Bioinformatics, 32(15), 2016, 2272–2280
doi: 10.1093/bioinformatics/btw150
Advance Access Publication Date: 24 March 2016
Original Paper

Sequence analysis

DUDes: a top-down taxonomic profiler for metagenomics

Vitor C. Piro[1,2], Martin S. Lindner[1,3] and Bernhard Y. Renard[1,*]

[1]Research Group Bioinformatics (NG4), Robert Koch Institute, Nordufer 20, Berlin 13353, Germany, [2]CAPES Foundation, Ministry of Education of Brazil, Brasília - DF, 70040-020 Brazil and [3]4-Antibody AG, Hochberger Strasse 60C, Basel 4057, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 24, 2015; revised on February 24, 2016; accepted on March 11, 2016

Abstract

Motivation: Species identification and quantification are common tasks in metagenomics and pathogen detection studies. The most recent techniques are built on mapping the sequenced reads against a reference database (e.g. whole genomes, marker genes, proteins) followed by application-dependent analysis steps. Although these methods have been proven to be useful in many scenarios, there is still room for improvement in species and strain level detection, mainly for low abundant organisms.

Results: We propose a new method: DUDes, a reference-based taxonomic profiler that introduces a novel top-down approach to analyze metagenomic Next-generation sequencing (NGS) samples. Rather than predicting an organism presence in the sample based only on relative abundances, DUDes first identifies possible candidates by comparing the strength of the read mapping in each node of the taxonomic tree in an iterative manner. Instead of using the lowest common ancestor we propose a new approach: the deepest uncommon descendent. We showed in experiments that DUDes works for single and multiple organisms and can identify low abundant taxonomic groups with high precision.

Availability and Implementation: DUDes is open source and it is available at http://sf.net/p/dudes

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: renardB@rki.de

1 Introduction

The fast increase of complete genome sequences available on public databases has allowed better predictions of the microbial content from sequenced environmental and clinical samples. In addition, the fast evolution and decreasing costs of high-throughput sequencing as well as the development of fast and precise bioinformatics tools to handle huge amounts of data (e.g. read mappers, assemblers) are enabling the integration of automated computational methods in clinical practice (Köser *et al.*, 2012; Pallen, 2014).

Taxonomic or community profiling are common terms to define the process of identification of organisms and their quantification given a targeted or whole shotgun metagenomic sequencing (Mande *et al.*, 2012). Characterizing the taxonomic diversity is an initial and fundamental step to understand complex biological processes, diversity and functions of a microbial community and it can be applied for pathogen detection studies. Several tools have been recently developed for this characterization, with different approaches and applications (Lindgreen *et al.*, 2016). Considering only the reference-based methods, that is, methods that use reference sequences to guide their analysis and classify sequences in the sample, community profiling is categorized in two sub-groups: composition and similarity-based. Composition-based methods extract information from the composition of sequences from both sample and database (e.g. GC content, codon usage) and search for similarities between reads and references. Similarity or homology-based techniques are built on mapping or aligning the sequenced reads against reference databases and performing further analysis. Despite significant performance improvements in the last years, similarity analysis is still computationally challenging due to the extremely high throughput of modern sequencing machines and the accelerated growth of available genomic sequences (Benson *et al.*, 2013). Some tools address this problem by reducing the database space, selecting only marker genes or a specific subset of sequences (Freitas *et al.*, 2015; Segata *et al.*, 2012). This simplification can speed up analysis but at the same time reduces the complexity and diversification of the references, decreasing precision for more specific identifications. Other methods use custom databases of partial or whole genome sequences (Francis *et al.*, 2013; Huson *et al.*, 2007; Lindner and Renard, 2015). Additionally, very efficient k-mer based read binning tools (Ounit *et al.*, 2015; Wood and Salzberg, 2014) can also be used in some extension for profiling communities, by selecting the targets with more associated sequences.

Despite the remarkable recent advances in this area and a vast number of tools available, there is still a number of challenges from sequencing methods to bioinformatics tools to integrate computational and automated approaches into molecular and metagenomics diagnostics (Fricke and Rasko, 2014; Klymiuk *et al.*, 2014). Specifically for community profiling, there is still room for improvement in species and strain level detection, which can have very similar genomic content and at the same time low abundances. The high discordant number of available sequenced genome sequences among several taxonomic groups (TGs) poses another common problem, where some organisms (e.g. pathogens, model organisms) are more studied and therefore overrepresented in the public sequence databases.

Aiming to solve those limitations, we propose a new method: DUDes, a reference-based taxonomic profiler, which introduces a novel top-down approach to analyze metagenomics NGS samples. Rather than predicting an organism presence in the sample based only on relative abundances, DUDes first identifies possible candidates by comparing the strength of the read mapping in each node of the taxonomic tree in an iterative manner. Instead of using the lowest common ancestor (LCA) (Huson *et al.*, 2007), a commonly used bottom-up approach to solve ambiguities in identifications, we propose a new approach: the deepest uncommon descendent (DUD). While the LCA method solves ambiguous identifications by going back in the taxonomic tree to the LCA, the DUD approach starts at the root node and tries to go for deeper taxonomic levels, even when ambiguities are found. That way it is possible to have less conservative identifications in higher taxonomic levels. Besides, when the provided data does not allow a specific identification on higher levels, our method identifies a small set of probable candidates among dozens of possibilities (e.g. instead of stopping at species identification, DUDes will provide 5 highly likely strains out of 150). Permutation tests are performed to estimate *P*-values between TGs and to identify the presence of them on each level. We show in experiments and comparisons with state-of-the-art tools that DUDes works well for single and multiple organism detection, can handle unequally represented references in the database and identifies low abundant TGs with high precision. DUDes is open source and it is available at http://sf.net/p/dudes

2 Methods

The DUDes workflow requires three main inputs: a set of reads, references sequences and a taxonomic tree structure (Fig. 1). By mapping the reads against the reference sequences, a SAM file (Li *et al.*, 2009) is created. The linkage between the reference sequences and the taxonomic information is performed by DUDesDB and stored in a database file. Both files serve as input to DUDes profiler which performs an iterative top-down analysis on the taxonomic tree structure.

