all solutions are prepared and kept uncontaminated; work under UV decontaminated sterile benches for initial sample preparation and in strictly UV decontaminated PCR cabinets; use clean pipettes and pipette tips that are only used for DNA extraction and separate pipette sets for the first PCR and the nested PCR, respectively, which are used only for this purpose and decontaminated from DNA regularly. An overview of the steps for the molecular characterization and identification of AMF by using Sanger- and/or 454 GS-FLX+ sequencing is shown in Fig. 2.

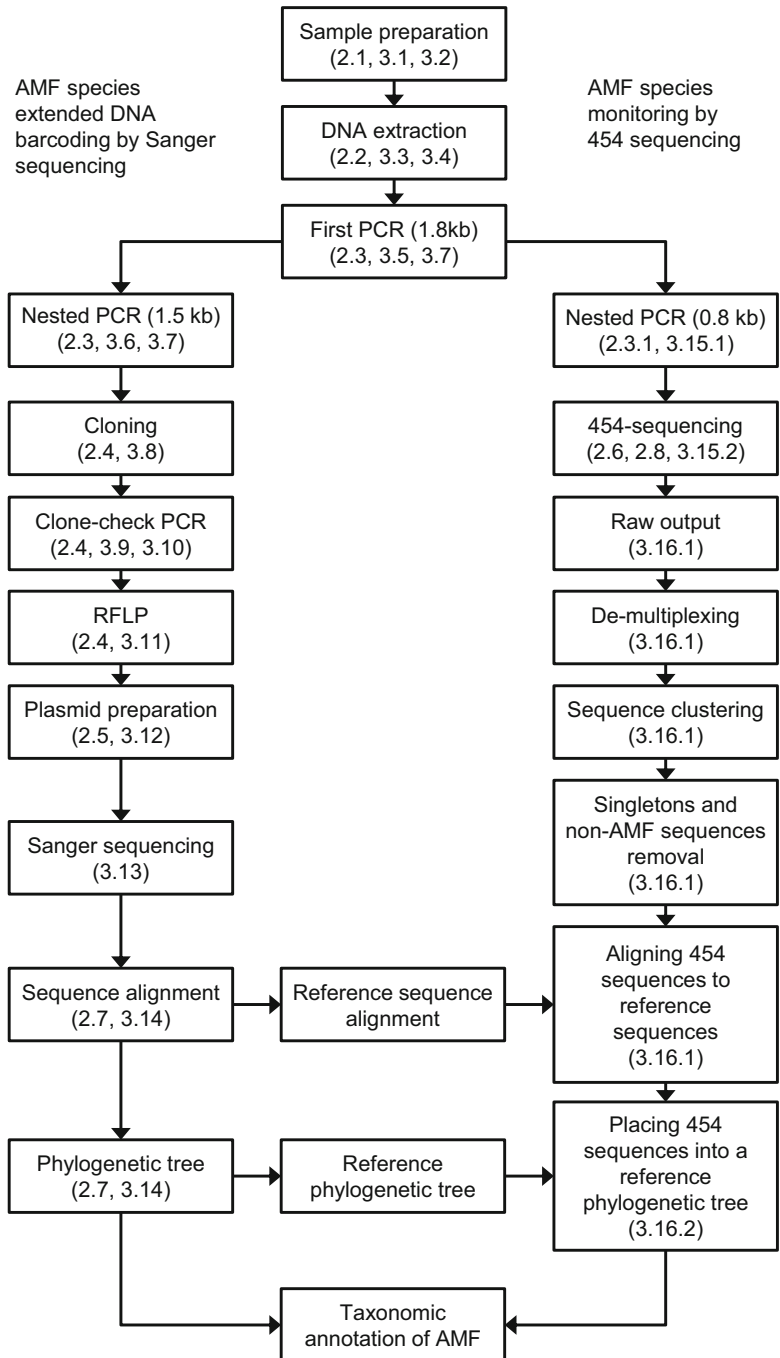


Fig. 2 Diagram of the steps for the molecular characterization and identification of AMF by using Sanger- and/or 454 GS-FLX+ sequencing

3.1 Processing Root Material

1. Obtain the entire root system of the plant and wash with tap water to remove adherent soil.
2. Cut root fragments by using a scalpel. To avoid contaminations, flame the blade or change it before cutting a new sample. Cut ten pieces of 0.5–1 cm length, depending on root thickness (0.5 cm is sufficient for thick and well-colonized roots). For weakly colonized roots, use 20 pieces of 1 cm. If a major goal is to obtain highly representative samples for a root system, the root material may be blended into small pieces and an amount no larger than 500 mg of material corresponding to a total of 5–20 cm root length can be taken, depending on thickness and colonization level.
3. To store samples, put them in ethanol (>80 % end concentration, consider the dilution effect by the samples and replace ethanol if there is too much dilution) in 2 mL microcentrifuge tubes. If possible, place samples at -20°C until DNA extraction (*see Note 5*).

3.2 Processing Root Material Stored in Ethanol

1. Wash the sample once with clean 100 % ethanol (*see Note 6*). Transfer the root fragments into a new 2 mL microcentrifuge tube.
2. Dry at 60°C in a clean environment (e.g., put open vials in a sterile, closed plastic bag such as a “sunbag”; Sigma-Aldrich, Munich, Germany). The time depends on root thickness and the amount of root pieces. It is important that there is no ethanol left in the samples.
3. Before DNA extraction, add 100 μL of water to the dried roots for 1 min (*see Note 7*).
4. Remove excess water with a clean pipette and proceed with DNA extraction.

3.3 DNA Extraction with the FastDNA[®] SPIN Kit for Soil

1. Add the processed root material to a Lysing Matrix A tube (in a contamination-free environment).
2. Add 978 μL of sodium phosphate buffer.
3. Add 122 μL of MT buffer.
4. Homogenize for 40 s at a speed setting of 6.0 using the FastPrep[®] Instrument. If roots are not completely disrupted, repeat the step. Then, while the instrument cools down, place the samples on ice or keep them at 4°C .
5. Centrifuge at $14,000\times g$ for 15 min to pellet debris (*see Note 8*).
6. Transfer the supernatant to a new 2 mL microcentrifuge tube. Add 250 μL of protein precipitation solution (PPS) and mix by inverting the tube 10 times.
7. Centrifuge for 5 min at $14,000\times g$ and transfer the supernatant to a clean 15 mL tube.

8. Resuspend the binding matrix suspension and add 1 mL to the supernatant in the 15 mL tube.
9. Invert by hand for 2 min. Place the tube in a rack until the silica matrix is precipitated (approx. 3 min).
10. Remove and discard 500 μ L of the supernatant avoiding the settled matrix.
11. Resuspend the binding matrix in the remaining amount of supernatant. Transfer 600 μ L of the mixture to a SPIN™ filter and centrifuge for 1 min at 14,000 $\times g$. Empty the catch tube and repeat this step until there is no binding matrix left in the 15 mL tube.
12. Add 500 μ L of prepared SEWS-M and resuspend the pellet by pipetting.
13. Centrifuge for 1 min at 14,000 $\times g$. Empty the catch tube and replace.
14. Without adding any liquid, centrifuge again for 2 min at 14,000 $\times g$ to dry the matrix. Replace the catch tube with a new one.
15. Air-dry the tube for 5 min at room temperature.
16. Add 50–80 μ L of DNase/pyrogen free water (DES) (*see Note 9*).
17. Centrifuge for 1 min at 14,000 $\times g$ to elute the DNA.
18. Discard the filter and store the DNA at 4 °C until use or at –20 °C for storage.

3.4 DNA Extraction with CTAB Method

1. Pre-cool the holders/adaptors of the tissue lyser by placing them in liquid N₂.
2. Add a single tungsten carbide bead (3 mm, DNA free) to the sample tube and freeze the tube with roots (water re-hydrated) in liquid N₂ for at least 30 s.
3. Disrupt the frozen samples, e.g., for 3 min at 30 Hz in a Tissue Lyser II bead mill (Qiagen, Leipzig, Germany). Repeat this step if needed until the result is a fine powder.
4. Add 1 mL of warm (60 °C) 2 \times CTAB buffer to the frozen fine powder and homogenize by vortexing (*see Note 10*).
5. Incubate 30 min at 60 °C.
6. Add one volume (1 mL) of 24:1 chloroform/isoamylalcohol and vortex.
7. Centrifuge for 5 min at 2500 $\times g$ and transfer the supernatant (aqueous upper layer, should be approx. 800 μ L) to a new 2 mL tube.
8. Add 2.5 μ L of 10 mg/mL RNaseA and incubate at 37 °C for 30 min.

9. Add one volume (approx. 800 μL) of 24:1 chloroform/isoamylalcohol and vortex.
10. Centrifuge for 5 min at $2500\times g$ and transfer the supernatant (aqueous upper layer, should be approx. 600 μL) to a new 2 mL tube.
11. Add 2/3 of the volume (approx. 400 μL) of isopropanol. Mix by inverting the tube 8 times and incubate at 4 °C for 15 min.
12. Centrifuge at $10,000\times g$ for 10 min and discard the supernatant.
13. Wash the pellet with 500 μL of 70 % ethanol (dilute with molecular grade water). Invert the tube once and centrifuge for 1 min at $2500\times g$. Discard the supernatant.
14. Air-dry the pellet for 15–30 min (in a clean, contamination-free environment).
15. Resuspend the pellet in 50 μL of water and dissolve at 4 °C overnight. TE buffer may also be used.
16. Store at 4 °C for direct use or freeze at -20 °C for later use.

3.5 First PCR

We use the 2 \times Phusion[®] High-Fidelity DNA polymerase Master Mix HF (NEB, Frankfurt, Germany). With the Phusion DNA polymerase relatively high melting temperatures and short elongation times are applied. If you intend to use *Taq* polymerase, several modifications have to be made, please refer to publications where this has been adopted. We, however, recommend using a proof-reading polymerase with DNA-binding domain such as the Phusion, as this reduces PCR errors and chimera formation.

1. Dependent on your sample number, you should prepare a master mix for all samples, e.g., for 15 samples use 165 μL master mix (15 times 15 μL plus 10 % as a buffer); calculate needed components accordingly. For an individual 15 μL PCR reaction, mix 7.5 μL of 2 \times Phusion[®] Master Mix, 0.75 μL of each, 10 μM forward and reverse primers (0.5 μM final concentration for each), 0.075 μL of 10 mg/mL BSA (final concentration of 50 $\mu\text{g}/\text{mL}$), 5.725 μL of water, and 0.2 μL of DNA (*see Note 11*).

3.6 Nested PCR

1. For an individual 20 μL PCR reaction, mix 10 μL of 2 \times Phusion[®] Master Mix, 1 μL of each, 10 μM forward and reverse primers (0.5 μM final concentration for each), 0.1 μL of 10 mg/mL BSA (final concentration of 50 $\mu\text{g}/\text{mL}$), 7.7 μL of water, and 0.2 μL of template from the first PCR.

3.7 PCR Conditions

The following is an example of the thermal cycling conditions to amplify an approx. 1.8 kb fragment. The annealing temperature in the protocol is tested for the AMF-specific primers we use; when using other primers you need to adjust the protocol accordingly.

1. Maintain your tubes always at cold temperature. Pre-heat the thermal cycler lid and place your tubes inside only when the program is about to start, to avoid unspecific reactions.
2. Run an initial denaturation step for 5 min at 99 °C.
3. For the first PCR, run 35 (30–40) cycles of 10 s of denaturation step at 99 °C, 30 s of annealing step at 60 °C, and 1 min of elongation step at 72 °C. For the nested 1.5 kb PCR, the annealing temperature is 63 °C and we usually run 30 cycles; other parameters are identical (*see Note 12*).
4. Run a final elongation step of 10 min at 72 °C.
5. Check that a 1.5 kb fragment from the nested PCR was amplified by electrophoresis in a 1 % agarose gel in 1× TAE buffer (*see Note 13*).

3.8 Cloning

1. Work under sterile conditions and maintain the cloning vector in a cold rack.
2. Set up the cloning reaction by adding 1 µL of PCR product into a 200 µL tube.
3. Add 0.3 µL of each salt, water, and plasmid vector (provided with the kit).
4. Spin the tubes and incubate at room temperature for 30 min.
5. From here onwards keep working under sterile conditions and place reactions on ice or in a cold rack.
6. Place the cells on ice and let them thaw for 2 min. Divide the 50 µL aliquot of cells into two aliquots of 25 µL in a clean, sterile 2 mL tube.
7. Add 2 µL of the cloning reaction, mix very gently by moving the pipette tip, do not pipette up and down to resuspend bacteria. Incubate for 10 min on ice.
8. Warm up a heat block to 42 °C and place your tubes for 30 s.
9. Place the tubes on ice.
10. Add 250 µL of SOC media (provided with the kit).
11. Incubate in a shaker for 1 h at 37 °C.
12. Plate 125 µL in each of two LB plates with kanamycin. Distribute evenly to obtain clone colonies that are separated one from another.
13. Incubate the plates overnight at 37 °C.

3.9 Analysis for Positive Clones by Using PCR

1. Identify and mark the desired number of clones on your LB plates.
2. For each clone to be analyzed prepare a 25 µL PCR reaction, composed of 5 µL of the 5× Green GoTaq® Reaction Buffer (included with the DNA GoTaq® Polymerase), 0.75 µL of

each of the 10 μM M13F (-20) and M13R primers, 0.5 μL of dNTPs, 0.125 μL of *Taq* DNA polymerase, and 17.88 μL of water. The reactions should be set up as a homogenous master mix from which 25 μL reaction aliquots are later pipetted into 200 μL tubes.

3. Pick the clone with a sterile pipette tip (*see* **Note 14**).
4. Place the tip inside the 200 μL tube, just so it touches the liquid, move slightly and very gently to wash bacteria into the solution.
5. Take the same tip and streak out bacteria on an LB plate with the drawn grid. Use one drawn square of the grid as a boundary to define where to streak one individual clone. Dispose the tip.
6. Repeat the previous step for every clone. Each clone is thus analyzed by a single PCR reaction and streaked out in an individual square of the grid on an LB plate.
7. Place the LB plate at 37 °C overnight. Afterwards store at 4 °C.

3.10 PCR Conditions to Analyze the Clones

1. Maintain your tubes always at a cold temperature. Pre-heat the thermal cycler lid and place your tubes inside only when the program is about to start.
2. Run an initial denaturation step for 5 min at 95 °C.
3. Run 35 cycles of 30 s of denaturation step at 95 °C, 30 s of annealing step at 65 °C, and 1 min of elongation step at 72 °C.
4. Run a final elongation step of 10 min at 72 °C.
5. Check that a 1.5 kb fragment was amplified by electrophoresis of a 1 % agarose gel in 1 \times TAE buffer.

3.11 Restriction Fragment Length Polymorphism (RFLP)

1. Prepare reactions for every positive clone using three restriction enzymes (*Rsa I*, *Hinf I*, and *Mbo I*). The reactions have to be done each in a separate 200 μL tube.
2. Add 3.9 μL of water, 1 μL of buffer, 0.1 μL of the enzyme, and 5 μL of the clone-check PCR product.
3. Spin and incubate for 1 h at 37 °C.
4. Run a 1.5 % agarose gel in 1 \times TAE buffer. The full amount (10 μL) of the RFLP reaction should be analyzed on the gel.
5. Compare the RFLP patterns and select clones that differ among each other. This is to obtain different sequence variants from a sample.
6. Pick the clones that correspond to the different RFLP patterns from the grid LB plate and grow each clone in a 15 mL tube containing 2 mL of LB medium with kanamycin.
7. Incubate the tubes in a shaker at 37 °C overnight.

**3.12 Plasmid
Preparation
("miniprep")
with NucleoSpin® 8
Plasmid and Vacuum**

1. Prepare column holders: insert NucleoSpin® Plasmid Binding Strips in the first column holder and NucleoSpin® Plasmid Filter Strips into the second column holder. Close unused wells of each column holder with NucleoSpin® Dummy Strips.
2. Centrifuge the bacteria cultures that were grown overnight in Subheading 3.11 for 10 min at $1000\times g$ and discard the supernatant.
3. Add 250 μL of Buffer A1 with RNase A to each sample. Resuspend the bacteria pellet completely by vortexing.
4. Add 250 μL of Buffer A2 and mix by inverting the tube 10 times. Incubate at room temperature for a maximum of 5 min.
5. Add 350 μL of Buffer A3 and mix by inverting the tube 10 times.
6. Transfer the crude lysates completely into the wells of the NucleoSpin® Plasmid Filter Strips.
7. Apply vacuum (-0.4 bar) for 1–5 min. Release the vacuum when the crude lysate has passed completely through the filter strips.
8. Remove and discard the NucleoSpin® Plasmid Filter Strips. Open the manifold lid. Remove the Column Holder A with the NucleoSpin® Plasmid Binding Strips with cleared lysates.
9. Insert the MN Wash Plate on the spacers inside the manifold base. Close the manifold base with the manifold lid. Place the column holder with the binding strips on top of the manifold.
10. Apply vacuum (-0.4 bar) for 1 min. Release the vacuum when the cleared lysate has drained off.
11. Add 600 μL of Buffer AW to each well of the binding strips. Apply vacuum (-0.4 bar) for 1 min.
12. Add 900 μL of Buffer A4 to each well. Apply vacuum (-0.4 bar) for 1 min. Repeat this step once.
13. Remove the Column Holder A with the inserted NucleoSpin® Plasmid Binding Strips. Remove the manifold lid, MN Wash Plate, and waste container from the vacuum manifold.
14. Remove any residual wash buffer from the binding strips by placing the outlets of the strips in soft tissue until no further drops come out.
15. Close the manifold base with the manifold lid. Place the column holder with the binding strips on top of the manifold. Apply vacuum (-0.4 bar) for 15 min until the membrane is completely dry.
16. Remove the manifold lid with the Column Holder A from the vacuum manifold. Insert spacers "Microtube rack," notched side up, into the grooves located at the short sides of the vacuum manifold.

17. Insert the rack of tube strips on the spacers inside the manifold base and reinstall the vacuum manifold as previously described.
18. Elute DNA by adding 100 μ L of sterile distilled water to each well of the binding strips. Incubate at room temperature for 3 min. Apply vacuum (-0.4 bar) for 1 min.
19. Remove the rack of tube strips, seal with cap strips, and store at -20 °C.

3.13 Sequencing

1. Sequence your reactions according to your sequencing service provider (we do not perform the cycle-sequencing ourselves). The amount of template and primer concentration will depend on the sequence provider.
2. We use three sequencing reactions to sequence the full 1.5 kb fragment: besides the primers M13F (-20) and M13R we use a modified version of the reverse primer R377 to cover the entire fragment with sufficient overlap of the sequencing reads. If needed, you may use additional primers to completely sequence both strands of the template.

3.14 Bioinformatics for Sanger Sequence Analyses

3.14.1 Basic Analysis

1. We use SEQASSEM (www.sequentix.de) to assemble and proofread the sequences, but any other similar programs can be used.
2. Assemble your sequences and arrange them in 5'–3' (forward) orientation.
3. Proofread the chromatograms for ambiguous bases, which may be facilitated by tools your software offers (e.g., “refine mode” in SEQASSEM).
4. Remove the primer-binding site sequences and assemble the consensus sequence.
5. Export the consensus sequences as a FASTA file.
6. BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) the sequences against a nucleotide sequence database (e.g., NCBI). To interpret similar sequences, take into account the length of the resulting sequences, the E-value, and the percentage of similarity. Sequences can be roughly annotated based on such BLAST results; however be aware that many database sequences are not correctly annotated. Therefore, consider basing your annotations on reference sequences from defined isolates. To annotate sequences to the species level, computing a phylogenetic tree based on reference sequences is necessary.

3.14.2 Phylogenetic Analysis

1. Obtain 1.5 kb SSU-ITS-LSU reference sequences from a sequence database, such as NCBI. For published datasets on AMF sequences you may use as a baseline the sequences analyzed in [15], available from www.amf-phylogeny.com, but note that recently published sequences should be implemented

to obtain an up to date reference sequences dataset. If there is no indication on the identity of the sample, use sequences from species of all AMF genera for your analysis. If you have such indication, e.g., after BLAST, and you are only interested in the placement within the group of close relatives, you should construct the tree based on sequences from species of the same genus or family only.

2. Align your sequences together with the reference sequences. We use MAFFT (<http://mafft.cbrc.jp/alignment/software/>). We usually select the alignment strategy based on the amount of sequences, while prioritizing the accuracy of the results in the settings of the program.
3. If you already have an alignment and you want to introduce new and unaligned sequences, you should use the “--add” command in MAFFT, which allows aligning individual sequences against a fixed alignment (see http://mafft.cbrc.jp/alignment/server/add_sequences.html).
4. Manually check and optimize the alignment of your sequences. We use ALIGN (www.sequentix.de) but any similar program can be used (see **Note 15**).
5. Save your alignment file as a relaxed PHYLIP format and upload it to the CIPRES Science Gateway (<https://www.phylo.org/portal2/>) in your data folder.
6. Go to your tasks folder and create a new task. Select your alignment file as input data and select RAxML-HPC2 on XSEDE as a tool. You can adjust the parameters depending on your data (see the advanced help section when setting up your task parameters). We use the option GTR+ optimization of substitution rates + GAMMA model of rate heterogeneity (alpha parameter will be estimated) for bootstrapping and final tree evaluation.
7. When RAxML is finished, download the file RAxML_bipartitions.result.
8. Change the name of the RAxML_bipartitions.result to RAxML_bipartitions.tre (or chose a *.tre filename suiting your data analysis pipeline) and open the Newick format tree file in your preferred software for tree visualization, e.g., FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>). To facilitate tree visualization and posterior editing, we normally replace the sequence annotations (usually the accession or clone number) in the RAxML_bipartitions.tre file with the desired information (usually species, isolate identifier, sampling site, etc.), which are implemented in our alignment file by using a batch replacement tool available in ALIGN. You may also use other programs for annotation of taxa in a tree file.

9. Once you open your tree, select a lineage to root the tree. When studying non-paraglomeralean AMF we normally use members of the *Paraglomerales* as outgroup, as this order is likely the most ancestral currently known lineage in the AM fungi [15]. However, if you analyze only a certain family or genus, you should use outgroups that are more closely related to that ingroup.
10. We usually edit the text labels, fonts, line thickness, and the distance bar in MS Power Point after exporting the tree file as a vector graph (e.g., as an *.emf file) from FigTree.

**3.15 Methods
for 454 GS-FLX+
Based High
Throughput
Monitoring of AMF**

We use an “evolutionary placement algorithm” (EPA) for the assignment of short reads to the edges of a reference phylogenetic tree under the maximum-likelihood model. EPA is integrated in RAxML, a frequently used and fast maximum-likelihood program [16, 17]. Briefly, we perform two steps. In the first, a reference phylogenetic tree based on 1.5 kb sequences is computed with RAxML in the CIPRES web portal (<https://www.phylo.org/portal2/>) and in the second, reference sequences of a 98 % similarity cluster are individually placed into this reference tree and annotated to species by using EPA through a web server (<http://epa.hits.org/raxml>) (see **Note 16**).

**3.15.1 454 Sequencing
PCR Amplification**

The first PCR is performed as described in Subheading 3.5. A 1.8 kb fragment is amplified and used as template for amplifying a nested PCR product that will be pyrosequenced.

The protocol below is adjusted for the LSU primers used by [12] to amplify a 0.8 kb LSU rRNA gene fragment; if using other primers, the protocol needs to be adjusted accordingly.

1. Maintain your tubes always at cold temperature. Pre-heat the thermal cycler lid and place your tubes inside only when the program is about to start, to avoid unspecific reactions.
2. Run an initial denaturation step for 5 min at 99 °C.
3. Run 25 cycles of 10 s of denaturation step at 99 °C, 30 s of annealing step at 63 °C, and 30 s of elongation step at 72 °C. The cycle numbers can be increased to 30 or 35 cycles, if no product was visible after 25 cycles.
4. Run a final elongation step of 10 min at 72 °C.
5. Check that an approx. 0.8 kb fragment was amplified by electrophoresis of a 1 % agarose gel in 1× TAE buffer.

**3.15.2 Processing
of PCR Amplicons for 454
GS-FLX+ Pyrosequencing**

In our pipeline, we rely on a commercial service to perform 454 GS-FLX+ sequencing of PCR amplicons. In brief, PCR samples are handed over to IMG Laboratory (Martinsried, Germany) where each amplicon is separately purified using the Agencourt AMPure

XP beads and quantified using PicoGreen dsDNA Assay Kit. Libraries are generated containing the amplicon samples (pooled equimolarly) each coded with different MIDs. Each library is purified three times applying two different methods. First, a gel extraction, followed by a size selection step (>250 bp) using the Agencourt AMPure XP beads, is performed twice. Sequencing is done using the GS FLX+ Titanium Sequencing Kit (Roche, Basel, Switzerland). Image and signal raw pyrosequencing data are processed by the Roche 454 GS-FLX+ inherent software packages applying the LongAmplicon3 processing pipeline which allows for 3' end trimming, recommended for processing long amplicon reads.

3.16 Bioinformatics Pipeline for Analyzing 454 Sequence Reads

As a baseline dataset the alignment described in Subheading 3.14 is used to construct a “phylogenetic backbone” maximum-likelihood phylogenetic tree. This tree later serves to affiliate the 0.8 kb 454-sequences. The tree is computed as described in Subheading 3.14.2 at the CIPRES web portal (<https://www.phylo.org/portal2/>). Download the RAxML_bestTree.result file, not the “bipartition file,” for later input into EPA (<http://epa.hits.org/raxml>). For this section, detailed examples of command lines are given in [Appendix](#).

3.16.1 Preparing the 454 Sequences

1. We use QIIME (<http://qiime.org/>) to reduce the 454-sequence dataset from the raw output and to obtain representative sequences, after a 98 % clustering step. We do not use a 97 % clustering as such clusters may contain sequences from different closely related species. QIIME may also be used for other steps, from processing raw sequences to obtaining a phylogenetic tree; however we use different programs along our pipeline.
2. QIIME already has detailed instructions in the 454 overview tutorial (<http://qiime.org/tutorials/tutorial.html>); therefore we only describe our pipeline in a general manner. For more detail information, refer to the QIIME tutorial.
3. Create your mapping file. This is a text file that contains basic information about the samples.
4. De-multiplex. In this step you assign the samples to the reads.
5. Combine your de-multiplexed sequences in a single file.
6. Cluster your sequences.
7. Remove singletons.
8. Remove non-AMF sequences.
9. Obtain a tabulator delimited table with information about the samples, AMF “OTUs” (98 % similarity clusters), and read amounts. This file will be used later in Subheading 3.16.2.

10. Prepare a FASTA file without singletons and without non-AMF sequences. QIIME provides means to do further analyses; however we continue with the aligning step using MAFFT.
11. Align the 454 sequences to the reference sequence alignment. We do this in MAFFT by using the `--add` command previously described in Subheading 3.14.2. Save the file containing your aligned reference sequences and the 454 sequences as a relaxed PHYLIP format file.

3.16.2 *Evolutionary Placement Algorithm for Affiliation of Sequences and Their Annotation to AMF Species*

1. Before running EPA, make sure that the reference sequence alignment does not contain undetermined values and sequences that are equal (*see Note 17*). There is an option in EPA to upload unaligned 454 sequences, but we prefer to align them and check the alignment before submission.
2. Check that none of the sequences names is repeated, including the “OTUs” (98 % similarity clusters) as identical names cannot be processed by the software.
3. Upload your alignment with both, reference sequences and the “OTUs” (98 % similarity clusters) along with the RAxML_bestTree.result file to EPA (<http://epa.h-its.org/raxml>). The reference sequence alignment has to correspond to the reference tree entirely, which means that corresponding sequence names in the alignment and in the tree have to be identical and every reference sequence present in the alignment has to be found in the reference phylogenetic tree and vice versa. Any deviation will lead to termination of the EPA analysis.
4. The EPA analysis results in several output files. Download them in a separate folder. We take the RAxML_classification and the original_Labeled tree files to generate a phyloXML file by using the phyloXML converter in the GUI (graphical user interface) of the RAxML Workbench in Linux. With this file you can visualize your reference phylogenetic tree and the branches in which the 454 sequences have been placed.
5. Open your phyloXML file in Archaeopteryx (<https://sites.google.com/site/cmzmasek/home/software/archaeopteryx>).
6. Open the RAxML_classification file in Excel. This file contains the names of your 454 sequences and the optimal reference tree branch to which the 454 sequences were inserted. You will see all the information together in a single column, split the information into three columns, one that contains the name of your sequence, one that contains the branch name, and one that contains the weight value (*see Note 18*). Sort the column of branch names alphabetically.
7. In Archaeopteryx find the branch name from the Excel document. Once you located the branch name, you can make a taxonomic annotation based on the location of the branch in the phylogenetic tree.

8. Make a MS Excel spreadsheet where you annotate each of the “OTUs” (98 % similarity clusters) to their related taxonomy, make a column for the name and a column for the annotation. We find it easier to make a list of “species,” so we can just give a species number to each 454 sequence. Dependent on the comprehensiveness of the baseline dataset and, e.g., the ecosystem studied, EPA cannot annotate all of the sequences to branches that correspond to the species level; some can only be annotated to genus or even family, at deeper nodes in the reference phylogenetic tree (*see* **Note 19**).
9. Once all data are annotated, open the tabulator delimited table file that contains information about the samples, “OTUs,” and reads (from Subheading 3.16.1, **step 9** and **Appendix, step 6**) in Excel. In this file you have to replace the “OTUs,” which are your 98 % representative 454 sequences, with your species annotations. To do this, we sort the 454 sequences names from both documents alphabetically. The order within this column has to be identical in both documents. Then just copy the species numbers from your taxonomic annotations document and paste it on top of the 454 sequences names. You will now have a new table giving species names, sample information, and read amounts.

In general, be aware that there are other ways of running the RAxML-EPA pipeline and also visualizing the tree results, e.g., running RAxML-EPA at your computer or using the .jplace output file to visualize the tree. Check the Google support group for both EPA and RAxML (<https://groups.google.com/forum/#!forum/raxml>) which cover all these related topics.

4 Notes

1. For DNA extraction from roots, we routinely use the FastDNA[®] SPIN Kit for soil (MP Biomedicals, Heidelberg, Germany) with Lysing Matrix A tubes with an extra big bead instead of Lysing Matrix E, because this modification performed better for weakly colonized roots. A FastPrep[®] Instrument (MP Biomedicals, Heidelberg, Germany) is needed to disrupt the roots. Alternatively, a cetyl trimethylammonium bromide (CTAB) protocol can be employed to extract DNA, by using a tissue lyser to disrupt the roots.
2. Add 1–2 % (w/v) polyvinylpyrrolidone (PVPP) to the 2× CTAB buffer if polyphenolic compounds have to be removed. Hydrate the added PVPP in the 2× buffer at least for 2 h by using a magnetic stirrer, before using the buffer for DNA extraction.
3. If using the Go Taq[®] DNA Polymerase, you have to additionally add a solution mix of dNTP (NEB, Frankfurt, Germany)

in your PCR reaction. In the Go Taq[®] Green Master Mix, the dNTPs are included.

4. It is important to pre-arrange a 454 sequencing strategy with the company or sequencing facility that will carry out the 454 run. We outsource these steps to a company, but the following is an example of material that is needed for our sequencing strategy.
5. Label your vials with ethanol-resistant markers and use vials with tight screw lids, if samples are to be transported, e.g., after remote field samplings, or over larger distances (leaking of 1 vial can spoil all sample labelings!).
6. If the storage ethanol is heavily colored, the material should be washed with fresh 100 % ethanol twice.
7. The re-hydrated root samples can be stored at -20°C until further usage.
8. Even when extending the homogenization time, there are always small amounts of leftover debris.
9. We do not resuspend the binding matrix because a lot of the material may get trapped in the pipette tip.
10. Do not use more than 75 mg tissue (fresh weight) per 750 μL 2 \times CTAB extraction buffer to avoid overloading the buffer.
11. Higher template concentrations can cause PCR inhibition but may be used if the DNA concentration is too low.
12. The Phusion DNA polymerase amplifies 1 kb in 30 s even from complex DNA templates.
13. Frequently, the DNA concentration in the PCR product from the first PCR is too low to be visible after agarose gel electrophoresis; therefore we usually only run a gel for the nested PCR.
14. We usually do not use wooden toothpicks because they may absorb a significant amount of liquid.
15. The ITS1 and ITS2 regions are difficult to align by standard programs. These regions are highly variable and it is recommended to check and optimize automatic alignments manually to avoid misalignments.
16. For more details visit the webpage of the Exelixis Lab, where RAxML, EPA, and other tools were developed (<http://sco.h-its.org/exelixis/web/software/epa/index.html>).
17. When you run a phylogenetic tree with RAxML, a file named infile.reduced without undetermined values and identical sequences is automatically created.
18. In addition you can see the other likelihood weight values of branch placements in the RAxML_classificationLikelihoodWeights file.

19. It is an advantage of EPA over, e.g., similarity-based methods like BLAST that it can give such placements also to nonterminal nodes, together with likelihood values for the placements.

Acknowledgements

This work was supported by the European Community's Seventh Framework Programme FP7/2007 under grant agreement no. 227522.

5 Appendix: Examples of Command Lines for 454 Sequencing Data Analysis

Bioinformatics pipeline for analyzing 454 sequence reads

1. De-multiplex.

For example, when sequencing a full 454-plate split into four gaskets (physically separated compartments), we use the following command to de-multiplex the first gasket:

```
split_libraries.py -m Mapping1.txt -f 1.TCA.454Reads.fna -q
1.TCA.454Reads.qual -l 500 -o split_Library_Run1_Output/
-n 1000000
```

and this command to de-multiplex the second gasket:

```
split_libraries.py -m Mapping2.txt -f 2.TCA.454Reads.fna -q
2.TCA.454Reads.qual -l 500 -o split_Library_Run2_Output/
-n 2000000
```

Consider that the parameters (such as sequence length) can be modified according to your needs; in the previous example we set the minimum length of sequences to be implemented in the clustering to 500 bp.

2. Combine your de-multiplexed sequences in a single file:

```
cat split_Library_Run1_Output/seqs.fna split_Library_Run2_
Output/seqs.fna > Combined_seqs.fna
```

3. Cluster your sequences.

First you have to prepare a text file containing the parameters of the clustering. We use the following parameters and save the text file as parameters.txt:

```
pick_otus:otu_picking_method uclust
pick_otus:similarity 0.98
pick_otus:enable_rev_strand_match True
```

We afterwards perform the clustering by using the following command:

```
pick_de_novo_otus.py -i combined_seqs.fna -p parameters.txt
-o uclust_picked_otus/
```

4. Remove singletons.

After clustering, you obtain a biom table with your “OTUs” (representative sequences of 98 % similarity clusters).

First remove the singletons (sequences represented only once) from the table:

```
filter_otus_from_otu_table.py -i otu_table.biom -o otu_table_no_singletons.biom -n2
```

Afterwards remove singletons from the fasta file:

```
filter_fasta.py -f combined_seqs.fasta -o biom_filtered_seqs.fasta -b otu_table_no_singletons.biom
```

5. Remove non-AMF sequences.

The previously created file “biom_filtered_seqs.fasta” contains your combined sequences without singletons. However, it still contains non-AMF sequences which in most cases have to be removed before further analysis.

To remove these sequences we use Blast2GO (<https://www.blast2go.com/b2ghome>) which takes individual sequences and finds similar sequences in NCBI.

The output of Blast2GO is an Excel-format file with the hits of your query sequences. We normally order the hits alphabetically and simply manually delete the non-AMF rows from the Excel table. After deleting the non-AMF rows, copy the remaining names of sequences, which will be kept for further analysis, and paste them into a text file. Name the text file as seqs_to_keep.txt (or according to your naming system).

To remove the non-AMF sequences from the FASTA file write in QIIME:

```
filter_fasta.py -f seqs_no_singletons.fasta -o seqs_no_cont.fasta -s seqs_to_keep.txt
```

To remove the non-AMF sequences from the OTU table write:

```
filter_otus_from_otu_table.py -i otu_table_no_singletons.biom -o otu_table_nosingletons_nocontaminants.biom -e seqs_to_keep.txt --negate_ids_to_exclude
```

6. Convert the file otu_table_nosingletons_nocontaminants.biom into a tabulator delimited table:

```
convert_biom.py -i otu_table_nosingletons_nocontaminants.biom -o otu_table_nosingletons_nocontaminants.txt -b
```

This table contains information about the samples, AMF “OTUs” (98 % similarity clusters), and read amounts.

References

- Schüßler A, Walker C (2011) Evolution of the ‘plant-symbiotic’ fungal phylum, *Glomeromycota*. In: Pöggeler S, Wöstemeyer J (eds) Evolution of fungi and fungal-like organisms, vol XIV, The Mycota. Springer, Berlin Heidelberg, pp 163–185
- Hempel S, Renker C, Buscot F (2007) Differences in the species composition of arbuscular mycorrhizal fungi in spore, root and soil communities in a grassland ecosystem. *Environ Microbiol* 9:1930–1938

3. Lee J, Lee S, Young JPW (2008) Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiol Ecol* 65:339–349
4. Redecker D (2000) Specific PCR primers to identify arbuscular mycorrhizal fungi within colonized roots. *Mycorrhiza* 10:73–80
5. Mummey DL, Rillig MC (2007) Evaluation of LSU rRNA-gene PCR primers for analysis of arbuscular mycorrhizal fungal communities via terminal restriction fragment length polymorphism analysis. *J Microbiol Methods* 70:200–204
6. Stockinger H, Krüger M, Schüßler A (2010) DNA barcoding of arbuscular mycorrhizal fungi. *New Phytol* 187:461–474
7. Krüger C, Walker C, Schüßler A (2014) *Scutellospora savannicola*: redescription, epitypification, DNA barcoding and transfer to *Dentiscutata*. *Mycol Prog* 13:1165–1178
8. Stockinger H, Walker C, Schüßler A (2009) ‘*Glomus intraradices* DAOM197198’, a model fungus in arbuscular mycorrhiza research, is not *Glomus intraradices*. *New Phytol* 183:1176–1187
9. Krüger M, Stockinger H, Krüger C et al (2009) DNA-based species level detection of *Glomeromycota*: one PCR primer set for all arbuscular mycorrhizal fungi. *New Phytol* 183:212–223
10. Kohout P, Sudová R, Janoušková M et al (2014) Comparison of commonly used primer sets for evaluating arbuscular mycorrhizal fungal communities: is there a universal solution? *Soil Biol Biochem* 68:482–493
11. Senés-Guerrero C, Torres-Cortés G, Pfeiffer S et al (2014) Potato-associated arbuscular mycorrhizal fungal communities in the Peruvian Andes. *Mycorrhiza* 24:405–417
12. Senés-Guerrero C, Schüßler A (2015) A conserved arbuscular mycorrhizal fungal core-species community structure in potato roots from the Andes. *Fungal Divers* (in press): online first, DOI: 10.1007/s13225-015-0328-7
13. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acids Res* 30:3059–3066
14. Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336
15. Krüger M, Krüger C, Walker C et al (2012) Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. *New Phytol* 193:970–984
16. Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum-likelihood. *Systematic Biol* 60:291–302
17. Berger SA, Stamatakis A (2011) Aligning short reads to reference alignments and trees. *Bioinformatics* 27:2068–2075

Molecular Identification of Soil Eukaryotes and Focused Approaches Targeting Protist and Faunal Groups Using High-Throughput Metabarcoding

G. Arjen de Groot, Ivo Laros, and Stefan Geisen

Abstract

While until recently the application of high-throughput sequencing approaches has mostly been restricted to bacteria and fungi, these methods have now also become available to less often studied (eukaryotic) groups, such as fauna and protists. Such approaches allow routine diversity screening for large numbers of samples via DNA metabarcoding. Given the enormous taxonomic diversity within the eukaryote tree of life, metabarcoding approaches targeting a single specific DNA region do not allow to discriminate members of all eukaryote clades at high taxonomic resolution. Here, we report on protocols that enable studying the diversity of soil eukaryotes and, at high taxonomic resolution, of individual faunal and protist groups therein using a tiered approach: first, the use of a general eukaryotic primer set targeting a wide range of eukaryotes provides a rough impression on the entire diversity of protists and faunal groups. Second, more focused approaches enable deciphering subsets of soil eukaryotes in higher taxonomic detail. We provide primers and protocols for two examples: soil microarthropods and cercozoan protists.

Key words 454 Metabarcoding, High-throughput sequencing, Soil metazoa, Soil protists, Soil microarthropods, 18S rDNA, CO1

1 Introduction

Little is known about the distribution of the entity of soil eukaryotes across landscapes and ecosystems, largely owing to the enormous diversity of those largely tiny organisms. With the advent of molecular techniques, especially high-throughput sequencing technologies, fungi have widely been targeted due to their high abundance in soils and ease of extracting molecular marker molecules [1, 2].

Among fauna, only representatives of the macrofauna, especially earthworms have been well studied. Mesofauna (e.g., nematodes (Nematoda), potworms (Enchytraeidae) and microarthropods such as springtails (Collembola) and mites (Acari)) and microfauna (e.g., rotifers) have received little attention, although some attention has

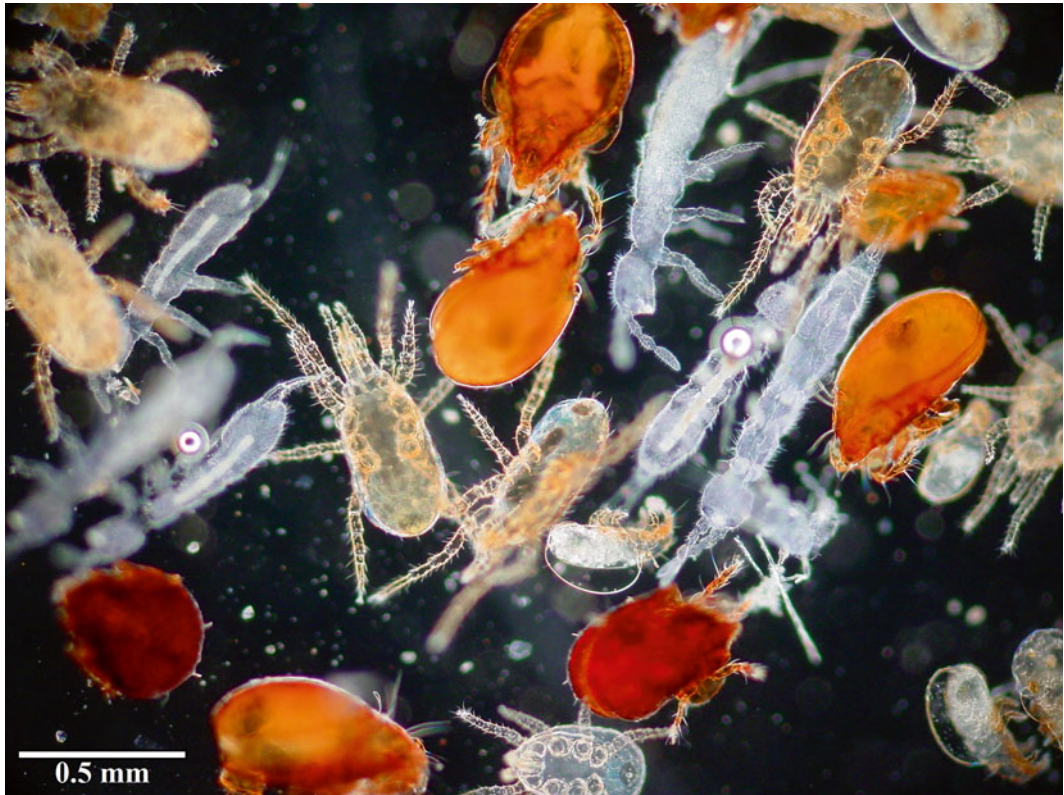


Fig. 1 Microscopic photograph of multiple microarthropod specimens (soil mites and springtails) extracted from a soil sample. Photographer: Wim Dimmers

been given to nematodes [3, 4]. Most of the work that has been done was based on conventional barcoding: sequencing individual specimens for identification or phylogenetic purposes, and the establishment of reference databases for that purpose. Now that availability of such reference data is rising, community screening by metabarcoding approaches is within reach also for soil fauna. General metazoan barcodes targeting the mitochondrial encoded cytochrome c oxidase I (CO1) region or the 18S ribosomal DNA region of the majority of metazoan exist but miss part of the soil faunal diversity (e.g., [5]) as the targeted regions are often too conserved to differentiate taxa [6]. To catch the full diversity, different faunal groups thus often need to be approached with different primers, targeting different genes and thus using distinct protocols.

Studying soil fauna, such as microarthropods (soil mites and springtails; Fig. 1), is of academic as well as applied interest, as they yield valuable indicator organisms for soil quality. Yet their use as bio-indicators depends on a classification into functional groups [7], which requires identification to family or even species level. Morphological identification is tedious and time consuming, and a molecular alternative is therefore needed for high-throughput analyses.

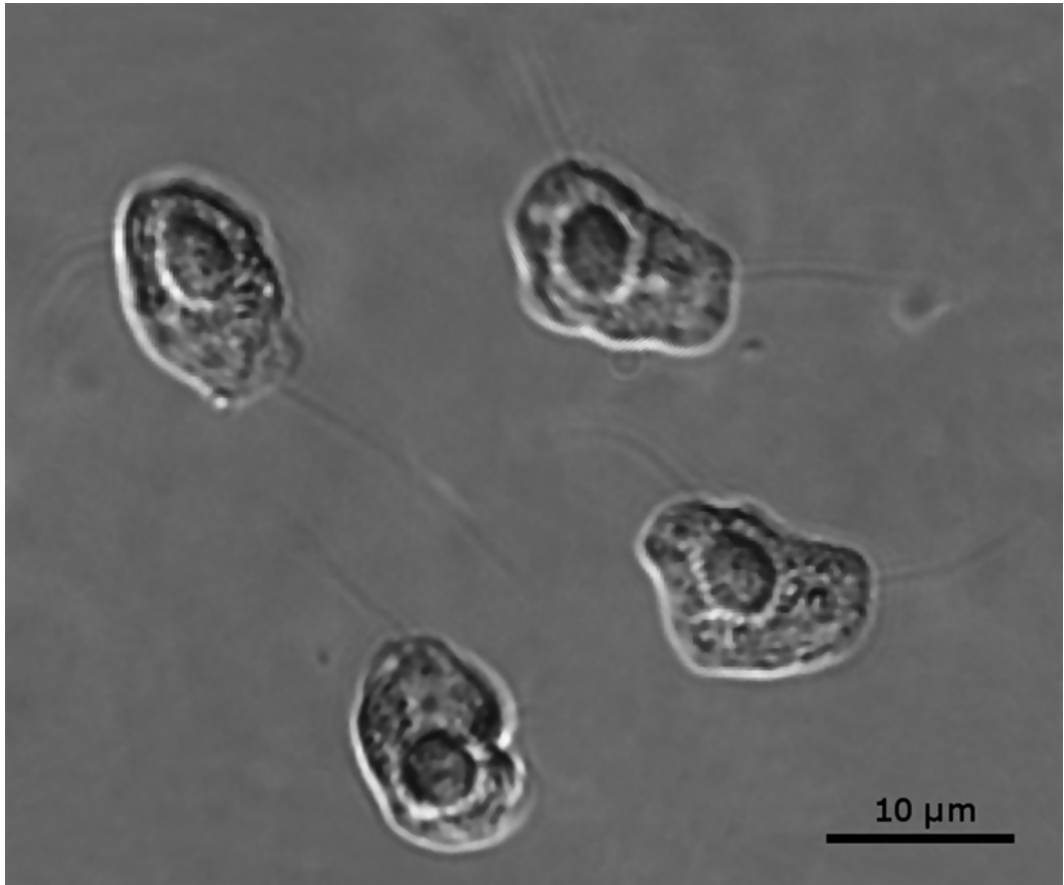


Fig. 2 Light microscopic image of *Cercomonas* sp., a common cercozoan amoeboflagellate, with ingested yeast (*Saccharomyces cerevisiae*). Scale bar: 10 μm . Pictures taken by Kenneth Dumack, assembled by Stefan Geisen

While soil fauna represents a single monophyletic eukaryotic clade of multicellular organisms, soil protists spread across the entire eukaryotic tree of life and host a huge variety of mostly heterotrophic, single celled organisms [8]. Nevertheless, protists represent the least studied soil organisms largely due to limiting methodological approaches attributed to the tiny size, transparency, and close attachment to soil particles, making them largely invisible in direct observations [9]. Morphological studies on the numerically most dominant protists, i.e., flagellates and naked amoebae, rely on tedious enrichment cultivation methods [10, 11]. Furthermore, subsequent morphological identification of especially amoeboid groups such as cercozoan flagellates (Fig. 2) is nearly impossible [12].

These methods, however, are biased towards cultivable taxa which likely present only a fraction of all known protists [9]. Using molecular tools targeting protists without prior cultivation circumvent most of these problems and have revealed a wide range of

formerly unknown soil taxa [13, 14]. This broadened knowledge on the diversity of protists revealed, e.g., Foraminifera, Dinoflagellata, and Apicomplexa that were formerly unknown from soils [14, 15]. Molecular techniques also revolutionized the view on the community structure of soil protists showing profound differences of major protist clades between soils [13, 15].

Next to analyses of functional genes, metagenomic and metatranscriptomic approaches can be applied to characterize the full eukaryote biodiversity in soil samples [13, 16]. However, these methods still require an immense sequencing effort, as the entirety of soil DNA or RNA is being sequenced (thus including non-target regions and organisms, e.g., non-barcoding genes and bacteria, respectively). Here we report on high-throughput sequencing metabarcoding approaches allowing easy and simultaneous analyses of tens to hundreds of samples to target the entirety of soil eukaryotes, specifically focusing on fauna and protists (Fig. 3). The first approach targets a wide range of eukaryotes and consequently provides a cumulative overview of the diversity and community structure of all dominant taxa of eukaryotes. This includes major protist groups, and thus will also be the method of choice for targeting the full protist diversity. The second and third approach illustrates how specific eukaryotic groups can be targeted using specific primers that enable a more detailed and more complete overview of the local taxonomic composition in this group. We show examples for a highly diverse group of protists, the phylum Cercozoa, and a highly diverse group of fauna, the microarthropods (Fig. 3).

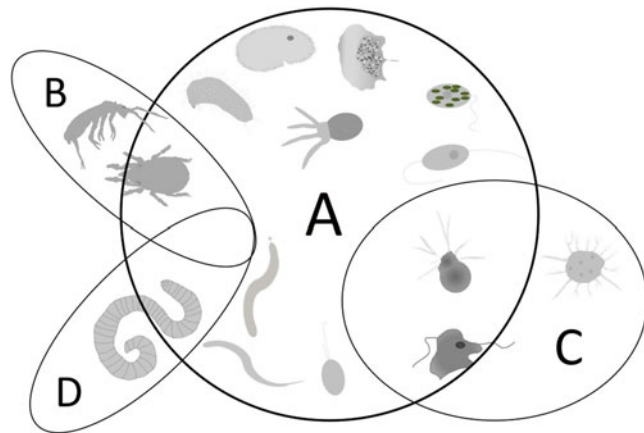


Fig. 3 Conceptual representation of a tiered metabarcoding approach using both a general 18S primer set to target a broad range of eukaryotes [18] (a) and multiple more focused primer sets to target the diversity in a particular taxonomic group in high resolution: microarthropods (b; G.A. de Groot, personal communication), Cercozoa (c; S. Geisen, personal communication) and enchytraeids (d; R. Schmeltz, personal communication). The general set captures most, but not all of the diversity captured by the focused sets

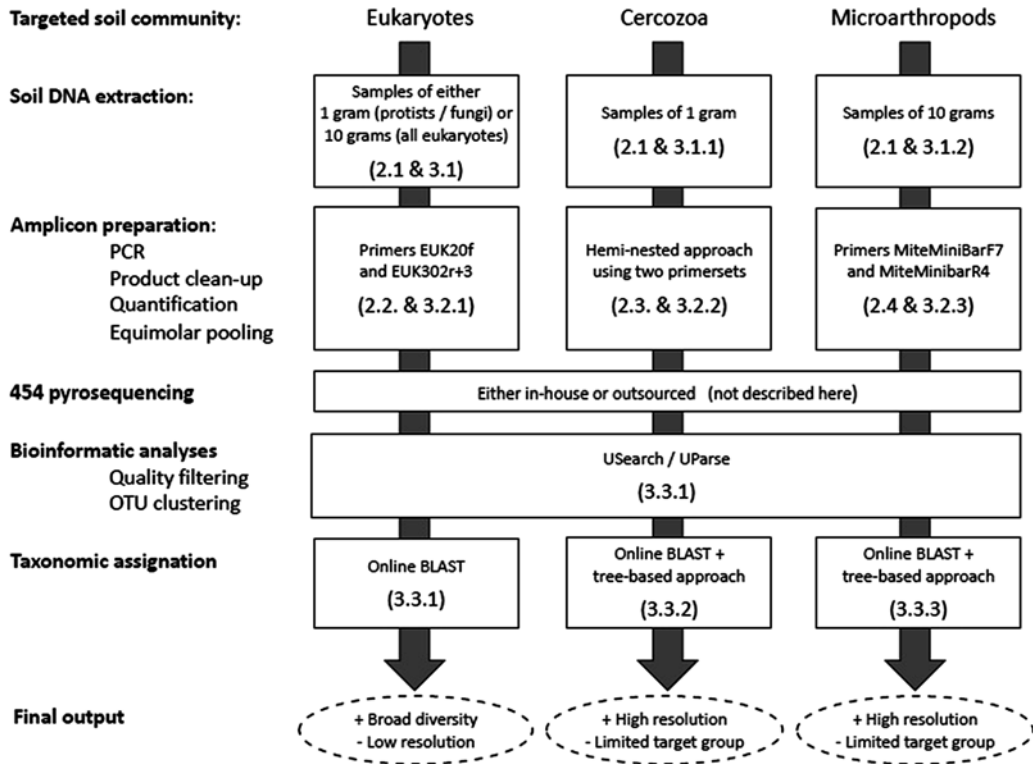


Fig. 4 Schematic overview of metabarcoding strategy targeting the entity of soil eukaryotes (a) or individual groups in more focused analyses such as mites (b) and cercozoan protists (c)

We mainly cover the process of amplicon preparation, sequencing and taxonomic identification (Fig. 4). The template DNA used for amplicon preparation can either have resulted from direct extraction of (extra)cellular soil DNA, or from an indirect method in which the community of organisms is first extracted from the soil and then a metagenomic DNA is extracted from them. As we aim for the most time-efficient method, we here describe only the direct extraction.

2 Materials

2.1 Soil DNA Extraction

2.1.1 Soil DNA Extraction for Microarthropods or Entire Eukaryotic Communities

1. Split soil core sampler ($\phi=6$ cm; length depending on the desired sampling depth).
2. PVC rings (1 per sample; 2.5 cm high, $\phi=5.8$ cm).
3. Plastic bags for carrying.
4. Hammer.
5. Sharp knife.

6. Cooling box.
7. Water and brush to clean sampler.
8. PowerMax[®] Soil DNA Isolation Kit (*Mobio Inc.*).

2.1.2 Soil DNA Extraction for Protists

Materials for extraction of protist DNA follow the ISOm standard protocol, as presented in Plassart et al. [17].

2.2 Amplicon Preparation for Metabarcoding of Eukaryotes at Low Taxonomic Resolution

1. (Environmental) template DNA with concentration standardized to 10 ng/ μ L.
2. Primers EUK20f (5'-TGC CAG TAG TCA TAT GCT TGT-3') and EUK302r+3 (5'-ACC AGA CTT GYC CTC CAA T-3') [18], preceded with 5 bp mid-tags, concentration: 10 μ M.
3. PCR Master Mix (*Supreme NZYTaQ Green PCR Master Mix; NZYTech Ld.*; containing 0.2 U/ μ L Taq polymerase, 200 μ M of each dATP, and 2.5 μ M MgCl₂); concentration: 1 \times .
4. PCR equipment (Thermocycler, 50 μ L PCR tubes).
5. PCR product quantification device (*Picogreen*[®] or *Qubit*[®]).
6. PCR cleanup kit (*Agarose GelExtract Mini kit*, 5 Prime).

2.3 Amplicon Preparation for High-Resolution Metabarcoding of Cercozoan Protists

1. (Environmental) template DNA with concentration standardized to 2 ng/ μ L.
2. Primers 25 F (5'-CAT ATG CTT GTC TCA AAG ATT AAG CCA-3') and 1256R (5'-GCA CCA CCA CCC AYA GAA TCA AGA AAG AWC TTC-3'; [19]); concentration: 10 μ M.
3. 5 bp mid-tagged primers 1256R and PreV4 (5'-GYT GCA GTT AAA AAG CTC GTA GTT G-3'; concentration: 10 μ M (Geisen et al., personal communication)).
4. PCR ingredients (nucleotides (10 mM), ddH₂O, GreenTaq Buffer (10 \times), and GreenTaq polymerase (5 U/ μ L), Fermentas, St. Leon-Rot, Germany).
5. PCR equipment (Thermocycler, 50 μ L tubes).
6. PCR product quantification device (*Picogreen*[®] or *Qubit*[®]).
7. PCR cleanup kit (*Agarose GelExtract Mini kit*, 5 Prime).

2.4 Amplicon Preparation for High-Resolution Metabarcoding of Microarthropods

- 1 (Environmental) template DNA with concentration standardized to 10 ng/ μ L.
- 2 Primers MiteMinBarF7 (5'-CAT CGI TTY RTI ATR ATT TTT TTY ATA G-3') and MiteMinBarR4 (5'-GAT AHA CWG TTC AHC CWG TSC C-3'); concentration: 10 μ M; (De Groot et al., personal communication), preceded with 5 bp mid-tags.
- 3 PCR Master Mix (*Supreme NZYTaQ Green PCR Master Mix; NZYTech Ld.*; see Subheading 2.2 for contents); concentration: 1 \times .
- 4 PCR equipment (Thermocycler, 50 μ L PCR tubes).

5 PCR product quantification device (*Picogreen*[®] or *Qubit*[®]).

6 PCR cleanup kit (*Agarose GelExtract Mini kit*, 5 Prime).

3 Methods

3.1 Soil DNA Extraction

3.1.1 Soil DNA Extraction for Microarthropods or Entire Eukaryotic Communities

Standard processing of 1 g of soil as common for microbes will not result in a proper description of the community diversity, when eukaryotes with body sizes that strongly exceed those of microbes such as the majority of soil fauna are targeted. Therefore, we advise the use of an alternative approach based on processing of larger volumes of soil.

1. Using a split soil corer (ISO 23611 2-4), take a total of ten samples, evenly distributed along the outer edge of a circular plot with a 1 m radius. Sampling depth depends on habitat type and targeted community.
2. Seal the samples individually in plastic bags and transported to the lab. Keep samples in a cool box during transport and store at 4 °C. In the lab, each individual soil sample is homogenized and a subsample of 10 g is taken for further processing.
3. Separate DNA extractions are performed for each of the ten soil samples of 10 g obtained in **step 2**. The Power Max Soil DNA Isolation Kit is used for this purpose, as this kit allows the processing of up to 10 g of soil in a single DNA extraction and includes a lysis step.
4. Obtained extracts may or may not be pooled in order to limit the samples for amplicon sequencing (Subheading [3.2](#)), depending on the desire to test for heterogeneity within the sampled plot.

3.1.2 Soil DNA Extraction for Protists

Different soil DNA extraction protocols can be adopted to study soil protists. Common DNA extraction methods used by microbiologists can be applied as microbial protists are highly abundant in tiny amounts of soil. We advise the use of the following strategy, based on Plassart et al. [[17](#)]:

1. Take five individual soil cores (20 cm depth) per site, evenly distributed along the outer edge of a circular plot with a 1 m radius. Pool replicated soil cores to obtain a composite sample for each site.
2. Sieve the composite soil samples to <4 mm.
3. Take an aliquot of 50 g from each sieved sample, and store at -40 °C prior to DNA extraction.
4. For subsequent DNA extraction from the soil samples, follow the modified ISO standard procedure (ISOM) as described in

Plassart et al. [17]. Perform three replicate extractions per aliquot (= per site).

3.2 Amplicon Preparation

Perform all steps in sterile conditions to prevent contamination.

3.2.1 Metabarcoding of Eukaryotes at Low Taxonomic Resolution

1. Use primers EUK20f and EUK302r+3 to target a 500–700 bp long region of the 18S rDNA of a wide range of eukaryotes including protists and fauna (*see Note 1*).
2. Carry out PCR reactions in 50 μ L PCR tubes in 25 μ L volume consisting of 1.5 μ L of each primer (10 μ M), 12.50 μ L PCR Master Mix, 2.5 μ L template DNA (*see Note 2*).
3. Apply the following PCR setup: Initial denaturation at 95 °C for 5 min, followed by 35 cycles of denaturation at 95 °C for 30 s, annealing at 48 °C for 60 s, and elongation at 72 °C for 2 min with a final extension for 5 min at 72 °C (*see Note 3*). Use forward and reverse primers containing the same mid-tag for individual samples and randomly choose one of the remaining primer pairs for subsequent samples (*see Note 4*).
4. Replicate **steps 1–3** and pool both pseudoreplicates (*see Note 5*).
5. Purify individual PCR products using a standard column-based purification kit (*see Note 6*).
6. Quantify purified PCR products using a fluorometric quantification device (*see Subheading 2 and Note 7*).
7. Pool the amplicons in equimolar concentrations to form a library. In case of >16 samples, allocate the amplicons to multiple libraries (*see Note 4*).
8. Send pooled libraries for pyrosequencing using the company's standard protocol.

3.2.2 High-Resolution Metabarcoding Example 1: Cercozoan Protists

1. Cercozoa are specifically being amplified using a hemi-nested PCR approach. Carry out all PCR reactions in 50 μ L PCR tubes in 31 μ L volume consisting of 0.6 μ L of each primer (10 μ M), 0.6 μ L nucleotides (10 mM), 1.0 μ L template DNA, 24.5 μ L H₂O, 3 μ L PCR Buffer, and 0.15 μ L polymerase (5 U/ μ L) (*see Note 2*).
2. In a first PCR round, use the cercozoan-specific primer combination 25F and 1256R to generate amplicons of ~1200 bp from each sample [19]. Apply the following PCR setup: Initial denaturation at 95 °C for 1 min, followed by 35 cycles of denaturation at 95 °C for 30 s, annealing at 70 °C for 60s, and elongation at 72 °C for 2 min with a final extension for 5 min at 72 °C (*see Note 3*).
3. Replicate **steps 1–3** and pool both pseudoreplicates (*see Note 5*).

4. Take 1.0 μL aliquots of the resulting PCR product mixtures as template for a hemi-nested PCR step using mid-tagged primers PreV4 and 1256R to produce amplicons of ~ 500 bp, appropriate for 454 sequencing. Apply the same PCR conditions as above but decrease the annealing temperature to 66 $^{\circ}\text{C}$ and reduce the elongation to 90 s. Use primers with the same mid-tag for individual samples and randomly use one of the remaining primer pairs for subsequent samples (*see Note 4*).
5. Replicate **steps 1–3** and pool both pseudoreplicates (*see Note 5*).
6. Follow **steps 5–8** in Subheading [3.2.1](#).

