MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1415–1427 doi: 10.1111/1755-0998.12546

metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data

GUILLAUME LOUVEL,*[1] CLIO DER SARKISSIAN,*[1] KRISTIAN HANGHØJ* and LUDOVIC ORLANDO*†

*Centre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark, †Université de Toulouse, University Paul Sabatier (UPS), Laboratoire AMIS, CNRS UMR 5288, 37 allées Jules Guesde, 31000 Toulouse, France

Abstract

Micro-organisms account for most of the Earth's biodiversity and yet remain largely unknown. The complexity and diversity of microbial communities present in clinical and environmental samples can now be robustly investigated in record times and prices thanks to recent advances in high-throughput DNA sequencing (HTS). Here, we develop metaBIT, an open-source computational pipeline automatizing routine microbial profiling of shotgun HTS data. Customizable by the user at different stringency levels, it performs robust taxonomy-based assignment and relative abundance calculation of microbial taxa, as well as cross-sample statistical analyses of microbial diversity distributions. We demonstrate the versatility of metaBIT within a range of published HTS data sets sampled from the environment (soil and seawater) and the human body (skin and gut), but also from archaeological specimens. We present the diversity of outputs provided by the pipeline for the visualization of microbial profiles (barplots, heatmaps) and for their characterization and comparison (diversity indices, hierarchical clustering and principal coordinates analyses). We show that metaBIT allows an automatic, fast and user-friendly profiling of the microbial DNA present in HTS shotgun data sets. The applications of metaBIT are vast, from monitoring of laboratory errors and contaminations, to the reconstruction of past and present microbiota, and the detection of candidate species, including pathogens.

Keywords: ancient DNA, metagenomics, microbial profiling, microbiome, shotgun sequencing

Received 6 October 2015; revision received 3 May 2016; accepted 13 May 2016

Introduction

With up to billions of short DNA reads provided in a both cost- and time-effective manner by high-throughput DNA sequencing (HTS) technologies (Metzker 2010), microbial ecology no longer relies on culture-dependent methods for characterizing the diversity, structure and function of microbial communities. HTS motivated massive international collaborations to unveil the most abundant, but yet largely unknown, environmental microbes present in virtually all ecosystems on Earth [e.g. the Earth Microbiome Project (Gilbert *et al.* 2014), the Tara Oceans Expedition (Sunagawa *et al.* 2015)]. It also unleashed an unprecedented characterization of the collective population of micro-organisms living both on and within us [e.g. the Metagenomic Human Intestinal Tract (MetaHIT) project (Qin *et al.* 2010) or the Human Microbiome Project (HMP) (Human Microbiome Project Consortium 2012a,b)]. Human-associated microbes outnumber by 10-fold the number of human cells in each individual and are increasingly recognized to influence the development of major diseases and disorders, such as obesity (Turnbaugh *et al.* 2009), type 2 diabetes (Qin *et al.* 2012) and schizophrenia (Dinan *et al.* 2014).

The gold standard method for microbiota profiling is based on the molecular diversity of the 16S ribosomal DNA (16S rDNA), building on the growing number of 16S rDNA reference sequences available in public databases (e.g. the April 2016 release of the SILVA (SSURef 126) database included 5 366 469 bacterial and archaeal sequences; Quast *et al.* 2013). This approach has shown great success on a wide panel of environmental and biological material, ranging from soils (e.g. Lauber *et al.* 2009; Fierer *et al.* 2012), seawater (e.g. Sogin *et al.* 2006) to

Correspondence: Ludovic Orlando, Fax: +4535322325; E-mail: lorlando@snm.ku.dk

[1]These authors contributed equally to this work.

the human mouth, gut and skin (e.g. Costello *et al.* 2009; Qin *et al.* 2010; Human Microbiome Project Consortium 2012a). DNA extracted from ancient dental plaque also revealed the persistence of oral microbiomes in the archaeological record, adding a temporal dimension to microbiome analyses (Adler *et al.* 2013; Warinner *et al.* 2014). 16S rDNA-based microbial profiling has, however, a number of limitations, reviewed in Pedersen *et al.* (2015) and Brooks *et al.* (2015), including bias during PCR amplification (e.g. taxa-dependent primer binding efficiency or %GC-dependent amplification success), limitations in taxonomic resolution and calculation of relative abundances, due to variable copy numbers of the 16S rRNA gene per-cell among microbial species (Ziesemer *et al.* 2015).

Metagenomic shotgun sequencing of the entire population of DNA molecules in a sample has recently emerged as a robust alternative to 16S rDNA sequencing for microbiota profiling, both in terms of taxonomic and functional diversity (Segata *et al.* 2013). This approach provides fundamental knowledge on microbial communities and their ecology, but can also help guide restoration efforts of natural environments and suggest novel therapeutic strategies.

HTS metagenomic analyses face many challenges due to the inherent complexity of metagenomic data sets, which consist of a mixture of DNA sequences from multiple origins. To ensure that microbial profiles truly represent the community of interest, sampling should be carried out with extreme care, (i) including sample replicates, (ii) avoiding contamination by environmental microbes and (iii) preserving environmental conditions precluding DNA and/or microbial degradation until DNA extraction. The latter can be biased towards particular microbes, as differences in bacterial cell wall structures can lead to differential cell lysis (Ferrand *et al.* 2014). Moreover, the molecular tools underlying microbial profiling, for example, the types of DNA polymerases used during DNA library amplification (Dabney & Meyer 2012) and/or the DNA sequencing platforms, can favour DNA templates showing particular base compositions and skew the molecular representation originally present in the extract. Another limitation is that metagenomic shotgun analyses are computationally heavy and require large numbers of DNA sequences, typically at least in the range of a million reads, to capture the full diversity within a sample. Importantly, the uneven representation of microbial organisms in the reference databases used to perform taxonomic assignment based on sequence similarity represents one of the main limitations. As most micro-organisms still remain either unknown, uncharacterized, uncultured and/or unsequenced, most assignment programs show significant false-negative rates. This has two important consequences. Firstly, at the species or strain levels, reads from a yet unsequenced species/strain can be inaccurately assigned to the sole closely related species/strain currently known. Secondly, relative abundances can be distorted by the presence of unassigned reads originating from uncharacterized organisms absent from the comparative database. This bias is further magnified when identification markers are present in different copy numbers across taxa, when genetic material is shared across taxa due to horizontal transfers, and when biases in genome sizes and database representativeness are not accounted for. Finally, the short-read length generated by most current HTS platforms, which is even more limited in highly degraded (ancient) samples (Orlando *et al.* 2015), further increases the complexity of taxonomic assignment.

The metagenomic module of PALEOMIX, an open-source and user-friendly pipeline for whole-genome resequencing analyses (Schubert *et al.* 2014a), was originally developed to facilitate the characterization of the microbial content in HTS data sets obtained from both modern and ancient specimens. It uses the program METAPHLAN [metagenomic phylogenetic analysis (Segata *et al.* 2012; Truong *et al.* 2015)], which exploits a balanced database of unique clade-specific markers to carry out microbial taxonomic assignment and calculate taxon relative abundances. By construction, the reduced and curated METAPHLAN database addresses the limitations of meta-taxonomic assignment caused by horizontal transfer and gene duplication, while increasing processing speed. METAPHLAN constitutes a useful tool for characterizing and comparing the microbial taxonomic content of complex samples and allows initial screening for identifying candidate groups of interest (e.g. pathogenic species/strains). However, due to the limitations of currently available comparative databases highlighted above, assignments at the species and strain levels should be considered carefully and independently validated. Possible validation approaches include sequence alignment against candidate reference genomes followed by phylogenetic placement within the larger group of related bacteria. Whole-genome targeted enrichment approaches can be used to increase the coverage of specific candidate genomes (e.g. pathogens) available for phylogenetic reconstruction (Bos *et al.* 2011). Additionally, de novo genome assemblies can be attempted following existing methods tailor-made for metagenomic mixtures (Lai *et al.* 2015).