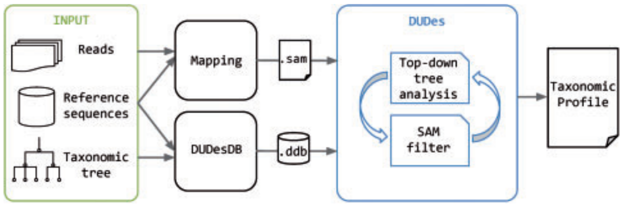

Fig. 1. DUDes workflow. A set of reads, references sequences and taxonomic tree structure are the pre-requisites for the complete workflow. A SAM file from mapping the reads against references sequences and a database file generated by DUDesDB are the input files required for DUDes profiler, which will perform the top-down tree analysis and generate the taxonomic profile

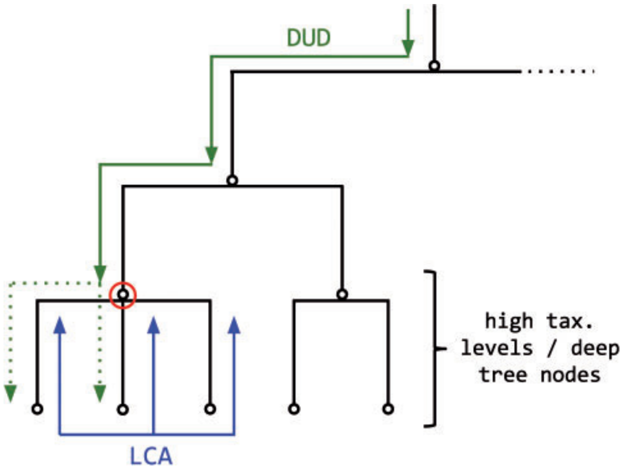

Fig. 2. Comparison between DUD and LCA. DUD is a top-down approach that starts from low taxonomic levels. LCA is a bottom-up method that solves ambiguities from high taxonomic levels. The node marked with a circle is considered the LCA for its children nodes. Dotted arrows represents how DUD can be more specific in higher taxonomic levels pointing to candidates nodes while LCA is more conservative and goes back to the LCA

The first step of the DUDes algorithm is to assign a set of reference sequences and read matches for each node of the taxonomic tree (e.g. the root node contains all sequences and matches, while the *Escherichia* node (genus) will contain only the sequences corresponding to the organisms that belong to this classification and their respective matches). The general idea is to start at the root node and go deeper into the tree (Fig. 2), evaluating all taxonomic levels and identifying nodes with significant matches that are the ones that will be considered present in the sample. The presence of a node in a certain taxonomic level is defined by the following steps (Fig. 3):

2.1 Bin generation

First, we create a set of bins for each node of the tree. Bins are subsequences from the reference sequences assigned to the node (Fig. 3b). They are non-overlapping and equal-sized. The bin size is fixed for all nodes and it is defined by default as the 25th percentile of the sequence lengths of the whole reference database. The bin size should be a balance point between the number of small sequences in the database and speed requirements: the smaller the bin size, the more calculations are needed. The larger the bin size, the bigger the chance that higher nodes lack of full bins. In our analysis the 25th percentile provides a good trade-off between those, ensuring that the

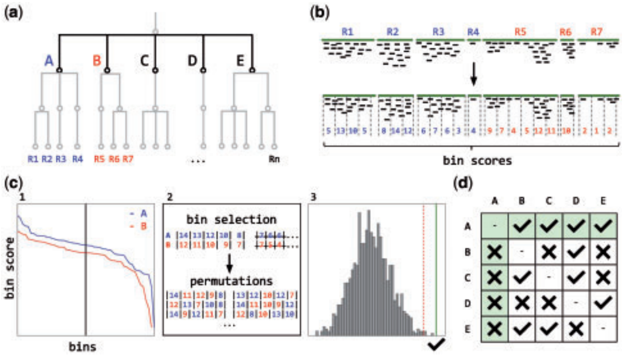

Fig. 3. Steps for identifications on a taxonomic level. **(a)** Taxonomic tree structure with the tested level in black and the reference sequences of each node (R1-Rn). Nodes A and B are the first ones to be compared against each other **(b)** Bin generation: reference sequences (green) and their respective read matches are subdivided into bins. For each bin, a match-based score is calculated **(c)** *P*-value estimation: 1—Bin scores of the node A (blue, references R1-R4) are compared against bin scores of node B (red, references R5-R7). The black line represents the cutoff value based on the top 50% of bin scores from node A. 2—Only the best bins are selected based on the cutoff. Permutations are then performed among the selected bins. 3—The distribution is generated by the randomly permuted difference of means between bin scores. The red dotted line represents the critical value and the green the observed difference of means. In this example node A is therefore significant against node B (*P*-value $\leq$ critical value). **(d)** Identification: steps b and c are then repeated for all pair of nodes. In this example, node A is identified because it was the only significant node against all the others (rows) and any other node could be significant against it (columns)

majority of reference sequences will have at least one full bin. Each one of these sub-regions will have a bin score based on read matches, defined as the sum of all match scores (*ms*). The match score is defined for each match *m* as:

$$\mathrm{ms_m} = l - e \tag{1}$$

where *l* is the length and *e* the edit distance of the match (minimum number of edit operations—insertions, deletions and substitutions of letters—transforming one sequence into another (Levenshtein, 1966)). Bin scores do not only take the number of matches to the bins into account, but also the number of mismatches and indels in those sub-regions.

