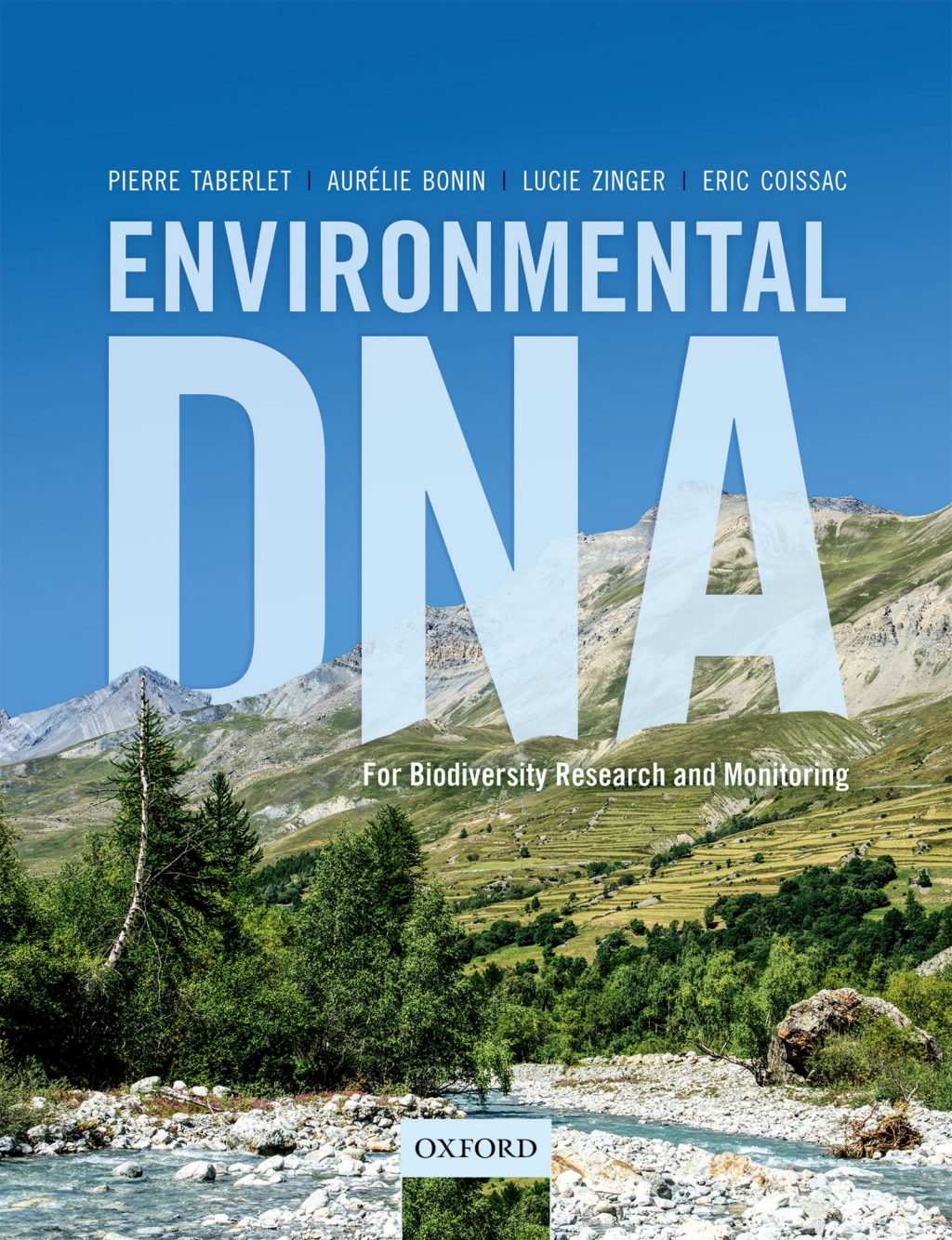
PIERRE TABERLET | AURÉLIE BONIN | LUCIE ZINGER | ERIC COISSAC

ENVIRONMENTAL DNA

For Biodiversity Research and Monitoring



OXFORD

Environmental DNA

Environmental DNA

For Biodiversity Research and Monitoring

Pierre Taberlet
Centre National de la Recherche Scientifique and Université Grenoble Alpes, France

Aurélie Bonin
Centre National de la Recherche Scientifique and Université Grenoble Alpes, France

Lucie Zinger
Ecole Normale Supérieure de Paris, France

Eric Coissac
Université Grenoble Alpes, France

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2018

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2017959944

ISBN 978–0–19–876722–0 (hbk.)
ISBN 978–0–19–876728–2 (pbk.)

DOI: 10.1093/oso/9780198767220.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Preface

For a few years, the scientific community has acknowledged that environmental DNA contains a huge amount of information about all types of organisms found in ecosystems. It is thus very attractive to try to at least partially read this information encoded in DNA molecules. The development of next-generation DNA sequencing boosted this research area, and opened the possibility of high-throughput data acquisition. However, implementing an eDNA study is not as simple as carrying out a DNA analysis on a single species. It also requires putting together many different skills, including obviously all the classical corpus of knowledge in ecology, but also skills in the field for sampling, skills at the bench for producing the sequence results, and skills in bioinformatics for dealing with massive amounts of sequence data.

As for any emerging scientific area, many preliminary or pilot experiments have been published thus far, as well as many reviews highlighting the potential of eDNA in ecology, paleoecology, archaeology, forensics, and biodiversity management. Unfortunately, despite this literature, extracting pertinent information stored in eDNA is not straightforward, and not exempt from biases. There are many potential traps in eDNA studies. The experiments must be designed very carefully, taking into account all the problems likely to occur at any step of the analysis, and include appropriate controls to ensure accurate final results. In light of these observations, we decided a few years ago to organize DNA metabarcoding schools, with now at least one session per year. Seven editions have occurred since 2012. At the beginning, the number of applicants was reasonable, however it is no longer the case due to the recent increasing interest in eDNA. The number of applicants is now far more than our teaching capacities, and we do not feel comfortable rejecting applications from many motivated young scientists. This book is an attempt to cope with the high demand from ecologists for developing research involving eDNA.

Our aim was to write a book for ecologists, who do not necessarily have a strong background in molecular genetics. As a result, some parts might look naive to experienced molecular ecologists. Moreover, it was our wish not to give very precise protocols, but to provide the background information that will ultimately enable the design of sound experiments. This book is technically oriented, as the difficulties in eDNA studies mainly derive from the technical aspects, with which ecologists are usually not familiar. When working with eDNA in ecology, the questions and concepts remain the same as in any ecological study, and we did not emphasize these aspects, as many excellent ecological textbooks are already available. Incidentally, it is interesting to note that eDNA now allows a few questions that could not be previously addressed to be tackled. Another objective of this book was to deliver a relatively comprehensive bibliographic orientation, allowing the readers to refer to the primary literature. However, due to the recent burst of eDNA studies, it becomes very difficult to be complete, and clearly, we did not cite all pertinent papers.

Finally, we must also recognize that this book reflects our view of eDNA analysis, and does not necessarily represent all the different opinions expressed in the scientific community. Globally, our objective was to favor simple and robust solutions that might not be optimal in terms of accuracy, but which can be implemented at large scales for analyzing hundreds or thousands of environmental samples.

