
EMG-docs Documentation

Release 1.0

SCP

Nov 13, 2017

Contents:

1	About	1
1.1	The EBI Metagenomics service	1
1.2	What does EBI Metagenomics offer	1
1.3	Funding	1
1.4	How to cite	2
1.5	Contact	2
2	Data flow from submission to results	3
3	Analysis pipeline v4.0	5
3.1	Overview	5
3.2	Taxonomic analysis	6
3.2.1	Other non-coding RNAs	7
3.3	Functional analysis	7
3.3.1	Protein signatures	7
3.3.2	Assigning GO terms to metagenomic sequences	8
4	Website and portal	11
4.1	Content of the ‘Associated runs’ table on project page	11
4.2	Finding quality control information about runs on the EBI Metagenomics website	11
4.3	Finding functional information about runs on the EBI Metagenomics website	13
4.4	Finding taxonomic information about runs on the EBI Metagenomics website	14
4.5	Files available to download on the EBI Metagenomics website	15
4.5.1	Description of fasta files available to download	17
4.5.2	Description of functional annotation files available to download	17
4.5.3	Description of taxonomic assignment files available to download	18
4.6	Summary files	18
4.6.1	functional summary files	18
4.6.2	taxonomy summary files	19
4.7	Comparison tool	19
4.8	Data discovery on EBI Metagenomics portal	20
4.8.1	Search tool	20
4.8.2	Browsing options	21
4.9	Privat area	21
5	Guides and Tutorials	23
5.1	EBI Metagenomics online tutorials	23

5.2	ENA online guides	23
6	FAQs	25
6.1	What kind of sequence data does the service accept?	25
6.2	Can I submit assembled metagenomic sequences for analysis?	25
6.3	Can I submit 18S rRNA or ITS amplicon sequences?	25
6.4	Can I submit viral sequences?	25
6.5	How do I run a BLAST search against the metagenomics datasets?	26
6.6	Can I change the release date of my project?	26
6.7	How long will it take for my data to be analyzed?	26
6.8	I have submitted my data - how do I trigger the analysis?	26
6.9	Do you have an API?	26
6.10	How can I download several sets of data?	26
6.11	How can I bulk download meta-data?	26
6.12	How can I re-analyse my data with a different version of the pipeline?	27
6.13	Can I request that a dataset is analyzed if I am not the original submitter?	27
6.14	Can I request my data to not be analyzed by EBI Metagenomics?	27
6.15	Can I compare the taxonomic assignments between runs of a project?	27
6.16	Can I know which bacteria encodes particular pCDS in my dataset?	27
7	Indices and tables	29

1.1 The EBI Metagenomics service

EBI metagenomics is a freely available hub for the analysis and exploration of metagenomic, metatranscriptomic, amplicon and assembly data. The resource provides rich functional and taxonomic analyses of user-submitted sequences, as well as analysis of publicly available metagenomic datasets held within the European Nucleotide Archive ([ENA](#)).

1.2 What does EBI Metagenomics offer

Standards-compliant data The service provides submission tools that help ensure sequence data and metadata comply with the European Nucleotide Archive (ENA) data schemes and the Genomic Standards Consortium ([GSC](#)) metadata guidelines, allowing harmonisation of analysis efforts across the wider genomics community.

Powerful analysis The service provides powerful taxonomic and functional analysis of sequence data. Functional analysis results can be summarized and compared using Gene Ontology ([GO](#)) terms.

Data Archiving Data submitted to the service is automatically archived in the ENA. Accession numbers are supplied for sequence data as part of the archiving process, which is a prerequisite for publication in many journals.

1.3 Funding

EBI Metagenomics was initiated with funding from EMBL. It continues to be developed with EMBL support and funding gratefully received from the Biotechnology and Biological Sciences Research Council (BBSRC grant BB/M011755/1) and [ELIXIR GRANT INFO]. It previously received funding from BBSRC grant BB/I02612X/1 and the EU's Seventh Framework Programme for Research (FP7 grant MICROB3).

1.4 How to cite

If you find this resource useful, please consider citing the following publication:

EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data.
A Mitchell, F Bucchini, G Cochrane, H Denise, P ten Hoopen, M Fraser, S Pesseat, S Potter, M Scheremetjew, P
Serk and R D Finn.
Nucleic Acids Research (2016) Database Issue 44:D595-D603. [10.1093/nar/gkv1195](https://doi.org/10.1093/nar/gkv1195) (PDF)

1.5 Contact

If you would like to get in touch, please contact us using the [helpdesk](#). We will respond as quickly as possible, but please bear in mind that we do not have a member of staff dedicated to running the helpesk service.

Please, check our [FAQ](#) prior to contacting us.

If you would like to keep up to date with developments with the EBI Metagenomics, please follow us on Twitter ([@EBImetagenomics](#)).

Data flow from submission to results

The graph below summarize the EBI Metagenomics data flow from submission to analysis results:

(1) Submissions are handled by the [European Nucleotide Archive \(ENA\)](#) and therefore users have to have an [ENA Webin account](#) .

In addition, users submitting private data have to provide an expressed agreement that EBI Metagenomics can access their data for analysis, as described [here](#). Otherwise, we will not be able to access their data. EBI Metagenomics will, of course, handle these data confidentially.

(2) Access to the [ENA submission page](#) requires login in using a registered email address or a Webin identifier (Webin-XXXX).

(3 and 4): upload and submission.

These steps are described in details in the [ENA online guides](#). EBI Metagenomics is providing a step by step guide to submission ([EBI Metagenomics online tutorials](#)). Please also check our [FAQs](#).

Note that all queries concerning data submission should be directed to [ENA dedicated help desk](#)

Following completion of these two steps, and after data validation by ENA, we will be able to access the submitted data and they will be queued for analysis (more details [here](#)).

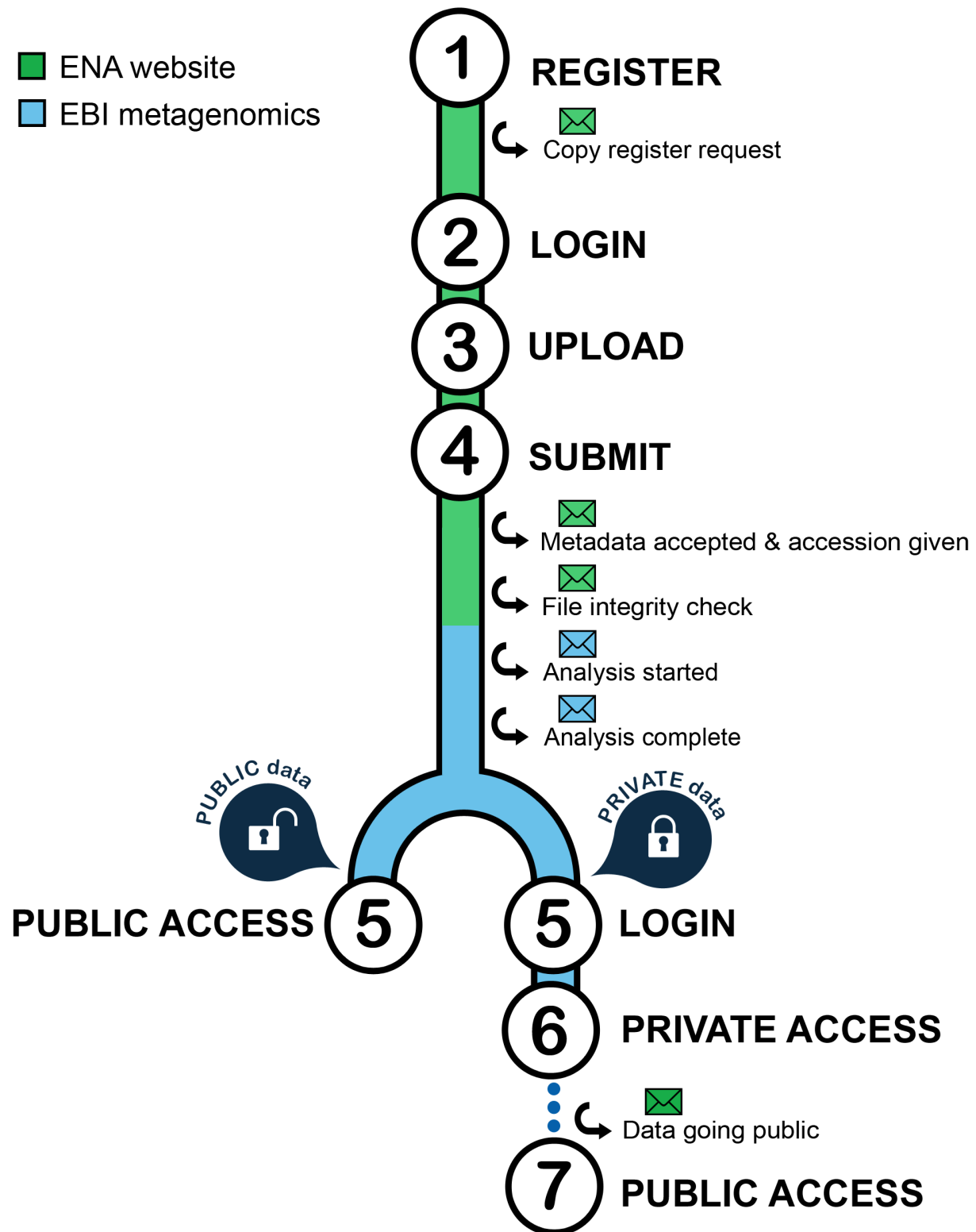
The length of time requires for analysis varies according to the number of projects in the queue, the nature and the number of runs in the submission however we aim to have most analysis completed in less than a week once validated by ENA.

