MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 176–182 doi: 10.1111/1755-0998.12428

OBITOOLS: a UNIX-inspired software package for DNA metabarcoding

FRÉDÉRIC BOYER,* CÉLINE MERCIER,* AURÉLIE BONIN,* YVAN LE BRAS,† PIERRE TABERLET* and ERIC COISSAC*

*Laboratoire d'Écologie Alpine (LECA), CNRS UMR 5553, Université Joseph Fourrier, BP 53, 38041 Grenoble Cedex-9, France, †GenOuest Core Facility, CNRS UMR 6074 IRISA-INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France

Abstract

DNA metabarcoding offers new perspectives in biodiversity research. This recently developed approach to ecosystem study relies heavily on the use of next-generation sequencing (NGS) and thus calls upon the ability to deal with huge sequence data sets. The OBITOOLS package satisfies this requirement thanks to a set of programs specifically designed for analysing NGS data in a DNA metabarcoding context. Their capacity to filter and edit sequences while taking into account taxonomic annotation helps to set up tailor-made analysis pipelines for a broad range of DNA metabarcoding applications, including biodiversity surveys or diet analyses. The OBITOOLS package is distributed as an open source software available on the following website: http://metabarcoding.org/obitools. A Galaxy wrapper is available on the GenOuest core facility toolshed: http://toolshed.genouest.org.

Keywords: biodiversity, next-generation sequencing, PCR errors, sequence analysis, taxonomic annotation

Received 16 December 2014; accepted 5 May 2015

Introduction

DNA metabarcoding is an emerging approach for biodiversity studies (Taberlet *et al.* 2012). Originally mainly developed by microbiologists (e.g. Sogin *et al.* 2006), it is now widely used for plants (e.g. Sønstebø *et al.* 2010; Parducci *et al.* 2012; Yoccoz *et al.* 2012) and animals from meiofauna (e.g. Chariton *et al.* 2010; Baldwin *et al.* 2013) to larger organisms (e.g. Andersen *et al.* 2012; Thomsen *et al.* 2012). Interestingly, this method is not limited to *sensu stricto* biodiversity surveys, but it can also be implemented in other ecological contexts such as for herbivore (e.g. Valentini *et al.* 2009; Kowalczyk *et al.* 2011) or carnivore (e.g. Deagle *et al.* 2009; Shehzad *et al.* 2012) diet analyses.

Whatever the biological question under consideration, the DNA metabarcoding methodology relies heavily on next-generation sequencing (NGS) and generates considerable numbers of DNA sequence reads (typically million of reads). Manipulation of such large data sets requires dedicated programs usually running on a Unix system.

Since its early stages, UNIX is an operating system dedicated to scientific computing that includes a large set of simple tools to efficiently process text files. Most of those programs can be viewed as filters extracting information from a text file to create a new text file. These programs process text files as streams, line per line, therefore allowing computation on a huge data set without requiring a large memory. UNIX programs usually print their results to their standard output (stdout). The main philosophy of the UNIX environment is to allow easy redirection of the stdout either to a file, for saving the results, or to the standard input (stdin) of a second program, thus allowing to easily create complex processing from simple base commands. Access to UNIX computers is increasingly easier for scientists nowadays as LINUX, an open source version of UNIX, can be freely installed on every PC machine, and the MACOSX operating system, running on Apple computers, is also a UNIX system.

The OBITOOLS programs imitate UNIX standard programs because they usually act as filters, reading their data from text files or the stdin and writing their results to the stdout. The main difference with classical UNIX programs is that text files are not analysed line per line but sequence record per sequence record (see below for a detailed description of a sequence record).

