

INVITED REVIEW

Bioinformatic challenges for DNA metabarcoding of plants and animals

ERIC COISSAC,* TIAYYBA RIAZ* and NICOLAS PUILLANDRE†‡

*Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France,

†Atheris Laboratories, Case postale 314, CH-1233 Bernex-Geneva, Switzerland, ‡Museum National d'Histoire Naturelle, Département Systematique et Evolution, UMR 7138, 43, Rue Cuvier, 75231 Paris, France

Abstract

Almost all empirical studies in ecology have to identify the species involved in the ecological process under examination. DNA metabarcoding, which couples the principles of DNA barcoding with next generation sequencing technology, provides an opportunity to easily produce large amounts of data on biodiversity. Microbiologists have long used metabarcoding approaches, but use of this technique in the assessment of biodiversity in plant and animal communities is under-explored. Despite its relationship with DNA barcoding, several unique features of DNA metabarcoding justify the development of specific data analysis methodologies. In this review, we describe the bioinformatics tools available for DNA metabarcoding of plants and animals, and we revisit others developed for DNA barcoding or microbial metabarcoding. We also discuss the principles and associated tools for evaluating and comparing DNA barcodes in the context of DNA metabarcoding, for designing new custom-made barcodes adapted to specific ecological question, for dealing with PCR and sequencing errors, and for inferring taxonomical data from sequences.

Keywords: bioinformatics, data analysis, DNA metabarcoding, primer design

Received 2 October 2011; revision received 9 February 2012; accepted 15 February 2012

Introduction

Ecological studies typically require a determination of the species involved in the ecological processes. Acquisition of such biodiversity data for plants and animals using morphological characteristics to identify field collected samples requires both a significant sampling effort and a range of taxonomic expertise that is rarely available within a single scientific group. The recent development of DNA-based methods for species identification, known as DNA barcoding (Hebert *et al.* 2003a), has drastically simplified this identification step. It is especially useful for those cases in which only the remains of the specimens are available, or when the morphological stage of the studied specimens is not conducive to making a reliable identification. DNA barcoding relies on a small piece of the genome found in a broad range of species and usually located on the mito-

chondrial genome for animals or the chloroplast genome for plants. The DNA sequence of the selected region is used as a supplementary characteristic for identification and is called, by way of analogy, a barcode. The Consortium for the Barcode of Life (CBOL <http://www.barcodeoflife.org/>) has standardized this method by progressively defining the barcode markers to be used for each taxonomic group. The standardized barcode markers for animals and plants are, respectively, a fragment of the cytochrome oxidase I gene (COI) (Hebert *et al.* 2003b) and a fragment of the ribulose 1,5-bisphosphate carboxylase gene (*rbcL*), combined with a fragment of the maturase gene (*matK*) (Lahaye *et al.* 2008, CBOL Plant Working Group 2009). For fungi, CBOL recommends the internal transcribed spacer (ITS) of the ribosomal DNA. CBOL also contributed to this method by developing the corresponding sequence reference database for these markers. It is accessible via the barcode of life data system (BOLD—Ratnasingham & Hebert 2007) web site (<http://www.boldsystems.org>). Collecting DNA

Correspondence: Eric Coissac, Fax: +33(0)4 76 51 42 79; E-mail: eric.coissac@inrialpes.fr

sequences of these markers, this curated database makes the link between a species and the specimen from which the sequence has been obtained. Although useful, DNA barcoding simplifies only the taxonomic aspect of establishing the list of species present in an ecosystem; it does not help to reduce the sampling effort. The development of next generation sequencing (NGS) offers an opportunity to address this problem. If we accept that all living organisms contaminate their environment with their DNA (Willerslev *et al.* 2003; Andersen *et al.* 2012; Bienert *et al.* 2012), it should be possible to collect this DNA, amplify barcode markers from it with PCR, and sequence the resulting amplicons using NGS. The set of DNA sequences produced can then serve as a proxy for the biodiversity present in the collected samples. This approach, called DNA metabarcoding, relies on the same principles underlying that of DNA barcoding, (i.e. linking a DNA sequence to a species). Microbiologists have used this approach for several years (Sogin *et al.* 2006), and it has allowed insights into bacterial and fungi diversities not otherwise easily obtainable. However, for plants and animals, it is often only remains or traces rather than complete specimens that are present in environmental samples, and thus, DNA sequences may be the only evidence that they are in the study area. By relaxing the need for an intensive sampling effort, DNA metabarcoding opens the door to high-throughput biodiversity assessment for plants and animals (Taberlet *et al.* 2012a,b). In this context, DNA metabarcoding has already been implemented for several ecological applications, including the reconstruction of past plant and animal communities (Willerslev *et al.* 2007; Sønstebo *et al.* 2010; Epp *et al.* 2012), animal diet assessment (see review in Pompanon *et al.* 2012) and earthworm surveys (Bienert *et al.* 2012). Relying as it does on the DNA shed by living organisms into their environment, DNA metabarcoding for plants and animals is subject to some specific constraints. The two main ones relate to the initial quality of the DNA used (Valentini *et al.* 2009a) and also the need to identify several organisms simultaneously from a unique PCR product that derived from one environmental sample (Valentini *et al.* 2009b). These constraints affect both the choice of barcode markers and the approach to analysing the results. While DNA barcoding is a well-established research field with an abundant literature available on the different strategies and methods employed for analysing data, the field of DNA metabarcoding is of more recent origin (Valentini *et al.* 2009a). Our objective in this review is to present a survey of the bioinformatics tools that have been developed specifically for plant and animal biodiversity assessment using metabarcoding, and also to revisit those developed for DNA barcoding or

microbial metabarcoding. Bioinformatics tools are useful for DNA metabarcoding approaches from the very conception of the experiment, where they can be used to select and design the most appropriate markers. These tools can then be used to identify sets of tags that allow multiplexing of samples, and finally for the post-sequencing filtering of data and determination of taxonomical identities.

Finding and testing suitable DNA metabarcodes

New markers are required for metabarcoding applications

The primary goal of DNA metabarcoding is the simultaneous identification of large sets of taxa present in single environmental samples (Taberlet *et al.* 2012a). This requires the careful selection of barcode markers that accord with the aims of the study. Various studies have shown that the standard barcode markers defined by CBOL for taxonomical identification of plants and animal specimens are not entirely adaptable for environmental applications (Valentini *et al.* 2009b; Ficetola *et al.* 2010). Two properties of conventional DNA barcodes limit their effectiveness in DNA metabarcoding contexts. First, classical barcoding approaches favour those barcode regions that can identify closely related species. To achieve a high resolution at the species level, standard markers usually need to be longer than 500 bp (Hebert *et al.* 2003b). This is not a problem for conventional barcoding applications that use a single pure specimen from which the DNA extracted is of high quality and sufficient quantity for analysis (Hebert *et al.* 2003a) or for microorganism DNA metabarcoding studies that extract DNA from a pool of living organisms (Sogin *et al.* 2006). Unfortunately, for plant and animal metabarcoding applications, researchers are usually working from organism remains. DNA degradation in the environment often prevents the recovery of PCR fragments longer than 150 bp, impeding barcode recovery (Goldstein & Desalle 2003; Wandeler *et al.* 2007). Second, the DNA metabarcoding requires simultaneously amplifying DNA from all of the species in the same tube. If these amplicons are to be representative of the species present in the environment, it is important to limit the risk of over-amplifying some species relative to others; even more so if we wish the amplicons to be representative of species' relative abundance (Murray *et al.* 2011). Standard barcodes proposed by CBOL belong to parts of genes that encode for proteins. Because of the degeneration of the genetic code, a high variability exists at the DNA sequence level and thus at the primer binding site. To address these phenomena, standard markers require that distinct primer sets be

used for each major taxonomic group (Meusnier *et al.* 2008). This is incompatible with the demands of comprehensive analyses of environmental samples, which call for equilibrated co-amplification of all species. Because of these limitations, nonstandard barcode markers need to be selected for metabarcoding applications. This requires comparing several candidate barcodes and identifying new DNA regions that respect all of the constraints imposed by the DNA metabarcoding method.