(5) Upon completion of analysis, data will be uploaded on the website

EBI Metagenomics pipeline will generate a number of charts and downloadable files (fully described [Files available to download on the EBI Metagenomics website](#)).

(6) For private data, users will have to login on the EBI Metagenomics website to access their data until they become public

(7) Private data will become public after an intital confidential period of two years. Submitters will receive an email from ENA prior to public release giving them the opportunity to extend the confidential period which is set to two years per default (as indicated at [Can I change the release date of my project?](#)).



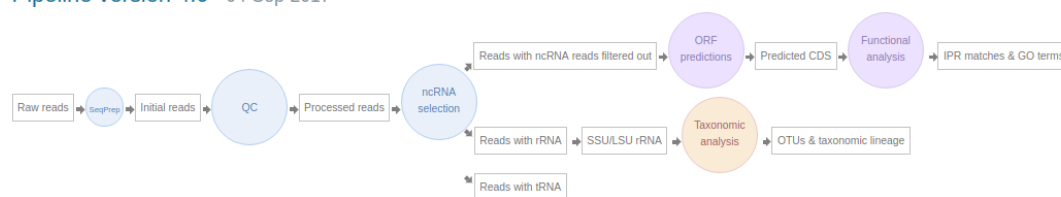
3.1 Overview

Version 4.0 of the pipeline was released in September 2017 and includes the following updates and changes:

- Updated tools: InterProScan to version 5.25-64.0
- rRNASelector (used to identify 16S rRNA genes) was replaced with Infernal for SSU and LSU gene identification
- The QIIME taxonomic classification component was replaced with MAPseq
- The Greengenes reference database was replaced with SILVA SSU / LSU version 128, enabling classification of eukaryotes, remapped to a 8-level taxonomy
- Prodigal was added to run alongside FragGeneScan as part of a combined gene caller when processing assembled sequences

Figure 1 gives a visual overview of the main steps and tools included in this version:

Pipeline version 4.0 - 04-Sep-2017



Pipeline tools & steps

Tools	Version	Description	How we use it
1 SeqPrep	1.1	A program to merge paired end Illumina reads that are overlapping into a single longer read.	Paired-end overlapping reads are merged - we do not perform assembly.
2.1 Trimmomatic	0.35	A flexible read trimming tool.	Low quality trimming (low quality ends and sequences with > 10% undetermined nucleotides removed). Adapter sequences removed using Biopython SeqIO package.
2.2 Biopython	1.65	A set of freely available tools for biological computation written in Python.	Sequences < 100 nucleotides in length removed.
3.1 Infernal	1.1.2	Infernal ("INference of RNA Alignment") is for searching DNA sequence databases for RNA structure and sequence similarities. It is an implementation of a special case of profile stochastic context-free grammars called covariance models (CMs). A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus, so in many cases, it is more capable of identifying RNA homologs that conserve their secondary structure more than their primary sequence.	Identification of ncRNAs.
3.2 cmsearch deoverlap script	1.0	A tool, which removes lower scoring overlaps from cmsearch --tblout files.	Removes lower scoring overlaps from cmsearch --tblout files.
4.1 FragGeneScan	1.20	An application for finding (fragmented) genes in short reads.	Run as a combined gene caller component, giving priority to Prodigal predictions in the case of assembled sequences or FragGeneScan for short reads (all predictions from the higher priority caller are used, supplemented by any non-overlapping regions predicted by the other).
4.2 Prodigal	2.6.3	Prodigal (Prokaryotic Dynamic Programming Gene-finding Algorithm) is a microbial (bacterial and archaeal) gene finding program.	
5 InterProScan	5.25-64.0	A sequence analysis application (nucleotide and protein sequences) that combines different protein signature recognition methods into one resource.	Matches are generated against predicted CDS, using a sub set of databases (Pfam, TIGRFAM, PRINTS, PROSITE patterns, Gene3d) from InterPro release 64.0. A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided. It is generated using a reduced list of GO terms called GO slim (version goslim_goa).
6 MAPseq	1.2	MAPseq is a set of fast and accurate sequence read classification tools designed to assign taxonomy and OTU classifications to ribosomal RNA sequences.	SSU and LSU rRNA are annotated using SILVA's SSU/LSU version 128 reference database, enabling classification of eukaryotes, remapped to a 7-level taxonomy.

Figure 1. Overview of steps and tools included in pipeline v4.0

3.2 Taxonomic analysis

The analysis pipeline underwent a substantial update in August 2017 to version 4.0, with the entire taxonomic profiling section replaced. The [rRNASelector](#) based component, which was previously used to identify 16S rRNA genes, was replaced with [Infernal](#) (running in hmm-only mode) using a library of ribosomal RNA hidden Markov models from [Rfam](#) 12.2. This allows accurate identification of both large and small subunit (LSU and SSU) ribosomal ribonucleic acid genes, including the eukaryotic 18S rRNA gene. To identify those we use the families found in the following clans: CL00111 (SSU) and CL00112 (LSU).

The QIIME taxonomic classification component was replaced with [MAPseq](#) version 1.2, which offers fast and accurate classification of reads, and provides corresponding confidence scores for assignment at each taxonomic level. The Greengenes reference database was replaced with [SILVA](#) SSU / LSU version 128, which includes eukaryotic as well as prokaryotic sequences, thus enabling eukaryotic taxonomic classification. In order to make it compatible with MAPseq, the SILVA database was remapped to a flat, 8-level taxonomy, using in house scripts. The resulting classification system was compared to QIIME/Greengenes and benchmarked using both mock community and real world datasets to confirm accuracy of results.

3.2.1 Other non-coding RNAs

In addition to the ribosomal subunit RNAs we also identify other non-coding RNAs (ncRNAs) such as SRP RNA, tRNA, tmRNA and RNase. The following clans are used for the ncRNAs: CL00001 (tRNA), CL00002 (RNase P) and CL00003 (SRP RNA).

3.3 Functional analysis

Functional analysis of predicted coding sequences (pCDS) from metagenomic data is provided using the [InterProScan](#) software provided by the [InterPro](#) database. InterPro is a sequence analysis resource that predicts protein family membership, along with the presence of important domains and sites. It does this by combining predictive models known as protein signatures from a number of different databases into a single searchable resource. InterPro curators manually integrate the different signatures, provide names and descriptive abstracts and, whenever possible, add Gene Ontology (GO) terms.

3.3.1 Protein signatures

Protein signatures are obtained by modelling the conservation of amino acids at specific positions within a group of related proteins (i.e., a protein family), or within the domains/sites shared by a group of proteins. InterPro's different member databases use different computational methods to produce protein signatures, and they each have their own particular focus of interest: structural and/or functional domains, protein families, or protein features, such as active sites or binding sites (see below).