Compared to packages for similar purposes like MOTHUR (Schloss *et al.* 2009) or QIIME (Caporaso *et al.* 2010), OBITOOLS mainly relies on filtering and sorting algorithms. This allows users to set up versatile data analysis pipelines (Fig. 1), adjustable to the broad range of DNA

Correspondence: Eric Coissac, Fax: +33 476514279;
E-mail: eric.coissac@inria.fr

metabarcoding applications. The innovation of the OBITOOLS is their ability to take into account the taxonomic annotations, ultimately allowing sorting and filtering of sequence records based on the taxonomy. The OBITOOLS have already been used in many published studies (e.g. Ficetola *et al.* 2010; Kowalczyk *et al.* 2011; Yoccoz *et al.* 2012; Bellemain *et al.* 2013), demonstrating their usefulness for various applications related to metabarcoding. The purpose of this study was to provide a full overview of their functionalities. The OBITOOLS package is freely available on the following website: http://metabarcoding.org/obitools.

Basic concepts of OBITOOLS

Once installed, the OBITOOLS enrich the UNIX command line interface with a set of new commands dedicated to NGS data processing. Most of them have a name starting with the `obi` prefix. They automatically recognize the input file format amongst most of the standard sequence file formats (i.e. FASTA, FASTQ, EMBL and GENBANK formats). Nevertheless, options are available to enforce some format specificity such as the encoding system used in FASTQ files for quality codes. Most of the basic UNIX commands have their OBITOOLS equivalent (e.g. obihead vs. head, obitail vs. tail, obigrep vs. grep), which is convenient for