The important properties of a suitable metabarcode

DNA metabarcoding applications require short amplicons and robust PCR conditions to achieve unbiased amplification from a mixture of several DNA templates. These restrictions mean that the selected markers must correspond to a genomic loci flanked by two highly conserved priming sites to simplify PCR conditions and reduce differences in amplification among the different DNA templates (Bellemain *et al.* 2010). Moreover, the chosen region should be able to discriminate among most of the amplified taxa. In contrast to conventional barcoding, species-level identification is not always the most important quality for metabarcoding applications; identification at a higher taxonomic level (e.g. family, order, etc.) is often sufficient (Valentini *et al.* 2009a). A short barcode marker that assigns most of the individuals to a reasonable taxonomic level (species, genus or even family) can thus be considered to be a good metabarcode (Riaz *et al.* 2011).

How to test metabarcodes?

Several DNA regions have been proposed as barcodes; the region selected strongly influences the output of a study. It is therefore important to select those markers most suitable for any given study. In this context, an a priori knowledge of the relative quality of a barcode region is very important, and a method to perform a formal comparison between barcodes is desirable. Some studies have already compared loci potentially usable as barcodes; however, they usually inspect barcodes for a particular clade (Vences *et al.* 2005; Nijman & Aliabadian 2010; Luo *et al.* 2011). A more standardized method to test barcode regions was recently proposed by Ficetola *et al.* (2010). This method relies on an 'electronic PCR' application (ecoPCR) and two formal indices—barcode coverage (B_c) and barcode specificity (B_s)—to measure, respectively, the conservation of the primers and the capacity of the amplified region to discriminate between taxa. Results of *in silico* and *in vitro* PCRs can differ somewhat; ecoPCR results are nevertheless useful for improving the performance of a study, in

that they facilitate a preliminary comparison of several DNA regions to identify the most appropriate barcodes for the study's aims. Bellemain *et al.* (2010) successfully used this approach to compare fungal ITS primers. This study showed that some ITS primers, when used with higher numbers of mismatches, potentially introduce bias during PCR amplification and that different primer combinations or different parts of the ITS region should be analysed in parallel, or that alternative ITS primers should be searched for.

Identifying new metabarcode regions

Finding new primer pairs (and their associated barcode regions) that are suitable for particular environmental applications is one of the major challenges of metabarcoding. A number of programs have been developed to help biologists with the primer design process. Although most of these were not designed for DNA metabarcoding, they are suitable for *sensu stricto* barcoding. The best-known programs for primer design are Primer3 (Rozen & Skaletsky 2000; Koressaar & Remm 2007) and QPRIMER (Kim & Lee 2007), both of which optimize primers by considering thermodynamic properties for a single target DNA sequence. As these programs optimize primers according to only one sequence, they cannot take into account the interspecific variability in priming sites; they therefore cannot guarantee the universality of the designed primers. For the purposes of DNA metabarcoding primer design, this software family is interesting largely for its utility during the final step of the primer design in, in which it can be used to check the thermodynamical properties of primers designed by other programs. UniPrime (Bekaert & Teeling 2008) and Primaclade (Gadberry *et al.* 2005) work on multiple alignments of short sequences (i.e. gene sequences) and allow the design of primers that amplify several homologous sequences. Similarly, the Amplicon program (Jarman 2004) also allows the selection of primers specific to a group of aligned sequences and the exclusion of a counterexample data set. Even if it is possible to design efficient algorithms to analyse multiple sequence alignments, the main problem with such an approach is the alignment computation time, which increases quickly with the data set size despite the heuristics used by all multiple alignment software programs (see Thompson *et al.* 1994; Notredame *et al.* 2000; Edgar 2004). This restrains the use of these tools to a set of short regions compatible with multi-alignment software capacity. Nevertheless, some pre-aligned data sets exist, which allow these programs to be applied efficiently. Among these pre-aligned data sets, we can cite RDPII, which collects 16S bacterial sequences (Cole *et al.* 2009), UNITE, which is dedicated

to fungi ITS (Abarenkov *et al.* 2010), and the ARB Silva initiative (Pruesse *et al.* 2007), which collects LSU and SSU sequences from all organisms, allowing the design of primers for other groups.

Some other programs were developed to design barcode markers for specific applications. Among them are PrimerHunter (Duitama *et al.* 2009) and Greene SCPrimer (Jabado *et al.* 2006), which design primer pairs specific to virus subtypes. PrimerHunter designs primer pairs by analysing multiple target sequences. However, its thermodynamical model for primer selection, which allows mismatches, limits its efficiency in that it can only be run on a very small sequence database. Greene SCPrimer relies on the processing of a multiple sequence alignments. It constructs phylogenetic trees to identify candidate primers and uses a greedy algorithm to identify the minimum sets of primers that amplify all of the aligned sequences.

Although there exist a number of such primer design programs, most of them suffer from the limitations of computational efficiency when required to run on the large number of sequences available in public databases. ecoPrimers (Riaz *et al.* 2011) addresses such problems by coupling a simple and highly efficient syntactic algorithm and the two quality indices developed by Ficetola *et al.* (2010). Looking for conserved regions among a set of sequences is equivalent to looking for repeats in a long sequence corresponding to the concatenation of the data set. The main advantage of this approach is to allow ecoPrimers to work on an unaligned data set. As drawback, efficiency of the ecoPrimers' algorithm is highly related to the composition of the sequence set. Thus, the computation time increases significantly if the set of sequences is only constituted of few sequences or of highly similar sequences. A second potential limitation of ecoPrimers is that no thermodynamic model is used during primer selection, and therefore, it would be useful to check whether the proposed primer pairs exhibit all required thermodynamic properties for a good PCR amplification. As with the Amplicon program (Jarman 2004), ecoPrimers takes into account a set of example sequences and eventually a set of counterexample sequences.

How to reliably multiplex samples?