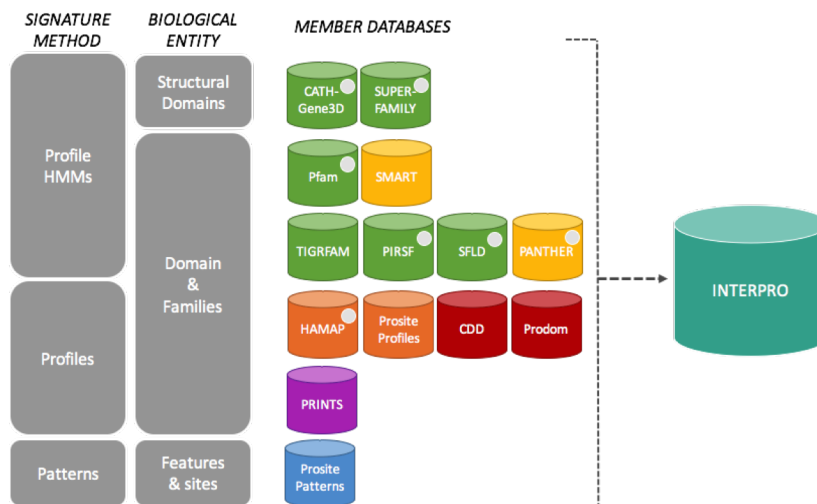


Fig. 3.1: InterPro member databases grouped by the methods used to construct their signatures and focus of interest.

Only a subset of the InterPro member databases are used by EBI Metagenomics: Gene3D, TIGRFAMs, Pfam, PRINTS and PROSITE patterns. These databases were selected since, together, they provide both high coverage and offer detailed functional analysis, and have underlying algorithms that can cope with the vast amounts of fragmentary sequence data found in metagenomic datasets.

3.3.2 Assigning GO terms to metagenomic sequences

While InterPro matches to metagenomic sequence sets are informative in their own right, EBI Metagenomics offers an additional type of annotation in the form of Gene Ontology (GO) terms.

The GO is made up of 3 structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. By using GO terms, scientists working on different species or using different databases can compare datasets, since they have a precisely defined name and meaning for a particular concept.

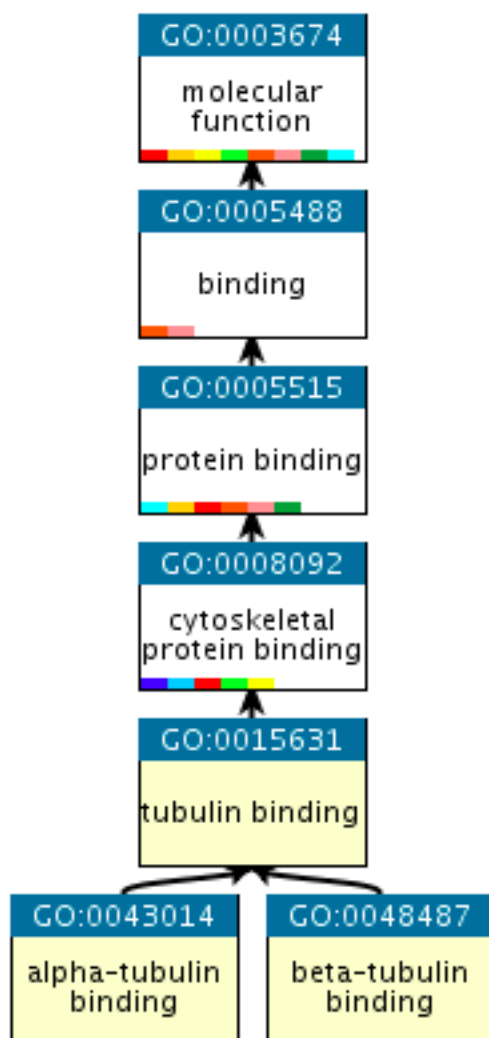


Fig. 3.2: An example of GO terms organised into a hierarchy.


Terms in the GO are ordered into hierarchies, with less specific terms towards the top and more specific terms towards the bottom. (e.g., alpha-tubulin binding is a type of cytoskeletal binding, which is a type of protein binding). Note that a GO term can have more than one parent term. The Gene Ontology also allows for different types of relationships between terms (such as ‘has part of’ or ‘regulates’). The EMG analysis pipeline only uses the straightforward ‘is a’ relationships. More information about the GO can be found on the GO consortium [documentation page](#).

As part of the metagenomic analysis pipeline, GO terms for molecular function, biological process and cellular component are assigned to pCDS in a sample by via the InterPro2GO mapping service. This works as follows: InterPro

entries are given GO terms by curators if the terms can be accurately applied to all of the proteins matching that entry. Sequences searched against InterPro are then associated with GO terms by virtue of the entries they match - a protein that matches one InterPro entry with the GO term 'kinase activity' and another InterPro entry with the GO term 'zinc ion binding' will be annotated with both GO terms.

4.1 Content of the ‘Associated runs’ table on project page

This table lists all samples and runs associated with a project as well as the experiment type (‘Amplicon’, ‘Assembly’, ‘Metabarcoding’, ‘Metagenomic’ or ‘Metatranscriptomic’), sequencing instrument model and pipeline version for each individual run. In addition, the last field displays links to analysis results and download pages (the latter being

represented by the  icon).

The ‘Analysis results’ field could also displays two types of messages:

- ‘QC not passed’: this message indicates that no sequences survived the filtering occurring during the QC steps. This could be due to base quality filtering, ambiguous base filtering or length filtering.
- ‘Unable to process’: this message indicates that no data suitable for analysis were available for this run. The sequences may not be available in ENA, failed to merge, in the case of pair-end reads, or be in an unsuitable format.

4.2 Finding quality control information about runs on the EBI Metagenomics website

Quality control (Qc) analysis of runs within projects on the [EBI Metagenomics website](#) can be accessed by selecting the ‘Quality control’ tab found toward the top of any run page (see Figure 1 below).

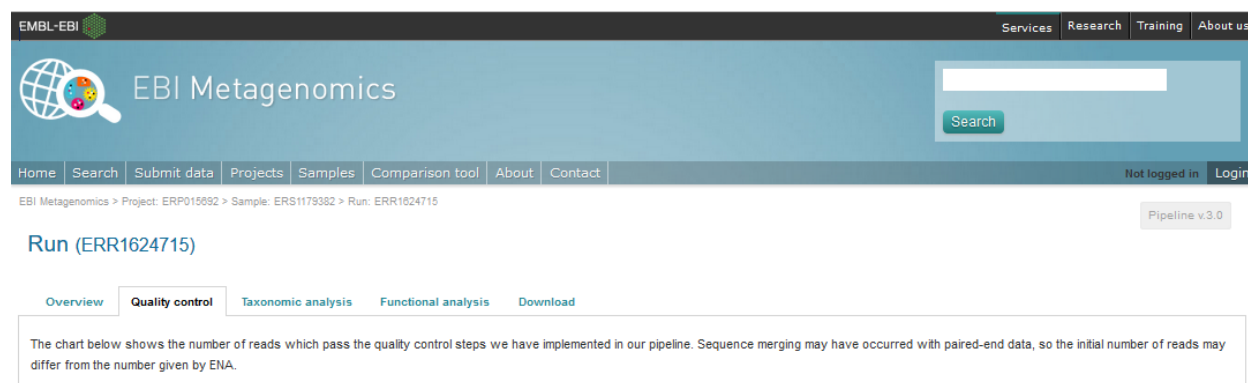


Figure 1. A ‘Quality control’ tab can be found towards the top of each run page.

Selecting this tab brings a page containing four graphical representations giving the number of reads remaining, and filtered, after each QC step as well as the length, GC content and nucleotide distributions of the reads having passed the QC referred as processed reads. These are available to download via the ‘Download’ tab found toward the top of any run page (see Figure 8 below).

An histogram is used to represent the nucleotide distribution for the first 500 nucleotides of the processed reads; while metagenome, metatranscriptome and assembly chart should indicate an even distribution (Figure 2 below), amplicon graph should indicate a clearly uneven pattern.

