Computational and Structural Biotechnology Journal 20 (2022) 1798–1810

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

UPIMAPI, reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets

João C. Sequeira [a,b], Miguel Rocha [a,b], M. Madalena Alves [a,b], Andreia F. Salvador [a,b,*]

[a] CEB - Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal
[b] LABBELS – Associate Laboratory, Braga/Guimarães, Portugal

ARTICLE INFO

Article history:
Received 1 July 2021
Received in revised form 31 March 2022
Accepted 31 March 2022
Available online 9 April 2022

Keywords:
Genomics
Metagenomics
Metatranscriptomics
Functional annotation
Metabolic pathways mapping
Differential expression analysis

ABSTRACT

Omics and meta-omics technologies are powerful approaches to explore microorganisms' functions, but the sheer size and complexity of omics datasets often turn the analysis into a challenging task. Software developed for omics and *meta*-omics analyses, together with knowledgebases encompassing information on genes, proteins, taxonomic and functional annotation, among other types of information, are valuable resources for analyzing omics data. Although several bioinformatics resources are available for *meta*-omics analyses, many require significant computational expertise. Web interfaces are more user-friendly, but often struggle to handle large data files, such as those obtained in metagenomics, metatranscriptomics, or metaproteomics experiments.

In this work, we present three novel bioinformatics tools, which are available through user-friendly command-line interfaces, can be run sequentially or stand-alone, and combine popular resources for functional annotation. UPIMAPI performs sequence homology-based annotation and obtains data from UniProtKB (e.g., protein names, EC numbers, Gene Ontology, Taxonomy, cross-references to external databases). reCOGnizer performs multithreaded domain homology-based annotation of protein sequences with several functional databases (i.e., CDD, NCBIfam, Pfam, Protein Clusters, SMART, TIGRFAM, COG and KOG) and in addition, obtains information on domain names and descriptions and EC numbers. KEGGCharter represents omics results, including differential gene expression, in KEGG metabolic pathways. In addition, it shows the taxonomic assignment of the enzymes represented, which is particularly useful in metagenomics studies in which several microorganisms are present.

reCOGnizer, UPIMAPI and KEGGCharter together provide a comprehensive and complete functional characterization of large datasets, facilitating the interpretation of microbial activities in nature and in biotechnological processes.

1. Introduction

Current high-throughput sequencing technologies produce valuable information to characterize microbial communities in terms of taxonomic composition and functional behavior. The biological interpretation of such big datasets benefits from easy-to-use tools, that automatically convert raw data into comprehensive information, ideally with minor input from users and reduced time of analysis.

Generally, bioinformatics tools used for metagenomics (MG) and other *meta*-omics analyses (e.g., metatranscriptomics (MT) and metaproteomics (MP)) perform taxonomic and functional annotations, thus linking microbial identity to function in complex microbial communities. The most popular methodology for gene/protein annotation consists in inferring function based on sequence homology, using sequence alignment tools such as BLAST, DIAMOND, lastal, or MMSeq2, among others [1]. These tools find similarities between gene or protein sequences present in biological samples, and those in reference databases. The most popular reference databases for homology-based annotation are UniProtKB [2] and RefSeq [3]. An advantage of using the UniProtKB as the reference database for protein annotation is the possibility of obtaining complementary information from several other databases (e.g., BioCyc, BRENDA, PDB, RefSeq, CDD, KEGG, InterPro, PRIDE, eggNOG, Ensembl), by using the UniProt ID mapping service (available at https://www.uniprot.org/uploadlists/).

* Corresponding author at: Universidade do Minho, Centro de Engenharia Biológica, Campus de Gualtar, 4710-057 Braga, Portugal.
E-mail address: asalvador@ceb.uminho.pt (A.F. Salvador).

https://doi.org/10.1016/j.csbj.2022.03.042

In *meta*-omics experiments, it is common that the majority of genes and proteins are not annotated, as a significant number of genomes are not sequenced, or are poorly annotated, which results in a significant amount of proteins identified as "putative" or "uncharacterized" [4]. Using sequence homology annotation, proteins with low sequence similarity, but with the same function, may not be annotated [4]. Allying sequence homology-based annotation to, for example, annotation based in protein conserved domains maximizes function assignment and better reflects microbial functions in complex communities [4,5]. A well-known protein domain database is the Conserved Domains Database (CDD), which includes the Cluster of Orthologous Groups of proteins (COG), and is a manually curated collection of protein profiles and domains, represented as Hidden Markov Models (HMMs) [6]. COG is one of the most used resources for protein functional characterization and is recommended by the Genome Standards Consortium to characterize newly published genomes [4].

Web-based applications performing functional annotation using the COG database as reference include WebMGA and the NCBI's Batch CD-search [7]. WebMGA performs annotation with HMMER3, and PFAM and TIGRFAM databases as reference, and with RPS-BLAST using COG, KOG and Protein Clusters databases as reference [8]. The Batch CD-search service performs annotation with the models present in the CDD database, which include PFAM, TIGRFAM, Protein Clusters, NCBIfam, SMART, COG and KOG. Web-based applications depend on webservers that limit the number of sequences that can be submitted simultaneously, turning the process slower, and requiring that the datasets are split before analysis. Thus, using command-line tools may be advantageous, as it is the case of DFAST, which uses GHOSTX or BLASTP for protein alignment against a reference database containing 124 well-curated prokaryotic genomes, and RPS-BLAST and HMMER for protein functional annotation based on the COG and TIGRFAM databases [9]. Prokka, another command-line tool, uses BLAST+ blastp and HMMER for annotation based in sequence and domain homology, respectively, using a series of databases as reference, which includes user-set databases, good-quality protein sequences from UniProt and RefSeq, Pfam and TIGRFAM [10]. eggNOG-mapper [11] and Mantis [12] are also command-line tools that perform functional annotation by using HMMER and DIAMOND. eggNOG-mapper annotates protein sequences with reference to the eggNOG database, while Mantis performs annotation with a larger set of reference databases, i.e., eggNOG, Pfam, KOfam, TIGRfam, and NCBIfam [12].

Even though these tools provide diverse methods for gene/protein annotation, there is no tool allowing complete retrieval of information from UniProt's ID mapping, preceded by sequence homology annotation. Also, none of these tools include all the databases available at CDD, while avoiding the use of web servers.

Several other tools have been developed to facilitate the interpretation of functional annotation results and represent the identified genes or proteins in metabolic pathways. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database provides many functionalities to explore metabolic and regulatory networks through interactions between enzymes and metabolites [13]. Most resources are accessed through the KEGG Orthology database, where for each function a K number, or a KEGG Ortholog (KO), is assigned [13]. KEGG Pathway provides hundreds of hand-drawn maps that represent different metabolic pathways, and that can be customized to represent functions of interest, with these functions identified by their KOs. This visualization helps to understand the intricate relations between different enzymes [14].

Some tools have been developed to represent functions of interest in KEGG's metabolic maps. For example, KEGG Mapper offers a collection of tools that connect the databases of KEGG, e.g., "Reconstruct Pathway", which maps KOs to PATHWAY's maps, and "Color Pathway" which expands the mapping of KOs with the option to color differently the identified boxes [15]. iPath3.0 is a tool that produces interactive plots, where detailed descriptions of the pathway reactions can be accessed by hovering the corresponding elements in the metabolic maps [16]. The Pathway Projector tool provides a Zoomable User Interface that allows the visualization of the entire global metabolic map of KEGG [17]. PathCase is another web service, which integrates information from several different sources, including KEGG and BioCyc, to build custom functional maps. These custom maps allow PathCase to provide querying for different levels of detail (at the organisms, biochemistry, molecular or genetic level) directly on the maps [18]. SQMTools is an R package that analyzes results from SqueezeMeta, a MG and MT data analysis pipeline, and can be used to represent differential gene expression between two samples through a color gradient [19]. KEGGprofile is an R package that represents expression profiles in KEGG maps [20]. MEGAN, a popular MG annotation pipeline, also represents results in KEGG Pathways [21]. However, there is no automated tool that, from a simple table of information, represents the functional characterization of MG/MT/MP datasets in metabolic pathways.

In this work, we present three command-line bioinformatics tools, UPIMAPI, reCOGnizer and KEGGCharter, developed for functional annotation with multiple databases, and representation of *meta*-omics data in metabolic pathways. After performing sequence-based annotation, UPIMAPI allows to obtain information on UniProt IDs, and after performing domain-based annotation with several databases, reCOGnizer obtains taxonomic classifications, EC numbers, and COG categories. Finally, KEGGCharter represents functional potential, taxonomy and differential gene expression in KEGG metabolic maps. All these functionalities are accessed through completely automated workflows.

2. Methods

2.1. Development of UPIMAPI tool for taxonomic and functional annotation

UPIMAPI (UniProt's Id Mapping through API) was developed with Python 3 and combines automatic construction of reference databases, sequence-based annotation with DIAMOND, and retrieval of information from UniProt (Fig. 1). UPIMAPI constructs reference databases in three different ways (Fig. 1, step 1), which should be chosen by the user through the "--database" parameter: if "uniprot" is chosen, UPIMAPI will download the entire UniProt database, and use it as the reference; similarly, if "swissprot" is chosen, the SwissProt database will be downloaded and used; finally, if "taxids" is chosen, a selection of tax IDs must be inputted through the "--taxids" parameter, and UPIMAPI will download the reference proteomes corresponding to those tax IDs through UniProt's Taxonomy database and use it as the reference database. A custom database can also be inputted directly to UPIMAPI in FASTA or DMND (DIAMOND binary) format, but it must contain UniProt IDs in the headers, in the right format (e.g., "sp|Q74FU6|SFRA_GEOSL"). UPIMAPI also downloads the versions of the UniProt Knowledgebase, the UniProtKB/SwissProt, and the UniProtKB/TrEMBL.