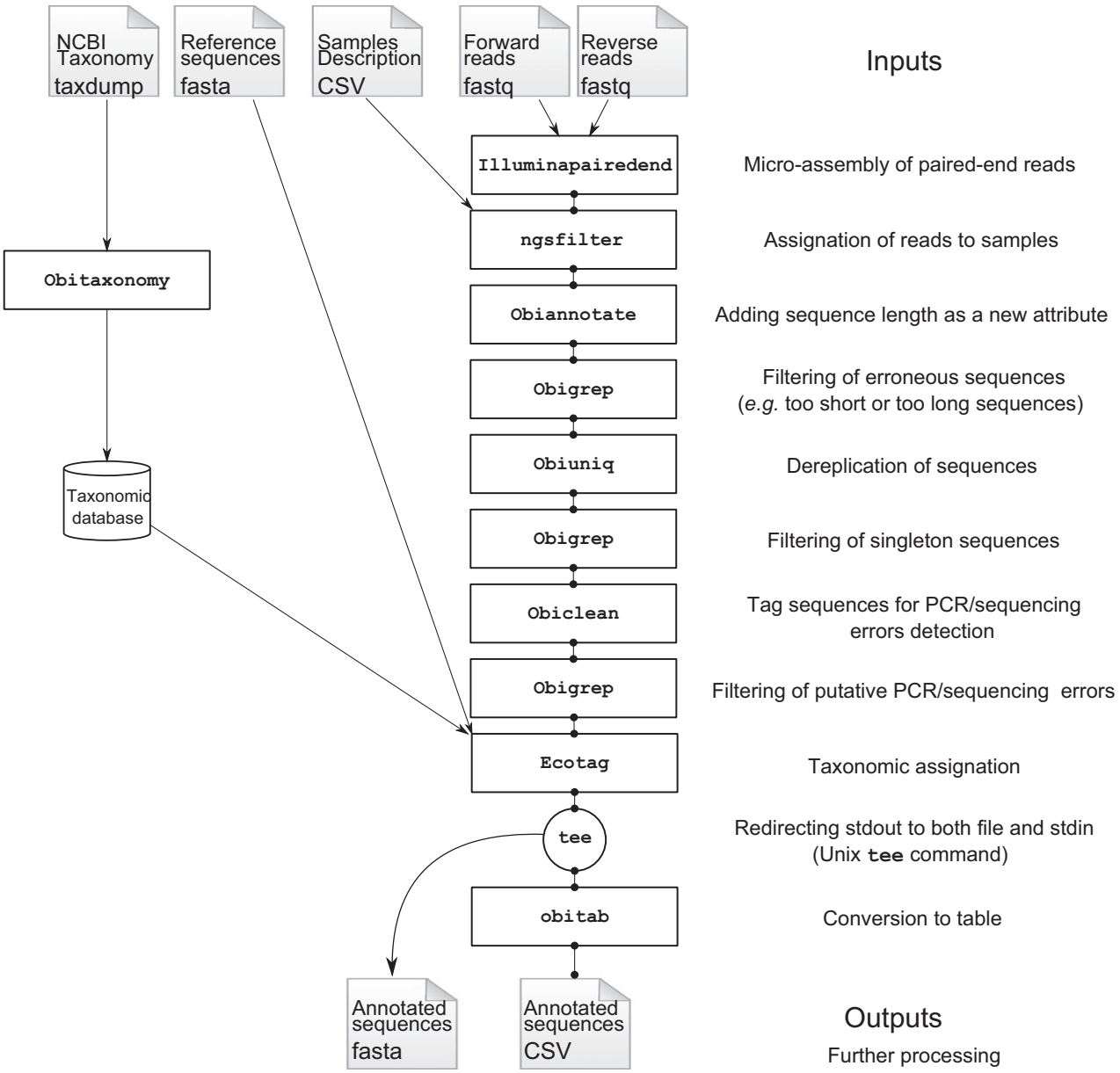

Fig. 1 Pipeline example for a standard biodiversity survey.

scientists familiar with UNIX. The main difference between any standard UNIX command and its OBITOOLS counterpart is that the treatment unit is no longer the text line but the sequence record. As a sequence record is more complex than a single text line, the OBITOOLS programs have many supplementary options compared to their UNIX equivalents.

The structure of a sequence record

OBITOOLS commands consider a sequence record as an entity composed of five distinct elements. Two of them are mandatory, the identifier (id) and the DNA or protein sequence itself. The id is a single word composed of characters, digits and other symbols like dots or underscores excluding spaces. Formally, the ids should be unique within a data set and should identify each sequence record unambiguously, but only a few OBITOOLS actually rely on this property. The sequence is an ordered set of characters corresponding to nucleotides or amino acids according to the International Union of Pure and Applied Chemistry (IUPAC) nomenclature (IUPAC-IUB Commission on Biochemical Nomenclature (CBN) 1968; Cornish-Bowden 1985). The three other elements composing a sequence record are optional. They consist in a sequence definition, a quality vector and a set of attributes. The sequence definition is a free text describing the sequence briefly. The quality vector associates a quality score to each nucleotide or amino acid. Usually, this quality score is the result of the base-calling process by the sequencer. The last element is a set of attributes qualifying the sequence, each attribute being described by a 'key=value' pair. The set of attributes is the central concept of the OBITOOLS system. When an OBITOOLS command is run on the sequence records included in a data set, the result of the computation often consists in the addition of new attributes completing the annotation of each sequence record. This strategy of sequence annotation allows the OBITOOLS to return their results as a new sequence record file that can be used as the input of another OBITOOLS, ultimately creating complex pipelines.

Managed sequence file formats

Most of the OBITOOLS commands read sequence records from a file or from the stdin, make some computations on the sequence records and output annotated sequence records. As inputs, the OBITOOLS are able to automatically recognize the most common sequence file formats (i.e. FASTA, FASTQ, EMBL and GENBANK). They are also able to read ECOPCR (Ficetola *et al.* 2010) result files and ECOPCR/ECOPRIMERS formatted sequence databases (Riaz *et al.* 2011) as ordinary sequence files. File format outputs are more limited. By default, sequences without and with quality information are written in FASTA and SANGER FASTQ formats, respectively. However, dedicated options allow enforcing the output format, and the OBITOOLS are also able to write sequences in the ECOPCR/ECOPRIMERS database format, to produce reference databases for these programs. In the FASTA or FASTQ format, the attributes are written in the header line just after the id, following a `key=value;` format (Fig. 2).