Although applied to shotgun HTS data from a range of archaeological, herbarium and museum specimens (Der Sarkissian *et al.* 2014; Schubert *et al.* 2014a,b; Der Sarkissian *et al.* 2015; Librado *et al.* 2015; Seguin-Orlando *et al.* 2015), the PALEOMIX metagenomic module only consisted of a long series of individual command lines (Schubert *et al.* 2014a). Here, we introduce metaBIT

(metagenomic Biomarker Inferred Taxonomy), a Python and R pipeline representing an automatization of the PALEOMIX metagenomic module. The metaBIT pipeline is publicly available at https://bitbucket.org/Glouvel/metabit. It can be run easily using the default parameters described in the 'Material and Methods' section and is provided with an extensive documentation and examples. Importantly, each step of metaBIT is customizable by the user, who has the possibility to select the profiling accuracy/sensitivity, the type of analysis and the most appropriate parameters given the nature of the data analysed. We should therefore caution that it is essential that metaBIT outputs are critically interpreted with regard to the parameters chosen by the user. With the objective to illustrate how a whole range of scientific questions can be addressed through metaBIT, we present the diversity of output produced by the pipeline, processing publicly available HTS data sets of environmental seawater samples (Hingamp *et al.* 2013), skin and gut human clinical samples (HMP, Human Microbiome Project Consortium 2012a; MetaHIT, Qin *et al.* 2010), ~1000-year-old human dental calculus (Warinner *et al.* 2014) and ~300-year-old human mummies showing evidence of tuberculosis infection (Kay *et al.* 2015).

Material and methods

metaBIT analyses the microbial diversity present in complex samples that have been shotgun sequenced using a variety of HTS platforms. The pipeline follows four main analytical steps (Fig. 1). HTS reads are first mapped against METAPHLAN databases and taxonomically assigned following removal of PCR duplicates. The last step groups visualization of the microbial profiles and a variety of statistical analyses in the R statistical environment (version ≥3), including calculation of diversity indices, principal coordinates analyses (PCoA), hierarchical clustering, and eventually, biomarker identification, as implemented in LEfSe [linear discriminant analysis effect size (Segata *et al.* 2011)]. We strongly recommend that users refer to the documentation of the respective programs and packages that are part of metaBIT to ensure that the selected analytical parameters correspond to the users' objectives.

Input data

As input, metaBIT takes compressed fastq files containing reads trimmed for adapter sequences using programs such as ADAPTERREMOVAL v2 (Schubert *et al.* 2016) implemented in PALEOMIX (Schubert *et al.* 2014a). metaBIT can handle both single-end or paired-end reads: overlapping mate reads transformed into single 'collapsed' reads, nonoverlapping pair mates (subsequently referred to as 'paired' reads) and singletons ('singles'), for which one mate was lost following quality filtering (Orlando *et al.* 2013; Schubert *et al.* 2014b).

Read alignment against the METAPHLAN database

Using metaBIT, DNA reads are first mapped to the BOWTIE2 METAPHLAN database version 1.7.8 (Segata *et al.* 2012) or 2.0 (Truong *et al.* 2015) using the BOWTIE2 v2.1.0 ALIGNER (Langmead & Salzberg 2012). To ensure accurate abundance estimation when analysing paired-end reads, only one mate is counted when both mates of a read-pair map to the same METAPHLAN marker, while both mates are disregarded when mapping to different markers. Following read mapping, metaBIT produces the two output files of BOWTIE2, consisting of an unsorted SAM alignment file 'reads.*paired*/*collapsed*/*singles*.bowtie2out.sam' and an alignment statistic file 'reads.*paired*/*collapsed*/*singles*.bowtie2out.stats' for *paired*-end, *collapsed* or *single*-end reads, which includes the total number of reads considered and the number of reads with no or at least one hit. For paired-end data, the number of pairs where both mates map to the same marker is also indicated and an additional 'reads.*paired*.bowtie2out.filter.sam' SAM alignment file is produced after filtering.

PCR duplicate removal

Aligned reads are sorted with the SAMtools *sort* utility version ≥0.1.8 (Li *et al.* 2009) before PCR duplicates are removed. For noncollapsed reads, duplicates are identified and removed (option REMOVE_DUPLICATES = true) with the MarkDuplicates function of PICARD TOOLS version 1.119 (http://broadinstitute.github.io/picard/). For collapsed reads, duplicates are filtered using the '*rmdup_collapsed.py*' Python script from PALEOMIX (Schubert *et al.* 2014a). Three output files are generated, including a sorted BAM alignment (binary SAM file), 'reads.*paired*/*collapsed*/*singles*.bowtie2out.filter.sorted.bam', a 'reads.*paired*/*collapsed*/*singles*.bowtie2out.filter.sorted.no_dup.bam' BAM alignment filtered for PCR duplicates and a 'reads.*paired*/*singles*.bowtie2out.filter.sorted.no_dup.metrics' file providing clonality statistics.

Microbial profiling using METAPHLAN

From the duplicate-filtered BAM alignments, the query template names and reference sequence names are used as input for the Python program METAPHLAN version 1.7.8 or 2.0. This generates, for each sample, tabulated tables ('taxa.tsv') of relative abundances for all identified taxa at all taxonomic levels from kingdom to species (or strain in METAPHLAN version 2.0). The proportion of unidentified taxa is recorded as 'unclassified' for each taxonomic level considered. A 'summary_readcounts.tsv' table is also