3.2.3 High-Resolution Metabarcoding Example 2: Soil Microarthropods

1. Use primers MiteMinBarF7 and MiteMinBarR4 to target a ~ 200 bp minibarcode located within the cytochrome oxidase subunit 1 (CO1) region. This minibarcode was designed to target all clades of Acari (Acariformes and Parasitiformes).
2. Carry out PCR reactions in 50 μL PCR tubes in 25 μL volume consisting of 1.5 μL of each primer (10 μM), 12.5 μL PCR Master Mix, 2.5 μL template DNA (*see Note 2*).
3. Apply the following PCR setup: Initial denaturation at 95 $^{\circ}\text{C}$ for 5 min, followed by five cycles of denaturation at 95 $^{\circ}\text{C}$ for 40 s, annealing at 43 $^{\circ}\text{C}$ for 40 s, and elongation at 72 $^{\circ}\text{C}$ for 1 min, then followed by 35 cycles of denaturation at 95 $^{\circ}\text{C}$ for 40 s, annealing at 49 $^{\circ}\text{C}$ for 40 s, and elongation at 72 $^{\circ}\text{C}$ for 1 min, with a final extension for 5 min at 72 $^{\circ}\text{C}$ (*see Note 3*). Use forward and reverse primers containing the same mid-tag for individual samples and randomly choose one of the remaining primer pairs for subsequent samples (*see Note 8*).
4. Replicate **steps 1–3** and pool both pseudoreplicates (*see Note 5*).
5. Follow **steps 5–8** in Subheading [3.2.1](#).

3.3 Bioinformatics Analyses

Below we describe pipeline for bioinformatic analysis of the sequencing output, largely making use of the USearch software package [20, 21], and in some cases the Mothur [22] package. [Appendix](#) provides an example of the exact list of scripting command lines as applied on sequencing data for the microarthropod primer set (Subheading [3.3.3](#)).

3.3.1 Metabarcoding of All Eukaryotes at Low Taxonomic Resolution

1. Convert the raw output of the sequencer (sff-files) to fasta and quality files using the using the sffinfo command of Mothur v.1.22.2 [22].
2. Convert fasta and quality files into fastq file using the faqual-2fastq.py script of Usearch v7.0.1001 [16].
3. Use the fastq_strip_barcode_relabel2.py script (available in USearch package) to sort reads according to primer sequence, label reads with their multiplex identifier (MID) tag and strip

primer and MID sequence from read. Use standard stringency, allowing up to two mismatches in the primer sequence (*see* **Note 9**) and no mismatches in the MID sequence.

In this step reads not matching the barcode or primer sequence, with mentioned stringency, are discarded. This step is performed for the forward and the reverse primer separately, resulting in two separate files (*see* **Note 10**).

4. For the file with reads that include the forward primer, filter out reads of low quality; maximum expected error allowed is 0.5 (*see* **Note 11**). And truncate all reads to one and the same length of 250 bp (*see* **Note 12**), using Usearch command `fastq_filter`.
5. Dereplicate truncated sequences to remove duplicated sequences using the `derep_fulllength` command of Usearch.
6. Sort dereplicated sequences by decreasing abundance and discard singletons using `sortbysize` command of Usearch.
7. Generate operational taxonomic units (OTUs; an operational definition of a species or a group of species that is entirely based on sequence information), from abundance-sorted sequences using the `cluster_otus` command of Usearch for 97 % similarity thresholds (*see* **Note 13**).
8. Label OTUs using the `fasta_number.py` script (available in USearch package).
9. Map trimmed sequences (including singletons) against the OTU representative sequences using the `usearch_global` command of Usearch. Identity threshold is 97 %.
10. Generate matrices containing the sequence abundances of different OTUs in each soil sample based on these mapping results using the `uc2otutab.py` script of Usearch (*see* **Note 14**).
11. Determine taxonomic assignation for each OTU using the Basic Local Alignment Search Tool (BLAST) algorithm v 2.2.23 [23] against the Protist Ribosomal Reference Database PR2 [24] using an e -value cutoff of $1e^{-5}$, an identity cutoff of 90 %, and a coverage cutoff of 80 % of the query sequence covered in the alignments (*see* **Note 15**).
12. Assign OTUs to different taxonomic levels (class, order, family, genus, species, and OTU level) (*see* **Note 16**).
13. Repeat **steps 4–12** for the file with reads that include the reverse primer.

3.3.2 High-Resolution Metabarcoding Example 1: Cerczoan Protists

1. Follow **steps 1–12** from Subheading 3.3.1.
2. Determine a rough taxonomic assignation for each OTU using the Basic Local Alignment Search Tool (BLAST) algorithm v 2.2.23 [23] against the Protist Ribosomal Reference Database PR2 [24] using an e -value cutoff of $1e^{-5}$, an identity cutoff of

90 %, and a coverage cutoff of 80 % of the query sequence covered in the alignments (*see* **Note 15**).

3. Assign OTUs to higher taxonomic levels (class, order, family) (*see* **Note 16**). Discard any OTUs not assigned as Cercozoa.
4. For assignation to genus or species level, we advise to also use a phylogenetic tree-based approach. For this purpose, download a representative set of cercozoan sequences from the online databases. Align these with the newly gained OTUs and construct a maximum-likelihood tree.

3.3.3 High-Resolution Metabarcoding Example 2: Soil Microarthropods

1. Follow **steps 1–3** from Subheading **3.3.1**.
2. Reads produced by sequencing from the reverse primer are converted into reverse complements using the `reverse.seqs` command in Mothur. And forward and reverse sequences are combined into one fasta file using the `Mothur merge.files` command.
3. Using Usearch command `fastq_filter`, filter out reads of low quality, max expected error allowed is 0.5 (*see* **Note 11**) and truncate all reads to one and the same length of 158 bp (*see* **Note 12**).
4. Follow **steps 5–10** from Subheading **3.3.1**.
5. Determine a rough taxonomic assignation for each OTU using the Basic Local Alignment Search Tool (BLAST) algorithm v 2.2.23 [23] against the Protist Ribosomal Reference Database PR2 [24] using an e -value cutoff of $1e^{-5}$, an identity cutoff of 90 %, and a coverage cutoff of 80 % of the query sequence covered in the alignments (*see* **Note 15**).
6. Assign OTUs to higher taxonomic levels (class, order, family) (*see* **Note 16**). Discard any OTUs not assigned as micro arthropods.
7. For assignation to genus or species level, we advise to also use a phylogenetic tree-based approach. For this purpose, download a representative set of Acari sequences from the online databases. Align these with the newly gained OTUs and construct a maximum-likelihood tree.

4 Notes

1. A wide range of so called general eukaryotic primers have been tested and used in recent studies [15, 25–28]. All primer combinations will provide a different picture of the eukaryotic community due to primer biases and differences in amplicon lengths, leading to over-/underrepresented of some taxa [29, 30]. Another issue is raised by profound differences in copy numbers of even closely related eukaryotes ruling out quantitative information of the obtained sequence data [31–33].

Therefore, relative abundance information needs to be used to investigate eukaryotic communities.

2. Higher or lower volumes can be used in case more or less product is needed for subsequent sequencing.
3. PCR conditions might depend on the polymerase and thermocycler being used and should be tested using a gradient PCR before application. An altered annealing temperature will, however, provide a different picture of the resulting community as higher annealing temperatures will benefit those targets that optimally bind primers, while lower temperature will also amplify less-specific targets [34, 35]. The approach taken depends on the experimental question but to compare different studies, adopting an identical protocol is inevitable.
4. If more libraries are to be created (*see step 7* of Subheading 3.2.1), the same primers can be reused in different libraries, thereby reducing primer costs. In this case, consider the distribution of samples over libraries already before the PCR, so that samples can randomly be allocated to a library.
5. Even more independent PCR replicates can be conducted and pooled to decrease the error rate occurring in individual PCR reactions [36].
6. Running a subsample on gel may be worthwhile in order to check for additional unwanted products of different size. Especially products with smaller fragment size than the target product may consume many sequencing reads, thus lowering the coverage of the targeted organisms. In case unwanted products are observed, gel extraction may be used to retain only the targeted product.
7. DNA quantification methods via NanoDrop is not recommended as this does not allow specific product quantification and consequently rules out subsequent equimolar pooling.
8. Amplification success is known to vary among primer-tag combinations. Based on tests with a limited set of tags, we can report good results for our primers using the following tags: AGTCT, AGCGA, ATCGT, ATAGT, ACGTA, CAGTA, CAGCT, CTAGT, CTCGA, TAGAT, TGAGT, and TACGT.
9. For a more conservative approach, zero mismatches in the primer region can be used
10. Given the length of the fragment amplified here (500–700 bp), the forward and reverse reads show insufficient overlap to be pooled into a single file with stringent quality filtering and sequence cutoff resulting in short, but high-quality sequences. Therefore the next steps are done separately for the datasets with reads including the forward and reverse primer respectively. Note that pooling is possible when using the focused microarthropod primer set (Subheading 3.3.3).

11. For a more conservative approach, higher quality values can be applied.
12. Truncation of the sequencing length strongly depends on the sequencing platform, chemistry used, and data output. In case the majority of sequences remaining after the first steps of quality filtering is much higher than 300 bp, truncation to longer amplicons is recommended as it allows deeper and more reliable taxonomic OTU assignments in subsequent steps [37–39].
13. Investigating different OTU clustering levels give an overview of an appropriate OTU clustering level for subsequent analyses. No standard clustering level is recommended as profound differences between and within eukaryotic groups exist [40–43]. Also differences in error rates in the respective sequencing approach need to be considered [44, 45]. Therefore, OTU delineation depends on the group investigated and the sequencing approach taken. Yet, at similarity levels >98 %, the clustering algorithm might be less effective in filtering out chimeras (USearch Manual: http://www.drive5.com/usearch/manual/cluster_otus.html).
14. Rarefaction of the obtained OTU matrix is current practice in order to normalize read numbers among samples, thereby avoiding biased diversity differences due to unequal sequence depth. However, recent studies strongly advise to avoid rarefaction. *See* [46] for more information and potential alternatives.
15. Alternative databases such as GenBank [47] or Silva [48] can be used, but PR2 has recently been introduced, providing quality filtered data especially suitable for protists and other eukaryotes [24].
16. Automated assignment to deeper taxonomic levels, such as genus or species should be done with great care, as the species concept in many protist groups is far from common acceptance [42, 49, 50]. Many commonly defined species cannot be distinguished using short sequences due too very similar or even identical barcode sequences and clustering algorithms needed to compensate for sequencing errors often artificially lump together different species, while less stringent clustering inflates species diversity [40, 42, 51, 52]. Therefore, we recommend applying stringent OTU clustering and keep the OTU level as a unique taxonomic level unless specific facts about certain genera or species are targeted. In that case, researchers are recommended to carefully evaluate all automatically assigned OTUs, e.g., using manual BLAST searches (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) and/or phylogenetic analyses of respective sequences with those of reference taxa.

Acknowledgements

This work was supported by the European Commission within the FP7 project EcoFINDERS (FP7-264465). The authors would like to sincerely thank the colleagues involved in planning and contributing to parts of the described methods, especially Marc Bueé, Dalila Costa, Bryan Griffiths, Francis Martin, Rüdiger Schmelz, Dorothy Stone, Antón Vizcaíno, David Bass, and Michael Bonkowski.

5 Appendix: Example of Scripting Command lines as Applied for Bioinformatic Analysis of 454 Sequencing Data Gained for Microarthropods

Appendix provides an example of the exact list of scripting command lines.

1. Conversion of the sff file to a fastq and a quality file is often already done by the sequencing service provider. If not, apply the `.sffinfo` command in Mothur.
2. `python faqual2fastq2.py 454_LibraryX_Reads.fna 454_LibraryX_Reads.qual > 454_LibraryX_Reads+Qual.fq`
3. (a) `python fastq_strip_barcode_relabel2.py 454_LibraryX_Reads+Qual.fq CATGCNTTYRTNATRATTTTTTYATAG BarcodesX.fas FW > 454_LibraryX_Reads+Qual_FW.fq`
 (b) `python fastq_strip_barcode_relabel2.py 454_LibraryX_Reads+Qual.fq GATAHACWGTTCACAHCCWGTSCC BarcodesX.fas RV > 454_LibraryX_Reads+Qual_RV.fq`
4. (a) `usearch -fastq_filter 454_LibraryX_Reads+Qual_FW.fq -fastq_maxee 0.5 -fastq_trunclen 158 -fastaout 454_LibraryX_Reads_FW_158bp.fa`
 (b) `usearch -fastq_filter 454_LibraryX_Reads+Qual_RV.fq -fastq_maxee 0.5 -fastq_trunclen 158 -fastaout 454_LibraryX_Reads_RV_158bp.fa`
5. `mothur > reverse.seqs(fasta=454_LibraryX_Reads_RV_158bp.fa)`
6. `mothur > merge.files(input=454_LibraryX_Reads_RV_158bp.rc.fa-454_LibraryX_Reads_FW_158bp.fa, output=454_LibraryX_Reads_FW+RVrc_158bp.fa)`
7. `usearch -derep_fulllength 454_LibraryX_Reads_FW+RVrc_158bp.fa -output 454_LibraryX_Reads_derep.fa -sizeout`
8. `usearch -sortbysize 454_LibraryX_Reads_derep.fa -output 454_LibraryX_Reads_sorted.fa -minsize 2`
9. `usearch -cluster_otus 454_LibraryX_Reads_sorted.fa -otus 454_LibraryX_Reads_otus1.fa`

10. python fasta_number.py 454_LibraryX_Reads_otus1.fa OTU_> 454_LibraryX_Reads_otus.fa
11. usearch -usearch_global 454_LibraryX_Reads_FW+RVrc_158bp.fa -db 454_LibraryX_Reads_otus.fa -strand plus -id 0.97 -uc 454_LibraryX_Reads_map97.uc
12. python uc2otutab.py 454_LibraryX_Reads_map97.uc > 454_LibraryX_Reads_otu97_table.txt

References

1. Bálint M, Schmidt P-A, Sharma R et al (2014) An Illumina metabarcoding pipeline for fungi. *Ecol Evol* 4(13):2642–2653
2. Buée M, Reich M, Murat C et al (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytol* 184(2):449–456
3. Chen X, Daniell T, Neilson R et al (2010) A comparison of molecular methods for monitoring soil nematodes and their use as biological indicators. *Eur J Soil Biol* 46(5):319–324
4. Porazinska DL, Giblin-Davis RM, Faller L et al (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Mol Ecol Res* 9(6):1439–1450
5. Tang CQ, Leasi F, Obertegger U et al (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc Natl Acad Sci U S A* 109(40):16208–16212
6. Bass D, Richards T, Matthai L et al (2007) DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* 7(1):162
7. Siepel H (1995) Applications of microarthropod life-history tactics in nature management and ecotoxicology. *Biol Fert Soils* 19(1):75–83
8. Adl SM, Simpson AGB, Lane CE et al (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol* 59(5):429–514
9. Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agr Ecosyst Environ* 74(1-3):95–112
10. Darbyshire JF, Whitley RE, Graebes MP et al (1974) A rapid micromethod for estimating bacterial and protozoan populations in soil. *Rev Ecol Biol Sol* 11:465–475
11. Geisen S, Bandow C, Römbke J et al (2014) Soil water availability strongly alters the community composition of soil protists. *Pedobiologia* 57(4-6):205–213
12. Bass D, Howe AT, Mylnikov AP et al (2009) Phylogeny and classification of Cercomonadida (Protozoa, Cercozoa): *Cercomonas*, *Eocercomonas*, *Paracercomonas*, and *Cavernomonas* gen. nov. *Protist* 160(4):483–521
13. Geisen S, Tveit AT, Clark IM et al. Metatranscriptomic census of active protists in soils. *ISME J* 9(10):2178–2190
14. Lejzerowicz F, Pawlowski J, Fraissinet-Tachet L et al (2010) Molecular evidence for widespread occurrence of Foraminifera in soils. *Environ Microbiol* 12(9):2518–2526
15. Bates ST, Clemente JC, Flores GE et al (2013) Global biogeography of highly diverse protistan communities in soil. *ISME J* 7(3):652–659
16. Urich T, Lanzén A, Qi J et al (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* 3(6), e2527
17. Plassart P, Terrat S, Thomson B et al (2012) Evaluation of the ISO standard 11063 DNA extraction procedure for assessing soil microbial abundance and community structure. *PLoS One* 7(9), e44279
18. Euringer K, Lueders T (2008) An optimised PCR/T-RFLP fingerprinting approach for the investigation of protistan communities in groundwater environments. *J Microbiol Methods* 75(2):262–268
19. Bass D, Cavalier-Smith T (2004) Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). *Int J Syst Evol Microbiol* 54(6):2393–2404
20. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461
21. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10(10):996–998
22. Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537–7541
23. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410

24. Guillou L, Bachar D, Audic S et al (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41(D1):D597–D604
25. Baldwin DS, Colloff MJ, Rees GN et al (2013) Impacts of inundation and drought on eukaryote biodiversity in semi-arid floodplain soils. *Mol Ecol* 22(6):1746–1758
26. Wang Y, Tian RM, Gao ZM et al (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PLoS One* 9(3), e90053
27. Hugerth LW, Muller EEL, Hu YOO et al (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* 9(4), e95567
28. Adl SM, Habura A, Eglit Y (2014) Amplification primers of SSU rDNA for soil protists. *Soil Biol Biochem* 69:328–342
29. Epstein S, López-García P (2008) “Missing” protists: a molecular prospective. *Biodivers Conserv* 17(2):261–276
30. Stoeck T, Hayward B, Taylor GT et al (2006) A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* 157(1):31–43
31. Zhu F, Massana R, Not F et al (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* 52(1):79–92
32. Gong J, Dong J, Liu X et al (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* 164(3):369–379
33. Medinger R, Nolte V, Pandey RV et al (2010) Diversity in a hidden world: potential and limitation of next generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* 19(Supplement s1):32–40
34. Prosser J, Jansson JK, Liu WT (2010) Nucleic-acid-based characterization of community structure and function. *Environ Mol Microbiol* 63
35. Schmidt P-A, Bálint M, Greshake B et al (2013) Illumina metabarcoding of a soil fungal community. *Soil Biol Biochem* 65:128–132
36. von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21(3):213–229
37. Corsaro D, Venditti D (2011) More *Acanthamoeba* genotypes: limits to the use rDNA fragments to describe new genotypes. *Acta Protozool* 50(1):49
38. Nowrousian M (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell* 9(9):1300–1310
39. Shokralla S, Spall JL, Gibson JF et al (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21(8):1794–1805
40. Geisen S, Kudryavtsev A, Bonkowski M et al (2014) Discrepancy between species borders at morphological and molecular levels in the genus *Cochliopodium* (Amoebozoa, Himatismenida), with the description of *Cochliopodium plurinucleolum* n. sp. *Protist* 165(3):364–383
41. Geisen S, Fiore-Donno AM, Walochnik J et al (2014) *Acanthamoeba* everywhere: high diversity of *Acanthamoeba* in soils. *Parasitol Res* 113(9):3151–3158
42. Boenigk J, Ereshefsky M, Hoef-Emden K et al (2012) Concepts in protistology: species definitions and boundaries. *Eur J Protistol* 48(2):96–102
43. Caron DA (2013) Towards a molecular taxonomy for protists: benefits, risks, and applications in plankton ecology. *J Eukaryot Microbiol* 60(4):407–413
44. Quail M, Smith M, Coupland P et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13(1):341
45. Behnke A, Engel M, Christen R et al (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microbiol* 13(2):340–349
46. McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10(4), e1003531
47. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2005) GenBank. *Nucleic Acids Res* 33(Database issue):D34–D38
48. Pruesse E, Quast C, Knittel K et al (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35(21):7188–7196
49. Coleman AW (2002) Microbial eukaryote species. *Science* 297:337
50. Caron DA, Worden AZ, Countway PD et al (2008) Protists are microbes too: a perspective. *ISME J* 3(1):4–12
51. Quince C, Lanzén A, Curtis TP et al (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6(9):639
52. Quince C, Lanzen A, Davenport R et al (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12(1):38

Identification and In Situ Distribution of a Fungal Gene Marker: The Mating Type Genes of the Black Truffle

Herminia De la Varga and Claude Murat

Abstract

Truffles are ectomycorrhizal fungi harvested mainly in human managed agroforestry ecosystems. Truffle production in truffle orchards faces two important bottlenecks or challenges: the initiation of the sexual reproduction and the growth of the ascocarps during several months. The black Périgord truffle, *Tuber melanosporum*, is a heterothallic species and the mating type genes (MAT1-1 and MIT1-2) have been characterized. In this context, the unraveling of the *T. melanosporum* mating type strains distribution in truffle orchards is a critical starting point to provide new insights into its sexual reproduction. The aim of this chapter is to present the protocol used to characterize the *T. melanosporum* mating type present in a truffle orchard from ascocarps, hazel mycorrhizal root tips, and/or soil samples, by polymerase chain reactions using specific primers for those genes, but it can be adapted for other fungal species.

Key words *Tuber melanosporum*, Mating type genes, Polymerase chain reaction (PCR), Ascocarps, Ectomycorrhiza, Soil, DNA

1 Introduction

Truffles are soil fungi that associate with the roots of certain species of trees and shrubs to form a dual symbiotic organ called an ectomycorrhiza. This ectomycorrhizal association is a mutualistic symbiosis occurring between fungi and the root of trees, wherein plants provide sugar and fungi help the tree with its mineral and water uptakes. When sporulating, truffles form a fleshy and scented structure called an ascocarp (i.e., truffle fructification), which attracts animals and disperses spores following ascocarp ingestion. In Europe, 32 species of truffles have been identified but few have been successfully commercialized. The Périgord black truffle (*Tuber melanosporum*) grows naturally in Southern Europe and in contrasting climates such as the warmer climate of the Mediterranean in southern Spain and Italy as well as in colder continental climates in northeastern France. *T. melanosporum* is harvested in different environments ranging from managed plantations to natural forests.

The first inoculated seedlings with truffles were commercialized 40 years ago under an INRA/ANVAR (“*Institut National de la Recherche Agronomique/Agence National de Valorisation de la Recherche*”) know-how license [1]. The inoculation of seedlings with truffle species is performed using ascocarps as inoculum containing meiotic spores, which after germination form ectomycorrhizas on seedling roots. Truffle production was moved in the twentieth century from sub-natural woodlands to truffle orchards. In France, more than 80 % of black truffle production is produced in truffle orchards, i.e., in managed agroforestry ecosystems where seedlings, previously inoculated in nurseries with different *Tuber* species, are implanted [2]. Truffle production in truffle orchards faces two important bottlenecks or challenges: the initiation of the sexual reproduction and the growth of the ascocarps during several months. Thank to the sequencing of its genome, it has been demonstrated that *T. melanosporum* is a heterothallic species and both mating type idiomorphs (MAT1-1 and MAT1-2) have been characterized [3, 4]. The initiation of the sexual reproduction required therefore two compatibles mycelium. In order to better understand how this sexual reproduction occurs we characterized the *T. melanosporum* small-scale genetic structure [5]. This highlighted the fact that *T. melanosporum* truffle orchards are, in themselves, dynamic ecosystems which can contain up to 13 small black truffle genets (group of genetically identical individuals) in 30 m². Moreover, a nonrandom distribution pattern of *T. melanosporum* was observed, resulting in field patches exclusively colonized by genets of the same mating types (Fig. 1). The aim of this chapter is to present the protocol used to characterize the *T. melanosporum* mating type present in the truffle orchard from ascocarps, mycorrhizal root tips, and/or soil samples. The identification of mating types for the most profitable truffle species (i.e., *T. aestivum*, *T. borchii*, *T. indicum*, *T. magnatum*, and *T. melanosporum*) is protected by an international patent (n°WO2012/032098). The protocol presented here is therefore only useable for noncommercial purpose. This methodology can be extended to other fungal species whether the sequences of their mating type genes are available and by adjusting the molecular markers.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise).

5.5 M Solution of guanidine thiocyanate: pH 7. Dissolve 324.94 g of guanidine thiocyanate in 350 mL of ultrapure water. Mix and adjust pH to 7. Make up to 1 L with ultrapure water. Store at room temperature in the dark (light sensitive).

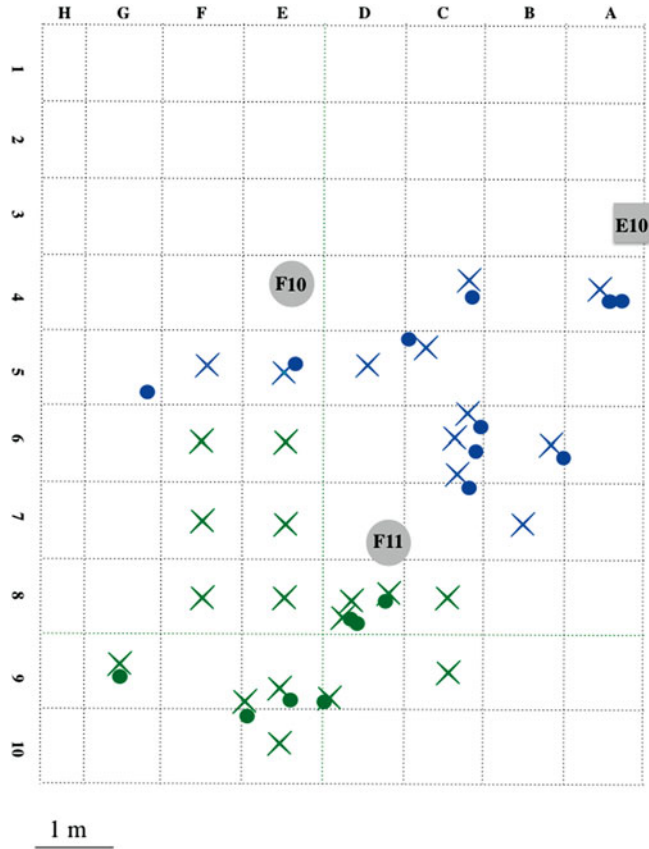


Fig. 1 Mating type mapping under three trees (F10/F11/E10) in a truffle orchard (modified from ref. 5). The positions of the genotyped ascocarps (*dots*) and ectomycorrhizas (ECMs) (*crosses*) are indicated. Samples that displayed the MAT1-1 and MAT1-2 mating types are indicated in *green* and *blue*, respectively

10× TBE or Tris/Borate/EDTA buffer: dissolve 81 g of Tris base and 41.25 g of boric acid in 500 mL of ultrapure water. Add 30 ml of 0.5 M EDTA pH 8. Make up to 750 mL with ultrapure water. Autoclave and store at room temperature.

Stereomicroscope Zeiss Stemi 2000-C, with 20- to 40-fold zoom factor and a cold light microscope KL200 led.

DNA extraction kits (examples):

- Dneasy Plant Mini Kit (Qiagen SA, Courtaboeuf, France).
- REDExtract-N-Amp™ Plant PCR Kit (Sigma-Aldrich Co. LLC, St Louis, MO, USA).
- Fast DNA Spin kit for soil (MP Biomedicals, Illkirch, France).
- Power Soil® DNA isolation Kit (MoBio Laboratories, Carlsbad, CA).

3 Methods

3.1 Sampling

3.1.1 *Ascocarps* Samples

1. Collect the truffle and map their location, host species, and date.
2. Clean truffles with tap water and soft brush to remove all soil attached to the surface.
3. With a sterile scalpel, remove the peridium and cut the truffle in small pieces (0.5–1 cm).
4. Store the truffle pieces in microcentrifuge tubes at $-20\text{ }^{\circ}\text{C}$ for molecular analyses.

3.1.2 *Mycorrhizal* Samples

1. Retrieve tree fine roots carefully from the first 10–20 cm of soil layer, at a minimum distance of 30 cm from trunk trees. Map their location, host species, and date.
2. Root pieces are washed in water, leaving the roots in water bath to allow the remaining soil to go down.
3. Root samples are transferred to glass Petri dishes or containers with clean water and observed under stereomicroscope and a cold light microscope (KL200 led).
4. With the help of fine forceps single *T. melanosporum* ectomycorrhizal tips are selected. The mycorrhizae are identified by morphotyping on the basis of color, mantle shape, and surface texture [6, 7] (*see Note 1*).
5. Each single mycorrhiza is stored individually in microcentrifuge tubes at $-20\text{ }^{\circ}\text{C}$ for molecular analyses.

3.1.3 *Soil Samples*

1. Collect soil samples in the first 10–20 cm of soil layer, at a minimum distance of 30 cm from trunk trees and map their location, tree species, and date (*see Note 2*).
2. Classify soil samples in plastic bags or tubes.
3. Air-dry each soil sample at room temperature and then sieve it through 1 mm mesh to eliminate plant debris, stones, and roots.
4. Keep soil samples at $-20\text{ }^{\circ}\text{C}$ until processing for molecular analyses.

3.2 DNA Extractions

3.2.1 *Ascocarps* and *Mycorrhizal Samples*

Genomic DNA can be isolated from single mycorrhizal root tips and ascocarps (100 mg) using any commercial kit, such as the Dneasy Plant Mini Kit (Qiagen SA, Courtaboeuf, France) following the manufacturer's instructions, or faster ones such as the REDExtract-N-Amp™ Plant PCR Kit (Sigma-Aldrich Co. LLC, St Louis, MO, USA) (*see Note 3*).

3.2.2 *Soil Samples*

Genomic DNA can be isolated from 0.25 or 0.5 g of soil using any commercial kit, as the Power Soil® DNA isolation Kit (MoBio Laboratories, Carlsbad, CA), following the manufacturer's instructions, or the Fast DNA Spin kit for soil (MP Biomedicals, Illkirch, France) with some modifications (*see Note 4*).

3.3 PCR Amplifications

The identity of *T. melanosporum* ectomycorrhizas (ECM), ascocarps, and *T. melanosporum* structures in soil samples (*see Note 5*) is assessed by PCR amplification using species-specific internal transcribed spacer (ITS) primers of the nuclear ribosomal-DNA [8, 9]: ITS4LNG (5'-TGA TAT GCT TAA GTT CAG CGG G-3') and ITSML (5'-TGG CCA TGT GTC AGA TTT AGT A-3').

For the identification of the mating type genes, PCR reactions are carried out by using the specific primers for the two genes described by Rubini et al. [4].

MAT1-2-1 primers: P1 (5'-CAG GTC CGT CAT CTC CTT CCA GCAG-3') and P2 (5'-CCA CAT GCG ACC GAG AAT CTT GGC TA-3').

MAT1-1-1 primers: P19 (5'-CAA TCT CAC TCG TGA TGT CTG GGT C-3') and P20 (5'-TCT CGG GCT GGA GGT GCG GGT CGA GT-3').

Prepare PCR reaction mixtures for DNA amplification. PCRs are performed in 10 μ L volume reactions mixture. Prepare enough mix for the number of reactions plus 5 %. (e.g., 3 samples + 1 negative control + 1 positive control = 6 \times). *Keep the tubes on ice.*

1. Mix 5 μ L of REDExtract-N-Amp™ PCR ReadyMix™ (Sigma-Aldrich Co. LLC, St Louis, MO, USA)—a 2 \times PCR mix containing—200 nM of each primer (0.2 μ L of a 10 μ M stock solution); 0.35 μ L of BSA (16 μ g/ μ L) for ectomycorrhizal and soil samples. Finally adjust volume to 8 μ L/reaction with ultra-pure water (*see Notes 6 and 7*). Mix by vortexing.
2. Add 8 μ L of reaction mixture to each PCR tube. Pipette so that the mix is on the bottom of the tube.
3. Vortex the DNA tubes (2–3 s) or mix by pipetting, centrifuge them (short spin), and add 2 μ L (5–20 ng) of the appropriate template DNA extract to each reaction tube. Pipette reaction mix so the DNA template is mixed with the Master Mix in the reaction tube. Close the tubes. If needed, perform a short spin of the sample tubes to bring the liquid down to the bottom.
4. Place tubes into the thermal cycler and begin cycling according to the following parameters described in Table 1.

3.4 Electrophoresis

1. Prepare a 2 % agarose gel, i.e., add 2 g of agarose to 100 mL of 1 \times TBE buffer in a glass bottle or Erlenmeyer. Microwave the solution for about 2 min, until agarose is completely dissolved. Cool the agarose solution to about 55 °C.
2. Prepare the gel tray and place the well comb with the desired number of wells (samples) in the slots at the top of the gel.
3. Pour the agarose into the middle of the tray until it is about half way up the teeth of the comb and has filled the tray to the corners. Let the agarose solidify during 20 min.

Table 1

Thermal profiles (temperature, time, and cycle numbers) for the different primer pairs used and the different samples

	ITS1F-ITS4		ITS4LN-ITSML/MAT		Ascocarps	ECM/soil
Initialization	94 °C	4 m	94 °C	4 m		
Denaturation	94 °C	30 s	94 °C	30 s		
Annealing	55 °C	30 s	60 °C	30 s	30 cycles	35 cycles
Extension	72 °C	1 m 30 s	72 °C	1 m		
Final extension	72 °C	5 m	72 °C	5 m		

4. When the gel is solidified, put it into the electrophoresis chamber, with the comb closest to the black electrode (on the top). Fill the chamber with 1× TBE buffer, covering the gel. Remove the comb carefully.
5. Load 5 µL of the samples into the gel (one sample in each well of the gel). Load also a 100 bp DNA ladder in one well (following manufacturer's protocol for quantities) (*see Note 8*).
6. Place the lid on the gel box; connect the electrodes to the power supply. Make sure that the black wire goes into the black plug and the red into red.
7. Turn on the power supply and set it at 110 V. Run the gel for 60–70 min.
8. After the gel has run, remove it from the gel box.
9. Put the gel into a 0.5 µg/mL ethidium bromide solution bath for 5 min. Then wash it in a water bath for at least 10 min.
10. Place the gel on the transilluminator Chemidoc® UV light box to visualize the DNA and to photograph it. Examples of the expected results in Fig. 2.

4 Notes

1. From each root sample, at least five single *T. melanosporum* ectomycorrhizas are selected along a root piece, to be sure that we have enough material to work with. It is better to collect young mycorrhiza.
2. The sampling can be done by taking soil cylinders with a soil borer (e.g., 200 mL volume, 3.2 cm diameter and 20 cm deep) or, alternatively, directly by digging to that depth and sampling a similar volume with a small shovel.
3. When using the REExtract-N-Amp™ Plant PCR Kit to extract DNA from ascocarps the volume of buffer used can be

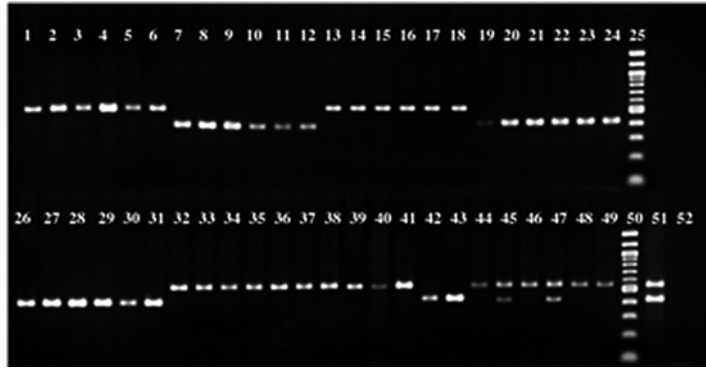


Fig. 2 Results of the amplification by PCR of the *T. melanosporum* mating type genes of different samples in multiplex PCR reactions. The expected sizes for the different mating type idiomorphs are 550 bp for *Mat 1-1* and 421 bp for *MAT 1-2*. The wells 1–24 present results for ectomycorrhiza samples; 26–39 present PCR results for ascocarps; and wells 40–49 present soil samples results; in wells 25 and 50 is shown a 100 bp DNA Ladder, which presents size bands of (bp): 100, 200, 300, 400, 500/517, 600, 800, 900, 1000, 1200, 1517. Well 51 presents the results of the positive control and well 52 the negative control (water). For ectomycorrhizal and ascocarps samples, we expect to find only amplification for one mating type. In the case of soil samples we can find one or both mating types, as free-living soil mycelium of different strains can be found in the same sample (as in wells 45 and 47)

reduced to 50 μ L of each solution and to 25 μ L when extracting DNA from ectomycorrhizal tips.

4. When extracting soil DNA by using the Fast DNA Spin kit for soil (MP Biomedicals, Illkirch, France), we recommend doing it with some modifications in the manufacturer's instructions:
 - (a) Add 4–6 mg PVPP (polyvinylpolypyrrolidone) when adding MT Buffer (**step 3**).
 - (b) In **steps 7** and **8** use a 2 ml microcentrifuge tube.
 - (c) Between **steps 9** and **10** wash each sample 10 times with a solution of 5.5 M guanidine thiocyanate (pH 7) by adding 1 mL of solution, mix by vortexing for 2 s, followed by a centrifugation of 5 s. Remove the supernatant except on the last washing and continue with the protocol [10].
5. For soil samples usually it is better to check if the DNA extraction worked by checking the presence of fungal species in soil samples by amplifying soil DNA with the fungal universal primers ITS1f (5'-CTT GGT CAT TTA GAG GAA GTA A-3') and ITS4 (5'-TCC TCC GCT TAT TGA TAT GC-3') with a denaturation step at 94 $^{\circ}$ C for 4 min, followed by 35 cycles of denaturation at 94 $^{\circ}$ C for 30 s, annealing at 55 $^{\circ}$ C for 30 s and extension at 72 $^{\circ}$ C for 1 min and 30 s, and a final extension step of 5 min at 72 $^{\circ}$ C. Using the same PCR mix conditions described in point 3.3.

6. PCR reactions can be performed by using other buffers, i.e., 25 μL reaction mixture containing 2.5 μL of 10 \times reaction buffer (Sigma-Aldrich Co. LLC, St Louis, MO, USA), 2.5 mM MgCl_2 , 200 μM of each dNTP, 200 nM of each primer, 1 U of Taq polymerase (Sigma-Aldrich Co. LLC, St. Louis, MO, USA), 16.55 μL of deionized sterile water, and 5–20 ng of DNA.
7. For the free-living soil mycelia, sometimes it is better to perform two independent PCR reactions for each MAT primer pairs (200 nM of each primer). Each reaction is done with the same conditions as the reaction described before.
8. The REDExtract-N-AmpTM PCR ReadyMixTM contains a loading dye buffer (pink), so that no addition of it is necessary for electrophoresis. If using different PCR buffer, add 1/6 volume of 6 \times Gel DNA Loading Dye (New England Biolabs, Ipswich, Massachusetts, USA) to your PCR sample before loading it into the gel. E.g.: 1 μL of 6 \times gel loading dye + 5 μL of PCR product.

Acknowledgments

The UMR1136 is supported by a grant overseen by the French National Research Agency (ANR) as part of the Investments for the Future Programme (ANR-11-LABX-0002-01, Lab of Excellence ARBRE). This study benefited from ANR SYSTERRA SYSTRUF (ANR-09-STRA-10). We would like to thank Francesco Paolocci and his team at the CNR of Perugia (Italy), and Francis Martin and François Le Tacon for their constructive advice and helpful discussions. We are grateful to Christophe Robin for allowing us to work in his truffle orchards.

References

1. Murat C (2015) Forty years of inoculating seedlings with truffle fungi: past and future perspectives. *Mycorrhiza* 25:77–81. doi:[10.1007/s00572-014-0593-4](https://doi.org/10.1007/s00572-014-0593-4)
2. Olivier J, Savignac J, Sourzat P (2012) Truffe et trufficulture. FANLAC Editions, Périgueux, France
3. Martin F, Kohler A, Murat C et al (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* 464:1033–1038
4. Rubini A, Belfiori B, Riccioni C et al (2011) Isolation and characterization of MAT genes in the symbiotic ascomycete *Tuber melanosporum*. *New Phytol* 189:710–722
5. Murat C, Rubini A, Riccioni C et al (2013) Fine-scale spatial genetic structure of the black truffle (*Tuber melanosporum*) investigated with neutral microsatellites and functional mating type genes. *New Phytol* 199:176–187. doi:[10.1111/nph.12264](https://doi.org/10.1111/nph.12264)
6. Zambonelli A, Salomoni S, Pisi A (1993) Caratterizzazione anatomo-morfologica delle micorrize di *Tuber* spp. su *Quercus pubescens* Will. *Micol Ital* 3:73–90
7. Rauscher T, Agerer R, Chevalier G (1995) Ektomykorrhizen von *Tuber melanosporum*, *Tuber mesentericum* und *Tuber rufum* (Tuberales) an *Corylus avellana*. *Nova Hedwigia* 61:281–322

8. Rubini A, Paolocci F, Granetti B et al (1998) Single step molecular characterization of morphologically similar black truffle species. *FEMS Microbiol Lett* 164:7–12. doi:[10.1016/S0378-1097\(98\)00183-9](https://doi.org/10.1016/S0378-1097(98)00183-9)
9. Paolocci F, Rubini A, Granetti B et al (1999) Rapid molecular approach for a reliable identification of *Tuber* spp. ectomycorrhizae. *FEMS Microbiol Ecol* 28:23–30
10. Luis P, Walther G, Kellner H et al (2004) Diversity of laccase genes from basidiomycetes in a forest soil. *Soil Biol Biochem* 36:1025–1036

Stable-Isotope Probing RNA to Study Plant/Fungus Interactions

Amandine Lê Van, Marie Duhamel, Achim Quaiser,
and Philippe Vandenkoornhuys

Abstract

The use of stable-isotope probing (SIP) allows tracing specific labeled substrates into fungi leading to a better understanding of their role in biogeochemical cycles and their relationship with their environment. Stable-isotope probing combined with ribosomal RNA molecule, conserved in the three kingdoms of life, and messenger RNA analysis permits the linkage of diversity and function. Here, we describe two methods designed to investigate the interactions between plant and its associated mycorrhizal compartment by tracing carbon flux from the host plant to its symbionts.

Key words Stable-isotope probing (SIP), RNA, qRT-PCR, Fungal plant symbiont, Carbon thirteen, Carbon transfer

1 Introduction

To identify the actors of a particular biogeochemical process performed *in natura*, different strategies have been developed during the last decade. While metagenomic analyses allow to address the functional potential of a microbial community based on massive sequencing of nucleic acids extracted from an environmental sample, more focused molecular tools have been successfully developed. Among these methods, stable-isotope probing (SIP) combined with high-throughput sequencing represents one of the most powerful tools [1]. The principle is simple (Fig. 1). The consumption of a stable isotope labeled substrate is reflected within the cellular compounds (i.e. DNA, RNA, proteins, lipids) modifying their densities and allowing the fractionation of metabolized compounds. Subsequently, the microorganisms involved in a targeted function can be identified from a complex environmental sample.

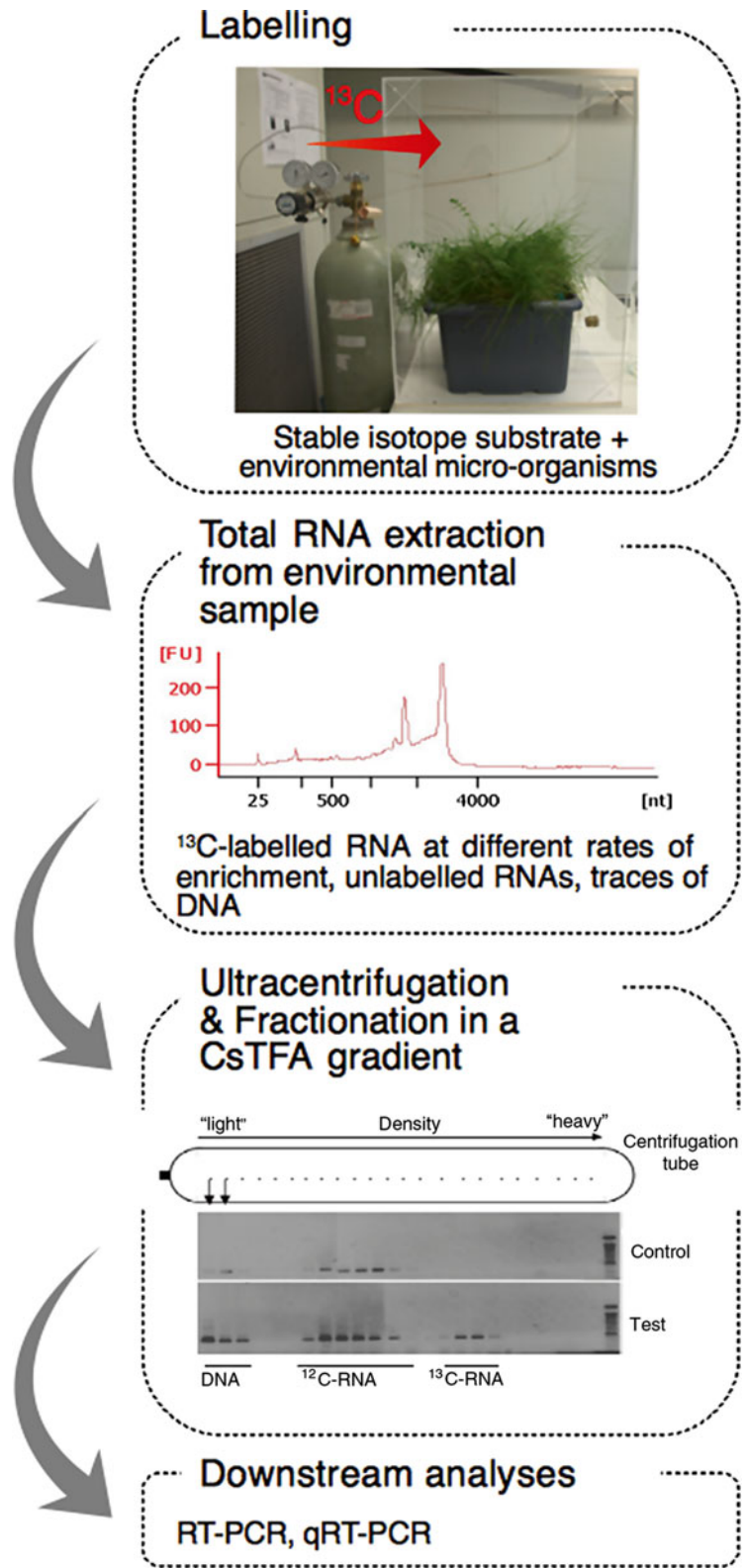


Fig. 1 The consumption of a stable-isotope labeled substrate is reflected within the cellular compounds

1.1 Advantages and Limits among SIP Strategies

One major challenge of SIP studies is finding a compromise between the quantity of labeled substrate used and incubation time that corresponds best to environmental conditions. SIP was first applied using phospholipid-derived fatty acids (PLFA) [2] as well as DNA-labeled analysis [1, 3]. In these cases, cells must undergo lipid biosynthesis or DNA synthesis respectively to incorporate labeled molecules. Because of the limited resolution of taxonomic assignment of PLFAs, SIP-PLFAs method was mainly used to highlight groups of microorganisms involved in a process. Conversely, DNA labeling combined with ribosomal RNA gene analysis allows detailed taxonomic affiliation, but requires relatively long incubation times due to the need of cell division and high levels of enrichment [4, 5]. A more direct method that allows mitigating these technical problems is the use of SIP-RNA [6].

1.2 The Different SIP-RNA Applications

SIP analyses have mostly used ^{13}C -enriched compounds such as methanol [2, 3], phenol [6] and trace molecules in soils such as atrazine [7]. Virtually all organic molecules that can be enriched in ^{13}C when chemically synthesized or biologically produced in vitro can be used within SIP-RNA-based study. SIP-RNA approaches have been used to analyze interactions and behavior between plants and root symbionts [8, 9]. In these two studies, after a $^{13}\text{CO}_2$ pulse labeling at atmospheric concentration, the carbon flux from the host plant to its symbionts (i.e. through ^{13}C -enriched photosynthates) was estimated under the assumption that the more the symbiont receives photosynthates, the more it cooperates with its host plant. The two strategies and related methodologies developed and validated are provided below.

2 Material

2.1 Material Common to Methods 1 and 2

2.1.1 General Supplies

1. Ultrapure nuclease-free water (Sigma-Aldrich).
2. Filter tips for RNA manipulation DNase/RNase-free.
3. Microtubes DNase/RNase-free.
4. 70 and 96 % ethanol.
5. Ice-cold isopropanol (>99 %).
6. Liquid nitrogen.
7. Precision balance (various manufacturers).
8. Labtop centrifuge (Eppendorf 5417R).
9. Thermocycler.

2.1.2 Isopycnic Ultracentrifugation

1. Cesium trifluoroacetate solution, CsTFA 2 g/mL (GE Healthcare), store at 4 °C.

2.1.3 RNA Extraction and Quantification

1. Micropestle (Eppendorf® micropestle for 1.2–2 mL tubes) or bead beater and beads.
2. RNeasy Plant Mini kit (Qiagen) (*see Note 1*).
3. Electrophoresis supplies (Ladder, 6× loading dye, 1 % Agarose gel, 0.5× Tris-Borate-EDTA, Ethidium Bromide) and electrophoresis machine or alternatively a Bioanalyser RNA 6000 Pico chip, RNA 6000 Pico kit (Agilent).
4. Nanodrop (Thermoscientific).

2.2 Specific Material of Method 1

2.2.1 ¹³C Labeling and Sampling

1. ¹³C-labeled substrate: ¹³C-CO₂/N₂/O₂ gas mix, ratio 0.033/78/21.967, 99 % ¹³C (CortecNet). In our conditions approximately 2 m³ of gas was needed for one labeling experiment.
2. A hermetic box (W50×L50×h70 cm) connected to the gas cylinder by a two-stage gas pressure regulator on one side (high position) and with an opening on the opposite side (low position) (Fig. 1) (*see Note 2*).
3. Triton X100.

2.2.2 Isopycnic Ultracentrifugation

1. Beckman Coulter Optima L-90k preparative ultracentrifuge.
2. Beckman 90Ti rotor.
3. Quick seal Ultra-clear centrifugation tubes, 13.5 mL (Beckman).
4. OptiSeal™ tube kit (Beckman).
5. Beckman tube topper (Beckman).
6. Beckman tube caps (Beckman).
7. Clamp attached on a retort stand (Fig. 2).
8. 5 mL syringe.
9. 1 mL syringe.
10. Needles 0.5×16 mm.
11. Custom-made guide punctuated every 0.1 in (2.54 mm) (Fig. 2).

2.2.3 RT-PCR

1. Titan One tube RT-PCR kit (Roche).
2. Primers NS31 and AM1 0.4 μM (each) targeting SSU rDNA of arbuscular mycorrhizal (AM) fungi (*see Note 3*) [10].

2.3 Specific Material of Method 2

2.3.1 ¹³C Labeling and Sampling

1. Biological material: Plants colonized by AM fungi.
2. Labeling chamber computer-controlled closed-system climate chambers (“Espas”).
3. CO₂ scrubber.
4. ¹³CO₂ in a pressurized cylinder (99 atom % ¹³C, 1 atom % ¹²C; Isotec).

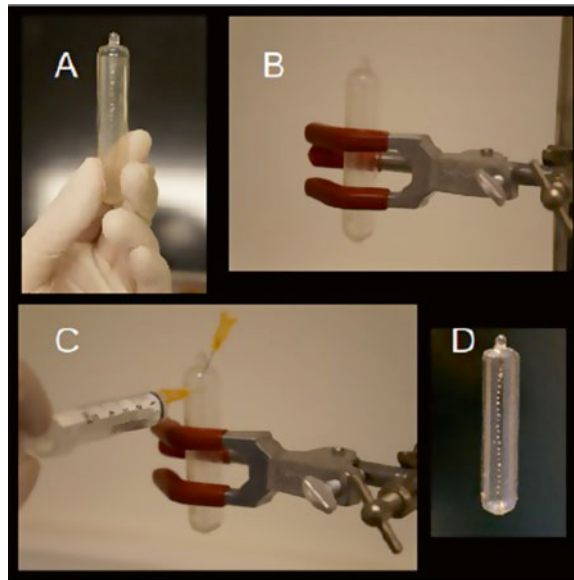


Fig. 2 Clamp attached on a retort stand

5. $^{12}\text{CO}_2$ in a pressurized cylinder.
6. Sieve.

2.3.2 Ultracentrifugation

1. De-ionized 100 % formamide (Sigma—store at $-20\text{ }^\circ\text{C}$) (*see Note 4*).
2. Ultracentrifuge (Sorvall discovery m120 SE micro-ultracentrifuge (Thermo Fisher Scientific), S120 VT fixed angle titanium vertical rotor).
3. 2 mL ultracentrifuge vials (Sysmex).
4. Capping system for ultracentrifuge vials (plastic and metallic caps).

2.3.3 Fractionation

1. 10 mL syringe.
2. Flexible plastic tubing (about 60 cm long).
3. Green 21 gauge 40 mm (1.5 in.) needles and Blue 23 gauge 25 mm (1 in.) needles.
4. Fractionator (Harvard Apparatus).
5. Vial carrier.
6. Dry ice.

2.3.4 cDNA Synthesis by Reverse Transcription

1. $5\times$ Buffer.
2. 10 mM dNTP mix.
3. $1\text{ }\mu\text{g}/\mu\text{L}$ (microgram/microliter).
4. M-MLV (Moloney Murine Leukemia Virus) reverse transcriptase ($200\text{ U}/\mu\text{L}$, Promega).

2.3.5 *Real-Time
Quantitative PCR (qPCR)*

1. Specific primers for qPCR.
2. LightCycler 2.0 instrument (Roche).
3. LightCycler TaqMan chemistry (LightCycler TaqMan Master).
4. 20 μ L Lightcycler glass capillaries.

3 Method

**3.1 Method 1:
Identification of Fungi
Interacting with Their
Host Plant by SIP-RNA**

The aim of this method is to identify potential symbionts in plant roots by discriminating active fungi that received labeled plant photosynthates from other facultative transient endophytes.

3.1.1 ¹³C Labeling

1. Before labeling, take a control sample of your plants to check the natural presence of ¹³C in roots and to check fractionation in the isopycnic ultracentrifugation gradient (see below).
2. Place plants in the box and apply an air flush during 5 min then decrease the air-flow at 25 pound per square inch (psi). After 1 h of labeling decrease air-flow at 15 psi for 4 h. If available, an infrared gas analyzer can be used to accurately determine the air-flow (\approx 5 L/min) and CO₂ delivery by measuring the CO₂ concentration in the vent gas. Time duration of labeling should be adapted for each experiment accordingly to the incorporation rate of your system and your target (*see Note 5*).
3. Immediately after labeling take core samples. Roots are washed in tap water, three times in 0.1 % Triton X100, and five times in sterilized distilled water. For each plant, all the root system is sampled that represents approximately the volume of half an eppendorf tube (1.5 mL) or 200 mg. After the washing, roots are frozen in liquid nitrogen and stored at -80 °C until used.
4. We recommend determining isotopic signature ($\delta^{13}\text{C}$) by isotopic ratio mass spectrometry in dried roots before to go further. This analysis can also be used to determine the kinetic of carbon incorporation and accurately choose sampling times.

3.1.2 *RNA Extraction
and Quantification*

1. Cleaned roots are grinded to powder either using liquid nitrogen and micro pestles or using a bead beater. The material has to keep frozen.
2. Extract total RNA from plant roots following the provider's instructions (RNeasy Plant Mini Kit, "Purification of total RNA from plant cells and tissues and filamentous fungi" protocol) or use any other validated protocol. For one RNA extraction, approximately 30 mg (fresh weight) of roots are needed. Skip the DNA-digestion step if you want to analyze the fungal community diversity from DNA and RNA (Fig. 1).

3. Quantify your total RNAs using any method of your convenience. The use of the Bioanalyser RNA 6000 pico chip allows checking for RNA quality using 2 μ L only of your extraction.
4. Quantify RNA using Nanodrop (Thermo Scientific).
5. Store RNA at -80°C until use.

3.1.3 Isopycnic Ultracentrifugation

Follow carefully all safety instructions for ultracentrifugation. A cleaning of the rotor might be done before starting. To do so, see the instructions of the rotor provider.

1. Cool down the ultracentrifuge and the rotor at 4°C .
2. Dilute your CsTFA solution with nuclease-free water to obtain a starting density of 1.8 g/mL. Prepare a solution at the desired density for all tubes and then aliquot your working solution. For one tube add 2.993 mL of water to 12 mL of CsTFA at 2 g/mL.
3. While the ultracentrifuge cool down, fill an even number of ultracentrifuge tubes with CsTFA (store at 4°C) using a 5 mL syringe and needle. Fill the tube until the base of the dome-top (*see Note 6*). Avoid air bubbles by slowly filling; any air bubble should be removed. Work on ice. Avoid preparing more than six tubes at a time, as RNA is fragile and must be handled quickly after ultracentrifugation.
4. Load ~ 50 ng of RNA on the surface of the tube (*see Note 7*).
5. Add about 1 mL of CsTFA.
6. Balance your tubes pair-wise using a precision balance and add CsTFA solution at 1.8 g/mL to equal weights. Balance tubes to the nearest 10 mg using micropipettes. The volume should not exceed the base of the tube neck.
7. Seal your tubes with the heat sealer following the manufacturer's instructions. Insure that your tubes are well sealed by gently pressing them. The seal must be straight to allow correct positioning of the tube cap.
8. Place the tubes in the rotor and note their position. Place the tube caps.
9. Centrifuge for 48 h at 4°C and at 45,000 rpm ($173,192\times g$ at the maximum radius— $77,427\times g$ at the minimum radius) in a 90Ti rotor with maximum acceleration and no brake (*see Note 8*).

3.1.4 Fraction Recovery

1. Proceed to fraction recovery immediately after ultracentrifugation. Manipulate the tubes gently to not disturb the gradient. Perform only one tube at a time. Remove the tube from the rotor with forceps. Clean the tube wall with ethanol 70 %. Place the guide on the tube and firmly hold them in the clamp (Fig. 2). Place a beaker under your tube to collect wastes.

2. Pierce the top of the tube with a needle not connected to a syringe to depress the tube and allow air influx. Do not remove this needle.
3. Puncture the wall with the needle by starting from the top of the tube. The aperture of the needle should be oriented toward the top. Collect about 1 mL of the first fraction that contains DNA (*see Note 9*). Then, collect the other 23 fractions (approximately 0.5 mL per fraction) from the top to the bottom of the tube. Make sure to collect the meniscus (*see Note 9*) to limit contaminations between fractions. Progressively remove needle after sampling and place each fraction into a nuclease-free microfuge tube of 1.5 mL. The 10th and 17th fractions contain ^{12}C RNA and ^{13}C RNA respectively (*see Note 10*).
4. Add two volumes of ice-cold isopropanol. Mix by inversion and place the tubes at $-20\text{ }^{\circ}\text{C}$ for 3 h. The tubes can be kept overnight at $-20\text{ }^{\circ}\text{C}$.
5. Check you gradient fractionation by collecting all fractions of the blank tube (without nucleic acid). Weight each fraction using a precision balance. Labeled and unlabeled RNA are expected at buoyant densities of 1.82–1.85 g/mL and 1.78–1.80 g/mL respectively. Notice that a gradient of enrichments in ^{13}C usually exists within the RNA population.

3.1.5 RNA Precipitation (Work on Ice)

1. Centrifuge the tubes for 20 min at maximum speed ($20,000\times g$) at $0\text{ }^{\circ}\text{C}$ using the labtop centrifuge. Note the position of your tube to know the pellet location as it will not be visible. Carefully remove the supernatant without touching the pellet side.
2. Wash the pellet with 180 μL of ice-cold isopropanol. Centrifuge for 15 min at maximum speed ($20,000\times g$). Remove the supernatant. Centrifuge at maximum speed for 5 min and remove the last drops with a micropipette. Air dry at room temperature for maximum 5 min. Add 25 μL of ultrapure nuclease-free water. Immediately proceed to the RT-PCR. Alternatively, do not add water and store your dried pellets at $-80\text{ }^{\circ}\text{C}$.

3.1.6 PCR and RT-PCR

Run a PCR if you analyze the DNA fraction (mixture of ^{12}C and ^{13}C DNA) using any validated protocol. For RT-PCR use 4 μL of RNA in a final volume of 50 μL . Follow Titan-One tube manufacturer's instructions (Roche) (*see Note 11*). Annealing temperature for our primers is $58\text{ }^{\circ}\text{C}$ for 1 min.

3.2 Method 2: Analysis of the Carbon Transfer Intensity from the Plant to Fungi by SIP-RNA

The aim of this method is to assess which arbuscular mycorrhizal (AM) fungus receives more carbon from the plant when several fungi are competing for carbon resource within the same root system. It involves the need of specific primers for each of the fungal strains.

3.2.1 $^{13}\text{CO}_2$ Labeling

After sterilizing and germinating plant seeds, inoculate them with several AM fungi competing for plant carbohydrate resources. Time of growth will depend on the plant used. For our SIP experiments [9], *Medicago truncatula* had to be grown for 10 weeks.

1. Acclimate the plants colonized by AM fungi into the labeling chamber for 48 h before labeling.
2. During the night period before labeling and in accordance with the $^{12}\text{CO}_2$ respiration of the plant used in the experiment, remove $^{12}\text{CO}_2$ using a CO_2 scrubber.
3. One hour before the start of the day period, inject $^{13}\text{CO}_2$ using a pressurized cylinder (99 atom % ^{13}C , 1 atom % ^{12}C ; Isotec).
4. Introduce $^{13}\text{CO}_2$ at the atmospheric concentration, day/night period: 16/8 h, day/night temperature: 21 °C/17 °C, irradiation at plant height: 700 $\mu\text{mol}/\text{m}^2/\text{s}$, 80 % relative humidity. These parameters should be adapted for each experiment accordingly to the incorporation rate of the biological system.
5. Maintain the CO_2 level in the chamber at 400 $\mu\text{L}/\text{L}$ by injecting $^{12}\text{CO}_2$ from a pressurized cylinder. For 6 h, a total CO_2 level ($^{12}\text{CO}_2 + ^{13}\text{CO}_2$) of 400 $\mu\text{L}/\text{L}$ CO_2 should be maintained.
6. After 6 h, open and flush the labeling chamber with fresh air to remove the labeled $^{13}\text{CO}_2$.
7. Close the labeling chamber and maintain the $^{12}\text{CO}_2$ level at 400 $\mu\text{L}/\text{L}$.

3.2.2 Root Harvesting

1. Harvest plants at the 6 h flushing period, at the 12 h and at the 24 h time point.
2. At each harvest, remove the aboveground plant parts.
3. Gently wash the root systems using sieves and distilled water.
4. Put roots onto towel paper to remove the water excess.
5. Homogenize, weigh, and place root aliquots in Eppendorf tubes.
6. Freeze them with liquid nitrogen.

RNA Extraction

see Method 1

3.2.3 Ultracentrifugation

1. CsTFA solution:
RNA gradient density should be a 1.8 g/mL starting density. For 1 \times 2.2 mL gradient (one ultracentrifugation vial/one RNA sample), in a falcon tube, mix: —1.86 mL of 1.99 g/mL CsTFA (Amersham) (always check the density of CsTFA by weighing 100 μL using a water calibrated pipette) (*see Note 12*).
 - 375 μL ultrapure nuclease-free water.
 - 75 μL of formamide (aliquot stock at -20 °C). The formamide should be added at the end and mixed before use.

2. Once the CsTFA solution is ready, check its density by weighing if 100 μL is 1.8 g.
3. Transfer 500 ng of RNA in 2 mL ultracentrifuge vials (Sysmex) pre-filled with the CsTFA volume needed to fill 2.2 mL of 1.8 g/mL CsTFA solution. The quantity of RNA and the quantity of CsTFA solution must be adjusted depending on the RNA concentration to obtain the exact same weight in each vial used in the same ultracentrifugation run. Avoid any air bubble. All the vials should be filled equally. Deposit the amount of RNA in the vial at the top of the gradient.
4. Include an extra vial without RNA in each ultracentrifugation batch for gravimetric estimation of density of each gradient fraction.
5. Put the plastic plug and the metallic cap on the vials. Seal the vials following the manufacturer's instructions (it should be really vertical). Sealed vials have to be cleaned using 70 % ethanol otherwise it can collapse in the centrifuge.
6. Put the sealed vials in the S120VT fixed angle titanium vertical rotor. Place the rotor vial caps.
7. Put the rotor in the Sorvall discovery m120 SE micro ultracentrifuge (Thermo FisherScientific).
8. Centrifuge vials for 48 h at 20 °C at a speed of 64,000 rpm (142,417 $\times g$ at the maximum radius—91,128 $\times g$ at the minimum radius) with minimum acceleration (4 min from rest to 5000 rpm) (869 $\times g$ at the maximum radius—556 $\times g$ at the minimum radius) and minimum deceleration (8 min from 5000 rpm to rest).