Management of the taxonomy

Filtering and annotation steps in the processing of DNA metabarcoding sequence data are greatly eased by the explicit association of taxonomic information to sequences together with an easy access to the taxonomy. A taxonomic information can be associated with each sequence record through a numerical taxonomic identifier (taxid) stored in an attribute named `taxid`. When querying taxonomic information of a sequence record, OBITOOLS rely only on this `taxid` attribute; nevertheless, several OBITOOLS commands can annotate sequence records with text taxonomy-related attributes for the user's convenience. The value of the `taxid` attribute must be a unique integer referring unambiguously to one taxon in the taxonomic associated database. Although this is not mandatory, the NCBI taxonomy is

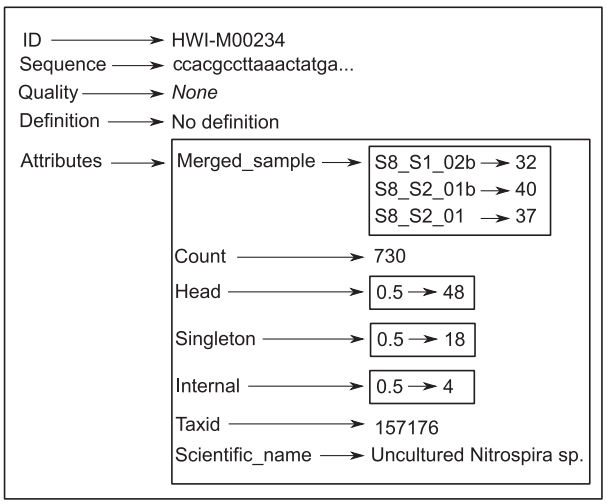

Fig. 2 The structure of an OBITOOLS sequence record and its representation in FASTA and FASTQ formats. The numbers `32`, `40` and `37` correspond to the counts of the relevant sequence obtained for samples `S8_S1_02b`, `S8_S2_01b` and `S8_S2_01`, respectively. The number `730` corresponds to the total count of the relevant sequence obtained in the whole data set. The numbers `48`, `18` and `4` correspond to the number of times amongst all samples that the relevant sequence is considered as `head`, `singleton` and `internal`, respectively.

the preferred source of taxonomic information as it provides a coherent taxonomy description covering the whole tree of life. In addition to OBITAXONOMY, which is able to reformat the NCBI taxonomy database (downloadable at the following URL: ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz) into the OBITOOLS format, two other OBITOOLSOBISILVA and OBIPR2 are dedicated to reformat the SILVA (Pruesse *et al*. 2007) and PR2 (Guillou *et al*. 2013) databases into the OBITOOLS format, respectively. Users deciding to use these customized taxonomies have to take care that the SILVA and PR2 `taxids` are not compatible with the NCBI `taxids`. Moreover, the LSU Silva `taxids` are not compatible with the SSU Silva `taxids`. These nonstandard `taxids` forbid to merge easily results obtained from several markers. The OBITAXONOMY command can enrich an existing taxonomy, whatever its origin (e.g. NCBI, SILVA or PR2) with private taxa, therefore enabling to associate sequence records to taxa not initially present in the reference taxonomic database. As the OBITOOLS have access to the full taxonomic tree topology, they are able to inform higher taxonomic levels from a taxon identifier (e.g. the family, order, class and phylum, corresponding to a genus) leading to efficient and simple annotation and querying of taxonomic information.

Implemented algorithms

Most of the algorithms implemented in the OBITOOLS (Table 1) are basic algorithms allowing sampling, filtering and annotation of sequence records based on their associated attribute set or sequence (e.g. OBISAMPLE, OBIGREP, OBIANNOTATE). Some others implement algorithms directly related to NGS or to DNA metabarcoding (e.g. ILLUMINAPAIREDEND, NGSFILTER, ECOTAG). Finally, a few of them do not run on sequence records and/or do not provide their results as sequence records. Amongst them, OLIGOTAG (Coissac 2012) generates a set of short oligonucleotide sequences (hereafter referred to as *tags*) useful to uniquely identify individual samples within a single NGS library containing many samples. Hereby, we will describe some of the implemented algorithms pertaining directly to DNA metabarcoding, as well as the corresponding programs. A full description of all programs included in the OBITOOLS suite is available on the web (http://metabarcoding.org/obitools/doc).