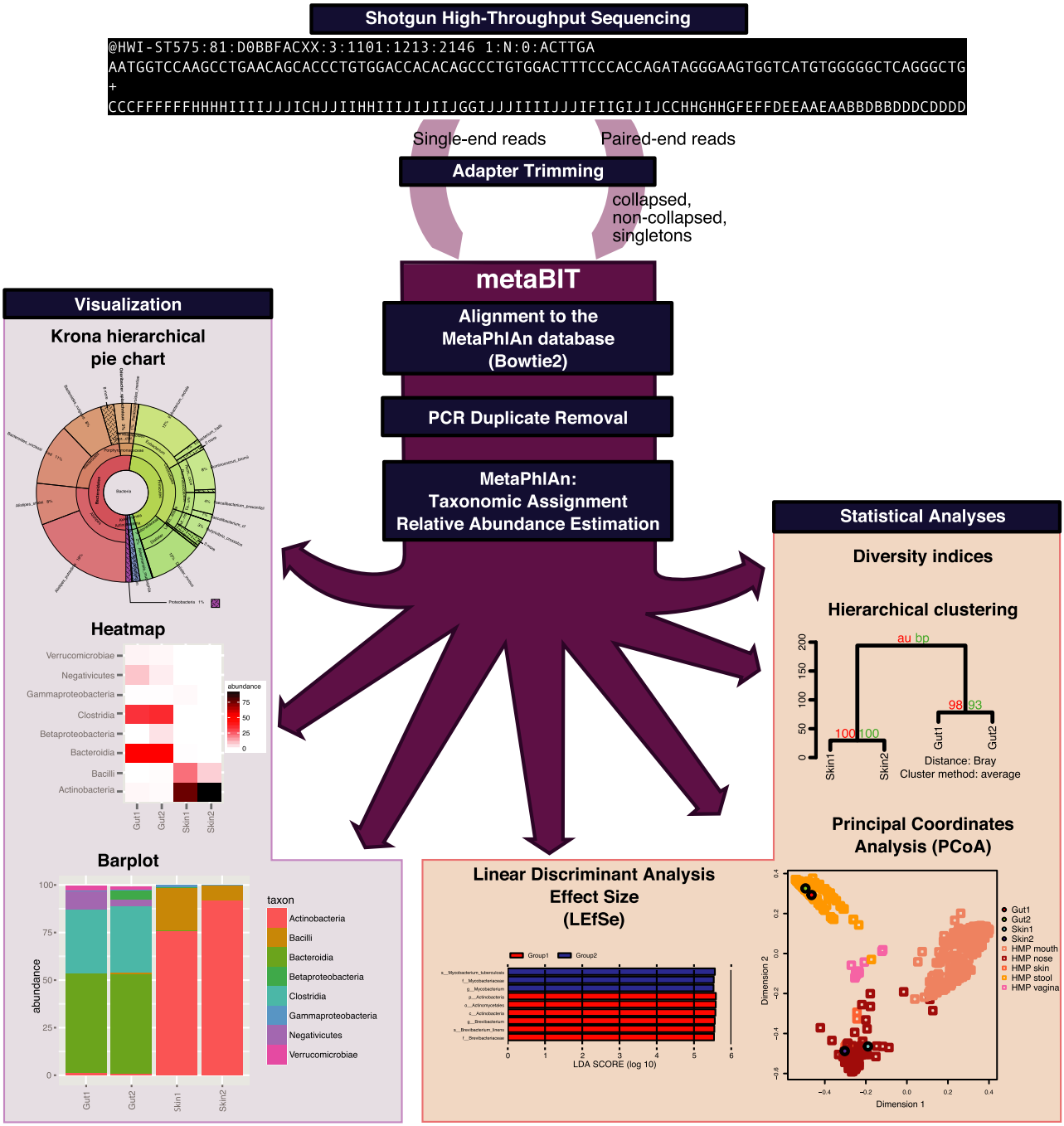

Fig. 1 metaBIT flow chart.

produced, providing the total number of reads processed, mapped and uniquely mapped.

Merging relative abundance tables

For all samples considered in subsequent analyses, the individual tables of relative taxonomic abundances are merged into a 'all_taxa.tsv' table, which is further processed for a single or multiple sets of analyses, each defined by the user under a specific *name*. This table can be filtered for low-abundance taxa, according to a user-defined threshold, in order to reduce the impact of false-positive identifications. This produces, for each set *name* of statistical analysis, one '*name*_filtered_*level*.tsv' table per taxonomic *level*, which reports relative abundances post-filtering.

Data visualization and diversity indices

Relative taxon abundances for individual profiles at all taxonomic levels can be visualized using dynamic Krona pie charts (Ondov *et al.* 2011) produced by the Krona Tools scripts (http://sourceforge.net/projects/krona/). Conversions from METAPHLAN to Krona formats (saved as files with the 'krona.in' extension) are carried out automatically by an integrated version of the METAPHLAN 'metaphlan2krona.py' script (Segata *et al.* 2012), and resulting Krona pie charts are recorded in a single file named 'all_taxa.krona.html'.

Relative abundances can also be visualized at all taxonomic levels by stacked barplots ('*name*_barplot_*level*.pdf' files) and heatmaps ('*name*_heatmap_*level*.pdf' files), using the *ggplot2* package (http://cran.r-project.org/package=ggplot2) and the *hclust* R function. Diversity indices are calculated at all taxonomic levels using the function *diversity* of the R package *vegan* (http://cran.r-project.org/package=vegan) and are provided in a '*name*_diversities.tsv' tabulated file (see corresponding documentation for a full list of supported indices).

Principal coordinates analyses

Pairwise ecological distances among profiles can be computed at all taxonomic levels using taxon relative abundances and the *vegdist* function from the *vegan* package (see corresponding documentation for a full list of supported distances). Ecological distances are compiled within individual tabulated tables for each taxonomic level and written in individual files '*name*_pcoa_*levels*.tsv'. These are used for PCoA with the R function *pcoa*, which are plotted as '*name*_pcoa_*level*.pdf' files.

Hierarchical clustering

Hierarchical clustering analyses can be automatically carried out, using a modified version of the R package *pvclust* (Suzuki & Shimodaira 2006) to allow the use of Bray–Curtis distances. Cluster supports are estimated by approximately unbiased *P*-values and bootstrap probabilities. The method used for distance calculation and for clustering (a full list of supported distances and clustering methods is provided in the documentation), as well as the number of bootstrap iterations, is user-defined. Distance tables are saved as '*name*_clust_*level*.tsv' and hierarchical clustering trees as '*name*_clust_*level*.pdf'.

LEfSe analyses

Visualization (barplot, heatmap) and statistical analyses (PCoA, hierarchical clustering) of the taxonomic profiles can help detect potentially relevant structures within the microbial diversity of the samples analysed. The statistical support for these structures can then be assessed in a second step by linear discriminant analysis (LDA), using LEfSe (Segata *et al.* 2011), which also identifies and quantifies taxa (biomarkers) driving statistically significant differences between these user-defined sample groups.

Comparative data sets

As part of the metaBIT distribution package, we provide merged tables of taxon relative abundances at all taxonomic levels (using both METAPHLAN versions 1.7.8 and 2.0) for soil and human-associated samples to facilitate future cross-environment and cross-tissue comparisons. The 15 soil microbial profiles were generated in Der Sarkissian *et al.* (2014) using the data described in Fierer *et al.* (2012) and obtained from cold deserts ($N = 5$), hot deserts ($N = 3$), tropical forests ($N = 2$), temperate forests ($N = 2$), the temperate grassland, the boreal forest and the Arctic tundra. The human microbiome (HMP) data set is composed of 689 profiles published in Human Microbiome Project Consortium (2012a) and obtained from samples of the mouth ($N = 382$), skin ($N = 26$), nose ($N = 87$), vagina ($N = 56$) and stools ($N = 138$).

Implementation and makefile preparation

The pipeline performs all analyses described above automatically, allowing multiple samples to be processed in parallel. It uses as input a single makefile in yaml format (http://yaml.org/), which contains the fastq file locations, the data and sample structure, the type of statistical analyses to be performed, as well as their options and parameters.

In the section 'Samples' of the metaBIT makefile, the user defines the names of the samples and indicates the location of the reads (see examples in Fig. 2 and Fig. S1, Supporting information). If several libraries have been generated and sequenced for a given sample, the user has the possibility to run the metaBIT pipeline on a per-library basis in a first instance and to pool metaBIT profiles per sample in a subsequent run (see the option 'Pool:' in the makefile section 'Taxonomic profiling/Metaphlan' in Fig. S1, Supporting information). Examples of how to define samples/libraries and indicate locations for different sequence read types ('Collapsed', 'Paired' or 'Singles') are given in Fig. 2 and Fig. S1 (Supporting information).