Next generation sequencing platforms produce too many sequences per run for just one environmental sample (i.e. up to 1 million and up to 6 billion reads, respectively, for the 454 GS FLX/Roche machine and for the Illumina HiSeq 2000 machine). Physical procedures allow the splitting of a 454 GS FLX/Roche run into as many as eight regions, and a HiSeq/Illumina run can analyse up to 16 different lanes instar of

samples. Despite this, the number of reads per sample is usually still too high for common usage; this waste can enormously increase the cost of analysis. Thus, it is necessary to multiplex several independent samples in one run. The number of reads per sample can be adjusted by selecting the appropriate number of samples to mix. Sequence reads must be separated according to their corresponding samples after sequencing. To associate a sequence read to a sample, we add a small DNA word at one or both of its extremities. Depending on the authors, these DNA words can be called MID (for Multiplex Identifier), 'barcodes', or tags. To limit confusion in this review, we call them tags. Soon after the development of pyrosequencing (Margulies *et al.* 2005), several publications used this strategy to multiplex samples (Binladen *et al.* 2007, Hoffmann *et al.* 2007). NGS manufacturers now offer kits for multiplexing, but they are limited to a small number of samples (e.g. the Multiplexing Sample Preparation Oligonucleotide Kit from Illumina (Illumina, Inc., San Diego, CA, USA) contains 12 tags which allow the multiplexing of 96 samples). It is possible, if certain guidelines are observed, to produce sets of tags to index as many samples as needed. The goal is to design tags so that they cannot be confused with anything else, even given sequencing errors (Parameswaran *et al.* 2007, Hamady *et al.* 2008). This means that at least '*n*' differences must exist between two tags, where '*n*' is determined according to the tagging strategy and the number of reads produced by the sequencer (Coissac 2012). To elaborate a list of tags, several strategies have been explored. Parameswaran *et al.* (2007) defined the constraints to design a set of 64 decamers with at least four differences. Using an approximation based on auto-corrective binary Hamming code, Hamady *et al.* (2008) proposed a solution to efficiently compute lists of tags. In their article, they proposed a list of 1544 octamer tags with at least three differences. Because of the approximation used, a small number of tag-pairs have <3 differences, but the efficiency of the method justifies its consideration. Frank (2009) published BARCRAWL, a program with a naïve algorithm dedicated to designing tagged primers that combine the sequencing primers for the 454 GS FLX/Roche, the tag, and the amplification primer. In the OBITools package (<http://www.grenoble.prabi.fr/trac/OBITools>), the oligoTag program allows a set of tags to be designed according to user-defined lexical constraints. Based on graph theory and on the notion of clique (i.e. complete sub-graph), it produces exact solutions, but as clique searching is a difficult problem in computer science, the computation time and the memory required for some sets of parameters can be huge. Using this program, we were able to generate a set of 36 octamers with at least five differences between them

that allowed the tagging of $36^2 = 1296$ samples (Table 1). When standard multiplexing sample kits proposed by the manufacturers are used, the assignment of each read to a sample is assumed by the sequencing platform. For all other cases, the operation has to be performed later. The *BARCRAWL* package includes a second program, *BARTAB* (Frank 2009), for demultiplexing reads after the sequencing procedure. *OBITools* includes *ngsfilter*, a generic program to perform this task. *ngsfilter* allows the sorting of reads for all common tagging strategies: tagging on one or both sides of the amplicon, and with identical or different tags at each extremity.

How to analyse very large data sets?

Initial environmental barcoding studies were conducted using the Sanger sequencing method and the largest data set had a few thousand sequences (e.g. Zinger *et al.* 2009). Today, with NGS, the size of data sets has increased considerably. A run of Illumina HiSeq 2000 produces approximately 6 billion reads. This number is not by itself meaningful in terms of the computation required to analyse it. A naïve calculation of the number of pages needed to print such a data set with a standard word processor indicates that the amount of paper required would create a stack 48 km high. There is clearly no hope of coping with the raw output of these new sequencers with standard spreadsheets within even the basic R environment (R Development Core Team 2008). The easiest way to manipulate these amounts of data is to use computers running UNIX. This is not prohibitive, as LINUX can be installed on all personal computers, and MacOSX (Apple system) is a UNIX system. All of the major sequence analysis programs have a version that runs on this family of systems. Even though standard UNIX tools are convenient, it is also helpful to instal some packages that are dedicated to sequence manipulation. Among them is The European Molecular Biology Open Software Suite (*EMBOSS*—<http://emboss.sourceforge.net>) developed by the European Bioinformatics Institute (Rice *et al.* 2000). *EMBOSS* is a generic sequence analysis package. It provides basic sequence manipulation, including sequence format conversion, sequence trim-

ming, and alignment. Recently, several packages dedicated to NGS output and metabarcoding analysis have been developed. *Qiime* (Quantitative Insights Into Microbial Ecology: Caporaso *et al.* 2010) is a set of programs wrapping several algorithms in a way that allows the development of *ad hoc* pipelines to automate data analysis from raw data processing to taxonomic analysis. Each program can be run from the UNIX command line, and several examples of their use can be found in the literature (e.g. Brinkman *et al.* 2011; Flores *et al.* 2011; Kostka *et al.* 2011). Initially developed for microbial DNA metabarcoding, *Qiime* provides a set of data analysis algorithms commonly used in this community, including graphical outputs. Similarly, *OBITools* is another package for sequence analysis dedicated to DNA metabarcoding that has also been used in several published studies (e.g. Pegard *et al.* 2009; Soininen *et al.* 2009; Sønstebo *et al.* 2010). *OBITools* is composed of a set of programs that allow treatment of the raw sequence files from the sequencer up to the taxon assignment. The chief advantage of *OBITools* is that it can account for taxonomic information in filtering or annotating data. It works with FASTA and FASTQ files (Fig. 1) and is fully compatible with the *ecoPCR* and *ecoPrimers* programs. In addition to the previously described *oligoTag* and *ngsfilter*, other interesting programs are *solexaPairEnd*, which aligns both pair-ended reads issued from Illumina GA IIx or HiSeq to rebuild full-length amplicons, taking into account sequence quality; *obigrep* for filtering sequence on sequence patterns and sequence annotations including full taxonomy; *obiuniq* for clustering strictly identical sequences; *obistat* for summarizing data about sequences; and *ecotag* for basic taxonomic assignment. *Qiime* and *OBITools* are both developed in Python. This allows more advanced users to extend them easily. *Prinseq* (Schmieder & Edwards 2011) offers similar functionalities to *Qiime* and the *OBITools*, but through a web interface. Although a 'lite' version of *Prinseq* exists that users can instal locally, using the full version of *Prinseq* requires that the relevant data be transferred to their servers. It is difficult to predict how this feature will evolve as data volumes continue to increase.

Table 1 Set of 36 octamers with at least five differences between them

acacacac	actagatc	agactatg	acgacgag	agcgacta	agcacagt
acagcaca	gatcgcca	gcgtcagc	catcagtc	tcagtgtc	tagctagt
gtgtacat	cgctctcg	tgacatca	atcagtca	actctgct	agtgctac
tatgtcag	gtcgtaga	acatgtgt	tctactga	atatagcg	cgtataca
tagtcgca	gtcacgtc	gtacgact	gatgatct	ctatgcta	cgagtcgt
tactatac	gactgatg	atgatcgc	ctgcgtac	tcgcgctg	cacatgat

```
>HWUSI-EAS1510_0005 tag_length=9; reverse_primer=ccattgagtctctgcacctatc;
alignment=left; merged_sample={'EH0997': 1030, 'EH1506': 34, 'EH0995': 10784,
'EH0994': 3551}; count=14169; direct_primer=gggcaatcctgagccaa;
experiment=Perm-gh_R1;
atcctatattttcgaaaacaacaaggttcataaagacagaataagaatacaaaaag
```

Fig. 1 A sample sequence in FASTA format annotated following the OBITools convention. OBITools mainly manipulates sequences formatted in FASTA or FASTQ. These are traditional formats for storing un-annotated sequences. OBITools relies on a special way to format the header line of each sequence to store some basic annotation following a key, value principle. A header line of a FASTA or FASTQ formatted sequence is divided by OBITools in three parts. The first word (first set of characters without a blank) is the sequence identifier. This identifier is optionally followed by a set of key, value pairs and finally by the sequence definition. Those key, value pairs are intensively used by OBITools for annotating the sequences and for filtering them.