Table 1 List of OBITOOLS programs

Metabarcode design and quality assessment	
ECOTAXSPECIFICITY	Evaluates barcode resolution
File format conversions	
OBICONVERT	Converts sequence files to different output formats
OBIPR2	Converts PR2 database into an ECOPCR database
OBISILVA	Converts SILVA database into an ECOPCR database
OBITAB	Converts a sequence file to a tabular file
Sequence annotations	
ECOTAG	Assigns sequences to taxa
OBIANNOTATE	Adds/edits sequence record annotations
OBIADDTAXIDS	Adds taxids to sequence records using an ECOPCR database
Computations on sequences	
ILLUMINAPAIREDEND	Aligns paired-end Illumina reads
NGSFILTER	Assigns PCR product sequence records to their experiments/samples based on DNA tags and primers
OBICOMPLEMENT	Produces reverse complement sequences
OBICLEAN	Tags a set of sequences for PCR/sequencing errors identification
OBICUT	Trims sequences
OBIJOINPAIREDEND	Joins paired-end reads
OBIUNIQ	Groups and dereplicates sequences
Sequence sampling and filtering	
OBIEXTRACT	Extract samples from a data set
OBIGREP	Filters sequence file
OBIHEAD	Extracts the first sequence records
OBISAMPLE	Randomly resamples sequence records
OBISELECT	Selects representative sequence records
OBISPLIT	Splits a sequence file in a set of subfiles
OBISELECT	Selects representative sequence records
OBITAIL	Extracts the last sequence records
Statistics over sequence file	
ECODBTAXSTAT	Gives taxonomic rank frequency of a given ECOPCR database
OBICOUNT	Counts the number of sequence records
OBISTAT	Computes basic statistics for attribute values
Utilities	
OLIGOTAG	Designs a set of oligonucleotides with specified properties
OBISORT	Sorts sequence records according to the value of a given attribute
OBITAXONOMY	Manages taxonomic databases

Pairwise alignment of Illumina paired-end reads

Illumina sequencers have incomparable high sequencing capacity compared to other sequencing machines currently available on the market, but they produce relatively short sequence reads: 100 bases for the Illumina HiSeq 2000, 150 bases for the HiSeq 2500 and up to 300 bases for the MiSeq. To circumvent this limitation, a paired-end approach can be adopted that relies on the alignment of the forward and reverse reads to reconstruct the full-length amplicon consensus sequence. For this purpose, the ILLUMINAPAIREDEND program implements an exact dynamic programming alignment algorithm searching for the best 3′ end of the forward read matching the 3′ end of the reverse-completed reverse read. It takes into account quality scores associated with each

read avoiding the low-quality-based trimming of the read ends usually done. In our algorithm, each paired nucleotide is considered during the alignment process as partly a match and partly a mismatch. The proportion of matches and mismatches is estimated from the quality score of both paired nucleotides. Assuming a quality score Q for the nucleotide $N \in \{a, c, g, t\}$, the probability P_N that the read nucleotide N is actually a N is estimated using formula 1. The probability $P_{\bar{N}}$ to be one of the three other nucleotides is estimated using formula 2.

$$P_N = 1 - 10^{\frac{Q}{10}} \quad (1)$$

$$P_{\bar{N}} = \frac{1 - P_N}{3} \quad (2)$$

Using the two above formula, probabilities of a true match P_m and of a true mismatch $P_{\bar{m}}$ can be computed. The score associated with a pair of nucleotides in the alignment is a linear combination of a match reward and of a mismatch penalty weighted by P_m and $P_{\bar{m}}$. As inputs, ILLUMINAPAIREDEND takes the two FASTQ files corresponding to the forward and reverse reads and returns on the standard output a FASTQ file containing for each input pair of reads the consensus sequence of the alignment maximizing the likelihood of the sequence based on the probabilities estimated by formulas 1 and 2. The consensus sequence is annotated by a set of attributes describing the alignment statistics.