In the section 'Taxonomic profiling', the user has the possibility to set parameters for mapping sequence reads to the METAPHLAN databases using BOWTIE2 and for taxon assignment/abundance calculation using METAPHLAN (see the documentation of each program for a comprehensive list of options). For example, when using METAPHLAN

```
#Samples:
Samples:
  Sample1:
    Lib1PE:
      Collapsed: folder/Lib1PE/date1/lane*/reads.collapsed.gz
      Paired:
        - folder/Lib1PE/date1/lane*/reads.pair{Pair}.gz
        - folder/Lib1PE/date2/lane*/reads.pair{Pair}.gz
      Singles: folder/Lib1PE/date1/lane*/reads.singleton.gz
    Lib2PE: folder/Lib2PE/date1

  Sample2:
    Lib1SE:
      Singles: folder/Lib1SE/date1/lane*/reads.singleton.gz

#Taxonomic Profiling
Bowtie2:
  --end-to-end: yes
  --sensitive: yes

Metaphlan:
  --ignore_eukaryotes:
  --ignore_viruses:

#Statistical analysis of taxonomic profiles
Krona:
  run: yes

Statax:
  MyAnalysis1:
    merge:
      - Path/to/previous/all_taxa.tsv
    taxlevels: g
    filterout: 1
    doDiv:
      --index: shannon
    doBarplot:
    doHeatmap:
    doPcoa:
      --distance: bray
    doClust:
      --method.dist: bray
      --method.hclust: average
      --nboot: 10000

  MyAnalysis2:
    doDiv:
      --index: simpson

Lefse:
  run: yes
  merge:
    - Comparative_all_taxa.tsv
  Groups:
    ComparativeGroup1:
      - Sample1_Lib1PE
      - Sample1_Lib2PE
      - Comparative_Sample1
    ComparativeGroup2:
      - Sample2_Lib1SE
      - Comparative_Sample2
      - Comparative_Sample3
      - Comparative_Sample4
```

Fig. 2 Main functions and set-up of the metaBIT makefile. Built-in structure text of the metaBIT makefile is coloured in white, user-defined names in orange and user-defined paths/parameters in black. Lines starting with '#' are comments and are ignored.

version 2.0, the user can choose to include or exclude eukaryotes and viruses from taxonomic profiles (Fig. 2 and Fig. S1, Supporting information).

In the section 'Statistical analysis of taxonomic profiles', the user defines visualization tools and statistical methods for profiling microbes (Fig. 2 and Fig. S1, Supporting information), including Krona tables, the rest being piloted by the 'Statax' subsection. There, multiple sets of analyses (for example, computing different diversity indices) can be specified and run simultaneously, each set identified by a unique identifier, subsequently used as a folder name containing output files. This makes metaBIT a useful tool for parameter exploration and selection. In the case where other samples were previously profiled using the same version of METAPHLAN, the merged table of taxon relative abundances at all taxonomic levels (all_taxa.tsv) can be used for comparison. This is performed by providing the path to the comparative table through the 'merge' option (Fig. 2 and Fig. S1, Supporting information). The taxonomic levels at which the analyses should be performed are defined with the option 'taxlevels:', which accepts a combination of the following: 'k' for kingdom, 'p' for phylum, 'c' for class, 'o' for order, 'f' for family, 'g' for genus, 's' for species and 't' for strain. The 'filterout:' option allows the user to set a minimal value for relative abundances for the taxa to be considered in downstream statistical analyses. The user can then select which statistical analyses should be run by calling the options 'doDiv' for diversity index calculation, 'doBarplot' for barplot visualization of taxon relative abundances, 'doHeatmap' for heatmap visualization, 'doPcoa' for PCoA and 'doClust' for hierarchical clustering (Fig. 2 and Fig. S1, Supporting information). All parameters and options for these analyses are detailed in the help menu of the corresponding R scripts distributed in the metaBIT package. From these analyses, patterns of differentiation among groups of samples can emerge, which the user can test statistically with LEfSe by defining groups (and potentially subgroups) of samples in the subsection 'Lefse' (disabled by default) (Fig. 2 and Fig. S1, Supporting information). This is typically performed in a second analysis after a first analysis has helped identify interesting groups/subgroups potentially structuring the communities investigated.

Example data sets

We have selected five previously published data sets to showcase the most common features of metaBIT. These include shotgun HTS reads of seawater samples (Hingamp *et al.* 2013), human skin (Human Microbiome Project Consortium 2012a), human gut (Qin *et al.* 2010), ~1000-year-old human dental calculus (Warinner *et al.* 2014) and ~300-year-old human remains, including

individuals showing evidence of tuberculosis infection (Kay *et al.* 2015) (Table S1, Supporting information).

Seawater data sets are part of the Tara Oceans Expedition and were selected to represent the marine environment of the Antarctic, the South Atlantic Ocean and the Arabian Sea, including for each location, two surface water and two mesopelagic samples. Human skin data sets were generated by the HMP from two swabs of the right retro-auricular crease of two healthy individuals (Qin *et al.* 2010), while human gut data sets were obtained from metaHIT faeces samples of two healthy Danish individuals (Human Microbiome Project Consortium 2012a). Ancient dental calculus data sets were derived from two adult individuals, B61 and G12, excavated in the St Petri convent complex of Dalheim, Germany, and dated to 950–1200 CE (Warinner *et al.* 2014). For B61 and G12, a total of four and five data sets were, respectively, generated from independent extracts, following different decontamination and extraction protocols. Finally, we analysed 42 data sets from 26 ancient human remains, dated to the 18th century CE and found in the crypt of the Vác Dominican church, Hungary (Kay *et al.* 2015).

Microbial profiling was performed using all single-end reads for ancient dental calculus samples and all paired-end reads (read pairs and single reads) for modern seawater, human skin, and human gut samples. For the ancient human samples from Kay *et al.* (2015), we only used collapsed read pairs, to limit the possible impact of contamination by modern microbial DNA, which is expected to exhibit longer fragment lengths than damaged endogenous DNA fragments. For all paired-end data, read collapsing was achieved using ADAPTERREMOVAL v2 (Schubert *et al.* 2016) as implemented in the PALEOMIX pipeline (Schubert *et al.* 2014a). The microbial profiles were obtained by running the metaBIT pipeline with the '–end-to-end' mode and a sensitive alignment strategy (default) for BOWTIE2 read mapping to the METAPHLAN databases. Taxa showing abundances <1% (default) were not taken into consideration in further analyses. We then summarized taxonomic richness and evenness across profiles by computing Shannon alpha diversity indices (Shannon 1948). PCoA was performed on the commonly applied Bray–Curtis distances among profiles (Bray & Curtis 1957), on which hierarchical clustering was also based using an average linkage clustering method and 10 000 bootstrap iterations (default), following Schubert *et al.* (2014a). LEfSe was applied using the nonparametric Kruskal–Wallis rank-sum test and the unpaired Wilcoxon test with *P*-values of 0.05, and threshold values of 2 on the logarithmic LDA score (default). We compared the seawater microbial profiles against the soil comparative data set and those from human skin, gut and dental calculus samples against the HMP comparative data set. Typical running times for all analyses performed here are provided in Table S2 (Supporting information).

Results

Environmental samples

We first showcase the application of metaBIT to environmental samples, here represented by shotgun HTS data sets from twelve seawater samples (Sunagawa *et al.* 2015) and the comparative soil data set. For these analyses, we used the METAPHLAN database version 1.7.8 for taxonomic profiling, as version 2.0 showed conflicting taxon annotations, with different markers for species *Candidatus Pelagibacter ubique* and *Citromicrobium bathyomarinum*, and order Nitrosopumilales displaying different annotations at higher taxonomic levels (e.g. Erythrobacteraceae and Sphingomonadaceae at the family level for *Citromicrobium bathyomarinum*; see other examples in Table S3, Supporting information). Hierarchical clustering at the genus level (Fig. 3A) showed maximal approximately unbiased *P*-values (au) and bootstrap probabilities (bp) supporting a differentiation between 'seawater' (31 biomarkers using LEfSe) and 'soil' (7 biomarkers; Fig. 3B) samples on the one hand and between 'non-Antarctic Mesopelagic' (15 biomarkers) and 'non-Antarctic Surface Water/All Antarctic' (31 biomarkers; Fig. 3C) seawater samples on the other hand. Seawater sample replicates from the same geographical environment and sampling zone were found to cluster together, in support for the robustness of the profiling method (Fig. 3A). These results recapitulate previous findings (Fierer *et al.* 2012; Sunagawa *et al.* 2015) and show that metaBIT can be effectively used to describe environmental microbial communities, despite the fact that environmental micro-organisms are notoriously less well described than human microbes, in particular pathogens (Fig. 3C). We encourage environmental microbial ecologists to take advantage of the soil comparative data set made available as part of the metaBIT package for a fast and easy comparison of their study microenvironments. We, however, acknowledge that the sensitivity of the method for profiling soil and seawater samples is limited due to the incompleteness of the current reference database.