3.2.4 Fractionation

Once the centrifuge is stopped, remove very carefully the vials from the rotor. Do not shake them, minimize movement as much as you can to not disturb the gradient.

1. Clean the flexible plastic tubing and syringe first with 96 % ethanol and then with ultrapure nuclease-free water. Attach the 10 mL syringe to the first end of the flexible plastic tubing and a blue needle to the other end. Put the syringe filled with ultrapure nuclease-free water in the fractionator (syringe pump, Harvard Apparatus) (Fig. 3).
2. Fix the vial to the carrier.
3. Before putting the blue needle to the top of the vial, let the water go through the plastic tubing until it drops from the needle to avoid any air bubbles in the tube. Connect the upper needle (blue) horizontally to the top of the vial. Then plug a green needle vertically at the bottom in the middle of the vial (Fig. 3).
4. Start the fractionator (speed 120–320 $\mu\text{L}/\text{min}$, depending on how experienced you are). This leads to a continuous flow of fractions from the lower needle.

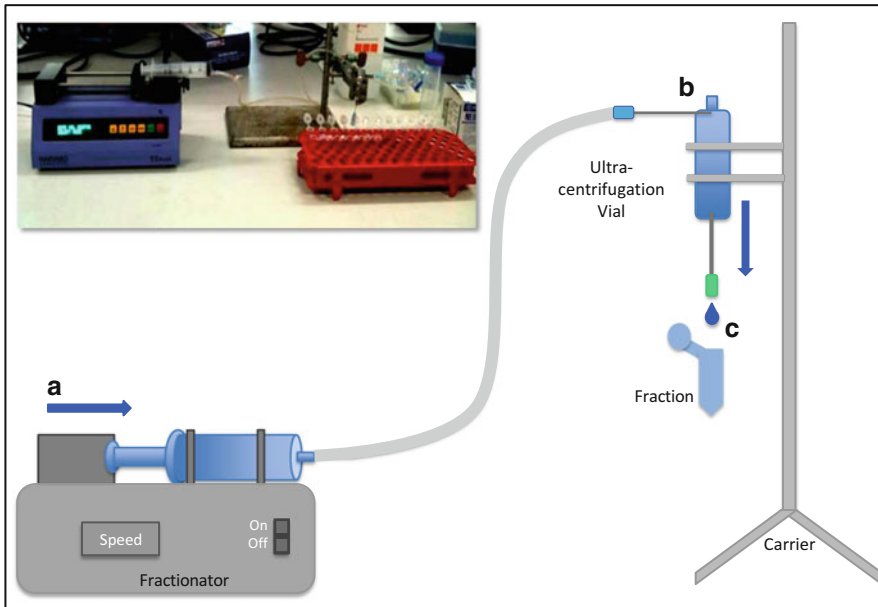


Fig. 3 Syringe pump, Harvard Apparatus

5. Collect 20 fractions, amounting to 100 μL per fraction in Eppendorf tubes. You can see that the entire gradient has been fractionated, as the water drops are bigger than the CsTFA drops.
6. Add 200 μL of ice-cold isopropanol before doing the next sample (first step from precipitation) and put the collected fractions on dry ice.
7. Always check that there is enough water in the syringe for the next sample. Take two new needles (one blue and one green) for each new sample.
8. Fractionate the extra vial for gravimetric estimation of density as above and weigh 100 μL aliquots of each fraction to determine the gradient. Beforehand, always calibrate the pipette with RNase-free water.

Do the fractionation as quick as possible as the gradient relaxes through time.

3.2.5 RNA Precipitation (Work on Ice)

1. Add 200 μL of ice-cold isopropanol to each 100 μL fractions (this step is already done just after the fractionation).
2. Incubate at $-20\text{ }^{\circ}\text{C}$ for 30 min at least.
3. Centrifuge for 20 min at maximum speed at $20,000\times g$ at $4\text{ }^{\circ}\text{C}$.
4. Remove supernatant with pipette and add a further 150 μL of ice-cold isopropanol.
5. Spin at $20,000\times g$ for 5 min at $4\text{ }^{\circ}\text{C}$ and remove the supernatant with a pipette.

6. Air-dry the samples in a laminar flow cabinet or in a vacuum desiccator at 4 °C.
7. Resuspend RNA pellets (30 min at 30 °C) in 10 µL of ultra-pure nuclease-free water.

3.2.6 *cDNA Synthesis by Reverse Transcription*

Do a reverse transcription on RNA from each fraction:

1. Incubate 5 µL of RNA template at 70 °C for 5 min. Chill on Ice.
2. Make a mix with, for each sample:
 - 5 µL of 5× reaction buffer.
 - 1.5 µL of 10 mM dNTP mix.
 - 0.5 µL of 1 µg/µL random hexamers.
 - 1 µL of M-MLV reverse transcriptase (200 U/µL).
 - 12 µL of ultrapure nuclease-free water.
3. Add this mix to each RNA template to obtain a final volume of 25 µL.
4. Incubate for 5 min at 25 °C followed by 60 min at 42 °C.
5. Terminate the reaction by heating at 70 °C for 15 min.
6. Check the cDNA on electrophoresis gel or with a Bioanalyser (Agilent).

3.2.7 *Real-Time Quantitative PCR (qPCR)*

Do a qPCR on each fraction:

1. Perform qPCR in 9 µL reactions, using the LightCycler 2.0 instrument, LightCycler TaqMan chemistry (LightCycler TaqMan Master) and 20 µL-Lightcycler glass capillaries.
2. Use a final concentration of 0.5 µM of primers, 0.11 µM of hydrolysis probe and 1.8 µL of Roche Master Mix.
3. Include 2.25 µL of cDNA template in each reaction.

3.2.8 *Statistical Analyses of Peak Fronts*

Variation in host plant C allocation is calculated based on differences in peak front among the inoculated AM fungal species. Peak front is the density (in mg/mL) of the heaviest RNA fraction of each of the AM fungal species. Peak front in the heavier fractions of the density gradient means a higher ¹³C enrichment, indicating a preferential C allocation to that particular AM fungal species. These peak front positions can be compared to each other. For this particular application, the number of replicates should be above 10 to get enough statistical powerfulness.

1. Determine peak front for each sample:

To determine peak front differences among the AM fungal species within each plant root sample, first measure the abundance of each AM fungal species (targeted gene copy number)

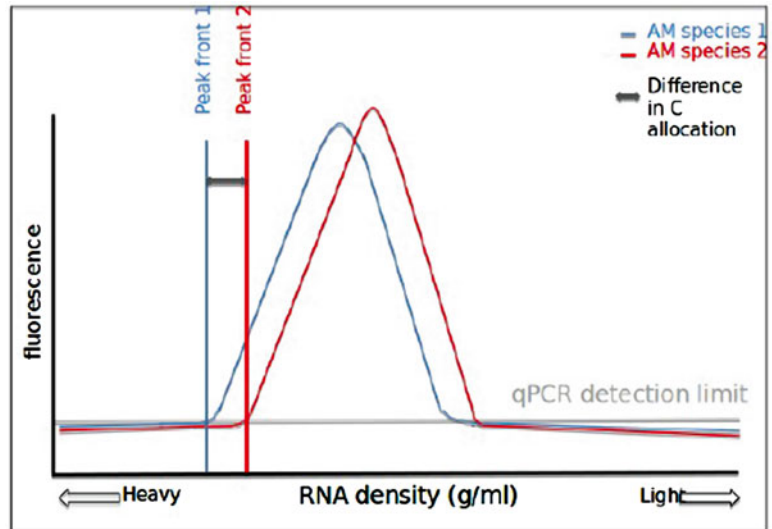


Fig. 4 Peak front is the fraction where the Gaussian regression curves cut through the detection limit of the qPCR assay

in each RNA density fraction by using qPCR with species-specific markers. Then, construct Gaussian regressions across the different fractions for each AM fungal species. Peak front is the fraction where the Gaussian regression curves cut through the detection limit of the qPCR assay (Fig. 4).

2. Measure the preferential C allocation:

To determine differences in ^{13}C enrichment of the AM fungal species, perform pair-wise comparisons of peak front position for all pairs of AM fungal species. Using non-parametric sign test, calculate the differences in peak front positions on the density gradient. Values different from zero indicate a preferential C allocation to one of the AM fungal species. Values equal to zero indicate no preferential allocation.

3. Data analysis of the preferential allocation:

To further confirm the results, a parametric generalized linear model (GLM) analysis can be performed. For each replicate and each AM fungal species pair, first calculate differences in peak front positions between AM fungal species, and then produce a GLM to test the modalities. Use the Akaike criteria (AIC) to select the best possible GLM and the best data transformation. Data are modeled by a saturated model. The relative importance of a given interaction term is estimated after removal of this term from the saturated model. Fisher tests are used to assess the statistical significance of a given interaction term. These statistical analyses can be done using R (<http://www.r-project.org/>) or other statistical tools.

4 Notes (Table 1)

1. Use any validated protocol or kit for RNA extraction.
2. The labeling box can be custom made or an existing air-flow chamber can be used. A light meter can be used to check whether the box does not filtrate light that could reduce photosynthesis and jeopardize the experiment.
3. Other specific primers or universal primers targeting fungal 18S rDNA can be used.

Table 1
Troubleshooting table

Problems and notes	Comments and suggestions
RNA manipulation	Always clean pipettes, bench and vessels with ethanol and RNase decontamination solution beforehand. Avoid RNAase contamination by wearing RNase-free gloves and changing them regularly. Work on ice.
Tubes collapsing	Clean the rotor and the vials using 96 % ethanol. Pay attention to seal the vial really vertically. Samples placed in opposite rotor compartments should have the exact same specific gravity.
Disturbed gradient	Minimize the fractionation time thus, do work with limited number of tubes. Avoid sudden movements.
Amplification occurs in all fractions	Insure that no contamination occurs during the fractionation between fractions. To be sure that no contact between needles occurred remove the needle from the syringe before transferring the collected fraction in a nuclease-free microfuge tube (protocol of Subheading 2.2). Decrease the RNA-loaded quantity on the gradient.
Reverse transcription fails	Could be due to RNA degradation. Avoid any RNase contamination (see above). Check that all isopropanol from RNA precipitation step is completely evaporated before elution otherwise it can inhibit the RT reaction.
No amplification of labeled RNA occurs	Check if labeling was efficient by isotopic ratio mass spectrometry (IRMS) of dried roots. In our case a $\delta^{13}\text{C}$ of 30‰ was high enough to successfully extract, fractionate, and amplify the ^{13}C -enriched RNA. Alternatively, you can assess IRMS of purified RNA.
No amplification occurs following RT-PCR	Insure the RNA integrity. Increase the RNA quantity loaded on the gradient up to 250 ng (protocol of Subheading 2.2). Make sure your RT-PCR protocol is efficient with low RNA quantities. An increase of the number of cycle could improve the result. Do not forget to include a negative control.

These methods are not easy to handle. Training before starting should help.

4. Aliquot the formamide stock and freeze at $-20\text{ }^{\circ}\text{C}$. Once defrosted, the formamide is no longer deionized. Use under fumehood cabinet. Formamide is highly toxic. Read carefully the formamide safety rules.
5. Time duration of the experiment was determined based on previous experiments that showed maximum enrichment of microbial RNA 3 h after the end of a 6 h pulse [11]. To focus on the primary consumers of plant photosynthates the pulse was reduced to 5 h.
6. Do not overfill your tubes because you will have to add CsTFA to have equal weights among tube pairs before centrifugation.
7. We strongly recommend performing an unlabeled control tube without ^{13}C RNA to check if heavy fractions are not contaminated by ^{12}C RNA.
8. Without brake the deceleration takes approximately 50 min.
9. During the collection always keep your needle and syringe at a 90° angle with your tube. Proceed to small movements to explore the whole surface while you collect the fraction. At the end of the fraction sampling, turn the needle aperture toward the bottom and collect the meniscus. If the sampling is too hard check that the first needle allowing air influx is not obstructed.
10. Insure that ^{12}C RNA and ^{13}C RNA are located in these fractions in your conditions by collecting all fractions for the first assays. Amplification of each fraction will allow you to determine fraction of interest. After experimental settings, only fractions of interest can be collected. Collect approximately 0.5 mL of each fraction. Do not remove the needle until all the targeted fractions have been collected to avoid contaminations between fractions. Take over all your syringes and place each fraction into a nuclease-free microfuge tube of 1.5 mL.
11. The RT-PCR and PCR conditions must be optimized with low quantities of RNA templates.
12. If the density of CsTFA is different from 1.99 g/mL, use the following formula to calculate the volume of original CsTFA to use:

$$V_o = \frac{V_f}{1 + \frac{D_o - D}{D - D_x}}$$

where:

V_f : Final volume

V_o : Volume of original CsTFA in g/mL

D_0 : Density of original CsTFA in g/mL

D_x : Density of water in g/mL (0.998 g/mL at 25 °C)

D : Density of desired solution produced in g/mL (1.8 g/mL for RNA)

5 Conclusions and Prospects

These SIP-RNA methods will likely be developed within the incoming years to address questions related to the link between diversity and functions possibly to discover new metabolic pathways or processes mediated by microorganisms. One of the main technological prospects of SIP-RNA-based methods is likely the possibility to develop subtractive meta-transcriptomic analyses.

References

1. Radajewski S, Ineson P, Parekh NR et al (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403:646–649
2. Bull ID, Parekh NR, Hall GH et al (2000) Detection and classification of atmospheric methane oxidizing bacteria in soil. *Nature* 405:175–178
3. Morris SA, Radajewski S et al (2002) Identification of the functionally active methanotroph population in a peat soil microcosm by stable-isotope probing. *Appl Environ Microbiol* 68:1446–1453
4. Kalyuzhnaya MG, Lapidus A, Ivanova N et al (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol* 26:1029–1034
5. Kalyuzhnaya MG, Beck DAC, Chistoserdova L (2011) Functional metagenomics of methylo-trophs, vol 495, 1st edn, *Methods in methane metabolism*, Part B. Elsevier Inc, Amsterdam, pp 81–98
6. Manefield M, Whiteley AS et al (2002) RNA stable isotope probing, a novel means of linking microbial community function to phylogeny. *Appl Environ Microbiol* 68:5367–5373
7. Monard C, Vandenkoornhuysse P et al (2011) Relationship between bacterial diversity and function under biotic control: the soil pesticide degraders as a case study. *ISME J* 5:1048–1056
8. Vandenkoornhuysse P, Mahe S, Ineson P et al (2007) Active root-inhabiting microbes identified by rapid incorporation of plant-derived carbon into RNA. *Proc Natl Acad Sci U S A* 104:16970–16975
9. Kiers ET, Duhamel M, Beesetty Y et al (2011) Reciprocal Rewards Stabilize Cooperation in the Mycorrhizal Symbiosis. *Science* 333:880–882
10. Helgason T, Daniell TJ, Husband R et al (1998) Ploughing up the wood-wide web? *Nature* 394:431
11. Rangel-Castro JI, Killham K, Ostle N et al (2005) Stable isotope probing analysis of the influence of liming on root exudate utilization by soil microorganisms. *Environ Microbiol* 7:828–838

Chapter 10

Targeted Gene Capture by Hybridization to Illuminate Ecosystem Functioning

Céline Ribière, Réjane Beugnot, Nicolas Parisot, Cyrielle Gasc, Clémence Defois, Jérémie Denonfoux, Delphine Boucher, Eric Peyretailade, and Pierre Peyret

Abstract

Microbial communities are extremely abundant and diverse on earth surface and play key role in the ecosystem functioning. Thus, although next-generation sequencing (NGS) technologies have greatly improved knowledge on microbial diversity, it is necessary to reduce the biological complexity to better understand the microorganism functions. To achieve this goal, we describe a promising approach, based on the solution hybrid selection (SHS) method for the selective enrichment in a target-specific biomarker from metagenomic and metatranscriptomic samples. The success of this method strongly depends on the determination of sensitive, specific, and explorative probes to assess the complete targeted gene repertoire. Indeed, in this method, RNA probes were used to capture large DNA or RNA fragments harboring biomarkers of interest that potentially allow to link structure and function of communities of interest.

Key words Solution hybrid selection, Metagenomics, Metatranscriptomics, Microbial diversity, RNA probes, Next-generation sequencing

1 Introduction

Microbial communities show the greatest organisms diversity on earth and are key players for the functioning of all the ecosystems. For example, 1 g of soil may contain up to 10^9 bacterial cells [1] and assuming 3000 genes per single bacteria genome [2] and an average of 1000 bp per gene, such cells will thus represent up to 3×10^{15} bp. To explore such diversity, next-generation sequencing (NGS) technologies, especially Illumina systems, produce a great amount of sequence information (e.g. HiSeq 2500 produces six billion paired-end reads corresponding to 600 Gb of data). High-throughput sequencing greatly improved the resolution for microbial diversity description [3]. However, a substantial number of runs (6000) must be realized with a global cost of \$267 million to

produce a dataset representing onefold coverage of the microbial from 1 g of soil [4].

To reduce this biological complexity, barcoding is an efficient method [5] but cannot establish the link between the microbial structure and the realized functions limiting the understanding levels [6]. Furthermore, various PCR biases could alter these descriptions [7]. Recently, promising approaches, based on the SHS (Solution Hybrid Selection) capture method for the selective enrichment in a target-specific biomarker from metagenomic [8] and metatranscriptomic [9] (*see* Chapter 14 for metatranscriptomic application) samples have been developed (Fig. 1). First results have showed that this technology allows the identification of rare populations within the studied environment but also to participate to large DNA fragments reconstruction potentially allowing to link structure and function in microbial communities. The success of this innovating gene capture approach, however strongly depends on the determination of the best probe set while taking the biological question into account [10]. Consequently, capture probe design is of critical importance and should therefore consider multiple parameters in order to assess the complete targeted gene repertoire. In addition, to being sensitive and specific, probes must also anticipate genetic variations and thus must be able to detect known and unknown sequences in environmental samples. To design such explorative probes, three algorithms, PhylGrid, KASpOD and HiSpOD have been developed. PhylGrid is a large-scale probe design software linked to the EGI (European Grid Infrastructure) [11]. It is an improvement of the PhylArray algorithm presented in Milton et al. [12] which relies on initial multiple sequence alignments to define regular and explorative oligonucleotide probes for SSU rRNA genes. KASpOD is a web service dedicated to the design of signature sequences using a k-mer-based algorithm [13, 14]. PhylGrid and KASpOD software were used to define 74,003 probes of 25 mer targeting SSU rRNA genes from 2178 genera including Bacteria and Archaea. These probes are available using PhylOPDb, an online resource for a comprehensive phylogenetic oligonucleotide probe database [15]. Finally, the HiSpOD program allows designing both gene-specific and sequence-specific probes to target any functional biomarker [16]. All these software, developed in the context of microbial ecology, are then particularly appropriate for the design of highly sensitive, specific, and explorative probes in the context of gene capture by hybridization. Indeed, these probes, used for the gene capture molecular approach described below will ensure the selective enrichment of DNA or RNA (*see* Chapter 14 for metatranscriptomic application) from targeted phylogenetic or functional biomarker genes of interest in complex environments.

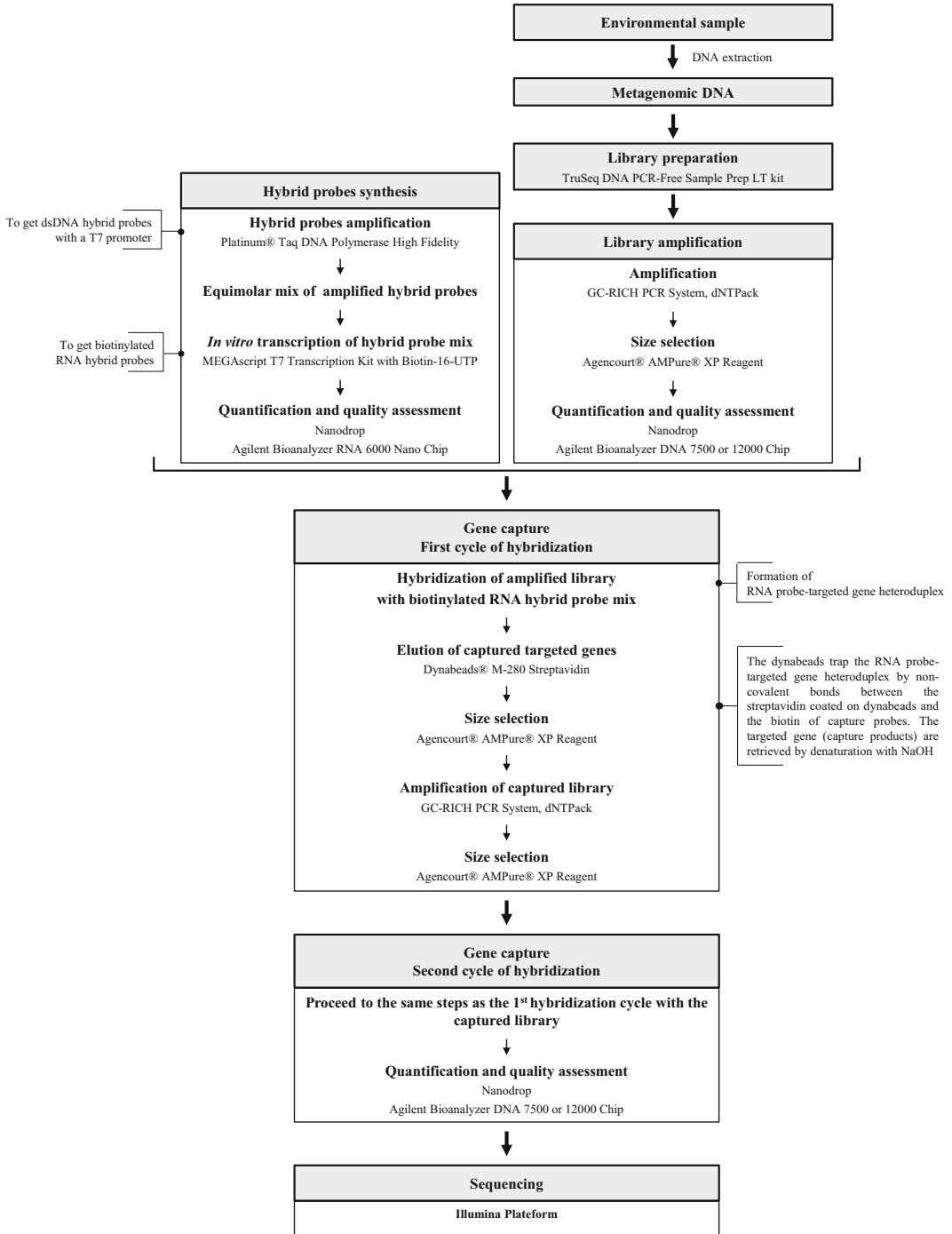


Fig. 1 Process workflow of targeted gene capture by hybridization

2 Materials

2.1 Reagents and Kits

1. Agencourt® AMPure® XP Reagent (Beckman Coulter, Brea, CA, USA).
2. Agilent DNA 7500 or 12000 Kit (Agilent Technologies, Santa Clara, CA, USA).
3. Agilent RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA, USA).
4. Biotin-16-UTP (Epicentre, Madison, WI, USA).
5. Dynabeads® M-280 Streptavidin (Life Technologies, Carlsbad, CA, USA).
6. GC-RICH PCR System, dNTPack (Roche Applied Science, Basel, Switzerland).
7. Glycogen (molecular biology grade).
8. MEGAscript® T7 Transcription Kit (Life Technologies, Carlsbad, CA, USA).
9. MinElute Gel Extraction Kit (Qiagen, Hilden, Germany).
10. MinElute PCR Purification Kit (Qiagen, Hilden, Germany).
11. Platinum® Taq DNA Polymerase High Fidelity (Life Technologies, Carlsbad, CA, USA).
12. QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).
13. RNeasy Plus Mini Kit (Qiagen, Hilden, Germany).
14. 10 mg/mL sheared salmon sperm DNA (Life Technologies, Carlsbad, CA, USA).
15. TruSeq DNA PCR-Free Sample Prep LT Set A or Set B (Illumina, San Diego, CA, USA).

2.2 Buffers and Solutions

All buffers and solutions could be prepared in laboratory under DNase/RNase-free conditions or purchased in general lab supplier. Prepare all solutions using ultrapure water (prepared by purifying deionized water to attain a sensitivity of 18 MΩ cm at 25 °C) and molecular biology grade reagents. Prepare and store all solutions and buffers at room temperature (unless indicated otherwise).

1. 100× Denhardt's solution: 2 % bovine serum albumin (BSA) (Fraction V), 2 % Ficoll 400, 2 % polyvinylpyrrolidone. Weigh 1 g BSA, 1 g Ficoll 400 and 1 g polyvinylpyrrolidone and transfer to a 50 mL graduated cylinder. Add water to a volume of 50 mL. Filter through a 0.2 μm syringe filter to sterilize. Divide into aliquots of 2 mL, and store at -20 °C.
2. 0.5 M ethylenediaminetetraacetic acid (EDTA): pH 8.0. Weigh 46.53 g EDTA and transfer to a 250 mL graduated cylinder. Add water to a volume of 150 mL. Mix and adjust pH with

NaOH. Make up to 250 mL with water. Sterilize by autoclaving (*see Note 1*).

3. 80 % Ethanol (*see Note 2*).
4. 5 M NaCl. Weigh 73.08 g NaCl and transfer to a 250 mL graduated cylinder. Add water to a volume of 250 mL. Sterilize by autoclaving.
5. 1 M NaOH. Weigh 1 g NaOH and transfer into the plastic beaker. Add water to a volume of 25 mL. Stir vigorously and as precaution, place the beaker on ice (*see Note 2*).
6. 0.1 M NaOH. Make a dilution at $1/10^6$ in nuclease-free water of 1 M NaOH solution (*see Note 2*).
7. 3 M sodium acetate: pH 5.2. Weigh 40.83 g sodium acetate and transfer to a 100 mL graduated cylinder. Add water to a volume of 80 mL. Adjust pH with glacial acetic acid. Make up to 100 mL with water. Filter through a 0.2 μm syringe filter to sterilize.
8. 10 % sodium dodecyl sulfate (SDS). Weigh 10 g SDS and transfer to a 100 mL graduated cylinder. Add water to a volume of 85 mL. Heat to 68 °C and stir with a magnetic stirrer to assist dissolution. Adjust pH to 7.2 by adding a few drop of concentrated HCl (36 %). Make up to 100 mL with water. Filter through a 0.2 μm syringe filter to sterilize.
9. 10 \times Tris-Borate-EDTA (TBE): 450 mM Tris-borate, 10 mM EDTA. Weigh 108 g Tris base, 27.5 g boric acid and transfer to a 1 L graduated cylinder. Add 10 mL of 0.5 M EDTA (pH 8.0) and water to a volume of 800 mL. Mix and adjust pH to 8. Make up to 1 L with water. Sterilize by autoclaving.
10. 10 \times Tris-EDTA (TE): 100 mM Tris-HCl, 10 mM EDTA. Weigh 0.24 g Tris base and 0.23 g EDTA and transfer to a 50 mL graduated cylinder. Add water to a volume of 35 mL. Adjust pH to 7.5 with HCl. Make up to 50 mL with water. Sterilize by autoclaving.
11. 1 M Tris-HCl: pH 7.5. Weigh 60.57 g Tris base and transfer to a 500 mL graduated cylinder. Add water to a volume of 350 mL. Adjust pH to 7.5 with HCl. Make up to 500 mL with water. Sterilize by autoclaving.
12. 20 \times SSC: 3 M NaCl, 0.3 M trisodium citrate. Weigh 17.53 g NaCl and 8.82 g trisodium citrate and transfer to a 100 mL graduated cylinder. Add water to a volume of 80 mL. Adjust pH to 7.0 by adding HCl. Make up to 100 mL with water. Sterilize by autoclaving.
13. 20 \times SSPE: 3 M NaCl, 0.2 M NaH_2PO_4 , 0.02 M EDTA. Weigh 17.53 g NaCl, 2.76 g NaH_2PO_4 , and 0.74 g EDTA and transfer to the cylinder. Add water to a volume of 80 mL. Adjust pH to 7.4 with NaOH. Make up to 100 mL with water. Sterilize by autoclaving.

14. Binding buffer: 1 M NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA. Transfer 10 mL of 5 M NaCl, 500 μ L of 1 M Tris-HCl (pH 7.5) and 100 μ L of 0.5 M EDTA (pH 8.0) to the cylinder. Make up to 50 mL with water. Filter through a 0.2 μ m syringe filter to sterilize (*see Note 2*).
15. 2 \times Hybridization buffer: 10 \times SSPE, 10 \times Denhardt's solution, 10 mM EDTA, 0.2 % SDS. Transfer 10 mL of 20 \times SSPE, 2 mL of 100 \times Denhardt's solution, 400 μ L of 0.5 M EDTA (pH 8.0), and 400 μ L of 10 % SDS to a 20 mL graduated cylinder. Make up to 20 mL with water. Filter through a 0.2 μ m syringe filter to sterilize. Divide into aliquots of 2 mL, and store at -20 °C.
16. Wash buffer n°1: 1 \times SSC, 0.1 % SDS. Transfer 2.5 mL of 20 \times SSC and 500 μ L of 10 % SDS to a 50 mL graduated cylinder. Make up to 50 mL with water. Filter through a 0.2 μ m syringe filter to sterilize (*see Note 2*).
17. Wash buffer n°2: 0.1 \times SSC, 0.1 % SDS. Transfer 250 μ L of 20 \times SSC and 500 μ L of 10 % SDS to a 50 mL graduated cylinder. Make up to 50 mL with water. Filter through a 0.2 μ m syringe filter to sterilize (*see Note 2*).

2.3 Oligonucleotides (Probes and Primers)

1. Hybrid probes. Purchase hybrid probes at 100 μ M. Adaptor sequences must be added to the 5' and 3' ends of the specific capture probes. These hybrid probes consist of 5'-ATCGCA CCAGCGTGT(X)CACTGCGGCTCCTCA-3', with X indicating the specific capture probe.
2. Primers for probe amplification. Purchase oligonucleotides at 100 μ M, T7-A 5'-GGATTCTAATACGACTCACTATAGG GATCGCACCAGCGTGT-3' and B 5'-CGTGGATGAGGAG CCGCAGTG-3'.
3. Primers for library amplification. Purchase oligonucleotides at 100 μ M, TS-PCR Oligo 1 5'-AATGATACGGCGACCAC CGAGA-3' and TS-PCR Oligo 2 5'-CAAGCAGAAGACGGCA TACGAG-3'.

2.4 Equipments

1. Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).
2. AFA System (Covaris, Woburn, MA, USA). One of these following items: M220, S220, S2 or E210 Focused-Ultrasonicator with the corresponding AFA Tubes (*see Note 3*).
3. DynaMagTM-2 Magnet (Life Technologies, Carlsbad, CA, USA).
4. HulaMixer[®] Sample Mixer (Life Technologies, Carlsbad, CA, USA) (optional).
5. Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA) or other systems for DNA quantification (e.g.:

Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) or other fluorometers).

6. Speed vacuum.
7. Thermal cycler (with heated lid).

3 Methods

3.1 Hybrid Probe Synthesis

1. First step of hybrid probe synthesis consists in amplification of oligonucleotide to obtain double-stranded DNA (dsDNA). Each amplification reaction should contain 5 μL of 10 \times high fidelity buffer, 1 μL of dNTPs (10 mM), 2 μL of MgSO_4 (50 mM), 1 μL of primer T7-A (10 μM), 1 μL of primer B (10 μM), 0.2 μL of Platinum® Taq DNA polymerase high fidelity, 38.8 μL of nuclease-free water and 1 μL of hybrid probe diluted at 10 μM (*see Note 4*). Include a negative control with 1 μL of nuclease-free water instead of 1 μL of hybrid probe. Use a thermal cycler with the following conditions: 2 min at 94 °C then 35 cycles of 30 s at 94 °C, 30 s at 58 °C and 20 s at 68 °C and a final elongation step at 68 °C for 5 min.
2. Check the probe amplification by electrophoresis on a 2 % agarose-TBE gel containing 0.5 \times syber safe (or comparable nucleic acid stain). Deposit 5 μL of amplified product (with loading buffer) (*see Note 5*). One lane is reserved for 100 bp DNA ladder. The gel is run in TBE buffer at 100 V for 45 min. The DNA is visualized on a UV transilluminator.
3. If one band at the expected size is observed, proceed to the purification of the remaining 45 μL of amplified products using the MinElute PCR Purification Kit, following the manufacturer's instructions. If two amplification bands are observed, deposit the remaining PCR product (i.e. 45 μL), excise with a clean razor blade or scalpel the band corresponding to the size of hybrid probes and proceed to their purification using the MinElute Gel Extraction Kit, following the manufacturer's instructions. The purified product is eluted in 15 μL of nuclease-free water (Fig. 2).
4. Evaluate the concentration of purified amplified hybrid probes with Nanodrop spectrophotometer.
5. For RNA synthesis, mix all hybrid probes in an equimolar amount taking into account the degeneracy of each probe. Each probe combination must be present in the same molecular amount (*see Note 6*). Validate the concentration of hybrid probe mix with Nanodrop spectrophotometer. Take 150 ng of hybrid probe mix, evaporate to dryness with a speed vacuum and resuspend in 4.75 μL of nuclease-free water. If the 150 ng

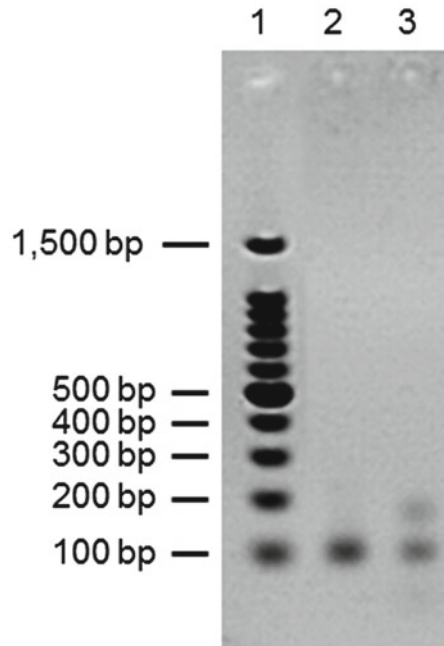


Fig. 2 Amplified hybrid probes electrophoresis on a 2 % agarose-TBE gel. *Lane 1* shows DNA ladder (100 bp). *Lane 2* shows one amplification band at the expected size, in this case at 112 bp corresponding to a hybrid probe of 50 bp with amplification primers T7-A (41 bp) and B (21 bp). *Lane 3* shows two amplification bands, one at the expected size (same as *lane 2*) and another one (at 150 bp) corresponding to aberrant amplification of the T7 promoter

of hybrid probe mix is in a volume lower than 4.75 μL , do not evaporate and adjust the volume to 4.75 μL with nuclease-free water.

6. The *in vitro* transcription (IVT) is realized with the MEGAscript[®] T7 Transcription Kit and using Biotin-16-UTP to produce biotinylated RNA. Each IVT reaction should contain 2 μL of 10 \times reaction buffer, 2 μL of ATP solution (75 mM), 2 μL of CTP solution (75 mM), 2 μL of GTP solution (75 mM), 1.5 μL of UTP solution (75 mM), 3.75 μL of biotin-16-UTP (10 mM), 2 μL of T7 enzyme mix, and the 4.75 μL of previously prepared hybrid probe mix (*see Notes 4 and 7*). Incubate at 37 $^{\circ}\text{C}$ for at least 6 h (or overnight).
7. Add 1 μL of TURBO DNase (include in the MEGAscript[®] T7 Transcription Kit) to each IVT reaction and incubate at 37 $^{\circ}\text{C}$ for 30 min.
8. For RNA precipitation, transfer the IVT reaction mix in a 1.5 mL microcentrifuge tube. Add 1/10^e volume of 3 M sodium acetate (pH 5.2), 3 volumes of cold 100 % ethanol, and 1 μL of glycogen (20 $\mu\text{g}/\mu\text{L}$). The reaction is incubated at

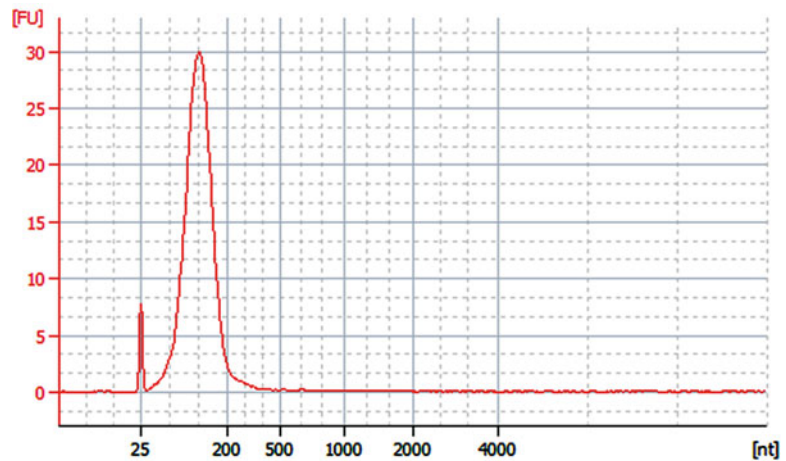


Fig. 3 Quality of biotinylated RNA probes assessed on Agilent Bioanalyzer RNA 6000 Nano chip. The electropherogram shows a resolved peak at the expected size of a hybrid probe of 50 bp with amplification primers T7-A (41 bp) and B (21 bp) (same probe as Fig. 2)

–80 °C for 30 min. Centrifuge at $18,000\times g$ for 15 min at 4 °C. Discard the supernatant and wash two times the pellet as following: add 500 μ L of cold 70 % ethanol, centrifuge at $18,000\times g$ for 10 min at 4 °C and discard the supernatant. Dry the pellet with a speed vacuum. Add 100 μ L of TE for pellet resuspension.

9. Proceed to the purification of biotinylated RNA probe mix with the RNeasy Plus Mini Kit following the “Appendix D: Purification of Total RNA Containing Small RNAs from Cells” instructions excepted the step D2 with the gDNA Eliminator spin column. Make two RNeasy Mini spin columns per hybrid probe mix (apply 50 μ L of biotinylated RNA probe mix on each column), elute the product in 40 μ L of nuclease-free water and pool the two obtained eluates.
10. Evaluate the concentration of purified biotinylated RNA probe mix with Nanodrop spectrophotometer. Assess their quality on an Agilent Bioanalyzer RNA 6000 Nano chip, according to the manufacturer’s instructions (Fig. 3).
11. Store at –80 °C.

3.2 Library Preparation (550 bp Insert)

The library is prepared for 550 bp insert using the TruSeq DNA PCR-Free Sample Prep LT kit by Illumina following the manufacturer’s instructions.

3.3 Library Amplification

1. Add 30 μ L of nuclease-free water to the library.
2. Proceed to the library amplification with the GC-RICH PCR System, dNTPack. Realize ten 50 μ L PCR reactions per library.

Each amplification reaction should contain 10 μL of 5 \times GC-RICH PCR reaction buffer, 2 μL of 25 mM MgCl_2 , 1 μL of PCR grade nucleotide mix, 1 μL of 25 μM TS-PCR Oligo 1, 1 μL of 25 μM TS-PCR Oligo 2, 29 μL of PCR grade water, 1 μL of GC-RICH enzyme mix and 5 μL of prepared library (*see Note 4*). Use the following thermal conditions: 4 min at 94 $^\circ\text{C}$ then 20 cycles of 30 s at 94 $^\circ\text{C}$, 1 min at 58 $^\circ\text{C}$ and 1 min 30 s at 68 $^\circ\text{C}$ and a final elongation step at 68 $^\circ\text{C}$ for 3 min.

3. Purify the amplified library using the QIAquick PCR Purification Kit following the manufacturer's instructions. Use one column of the kit for two PCR reactions pooled from a same library (i.e. five columns per library). The purified product is eluted in 50 μL of nuclease-free water.
4. Select the DNA fragments size with the Agencourt[®] AMPure[®] XP Reagent. Check that the eluate volume is equal to 50 μL . If necessary make up to 50 μL with nuclease-free water. Add 50 μL of AMPure beads, gently mix by pipetting and incubate for 5 min at room temperature (*see Note 8*). Place the tubes on the magnetic stand for at least 5 min at room temperature (until the supernatant is clear). Remove and discard the supernatant from each tube. Keep the tubes on the magnetic stand and wash two times the beads like following: add 500 μL of 80 % ethanol to each tube without disturbing them, incubate for 30 s at room temperature, and then remove and discard all of the supernatant from each tube. Take care not to disturb the beads. Remove and discard any remaining ethanol with a 10 μL pipette and let the beads air-dry for 5 min at room temperature. Add 50 μL of nuclease-free water to each tube. Remove the tubes from the magnetic stand. By pipetting, resuspend the beads by repeatedly dispensing the water over the bead pellet until it is immersed in the solution. Incubate for 2 min at room temperature. Place the tubes on the magnetic stand for at least 5 min at room temperature (until the supernatant is clear). Transfer all of the supernatant from each of the five tubes into a new 1.5 mL microcentrifuge tube (*see Note 9*).
5. Evaluate the concentration of purified amplified library with NanoDrop spectrophotometer. Assess its quality on an Agilent DNA 7500 or 12000 chip, according to the manufacturer's instructions (Fig. 4).
6. Store the purified amplified libraries at $-20\text{ }^\circ\text{C}$.

3.4 Gene Capture by Hybridization

1. Transfer 2.5 μg of sheared salmon sperm DNA and 500 ng of purified amplified library into a 0.2 mL PCR tube (*see Note 10*). Evaporate to dryness with a speed vacuum and resuspend in 7 μL of nuclease-free water.
2. Thaw an aliquot of 2 \times hybridization buffer, prewarmed it at 65 $^\circ\text{C}$ and transfer 20 μL into a 0.2 mL PCR tube (*see Note 11*).

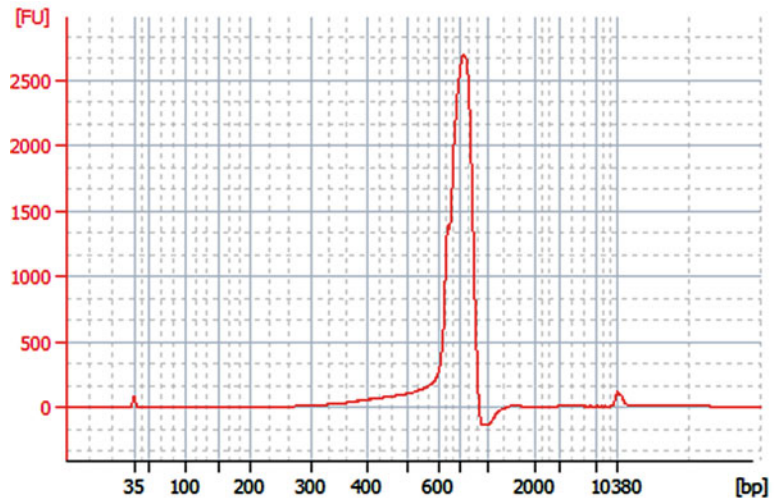


Fig. 4 Quality of amplified library (prepared for 650 bp insert) assess on an Agilent Bioanalyzer High Sensitivity DNA chip. The electropherogram shows a peak focused on 770 bp for the case of library prepared for 650 bp insert with 120 bp Illumina adaptors. With a library prepared for 550 bp, the same profile will be observed but with a peak focused on 670 bp

3. Thaw the biotinylated RNA probe mix on ice, transfer 500 ng into a 0.2 mL PCR tube and adjust the volume to 6 μ L with nuclease-free water (*see* **Notes 11** and **12**).
4. Incubate the salmon sperm DNA/purified amplified library (SL) mix in a thermal cycler with the following conditions: 95 $^{\circ}$ C for 5 min and 65 $^{\circ}$ C at 5 min.
5. Without removing SL mix from the thermal cycler, incubate at 65 $^{\circ}$ C the 0.2 mL PCR tube with 2 \times hybridization buffer. Add quickly 13 μ L of prewarmed 2 \times hybridization buffer to the tube containing the SL mix and homogenize by pipetting.
6. Always without removing SL mix from the thermal cycler, incubate at 65 $^{\circ}$ C the 0.2 mL PCR tube with the biotinylated capture probes mix. Add quickly 6 μ L of probe mix to SL mix (hybridization mix) and homogenize by pipetting. Incubate at 65 $^{\circ}$ C for the obtained hybridization mix 24 h in the thermal cycler.
7. Prior to removing the hybridization mix from the thermal cycler, prepare the Dynabeads[®] M-280 Streptavidin as following: transfer 50 μ L of dynabeads into a 1.5 mL microcentrifuge tube (*see* **Note 11**), place the tube on the magnetic stand until the supernatant is clear, remove and discard it. Wash three times the dynabeads as following: add 200 μ L of binding buffer, gently tap the tube to resuspend the dynabeads, place the tube on the magnetic stand until the supernatant is clear, remove and discard the supernatant. Take care not to disturb

the dynabeads. After the three washes, resuspend the dynabeads in 200 μL of binding buffer.

8. Add the 26 μL of hybridization mix to the washed dynabeads. Gently tap the tube to resuspend the dynabeads and incubate for 30 min at room temperature (off the magnetic stand). Regularly resuspend the dynabeads during the incubation by gently taping the tube (*see Note 13*).
9. During the incubation, pre-warm the wash buffer n°2 at 65 °C (at least 1.5 mL per captured library).
10. After the incubation, place the tube on the magnetic stand until the supernatant is clear. Remove and discard the supernatant. Take care not to disturb the dynabeads. Add 500 μL of wash buffer n°1 and resuspend the dynabeads by gently taping the tube. Incubate for 15 min at room temperature (off the magnetic stand). Regularly resuspend the dynabeads during the incubation by gently taping the tube (*see Note 13*).
11. After the incubation, place the tube on the magnetic stand until the supernatant is clear, remove and discard it. Take care not to disturb the dynabeads. Wash three times the dynabeads as following: resuspend the dynabeads in 500 μL of pre-warmed wash buffer n°2, incubate for 10 min at 65 °C (off the magnetic stand). Regularly resuspend the dynabeads during the incubation by gently taping the tube. Place the tube on the magnetic stand until the supernatant is clear. Remove and discard the supernatant. Take care not to disturb the dynabeads.
12. Resuspend the dynabeads in 50 μL of 0.1 M NaOH by vortexing the tube for 5 s (*see Note 2*). Incubate for 10 min at room temperature (off the magnetic stand).
13. Place the tube on the magnetic stand until the supernatant is clear and transfer it to a 1.5 mL microcentrifuge tube containing 70 μL of 1 M Tris-HCl (pH 7.5). Take care not to disturb the Dynabeads.
14. Purify the captured library using the QIAquick PCR Purification Kit following the manufacturer's instructions. The purified product is eluted in 50 μL of nuclease-free water.
15. Select the DNA fragment size with the Agencourt® AMPure® XP Reagent as indicated in **step 4** of Subheading 3.3.
16. Amplify the captured library using the GC-RICH PCR System, dNTPack. Make five 50 μL PCR reactions per captured library and proceed in the same way as the **step 2** of Subheading 3.3 but realize 25 amplification cycles instead of 20 in the thermal conditions.
17. Purify the amplified captured library using the QIAquick PCR Purification Kit following the manufacturer's instructions. Realize one column for 2.5 amplification reactions from a same

library (i.e. two columns per library (125 μ L)). The purified product is eluted in 50 μ L of nuclease-free water.

18. Select the DNA fragment size with the Agencourt® AMPure® XP Reagent as indicated in **step 4** of Subheading **3.3**.
19. Evaluate the concentration of purified amplified captured library with Nanodrop spectrophotometer.
20. Proceed to a second cycle of hybridization by repeating the **steps 1–15** with the purified amplified capture products obtained previously (*see Notes 14 and 15*).
21. Amplify the captured library using the GC-RICH PCR System, dNTPack. Make ten PCR reactions per library and proceed in the same way as the **step 2** of Subheading **3.3** but realize 25 amplification cycles instead of 20 in the thermal conditions.
22. Purify the amplified captured library using the QIAquick PCR Purification Kit following the manufacturer's instructions. Realize one column for two amplification reactions from a same library (i.e. five columns per library). The purified product is eluted in 50 μ L of nuclease-free water.
23. Select the DNA fragment size with the Agencourt® AMPure® XP Reagent as indicated in **step 4** of Subheading **3.3**.
24. Evaluate the concentration of purified amplified captured library with Nanodrop spectrophotometer. Assess its quality on an Agilent DNA 7500 or 12000 chip, according to the manufacturer's instructions (Fig. 5).

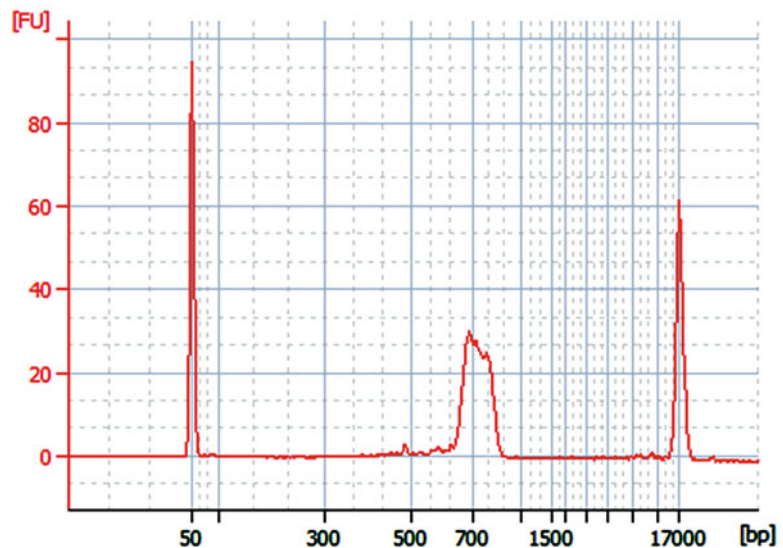


Fig. 5 Quality of captured library (prepared for 650 bp insert) assess on an Agilent Bioanalyzer 12000 DNA chip. The electropherogram shows a peak focused on 770 bp for the case of library prepared for 650 bp insert with 120 bp Illumina adaptor. With a library prepared for 550 bp, the same profile will be observed but with a peak focused on 670 bp

25. Store the purified amplified captured library products at -20°C .
26. Proceed to the sequencing of captured library on an Illumina sequencer compatible with the kit use to prepare the library.

4 Notes

1. EDTA will not go into solution until the pH of the solution is adjusted to ~ 8.0 by the addition of 1 M NaOH.
2. Buffers and solutions must be extemporaneously prepared.
3. For the fragmentation of gDNA, we recommend, as Illumina, to use Covaris microTUBES with a focused-ultrasonicator. Covaris offers different models of focused-ultrasonicator and the material necessary for the fragmentation depends on the focused-ultrasonicator used.
4. Include 10 % excess for multiple samples.
5. The verification of the amplified hybrid probe size by electrophoresis on a 2 % agarose-TBE gel is absolutely necessary. It is possible to get two amplification bands, one at the expected size (i.e. size of hybrid probe with amplification primers T7-A and B) and another due to aberrant amplification of the T7 promoter. Only the correct band at the expected size must be excised and purified.
6. For example, three hybrid probes (specific capture probe and adaptor sequences) are necessary for the mix with specific capture probe as following: probe A 5'-CCCAGGATWAGATACCKCCYAGTTTAYRC-3', probe B 5'-TTCAGAAGTAGATATGCTGGTAGTCTACCA-3' and probe C 5'-TGCACAATCAGATAGTYTGGYAGTGACCGC-3'. The probe A has one W base (two combinations, T or A), one K base (two combinations, T or G), two Y bases (two combinations, C or T), and one R base (two combinations, A or G), therefore there is 32 possible combinations for the probe A ($1\text{ W} \times 1\text{ K} \times 2\text{ Y} \times 1\text{ R} = 2 \times 2 \times (2 \times 2) \times 2 = 32$). In the same way, the probes B and C have a degeneracy of 1 (none degenerated base) and 4 (2×2) respectively. To make the hybrid probe mix, use $1/32^{\text{e}}$ in quantity of the probe B and $1/8^{\text{e}}$ in quantity of the probe C compared to the probe A. Thus, for 500 ng of probe A, add 15.6 ng of probe B and 62.5 ng of probe C. In this example, the three probes have the same size, otherwise you must calculate a number of molecules (taking into account the size of each probe) to realize an equimolar mix.
7. The IVT reaction mix must be realized at room temperature. Indeed the spermidine in the $10\times$ reaction buffer can coprecipitate the template DNA if the reaction is assembled on ice.

8. Remove the AMPure beads from 2 to 8 °C storage and let stand for at least 30 min to bring them to room temperature. Vortex the room temperature AMPure beads for at least 1 min or until they are well dispersed. Vortex the beads frequently and collect them by slowly pipetting (due to their viscosity) to make sure that they are evenly distributed.
9. Be sure not to retrieve beads with the supernatant because the beads can interfere with the following steps. It is better to retrieve less volume of clear supernatant and if necessary to adjust the volume by adding nuclease-free water to the purified product.
10. The sheared salmon sperm DNA commercial solution is concentrated at 10 mg/mL. Make a dilution at 1/10^e in nuclease-free water and pipette 2.5 µL to get 2.5 µg of sheared salmon sperm DNA.
11. Make as many aliquots as captured library.
12. It could be necessary to dilute the biotinylated RNA probe mix. The 500 ng of the biotinylated RNA probe mix should be in a volume less than or equal to 6 µL.
13. Alternatively, incubate the appropriate time at room temperature with tilting and gentle rotation on a HulaMixer Sample Mixer or other similar device.
14. If you do not plan to proceed immediately to the second cycle of hybridization, the protocol can be safely stopped here. If you are stopping, store the tubes at -20 °C.
15. If the quantity of purified amplified capture products, obtained after the first cycle of hybridization, is less than 500 ng, proceed to the second cycle of hybridization with the totality of the sample.

References

1. Schloss PD, Handelsman J (2006) Toward a census of bacteria in soil. *PLoS Comput Biol*. doi:[10.1371/journal.pcbi.0020092](https://doi.org/10.1371/journal.pcbi.0020092)
2. Vieites JM, Guazzaroni M-E, Beloqui A et al (2009) Metagenomics approaches in systems microbiology. *FEMS Microbiol Rev* 33:236–255
3. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21:1794–1805
4. Desai N, Antonopoulos D, Gilbert JA et al (2012) From genomics to metagenomics. *Curr Opin Biotechnol* 23:72–76
5. Klindworth A, Pruesse E, Schweer T et al (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. doi:[10.1093/nar/gks808](https://doi.org/10.1093/nar/gks808)
6. Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 3:489–498
7. Hong S, Bunge J, Leslin C et al (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* 3:1365–1373
8. Denonfoux J, Parisot N, Dugat-Bony E et al (2013) Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Res* 20:185–196
9. Bragalini C, Ribière C, Parisot N et al (2014) Solution hybrid selection capture for the recovery of functional full-length eukaryotic cDNAs

- from complex environmental samples. *DNA Res* 21:685–694
10. Dugat-Bony E, Peyretailade E, Parisot N et al (2012) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environ Microbiol* 14:356–371
 11. Jaziri F, Peyretailade E, Missaoui M et al (2014) Large scale explorative oligonucleotide probe selection for thousands of genetic groups on a computing grid: application to phylogenetic probe design using a curated small subunit ribosomal RNA gene database. *Sci World J*. doi:[10.1155/2014/350487](https://doi.org/10.1155/2014/350487)
 12. Militon C, Rimour S, Missaoui M et al (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics* 23:2550–2557
 13. Parisot N, Denonfoux J, Dugat-Bony E et al (2012) KASpOD—a web service for highly specific and explorative oligonucleotide design. *Bioinformatics* 28:3161–3162
 14. Parisot N, Denonfoux J, Dugat-Bony E et al (2014) Software tools for the selection of oligonucleotide probes for microarrays. In: He Z (ed) *Current technology, innovations and applications*. Academic, New York
 15. Jaziri F, Parisot N, Abid A et al (2014) PhyLOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification. Database. doi:[10.1093/database/bau036](https://doi.org/10.1093/database/bau036)
 16. Dugat-Bony E, Missaoui M, Peyretailade E et al (2011) HiSpOD: probe design for functional DNA microarrays. *Bioinformatics* 27:641–648

Chapter 11

Hybridization of Environmental Microbial Community Nucleic Acids by GeoChip

Joy D. Van Nostrand, Huaqin Yin, Liyou Wu, Tong Yuan, and Jizhong Zhou

Abstract

Functional gene arrays, like the GeoChip, allow for the study of tens of thousands of genes in a single assay. The GeoChip array (5.0) contains probes for genes involved in geochemical cycling (N, C, S, and P), metal homeostasis, stress response, organic contaminant degradation, antibiotic resistance, secondary metabolism, and virulence factors as well as genes specific for fungi, protists, and viruses. Here, we briefly describe GeoChip design strategies (gene selection and probe design) and discuss minimum quantity and quality requirements for nucleic acids. We then provide detailed protocols for amplification, labeling, and hybridization of samples to the GeoChip.

Key words GeoChip, Functional gene array, Microbial communities, Microbial ecology, Hybridization, Fluorescent labeling, Whole community genome amplification

1 Introduction

Microorganisms are critical for global biogeochemical cycling of N, C, S, and metals; however, a large percentage of microorganisms (>99 %) remain uncultured [1–3] making it difficult to fully examine microbial community activity. To overcome this limitation, culture-independent approaches are needed. Examination of functional genes, those genes that are involved in processes of interest (e.g., *nifH* for N fixation, *dsrA* for sulfite reduction, etc.), can shed light on the functional abilities of microbial communities. While there are a wide variety of culture-independent approaches that could be used to examine functional genes, most can provide information on only a small number of functional genes. Microarrays, on the other hand, allow the examination of tens of thousands of genes at one time, thus allowing for a comprehensive examination of a wide range of genes.

Functional gene microarrays (FGAs) probe for multiple functional genes involved in microbial functional processes of interest at one time [4–6]. Because FGAs can shed light on the

functional potential of microbial communities, they are ideal linking microbial functional structure with environmental processes. The most comprehensive FGA to date is the GeoChip microarray [5–7]. The GeoChip was designed to address two major challenges in using FGAs for microbial community analysis: (1) the need for adequate oligonucleotide probe specificity for genes that have high homologies and (2) the lack of a truly comprehensive probe set [5].

The following chapter briefly describes the process of gene selection and probe design for the GeoChip microarrays and describes the sample preparation and hybridization protocols for the GeoChip arrays. The GeoChip is currently manufactured by Agilent Technologies (Santa Clara, CA, USA), but there are other companies that manufacture arrays and they can also be printed in-house. These hybridization methods included are specific for Agilent arrays, but other methods can be used for other array types and there are methods for using Agilent arrays with automated hybridization stations. A general overview of the GeoChip analysis protocol is shown in Fig. 1.

The first step for GeoChip design is the selection of genes representing processes of interest. GeoChip was designed to

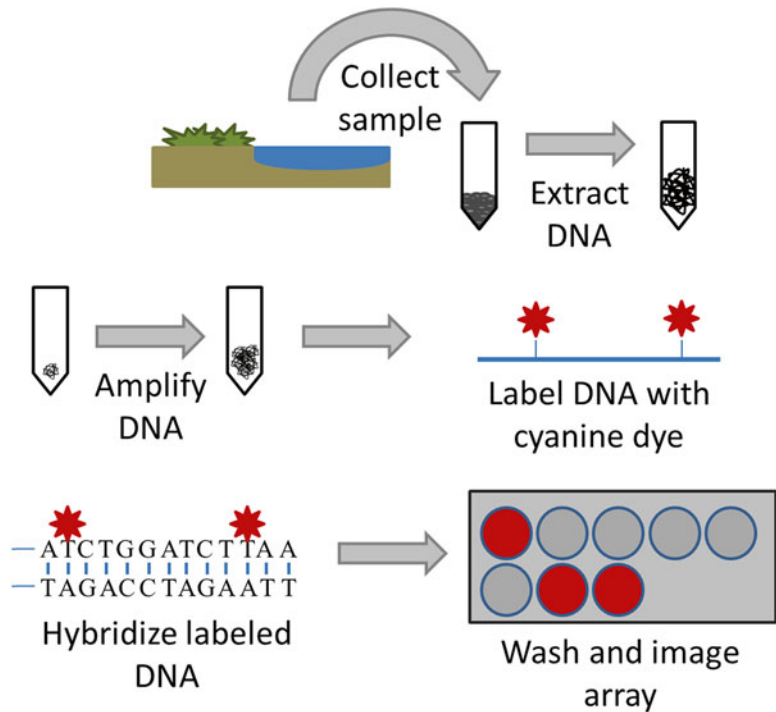


Fig. 1 Overview of the GeoChip protocol. Samples from the environment of interest are collected and DNA is extracted. If the yield of DNA is insufficient, whole community genome amplification can be performed to increase the quantity of DNA. The DNA is then labeled with a cyanine dye and hybridized to the GeoChip. Any unhybridized DNA is then washed off and the array is imaged

examine microbial community functional potential, so only genes for those proteins actively involved in a process of interest, i.e., those containing catalytic subunits or active sites, are included. Next, public databases are searched using keywords that are chosen to select a wide range of sequences that may be the gene of interest. These sequences are then confirmed by HMMER alignment (<http://hmmmer.wustl.edu/>) with seed sequences that have been experimentally confirmed. These confirmed sequences are then used for probe design (50mer) using experimentally determined criteria based on sequence homology (≤ 90 % identity for gene-specific probes and ≥ 96 % for group-specific probes), continuous stretch length (≤ 20 bases for gene-specific probes and ≥ 35 for group-specific probes), and free energy (≥ -35 kJ/mol for gene-specific probes and ≤ -60 kJ/mol for group-specific probes) [8, 9] with new versions of the CommOligo software [10]. The final step is to confirm specificity by BLASTing the probe sequences against the GenBank database.

There are a number of DNA extraction and purification kits and methods available for microbial community samples. As long as the resultant nucleic acid quantity and quality is sufficient and the fragment length is fairly large, any method can be used. The absorbance ratio (A260:A280) should be ~ 1.8 and > 1.9 for DNA and RNA, respectively, and an A260:A320 > 1.7 . The A260:A230 ratio is most important for microarray success [11]. A high A230 value may indicate contamination with EDTA, carbohydrates, phenol, or guanidine HCl [12]. For samples with high humics, a gel purification strategy followed by a phenol-chloroform-butanol extraction [13, 14] may be needed. If a low yield of nucleic acid is expected, large nucleic acid fragments are needed for efficient amplification.

GeoChip requires 0.5–1 μg of DNA or 10 μg of RNA. If this amount cannot be obtained, an amplification step will be needed. Whole community genome amplification (WCGA) using the Templphi 500 amplification kit (phi 29 DNA polymerase, GE Healthcare, Piscataway, NJ) and a modified reaction buffer can be used. This method provides a sensitive (10 fg detection limit) and representative amplification (< 0.5 % of amplified genes showed > 2 -fold different from unamplified) from 1 to 100 ng template DNA [15]. Whole community RNA amplification (WCRA) provides a representative amplification with 50–100 ng of starting material [16].

2 Materials

2.1 Reagents and Kits

1. Templphi 500 amplification kit (GE Healthcare, Piscataway, NJ, USA).
2. Single-stranded binding protein (SSB) (Affymetrix, Santa Clara, CA, USA).