Annotation with DIAMOND takes as input the query protein sequences (in FASTA format) and the reference database (Fig. 1, step 2). An additional perk of UPIMAPI is the automatic determination of the optimal parameter values for running DIAMOND, i.e., the number of threads, block size, and index chunks – with the last two increasing annotation speed at the cost of higher memory usage. The automatic number of threads is the number of threads available minus 2, the automatic block size is set as the amount of memory available in Gb divided by 20, and the automatic number

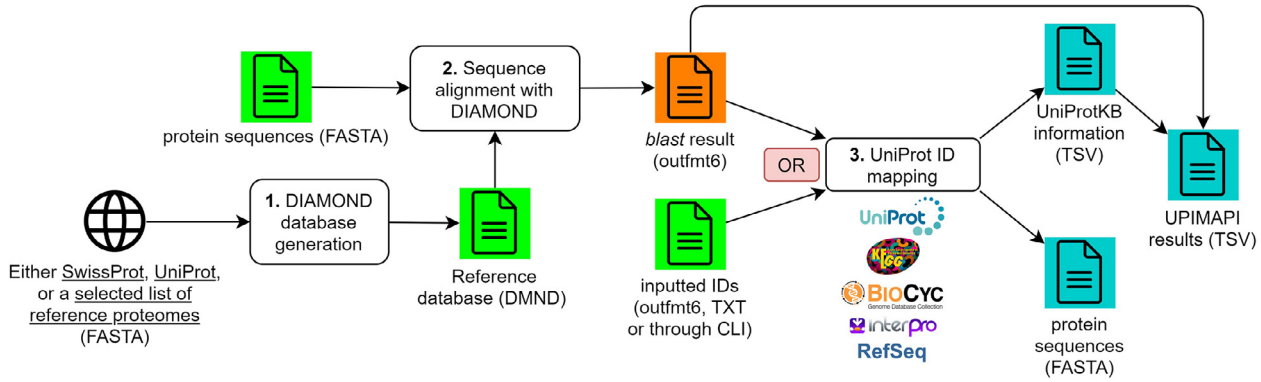

Fig. 1. Workflow of UPIMAPI, which includes the first step of sequence-based annotation with DIAMOND and a second step where functional information is obtained from the UniProtKB, including cross-references to other databases. Green icons represent input files (i.e., FASTA protein sequences, reference database, and UniProt IDs either through a BLAST result file, a TXT file, or directly through the CLI), orange icons represent intermediate files (BLAST result file), and blue icons represent output files, i.e., TSV files containing information from UniProtKB (including cross-references to several databases), and FASTA files containing the annotated protein sequences. Database logos identify some databases from which information can be obtained through UPIMAPI. The pink box identifies the different possibilities to submit the IDs to the UniProt ID mapping.

of index chunks is set as 1, 2, 3, or 4 depending on if the block size is over 3, between 2 and 3, between 1 and 2, or 1 or less, respectively. The default E-value is 1e-3, determined as described in SI (Section 1.6). If the user wants to manually alter these parameters, this can be done directly through the CLI. UniProt IDs are obtained after annotation with DIAMOND (Fig. 1, step 2) and are automatically submitted to the UniProt ID mapping (Fig. 1, step 3).

Mapping of IDs follows two different implementations: (a) mapping through UniProt's API, using urllib to sequentially submit batch requests, and parsing and storing the information obtained in local storage; and (b) local mapping of SwissProt IDs (to use if the input data have SwissProt IDs), that does the download of the SwissProt's DAT file from UniProt's FTP site (Table S2), which is then parsed and queried with BioPython [22]. The information obtained with this last option is parsed to be as close as possible to the information obtained through the API and then it is also stored locally. UPIMAPI offers the possibility of submitting UniProt IDs directly to the ID mapping, skipping the annotation step, i.e., outfmt6 files resulting from sequence alignment, either with DIAMOND (without using UPIMAPI) or BLASTp, TXT files with UniProt IDs separated by commas, or UniProt IDs directly inputted through the command line (Fig. 1, pink box). Requests to the Uniprot API are performed by using the following default parameters: 10,000 IDs per request, and 3 s between requests. The user can also choose the fields of information to be retrieved from the UniProtKB through the command line. Otherwise, UPIMAPI will use a default list of fields to be outputted (Table S1). UPIMAPI organizes the information in "columns" and "databases": "Columns" correspond to functional, taxonomic, expression, structural, and additional information stored at UniProt (e.g., "Gene names", "Protein names", "EC number", "Pathway", "Gene ontology (GO)", "Taxonomic lineage"); "Databases" correspond to cross-references between UniProt and external databases, such as BioCyc, BRENDA, PDB, RefSeq, CDD, KEGG, InterPro, PRIDE, eggNOG, Ensembl, among others. These parameters may be changed by the user if necessary, through the command line. Examples and explanations of commands to install and run UPIMAPI are given in SI (Section 1.4.1).

2.2. Development of reCOGnizer tool for domain-based functional annotation

reCOGnizer combines HMM databases construction with domain-based annotation and retrieves the information obtained from the annotation with different databases (Fig. 2).

For database construction, the Position Specific Scoring Matrices (PSSMs) are first retrieved from CDD through NCBI's FTP site, and then two different workflows can be followed: (a) a taxonomic workflow (Fig. 2, step 1a), which is applied to Pfam, NCBIfam, Protein Clusters, TIGRFAM, and COG databases, and which splits the SMP files by the respective taxonomy assigned as detailed in the hmm_PGAP and NOG.members relation tables (Table S2); and (b) a more general workflow (Fig. 2, step 1b), applied to all databases, which builds the databases in a similar structure as the one available in the "little_endian" directory (https://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/little_endian/), containing all the SMPs of each database. These databases are built using the makeprofiledb tool with default parameters. reCOGnizer also downloads the version of the CDD database. Domain-based annotation is performed with Reverse PSI-BLAST (RPS-BLAST), a variant of PSI-BLAST [23], and may follow two different workflows, depending on the availability of taxonomic information. A taxonomic file can be used to detail the tax IDs for each protein and is specified with the parameter "--tax-file", while the column with the proteins IDs (as they are in the FASTA file) and corresponding tax IDs must be specified with the "--protein-id-col" and "--tax-col" parameters, respectively. In this case, reCOGnizer will apply the taxonomic workflow for annotation with Pfam, NCBIfam, Protein Clusters, TIGRFAM, and COG databases. For this purpose, reCOGnizer splits the FASTA file into the different tax IDs and then, for each tax ID specified in the taxonomic file, reCOGnizer retrieves the tax IDs of the entire lineage (from the taxonomy.rdf file, Table S2) and uses the partial databases corresponding to those tax IDs together as reference (Fig. 2, step 3a). A more general workflow, which annotates all sequences against all HMMs present in the database, is applied for annotation with CDD, KOG and SMART databases, and for the remaining databases if taxonomic information is not provided by the user. In this taxonomy-independent annotation workflow, the entire databases are used as the reference for annotation (Fig. 2, step 3b). Before the annotation step, reCOGnizer splits the FASTA files into the number of threads to be used to achieve parallelization, which increases the speed of the analysis (Fig. 2, step 2). The default E-value for annotation is 1e-3, determined as described in SI (Section 1.6), which can be changed through the CLI. There is also the option of annotating only with a specified set of databases, by using the "--databases" parameter (e.g., --databases COG,Pfam – to annotate only with COG and Pfam databases).

The annotation results are obtained in ASN (outfmt11) reports, which are inputted to the rpsbproc tool for multi-domain solving