Human clinical samples

We next demonstrated the application of metaBIT and the METAPHLAN database version 2 to the analysis of the microbial diversity in human faecal and skin clinical samples. Taxon distributions were in accordance with previous descriptions of the healthy skin (Grice & Segre

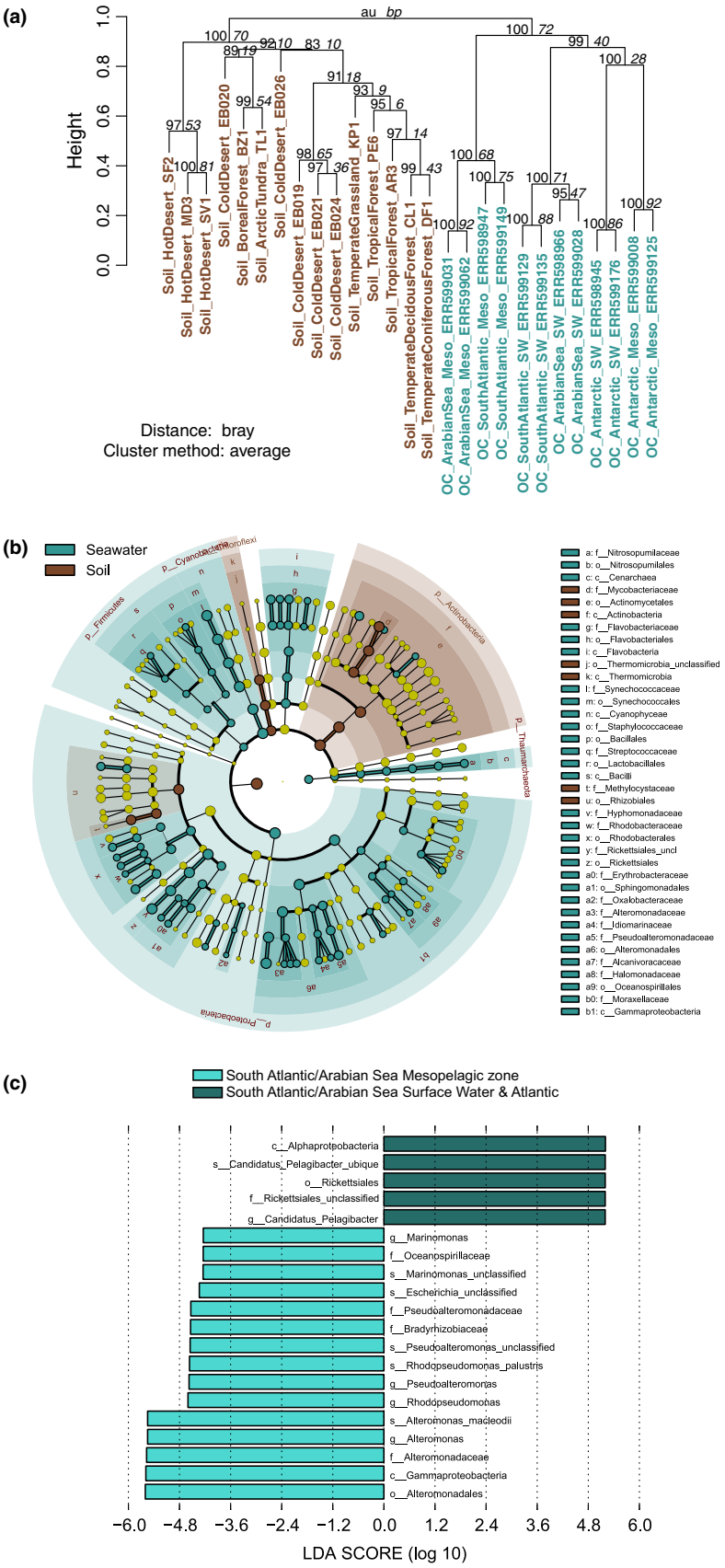

Fig. 3 Analyses of soil and seawater environmental samples. (A) Hierarchical clustering tree at the genus level. Soil samples in brown are identified by the label prefix 'Soil_', and seawater samples in blue by the prefix 'OC_'. (B) LEfSe cladogram of biomarkers for soil and seawater microbial profiles. (C) LEfSe barplot of biomarkers LDA effect in south Atlantic/Arabian Sea Mesopelagic zone vs. south Atlantic/Arabian Sea Surface Water and Atlantic samples. Taxonomic levels: 'k__', kingdom; 'p__', phylum; 'c__', class; 'f__', family; 'g__', genus; 's__', species.

2011; Bouslimani *et al.* 2015) and gut (Eckburg *et al.* 2005; Tremaroli & Bäckhed 2012) microenvironments (Fig. 4A, B). Exploiting the 689 HMP human-associated microbial profiles made available as part of the metaBIT package, PCoA revealed the gut samples to fall within the diversity of the human stool, and the skin samples within the diversity of the human nose and skin (Fig. 4C). We found more similarity in the taxonomic profiles within sample types (Skin, Gut) than between, with no taxon overlap being observed at the phylum level (Fig. 4A). The gut microbiota also appeared to be more diverse than the skin microbiota (Fig. 4B, Table S1, Supporting information and genus-level Shannon alpha diversity indices: 2.31–2.43 vs. 0.37–1.19), in line with earlier findings (Human Microbiome Project Consortium 2012a). Altogether, this demonstrates that metaBIT can successfully characterize patterns of microbial composition across present-day human microbiomes.

Ancient oral microbiomes

We next investigated whether metaBIT could profile ancient microbiomes, using nine shotgun HTS data sets generated from two ~1000-year-old dental calculus specimens (Warinner *et al.* 2014). These exhibited highly diverse microbial compositions (Fig. 5A,B, Table S1, Supporting information, and average species-level Shannon alpha diversity indices = 3.63 and 3.37), which were found to fall within the diversity of modern oral microbiomes (Fig. 5C), thereby confirming that ancient microbiomes can be reconstructed. Although the ancient dental calculus data sets yielded profiles similar to modern oral microbiomes, which are better described than those of environmental samples, we note that we identified organisms belonging to genera not yet fully described at the species level, such as, for example *Methanobrevibacter*, *Capnocytophaga*, *Aggregatibacter* and TM7 (Fig. 5B). Overall, we detected the 40 species identified in Warinner *et al.* (2014) as potential pathogens due to their association with periodontitis, oral infection or endocarditis in immune-depressed subjects (Fig. 5B). Importantly, when realigning all reads against the genome sequences of the most abundant species and analysing the alignments for the four species showing the highest mean depth-of-coverage (≥10.0X) with MAPDAMAGE2 (Jónsson *et al.* 2013), we found nucleotide misincorporation patterns typical of post-mortem DNA damage in A-tailed libraries (Seguin-Orlando *et al.* 2013) (Fig. S2, Supporting information). This suggests that microbial DNA originated from ancient microbes rather than modern contaminants. Overall, our results are in agreement with those reported in Warinner *et al.* (2014) and show the proficiency of metaBIT in recovering ancient microbial profiles from dental calculus.