3. Quant-iT™ PicoGreen® dsDNA kit (Life Technologies, Carlsbad, CA, USA).
4. dNTP mix: 5 mM (2.5 mM dTTP). Combine 5 µL each of 100 mM dATP, dGTP, and dCTP and 2.5 µL 100 mM dTTP. Bring to 100 µL with 82.5 µL nuclease-free water.
5. 40 U/µL Klenow (IMER, Inc., San Diego, CA, USA).
6. 25 nM Cy3-dUTP (GE Healthcare).
7. QIAquick PCR Purification kit (Qiagen, Valencia, C, USA).
8. Cot-1 Human DNA (catalog number 5190-3393) (Agilent).

2.2 Buffers and Solutions

1. 2.4 mM spermidine. Weight out 1.66 g spermidine and make up to 100 mL with water.
2. Hybridization buffer: Part of the Oligo aCGH hybridization kit, large, catalog number 5188-5380 (Agilent).
3. Blocking agent: Add 1350 µL water to the 10× aCGH Blocking Agent (Part of the Oligo aCGH hybridization kit, large, catalog number 5188-5380) (Agilent). Incubate at room temperature for at least 6 h before using.
4. Wash Buffers: Oligo aCGH/ChIP-on-Chip Wash Buffer kit (catalog number 5188-5226, contains Wash Buffers 1 and 2) (Agilent).

2.3 Probes and Primers

1. Random primers (mostly hexamers, 3 µg/µL, catalog number 48190-011) (Life Technologies).
2. Universal standard. The universal standard or common oligo reference standard (CORS) target is complimentary to a CORS probe printed onto the GeoChip microarray. The CORS target can be synthesized by an array of manufacturers. We use Eurofins MWG Operon for target synthesis. The nucleotide sequence contains Cy-5 label on the 5' with purification by desalting. [Cy5]GCCAGCACAGCTACACGTCCTCAAACG ATTGTGTGCGGTCCGAGGTGCGG. See Liang et al. [13], for full details of the CORS development.

2.4 Equipment

1. NanoDrop spectrometer (Thermo Scientific, Wilmington, DE, USA) or similar instrument capable of measuring cy-dye incorporation (568 and 647 nm).
2. General molecular lab equipment: thermocyclers, fluorescent plate reader (for PicoGreen measurement), vacuum concentrator for drying samples.
3. Microarrays: GeoChip 5 uses Agilent arrays (SurePrint G3 custom array) in two sizes: 8×60K, 8 arrays of 60,000 probes per slide or 4×180K, 4 arrays of 180,000 probes per slide.

4. Gasket slides (catalog number G2534-60013 for four arrays per slide or G2534-60016 for eight arrays per slide) (Agilent).
5. Hybridization equipment: hybridization oven, hybridization oven rotator rack, hybridization oven conversion rod, hybridization chamber (Agilent), magnetic stir plate with heating element.
6. Microarray scanner: The SureScan Microarray Scanner (Agilent Technologies, Santa Clara, CA) is ideal for Agilent microarrays, but other microarray scanners having a 2–3 μm resolution and the ability to image Cyanine 3 and 5 can be used.

3 Methods

3.1 Amplification

The GeoChip requires 1 μg DNA for hybridization. While many samples, such as soil, can easily meet this criterion, sites with low microbial abundance or that have restrictions on the amount of sample that can be taken, may not yield enough DNA. In these cases, WCGA can be used to increase the amount of DNA available. WCGA uses a modification of the Templiphi 500 amplification kit (phi 29 DNA polymerase) [15]. The amplification buffer is supplemented with single-stranded DNA binding protein (SSB) and spermidine and a larger amount of enzyme to increase sensitivity and representative amplification. The SSB and spermidine likely assist with DNA replication and bind inhibitors [17, 18]. Using 1–100 ng DNA provides a sensitive (10 fg detection limit) and representative amplification (<0.5 % of amplified genes showed >2-fold different from unamplified) [15].

This amplification reaction is very sensitive and will amplify any contaminating DNA. As such, all steps should be performed in a PCR hood. Zhang et al. [19] have outlined additional steps that should be followed to minimize contamination. These include UV irradiation of the hood and all items to be used in the protocol (i.e., tips, tubes, pipettors, tube racks, ice and ice bucket, etc.). Negative controls should always be run alongside the samples.

1. Add 10 μL sample buffer to a 0.2 mL PCR tube (*see Note 1*).
2. Add DNA (preferably 100 ng) (*see Notes 2 and 3*).
3. Mix DNA and buffer thoroughly and incubate at room temperature (RT) for 10 min.
4. While DNA and buffer are incubating, prepare Templiphi premix in a 1.7 mL tube (*see Note 1*). [For each reaction: 10 μL of reaction buffer, 0.6 μL enzyme mixture (both supplied in the Templiphi kit), 1.25 μL of 5 $\mu\text{g}/\mu\text{L}$ SSB (USB; Cleveland, OH), and 1 μL of 2.4 mM spermidine stock.]

5. Transfer 12.85 μL of the premix to the DNA/buffer mixture, mix well and spin twice.
6. Incubate the reaction at 30 $^{\circ}\text{C}$ for 6 h then heat-inactivate the enzyme at 65 $^{\circ}\text{C}$ for 10 min.
7. Run an aliquot ($\sim 2 \mu\text{L}$) of the product on a gel. The product should produce a smear rather than a single band (*see Note 4*).
8. Quantify the amplified DNA using a dye-binding assay, such as PicoGreen (Quant-iT™ PicoGreen® dsDNA kit). Using the 260/280 ratio will result in an erroneous quantity due to the primers and dNTPs remaining in the sample.

3.2 Labeling

DNA for microarray hybridization is generally labeled using fluorescent dyes such as cyanine dyes. The DNA can be labeled directly (dyes are directly integrated into the target DNA) or indirectly (targets are labeled after hybridization). A direct labeling approach is detailed here. Either amplified or unamplified DNA can be used.

1. Combine 5.5 μL random primers (3 $\mu\text{L}/\mu\text{L}$) and 1 μg target DNA in a 0.2 mL PCR tube and bring to 35 μL with nuclease-free water (*see Note 5*).
2. Mix well and incubate at 99.9 $^{\circ}\text{C}$ for 5 min, then immediately chill on ice.
3. In a separate 1.7 mL microcentrifuge tube, prepare a master mix. [For each sample: 6 μL nuclease-free water, 5 μL 10 \times reaction buffer (included with enzyme), 2.5 μL 5 mM dNTP mix (2.5 mM dTTP), 1 μL klenow (40 U/ μL), and 0.5 μL CyDye (25 nM, Cy3-dUTP)] (*see Notes 5 and 6*).
4. Add 15 μL of the master mix to primer/target DNA tube and mix well.
5. Incubate the reaction at 37 $^{\circ}\text{C}$ for 6 h, heat-inactivate the enzyme at 95 $^{\circ}\text{C}$ for 3 min and then cool to 4 $^{\circ}\text{C}$.
6. Purify the labeled DNA with a Qiagen QIAquick Kit as described by the manufacturer.
7. Elute the DNA using 100 μL H₂O and check the CyDye incorporation using NanoDrop (*see Note 7*).
8. Dry the labeled DNA using a Speed Vac at Vacuum Level 5.1 for 2 h at 45 $^{\circ}\text{C}$.

3.3 Hybridization

There are currently two versions of GeoChip 5.0. The smaller array has $\sim 60,000$ probes (60K) and is more for general microbial ecology studies ($\sim 60\text{K}$ probes) and was designed to cover the core biogeochemical cycles (C, N, S, and P) as well as degradation genes for the more common contaminants (such as BTEX) and metals and antibiotic resistance genes as long as they changed the metal or antibiotic (oxidation, reduction, degradation) and not just acted as

a pump. The larger version has ~180,000 probes (180K) and covers all the genes on the smaller array plus more contaminant degradation genes and includes genes for non-transformative metal resistance mechanisms (e.g., pumps) as well as general metal homeostasis genes, and several other categories (viral and protist genes, stress response, etc.). A complete list of genes covered by both GeoChip 5.0 versions can be found by clicking the GeoChip summary links at <http://ieg.ou.edu/entrance.html>.

GeoChip 5.0 uses an Agilent format and the following protocol is based on one of Agilent's hybridization protocols [20] but has some modifications. The 60K array has eight arrays per slide and there are four 180K arrays per slide; because of this, the 60K and 180K arrays use different volumes of hybridization buffer. All quantities below are for the 60K array with the volumes for the 180K in parentheses.

3.3.1 Sample Preparation

1. Prepare hybridization buffer (*see Note 8*). (Per sample: 27.5 μL 2 \times HI-RPM hybridization buffer, 5.5 μL blocking agent, 2.4 μL Cot-1 DNA, 1.1 μL universal standard [13], and 5.5 μL formamide.)
2. Add 13 μL nuclease-free water to the labeled DNA.
3. Add 42 μL of the hybridization buffer to the DNA and mix well by pipetting up and down and then spin to collect liquid in the bottom of the tube.
4. Heat samples at 95 $^{\circ}\text{C}$ for 3 min, then immediately transfer samples to 37 $^{\circ}\text{C}$ and incubate for another 30 min (*see Note 9*).
5. Centrifuge samples for 1 min at 6000 $\times g$ to collect the sample in the bottom of the tube.

3.3.2 Array Assembly

1. Place a new gasket slide (gasket side up) into the Agilent SureHyb chamber. Make sure the slide is aligned properly and is flush with the chamber base.
2. Slowly pipette 48 μL of the sample into a gasket well, avoiding touching the slide and making sure the liquid does not touch the gasket (*see Note 10*). Repeat with next sample until all gasket wells have been filled.
3. Place the microarray onto the gasket slide, making sure the active side (the text Agilent is printed on the active side) is down and that the array slide is properly aligned with the gasket slide.
4. Place the chamber cover on the slides, slide the clamp into place, and then firmly tighten the clamp (*see Note 11*).
5. Lift the assembled chamber and rotate to wet the slides and confirm that the air bubble moves freely and that there are no small bubbles that may inhibit mixing (*see Note 12*).

3.3.3 Hybridization Protocol

1. Put the slide chambers in the hybridization oven's rotator rack (*see Note 13*). Use empty slide chambers to keep the rotator balanced, if necessary.
2. Hybridize arrays at 67 °C for 24 h with a rotation speed of 20 rpm.

3.3.4 Washing

1. Prepare the wash buffers in three separate wash dishes (*see Notes 14 and 15*).
 - (a) Wash 1—The Wash Buffer 1 (WB1) dish should be placed on the benchtop and maintained at room temperature.
 - (b) Wash 2—A second room temperature WB1 dish should be placed on a magnetic stir plate and contain a slide rack and a stirbar. There should be sufficient WB1 to completely cover the slide rack.
 - (c) Wash 3—The third dish contains the prewarmed Wash Buffer 2 (WB2) (*see Note 16*). This dish should be placed on a magnetic stir plate with heating element to maintain the buffer at 37 °C. There should be sufficient WB2 to completely cover the slide rack. Insert a stir bar.
2. Remove one hybridization chamber from the incubator (*see Note 17*).
3. Check to determine if any bubbles formed, if there was any leakage, and if the sample is still able to rotate freely (*see Note 18*).
4. Place the chamber on a flat surface, loosen the screw and remove the clamp and chamber cover. Carefully lift one end of the array and gasket slide and then hold the sides of the array, maintaining its current orientation (array on top, gasket slide on bottom) quickly place it into Wash 1.
5. Make sure the slide is completely submerged and then using forceps to make a gentle twisting motion, pry the array slide off the gasket slide. Leave the gasket slide and quickly place it in the slide rack in Wash 2.
6. Repeat **steps 2–4** until all slides have been transferred or there are five slides in Wash 2, whichever comes first (*see Note 19*).
7. Turn on the magnetic stirrer in Wash 2, such that there is a depression on the surface of the buffer without creating a vortex. We were able to obtain this with a speed of 250–300 rpm on a VWR (Radnor, PA) model magnetic stirplate with a maximum speed of 1600 rpm. Incubate for 5 min.
8. Transfer the slide rack to Wash 3 and then turn on stirrer as described in **step 6**. Incubate for 1.5 min.
9. Slowly remove the slide rack from Wash 3. It should take about 10 s to remove. No additional spinning or drying is needed. The slides are hydrophobic and should shed the buffer.
10. Discard used buffers and replenish if additional slides need to be washed (*see Note 14*).

3.3.5 Scanning

1. For the Agilent SureScan microarray scanner, the array slides must be placed into slide holders. The array should be placed array-side up. Close the holders, making sure you hear a click.
2. Place the slide holders containing microarray slides into the scanner cassette.
3. Select the appropriate scanning protocol and check the settings (*see Note 20*). For GeoChip arrays, which use the Agilent platform, using the SureScan Microarray Scanner, scanning is done in red and green channels (lasers with excitation wavelengths at 640 and 532 nm, respectively), 3 μm resolution, 20 bit Tiff dynamic range ($>10^5$), and 100 % photomultiplier tube sensitivity for both channels.
4. Scan the slides and then extract the data using Agilent's Feature Extraction program. Select the appropriate grid template (each array design should have its own specific template file generated by the array manufacturer) and the Feature Extraction program will automatically place and optimize the grid placement. Array features are automatically selected and mean pixel intensity is scored using the program's default settings.
5. Evaluate hybridization quality of the array. The slide image can be displayed in the Feature Extraction program once scanning is complete. Examine the images to make sure positive control spots are present. GeoChip 5 contains a series of 16S rRNA gene probes across the array in an easily observable pattern (green channel) and the CORS probes should also be visible across the array (red channel) (Fig. 2). Also make sure the arrays have even hybridization and no obvious problems [areas with no positive probes, very bright or dim areas, or "flares" where spots are obscured by dust or other fluorescent contaminants (Fig. 3)]. Make sure the background is even and that none of the arrays have a obviously higher background by switching to log scale.

3.4 Data Processing and Analysis

There are a large variety of microarray analysis software available to use. Select the software that best meets your particular needs. The GeoChip microarrays use an in-house developed data analysis pipeline. The pipeline allows the user to select normalization protocols, the method to determine signal cutoff and what controls to use.

GeoChip data normalization and quality filtering is performed with multiple steps [13, 21]. As a general rule, the following protocol is followed although other settings may be used depending on the samples (*see Note 21*). First, the average signal intensity of common oligo reference standard is calculated for each array, and the maximum average value is applied to normalize the signal intensity of samples in each array. Second, the sum of the signal intensity of samples is calculated for each array, and the maximum

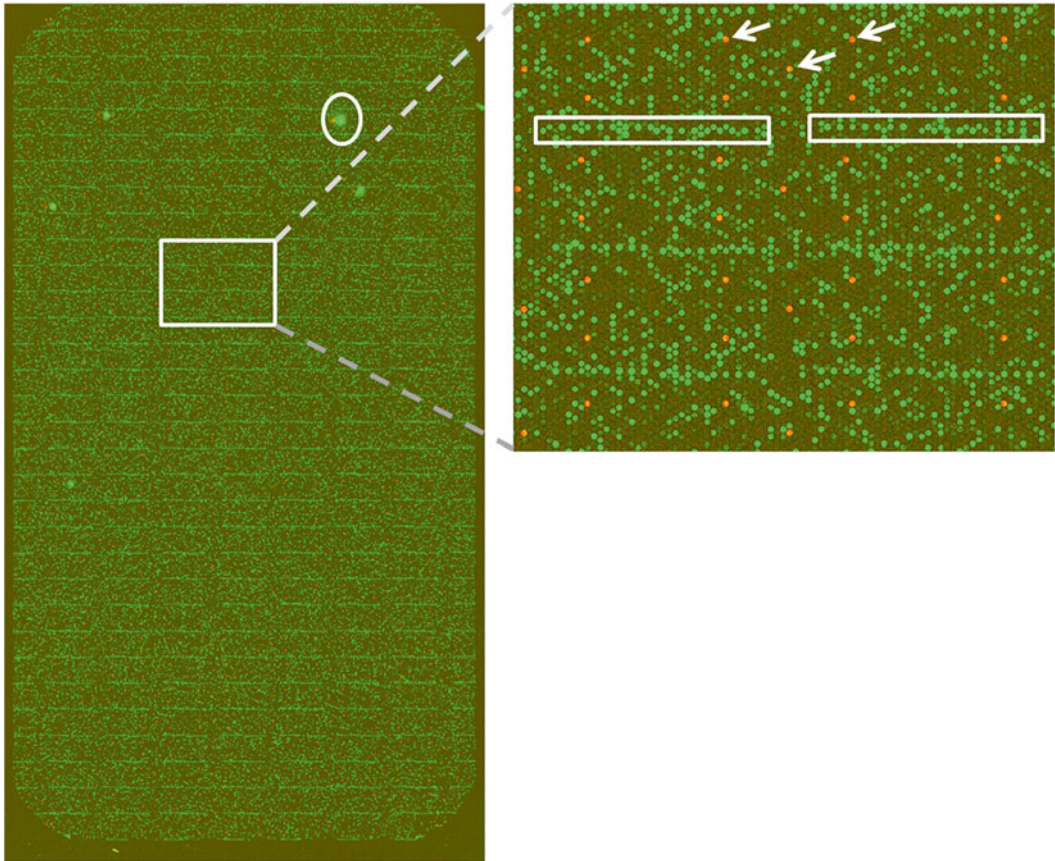


Fig. 2 GeoChip 5 microarray image. Image to the *left* is a hybridized GeoChip array. The *circle spot* is fluorescence from a piece of dust or other debris. The *boxed area* is enlarged to the *right*. The *arrows* point to a few of the CORS control probes (shown as *red spots*). The *boxed area* in the *right image* shows the 16S rRNA control probes

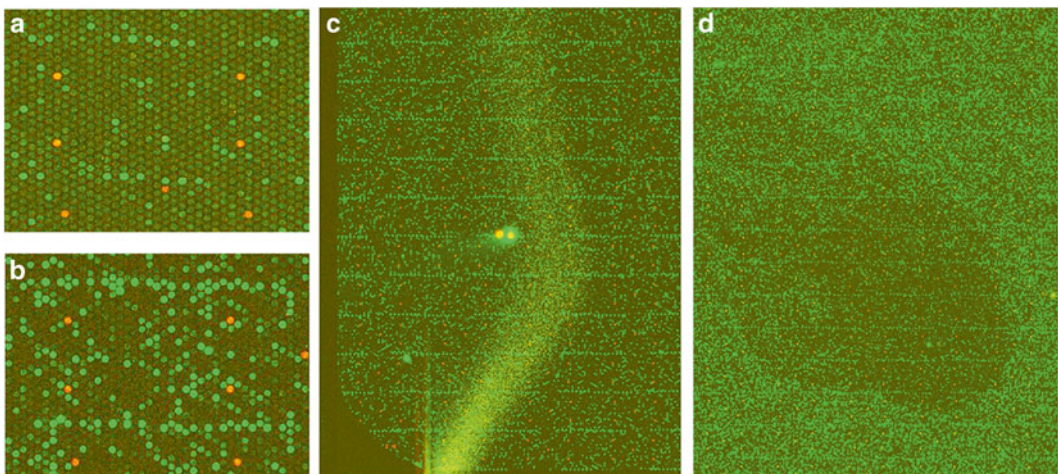


Fig. 3 Examples of poor quality hybridization or artifacts. **(a)** Low or no 16S rRNA gene probe signal and poor sample hybridization (compare with **b**). **(c)** A “flare” likely from incomplete washing, and **(d)** an area of no/poor hybridization

sum value is applied to normalize the signal intensity of all spots in an array, producing a normalized value for each spot in each array. Spots are then scored as positive and retained if the signal-to-noise ratio [$SNR = (\text{signal mean} - \text{background mean}) / \text{background standard deviation}$] is ≥ 2.0 , the coefficient of variation (CV) of the background is < 0.8 , and the signal intensity is at least 1.3 times the background. In addition, spots with signal intensities less than ~ 200 are discarded. The minimum signal value chosen should be at least twice the average background signal for a given set of samples. Spots that were detected in less than two samples (either within a replicate group or across all samples) are also removed. Before statistical analysis, logarithmic transformation is carried out for the remaining spots, and the signals of all spots are transferred into relative abundances.

Microarray data analysis can be challenging due to the large amount of data generated and its multivariate structure. The GeoChip data analysis pipeline has a variety of analysis tools for microarray data. These include relative abundance of genes or gene categories or subcategories, richness and α and β diversity of functional genes, and gene overlap between individual samples or sample groups. To look at differences between conditions, hierarchical cluster analysis, *T*-tests, analysis of variance (ANOVA), and dissimilarity tests could be used. Response ratios can be used to compare gene levels or signal intensity between conditions (e.g., treatment vs. control, contaminated vs. uncontaminated) [22]. Unconstrained ordination methods, such as principal component analysis (PCA) and correspondence analysis (CA) maximize the visible variability of data sets by reducing the dimensionality of variables. Due to the differences in the assumed data structures of ecological studies, CA may be preferred over PCA. Non-metric multidimensional scaling (NMDS), which represents the relative inter-relatedness of samples on a priori dimensions, could also be used.

Interrelationships among functional genes detected and other abiotic and biotic factors can be examined by constrained ordination, such as canonical correspondence analysis (CCA) [23], distance-based redundancy analysis (db-RDA) [24], variation partitioning analysis (VPA) [25, 26], and Mantel test. CCA is commonly used for GeoChip-based studies to better understand how environmental factors impact and drive community structure, while canonical VPA, based on the results of CCA and partial CCA, provides information on the relative influence of individual parameters on the microbial community structure. The Mantel test can be used to compare environmental factors with functional genes detected by GeoChip.

4 Notes

1. Before use, mix reagents well, then briefly spin them in a microcentrifuge or microfuge to collect all the liquid in the bottom of the tube. Repeat to make sure all reagents are well mixed.
2. Make sure the same amount of DNA is used for all samples within an experimental group.
3. Efficient amplification requires high molecular weight DNA of the highest quality obtainable.
4. No or minimal amplification is likely due to inhibitors in the sample. To overcome this, decrease the amount of sample volume used or run the sample through a serial dilution (two to three dilution steps) to reduce the concentration of inhibitor or “wash” the DNA. Inhibitors may be present even if the DNA quality is high. Additionally, the DNA can be re-precipitated (using ethanol or isopropanol protocols [27]) to reduce inhibitors. If there are still problems with the amplification, the incubation time can be increased or multiple amplification products can be combined to obtain sufficient DNA.
5. For best results, high quality, fresh reagents should be used [4].
6. CyDyes are light sensitive so the samples should be protected from the light as much as possible after the dye is added.
7. Minimum dye incorporation should be >50 pmol (pmol/ $\mu\text{L} \times \text{total } \mu\text{L}$).
8. Prepare blocking agent ahead of time by adding 1350 μL nuclease-free water to the 10 \times aCGH Blocking Agent and keep at room temperature for at least 6 h.
9. Prior to starting, preheat two heat blocks or thermocyclers to 95 and 37 $^{\circ}\text{C}$.
10. The Agilent arrays use bubble mixing, so the hybridization buffer will not completely fill the array chambers.
11. Make sure to tighten the clamp as firmly as possible using your hand. The clamp is designed to prevent damage to the slide so over tightening will not break the slide.
12. If small bubbles are present, firmly tap the assembly on a hard surface to dislodge the bubbles.
13. Preheat hybridization oven to 67 $^{\circ}\text{C}$ and allow it to equilibrate for at least 6 h.
14. Prior to use and after use, thoroughly rinse all glassware and stirbars used in the washing steps with copious amounts of Milli-Q water or equivalent and allow to air dry.

15. Washing should be performed in a low ozone environment such as an ozone-free hood as ozone can degrade cy-dye signal [28, 29]. The most critical time for cy-dye signal loss is during the transition from wet to dry [30], so it is important to limit ozone exposure during washing.
16. Preheat Wash Buffer 2 to 37 °C in an incubator. The Wash Buffer should be preheated in the wash dish and remain at 37 °C overnight.
17. Maintain rotation and temperature for the remaining chambers.
18. Any of these issues may result in inefficient hybridization. Make note if any of these issues occurred and continue with washing. Once the slide is scanned, hybridization quality can be assessed.
19. This setup is sufficient for five slides. If more slides need to be washed, fresh buffers should be used.
20. Other microarray scanners can be used for scanning the GeoChip as long as the scanner has at least a 3 µm resolution, can scan both the red and green channels and has at least a 16 bit dynamic range.
21. Based on results from hundreds to thousands of samples analyzed by various GeoChip versions, we anticipate about 30 % of probes to be positive for most soil samples. Other sample types, such as groundwater or bioreactors, may have fewer positive probes. For samples with more positive probes than expected, settings can be adjusted by increasing the SNR cutoff used or increasing the minimum signal required. Removing singletons (those probes detected in only one sample) can also reduce noise in the data.

Acknowledgement

Efforts for writing this Chapter were supported by the Department of Energy's Carbon Cycling program (DE-SC0004601 and DE-SC0010715).

References

1. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–169
2. Furrhman JA, Campbell L (1998) Marine ecology: microbial microdiversity. *Nature* 393: 410–411
3. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95:6578–6583
4. Wu L, Thompson DK, Li G et al (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol* 67:5780–5790

5. He Z, Gentry TJ, Schadt CW et al (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* 1:67–77
6. He Z, Deng Y, Van Nostrand JD et al (2010) GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *ISME J* 4:1167–1179
7. Tu Q, Yu H, He Z et al (2014) GeoChip 4: a functional gene arrays-based high throughput environmental technology for microbial community analysis. *Mol Ecol Resour* 14:914–928
8. He Z, Wu LY, Li XY et al (2005) Empirical establishment of oligonucleotide probe design criteria. *Appl Environ Microbiol* 71:3753–3760
9. Liebich J, Schadt CW, Chong SC et al (2006) Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol* 72:1688–1691
10. Li X, He Z, Zhou J (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res* 33: 6114–6123
11. Ning J, Liebich J, Kästner M et al (2009) Different influences of DNA purity indices and quantity on PCR-based DGGE and functional gene microarray in soil microbial community study. *Appl Microbiol Biotechnol* 82: 983–993
12. NanoDrop (2007) 260/280 and 260/230 Ratios NanoDrop® ND-1000 and ND-8000 8-Sample Spectrophotometers. Technical Support Bulletin T009
13. Liang Y, He Z, Wu L et al (2010) Development of a common oligonucleotide reference standard (CORS) for microarray data normalization and comparison across different microbial communities. *Appl Environ Microbiol* 76:1088–1094
14. Xie J, Wu L, Van Nostrand JD et al (2012) Improvements on environmental DNA extraction and purification procedures for metagenomic analysis. *J Cent South Univ* 19:3055–3063
15. Wu L, Liu X, Schadt CW et al (2006) Microarray-based analysis of submicrogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* 72: 4931–4941
16. Gao H, Yang ZK, Gentry TJ et al (2007) Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. *Appl Environ Microbiol* 73:563–571
17. Zhang K, Martiny AC, Reppas NB et al (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24:680–686
18. Khan AU, Mei YH, Wilson T (1992) A proposed function for spermine and spermidine: protection of replicating DNA against damage by singlet oxygen. *Proc Natl Acad Sci U S A* 89:11426–11427
19. Marceau AH (2012) Functions of single-strand DNA-binding proteins in DNA replication, recombination, and repair. *Methods Mol Biol* 922:1–21
20. Agilent (2012) Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis. Version 3.4, July 2012. Agilent Technologies
21. Deng Y, He Z (2014) Microarray data analysis. In: He Z (ed) *Microarrays: current technology, innovations and applications*. Caister Academic Press, Norwich, UK, <http://www.horizonpress.com/microarrays2>
22. Luo Y, Hui D, Zhang D (2006) Elevated CO₂ stimulates net accumulations of carbon and nitrogen in land ecosystems: a meta-analysis. *Ecology* 87:53–63
23. ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67:1167–1179
24. Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. *Ecol Monogr* 69:1–24
25. Økland RH, Eilertsen O (1994) Canonical correspondence analysis with variation partitioning: some comments and an application. *J Veg Sci* 5:117–126
26. Ramette A, Tiedje JM (2007) Multiscale responses of microbial life in spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc Natl Acad Sci U S A* 104:2761–2766
27. Sambrook J, Russell DW (2001) *Molecular cloning a laboratory manual*, vol 1. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
28. Fare TL, Coffey EM, Dai H (2003) Effects of atmospheric ozone on microarray data quality. *Anal Chem* 75:4672–4675
29. Branham WS, Melvin CD, Han T et al (2007) Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. *BMC Biotechnol* 7:8
30. Byerly S, Sundin K, Raja R et al (2009) Effects of ozone exposure during microarray posthybridization washes and scanning. *J Mol Diagn* 11:590–597

Chapter 12

Reconstruction of Transformation Processes Catalyzed by the Soil Microbiome Using Metagenomic Approaches

Anne Schöler, Maria de Vries, Gisle Vestergaard,
and Michael Schloter

Abstract

Microorganisms are central players in the turnover of nutrients in soil and drive the decomposition of complex organic materials into simpler forms that can be utilized by other biota. Therefore microbes strongly drive soil quality and ecosystem services provided by soils, including plant yield and quality. Thus it is one of the major goals of soil sciences to describe the most relevant enzymes that are involved in nutrient mobilization and to understand the regulation of gene expression of the corresponding genes. This task is however impeded by the enormous microbial diversity in soils. Indeed, we are far to appreciate the number of species present in 1 g of soil, as well as the major functional traits they carry. Here, also most next-generation sequencing (NGS) approaches fail as immense sequencing efforts are needed to fully uncover the functional diversity of soils. Thus even if a gene of interest can be identified by BLAST similarity analysis, the obtained number of reads by NGS is too low for a quantitative assessment of the gene or for a description of its taxonomic diversity. Here we present an integrated approach, which we termed the second-generation full cycle approach, to quantify the abundance and diversity of key enzymes involved in nutrient mobilization. This approach involves the functional annotation of metagenomic data with a relative low coverage (5 Gbases or less) and the design of highly targeted primer systems to assess the abundance or diversity of enzyme-coding genes that are drivers for a particular transformation step in nutrient turnover.

Key words Soil microbiome, Metagenomics, Next-generation sequencing, Bioinformatics, Primer design, Amplicon libraries, Quantitative real-time PCR, Nutrient cycles, Ecosystem service

1 Introduction

Microbes can be considered as architects of soil quality and drive most ecosystem services provided by soils, including the promotion of plant growth, the safeguarding of drinking water and the sequestration of carbon [1].

Specific ecosystem services can be linked to unique functional traits of bacteria, archaea, and fungi. These include the degradation of complex organic compounds like xenobiotics or natural compounds like lignin, chitin or cellulose transformed into simpler

forms, which can be then used by other biota as essential nutrients [2]. In this context, the reconstruction of major soil nutrient cycle processes as well as the description of the related food web structures became a major issue in microbial ecology. In this respect research aiming to identify the key microbial players is in progress, combining network analyses with functional characterization. In the light of the ongoing climate change and the related threats for soil quality, many authors have claimed that by improving our understanding on soil microbes, we will be able to develop sustainable mitigation strategies, which ensure a protection of soil quality also for future generations [3].

However, the development of such strategies requires the answers of very important questions, for example (1) Is there a core microbiome in soils that ensures the continuous mobilization of nutrients from biomass? (2) What is the role of the rare biosphere for the resilience of soils after disturbance? (3) How can potential functional traits be induced and the corresponding genes expressed? (4) What is the best scale to study microbial processes? Although at a first glance the answer to such questions seems easy, taking a second look it can be realized that we are still far away from being able to give answers. This is closely related to the fact that soils show the highest microbial biodiversity on earth [4]. Even today we are unable to describe the microbiome of 1 g of soil. Thus, we only have rough estimations, ranging from 5×10^4 to 10^6 different species that live in 1 g of soil [5]. Taking even conservative estimations on the microbial diversity in soils into account, we would need to sequence 10^{11} bases per gram soil to cover all the diversity if an average genome size of 2 Mbases per organism is assumed. Taking into account that for statistical reasons and for the sake of building up larger contigs a coverage of at least 50-fold is needed, we end up with more than 5×10^{12} bases that need to be sequenced to assess the metagenome of such a small amount of soil. Even most recent advances in next-generation sequencing technologies make it a very challenging and time-consuming exercise.

Thus, considering the overall efforts needed to completely sequence 1 g of soil in relation to the well-known heterogeneity of microbes [6] in space and the dynamics in time, it might be worth to consider other approaches besides ultra-deep sequencing, mainly if the aim of a study is to understand the link between abiotic soil factors and a specific response of the soil microbiome. In this respect two approaches can be considered: One possibility is to couple the analysis of soil metagenomes with stable isotope probing (SIP) to identify functional communities for certain functional traits [7]. This method will be explained in detail in Chapter 14 of this book. The other possibility is to combine shotgun metagenomics with subsequent PCR-based techniques like qPCR or amplicon sequencing. In the second approach, the metagenomic

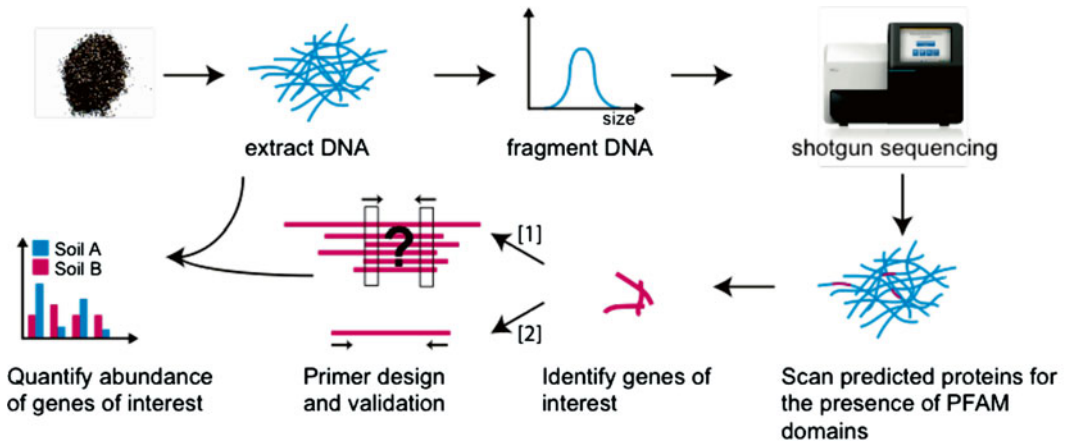


Fig. 1 Depicted is the principle of second-generation full cycle approach. DNA is extracted from soil, fragmented, and shotgun sequenced using next-generation sequencing (e.g. MiSeq). Open-reading frames (ORFs) are predicted from the reads and scanned for the presence of a protein family (PFAM) domain of interest. Subsequently, primers are designed that either detect several similar sequences (1) or one specific sequence (2). After careful validation, these primers can be used to quantify the abundance of genes or transcripts of interest. MiSeq machine image is Courtesy of Illumina, Inc

data are first functionally annotated to identify genes or gene networks to then permit the design of specific primers based on *in silico* analysis. Using these primers, a PCR-based assessment of functional genes is possible as well as a quantitative analysis of the dynamics of the respective populations in time and space. This approach has been named second-generation full cycle approach (Fig. 1) and will be described in the following in more detail focusing as a specific example on the identification of genes involved in cellulose degradation.

2 Materials

1. DNA-isolation kit “Genomic DNA from soil” NucleoSpin Soil Kit (Machery–Nagel, Germany).
2. 2100 Bioanalyzer (Agilent Technologies, United States).
3. *Agilent DNA 1000 Kit* and *Agilent High Sensitivity DNA Kit* (Agilent Technologies, United States).
4. NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, United States).
5. Covaris® Ultrasonicator (Covaris, Inc., United States).
6. Pippin Prep (Sage Science, United States).
7. Illumina® MiSeq™ instrument (Illumina, Inc., United States).
8. MiSeq™ Reagent Kit v3 (Illumina, Inc., United States).
9. Agencourt AMPure XP (Beckman Coulter, United States).

10. Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies, United States).
11. Adequate computing solution (http://support.illumina.com/sequencing/sequencing_software/casava/computing_requirements.html).

3 Methods

The described method is based on the use of the MiSeq (Illumina, San Diego, USA) sequencing platform, which is the most frequently used benchtop solution at the moment. If other instruments are used (e.g. Roche 454 Sequencing, Branford, USA), the workflow needs to be adapted according to the manufacturer. As the progress in NGS is very fast it might be useful to check also most recent developments of new kits, even if the described MiSeq is used.

3.1 *Metagenomic Library Preparation and Sequencing*

1. Total DNA is extracted from 300 mg of soil using a suitable DNA extraction method such as the DNA-isolation kit “Genomic DNA from soil” NucleoSpin Soil Kit (Macherey–Nagel, Düren, Germany). This amount may vary depending on the microbial biomass of the soil. 300 mg are sufficient for most soils under agricultural use in middle Europe. Amounts up to 5 g may be used if soils from low biomass environments like sand dunes or deserts are studied. The extracted DNA should be free from contaminating compounds like humic acids to avoid problems with any amplification step needed.
2. 1–2 µg of the extracted DNA is fragmented by ultra-sonication for approximately 80 s using an Ultrasonicator (Covaris, Woburn, USA; *see Note 1*). The success of the fragmentation can be accessed using a Bioanalyzer with the Agilent DNA 1000 Kit. The fragmented DNA should have a broad length distribution with a peak at around 600 bps (*see Note 2*).
3. The library for DNA sequencing is prepared using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Frankfurt, Germany) according to the protocol of the manufacturer without size selection.
4. Sequencing primers and barcodes are incorporated during the library preparation and allow multiplexing of many samples in one MiSeq™ run. These steps should be performed according to the protocol provided by Illumina.
5. Size selection is carried out using Pippin Prep (Sage Science, US) to select a size range between 686 and 786 bps corresponding to an insert size between 550 and 650 bps.

6. DNA is quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies, Darmstadt, Germany) or other suitable methods and pooled to sequence multiple samples in one MiSeq™ run.
7. DNA libraries are denatured and diluted to a final concentration of 15 pM and sequenced using the MiSeq™ Reagent Kit v3 to generate long high-quality reads using paired end sequencing according to the protocol provided by Illumina. This results in typical read length of 600 bps (2 × 300 bps). Overall the MiSeq instruments consistently generate 12–15 Gb of data, which can be used for further analysis.

3.2 Using HMMER to Identify Reads that Are Homologous to Genes Involved in Nutrient Turnover

The described pipeline can be used to assign reads to any microbial gene of interest. It uses hidden Markov models (HMMs) to identify homologs of known protein sequences that have the function of interest. Functional predictions using HMMs are more accurate and more sensitive compared to BLAST searches because of the strength of the underlying mathematical models. To demonstrate the practicability of the approach, wherever useful we refer to a study where putative cellulases were identified from agricultural soils. The size of the metagenome was 0.5 Gbases.

1. An overview of the entire workflow is depicted in Fig. 2. Prior to analysis several quality-filtering steps need to be carried out first to remove low-quality bases (e.g. trim read when at least three successive bases have a Phred score below 20), filter out primer sequences and remove sequences shorter than 50 bps. Several packages exist to this end, including Biopieces (www.biopieces.org) and Trimmomatic [8] (*see Note 3*). For a well-operated sequencing only few percent of reads will be removed at this step, but the reads are typically shortened to about 250 bp average length.
2. If a reasonable sequencing depth is obtained (>10 Gbases), assembly of the sequences into contigs may be more informative but is not required.
3. Next, open reading frames are predicted using for instance Frag Gene Scan [9] to obtain predicted amino acid sequences.
4. The predicted proteins are then scanned for the presence of a protein family (PFAM) motif taken from the PFAM-A database [10] using HMMER 3 [11] (*see Note 4*). For the identification of genes involved in cellulose degradation 32 different motifs exist for glycoside hydrolase (GH) families, carbohydrate binding modules (CBM) and auxiliary activities (AA). 0.3 % of all predicted open reading frames were associated with cellulose degradation.
5. Depending on the gene or group of genes of interest the PFAM motif can also be refined or generated de novo using a

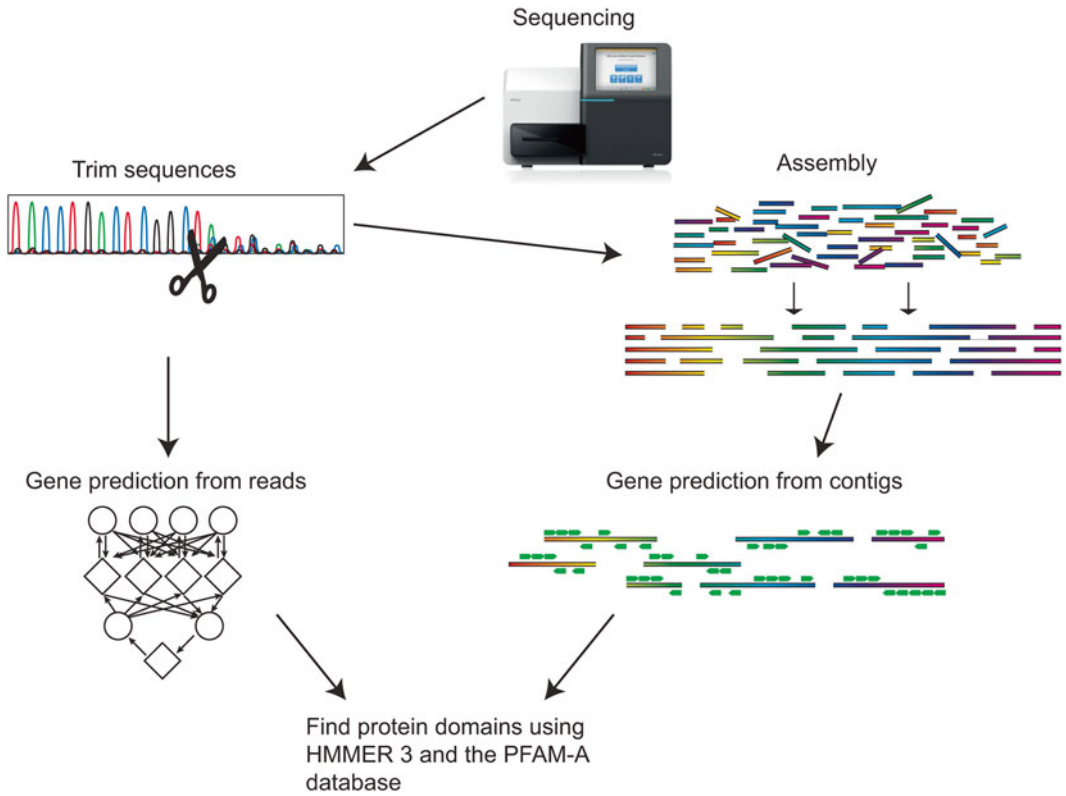


Fig. 2 Depicted is the workflow of using protein family (PFAM) motifs for the identification of genes of interest from metagenomic data. The sequencing data are quality filtered and can contain an optional step of assembly, if desired. Gene prediction is subsequently performed on either reads or contigs. These predicted proteins are then scanned (using HMMER 3) for the presence of a PFAM motif (taken from the PFAM-A database) to identify genes of interest. Miseq machine image is Courtesy of Illumina, Inc

seed of conserved sequences and HMMER 3 (*see Note 5*). For the identification of cellulose degrading enzymes the CAZy database provides a valuable resource containing protein sequences of GH, CBM, and AA families [12]. These can be used to generate new HMMs.

6. If a large number of PFAM domains are of interest, the motifs with the highest number of reads might be considered as the most important ones (*see Note 6*).

3.3 Design of Primer Systems that Amplify the Genes Identified by HMMER 3

1. After selection of the sequences of interest, primer design can be based on either conserved nucleotide or amino acid sequences shared between the sequences. In the case of our study, 32 sequences corresponding to three different GH families were selected. Note that whereas database sequences derived from complete protein sequencing often include the catalytic or functional domain, metagenome sequences are mostly shorter and might code for a part of the sequence of the

gene of interest which is not the catalytic or functional domain and thus might be highly variable, which makes the finding of a conserved motif difficult, and targeting-success cannot be guaranteed. Thus, if possible, sequences should be used which code for the functional domain of the enzyme.

2. If the goal is to design a primer pair that can identify a group of sequences with the same function (1), a clustering and alignment of the sequences within this group should be performed. If a primer pair is supposed to identify a single sequence of high importance (specific for a function) (2), the primers can be designed based on this sequence alone. Table 1 gives a short overview about both strategies.
3. For designing primers with the first goal (1), proceed as follows:
 - The sequences that are predicted by HMMER 3 to have the desired function can be clustered using cd-hit [13]. Cd-hit can cluster sequences down to a similarity of 80 %. As a comparison or reference, one or more known sequences from relevant databases can be included in the alignment for correct localization of the functional domains. For cellulase sequences, clustering resulted in very few clusters per GH family. The largest clusters contained two to three reads. If conserved regions exist, these can be used as templates for the primer design, if needed with degeneracy, i.e. including positions with variable nucleotides.

Table 1
Pros and Cons of two different primer design strategies

Strategy	Short description	Pro	Con
1—Coverage of a group of sequences coding for the same function of interest	Alignment of several sequences of interest and identification of two (relatively) conserved regions	One primer pair can identify a high diversity and obtain a more complete overview of the investigated function	There are chances that the primer pair is not specific enough and will amplify undesired sequences/false positives
2—Amplification of one specific sequence of interest	Choice of two nucleotide regions in the sequence of interest. Check for specificity against the NCBI nucleotide database	The primer pair has a high specificity	If the target sequence is not very abundant it might be difficult to amplify from metagenomic DNA. Detection of one specific sequence might have limited relevance for the entire pathway

4. For designing primers with high specificity (2) proceed as follows:
 - Primers can be designed that amplify specific sequences from a metagenome using a primer design tool from NCBI (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) selecting a size range of 60–120 bp for qPCR primers or 100–400 bp for amplicon sequencing primers.
5. If the obtained sequences from the metagenome show little conservation, public sequencing data from other metagenomes can also be queried to identify a larger number of the gene of interest and to assess if this might aid with the primer design (*see Note 7*).
6. Finally, the primer melting temperature should also be taken into account (usually around 60–62 °C); too low melting temperatures (caused by too short oligonucleotides or too high amount of A's and T's) will result in more unspecific binding. Furthermore, one can take into account the fact that binding is stronger if Gs or Cs are contained at the 3'-end of the primer (the so-called GC-clamp). In general, primers are 18–25 bps long, with 1–3 G's or C's at the 3'-end. The specificity of all designed primers should be assessed with several tools:
7. An *in silico* PCR, using for instance DeMETA-ST [14] together with a metagenome of choice or the whole NCBI database as template, is highly informative (*see Note 8*). Functional annotation of the *in silico* PCR products gives an indication of the predicted specificity. If many predicted PCR products have an unspecific function, the primer sequences should be adjusted or reconsidered. Adjusting the degeneracy of single bases by base substitution can lead to increased primer specificity (*see Note 9*). Besides *in silico* analysis, the specificity of primers should also be assessed by cloning or next-generation sequencing of the PCR products.

4 Notes

1. The duration of the ultra-sonication that is necessary to fragment the DNA sufficiently can differ with each soil and should be optimized before the experiment. Good sonication results for DNA extracted from several agricultural soils were achieved using the following settings on a Covaris E220: Duty Factor: 10 %, Cycles per Burst: 200, and Peak Incident Power: 175.
2. The proposed insert size of around 600 bp ensures that a maximal amount of the metagenome will be sequenced. If it is desired that the forward and reverse reads are to be merged in a later step the insert size should be reduced to 400–500 bps results, which will result in an overlap of around 100–200 bps of the paired end reads.

3. Useful Biopieces commands include: `trim_seq` for trimming reads and `find_adaptor` and `clip_adaptor` for removing the sequencing primer sequence.
4. The PFAM database is a large collection of protein families, and represents a good starting point for functional annotation. However, depending on the function of interest more specific databases might exist, that contain carefully curated annotations about genes of interest. Using this information to generate new and possibly more specific hidden Markov models is highly recommended.
5. It is important to keep in mind that PFAM motifs are not necessarily specific for the function of interest and may contain many other enzymes with similar structure but different function. Therefore, careful assessment of motifs is suggested. Depending on the specificity of PFAM motifs for the desired function, further filtering steps might be required. One option is to assess the specificity of PFAM motifs by blasting the output from HMMER 3 against the NCBI nr database and perform a keyword search of the best hits of the Blastp output for the desired function. This will give an indication about the specificity of each PFAM motif for a specific function.
6. If the original size of the DNA library was around 400 bp, this will result in an overlapping sequence in the forward and reverse reads. This overlap can result in a quantification bias of the motifs that lie within the overlapping region as this region is sequenced twice as many times as the not overlapping region.
7. The diversity of genes involved in nutrient cycling or any other process of interest can vary greatly. Whereas genes involved in the nitrogen cycle tend to be conserved this is not the case for cellulases, key players in the carbon cycle.
8. It is important to keep in mind that a large extent of microbial diversity remains undiscovered and that public databases are incomplete, therefore *in silico* PCR can be informative but are not sufficient to assess primer specificity.
9. The *in silico* PCR using the NCBI database as template can be informative, but is not a realistic scenario. PCR products of organisms that most probably will not be present in the analyzed samples might give a false indication of primers.

References

1. Power AG (2010) Ecosystem services and agriculture: tradeoffs and synergies. *Philos Trans R Soc Lond B Biol Sci* 365:2959–2971
2. Maron PA, Mougél C, Ranjard L (2011) Soil microbial diversity: methodological strategy, spatial overview and functional interest. *C R Biol* 334:403–411
3. Griffiths BS, Philippot L (2013) Insights into the resistance and resilience of the soil microbial community. *FEMS Microbiol Rev* 37:112–129
4. Lee MH, Lee SW (2013) Bioprospecting potential of the soil metagenome: novel enzymes and bioactivities. *Genomics Inform* 11:114–120

5. Torsvik V, Øvreås L (2012) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* 5:240–245
6. Shi Y, Yang H, Zhang T, Sun J, Lou K (2014) Illumina-based analysis of endophytic bacterial diversity and space-time dynamics in sugar beet on the north slope of Tianshan mountain. *Appl Microbiol Biotechnol* 98:6375–6385
7. Neufeld JD et al (2007) DNA stable-isotope probing. *Nat Protoc* 2:860–866
8. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120
9. Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38:e191
10. Finn RD et al (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230. doi:[10.1093/nar/gkt1223](https://doi.org/10.1093/nar/gkt1223)
11. Eddy SR (2011) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23:205–211
12. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42(Database issue):D490–D495. doi:[10.1093/nar/gkt1178](https://doi.org/10.1093/nar/gkt1178)
13. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152
14. Gulvik CA, Effler TC, Wilhelm SW, Buchan A (2012) De-MetaST-BLAST: a tool for the validation of degenerate primer sets and data mining of publicly available metagenomes. *PLoS One* 7:e50362. doi:[10.1371/journal.pone.0050362](https://doi.org/10.1371/journal.pone.0050362)

Chapter 13

MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function

Kevin P. Keegan, Elizabeth M. Glass, and Folker Meyer

Abstract

Approaches in molecular biology, particularly those that deal with high-throughput sequencing of entire microbial communities (the field of metagenomics), are rapidly advancing our understanding of the composition and functional content of microbial communities involved in climate change, environmental pollution, human health, biotechnology, etc. Metagenomics provides researchers with the most complete picture of the taxonomic (i.e., *what organisms are there*) and functional (i.e., *what are those organisms doing*) composition of natively sampled microbial communities, making it possible to perform investigations that include organisms that were previously intractable to laboratory-controlled culturing; currently, these constitute the vast majority of all microbes on the planet. All organisms contained in environmental samples are sequenced in a culture-independent manner, most often with 16S ribosomal amplicon methods to investigate the taxonomic or whole-genome shotgun-based methods to investigate the functional content of sampled communities. Metagenomics allows researchers to characterize the community composition and functional content of microbial communities, but it cannot show which functional processes are active; however, near parallel developments in transcriptomics promise a dramatic increase in our knowledge in this area as well.

Since 2008, MG-RAST (Meyer et al., BMC Bioinformatics 9:386, 2008) has served as a public resource for annotation and analysis of metagenomic sequence data, providing a repository that currently houses more than 150,000 data sets (containing 60+ tera-base-pairs) with more than 23,000 publically available. MG-RAST, or the metagenomics RAST (rapid annotation using subsystems technology) server makes it possible for users to upload raw metagenomic sequence data in (preferably) fastq or fasta format. Assessments of sequence quality, annotation with respect to multiple reference databases, are performed automatically with minimal input from the user (*see* Subheading 4 at the end of this chapter for more details). Post-annotation analysis and visualization are also possible, directly through the web interface, or with tools like matR (metagenomic analysis tools for R, covered later in this chapter) that utilize the MG-RAST API (<http://api.metagenomics.anl.gov/api.html>) to easily download data from any stage in the MG-RAST processing pipeline. Over the years, MG-RAST has undergone substantial revisions to keep pace with the dramatic growth in the number, size, and types of sequence data that accompany constantly evolving developments in metagenomics and related -omic sciences (e.g., metatranscriptomics).

Key words Metagenomics, Comparative analysis, Sequence quality, Automated pipeline, High-throughput, matR

1 Introduction

Developing approaches in molecular biology are rapidly advancing our understanding of the composition and functional content of microbial communities. This has led to a much clearer picture of the pivotal role these communities play in phenomena as diverse as climate change, environmental pollution, human health, and developments in biotechnology. Metagenomics utilizes cutting-edge technology in sequencing and sequence characterization to create an inventory of microbial genes that are present in any given environment, including those contained in microbes intransigent to classical culture-based methods; currently, these constitute the vast majority (estimates typically report 99 % or more) of all microbes on the planet. All organisms contained in native microbial communities (also referred to as assemblages) are sequenced in a culture-independent manner, most often with 16S ribosomal amplicon methods to investigate the taxonomic or whole-genome shotgun-based methods to investigate the functional content of sampled communities. This makes it possible to create a much clearer picture of the composition (*who is there*) and potential functional content (*what can they do*) of microbial communities than was possible with previous methods. Metatranscriptomics extends this knowledge by providing us with a catalog that can link functions active in a community (*what are they doing*) to the temporal and conditional variables that drive them (*why are they doing ...*).

For these kinds of sequence-dependent studies, the underlying quality of sequence data is a fundamental concern, complicated by the use of an ever-expanding assortment of methods, equipment, and software. Metagenomic analyses rely on the use of highly automated analysis tools; therefore, early identification of quality-related problems is essential to avoid wasteful use of limited computational resources as well as interpretation of fundamentally flawed data that can lead to erroneous biological inferences.

In regards to sequence quality, the scientific community faces another hurdle with the study of metagenomics data. Most researchers understand how critical it is for sequence data to exhibit the highest possible quality—especially in applications where a high level of functional or taxonomic resolution is desired—but do not possess the technical expertise to assess quality (i.e., independently from metrics provided by black-box vendor-specific software and/or sequencing centers that may not be using the most current or best practices). MG-RAST possesses multiple features that make it easy for users to assess sequence quality and address some of the most common issues (e.g., high error rates, contamination with adapter sequences, contamination with artificial duplicate reads, etc.).

In recent years, the sequencing costs have dramatically reduced whereas costs of computing have remained relatively stable. This has shifted the limiting factor in sequence-dependent investigations from data generation (i.e., sequencing) to data analysis (annotation and post-annotation analyses). Wilkening et al. [1] provide a real currency cost for the analysis of 100 giga-base-pairs of DNA sequence data using BLASTX on Amazon's EC2 service, \$300,000. A more recent study by University of Maryland researchers [2] estimates the computation for a terabase of DNA shotgun data using their CLOVR metagenome analysis pipeline at over \$5 million per terabase. As the size and number of sequence data sets continue to increase, costs related to their analysis will continue to rise.

In addition, metadata (data describing data—e.g., data that describe the temporal and environmental parameters for a sampled microbial community) provide an essential complement to experimental data, helping to answer questions about its source, mode of collection, and reliability as well as making it possible answer meaningful biological questions (e.g., what factor(s) cause a shift in the composition or functional content of a microbial community in a particular environment). Metadata collection and interpretation have become vital to the genomics and metagenomics community, but considerable challenges remain, including exchange, curation, and distribution.

Since 2008, MG-RAST [3] has served as both a repository and tool for the analysis of metagenomic data (and metadata)—annotation and post-annotation analyses. Currently, the system has analyzed more than 60 tera-base-pairs of data from more than 150,000 data sets, with more than 23,000 available to the public. Over the years, MG-RAST has undergone substantial revisions to keep pace with the dramatic growth in the number, size, and types of sequence data that accompany constantly evolving developments in metagenomics and related -omic sciences (e.g., metatranscriptomics). These include innovations in engineering as well as modifications to our bioinformatics pipeline to accommodate the evolving needs of novel sequencing technologies and growing data volumes. The MG-RAST system has been an early adopter of the minimal checklist standards and the expanded biome-specific environmental packages devised by the Genomics Standards Consortium [4] and provides an easy-to-use uploader for metadata capture at the time of data submission.

2 Materials

2.1 Database

The MG-RAST automated analysis pipeline uses the M5nr (MD5-based non-redundant protein database) [5] for annotation. The M5nr is an integration of many sequence databases into one single,

indexed, searchable database. A single similarity search (using BLAST [6] or BLAT [7]) allows the user to retrieve similarities to several databases,

- EBI, European Bioinformatics Institute [8]
- GO, Gene Ontology [9]
- JGI, Joint Genome Institute [10]
- KEGG, Kyoto Encyclopedia of Genes and Genomes [11]
- NCBI, National Center for Biotechnology Information [12]
- Phantome, Phage Annotation Tools and Methods [13]
- SEED, The SEED Project [14]
- UniProt, UniProt Knowledgebase [15]
- VBI, Virginia Bioinformatics Institute [16]
- eggNOG, evolutionary genealogy of genes, Non-supervised Orthologous Groups [17]

Computation of sequence similarities is becoming a limiting factor in metagenomic analyses. Sequence similarity search results encoded in an open, exchangeable format distributed with the sequence sets have the potential to limit the needs for computational re-analysis of data sets. A prerequisite for sharing of similarity results is a common reference—this is exactly what the M5nr provides, a commonly indexed database that contains all of databases noted above.

M5nr mechanisms are used for automatically maintaining this comprehensive non-redundant protein database and creating a quarterly release of this resource. In addition, MG-RAST provides tools for translating similarity searches into many namespaces, e.g., KEGG, NOG, SEED Subsystems, NCBI's GenBank, etc.

2.2 Analysis Pipeline

The pipeline shown in Fig. 1 contains a number of improvements to previous MG-RAST versions. Several key algorithmic improvements were needed to support the flood of user-generated data. Initial analysis differentiates amplicon-based ribosomal 16S from whole-genome shotgun (WGS) samples and processed them separately (*see* Subheading 3.1 below for processing of WGS data and Subheading 3.2 for processing of amplicon 26s data). WGS samples are analyzed with dedicated software to perform gene prediction on nucleotide data prior to protein similarity-based annotation. This provides drastic runtime improvements over nucleotide similarity-based approaches. Clustering of predicted proteins at 90 % identity provides additional performance improvement while preserving biological signals. While protein-based annotation is used for proteins predicted from WGS samples, samples detected as 16S ribosomal data are annotated with nucleotide-based similarity.

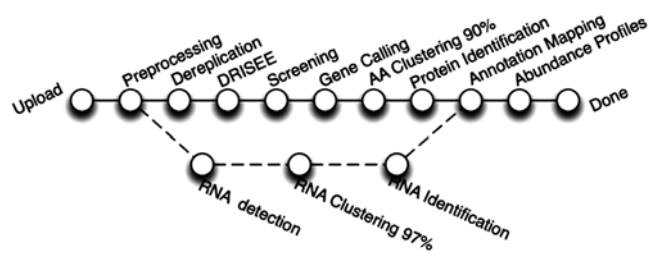


Fig. 1 Details of the analysis pipeline for MG-RAST. After upload, the pipeline diverges for amplicon and WGS data sets. Amplicon samples run through RNA detection, clustering, and identification. WGS samples undergo a number of additional processing steps to assess data quality prior to annotation

In Subheading 3, we describe each step of the pipeline in detail. All data sets generated by the individual stages of the processing pipeline are made available as downloads.

2.3 Compute Resources

While MG-RAST was originally built as a traditional, centrally located, cluster-based bioinformatics system, the most recent version embraces novel technologies that make it possible for it to utilize local and remote compute resources. MG-RAST data are stored in SHOCK [18] and computing is orchestrated by AWE [19]. These technologies were developed to enable execution on a variety of computational platforms; currently, computational resources are contributed by the DOE Magellan cloud at Argonne National Laboratory, Amazon EC2 Web services provided by individual users, and a number of traditional clusters. An installation of the pipeline exists at DOE's NERSC supercomputing center. In recent months, the system handles over 4 tera-base-pairs of data per month. The use of Skyport [38] has enabled multi cloud workflows without introducing management overhead.

3 Methods

The pipeline diverges after upload for 16S ribosomal amplicon and whole-genome shotgun (WGS) samples. The WGS pipeline is composed of several steps from the removal of low-quality reads, dereplication, gene calling, and annotation to creation of functional abundance profiles. rRNA samples run through RNA detection, clustering, and identification, and the production of taxonomic abundance profiles. Subheading 4 found at the end of this chapter includes additional details.

3.1 The WGS Pipeline

1. *Preprocessing.* After upload, data are preprocessed by using SolexaQA [20] to trim low-quality regions from FASTQ data. Platform-specific approaches are used for 454 data submitted in FASTA format, reads more than two standard deviations away from the mean read length are discarded [21]. All sequences submitted to the system are available, but discarded reads are not analyzed further.

2. *Dereplication.* For shotgun metagenome and shotgun metatranscriptome data sets, we perform a dereplication step. We use a simple k-mer approach to rapidly identify all 20 character prefix identical sequences. This step is required in order to remove artificial duplicate reads (ADRs) [22]. Instead of simply discarding the ADRs, we set them aside and use them later as a means to assess sample quality. We note that dereplication is not suitable for amplicon data sets that are likely to share common prefixes.
3. *DRISEE.* MG-RAST v3 uses DRISEE (Duplicate Read Inferred Sequencing Error Estimation) [23] to analyze the sets of ADRs and determine the degree of variation among prefix-identical sequences derived from the same template. See below for details.
4. *Screening.* The pipeline provides the option of removing reads that are near-exact matches to the genomes of a handful of model organisms, including fly, mouse, cow, and human. The screening stage uses Bowtie [24] (a fast, memory-efficient, short read aligner), and only reads that do not match the model organisms pass into the next stage of the annotation pipeline.

Note that this option will remove all reads similar to the human genome and render them inaccessible. This decision was made in order to avoid storing any human DNA on MG-RAST.

5. *Gene calling.* The previous version of MG-RAST used nucleotide-based similarity for annotation of WGS data, an approach that is significantly more expensive computationally than de novo gene prediction followed by protein similarity-based annotation. After an in-depth investigation of tool performance [25], we have moved to a machine learning approach that utilizes FragGeneScan [26] to predict proteins/protein fragments from de novo sequence data (FragGeneScan uses a well tested algorithm [25] to perform in silico translation of predicted protein coding nucleic acid sequences). Utilizing this approach, we can predict coding regions in DNA sequences that are 75 base pairs or longer. Our novel approach also enables the analysis of user-provided assembled contigs. We note that FragGeneScan is trained for prokaryotes only. While it will identify proteins for eukaryotic sequences, the results should be viewed critically.
6. *AA clustering.* MG-RAST builds clusters of proteins at the 90 % identity level using the uclust [27] implementation in QIIME [28], preserving the relative abundances. These clusters greatly reduce the computational burden of comparing all pairs of short reads, while clustering at 90 % identity preserves sufficient biological signals.

7. *Protein identification.* Once created, a representative (the longest sequence) for each cluster is subjected to similarity analysis. Functional identification of representative sequences does not use BLAST, instead we use a much more efficient algorithm, sBLAT, an implementation of the BLAT algorithm, which we parallelized using OpenMPI. We reconstruct the putative species composition of WGS data by looking at the phylogenetic origin of the database sequences hit by the protein-based similarity searches. Note that processing of rRNA 16S amplicon data is covered in Subheading 3.2 below.
8. *Annotation mapping.* Sequence similarity searches are computed against a protein database derived from the M5NR, which provides nonredundant integration of many databases. Users can easily change views without recomputation. For example, COG and KEGG views can be displayed, which both show the relative abundances of histidine biosynthesis in a data set of four cow rumen metagenomes.