Contents

Acknowledgments xiii

1 Introduction to environmental DNA (eDNA) 1

1.1 Definitions 1
1.2 A brief history of eDNA analysis 2
1.3 Constraints when working with eDNA 3
1.4 Workflow in eDNA studies and main methods used 4
1.5 Environmental DNA as a monitoring tool 5

2 DNA metabarcode choice and design 7

2.1 Which DNA metabarcode? 7
2.2 Properties of the ideal DNA metabarcode 8
2.3 *In silico* primer design and testing 9
2.3.1 Prerequisites 10
2.3.2 Reference sequences: description, filtering, and formatting for `ecoPrimers` 10
2.3.3 *In silico* primer design with `ecoPrimers` 11
2.3.3.1 The `ecoPrimers` output 11
2.3.4 *In silico* primer testing with `ecoPCR` 11
2.3.4.1 The `ecoPCR` output 14
2.3.4.2 Filtering of the `ecoPCR` output 16
2.3.4.3 Evaluation of primer conservation 16
2.3.4.4 Taxonomic resolution and *Bs* index 17
2.4 Examples of primer pairs available for DNA metabarcoding 19

3 Reference databases 21

3.1 Extracting reference databases from EMBL/GenBank/DDBJ 21
3.1.1 Downloading a local copy of EMBL 21
3.1.2 Identifying sequences corresponding to the relevant metabarcode 23
3.2 Marker-specific reference databases 23
3.2.1 Nuclear rRNA gene reference databases 23
3.2.2 Eukaryote-specific databases 24
3.3 Building a local reference database 25
3.3.1 PCR-based local reference database 26
3.3.2 Shotgun-based local reference database 27
3.4 Current challenges and future directions 27

4 Sampling **28**

4.1 The cycle of eDNA in the environment 28
4.1.1 State and origin 28
4.1.2 Fate 29
4.1.3 Transport 29
4.2 Sampling design 30
4.2.1 Focusing on the appropriate DNA population 31
4.2.2 Defining the sampling strategy 32
4.3 Sample preservation 33

5 DNA extraction **35**

5.1 From soil samples 35
5.2 From sediment 39
5.3 From litter 39
5.4 From fecal samples 39
5.5 From water samples 40

6 DNA amplification and multiplexing **41**

6.1 Principle of the PCR 41
6.2 Which polymerase to choose? 43
6.3 The standard PCR reaction 44
6.4 The importance of including appropriate controls 45
6.4.1 Extraction negative controls 45
6.4.2 PCR negative controls 45
6.4.3 PCR positive controls 46
6.4.4 Tagging system controls 46
6.4.5 Internal controls 46
6.5 PCR optimization 46
6.6 How to limit the risk of contamination? 48
6.7 Blocking oligonucleotides for reducing the amplification of undesirable sequences 50
6.8 How many PCR replicates? 51
6.9 Multiplexing several metabarcodes within the same PCR 52
6.10 Multiplexing many samples on the same sequencing lane 52
6.10.1 Overview of the problem 52
6.10.2 Strategy 1: single-step PCR with Illumina adapters 54
6.10.3 Strategy 2: two-step PCR with Illumina adapters 55
6.10.4 Strategy 3: single-step PCR with tagged primers 55

7 DNA sequencing **58**

7.1 Overview of the first, second, and third generations of sequencing technologies 58
7.2 The Illumina technology 59
7.2.1 Library preparation 59
7.2.2 Flow cell, bridge PCR, and clusters 60

7.2.3 Sequencing by synthesis 62
7.2.4 Quality scores of the sequence reads 63

8 DNA metabarcoding data analysis 65

8.1 Basic sequence handling and curation 65
8.1.1 Sequencing quality 65
8.1.1.1 The pros and cons of read quality-based filtering 65
8.1.1.2 Quality trimming software 68
8.1.2 Paired-end read pairing 68
8.1.3 Sequence demultiplexing 69
8.1.4 Sequence dereplication 69
8.1.5 Rough sequence curation 70
8.2 Sequence classification 70
8.2.1 Taxonomic classification 71
8.2.2 Unsupervised classification 73
8.2.3 Chimera identification 75
8.3 Taking advantages of experimental controls 76
8.3.1 Filtering out potential contaminants 76
8.3.2 Removing dysfunctional PCRs 78
8.4 General considerations on ecological analyses 80
8.4.1 Sampling effort and representativeness 81
8.4.1.1 Evaluating representativeness of the sequencing per PCR 81
8.4.1.2 Evaluating representativeness at the sampling unit or site level 81
8.4.2 Handling samples with varying sequencing depth 83
8.4.3 Going further and adapting the ecological models to metabarcoding 84

9 Single-species detection 85

9.1 Principle of the quantitative PCR (qPCR) 85
9.1.1 Recording amplicon accumulation in real time via fluorescence measurement 85
9.1.2 The typical amplification curve 86
9.1.3 Quantification of target sequences with the Ct method 86
9.2 Design and testing of qPCR barcodes targeting a single species 87
9.2.1 The problem of specificity 87
9.2.2 qPCR primers and probe 88
9.2.3 Candidate qPCR barcodes 88
9.3 Additional experimental considerations 88
9.3.1 General issues associated with sampling, extraction, and PCR amplification 88
9.3.2 The particular concerns of contamination and inhibition 88

10 Environmental DNA for functional diversity 90

10.1 Functional diversity from DNA metabarcoding 90
10.1.1 Functional inferences 90
10.1.2 Targeting active populations 92

10.2 Metagenomics and metatranscriptomics: sequencing more than a barcode 93
10.2.1 General sampling constraints 94
10.2.1.1 Optimization of the number of samples 94
10.2.1.2 Enrichment in target organisms 94
10.2.1.3 Enrichment in functional information 95
10.2.2 General molecular constraints 96
10.2.3 From sequences to functions 96
10.2.3.1 Assembling (or not) a metagenome 97
10.2.3.2 Sorting contigs or reads in broad categories 97
10.2.3.3 Extracting functional information via taxonomic inferences 98
10.2.3.4 Functional annotation of metagenomes 98

11 Some early landmark studies 99

11.1 Emergence of the concept of eDNA and first results on microorganisms 99
11.2 Examining metagenomes to explore the functional information carried by eDNA 100
11.3 Extension to macroorganisms 102

12 Freshwater ecosystems 104

12.1 Production, persistence, transport, and detectability of eDNA in freshwater ecosystems 104
12.1.1 Production 104
12.1.2 Persistence 104
12.1.3 Transport/diffusion distance 105
12.1.4 Detectability 106
12.2 Macroinvertebrates 106
12.3 Diatoms and microeukaryotes 106
12.4 Aquatic plants 107
12.5 Fish, amphibians, and other vertebrates 107
12.5.1 Species detection 107
12.5.2 Biomass estimates 108
12.6 Are rivers conveyer belts of biodiversity information? 108

13 Marine environments 110

13.1 Environmental DNA cycle and transport in marine ecosystems 110
13.2 Marine microbial diversity 111
13.3 Environmental DNA for marine macroorganisms 112

14 Terrestrial ecosystems 114

14.1 Detectability, persistence, and mobility of eDNA in soil 114
14.2 Plant community characterization 116
14.3 Earthworm community characterization 117
14.4 Bacterial community or metagenome characterization 117
14.5 Multitaxa diversity surveys 119

15 Paleoenvironments **121**

15.1 Lake sediments 121
15.1.1 Pollen, macrofossils, and DNA metabarcoding 121
15.1.2 Plants and mammals from Lake Anterne 121
15.1.3 Viability in the ice-free corridor in North America 122
15.2 Permafrost 125
15.2.1 Overview of the emergence of permafrost as a source of eDNA 125
15.2.2 Large-scale analysis of permafrost samples for reconstructing past plant communities 125
15.3 Archaeological midden material 126
15.3.1 Bulk archaeological fish bones from Madagascar 126
15.3.2 Midden from Greenland to assess past human diet 126