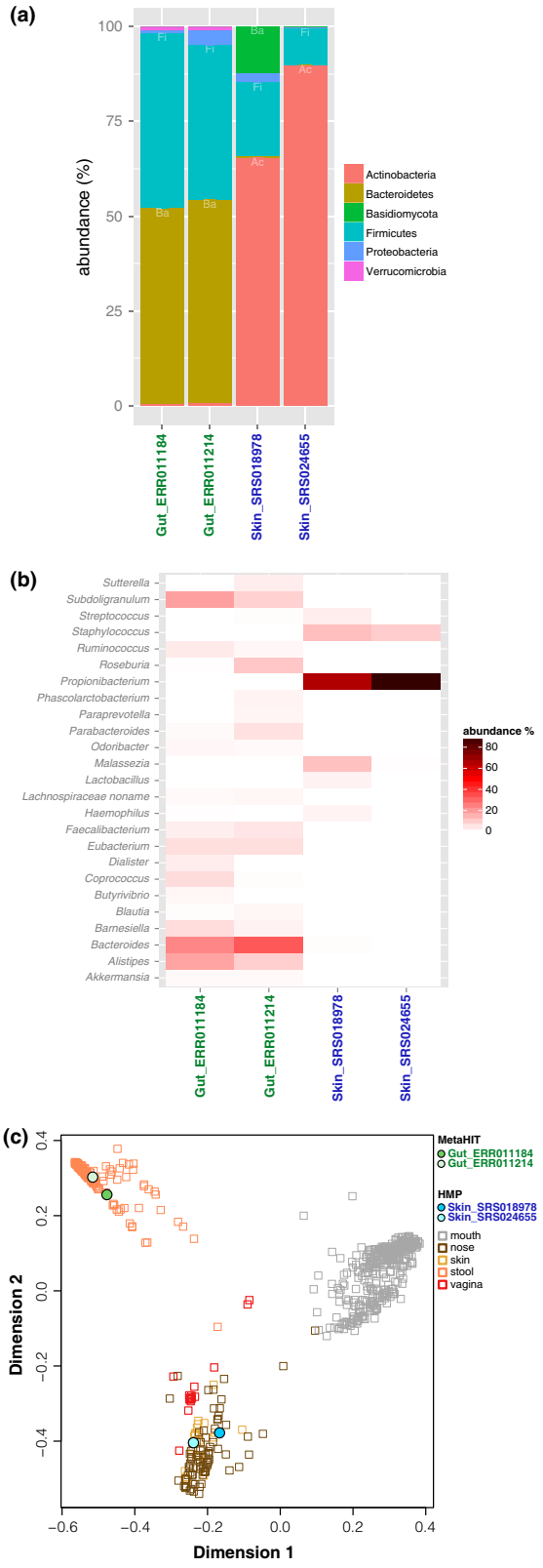

Fig. 4 Analyses of human clinical samples. (A) Barplot at the phylum level. (B) Heatmap at the genus level. (C) PCoA (genus level) of samples and HMP microbial profiles of the human body.

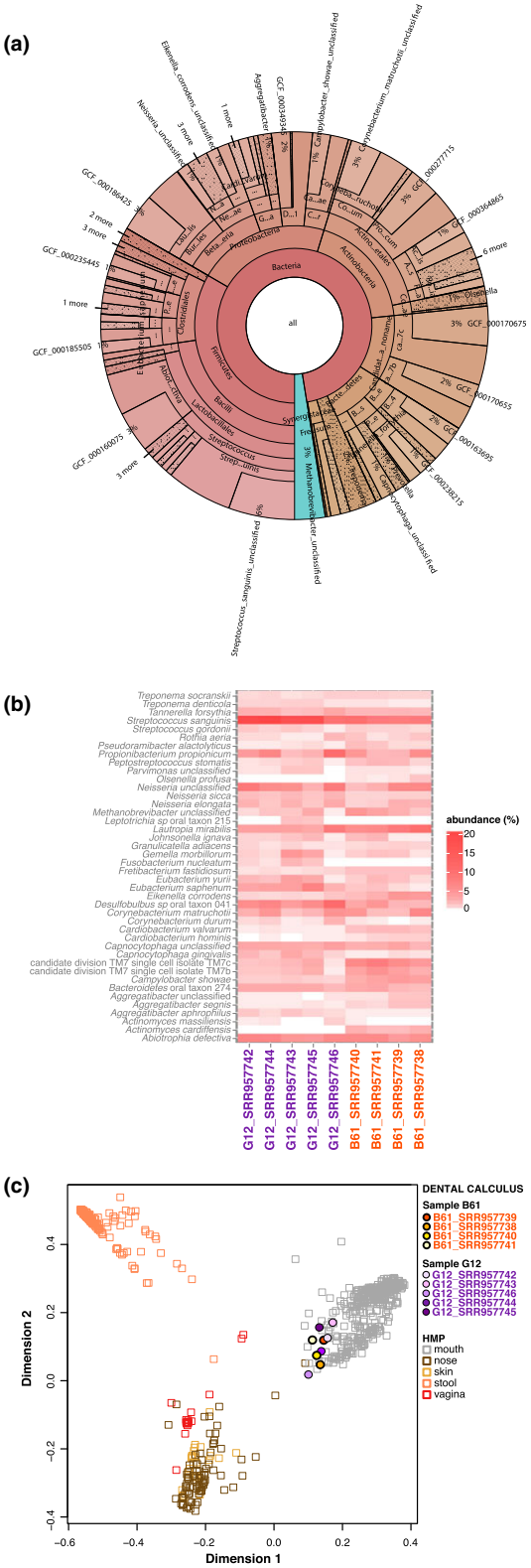

Fig. 5 Analyses of ancient dental calculus samples. (A) Krona graph for library G12_SRR957746. (B) Barplot at the species level. (C) PCoA (genus level) of samples and HMP microbial profiles of the human body.

Ancient human tuberculosis-infected samples

We finally applied metaBIT to 42 shotgun HTS data sets generated from the mummified tissues of 26 individuals buried in a Hungarian crypt in the 18th century CE (Kay *et al.* 2015). The tissues collected include samples from the chest, abdomen, pelvis, pleura, lungs and the ribs of individuals showing signs of tuberculosis infection. In six samples, no taxa could be identified (Table S1, Supporting information). Hierarchical clustering at the species level revealed three main clusters (Fig. 6A), with different tissues from the same individual belonging to different clusters. This shows that the microbial diversity is not structured by individuals and/or collection sites. In support for this and for a limited impact of human-driven sample contamination, the Vác profiles did not cluster with skin, stool and oral microbiomes (Fig. S3A, Supporting information). Samples 39.Pleura.-B_ERR650991 and 37.PAbdomenWall.A_ERR650989 showing high abundance of skin-specific *Propionibacterium* and *Corynebacterium*, respectively, stand as exceptions, suggesting skin bacterial contamination from handling of the specimens. The other samples fell within the diversity of soil samples, due to their similarly low detected microbial diversity, elevated *Mycobacterium* genus relative abundances and/or potential contamination from the depositional environment (Fig. S3A, Supporting information).