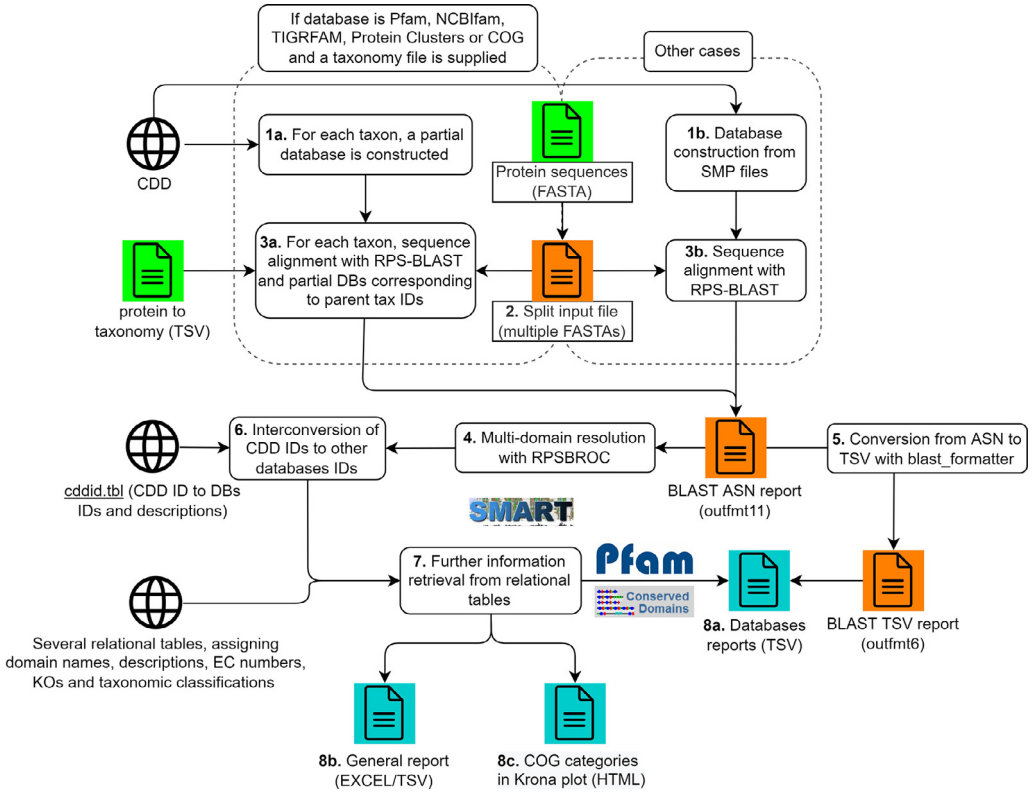

Fig. 2. Workflow of reGOGnizer: construction of databases (step 1); domain-based annotation of inputted protein sequences with the reference databases (step 3), after splitting the input sequences into multiple FASTA files (step 2); post-processing of annotation results (steps 4 and 5); interconversion of CDD IDs to other databases IDs (step 6) and retrieval of information from several databases, by using relational tables (step 7); output the annotation information in TSV, Excel and HTML report files (step 8). Web icons represent resources retrieved from the web, specifically the domain models from CDD and the relational tables from NCBI, COG, eggNOG, and SMART. The green icons represent the input files (FASTA protein sequences and the file containing the taxonomic information), the orange icons represent intermediate files generated during the analyses (input protein sequences split into multiple FASTA files and the annotation results in ASN and TSV formats), and the blue icons represent the output files.

(Fig. 2, step 4). rpsbproc processes the annotation results providing a non-redundant list of conserved domains that do not overlap in the query sequences, which bypasses the problems of considering only the first annotation for proteins with several domains for several functions. ASN reports are converted into TSV (outfmt6) reports, with the blast_formatter tool (Fig. 2, step 5), to facilitate the visualization of the annotation results and respective metrics. At the end of the workflow, CDD IDs, obtained from the annotation in the previous steps, are converted to the IDs of the other databases (i.e., Pfam, NCBIfam, Protein Clusters, TIGRFAM, SMART, CDD, COG, and KOG databases) using the "cddid.tbl" relational table (Fig. 2, step 6). From this relational table, a domain description from CDD is also obtained.

Once the protein IDs from the 8 databases are obtained, additional functional information is collected (Fig. 2, step 7), namely: COG functional categories and protein descriptions, by ID mapping of COG and KOG IDs with "cog-20.def.tab" and "kog" files, respectively; COG general functional categories, by mapping the obtained categories with "fun-20.tab"; EC numbers and KOs, by mapping COG IDs with eggNOG's (Huerta-cepas et al., 2019) files "NOG.members.tsv" and "egg-nog4.protein_id_conversion.tsv"; domain names, EC numbers and taxonomic classifications, obtained by mapping NCBIfam, Pfam, Protein Clusters and TIGRFAM IDs with the "hmm_PGAP.tsv" file; and SMART descriptions, by mapping SMART IDs with the "descriptions.pl" file.

Web locations of relational tables used by reCOGnizer, and the CDD tarball containing the HMM models of the databases, are given in Table S2.

Regarding the assignment of EC numbers to COG IDs, these are only assigned when at least 50 % of the EC numbers that match the COD ID are concordant. On the other hand, all the KOs assigned are listed in the final output.

Examples and explanations of commands to install and run reCOGnizer are given in SI (Section 1.4.2).

2.3. Development of KEGGCharter for visualization of functional annotation results in metabolic maps

KEGGCharter is a command-line tool that accesses the following KEGG Pathway's functionalities: the interconversion of KEGG IDs, KOs and EC numbers, and the representation of KOs in KEGG metabolic maps, together with taxonomy and gene expression information.

The input to KEGGCharter is a table, in either TSV or Excel format (Fig. 3), containing:

- a column with either KEGG IDs, KOs, or EC numbers;
- columns with either MG and/or MT quantification, i.e., abundance of different taxa in MG datasets and gene expression analysis, respectively. If no quantification is available, the "--input-quantification" parameter may be used to input a placeholder quantification, assigning a value of 1 to all rows;
- a column with taxonomic information, for the MG workflow. If no taxonomic classification is available, the "--input-taxonomy" parameter may be used to input a value directly from the command line.

KEGGCharter's workflow begins with the interconversion of input IDs (from KEGG IDs or EC numbers to KOs and from KOs to EC numbers) (Fig. 3, step 1). These interconversions are performed

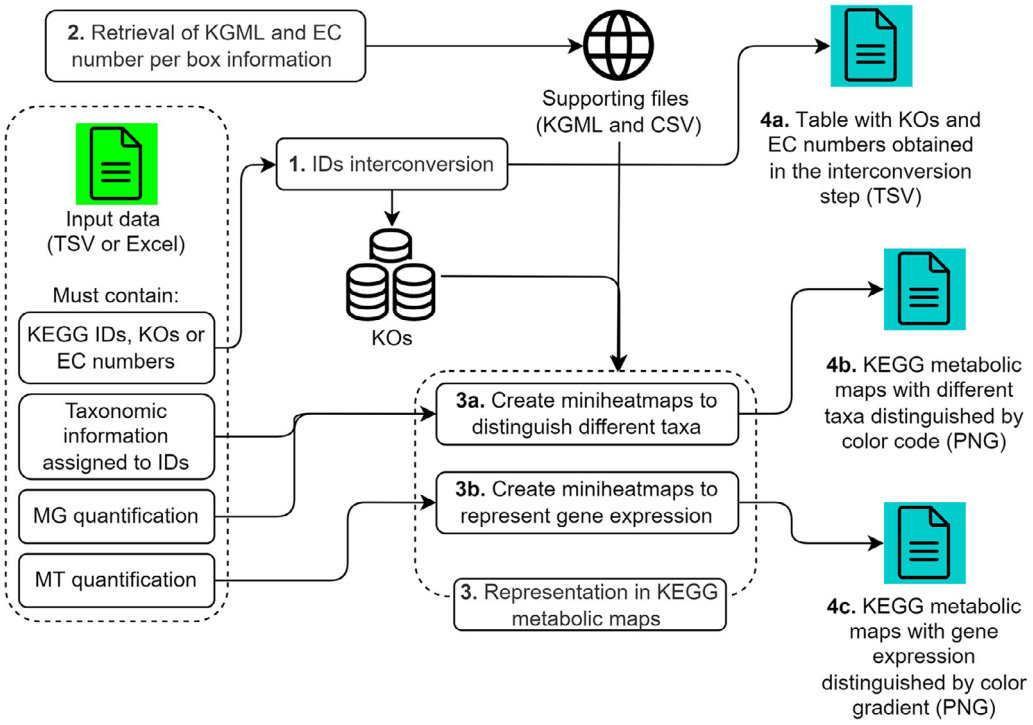

Fig. 3. Workflow of KEGGCharter, which begins with the interconversion between KEGG IDs, KOS, and EC numbers, followed by the retrieval of KGMLs and EC number to box relation from the KEGG API, and finally by the generation of metabolic maps representing the genomic potential (MG analysis) or the gene expression (MT analysis) information. The web icon represents the retrieval of KGMLs and EC number to box relations; the green file icon represents the input file containing KEEG IDs, KOs, or EC numbers, together with quantification and taxonomic information; and the blue file icons represent the two outputs from KEGGCharter: TSV files with tables containing the IDs obtained from the interconversion step, and the KEGG metabolic maps in PNG format, containing taxonomic and quantitative information. For each map, a KGML file is obtained, which is the XML description of the metabolic map. A CSV file is also built by KEGGCharter, where the KOs of each box are stored.

to obtain as much information as possible from the KEGG database. The interconversions between IDs use the "KEGG.REST.kegg_link" method of BioPython [22]. Results from these interconversions are outputted in TSV format (Fig. 3, step 4a). The KOs obtained are then mapped in KEGG metabolic maps.

The list of metabolic maps to be represented can be inputted through the command line, otherwise, a default selection of maps is used (Table S3). If the user wants to know which maps are available, the following command can be run: *keggcharter.py --show-available-maps*.

For each metabolic map, KEGGCharter will obtain information necessary for mapping function into maps and store it into two files: a Kegg Markup Language (KGML) file with the detailed description of the map pulled directly from KEGG, and a Comma-Separated Values (CSV) file relating KOs to corresponding boxes in the maps, built by extracting the information of EC numbers to boxes from the KGML and converting those EC numbers to KOs (Fig. 3, step 2). KGMLs are obtained with the "KEGG.REST.kegg_get" method of BioPython.

The main feature of KEGG metabolic maps is the presence of boxes representing reactions. The maps and the corresponding boxes are described in XML format on KGML files, obtained in step 2. In these KGMLs, each box contains a list of orthologs, which are the corresponding KOs. KEGGCharter identifies the specific boxes to manipulate, superimposes each box with a new box, and colorizes according to the information to be represented. The label of these boxes is always an EC number, in the original version of the maps, but when new boxes are superimposed, that information is hidden. KEGGCharter retrieves these EC numbers by converting the KOs of each box (obtained from the KGML) to EC numbers and storing that information in CSV files. When generating new maps, for each box the most abundant EC number is set as the label for the new boxes.