Help in interpreting results, MG-RAST searches the nonredundant M5NR and M5RNA databases in which each sequence is unique. These two databases are built from multiple sequence database sources, and the individual sequences may occur multiple times in different strains and species (and sometimes genera) with 100 % identity. In these circumstances, choosing the “right” taxonomic information is not a straightforward process. To optimally serve a number of different use cases, we have implemented three methods for end users to determine the number of hits (occurrences of the input sequence in the database) in their samples.

- *Best hit,* The best hit classification reports the functional and taxonomic annotation of the best hit in the M5NR for each feature. In those cases where the similarity search yields multiple same-scoring hits for a feature, we do not choose any single “correct” label. For this reason MG-RAST double counts all annotations with identical match properties and leaves determination of truth to our users. While this approach aims to inform about the functional and taxonomic potential of a microbial community by preserving all information, subsequent analysis can be biased (e.g., a single feature may have multiple annotations), leading to inflated hit counts. For users looking for a specific species or function in their results, the best hit classification is likely what is wanted.
- *Representative hit,* The representative hit classification selects a single, unambiguous annotation for each feature. The annotation is based on the first hit in the homology search and the first annotation for that hit in the database. This approach makes counts additive across functional and

taxonomic levels and is better suited for comparisons of functional and taxonomic profiles of multiple metagenomes.

- *Lowest Common Ancestor (LCA)*, To avoid the problem of multiple taxonomic annotations for a single feature, MG-RAST provides taxonomic annotations based on the widely used LCA method introduced by MEGAN [29]. In this method, all hits are collected that have a bit score close to the bit score of the best hit. The taxonomic annotation of the feature is then determined by computing the LCA of all species in this set. This replaces all taxonomic annotations from ambiguous hits with a single higher-level annotation in the NCBI taxonomy tree.
9. *Abundance profiles*. Abundance profiles (essentially tables that indicate detected taxa or functions and their relative abundance as determined by the methods described in Subheading 3.1, **step 8**—examples can be found in the MG-RAST user manual, *see* the “additional documentation” in Subheading 4 found at the end of this chapter) are the primary data product that the MG-RAST’s user interface uses to display information in annotated data sets. Using the abundance profiles, the MG-RAST system defers to the user to select several parameters that will define their abundance data, *e*-value, percent identity, and minimal alignment length. As it is not possible to arbitrarily select thresholds suitable for all use cases, users can select their own thresholds for each of these values.

Taxonomic profiles use the NCBI taxonomy. All taxonomic information is projected against the NCBI taxonomy.

Functional profiles are available for data sources that provide hierarchical information. These currently include SEED subsystems, KEGG orthologs, and COGs. SEED subsystems represent an independent reannotation effort utilized by RAST [30] and MG-RAST. Manual curation of subsystems makes them an extremely valuable data source. The current subsystems hierarchy can be viewed at <http://pubseed.theseed.org//SubsysEditor.cgi> which allows browsing the subsystems.

Subsystems represent a four-level hierarchy,

1. Subsystem level 1—highest level
2. Subsystem level 2—intermediate level
3. Subsystem level 3—similar to a KEGG pathway
4. Subsystem level 4—actual functional assignment to the feature in question

KEGG Orthologs. MG-RAST uses the KEGG enzyme number to implement a four-level hierarchy. We note that KEGG data are no longer available for free download; therefore, we rely on the latest freely downloadable version of these data.

1. KEGG level 1—first digit of the EC number (EC,X.*.*.*)
2. KEGG level 2—first two digits of the EC number (EC,X.Y.*.*)
3. KEGG level 3—first three digits of the EC number (EC,X,Y,Z.*)
4. KEGG level 4—entire four digits of the EC number

The high-level KEGG categories are as follows:

1. Cellular Processes
2. Environmental Information Processing
3. Genetic Information Processing
4. Human Diseases
5. Metabolism
6. Organizational Systems

COG and EGGNOG Categories. The high-level COG and EGGNOG categories are as follows:

1. Cellular Processes
2. Information Storage and Processing
3. Metabolism
4. Poorly Characterized

3.2 The rRNA Pipeline

The rRNA pipeline starts with upload of rRNA reads and proceeds through the following steps:

1. *rRNA detection*. Reads are identified as rRNA through a simple rRNA detection. An initial BLAT search against a reduced RNA database efficiently identifies RNA. The reduced database is a 90 % identity clustered version of the SILVA database and is used merely to differentiate samples containing solely rRNA data from other samples (e.g., WGS or transcriptomic samples).
2. *rRNA clustering*. The rRNA-similar reads are clustered at 97 % identity, and the longest sequence is picked as the cluster representative.
3. *rRNA identification*. A nucleotide BLAT similarity search for the longest cluster representative is performed against the M5rna database, integrating SILVA [31], Greengenes [32], and RDP [33].

3.3 Using the MG-RAST User Interface

The MG-RAST system provides a rich web user interface that covers all aspects of metagenome analysis, from data upload to ordination analysis of annotation abundances. The web interface can also be used for data discovery. Metagenomic data sets can be easily selected individually or on the basis of filters such as technology (including read length), quality, sample type, and keyword, with dynamic filtering of results based on similarity to known reference

proteins or taxonomy. For example, a user may want to perform a search such as “phylum eq ‘actinobacteria’ and function in KEGG pathway Lysine Biosynthesis and sample in ‘Ocean’” to extract sets of reads matching the appropriate functions and taxa across metagenomes. The results can be displayed in familiar formats, including bar charts, trees that incorporate abundance information, heatmaps, principal component analyses, or raw abundance tables exported in tabular form. The raw or processed data can be recovered via download pages or with the `matR` package for R (*see* Subheading 4 below). Metabolic reconstructions based on mapping to KEGG pathways are also provided.

Sample selection is crucial for understanding large-scale patterns when multiple metagenomes are compared. Accordingly, MG-RAST supports MIXS and MIMARKS [34] (as well as domain-specific plug-ins for specialized environments not extending the minimal GSC standards); several projects, including TerraGenome, HMP, TARA, and EMP, use these GSC standards, enabling standardized queries that integrate new samples into these massive data sets.

One key aspect of the MG-RAST approach is the creation of smart data products enabling the user at the time of analysis to determine the best parameters for, for example, a comparison between samples. This is done without the need for recomputation of results.

3.3.1 Navigation

The MG-RAST website is rich with functionality and offers several options. The site at <http://metagenomics.anl.gov> has five main pages and a home page, shown in blue in Fig. 2.

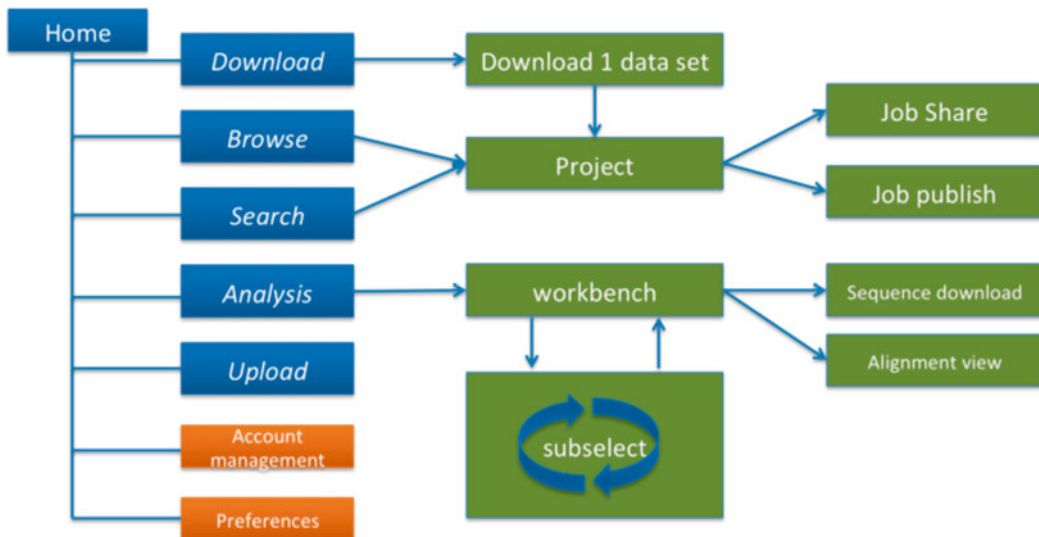


Fig. 2 Sitemap for the MG-RAST version 3 website. On the site map the main pages are shown in *blue*, management pages in *orange*. The *green boxes* represent pages that are not directly accessible from the home page

- Download page—lists all publicly available data for download. The data are structured into projects.
- Browse page—allows interactive browsing of all data sets and is powered by metadata.
- Search page—allows identifier, taxonomy, and function-driven searches against all public data.
- Analysis page—enables in-depth analyses and comparisons between data sets.
- Upload page—allows users to provide their samples and metadata to MG-RAST.
- Home (Metagene Overview) page—provides an overview for each individual data set.

3.3.2 Upload Page

Data and metadata can be uploaded in the form of spreadsheets along with the sequence data by using both the ftp and the http protocols. The web uploader will automatically split large files and also allows parallel uploads. MG-RAST supports data sets that are augmented with rich metadata using the standards and technology developed by the GSC. Each user has a temporary storage location inside the MG-RAST system. This inbox provides temporary storage for data and metadata to be submitted to the system. Using the inbox, users can extract compressed files, convert a number of vendor-specific formats to MG-RAST submission-compliant formats, and obtain an MD5 checksum for verifying that transmission to MG-RAST has not altered the data. The web uploader has been optimized for large data sets of over 100 giga-base-pairs, often resulting in file sizes in excess of 150 GB.

3.3.3 Browse Page: Metadata-Enabled Data Discovery

The Browse page lists all data sets visible to the user (the users own data sets as well as all public data and all data shared by other users). This page also provides an overview of the nonpublic data sets submitted by the user or shared with users. The interactive metagenome browse table provides an interactive graphical means to discover data based on technical data (e.g., sequence type or data set size) or metadata (e.g., location or biome).

3.3.4 Project Page

The project page provides a list of data sets and metadata for a project. The table at the bottom of the Project page provides access to the individual metagenomes by clicking on the identifiers in the first column. In addition, the final column provides downloads for metadata, submitted data, and the analysis results via the three labeled arrows. For the data set owners, the Project page provides an editing capability using a number of menu entries at the top of the page. Figure 3 shows the available options.

- Share Project—make the data in this project available to third parties via sending them access tokens.
- Add Jobs—add additional data sets to this project.

THE ORAL METAGENOME IN HEALTH AND DISEASE (ID 128) [metagenomes](#) [project metadata](#)

Visibility: Public
 Static Link: <http://metagenomics.afl.gov.au/inx.cgi?project=128>

[Share Project](#) [Add Jobs](#) [Edit Project Data](#) [Upload Info](#) [Upload MetaData](#) [Export MetaData](#)

DESCRIPTION

The oral cavity of humans is inhabited by hundreds of bacterial species and some of them have a key role in the development of oral diseases, mainly dental caries and periodontitis. We describe for the first time the metagenome of the human oral cavity under health and diseased conditions, with a focus on supragingival dental plaque and cavities. Direct pyrosequencing of eight samples with different oral-health status produced 1 Gbp of sequence without the biases imposed by PCR or cloning. These data show that cavities are not dominated by *Streptococcus mutans* (the species originally identified as the ethiological agent of dental caries) but are in fact a complex community formed by tens of bacterial species, in agreement with the view that caries is a polymicrobial disease. The analysis of the reads indicated that the oral cavity is functionally a different environment from the gut, with many functional categories enriched in one of the two environments and depleted in the other. Individuals who had never suffered from dental caries showed an over-representation of several functional categories, like genes for antimicrobial peptides and quorum sensing. In addition, they did not have *mutans streptococci* but displayed high recruitment of other species. Several isolates belonging to these dominant bacteria in healthy individuals were cultured and shown to inhibit the growth of cariogenic bacteria, suggesting the use of these commensal bacterial strains as probiotics to promote oral health and prevent dental caries.

FUNDING SOURCE

Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider- Ingenio program.

CONTACT

Administrative
 Alex Mira (CSISP)
 Avda. Cataluña, 21. Valencia, Spain

Technical
 Pedro Belda-Ferre (Center for Advanced Research in Public Health, Department of Genomics and Health)
 Avda. Cataluña, 21 ; 46020 ; Valencia ; Comunidad Valenciana, Spain

ADDITIONAL DATA

administrative-contact_PI_lastname	Mira
project-description_internal_project_ID	The oral metagenome in health and disease
administrative-contact_PI_email	mira_ale@gva.es
administrative-contact_PI_organization	Center for Advanced Research in Public Health, Department of Genomics and Health
administrative-contact_PI_organization_country	Spain
administrative-contact_PI_organization_address	Avda. Cataluña, 21 ; 46020 ; Valencia ; Comunidad Valenciana
administrative-contact_PI_organization_uri	www.csisp.gva.es/web/csisp
administrative-contact_PI_firstname	Alex

METAGENOMES

There are 8 metagenomes in this project.

[Export Jobs Table](#)

MG-RAST ID	Metagenome Name	bp Count	Sequence Count	Biome	Feature	Material	Location	Country	Coordinates	Sequence Type	Sequence Method	Download
		<	<	humar	human-	human-				WGS	454	
4447843.3	CA_04P	142,374,233	339,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447192.3	NOCA_01P	77,538,485	204,218	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447103.3	CA1_01P	203,711,161	464,594	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447103.3	NOCA_03P	100,126,119	244,881	human-	human-	human-	Valencia	Spain	39.481448	WGS	454	metadata submitted analysis

Fig. 3 Project page, providing a summary of all data in the project and an interface for downloads

- Edit Project Data—edit the contents of this page.
- Upload Info—upload information to be displayed on this page.
- Upload MetaData—upload a metadata spreadsheet for the project.
- Export MetaData2—export the metadata spreadsheet for this project.

3.3.5 Overview Page

MG-RAST automatically creates an individual summary page for each data set. This metagenome overview page provides a summary of the annotations for a single data set. The page is made available

by the automated pipeline once the computation is finished. This page is a good starting point for looking at a particular data set. It provides information regarding technical detail and biological content. The page is intended as a single point of reference for metadata, quality, and data. It also provides an initial overview of the analysis results for individual data sets with default parameters. Further analyses are available on the Analysis page.

Technical Details on Sequencing and Analysis

The Overview page provides the MG-RAST ID for a data set, a unique identifier that is usable as an accession number for publications. Additional information, such as the name of the submitting PI, organization, and a user-provided metagenome name are displayed at the top of the page. A static URL for linking to the system that will be stable across changes to the MG-RAST web interface is provided as additional information (Fig. 7).

MG-RAST provides an automatically generated paragraph of text describing the submitted data and the results computed by the pipeline. By means of the project information, we display additional information provided by the data submitters at the time of submission or later.

One of the key diagrams in MG-RAST is the sequence breakdown pie chart (Fig. 4) classifying the submitted sequences into several categories according to their annotation status. As detailed in the description of the MG-RAST v3 pipeline above, the features annotated in MG-RAST are protein coding genes and ribosomal proteins.

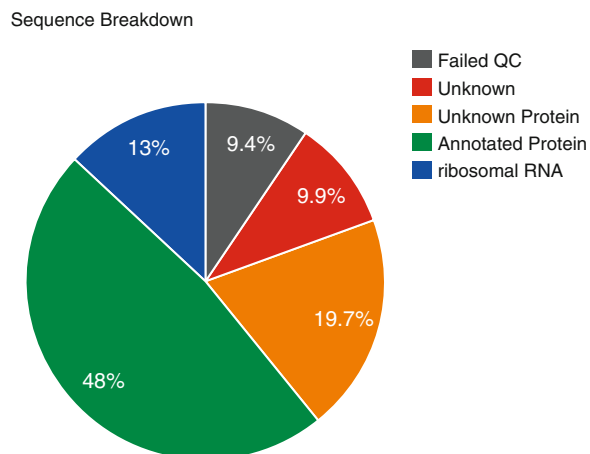


Fig. 4 Sequences to the pipeline are classified into one of five categories: *grey*=failed the QC, *red*=unknown sequences, *yellow*=unknown function but protein coding, *green*=protein coding with known function, and *blue*=ribosomal RNA. For this example, over 50 % of sequences were either filtered by QC or failed to be recognized as either protein coding or ribosomal

Note that for performance reasons no other sequence features are annotated by the default pipeline. Other feature types such as small RNAs or regulatory motifs (e.g., CRISPRs [35]) not only will require significantly higher computational resources but also are frequently not supported by the unassembled short reads that constitute the vast majority of today's metagenomic data in MG-RAST.

The quality of the sequence data coming from next-generation instruments requires careful design of experiments, lest the sensitivity of the methods is greater than the signal-to-noise ratio the data supports.

The overview page also provides metadata for each data set to the extent that such information has been made available. Metadata enables other researchers to discover data sets and compare annotations. MG-RAST requires standard metadata for data sharing and data publication. This is implemented using the standards developed by the Genomics Standards Consortium.

All metadata stored for a specific data set is available in MG-RAST; we merely display a standardized subset in this table. A link at the bottom of the table ("More Metadata") provides access to a table with the complete metadata. This enables users to provide extended metadata going beyond the GSC minimal standards. A mechanism to provide community consensus extensions to the minimal checklists and the environmental packages are explicitly encouraged but not required when using MG-RAST.

Metagenome Quality Control

The analysis flowchart and analysis statistics provide an overview of the number of sequences at each stage in the pipeline. The text block next to the analysis flowchart presents the numbers next to their definitions.

Source Hits Distribution

The source hits distribution shows the percentage of the predicted protein features annotated with similarity to a protein of known function per source database. In addition, ribosomal RNA genes are mapped to the rRNA databases.

In addition, this display will print the number of records in the M5NR protein database and in the M5RNA ribosomal databases.

Other Statistics

MG-RAST also provides a quick link to other statistics. For example, the Analysis Statistics and Analysis Flowchart provide sequence statistics for the main steps in the pipeline from raw data to annotation, describing the transformation of the data between steps. Sequence length and GC histograms display the distribution before and after quality control steps. Metadata is presented in a searchable table that contains contextual metadata describing sample location, acquisition, library construction, and sequencing using GSC compliant metadata. All metadata can be downloaded from the table.

3.3.6 Biological Part of the Overview Page

The taxonomic hit distribution display divides taxonomic units into a series of pie charts of all the annotations grouped at various taxonomic ranks (domain, phylum, class, order, family, genus). The subsets are selectable for downstream analysis; this also enables downloads of subsets of reads, for example, those hitting a specific taxonomic unit.

Rank Abundance

The rank abundance plot provides a rank-ordered list of taxonomic units at a user-defined taxonomic level, ordered by their abundance in the annotations.

Rarefaction

The rarefaction curve of annotated species richness is a plot (*see* Fig. 5) of the total number of distinct species annotations as a function of the number of sequences sampled. The slope of the right-hand part of the curve is related to the fraction of sampled species that are rare. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled, more intensive sampling is likely to yield only few additional species. Sampling curves generally rise quickly at first and then level off toward an asymptote as fewer new species are found per unit of individuals collected.

The rarefaction curve is derived from the protein taxonomic annotations and is subject to problems stemming from technical artifacts. These artifacts can be similar to the ones affecting amplicon sequencing [36], but the process of inferring species from protein similarities may introduce additional uncertainty.

Alpha Diversity

In this section, we display an estimate of the alpha diversity based on the taxonomic annotations for the predicted proteins. The alpha diversity is presented in context of other metagenomes in the same project (*see* Fig. 6).

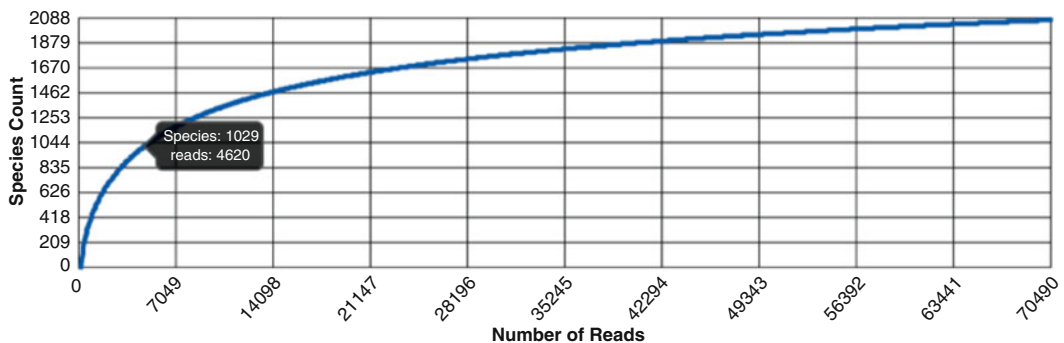


Fig. 5 Rarefaction plot showing a curve of annotated species richness (i.e., the number of unique species). This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled

α -Diversity = 377.113 species

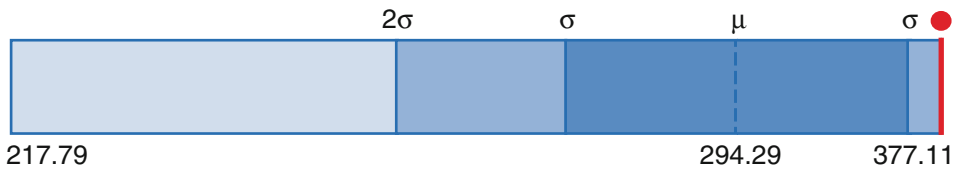


Fig. 6 Alpha diversity plot showing the range of α -diversity values in the project the data set belongs to. The min, max, and mean values are shown, with the standard deviation ranges in different shades. The alpha-diversity of this metagenome is shown in red. The species-level annotations are from all the annotation source databases used by MG-RAST

The alpha diversity estimate is a single number that summarizes the distribution of species-level annotations in a data set. The Shannon diversity index is an abundance-weighted average of the logarithm of the relative abundances of annotated species. We compute the species richness σ as the antilog of the Shannon diversity.

Functional Categories

This section contains four pie charts providing a breakdown of the functional categories for KEGG, COG, SEED Subsystems, and EggNOGs. Clicking on the individual pie chart slices will save the respective sequences to the workbench. The relative abundance of sequences per functional category can be downloaded as a spreadsheet, and users can browse the functional breakdowns.

A more detailed functional analysis, allowing the user to manipulate parameters for sequence similarity matches, is available from the Analysis page.

3.3.7 Analysis Page

The MG-RAST annotation pipeline produces a set of annotations for each sample; these annotations can be interpreted as functional or taxonomic abundance profiles. The analysis page can be used to view these profiles for a single metagenome or to compare profiles from multiple metagenomes using various visualizations (e.g., heatmap) and statistics (e.g., PCoA, normalization).

The page is divided into three parts following a typical workflow (Fig. 7).

1. Data type

Selection of an MG-RAST analysis scheme, that is, selection of a particular taxonomic or functional abundance profile mapping. For taxonomic annotations, since there is not always a unique mapping from hit to annotation, we provide three interpretations: best hit, representative hit, and lowest common ancestor. When choosing the LCA annotations, not all downstream tools are available. The reason is the fact that for the LCA annotations not all sequences will be annotated to the same level, classifications are returned on different taxonomic levels.

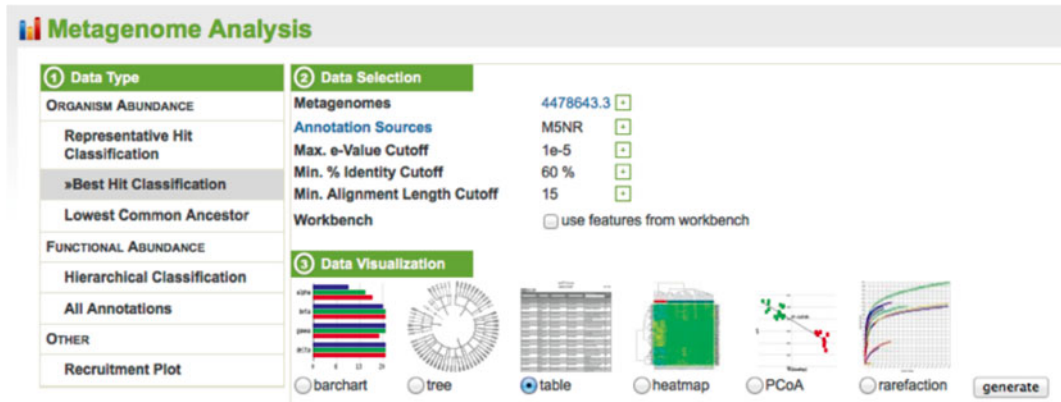


Fig. 7 Three-step process in using the Analysis page: (1) select a profile and hit (see text) type; (2) select a list of metagenomes and set annotation source and similarity parameters; (3) choose a comparison

Functional annotations can be grouped into mappings to functional hierarchies or can be displayed without a hierarchy. In addition, the recruitment plot displays the recruitment of protein sequences against a reference genome. Each selected data type has data selections and data visualizations specific for it.

2. Data selection

Selection of sample and parameters. This dialog allows the selection of multiple metagenomes that can be compared individually or selected and compared as groups. Comparison is always relative to the annotation source, e -value, and percent identity cut-offs selectable in this section. In addition to the metagenomes available in MG-RAST, sets of sequences previously saved in the workbench can be selected for visualization.

3. Data visualization

Data visualization and comparison. Depending on the selected profile type, the profiles for the metagenomes can be visualized and compared by using barcharts, trees, spreadsheet-like tables, heatmaps, PCoA, rarefaction plots, circular recruitment plot, and KEGG maps.

The data selection dialog provides access to data sets in four ways. The four categories can be selected from a pull-down menu.

- private data—list of private or shared data sets for browsing under available metagenomes.
- collections—defined sets of metagenomes grouped for easier analysis. This is the recommended way of working with the analysis page.
- projects—global groups of data sets grouped by the submitting user. The project name will be displayed.
- public data—display of all public data sets.

When using collections or projects, data can also be grouped into one set per collection or project and subsequently compared or added.

Normalization

Normalization refers to a transformation that attempts to reshape an underlying distribution. A large number of biological variables exhibit a log-normal distribution, meaning that when the data are transformed with a log transformation, the values exhibit a normal distribution. Log transformation of the counts data makes a normalized data product that is more likely to satisfy the assumptions of additional downstream tests such as ANOVA or *t*-tests. Standardization is a transformation applied to each distribution in a group of distributions so that all distributions exhibit the same mean and the same standard deviation. This removes some aspects of inter-sample variability and can make data more comparable. This sort of procedure is analogous to commonly practiced scaling procedures but is more robust in that it controls for both scale and location.

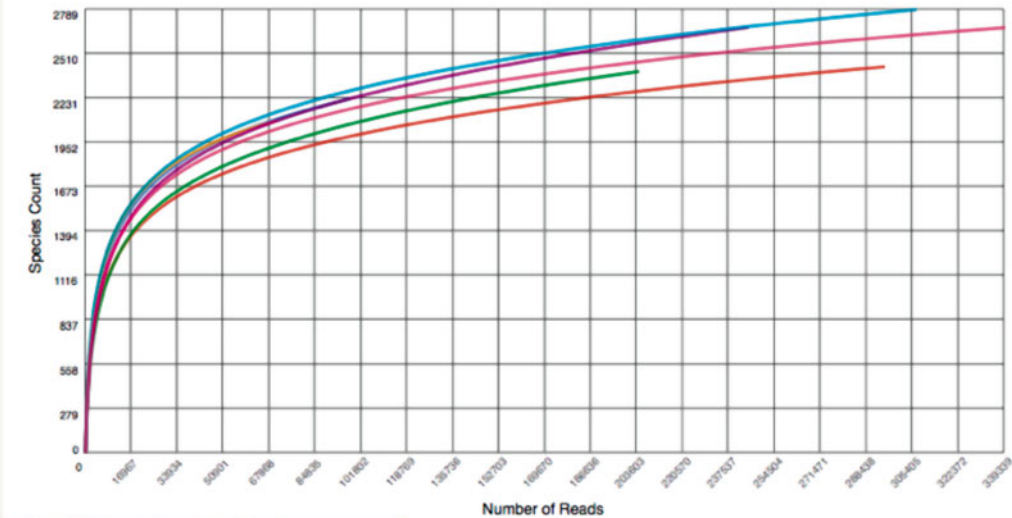
Rarefaction

The rarefaction view is available only for taxonomic data. The rarefaction curve of annotated species richness is a plot (*see* Fig. 8)

This data was calculated for metagenomes 4447970.3, 4447943.3, 4447192.3, 4447103.3, 4447102.3, 4447101.3, 4447971.3 and 4447903.3. The data was compared to MSNR using a maximum e-value of 1e-5, a minimum identity of 80 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.

Metagenome 4447103.3 contains no organism data for the above selected sources and cutoffs. They are being excluded from the analysis.

The image is currently dynamic. To be able to right-click/save the image, please click the static button










rarefaction curve	metagenome	alpha diversity
	4447971.3	356.56
	4447101.3	299.15
	4447970.3	377.11
	4447192.3	217.79
	4447102.3	217.81
	4447903.3	369.26
	4447943.3	222.36

Fig. 8 Rarefaction plot showing a curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled

of the total number of distinct species annotations as a function of the number of sequences sampled. As shown in the figure, multiple data sets can be included.

The slope of the right-hand part of the curve is related to the fraction of sampled species that are rare. When the rarefaction curve is flat, more intensive sampling is likely to yield only a few additional species. The rarefaction curve is derived from the protein taxonomic annotations and is subject to problems stemming from technical artifacts. These artifacts can be similar to the ones affecting amplicon sequencing [31], but the process of inferring species from protein similarities may introduce additional uncertainty.

On the Analysis page, the rarefaction plot serves as a means of comparing species richness between samples in a way independent of the sampling depth. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled, more intensive sampling is likely to yield only a few additional species. Sampling curves generally rise very quickly at first and then level off toward an asymptote as fewer new species are found per unit of individuals collected. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species annotations for sub-samples of the complete data set.

Heatmap/Dendrogram

The heatmap/dendrogram allows an enormous amount of information to be presented in a visual form that is amenable to human interpretation. Dendrograms are trees that indicate similarities between annotation vectors. The MG-RAST heatmap/dendrogram has two dendrograms, one indicating the similarity/dissimilarity among metagenomic samples (x -axis dendrogram) and another indicating the similarity/dissimilarity among annotation categories (e.g., functional roles; the y -axis dendrogram). A distance metric is evaluated between every possible pair of sample abundance profiles. A clustering algorithm (e.g., ward-based clustering) then produces the dendrogram trees. Each square in the heatmap dendrogram represents the abundance level of a single category in a single sample. The values used to generate the heatmap/dendrogram figure can be downloaded as a table by clicking on the download button.

Ordination

MG-RAST uses Principle Coordinate Analysis (PCoA) to reduce the dimensionality of comparisons of multiple samples that consider functional or taxonomic annotations. Dimensionality reduction is a process that allows the complex variation found in a large data sets (e.g., the abundance values of thousands of functional roles or annotated species across dozens of metagenomic samples) to be reduced to a much smaller number of variables that can be visualized as simple two or three-dimensional

scatter plots. The plots enable interpretation of the multidimensional data in a human-friendly presentation. Samples that exhibit similar abundance profiles (taxonomic or functional) group together, whereas those that differ are found farther apart.

A key feature of PCoA-based analyses is that users can compare components not just to each other but to metadata recorded variables (e.g., sample pH, biome, DNA extraction protocol) to reveal correlations between extracted variation and metadata-defined characteristics of the samples.

It is also possible to couple PCoA with higher-resolution statistical methods in order to identify individual sample features (taxa or functions) that drive correlations observed in PCoA visualizations.

This coupling can be accomplished with permutation-based statistics applied directly to the data before calculation of distance measures used to produce PCoAs; alternatively, one can apply conventional statistical approaches (e.g., ANOVA or Kruskal–Wallis test) to groups observed in PCoA-based visualizations.

Bar Charts

The bar chart visualization option on the Analysis page has a built-in ability to drill down by clicking on a specific category. You can expand the categories to show the normalized abundance (adjusted for sample sizes) at various levels. The abundance information displayed can be downloaded into a local spreadsheet. Once a sub-selection has been made (e.g., the domain Bacteria selected), data can be sent to the workbench for detailed analysis. In addition, reads from a specific level can be added into the workbench.

Tree Diagram

The tree diagram allows comparison of data sets against a hierarchy (e.g., Subsystems or the NCBI taxonomy). The hierarchy is displayed as a rooted tree, and the abundance (normalized for data set size or raw) for each data set in the various categories is displayed as a bar chart for each category. By clicking on a category (inside the circle), detailed information can be requested for that node

Table

The table tool creates a spreadsheet-based abundance table that can be searched and restricted by the user. Tables can be generated at user-selected levels of phylogenetic or functional resolution. Table data can be visualized by using Krona [37] or can be exported in BIOM [24] format to be used in other tools (e.g., QIIME). The tables also can be exported as tab-separated text. Abundance tables serve as the basis for all comparative analysis tools in MG-RAST, from PCoA to heatmap/dendrograms.

Workbench

The workbench was designed to allow users to select subsets of the data for comparison or export. Specifically, the workbench supports selecting sequence features and submitting them to further analysis or other analysis. A number of use cases are described below. An important limitation with the current implementation is that data sent to the workbench exist only until the current session is closed.

3.3.8 Metadata, Publishing, and Sharing

MG-RAST is both an analytical platform and a data integration system. To enable data reuse, for example for meta-analyses, we require that all data being made available to third parties contain at least minimal metadata. The MG-RAST team has decided to follow the minimal checklist approach used by the GSC.

MG-RAST provides a mechanism to make data and analyses publicly accessible. Only the submitting user can make data public on MG-RAST. As stated above, metadata is mandatory for data set publication. Metazen [39] is a web based tool for assisting end-users in the creation of metadata with the correct controlled vocabularies and in the correct format.

In addition to publishing, data and analysis can also be shared with specific users (Fig. 9). To share data, users simply enter their email address via clicking sharing on the Overview page.

4 matR, Metagenomic analysis tools for R

We have recently produced a package for the R environment for statistical computing (www.r-project.org/) that provides accessory analytical capabilities to complement those already available through the MG-RAST website. The matR package is primarily designed for download and analysis of MG-RAST-based annotation abundance profiles. It makes it possible to download annotation abundance data from MG-RAST into R friendly data objects suitable for analysis with included analysis functions. We note that matR has been specifically designed to perform large-scale analyses on abundance profiles from dozens to thousands of data sets with suitable pre-processing, normalization, statistics, and visualization

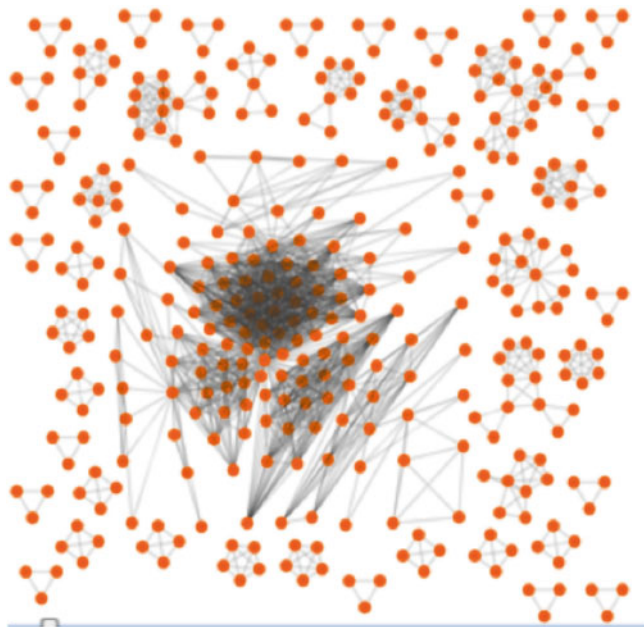


Fig. 9 Data sets shared in MG-RAST by users (*orange dots*), shown as connecting edges

tools. Users can utilize these built-in tools or any of the enormous variety of tools available within the R universe. The release version of matR is available through CRAN (<http://cran.r-project.org/web/packages/matR/>); pre-release and development versions are available on github (<https://github.com/MG-RAST/matR/>); a google group is available (<https://groups.google.com/forum/#!forum/matr-forum>); a publication demonstrating the ease with which matR can be used to conduct large-scale analyses is forthcoming.

4 Notes

Typical analysis parameters

MG-RAST utilizes a number of tools and analyses to generate annotation abundance data, and subsequent visualizations of abundance data, from raw sequence data. Users have the option to vary several of the parameters that define several aspects of how MG-RAST performs. While the default settings have been selected to perform *well* in most circumstances, it is not possible to find a single collection of analysis parameter values that will perform optimally on all data submitted to MG-RAST. Here we briefly mention some of the most important parameters, discussion when a user may want to alter the default values, and how they can go about selecting the best values for their analyses in a methodological fashion. We divide these parameters into two sections—pipeline options that users must be specific before their data are processed through MG-RAST and data options that define the annotation abundance data that are returned to the user,

Pipeline options

Pipeline options are specified by the user during upload and prior to annotations with MG-RAST. These options are used to filter data with a number of metrics that characterize the quality of each individual read in a sample. Users can modify these key parameters from their default values,

Assembled (NOT checked by default)

Select this option if your data are the product of any assembly-based tools applied to the data before upload to MG-RAST. *Note that we recommend upload of raw reads with their accompanying quality data (i.e., fastq files).* This allows MG-RAST to directly sequence QC information when processing reads. When assembled data are used, a great deal of abundance information is lost; with assembled data MG-RAST can only provide abundance that indicates the number of times a feature is observed in the assembly. Relative abundance information contained in the original reads is lost. In addition, assembly might introduce chimeric artifacts that will con-

found subsequent annotation and analysis. Users should select this option if their data have undergone any assembly prior to upload.

Dereplication (CHECKED by default)

A well-known artifact in NGS sequencing is the production of artificial replicate sequences. These are identical (or nearly identical) reads that occur with extremely high abundances (see <http://www.nature.com/ismej/journal/v3/n11/full/ismej200972a.html>). While the exact causes for such sequences are not well understood, it has been posited that they are due in part to inclusion of low complexity and/or adapter sequences (sequences ligated onto reads to facilitate processing with NGS that SHOULD NEVER appear in output). Artificial duplicates are utilized by MG-RAST as the basis for DRISSE-based error estimates (<http://www.ncbi.nlm.nih.gov/pubmed/22685393>); we maintain that the inclusion of such sequences constitutes a clear sequencing error/artifact, but this is a contested notion that remains to be resolved (<http://www.ncbi.nlm.nih.gov/pubmed/23698723>). However, what metagenomicists can agree on is that such sequences appear frequently, when they are not expected, and are present even after raw sequence data have been treated with vendor-specific tools to remove them. Dereplication should always be turned on for WGS sequencing—but, users may want to deselect this option for amplicon-based data (if reads start with a high conserved region, they could be misinterpreted as artificial replicates) or any other data sets where an extremely high level of replication among reads is expected.

Screening (set to *H. sapiens*, NCBI v36 by default)

It is common for metagenomic NGS data to contain contaminant sequences from an undesired organism (e.g., human sequences in a human gut sampled microbiome). Screening makes it possible for these sequences to be identified and removed from subsequent analysis. Users have a number of other organisms that can be used to screen the data. Currently, MG-RAST supports filtering against a single contaminant organism. Users can select the most appropriate organism, or select “none.” If a user wants to maintain all sequence data, they should select “none.” Note that MG-RAST is not designed to annotate eukaryotic data, screening is used as a means to remove such data (assumed to be host sequence data) from a number of known sources. Note that the currently collection of organisms against which screening is possible can be expanded. Users should contact MG-RAST if they want add additional organisms to those that can be screened.

Dynamic Trimming (CHECKED by default with 15 as the minimum retained phred score and 5 as the maximum number of bases allowed in each read that can contain a base lower than the phred minimum)

Dynamic trimming is only possible if users have uploaded data in fastq format that contains sequence quality information. Dynamic trimming is used in place of length filtering and ambiguous-based filtering when a fastq is uploaded. More stringent values (e.g., higher phred, and lower allowance for base that do not meet the phred threshold) will reduce the length and amount of reads that MG-RAST processes, but can increase their overall quality. Users should increase stringency if they want to reduce annotated data to more constrained, higher quality sequences. We do not recommend reducing stringency; this will lead to the inclusion of low-quality data that will most likely produce no or extremely unreliable annotation.

Length filtering (CHECKED by default, with a standard deviation multiplier set to 2; only applies to fasta data; it is not used on fastq data)

Only applied to fasta data, or fastq data with no quality information (essentially, a surrogate for the dynamic trimming applied to fastq data that include quality information), length trimming calculates the average sequence length for all reads in a data set and removes those that are,

Longer than $\text{sample_mean} + (\text{standard_deviation_multiplier} * \text{standard_deviation})$

or shorter than $\text{sample_mean} - (\text{standard_deviation_multiplier} * \text{standard_deviation})$

This is an attempt to remove sequences that exhibit outlier lengths and are likely to be sequencing artifacts.

For fasta data we recommend that users always use length filtering, and that they do not use a standard deviation multiplier less than the default of 2. Users may want to increase the standard deviation multiplier if their reads exhibit a large degree of variation with respect to read length.

Ambiguous base filtering (CHECKED by default with a maximum number of allowed ambiguous bases per read set to 5; only applies to fasta data or fastq without quality information)

Sequences frequently produce “ambiguous bases” (bases that are not A, T, C, or G), these represent bases for which the sequences could not make a definitive call for the identity of the base. Ambiguous bases are expected at the end of sequenced reads, and are generally considered to be an indication of low quality if they are found anywhere else in the read, particularly at the start. MG-RAST will reject any read that contains more than the specified number of ambiguous bases. We recommend that users should not use values less stringent (i.e., larger) than the default of 5. Users are free to specify more stringent criteria (smaller number of allowed ambiguous bases)—this will reduce the number of anno-

tated reads, but will produce a set of reads that have a higher overall quality (less likely to contain artifacts that could lead to erroneous annotations).

Data options

After processing through MG-RAST, users have a number of options that they can use to filter annotation abundance data (accessed via the). Chief among these are the parameters described briefly below,

Annotation source (default m5NR)

MG-RAST provides users with the unique ability to provide annotations for analyzed data from multiple source databases. By default, the m5NR is used—but users are free to use any one of the numerous additional annotation sources. As the m5NR represents a non-redundant union of all annotation databases contained within MG-RAST, we generally recommend its use over any of the individual databases.

Max e -value cutoff (default $1e-5$)

The max e -value cutoff indicates the largest (least stringent) e -value for an annotation to be included in the output annotations. We generally recommend that users not use larger (less stringent) e -value cutoff. The use of smaller (more stringent) e -values will ensure that annotations exhibit higher statistical fidelity; however, this will come at the cost of a smaller overall number of annotated features. We suggest that users experiment with multiple e -values until they arrive at one that produces enough annotations to address the hypothesis(es) in question while minimizing the number of spurious (false positive) annotations.

Min % identity cutoff (default 60 %)

The minimum percent identity represents the lower bound threshold for annotations to be returned to the user. Matches between the query and selected annotation that match or exceed this value are retained, all others are rejected. We recommend that users not select a value any lower than the default. Users may choose larger values to return annotations only if they meet more stringent match criteria. An increase in the minimum percent identity cutoff will produce annotations that exhibit closer matches to the reference database at the cost of a lower overall number of annotations. Once again, we encourage users to experiment with this threshold until it produce the desired number of annotations at an acceptable level of stringency.

Min alignment length cutoff (default 15)

The minimum length cutoff represent the lower bound threshold for alignment lengths to be included in output annotations. As with the e -value and percent identity cutoffs, we recommend that

users do not attempt to select a less stringent (smaller) value. They can select larger (more stringent) values to produce a smaller set of longer matches (generally considered to be synonymous with higher quality and increased accurate). Once again, users need to experiment with this value to find an optimal balance between stringency and number of annotations.

Additional Documentation

This chapter is intended as an introduction to MG-RAST, and necessarily treats topics as concisely as possible; other topics were omitted entirely for the sake of brevity. Users interested in a much more detailed description of MG-RAST are referred to the MG-RAST user manual. The manual can be downloaded from metagenomics.anl.gov/; simply click on the link for the *MG-RAST manual*.

References

1. Wilkening J, Wilke A, Desai N et al (2009) Using clouds for metagenomics. A case study. In: IEEE Cluster, 2009
2. Angiuoli S, Matalaka M, Gussman A et al (2011) Clovr, a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12:356
3. Meyer F, Paarmann D, D'Souza M et al (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386
4. Field D, Amaral-Zettler L, Cochrane G et al (2011) The genomic standards consortium. *PLoS Biol* 9:e1001088
5. Wilke A, Harrison T, Wilkening J et al (2012) The m5nr, a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 13:141
6. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
7. Kent WJ (2002) Blat—the blast-like alignment tool. *Genome Res* 12:656–664
8. Brooksbank C, Bergman MT, Apweiler R et al (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res* 42 (Database issue):D18–D25
9. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* 5:e1000431
10. Markowitz VM, Ivanova NN, Szeto E et al (2008) IMG/M, a data management and analysis system for metagenomes. *Nucleic Acids Res* 36(Database issue):D534–D538
11. Kanehisa M (2002) The KEGG database. *Novartis Found Symp* 247:91–101
12. Benson DA, Cavanaugh M, Clark K (2013) Genbank. *Nucleic Acids Res* 41(Database issue):D36–D42
13. Dwivedi B, Schmieder R, Goldsmith DB et al (2012) PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. *BMC Bioinformatics* 13:37
14. Overbeek R, Begley T, Butler RM et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702
15. Magrane M, Uniprot Consortium (2011) UniProt knowledgebase: a hub of integrated protein data. *Database (Oxford)*. doi:10.1093/database/bar009
16. Snyder EE, Kampanya N, Lu J et al (2007) PATRIC: the VBI PathoSystems resource integration center. *Nucleic Acids Res* 35(Database issue):D401–D406
17. Jensen LJ, Julien P, Kuhn M et al (2008) EggNog: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36(Database issue):D250–D254
18. Tang W, Wilkening J, Desai N, Gerlach W, Wilke A, Meyer F (2013) A scalable data analysis platform for metagenomics. *Proceedings of the 2013 International Conference on Big Data*
19. Bischof, J., Wilke, A., Gerlach, W., Harrison, T., Paczian, T., Tang, W., Trimble, W., Wilkening, J., Desai, N. and Meyer, F. (2015), Shock: Active Storage for Multicloud Streaming

- Data Analysis, 2nd IEEE/ACM International Symposium on Big Data Computing, Limassol, Cyprus, 2015
20. Cox MP, Peterson DA, Biggs PJ (2010) Solexaqa: at-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinformatics* 11:485
 21. Huse SM, Huber JA, Morrison HG et al (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8:R143
 22. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3:1314–1317
 23. Keegan KP, Trimble WL, Wilkening J et al (2012) A platform-independent method for detecting errors in metagenomic sequencing data, Drisee. *PLoS Comput Biol* 8:e1002541
 24. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
 25. Trimble WL, Keegan KP, D'Souza M et al (2012) Short-read reading-frame predictors are not created equal, sequence error causes loss of signal. *BMC Bioinformatics* 13:183
 26. Rho M, Tang H, Ye Y (2009) Fraggenscan, Predicting genes in short and error prone reads. *Nucleic Acids Res* 38:e191
 27. Edgar RC (2010) Search and clustering orders of magnitude faster than blast. *Bioinformatics* 26:2460–2461
 28. Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336
 29. Huson DH, Auch AF, Qi J et al (2007) Megan analysis of metagenomic data. *Genome Res* 17:377–386
 30. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi:10.1186/1471-2164-9-75
 31. Pruesse E, Quast C, Knittel K et al (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196
 32. DeSantis TZ, Hugenholtz P, Larsen N et al (2006) Greengenes: a Chimera-Checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072
 33. Cole JR, Chai B, Marsh TL et al (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res* 31:442–443
 34. Yilmaz P, Kottmann R, Field D et al (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 29:415–420
 35. Bolotin A, Quinquis B, Sorokin A et al (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151:2551–2561
 36. Reeder J, Knight R (2009) The 'rare biosphere', a reality check. *Nat Methods* 6:636–637
 37. Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 12:385
 38. Gerlach, W., Tang, W., Keegan, K., Harrison, T., Wilke, A., Bischof, J., D'Souza, M., Devold, S., Murphy-Olson, D., and Desai, N. (2014) Skyport – Container-based execution environment management for multi-cloud scientific workflows. In Proc. 5th Int'l Workshop on Data-Intensive Computing in the Clouds. IEEE Press, pp. 25–32

Chapter 14

Analysis of Active Methylo-trophic Communities: When DNA-SIP Meets High-Throughput Technologies

Martin Taubert, Carolina Grob, Alexandra M. Howat, Oliver J. Burns, Yin Chen, Josh D. Neufeld, and J. Colin Murrell

Abstract

Methylo-trophs are microorganisms ubiquitous in the environment that can metabolize one-carbon (C1) compounds as carbon and/or energy sources. The activity of these prokaryotes impacts biogeochemical cycles within their respective habitats and can determine whether these habitats act as sources or sinks of C1 compounds. Due to the high importance of C1 compounds, not only in biogeochemical cycles, but also for climatic processes, it is vital to understand the contributions of these microorganisms to carbon cycling in different environments. One of the most challenging questions when investigating methylo-trophs, but also in environmental microbiology in general, is which species contribute to the environmental processes of interest, or “who does what, where and when?” Metabolic labeling with C1 compounds substituted with ^{13}C , a technique called stable isotope probing, is a key method to trace carbon fluxes within methylo-trophic communities. The incorporation of ^{13}C into the biomass of active methylo-trophs leads to an increase in the molecular mass of their biomolecules. For DNA-based stable isotope probing (DNA-SIP), labeled and unlabeled DNA is separated by isopycnic ultracentrifugation. The ability to specifically analyze DNA of active methylo-trophs from a complex background community by high-throughput sequencing techniques, i.e. targeted metagenomics, is the hallmark strength of DNA-SIP for elucidating ecosystem functioning, and a protocol is detailed in this chapter.

Key words Carbon-13, DNA stable isotope probing, DNA-SIP, High-throughput sequencing, Isotopic labeling, Methylo-trophy, Metagenomics, One-carbon compounds

1 Introduction

One carbon (C1) compounds, as well as compounds with multiple carbons but no carbon–carbon bonds, such as methylated amines, are diverse and widespread in the environment. These compounds play key roles in the biogeochemical cycles of carbon, and also nitrogen, sulfur, and phosphorus [1–3]. Some of these compounds have an influence on climatic processes through their release to the atmosphere [1], and thus a direct relevance for global ecology. Microorganisms that can metabolize these compounds, called

methylotrophs, are ubiquitous in the environment. Next to physicochemical reactions, microbial activities often are the only major processes involved in C1 compound conversion [3–5]. Thus, the composition and activity of the microbial community in a specific habitat is a major factor that modulates the release or uptake of C1 compounds into and from the atmosphere. Consequently, investigation of these microorganisms in different habitats, as well as assessment of their activity and contribution to biogeochemical cycles, is essential for understanding and modeling the environmental processes that shape and sustain our planet.

Most knowledge of C1 compound metabolism was obtained from isolation and characterization of pure cultures of methylotrophs [6]. However, insights deduced by these cultivation-dependent approaches are difficult to transfer directly to environmental systems, where microorganisms are tightly integrated in metabolic networks and potentially dissimilar to those readily cultivated microorganisms. Actual microbial communities that catalyze processes of interest often remain “black boxes” for the environmental microbiologist, making it difficult to answer the key question of “who is doing what, where and when?” [7] in a particular environment.

Classical approaches for environmental studies of methylotrophs rely on the analysis of specific biomarkers. The detection of 16S rRNA genes similar to those of known and characterized methylotrophs in environmental samples is often used to infer a corresponding function to these detected organisms. In addition, structural genes can be used to identify environmental distribution of key enzymes for the conversion of C1 compounds, including a range of dehydrogenases, monooxygenases, and methyl transferases [8]. Various PCR primer sets have been introduced to target these genes in environmental surveys [9–15]. For example, *pmoA* and *mmoX*, encoding subunits of the particulate and soluble methane monooxygenase, have been used to target methanotrophs, and *mxnF*, encoding the large subunit of methanol dehydrogenase, to target methylotrophs [9, 10, 13]. High-throughput sequencing technologies have improved rapidly over the past decade, allowing much deeper sequencing of environmental samples. Pyrosequencing, reversible dye terminator sequencing, or ion semiconductor sequencing [16] are often used in combination with biomarker approaches. Selection for biomarkers of interest can either be done prior to sequencing, for example by using PCR amplicon pyrosequencing [17, 18], or by screening of shotgun metagenomic datasets [19, 20]. However, these approaches do not provide information on the real metabolic activities of the microbial communities being investigated.

In order to unravel the functional contributions of methylotrophs in microbial communities, cultivation-independent approaches are needed that can establish a direct link between phylogeny and function. Stable isotope probing (SIP), a metabolic labeling approach with

substrates enriched with heavy, nonradioactive isotopes, can fulfill these requirements. In a SIP experiment targeting methylootrophs, environmental material (e.g. water, soil or sediment) is incubated with a ^{13}C -labeled C1 compound. Active methylootrophs that use this compound as a carbon source incorporate the heavy carbon atoms into their biomass, including and notably their DNA. Detection of ^{13}C enrichment in biomarkers of specific organisms is therefore evidence for methylootrophic activity resulting in substrate assimilation. This approach was first described in combination with the investigation of microbial polar lipid derived fatty acids (PLFA), using isotope ratio mass spectrometry to detect the heavy isotopes [21]. The combination with metagenomics (DNA-SIP) followed 2 years later, and allowed the implementation of SIP with the classical approaches described above to detect active methylootrophs in the environment [22–26]. Compared to a PLFA-based approach, DNA-SIP offers better phylogenetic resolution and provides substantial functional information from the labeled DNA sequences (*see e.g.* [14, 15, 27]). Even the retrieval of whole genomes of the active methylootrophs is possible [28].

A SIP experiment employing ^{13}C -labeled C1 compounds, followed by DNA extraction, typically results in a mix of heavy (^{13}C -labeled) DNA from active methylootrophs and unlabeled light (^{12}C) DNA from other organisms, including inactive methylootrophs. In this chapter, we outline requirements for a DNA-SIP experiment, describe the methods necessary for isolation and identification of the labeled DNA and give advice for troubleshooting and interpretation of subsequent results. In addition, we highlight strategies for the analysis of metagenomics DNA by high-throughput sequencing.

Separation of heavy and light DNA is achieved in a density gradient because the substitution of ^{12}C with ^{13}C proportionally increases the density of DNA. Ultracentrifugation of the extracted DNA mix in a cesium chloride solution results in the migration of DNA according to its density within the gradient, forming bands of increasingly labeled DNA down the gradient. The density gradient is partitioned into a number of fractions, and the 16S rRNA gene profiles of the DNA recovered from each of these fractions are investigated via denaturing gradient gel electrophoresis (DGGE) [29]. This fingerprinting technique represents a straightforward method that separates PCR amplicons based on their GC content and sequence [14, 15].

The rRNA gene fingerprints are important to rapidly identify the fractions containing ^{13}C -labeled DNA by comparison with corresponding fraction profiles from an unlabeled (^{12}C) control incubation. These fractions containing DNA enriched with genetic material of the active methylootrophs can subsequently be used for sequence analysis, starting with amplicon sequencing targeting 16S rRNA genes and functional genes (e.g. *pmoA*, *mxnF*), to obtain phylogenetic and functional information. If necessary,

labeled DNA can be amplified by multiple displacement amplification (MDA) to obtain sufficient material prior to shotgun metagenomics [30], enabling a more in-depth functional investigation of active methylotrophs, including the potential for genome assembly even with very low quantities of labeled material.

2 Materials

Use analytical grade reagents and ultrapure water for the preparation of all solutions. For suspending DNA, use nuclease-free water. All solutions should be prepared and stored at room temperature, unless otherwise indicated.

2.1 Density Gradient Centrifugation Components

1. EDTA solution: 0.5 M EDTA, pH 8.0. Dissolve 186.1 g of disodium ethylenediamine tetraacetate dihydrate (EDTA) in 900 mL of water. Add 2 M NaOH to adjust the pH to 8.0 (*see Note 1*) and make up to 1 L with water. Sterilize in an autoclave.
2. Tris-EDTA (TE) buffer: 10 mM Tris-HCl, 1 mM EDTA, pH 8.0. Dissolve 60.6 mg of Tris in 40 mL of water. Add 100 μ L of a 0.5 M EDTA solution, mix and adjust pH to 8.0 with 0.5 M HCl. Make up to 50 mL with water. Filter sterilize (0.22 μ m) or autoclave.
3. DNA from a metabolic labeling experiment using the 13 C-labeled C1 compound of interest and DNA from a control treatment with the same 12 C compound, in TE buffer or water (*see Note 2*), with known DNA concentrations.
4. CsCl solution: Dissolve 603.0 g of CsCl in water to a final volume of 500 mL, resulting in a 7.163 M CsCl solution (*see Note 3*). Adjust the density to a final value between 1.88 and 1.89 g/mL at 20 °C (*see Note 4*).
5. Gradient buffer (GB): 0.1 M Tris-HCl, 0.1 M KCl, 1 mM EDTA, pH 8.0. Dissolve 12.11 g of Tris and 7.46 g of KCl in 900 mL of water. Add 2 mL of a 0.5 M EDTA solution, mix and adjust pH to 8.0 with HCl. Make up to 1 L with water. Sterilize in an autoclave.
6. Ultracentrifuge tubes: 5.1 mL, 13 mm \times 51 mm Polyallomer Quick-Seal Centrifuge Tubes (Beckman Coulter Ltd., High Wycombe, UK).
7. Ultracentrifuge rotor capable of withstanding 177,087 $\times g$ average: e.g. VTi 65.2 Beckman Coulter Vertical (Beckman Coulter Ltd., High Wycombe, UK).
8. Pump for fractionation: Syringe pump or peristaltic pump able to deliver a constant flow of 425 μ L/min.
9. Digital refractometer: e.g. AR200 Digital Handheld Refractometer (Reichert Technologies, Depew, NY, USA).

10. APS solution: 10 % ammonium persulfate (w/v). Dissolve 1 g of ammonium persulfate (APS) in 10 mL of water. Aliquot in 1 mL portions and store at -20°C . Frozen APS solution can be used for several months.
11. Linear Polyacrylamide (LPA): Mix (in order) 250 mg of acrylamide, 4.25 mL of water, 200 μL of 1 M Tris-HCl pH 8.0, 33 μL of 3 M sodium acetate pH 7.5, 10 μL of 0.5 M EDTA solution, 50 μL of 10 % ammonium persulfate solution and 5 μL of tetramethylethylenediamine (TEMED) in a 50-mL tube, leave at room temperature for 30 min. Add 12.5 mL of 95 % ethanol to precipitate for 5 min. Remove liquid (squeeze pellet), wash with 70 % ethanol and remove liquid again. Air dry for 10 min. Suspend pellet overnight in 50 mL of water, aliquot and store at -20°C .
12. Polyethylene glycol-NaCl (PEG-NaCl) solution: 30 % PEG 6000, 1.6 M NaCl. Dissolve 150 g of PEG 6000 and 46.8 g of NaCl into a final volume of 500 mL. Sterilize in an autoclave. Two phases may form after autoclaving or prolonged storage, so mix well before each use.

2.2 DGGE Components

1. PCR primers for DGGE: Primer set 341f_GC (CGCCCCG CCGC GCGCGGCGGG CGGGGCGGGG GCACGGGG GG CCTACGGGAG GCAGCAG) and 518r (ATTACCGCGG CTGCTGG) targeting bacterial 16S rRNA genes [31].
2. 50 \times Tris-Acetate-EDTA (TAE) buffer: 2 M Tris-HCl, 1 M Acetic acid, 0.05 M EDTA. Dissolve 242 g of Tris in 800 mL of water. Add 57.1 mL of 100 % acetic acid and 100 mL of 0.5 M EDTA solution. Make up to 1 L with water.
3. 30 % DGGE solution: 1 \times TAE, 10 % acrylamide/bis-acrylamide, 12 % formamide (v/v), 12.6 % urea (w/v). Dissolve 6.3 g of urea in 10 mL of water. Add 6 mL of formamide, 1 mL of 50 \times TAE buffer and 12.5 mL of 40 % acrylamide/bis (37.5:1). Make up to 50 mL with water while the remaining urea dissolves.
4. 70 % DGGE solution: 1 \times TAE, 10 % acrylamide/bis-acrylamide, 28 % formamide (v/v), 29.4 % urea (w/v). Dissolve 14.7 g of urea in 10 mL of water. Add 14 mL of formamide, 1 mL of 50 \times TAE buffer and 12.5 mL of 40 % acrylamide/bis (37.5:1). Make up to 50 mL with water while the remaining urea dissolves. 5 mg of bromophenol blue can be added for visual differentiation from the 30 % DGGE solution.
5. 5 \times DGGE loading dye: 50 % glycerol (v/v), 0.2 M EDTA, 0.05 % bromophenol blue (w/v). Mix 2.5 mL of glycerol, 2 mL of 0.5 M EDTA solution and 2.5 mg of bromophenol blue. Make up to 5 mL with water.
6. DGGE system: e.g. DCode Universal Mutation Detection System (Bio Rad, Hemel Hempstead, UK) or DGGEK-2001-110 (C.B.S. Scientific, San Diego, CA, USA).

Table 1
PCR primer sets for functional genes involved in methylotrophy

Gene	Name	Sequence	Reference
<i>pmoA/amoA</i>	pmoA189F	GGNGACTGGGACTTCTGG	Holmes et al. [9]
<i>pmoA/amoA</i>	pmoA682R	GAASGCNGAGAAGAASGC	
<i>pmoA</i>	mb661R ^a	CCGGMGCAACGTCYTTACC	Costello and Lidstrom [11]
<i>msxA</i>	1003F	GCGGCACCAACTGGGGCTGGT	McDonald and Murrell [10]
<i>msxA</i>	1561R	GGGCAGCATGAAGGGCTCCC	
<i>msxA</i>	1555R	CATGAABGGCTCCCARTCCAT	Neufeld et al. [14]
<i>mmoX</i>	206F	ATCGCBAARGAATAYGCSCG	Hutchens et al. [13]
<i>mmoX</i>	886R	ACCCANGGCTCGACYTTGAA	
<i>mmoX</i>	mmoX166F	ACCAAGGARCARTTCAAG	Auman et al. [12]
<i>mmoX</i>	mmoX1401R	TGGCACTCRTARCGCTC	
<i>mauA</i>	mauAf1	ARKCYTGYGABTAYTGGCG	Neufeld et al. [14]
<i>mauA</i>	mauAr1	GARAYVGTGCARTGRTARGTC	
<i>gmaS</i>	557F	GARGAYGCSAACGGYCAGTT	Wischer et al. [15]
<i>gmaS</i>	1332R ^b	GTAMTCSAYCCAYTCCATG	
<i>gmaS</i>	970R ^c	TGGGTSCGRTRTRTGCCSG	

^aFor nested PCR^bBeta- and Gammaproteobacteria^cAlphaproteobacteria

2.3 DNA Amplification Components and Bioinformatics Tools

1. PCR primer sets targeting functional genes (Table 1).
2. Multiple displacement amplification (MDA) kit: e.g. REPLI-g Mini Kit (QIAGEN Ltd., Manchester, UK).
3. Software package mothur: www.mothur.org [32].
4. Software package USEARCH: www.drive5.com/usearch/ [33].

3 Methods

Carry out all procedures at room temperature unless otherwise specified.