Using LEfSe, we found that taxa belonging to the *M. tuberculosis* complex, that is, *M. tuberculosis/bovis/africanum/canetti* (2.5–99.9%), represented biomarkers of cluster 1 (Fig. 6B,C). This is in line with the fact that all the samples for which Kay *et al.* (2015) could characterize *M. tuberculosis* draft genomes at 0.4-332.0X average coverage belonged to this cluster, except the abdomen samples from Body 25. In one sample from cluster 1 (121.Rib.B_ERR651009), we also identified the species *M. orygis* (33.8%), which is part of the *M. tuberculosis* complex and an etiological agent of tuberculosis in humans (van Ingen *et al.* 2012). Shotgun read alignment to the *M. tuberculosis* genome reference sequence (H37v, GCF_000195955.2) and to the genome sequence of *M. orygis* (GCF_12400015) followed by damage analyses using MAPDAMAGE2 (Jónsson *et al.* 2013) confirms the presence of ancient *Mycobacterium* sequences, as we observed nucleotide misincorporation patterns typical of post-mortem DNA damage in A-tailed libraries (Seguin-Orlando *et al.* 2013) (Fig. S4, Supporting information). Conversely, clusters 2 and 3 were characterized by environmental taxa (Fig. S3B, Supporting information). Altogether, our analyses demonstrate that metaBIT can identify the presence of *M. tuberculosis* and discriminate between the various members of *M. tuberculosis*. As pathogenic bacteria rarely leave morphological traces in

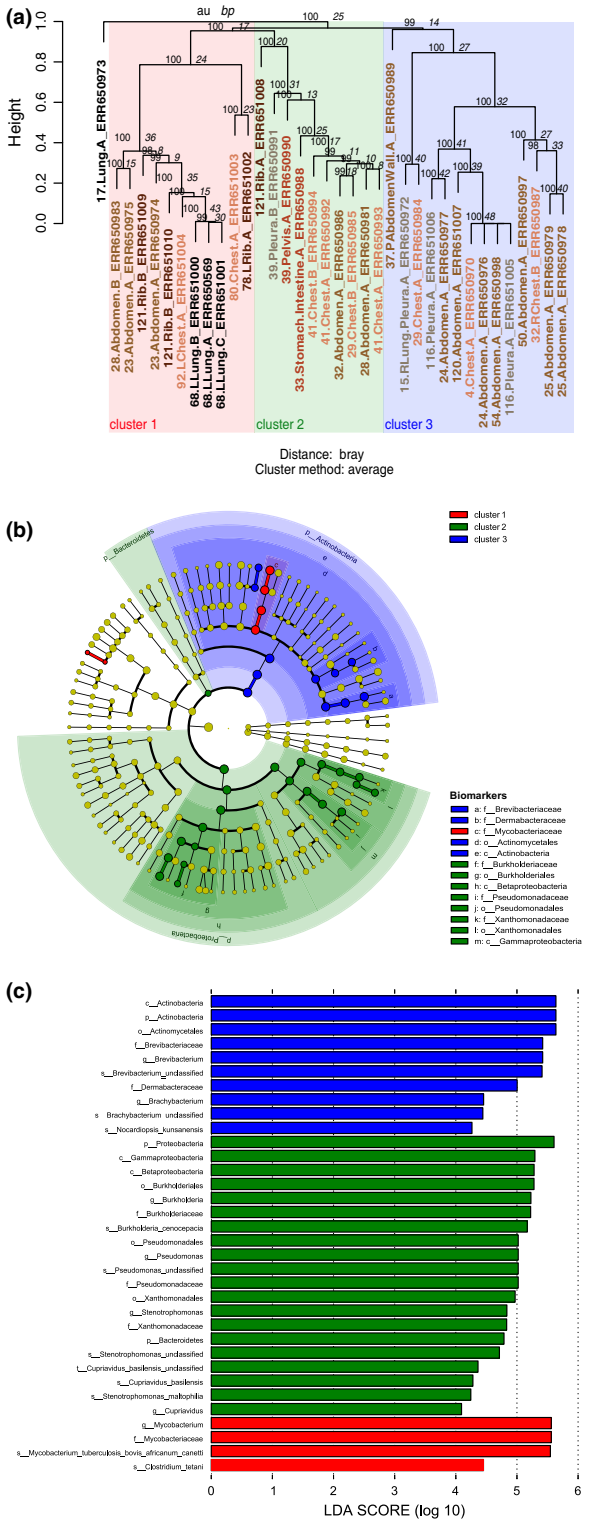

Fig. 6 Analyses of ancient tuberculosis-infected samples. (A) Hierarchical clustering tree at the species level. (B) LEfSe cladogram of biomarkers for clusters 1, 2 and 3. (C) LEfSe barplot of biomarkers LDA effect for clusters 1, 2 and 3. Taxonomic levels: 'k__', kingdom; 'p__', phylum; 'c__', class; 'f__', family; 'g__', genus; 's__', species.

the archaeological osseous material, metaBIT might prove instrumental for screening the etiological agents of past epidemics and complement current approaches based on target enrichment (Devault *et al.* 2014; Bos *et al.* 2015).

Discussion

In this study, we demonstrate the versatility of metaBIT for characterizing environmental, clinical and ancient microbiota from shotgun HTS data. Typical running times for the analyses presented involve the processing of ~1.1–1.8 million reads per minute (Table S2, Supporting information). We successfully retrieved expected patterns of microbial distribution in environmental samples representing different ecological niches, such as soil and seawater samples, and mesopelagic and surface seawaters. We also show that metaBIT can be effectively used to characterize microenvironments of the human host or to conduct a preliminary screen for pathogens, providing candidates for more targeted analyses, for example, whole-genome enrichment of particular pathogens.

We found metaBIT appropriate for a fast and automated analysis of shotgun HTS data generated from ancient specimens. Although archaeological hair is resistant to microbial attack (Gilbert *et al.* 2006) and generally shows high endogenous DNA content (Gilbert *et al.* 2007, 2008; Miller *et al.* 2008; Rasmussen *et al.* 2010, 2011), DNA extracts from ancient bones and teeth are generally dominated by environmental microbes. The metaBIT pipeline provides a fast and user-friendly approach to routinely retrieve valuable information about the microbial diversity present in such data sets at no additional experimental costs. This potentially allows detecting pathogen candidates, or characterizing preservation conditions and depositional environments, which could reveal microbial signatures for ancient DNA quality (Der Sarkissian *et al.* 2014) or past ecosystems (Willerslev *et al.* 2004). Conversely, when analysing dental calculus and coprolite samples, describing microbial communities is the primary aim and metaBIT provides an effective platform to recover first insights on the impact of dietary changes on past health status.

As metaBIT makes use of trimmed reads, it is directly compatible with other pipelines, such a PALEOMIX (Schubert *et al.* 2014a) for reconstructing and characterizing complete genomes. We advocate that metaBIT could also be advantageous as a quality control, for example, to monitor experimental errors, such as sample mislabelling when several libraries of the same extract are available, or to control for contamination from other samples during laboratory work or from human skin/saliva during sample manipulation (Der Sarkissian *et al.*

2015; Librado *et al.* 2015). We would like to stress that this is facilitated by the two soil and human microbiome profile comparative data sets made available as part of the metaBIT package.

We should, however, caution that the user must pay particular attention to the choice of parameters used at each step of the pipeline and interpret results accordingly. While METAPHLAN addresses some limitations of metagenomic analyses, the user should also be aware that biases in sampling, DNA extraction and sequencing can still impact the accuracy of the microbial profiles obtained with metaBIT. Importantly, the robustness and resolution of metaBIT profiles entirely relies on the sensitivity of METAPHLAN in taxonomic identification, which is mostly driven by the taxonomic representation in the reference database. This is a known current limitation of metagenomic analyses, as the vast majority of microbial taxa is presently unknown or uncharacterized at the complete genome level. Species/strain assignment is particularly sensitive to representativeness biases within reference databases and should be considered critically and/or independently replicated. The metaBIT pipeline might also show higher sensitivity for analysing clinical samples or screening for pathogens, given that human microbiota and pathogens are well characterized at the sequence level. With the ever-increasing number of complete genome sequences made publicly available and the curation of the METAPHLAN database, we are confident that metaBIT will allow a microbial profiling of ever-increasing quality.