Some dedicated packages also include complementary programs to assist in data manipulation. In the PyroNoise package (Quince *et al.* 2011), several utilities are included, such as FCluster, which implements a hierarchical clustering algorithm from a distance matrix, FastaUnique, which clusters strictly identical sequences, and NDist, which computes a matrix distance using a global NWS algorithm (Needleman & Wunsch 1970).

How to deal with DNA degradation/amplification/sequencing errors?

Errors bias diversity estimates

The DNA metabarcoding approach is not without its drawbacks. For example, the 'rare biosphere' elements of microbial communities that were emphasized by the first NGS analysis of microbial diversity (Sogin *et al.* 2006), but this finding has since been moderated by a large number of subsequent studies showing that the initial study's estimation of microbial diversity was highly biased because of errors (Quince *et al.* 2009, 2011; Kunin *et al.* 2010). The lack of both universal primers and a reference database (Stoeck *et al.* 2010) is a significant problem for the microbiologist community. Artefactual sequences generated during PCR or sequencing are often mistaken for rare MOTUs. Biodiversity signals are also obscured by PCR and/or sequencing errors in DNA metabarcoding on vascular plants or animals (Kristina *et al.* 2006). Raw data must therefore be filtered to limit false positives during the determination of the species list.

At least three sources of error can be identified: DNA degradation before sampling, PCR-generated errors, and sequencing errors. PCR-generated errors include point mutations and formation of chimeric molecules (Acinas *et al.* 2005). However, it is commonly understood that most erroneous reads are generated by sequencing chemistry. Kunin *et al.* (2010) have shown that sequencing errors inflate estimates of actual diversity by two orders of magnitude when considering unique reads. Many of these artefacts are attributable to miscounted homopoly-

meric runs that occur in high-quality regions of the read; they are therefore not removed by end-trimming based on quality scores. Similarly, Agogue *et al.* (2011), Gilbert *et al.* (2009) and Pommier *et al.* (2010) have shown that more than 50% of the MOTUs obtained by 454 pyrosequencing are, after applying sequence filtering, represented by only a few or even just a single sequence. These low frequency reads are often suspected to be artefacts and may be discarded from further analyses (Reeder & Knight 2010). To remove noisy reads from 454 pyrosequencing runs, a number of programs have been developed. The best known of these algorithms is PyroNoise (Quince *et al.* 2009). PyroNoise is based on flowgram clustering; it uses light intensities associated with each flowgram to calculate the probability that a flowgram was generated by a given sequence. The program then reconstructs true sequences and frequencies in the sample prior to MOTU construction. Unfortunately, this Bayesian approach requires intensive computations that are not compatible with the capabilities of computers accessible to most laboratories. Another similar algorithm based on flowgram denoising is DeNoiser, developed by Reeder & Knight (2010). The computational efficiency of DeNoiser is better than that of PyroNoise, because DeNoiser uses a greedy agglomerative clustering approach instead of the iterative approach used by PyroNoise. The algorithm of DeNoiser is based on calculating the distance of each flowgram sequence with the most abundant flowgram sequence. However, this approach is liable to mis-assign reads when true sequences are very similar, a failing that increases the possibility that MOTUs will not be accurately reconstructed. PyroNoise and DeNoiser rely a 454 apparatus, which is their main drawback relative to the more versatile approach based on sequence distances rather than flowgram-based distances. This strategy is called single linkage preclustering (Huse *et al.* 2010) and is used in the PyroTagger program, developed by Kunin & Hugenholtz (2010).

Errors also originate during PCR amplification steps. The rate of PCR errors is substantial: approximately 10% of all sequences contain one or more PCR errors when a typical 250-bp sequence is amplified, according

to Kobayashi *et al.* (1999). A new program in the AmpliconNoise package (Quince *et al.* 2011) has been developed to remove 454 sequencing errors and PCR errors. AmpliconNoise is actually an extension of PyroNoise, which has been developed by adding new filtration to the initial clustering of flowgrams; this removes PCR errors and identifies PCR chimera product. A comparison of these programs in the context of microbial biodiversity estimation is given in Zinger *et al.* (2012). Microbiologists, who are today the community most actively using DNA metabarcoding, usually work with long markers (several hundreds of base pairs) and thus need to use the 454 platform. However, for short markers (<150 bp) that are favoured for metabarcoding environmental DNA from plants and animals, the Illumina/Solexa technology is perfectly adapted and is also less expensive. As the use of the Illumina/Solexa platforms is less common, no similar programs are yet available to filter errors generated by it.

Some programs have also been specifically developed to detect PCR-generated chimeric sequences; these use different methodologies. For example, the Ribosomal Database Project (RDP-II) developed by Cole *et al.* (2009) provides a program called *Chimera Check*, while Komatsoulis & Waterman (1997) have developed an application called *Chimeric Alignment* to detect chimeric sequences. Both of these programs rely on direct comparisons of individual sequences to one or two putative parent sequences at a time. Moreover, *Chimera Check* is embedded in the RDP-II platform and is thus mainly dedicated to the 16S rRNA bacterial data set. Other existing algorithms include *Pintail* by Ashelford *et al.* (2005) and *Bellerophon* developed by Huber *et al.* (2004). These two programs were developed to remove chimeras from full-length clone sequences and lack the sensitivity for short sequence reads. More recently developed applications for detecting chimeric reads include *ChimeraSlayer* (Haas *et al.* 2011), *Uchime* (Edgar *et al.* 2011) and *Perseus* (Quince *et al.* 2011). These three applications were developed to detect chimeras from short pyrosequencing reads. *ChimeraSlayer* requires a reference data set of sequences that are known to be nonchimeric. The *Uchime* algorithm can work either with or without a reference database. In the latter mode, the generation of the reference database from sequences classified as nonchimeric occurs during the data analysis. Finally, *Perseus* treats the problem of chimeric detection as a 'classification' or 'supervised learning' problem but does not require a set of reference sequences.

Error process is biased

To extend our understanding of error filtering, it could be interesting to analyse the behaviour of errors. An

opportunity to observe error bias is furnished by an analysis of faeces samples from snow leopards (*Uncia uncia*—UU) to characterize diet (Shehzad *et al.* 2012). The barcode used here is specific to vertebrates and corresponds to a 100 bp/108 bp fragment of the 12S RNA gene (Riaz *et al.* 2011). The sequencing was carried out on an Illumina GA IIX. Snow leopards usually eat one prey during a meal (Shehzad *et al.* 2012), and so the vertebrate DNA present in the faeces of a snow leopard usually corresponds to two species: the snow leopard itself and the prey. To expose some behaviour of errors, we randomly selected three samples with *Capra sibirica* (Siberian Ibex) (the mountain goat—CS) as the unique prey. Two simple analyses were performed on this data set. For both analyses, strictly identical sequences were first clustered and weighted by the number of grouped sequence reads. In Fig. 2, the weight of each one of these clusters is plotted as a function of its distance (number of sequence differences) from the true *Uncia uncia* sequence (identical to entry EMBL:EF551004). The two true sequences of UU and CS (identical to entry EMBL:AF363779) are identified on the plot and are surrounded by many other clusters. The most probable explanation is that the latter all correspond to errors. The importance of all these errors is obvious, as is the necessity to account for them during data processing. Sequence errors are usually thought to be the result of a homogeneous random process. With this in mind, several authors have proposed reducing the effect of this noise by mixing several independent PCR results obtained from the same sample. To characterize sequence errors more precisely, errors were categorized together as a function of their position and type (e.g. $A \rightarrow T$ at position 3 or deletion of a T at position 10). A contingency table was computed for each sample by counting how many times each category of error occurs in the considered sample. Figure 3 presents the scatter plots of pairwise comparisons of these contingency tables. Correlations observed between any two PCR products were tested using the Kendall Tau test. After a Bonferroni correction for multiple tests (Bonferroni 1935), all P -values were estimated to be close to 0, demonstrating a high similarity between error patterns. While these observations represent raw analysis, the insights they provide on error abundance and, more particularly, error bias will have to be taken into account during sequence denoising for metabarcoding applications to quantify diversity realistically.