2.2 *P*-value estimation

DUDes performs a pairwise comparison between all nodes on a taxonomic level, estimating a *P*-value for each pair with permutation tests. The permutations occur between the nodes' bin scores. Only bins with scores higher than zero are considered. Additionally, only the best bins are permuted: a cutoff is chosen, based on the number of bins representing the top 50% scores of the main node (Fig. 3c-1). This cutoff is useful in order not to prioritize nodes with larger or more references sequences, meaning that the comparison of a certain node X against node Y can have a different cutoff value from the comparison of the same node Y against X. For example: consider a node X with 100 bins and a node Y with only 70 bins. When comparing X against Y, only the first 50 bins from node X are going to be compared against the first 50 bins of node Y. When comparing Y against X, only 35 bins from both nodes are going to be considered (Supplementary Figure S1). When one of the nodes does not have enough bins to be compared, the cutoff will be reset to the total number of bins of the node with fewer bins (Supplementary Figure S2). The cutoff is also useful to remove poorly mapped bins with low scores and to normalize the number of bins between the two compared nodes, allowing a fair comparison between them. A limit of five bins is required for each node to allow a significant permutation. Permutations between the selected bin scores are performed 1000 times by default. The values are randomly shuffled and separated in two groups based on the cutoff value. The random difference of means between the groups is calculated, generating a distribution like the one shown in Figure 3c-3. A one-sided *P*-value is then estimated based on the observed difference of means between their actual bin scores (Supplementary Information—*P*-value estimation).

The estimated *P*-value is considered significant if it is lower than a certain critical value. Because many hypotheses are being tested, multiple testing correction is necessary to control the type I error, that is, the risk of falsely rejecting a hypothesis that is true (Goeman and Solari, 2014). Here we applied two methods: Bonferroni correction locally (taxonomic level) and the Meinshausen procedure globally (tree level) (Meinshausen, 2008). Two methods were applied together because multiple tests occur in two ways: several taxonomic levels on the tree and several nodes for each taxonomic level are tested. The Meinshausen procedure was chosen for being a hierarchical approach that can achieve larger power for coarser resolution levels. It was applied to control the multiple testing error generated by several comparisons on each taxonomic level of the tree. Bonferroni correction was applied locally to correct for the multiple tests performed among the nodes in a taxonomic level. Although the Bonferroni method is highly conservative in general, this is not critical in our application since the number of nodes on a taxonomic level is usually fairly small. The critical value *cv* for each node *n* is calculated as:

$$\mathrm{cv_n} = \frac{\alpha}{N-1}\frac{L_\mathrm{n}}{L} \tag{2}$$

where α is the significance level threshold (default 0.05) and *N* is the total number of nodes in the tested taxonomic level (minus itself), comprising the Bonferroni correction. Additionally, *L* is the total number of leaves of the tree and L_n the total number of leaves below node *n*. They stand for the Meinshausen procedure correction. All nodes are tested against each other at the same level and, if the *P*-value of the comparison is below the critical value, the comparison is set as significant as pictured in the table in Figure 3d.

2.3 Identification

After all nodes within a taxonomic level have been compared against each other, they are evaluated using two criteria: how many times they are not significant and how many times other nodes are significant against them (columns and rows in Fig. 3d, respectively). The number of occurrences of those two metric are summed for each node and DUDes selects the one with the minimum value as identified, therefore present on the sample. In a metagenomic dataset it is possible that more than one node is identified at once, when their bins have similar abundances. In this case, two or more nodes with sufficient number of bins and matches that could not be significant against each other are going to be identified concurrently. That means that more than one node is going to be considered present in the current taxonomic level, leading to more than one path on the following tree analysis.

The three-step-algorithm (consisting of bin generation, *P*-value estimation and identification) continues to the next taxonomic level only for the children of the identified nodes. The process is iterated until it reaches the leaf nodes of the tree (here considered the deepest possible uncommon descendent—at species level by default) or until no more identifications are possible, due to lack of minimum matches or bins support. Here we can point out an advantage of the DUD method over the LCA. When the LCA is applied, the results tend to be very conservative because it solves ambiguous identifications by going back one taxonomic level to the LCA. Instead, the DUD approach will always try to go for a deeper taxonomic level, even when ambiguities are found (Fig. 2). That way it is possible to have identifications in higher taxonomic levels. Besides, when the provided data does not allow a specific identification on higher levels, it is still possible to propose a set of likely candidates based on the concurrent identification, being more specific than going back in the taxonomic tree.

At the end of the tree iteration, one or more paths on the tree and their leaf nodes are identified as candidate TGs to be present on the sample. Because metagenomic samples can contain hundreds to thousands of organisms (Handelsmanl *et al.*, 1998), a filtering step is performed to remove identified matches and allow more iterations. DUDes perform this step by filtering out the direct matches on the identified candidates' references sequences. Furthermore, all matches from the reads that had at least one direct match are analyzed. If those read matches have a *ms* lower than the direct match, they are considered indirect matches, and are filtered out as well. With this new set of matches, a new iteration is started from the root node. Several iterations are performed until the number of matches is below a certain threshold or until all matches were filtered.