16 Host-associated microbiota **127**

16.1 DNA dynamics 127
16.2 Early molecular-based works 127
16.3 Post-holobiont works 128

17 Diet analysis **131**

17.1 Some seminal diet studies 131
17.1.1 Proof of concept—analyzing herbivore diet using next-generation sequencing 131
17.1.2 Assessing the efficiency of conservation actions in Białowieża forest 132
17.1.3 Characterizing carnivore diet, or how to disentangle predator and prey eDNA 133
17.1.4 Analyzing an omnivorous diet, or integrating several diets in a single one 133
17.2 Methodological and experimental specificities of eDNA diet analyses 135
17.2.1 eDNA sources 135
17.2.1.1 Feces 135
17.2.1.2 Gut content 135
17.2.1.3 Whole body 135
17.2.2 Quantitative aspects 136
17.2.2.1 Relationship between the amount of ingested food and DNA quantity in the sample 136
17.2.2.2 Quantifying DNA with PCR and next-generation sequencing 137
17.2.2.3 Empirical correction of abundances 138
17.2.3 Diet as a sample of the existing biodiversity 138
17.2.4 Problematic diets 139

18 Analysis of bulk samples **140**

18.1 What is a bulk sample? 140
18.2 Case studies 140

18.2.1 Bulk insect samples for biodiversity monitoring 140
18.2.2 Nematode diversity in tropical rainforest 141
18.2.3 Marine metazoan diversity in benthic ecosystems 141
18.3 Metabarcoding markers for bulk samples 141
18.4 Alternative strategies 143

19 The future of eDNA metabarcoding 144

19.1 PCR-based approaches 144
19.1.1 Single-marker approach 144
19.1.2 Multiplex approach 144
19.2 Shotgun-based metabarcoding 145
19.2.1 Without enrichment by capture 145
19.2.2 With enrichment by capture 146
19.3 Toward more standardization 146
19.3.1 For sound comparisons across studies 146
19.3.2 For environmental monitoring 147
19.4 Next-generation reference databases 148
19.5 Open questions 148
19.5.1 What will be the impact of new sequencing technologies on eDNA analysis? 148
19.5.2 Will some specific repositories be developed for DNA metabarcoding? 148
19.5.3 Will metabarcoding provide quantitative results? 149
19.5.4 Will metabarcoding be fully integrated into ecological models and theories? 150
19.5.5 How do we train students and managers to effectively integrate this tool into academic and operational ecological research and monitoring? 150

Appendix 1 151
Appendix 2 217
Appendix 3 220
References 223
Index 247

Acknowledgments

It would not have been possible to write this book without the fruitful discussions about environmental DNA and metabarcoding we have had over the years with colleagues and participants in the different metabarcoding schools. Among them, we are particularly thankful to Miklós Bálint, Guillaume Besnard, Julien Bessière, Kristine Bohmann, Frédéric Boyer, Christian Brochmann, Anthony Chariton, Jérôme Chave, Corinne Cruaud, Bruce Deagle, Marta De Barba, Tony Dejean, Mary Edwards, Francesco Ficetola, Roberto Geremia, Simon Jarman, Stefaniya Kamenova, Håvard Kauserud, Carla Martin Lopes, Christelle Melo de Lima, Céline Mercier, Ludovic Orlando, Johan Pansu, Jan Pawlowski, François Pompanon, Dorota Porazinska, Gilles Rayé, Tiayyba Riaz, Maurizio Rossetto, Heidy Schimann, Wasim Shehzad, Min Tang, Philippe Thomsen, Wilfried Thuiller, Philippe Usseglio-Polatera, Alice Valentini, Alain Viari, Eske Willerslev, Patrick Wincker, Meng Yao, Nigel Yoccoz, Douglas Yu, and Xin Zhou. We gratefully acknowledge the help of Amaia Iribar, Philippe Gaucher, Ludovic Gielly, Christian Miquel, Delphine Rioux, and Marie-Odile Taberlet in the field or at the bench. We are also very grateful to Bryony Taberlet for her help with language proofreading. Finally, we must acknowledge that we greatly appreciated the help and encouragement of Ian Sherman, Bethany Kershaw, and Lucy Nash from Oxford University Press throughout the process of completing this book.

July 2017
Aurélie Bonin
Eric Coissac
Pierre Taberlet
Lucie Zinger

CHAPTER 1

Introduction to environmental DNA (eDNA)

Environmental DNA (eDNA) is becoming a key component of the ecologists' and environmental managers' toolbox. Such a strong enthusiasm was recently sparked by the development of next-generation sequencers. In 10 years, our sequencing capabilities have indeed been multiplied by at least four orders of magnitude.

Environmental microbiology aside, the emergence of eDNA studies has been surprisingly slow when considering the small number of articles published during the few years following the commercialization of the first next-generation sequencers in 2005. This is probably due to the relatively high costs associated with this new type of sequencing and to the characteristics of eDNA. As a complex mixture of DNA from different organisms, possibly degraded and in low concentrations, eDNA can indeed be more difficult to analyze than DNA originating from the fresh tissues of a single organism. Another impediment is that eDNA analysis requires the combination of many different skills, from classical ecology to bioinformatics, including molecular biology techniques.

The aim of this introductory chapter is to give definitions of what eDNA is, to present its history, to highlight the different steps of an eDNA study, and to give an overview of the different types of eDNA methods implemented in research or biodiversity management. It also points the reader to the chapter(s) where the important aspects of eDNA analyses are addressed in detail.

1.1 Definitions

Environmental DNA is a complex mixture of genomic DNA from many different organisms found in an environmental sample (Taberlet *et al.* 2012a). Soil, sediment, water, or even feces are considered as environmental samples, which can also include the material resulting from filtering air or water, from sifting sediments, or from bulk samples (e.g., the whole insect content of a Malaise trap). Alternatively, environmental DNA can be defined from another perspective (i.e., the objective of the study). In this case, eDNA corresponds to DNA extracted from an environmental sample with the aim of obtaining the most comprehensive DNA-based taxonomic or functional information as possible for the ecosystem under consideration. Total eDNA contains both intracellular and extracellular DNA (Levy-Booth *et al.* 2007; Pietramellara *et al.* 2009). Intracellular DNA originates from living cells or living multicellular organisms that are present in the environmental sample. Extracellular DNA results from cell death and subsequent destruction of cell structures, and can be degraded through physical, chemical, or biological processes. For example, DNA molecules can be cut into smaller fragments by nucleases. After its release, extracellular eDNA may be adsorbed by inorganic or organic surface-reactive particles such as clay, sand, silt, and humic substances.

If we are to identify the taxa present in an eDNA sample, two approaches can be considered, mainly based on PCR (polymerase chain reaction; Mullis & Faloona 1987; Saiki *et al.* 1985, 1988). When the aim is to determine the presence or absence of a single species, the best solution is often to favor a species-specific approach, generally based on quantitative PCR (e.g., Logan *et al.* 2009). Alternatively, a more general approach based on targeted PCR (Saiki

Environmental DNA for Biodiversity Research and Monitoring. Pierre Taberlet, Aurélie Bonin, Lucie Zinger, & Eric Coissac, Oxford University Press (2018).
DOI: 10.1093/oso/9780198767220.001.0001

et al. 1988; White *et al.* 1989) or on shotgun sequencing (Deininger 1983) has the potential of revealing the presence of all species within a clade. This last approach is called "DNA metabarcoding."