Taxonomic assignment of metabarcode sequences

Two main categories of algorithms for the taxonomic assignment of a DNA-barcode sequence exist. They

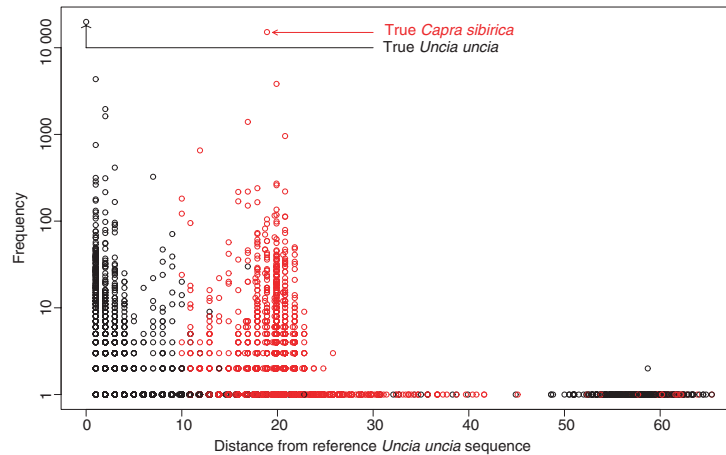


Fig. 2 Evaluation of the sequence errors in the Snow Leopard diet analysis. Figure shows the distance of all sequences in one PCR from true *Uncia uncia* (UU) sequence. Each dot corresponds to one sequence. On the *x*-axis is the distance of the sequences from true UU sequence, whereas on the *y*-axis is the occurrence count of that sequence. Colour is black if the distance of the observed sequence from the true UU sequence is less than its distance from *Capra sibirica* (CS) sequence, and is otherwise red.

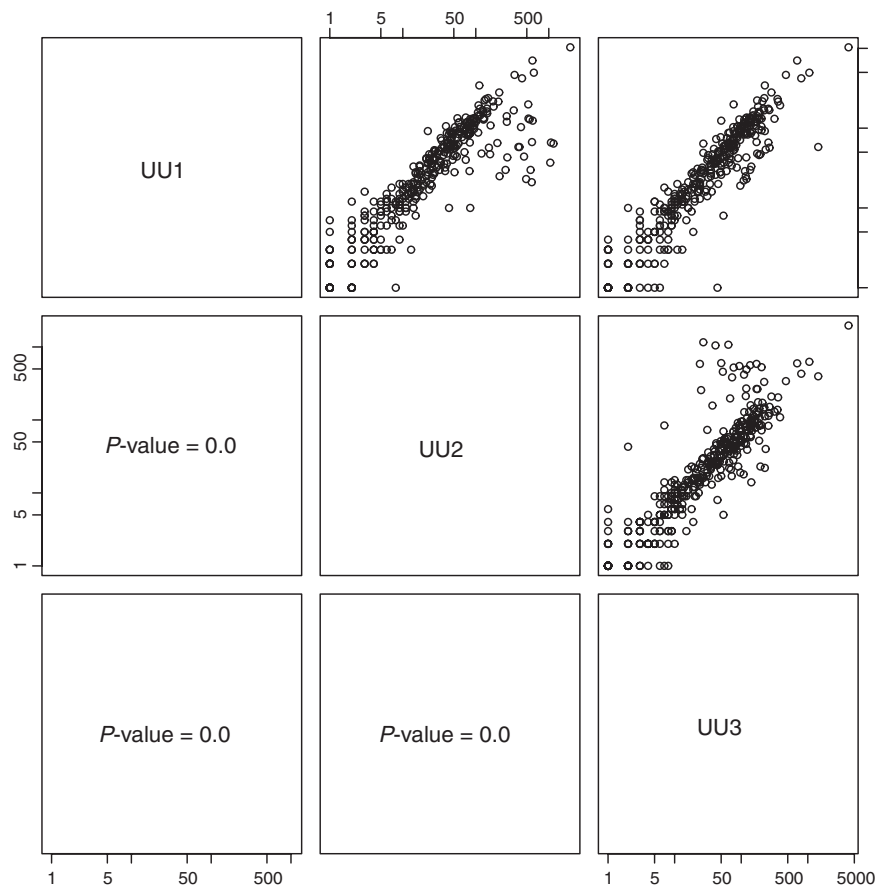


Fig. 3 Bias in observed sequence error. Figure shows the Kendall Tau correlation test applied on three PCRs by selecting sequences having a distance of one nucleotide from the true UU sequence. Upper triangle shows correlation plots and lower triangle shows *P*-values obtained with Kendall Tau test. A positive correlation was observed between any two PCR pairs depicting that the two PCRs show same error pattern, that is, the same errors occur with same frequency in both PCRs.

correspond to two distinct problems: assigning a taxon to a sequence and helping to delimit species. The first set of methods relies on the comparison of unknown sequences with a reference data set that includes DNA barcodes of identified specimens. The second set of methods is designed to propose groups of sequences, without relying on a pre-existing data set of identified sequences.

Assigning specimens to species by comparing unknown sequences to sets of identified sequences

The DNA-barcode community has argued that the primary goal of DNA barcoding is to link an unknown DNA sequence to a taxonomic name by comparing the sequence with a reference database that includes voucher specimens, properly identified by a taxonomist, as well as their DNA barcodes. Most of the available methods to do this have been reviewed and discussed by Frezal & Leblois (2008), and we confine ourselves here to present a brief update of this list.