It is also important to note that a given KEGG ID can correspond to more than one KO, which can lead to an overestimation of the protein expression. To overcome this, KEGGCharter determines how many KOs are assigned to each protein and divides the protein quantification by the number of KOs, before drawing the maps.

Representation of (meta)genomics and (meta)transcriptomics data in KEGGCharter follows two distinct workflows depending on the type of analysis, i.e., MG or MT analysis (Fig. 3, step 3). Representation of MG results distinguishes between different taxa by attributing a different color to each taxon. By default, KEGGCharter represents the 10 most abundant microorganisms from the metagenome in the metabolic maps. Nevertheless, a different number of microorganisms can be represented simultaneously by inputting the desired number in the "--number-of-taxa" parameter. In addition, the user might select the specific taxa to be represented by using the "--taxa-list" parameter. For example, if the user intends to compare the genomic potential between two microorganisms present in the metagenome, only their taxa names should be indicated. Because this representation is made by changing the color of the boxes in metabolic maps, if different taxa have KOs corresponding to the same box, the new box will be split between the colors of the different corresponding taxa. On the other hand, representation of MT results describes the gene expression of the community as a whole. MT quantification is first summed by KO, and for each box in each metabolic map, KEGGCharter will sum the quantifications of the several KOs mapping to that box and represent that quantification with a colored gradient. The new boxes will be divided by the number of samples represented, up to a maximum of 10 samples.

KEGGCharter obtains information from KEGG on the correspondence of different taxa to specific functions and specific maps. This avoids the representation of functions associated with organisms that are not able to perform those functions, which can happen when using KOs for mapping functions, since KOs are specific to functions but not to taxa. Metabolic maps are outputted in PDF format with the "KEGG.KGML.KGML_pathway.to_pdf" method of Biopython. The colors for both different taxa and different levels of expression are obtained with matplotlib, which is also used to export the legends explaining the color codes as PNG files. pdftoppm (available at https://man.archlinux.org/man/pdftoppm) is used to convert the colorized metabolic maps from PDF to PNG format, and PIL is used to resize the images and join the metabolic maps with the respective labels, outputting the joined information in PNG format (Fig. 3, steps 4b and 4c).

KEGGCharter can be run using the outputs from UPIMAPI or reCOGnizer (Fig. 4). The commands to automatically perform this analysis are given in SI (Section 1.4.3).

2.4. Benchmarking methodology

UPIMAPI (version 1.6.4) and reCOGnizer (version 1.6.4) were compared with state-of-the-art functional annotation tools, namely Prokka (version 1.14.6), DFAST (version 1.2.15), Mantis (version 1.4.5) and eggNOG-mapper (version 2.1.6). For that purpose, different datasets were used, one real metagenomics dataset (described in SI, Section 1.5.3), and five datasets each one consisting of seven genomes, obtained from Ensembl Genomes database (SI, Section 1.5.1), and hereafter named as artificial metagenomes (artificial metagenome #1 to #5). The real dataset was used to evaluate the performance regarding the analysis of real metagenomics data, while the artificial metagenomes were used to determine the accuracy of the annotation by calculating quality metrics.

Because all the functional annotation tools run for benchmarking, i.e., Mantis, eggNOG-mapper, Prokka and DFAST, combine sequence and domain homology annotation, the results of UPIMAPI (performing sequence homology annotation) and reCOGnizer (performing domain homology annotation) were combined for comparative purposes, i.e., the proteins without EC number or orthologous groups (OGs) assigned by UPIMAPI received the annotation obtained with reCOGnizer.

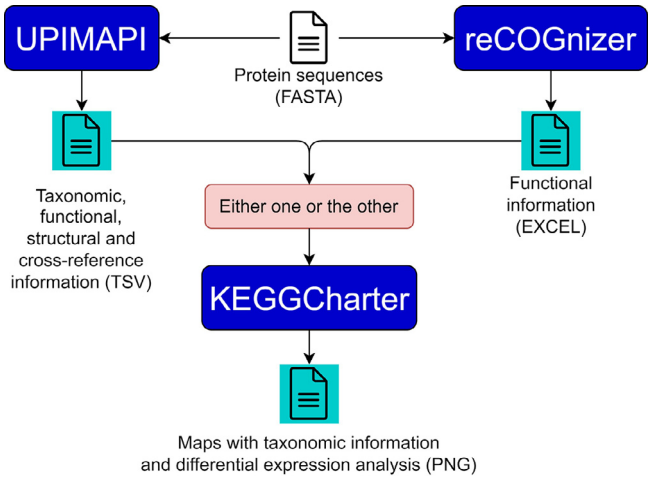

Fig. 4. Interconnection between UPIMAPI, reCOGnizer and KEGGCharter. UPIMAPI and reCOGnizer annotate protein sequences inputted in FASTA format and obtain functional information, i.e., KEGG IDs, KOs, and EC numbers, that can be used as input for KEGGCharter, which represents proteins in KEGG metabolic maps. Blue file icons represent the main output files obtained by the three tools.

All tools were run with default parameters, with the following exceptions that were necessary to run the tools in the most similar conditions: (1) the UniProt database was added to the reference databases used to run Mantis; (2) Prokka was run using the "--metagenome" parameter; and (3) protein sequences from the organisms in the artificial metagenomes were removed from the UniProt database, to guarantee that the datasets were independent of the reference database (SI, Section 1.5.1).

Prokka and DFAST used as input the reference genomes (in the case of the artificial metagenomes) and the contigs (in the case of the real metagenome) because these tools perform their gene calling before functional annotation. On the other hand, UPIMAPI, reCOGnizer, Mantis and eggNOG-mapper used protein FASTA files as input, containing protein sequences obtained after gene calling (with Prodigal, using default parameters) of the reference genomes and the contigs, in the case of the artificial metagenomes and the real metagenome, respectively.

The quality metrics precision, recall and F1 score, calculated as described in SI (Section 1.6), were used to assess the performance of the tools. The number of proteins with OGs and EC number assigned with each one of the annotation tools was also taken into account to perform the comparison.

All analyses were run on a #116-Ubuntu SMP kernel, with 126 Gb of memory and 16 threads.

3. Results

The source codes of UPIMAPI, reCOGnizer and KEGGCharter are available at GitHub at https://github.com/iquasere/UPIMAPI, https://github.com/iquasere/reCOGnizer, and https://github.com/iquasere/KEGGCharter, respectively.

3.1. Functional annotation with UPIMAPI and reCOGnizer

UPIMAPI provides a Command Line Interface (CLI) to perform protein sequence annotation and to retrieve functional information from the UniProtKB, through UniProt's ID mapping service [24]. The annotation step runs local sequence alignment with DIAMOND (Fig. 1, step 2) against a reference database that can be automatically downloaded (Fig. 1, step 1). In the UniProt ID mapping step, UPIMAPI uses the UniProt API for obtaining diverse information about the proteins/IDs analyzed (Fig. 1, step 3). This is an iterative process, repeated until functional information is acquired for the highest number of UniProt IDs. As a result, complete information assigned to UniProt IDs, including those from cross-reference databases linked to UniProtKB and taxonomic assignment, is obtained.

Two types of output can be obtained with UPIMAPI: the protein sequences, in FASTA format, or the data available at the UniProt knowledgebase, in a TSV table (Fig. 1). Besides these two main outputs, files resulting from annotation with DIAMOND are also generated, i.e., an outfmt6 file with the annotation result and metrics of the sequence alignments, and a FASTA file containing query sequences that were not annotated (because they did not show significant homology to those in the reference database).

In addition to the ID mapping through the UniProt's API, UPIMAPI performs the local ID mapping of SwissProt IDs. The information outputted after local ID mapping is organized differently from the output obtained through the UniProt's API, and UPIMAPI resolved this issue by parsing the information into the formats of the API for almost all columns. However, ID mapping with UPIMAPI is still largely dependent on UniProt's API for mapping TrEMBL IDs, and this information could potentially be obtained with a fully local method. Nevertheless, we noticed that local TrEMBL ID mapping would require the storage of ~1 Tb of data, which could be a problem for the users who have no access to servers with enough

space. UPIMAPI retrieves the same information as UniProt's ID mapping web service but can handle millions of IDs with a single command, which is a major improvement for the analysis of *meta*-omics datasets (containing typically hundreds of thousands of IDs). Comparatively, through the UniProt's API, a maximum of 50,000 protein IDs can be uploaded at a time (from https://www.uniprot.org/help/uploadlists).

While UPIMAPI performs sequence-based functional annotation, reCOGnizer performs domain-based annotation of large datasets (protein FASTA files), using as reference the following databases: CDD, COG, KOG, NCBIfam, Protein Clusters, Pfam, SMART, and TIGRFAM. The general steps of reCOGnizer are the following: database construction for general and taxonomic workflows (Fig. 2, step 1), sequence alignment with RPS-BLAST (Fig. 2, step 2), interconversion of CDD IDs to other database IDs (Fig. 2, step 6), retrieval of information from databases (Fig. 2, step 7) and generation of output reports (Fig. 2, step 8).