At the end, a relative abundance value is calculated for the final candidates. These are based on the direct matches of the identified leaf nodes and normalized by the length of their respective reference sequences. Each identified leaf node *n* has an abundance ab calculated as:

$$\mathrm{ab_n} = \sum_{i=1}^{r} \frac{\sum_{j=1}^{t} \mathrm{ms}}{l} \tag{3}$$

where *r* is the number of references sequences belonging to the node *n*, *t* are the matches belonging to the reference *i*, *ms* is the match score and *l* is the length of the reference *i*. The abundance of the parent nodes are based on the cumulative sum of their children nodes' abundance.

DUDes outputs a file with a set of final candidates in the BioBoxes Profiling Output Format v0.9.3 (https://github.com/bioboxes/rfc/blob/master/data-format/profiling.mkd). When strain identification is selected, DUDes outputs an additional file with all identified strains and their relative abundances.

2.4 Strain identification

Optionally, DUDes will try to extend the species identification and provide a set of probable strains present in the sample. The process of strain identification works identically as the three-step-algorithm but starting the analysis from each one of the identified species nodes. Sequences among strains usually have high similarity in their composition. This makes the identification process more challenging. For that reason we implemented a post-filtering process to better select a candidate strain. Given a set of identified strains by the three-step method, we choose one representative candidate, which has the maximum summed value of *ms* in this set. Alternatively we provide a second output, reporting all other strains identified and their relative abundance.

2.5 DUDesDB

DUDesDB pre-processes the taxonomic tree structure and the reference sequences, generating a database file to be used by DUDes profiler. The current version of DUDesDB supports the NCBI taxonomic tree (NCBI, 2015) and uses the GI or accession.version identifier to make the link between reference sequences and tree nodes. Because the *strain* level is not directly defined in the NCBI taxonomic tree structure, we considered any unclassified node with the tag *no rank* after the species level as a strain node.

2.6 Mapping

DUDes can handle multiple matches and account for the mapping quality with match scores, improving its identification capabilities. By default, the number of allowed matches should be as high as possible, allowing all matches when feasible. Since that can be computational impracticable, we used a default value of 60 matches for each read. Other mapping parameters can be found in the Supplementary information—Tools and parameters.

2.7 Experiments

DUDes evaluation was performed in four distinct datasets: two synthetic communities and two real metagenomic samples.

First we analyzed synthetic metagenomic data with available ground truth to evaluate how precise is our identification method. We chose a common set for metagenomics evaluations—the Human Microbiome Project (HMP) mock community (Turnbaugh *et al.*, 2007), an *in vitro* synthetic mixture of 22 organisms (20 Bacteria, 1 Archaea and 1 Eukaryote) that mimics errors and organism abundances from real metagenomics samples. Only Bacteria and Archaea were considered in this evaluation. Further, this set was also divided in sub-sets of different percentages of reads and compared against other metagenomics analysis tools: kraken (Wood and Salzberg, 2014), GOTTCHA (Freitas *et al.*, 2015) and MetaPhlAn2 (Truong *et al.*, 2015). They were chosen for having rather different approaches to solve the taxonomic profiling problem and for having good results in recent metagenomics studies (Lindgreen *et al.*, 2016). Kraken is a read binning tool that uses a k-mer approach to classify each read in a given sample with focus in high performance. GOTTCHA is a taxonomic profiler that uses non-redundant signature databases and aims for lower false discovery rates. MetaPhlAn2 relies on a curated database of approximately 1 Million unique clade-specific marker genes for profiling metagenomic samples. A second synthetic community consisted of 64 laboratory-mixed microbial genomic DNAs was also evaluated (Shakya *et al.*, 2013). This community made of organisms of known sequences has a very broad diversity among bacteria and archaea and a wide range of genetic variation at different taxonomic levels. At the same time, this dataset provides a large number of sequenced reads (∼110 M), allowing a more realistic performance evaluation.

We also applied DUDes to real metagenomic samples of gut microbiomes from the outbreak of Shiga-toxigenic *Escherichia coli* (STEC) in Germany in 2011 (Loman *et al.*, 2013). With this dataset we evaluated how well DUDes performs in a real scenario to profile a pathogenic sample, and compared the results with the previously known experiments. Furthermore, we evaluated this set based on previous known information (e.g. lab experiments, other tools based on the LCA approach) performing a more specific profiling. Lastly, we profile a marine dataset from Tara Oceans (Sunagawa *et al.*, 2015) with Bacteria, Archaea, Virus and Eukaryotes present on the sample, showing the versatility of the tool. Datasets' details are shown in Supplementary Table S1.

The HMP and STEC samples were pre-processed with the digital normalization algorithm (Brown *et al.*, 2012) for decreasing sampling variation and for error correction. In both analysis, the reference database for DUDes and kraken was generated with the set of complete genomes sequences (Bacteria and Archaea) together with the taxonomic tree structure, both from NCBI (NCBI, 2015) from March 26, 2015. The 64-organim set was used without any filter. For this set we used the above database with the addition of four non-complete genome sequences [taxid: 901, 52598, 314267, 304736] to have all species in the sample available. For the Tara dataset we made a custom database, containing only expected marine organisms. Bacterial, Archaeal and Viral taxons were obtained from the references sequences used in the Tara Oceans Project (Sunagawa *et al.*, 2015) and the Eukaryotic set of taxons was obtained from the MMETSP (Keeling *et al.*, 2014). All NCBI refseq sequences relative to those taxons were collected to generate the database (from January 31, 2016). For MetaPhlAn2 and GOTTCHA (all sets) we used their provided database, v20 and v20150825, respectively. Bowtie2 (Langmead and Salzberg, 2012) was used for read mapping. Parameters and usage details of each tool can be found in the Supplementary Materials—Tools and parameters.

We evaluated the output from each tool based on a binary classification of the sorted taxonomic profile. The binary classification is valid for TGs of a certain taxonomic level. True positives are all TGs present in the sample and correctly identified, false positives are the identified TGs known to not be present in the sample, false negatives are the TGs that are present but could not be identified and true negatives are all TGs on the database not identified and known not to be present in the sample.