Methods can be classified according to five partly overlapping categories: similarity, phylogeny, character-based, classification algorithms and coalescent-based. Similarity methods are based either on the comparison of sequences (BLAST—Altschul *et al.* 1990; BOLD—Ratnasingham & Hebert 2007), on the use of a genetic distance thresholds (*TaxI*—Steinke *et al.* 2005; TaxonDNA—Meier *et al.* 2006), or on the recognition of short sequences ('words' or 'probes') (DNA-BAR—Das-Gupta *et al.* 2005; DOME-ID—Little, Little & Stevenson 2007; *Googling DNA*—Hajibabaei & Singer 2009; BRONX—Little 2011). Recently, Zhang *et al.* (2011) coupled a distance method with the fuzzy membership function to assign sequences as a fuzzy member of a named species. Phylogenetic methods commonly used in barcoding are Neighbour-Joining and Maximum Likelihood; they identify query sequences by their position within a clade (see e.g. the Statistic Assignment Package, Munch *et al.* 2008; pplacer, Matsen *et al.* 2010; or the Evolutionary Placement Algorithm, Berger *et al.* 2011) or their status as sister-branch (see Ross *et al.* 2008). Chu *et al.* (2009) proposed a phylogeny-based method, but without alignment and using composition vectors. Generally presented as an alternative to the distance-based methods, character-based methods do not reduce DNA barcodes to distances but rely instead on the presence of diagnostic character states to assign sequences to species. This technique was first advocated by DeSalle *et al.* (2005) and DeSalle (2006) in the barcoding context, and then formalized by several authors (Anderson 2009; Bergmann *et al.* 2009; Bertolazzi *et al.* 2009; Rach *et al.* 2008; —see also Sites & Marshall 2003). Classification algorithms have also been widely used in this context, though they are not primarily designed for

DNA barcode assignment. Supervised classification methods such as *k*-NN, CART, RF and Kernel are compared in Austerlitz *et al.* 2009; see also Kuksa & Pavlovic 2009; or Robust Discrete Discriminant Analysis—Bouveyron *et al.* 2009; neuronal networks—Bai & Kremer 2011; Zhang *et al.* 2008; Klee diagrams using indicator vectors—Sirovich *et al.* 2009, 2010. Finally, the coalescent theory has also been used to evaluate the quality of the assignments in a likelihood or bayesian framework (Abdo & Golding 2007; but see also Lou & Golding 2010 for a similar method that uses the theory of segregating sites [Nielsen & Matz 2006] instead of a coalescent approach).

Comparative studies (Little & Stevenson 2007; Ross *et al.* 2008; Austerlitz *et al.* 2009; Pettengill & Neel 2010; Goldstein & DeSalle 2011) generally conclude that none of these methods outperform the others in all conditions. This is because their results depend on a number of different parameters, including the quality of the sampling (both intra- and interspecific) and the level of genetic variability. These methods also differ in computation times required (distance-based methods are generally faster than phylogenetic reconstruction), but also in their propensity to follow one or another 'species concept' (e.g. character-based methods follow the Phylogenetic Species Concept, distance-based methods do not). Finally, their capacity to quantify the quality of the assignment (i.e. detect false positives and negatives) has also been identified as being an important performance criteria (Frezal & Leblois 2008).

Delimiting species

DNA barcodes can also provide useful characters for proposing species delimitation. In contrast to sequence assignment, the aim here is not to compare unknown sequences with an identified set of sequences, but to partition the sequences (including both identified and non-identified sequences) into groups, following given criteria. Here again, different methodological categories are available. First, several methods rely on the application of an arbitrary criteria, such as the 10× rule (interspecific distances are 10 times greater than intraspecific distances—Hebert *et al.* 2004) or any threshold of genetic distances (Blaxter *et al.* 2005; Jones *et al.* 2011), or the 4× rule, where species are defined as reciprocally monophyletic and separated by a mean genetic distance greater than $4 \times \theta$ (where $\theta = 2N_e\mu$, with N_e the effective population size and μ the mutation rate per base pair per generation; Birky *et al.* 2010). Other methods (generally referred as clustering methods) also apply arbitrary criteria to delimit groups (such as minimizing the intra-group variability compared with the overall variability), which either require specifying the number

of expected groups or not (e.g. the K-means method—Steinley & Brusco 2007 or the Markov Chain Clustering—Zinger *et al.* 2009), but they have only rarely been applied to species delimitation. Older methods, reviewed in Sites & Marshall (2003), are specifically designed to delimit species using DNA data, but they rely on the a priori definition of populations (using morphological and geographical data). This category could also include methods that aim to test the validity of previously delimited species (e.g. Ence & Carstens 2011; Sakalidis *et al.* 2011). Another widely used method is simply to reconstruct a phylogenetic tree on which species are then defined as terminal clades. However, these 'terminal clades' rarely correspond to a formalized criterion. This method is applicable and reproducible only when a few species are involved; it becomes harder to objectively delimit species when numerous taxa are involved. Finally, there are only a few formalized methods specifically designed to delimit species solely using DNA sequences (and they are thus referred as exploratory methods). One, General Mixed Yule Coalescent (GMYC), was proposed by Pons *et al.* (2006), updated in Monaghan *et al.* (2009): using a phylogenetic tree, the limit between inter- and intraspecific branching events is identified by estimating the likelihood of each branching events under a speciation (Yule) or a coalescent model. Another, Automatic Barcode Gap Discovery (ABGD), is described in this volume (Puillandre *et al.* 2012).

While barcoding data sets are generally composed of only one or very few loci, another field of research in species delimitation, based on the use of multiple loci, is developing in the framework of the coalescent theory (Carstens & Knowles 2007; Knowles & Carstens 2007; Knowles 2009; Kubatko *et al.* 2009; Carstens & Dewey 2010; Hausdorf & Hennig 2010; Leaché & Fujita 2010; O'Meara 2010; Ross *et al.* 2010). Finally, it should also be noted that a consensus is now arising: species delimitation methods relying exclusively or primarily on DNA sequences should be combined with other characteristics and criteria in an integrative context (e.g. Padial *et al.* 2010; Yeates *et al.* 2010; Goldstein & DeSalle 2011; Zhang *et al.* 2011).

Unsupervised approaches in metabarcoding

One of the main challenges in environmental metabarcoding is to link the DNA sequences obtained to their appropriate taxonomic names (either at order, familial, genus or species level, depending on the aim of the study). While this problem is similar to classic DNA barcoding approaches in theory, several differences exist in practice. The genetic markers used are generally nonstandard (as previously explained), which prevents the use of the Barcode of Life database (BOLD), which

is based mainly on animal and plant barcodes. A solution is to build specific reference databases for each study (Valentini *et al.* 2009a; Hajibabaei *et al.* 2011). That way, most of the assignment methods reviewed above may be used to identify sequences. Conversely, when no database reference is available, species delimitation methods can be applied. In this case, sequences would not be linked to a taxonomic name, but would be clustered in MOTUs that can be compared in different studies, for example, comparing the diversity of MOTUs in different localities or under different parameters in the same locality. However, most of the methods primarily designed for DNA barcoding have never been tested in a metabarcoding context, and some of them might be of little use. For example, phylogeny-based methods require well-aligned reference sequences and a robust phylogenetic tree that might not be available in every case. Additionally, computational limitations might also prevent the use of these methods for very large data sets. For example, the GMYC approach requires the prior construction of a phylogenetic tree including all the analysed sequences, which might be impossible when the data set includes more than several thousand sequences.

In addition to the tools developed for classic DNA barcoding, several software programs designed specifically for metabarcoding and presented previously (e.g. OBITools, PyroTagger, or FCluster, FastaUnique and NDist in the PyroNoise package) can be applied to assign taxonomic names to sequences when a taxonomic database exists or to cluster sequences in MOTUs when no taxonomic data are available.

Another limitation of metabarcoding in regard to determining taxonomic assignment is the length and variability of the amplified fragments: they can be very short (i.e. <100 bp—Valentini *et al.* 2009b; see Hajibabaei *et al.* 2006; Meusnier *et al.* 2008 for an introduction to mini-barcodes). Moreover, certain markers like the tRNA-Leu P6-loop (Taberlet *et al.* 2007) are highly variable in size, and so alignment is almost impossible. Consequently, methods based on multiple alignments of the sequences, and eventually robust phylogenetic trees, are not applicable. Finally, the error rate of the PCR amplification and of the NGS can be high, leading to the generation of artificial DNA sequences. This can have important consequences for taxonomic assignment or MOTUs clustering: depending on the criterion chosen to assign a sequence to a taxon or to define the MOTUs, same-taxon sequences could wrongly be classified as belonging to different groups, potentially increasing the overall genetic variability and species diversity recorded. In these cases, methods that provide a probability of assignment to a given species should be preferred.