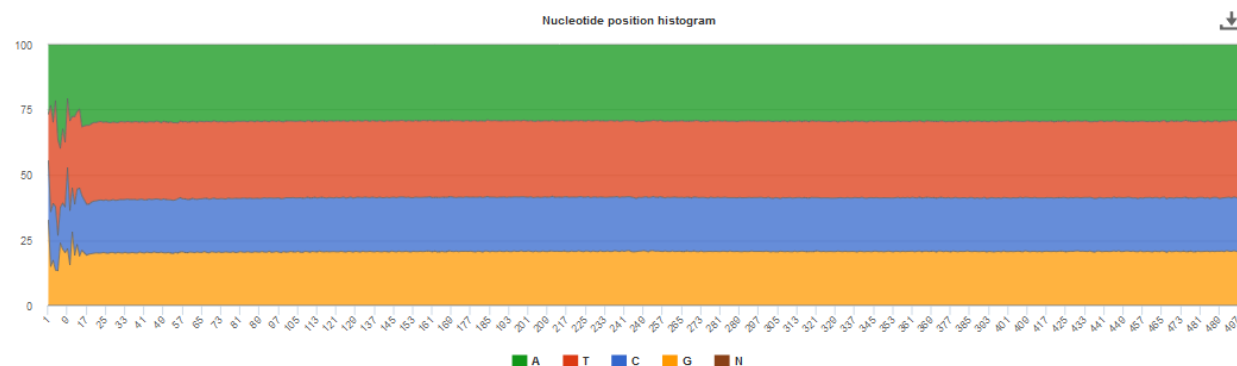


Figure 2. Typical even nucleotide distribution expected for metagenome, metatranscriptome and assembly. Note that the stretch of uneven distribution observed until position 20 are indicative that the sequencing adapters had not been completely removed in the submitted reads.

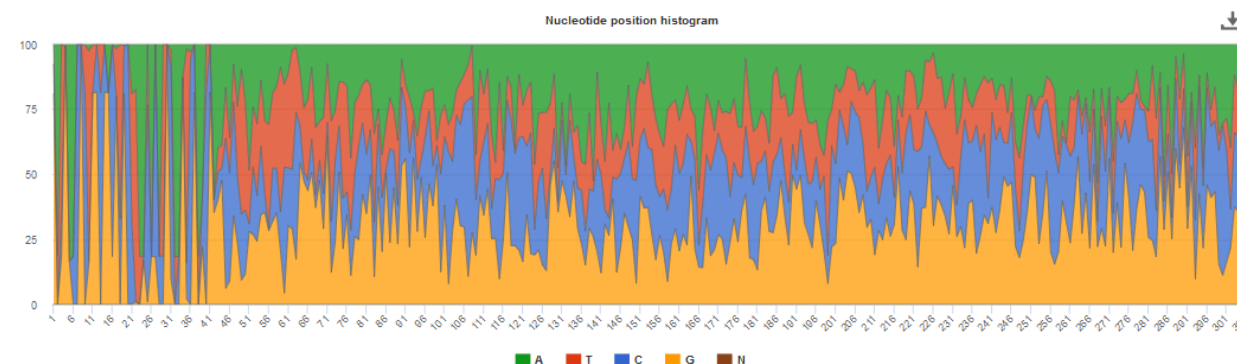


Figure 3. Typical uneven nucleotide distribution expected for amplicon.

4.3 Finding functional information about runs on the EBI Metagenomics website

Functional analysis of runs within projects on the [EBI Metagenomics website](#) can be accessed by selecting the ‘Functional Analysis’ tab found toward the top of any run page (see Figure 4 below). Note that this tab will be greyed out for amplicon runs that have no functional results.

Run (ERR2014354)

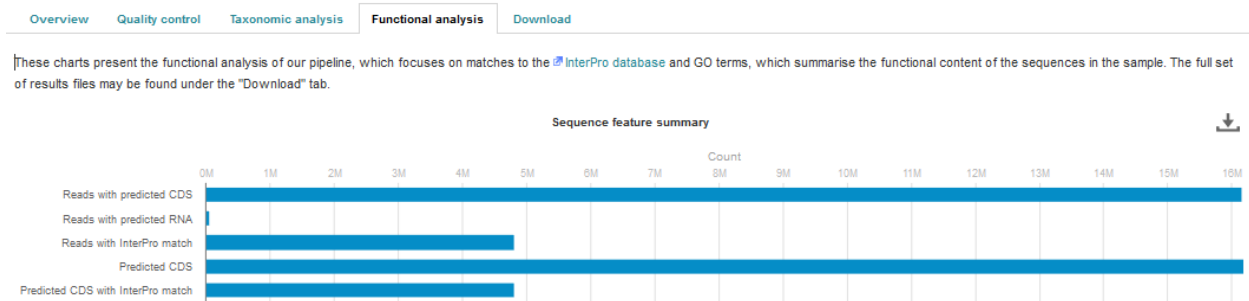
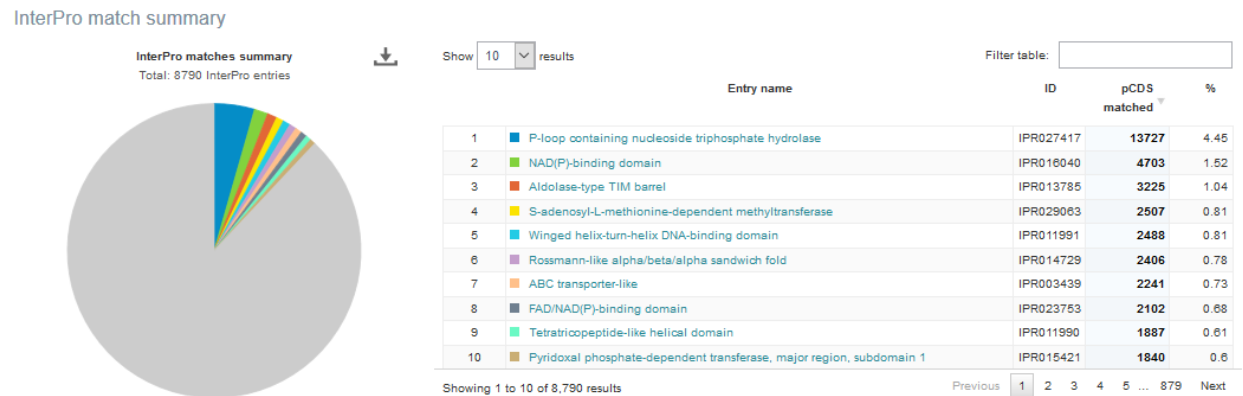


Figure 4. A Functional analysis tab can be found towards the top of each run page. Selecting this tab brings up a page displaying sequence features (‘number of reads with predicted CDS (pCDS)’, ‘number of reads with predicted RNA’, ‘number of reads with InterPro matches’, ‘number of pCDS’ and ‘number of pCDS with InterPro match’).

Below this first bar chart, two other charts display the InterPro match information and GO term annotation for the run, as shown in Figure 5A and 5B below.

A



B

GO Terms annotation

A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.



Figure 5. Functional analysis of metagenomics data, as shown on the EBI Metagenomics website. A) InterPro match information for the predicted coding sequences in the run is shown. The number of InterPro matches are displayed graphically, and as a table that has a text search facility. B) The GO terms predicted for the sample are displayed. Different graphical representations are available, and can be selected by clicking on the ‘Switch view’ icons.

The Gene Ontology terms displayed graphically on the web site have been ‘slimmed’ with a special GO slim developed for metagenomic data sets. GO slims are cut-down versions of the Gene Ontology, containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine-grained terms.

The full data sets used to generate both the InterPro and GO overview charts, along with a host of additional data and intermediate files (processed reads, pCDS, reads encoding RNA and taxonomic analysis results) can be downloaded for further analysis by clicking the Download tab, found towards the top of the page (see complete description here: [Files available to download on the EBI Metagenomics website](#))

4.4 Finding taxonomic information about runs on the EBI Metagenomics website

Taxonomic analysis of runs within projects on the [EBI Metagenomics website](#) can be accessed by selecting the ‘Taxonomic analysis’ tab found toward the top of any run page (see Figure 7 below).