3 Results

3.1 HMP mock community

We first assessed DUDes' taxonomic profiling capabilities with the set of Illumina reads from the HMP staggered mock community. Digital normalization was applied to correct errors, reduce data size and decrease sample variation. The normalization reduced around 18% from the complete set of 7.932.819 reads. Then, we mapped this normalized set against the reference database using Bowtie2 and the resulting mapping file was used in DUDes to profile the mock community sample.

The performance in terms of sensitivity and specificity for the identifications at each taxonomic level is shown in Table 1. DUDes successfully identified 20 of the 21 known species present in the sample (Bacteria and Archaea). *Actinomyces odontolyticus* does not have reference sequences in our database, therefore it could not be identified directly as species. However, at lower levels—Actinomycetales (order), Actinobacteria (class), Actinobacteria (phylum), Bacteria (superkingdom)—all present TGs could be identified, reaching a true positive rate of 1. Even in deeper levels—family, genus and species—DUDes achieved a sensitivity level of ~0.95. Furthermore, high specificity values show that DUDes can precisely select organisms present in a given metagenomic sample even with a large number of references sequences in the database.

We further analyze the same dataset, comparing the results against the established metagenomics tools kraken, GOTTCHA and MetaPhlAn2—at species level. In addition, we evaluated random subsets of the normalized set of reads, in a range from 1 to 100%. The sub-set analysis allowed us to show how the tools perform over a wide range of coverages: from the set where all data is available and each organism is well represented to a scenario where very few reads are

Table 1. DUDes' sensitivity and specificity values of the HMP mock community evaluation

	S.kingdom	Phylum	Class	Order	Family	Genus	Species
Sensitivity	1	1	1	1	0.944	0.944	0.952
Specificity	1	1	1	0.973	0.987	0.991	0.994

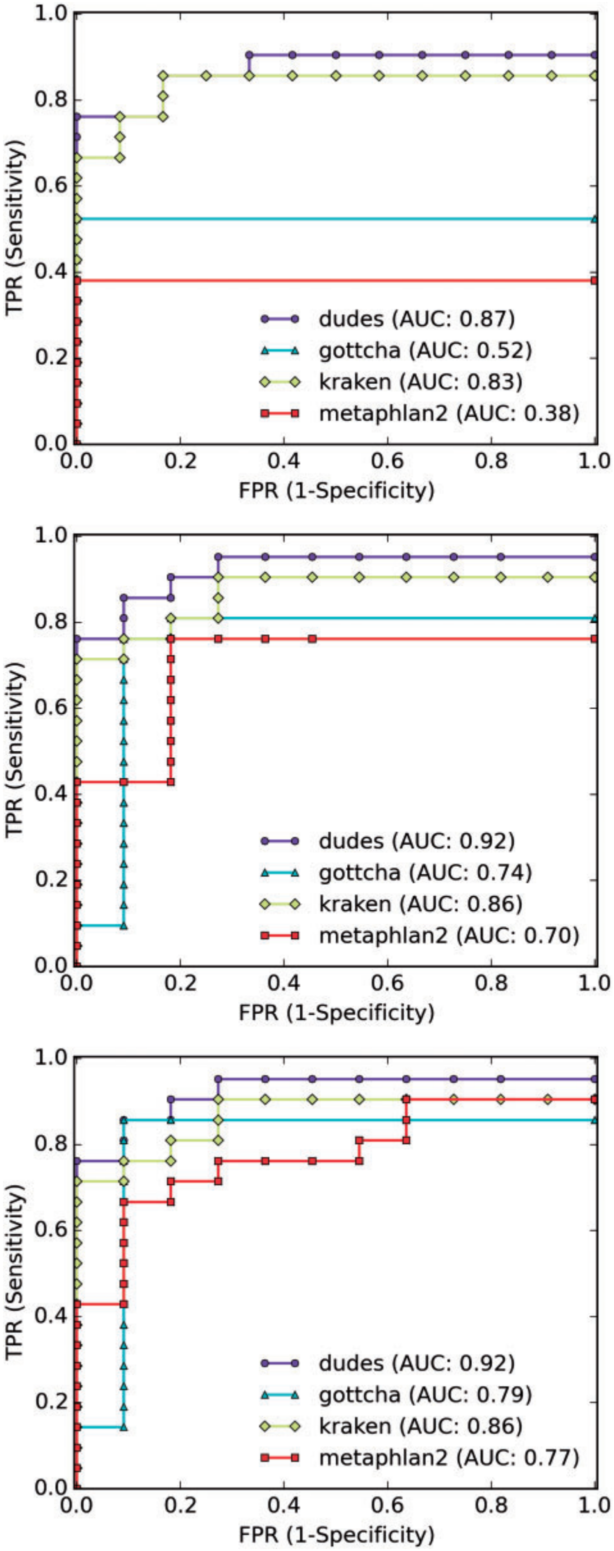

Fig. 4. ROC curves comparing results based on three sub-sets (1, 50 and 100%) of the normalized HMP Illumina set of reads