Extracting and demultiplexing amplicon sequences

Multiplexing of several samples in a single sequencing lane can be done using the indexing system of the DNA library offered by library preparation kits. To increase the number of multiplexed samples, it is possible to boost this basic system by the addition of tags on the 5′ end of the primers used for the PCR amplification. This additional tagging system, which can be set up using the OLIGOTAG program (Coissac 2012), is not processed by the sequencer software, so the end user has to decode those tags by himself during the first steps of raw data processing. In the OBITOOLS package, the sample tag decoding and amplification primer trimming are realized by the NGSFILTER program. This program uses a file describing the primer pair and tag(s) used for labelling each PCR product mixed into a library. NGSFILTER first identifies the primers using the Needleman and Wunsch algorithm implemented with the free-end-gap cost function (Erickson & Sellers 1983). The maximum number of allowed mismatches between primers and sequences can be set up using the `-e` option. Once primer pairs are identified, NGSFILTER searches for the tag on the 5′ end of each primer. No mismatches are allowed in tags identifying samples. NGSFILTER can deal with tagging systems on one of both ends of the amplicon, with the same or different tags on both extremities. As inputs, NGSFILTER requires a raw sequence file (potentially the output of ILLUMINAPAIREDEND) and a file describing the samples. As outputs, NGSFILTER writes the trimmed sequences annotated by a set of attributes indicating the primer pair found as well as the associated sample and experiment. Nonassigned sequences can be saved in another file using the `-u` option. Those sequences are annotated by a set of attributes describing the reason of their rejection.

Detection of PCR errors

Sequencing of PCR amplicons using NGS typically reveals a lot of variants for each sequence present in the amplified mix, and these correspond mainly to PCR errors (Schloss *et al.* 2011; Coissac *et al.* 2012). Two kinds of PCR errors exist: punctual errors (nucleotide substitutions or small indels) and chimeric sequences. This latter category can be identified using dedicated programs (e.g. Edgar *et al.* 2011; Quince *et al.* 2011; Wright *et al.* 2012). The OBICLEAN program aims to identify punctual PCR errors based on two underlying hypotheses. First, the error probability is small enough to assume that no more than one error occurs per DNA molecule and per PCR cycle. Consequently, if an erroneous sequence differs from a true sequence by more than one error, intermediate erroneous sequences must be present in the PCR product. Second, an erroneous sequence is always less abundant than the original one in the PCR product because it is created from a molecule already existing in the previous cycle. A property emerging from those two hypotheses is that a cascade of errors will generate a set of sequences with a decreasing abundance from the true sequence to the most erroneous ones. Based on this model, OBICLEAN builds a directed acyclic graph (DAG), where vertices correspond to all the sequence variants observed in a single PCR (Fig. 3). Each vertex is weighted by its abundance estimated by the read count corresponding to this sequence. Edges link sequences exhibiting just a single difference (substitution, insertion or deletion of one nucleotide) under the condition that $R = A_e/A_o$ is smaller than a given threshold, fixed empirically by default to 0.5 (with A_e the abundance of the erroneous sequence and A_o the abundance of the original sequence). Edges are oriented from the heaviest vertex to the lightest one. According to their position in the DAG, sequences are labelled using one of three status: `head`, `internal` or `singleton`. The sequences corresponding to summits of the DAG are labelled as `head` and ideally correspond to true sequences. Sequences related to no other sequences in the DAG are labelled as `singleton`. This status may

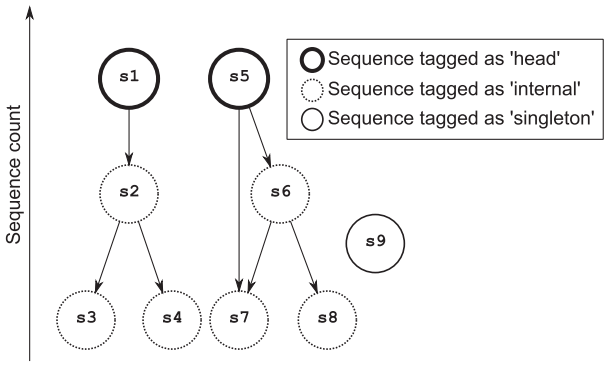

Fig. 3 An example of the topology of the OBICLEAN directed acyclic graph (DAG) used to identify PCR errors – circles represent sequences, and arrows represent sequence similarities.