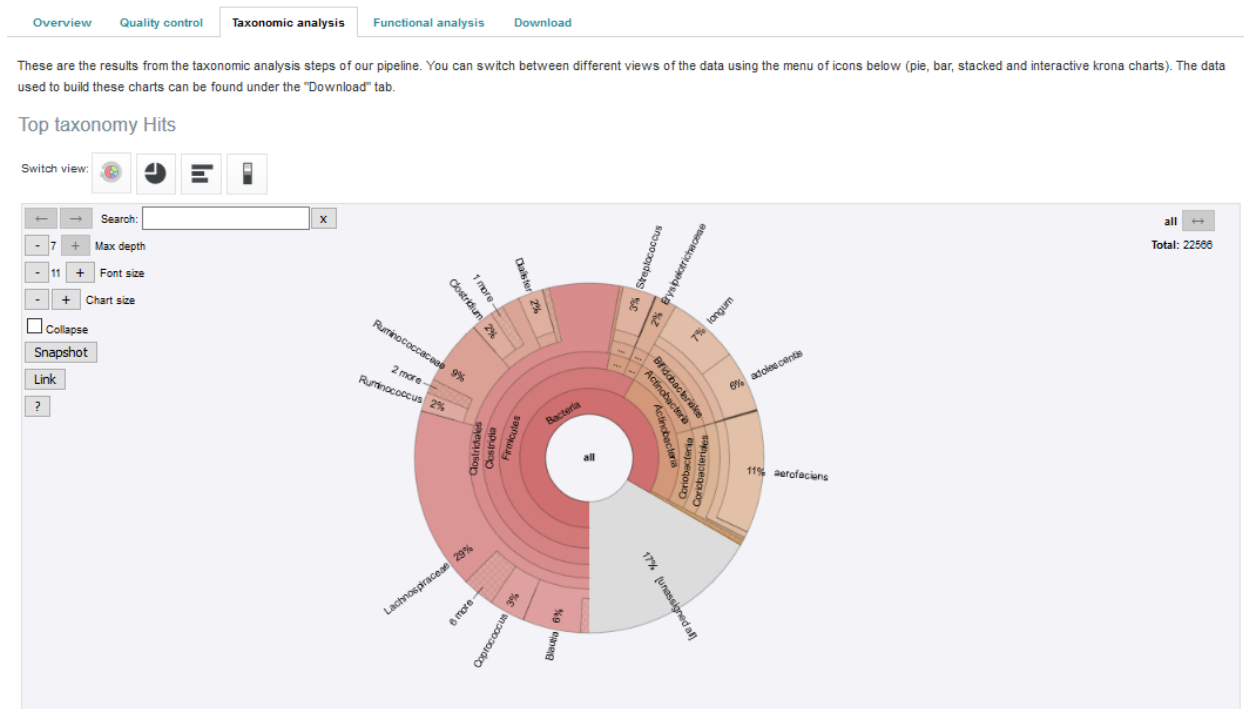


Figure 7. A ‘Taxonomic analysis’ tab can be found towards the top of each run page. Selecting this tab brings up a page displaying the taxonomic results displayed as an interactive [Krona plot](#).

The taxonomic analysis results are displayed as Krona plot. This feature allows users to explore their taxonomic results and to zoom in on a particular taxonomic level by double clicking on it. The corresponding distribution charts are displayed on the right hand side of the panel.

Alternative pie, bar and stacked chart representations can be generated by clicking on the ‘Switch view’ icons located above the Krona plot however data are then presented at the phylum level for clarity.

The full data sets used to generate both Krona and other charts, along with a host of additional data and intermediate files (processed reads, pCDS, reads encoding RNA and functional analysis results) can be downloaded for further analysis by clicking the Download tab, found towards the top of the page (see complete description here: [Files available to download on the EBI Metagenomics website](#))

4.5 Files available to download on the EBI Metagenomics website

EBI Metagenomics analysis pipeline produces a number of files underlying the charts displayed on the website. These files are available via the ‘Download’ tab found toward the top of any run page (see Figure 8 below).

Overview	Quality control	Taxonomic analysis	Functional analysis	Download
Here you may download the full set of analysis results files. For the original raw sequence reads, follow the link to ENA. Multi-part files should be concatenated after unzipping.				
Sequence data				
Name	Data type	Compression	Format	
Submitted nucleotide reads	XXX			(ENA website)
Processed nucleotide reads	XXX	GZIP	FASTA	(68 MB)
Processed reads with pCDS	XXX	GZIP	FASTA	(66 MB)
Processed reads with annotation	XXX	GZIP	FASTA	(34 MB)
Processed reads without annotation	XXX	GZIP	FASTA	(32 MB)
Predicted CDS with annotation	XXX	GZIP	FASTA	(26 MB)
Predicted CDS without annotation	XXX	GZIP	FASTA	(20 MB)
Predicted ORF without annotation	XXX	GZIP	FASTA	(29 MB)
Predicted tRNAs	XXX	-	FASTA	(91 KB)
Functional analysis				
Name	Data type	Compression	Format	
InterPro matches	XXX	GZIP	TSV	(23 MB)
Complete GO annotation	XXX	-	CSV	(159 KB)
GO slim annotation	XXX	-	CSV	(7 KB)
Taxonomic analysis				
Name	Data type	Compression	Format	
Reads encoding 5S rRNA	XXX	-	FASTA	(4 KB)
Reads encoding 16S rRNA	XXX	-	FASTA	(36 KB)
Reads encoding 23S rRNA	XXX	-	FASTA	(105 KB)
OTUs, reads and taxonomic assignments	XXX	-	TSV	(3 KB)
OTUs, reads and taxonomic assignments	XXX	-	HDF5 Biom	(6 KB)
OTUs, reads and taxonomic assignments	XXX	-	JSON Biom	(6 KB)
Phylogenetic tree	XXX	-	Newick format	(915 bytes)

Figure 8. The Download tab is organised in 3 sections: ‘Sequence data’, ‘Functional analysis’ (not available in the case of amplicon runs) and ‘Taxonomic analysis’.

Some of the files, particularly the sequence files in FASTA format, can be large. To facilitate their download process, these files are compressed with GZIP and when too large to be easily transferable, chunked in manageable size. If it is the case for your runs, please download all chunks, decompress them and concatenate them to reconstitute the full files.

4.5.1 Description of fasta files available to download

- Processed nucleotide reads: this file contains all reads having passed the quality control (QC) step.
- Processed reads with pCDS: this file contains all processed reads having predicted CDS(s) (pCDS). The CDS prediction is performed using [FragGenScan](#) on the reads having passed the QC after masking of predicted rRNA and tRNA.
- Processed reads with annotation: this file contains all processed reads containing pCDS(s) annotated by [InterProScan](#).
- Processed reads without annotation: this file contains all processed reads having pCDS(s) not annotated by InterProScan
- Predicted CDS with annotation : this file contains all the predicted proteins having been annotated by InterProScan. The sequence headers are: <run_id>_<start of pCDS>_<end of pCDS>_<strand of pCDS><space><InterPro term>/<member database ID>/<start of hit in predicted protein>-<end of hit in predicted protein>.
- Predicted CDS without annotation: this file contains all the predicted proteins not annotated by InterProScan. The sequence headers are <run_id>_<start of pCDS>_<end of pCDS>_<strand of pCDS>.
- Predicted ORF without annotation: this file contains all the pCDS coding for predicted proteins that were not annotated by InterProScan. The sequence headers are <run_id>_<start of pCDS>_<end of pCDS>_<strand of pCDS>.
- Predicted tRNAs: this file contains all the sequences predicted to encode tRNAs. The prediction was done using models from [Rfam](#) with [Hmmer tools](#).
- Reads encoding 5S rRNA: this file contains all reads predicted to encode for 5S rRNA by rRNASelector.
- Reads encoding 16S rRNA: this file contains all reads predicted to encode for 16S rRNA by rRNASelector.
- Reads encoding 23S rRNA: this file contains all reads predicted to encode for 23S rRNA by rRNASelector.