The expression "DNA metabarcoding" was first used in 2011 (Pompanon *et al.* 2011; Riaz *et al.* 2011). It corresponds to the simultaneous DNA-based identification of many taxa found in the same environmental sample. Generally, DNA metabarcoding involves examining metabarcode sequences amplified from eDNA. A metabarcode consists of a short and taxonomically informative DNA region flanked by two conserved regions serving as primer anchors for the PCR (see Chapter 2). DNA metabarcoding can also be performed by shotgun sequencing of eDNA, without any metabarcode amplification (Taberlet *et al.* 2012b). Shotgun sequencing involves the sequencing of random DNA fragments from a DNA extract, generally using next-generation sequencers (Glenn 2011). However, taxa identification based on shotgun sequencing is difficult to achieve as it requires high sequencing depths and extensive reference databases for taxonomic assignment.

In the early days of DNA metabarcoding, many different terminologies were coined to designate PCR-based identification of multiple taxa at the same time: ecometagenetics (Porazinska *et al.* 2010), ecogenomics (Chariton *et al.* 2010), environmental barcoding (Hajibabaei *et al.* 2011), metataxogenomics (Terrat *et al.* 2012), and metasystematics (Hajibabaei 2012). Microbiologists, who first routinely analyzed eDNA to have access to uncultivable microorganisms, often improperly used metagenomics to mean DNA metabarcoding. Indeed, when working with eDNA, microbiologists have three main objectives: (i) identifying the microbial taxa present in environmental samples; (ii) identifying their potential biochemical functions via the analysis of coding genes; and (iii) assembling whole genomes of uncultivable microorganisms. The expression "DNA metabarcoding" is more appropriate when referring to taxa identification, while the two last objectives (i.e., the functional aspects and the assembly of genomes) are more clearly called to mind by the word "metagenomics." The confusion between these two terminologies has arisen from a seminal article with "metagenomics" in the title, which combined shotgun sequencing and 16S rDNA-based taxonomic identification (Tringe *et al.* 2005).

Aside from these concepts, metatranscriptomics is the analysis of a complete set of ribonucleic acid (RNA) molecules extracted from environmental samples to examine gene expression and regulation at the sampling time. This methodology is hence usually employed to assess both the taxonomic and functional components of the examined sample. Metatranscriptomics remains challenging, primarily because the half-life of messenger RNA (mRNA) is short (Selinger *et al.* 2003) and because total RNA is mainly composed of ribosomal RNA (rRNA) that does not provide direct information about functional aspects.

1.2 A brief history of eDNA analysis

Figure 1.1 is a chronology illustrating the brief history of eDNA analysis and its milestones. The history of eDNA started in 1987, with the report of an extraction protocol for eDNA found in sediments (Ogram *et al.* 1987). Only three years later, and surprisingly about 10 years earlier than the subsequent papers, the first DNA metabarcoding study was published (Giovannoni *et al.* 1990). This pioneering work analyzed the diversity of the 16S rRNA gene in bacterioplankton sampled in the Sargasso Sea, using PCR followed by cloning (but see also Ward *et al.* 1990 using a similar idea, but starting from rRNA). Metagenomics was initiated in 1998, by cloning and sequencing fragments of soil eDNA for identifying new pathways for the synthesis of bioactive molecules in uncultivated microorganisms (Handelsman *et al.* 1998). At the beginning of the 2000s, metagenomics and DNA metabarcoding based on cloning became commonplace in microbiology. In 2003, the first DNA metabarcoding article focusing on macroorganisms showed that it was possible to retrieve megafaunal (mammoth, bison, horse) and ancient plant DNA from permafrost, and DNA of extinct ratite moa from cave sediments (Willerslev *et al.* 2003).

The first metagenomics and metabarcoding studies relied on cloning to isolate single DNA fragments from a complex mixture prior to Sanger sequencing. DNA fragments were inserted into cloning vectors (e.g., plasmids or bacteriophages), which allowed for their isolation and multiplication

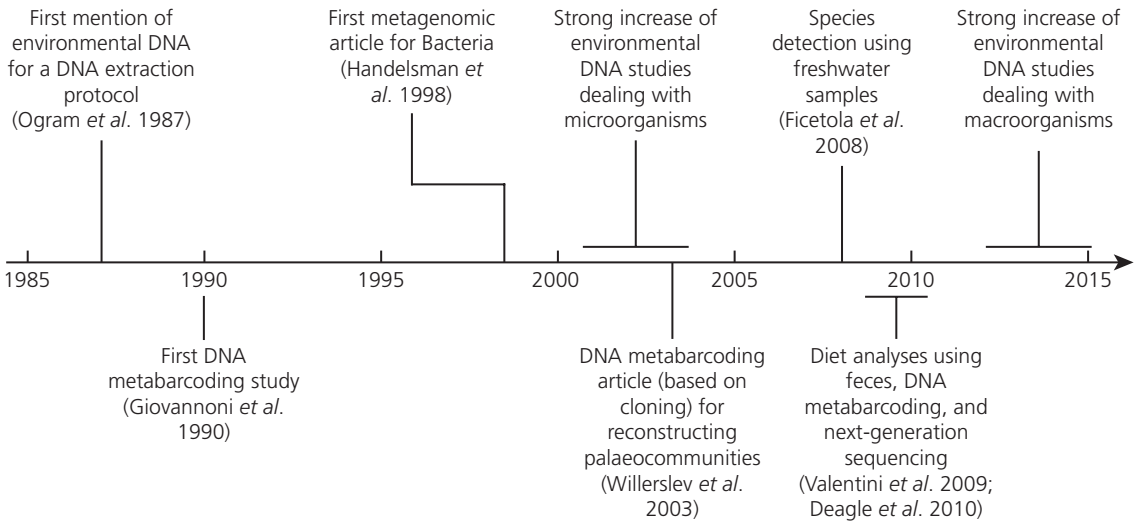


Fig. 1.1 Overview of the emergence of eDNA studies.

in the cells of a suitable host, such as *Escherichia coli*. Fortunately, this expensive and time-consuming cloning step was made unnecessary by the outbreak of next-generation sequencing after 2005 (Shendure & Ji 2008), which further stimulated metabarcoding and metagenomics. By the end of the 2000s, application of DNA metabarcoding was extended to macroorganisms, first for diet analyses using feces as a source of DNA (Deagle *et al.* 2010; Pompanon *et al.* 2012; Valentini *et al.* 2009), and then for water and soil eDNA studies (Ficetola *et al.* 2008; Sønstebø *et al.* 2010). More recently, eDNA analysis from freshwater macroorganisms led to many publications, dealing mainly with single-species detection (Dejean *et al.* 2012; Goldberg *et al.* 2011; Jerde *et al.* 2011), but also aiming at identifying multiple taxa through metabarcoding (Thomsen *et al.* 2012b; Valentini *et al.* 2016). In parallel, marine macrofauna and meiofauna were also analyzed using either water samples (Thomsen *et al.* 2012a), multilayered settlement surfaces (autonomous reef monitoring structures; Leray & Knowlton 2015), or sifted sediments (Chariton *et al.* 2010, 2015). If studies of soil microorganisms using eDNA can now be considered as routine, this is different for soil meio and macrofauna, for which only a handful of studies have been completed (Baldwin *et al.* 2013; Bienert *et al.* 2012; Pansu *et al.* 2015a; Wu *et al.* 2011). However, there is no doubt that soil macroorganisms will also be extensively investigated via eDNA in the near future.

After Willerslev *et al.'s* seminal study in 2003, analysis of ancient eDNA has also begun to be a very attractive approach for gaining insight into past communities, using either DNA metabarcoding (Epp *et al.* 2015; Giguet-Covex *et al.* 2014; Pansu *et al.* 2015b) or metagenomics (Smith *et al.* 2015). However, despite this recent enthusiasm of the scientific community for eDNA analysis, some of the technical aspects remain challenging, both at the bench and bioinformatics levels.