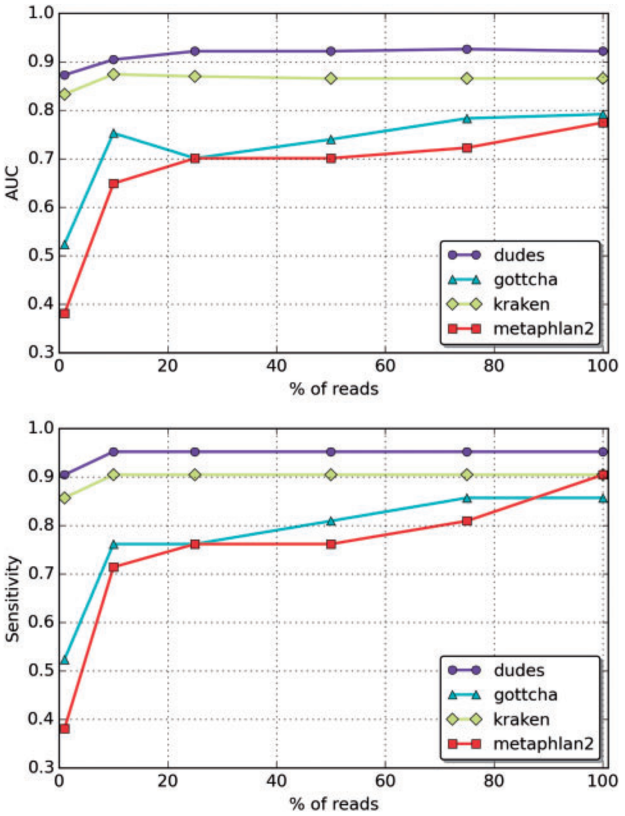

Fig. 5. AUC and Sensitivity values for the HMP sets. 6 data points were evaluated for each tool (1, 10, 25, 50, 75 and 100%) representing the percentage of reads in each sub-set

Table 2. Running time of the evaluated tools ([hh:]mm:ss)

Sub-set	DUDes	GOTTCHA	kraken + kraken_report	MetaPhlAn2
HMP 1%	01:29	02:30	06:42 + 00:26	01:49
HMP 10%	04:13	03:13	09:08 + 00:28	02:00
HMP 25%	08:06	04:28	09:53 + 00:29	02:04
HMP 50%	15:38	06:24	10:50 + 00:35	02:16
HMP 75%	23:42	08:46	11:56 + 00:39	02:47
HMP 100%	32:35	15:37	27:13 + 00:47	03:03
64-organism	02:46:14	05:09:50	01:05:33 + 05:05	39:52

Computer specifications: AMD Opteron(tm) Processor 6174 2.2–3.2 GHz, 48 cores, 256 GB RAM.

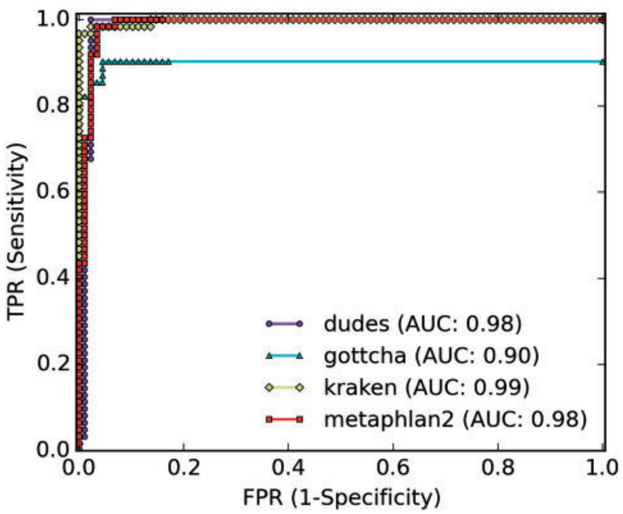

Fig. 6. ROC curves comparing results of the 64-organism Archaea and Bacteria synthetic community

present for some low abundant organisms. All results were based on the ordered output provided by each tool. Kraken is a read binning tool and outputs an unsorted report with all identified organisms. From this output we selected only species level identifications and sorted the candidates by the percentage of reads assigned. We limited the output from all tools by the first 30 entries when necessary.

Figure 4 shows ROC curves comparing the output from each tool on three sub-sets: 1, 50 and 100% (all sub-sets in the Supplementary Figure S3). DUDes and kraken had stable and similar performances in all sets, with DUDes achieving the best AUC values. When only 1% of the reads was used, MetaPhlAn2 and GOTTCHA did not achieve good results. That can be explained by the fact that both tools use specific gene/signature databases, decreasing the chance of finding matches when the number of reads is very low. Both DUDes and kraken use the whole genome sequence as database and therefore achieved good predictions, even at extremely low coverages. DUDes could successfully identify *Bacteroides vulgatus* at species level although there were only 14 matches on the references. Since kraken is not a taxonomic profiler, it outputs each organism that has at least one read, resulting in a long tail of false positives. DUDes produced fewer false positives. With more reads, all tools improved their results (Fig. 5), with DUDes being slightly superior, followed by kraken and GOTTCHA. Both DUDes and kraken identified 20 of 21 organisms present in the sample using only 10% of the reads, keeping the results stable for the next sub-sets. GOTTCHA needed at least 50% of the data to reach the same level of identification. MetaPhlAn2 performed poorly in this scenario, even with the whole set of normalized reads, but it had the best overall running time ranging from 1 min 49 s with the 1% dataset to 3 min 03 s for the complete dataset, disregarding database generation (Table 2). DUDes had a very similar time for the smaller dataset but an increase in running time for bigger datasets (15 min 38 s for 50% and 32 min 35 s for 100%), since it had to deal with larger SAM files (Supplementary Table S2).