For each database, reCOGnizer generates a TSV report, containing the BLAST metrics and annotation results obtained with RPS-BLAST, as well as extra information including functional sites and motifs, protein superfamilies, domain names, and descriptions, EC numbers, taxonomic classifications, and KOs (Fig. 2, step 8a). All these reports are gathered into general reports in TSV and Excel formats (Fig. 2, step 8b). For the results obtained from COG and KOG databases, another report is generated, quantifying the occurrence of each COG/KOG and organizing the results by the respective COGs' categories. This information is represented in interactive pie charts (Krona plots) (Fig. 2, step 8c).

Some of the IDs attributed by reCOGnizer, i.e., those from NCBIfam, Pfam, Protein Clusters and TIGRFAM, have taxonomy assigned, and this is the taxonomy that is reported by reCOGnizer. Sequences of the proteins annotated are also included in the reports, but to avoid considerably larger outputs when analyzing big datasets, reCOGnizer provides an option "--no-output-sequences" to exclude this information.

The analysis of the artificial metagenomics dataset #3 is given as example. UPIMAPI annotated 26,634 proteins from a total of 26,920 proteins, i.e., 99 % of the total proteins (Table 1), from which 14 % were assigned to a KEGG ID and 17 % were assigned to an EC number (Table 1, Table S5). These results correspond to the information present in the UniProtKB.

After the annotation with UPIMAPI, 15 to 21 % of the proteins from the artificial metagenomes were identified as uncharacterized (Table 1). The high number of uncharacterized proteins makes it more difficult the evaluation of the overall functional potential of microbial communities. In this study, domain-based annotation with reCOGnizer increased significantly the functional information retrieved from the artificial metagenome (Table 2, Table S9). Table 2 contains the functional categories obtained by reCOGnizer for proteins identified by UPIMAPI as "uncharacterized" in the artificial metagenomes. reCOGnizer could obtain information for 19.44 ± 0.97 % (excluding the ones annotated as "poorly characterized" and with "no COG ID") of the proteins annotated as "uncharacterized proteins" by UPIMAPI.

In total, reCOGnizer retrieved functional information for 83.76 ± 2.00 % of the proteins inputted (Table 3, Table S9), and EC numbers were obtained for 41.00 ± 1.57 % of the proteins, considering the analysis of all the artificial metagenomes. Our results clearly show that these two methodologies (i.e., sequence- and domain-based annotation) are complementary and useful to extract the most information possible from the analyzed datasets. Fig. 5 shows the default interactive Krona plot generated by reCOGnizer, representing the functional characterization of the artificial metagenome #3, which is organized by COG IDs and respective COG categories.

3.2. Representation of metabolic functions in KEGG maps with KEGGCharter

KEGGCharter provides an elegant view of (meta)genomics and (meta)transcriptomics/ (meta)proteomics results, by showing the functions each taxon can perform, and by showing the gene/protein expression of the collective community. This tool is useful to visualize, in metabolic maps, the functions that microbial communities can perform (i.e., the genomic potential of microbial communities, obtained by metagenomics studies) and to identify the microorganisms inside the communities that can perform that function. In addition, KEGGCharter can represent gene expression, the result of (meta)transcriptomics and (meta)proteomics studies, by showing the metabolic functions expressed in heatmaps.

KEGGCharter accesses the following KEGG Pathway's functionalities: the interconversion of KEGG IDs, KOs, and EC numbers and the representation of taxonomy and gene expression in KEGG metabolic maps. To show the functionalities of KEGGCharter, MT datasets were simulated as described in SI, Section 1.5.2, to simulate a gene expression experiment. The TSV report obtained with UPIMAPI, together with gene expression quantification data obtained from MT analysis, was used as the input for KEGGCharter (Fig. 6). Comparative gene expression maps produced by KEGGCharter show the quantification of each function for the

Table 1
Results obtained from the analysis of the artificial metagenomes with UPIMAPI.

	Number of proteins and respective percentage [a]				
	Artificial metagenome #1	Artificial metagenome #2	Artificial metagenome #3	Artificial metagenome #4	Artificial metagenome #5
Genes identified after gene calling	23,283	22,817	26,920	28,908	41,511
Annotated proteins	22,489 (96.59 %)	22,397 (98.16 %)	26,634 (98.94 %)	27,912 (96.55 %)	40,865 (98.44 %)
Unique UniProt IDs	21,485 (92.28 %)	22,118 (96.94 %)	26,364 (97.93 %)	27,066 (93.63 %)	40,655 (97.94 %)
Uncharacterized proteins	4333 (18.61 %)	3608 (15.81 %)	4073 (15.13 %)	6129 (21.2 %)	7681 (18.5 %)
Proteins with assigned KEGG ID	2931 (12.59 %)	1654 (7.25 %)	3861 (14.34 %)	1102 (3.81 %)	1660 (4.0 %)
Proteins with assigned EC number	4205 (18.06 %)	4349 (19.06 %)	4602 (17.1 %)	4202 (14.54 %)	6434 (15.5 %)
Proteins with assigned OG	493 (2.12 %)	1318 (5.78 %)	5014 (18.63 %)	356 (1.23 %)	1185 (2.85 %)

[a] The percentages were calculated relatively to the number of proteins identified by gene calling.

Table 2
Number of proteins assigned to COG categories by reCOGnizer that were classified as "uncharacterized" by sequence-based annotation with UPIMAPI.

		# of proteins				
		Artificial metagenome #1	Artificial metagenome #2	Artificial metagenome #3	Artificial metagenome #4	Artificial metagenome #5
CELLULAR PROCESSES AND SIGNALING	Cell cycle control, cell division, chromosome partitioning	70	24	52	108	72
	Cell motility	29	9	7	46	22
	Cell wall/membrane/envelope biogenesis	81	52	41	171	99
	Cytoskeleton	9	0	0	5	14
	Defense mechanisms	41	38	27	65	42
	Extracellular structures	34	7	9	61	28
	Intracellular trafficking, secretion, and vesicular transport	20	13	14	40	22
	Mobilome: prophages, transposons	16	12	11	6	18
	Posttranslational modification, protein turnover, chaperones	51	19	17	73	53
	Signal transduction mechanisms	216	29	46	272	151
INFORMATION STORAGE AND PROCESSING	RNA processing and modification	0	1	0	0	1
	Replication, recombination and repair	23	9	11	23	31
	Transcription	43	43	33	66	79
	Translation, ribosomal structure and biogenesis	22	16	14	12	24
METABOLISM	Amino acid transport and metabolism	12	24	18	26	54
	Carbohydrate transport and metabolism	13	26	26	31	77
	Coenzyme transport and metabolism	28	27	14	38	27
	Energy production and conversion	26	21	10	41	40
	Inorganic ion transport and metabolism	35	19	31	35	68
	Lipid transport and metabolism	14	18	7	35	52
	Nucleotide transport and metabolism	4	3	7	11	17
	Secondary metabolites biosynthesis, transport and catabolism	3	3	30	12	33
No COG ID	–	3715	3192	3655	5303	6764
POORLY CHARACTERIZED	Function unknown	84	86	67	59	175
	General function prediction only	205	81	102	296	183

Table 3
Number of proteins annotated by reCOGnizer, with different reference databases, and respective percentages relatively to the total number of proteins in the artificial metagenomes, for E-value = 0.001.

		# of proteins annotated with reference to the databases included in reCOGnizer						
		CDD	NCBIfam	Protein Clusters	TIGRFAM	Pfam	Smart	COG
Artificial metagenome #1	Annotated proteins	11,119 (47.76 %)	4108 (17.64 %)	12,728 (54.67 %)	10,162 (43.65 %)	18,332 (78.74 %)	5362 (23.03 %)	17,292 (74.27 %)
	Proteins with assigned EC number	0 (0.0 %)	73 (0.31 %)	3795 (16.3 %)	2743 (11.78 %)	30 (0.13 %)	0 (0 %)	8199 (35.21 %)
Artificial metagenome #2	Annotated proteins	11,119 (48.73 %)	4108 (18.0 %)	12,728 (55.78 %)	10,162 (44.54 %)	18,332 (80.34 %)	5362 (23.5 %)	17,292 (75.79 %)
	Proteins with assigned EC number	0 (0.0 %)	73 (0.32 %)	3795 (16.63 %)	2743 (12.02 %)	30 (0.13 %)	0 (0 %)	8199 (35.93 %)
Artificial metagenome #3	Annotated proteins	11,119 (41.3 %)	4108 (15.26 %)	12,728 (47.28 %)	10,162 (37.75 %)	18,332 (68.1 %)	5362 (19.92 %)	17,292 (64.23 %)
	Proteins with assigned EC number	0 (0.0 %)	73 (0.27 %)	3795 (14.1 %)	2743 (10.19 %)	30 (0.11 %)	0 (0 %)	8199 (30.46 %)
Artificial metagenome #4	Annotated proteins	11,119 (38.46 %)	4108 (14.21 %)	12,728 (44.03 %)	10,162 (35.15 %)	18,332 (63.41 %)	5362 (18.55 %)	17,292 (59.82 %)
	Proteins with assigned EC number	0 (0.0 %)	73 (0.25 %)	3795 (13.13 %)	2743 (9.49 %)	30 (0.1 %)	0 (0 %)	8199 (28.36 %)
Artificial metagenome #5	Annotated proteins	11,119 (26.79 %)	4108 (9.9 %)	12,728 (30.66 %)	10,162 (24.48 %)	18,332 (44.16 %)	5362 (12.92 %)	17,292 (41.66 %)
	Proteins with assigned EC number	0 (0.0 %)	73 (0.18 %)	3795 (9.14 %)	2743 (6.61 %)	30 (0.07 %)	0 (0 %)	8199 (19.75 %)

entire community under different conditions. An example of KEGGCharter output with differential gene expression is given in SI (Fig. S1). Fig. 6 represents one of the maps obtained with KEGGCharter from the analysis of the real metagenome, after functional annotation with UPIMAPI.