1.3 Constraints when working with eDNA

One of the main characteristics of eDNA is the heterogeneity of the extracts obtained from environmental samples. Thus, working with eDNA is usually not as straightforward as when working with DNA extracted from a tissue sample of a known plant or animal species. A wide range of situations can be encountered: from concentrated high-quality DNA without enzyme inhibitors (comparable to DNA extracted from tissues), to highly diluted and degraded DNA (similar to the extracts obtained in ancient DNA studies). Furthermore, it has been shown that different DNA extraction protocols are not as

equally efficient at removing PCR inhibitors depending on the sample type (e.g., varying amounts of humic substances in soil samples) and that this can lead to different results (Frostegård *et al.* 1999; Martin-Laurent *et al.* 2001). As a consequence, there is no simple and standard protocol suitable for the analysis of all types of eDNA. In this context, the objective of the following chapters is to help adjust the experimental protocols according to the question and experimental constraints, in order to carry out a sound eDNA study.

1.4 Workflow in eDNA studies and main methods used

Figure 1.2 describes the main steps of an eDNA study and the alternative strategies that can be adopted for eDNA analysis. The first one consists in targeting a single species using standard or quantitative PCR and is already popular for many taxa (e.g., Ficetola *et al.* 2008; Goldberg *et al.* 2011; Jerde *et al.* 2011; Thomsen *et al.* 2012b). The second strategy relies on PCR-based assays aiming at detecting all taxa from a given taxonomic group such as Bacteria (e.g., Tringe *et al.* 2005; Sogin *et al.* 2006), Fungi (e.g., Blaalid *et al.* 2012; Tedersoo *et al.* 2014), plants (e.g., Kartzinel *et al.* 2015; Sønstebø *et al.* 2010; Yoccoz *et al.* 2012), eukaryotes (e.g., Baldwin *et al.* 2013; Chariton *et al.* 2010, 2015; De Vargas *et al.* 2015), earthworms (Bienert *et al.* 2012; Pansu *et al.* 2015a), fish (e.g., Kelly *et al.* 2014a; Thomsen *et al.* 2012a; Valentini *et al.* 2016), and so on. Metagenomics, the third strategy which is based on shotgun sequencing of eDNA without any targeted PCR, is extensively employed for studying the functional characteristics of genomes, mainly of microorganisms (see review in Simon & Daniel 2011).

At the beginning of an eDNA study, when uncertainties remain about the experimental protocols to be implemented, it is highly advisable to carry out pilot experiments. Such exploratory trials allow for the adjustment of parameters in order to design a reliable full experiment. At this stage, it is important to foresee all possible errors and artifacts that can happen during the course of the full experiment, including problems during sampling, eDNA extraction, PCR, sequencing, and data analysis. With these potential problems in mind, the full experiment must be designed in a way that will prove that the obtained results do not originate from errors or artifacts. Ideally, it should include negative extraction controls, negative PCR controls, and positive PCR controls. For example, an inappropriate bioinformatic treatment of eDNA sequences might induce erroneous results. Therefore, it is extremely important to be able to detect these during the data analysis step using a few positive controls, so as to modify the bioinformatic pipeline accordingly if this is necessary.

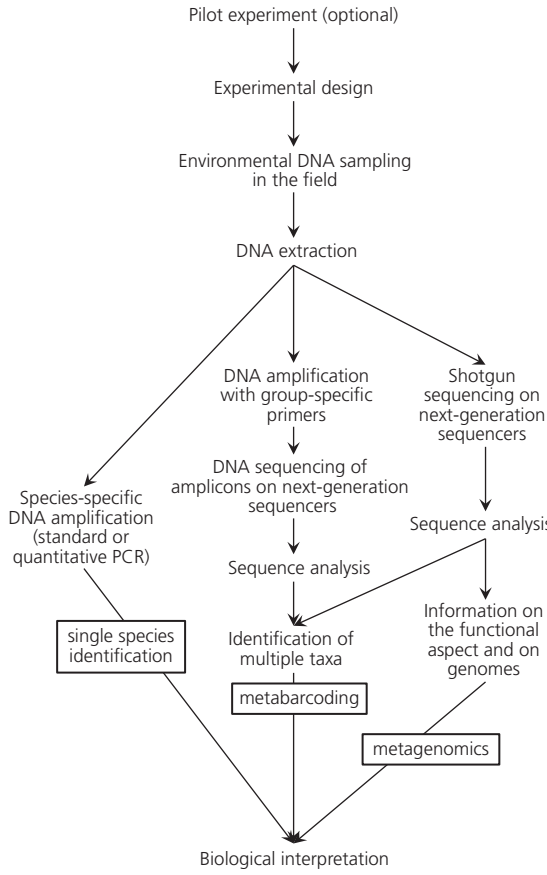


Fig. 1.2 The main steps of an eDNA study, showing the three possible approaches: single-species identification, metabarcoding, and metagenomics. The same molecular method, shotgun sequencing, can lead to metabarcoding or to metagenomics (if the objective is the taxonomic or functional aspects, respectively).

Planning the full eDNA experiment is obviously crucial and any mistake at this stage can strongly compromise the entire study (see Appendix 3 for a checklist concerning the experimental design). Key parameters to be decided are:

(i) the different controls to be included at various steps of the process (extraction negative controls, PCR negative controls, PCR positive controls of known composition, replicated samples, blind samples, and so on);
(ii) the sampling strategy (how many samples, how many sample replicates, how to spatially distribute the samples, at which time of the year, and so on);
(iii) the sample preservation method and the DNA extraction protocol (should the samples be preserved before DNA extraction, or do they have to be extracted immediately in the field to avoid degradation and/or microorganism development; which extraction protocol to choose according to the scientific question, logistic constraints, financial aspects, and so on);
(iv) the protocol for DNA amplification in DNA metabarcoding (which metabarcode(s) to analyze, which multiplexing strategy to adopt according to the number of samples and the sequencing platform, and so on);
(v) the sequencing strategy (should the sequencing be done in-house, or should it be outsourced; with which sequencing platform, and so on);
(vi) the strategy for data analysis.

The following chapters provide key information for each step of an eDNA study, including design of new metabarcodes (Chapter 2), choice of reference databases for taxonomic assignment (Chapter 3), sampling design (Chapter 4), DNA extraction (Chapter 5), DNA amplification (Chapter 6), DNA sequencing (Chapter 7), bioinformatic analysis of metabarcoding data (Chapter 8), methods for single-species identification (Chapter 9), and all the different aspects of metagenomics (Chapter 10).

Outsourcing one or several steps of the eDNA workflow is sometimes a realistic option to consider. Sequencing can now be easily and relatively cheaply outsourced to one of the many companies or common facilities proposing next-generation sequencing services. A few structures already offer full eDNA analyses, either including or excluding the sampling step, but obviously without the biological interpretation. In any case, any outsourced eDNA study should include blind samples, for which the provenance or any other characteristics are not known by the person in charge of the experiments. These are very helpful for assessing the reliability and reproducibility of the process.

1.5 Environmental DNA as a monitoring tool

Beyond research in ecology, eDNA has also proven to be a useful material for biodiversity monitoring purposes. For example, single-species detection is often applied to track rare species, or invasive species in the early stages of invasions, mainly in aquatic ecosystems (e.g., Deagle *et al.* 2003; Dejean *et al.* 2012; Jerde *et al.* 2011; Mächler *et al.* 2014; Nathan *et al.* 2015; Tréguier *et al.* 2014).