3.2 64-Organism Archaea and Bacteria synthetic community

In addition to the HMP dataset, we also evaluated another synthetic community. Shakya *et al.* (2013) created this set not to simulate any specific environment but to represent phylogenetic and genomic heterogeneity within Bacteria and Archaea usually encountered in communities. It comprises 64 organisms from 62 species. Similarly to the HMP mock set, we evaluated the output and performance of DUDes and three other tools with this dataset. Figure 6 shows ROC curves comparing the results of all tools. Except GOTTCHA, all tools had very similar results, identifying all 62 species in the sample (Sensitivity 0.9 and 1, respectively). Kraken achieved the best AUC value, but with the drawback of a long list (962) of false positives (showing only top 150 entries). DUDes performed as well as MetaPhlAn2 in terms of AUC but with only 2 false positives against 14, respectively. MetaPhlAn2 was the fastest tool among all (Table 2). GOTTCHA performs poorly in this set, with 15 false positives among its 56 identified organisms.

Table 3. STEC samples evaluated

Sample	Pathogens	Expected abundance	DUDes estimated abundance
1122	*C. difficile*	low	0.35%
1253	*C. difficile*	low	0.15%
	C. concisus	low	0.32%
2535	STEC	medium	10%
2638	STEC	high	26%

Known pathogens were discovered from conventional microbiology and computational metagenomics analysis (Loman *et al.*, 2013). All samples were generated with a single Illumina MiSeq run (2 × 151 paired-end sequencing).

3.3 Shiga-toxigenic *E. coli*

In real metagenomics applications, direct performance as with the HMP mock community is typically not possible due to lack of ground truth information. To simulate those scenarios, we analyzed samples where some information about their compositions is already known based on previous conventional microbiology and metagenomics analysis, but the true composition is not completely known. The data were obtained from stool samples of diarrhea patients during the STEC outbreak in Germany (Loman *et al.*, 2013). We analyzed four samples, described in the Table 3. In this evaluation we try to verify if DUDes is capable of identify previously known pathogens. We opted to frame the discussion here in terms of relative abundance given the lack of complete ground truth, just estimations based on the evaluations provided by Loman *et al.* (2013).

We first applied digital normalization to all datasets and mapped the normalized reads against the reference database. We then performed DUDes' top-down analysis, this time allowing strain identifications. In this mode, DUDes will always try to find a set of possible candidate strains in the sample and choose one among them to be the representative strain on the final output. In the sample 1122, DUDes identified 53 likely strains with relative abundance higher than 0.01%, with *Clostridium difficile* (strain M68) present in a low abundance (~0.35%) (Table 3). In sample 1253, DUDes identified 47 strains with relative abundance higher than 0.01%, with *C. difficile* (strain 630) and *Campylobacter concisus* (strain 13826) among them, with abundances of ~0.15 and 0.32%, respectively. For samples 2535 and 2638 DUDes identified STEC (*E. coli* O104:H4 strain 2011C-3493) as candidate strain with a medium and high relative abundance values (~10 and 26%, respectively), matching the previous analysis results and selecting pathogenic strains as candidates.

DUDes can also provide a guided identification, starting the analysis from a defined node of the taxonomic tree. It can be applied when a certain TG is known to be present in the sample from previous analyses or other approaches (e.g. LCA-based tools) but a specific identification in deeper levels is necessary. For example, conventional tests can identify that a certain family is highly abundant on the sample. Starting the analysis from a specific family node, DUDes will provide a more accurate profile for the following taxonomic levels, giving abundance values relative to the specified starting node. We applied the guided identification to samples 2535 and 2638, starting the analysis from previous MetaPhlAn2 results, aiming at strain identification. MetaPhlAn2 identified the *E. coli* species as the most abundant organism in both samples but it did not identify any specific strain. Starting the analysis from the *E. coli* species level (taxid:562) DUDes precisely identified the STEC (*E. coli* O104:H4 strain 2011C-3493) as the first candidate and most likely strain as well as other probable strains in smaller abundance. Further, we performed the same guided identification at each taxonomic rank above STEC (Supplementary Table S3). The higher the starting taxonomic level, the more precise are the identifications with less strain candidates. DUDes' specific identification method can be applied as a complement for other methods of identification, improving their results and providing a deeper classification.

3.4 Tara oceans

We choose one marine sample from the Tara Oceans Project (Sunagawa *et al.*, 2015) to evaluate DUDes' performance on a mixed and complex environment. Differently from human host-associated data, marine environments communities are very challenging to be analyzed, mainly for the lack of reference sequences available. We built a custom database, focused on organisms commonly found in marine environments, composed of Bacterial, Archaeal, Viral and Eukaryotic organisms.

From ~289 million reads on our selected sample (Supplementary Table 1), only 2.2% (<5 million reads) could be mapped against the reference database. That shows how difficult it is to profile metagenomic samples in a whole genome fashion lacking known references. Still, we evaluated DUDes output from this set and compared against the 16S rRNA evaluation provides by the Tara Oceans Project -http://ocean-microbiome.embl.de (only Bacteria and Archaea). At the Kingdom level, DUDes estimated Bacteria as the most common organism (>90%) with the Proteobacteria phylum being the most representative (55%), as indicated in the 16S-based study. DUDes profiled Archaeal organisms representing ~2% of the set. Viruses and Eukaryotes sum up to ~4%, low as expected from such environment.

Since only 2% of the reads could be analyzed, this analysis is still biased towards the references. But it is shown here that DUDes can cope with mixed non-human host-associated environments and that it could profile such samples with a wider range of reference sequences.