correspond to rare true sequences. All other sequences are labelled as `internal` and consequently considered as erroneous sequences. Sequence records submitted to OBICLEAN are usually the result of the sequencing of several PCR products; OBICLEAN efficiently handles such cases by comparing only once the whole set of sequences and then labels each sequence record by one of the three possible status for every PCR. Even if sequencing errors are not explicitly represented in the OBICLEAN model, they will be treated as PCR errors, connected in a terminal position of the DAGs and labelled with the internal status. Chimeric sequences usually differ from true sequences by more than one difference accumulated in one single step because of a recombination event. As a result, chimeric sequences are usually labelled as `head` or `singleton`. Sequences annotated by OBICLEAN can be further postprocessed using chimeric sequences identification program like PERSEUS (Quince *et al.* 2011) to point out the most evident chimeric sequences. As input, OBICLEAN requires a dereplicated sequence file like those produced by OBIUNIQ. The output is the same set of sequences with new attributes indicating their OBICLEAN status in each PCR and the number of times they get each one of the three statuses over all the PCR.

Taxonomic assignation of sequences

In many instances, assigning a sequence to a taxon is the ultimate step of the DNA metabarcoding sequence analysis process. This task is addressed by the ECOTAG program, which can be considered as a supervised classification algorithm. ECOTAG uses three inputs: (i) the data set of sequences to be annotated, (ii) a taxonomy database defining relationship between taxa and (iii) a reference sequence database containing a set of sequences annotated by a taxid linking them to the taxonomy database (see Fig. 2 for an example). ECOTAG compares each sequence of the data set (the query) to the reference database. ECOTAG makes the assumption that both the query and the reference sequences are full-length barcodes. The similarity between a query sequence and a reference sequence is thus measured as the ratio between the length of the longest common substring (LCS) and the length of the shortest alignment corresponding to this LCS. This ratio corresponds to the fraction of identity between the query and reference sequences and can be easily computed using a global alignment algorithm derived from Needleman & Wunsch (1970). The ECOTAG program runs in three phases. First, for each query sequence, it searches for the most similar sequence B_{seq} in the reference database and keeps the similarity ratio S_{query} in memory. In a second step, it builds a set B_{ref} composed of all the sequences in the reference database exhibiting a similarity ratio higher than S_{query} with B_{seq}. In the last step, it uses the taxonomic database tree to determine the last common ancestor of all the B_{ref} sequences. The corresponding taxid is then assigned to the query sequence. The output of ECOTAG consists in the input data set of sequences annotated by a taxid.

Implementation

The OBITOOLS are a set of PYTHON programs relying on an eponym PYTHON library. The OBITOOLS library is mainly developed in PYTHON (version 2.7, see http://www.python.org). For increasing the speed of execution, many parts of the OBITOOLS library are developed using CYTHON (http://cython.org/, a PYTHON to C compiler) or the C language directly. For optimizing computation time on huge data set, the OBIDISTRIBUTE program by splitting data set in many chunks allows to easily build massively parallel pipeline able to run on computer grid very efficiently. The OBITOOLS compile on UNIX systems including LINUX and MACOSX.