3.1 Metabolic Labeling with ¹³C-Labeled C1 Compounds

Setup conditions for metabolic labeling experiments are complex and depend on many factors, including the composition of the microbial community, type of heavy isotope substrate used, metabolic activity of the target population, conversion efficiency and biochemical processes of interest. Thus, no comprehensive protocol can be given for this part of the experiment (*see Note 5*).

The following section gives a basic guideline highlighting key steps and crucial points of a metabolic labeling experiment.

1. Obtain environmental material containing the microbial community of interest, e.g. soil, sediment, sludge, biofilm, or aquatic material. Ensure enough material to obtain sufficient DNA after incubation: 5 μg of genomic DNA are required. Furthermore, process the environmental material as soon as possible after sampling. Excessive transport or storage times might influence the microbial community and bias the experimental outcome.
2. Mix the environmental sample to avoid experimental inconsistencies due to sample heterogeneity. Split into individual batches (e.g. bottles, microcosms) for incubation. Prepare all incubations in duplicates or triplicates. In addition to incubations with the ^{13}C -labeled CI compound, incubations with the corresponding ^{12}C compound are also required. This is critical to identify ^{13}C -labeled DNA later on. Also prepare controls without substrate and sterile controls as necessary.
3. Select the incubation time(s) for your experiment. This depends largely on the metabolic activity of the microbial community of interest. Based on that, an incubation time that is too short will result in insufficient labeling; an incubation time that is too long results in unspecific labeling (i.e. cross-feeding). A preliminary experiment to assess the microbial activity can be useful. Furthermore, performing a time series experiment can give additional information about the carbon flux through the microbial community.
4. Choose the substrate concentration and incubation conditions. The concentration of the added CI compound should be as close as possible to the concentration present in the environment. Too low a substrate concentration can result in insufficient labeling. Aim for incorporation of 5–500 μmol of ^{13}C per gram of soil or sediment and 1–100 μmol of ^{13}C per liter of water. Incubation conditions (i.e. temperature, light level, nutrient and oxygen concentration) should be as close to natural conditions as possible to reduce biases on the active microbial community detected [34].
5. Monitor substrate consumption. This will allow quantification of incorporation and facilitate selection of the most suitable sampling times. If no reliable method for determination of substrate concentrations is available, consider monitoring $^{13}\text{CO}_2$ production or enrichment in biomass (e.g. using isotope ratio mass spectrometry) to have a proxy for microbial activity.

3.2 Preparation and Setup

1. Prepare a calibration curve for calculation of the density of mixtures of the CsCl solution and GB from refractive indices. Mix 450 μL of CsCl solution with 0, 10, 20, 35, 50, 65, 80,

100, 120, and 140 μL GB. Measure the density of the mixtures (*see Note 4*). Measure refractive indices with a digital refractometer with a resolution of at least 0.0001 and temperature correction (nD-TC) to 20 °C. Plot density versus nD-TC and calculate a linear regression. The calibration curve is required to convert nD-TC readings to density to set up samples of the correct density for density gradient centrifugation and to verify gradient formation afterward (*see Note 6*).

2. Calculate the required amount of GB to get to the desired starting density for density gradient centrifugation of 1.725 g/mL. This can be done using the formula:

$$\text{Required volume} = (\text{CsCl stock density} - 1.725 \text{ g/mL}) \\ \times \text{volume of CsCl stock added} \times 1.52 \text{ mL/g} \text{ (see Note 7)}.$$

3. Based on the DNA concentrations of each sample, calculate the volume required to obtain 5 μg of DNA per sample. The amount of GB for each sample needs to be corrected by this volume.
4. Prepare a 15 mL tube for each DNA sample with 4.8 mL of CsCl stock solution. Add 5 μg of DNA for each sample. Add the calculated volume of GB that is reduced by the volume of DNA solution you added for each sample. Calculate the targeted refractive index based on the calibration curve prepared in step 1 for a desired final density of 1.725 g/mL. This typically will be around an nD-TC of 1.4040, but can vary slightly for different stock solutions. Add small amounts of GB and CsCl stock solution to reach the desired refractive index, mix well after each addition (*see Note 8*). Samples should be within ± 0.0002 of the targeted refractive index.
5. Fill ultracentrifuge tubes with the prepared CsCl/GB/DNA mixtures. Use disposable Pasteur pipettes for convenience. To remove air bubbles that stick to the tube walls, fill the tubes up to 1 cm below the top, then gently tilt and rotate the tube, allowing the remaining air to run over the tube walls to gather any smaller air bubbles. Carefully top up the tubes to the tube stem.
6. Balance pairs of ultracentrifuge tubes using an analytical balance. Weight differences below 2 mg are essential for each pair. Heat seal the tubes according to the manufacturer's instruction. Squeeze tubes firmly to make sure that they are properly sealed. Reweigh the paired tubes to ensure that they remain balanced (*see Note 9*).
7. Load tubes into the ultracentrifuge rotor, taking care to position balanced tube pairs opposite each other. Note sample

names and rotor positions; tube labels can come off during ultracentrifugation. Prepare the rotor according to the manufacturer's instructions.

3.3 Ultra centrifugation and Fractionation

1. Ultracentrifugation should be carried out for at least 40 h to ensure proper gradient formation and focused migration of DNA to the corresponding densities. Extended run times of 60–72 h, i.e. over weekends, can also be used. Set the centrifuge to a speed according to $177,087 \times g$ average (e.g., 44,100 rpm for the VTi 65.2 Beckman Coulter Vertical rotor; *see Note 10*) and a temperature of 20 °C. Note that temperature influences density directly. Set the centrifuge to maximum acceleration and select the “no brake” option for deceleration. Calculate between 1.5 and 2 h of additional run time until the centrifuge has stopped. Follow the manufacturer's instructions when operating the ultracentrifuge.
2. Collect all tubes carefully from the rotor, keeping each tube vertical at all times. A pump with adjustable speed and uniform flow rate is needed to fractionate each SIP gradient. The required flow rate is 425 $\mu\text{L}/\text{min}$. A syringe pump should be used for best results, or a peristaltic pump instead. Adjust the speed of the pump by running it with water for 10 min and measuring the volume of the flow-through to get to the desired flow rate. Make sure that the tubing connected to the pump is fitted with a male Luer fitting. Before fractionating the first tube, rinse and fill tubing with water (*see Note 11*).
3. After ultracentrifugation, fit the ultracentrifuge tube in a stand with a suitable clamp for fractionation. Handle the tube carefully to prevent disturbing the density gradient. The clamp should be only tight enough to hold the tube securely, without squeezing it. Connect a 23-gauge (0.6 \times 25 mm) needle to the tubing of the pump. Run the pump momentarily to remove all air from the needle. Carefully pierce the top of the ultracentrifuge tube with the needle, adjacent to the tube stem (*see Note 12*). Ensure that the needle and tubing are secured and cannot slip away during fractionation. Use a second needle to pierce the tube at the bottom, then remove this needle again (*see Note 13*).
4. Prepare a series of 12 tubes (1.5 mL) to capture all sample fractions. Activate the prepared pump to fill the ultracentrifuge tube with water, replacing the CsCl solution, together with a timer (*see Fig. 1*). Collect the CsCl solution at the bottom of the tube in the prepared tubes. Collect 425 μL per fraction (i.e. 1 min per fraction). An automated fraction collector might be used, but is not necessary. A full ultracentrifuge tube will result in 12 fractions with a flow rate of 425 $\mu\text{L}/\text{min}$, but keep additional 1.5 mL tubes ready for a potential 13th fraction.

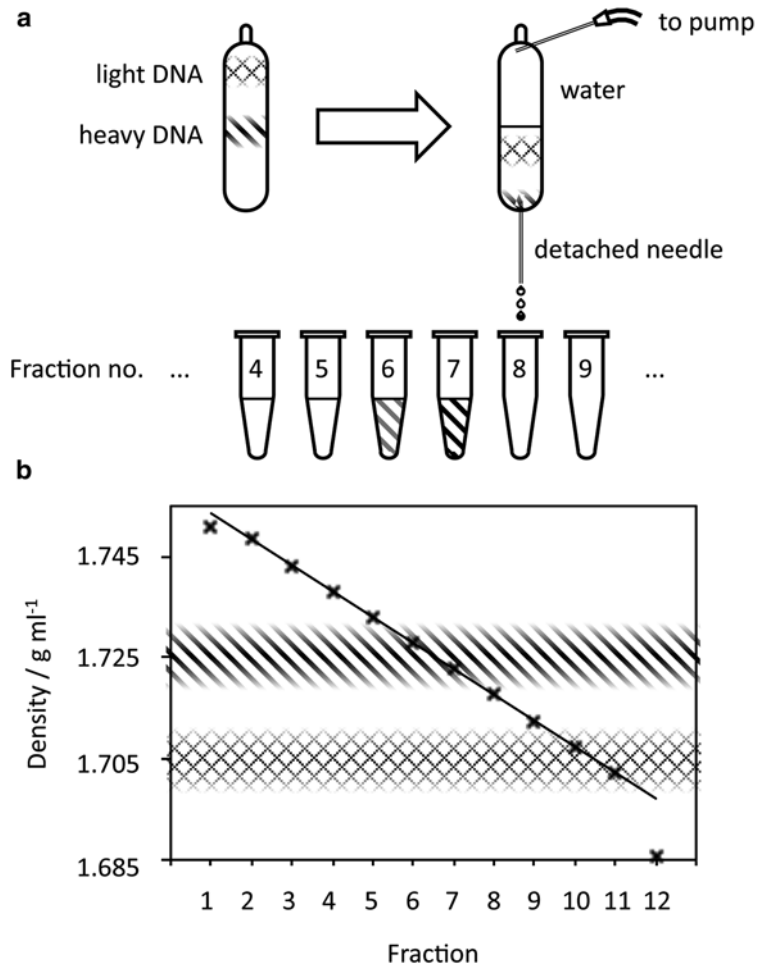


Fig. 1 Illustration of the density gradient fractionation process after separation of labeled and unlabeled DNA by ultracentrifugation. (a) The ultracentrifuge tube is pierced at the top and bottom and the CsCl solution is replaced by water and collected in 1.5 mL tubes. The refractive indices of individual fractions are determined for calculation of densities. ¹³C-labeled DNA is typically found in fractions 6–8, at a density of around 1.725 g/mL. (b) The density curve typically shows a deviation from linearity for the first fraction (due to diffusion) and the last fraction (due to mixing with water). Labeled DNA is indicated by diagonal line pattern, unlabeled DNA by a checked pattern

Label the fractions in the order of collection, 1–12. Repeat the fractionation process with the next sample (*see Note 14*).

5. Measure the refractive indices of all individual fractions to ensure proper gradient formation (*see Note 15*). The refractive indices typically are ± 0.0025 around the refractive index measured before ultracentrifugation, with the first fractions having the highest refractive index and the last fractions the lowest. With the refractive indices, the densities of the fractions

can be calculated based on the calibration curve prepared in Subheading 3.1, **step 1**. On average, ^{13}C -labeled DNA has a density of 1.725 g/mL; unlabeled DNA has an average density of 1.705 g/mL.

6. In order to purify DNA from the CsCl solution, precipitate DNA by adding 5 μL LPA (5 mg/mL) per fraction and mix well (*see Note 16*). Add 850 μL of PEG-NaCl solution and mix well. Leave at room temperature for at least 2 h to allow precipitation. Incubation overnight is also possible. Centrifuge at $13,000\times g$ for 30 min and withdraw supernatant with a 1 mL pipette. A transparent pellet should be visible after the supernatant is removed. Wash with 400 μL of 70 % ethanol, centrifuging at $13,000\times g$ for 10 min. Discard supernatant as before. A white pellet should be visible now, which can easily become detached from the tube wall. Air-dry for 15 min, then suspend in 50 μL TE buffer for 30 min on ice, tapping the tube every 5–10 min to ensure that DNA dissolves fully.

3.4 Identification of Labeled DNA by DGGE Fingerprinting

1. Check retrieval and quality of DNA obtained from individual fractions by applying 5 μL to 1 % (w/v) agarose gel electrophoresis following standard laboratory procedures. Quantification of DNA is possible by fluorometric assays. Do not use photometric DNA quantification based on absorbance in the UV range, because this is usually not sensitive enough to detect the low amounts of DNA that might be present. High molecular mass DNA bands should be visible under UV light after staining with ethidium bromide (0.5 $\mu\text{g}/\text{mL}$ gel) in at least some of the fractions, typically between fractions 6 and 12. For troubleshooting on DNA retrieval, *see Table 2*.
2. Perform a PCR with primers targeting rRNA genes of the organisms of interest, including a GC clamp. To target bacterial 16S rRNA genes, we typically use the primer set 341f_GC/518r which amplifies a ~230 bp portion of the gene, spanning the V3 hypervariable region. The PCR conditions are: 95 °C for 5 min, 30 cycles of 94 °C for 1 min, 55 °C for 1 min, 72 °C for 1 min, followed by a final extension of 72 °C for 5 min [31]. The final reaction volume is 50 μL . Check the PCR products by applying 5 μL to 1 % (w/v) agarose gel electrophoresis. Prepare samples for DGGE by mixing 4–40 μL of the PCR product, according to band intensity on the agarose gel, with DGGE loading dye to achieve a $1\times$ final concentration.
3. Prepare a gel for denaturing gradient gel electrophoresis. The following volumes are given for DGGE equipment supporting 20×20 cm glass plates in a 6.5 L tank. Transfer 12.5 mL of the 30 % and 70 % DGGE solution to two 15 mL falcon tubes and keep them on ice. Add 12.6 μL of TEMED and 126 μL of APS solution to each tube and transfer to a gradient mixer. Cast

Table 2
Potential sources of problems in fractionation of DNA from SIP experiments and recommended solutions

Problem	Potential reason	Solution
No gradient formation	Problems with ultracentrifugation	Repeat ultracentrifugation, ensure no brakes are applied for deceleration and no errors during run
No DNA recovery	Wrong density (DNA sticking on side of tube)	Check correct starting density
	DNA amount too low	Use >5 µg of DNA as starting material
	Loss during DNA precipitation	Make sure to use carrier. Make sure to visualize pellet
DNA at unexpected densities	Incorrect calibration density/nD-TC	Repeat density calibration
	Temperature deviation during ultracentrifugation	Do ultracentrifugation at 20 °C
No difference between ¹² C and ¹³ C experiment <i>or</i> Low amount of ¹³ C DNA	Insufficient labeling	Increase substrate concentrations and/or incubation times ^a
		Try to amplify by MDA
DNA at intermediate densities (partially labeled)	Insufficient labeling (partially and unlabeled DNA)	Increase incubation time
	Crossfeeding (partially and fully labeled DNA)	Reduce incubation time
	Organisms use alternative carbon source (e.g. CO ₂)	Perform additional metabolic labeling experiment with ¹³ C-CO ₂ for confirmation
Only labeled DNA in ¹³ C experiment	High enrichment of active organisms	Reduce substrate concentrations and/or incubation times
	Crossfeeding	Reduce incubation time
Same genotypes in all fractions	Contamination during/ after fractionation	Change solutions, repeat ultracentrifugation with fresh DNA

^aOrganisms that metabolize a C1 compound without using it as carbon source cannot be detected

gradient gel according to standard laboratory protocols, with the 70 % solution at the bottom and the 30 % solution on top. Overlay the gel with 0.5 mL of isopropanol to achieve an even surface. Wait 45 min for the gel to polymerize.

4. Prepare the DGGE tank with 6.4 L of water and add 130 mL of 50× TAE buffer to a final concentration of 1× TAE, heat up to 60 °C. Remove the isopropanol from the polymerized gel and rinse the surface with water three times. Cast top-up gel with 5 mL of 0 % DGGE solution, 5 µL of TEMED and 50 µL

- of 10 % APS solution. Insert a 16-well comb without introducing air bubbles. Wait 30 min for the top-up gel to polymerize.
- Submerge the gels in the DGGE tank and rinse the wells with buffer. Load the samples prepared in **step 3**. Load DGGE ladder if available and load empty wells with 1× DGGE loading dye. Run the DGGE at 75 V for 16 h overnight. Ideally, run all fractions of a ^{13}C sample and the corresponding ^{12}C sample on two gels in parallel. After electrophoresis, stain the gels according to standard laboratory protocols (e.g. with SYBR Gold) and image the gel for evaluating fractionation results.
 - Check band patterns to identify fractions of the ^{13}C sample containing labeled DNA by careful comparison with the gel of the corresponding ^{12}C sample. Unlabeled DNA typically is found in fractions 10–12, fully labeled DNA in fractions 6–8. Ignore bands that are present in all fractions, as these are not likely to have originated from ^{13}C -DNA alone. Look for bands that are consistently present in the light fractions of both the ^{12}C and ^{13}C sample to identify unlabeled DNA. Then look for bands that change their position in relation to the unlabeled DNA to identify the labeled DNA (*see Note 17*). Select the appropriate fractions for further experimentation (*see Fig. 2*). *See also Table 2* for troubleshooting advice.

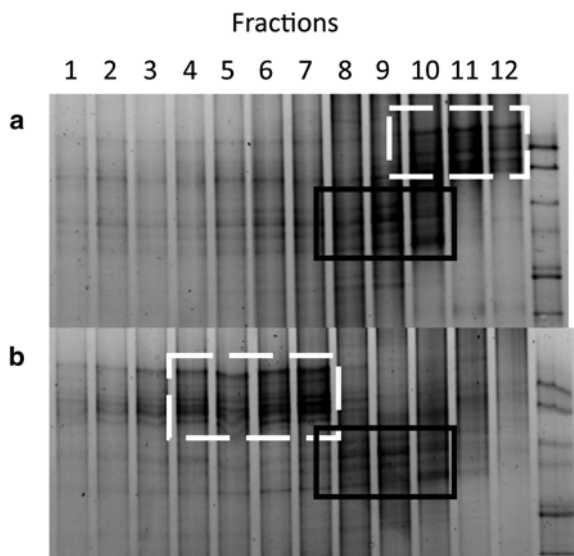


Fig. 2 DGGE gels obtained from fractionated DNA of (a) ^{12}C and (b) ^{13}C incubations on ^{13}C -labeled methanol after electrophoresis for 16 h at 75 V. *Black box*: bands occurring in the same fractions in ^{12}C and ^{13}C incubations representing unlabeled DNA. *White box*: bands enriched in the heavy fractions of the ^{13}C incubation due to labeling of DNA by methylotrophic activity

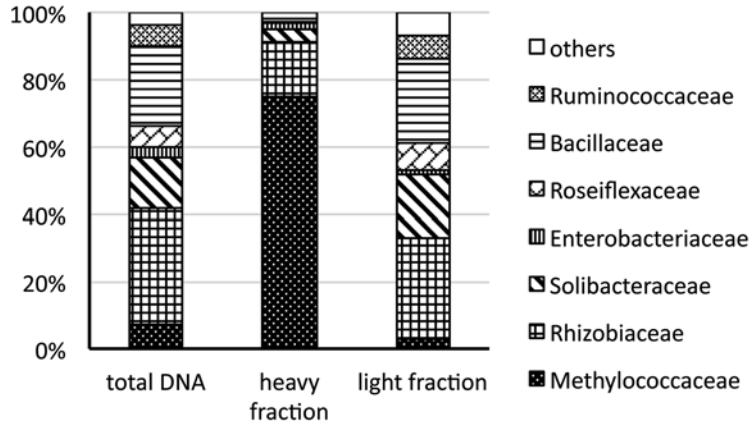


Fig. 3 Theoretical expected results of 454 amplicon pyrosequencing data targeting 16S rRNA genes in unfractionated DNA, DNA from heavy fractions and from light fractions. The heavy fractions show a strong enrichment, compared to unfractionated DNA, of the putative active methylotroph of the family *Methylococcaceae*, while the light fractions show sequences of the remaining, non-methylotrophic/inactive bacteria also detected in the total DNA

3.5 Analysis of Labeled DNA

1. Perform PCR assays with primers targeting bacterial 16S rRNA genes on the labeled DNA. Purify PCR products obtained using PEG-NaCl precipitation as described in Subheading 3.2, **step 6**. Perform sequencing of 16S rRNA gene PCR amplicons to acquire an overview of the phylogenetic composition of the labeled DNA and to identify putative methylotrophs (*see Note 18*). This may also be done with the unlabeled (light) DNA of the ^{13}C sample for comparison, to illustrate the relative enrichment of genes from methylotrophic organisms in the labeled DNA (*see Fig. 3*).
2. Screen for functional genes encoding key enzymes for methylotrophy by PCR. Depending on the investigated processes and applied substrates, different genes can be of interest. Commonly targeted are *mxnF*, encoding the large subunit of methanol dehydrogenase, *pmoA* and *mmoX*, encoding subunits of the particulate and soluble methane monooxygenase, as well as *mauA* and *gmaS*, encoding genes for alternative pathways of methylamine degradation (*see Table 1* for PCR primers and references). Purify obtained PCR products using PEG-NaCl precipitation as described in Subheading 3.2, **step 6**.
3. Sequence functional gene PCR amplicons by 454 pyrosequencing. We commonly use the software packages mothur and USEARCH when analyzing data from a GS FLX Titanium system (*see Note 19*). Use Mothur to extract flowgrams from raw *.sff data files with the sffinfo() command. Discard flowgrams with less than 450 usable flows and cut remaining flowgrams to 720 flows with trim.flows(). Denoise flowgrams and translate to

nucleic acid sequences using `shhh.flows()`. Use the `trim.seqs()` command to demultiplex sequences and remove barcode and primer sequences, to discard sequences with errors in the barcode or primer region, with ambiguous bases or homopolymer runs >6 bp and to filter sequences by length, depending on the expected product size. The `count.seqs()` command can be used to obtain quantitative information. Use USEARCH to sort sequences by abundance (`-sortbysize`) and for binning of operational taxonomic units (OUT), chimera removal and singleton removal (`-cluster.otus`). Use a 90 % identity threshold for this step (*see* **Note 20**).

4. The obtained OTUs can be analyzed either by approaches based on the basic local alignment search tool (BLAST, [35]) or by generating phylogenetic trees after aligning with reference sequences (*see* **Note 21**). The resulting phylogenetic affiliation of the functional genes of interest can be compared to the data obtained by 16S rRNA gene sequencing to confirm the presence of putative methylootrophs.
5. For a more comprehensive analysis of the enriched DNA, shotgun metagenomic sequencing can be used. Due to the low DNA amounts typically present in the fractions, multiple displacement amplification (MDA) can be used to obtain sufficient material for sequencing. Use a commercially available MDA kit and follow the manufacturer's instructions. We commonly use the REPLI-g Mini Kit (QIAGEN) with 1–10 ng of DNA as template, incubating for 16 h overnight at 30 °C, followed by heat inactivation for 3 min at 65 °C. Perform amplification in replicates and check fidelity of the amplified DNA by DGGE (*see* Subheading 3.3, steps 2–5; *see* **Note 22**). Merge and purify amplified DNA (*see* Subheading 3.2, step 6) before shotgun metagenomic sequencing.
6. Perform shotgun metagenomic sequencing using in-house protocols or a commercially available service (also *see* **Note 19**). First analysis of the sequences can be done by using the metagenomics Rapid Annotation using Subsystem Technology (MG-RAST) analysis server (metagenomics.anl.gov, [36]). This platform is designed to call and annotate the genes in a large set of short DNA sequence reads by comparison with DNA and protein databases. This allows an in-depth phylogenetic and functional analysis of the reads, as well as screening for functional genes of interest. *See* Chapter 4 “MG-RAST” for more information. If one or a few species are specifically enriched, assembly of the reads can be used to obtain larger DNA sequence fragments or even nearly complete genomes of the investigated methylootrophs, leading to additional information about organization of gene clusters and allowing reconstruction of bacterial metabolism.

4 Notes

1. EDTA will slowly dissolve as the pH gets near 8.0. When using solid NaOH pellets, around 18–20 g are required. Use a 2 M NaOH solution for more precise adjustment of the pH.
2. DNA extraction protocols will differ based on the source material (e.g. soil, sediment, sludge, biofilm, or aquatic samples) and, consequently, no specific instructions can be given. Do test extractions from source material obtained directly from the environment to establish a suitable DNA extraction method before starting a metabolic labeling experiment.
3. The high amount of CsCl leads to an increase in volume when dissolving. Make sure not to add too much water initially. Stirring and gently warming in a water bath will help to dissolve the CsCl more quickly.
4. For measuring density, use a digital density meter or carefully weigh 1-mL aliquots in triplicate. Make sure the solution is at 20 °C before beginning this process. If the density is too low, add more CsCl. Adding 5–10 g of CsCl increases the density by ~0.01 g/mL. A density above 1.89 g/mL can still be used if adjustments are done when setting up samples for ultracentrifugation (*see* Subheading 3.1, **step 4**).
5. This is discussed in more detail elsewhere [24, 25].
6. Density measurement using an analytical balance is tedious and much less accurate than refractive index measurement, and also provides a higher chance for sample loss or DNA contamination.
7. For example, when using 4.8 mL of a stock solution with a density of 1.890 g/mL, this equates to:
$$\text{Required volume} = (1.890 - 1.725 \text{ g/mL}) \times 4.8 \text{ mL} \times 1.52 \text{ mL/g}$$
$$\text{Required volume} = 1.20 \text{ mL of GB}$$
8. Before starting with your samples, prepare a sterile 15 mL tube with 4.8 mL of CsCl stock solution and the calculated volume of GB; mix well by inversion. Measure the refractive index and adjust as described. Addition of 10 μL of GB will decrease the refractive index by ~0.0001, and addition of 40 μL of CsCl solution will increase it by ~0.0001. Keep track of the additions to correct the required volume of GB calculated in **step 2**. The prepared solution can later be used to top up ultracentrifuge tubes in case there is too little solution for a sample, or for balancing tubes.
9. Sometimes the sealing process leads to a change in tube weight. If this occurs, or if you are in doubt about the sealing on a tube, it is best to prepare a completely new ultracentrifuge tube. For recovery of the sample, cut off the top of the suspicious tube

and empty it into the 15 mL tube used to prepare that sample by holding the ultracentrifuge tube upside down and squeezing repeatedly.

10. Differences in centrifugation speed and thus centrifugal force will lead to differences in gradient formation. Higher centrifugal forces result in a steeper gradients and thus in poorer sample separation. Lower centrifugal forces result in shallower gradients. Although this can increase sample separation slightly, lower centrifugal forces also require highly extended run times. The proposed centrifugal force of $177,087 \times g$ average is the best trade-off between sample separation and run time.
11. Previous protocols suggested the use of mineral oil for this purpose, but we found that water can be used to simplify the process of fractionation. Due to the high density difference between the CsCl solution in the tube and the water, only limited mixing will occur. For improved visualization, bromophenol blue or another dye can be added to the water.
12. Setting up the fractionation and piercing a tube can be difficult to do correctly for the first time. Prepare sealed ultracentrifuge tubes with water to test this process beforehand to ensure that it is working smoothly before processing the samples. Hold the tube with one hand to fix it securely in the clamp, otherwise it might slide down when you apply force with the needle. Put your thumb on top of the tube next to the tube stem and two other fingers under the tube. This ensures that you have the best control of the tube without having any fingers in line with the needle when piercing (potential danger of injury!). Make sure to apply controlled force to prevent the needle from entering too deep. Twisting the needle slightly can help to “drill” through the tube wall. The sharpened tip of the needle has to penetrate the tube wall completely to avoid spillage later on. This means that the first few millimeters of the needle will be inside the tube, but not more. Once through the tube wall, the needle will move much easier than before. To prevent deeper entry, you can wrap sticky tape around the needle or put a short piece of tubing over it beforehand, so the needle is blocked by the tube wall from going any deeper. If liquid from the tube is forced out once the tube is pierced, reduce the pressure from the clamp holding the tube.
13. Use the same precautions as when piercing the top of the tube. Under rare circumstances, when the needle at the top of the tube is not sealing properly, the tube can run out very suddenly at this step. Have a 15 mL tube ready to catch the CsCl solution in case this happens, so you can use the sample for a new ultracentrifugation. If a low amount of leaking occurs, a small drop of mineral oil applied to the top puncture hole can help prevent further sample loss. The following fractionation with

the pierced bottom of the tube will result in relatively large drops, and thus differences in fraction size. To create smaller drops and to allow easier fractionation, a detached 23-gauge needle can be fitted into the hole at the bottom of the tube. To do so, break a needle from the Luer slip (plastic part) by gently bending it left and right with tweezers, then carefully push it into the prepared hole.

14. The CsCl gradient is not stable over time and will mix again through diffusion. This will first be noticeable at the top and bottom of the gradient when measuring the refractive indices. Thus it is recommended to carry out the fractionation of all samples in a row as soon as the ultracentrifuge run has ended. Calculate roughly 20 min per sample (12 min of fractionation and 8 min of preparation). Fractionating eight samples in 3 h usually gives optimal results. If multiple pumps are available, fractionation in parallel is an option.
15. Although manufacturers of digital refractometers usually recommend covering the entire prism before measurement, a single drop in the centre of the prism is often sufficient for an accurate measurement. Depending on the model of refractometer used, accurate measurements can be obtained with volumes as small as 20 μL . This greatly reduces the loss of material at this step. Consistency of measurements should be checked before attempting to work with actual samples.
16. The addition of a carrier substance like LPA or glycogen is essential for the recovery of the small DNA amounts that might be present in the fractions (often <100 ng). Due to contamination issues with commercially available glycogen [37], we recommend LPA, which can be easily prepared in-house for a fraction of the cost of the commercially available product. UV treatment prior to use can ensure nucleic acid contamination will not affect downstream analysis.
17. DNA density is not only influenced by ^{13}C incorporation, but also by GC content: DNA with low GC content has a lower density than DNA with high GC content. This can lead to unlabeled genomic DNA spanning ± 2 fractions in the described protocol. Hence it is essential to have fractions of a ^{12}C control experiment as reference for identifying labeled DNA bands; selecting fractions based only on density can be misleading. *See* [14, 15] for examples on identifying labeled DNA bands.
18. A variety of primer sets targeting different regions of bacterial 16S rRNA gene sequences have been described and can be used to acquire amplicons for sequencing. Likewise, different high-throughput sequencing methods are available for this purpose, also as commercial services that include bioinformatics analysis of the obtained sequences. If no in-house sequencing and analysis pipeline is available, use of such a service is recommended.

19. Alternative sequencing methods (Illumina dye sequencing, Ion semiconductor sequencing) can be used instead. Be aware that methods producing reads from a defined position of the gene of interest, i.e. the primer sequence, can be investigated in the way described and binned to OTUs. Methods producing random reads from the amplicons cannot be binned to OTUs with the tools described, but can be analyzed using an approach based on the basic local alignment search tool (BLAST) using Megan [38]. See [39] for an alternative approach employing Megan for 454 pyrosequencing data. Instead of mothur, also the software package QIIME can be used [40].
20. Higher identity thresholds can be used, but be aware that sequence diversity for functional genes can be rather different than for 16S rRNA genes when trying to assign OTUs to different phylogenetic levels. Furthermore, 454 pyrosequencing and ion semiconductor sequencing is prone to errors on homopolymer repeats, sometimes introducing up to 5 % sequencing errors. Thus, while an identity threshold of 90 % might lead to the loss of resolution on the highest taxonomic levels, it will effectively reduce artificial diversity introduced by sequencing errors.
21. Be aware that different algorithms can lead to different results, especially when only distantly related reference sequences are available. This is true for different clustering algorithms when constructing phylogenetic trees as well as for different BLAST algorithms. Also be aware that reference data for functional genes are usually much more limited than for 16S rRNA genes, and environmental samples can often yield sequences that can only be classified on lower taxonomic levels due to the lack of matching reference sequences.
22. MDA is highly prone to contamination and most available kits can introduce an amplification bias [41]. Thus, if somehow possible, it should be avoided. If very low DNA amounts are retrieved, and amplification before sequencing is essential, there are several possibilities to improve product quality. Split your sample into replicates (5–10) before amplification to reduce bias. Reduce the volume of each reaction as far as possible and reduce the incubation time to avoid contamination and unspecific amplification.

Acknowledgement

This work was possible thanks to financial support from the Gordon and Betty Moore Foundation Marine Microbiology Initiative Grant GBMF3303 to J. Colin Murrell and Yin Chen and through the Earth and Life Systems Alliance, Norwich Research Park, Norwich, UK.

References

1. Carpenter LJ, Archer SD, Beale R (2012) Ocean-atmosphere trace gas exchange. *Chem Soc Rev* 41:6473–6506
2. Heikes BG, Chang WN, Pilson MEQ, Swift E, Singh HB, Guenther A, Jacob DJ, Field BD, Fall R, Riemer D, Brand L (2002) Atmospheric methanol budget and ocean implication. *Global Biogeochem Cycles* 16:8001–8013
3. Carini P, White AE, Campbell EO, Giovannoni SJ (2014) Methane production by phosphate-starved SAR11 chemoheterotrophic marine bacteria. *Nat Commun* 5:4346
4. Chen Y, McAleer KL, Murrell JC (2010) Monomethylamine as a nitrogen source for a nonmethylotrophic bacterium, *Agrobacterium tumefaciens*. *Appl Environ Microbiol* 76:4102–4104
5. Kiene RP, Linn LJ, Bruton JA (2000) New and important roles for DMSP in marine microbial communities. *J Sea Res* 43:209–224
6. Anthony C (1982) The biochemistry of methylotrophs. Academic, New York
7. Neufeld JD, Wagner M, Murrell JC (2007) Who eats what, where and when? Isotope labelling experiments are coming of age. *ISME J* 1:103–110
8. Chistoserdova L (2011) Modularity of methylotrophy, revisited. *Environ Microbiol* 13:2603–2622
9. Holmes AJ, Costello A, Lidstrom ME, Murrell JC (1995) Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiol Lett* 132:203–208
10. McDonald IR, Murrell JC (1997) The methanol dehydrogenase structural gene *mxnA* and its use as a functional gene probe for methanotrophs and methylotrophs. *Appl Environ Microbiol* 63:3218–3224
11. Costello AM, Lidstrom ME (1999) Molecular characterization of functional and phylogenetic genes from natural populations of methanotrophs in lake sediments. *Appl Environ Microbiol* 65:5066–5074
12. Auman AJ, Stolyar S, Costello AM, Lidstrom ME (2000) Molecular characterization of methanotrophic isolates from freshwater lake sediment. *Appl Environ Microbiol* 66:5259–5266
13. Hutchens E, Radajewski S, Dumont MG, McDonald IR, Murrell JC (2004) Analysis of methanotrophic bacteria in Movile Cave by stable isotope probing. *Environ Microbiol* 6:111–120
14. Neufeld JD, Schafer H, Cox MJ, Boden R, McDonald IR, Murrell JC (2007) Stable-isotope probing implicates *Methylophaga* spp. and novel Gammaproteobacteria in marine methanol and methylamine metabolism. *ISME J* 1:480–491
15. Wischer D, Kumaresan D, Johnston A, El Khawand M, Stephenson J, Hillebrand-Voiculescu AM, Chen Y, Colin Murrell J (2014) Bacterial metabolism of methylated amines and identification of novel methylotrophs in Movile Cave. *ISME J* 9:195–206
16. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21:1794–1805
17. Lüke C, Frenzel P (2011) Potential of *pmoA* amplicon pyrosequencing for methanotroph diversity studies. *Appl Environ Microbiol* 77:6305–6309
18. Kolb S, Stacheter A (2013) Prerequisites for amplicon pyrosequencing of microbial methanol utilizers in the environment. *Front Microbiol* 4:1–12
19. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74
20. Chistoserdova L (2014) Is metagenomics resolving identification of functions in microbial communities? *Microb Biotechnol* 7:1–4
21. Boschker H, Nold S, Wellsbury P, Bos D, De Graaf W, Pel R, Parkes R, Cappenberg T (1998) Direct linking of microbial populations to specific biogeochemical processes by ¹³C-labelling of biomarkers. *Nature* 392:801–805
22. Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403:646–649
23. Neufeld JD, Vohra J, Dumont MG, Lueders T, Manfield M, Friedrich MW, Murrell JC (2007) DNA stable-isotope probing. *Nat Protoc* 2:860–866
24. Murrell JC, Whiteley AS (2011) Stable isotope probing and related technologies. ASM, Washington, DC
25. Neufeld JD, Dumont MG, Vohra J, Murrell JC (2007) Methodological considerations for the

- use of stable isotope probing in microbial ecology. *Microb Ecol* 53:435–442:2027
26. Dunford EA, Neufeld JD (2010) DNA stable-isotope probing (DNA-SIP). *J Vis Exp* 42:2027
 27. Neufeld JD, Chen Y, Dumont MG, Murrell JC (2008) Marine methylootrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ Microbiol* 10:1526–1535
 28. Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suci D, Levine SR, Markowitz VM, Rigoutsos I, Tringe SG, Bruce DC, Richardson PM, Lidstrom ME, Chistoserdova L (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol* 26:1029–1034
 29. Green SJ, Leigh MB, Neufeld JD (2010) Denaturing Gradient Gel Electrophoresis (DGGE) for microbial community analysis. In: Timmis K (ed) *Handbook of hydrocarbon and lipid microbiology*. Springer, Berlin Heidelberg, pp 4137–4158
 30. Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* 2:233–241
 31. Muyzer G, De Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* 59:695–700
 32. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541
 33. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10:996–998
 34. Cebron A, Bodrossy L, Stralis-Pavese N, Singer AC, Thompson IP, Prosser JI, Murrell JC (2007) Nutrient amendments in soil DNA stable isotope probing experiments reduce the observed methanotroph diversity. *Appl Environ Microbiol* 73:798–807
 35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
 36. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386
 37. Bartram A, Poon C, Neufeld J (2009) Nucleic acid contamination of glycogen used in nucleic acid precipitation and assessment of linear polyacrylamide as an alternative co-precipitant. *Biotechniques* 47:1019–1022
 38. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21:1552–1560
 39. Dumont MG, Lüke C, Deng YC, Frenzel P (2014) Classification of *pmoA* amplicon pyrosequences using BLAST and the lowest common ancestor method in MEGAN. *Front Microbiol* 5:1–11
 40. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenkov T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336
 41. Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* 7:943–944

Functional Metagenomics: Construction and High-Throughput Screening of Fosmid Libraries for Discovery of Novel Carbohydrate-Active Enzymes

Lisa Ufarté, Sophie Bozonnet, Elisabeth Laville, Davide A. Cecchini, Sandra Pizzut-Serin, Samuel Jacquiod, Sandrine Demanèche, Pascal Simonet, Laure Franqueville, and Gabrielle Potocki Veronese

Abstract

Activity-based metagenomics is one of the most efficient approaches to boost the discovery of novel biocatalysts from the huge reservoir of uncultivated bacteria. In this chapter, we describe a highly generic procedure of metagenomic library construction and high-throughput screening for carbohydrate-active enzymes. Applicable to any bacterial ecosystem, it enables the swift identification of functional enzymes that are highly efficient, alone or acting in synergy, to break down polysaccharides and oligosaccharides.

Key words Metagenomic DNA, Fosmidic libraries, High-throughput screening, Carbohydrate-active enzymes, Complex glycans

1 Introduction

Early metagenomic studies focused on exploring microbial diversity through sequencing of ribosomal RNA sequences and later, with the emergence of powerful sequencing technologies, of functional DNA recovered from environmental samples. Generated data usually cover several gigabases of sequence information in the form of short sequences, which need to undergo an assembly pipeline in order to extract useful gene information. However, gene annotation provides only a functional potential to the annotated genes, according to their sequence homology, the real activity requiring an experimental demonstration.

Activity-based metagenomics allows to by-pass these challenges, as proven by numerous studies dedicated to the discovery of novel enzymes, in particular Carbohydrate-Active Enzymes or CAZymes [1]. Indeed, carbohydrates, in particular glycans, assure key and highly versatile functions in the living world, for energy storage,

cell signaling, recognition, or shape maintain. Carbohydrate metabolism is thus crucial for all organisms, and requires a large panel of CAZymes to cleave, modify, or create osidic linkages. Moreover, as plant polysaccharides constitute the main source of renewable carbon, the extraordinary diversity and natural efficiency of microbial CAZymes can be exploited to develop green processes of plant biomass conversion into biofuels or sugar-based materials including surfactants, fine-chemicals, secondary metabolites, drugs, vaccines, among others.

As a result, in the last years, numerous studies have been published, exploiting the immense potential of functional metagenomics to explore various bacterial ecosystems. Indeed, many ecosystems like mammal and insect guts, soil and composts, are more or less specialized in polysaccharide breakdown, enabling the discovery of carbohydrate catabolism-related enzymes, like glycoside hydrolases.

Functional metagenomics consists in (1) constructing large libraries of thousands to several hundred thousands of recombinant clones, carrying metagenomic DNA fragments sizing between 2 and 200 kbp, cloned into plasmids, cosmids, fosmids or even bacterial artificial chromosomes, (2) screening them for the targeted activities, and (3) sequencing the screening hits in order to identify the genes that are responsible for the observed phenotypes (Fig. 1). By using this approach, several hundreds of novel CAZymes were

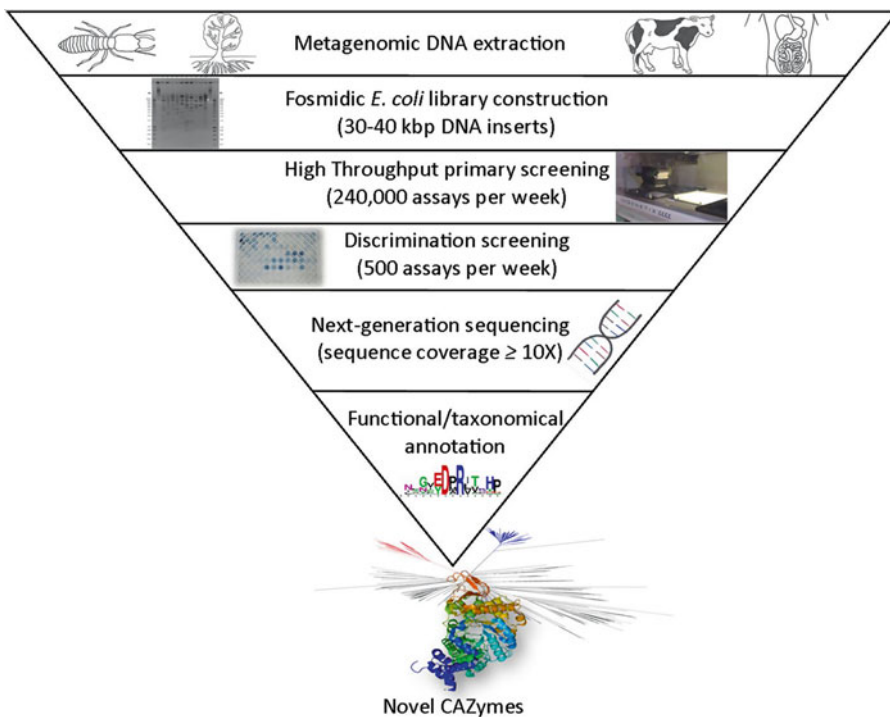


Fig. 1 Multistep activity-based metagenomic strategy for CAZyme discovery

retrieved from metagenomes these last few years [2, 3], most of them presenting very original sequences [4], sometimes belonging to novel protein families [5] and/or displaying inedited key functions of carbohydrate foraging [6] which would not have been predicted by genomic or metagenomic sequence analysis.

In this chapter, we describe a robust and inexpensive procedure of high-throughput functional exploration of bacterial ecosystems (soil being used as an example), to drive in-depth metagenome sequencing and focus on genes encoding catabolic CAZymes. Even if alternative strains with different expression and secretion capabilities can be used [7, 8], we detail the construction of *E. coli* fosmidic metagenomic libraries, as it allows to easily and rapidly explore extremely large sequence spaces, covering several Gbp of metagenomic DNA. Search for functional CAZyme encoding genes consists here in applying a multistep screening approach to (1) isolate clones producing catalysts with the desired specificity toward polysaccharidic and oligosaccharidic substrates, (2) discriminate endo- and exo-hydrolytic activities (Fig. 2a, b) and even discover enzyme cocktails, encoded by multigenic clusters that are frequently found on large metagenomic fosmidic inserts (sizing between 30 and 50 kbp) that are involved in the break-

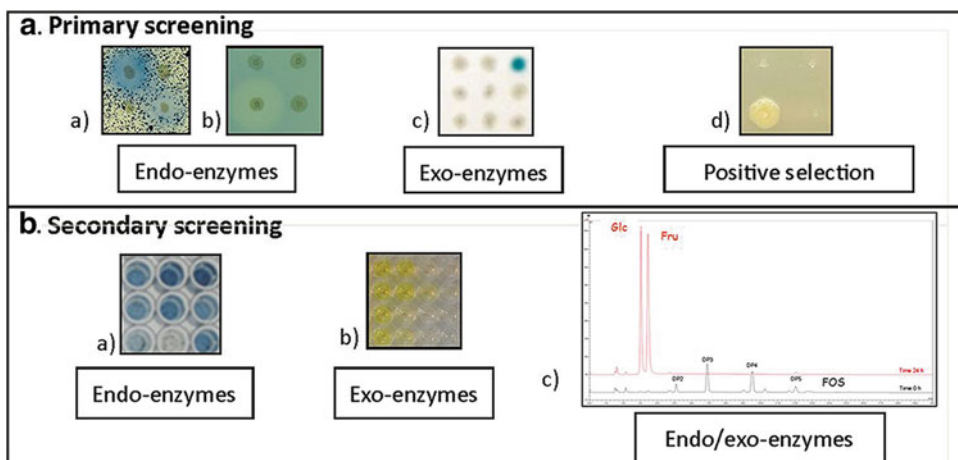


Fig. 2 High-throughput screening of *E. coli* metagenomic libraries for endo- and exo-acting glycanases. **(a)** Pictures of primary screening results (solid medium) on *(a)* insoluble AZCL-polysaccharides, *blue halos* showing the release of soluble AZCL-oligosaccharides; *(b)* solubilized Azo-polysaccharides, *clear halo* showing the degradation of the colored polymer; *(c)* X-mono/oligosaccharides, *blue color* showing the release of free X-compounds; *(d)* minimal medium supplemented with oligosaccharides as carbon source, the sole growing clone being able to degrade the targeted oligosaccharides. **(b)** Pictures of secondary screening results (liquid medium) on *(a)* AZCL-polysaccharides, *blue color* showing the release of soluble AZCL-oligosaccharides in the reaction medium; *(b)* *pNP*-mono/oligosaccharides, *yellow color* showing the release of free *pNP*-compounds in the reaction medium; *(c)* HPAEC-PAD analysis of, in *black*, oligosaccharides substrate (fructo-oligosaccharides as an example) before enzymatic hydrolysis; in *red*, hydrolysis reaction products (glucose and fructose as an example) after 24 h of reaction

down of plant cell wall. If the automated solid plate assays used in the primary screens can be easily carried out by only one person at a throughput of 240,000 assays per week (corresponding to the screening of 20,000 clones for 12 different activities) for only few k€, the discriminating assays in liquid media are used with a lower throughput of around 500 assays per week. However, they are highly recommended in order to avoid total or partial sequence redundancy between the hits presenting the same activities. Sequence redundancy can also be avoided by choosing to screen several little libraries (of few dozens of thousands clones) constructed from different samples rather than only a large one (of hundred thousand clones) issued from one unique sample.

This highly generic approach is applicable to mine all complex bacterial communities for novel catabolic CAZymes. Depending on their ability to face the structural complexity of plant cell wall, and to use it as the main carbon source for growing and maintaining themselves in their habitat, the hit rates will vary from less than 0.2‰ (for example for soil communities) to more than 4‰ (for highly specialized ecosystems like termite guts [9]). In any case, in order to increase hit yield, we recommend (1) increasing the number of primary screens by using a large diversity of polysaccharidic and oligosaccharidic substrates, varying in nature of glycosidic residues, type of osidic linkages, polymerization degree and ramification content; (2) increasing the library size, preferably constructed from several metagenomic DNA samples, in order to minimize sequence redundancy.

After hit recovery, fosmid sequencing with high coverage (more than 10×) allows to easily identify the genes, or the gene clusters, that are responsible for the screened activity, and their taxonomic origin, sometimes up to species level. As the functional and taxonomical annotation procedures do not differ from those developed for sequenced-based metagenomics and genomics, they will not be detailed in this chapter.

2 Materials

All plastics used are certified free of DNase and DNA. All tools and materials used should be washed and cleaned with 70 % ethanol solution (*see Note 1*). Glass or metal materials, as well as solutions when specified, are sterilized before use (121 °C, 20 min). Prepare all solutions using ultrapure water (18 MΩ cm at 25 °C) and analytical grade reagents. Follow all waste disposal regulations when disposing waste materials.

2.1 DNA Sampling

1. 4-mm sterilized glass beads.
2. 0.2 % Sodium hexametaphosphate (HMP): Add about 400 mL water to a 500-mL measuring tube. Weight 1.0 g HMP and

transfer to the cylinder. Add water to a volume of 500 mL. Mix and transfer to a 1-L glass bottle. Sterilize and store at room temperature.

3. Two 250-mL sterilized polypropylene Nalgene tubes.
4. Sterile gauze.
5. 0.8 % Sodium chloride (NaCl): Add about 700 mL water to a 1-L measuring tube. Weight 8.0 g NaCl and transfer to the cylinder. Add water up to 1 L. Mix and transfer 500 mL to two 1-L glass bottles. Sterilize and store at room temperature.
6. 1.3 g/mL 5-(*N*-2,3-dihydroxy propylacetamido)-2,4,6-tri-iodo-*N,N'*-bis (2,3 dihydroxypropyl) isophthalamide (Nycodenz[®]) (Axis-Shield): in order to obtain a density of 1.3 g/mL, mix 50 mL water and 40 g Nycodenz[®] in a 100-mL glass bottle with a magnetic stirrer. Stir and heat to 50 °C to dissolve Nycodenz[®]. Remove the stirrer, sterilize and store at room temperature.
7. Tris-HCl-EDTA buffer (TE): 50 mM Tris-HCl (pH 8.0) with 100 mM EDTA buffer.
8. InCert[®] agarose (BMA).
9. Plug molds (Bio-Rad).
10. Lysis buffer A (LA): 50 mM Tris-HCl (pH 8.0), 100 mM EDTA, 5 mg/mL lysozyme, 0.5 mg/mL achromopeptidase.
11. Lysis buffer B (LBB): 50 mM Tris-HCl (pH 8.0), 100 mM EDTA, 1 % lauryl sarcosyl, 2 mg/mL proteinase K.
12. Storage buffer: 10 mM Tris-HCl (pH 8.0), 1 mM EDTA.
13. 0.1 mM Phenylmethanesulfonyl fluoride (PMSF) (Sigma): dilute the weighed powder directly in storage buffer.

2.2 Fosmid Library Construction

1. Low-melting-temperature agarose (Bio-Rad).
2. Tris-acetate-EDTA (TAE): dilute 10× stock solution ten times (Promega).
3. PFGE ladder: lambda bacteriophage DNA (NEB).
4. 1 µg/mL Ethidium bromide: dilute in water.
5. GELase (Epicentre Technologies).
6. EpiFOS[™] Fosmid Library Production Kit (Epicentre, Illumina[®]).
7. 50 mg/mL Chloramphenicol: prepare stock solution in ethanol and filter-sterilize before aliquots storage at -20 °C.
8. Freezing medium: Luria-Bertani medium supplemented with 20 % (w/v) glycerol and 12.5 µg/mL chloramphenicol.

2.3 Media and Solutions for Functional Screening

For libraries using the pEpiFOS-5 Fosmid Vector (EPICENTRE).

1. 1000× Cm stock solution: 12.5 mg/mL chloramphenicol (Cm) in ethanol, stored at -20 °C.

2. 5× Salts stock solution: 18 g/L Na₂HPO₄, 12H₂O, 3.31 g/L KH₂PO₄, 0.53 g/L NaCl, 2.11 g/L NH₄Cl, deionized water (dH₂O) up to 1 L. Autoclave.
3. 500× MgSO₄ stock solution: 1 M MgSO₄ in dH₂O. Autoclave.
4. 333× CaCl₂ stock solution: 0.01 M in dH₂O. Autoclave.
5. 1000× Salts stock solution: 15 g/L Na₂EDTA-2H₂O, 4.5 g/L ZnSO₄-7H₂O, 3 g/L CoCl₂-6H₂O, 1 g/L MnCl₂-4H₂O, 1 g/L H₃BO₃, 0.4 g/L Na₂MoO₄-2H₂O, 3 g/L FeSO₄-7H₂O, 0.3 g/L CuSO₄-5H₂O. Dissolve EDTA and ZnSO₄ in 800 mL of deionized water, adjust pH to 6.0 with HCl/NaOH. Dissolve the other compounds one by one and keep the pH at 6.0. Adjust pH to 4 and the volume to 1 L (*see Note 2*). The solutions are filter-sterilized.
6. 1000× Leucine: 40 g/L in dH₂O (*see Note 3*). Filter-sterilize.
7. 100× Thiamine hypochloride: 10 g/L in dH₂O. After dissolution, adjust pH to 2.0 with 2 N HCl. Filter sterilization and preservation at 4 °C hidden from light.
8. Luria-Bertani-Chloramphenicol (LB-Cm) medium: 10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, 1 mL/L 1000× Cm stock solution. Autoclave LB medium and let it cool at 50 °C before adding Cm stock solution. Prepare 200 mL LB-Cm medium for the underlay and 100 mL for the overlay for each large agar plate (QTray, 24.5 cm × 24.5 cm) (*see Note 4*). For solid medium, add 15 g/L agar.
9. Minimal (M9) medium: 15 g/L agar, 200 mL/L salts (5×) stock solution, 2 mL/L MgSO₄ (500×) stock solution, 3 mL/L CaCl₂ (333×) stock solution, 1 mL/L salts (1000×) stock solution, 10 mL/L thiamine hypochloride (100×) stock solution, 1 mL/L leucine (1000×) stock solution, 1 mL/L Cm (1000×) stock solution. For each QTray (24.5 cm × 24.5 cm plates), prepare 200 mL medium for the underlay and 100 mL for the overlay, 24.5 cm × 24.5 cm (*see Notes 5 and 6*).
10. pEpiFOS-5 library: Set of *Escherichia coli* EPI100 clones arrayed in 384-well plates, each well containing one copy of a fosmid clone in LB + glycerol 8 %.
11. QTrays (24.5 cm × 24.5 cm), sterile (Corning Incorporated).
12. LB + 8 % glycerol: Autoclave separately 500 mL 2× LB and 500 mL glycerol at 16 % (w/v) in deionized water, cool to room temperature, mix and add 1 mL/L Cm stock solution.
13. Azurine-Crosslinked (AZCL)/Azo substrates (used for identification of endo-acting CAZymes): Autoclave separately 500 mL 2× LB agar and 500 mL 2× AZCL/Azo-substrate in water. Cool to 60 °C, and mix the two preparations. The final screening medium contains 2 g/L of these chromogenic substrates.

14. 1000× 5-Bromo-4-chloro-3-indolyl (X-) substrates (used for identification of exo-acting CAZymes): 60 mg/mL in dimethyl sulfoxide (DMSO). The final screening medium contains 60 mg/L of X-substrates.
15. Glycerol stock solution: 30 % glycerol (w/v) in deionized water, autoclaved.
16. Omnitrays (86 cm × 128 cm), sterile (Thermo scientific Nunc).
17. Clear 96-well microplates, sterile (Corning Incorporated).
18. Cryotubes, 2 mL sterile (Thermo scientific Nunc).
19. Dinitrosalicylic acid (DNS) solution: 10 g/L DNS, 300 g/L potassium sodium tartrate, 16 g/L NaOH in dH₂O (*see Notes 7 and 8*).
20. 10× *para*-Nitrophenyl (*p*NP-) substrates stock solution: 10 mM in dH₂O (*see Note 9*).
21. AZCL/Azo-substrate stock solution for discrimination screening: 0.2 % (w/v) final concentration in deionized water.
22. Lysozyme stock solution (10×): 5 g/L in activity buffer. Store at -20 °C.
23. 50 mM Potassium phosphate buffer, pH 7.0.
24. 1 M Na₂CO₃ stock solution.
25. 5× Oligosaccharide stock solution: 5 % oligosaccharide (w/v) in deionized water.
26. 1 M NaOH stock solution: in water.
27. Eluent A: 150 mM NaOH.
28. Eluent B: 150 mM NaOH, 500 mM CH₃COONa.

2.4 Equipments

1. Eppendorf centrifuge 5810 R, with swing-bucket rotor A-4-81, fixed angle-rotor F-34-6-38 (+ adaptors for 50 and 15 mL Falcon tubes), and a rotor for Eppendorf tube.
2. Pulsed-field CHEFDRII electrophoresis system (Bio-Rad).
3. Liquid handling automat operating in sterile conditions (Biomek 2000, Beckman, Fullerton, CA).
4. Pump PM600 Jouan.
5. Colony picker QPixII (Genetix, Hampshire, UK). Colony picking and microplate replicating.
6. Microtiterplate shaker incubator (Multitron, Infors, Massy, France).
7. Automated microplate gridder (K2, KBiosystem, Basildon, UK): replication of microtiterplate organized libraries (384 well plates) on solid agar plates (QTrays).
8. Bioblock Scientific Vibra-Cell 72412 ultrasonic processor.

9. Microplate spectrophotometer (e.g., Sunrise, TECAN, Männedorf, Switzerland).
10. Dionex ICS-3000 system (Dionex Corp., Sunnyvale, CA).
11. CarboPac PA100 46250 column and guard column (Dionex).

3 Methods

3.1 DNA Sampling

3.1.1 Soil Sampling

1. Collect soil core samples of 6 cm in diameter from surface soils (0–20 cm) by using geostatistical methods as described for example by Atteia [10] on a grid of 6.20 × 3.20 m (*see Note 10*).
2. Transfer soil cores as soon as possible to the laboratory in plastic bags.
3. Sieve soil at 2 mm and store it at 4 °C until rapid processing (within a week).

3.1.2 Bacterial Cells Recovery

1. Refrigerate HMP, NaCl and Nycodenz solutions at 4 °C and perform following procedures on ice unless otherwise specified. Mix the equivalent of 50 g of dry soil with 180 mL of HMP and about 20 glass beads and stir strongly (CATSSO stirrer, set to position 1/min) for 2 h at 22 °C.
2. Centrifuge in a swing rotor (Eppendorf A-4-81) at 18 × *g* for 1 min at 10 °C to eliminate coarse particles.
3. Filter supernatant on sterile gaze into a new 250-mL Nalgene tube and centrifuge in a swing rotor at 3,220 × *g* for 20 min at 10 °C.
4. Eliminate supernatant and suspend pellet in 35 mL NaCl (*see Note 11*).
5. Fill two 50-mL falcon tubes with 11 mL Nycodenz solution and carefully add on surface half of the soil suspension in both tubes (*see Note 12*) [11].
6. Centrifuge in a swing rotor at 3,220 × *g* for 40 min at 10 °C without acceleration and deceleration (set to 0).
7. Pipette the white bacterial ring (approximately 4 mL) without disturbing Nycodenz gradient at the interface between Nycodenz and NaCl (*see Note 13*), pool both rings in a single 50-mL falcon tube and fill with NaCl up to 40 mL.
8. Centrifuge the falcon tube with a fixed-angle rotor at 9,000 × *g* for 20 min at 10 °C (*see Note 14*).
9. Eliminate supernatant, wash pellet with 10 mL NaCl and transfer to a 15-mL Falcon tube.
10. Centrifuge with a fixed-angle rotor at 9,000 × *g* for 15 min at 10 °C.

11. Eliminate supernatant, wash pellet with 1 mL NaCl and transfer to a 1.5-mL Eppendorf tube.
12. Centrifuge at $13,000 \times g$ for 5 min at 10 °C.
13. Eliminate supernatant and suspend pellet in 50 μ L TE buffer.

3.1.3 High Molecular Weight DNA Extraction

1. Mix bacterial pellets with an equal volume of molten 1.6 % InCert[®] agarose (*see Note 15*) [12], transfer into disposable plug mold.
2. Let it stand at 4 °C until solidification, then unmold the solidified cell suspension and transfer into a 50-mL Falcon tube.
3. Add 45 mL of LA solution and incubate at 37 °C for 6 h.
4. Transfer the plug into a new 50-mL Falcon tube, add 45 mL of LBB solution and incubate at 55 °C for 24 h.
5. Repeat the operation: transfer plug in 45 mL of fresh LBB solution into a new 50-mL Falcon tube and incubate at 55 °C for 24 h.
6. Wash plug in 10 mL of storage buffer containing PMSF for 2×1 h at 50 °C (*see Note 16*).
7. Dialyze the plug in three successive 10 mL storage buffer baths for 8 h and store at 4 °C until use.

3.2 Fosmid Library Construction

1. Prepare 150 mL of 0.8 % low-melting-temperature agarose gel (*see Note 17*) in $1 \times$ TAE buffer (*see Note 18*) and wait for solidification.
2. Transfer high-molecular-weight bacterial DNA trapped in the agarose plug using a sterile pipette tip into wells of solidified gel, place in the pulsed-field electrophoresis system, load PFGE ladder (*see Note 19*) and fill it with $1 \times$ TAE buffer.
3. Migrate for 18 h at 4.5 V/cm with 5–40-s pulse times in $1 \times$ TAE buffer cooled at 12 °C for the whole migration time.
4. After electrophoresis, stain the gel in a solution of ethidium bromide for 30 min at room temperature.
5. Cut DNA fragments between 30 and 50 kbp and recover DNA with GELase following manufacturer's procedure.
6. Clone the extracted metagenomic DNA into fosmid and transform in the *E. coli* strain EPI100 as recommended by the manufacturer.
7. Using a colony picker, select transformants grown on plate supplemented with 12.5 μ g/mL chloramphenicol and transfer them to 384 multiwell plates containing 70 μ L freezing medium per well and incubate at 37 °C for 22 h (*see Note 20*).
8. Duplicate the library and store in two -80 °C different freezers for safety reasons.

3.3 Replication of the Metagenomic Library Prior to Functional Screening

1. The day before replication, place the metagenomic library at 4 °C to allow gentle thawing from -80 °C storage.
2. New 384-well microtiter plates are filled with LB + 8 % Glycerol solution, 70 µL per well, using an automated liquid handling station.
3. The metagenomic library is replicated, using a QPixII colony picker (3 h for 54 microplates, totalizing 20,736 clones).
4. The mother plates are stored back at -80 °C, covered with microplate aluminum sealing tapes. The copy plates are incubated overnight (about 16 h) at 37 °C, covered by porous adhesive membranes.

3.4 High-Throughput Primary Screening

3.4.1 QTray Preparation

1. Sterilize the autoclavable tubing of the PM600 Jouan pump, as well as deionized water to wash the tubing between two different media distribution. The quantity of water depends on the number of substrates (*see Note 21*).
2. Calibrate the pump using sterile water.
3. Pour 200 mL of underlay for each substrate using the pump (*see Note 22*). For chromogenic substrates, use LB medium; for selective growth, use M9 medium. Leave them to dry under the hood, lid off, for ~30 min.
4. Pour 100 mL of overlay medium for each substrate (*see Note 22*). For chromogenic substrates, use LB medium containing 2 g/L of AZCL/Azo-substrate, or 60 mg/L of X-substrates; for selective growth, use M9 medium containing a final concentration of 0.5 % (w/v) oligosaccharidic carbon source. Leave them to dry, stacked under the hood, lid off, for ~30 min (*see Notes 23 and 24*).
5. Until the day of the gridding, stock the plates at 4 °C, upside down.

3.4.2 Gridding

1. When storing the plates at 4 °C, place them at room temperature the day before the gridding.
2. Microtiter plates are gridded on large LB agar plates, using a K2 automated plate replication system. One QTray can accommodate the clones from six 384-well plates, for a total number of 2304 fosmid clones per plate. In 7 h, 54 microtiter plates can be gridded on 12 different substrates.
3. Plates containing arrayed clones are incubated at 37 °C.