4.5.2 Description of functional annotation files available to download

- InterPro matches file: it is a tab-delimited file containing 15 columns. They are fully described [here](#)
- Complete GO annotation file: it is a comma-separated file containing 4 columns. The first column lists the GO terms (labelled [GO:XXXXXXX](#)) having been associated with the predicted CDSs. The second gives the GO term description while the third indicates which category the GO term belong to. There is 3 category: 'biological process' (higher biological process such as 'rRNA modification'), 'molecular function' (individual catalytic activity such as 'mannosyltransferase activity') and 'cellular component' (cellular localisation of the activity such as 'mitochondrion'). The last column give the number of predicted CDSs having been annotated with the GO terms for the run.
- GO slim annotation file: this file is derived from the 'Complete GO annotation file' and has the same format. The GO slim set is a cut-down version of the GO terms containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the details of the specific fine grained terms. Go slim terms are used for visualisation on the website. To illustrate how the GO slim terms relates to the GO terms, the different metal binding GO terms present in the 'Complete GO annotation' file are summarized as one generic metal binding term in the 'GO slim annotation' file. The last column give the number of predicted CDSs having been annotated with the GO slim terms for the run.

4.5.3 Description of taxonomic assignment files available to download

- OTUs, reads and taxonomic assignments files: these files contain the same data presented in 3 different format : tab-separated file (TSV) and two Biom file (HD5F and JSON). The TSV file contains 3 columns which headers are in the second line of the file. The first column is the OTU Id. The second column indicates the number of predicted 16S sequences associated with each OTU. The third column contains the taxonomic lineages provided by [GreenGenes database](#). Note that the number of unannotated 16S sequences is not indicated in this file. This file can be directly imported into [Megan6](#) for visualisation and further analysis. The OTU id can be compared between runs for version 2 and 3 of the pipeline as they have been generated using [Qiime closed-reference protocol](#). The Biom files are [computer-readable files](#). The HD5F (Hierarchical Data Format) format can be imported into analysis and visualisation tools such as Matlab and R. A larger number of commercial and freely available tools, such as MEGAN6, can consume the JavaScript Object Notation (JSON) format.
- Phylogenetic tree (Newick format) file (only available up to version 3 of EBI Metagenomics pipeline): this file can be used to visualise the hierarchical distribution of the taxonomic lineages of each run. The [Newick format](#) is a computer-readable format to represent the tree and can be directly imported into freely-available viewers such as [FigTree](#) and [ITOL \(interactive Tree of Life\)](#).

4.6 Summary files

In addition to the output files for individual runs, described above, EBI Metagenomics provides a number of summary files available via the ‘Analysis summary’ tab on the project page (Figure 9 below). They summarize the counts per feature across all runs of a study and therefore provide an easy way to identify patterns. The summary files are split between functional (not available for amplicon-only study) and taxonomy sections.

Project ERP022201 (PRJEB20085)

Metagenomic study from hydrothermal sediments retrieved from Menez Gwen and Rainbow deep-sea vents.



Overview
Analysis summary

In this section you can download the different results matrix files summarising the project. Each downloadable file contains an aggregation of the analysis results from the individual project runs. To visualise and download the analysis results for individual runs, please access their respective pages.

Pipeline version 3.0

Functional analysis for the project

- [InterPro matches \(TSV\)](#) - 549 KB
- [Complete GO annotation \(TSV\)](#) - 171 KB
- [GO slim annotation \(TSV\)](#) - 7 KB

Taxonomic analysis for the project

- [Phylum level taxonomies \(TSV\)](#) - 458 bytes
- [Taxonomic assignments \(TSV\)](#) - 6 KB

Figure 9. The ‘Analysis summary’ tab is organised in 2 sections: ‘Functional analysis for the project’ and ‘Taxonomic analysis for the project’ (the former is not available in the case of amplicon runs).

4.6.1 functional summary files

- [InterPro matches\(TSV\)](#): this tab-separated file contains 2 designation columns followed by a column for each valid runs of the project. The first column lists the InterPro terms having been associated to the predicted CDSs. The second column gives the description of the InterPro terms. All columns labelled with a run identifier present the number of predicted CDSs having been annotated with each InterPro terms for this run.
- [Complete GO annotation \(TSV\)](#): this file contains 3 designation columns followed by a column for each valid runs of the project. The first column lists the GO terms (labelled [GO:XXXXXXXX](#)) having been associated to the predicted CDSs. The second column gives the GO term description while the third column indicates which

category the GO term belong to. All columns labelled with a run identifier present the number of predicted CDSs having been annotated with each GO terms for this run.

- The ‘GO slim annotation (TSV)’ file is derived from the ‘Complete GO annotation’ file and has the same format. The GO slim term set is a cut-down version of the GO terms containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms.

4.6.2 taxonomy summary files

- Taxonomic assignments (TSV): this file contains one ‘Taxonomy’ column followed by a column for each valid runs of the project. The ‘Taxonomy’ column list the taxonomic lineages having been associated with the predicted 16S sequences. All columns labelled with a run identifier present the number of predicted 16S sequences having been annotated with the taxonomic lineages for this run. This file can be directly imported into [Megan6](#) for visualisation and further analysis.
- The ‘Phylum level taxonomies (TSV)’ file is derived from the ‘Taxonomic assignments’ file and presents the assignments brought up to ‘phylum’ level in order to give a high level view of the taxonomic assignments. The two first columns of this file present the ‘kingdom’ and ‘phylum’ level assignments, respectively. All columns labelled with a run identifier present the number of predicted 16S sequences having been annotated with the ‘phylum’ level taxonomic lineages for this run.

4.7 Comparison tool

Comparing runs helps to identify feature associated with experimental factors. EBI Metagenomics has developed a Comparison Tool that allows user to compare the GO-slim terms associated with the runs of a project (see [Analysis pipeline](#)).

To use the current tool, select the corresponding tab from any EBI Metagenomics webpage (Figure 10 below):

Home Search Submit data Projects Samples **Comparison tool** About Contact Not logged in Login

EBI Metagenomics > Comparison tool

Comparison tool

The comparison is currently based on a summary of Gene Ontology (GO) terms derived from InterPro matches to the selected runs. It is therefore not possible to select studies that contain only amplicon data at present.

Project list

- 4 samples from Pacific Ocean uncultured phage metagenome
- 454Titanium barcoded pyrosequencing of the 16S rRNA and proteorhodopsin-containing phot...
- 5 bacterial species mock community genome sequencing
- A Catalogue of the Mouse Gut Metagenome
- A core gut microbiome in obese and lean twins
- A human gut microbial gene catalog established by deep metagenomic sequencing (MetaHIT)
- A longitudinal study of the feline faecal microbiome identifies changes into early adul...**
- A Metagenomic Analysis of a Winogradsky Column Supplemented with Molybdate
- A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratifi...
- A method for identifying metagenomic species and variable genetic elements by exhaustiv...
- A plaque on both your houses. Exploring the history of urbanisation and infectious dise...
- A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial re...
- Acid sulfate soil microbial profile - pre-investigation

[More info about selected project](#)

Run list (7 selected out of 88)

- 10 - ERR878216
- 12 - ERR878217**
- 15 - ERR878218
- 16 - ERR878219
- 18 - ERR878220
- 20 - ERR878221**
- 21 - ERR878222**
- 22 - ERR878223**
- 23 - ERR878224**
- 24 - ERR878225**
- 25 - ERR878226**
- 26 - ERR878227**
- 27 - ERR878228

Select all | Unselect all

Advanced settings

Show / hide advanced settings

Compare | Clear all

Figure 10. The ‘Comparison tool’ tab let the user select projects and associated runs to compare them based on the GO-slim distribution.

- The first step is to select the project of interest. They are listed by title in alphabetical order. You can search the project list by entering the first letters of the title from the project you’re interested in.

- Clicking on the ‘More info about selected project’ link, located below the Project list, after selecting a project, will open a new browser window displaying the project page.
- Upon project selection, the ‘Run list’ window will be populated with the list of runs associated with the project and suitable for comparison. You can select all runs (using the ‘Select all’ link below the window) or up to 30 runs (by using the Ctrl key for Windows PC or Command key on Mac).
- Users can select the ‘Advanced settings’ link to have the options to set the relative abundance threshold for the GO terms to appear in the stack columns, the format of heatmap generated and the number of GO terms with most variation to display in the representations.