In this metabolic map, it is possible to visualize which enzymes present in the metagenome are related to the methane metabolism pathway. In this metagenomics dataset, microorganisms assigned to *Methanosaeta concilii* and *Methanobacterium subterranum*, colored in orange and purple, are the ones with the higher genomic potential to produce methane, as several enzymes assigned to these microorganisms could be mapped by KEGGCharter in this metabolic map. Other microorganisms with less representation in the dataset (identified as "other taxa") also contain enzymes

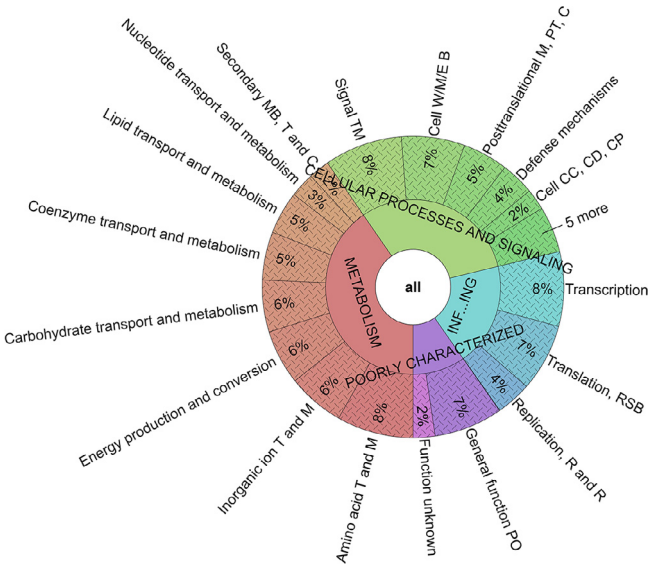

Fig. 5. Visual representation of COG categories in a Krona plot, obtained from the annotation of the artificial metagenome #3 with reCOGnizer. Percentages reflect the abundance of each COG category in the dataset. Cell CC, CD, CP represents Cell cycle control, cell division, chromosome partitioning; Cell W/M/E B represents Cell wall/membrane/envelope biogenesis; Posttranslational M, PT, C represents Post-translational modification, protein turnover, chaperones; Replication, R and R represents Replication, recombination and repair; Signal TM represents Signal transduction mechanisms; Translation, RSB represents Translation, ribosomal structure and biogenesis; Amino acid T and M represents Amino acid transport and metabolism; Inorganic ion T and M represents Inorganic ion transport and metabolism; General function PO represents General function prediction only; Secondary MB, T and C represents Secondary metabolites biosynthesis, transport and catabolism.

involved in the methane metabolism pathway. In addition, *Syntrophobacter fumaroxidans*, which is not a methanogen, has several enzymes identified in this map, that are not involved in methane production but participate in other reactions represented in the methane metabolism map.

3.3. Qualitative and quantitative benchmark

Several functional annotation tools were run to analyze the artificial metagenomes and the real metagenome. Regarding the artificial metagenomes, the number of genes annotated after gene calling by each one of the tools is presented in Table S11. UPIMAPI + reCOGnizer obtained functional annotation for 97.83 ± 0.93 % of the proteins, while 96.49 ± 1.70 % of the proteins were annotated with Mantis, 92.76 ± 2.21 % with eggNOG-mapper, 49.34 ± 3.65 % with Prokka and 40.63 ± 5.54 % with DFAST (Table S11). Functional annotation tools were compared regarding the assignment of EC numbers and OGs to the proteins present in the artificial metagenomes. Table 4 shows the average number (and standard deviation) of total EC numbers and OGs identified by each one of the tools, and respective quality metrics. Detailed results for the annotation of all datasets are given in Table S9. The highest F1 score for EC numbers' assignments was obtained with UPIMAPI + reCOGnizer (93.81 ± 1.26 %). For OGs assignments the highest F1 scores were obtained with eggNOG mapper and Mantis (89.47 ± 3.73 and 88.62 ± 3.10, respectively). These results show that the combination of UPIMAPI with reCOGnizer provides precise identifications for both OGs and EC numbers (precisions ranging between 89 % and 94 %, with the exception of OG assignment in artificial metagenome #1, which obtained a precision of 72 %), and assigns EC numbers to almost all proteins (recall higher than 95 %). On the other hand, assignment of OGs with UPIMAPI + reCOGnizer resulted in lower F1 scores (between 67 and 97 %) when compared with the other tools. eggNOG-mapper could assign more OG IDs than the remaining tools and presented a very low number of false negatives. Generally, regarding OG assignments, Mantis and eggNOG-mapper presented high F1 scores, but DFAST and Prokka showed F1 scores lower than 70 %. On the other hand, the F1 score obtained for the EC number assignment with Prokka was above 80 %, which was similar to the F1 scores obtained with Mantis and eggNOG-mapper (Table 4). Regarding the percentage of proteins that received a functional annotation beyond the information available at UniProt, UPIMAPI + reCOGnizer could retrieve the highest number of EC numbers (more than one third of the total number of proteins were assigned to an EC number) and eggNOG mapper assigned the most OGs (for over two-thirds of the total number of proteins). The concordance of the results obtained by the different annotation tools is presented in Table S10.

The performance of the annotation tools when analyzing the real MG dataset was similar to what was obtained with the artificial metagenomes, in respects to the number of functional assignments (Table 5). The highest number of proteins annotated was obtained with UPIMAPI + reCOGnizer (89 %), followed by Mantis (86 %), eggNOG-mapper (77 %), DFAST (38 %) and Prokka (27 %), but on the other hand, eggNOG-mapper could assign much more OGs (77 %), when compared with the other tools. Mantis assigned OGs to over 50 % of the proteins, UPIMAPI + reCOGnizer to 38 %, and DFAST and Prokka to lower percentages (28 % and 17 %, respectively) (Table 5). The highest number of EC numbers assigned was obtained by UPIMAPI + reCOGnizer (34 %), followed closely by Mantis (25 %) and eggNOG-mapper (24 %), and then by Prokka and DFAST with 17 % and 9 %, respectively.

All tools assigned more OG IDs than EC numbers, except Prokka that assigned OGs and EC numbers to approximately the same number of proteins (Tables 4 and 5). Generally, all tools could annotate a higher percentage of proteins from the artificial metagenomes (ranging from 99 and 35 %, depending on the tool, Table S11) than from the real metagenome (ranging from 89 and 27 %, depending on the tool, Tables 5).

4. Discussion

Functional annotation of *meta*-omics datasets is essential for understanding microbial behavior, and linking microbial identity to function in microbial communities. The combination of sequence-based functional annotation with domain-based functional annotation has been shown to provide a more complete picture of the data analyzed [25], and several tools have implemented this approach [9–12]. This work contributed to the improvement of the state of the art, by developing two novel command-line tools performing sequence (UPIMAPI) and domain (reCOGnizer) homology-based functional annotation of big datasets, and a tool to visualize the annotation results in metabolic maps (KEGGCharter). These tools offer several advantages for the end-users. For instance, they are fully automated, as they are run by using a single command. Also, they offer many options for customization, e.g., the choice of different reference databases and the content of the output information. In addition, functional annotation is obtained from 9 databases (UniProt, CDD, COG, KOG, Pfam, SMART, Protein Clusters, NCBIfam and TIGRFAM), providing a complete functional characterization. Both UPIMAPI and reCOGnizer are run with the most recent versions of these databases, providing up-to-date information. The annotation results of both metagenomics and gene expression analysis can be further visualized by using KEGGCharter, thus facilitating the interpretation of *meta*-omics results.