DNA metabarcoding also has a huge potential for the biomonitoring of different types of ecosystems. Several studies have already tried to adjust and standardize experimental protocols. In marine environments, for instance, it is now possible to assess the impact of pollution on eukaryotes in sediments (Chariton *et al.* 2015), or of fish farming on benthic Foraminifera communities (Pawlowski *et al.* 2014). In freshwater ecosystems, standardized metabarcoding protocols are already available for surveying fish and amphibians (Valentini *et al.* 2016), as well as diatoms (Apothéloz-Perret-Gentil *et al.* 2017; Visco *et al.* 2015; Zimmermann *et al.* 2015). Monitoring macroinvertebrates for assessing water quality is also a field of intense research (Hajibabaei *et al.* 2011, 2012; Thomsen *et al.* 2012b) that should soon lead to normalized approaches. In terrestrial ecosystems, metabarcoding can identify thousands of insects collected in Malaise traps in a single experiment, with the aim of taking decisions in restoration ecology and systematic conservation planning (Ji *et al.* 2013). More surprising is the application of metabarcoding to the limitation of birdstrike hazards at airports via the analysis of bird gut content (Coghlan *et al.* 2013). However, the development of environmental management via DNA metabar-

coding is highly dependent upon the availability of extensive reference databases for taxonomic assignment for the examined metabarcodes (Chapter 3), and upon the robustness of the experimental protocols (Chapter 19).

Decisions guiding environmental management can be based not only on taxonomic information, but also on functional capability. Metagenomics thus has a role to play in assessing the effects of anthropogenic pressures on the potential functions of microorganisms in ecosystems. Integration of metagenomics into environmental monitoring campaigns should therefore allow a better management of human impact on ecosystems (Kisand *et al.* 2012).

CHAPTER 2

DNA metabarcode choice and design

DNA barcodes are short, standardized genetic markers used for the taxonomic identification of isolated specimens (e.g., CBoL Plant Working Group 2009; Hebert *et al.* 2003a). Their characteristics are optimized for this purpose, and do not necessarily meet the requirements for a DNA metabarcoding experiment where many species must be identified simultaneously, often by analyzing low quality DNA. For this application, markers with other properties must be selected. To avoid confusion, hereafter we distinguish between DNA barcodes for classical taxonomic identification, and DNA metabarcodes for eDNA-based biodiversity surveys.

2.1 Which DNA metabarcode?

In any DNA metabarcoding study, the choice of the metabarcode is crucial and can greatly impact the end results. For example, when testing a metabarcoding protocol to assess freshwater invertebrate biodiversity, Elbrecht and Leese (2015) found that their metabarcode, derived from the standard cytochrome c oxidase subunit I (COI) barcode for animals, was unsuccessful in retrieving species abundance and biomass. This was due to differential polymerase chain reaction (PCR) efficiency among species. It is thus necessary to overcome the idea that a marker published for a given clade and application is always the best marker in a different context. Most importantly, such reflection should come early in the experimental process, as the decisions made at this stage can influence other aspects of the study (sampling, lab experiments, ecological interpretation, and so on).

Several elements should be carefully considered to make an informed choice. First, the taxonomic group of interest should be clearly defined, as well as the level of taxonomic resolution required to answer the question at hand. For example, in the diet analysis of the omnivorous brown bear, De Barba *et al.* (2014) characterized the overall plant component of the diet using a universal plant metabarcode. As it is not highly resolutive within a few plant families (Asteraceae, Cyperaceae, Poaceae, Rosaceae), they also resorted to family-specific metabarcodes to increase resolution within these families. The aim of the experiment is another point to consider. Preferential amplifications are more easily tolerated when it comes to detect a few indicator species, however, they are detrimental for exhaustive or quantitative biodiversity surveys. Another criterion to consider is the potential presence of DNA from non-target organisms in the collected samples, and whether its amplification is prejudicial. For example, several primers designed for amplifying fungal markers are well known to co-amplify plants. If one is conducting a study on anaerobic Fungi found in the rumen of ruminants, it might be better to select a specific primer pair that does not amplify plant DNA coming from the ruminant diet. Alternatively, when the target and non-target groups are too closely related, like in the analysis of a vertebrate predator's diet, it is possible to use blocking oligonucleotides that hinder amplification of the predator's DNA (see Chapter 6 for the design and implementation of blocking oligonucleotides, and Chapter 17 for examples). The expected level of DNA degradation will also constrain the size of the selected metabarcode: shorter metabarcodes (<100–150 bp) should be favored in case of highly degraded eDNA, such as eDNA originating from feces, while microbiome studies can usually accommodate longer metabarcodes (<250–300 bp), especially when targeting intracellular DNA. Finally, the better solution is to sometimes

Environmental DNA for Biodiversity Research and Monitoring. Pierre Taberlet, Aurélie Bonin, Lucie Zinger, & Eric Coissac, Oxford University Press (2018).
DOI: 10.1093/oso/9780198767220.001.0001

resort to a custom-made metabarcode especially designed for the examined taxon. However, this solution can only be contemplated if one has access to a sufficient number of reference sequences spanning the entire studied group, with no taxonomic bias, and where it is possible to identify a short variable fragment flanked by two conserved regions (i.e., a fragment displaying the characteristics of the ideal metabarcode; see Section 2.2). When it comes to metabarcode selection, all the aforementioned points will enter the equation to varying extents, and the final choice will ultimately be a matter of compromises, like many other experimental decisions.

2.2 Properties of the ideal DNA metabarcode

In an ideal world, the perfect DNA metabarcode is a DNA fragment as short as possible, displaying a highly variable sequence, and flanked by two conserved regions (Fig. 2.1). The central variable region is discriminative for all species of the target group, that is, its sequence is uniquely associated to a given species and not shared with others (Fig. 2.2). Note that this definition includes intraspecific polymorphism. On the contrary, the two flanking conserved regions are identical across the target group, but different in non-target taxa. These conserved regions correspond to the sites where the DNA metabarcoding primers will anneal perfectly, ensuring unbiased amplification of the targeted species, while preventing that of undesirable taxa. For this reason, it is usually a bad idea to design primers in protein-coding regions, where variation typically occurs every three nucleotides due to the redundancy of the genetic code. Indeed, DNA barcode markers, like COI for animals, have been shown to perform poorly and favor amplification of some target taxa over others (Clarke *et al.* 2014; Deagle *et al.* 2014). Furthermore, our experience shows that degenerating too many bases in primers is not a very satisfactory way to deal with variation in primer-annealing regions. Finally, an ideal metabarcode should be located in a genomic region that is exhaustively and accurately documented across all targeted taxa in reference databases (i.e., no missing species, no sequence or assignment errors) in order to obtain unambiguous taxonomic identifications.

The size of the ideal metabarcode is highly dependent on the number of target taxa to distinguish. In theory, a metabarcode of n nucleotides allows discriminating 4^n species. This means that 10 nucleotides should be sufficient to discriminate 1,048,576 different species (for the sake of comparison, this is close to the estimated number of Diptera on Earth). In practice, even if there is no metabarcode approaching this figure, a 16S rRNA fragment as short as 30 nucleotides is enough to differentiate all earthworm species across the world, and it even shows some intraspecific variability (Bienert *et al.* 2012).