To start the comparison for your selection, simply click on ‘Compare’. The page will now display the study and selected runs on top of 5 new comparison tabs:

- The first one is a barcharts representation with 3 dynamic graphs, corresponding to the 3 GO terms categories (see [Analysis pipeline](#)). On each, the GO terms and their relative abundance in each selected run is displayed. Hovering the mouse pointer above a bar will display the relative abundance values for this term in the corresponding run. You can export these barcharts representation in PNF, PDF or SVG format using the tool on the top right hand side.
- The second tab contains stacked column representations with the same dynamic properties than in the barcharts with the addition of the possibility to hide one or more terms of choice by selecting them from the list displayed below each category graph.
- The third tab presents heatmaps allowing to quickly identified patterns between the selected runs based on the relative abundance of the GO terms. There is currently no export function for this page although the images, being static, can be directly copied.
- The fourth tab contains dynamic Principal Component Analysis graphs which represent the amount of variance between runs, based on the relative abundance of the GO terms, between the runs for each GO category. Selecting a rectangular region with the mouse pointer will zoom in, which help to separate clustered run markers. The export function allows to download all or the enlarged region.
- The last tab is a searchable table where you can see the absolute and relative abundance of a given GO term for each run. It is based on the ‘Analysis_summary’ abundance table available from the project page. You can search the table using the run identifier, GO name, GO category, GO id or even absolute or relative abundance.

We are working with collaborators to develop this tool in order to be able to compare taxonomic annotations, provide statistical validations and compare runs between projects.

4.8 Data discovery on EBI Metagenomics portal

EBI metagenomics is the largest metagenomic resource of public datasets. In order to help users accessing the data present on the portal, EBI Metagenomics offers a powerful search tool and a range of browsing options.

4.8.1 Search tool

The Search tool is underpinned by [EBI search](#) and accessible via any EBI Metagenomics page (Figure 11 below).

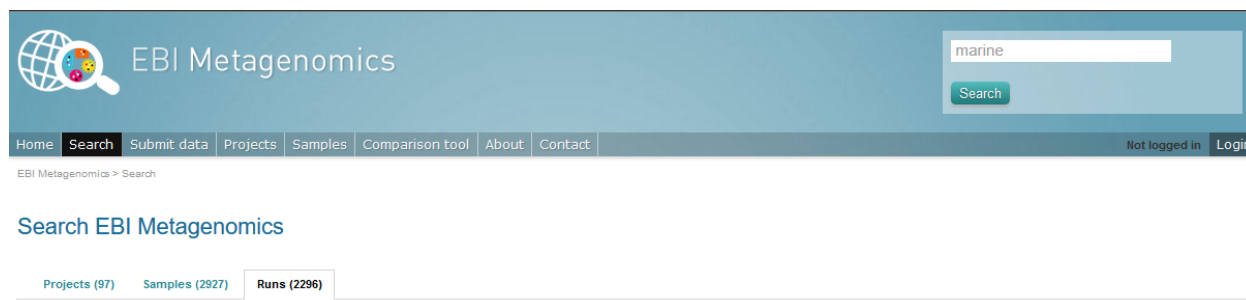


Figure 11. The ‘Search tool’ can be accessed using the ‘search’ tab or the ‘search’ button located on the right of the EBI Metagenomics banner. The search space can be restricted using the ‘search’ field located above the latter.

The search page contains 3 tabs allowing users to navigate between project, sample and run search levels. In each tab, the left hand side panel provide a number of facets that can be used to restrict the search space.

- at the project level, the search can be restricted by ‘biome’ and ‘centre name’. Selection of any of the facets will impact the search at sample and run level in order to be able to drill down into the results. Search results can be downloaded as tab-separated file.
- at the sample level, in addition to ‘biome’, the choice of facets includes ‘temperature’, ‘depth’, ‘sequencing method’, ‘sample origination’, ‘disease status’ and ‘phenotype’, when provided. Note that these metadata are provided by the data submitter and are not curated.
- at the run level, users can restrict their searches according to ‘biome’, ‘temperature’, ‘depth’, ‘pipeline version’, ‘organism’, ‘experiment type’ as well as Go and InterPro terms.

4.8.2 Browsing options

- Public project can be accessed using the links corresponding to the number of projects, samples and runs or experiment types located on the EBI Metagenomics home page below the main banner. Selecting one of those will redirect users to the corresponding EBI Metagenomics search page.
- Another way to discover data of interest is to browse the public projects by biome as displayed on the EBI Metagenomics homepage. The 10 biomes with higher number of projects are displayed by default however the list can be extended using the ‘See all biomes’ link. Upon selection, a table giving the hierarchical lineage according to [GOLD database classification](#) is provided. On the right hand-side of this table, the number of projects associated to the lineage in the strict sense or including sub-lineages are displayed as dynamic links giving access to the selected projects.
- Users can also access particular projects, or samples, using the corresponding tabs located above the EBI Metagenomics banner. The list of projects, or samples, can be restricted using the Biome drop-down menu and/or text search. The results of this filtering can be downloaded using the two spreadsheet icons located above the right hand-side of the tables.
- Finally, users have the option to access, from the EBI Metagenomics homepage, the latest public projects uploaded via the right side of the ‘Browse projects’ section.

4.9 Privat area

If you have given consent to the EBI Metagenomics team to analyse your data for which you have requested a pre-publication confidential hold, you can access the analysis results of those pre-published data sets by using your privat area. You can simply access this area by clicking on the ‘Login’ button, which you will find on the top right hand side of any page (see Figure 12 below).

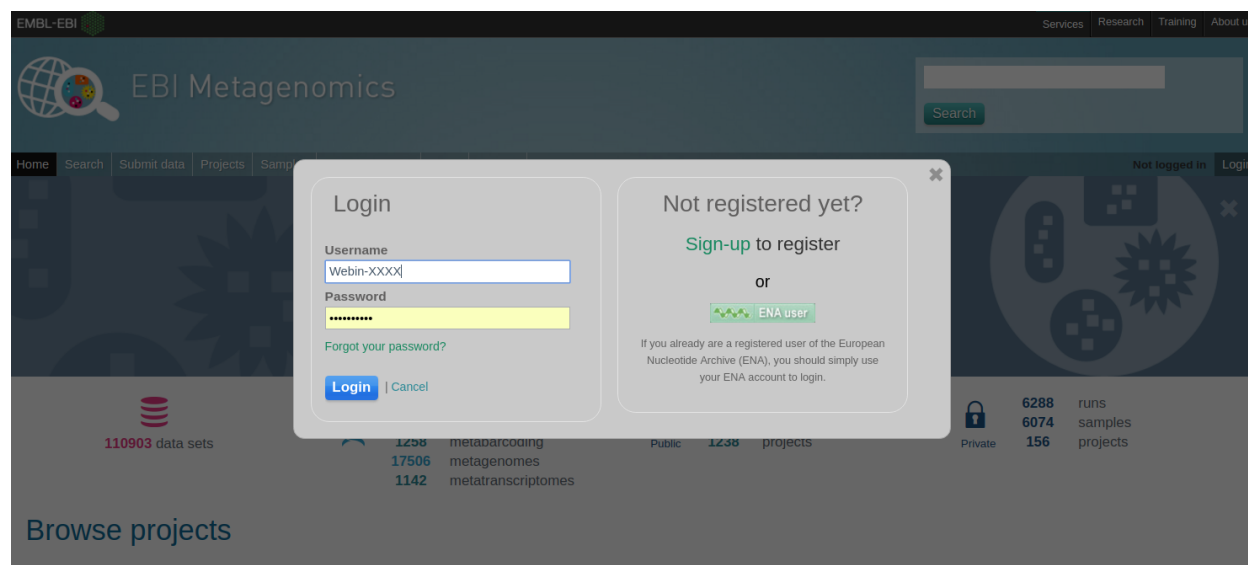


Figure 12. A login dialog will open once you have clicked on the ‘Login’ button, which can be found on the right top corner of each page.

After you have successfully logged into our system, you will have direct access to all your privately (and publicly) submitted projects and samples. You will find a list of your latest submissions (projects and samples) on the home page, but you have also access to all your submitted projects so far on the projects list view (Figure 13 below). On that page you will find a drop down filter item ‘My projects’, which allows you to list all your projects.