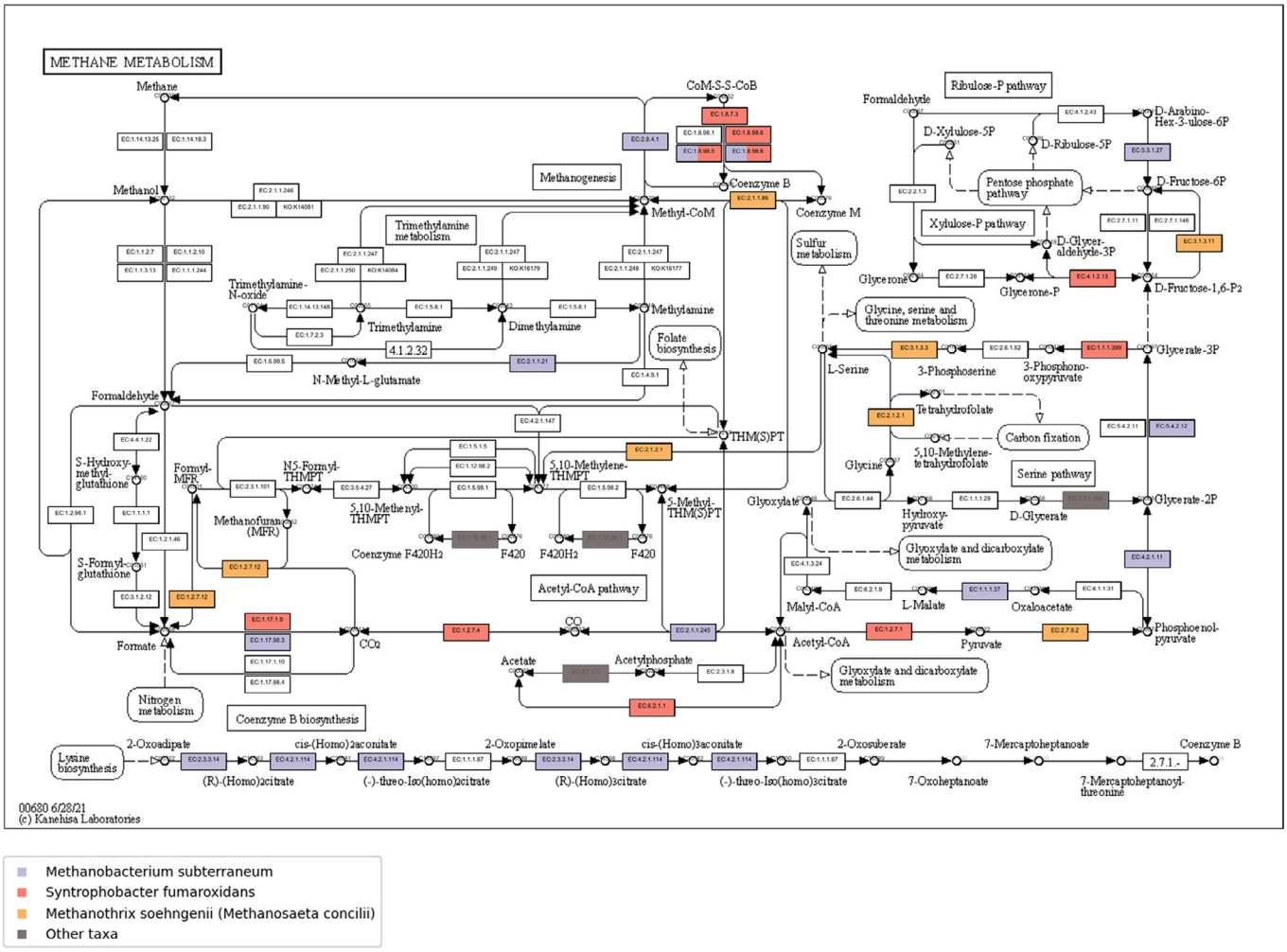

Fig. 6. Example of a KEGGCharter output map obtained from the analysis of the real metagenome. This map represents the genomic potential of the community for "Methane metabolism" (map00680 of KEGG Pathway). Enzymes assigned to different taxa are represented in different colors in the enzyme boxes, with the legend under the map identifying the taxa corresponding to each color. KEGGCharter was run from the results of UPIMAPI by setting the "input" parameter to the "UPIMAPI_results.tsv" table outputted by UPIMAPI. KEGG IDs column was set to "Cross-reference (KEGG)", taxa column set to "Taxonomic lineage (SPECIES)" and the input quantification parameter was set to true.

UPIMAPI and reCOGnizer offer some unique features for functional annotation. UPIMAPI retrieves information from the UniProt KB, which includes functional and taxonomic information from all proteins present in the UniProt database. It automatically builds customized databases for inputted tax IDs, which may be useful for studies relying only on a limited number of taxonomic groups. UPIMAPI could annotate a high percentage of proteins from the artificial (98 ± 0.9 %) and real (89 %) metagenomes (Table 1, 5 and Table S9), presenting high recall and sensitivity. Nevertheless, a significant percentage (18 ± 2.4 %) of the proteins annotated by UPIMAPI were uncharacterized proteins, for which functional information is not available in the UniProt database. The analysis with reCOGnizer decreased the number of proteins without information, thus increasing the functional characterization of the artificial metagenomes (Tables 2, 4 and 5 and Table S9). This happened because reCOGnizer uses different methodologies for annotation, i.e., sequence-based annotation with UPIMAPI is dependent on the availability of information about homologous sequences, while domain-based annotation with reCOGnizer tolerates less sequence homology, focusing on small but biologically relevant similarities [4]. These results reinforce the advantage of combining both methods for functional annotation.

When comparing the annotation metrics obtained after analysis of the artificial metagenomes and the real metagenome with different tools, UPIMAPI + reCOGnizer obtained the highest F1 score and highest number of functional assignments regarding EC numbers (Tables 4 and 5). This better performance may be related to the differences between UPIMAPI + reCOGnizer and the remaining tools such as the nature and number of the reference databases. UPIMAPI and reCOGnizer annotate with reference to the entire UniProt database (including SwissProt and TrEMBL), CDD, COG, KOG, Pfam, SMART, Protein Clusters, NCBIfam and TIGRFAM, while eggNOG-mapper uses only the eggNOG database, Mantis uses the eggNOG database, NCBIfam, KOfam, Pfam and offers the possibility to add customized databases (in this comparison the UniProt database was included), Prokka uses SwissProt, some reference genomes from RefSeq, Pfam and TIGRFAM, and finally, DFAST uses a database composed of 124 reference genomes, TIGRFAM and COG. While some tools were better at obtaining EC numbers (UPIMAPI and Prokka), others were better suited for OG annotation (eggNOG-mapper and Mantis). OG annotation with UPIMAPI was particularly affected, since the sequences in the artificial metagenomics datasets were removed from the reference database, to ensure that the benchmark data is independent from the reference

Table 4
Quality metrics of annotation of the artificial metagenomes, regarding the assignment of EC numbers and OGs, with UPIMAPI + reCOGnizer, Mantis, eggNOG-mapper, Prokka and DFAST. Only the proteins with EC number/OG assigned in the UniProt database were considered for this comparison, except for the last column, where it is shown the total number of proteins that were assigned OG/EC numbers. True positives (TP), false positives (FP), false negatives (FN), precision, recall, and F1 score were calculated as described in SI, Section 1.6. Because multiple iterations were performed, TPs, FPs and FNs were normalized by dividing by the total number of proteins with functional information (the sum of TPs. FPs and FNs).

Qualifier	Tool	TPs (%)	FPs (%)	FNs (%)	Precision (%)	Recall (%)	F1 score (%)	% of total identifications
EC number	UPIMAPI + reCOGnizer	**88.37** (±2.25)	**7.9** (±1.63)	**3.73** (±0.86)	**91.78** (±1.73)	**95.94** (±0.97)	**93.81** (±1.26)	**34.79** (±6.75)
	Prokka	**72.2** (±6.25)	**14.79** (±2.99)	**13.0** (±3.3)	**82.84** (±4.24)	**84.58** (±4.57)	**83.7** (±4.39)	**27.14** (±3.4)
	Mantis	**66.94** (±3.34)	**22.25** (±1.66)	**10.82** (±2.08)	**75.02** (±2.26)	**86.05** (±2.87)	**80.15** (±2.4)	**26.03** (±5.36)
	eggNOG-mapper	**65.96** (±2.86)	**17.9** (±1.0)	**16.13** (±2.0)	**78.62** (±1.62)	**80.32** (±2.62)	**79.46** (±2.07)	**20.22** (±4.18)
	DFAST	**59.18** (±8.72)	**9.06** (±1.1)	**31.76** (±9.55)	**86.62** (±1.15)	**65.16** (±10.03)	**73.95** (±7.36)	**14.75** (±3.03)
OG	eggNOG-mapper	**81.15** (±5.85)	**18.84** (±5.84)	**0.02** (±0.02)	**81.16** (±5.85)	**99.98** (±0.02)	**89.47** (±3.73)	**71.19** (±13.72)
	Mantis	**79.71** (±4.92)	**11.35** (±5.67)	**8.94** (±2.2)	**87.61** (±5.97)	**89.96** (±2.16)	**88.62** (±3.1)	**59.9** (±11.22)
	UPIMAPI + reCOGnizer	**66.32** (±15.11)	**8.14** (±5.94)	**25.54** (±12.19)	**88.52** (±8.78)	**71.79** (±13.57)	**78.83** (±10.13)	**48.8** (±9.17)
	DFAST	**53.37** (±6.28)	**7.86** (±4.64)	**38.77** (±3.64)	**87.07** (±7.92)	**57.77** (±4.87)	**69.37** (±5.51)	**48.92** (±6.51)
	Prokka	**36.33** (±4.55)	**3.75** (±1.73)	**59.93** (±2.92)	**90.28** (±5.47)	**37.67** (±4.17)	**53.12** (±5.1)	**30.09** (±5.39)

Table 5
Comparison of the results obtained with the functional annotation tools after analysis of the real metagenomics dataset, regarding the number of proteins annotated and the number of assigned EC numbers and OG IDs. Percentages were calculated with reference to the total number of genes obtained after gene calling.