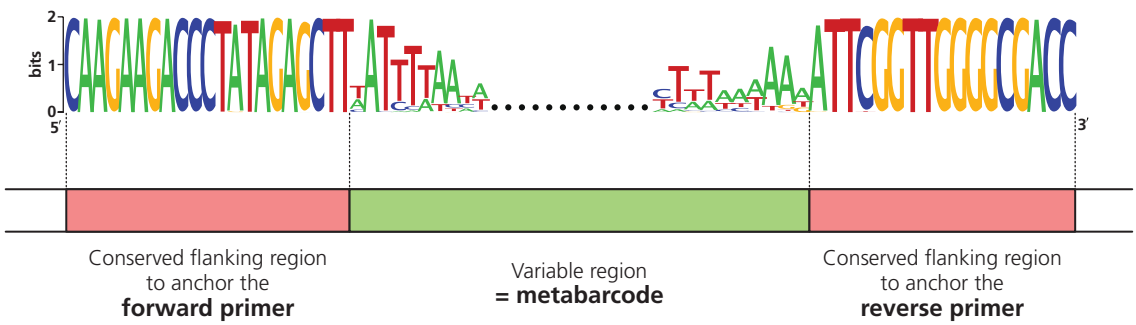


Fig. 2.1 Example of a variable metabarcode with its conserved flanking sequences. This example is based on the Lumb01 primer pair (see Bienert *et al.* 2012 and Appendix 1) targeting the suborder Lumbricina (earthworms). All Lumbricina sequences were extracted from the release 126 of EMBL using ecoPCR (1,973 sequences). Each logo consists of stacks of symbols (A, C, G, T), with one stack for each position in the nucleotide sequence. The overall height of the stack corresponds to the nucleotide conservation at that position across the Lumbricina lineage and is expressed in bits (a value of 2 indicates a perfect conservation, while 0 means the same probability for the four nucleotides). The height of each symbol within the stack indicates the relative frequency of each nucleotide at that position.

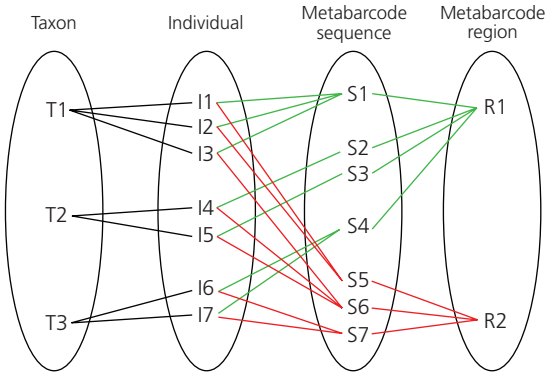


Fig. 2.2 Relationships between taxa, individuals, metabarcode sequences and metabarcode regions (adapted from Figure 1 in Ficetola *et al.* 2010). In a perfect metabarcode system, each metabarcode sequence is uniquely associated with one taxon, and not shared with others. In this figure, this is the case of the barcode system targeting region 1 (in green), and taxon 2 even shows some intragroup polymorphism. Conversely, the system targeting region 2 (in red) is not ideal, as metabarcode sequence S6 cannot discriminate between taxa 1 and 2.

In the real world, one always should compose with either lack of strict conservation of the flanking regions, or lack of taxonomic resolution for the chosen metabarcode, or more generally both. For a given metabarcoding system, two indexes are useful to evaluate the extent of departure from the perfect metabarcode: the coverage index (*Bc*), corresponding to the ratio between the number of amplified target taxa and the total number of target taxa, and the specificity index (*Bs*), defined as the ratio between the number of taxonomically discriminated taxa and the number of amplified taxa (Fig. 2.3; Ficetola *et al.* 2010). It must be noted that these two ratios are highly dependent upon the reference database they are estimated from. If this reference database contains sequencing errors, misassigned sequences, over or underrepresented taxa, this will influence the *Bc* and *Bs* values. As a result, some metabarcode systems are bound to evolve, as sequence databases are completed and/or curated constantly. For example, our initial primer pair for Fungi, originally published in Epp *et al.* (2012), was recently modified to take into account new Glomeromycota sequences published in the European Molecular Biology Laboratory (EMBL) database (see Section 2.4, and metabarcodes Fung01 and Fung02 in Appendix 1).

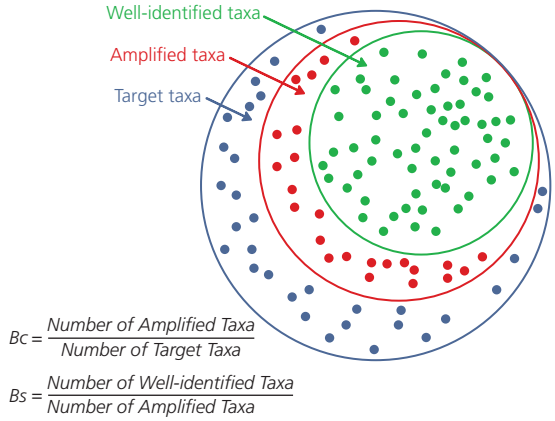


Fig. 2.3 Calculation of the *Bc* (coverage) and *Bs* (specificity) indexes.

2.3 *In silico* primer design and testing

Sometimes, metabarcodes that are already available and tested are not entirely satisfying for the purpose of an analysis. *In silico* primer design is then an alternative increasingly worth considering, especially nowadays where an exponential number of sequences is added every day to public sequence databases like the EMBL database, for a wider range of organisms. Indeed, one prerequisite for designing a robust and efficient metabarcoding system *in silico* is to have access to a set of reference sequences truly representative of the taxonomic group of interest, with reliable taxonomic annotation and low levels of sequencing errors. Additionally, it might be interesting to widen the set of examined sequences to non-target sequences, if the aim is to design primer pairs that preferentially amplify the target clade. It is unfortunate, however, that many of the sequences submitted to public databases do not include the conserved priming sites used for their amplification, which limits their value for primer design.

Obviously, the selected sequences need to correspond to the same locus or genomic regions. One can then decide to build on *a priori* knowledge of the studied organism (i.e., to focus on loci known to display interspecific variation while harboring conserved regions that can serve as primer anchors). It is the case, for instance, of the internal transcribed spacer (ITS) region in Fungi (Schoch *et al.* 2012), or

the 16S rRNA gene in Bacteria and Archaea (Fox *et al.* 1977; Pace 1997). Alternatively, one can choose an approach without *a priori*, i.e., without selecting a particular genomic region first. This entails working on whole mitochondrial, chloroplastic, or prokaryote genomes, and possibly even complete eukaryote genomes in the near future.

Once the working set of reference sequences has been obtained, two approaches can be adopted for the primer design itself. The first one relies on sequence alignments, like the method described by Walters *et al.* (2011), and is thus particularly appropriate for small sets of references with few indels. Primer anchoring regions are identified by screening the alignment for conserved regions. The second approach does not require the sequences to be aligned and it is based instead on a pattern search, like in the program `ecoPrimers` (Riaz *et al.* 2011). As there is no alignment step in this approach, it can work with any set of sequences, even whole or ganelle or prokaryote genomes. In the following paragraphs, we illustrate this approach by showing how to design new metabarcoding primers for Bacteria on whole bacterial genomes using `ecoPrimers`. We then test these new primers by running an *in silico* PCR with the `ecoPCR` program. This example is intended as a basic tutorial, so we also provide all the input and output files (available at http://www.oup.co.uk/companion/taberlet), and the associated Unix commands.

2.3.1 Prerequisites

The `ecoPrimers` (http://metabarcoding.org/ecoprimers) and `ecoPCR` (http://metabarcoding.org/ecopcr) programs are required to run this tutorial, more exactly to design and test primers *in silico*, respectively. In addition, the OBITools program suite (http://metabarcoding.org/obitools) is helpful for sequence handling, formatting, and filtering.