Acknowledgements

We thank Dr Luca Ermini for fruitful discussions and Dr Mikkel Schubert for technical support. This work was supported by the Danish Council for Independent Research, Natural Sciences (FNU, 4002-00152B); the Danish National Research Foundation (DNRF94); a Marie-Curie Career Integration Grant (FP7 CIG-293845); the 'Chaires d'Attractivité 2014' IDEX, University of Toulouse, France.

References

Adler CJ, Dobney K, Weyrich LS *et al.* (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and industrial revolutions. *Nature Genetics*, **45**, 450–455.

Bos KI, Schuenemann VJ, Golding GB *et al.* (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*, **478**, 506–510.

Bos KI, Jäger G, Schuenemann VJ *et al.* (2015) Parallel detection of ancient pathogens via array-based DNA capture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **370**, 20130375.

Bouslimani A, Porto C, Rath CM *et al.* (2015) Molecular cartography of the human skin surface in 3D. *Proceedings of the National Academy of Sciences*, **112**, E2120–E2129.

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 326–349.

Brooks JP, Edwards DJ, Harwich MD *et al.* (2015) The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, **15**, 66.

Costello EK, Lauber CL, Hamady M *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, **52**, 87–94.

Der Sarkissian C, Ermini L, Jónsson H *et al.* (2014) Shotgun microbial profiling of fossil remains. *Molecular Ecology*, **23**, 1780–1798.

Der Sarkissian C, Ermini L, Schubert M *et al.* (2015) Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, **25**, 2577–2583.

Devault AM, McLoughlin K, Jaing C *et al.* (2014) Ancient pathogen DNA in archaeological samples detected with a microbial detection array. *Scientific Reports*, **4**, 4245.

Dinan TG, Borre YE, Cryan JF (2014) Genomics of schizophrenia: time to consider the gut microbiome? *Molecular Psychiatry*, **19**, 1252–1257.

Eckburg PB, Bik EM, Bernstein CN *et al.* (2005) Diversity of the human intestinal microbial flora. *Science (New York, NY)*, **308**, 1635–1638.

Ferrand J, Patron K, Legrand-Frossi C *et al.* (2014) Comparison of seven methods for extraction of bacterial DNA from fecal and cecal samples of mice. *Journal of Microbiological Methods*, **105**, 180–185.

Fierer N, Leff JW, Adams BJ *et al.* (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, **109**, 21390–21395.

Gilbert MTP, Menez L, Janaway RC *et al.* (2006) Resistance of degraded hair shafts to contaminant DNA. *Forensic Science International*, **156**, 208–212.

Gilbert MTP, Tomsho LP, Rendulic S *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science (New York, NY)*, **317**, 1927–1930.

Gilbert MTP, Drautz DI, Lesk AM *et al.* (2008) Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proceedings of the National Academy of Sciences*, **105**, 8327–8332.

Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology*, **12**, 69.

Grice EA, Segre JA (2011) The skin microbiome. *Nature Reviews Microbiology*, **9**, 244–253.

Hingamp P, Grimsley N, Acinas SG *et al.* (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *The ISME Journal*, **7**, 1678–1695.

Human Microbiome Project Consortium (2012a) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Human Microbiome Project Consortium (2012b) A framework for human microbiome research. *Nature*, **486**, 215–221.

van Ingen J, Rahim Z, Mulder A *et al.* (2012) Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. *Emerging Infectious Diseases*, **18**, 653–655.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics (Oxford, England)*, **29**, 1682–1684.

Kay GL, Sergeant MJ, Zhou Z *et al.* (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature Communications*, **6**, 6717.

Lai B, Wang F, Wang X, Duan L, Zhu H (2015) InteMAP: integrated metagenomic assembly pipeline for NGS short reads. *BMC Bioinformatics*, **16**, 244.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111–5120.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Librado P, Der Sarkissian C, Ermini L *et al.* (2015) Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E6889–E6897.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Miller W, Drautz DI, Ratan A *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–390.

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Orlando L, Ginolhac A, Zhang G *et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499**, 74–78.

Orlando L, Gilbert MTP, Willerslev E (2015) Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, **16**, 395–408.

Pedersen MW, Overballe-Petersen S, Ermini L *et al.* (2015) Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **370**, 20130383.

Qin J, Li R, Raes J *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Qin J, Li Y, Cai Z *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

Rasmussen M, Li Y, Lindgreen S *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757–762.

Rasmussen M, Guo X, Wang Y *et al.* (2011) An aboriginal Australian genome reveals separate human dispersals into Asia. *Science (New York, NY)*, **334**, 94–98.

Schubert M, Ermini L, Der Sarkissian C *et al.* (2014a) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, **9**, 1056–1082.

Schubert M, Jónsson H, Chang D *et al.* (2014b) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, **111**, E5661–E5669.

Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, **9**, 88.

Segata N, Izard J, Waldron L *et al.* (2011) Metagenomic biomarker discovery and explanation. *Genome Biology*, **12**, R60.

Segata N, Waldron L, Ballarini A *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**, 811–814.

Segata N, Boernigen D, Tickle TL *et al.* (2013) Computational meta'omics for microbial community studies. *Molecular Systems Biology*, **9**, 666.

Seguin-Orlando A, Schubert M, Clary J *et al.* (2013) Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLoS One*, **8**, e78575.

Seguin-Orlando A, Gamba C, Der Sarkissian C *et al.* (2015) Pros and cons of methylation-based enrichment methods for ancient DNA. *Scientific Reports*, **5**, 11826.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.

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.

Sunagawa S, Coelho LP, Chaffron S *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

Tremaroli V, Bäckhed F (2012) Functional interactions between the gut microbiota and host metabolism. *Nature*, **489**, 242–249.

Truong DT, Franzosa EA, Tickle TL *et al.* (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, **12**, 902–903.

Turnbaugh PJ, Hamady M, Yatsunenko T *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Warinner C, Rodrigues JFM, Vyas R *et al.* (2014) Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics*, **46**, 336–344.

Willerslev E, Hansen AJ, Rønn R *et al.* (2004) Long-term persistence of bacterial DNA. *Current Biology: CB*, **14**, R9–R10.

Ziesemer KA, Mann AE, Sankaranarayanan K *et al.* (2015) Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Scientific Reports*, **5**, 16498.

L.O. designed the study. G.L. and K.H. wrote the metaBIT pipeline. G.L. and C.D. analysed data. C.D. and L.O. wrote the manuscript. L.O. provided analytical tools.

Data accessibility

The metaBIT pipeline, associated documentation, example data sets and a tutorial are available at https://bitbucket.org/Glouvel/metabit. DNA sequences analysed in this study: European Nucleotide Archive accessions ERR011184, ERR011214, ERR598945, ERR598947, ERR598966, ERR599008, ERR599028, ERR599031, ERR599062, ERR599125, ERR599129, ERR599135, ERR599149, ERR599176, and ERR650569-ERR651010; Sequence Read Archive accessions SRS018978, SRS024655, and SRR957738–SRR957746.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Main functions and set-up of the metaBIT makefile.

Fig. S2. Nucleotide misincorporation patterns of ancient dental calculus samples.

Fig. S3. PCoA (genus level) for ancient tuberculosis-infected human samples.

Fig. S4. Nucleotide misincorporation patterns of ancient tuberculosis-infected samples.

Table S1. Sample and microbial profile information.

Table S2. Running time (real time) to compute presented analyses using eight cores (AMD Opteron 6276, 2.3 GHz processors).

Table S3. METAPHLAN version 2 database inconsistencies detected in the analyses of the Tara Oceans Expedition dataset.