Availability of the OBITOOLS

The OBITOOLS are open source and protected by the CECILL 2.1 licence (http://www.cecill.info/licences/Licence_CeCILL_V2.1-en.html). All the sources can be downloaded from our git server (https://git.metabarcoding.org/obitools/obitools). The OBITOOLS are deposited on the PYTHON package index (https://pypi.python.org/pypi/OBITools) and therefore can be installed using the classical PYTHON package installer *pip* (https://pypi.python.org/pypi/pip) and also available on the metabarcoding.org web site (http://metabarcoding.org/obitools). The complete documentation is available at http://metabarcoding.org/obitools/doc. PYTHON 2.7, a C compiler and CYTHON have to be installed prior to the OBITOOLS. The OBITOOLS Galaxy (Giardine *et al.* 2005) wrapper can be downloaded from the GenOuest core facility

toolshed (http://toolshed.genouest.org) in the next- generation sequencing section.

Acknowledgements

We would like to thank Claudia Heriveau and Cyril Monjeaud of the GenOuest core facility, CNRS UMR 6074 IRISA-INRIA, Campus de Beaulieu, 35042 Rennes Cedex, for their help during the development of the OBITOOLS Galaxy wrapper; Harald Gruber-Vodicka of the Max Plank Institut, Bremen, Germany, for his help to integrate SILVA taxonomy in OBITOOLS. The authors acknowledge financial support from the Agence Nationale de la Recherche (ANR) through the METABAR project.

References

Andersen K, Bird KL, Rasmussen M *et al.* (2012) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Baldwin DS, Colloff MJ, Rees GN *et al.* (2013) Impacts of inundation and drought on eukaryote biodiversity in semi-arid floodplain soils. *Molecular Ecology*, **22**, 1746–1758.

Bellemain E, Davey ML, Kauserud H *et al.* (2013) Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from Arctic permafrost. *Environmental Microbiology*, **15**, 1176–1189.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Chariton AA, Court LN, Hartley DM, Colloff MJ, Hardy CM (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, **8**, 233–238.

Coissac E (2012) OligoTag: a program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in Molecular Biology*, **888**, 13–31.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, **13**, 3021–3030.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Erickson B, Sellers P (1983) Recognition of patterns in genetic sequences. In: *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (eds Sankoff D, Kruskal JB), pp. 55–91. Addison-Wesley, Reading, Massachusetts.

Ficetola GF, Coissac E, Zundel S *et al.* (2010) An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, **11**, 434.

Giardine B, Riemer C, Hardison RC *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.

Guillou L, Bachar D, Audic S *et al.* (2013) The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**, D597–D604.

IUPAC-IUB Commission on Biochemical Nomenclature (CBN) (1968) A one letter notation for amino acid sequence. *European Journal of Biochemistry*, **5**, 151–153.

Kowalczyk R, Taberlet P, Coissac E *et al.* (2011) Influence of management practices on large herbivore diet-case of European bison in Białowieża Primeval Forest (Poland). *Forest Ecology and Management*, **261**, 821–828.

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.

Parducci L, Jørgensen T, Tollefsrud MM *et al.* (2012) Glacial survival of boreal trees in northern Scandinavia. *Science*, **335**, 1083–1086.

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.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **39**, e145.

Sønstebø 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.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS One*, **6**, e27310.

Shehzad W, Riaz T, Nawaz MA *et al.* (2012) Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*prionailurus bengalensis*) in Pakistan. *Molecular Ecology*, **21**, 1951–1965.

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 USA*, **103**, 12115–12120.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Thomsen PF, Kielgast J, Iversen LL *et al.* (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Valentini A, Miquel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.

Wright ES, Yilmaz LS, Noguera DR (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Applied and Environmental Microbiology*, **78**, 717–725.

Yoccoz NG, Bråthen KA, Gielly L *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.

F.B. participated to the software development, to its maintenance, and contributed to draft the documentation, C.M. helped with the software development and with the documentation, A.B. and P.T. provided feedback on the tools and assisted with the manuscript and documentation drafting, Y.L. coordinated the development of the Galaxy wrapper, and E.C. initiated the OBITOOLS project, tackled the software development, wrote the first draft of the manuscript and contributed to draft the documentation.

Data accessibility

OBITOOLS software is available as described in the 'Availability of the OBITOOLS' section. A training data set can be downloaded from the documentation web site (http://metabarcoding.org/obitools/doc/wolves.html).