Tool	# of proteins annotated	# of proteins with EC number assigned	# of proteins with OG assigned
UPIMAPI + reCOGnizer	434,310 (88.77 %)	165,602 (33.85 %)	184,633 (37.74 %)
Mantis	419,527 (85.75 %)	123,960 (25.34 %)	248,109 (50.71 %)
eggNOG mapper	375,654 (76.78 %)	117,355 (23.99 %)	375,654 (76.78 %)
DFAST	43,726 (37.99 %)	10,318 (8.96 %)	32,540 (28.27 %)
Prokka	70,302 (27.31 %)	44,202 (17.17 %)	43,435 (16.87 %)

database. Therefore, the functional assignment was obtained from the proteins of closely related species. Because artificial metagenomes # 1 to 5 contain different microorganisms, with different closely related species, this explains the different results obtained for each one of the artificial metagenomes, i.e., the datasets containing species closely related to well characterized microorganisms obtained better annotation results with UPIMAPI. Since the sequences of the organisms studied were removed from UniProt, the results for OGs were mostly based on the results from reCOGnizer, which had consistently less recall than UPIMAPI (Table S9).

Generally, the tools with best performance in the functional annotation of the artificial metagenomes, were also the ones identifying the highest number of proteins, and attributing the highest number of EC numbers and OGs, in the real metagenome (Tables 4 and 5). Nevertheless, the results show that when analyzing real metagenomes, containing several unknown and uncharacterized proteins, it is important to use the broadest and most complete databases, as it is the case of the UniProt KB and eggNOG. For instance, UPIMAPI + reCOGnizer identified the highest number of EC numbers while eggNOG-mapper retrieved the highest number of OGs (Table 5). Thus, different tools may be the best choice depending on the objective of the study. UPIMAPI + reCOGnizer may be the best choice if further analysis is necessary, such as mapping metabolic functions or developing metabolic models, as the combination of these tools resulted in the highest number of proteins with EC numbers assigned, independently of the dataset used (Tables 4 and 5). On the other hand, a deeper study on the identification of OGs would benefit from the utilization of eggNOG-mapper or Mantis. Note that the performance of the annotation with Mantis was very close to the best performing tools, i.e., it presented high F1 scores for OGs and EC numbers assignments (Table 4), and annotated a high number of proteins from the real metagenome (Table 5). In addition, Mantis determines a consensus identification by using text mining and a hierarchical organization of the databases, which might be useful for certain studies [12]. The results obtained with UPIMAPI + reCOGnizer could have been further improved if the reference database used for the annotation contained only the proteins from the microorganisms present in the dataset, instead of using the entire UniProt database. This can be done by specifying the taxon IDs when constructing the reference database (by using the "--taxids" option), and may be useful when the microbial composition of the communities is previously known. However, from a practical viewpoint, metagenomics data are usually annotated with the entire UniProt database, since typically the taxonomic composition is unknown, and thus these results better reflect what can be obtained with real metagenomics data analyses. Results with reCOGnizer could also be improved by inputting taxonomic information with the "--tax-file" parameter (this information can be obtained for example from UPIMAPI's output).

A major difficulty in *meta*-omics studies is the interpretation of the data. To overcome this limitation, it is important that the outputs from the bioinformatics analysis are easily handled, organized and that can be easily visualized. KEGGCharter represents an alternative to already existing mapping options, offering the possibility to represent in metabolic maps both taxonomic assignment and differential gene expression up to 10 samples. The representation of differential expression for many samples can also be obtained through KEGGprofile, but it requires extensive input through R commands [20]. KEGG Mapper is another tool with similar purposes to KEGGCharter, but requires much more input from the users, and does not perform the automatic representation of

different taxonomies and differential gene expression results in metabolic maps. Still, KEGG Mapper provides extra information, for instance, it indicates gene IDs next to the corresponding enzyme boxes [15].

Visualization with KEGGCharter can still be improved by, for example, producing more interactive maps or combining taxonomy and expression levels in the same plot. However, the main limitation of applying KEGGCharter, and other mapping tools, to datasets containing poorly characterized proteins, is the quality of the input data, e.g., a protein without functional annotation in knowledge bases will not be mapped. This results in an underestimation of the functions that can be performed by a microbial community when visualizing the data in metabolic maps. Therefore, as functional annotation evolves, more reliable pathway enrichment results can be visualized. In this context, UPIMAPI + reCOGnizer brought a very important contribution, as they retrieve a high number of EC numbers and KEGG IDs (Tables 1, 3 and 4). But there is room for further improvements. Annotation with reCOGnizer could be expanded by parsing more information from the domain descriptions, such as EC numbers being obtained from CDD and SMART. Additionally, a methodology to reach a consensus annotation could be implemented, as it was done in other tools [12]. However, obtaining as much information as possible might be advantageous depending on the studies' objectives and requirements. Some users may want to use only the COG database because it is a known and trusted resource in bioinformatics and other users may want to annotate with the UniProt database only. For studies on single proteins, running a tool that obtains different matches in different databases may provide a higher level of trust in the results. ID mapping with UPIMAPI could also be improved, since it is still largely dependent on UniProt's API for mapping TrEMBL IDs, and this information could potentially be explored with a fully local method. Nevertheless, local TrEMBL ID mapping would require the storage of ~1 Tb of data, which could be a problem for the users who have no access to servers with enough space.

5. Conclusions

In this work, we present three new command-line tools, tailored for retrieving the maximum amount of information from protein sequences. reCOGnizer and UPIMAPI together compose a powerful approach for protein annotation by providing complementary functional information, which can then be represented in metabolic pathways by using KEGGCharter. A great advantage of these new tools is that they are fully automated, requiring minimum input from the user. The three tools can be easily installed through Bioconda.

Data availability

All the scripts and commands used in this paper are available at https://github.com/iquasere/annotation_paper.

The real dataset analysed in this paper was deposited in ENA database under the study accession PRJEB50269 (https://www.ebi.ac.uk/ena/browser/view/PRJEB50269).

CRediT authorship contribution statement

João C. Sequeira: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft, Data curation, Resources. **Miguel Rocha:** Funding acquisition, Resources, Software, Writing – review & editing, Supervision. **M. Madalena Alves:** Funding acquisition, Resources, Writing – review & editing, Supervision. **Andreia F. Salvador:** Conceptualization, Methodology, Investigation, Writing – review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research leading to these results has received funding from the European Union, Horizon 2020 innovation action programme under grant agreement No 952908—GLOMICAVE project.

This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/04469/2020 unit and the FCT grant SFRH/BD/147271/2019 (attributed to João Sequeira).

We would also like to acknowledge the contribution of Tiago Oliveira in the development of KEGGCharter.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.03.042.

References

[1] Hernández-Salmerón JE, Moreno-Hagelsieb G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genomics 2020;21:741. https://doi.org/10.1186/s12864-020-07132-6.
[2] The UniProt Consortium. UniProt: A hub for protein information. Nucleic Acids Res 2015;43:D204–12. https://doi.org/10.1093/nar/gku989.
[3] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.
[4] Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: The COG approach. Brief Bioinform 2019;20:1063–70. https://doi.org/10.1093/bib/bbx117.
[5] De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 2012;13:696–710. https://doi.org/10.1093/bib/bbs070.
[6] Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 2020;48:D265–8. https://doi.org/10.1093/nar/gkz991.
[7] Marchler-Bauer A, Bryant SH. CD-Search: Protein domain annotations on the fly. Nucleic Acids Res 2004;32:327–31. https://doi.org/10.1093/nar/gkh454.
[8] Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: A customizable web server for fast metagenomic sequence analysis. BMC Genomics 2011;12:444. https://doi.org/10.1186/1471-2164-12-444.
[9] Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 2018;34:1037–9. https://doi.org/10.1093/bioinformatics/btx713.
[10] Prokka ST. Rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153.
[11] Cantalapiedra CP, Hern Andez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021; msab293:1–5. 10.1093/molbev/msab293.
[12] Queirós P, Delogu F, Hickl O, May P, Wilmes P. Mantis: flexible and consensus-driven genome annotation. GigaScience 2021;10:1–14. https://doi.org/10.1093/gigascience/giab042.
[13] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62. https://doi.org/10.1093/nar/gkv1070.
[14] Klukas C, Schreiber F. Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics 2007;23:344–50. https://doi.org/10.1093/bioinformatics/btl611.
[15] Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci 2020;29:28–35. https://doi.org/10.1002/pro.3711.
[16] Darzi Y, Letunic I, Bork P, Yamada T. IPath3.0: Interactive pathways explorer v3. Nucleic Acids Res 2018;46:W510–3. https://doi.org/10.1093/nar/gky299.
[17] Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, et al. Pathway Projector: Web-Based Zoomable Pathway Browser Using KEGG Atlas and Google Maps API. PLoS ONE 2009;4:. https://doi.org/10.1371/journal.pone.0007710e7710.

[18] Elliott B, Kirac M, Cakmak A, Yavas G, Mayes S, Cheng E, et al. PathCase: Pathways database system. Bioinformatics 2008;24:2526–33. https://doi.org/10.1093/bioinformatics/btn459.
[19] Puente-Sánchez F, Garciá-Garciá N, Tamames J. SQMtools: Automated processing and visual analysis of 'omics data with R and anvi'o. BMC Bioinf 2020;21:358. https://doi.org/10.1186/s12859-020-03703-2.
[20] Zhao S. KEGGprofile: Application Examples. https://bioconductor.riken.jp/packages/3.2/bioc/vignettes/KEGGprofile/inst/doc/KEGGprofile.pdf. 2020.
[21] Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition-Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol 2016;12:. https://doi.org/10.1371/journal.pcbi.1004957e1004957.
[22] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
[23] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402. https://doi.org/10.1093/nar/25.17.3389.
[24] The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45:D158–69. https://doi.org/10.1093/nar/gkw1099.
[25] Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol 2009;10:207. https://doi.org/10.1186/gb-2009-10-2-207.