4 Discussion

We described here a new method for profiling metagenomics samples with a completely new approach to explore the taxonomic tree structure. In our experiments with mock communities, DUDes achieved high accuracy even in lower taxonomic levels. In an extreme scenario, using only 1% of the data with very few reads for some organisms, DUDes was the best tool overcoming GOTTCHA and MetaPhlAn2. Surprisingly, kraken, a read binning tool, had an excellent performance in this set, but with the disadvantage of a high number of false positives, as expected in this sort of application. DUDes achieved two times fewer false positives than kraken with the best AUC and sensitivity value among all tested tools in the HMP experiment. With 10% of the dataset, DUDes performed as well as with the full set of reads, showing that it can be very precise and stable with small sample sizes. In a broader synthetic community set with more than 100 million reads, DUDes performed equally well even with high organism diversity. From those results we see two main advantages of our tool: first, the method can be applied to profile datasets with low abundant organisms, identifying TGs with very few matches. Second, selecting only part of the dataset poses as a good approach to reduce sample size, allowing faster analysis. Smaller samples and the application of digital normalization generated good results in the HMP datasets without information loss (Supplementary Figure S4). This strategy can be useful with the increasing amount of data generated by NGS technologies, decreasing mapping and execution time as well as memory usage. In the meantime, a thorough analysis is necessary to better estimate the effects of such techniques. Digital normalization can skew abundances downwards and should be carefully used when there is no ground truth available.

DUDes also performed well when applied to outbreak samples for pathogen detection. Our method's results corroborated previous findings and could be used as a fast alternative or confirmation tool to the pathogen detection problem. DUDes can also be a fine-tuning tool for posterior analysis of LCA-based methods identifications when they cannot achieve high taxonomic levels.

The candidate selection approach used in DUDes is not unique to DUD method and it could also be applied in LCA tools. However DUDes provides this functionality out-of-the-box. In addition the choice is not based only in a presence of a certain taxon (given by counting read matches) but also based on a comparison against the other taxons of the same taxonomic level, giving more significance to the candidate selection.

By transforming the information from read mapping to bins of the same size, and subsequently selecting the same number of bins for comparisons we could ameliorate the fact that TGs are not evenly represented in the database. This technique provides a fair comparison among TGs regardless of their number of reference sequences.

The DUD method implemented in DUDes provides a new way to analyze the taxonomic tree structure. We see several advantages in this method: first it provides a reliable way to identify the presence of TGs, making a comparison on each taxonomic level. It also can solve ambiguities in a less conservative manner, first by allowing concurrent identifications and second by selecting a set of candidates when a specific identification is not possible. In comparison with LCA-based tools we can point out some methodological differences: kraken, like many LCA-based methods, applies LCA on a sequence level and set the presence of TGs after all sequences were classified in the LCA system. DUDes uses the taxonomic information to guide the analysis and it does not use the DUD algorithm directly on a sequence level but on a taxonomic group level. It is important to notice that those two methods have opposite approaches (Fig. 2) and at the same time have been applied in a different way. DUDes' implementation of the DUD method relies on permutations tests among bin scores of nodes of the tree with correction for multiple testing. That introduces statistical significance to our comparisons and decisions with a rigid control of type I errors values to avoid false identifications.

DUDes is a flexible tool that does not rely on a specific or custom-built databases or read mappers and it can run using any set of reference sequences (e.g. draft genomes, marker genes, proteins) not only whole genomes sequences as here presented. The creation of any other custom database is straightforward with DUDesDB. Our tool also provides a possibility to analyze sub-trees from the taxonomic tree structure by setting a start node and a final taxonomic level desired, giving a guided and fast identification.

In conclusion, DUDes propose a novel approach to the taxonomic profiling problem, with a top-down technique to analyze taxonomic tree structures. The DUD can be less conservative than current methods for solving ambiguities in the identifications, and showed superior performance in our experiments compared with recent tools, being very precise at low coverages. Additionally the tool provides a strain identification method that can propose one or more strains presents in a sample. We believe that DUDes can be useful for several applications, from complete metagenomics profiling to pathogen detection studies.

Acknowledgements

We thank Kathrin Trappe for helpful discussions and Enrico Seiler for technical support.

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Ciência sem Fronteiras [BEX 13472/13-5 to V.C.P.] and by the German Federal Ministry of Health [IIA5-2512-FSB-725 to B.Y.R.].

Conflict of Interest: none declared.

References

Benson,D.A. *et al.* (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.

Brown, C.T. *et al.* (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv*, **1203.4802**, 1–18.

Francis,O.E. *et al.* (2013) Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.*, **23**, 1721–1729.

Freitas,T.A.K. *et al.* (2015) Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.*, **43**, 1–14.

Fricke,W.F. and Rasko,D.A. (2014) Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genet.*, **15**, 49–55.

Goeman,J.J. and Solari,A. (2014) Multiple hypothesis testing in genomics. *Stat. Med.*, **33**, 1946–1978.

Handelsmanl,J. *et al.* (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, **5**, R245–R249.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Keeling,P.J. *et al.* (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.*, **12**, e1001889.

Klymiuk,I. *et al.* (2014) A physicians' wish list for the clinical application of intestinal metagenomics. *PLoS Med.*, **11**, e1001627.

Köser,C.U. *et al.* (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.*, **8**, e1002824.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Levenshtein,V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, **10**, 707–710.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindgreen,S. *et al.* (2016) An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.*, **6**, 19233.

Lindner,M.S. and Renard,B.Y. (2015) Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS. *PLoS One*, **10**, e0117711.

Loman,N.J. *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA*, **309**, 1502–1510.

Mande,S.S. *et al.* (2012) Classification of metagenomic sequences: Methods and challenges. *Brief. Bioinform.*, **13**, 669–681.

Meinshausen,N. (2008) Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.

NCBI. (2015) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **43**, D6.

Ounit,R. *et al.* (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**, 236.

Pallen, M.J. (2014) Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*, **141**, 1856–1862.

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

Shakya,M. *et al.* (2013) Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.*, **15**, 1882–1899.

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

Truong,D.T. *et al*. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.

Turnbaugh,P.J. *et al*. (2007) The human microbiome project. *Nature*, **449**, 804–810.

Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*., **15**, R46.