3.4.3 Hit Isolation, Selection, and Validation

1. Positive clones are recognized: (1) for AZCL-substrates, thanks to the blue-colored halo formed around the colonies (Fig. 2a.a); (2) on Azo-substrates by the appearance of a discoloration halo around positive clones (Fig. 2a.b); (3) as blue-colored colonies on X-substrates (Fig. 2a.c); (4) as the sole growing colonies on M9 media supplemented with targeted carbon source (*see Note 25*) (Fig. 2a.d).

2. Positive clones are picked from the QTray and streaked on Petri dishes containing LB and chloramphenicol, and grown overnight at 37 °C. For each selected clone, three isolated colonies are selected to inoculate three adjacent wells of a 96 microtiter plate, filled with 200 µL LB Cm (*see Note 26*). The plate is incubated at 37 °C with 200 rpm shaking (shaking throw 25 mm) for ~16 h.
3. This microplate is then gridded on omnitrays containing the same medium used for the primary screening, and incubated at 37 °C, until the awaited phenotype is observed.
4. Colonies from validated wells are streaked on fresh LB Cm plate, and after colony growth, two isolated colonies are picked to inoculate two 3 mL of liquid LB medium. Cells are incubated overnight at 37 °C, under shaking at 200 rpm (shaking throw 25 mm).
5. After overnight growth, 500 µL of culture are mixed into two cryotubes with 500 µL of glycerol stock solution (30 %). The two copies of each hit clones are stored at -80 °C in different freezers for safety concern.

3.5 Discrimination Screening of Validated Hits

3.5.1 Liquid Assays Using AZCL- Polysaccharides and pNP-Sugars

1. From an isolated colony, inoculate a 20 mL culture in liquid LB medium, and cultivate at 37 °C with shaking at 200 rpm overnight (shaking throw 25 mm) (*see Note 27*).
2. Measure the OD at 600 nm with a spectrophotometer.
3. Centrifuge the culture at $12,857 \times g$, for 5 min at 4 °C. Discard supernatant.
4. Suspend the pellet in activity buffer to obtain a final OD at 600 nm of 80 (*see Note 28*).
5. To break the cells, use the sonication method: with the probe at 30 % of the maximal power, do five cycles of 20 s separated by 4 min in ice.
6. Centrifuge the samples at $21,728 \times g$ for 10 min.
7. Filter the supernatant with a 0.2 µm filter (Minisart). The solution obtained is called enzymatic extract from now on.
8. To test the activity of enzymatic extracts on AZCL-substrates, mix in hemolysis tubes 500 µL of enzymatic extract, 100 µL of AZCL-substrate solution (0.2 % (w/v) final concentration), and 400 µL of activity buffer. Incubate at 37 °C with regular shaking. After reaction times of 0, 15, 30 min, 1 h and 24 h, transfer 120 µL of reaction in an Eppendorf tube, centrifuge for 1 min at $21,728 \times g$, transfer 100 µL of supernatant into the well of a polystyrene microplate and read the OD at 590 nm, with a plate reader. Positive hits present an increase of the OD over time (Fig. 2b.a).

9. To test the activity of enzymatic extracts on *p*NP-substrates, mix in hemolysis tubes 100 μ L of 10 mM *p*NP-substrate solution, 200 μ L of enzymatic extract, and 200 μ L of activity buffer (the same as suspension buffer). Incubate at 37 °C. After reaction time of 0, 10 and 30 min, mix 50 μ L of the reaction medium with 250 μ L of 1 M Na₂CO₃. Transfer 200 μ L of this medium into another polystyrene microplate and read the OD at 405 nm, with a plate reader. Positive hits present an increase of the OD over time (Fig. 2b.b).

3.5.2 HPAEC-PAD
Analysis to Analyze
Reaction Products
of Oligosaccharide
Degradation [13]

1. Clones are grown at 37 °C in 5 mL LB Cm medium, with shaking at 120 rpm for 24 h (shaking throw 25 mm).
2. Centrifuge the culture for 5 min at 3,214 $\times g$.
3. Resuspend in 1 mL activity buffer, containing 0.5 g/L of lysozyme. Incubate at 37 °C for an hour. Complete cell lysis with a freeze (–80 °C) and thaw (30 °C) cycle.
4. Centrifuge cell debris at 21,728 $\times g$ for 10 min and filter the cytoplasmic extracts with a 0.2 μ m filter (Minisart).
5. Enzymatic reaction medium contains 0.2 mL of the oligosaccharide stock solution and 0.8 mL of cytoplasmic extract. Incubate at 37 °C for 24 h.
6. After 30 s and 24 h of reaction, take a 100 μ L sample out of the reaction medium, and heat at 90 °C for 5 min to deactivate enzymes.
7. Dilute samples 200 times with ultrapure water.
8. Perform HPAEC-PAD analyses on a Dionex ICS-3000 system, equipped with a CarboPac PA100 4 \times 250 column connected to the corresponding guard column. Oligosaccharides are separated at 30 °C, with a flow rate of 1 mL/min with a multistep gradient: 0–30 min (0–60 % B), 30–32 min (60–90 % B), 32–36 min (90–0 % B), and 36–46 min (0 % B). Samples of monosaccharides and oligosaccharides at 5, 10, 15, and 20 mg/L are used as standards (Fig. 2b.c). One unit of activity is defined as the amount of enzyme releasing 1 μ mol of product per minute.

4 Notes

1. The main strategy for prokaryote cell segregation consists in applying a density gradient through centrifugation. Because of their size and density, bacteria will cluster apart from eukaryote cells into a specific fraction of the gradient which can easily be recovered. However, co-extraction of low-density eukaryote cells such as fungi spores and pico-eukaryotes (*see* ref. 14) is a possible source of contamination.

2. The solution is green and becomes dark yellow during preservation at 4 °C.
3. To help leucine dissolution, add 5 M NaOH.
4. Not only the system of overlay/underlay enable the use of less substrate, hence decreasing costs, but it is very important in the case of insoluble substrate such as AZCL-polysaccharide because they naturally sediment at the bottom of the plate during solidification of the agar medium. Pouring 200 or 300 mL of such medium leads to the accumulation of all the substrate far from the surface colonies, and functional enzymes liberated when *E. coli* cells die are located too far away from the substrate. A top layer of 100 mL AZCL-substrate medium brings the insoluble substrate much closer to the recombinant clone.
5. Water and agar are sterilized alone. Other sterile components are added one by one.
6. To avoid solidification of M9 medium before time, warm the bottle of 5× salts stock solution before mixing solutions.
7. Shake overnight at room temperature for better solubilization, inside a volumetric flask.
8. Cover the flask entirely with aluminum paper: the DNS solution is light-sensitive.
9. Some *p*NP substrates are difficult to solubilize. You might need as long as a night for them to be totally dissolved.
10. For a large field, a nested sampling of the kind devised and elaborated by Oliver and Webster (*see* ref. 15) may ensure that the important variation is captured.
11. Depending on soil type, pellet can be difficult to suspend. However, soil solution imperatively needs to be homogeneous without any fragment. Using an ultrasonic bath can help.
12. Nycodenz solution must not be disturbed: soil solution has to lay on surface. To avoid disturbance, use the gravity function of pipet-aid to add soil suspension.
13. Bacterial rings are sometimes difficult to see. Soil pellet usually fill the tube up to the 10-mL graduation, above the Nycodenz solution reach the 15-mL graduation. Bacterial cells are usually near the 15-mL graduation on Nycodenz surface and above cells is the NaCl solution.
14. Do not forget to readjust the centrifuge parameters to maximal acceleration and deceleration values.
15. Warm briefly the cell suspension at 37 °C to avoid premature gelation of agarose.
16. This step can also be performed overnight at room temperature (22 °C) with gentle shaking.
17. Use caution when handling low-melting agarose gel because it is very fragile.

18. TAE buffer is recommended for subsequent enzymatic reactions.
19. Use embedded ladder supplied in a GelSyringe dispenser.
20. This very high-throughput work can benefit from dedicated facilities, gathering all the automats useful for such experimentation: colony picking, microplate replication, liquid transfer, etc. The functional screening described in this chapter has been performed mainly using the ICEO facility (LISBP, INSA Toulouse) dedicated to enzyme screening and discovery, and part of the Integrated Screening Platform of Toulouse (PICT, IBiSA).
21. If you do not have a pump, use a measuring tube to pour the right volume of agar medium.
22. Make sure that the plate is fully horizontal to have an even overlay. Be careful not to pour the overlay on cold agar, as it will solidify too quickly and impedes the obtention of a regular layer. To overcome this, incubate your QTrays containing the solidified underlay at 37 °C for an hour, just before pouring the overlay.
23. Do not let the medium inside the pipe solidify: if you need to, make a closed system within the bottle containing the medium.
24. If you have some medium left, use it to pour omnitrays (media with substrates) and Petri dishes (LB agar) that will be used for validation.
25. Positive clones on Azo, AZCL and X-substrates are usually observed rapidly, between 2 and 7 days of incubation. Growth of the hits obtained by positive selection on oligosaccharides as sole carbon sources is visualized between 5 and 20 days.
26. It is easier to streak the Petri dishes and wait to pick the isolated colonies for a large group of positive clones, so that you can fill the microplates all at once and arrange them as you wish.
27. You can also make a pre-culture the day before in 3 mL of liquid LB medium from a freeze-d sample.
28. An OD at 600 nm of 80 corresponds to the most efficient concentration for sonication.

Acknowledgements

This research was funded by the European Union project MetaExplore, the French Research Agency (Agence Nationale de la Recherche) ANR project Metasoil, and the INRA metaprogramme M2E (project Metascreen).

References

1. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42:D490–D495
2. André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M (2014) CAZyme discovery and design for sweet dreams. *Curr Opin Chem Biol* 19:17–24
3. Li L-L, McCorkle SR, Monchy S, Taghavi S, van der Lelie D (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels* 2:10
4. Ferrer M, Golyshina OV, Chernikova TN, Khachane AN, Martins Dos Santos VAP, Yakimov MM, Timmis KN, Golyshin PN (2005) Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem Biol* 12:895–904
5. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL, Coutinho PM, Henrissat B, Leclerc M, Doré J, Monsan P, Remaud-Simeon M, Potocki-Veronese G (2010) Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res* 20:1605–1612
6. Ladevèze S, Tarquis L, Cecchini DA, Bercovici J, André I, Topham CM, Morel S, Laville E, Monsan P, Lombard V, Henrissat B, Potocki-Véronèse G (2013) Role of glycoside phosphorylases in mannose foraging by human gut bacteria. *J Biol Chem* 288:32370–32383
7. Ekkers DM, Cretoiu MS, Kielak AM, van Elsland JD (2012) The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Appl Microbiol Biotechnol* 93:1005–1020
8. Taupp M, Mewis K, Hallam SJ (2011) The art and design of functional metagenomic screens. *Curr Opin Biotechnol* 22:465–472
9. Bastien G, Arnal G, Bozonnet S, Laguerre S, Ferreira F, Fauré R, Henrissat B, Lefèvre F, Robe P, Bouchez O, Noirot C, Dumon C, O'Donohue M (2013) Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics. *Biotechnol Biofuels* 6:78
10. Atteia O, Dubois JP, Webster R (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environ Pollut (Barking Essex)* 86:315–327
11. Courtois S, Frostegård A, Göransson P, Depret G, Jeannin P, Simonet P (2001) Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ Microbiol* 3:431–439
12. Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K, Bertrand H, Vogel TM, Perriere G, Simonet P, Nalin R (2004) Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl Environ Microbiol* 70:5522–5527
13. Cecchini DA, Laville E, Laguerre S, Robe P, Leclerc M, Doré J, Henrissat B, Remaud-Siméon M, Monsan P, Potocki-Véronèse G (2013) Functional metagenomics reveals novel pathways of prebiotic breakdown by Human gut bacteria. *PLoS One* 8:e72766
14. Moreira D, López-García P (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol* 10:31–38
15. Oliver MA, Webster R (2010) Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geogr Anal* 18:227–242

Metatranscriptomics of Soil Eukaryotic Communities

Rajiv K. Yadav, Claudia Bragalini, Laurence Fraissinet-Tachet, Roland Marmeisse, and Patricia Luis

Abstract

Functions expressed by eukaryotic organisms in soil can be specifically studied by analyzing the pool of eukaryotic-specific polyadenylated mRNA directly extracted from environmental samples. In this chapter, we describe two alternative protocols for the extraction of high-quality RNA from soil samples. Total soil RNA or mRNA can be converted to cDNA for direct high-throughput sequencing. Polyadenylated mRNA-derived full-length cDNAs can also be cloned in expression plasmid vectors to constitute soil cDNA libraries, which can be subsequently screened for functional gene categories. Alternatively, the diversity of specific gene families can also be explored following cDNA sequence capture using exploratory oligonucleotide probes.

Key words Metatranscriptomics, Environmental RNA, cDNA synthesis, cDNA size fractionation, cDNA libraries, Sequence capture

1 Introduction

Molecular investigations on the taxonomic and functional diversity of microbial communities have initially focused on environment-extracted DNA [1]. Given the high diversity of microbial communities often dominated by uncultured species, metagenomic DNA also represents a large reservoir for new genes of potential interest in biotechnology [2, 3]. While the analysis of metagenomic DNA gives information on the microorganisms present in the environment as well as of the functions that they can potentially express, it cannot however be used as a proxy to infer actual microbial activities. They can instead be appreciated through the analysis of the more labile environmental RNA molecules, whose diversity and abundance reflect both the diversity and the transcription levels of expressed genes.

Metatranscriptomic RNA encompasses the transcriptomes of the different organisms present in the original environmental sample. As such it is dominated by ribosomal RNA (rRNA) which is

of little value to infer microbial activities as opposed to messenger RNA (mRNA). It is therefore desirable to obtain metatranscriptomic RNA fractions enriched in mRNA. This can be achieved by depleting environmental RNA from rRNA by subtractive hybridization capture of the latter molecules [4, 5]. rRNA subtraction enriches in mRNA from all organisms, bacteria, archaea as well as eukarya. Alternatively, eukaryotic mRNA can be selectively isolated thanks to their specific 3' poly-adenosine tail (poly-A mRNA) [6, 7]. This latter approach represents an elegant way to specifically study the activities expressed by eukarya in the environment. In soils, the eukaryotic biomass is generally dominated by fungi which play an essential role in plant organic matter degradation, a key step of the global terrestrial carbon cycle, as well as in the delivery of key soil nutrients to symbiotically associated plant roots. Soils also host numerous parasitic and/or free living small animals, constituting the so-called mesofauna, as well as phylogenetically diverse, mostly phagotrophic, unicellular species ("protists") which regulate bacterial biomass. Soil is however a complex matrix containing, in variable amounts, clay minerals and humified organic matter which can interfere with RNA extraction and purification. As a consequence, RNA extraction protocols often need to be adapted to each soil and the quantities of extracted RNA are often low (in the range of 10 ng–1 µg/g of soil).

This chapter presents protocols to extract total soil RNA and to work with the poly-A mRNA fraction (Fig. 1). Although the systematic sequencing and functional/taxonomic annotation of cDNAs derived from poly-A mRNA are susceptible to give

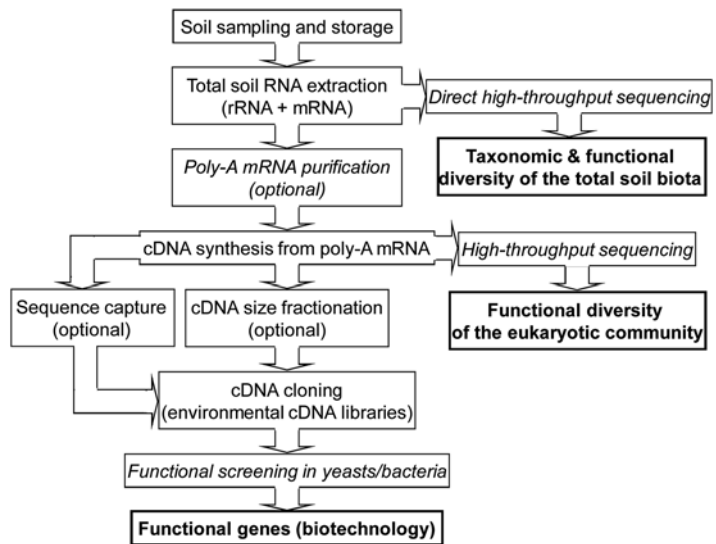


Fig. 1 Flow chart of the metatranscriptomic approach dedicated to the study of environmental eukaryotic mRNA. *Text in bold characters* refers to the objectives of the studies. *Protocols in italics* are not described in the present chapter

essential information on the functional diversity of eukaryotic communities [8], the cloning and functional characterization of full-length cDNAs are also of major interest to identify potentially novel and unsuspected processes carried out by soil eukarya [9, 10] as well as to use eukaryotic microbial communities as a source of novel biocatalysts in biotechnology [11, 12]. In this respect we also present optimized protocols for the generation and cloning of full-length cDNAs as well as the selection of specific gene families through subtractive hybridization capture on cDNAs [13].

2 Materials

2.1 Soil Sampling, Processing, and Storage

1. Soil sampling strategies need to be adapted to the scope and hypotheses to be tested in individual studies and therefore, only some general guidelines can be given in this chapter. The main factors to be taken into consideration regard sampling time and the number, spacing, and volume of soil samples. With respect to sampling time, soil microbial activity, and therefore metatranscriptomic RNA diversity, is highly responsive to environmental parameters, such as plant cover, temperature, and water content. Concerning sample numbers, soil microbial communities and more specifically fungal ones are spatially structured and therefore a single soil core, of a few centimeters in diameter, will only capture a small fraction of the taxonomic and functional diversity of the corresponding microbial community. Finally, most non-agricultural soils are vertically structured in discrete soil horizons colonized by often taxonomically and functionally distinct microbial species. As a consequence, when fine scale spatial structure does not represent an issue, it may be advisable to collect multiple soil cores, regularly distributed in a plot of interest, across different horizons which can be mixed together to constitute a composite soil sample.
2. As patterns of gene expression quickly change over time, it is advisable to process the soil samples as soon as possible (e.g. within 0–6 h) after sampling. Processing can be limited to sieving (e.g. by using a 2 mm-mesh sieve) to remove most plant roots, coarse plant debris and stones as well as the macro-fauna. Aliquots of soils are then quickly frozen in either dry ice or liquid nitrogen and stored at $-70/80$ °C. We have successfully extracted seemingly undegraded RNA from forest soil samples stored frozen for more than 5 years.

2.2 Equipment

1. A bead-beater instrument accepting 2 mL tubes.
2. Refrigerated microcentrifuge with rotors for 1.5 and 2 mL tubes.
3. At least two water baths or heating blocks.

4. A chemical hood for phenol and chloroform manipulation.
5. A thermal cycler to perform PCRs.
6. Two identical gel trays and a power supply for agarose gel electrophoresis.
7. A “blue light” transilluminator (*see Note 1*).
8. A spectrophotometer (e.g. NanoDrop™ from Thermo Scientific) and/or a spectrofluorimeter (e.g. Qubit® from Life Technologies) allowing measurements of nucleic acid concentrations using volumes in the μL range.
9. A capillary electrophoresis system (e.g. Bioanalyzer from Agilent Technologies) for RNA/DNA quantification and quality control.

2.3 Individual Chemicals and Molecular Biology Products

1. Chemicals of the highest grade must be used (the purchase of so-called “RNase-tested” or “RNase-free” chemicals is however usually not necessary): sodium dodecyl sulfate (SDS), LiCl, Na acetate, Tris-HCl, Na_2EDTA , NaCl, ethanol, isopropanol, isoamyl alcohol, chloroform, water-saturated phenol (at acidic pH or adjusted at pH 8.0, must be stored at 4 °C in the dark), beta-mercaptoethanol, orange G, xylene cyanol FF, ethidium bromide, diatomaceous earth (e.g. from Sigma Chemical company), acid-washed glass beads (106 μm in diameter from Sigma).
2. Molecular biology products include: agarose of molecular biology grade, low melting point agarose, yeast tRNA (e.g. 10 mg/mL from Ambion), *Sfi*I endonuclease (which recognizes and cuts the degenerate GGCCNNNNNGGCC restriction sites), T4 DNA ligase (and its buffer containing ATP), RNase-free DNase I.

2.4 Commercial Kits

It is advisable to use kits for:

1. For cDNA synthesis by the Reverse Transcriptase template switching protocol [14] and amplification, we used components of the Mint-2 kit (Evrogen, Russia).
2. RNA purification (e.g. Nucleobond RNA/DNA 90 kit from Macherey-Nagel).
3. Purification of PCR products (e.g. Qiaquick PCR purification kit from Qiagen).
4. Extraction of DNA fragments from agarose gels (e.g. QIAEXII kit from Qiagen).
5. Large-scale (maxipreparation) of plasmids.
6. Quantification of RNA by fluorimetry (Qubit RNA assay kit from Life Technologies).
7. Separation of DNA by capillary electrophoresis (cDNA Agilent 2100 Bioanalyzer DNA 12000 chip).

2.5 RNA Extraction and Manipulation

RNA extraction from soil can be performed using commercial kits, which can sometimes fail to give high-quality RNA. Alternatively, the two RNA extraction protocols described below (*see* Subheadings 3.1 and 3.2) can be used.

1. When working with RNA it is advisable to work on a dedicated bench with dedicated pipette sets and labware.
2. Filter pipette tips should be systemically used.
3. Gloves must be worn all time.
4. Glassware can be made free of RNase by baking for 2 h at 160 °C.
5. Reusable plasticware must be thoroughly washed with detergents, abundantly rinsed with deionized water and sterilized water before being autoclaved twice at 120 °C for 20 min.
6. Jars of microtubes (usual microbiology grade) must be filled with gloved hands and autoclaved twice at 120 °C for 20 min.
7. To prepare RNase-free water, pour deionized or ultrapure water in baked glass bottles and immediately sterilize twice by autoclaving (*see* Note 2).
8. All aqueous solutions or suspensions (20 % w/v SDS, 3 % diatomaceous earth in water, 4 M LiCl, 3 M Na acetate pH 4.8 or 5.2, Tris Borate EDTA (TBE) electrophoresis buffer) must be prepared in sterile RNase-free water and sterilized twice at 120 °C for 20 min.
9. pH measurements performed with thoroughly cleaned electrodes.
10. The denaturing and lysis solutions used in **step 2** in Subheading 3.3 contain (per L): denaturing solution: 472.64 g of guanidine thiocyanate, 1.21 g of Tris-HCl, 0.37 g of Na₂EDTA, pH 8.0; lysis solution: 12.44 g of Tris-HCl, 7.44 g of Na₂EDTA, 5.84 g of NaCl, 20 g of SDS, pH 9.0.

3 Methods

3.1 RNA Extraction Protocol 1

1. This protocol was originally described in [15]. Ten different soil samples can be extracted contemporaneously. To each 2 mL RNase-free screw-cap tube add 0.5 g of acid-washed glass beads (106 µm in diameter), 0.4 g of frozen soil and 350 µL of RNase-free water. Vortex mix to homogenize and incubate immediately for 1 h at -80 °C (*see* Note 3).
2. To the still frozen tubes add in the following order: 34 µL of 20 % SDS, 167 µL of homogenized 3 % diatomaceous earth and 583 µL of water-saturated phenol at pH 8.0.
3. Mix by shaking for 3 min at 1600 beats/min or 2.5 min at 2000 beats/min at room temperature using a bead-beater.

Centrifuge for 15 min at $18,000\times g$ and $4\text{ }^{\circ}\text{C}$ and transfer the aqueous upper-phase to a 1.5 mL RNase-free tube.

4. To each tube add $37\text{ }\mu\text{L}$ of 3 M Na acetate pH 5.2 and $478\text{ }\mu\text{L}$ of pure cold ($-20\text{ }^{\circ}\text{C}$) ethanol. Mix and incubate overnight at $-20\text{ }^{\circ}\text{C}$. Centrifuge for 15–25 min at $18,000\times g$ and $4\text{ }^{\circ}\text{C}$. Discard the supernatant with care and add to the pellet (usually visible) $100\text{ }\mu\text{L}$ of cold ($-20\text{ }^{\circ}\text{C}$) 70 % ethanol. Mix swiftly and centrifuge for 15–25 min at $18,000\times g$ and $4\text{ }^{\circ}\text{C}$. Remove the ethanol without dislodging the pellet and let it dry at room temperature for 10 min.
5. Resuspend each pellet in $20\text{ }\mu\text{L}$ of deionized RNase-free water and pool them by 2 ($2\times 20\text{ }\mu\text{L}=40\text{ }\mu\text{L}$). Selectively precipitate the RNA by adding to each tube, $65\text{ }\mu\text{L}$ of 4 M LiCl and mix gently by pipetting up and down (not by vortex mixing). Incubate overnight at $4\text{ }^{\circ}\text{C}$ and then centrifuge for 15–25 min at $18,000\times g$ and $4\text{ }^{\circ}\text{C}$.
6. Carefully remove the supernatant which contains the DNA and which can be kept for further purification (not described). Resuspend each RNA pellet in $30\text{ }\mu\text{L}$ of RNase-free water and immediately proceed to the DNase treatment by adding $4\text{ }\mu\text{L}$ of a $10\times$ concentrated DNase buffer and $6\text{ }\mu\text{L}$ of 1 U/ μL RNase-free DNase I. Incubate for 1 h 30 min at $37\text{ }^{\circ}\text{C}$.
7. Stop the reaction by adding $40\text{ }\mu\text{L}$ of cold ($-20\text{ }^{\circ}\text{C}$) isopropanol and incubate for 2 h at $4\text{ }^{\circ}\text{C}$. Centrifuge for 15–25 min at $18,000\times g$ and $4\text{ }^{\circ}\text{C}$ and remove the isopropanol. Dry the pellets at room temperature for 10 min and resuspend them in $10\text{ }\mu\text{L}$ of RNase-free water (*see Note 4*).
8. If all samples come from the same soil, pool all RNA extracts and proceed to the purification with the “Nucleobond RNA/DNA 80” kit according to the manufacturer’s instructions.
9. The cheapest way of controlling RNA quality is to run 5–10 μL of the extracted RNA on a RNase-free 1 % agarose gel and to visualize it under UV light after ethidium-bromide staining. Undegraded RNA is characterized by two sharp bands representing the large (LSU) and small (SSU) rRNA molecules (Fig. 2). Quality control can also be performed by running samples on an Agilent chip which also allows quantification of the RNA (*see Note 5*). Alternatively, RNA quantification can be performed by measuring absorbance at 260 nm or by fluorimetry (using the Qubit RNA Assay Kit) (*see Note 6*).

3.2 RNA Extraction, Alternative Protocol 2

1. This protocol was originally described in [7] and [16]. Ten different soil samples can be extracted contemporaneously. To each 2 mL RNase-free screw-cap tube quickly add in the following order: 0.5 g of $106\text{ }\mu\text{m}$ in diameter glass beads, 0.65 g of frozen soil, $950\text{ }\mu\text{L}$ of a mix of so-called denaturing and lysis

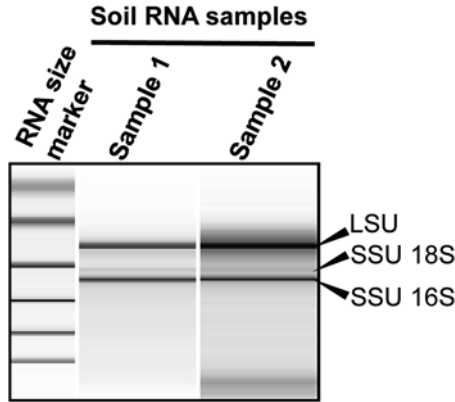


Fig. 2 Electrophoretic separation of soil total RNA. RNA was size fractionated on an Agilent Bioanalyser using a RNA 6000 nano kit. Soil rRNA small subunits (SSU) usually form two discrete bands: the smallest most intense one presumably of bacterial origin (16S), the largest and faintest one presumably of eukaryotic origin (18S). LSU large rRNA subunit (of both bacterial and eukaryotic origin)

solutions (*see item 10* in Subheading 2.5) (*see Note 7*), 50 μL of beta-mercaptoethanol and 4 μL of 10 mg/mL yeast tRNA. Proceed swiftly as to make sure that the soil samples do not thaw before mixing.

2. Mix by shaking 5 min at 1600 beats/min and room temperature using a multi-tube shaker and centrifuge for 5 min at $18,000\times g$ and 4 $^{\circ}\text{C}$. Transfer the aqueous upper-phase into a 2 mL RNase-free tube.
3. Add 1 mL of water-saturated 25:24:1 (v:v:v) phenol:chloroform:isoamyl alcohol solution and vortex mix at room temperature for 1 min. Centrifuge for 10 min at $18,000\times g$ and 4 $^{\circ}\text{C}$. Transfer the upper aqueous-phase into a 1.5 mL RNase-free tube. Add 500 μL of a 24:1 (v:v) chloroform:isoamyl alcohol solution and vortex mix at room temperature for 1 min. Centrifuge for 10 min at $18,000\times g$ and 4 $^{\circ}\text{C}$. Transfer the upper aqueous phase into a new 1.5 mL RNase-free tube and add 0.1 volume of 3 M Na acetate pH 5.2 and 2.5 volume of cold (-20°C) pure ethanol.
4. Incubate for 4 h at -80°C and centrifuge for 15–25 min at $18,000\times g$ and 4 $^{\circ}\text{C}$. Remove the supernatant and dry the pellet for 10 min at room temperature.
5. Resuspend each pellet in 20 μL RNase-free water and proceed according to **steps 5–9** in Subheading 3.1 (*see Note 5*).

3.3 Synthesis of First-Strand Eukaryotic cDNAs from Total Soil RNA

1. The protocol described here uses the components of the Mint-2 cDNA synthesis kit to synthesize eukaryotic cDNA starting from their 3' poly-A tails and to introduce *Sfi*IA (GGCCATTACGGCC) and *Sfi*IB (GGCCGCCTCGGCC) restriction sites at their 5' and 3' ends respectively [17].

2. Place 3 μg of total soil RNA (1 $\mu\text{g}/\mu\text{L}$) in a 0.2 mL microtube. Heat at 65 °C for 2 min in a thermal cycler with the heating lid on. Sequentially add 1 μL of 10 μM CDS-4M primer and 1 μL of 10 μM PlugOligo-3M adapter (*see Note 8*). Gently mix the components in the tube and centrifuge briefly. Incubate in a thermal cycler at 70 °C (heating lid on) for 2 min and then at 42 °C for 10 min.
3. Add 5 μL of reverse transcription master mix containing 2 μL of 5 \times first-strand buffer, 1 μL of 20 mM DTT, 1 μL of 10 mM dNTPs, 1 μL of Mint reverse transcriptase, and 0.5 μL of 20 U/ μL RNase inhibitor. Mix the content of the tube by pipetting up and down, centrifuge and immediately place the tube back in the thermal cycler. Incubate the tube at 42 °C for 30 min.
4. Add 5 μL of IP-solution and mix by pipetting (*see Note 9*). During the addition, do not remove the tube from thermal cycler. Incubate the tube at 42 °C for 1.5 h. Stop the reverse transcription by placing the tube on ice.

3.4 Double-Stranded cDNA Synthesis and Initial Amplification

1. Sequentially add to a 0.2 mL PCR tube, 36 μL of sterile RNase-free water, 5 μL of 10 \times Encyclo buffer (from the Mint-2 cDNA synthesis kit), 1 μL of 10 mM dNTP mixture, 2 μL of 10 μM PCR primer M1 (*see Note 10*), 5 μL (or even less) of first-strand cDNA (i.e. only 1/3 of the synthesized first strand cDNAs), and 1 μL of 50 \times Encyclo DNA polymerase mix. Mix the components in the tube by gently pipetting up and down.
2. Place the tube in a thermal cycler and apply a PCR amplification comprising an initial denaturation at 95 °C for 1 min, three cycles of 95 °C for 15 s, 66 °C for 20 s, and 72 °C for 3 min (*see Note 11*).

3.5 cDNA Size Fractionation by Two-Dimensional Agarose Gel Electrophoresis

1. This protocol was originally described by Wellenreuther et al. [18]. Melt the 0.7 % (w:v) agarose in half-strength (0.5 \times) standard Tris Borate EDTA (TBE) electrophoresis buffer. Cast two identical agarose gels (i.e. identical volumes) without ethidium bromide in two separate but identical electrophoresis trays. Add identical volumes of 0.5 \times TBE buffer to each of the trays.
2. Mix 50 μL of previously synthesized ds cDNA (Subheading 3.4, step 3) to 10 μL of loading buffer (*see Note 12*) and load in the first well of one of the two gels. Load in the first well of the second gel 10 μg of a DNA size marker (*see Note 13*). Connect both gel trays to the same power supply and run both gels for the same length of time at a low voltage (e.g. 3 V/cm). Stop the electrophoresis after the Orange G dye has run out of the gel that is when the 0.1 kb DNA size marker reaches the near end of the gel.
3. Without staining the gels, using a scalpel blade and a ruler, cut the gel lanes containing the cDNA and the DNA size marker.

Rotate the gel slices at 90° and place them at the upper end of two separate but identical gel trays. Pour in each tray identical volumes of 1.4 % low melting point agarose in 0.5× TBE buffer, enough as to cover the slices. As in **step 2** in Subheading 3.5, connect both trays to the same power supply and run the gels for the same time-span at a low voltage, as to allow the 0.1 kb size marker to reach the near end of the gel.

4. Stain the gel containing the DNA size marker for 30 min in a 0.5 µg/mL ethidium bromide solution while leaving unstained the gel containing the cDNAs. Soak the gel for 15 min in sterilized water to remove excess ethidium bromide. Place the DNA size marker-containing gel on a “Blue light transilluminator” (*see Note 1*) and superimpose the cDNA gel over the DNA marker gel. Using a scalpel blade, cut out from the upper cDNA gel different pieces of agarose corresponding to different cDNA size ranges (*see Fig. 3a*). Place the pieces of agarose in 1.5 mL tubes. Extract the cDNA from the agarose gel using a (commercial) gel extraction kit. Elute the cDNA in as little as 10 µL of elution buffer.

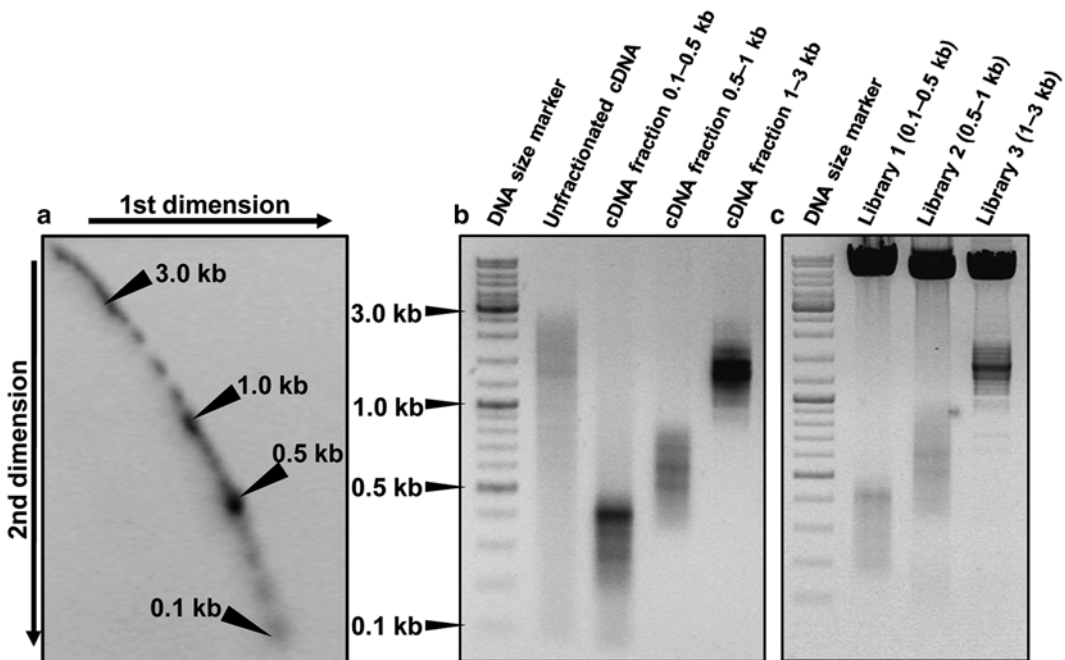


Fig. 3 cDNA size fractionation by two-dimensional agarose gel electrophoresis. **(a)** Separation of a DNA size marker by two-dimensional agarose gel electrophoresis. The unstained gel containing the cDNA is superimposed to the DNA size marker gel placed on a blue-light transilluminator to cut out the different cDNA size fractions. **(b)** Electrophoretic separation of three different PCR-amplified cDNA size fractions along a PCR-amplified non-fractionated cDNA sample. **(c)** Electrophoretic separation of *Sfi*I-digested samples of three sized cDNA libraries, the large intense band corresponds to the linearized plasmid vector

3.6 Amplification of the cDNA Size Fractions

1. As in Subheadings 3.3 and 3.4, this protocol makes use of components of the Mint-2 cDNA synthesis kit. In a 0.2 mL PCR tube, mix 36 μL of sterile RNase-free water, 5 μL of 10 \times Encyclo buffer, 1 μL of 10 mM dNTP mix, 2 μL of 10 μM PCR primer M1 (*see* Note 10), 5 μL of gel-eluted cDNA (from Subheading 3.5, step 4), and 1 μL of 50 \times Encyclo DNA polymerase. Mix the components by gently pipetting up and down.
2. Place the tube in a thermal cycler and apply a PCR cycle comprising an initial denaturation at 95 $^{\circ}\text{C}$ for 1 min, x cycles of 95 $^{\circ}\text{C}$ for 15 s, 66 $^{\circ}\text{C}$ for 20 s, and 72 $^{\circ}\text{C}$ for y min. Number of cycles (x) and extension time (y) must be adjusted for each cDNA size fraction. As an example, we used (*see* ref. 17) $x=30$ and $y=0.5$ min for size fractions between 100 and 500 bp, $x=26$ and $y=1$ min for size fractions between 500 and 1000 bp, and $x=22$ and $y=3$ min for size fractions between 1000 and 3000 bp. Control the amplification and the success of the size fragmentation by running 5 μL of the PCR reaction mix in a 1 % agarose gel (Fig. 3b).

3.7 cDNA Solution Hybrid Selection Capture

The aim of this protocol is to specifically select the different members of a specific gene family among the numerous and diverse environmental cDNA sequences. This protocol, first reported in [19], represents a specific use of the DNA hybrid selection capture detailed in [13]. As for the first-strand cDNA synthesis (*see* Subheading 3.3) and the second-strand cDNA synthesis and amplification (*see* Subheading 3.4), the protocol makes use of the Mint-2 cDNA synthesis kit.

1. Perform first-strand cDNA synthesis from 2 μg of total soil RNA as described in Subheading 3.3. Proceed to second-strand synthesis and cDNA amplification according to Subheading 3.4 by using a number of PCR cycles allowing recovery of μg amounts of cDNA (usually between 18 and 30 cycles depending on the RNA sample) (*see* Note 14).
2. Add 50 μL of water to 50 μL of amplified cDNA and perform a phenol:chloroform:isoamyl alcohol extraction followed by a chloroform:isoamyl alcohol extraction as described in step 3 in Subheading 3.2. Transfer the upper aqueous-phase into a new 1.5 mL tube and add 0.1 volume of 3 M Na acetate pH 5.2; 1.3 μL of 20 mg/mL glycogen and 2.5 volume of cold (-20°C) pure ethanol. Incubate overnight at -20°C and centrifuge for 20 min at $18,000\times g$ and 4 $^{\circ}\text{C}$. Remove the supernatant and wash the pellet with 100 μL of cold (-20°C) 70 % ethanol. Dry the pellet for 10 min at room temperature and dissolve the cDNA in 10 μL of water. Quantify DNA by spectrophotometry at 260 nm.
3. Perform a first round of hybrid selection capture using 500 ng of cDNA as described in steps 1–13 in Subheading 3.4 of

Chapter 10 (*see ref. 13*) and 500 ng of gene-specific biotinylated-RNA probes designed and synthesized as described in Subheading 3.1 of Chapter 10 (*see ref. 13*).

4. Perform a single purification of the captured cDNA using a PCR purification kit and recover the cDNA in 50 μL of water and amplify the cDNA using the Mint-2 cDNA synthesis kit. For each cDNA sample, prepare ten PCR reactions in 0.2 mL tubes as described in **step 1** in Subheading 3.4. After an initial denaturation at 95 °C for 1 min, perform 25 cycles of 95 °C for 15 s, 66 °C for 20 s, and 72 °C for 3 min. Perform a final elongation at 72 °C for 5 min.
5. Pool the PCR products 2 by 2 and purify each pool of cDNA using a PCR purification kit; recover the cDNAs in 50 μL of water. Pool all purified cDNA and quantify by spectrophotometry.
6. Perform a second round of hybrid selection capture on 500 ng of captured cDNA by repeating **steps 3–5** in Subheading 3.7.
7. Assess the quantity, quality, and size distribution of captured cDNA on an Agilent 2100 Bioanalyzer DNA 12000 chip. Captured cDNAs can be cloned as described in Subheading 3.8.

3.8 Cloning of the cDNA Size Fractions

1. Purify the amplified cDNA fractions by using a commercial kit (e.g. Qiagen Qiaquick PCR purification kit) and separately digest overnight the cDNA fractions and the cloning vector (*see Note 15*) with the *Sfi*I restriction enzyme at 50 °C.
2. Deactivate the enzyme by successive phenol:chloroform:isoamyl alcohol and chloroform:isoamyl alcohol extractions. Precipitate the DNA using 3 M Na acetate pH 4.8 and pure ethanol. Wash the pellets with cold 70 % ethanol and resuspend in a small volume (e.g. 20 μL) of water. Measure the concentration by spectrophotometry at 260 nm.
3. Ligate each of the cDNA fractions to the vector using the T4 DNA ligase and an approximate insert to plasmid molar ratio of 3:1 by following the instruction provided with the DNA ligase enzyme.
4. Perform an initial small-scale transformation of *E. coli* cells (*see Note 16*) using a small fraction of the ligation mix (5–10 %). Spread serial dilutions of the transformation mix on plates filled with a selective medium supplemented with the antibiotic corresponding to the plasmid antibiotic resistance gene. Incubate at 37 °C overnight and count the colonies.
5. Amplify, by colony PCR, the cDNA inserts present in *ca* 20 bacterial colonies using PCR primers located on each side of the plasmid cloning sites. Run the amplified products on a 1 % agarose gel to estimate the percentage of plasmids devoid of inserts and the average size of the cloned cDNAs (*see Note 17*).

6. Scale up the transformation to obtain a library with the desired number of independent clones. Plate the transformation mix as to obtain a high density of discrete colonies on the Petri dishes (*see Note 18*). After an overnight growth at 37 °C, pour 5 mL of liquid medium on each 140 mm Petri plate, scrap the colonies using a glass/plastic spreader. Mix all colonies from all plates in a single flask. Perform a large-scale plasmid extraction using either a standard alkaline lysis protocol [20] or a commercial kit.

4 Notes

1. Visualization of ethidium bromide-stained DNA by excitation with blue-light (in the 420–500 nm range, as performed by, e.g. the Dark Reader[®] instrument from Clare Chemicals) instead of UV light considerably minimizes DNA damage and increases both PCR amplification and cloning efficiency.
2. We do not recommend the use of diethylpyrocarbonate (DEPC), a hazardous compound, for preparing RNase-free solutions and plasticware.
3. At that stage where no RNase denaturing agent has yet been added, care must be taken to proceed swiftly, to leave the soil for a minimum of time at room temperature and to place the tubes at –80 °C immediately after vortexing.
4. At that stage, most samples still present a yellow to brown color indicative of the presence of non-RNA contaminants which need to be eliminated.
5. At that stage, pure, undegraded good-quality RNA can be sent to a sequencing platform. A minimum of 200 ng may be required for the sequencing of total soil RNA (rRNA + mRNA). Microgram quantities may be required for the specific sequencing of mRNA which requires the elimination of rRNA molecules.
6. Estimation of nucleic acid concentration by spectrophotometry often leads to overestimation of the actual concentration due to the overlooked presence of UV-absorbing chemical contaminants.
7. The relative proportions of “denaturing” and “lysis” solutions (*see item 10* in Subheading 2.5) need to be optimized for each type of soil. In a preliminary experiment, we recommend to perform a series of extractions using increasing amounts of denaturing solution (from 25 up to 150 µL) and decreasing amounts of the lysis one (from 925 down to 850 µL). Low ratios work usually better for soils poor in organic matter and high ratios for high organic matter contents.

8. CDS-4M primer contains an oligo-dT sequence that anneals to poly-A stretches of eukaryotic mRNA to initiate reverse transcription. PlugOligo-3M adapter contains an oligo-dG sequence at its 3' end. It pairs to complementary oligo-dC stretches "artificially" added at the 3'-end of first-strand cDNA by the reverse transcriptase (RT) when it reaches the 5' end of the mRNA. As a consequence, the RT continues first-strand cDNA synthesis to the end of the PlugOligo-3M incorporating its reverse complementary sequence at the 5' end of the cDNA. PlugOligo-3M and CDS-4M also contain asymmetric sites for *Sfi*I restriction endonuclease (*Sfi*IA and *Sfi*IB respectively). These sites allow directional cloning of the cDNA after their incorporation at 5' and 3' ends of synthesized cDNA.
9. The IP-solution is a specially tailored solution which increases the efficiency of PlugOligo-3M incorporation in the cDNA.
10. The sequence of primer M1 (AAGCAGTGGTATCAACGCAGAGT) is identical to the 5' end sequences of CDS-4M primer and PlugOligo-3M adapter. It therefore binds to both 5' and 3' ends of all first-strand cDNAs and allows synthesis and amplification of all ds-cDNAs.
11. Long-range PCR amplification of the cDNAs for a limited number of cycles (three) significantly increases the amount of cDNA that will be subsequently separated by agarose gel electrophoresis without significantly affecting the relative proportions of "long" versus "short" cDNAs in the original sample.
12. We use a 6× DNA loading buffer containing 0.15 % orange G, 0.03 % xylene cyanol FF and 60 % glycerol in 10 mM Tris-HCl pH 7.6; 60 mM EDTA.
13. It is advisable to use a DNA size marker with DNA fragments regularly distributed between 0.1 and 1 kb and between 1 and 5 kb.
14. The optimal number of cycles is selected by performing preliminary PCR amplifications with cycles ranging from 18 to 30. An optimal number of cycles leads to a visible smear of cDNA sequences on an agarose gel (between *ca* 0.5 and 3–5 kbp) without amplification of longer, artificial DNA fragments.
15. All amplified cDNA are bordered with a *Sfi*IA and a *Sfi*IB site at their 5' and 3' ends, respectively. Several *Saccharomyces cerevisiae* expression plasmids with these two sites placed downstream of a strong promoter sequence are available [7, 9, 11, 12].
16. It is advisable to use commercially available chemically or electrocompetent cells with transformation efficiencies of at least 10⁹ colony forming units per µg of transforming plasmid.

17. An unsatisfactory high percentage of empty plasmids may result either from partial digestion of the cDNA inserts and/or of the plasmid (*see step 1* in Subheading 3.8) which should be repeated or from an excess of plasmid respective of the insert in the ligation (*see step 3* in Subheading 3.8).
18. Up to 10,000–20,000 bacterial colonies can be plated on a single 140 mm Petri dish. Therefore, 50–100 Petri dishes may be required to obtain a plasmid library containing 10^6 independent clones.

Acknowledgements

Work on metatranscriptomics of soil eukaryotic microbial communities was supported by the CNRS (program INSU ECCO Microbien), the Agence Nationale de la Recherche (projects 06-BLAN-0088, 06-443-BDIV-006 and 09-GENM-033-001), INRA métaprogramme M2E and the Indo-French Centre for the Promotion of Advanced Research, grant 4709-1.

References

1. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* 66:2541–2547
2. Ferrer M, Belouqui A, Timmis KN, Golyshtin PN (2009) Metagenomics for mining new genetic resources of microbial communities. *J Mol Microbiol Biotechnol* 16:109–123
3. Chistoserdova L (2010) Recent progress and new challenges in metagenomics for biotechnology. *Biotechnol Lett* 32:1351–1359
4. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG et al (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* 10:807–812
5. Stewart FJ, Ottesen EA, DeLong EF (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* 4:896–907
6. Grant S, Grant WD, Cowan DA, Jones BE, Ma Y, Ventosa A, Heaphy S (2006) Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl Environ Microbiol* 72:135–143
7. Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, Wesolowski-Louvel M, Marmeisse R (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J* 1:632–642
8. Damon C, Lehembre F, Oger-Desfeux C, Luis P, Ranger J, Fraissinet-Tachet L, Marmeisse R (2012) Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PLoS One* 7:e28967
9. Damon C, Vallon L, Zimmermann S, Haider MZ, Galeote V, Dequin S et al (2011) A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes. *ISME J* 5:1871–1880
10. Lehembre F, Doillon D, David E, Perrotto S, Baude J, Foulon J et al (2013) Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes. *Environ Microbiol* 15:2829–2840
11. Kellner H, Luis P, Portetelle D, Vandenberg M (2011) Screening of a soil metatranscriptomic library by functional complementation of *Saccharomyces cerevisiae* mutants. *Microbiol Res* 166:360–368
12. Bragalini C, Ribière C, Parisot N, Vallon L, Prudent E, Peyretailade E et al (2014) Solution hybrid selection capture for the recovery of functional full-length eukaryotic

- cDNAs from complex environmental samples. *DNA Res* 21:685–694
13. Ribière C, Beugnot R, Parisot N, Gasc C, Defois C, Denonfoux J et al (2015) Targeted gene capture by hybridization to illuminate ecosystem functioning, *Methods in molecular biology*. Springer, New York
 14. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30:892–897
 15. Luis P, Kellner H, Martin F, Buscot F (2005) A molecular method to evaluate basidiomycete laccase gene expression in forest soils. *Geoderma* 128:18–27
 16. Damon C, Barroso G, Ferandon C, Ranger J, Fraissinet-Tachet F, Marmeisse R (2010) Performance of the *COX1* gene as a marker for the study of metabolically active Pezizomycotina and *Agaricomycetes* fungal communities from the analysis of soil RNA. *FEMS Microbiol Ecol* 74:693–705
 17. Yadav RK, Barbi F, Ziller A, Luis P, Marmeisse R, Reddy MS, Fraissinet-Tachet L (2014) Construction of sized eukaryotic cDNA libraries using low input of total environmental metatranscriptomic RNA. *BMC Biotechnol* 14:80
 18. Wellenreuther R, Schupp I, Poustka A, Wiemann S (2004) SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. *BMC Genomics* 5:36
 19. Denonfoux J, Parisot N, Dugat-Bony E, Biderre-Petit C, Boucher D, Morgavi DP et al (2013) Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration DNA. *DNA Res* 20:185–196
 20. Green MR, Sambrook J (2012) *Molecular cloning, a laboratory manual*, vol 1, 4th edn. Cold Spring Harbor Laboratory, New York

Chapter 17

Analysis of Ancient DNA in Microbial Ecology

Olivier Gorgé, E. Andrew Bennett, Diyendo Massilani, Julien Daligault, Melanie Pruvost, Eva-Maria Geigl, and Thierry Grange

Abstract

The development of next-generation sequencing has led to a breakthrough in the analysis of ancient genomes, and the subsequent genomic analyses of the skeletal remains of ancient humans have revolutionized the knowledge of the evolution of our species, including the discovery of a new hominin, and demonstrated admixtures with more distantly related archaic populations such as Neandertals and Denisovans. Moreover, it has also yielded novel insights into the evolution of ancient pathogens. The analysis of ancient microbial genomes allows the study of their recent evolution, presently over the last several millennia. These spectacular results have been attained despite the degradation of DNA after the death of the host, which results in very short DNA molecules that become increasingly damaged, only low quantities of which remain. The low quantity of ancient DNA molecules renders their analysis difficult and prone to contamination with modern DNA molecules, in particular via contamination from the reagents used in DNA purification and downstream analysis steps. Finally, the rare ancient molecules are diluted in environmental DNA originating from the soil microorganisms that colonize bones and teeth. Thus, ancient skeletal remains can share DNA profiles with environmental samples and identifying ancient microbial genomes among the more recent, presently poorly characterized, environmental microbiome is particularly challenging. Here, we describe the methods developed and/or in use in our laboratory to produce reliable and reproducible paleogenomic results from ancient skeletal remains that can be used to identify the presence of ancient microbiota.

Key words Ancient DNA, NGS, Double-stranded library, Single-stranded library, IonTorrent, Illumina, Contamination

1 Introduction

Ancient DNA (aDNA) preserved in skeletal remains from past organisms can be a rich source of information on the evolution of species, of both the organism itself and its pathogens (for a typical ancient skeleton *see* Fig. 1). aDNA, however, is often highly degraded and the techniques for its analysis need to be optimized in order to ensure the production of authentic results. Indeed,

*Author contributed equally with all other contributors.



Fig. 1 Skeletal remains from a 7000-year-old Neolithic burial from Berry-au-Bac “le Vieux Tordoir” (Aisne, France). Excavation and photograph: CNRS, UMR 8215 “Trajectoires”

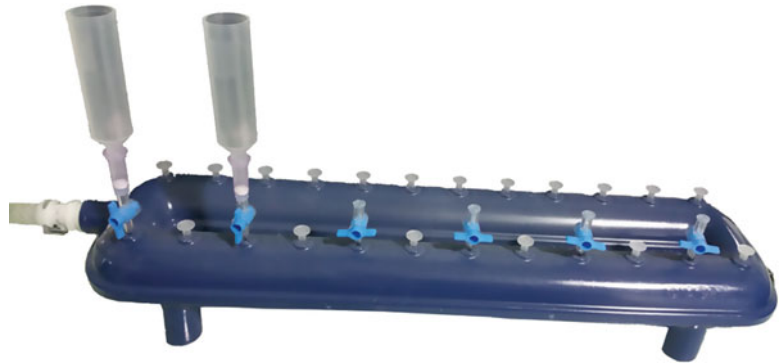


Fig. 2 Qiavac Manifold equipped with Qiaquick Spin columns and extenders to purify DNA from large-volume extracts

working with aDNA requires precautions that need to be applied even before samples enter the paleogenetic laboratory. Archaeologists should be taught the constraints of aDNA research so that they can adapt their working procedures to increase the likelihood of obtaining reliable aDNA results. For example, a common practice in archaeology is to wash bones after excavation. This washing leads to the dilution and degradation of the ancient DNA molecules as well as to contamination with environmental DNA [1, 2]. This is particularly problematic if only very low amounts of DNA are preserved in the skeletal remains. To increase the likelihood of DNA preservation in the sample, it is preferable to use freshly excavated remains for paleogenetic analyses [1]. This will be possible, however, only in a limited number of cases and does not concern previously excavated and curated remains. Moreover, DNA is heavily transformed after the death of an organism. Between several hours,

and possibly up to the first several years, after death, DNA is hydrolyzed enzymatically into small fragments leading to a median size of 50–70 bp (Guimaraes et al., unpublished) [3]. Over time, DNA bases become modified; in particular the cytosines become deaminated, which occurs preferentially close to the molecule ends [4]. The quantity of endogenous aDNA varies among samples, and even in different locations in the same skeletal remains. Although there may be other factors, temperature in particular has been characterized as playing a major role in aDNA preservation [5].

In order to study the genomes of pathogens that are associated with an animal (vertebrate) at the time of its death, one must consider all the events taking place during the diagenetic transformation of biomolecules, including DNA, following death. The body and its constituents will begin to decompose, mostly due to the action of microorganisms and insects. They not only metabolize biomolecules, but also deposit their own DNA, becoming the first contaminants of skeletal remains. Once the soft tissues and accessible organic parts of the bones have been consumed, the skeletal parts will enter a slow decay phase involving mostly chemical processes. When bones are buried, either intentionally at the time of death or simply due to natural burial of the skeleton over time, there will be a slow but regular exchange of biomolecules between bone and soil. Thus, at the time of excavation, the DNA that can be recovered may contain (1) DNA from the initial organism; (2) the DNA of the microbes, as well as parasites, that were associated with the organism during its lifetime (some of which may possibly have been the cause of its death); (3) the DNA of the organisms that have contributed to the decomposition of the body following its death; (4) and the DNA of the soil organisms that have penetrated into the bone. If at the time of excavation no special precautions are taken, and the bones are handled and washed as is routinely done, the bone can be further contaminated with fresh modern DNA, mostly of human and microbial origin, as well as from various other sources. The microbial composition of skeletal remains therefore reflects the microbial composition of the burial environment, showing that fossilizing skeletal remains resemble environmental samples. Indeed, when DNA retrieved from ancient bones is sequenced with a shotgun approach, typically only a few percent, or tenths of percent, of the sequenced DNA correspond to the initial organism, the rest being identified as “environmental DNA.” Most of this “environmental DNA” cannot be mapped to sequenced genomes, and remains as “unknown” [6]. Since older DNA is increasingly degraded until complete disappearance, one could expect that most of the environmental DNA recovered from the skeletal remains is of recent origin. The DNA decay rate depends on the environment; however, special “molecular niches” within the bone may offer more protected DNA-stabilizing microenvironments [7–9]. This can explain the exceptional preservation found in a limited number of

bones [10–13]. The more the microbiome is intimately associated with the bone and teeth matrix, the better the likelihood that it can also benefit from such “preserving molecular niches”. Thus, the DNA of pathogens that can be spread to bones and teeth with the bloodstream, like *Yersinia pestis*, *Mycobacterium leprae*, and *Mycobacterium tuberculosis*, can be retrieved from well-preserved ancient skeletal remains (e.g., [14–16]). Similarly, the DNA of ancient buccal microbiomes can be retrieved from dental plaque, which appears to offer a suitable mineralized environment for long-term DNA preservation [17]. It remains to be determined if other organisms of the microbiome can also deposit their DNA into favorable “preserving molecular niches” allowing long-term DNA preservation. Such microorganisms are the first players to colonize the body through the blood vessels and it is yet unknown whether most of the “environmental DNA” that can be retrieved from ancient bones is of recent or ancient origin. In fact, it is likely that the unique taphonomic history (the history of *postmortem* decay) of each bone, in both macro- and microniches, shows sufficient bone-to-bone diversity to allow very different outcomes in terms of DNA preservation, and of the age of the DNA that is recovered.

In order to optimize the recovery of DNA from the ancient microbiome, improved methods adapted to the preferential recovery of the most damaged molecules must be developed. Their use should prevent, as much as possible, the introduction and incorporation of modern DNA molecules into the sequencing libraries. After sequence production, bioinformatics methods adapted to analyze ancient molecules should be used. Since the analysis of ancient microbiomes is presently in its infancy, there are not yet reliable, established procedures available to ensure the recovery of authentic data. For the moment, one has to rely on the more established procedures developed to analyze ancient host DNA.

Here, we provide some guidelines designed for the analysis of ancient microbiomes. First, it is essential to use all possible means to minimize contamination with modern DNA because the minute quantities of DNA in the ancient bone and tooth extracts can readily be contaminated with traces of modern DNA from the same or other species. Indeed, the scarcer the endogenous DNA, the higher the ratio of contaminating DNA likely to be found in the extract. This requires a high-containment laboratory for extraction, purification, and library construction of aDNA. In addition, very strict protocols to avoid carryover contamination and to decontaminate reagents must be applied. Carryover contamination results from molecules produced during previous amplification or library construction steps being reintroduced into another sample. Another source of contamination is trace DNA molecules present in reagents such as DNA from domestic animals, the proteins of which are often used to stabilize enzymes,

human DNA from employees at biotech companies, or bacterial DNA from either the bacteria used for enzyme production or the bacteria introduced from the environment during the production process. In order to ensure the authenticity of the results, reagents must first be decontaminated to eliminate as much reagent-borne DNA as possible before use [18]. Here we describe the different protocols that we have developed to reduce the level of contaminating DNA found in reagents.

Second, because DNA is damaged and degraded, one must use library construction methods that allow the best possible recovery of the most damaged molecules, and, if possible, discriminate against the recovery of modern ones. We present herein experimental procedures allowing optimal recovery of short double-stranded and of highly damaged molecules, which are best recovered as single-stranded DNA [19]. In the case of amplicon sequencing (i.e., 16S rRNA gene), short regions must be targeted because of the reduced length of ancient DNA. When analyzing soils or sediments, the vast majority of DNA, and consequently 16S rRNA genes, is from modern organisms. As a consequence it is best to select for short DNA fragments, when possible, prior to amplifying targets.

Third, data produced must be analyzed using bioinformatic workflows designed to characterize ancient DNA and ancient microbiomes. We use *leeHom* [20], to quality trim and merge paired-end reads produced from short ancient DNA templates and *mapDamage 2.0* [21] to assess the authenticity of the mapped DNA. To analyze shotgun sequencing reads or 16S rRNA amplicons, we use both *MG-RAST* (presented in Chapter 4) and homemade dedicated pipelines with *leeHom* to pre-process reads, and *BWA* [22] to map them against an in-house reference sequence consisting of concatenated bacterial genomes, before taxonomic characterization.

2 Materials

Buffers are stored at room temperature while reaction mixes, primers, and most enzymes are stored at -20°C .

2.1 DNA Extraction Reagents

Prepare all solutions from autoclaved deionized water. We use household bleach (2.6 % sodium hypochlorite) and *RNase away* (Life Technologies, Carlsbad, CA, USA) as agents for decontamination and DNA removal.

1. Commercial soil extraction and purification kits are used, but the reagents are only opened in the high-containment laboratory, under controlled conditions to avoid contamination. We currently use *MoBio PowerMax Soil DNA Isolation kit* (MO

BIO, Carlsbad, CA, USA, ref. 12988), Qiagen Gel Extraction kit (Qiagen, Hilden, Germany, ref. 28704), and Qiagen PCR purification kit (ref. 28104).

2. Bone matrix disintegration and digestion buffer: 0.5 M EDTA pH 8, 0.25 M PO_4^{3-} , 0.14 M β -mercaptoethanol. 0.5 M EDTA pH 8 is prepared from EDTA powder and autoclaved water and pH is adjusted with NaOH pellets [2].
3. Buffer QG (Qiagen, ref. 19063), solubilization and binding buffer.
4. Buffer PE (Qiagen, ref. 19065), wash buffer.
5. Buffer EB (Qiagen, ref. 19086), elution buffer.