Acknowledgement

This work is financially supported by the European Commission, under the Sixth Framework Program (EcoChange project, contract no FP6-036866). NP was (partly) funded by CONCO, the cone snail genome project for health, funded by the European Commission: LIFESCIHEALTH-6 Integrated Project LSHB-CT-2007-037592.

Conflicts of interest

EC and TR are co-inventors of patents related to the g/h primers and the use of the V5 region of the 12S rRNA gene for vertebrate identification using degraded template DNA. These patents only restrict commercial applications and have no impact on the use of this locus by academic researchers.

References

- Abarenkov K, Henrik Nilsson R, Larsson KH *et al.* (2010) The unite database for molecular identification of fungi—recent updates and future perspectives. *New Phytologist*, **186**, 281–285.
- Abdo Z, Golding BG (2007) A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology*, **56**, 44–56.
- Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, **71**, 8966–8969.
- Agogue H, Lamy D, Neal PR, Sogin ML, Herndl GJ (2011) Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Molecular Ecology*, **20**, 258–274.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Andersen K, Bird KL, Rasmussen M *et al.* (2012) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.
- Anderson MP (2009) Bayesian classification of DNA barcodes. PhD thesis, Kansas State University, Manhattan, Kansas.
- Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, **71**, 7724–7736.
- Austerlitz F, David O, Schaeffer B *et al.* (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**, S10.
- Bai Z, Kremer SC (2011) Deep belief network: a step towards better DNA barcoding classification. *Proceedings of the 6th Canadian Student Conference on Biomedical Computing and Engineering*, **9**, 94.
- Bekaert M, Teeling EC (2008) UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Research*, **36**, e56.
- Bellemain E, Carlsen T, Brochmann C *et al.* (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189.
- Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, **60**, 291–302.
- Bergmann T, Hadrys H, Breves G, Schierwater B (2009) Character-based DNA barcoding: a superior tool for species classification. *Berliner und Münchener Tierärztliche Wochenschrift*, **122**, 446–450.
- Bertolazzi P, Felici G, Weitschek E (2009) Learning to classify species with barcodes. *BMC Bioinformatics*, **10**, S7.
- Bienert R, de Danieli S, Miquel C *et al.* (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, **21**, 2017–2030.
- Binladen J, Gilbert M, Thomas P *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.
- Birky JCW, Adams J, Gemmel M, Perry J (2010) Using population genetic theory and DNA sequences for species detection and identification in asexual organisms. *PLoS One*, **5**, e10609.
- Blaxter M, Mann J, Chapman T *et al.* (2005) Defining operational taxonomy units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1935–1943.
- Bonferroni CE (1935) Il calcolo delle assicurazioni sui gruppi di teste. In: *Studi in Onore del Professore Salvatore Ortu Carboni* (ed Carboni SO), pp. 13–60. Tipografia del Senato, Rome.
- Bouveyron C, Girard S, Olteanu M (2009) Supervised classification of categorical data with uncertain labels for DNA barcoding. *17th European Symposium on Artificial Neural Networks, (ESANN 2009)*, 29–34.
- Brinkman BM, Hildebrand F, Kubica M *et al.* (2011) Caspase deficiency alters the murine gut microbiome. *Cell Death and Disease*, **2**, e220.
- Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.
- Carstens BC, Dewey TA (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Systematic Biology*, **59**, 400–414.
- Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Systematic Biology*, **56**, 400–411.
- CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794–12797.
- Chu KH, Xu M, Li CP (2009) Rapid DNA barcoding analysis of large datasets using the composition vector method. *BMC Bioinformatics*, **10**, S8.
- Coissac E (2012) OligoTag a program for designing set of tags for next generation sequencing of multiplexed samples. In: *Data Production and Analysis in Population Genomics* (eds Bonin A and Pompanon F). Methods in Molecular Biology Series, Humana Press, New York.
- Cole JR, Wang Q, Cardenas E *et al.* (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, **37**, D141–D145.
- DasGupta B, Konwar M, Mandoiu II, Shvartsman AA (2005) DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*, **21**, 3424–3426.

- DeSalle R (2006) Species discovery versus species identification in DNA Barcoding efforts: response to Rubinoff. *Conservation Biology*, **20**, 1545–1547.
- DeSalle R, Egan MG, Sidall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1905–1916.
- Duitama J, Kumar DM, Hemphill E *et al.* (2009) PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Research*, **37**, 2483–2492.
- Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.
- Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, **11**, 473–480.
- Epp L, Boessenkool S, Bellemain E *et al.* (2012) New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, **21**, 1821–1833.
- Ficetola GF, Coissac E, Zundel S *et al.* (2010) An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, **11**, 434.
- Flores GE, Bates ST, Knights D *et al.* (2011) Microbial biogeography of public restroom surfaces. *PLoS One*, **6**, e28132.
- Frank DN (2009) BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics*, **10**, 362.
- Frezal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution*, **8**, 727–736.
- Gadberry MD, Malcomber ST, Doust AN, Kellogg EA (2005) Primaclade—a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21**, 1263–1264.
- Gilbert JA, Field D, Swift P *et al.* (2009) The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology*, **11**, 3132–3139.
- Goldstein PZ, Desalle R (2003) Calibrating phylogenetic species formation in a threatened insect using DNA from historical specimens. *Molecular Ecology*, **12**, 1993–1998.
- Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *BioEssays*, **33**, 135–147.
- Haas BJ, Gevers D, Earl AM *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, **21**, 494–504.
- Hajibabaei M, Singer GAC (2009) Googling DNA sequences on the World Wide Web. *BMC Bioinformatics*, **10**, S4.
- Hajibabaei M, Smith AM, Janzen DH *et al.* (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.
- Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.
- Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, **5**, 235–237.
- Hausdorf B, Hennig C (2010) Species delimitation using dominant and codominant multilocus markers. *Systematic Biology*, **59**, 491–503.
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. *Proceedings. Biological Sciences*, **270**, 313–321.
- Hebert PDN, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological Sciences*, **270**, 96–99.
- Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, 1657–1663.
- Hoffmann C, Minkah N, Leipzig J *et al.* (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Research*, **35**, e91.
- Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**, 2317–2319.
- Huse SM, Welch DMM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental microbiology*, **12**, 1889–1898.
- Jabado OJ, Palacios G, Kapoor V *et al.* (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, **34**, 6605–6611.
- Jarman SN (2004) Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, **20**, 1644–1645.
- Jones M, Ghoorah A, Blaxter M (2011) jMOTU and Taxonator: turning DNA barcode sequences into annotated operational taxonomic units. *PLoS One*, **6**, e19259.
- Kim N, Lee C (2007) QPRIMER: a quick web-based application for designing conserved PCR primers from multigenome alignments. *Bioinformatics*, **23**, 2331–2333.
- Knowles LL (2009) Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology*, **58**, 463–467.
- Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, **56**, 887–895.
- Kobayashi N, Tamura K, Aotsuka T (1999) PCR error and molecular population genetics. *Biochemical Genetics*, **37**, 317–321.
- Komatsoulis GA, Waterman MS (1997) A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations. *Applied and Environmental Microbiology*, **63**, 2338–2346.
- Koressaar T, Remm M (2007) Enhancements and modifications of primer design program primer3. *Bioinformatics*, **23**, 1289–1291.
- Kostka JE, Prakash O, Overholt WA *et al.* (2011) Hydrocarbon-degrading bacteria and the bacterial community response in gulf of mexico beach sands impacted by the deepwater horizon oil spill. *Applied and Environmental Microbiology*, **77**, 7962–7974.
- Kristina MS, Robert BP, Michael D (2006) Sorenson single base errors in PCR products from avian museum specimens and their effect on estimates of historical genetic diversity. *Conservation Genetics*, **8**, 879–884.

- Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.
- Kuksa P, Pavlovic V (2009) Efficient alignment-free DNA barcode analytics. *BMC Bioinformatics*, **10**, S9.
- Kunin V, Hugenholtz P (2010) Pyrotagger: a fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *The Open Journal*, **1**, 1.
- Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.
- Lahaye R, van der Bank M, Bogarin D *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 2923–2928.
- Leaché AD, Fujita MK (2010) Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society B: Biological Sciences*, **277**, 3071–3077.
- Little DP (2011) DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS One*, **6**, e20552.
- Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, **23**, 1–21.
- Lou M, Golding BG (2010) Assigning sequences to species in the absence of large interspecific differences. *Molecular Phylogenetics and Evolution*, **56**, 187–194.
- Luo A, Zhang A, Ho SY *et al.* (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics*, **12**, 84.
- Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Matsen FA, Kodner RB, Armbrust EV (2010) Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.
- Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.
- Meusnier I, Singer GAC, Landry JF *et al.* (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 214.
- Monaghan MT, Wild R, Elliot M *et al.* (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology*, **58**, 298–311.
- Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.
- Murray DC, Bunce M, Cannell BL *et al.* (2011) DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One*, **6**, e25776.
- Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.
- Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. *Systematic Biology*, **55**, 162–169.
- Nijman V, Aliabadian M (2010) Performance of distance-based DNA barcoding in the molecular identification of primates. *Comptes Rendus Biologies*, **333**, 11–16.
- Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–217.
- O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, **59**, 59–73.
- Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. *Frontiers in Zoology*, **7**, 16.
- Parameswaran P, Jalili R, Tao L *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**, e130.
- Pegard A, Miquel C, Valentini A *et al.* (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, **57**, 5700–5706.
- Pettengill JB, Neel MC (2010) An evaluation of candidate plant DNA barcodes and assignment methods in diagnosing 29 species in the genus *Agalinis* (Orobanchaceae). *American Journal of Botany*, **97**, 1391–1406.
- Pommier T, Neal PR, Gasol JM *et al.* (2010) Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquatic Microbial Ecology*, **61**, 212–224.
- Pompanon F, Deagle BE, Symondson WOC *et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931–1950.
- Pons J, Barraclough TG, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.
- 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 Research*, **35**, 7188–7196.
- Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.
- Quince C, Lanzén A, Curtis TP *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, **6**, 639–641.
- Quince C, Lanzen A, Davenport R, Turnbaugh P (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.
- R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. Available from <http://www.R-project.org>.
- Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 237–247.
- Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (<http://www.barcodingoflife.org>). *Molecular Ecology Notes*, **7**, 355–364.
- Reeder J, Knight R (2010) Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature Methods*, **7**, 668–669.

- Riaz T, Shehzad W, Viari A *et al.* (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **39**, e145.
- Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.
- Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, **57**, 216–230.
- Ross KG, Gotzek D, Ascunce MS, Shoemaker DD (2010) Species delimitation: a case study in a problematic ant taxon. *Systematic Biology*, **59**, 162–184.
- Rozen S, Skaletsky H (2000) Primer3 on the www for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.
- Sakalidis ML, St. J. Hardy GE, Burgess TI (2011) Use of the Genealogical Sorting Index (GSI) to delineate species boundaries in the *Neofusicoccum parvum*–*Neofusicoccum ribis* species complex. *Molecular Phylogenetics and Evolution*, **60**, 333–334.
- Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
- Shehzad W, McCarthy TM, Pompanon F *et al.* (2012) Prey preference of snow leopard (*Panthera uncia*) in South Gobi, Mongolia. *PLoS One*, **7**, e32104.
- Sirovich L, Stoeckle M, Zhang Y (2009) A scalable method for analysis and display of DNA sequences. *PLoS One*, **4**, e7051.
- Sirovich L, Stoeckle M, Zhang Y (2010) Structural analysis of biodiversity. *PLoS One*, **5**, e9266.
- Sites JW, Marshall JC (2003) Delimiting species: a renaissance issue in systematic biology. *Trends in Ecology and Evolution*, **19**, 462–470.
- Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.
- Soininen EM, Valentini A, Coissac E *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, **6**, 16.
- Sønstebo JH, Gielly L, Brysting AK *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past arctic vegetation and climate. *Molecular Ecology Resources*, **10**, 1009–1018.
- Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA barcoding using distance methods. *Proceedings of the Royal Society B: Biological Sciences*, **360**, 1975–1980.
- Steinley D, Brusco MJ (2007) Initializing K-means batch clustering: a critical evaluation of several techniques. *Journal of Classification*, **24**, 99–121.
- Stoeck T, Bass D, Nebel M *et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21–31.
- Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.
- Taberlet P, Prud’Homme S, Campione E *et al.* (2012a) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, **21**, 1816–1820.
- Taberlet P, Coissac E, Pompanon F *et al.* (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.
- Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.
- Valentini A, Miquel C, Ali Nawaz M *et al.* (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.
- Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. *Trends in Ecology and Evolution*, **24**, 110–117.
- Vences M, Thomas M, van der Meijden A, Chiari Y, Vieites DR (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, **2**, 5.
- Wandeler P, Hoeck PEA, Keller LF (2007) Back to the future: museum specimens in population genetics. *Trends in Ecology and Evolution*, **22**, 634–642.
- Willerslev E, Hansen AJ, Binladen J *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
- Willerslev E, Cappellini E, Boomsma W *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science*, **317**, 111–114.
- Yeates D, Seago A, Nelson L *et al.* (2010) Integrative taxonomy, or iterative taxonomy? *Systematic Entomology*, **36**, 209–217.
- Zhang AB, Sikes DS, Muster C, Li SQ (2008) Inferring species membership using DNA sequences with back-propagation neural networks. *Systematic Biology*, **57**, 202–215.
- Zhang AB, Muster C, Liang HB *et al.* (2011) A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology*, **21**, 1848–1863.
- Zinger L, Coissac E, Choler P, Geremia RA (2009) Assessment of microbial communities by graph partitioning in a study of soil fungi in two Alpine meadows. *Applied and Environmental Microbiology*, **75**, 5863–5870.
- Zinger L, Gobet A, Pmmier T (2012) Two decades of describing aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–1896.

E.C. is maître de conférences at Grenoble University (UJF) and is member of LECA. He is bioinformatician and is interested in biodiversity estimation using DNA based methods. T.R. was PhD student at LECA. She is computer scientist and interested in algorithmic dedicated to sequence analysis and more particularly to DNA barcoding. P.N. is a Post-Doc student at the Museum National d’Histoire Naturelle, Paris. He is interested in the taxonomy, diversification and evolution of marine gastropods, and in particular the Conoidea.