2.3.2 Reference sequences: description, filtering, and formatting for `ecoPrimers`

`ecoPrimers` and `ecoPCR` require a sequence database and a taxonomy database. The sequence database is constituted by the set of complete bacterial genomes downloaded from the European Bioinformatics Institute (EBI) Ensembl genome ftp site (ftp://ftp.ensemblgenomes.org/pub/bacteria/) in June 2012. The taxonomy database was downloaded from the National Center for Biotechnology Information (NCBI) ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz) at the same date.

In a taxonomy database, each taxon is designated by a unique integer value, commonly known as a taxid. A taxid identifies unambiguously a taxon, but is neither universal nor permanent. It is valid for a given taxonomy database (e.g., NCBI or SILVA) and a given version of this database. From a release of the database to another, some taxids can be added, changed, or removed. Working with different databases or different versions of the database can lead to inconsistencies. It is therefore important to keep track of the taxonomy database used for an analysis and of its version. When formatted for the OBITools, the taxonomy database consists of at least three files describing the tree structure of the taxonomy: Taxonomy.ndx, Taxonomy.rdx, and Taxonomy.tdx.

To avoid overrepresentation of well-studied bacterial genera (e.g., *Streptococcus* or *Clostridium*), the downloaded sequences were further subsampled to produce a database containing only one randomly selected sequence per bacterial genus. Ultimately, our set of reference sequences contained 517 whole-genome sequences in a FASTA format (file name: `ReferencesSequences.fasta`). Each sequence is annotated with a compulsory taxid.

First, the `ReferenceSequences.fasta` file was converted to a format combining sequence and taxonomic information, the `ecoPCR` database format, which is required to run `ecoPrimers` (Command 2.1).

Command 2.1

```
obiconvert -d Taxonomy --fasta --ecopcrdboutput=FILE1_ReferenceDatabase \
ReferenceSequences.fasta
```

This command produces five files whose names start with the "FILE1_ReferenceDatabase" prefix.

2.3.3 *In silico* primer design with `ecoPrimers`

In the second step, the `ecoPrimers` program was run on the `ecoPCR` database just created (Command 2.1). The name of the `ecoPCR` database was specified with the *-d* option (Command 2.2).

Command 2.2

```
ecoPrimers -d FILE1_ReferenceDatabase -e 3 -l 30 -L 280 -3 3 > \
FILE2_BacteriaPrimers.ecoprimer
```

The only two other mandatory parameters are the minimum and maximum metabarcode lengths (excluding primers) specified by the *-l* and *-L* options. All the other parameters have default values (see http://metabarcoding.org/obitools/doc/scripts/ecoPrimers.html for details), but it might be important to adjust them to fine-tune the primer design process. In particular, the default values can be too stringent to search for primers in a large taxonomic group like Bacteria. Here, we allowed each primer to exhibit at most three mismatches with the priming site on the amplified sequence (*-e* option). To ensure good amplification, we also disallowed mismatches within the three last nucleotides at the primer 3′-end of each primer (*-3* option) as this would strongly impede PCR efficiency (Wu *et al.* 2009). Another important parameter is the *-r* option that specifies the taxid of the clade for which the primer pair is optimized. The taxid of the clade whose amplification should be avoided can be defined via the *-i* option. This last option is interesting when there is a risk of amplifying a substantial proportion of non-target organisms.

2.3.3.1 The `ecoPrimers` output

The first part of the `ecoPrimers` output file summarizes the different parameter settings used for the primer design. The second part is a table listing the primer pairs identified with these settings, as well as the characteristics associated with the corresponding metabarcode system (Box 2.1).

A complete description of the `ecoPrimers` output can be found in http://metabarcoding.org/obitools/doc/scripts/ecoPrimers.html, so hereafter, we will focus only on the characteristics echoing the properties of the ideal metabarcode (see Section 2.2). The *Bc* and *Bs* indexes (columns 16 and 18, respectively, in red in Box 2.1) should of course be maximized, while the maximum and average metabarcode lengths (columns 20 and 21, respectively, in blue in Box 2.1) should be minimized.

Besides, it should be noted that `ecoPrimers` will identify all potential primers in a given conserved region, and combine them with all suitable primers from other conserved regions, as long as the resulting metabarcode fulfills the requirements. As a result, the different metabarcodes proposed by the program are often variants of the same metabarcode systems (Fig. 2.4). For example, there are only minor differences among primer pairs 0, 1, 2, 3, and 6 listed in Box 2.1 as regards their *Bc* and *Bs* indexes, or their maximum and average metabarcode lengths, because they all target the same variable region (Fig. 2.4).

This feature of the `ecoPrimers` output can actually be exploited to refine the selected primer pair. Indeed, it is sometimes advantageous to extend or shift the primer(s) from one or a few nucleotides in 5′ or 3′ to equilibrate the two primer melting temperatures (*Tm*). This can also be done to avoid primer dimers, 3′ complementarity, or secondary structures like hairpins (Fig. 2.5). Moreover, a good rule is, if possible, to avoid thymines at the 3′-end, to ensure good primer specificity (Kwok *et al.* 1990). For Bacteria, we ultimately selected two primers representing a good compromise in terms of *Bc, Bs*, metabarcode length, having similar *Tm*s, and being not susceptible to secondary structures (see the primer sequences in Fig. 2.4). These primers happen to target the V5–V6 regions of the 16S rRNA gene, a locus where conserved regions are known to be interspersed with highly variable ones (Schloss 2010).

2.3.4 *In silico* primer testing with `ecoPCR`

Now that bacterial primers have been designed, it can be useful to have a better indication of their

Box 2.1 Top of the `ecoPrimers` output file FILE2_PrimersForBacteria.ecoprimer. This corresponds to the first seven primer pairs suggested by `ecoPrimers`, out of a total of 683.

```
# ecoPrimer version 0.3
# Rank level optimisation: species
# max error count by oligonucleotide: 3
#
# strict primer quorum: 0.70
# example quorum: 0.90
#
# database: FILE1_ReferenceDatabase
# Database is constituted of 517 examples corresponding to 517 species
# and 0 counterexamples corresponding to 0 species
#
# amplifiat length between [30,280] bp
# DB sequences are considered as linear
# Pairs having specificity less than 0.60 will be ignored
#
0  ACGACACGAGCTGACGAC  GGATTAGATACCCTGGTA  60.5  44.8  49.5  25.3  11  8  GG  514  0  0.994  514  0  0.994  508  0.988  246  275  258.68
1  CACGACACGAGCTGACGA  GGATTAGATACCCTGGTA  60.5  44.8  49.5  25.3  11  8  GG  514  0  0.994  514  0  0.994  508  0.988  247  276  259.68
2  CACGACACGAGCTGACGA  GATTAGATACCCTGGTAG  60.5  44.8  48.2  30.4  11  8  GG  513  0  0.992  513  0  0.992  507  0.988  246  275  258.68
3  ACGACACGAGCTGACGAC  ATTAGATACCCTGGTAGT  60.5  44.8  48.5  35.7  11  7  GG  513  0  0.992  513  0  0.992  507  0.988  244  273  256.68
4  CGTGCCAGCAGCCGCGGT  GACTACCAGGGTATCTAA  69.6  53.6  49.5  25.6  14  8  GG  516  0  0.998  516  0  0.998  507  0.983  254  259  255.94
5  CGTGCCAGCAGCCGCGGT  GGACTACCAGGGTATCTA  69.6  53.6  51.6  28.3  14  9  GG  516  0  0.998  516  0  0.998  507  0.983  255  260  256.94
6  ACGACACGAGCTGACGAC  GATTAGATACCCTGGTAG  60.5  44.8  48.2  30.4  11  8  GG  513  0  0.992  513  0  0.992  507  0.988  245  274  257.68
7 . . .
```