2.2 qPCR Reagents

1. MixG (homemade qPCR mix) [23]: To prepare 100 μL of 10 \times mixG, mix 19.5 μL γ -irradiated water (*see Note 1*), 6.25 μL 10 mg/mL bovine or horse serum albumin (BSA or HSA), 3 μL 10 % Lubrol-17A17 (SERVA Electrophoresis GmbH, Heidelberg, Germany), 50 μL 50 % glycerol, 1.25 μL 5 M KCl, and 20 μL 2.5 M AMPD (2-amino-2-methyl-1,3-propanediol) pH 8.3 (*see Note 2*). For volumes higher than 200 μL , aliquot 200 μL each in UV-transparent tubes (Qubit Assay tubes, Life Technologies, ref. Q32856) and treat with UV (*see Note 3*), dilute 10,000 \times SYBR-Green I (Life Technologies, ref. S-7585) 1/40 in DMSO and add 1 μL diluted SYBR-Green I per 100 μL mix. Freeze overnight at -80°C (*see Note 4*).
2. BIOTEC buffer: To prepare 10 mL BIOTEC buffer, mix 200 μL 1 M Tris-HCl pH 7.5, 800 μL 25 mM MgCl_2 , 20 μL 5 M NaCl, 5 mL 50 % glycerol, 10 μL 10 % Triton x100, complete to 10 mL with γ -irradiated water. Aliquot 540 μL each in UV-transparent tubes, UV irradiate 300 s on each side on a UV cross-linker (*see Note 3*).
3. Thermolabile double-strand DNase (2 u/ μL hl-dsDNase) from ArcticZymes (Tromsø, Norway, ref. #70800). For a final activity of 0.02 u/ μL , add 1 μL 2 u/ μL of hl-dsDNase to 99 μL of BIOTEC buffer.
4. Decontaminated Taq DNA polymerase: 108 μL of 5 u/ μL Hot Start Taq polymerase in its storage buffer is supplemented with 6 μL premixed 200 mM MgCl_2 , 20 mM CaCl_2 , and 2.45 μL 50 mM DTT, then incubated with 6 μL 2 u/ μL hl-dsDNase for 30 min at 25°C followed by a 20-min inactivation step at 50°C . Aliquot to desired volumes. Final activity of decontaminated Taq is 4.4 u/ μL .
5. Decontaminated dNTPs: 20 μL of dATP, dCTP, dGTP, and 40 μL of dUTP (100 mM stock solutions each) are mixed with 100 μL of γ -irradiated water. 40 μL of this dNTP mix is mixed with 35 μL γ -irradiated water, 4 μL 50 mM DTT, 20 μL 250

mM Tris pH 8, 80 μL 50 mM MgCl_2 , and 20 μL 10 mM CaCl_2 and then incubated with 1 μL 0.02 u/ μL hl-dsDNase for 30 min at 25 °C followed by a 30-min inactivation at 55 °C. The final dNTP concentration is 2 mM (4 mM for dUTP).

6. 1u/ μL codUNG (ArcticZymes, ref. #70500): codUNG is a uracil-DNA glycosylase from Atlantic cod that is completely and irreversibly inactivated by moderate heat treatment [18].

2.3 Library

Preparation for PCR Products

1. Oligonucleotides
 - (a) Oligonucleotides are those proposed by Life Technologies for genomic DNA Fragment Library preparation¹ (*see Note 5*).
 - (b) Annealing buffer 10 \times
 - 25 μL 5 M NaCl
 - 100 μL 250 mM Tris
 - 50 μL 250 mM MgCl_2
 - 75 μL γ -irradiated water (*see Note 1*)
 - 40 μM annealed adapters (*see Note 6*)
2. End-repair enzymes (3 u/ μL T4 DNA polymerase and 10 u/ μL T4 polynucleotide kinase), such as NEBNext end repair module (ref. E6050, New England Biolabs (NEB), Ipswich, MA, USA)
3. Commercial purification kit, based on silica columns (Qiagen or Macherey-Nagel, Düren, Germany) or SPRI magnetic beads (Ampure XP, Agencourt Technologies, Beverly, MA, USA, ref. 16388 or NucleoMag NGS clean-up and size selection, Macherey Nagel ref. 744970)
 - (a) DNA purification kit for 96 samples (96 silica column plate or SPRI magnetic beads)
 - (b) Individual DNA purification kit (silica column or magnetic beads)
4. Quick ligase, such as NEBNext Quick ligation module (NEB, ref. E6056, 2000 u/ μL T4 DNA ligase)
5. Size selection reagents for E-gel or Caliper XT devices
6. OneTaq Hot Start 2 \times Master Mix with Standard Buffer (NEB, ref. M0484)

2.4 Double-Stranded DNA Library Preparation

Use γ -irradiated water for all solutions and buffer preparations as well as for any dilution or elution steps (unless otherwise indicated) (*see Note 1*).

¹Appendix E, p. 56–57, Publication Part Number MAN0009847; Revision C.0 Date 29 April 2014.

1. Oligonucleotides (*see* **Note 7**)
 - (a) 10× Annealing buffer (*see* Subheading 2.3, **item 1**) UV-irradiated (*see* **Note 3**)
 - (b) Annealed adapters 40 μM (*see* **Note 6**)
2. Deaminated cytosine repair
1 u/μL USER enzyme (NEB, ref. M5505)
3. End repair
NEBNext End Repair Module (NEB, ref. E6050)
4. Purification of repaired ancient DNA extract
 - (a) MinElute Column (Qiagen ref. 28604) (*see* **Note 8**)
 - (b) Qiaquick Gel Extraction Kit (Qiagen ref. 28704)
 - (c) QG Buffer (Qiagen ref. 19063)
 - (d) Isopropanol
5. Blunt-end double-stranded adapter ligation
 - (a) 40 μM Double-stranded adapters
 - (b) NEBNext Quick Ligation Module (NEB ref. E6056)
6. Elongation and pre-amplification
 - (a) OneTaq Hot Start 2× Master Mix with Standard Buffer (NEB, ref. M0484)
 - (b) 10 μM Illumina amplification forward and reverse modified primers P5s/P7s (*see* **Note 9**)
7. Purification and size selection
 - (a) NucleoMag NGS clean-up and size selection kit (Macherey Nagel ref. 744970)
 - (b) Freshly prepared 80 % ethanol
 - (c) DNase/RNase-free water
8. qPCR determination of the optimal number of cycles for library amplification
FastStart DNA Master^{PLUS} SYBR Green I mix (Roche Applied Science, ref. 035158)
9. Final library amplification
 - (a) FastStart Taq DNA Polymerase, dNTPack (Roche Applied Science, ref. 04738)
 - (b) DNase/RNase-free water
10. Library characterization and purification
Agilent Bioanalyzer high sensitivity DNA kit (Agilent ref. 5067-4626)

2.5 Single-Stranded DNA Library Preparation

Use γ -irradiated water for all solutions and buffer preparations as well as for any dilution or elution steps (unless indicated otherwise) (*see Note 1*).

1. Oligos (*see Note 10*)
 - (a) 10 \times Annealing buffer UV-irradiated (*see Note 3*)
 - (b) 40 μ M Annealed CL53/CL73 adapters (*see Note 6*)
2. DNA preparation
 - (a) 100 u/ μ L Circligase II ssDNA ligase with 10 \times Circligase buffer and 50 mM MnCl₂ solution (Epicentre, Chicago, IL, USA, ref. CL902)
 - (b) 10 u/ μ L Endonuclease VIII (NEB, ref. M0299)
 - (c) 1 u/ μ L codUNG from Artizymes (optional)
 - (d) 1 u/ μ L FastAP (Thermo Scientific, Waltham, MA, USA, ref. EF065), a thermosensitive alkaline phosphatase
3. First adapter ligation
 - (a) 50 % PEG 4000 (Sigma-Aldrich, St. Louis, MO, USA, ref. 95904)
 - (b) Dynabeads MyOne Streptavidin C1 (Life Technologies, ref. 6500)
 - (c) Bead binding buffer: 1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.05 % Tween-20, 0.5 % SDS. Prepare buffer just before use and discard immediately. Buffer has no shelf life after adding SDS.
 - (d) Wash buffer A: 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.05 % Tween-20, 0.5 % SDS. Can be stored at room temperature for a month.
 - (e) Wash buffer B: 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.05 % Tween-20. Can be stored at room temperature for a year.
 - (f) Stringency wash buffer: 0.1 \times SSC, 0.1 % SDS. Can be stored at room temperature for a month.
 - (g) 10 \times ThermoPol Buffer (NEB, ref. B9004)
 - (h) Bst 2.0 DNA Polymerase (NEB ref. M0537)
 - (i) 10 \times Tango buffer (Thermo Scientific, Waltham, MA, USA, ref. BY5)
 - (j) 1 % Tween-20 (Sigma-Aldrich, St. Louis, MO, USA, ref. P2287)
 - (k) T4 DNA polymerase (Thermo Scientific, ref. EP006)
 - (l) Stop solution: 0.5 M EDTA pH 8.0, 2 % Tween-20

4. Second adapter ligation
 - (a) T4 DNA ligase (Thermo Scientific, ref. EL001)
 - (b) EBT: 10 mM Tris-HCl pH 8.0, 0.05 % Tween-20
5. Library amplification
 - (a) qPCR master mix such as LightCycler FastStart DNA Master SYBR Green I (Roche Applied Science)
 - (b) Either MinElute PCR purification kit (Qiagen), AMPure XP (Agencourt Technologies), or NucleoMag NGS Clean-up and Size Select kit (Macherey-Nagel)

2.6 Equipment (other than common devices for molecular biology laboratories)

1. qPCR-capable thermocycler such as LightCycler 2.0 (Roche Applied Sciences)
2. UV-crosslinker such as Spectrolinker XL 1500 UV-crosslinker (Spectronics Corp., Westbury, NY, USA)
3. Multi-purpose rotating tool such as Dremel 9100 Fortiflex Heavy Duty Flex Shaft Tool (Robert Bosch GmbH, Stuttgart, Germany) with diamond cutting wheel (ref. 545) and high-speed cutting or drilling bits (e.g., ref. 194)
4. Freezer-mill such as Spex Certiprep 6770 Freezer/Mill® (SPEX, Metuchen, NJ, USA)
5. Electrophoresis system for DNA sizing and purification, such as E-Gel SizeSelect (Life Technologies), Caliper Labchip XT (Perkin-Elmer, Waltham, MA, USA) or Pippin Prep (Sage Science, Beverly, MA, USA)
6. Lab-on-chip electrophoresis system, such as Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA)
7. Fluorimeter for DNA quantification, such as Qubit 2.0 (Life Technologies)
8. Optional: Robotic platform, such as a TECAN EVO 100 (Tecan, Maennedorf, Switzerland), to facilitate high-throughput sample treatment, but all steps can be performed manually
9. Heating/cooling block with mixing capability, such as an Eppendorf Thermomixer Comfort (Eppendorf, Hamburg, Germany) (*see Note 11*)
10. Ice-water bath
11. Magnetic rack for 1.5 mL tubes

3 Methods (*See Note 12*)

3.1 DNA Extraction and Purification

1. DNA extraction from soil
To extract and purify DNA from soil (*see Note 13*), we found PowerMax Soil DNA Isolation kit from MoBio useful for a wide range of samples (*see Note 14*).

Typically, 10 g of soil are processed following the guidelines recommended by the manufacturer.

After extraction and purification, DNA is dissolved in 5 mL resuspension buffer. It is subsequently concentrated and repurified with a modified Qiagen Gel Extraction protocol as follows:

- (a) Set an incubator (e.g., heating block) or water bath to 50 °C.
- (b) Heat an aliquot of EB to 50 °C (150 µL per sample to be extracted).
- (c) Add 30 mL of QG and 20 mL of isopropanol to the DNA solution.
- (d) Pass through columns mounted on a manifold device (QIAvac 24 Plus, Qiagen, ref. 19413) equipped with extension tubes (Qiagen ref. 19587) (*see* Fig. 2).
- (e) Once all the samples have passed through the columns (*see* **Note 15**), remove extenders to wash them with autoclaved water and install them in place again.
- (f) Wash columns with 2 mL PE.
- (g) Break vacuum and stop vacuum pump.
- (h) Remove columns from the manifold and transfer them in clean 2 mL collection tubes.
- (i) Centrifuge at $9300\times g$ for 2 min to dry the columns.
- (j) Transfer columns to new clean 1.5 mL microcentrifuge tubes.
- (k) Carefully dispense 100 µL of preheated (50 °C) elution buffer (EB or molecular grade water) and incubate for 1 min at room temperature.
- (l) Centrifuge at $9300\times g$ for 1 min, discard columns, and store eluate (*see* **Note 16**).

2. DNA extraction from bone

Ancient bones and teeth can be considered as environmental samples (*see* **Note 17**), since they are a matrix in which bacteria, fungi, and other organisms reside. The dedicated procedures developed to access the endogenous DNA also extract the degraded bacterial (as well as fungal and eukaryotic) DNA present in the ancient bone (*see* **Note 18**).

The cleaning and powdering steps of the skeletal remains are performed in a high-containment laboratory [19] (*see* **Note 19**). The surface of the remains is removed in a UV-irradiated protective hood. Bone is then drilled or ground to fine powder in a freezer mill (Spex Certiprep 6750). Further processing of the bone powder is performed as described [19]. Blank extractions are carried out for each extraction series.

- (a) Decontaminate the hood and all tools with pure household bleach (use RNase away for decontaminating metal tools).
- (b) Cover the base of the working station with a sheet of aluminum foil.
- (c) Remove the outer layer of the sample with a scalpel. Use a new scalpel for each sample, and discard used scalpels.
- (d) Set up and turn on a vacuum cleaner during cutting to prevent the dispersal of bone powder in the working station.
- (e) Sample preparation in a freezer mill:
 - Use a Dremel multitool equipped with a diamond cutting wheel to cut pieces (up to 200 mg), adjusting the speed to the density/mineralization of the bone. Decontaminate the cutting wheel using the medium flame of a Bunsen burner at around 500 °C, but not above, to avoid damaging the diamond wheel.
 - Weigh the bone fragment(s) and pulverize it in a freezer mill in liquid nitrogen for 1 min (10 impacts per second).
 - After transferring the powder to a clean tube, clean freezer mill tubes. Metal parts: remove remaining bone powder in a bath of RNase away, rinse in a water bath, and dry under UV light. Plastic tubes: brush with water to remove remaining bone powder, rinse in a bleach bath, then with water and let them dry overnight on clean aluminum foil. Do not expose to UV light.
- (f) Sample preparation using a drill
 - Depending on the density of the bone, its shape, and the location of the sampling area, various drill bits can be chosen according to the user's needs. Assemble and decontaminate the drill bit using the medium flame of a Bunsen burner (ca. 500 °C). Drill at the lowest speed possible giving efficient bone powder production.
 - Transfer the bone powder in a pre-weighed tube and weigh it.
- (g) Cleaning
 - Between preparations of each sample, change the aluminum foil, clean the working station with bleach, and change gloves. Flame the drill bit or the cutting wheel using the medium flame of a Bunsen burner (ca. 500 °C).
 - After completion of the preparation series, decontaminate the Dremel tool with 70 % ethanol and RNase away (not bleach), and flame the drill bit or the cutting wheel using the medium flame of a Bunsen burner (ca. 500 °C). Clean and decontaminate with bleach the working station and the vacuum cleaner, place a UV

lightsource inside the working station so that it irradiates the surface of the bench at close proximity for at least 3 h. Clean the area around and wash floor with 10 % household bleach.

- (h) Once powdered, the bone is mixed with digestion buffer (1 mL per 100 mg of bone powder) and incubated at 37 °C in an orbital shaker (10 RPM) for 24/48 h or upon complete dissolution of the bone powder.
- (i) After disintegration of the bone matrix, DNA is purified on silica columns.
 - Centrifuge the suspension for 10 min at 13,000 $\times g$.
 - Prepare manifold with QiaAmp spin columns and extenders as in Subheading 3.1, step 1.
 - Transfer supernatant to 15 mL Falcon tubes and store pellets (-20 °C) for possible future re-extraction.
 - Follow the modified Qiagen Gel Extraction protocol described in Subheading 3.1, step 1.

3.2 DNA Amplification

1. To test whether the DNA extracts from the ancient bone samples inhibit the PCR (*see Note 20*), serial dilutions of the extracts spiked with a known quantity of positive internal control are amplified [24]. Subsequently, the Ct (crossing point at threshold) is analyzed for each serial dilution of the same sample. We dilute the extract two- and fourfold at the highest, considering that further dilution of ancient DNA extracts potentially containing only few molecules could cause the loss of targets.
2. For DNA amplification (*see Note 21*), we systematically use a home-made decontaminated qPCR mix (mixG). Endogenous DNA detection relies on the amplification of DNA fragments that are informative enough to discriminate between animals or, for bacteria, between phyla, classes, orders, genera, or even species depending on the targeted genes. To assess bacterial diversity, the 16S rRNA gene is a powerful tool, thanks to its ubiquity and structure, which make it an optimal marker for the characterization of environmental samples (*see Note 22*).

Amplification is performed with a LightCycler 2.0 (Roche Applied Sciences). Prepare a PCR mix using 1 μ L of DNA extract, 1 μ L of mixG, 1 mM MgCl₂, 1 μ M each of primer, 0.01 u of codUNG, 200 μ M dNTPs (with dUTP in place of dTTP), 0.1 u of Taq, and complete to 10 μ L with γ -irradiated water.

Hybridization times and temperatures are primer dependent and elongation times are defined according to user's needs.

3. With qPCR, amplification is reported by fluorescence emission of SYBR Green I intercalated in double-strand DNA. Non-specific products, which are sometimes synthesized during qPCR and are mainly primer-dimers, also lead to fluorescence

emission. A first control is the use of several non-template controls (NTCs) to monitor the cycle number required to amplify such dimers and to determine the T_m s of the various possible dimers. If similar C_t s and T_m s are observed for both samples and NTCs, the amplification of dimers can be suspected. Primer-dimers, however, may sometimes have a T_m close to that of the desired product. Gel visualization is then needed to discriminate between primer-dimers, other non-specific products, and the desired product. Direct sequencing of products is a mandatory step to authenticate ancient DNA PCR results. When amplifying a target gene present only in the species of interest, such direct sequencing can be done by Sanger sequencing (i.e., Eurofins Genomics, Ebersberg, Germany). When amplifying genes like the 16S rRNA, the PCR product is a mix of thousands of different sequences best analyzed through high-throughput sequencing (HTS). The IonTorrent platform is well suited for this.

3.3 Library Preparation for PCR Products

1. For each sample, mix 20 μ L of PCR product (diluted if necessary—*see Note 23*), 5 μ L of 10 \times enzyme buffer, 0.1 μ L of end repair enzyme, and 24.9 μ L of γ -irradiated water. Gently pipet the total volume up and down 1–2 times to mix and incubate for 30 min at 25 $^{\circ}$ C (*see Note 24*).
2. Purify end-repaired products using a 96-well plate system (for high-throughput) or in tubes by magnetic beads or silica columns according to the manufacturer's instructions (*see Note 25*). The final elution volume is usually 50 μ L. *Samples may be frozen at -20° C at this point.*
3. In a 96-well plate, add 1 μ L of a different barcoded adapter mix (A + P1) combination to each well to be used. Distribute a premix composed of 6 μ L of 5 \times ligation buffer, 1 μ L of Quick ligase, and 2 μ L of γ -irradiated water to each well and add 20 μ L of each purified end-repaired sample (*see Note 26*). Gently pipet up and down 1–2 times to mix and incubate for 30 min at 16 $^{\circ}$ C.
4. Immediately after the ligation, add binding buffer to each well according to the manufacturer's recommendation (60 μ L of NT buffer, Macherey-Nagel), mix well, and pool all samples before loading on a silica column. The binding, washing, and drying steps are performed as recommended. Elute in a volume between 30 and 50 μ L. *Samples may be frozen at -20° C at this point.*
5. Use an electrophoresis system for DNA sizing and purification to size-select your library and eliminate adapter dimers or multimers of amplicons. We use the E-Gel SizeSelect or the Caliper Labchip XT depending on the size range of amplicon size

(*see Note 27*). To determine the size range to select, add the length of the two adapters (85–87 bp depending on the barcode length, *see Note 5*) to your minimal and maximal amplification size.

6. Nick repair and amplification are made sequentially with the same reaction mix (*see Note 28*). Add 8 μL of size-selected sample, 1 μL of 10 μM Primer A, 1 μL of 10 μM Primer P1, and 10 μL of 2 \times OneTaq Hot Start Master Mix with buffer. Gently pipet up and down 1–2 times to mix, incubate for 20 min at 68 $^{\circ}\text{C}$ (nick repair step), and then amplify the library using the following program: initial denaturation at 94 $^{\circ}\text{C}$ for 5 min (94 $^{\circ}\text{C}$ for 15 s, 60 $^{\circ}\text{C}$ for 15 s, 68 $^{\circ}\text{C}$ for 40 s) for six cycles, final elongation at 68 $^{\circ}\text{C}$ for 5 min (*see Note 29*).
7. Purify the amplified libraries with a silica column or SPRI magnetic beads. Final elution volume is usually 30 μL . *Samples may be frozen at -20°C at this point.*
8. Qualitative analysis on a Bioanalyzer is recommended to check the final library product (*see Note 30*). Products saved at intermediate steps (sizing and amplification) can also be run on the same chip and compared. Libraries are quantified using a Qubit 2.0 to determine the concentration and adapt it to the emulsion PCR for IonTorrent PGM sequencing. A qPCR is also recommended to compare the new library to a known reference to ensure that the Qubit measurement corresponds to samples ligated with the two adapters.
9. Follow the manufacturer's recommendations to prepare the chip for sequencing. Depending on the heterogeneity of the PCR product(s) and the number of samples analyzed, either 314, 316, or 318 chips (V2) can be used.

3.4 Double-Stranded DNA Library Preparation

1. Mix in a 1.5 mL tube 1–500 ng of DNA extract, 3 μL of 10 \times NEBNext end repair buffer, 1.5 μL of 1 u/ μL USER enzyme, and complete to 28.5 μL with γ -irradiated water. Mix by pipetting and incubate for 1 h at 37 $^{\circ}\text{C}$ in a heating block.
2. After the cytosine deamination repair step, add 1.5 μL of NEBNext End Repair Enzyme Mix directly to the tube. Mix well by pipetting gently times and incubate for 30 min at 20 $^{\circ}\text{C}$.
3. The extract is purified with Qiagen MinElute kit, following the manufacturer's protocol, except that elution is done twice with 17 μL of γ -irradiated water (preheated at 50 $^{\circ}\text{C}$).
4. Add 10 μL of 5 \times Quick ligation reaction buffer, 1 μL of the 40 μM annealed P50X adapter, 1 μL of the 40 μM annealed P7XX adapter (*see Note 31*), 2 μL of Quick DNA ligase, and 6 μL of γ -irradiated water to the sample. Gently pipet 1–2 times to mix and incubate for 30 min at 20 $^{\circ}\text{C}$.

5. Add to the sample 3 μL of each of the 10 μM Illumina amplification forward and reverse primers P5s and P7s and 50 μL of OneTaq 2x Master Mix. Transfer the 100 μL reaction solution into a 0.2 mL PCR tube, and run in a thermocycler with the following program: 15 min OneTaq elongation step of the ligated products at 68 $^{\circ}\text{C}$, followed by six cycles involving denaturation for 20 s at 95 $^{\circ}\text{C}$, annealing for 35 s at 60 $^{\circ}\text{C}$, and primer extension for 70 s at 72 $^{\circ}\text{C}$.
6. The purification of the six-cycle-amplified library and the removal of potential artifacts (primer-dimers) are done using the NucleoMag NGS clean-up and Size Select kit:
 - (a) Add 130 μL (1.3 \times) Macherey-Nagel (MN) beads to the 100 μL of amplified library.
 - (b) Vortex and let sit for 5 min at RT.
 - (c) Quick spin and place on a magnetic rack; let sit for 2 min or until supernatant is clear.
 - (d) Discard liquid.
 - (e) Add 500 μL 80 % ethanol (freshly made).
 - (f) Twist tubes two or three times, until beads no longer stick to the tube's surface, and let sit for 2 min or until supernatant is clear.
 - (g) Discard ethanol with a pipet, being careful to remove as much as possible.
 - (h) Let dry for 2 min at RT on the magnetic rack.
 - (i) Remove from the magnetic rack, add 52 μL γ -irradiated water, and pipet ten times to mix well.
 - (j) Let stand for 2 min at RT away from the magnetic rack.
 - (k) Quick spin and place on the magnetic rack; let sit for 2 min or until supernatant is clear.
 - (l) Remove 50 μL to new tubes.
 - (m) Add 65 μL (1.3 \times) MN beads.
 - (n) Repeat **steps b–h**.
 - (o) Remove from the magnetic rack, add 25 μL γ -irradiated water or EBT, and pipet ten times to mix well.
 - (p) Let stand for 2 min at RT away from the magnetic rack.
 - (q) Quick spin and place on the magnetic rack; let sit for 2 min.
 - (r) Remove 22 μL of purified sample and place in a new tube.
7. To obtain a sufficient quantity of DNA while avoiding over-amplification (*see Note 30*), qPCR quantification of the pre-amplified library is performed using three serial dilutions (1:10, 1:100, 1:1000) and the modified Illumina P5 and P7 primers. For each sample, the amplification curves are analyzed and the

cycle number at the point between exponential phase and saturation is determined. This value is used to calculate how many cycles are needed to amplify the library. For a typical reaction, the correct amplification of the library will be 7–8 cycles less than the value determined with the 1/100 diluted sample (*see Note 32*).

8. To further amplify the library, mix 20 μL of the library, 61 μL of γ -irradiated water, 2 μL of 5 mM dNTPs (A,T,C,G), 3 μL of each 10 μM Illumina amplification primers P5s and P7s, 10 μL of the 10 \times PCR reaction buffer containing 20 mM MgCl_2 , and 1 μL of 5 u/ μL FastStart Taq DNA Polymerase. Gently pipet 1–2 times to mix and amplify the library using the following program: (95 $^\circ\text{C}$ for 20 s, 60 $^\circ\text{C}$ for 35 s, 72 $^\circ\text{C}$ for 70 s) for the appropriate number of cycles.
9. Purify the PCR using the Qiagen PCR purification kit.
10. Quantify the amplified libraries using a fluorescence-based quantification method; observe the size distribution of the amplified library and possible presence of artifacts on an Agilent Bioanalyzer 2100 (*see Note 33*).

3.5 Single-Stranded DNA Library Preparation

Prepare all enzymatic mixes prior to each step to avoid letting the beads dry between steps.

1. Dilute the purified DNA extract (between 1 fmol and 1 pmol of DNA) with γ -irradiated water to a final volume of 29 μL (*see Note 34*).
2. To optimize damaged DNA recovery prior to library preparation, add 29 μL of diluted DNA extract, 8 μL of 10 \times Circligase buffer, 4 μL of 50 mM MnCl_2 , and 0.5 μL 10 u/ μL Endonuclease VIII to a 1.5 mL tube and incubate for 1 h at 37 $^\circ\text{C}$ (*see Note 35*).
3. Add 1 u of FastAP to the above reaction and mix. Spin briefly and incubate for 10 min at 37 $^\circ\text{C}$. Incubate the reaction for 2 min at 95 $^\circ\text{C}$ to heat denature the DNA, then transfer tube to an ice-water bath, and leave for 1 min. Spin briefly.
4. To ligate the first adapter
 - (a) Add 32 μL 50 % PEG-4000, 1 μL 10 μM adapter oligo CL78, and 1 μL 100 u/ μL Circligase II. Incubate tube for 1.5–3 h at 60 $^\circ\text{C}$. *Samples may be frozen safely at -20 $^\circ\text{C}$ at this point.*
 - (b) For each sample, transfer 20 μL of MyOne C1 dynabeads into a 1.5 mL tube. Use additional tubes for more than five samples. Wash beads by placing tube on a magnetic rack for 2 min. Discard supernatant and wash beads twice with 500 μL bead binding buffer (*see Note 36*).

- (c) Resuspend beads in 250 μL bead binding buffer per sample (*see Note 37*), vortex, and transfer 250 μL of beads to each ligated sample.
 - (d) Rotate tubes slowly on a rotating wheel, making sure that the beads stay in suspension, for 20 min at room temperature.
 - (e) Spin tubes briefly, place on a magnetic rack for 2 min, and discard supernatant.
 - (f) Add 200 μL wash buffer A.
 - (g) Place tubes on the magnetic rack, wash by twisting tubes (*see Note 36*), and then leave tubes on the rack for 2 min. Discard supernatant.
 - (h) Add 200 μL wash buffer B.
 - (i) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on the rack for 2 min. Discard supernatant.
5. To create the second strand
- (a) Mix together 38.5 μL γ -irradiated water, 5 μL 10 \times Thermopol buffer, 2.5 μL dNTP mixture (5 mM of each), and 1 μL 100 μM extension primer CL9.
 - (b) Add mixture (47 μL) to beads in each tube. Vortex tubes and spin down briefly.
 - (c) Incubate tubes for 2 min at 65 $^{\circ}\text{C}$, and then cool for 1 min in an ice bath.
 - (d) Place tubes in an Eppendorf thermomixer set at 15 $^{\circ}\text{C}$, and then add 3 μL Bst 2.0.
 - (e) Incubate at 15 $^{\circ}\text{C}$ for 30 min, mixing 1000 RPM (*see Note 11*).
 - (f) Place tubes on the magnetic rack for 2 min, and discard supernatant.
 - (g) Add 200 μL wash buffer A.
 - (h) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on the rack for 2 min. Discard supernatant.
 - (i) Add 100 μL stringency wash buffer, and incubate for 3 min at 45 $^{\circ}\text{C}$ in a Thermomixer, mixing at 1000 RPM.
 - (j) Spin tubes briefly, then place on the magnetic rack for 2 min, and discard supernatant.
 - (k) Add 200 μL wash buffer B.
 - (l) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on the rack for 2 min. Discard supernatant.

6. Remove 3' overhangs remaining from the extension step
 - (a) Mix 84.5 μL γ -irradiated water, 10 μL 10 \times Tango buffer, 2 μL dNTP mix (5 mM each), 2.5 μL 1 % Tween-20, and 1 μL T4 DNA polymerase.
 - (b) Add mixture (100 μL) to beads in each tube, vortex, and spin down briefly.
 - (c) Incubate for 15 min at 25 $^{\circ}\text{C}$ in a Thermomixer, mixing at 1000 RPM.
 - (d) Add 10 μL stop solution to each tube.
 - (e) Place tubes on the magnetic rack for 2 min, and discard supernatant.
 - (f) Add 200 μL wash buffer A.
 - (g) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on rack for 2 min. Discard supernatant.
 - (h) Add 100 μL stringency wash buffer, and incubate for 3 min at 45 $^{\circ}\text{C}$ in a Thermomixer, mixing at 1000 RPM.
 - (i) Spin tubes briefly, place on the magnetic rack for 2 min, and discard supernatant.
 - (j) Add 200 μL wash buffer B.
 - (k) Place tubes on magnetic rack, wash by twisting tubes, and then leave tubes on rack for 2 min. Discard supernatant.
7. To ligate the second adapter
 - (a) Mix 71 μL γ -irradiated water, 10 μL 10 \times T4 DNA ligase buffer, 10 μL 50 % PEG-4000, and 2.5 μL 1 % Tween-20.
 - (b) Add mix (93.5 μL) to each tube, vortex, and spin down briefly.
 - (c) Add 5 μL 40 μM annealed adapters CL53/CL73 and 2 μL T4 DNA ligase to each tube, vortex, and spin down briefly.
 - (d) Incubate for 1 h at 25 $^{\circ}\text{C}$ in a Thermomixer, mixing at 1000 RPM.
 - (e) Place tubes on the magnetic rack for 2 min. Discard supernatant.
 - (f) Add 200 μL wash buffer A.
 - (g) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on rack for 2 min. Discard supernatant.
 - (h) Add 200 μL wash buffer B.
 - (i) Place tubes on the magnetic rack, wash by twisting tubes, and then leave tubes on rack for 2 min. Discard supernatant.

- (j) To elute the single-stranded library, add 50 μL EBT buffer to each tube and resuspend the beads by pipetting gently up and down.
 - (k) Incubate each tube for 1 min at 95 $^{\circ}\text{C}$ in a Thermomixer, mixing at 1000 RPM, and immediately move tube to the magnetic rack.
8. Transfer the supernatant, which contains the single-stranded library molecules, to a new tube. This non-indexed, single-stranded “proto”-library must be amplified with P7 and modified P5 indexing primers (preferably barcoded) before being ready for sequencing. Libraries can be stored at -20°C for several months.
 9. Follow the same protocol as for double-stranded library preparation (Subheading 3.4, steps 5–10) but using indexing primers P7 and modified P5.
 10. When setting up the Illumina sequencing run, be sure to replace the Read 1 sequencing primer with the custom primer CL72 using the manufacturer’s instructions for custom primers.

4 Notes

1. **Water was decontaminated by γ -Irradiation** (at least 2 kGy) using a ^{37}Cs source that had been calibrated by Fricke dosimetry [18].
2. MixG (home-made qPCR mastermix) is prepared with a high amount of BSA or HSA to allow its use with the LightCycler Carousel systems, since those systems use glass capillaries and BSA/HSA are needed to minimize DNA binding to the glass. We use the Carousel system for its sensitivity and its rapidity and to minimize carry-over contamination, each sample being in its own tube and capped after filling. The source of the albumin depends on the analyzed species. HSA is used for ancient bovine remains, and BSA in all other cases.
3. UV irradiation decontamination of reagents is performed in thin-wall, UV clear tubes for 300 s on each side in a Spectrolinker XL 1500 UV cross-linker device equipped with 254 nm UV light bulbs (Spectronics Corp., Westbury, NY, USA) corresponding to a total energy of 4.8 J/cm^2 [18]. The UV bulb must be at a close distance (we use 5 cm) from the tubes since the efficiency of the destruction of DNA molecules with UV light is a function of the square of the distance [18].
4. SYBR-Green I can be replaced by EvaGreen 25 mM diluted 1/80 in DMSO [18]. The -80°C freezing step is essential.

5. Primer sequences for IonTorrent library preparation

For sequencing on an Illumina sequencer, the adapters described in **Note 7** should be used instead.

(a) A barcoded adapter:

- Long strand: 5' CCATCTCATCCCTGCGTGTCTCC GACTCAGXXXXXXXXXXCGAT 3'
- Short strand: 5' ATCGXXXXXXXXXX 3'

(b) P1 adapter:

- Long strand: 5' CCACTACGCCTCCGCTTTCCTC TCTATGGGCAGTCGGTGAT 3'
- Short strand: 5' ATCACCGACTGCCC 3'

(c) Primer PCR A: 5' CCATCTCATCCCTGCGTGTCTC 3'

(d) Primer PCR P1: 5' CCACTACGCCTCCGCTTTCCT CTCT 3'

6. Hybridize both strands to make double-stranded adapters by mixing 20 μL of each adapter (100 μM), 5 μL of annealing buffer (10 \times), and 5 μL of γ -irradiated water. Incubate for 30 s at 95 $^{\circ}\text{C}$ in a heating block, then turn off the heat block, and allow the tubes to come to room temperature. Annealed oligos may be stored at -20°C . Final concentration of adapter is 40 μM . For IonTorrent adapters, since only the A adapter is bar-coded, it is more convenient to prepare a premix of equal volume of A and P1 adapters for each barcode.7. Primers for the double-strand library preparation are modified from the Illumina sequencing primers, Nextera or TruSeq barcodes, and amplification primers, to change the Y-shape adapter design back to the initial Solexa design with two different adapters, as this design minimizes dimer background when working with very low DNA amounts [19]. For sequencing on an IonTorrent sequencer, the adapters described in **Note 5** should be used instead. In the example below, amplification primers are in italics, barcodes are bold, and sequencing primers are in roman fonts.

(a) D7XX construct:

*CAAGCAGAAGACGGC***CATACGAGAT** **XXXXXXXXX**GT
GACTGGAGTTCAGACGTGTGCTCTTCCGATCT

(b) D50X construct:

*AATGATACGGCGAC***CCACCGAGATCTACAC**
XXXXXXXXXACTCTTTCCCTACACG
ACGCTCTTCCGATCT

(c) SLP5P7: 5'-AGATCGGAAGAG-3'

8. MinElute is used to allow small elution volume (here 17 μL). Columns are purchased independently of the purification kit.

9. P5s sequence CAAGCAGAAGACGGCATAACGAGAT
P7s sequence: AATGATACGGCGACCACCGAGAT
10. Oligonucleotides [10, 25]
- CL78: single-stranded adapter: 5'[Phosphate]-AGATCGGAAGXX-[TEG-biotin] (X=C18 spacer). Alternatively, ten C3 spacers can be used in the place of two C18 spacers.
 - CL9: 5'-tailed extension primer:
5'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT.
 - CL53: double-stranded adapter strand 1: 5'CGACGCTCTTC-[ddC] (ddC=dideoxy cytidine).
 - CL73: double-stranded adapter strand 2:
5'[Phosphate]-GGAAGAGCGTCGTGTAGGGAAA GAGTGTA.
Some protocols specify four phosphorothioate bonds at the 3' end of CL73 [25]. We obtain good results without this modification.
 - Modified P5 indexing primer:
5' AATGATACGGCGACCACCGAGATCTACAC[barcode]ACACTCTTTCCCTACACGACGCTCTTCC
Insert standard Illumina barcode of choice in the place of [barcode]. The modified P5 indexing primer with the P7 indexing primer is used to create, via PCR, full-length bar-coded Illumina adapters flanking the initial single-stranded library products.
IMPORTANT! Modifications in the P5 barcoded primer/adapter necessitate a corresponding custom sequencing primer, CL72, to be used in place of the Illumina Read 1 sequencing primer during sequencing (CL72: 5' ACACTCTTTCCCTACACGACGCTCTTCC [25]).
 - P7 indexing primer: Use a standard, non-annealed, bar-coded Illumina P7 adapter oligo as the P7 indexing primer.
11. Alternatively, a standard heating block or thermocycler with a heated lid can be used instead of Eppendorf Thermomixer Comfort, but tubes containing beads must be manually mixed periodically to prevent the beads from settling.
12. All pre-PCR work is carried out in a physically isolated high-containment laboratory in a part of the building where no DNA amplification is performed. Rooms are under positive air pressure with a gradient from low to high positive pressure from the airlock, through the extraction and purification rooms, to the pre-PCR room with the highest positive pressure. Incoming air is filtered. This installation minimizes contamination with airborne environmental DNA.

13. Many techniques exist to extract DNA from soil. Authors propose different buffers, but the general purpose is to homogenize and lyse the soil in a liquid buffer mechanically (Precellys, MoBio homogenizer, etc.) or using chemicals (SDS, NLS, etc.) while limiting the interactions between DNA and soil particles using a buffer of high ionic strength (such as NaCl 1.5 M). Different DNA extraction techniques have been found to modify the representation of bacterial populations [29, 30].
14. We use MoBio PowerSoil to obtain the least variable results and to be able to perform comparisons from one extract to another, since the various DNA purification protocols extract different bacterial populations [26].
15. Buffers usually pass through the column in a few minutes. With a vacuum pump capable of producing a vacuum of -800 to -900 mbar, buffers pass through the column in 5–10 min. Clogging of the silica membrane by particles remaining in the sample can sometimes occur during the DNA binding step. When this occurs, the remaining sample can be passed through one or more fresh columns and combined after elution. If clogging continues to occur, columns can be centrifuged multiple times ($10,000 \times g$ usually for 1 min) using 700 μL of buffer per centrifugation.
16. For maximal recovery, elution can be performed using 2×75 μL preheated (50 $^{\circ}\text{C}$) EB as follows:
 - (a) Load spin columns with 75 μL preheated EB and incubate for 2 min.
 - (b) Centrifuge for 1 min at $12,500 \times g$.
 - (c) Load spin columns with 75 μL preheated EB and incubate for 1 min in the thermo-block at 50 $^{\circ}\text{C}$.
 - (d) Centrifuge for 2 min at $12,500 \times g$.
 - (e) Pool the eluates and store at -20 $^{\circ}\text{C}$.
17. The situation is more complicated since the skeletal remains are chemically not homogeneous but rather consist of multiple chemical microenvironments, each with its specific chemistry, in which DNA preservation can be variable.
18. Since there is no specific lysis procedure to open up microbial cells, it is likely that the DNA recovery from live microbial cells is low.
19. A complete overview of the process can be seen in a movie, available at http://www.univ-paris-diderot.fr/Mediatheque/spip.php?article246&var_mode=calcul, especially after the seventh minute.
20. In environmental samples, many soil compounds can interfere with molecular techniques used downstream, such as humic and fulvic substances.

21. The 16S rRNA gene includes both conserved regions, which can be used for designing amplification primers across taxa, and nine hypervariable regions (VI–V9), which can be effectively used to discriminate between taxa [27]. Nearly every hypervariable region or combination thereof has been studied. Owing to the short size of DNA fragments in degraded ancient samples, we selected V5 (28 bp long in *Escherichia coli*). Among published primers, we selected the pair providing the best coverage according to SILVA TestPrime and SILVA SSU refNR r114 [28]. We selected a forward primer (E786F) from Baker et al. [29] and a reverse primer (926r) from Watanabe et al. [30], the pair producing a 141 bp long fragment [29].
22. To prevent carry-over contamination, we systematically use dUTP instead of dTTP in qPCR mixes [31]. Incorporation of dUTP during PCR allows for elimination of amplicons from previous PCR steps when incubation with uracil-N-glycosylase (UNG) precedes each PCR. We selected codUNG from ArcticZymes (Tromsø, Norway) for its enhanced thermostability and high efficiency [18].
23. 20 pmol of adapters are present in subsequent preparation steps and must be in excess with respect to PCR products. A maximum amount of 5 pmol of amplicon products is recommended per sample.
24. This step will create blunt-ended 5' phosphorylated DNA using two enzymes: T4 DNA polymerase and T4 polynucleotide kinase. T4 DNA polymerase fills in 5'-protruding ends and removes 3'-protruding ends, thus producing blunt ends. T4 polynucleotide kinase phosphorylates the 5'-ends of DNA.
25. We use a TECAN EVO 100 and NucleoSpin 96 PCR clean-up (Macherey-Nagel ref. 740658) to achieve automated 96-well purification.
26. This step ensures ligation of DNA fragments with adapters. We use Quick Ligase to increase the ligation reaction and reduce incubation time.
27. E-Gel is preferred when amplicons have similar sizes (within about 50 bp) whereas the Caliper XT is better suited when the amplicon sizes are more heterogeneous.
28. This step allows the removal of the small fragment of the adapters and the fill-in of the 3'-protruding end of the ligated adapter. We use OneTaq from NEB, a blend of Taq and Deep VentR™ DNA Polymerases.
29. Save 1 µL for bioanalyzer analysis if desired.
30. If libraries are amplified beyond the point at which PCR starts saturating, multimers of PCR products may form due to the

cross-hybridization of library molecules via their adapter sequences, both preventing the proper determination of fragment size distribution and DNA concentration using an electrophoresis-based system and causing abnormal chimeric sequences.

31. By using 1 μL of the 40 μM adapters, considering an average size of the ancient DNA fragment in the extract of 50 bp, the adapter excess is 2.5 \times if the starting input is 500 ng and 1250 \times if the starting input is 1 ng. The adapter concentration can be modified according to the quantity of the input material.
32. For example, considering a qPCR amplification plot obtained from quantifying the library with the 1:100 dilution: (1) qPCR was performed in a 10 μL reaction volume, whereas library amplification is performed in 100 μL . Thus, 3.5 cycles should be added to allow for ten times more end product. (2) One microliter of a 1:100 library dilution was used for measurement, whereas 20 μL of the library is used for the library amplification (2000 times more). This corresponds to roughly 11 cycles that should be deducted. Thus, 7.5 cycles should be deducted from the number of cycles just prior to the beginning of the saturation phase (around roughly 75 % of the plateau height) of the 1:100 dilution amplification curve. In this way we estimate the optimal cycle numbers for PCR.
33. If excessive adapter dimers or small inserts are present in your library, additional size selection may be desired following the library purification. If AMPure XP SPRI or NucleoMag beads are used, two rounds of purification using a bead volume of 1.3 \times sample volume are recommended to best remove adapter dimers while preserving inserts of 30 bp and above, as described in Subheading 3.4, step 6.
34. For optimal results, the purified DNA extract should first be quantified with a fluorescence-based quantification method, and length distribution can be observed in most cases using an Agilent 2100 BioAnalyzer. Positive and negative control libraries should be included with each procedure. A positive control oligonucleotide should be 5'-phosphorylated and have internal primers to allow quantification with qPCR. Use 29 μL of water for a negative control.
35. This step will remove DNA with abasic sites to maximize incorporation of damaged DNA molecules into the library. To additionally remove uracils, the result of DNA cytosine deamination, 0.5 μL codUNG (1 U) may also be added at this step. If uracils are not removed, some cytosines may be improperly rendered as thymines in the final sequence. 1 μL USER enzyme (NEB) may be used in place of Endonuclease VIII and UDG.

36. To wash beads, twist tube three turns while seated in the magnetic rack.
37. For example, if preparing three samples, resuspend beads in 750 µL.

Acknowledgments

EAB was supported by the Labex « Who am I ? ». The sequencing facility is supported by grants from the University Paris Diderot, the Fondation pour la Recherche Médicale, and the Région Ile-de-France.

References

1. Pruvost M, Schwarz R, Correia VB, Champlot S, Braguier S, Morel N et al (2007) Freshly excavated fossil bones are best for amplification of ancient DNA. *Proc Natl Acad Sci U S A* 104:739–744
2. Fortea J, de la Rasilla M, Garcia-Tabernero A, Gigli E, Rosas A, Lalueza-Fox C (2008) Excavation protocol of bone remains for Neandertal DNA analysis in El Sidron Cave (Asturias, Spain). *J Hum Evol* 55:353–357
3. Sawyer S, Krause J, Guschanski K, Savolainen V, Paabo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* 7:e34131
4. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K et al (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A* 104:14616–14621
5. Smith CI, Chamberlain AT, Riley MS, Cooper A, Stringer CB, Collins MJ (2001) Neandertal DNA. Not just old but old and cold? *Nature* 410:771–772
6. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G et al (2005) Genomic sequencing of Pleistocene cave bears. *Science* 309:597–599
7. Geigl EM (2002) On the circumstances surrounding the preservation and analysis of very old DNA. *Archaeometry* 44:337–342
8. Geigl EM (2005) Why ancient DNA research needs taphonomy. In: O'Connor T (ed) *Biosphere to lithosphere, new studies in vertebrate taphonomy*. Oxbow Books, Oxford, pp 79–86
9. Salamon M, Tuross N, Arensburg B, Weiner S (2005) Relatively well preserved DNA is present in the crystal aggregates of fossil bones. *Proc Natl Acad Sci U S A* 102:13783–13788
10. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S et al (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226
11. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M et al (2013) Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499:74–78
12. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S et al (2014) The complete genome sequence of a Neandertal from the Altai Mountains. *Nature* 505:43–49
13. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060
14. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I et al (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514:494–497
15. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK et al (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478:506–510
16. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jager G, Bos KI et al (2013) Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* 341:179–183
17. Warinner C, Rodrigues JF, Vyas R, Trachsel C, Shved N, Grossmann J et al (2014) Pathogens and host immunity in the ancient human oral cavity. *Nat Genet* 46:336–344
18. Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl EM (2010) An efficient

- multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One* 5:e13042
19. Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl EM, Grange T (2014) Library construction for ancient genomics, single strand or double strand? *Biotechniques* 56:289–290, 292–286, 298, passim
 20. Renaud G, Stenzel U, Kelso J (2015) leeHom, adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res* 42(18):e141
 21. Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L (2013) mapDamage2.0, fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29:1682–1684
 22. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
 23. Lutfalla G, Uze G (2006) Performing quantitative reverse-transcribed polymerase chain reaction experiments. *Methods Enzymol* 410:386–400
 24. Pruvost M, Geigl E-M (2004) Real-time quantitative PCR to assess the authenticity of ancient DNA amplification. *J Archaeol Sci* 31:1191–1197
 25. Gansauge MT, Meyer M (2013) Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc* 8:737–748
 26. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P et al (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* 77:1315–1324
 27. Clarridge JE 3rd (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 17:840–862
 28. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M et al (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 41:e1
 29. Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* 55:541–555
 30. Watanabe I, Kodama Y, Harayama S (2001) Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods* 44:253–262
 31. Pruvost M, Grange T, Geigl EM (2005) Minimizing DNA contamination by using UNG-coupled quantitative real-time PCR on degraded DNA samples: application to ancient DNA studies. *Biotechniques* 38 (4):569–575

INDEX

A

Absorbance ratio.....185
 Activity-based screening.....257–258
 aDNA. *See* Ancient DNA (aDNA)
 Algorithms 70, 137, 168, 253
 Alkaline phosphatase.....297
 Alpha-diversity.....222
 AMF. *See* Arbuscular mycorrhizal fungi (AMF)
 Amino acid clustering 19, 20, 201, 202
 Amoebae.....127
 Amoebozoa137
 Amplicon-based 16s ribosomal
 subunit sequencing.....210
 Amplicon mix.....80–82, 87
 Amplicon sequencing 198, 204, 221, 225, 237, 293
 Amplification bias 84, 253
 AnaEE-France 59
 Analysis parameters228
 Ancient DNA (aDNA)289–315
 Annotation abundance data..... 227, 228, 231
 Annotation automated annotation pipelines222
 Arbuscular mycorrhizal fungi
 (AMF)..... 29–52, 101–122
 community.....101
 Archaea..... 1–4, 16, 18, 24, 26, 197, 274
 Archaeologists290
 Ascocarps142–147
 AZCL-polysaccharides 267, 269
 Azo-polysaccharides.....259

B

Bacteria..... 2, 30, 47, 63, 111–113, 128,
 167, 168, 197, 226, 248, 268, 293, 299, 301
 Bacterial enrichment 36, 47, 51
 Barcoding 62, 70, 126, 168
 Basic local alignment search tool (BLAST) 20, 114, 115,
 121, 134, 135, 137, 201, 210, 213, 249, 253
 Bioinformatic analyses basic bioinformatics
 analyses106
 Bioinformatic analyses 454 sequencing bioinformatics
 analyses106
 Bioinformatics 3, 18–20, 22, 69, 114–116,
 133–135, 209–211, 240, 252, 292, 293
 pipeline..... 69, 117–119

Biomarker..... 168, 236, 237
 BLAST. *See* Basic local alignment search tool (BLAST)

C

Candidatus Glomeribacter gigasporarum
 (*CaGg*).....29, 49
 Carbohydrate-active enzymes (CAZymes).....257–271
 Carbon transfer 158–164
 Complementary DNA (cDNA)
 libraries281
 size fractionation280
 Contaminations.....42
 Cellulose..... 4, 9, 11, 32, 36, 38,
 39, 49, 89, 90, 94, 98, 197, 199, 201, 202
 Cercozoa..... 128, 132, 135
 Cesium trifluoroacetate 90, 153
 Circuligase II ssDNA ligase297
 Cloning..... 14–15, 25, 35, 50, 111,
 204, 275, 283–284
 Clusters of orthologous group (COG) 19, 20,
 213, 215, 222
 CO1 minibarcodes.....133
¹³CO₂..... 153, 154, 159, 241
 Collembola125
 Colony picker 263, 265, 266
 Community analysis 61–88, 184
 Community characterization 56, 59, 106, 107
 Confocal microscopy 33, 40–42
 Contaminations..... 3, 44, 47, 70–72, 74, 84,
 99, 108, 132, 158, 185, 187, 208, 246, 250, 252,
 253, 268, 290, 292, 293, 308, 310
 Culture-independent 183, 208
 Cytosine deamination..... 303, 313

D

Damaged DNA.....305, 313
 Data normalization.....191
 Data visualization.....223
 Databases..... 19–21, 59–60, 62, 70,
 102, 114, 126, 134, 135, 137, 168, 185, 202–205,
 209–210, 213, 215, 220, 231
 Decomposition 89, 291
 Denaturing gradient gel electrophoresis
 (DGGE)..... 237, 239, 245–249

Dereplication 211, 212, 229
 Diagnostic 59
 Directional cloning 285
 Diversity 1, 22, 62, 64, 125, 126,
 128, 131, 137, 156, 167, 193, 203, 205, 221–222,
 225, 253, 258, 260, 273, 275, 292, 301
DNA
 amplification
 dUTP 301
 hl-dsDNase 294, 295
 HVR 16S rRNA
 uracil-DNA glycosylase 295
 barcoding 102
 damage 284
 decay 291
 degradation 290
 DNA based species identification 101–122
 extraction
 CTAB method 73–74, 103, 109–110
 FastDNA® SPIN Kit for soil 103
 preparation in soil 70, 297
 purification
 from sediment samples 12
 for single-cell genomics 3, 4, 10, 15–16
 from soil samples 12, 56
 from water samples 11–12
 quantification
 bioanalyzer 86
 Nanodrop 90
 Qubit 172, 298
 traces 292
 DNA-stable isotope probing 80–81, 236
 DNase 17, 36, 47, 91, 97, 109, 153,
 170, 260, 276, 278, 294, 296
 Double-stranded DNA (dsDNA) 78–79, 87, 173,
 295, 296, 303
 Dremel multitool 300
 Duplicate ReadInferred Sequencing Error Estimation
 (DRISEE) 212, 229
E
 Earthworms 125
 Ecosystems 1, 55, 57, 59, 95, 119,
 125, 142, 167–181, 258–260
 Ecosystem services 55, 197
 Ectomycorrhizas 141–143, 145–147
 EGGNOG 215
 18S rDNA 164
 Electrophoresis 14, 72, 76, 85, 105,
 111, 112, 116, 145–146, 148, 154, 173, 174, 180,
 247, 263, 265, 276, 277, 285, 294, 298, 302, 313
 Enchytraeids 125
 Endobacteria 29–52
 Endonuclease VIII 297, 305, 313

Endophytes 156
 Enzymes 14, 25, 47, 50, 63, 71, 112,
 187, 188, 202, 203, 205, 236, 248, 257–259,
 268–270, 283, 292, 293, 295, 303, 312
 Eukaryotic microorganisms 135–137
 Evolutionary placement algorithm
 (EPA) 102, 117–119
 Expression plasmids 285
F
 Fluorescence in situ hybridization (FISH) 30, 35–36,
 45–47, 51
 Forests 90, 98, 141–142, 275
 Fosmids 7, 14, 15, 257–271
 454 GS-FLX+ based high throughput monitoring
 of AMF 116–117
 454 GS-FLX+ pyrosequencing 102, 116
 Freeze-drying 63, 80, 83
 Functional annotation 204, 205, 223
 Functional content 208, 209
 Functional gene arrays (FGAs) 183, 184
 Functional gene probing 236
 Functional traits 197, 198
 Functions 3, 50, 59, 90, 151, 168,
 201, 203–205, 207–232, 236, 237, 257, 259, 269,
 273, 308
 Fungal-bacterial interaction
 Fungal marker 94
 Fungal microbiota 30
 Fungal-specific primers 64
 Fungi 62, 63, 72, 89, 91–94, 141,
 154, 156–159, 197, 268, 274, 299
 Fungi-saprotrophic 89
G
 γ -Irradiation 308
 Gel electrophoresis 65, 76, 85, 120, 162,
 245, 276, 280
 Gene calling 211, 212
 Gene capture 167–181
 Gene prediction 210, 212
 General eukaryotic primers 135
 GenoSol 55–60
 GeoChip 183–195
Gigaspora margarita 30
 Glomeromycota 29–31
 Glycosyl hydrolase 90
 Gridding 266
 Group-specific primers 128
H
 Hemicelluloses 89
 Hidden Markov models (HMMs) 201, 205

High molecular weight DNA..... 194, 265
 High throughput annotation
 High-throughput screening.....257–271
 High-throughput sequencing (HTS) 3, 58, 61–88,
 97, 125, 128, 167, 236, 237, 252, 302
 HPLC
 Hybridization35, 45, 46, 51, 167–181,
 183–195, 274, 275, 301, 313

I

Illumina 17, 18, 64, 69, 167, 170,
 175, 177, 179, 180, 199–201, 253, 296, 304, 305,
 309, 310
 Inhibition51, 65, 84, 85, 120
 In silico PCR.....204, 205
 In situ 57, 141–149
 Internal transcribed spacer (ITS) region..... 62, 64,
 65, 69–71, 77–78, 84, 86, 101, 102, 145
 Isopycnic ultracentrifugation 153, 154, 156, 157
 ISO standard soil DNA extraction procedure58

K

Kyoto Encyclopedia of Genes and Genomes
 (KEGG) 19, 210, 213–216, 222, 223

L

LCA. *See* Lowest common ancestor (LCA)
 LeeHom293
 Lignin.....89, 197
 Litters.....62, 73, 83, 89–91, 95, 98
 Lowest common ancestor (LCA)214, 222

M

M5nr (non-redundant sequence database)209
 MapDamage.....293
 Mapping..... 117, 134, 143, 213, 216, 222
 Mating type genes141–149
 matR.....216, 227
Medicago truncatula..... 48, 159
 Metabarcoding 126, 128, 130–135
 Metagenomes 3, 21, 51, 52, 64, 198,
 201, 202, 204, 209, 212, 215, 217–220, 222, 259
 Metagenomics 3, 18, 19, 21, 22, 57, 58,
 91, 95–98, 198, 207–232, 237, 238, 249, 257–271
 libraries259
 recruitment 21, 22
 Metagenomics rapid annotation using subsystem
 technology (MG-RAST).....207–232,
 249, 293
 Metatranscriptomics.....3, 91, 95–98, 208,
 209, 273–287
 Methanol..... 153, 236, 247
 Methanol dehydrogenase..... 236, 248

Methylotroph248
 Microarray scanner 187, 191, 195
 Microarthropods.....125, 126, 128–131,
 133, 135, 136
 Microbial communities.....17, 55–57, 59, 90,
 167, 168, 183–185, 193, 208, 236, 240, 241, 275
 Microbial diversity..... 1, 55–60, 167, 198, 205, 257
 Microbiomes..... 198, 229, 292, 293
 Microscope 10, 33, 35, 36, 40, 42,
 45–47, 49, 51, 143, 144
 Microtiter plates 16, 266, 267
 Millipore sandwich method.....32, 38–39
 Mites 125, 126
 Mollicutes-related endobacteria29
 Multiplexing..... 17, 86, 200
mxnF..... 236, 237, 240, 248
 Mycelium 30, 38, 39, 61, 142, 147

N

NEBNext end repair module.....295, 296
 NEBNext Quick ligation module.....295, 296
 Nematodes..... 125, 126
 Next-generation sequencing (NGS).....17, 24, 90, 167,
 198, 204
 Nuclear ribosomal DNA (rDNA) 91, 101, 102,
 132, 145
 Nutrient cycling.....205
 Nycodenz.....264, 269

O

Oligonucleotide probes..... 35, 168, 172, 184
 Oligonucleotides91, 172, 173, 204,
 295, 296, 310, 313
 One carbon metabolism235
 Orchards..... 142, 143
 Organic matter83, 89–100, 274, 284

P

Paleogenetic.....290
 Para-Nitrophenyl-substrates.....263
 PCR. *See* Polymerase chain reaction (PCR)
 PFGE. *See* Pulsed-field electrophoresis system (PFGE)
 Photosynthates 153, 156
 Phylogenetic 1, 2, 19–21, 26, 102,
 106, 116–120, 126, 135, 168, 213, 226, 237,
 248, 249, 253
 analysis..... 21, 114–116
 Picking 15, 70, 263, 270
 Pipelines 58, 69, 70, 105, 115–117, 133,
 191, 201, 209–215, 219, 220, 228, 252, 257, 293
 Pipetting and handling robot
 Pixel intensity191
 Plant roots29, 156, 162, 274, 275

Plasmids 15, 44, 45, 50, 51, 104, 106,
111, 258, 276, 283, 285, 286
preparation..... 105, 113–114
Platform 17, 137, 191, 200, 227, 249,
270, 284, 298, 302
Polyadenylated mRNA
Polymerase chain reaction (PCR)
amplification..... 47, 64, 65, 94, 97, 103,
116, 145, 280, 284, 285
optimization 75–77
protocols 106
Preservation
for single cell genomics..... 10
sediment samples 10
soil samples 10
water samples 9, 10
Primer design 202–204
Probes 35, 46, 48, 51, 168, 172–175,
177, 180, 181, 184–186, 188, 189, 191, 195, 267
Prokaryote 70, 268
Proteinase K 6, 11, 33, 35, 42, 46, 261
Protein coding genes 219
Protein identification 213
Protists..... 125–137, 274
Protozoa
Pulsed-field electrophoresis system (PFGE) 261, 265

Q

qPCR. *See* Quantitative real-time PCR (qPCR)
QTrays..... 262, 263, 266, 267, 270
Quality control (of sequence data)..... 18
Quantitative real-time PCR (qPCR) 35, 45, 50,
51, 65, 85, 99, 156, 162, 163, 198, 204, 294, 295,
298, 301, 303, 304, 308, 313

R

Rarefaction 137, 221, 223–225
Real-time quantitative PCR (qPCR) 30, 44–45
Replicating 263
Réseau de mesures de la qualité des sols
(RMQS) 57, 59
Restriction fragment length polymorphism
(RFLP) 105, 112
Reverse transcription 91, 97, 155, 162, 280, 285
Ribonucleic acid (RNA) 211, 215, 219, 220
Ribosomal RNA (rRNA) genes..... 20, 24, 62,
101, 102, 153, 168, 211, 213, 215, 220, 236, 237,
245, 248, 249, 252, 253, 273, 274, 278, 284
annotation
clustering 215
detection 215
identification..... 215
RNA purification
from sediment samples 13

from soil samples 13
from water samples 12, 13
RNase..... 7, 10, 11, 17, 33, 42, 153,
161, 164, 277, 278, 280, 284, 293, 296, 300
RT-PCR..... 91, 94, 99, 154, 158

S

Saccharomyces cerevisiae..... 91, 285
Sample identification tag sequences 69
Sampling sediment 9, 10, 12, 13, 25
Sampling soil 10, 12, 13, 56, 59, 97, 126,
128, 131, 134, 142, 144, 145, 147, 195, 275,
277–279
Sampling water 9–12, 23
Sanger sequencing 16–17, 106–107, 114, 302
Screening 91, 126, 212, 229, 236,
249, 258–260, 262, 263, 266–267, 270
Sediment 4, 9, 10, 241, 269, 293
SEED subsystems..... 210, 214, 222
Sequence analysis
assembly..... 18–19
functional analysis..... 213, 222
gene detection..... 19–20
phylogenetic analysis 21
phylogenomic analysis 20, 21
quality control..... 18
statistics 19–20, 220
Sequence capture
Sequencing 454/Roche..... 200
Sequencing illumina 17, 308, 309
Sequencing sanger 16–17, 106–107, 114, 302
Sexual reproduction 62, 142
Shotgun libraries large-insert size 15
Shotgun libraries small-insert size 14
SHS. *See* Solution hybrid selection (SHS)
Single-cell genomics 3, 4, 15
Single-stranded DNA 187, 293, 297, 305–308
Small subunit rDNA 101, 279
SIP. *See* Stable isotope probing (SIP)
16S rRNA 11, 16, 24, 30, 50, 191,
239, 245, 248, 293, 301, 312
Software 22, 35, 45, 79, 84, 106,
114, 115, 117, 118, 133, 168, 185, 191, 208, 210,
240, 248, 253
Soil
DNA extraction..... 129–131
fungi 29, 141
microbiome..... 197–206
sampling 10, 12, 13, 56, 59, 97,
126, 128, 131, 134, 142, 144, 145, 147, 195, 275,
277–279
Solution hybrid selection (SHS)..... 168, 282–283
Species identification..... 62
Spore propagation 30, 37

Stable isotope probing (SIP) 90–94, 98,
 151–166, 198, 236, 237, 243, 246
 Structure..... 22, 40, 56, 128, 141, 142,
 145, 168, 184, 193, 198, 205, 207–232, 275, 301
 Substrates 37, 62, 63, 69, 73, 83,
 90, 91, 98, 151, 153, 154, 237, 240, 241, 246, 248,
 259, 260, 262, 266, 269, 270
 Symbionts..... 30, 153, 156
 Symbiosis..... 141

T

Taphonomy 292
 Targets 44, 45, 50, 51, 83, 93, 106,
 128, 132, 136, 156, 186, 188, 203, 240, 245, 293,
 301, 302
 Taxonomic annotation..... 118, 213, 214, 221, 222, 225
 Taxonomic content
 Taxonomic resolution 62, 132–134, 208
 Templphi 500 amplification kit..... 185, 187
 Traceability 56

Transmission electron microscopy 32–33, 40
 Truffle..... 141–149
Tuber melanosporum 141

U

Ultracentrifugation 94, 99, 157, 159–160,
 237, 243–246, 250, 251
 Unculturable bacteria..... 30, 50

W

Water samples 9, 10, 23
 Whole community genome amplification 185
 Whole genome shotgun sequencing..... 208, 210, 211

X

X-oligosaccharides 263

Y

Yeast 8, 105, 262, 276, 279