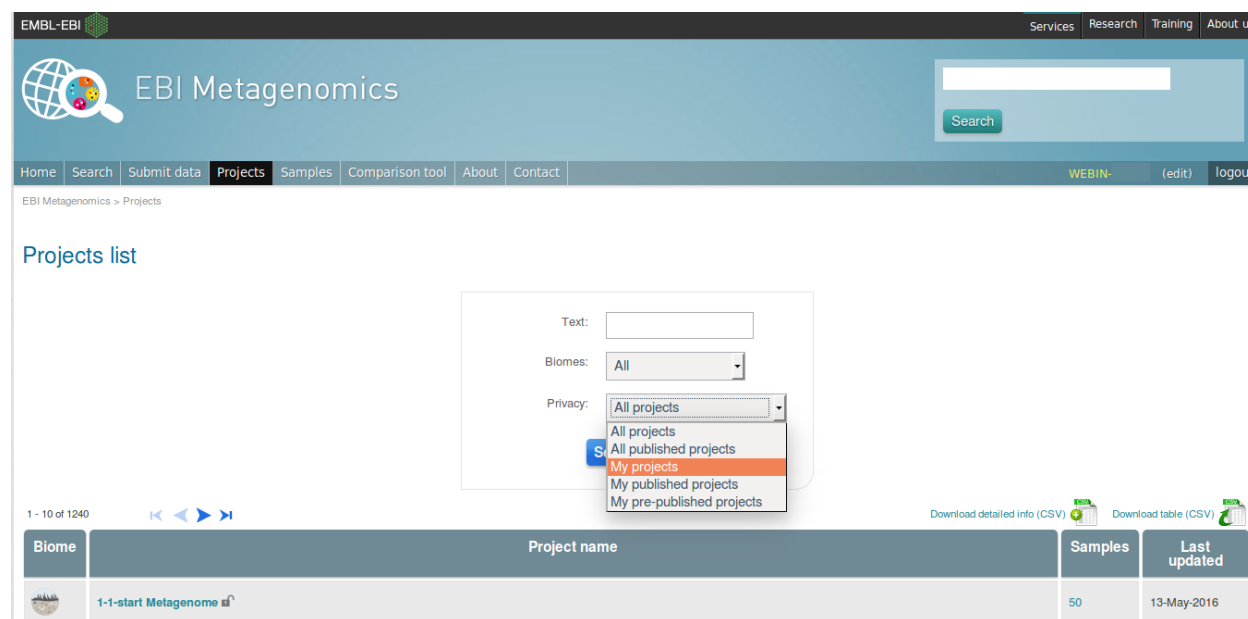


Figure 13. Filter options on the projects list view.

The comparison tool will list all your non-amplicon projects in the private area, for you ready to compare. Currently we do not permit the comparison of public and private data. In the public area you will find a list of all publicly available projects in EMG.

5.1 EBI Metagenomics online tutorials

A number of online tutorials relating to EBI Metagenomics are available:

[EBI Metagenomics portal: Quick tour](#)

[EBI Metagenomics portal: Submitting metagenomics data to the ENA](#)

5.2 ENA online guides

The ENA also provides relevant online help and information on how to submit metagenomic read data to the archive (file formats, uploading data files, etc) and the use of checklists to capture contextual meta-data:

[Submitting read data to ENA](#)

[Submitting environmental samples to ENA](#)

[Sample checklists](#)

6.1 What kind of sequence data does the service accept?

EBI Metagenomics accepts sequencing data from a wide range of platforms, including Roche 454, Illumina and Ion Torrent. In addition to analysis of whole-genome shotgun (WGS) sequenced metagenomic and metatranscriptomic samples, it also provides analysis of 16S ribosomal RNA (rRNA) amplicon data. If you would like to submit Oxford Nanopore sequences, we suggest you [contact us](#) prior to submission.

6.2 Can I submit assembled metagenomic sequences for analysis?

Yes, we welcome submission of assembled data.

6.3 Can I submit 18S rRNA or ITS amplicon sequences?

At the present moment, the pipeline does not provide taxonomic analysis of 18S rRNA or ITS sequences, so no meaningful analysis results will be returned for these data sets.

6.4 Can I submit viral sequences?

Although EBI Metagenomics does not currently provide taxonomic analysis of viral sequences, any reads submitted to the pipeline that encode predicted protein coding sequences (pCDS) undergo functional analysis using InterPro. Therefore, while no taxonomic data will be returned for viral sequences, it should be possible to obtain functional analysis results.

6.5 How do I run a BLAST search against the metagenomics datasets?

We don't offer BLAST searches against our metagenomic data sets via the web site. We do not have the resources to offer this service, owing to the size of the data. If you wish to perform such an analysis, You would will need to download the data locally and index it for BLAST. We do provide [scripts](#) for bulk download of results files, which may prove useful.

6.6 Can I change the release date of my project?

The date of release is set by you during the submission in ENA. If you do not explicitly choose a date, it is set to two years from the submission date by default. You can reduce or extend the release date by going to the 'Submit & update' page of the ENA website, logging in using your Webin account id, selecting the relevant study and then clicking the pen icon near the current release date to alter it.

6.7 How long will it take for my data to be analyzed?

We aim to analyze submitted data as quickly as possible. However, submitted data are only available to us once they have been validated and archived by ENA. This process takes at least 24 hours, and in some cases several days. Once the sequence files are made available, the analysis time depends on our current analysis backlog, the size and number of runs in the project you have submitted. Most studies will be validated, archived and analysed within one week. If you are concerned that your data is taking a long time to be analysed, please [contact us](#).

6.8 I have submitted my data - how do I trigger the analysis?

There is no manual way to trigger analysis. If you have provided [access agreement](#) for EBI Metagenomics, we will pick up your sequences from ENA automatically and queue them for analysis.

6.9 Do you have an API?

Not yet, but we are actively working on it. Watch this space.

6.10 How can I download several sets of data?

While we currently do not have an API, we have Python scripts allowing users to automatically download most processed files from the EBI Metagenomics website. The scripts and instructions for bulk downloading from the latter resource can be found [here](#).

6.11 How can I bulk download meta-data?

Most of the meta-data associated with projects is taken from the ENA. While EBI Metagenomics does not currently provide an API, the ENA do provide this service, which can be used as a stop gap until our API is in place ([ENA's API documentation](#))

6.12 How can I re-analyse my data with a different version of the pipeline?

It is possible to analyse data sets with different versions of our analysis pipeline. The original analyses are not deleted and are available side by side on our web site. Users interested in having data re-analysed should [contact us](#).

6.13 Can I request that a dataset is analyzed if I am not the original submitter?

We are currently working through the analysis of all publicly available metagenomic datasets in ENA, so if there is a publicly available study that you would like to see analysed in EBI Metagenomics, please get in touch and we will prioritise it.

6.14 Can I request my data to not be analyzed by EBI Metagenomics?

We can only access private data for analysis if you gave us agreement to do so. If, for any reasons, you do not want EBI Metagenomics to analyze one of your datasets, please [contact us](#) . If your data are public in ENA, then we can access them for analysis in any case.

6.15 Can I compare the taxonomic assignments between runs of a project?

The current version of the comparison tool let you only compare the GO annotations for runs of the same project. We are currently working on extending the functionality to taxonomy but this is not yet ready for release. In the meantime, please have a look at the summary files provided on the project page. They summarized the counts per feature across the runs and provide an easy way to identify patterns.

The ‘OTUs, reads and taxonomic assignments.tsv’ can be directly imported into [Megan 6](#) to perform comparison and visualisation. The Biom format can also be imported into third-party tools.

6.16 Can I know which bacteria encodes particular pCDS in my dataset?

The short answer is that it is generally not possible. The reason is that we annotate directly the reads and select the reads containing 16S for taxonomy assignments. The protein prediction is then performed on all reads after masking the tRNA and rRNA sequences. To link a predicted protein to a taxonomic assignments, the protein-coding gene would need to be on the same read than the annotated 16S sequence. It is possible to check if this is the case using the sequence headers from the ‘Interpro matches.tsv’ and ‘Reads encoding 16S rRNA.fasta’ files, both available on the ‘Download’ for each run. The same answer applies to assembly although, depending on the contig length, more protein-coding genes may be located near a 16S rRNA genes.

CHAPTER 7

Indices and tables

- `genindex`
- `modindex`
- `search`