

Functional and taxonomic analysis using EBI Metagenomics



Contents

Tutorial information	2
Tutorial learning objectives	2
An introduction to functional analysis using EMG.....	3
What are protein signatures?	3
Assigning functional information to metagenomic sequences.....	4
Finding functional information about samples on the EMG website.....	5
Exercises	8
Browsing analysis results on the EMG website	8
Comparing analyses	12
Visualising taxonomic data using MEGAN	13
Searching EBI Metagenomics using indexed metadata	15

Tutorial information

Course description	This tutorial provides an introduction to functional analysis of metagenomic data, using the EBI Metagenomics (EMG) resource. You will learn about the underlying analysis software and how it is used to infer information about the function of predicted coding sequences, and how to find and download data for use with 3 rd party analysis software.
Course level	Suitable for graduate-level scientists and above
Pre-requisites	Basic knowledge of biology and basic Unix skills
Subject area	Genes, Genomes and Metagenomes; Sequence Analysis
Target audience	Scientists interested in metagenome analysis
Resources required	Internet access (a current browser such as the latest version of Firefox, Chrome or Safari) and the MEGAN analysis package.
Approximate time needed	60 minutes

Tutorial learning objectives

After completing this course, you should:

- understand how EMG provides functional and taxonomic analysis of metagenomic data sets
- know where to find and how to interpret analysis results for samples on the EMG website
- understand how to compare data sets using the web site or 3rd party tools, such as MEGAN

An introduction to functional analysis using EMG

The EBI Metagenomics resource (EMG) provides functional analysis of predicted coding sequences (pCDS) from metagenomic data sets using 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, providing names and descriptive abstracts and, whenever possible, adding Gene Ontology (GO) terms. Links are also provided to pathway databases, such as KEGG, MetaCyc and Reactome, and to structural resources, such as SCOP, CATH and PDB.

What are 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 Figure 1).

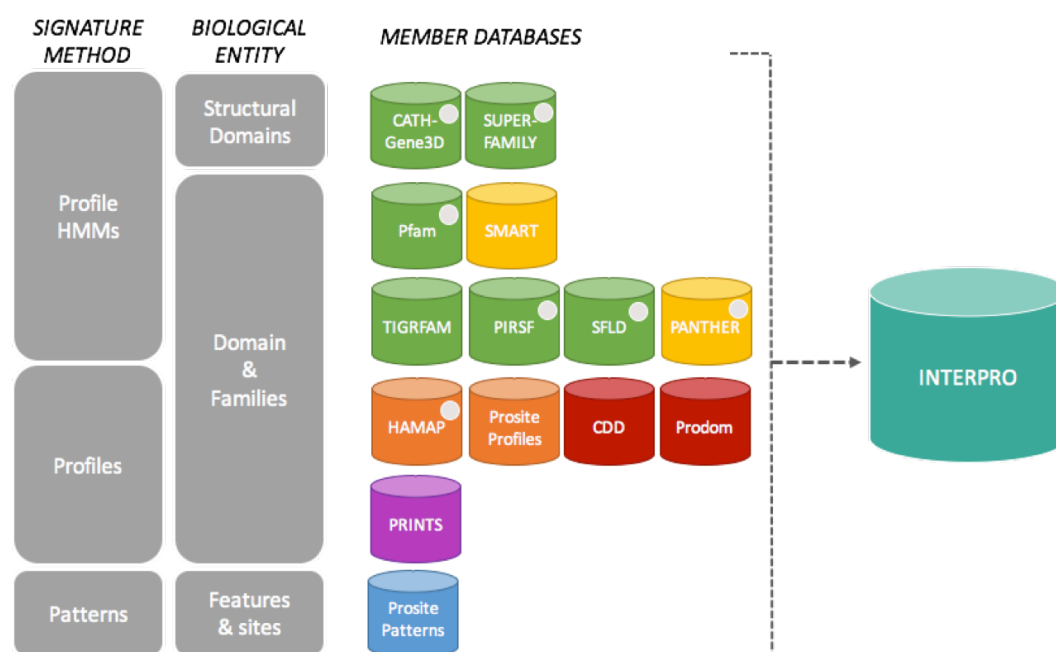


Figure 1. InterPro member databases grouped by the methods used to construct their signatures and focus of interest.

Only a subset of the InterPro member databases are used by EMG: 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.

Assigning functional information to metagenomic sequences

Whilst InterPro matches to metagenomic sequence sets are informative in their own right, EMG offers an additional type of analysis in the form of Gene Ontology (GO) terms.

The Gene Ontology 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. Terms in the Gene Ontology are ordered into hierarchies, with less specific terms towards the top and more specific terms towards the bottom (see Figure 2).

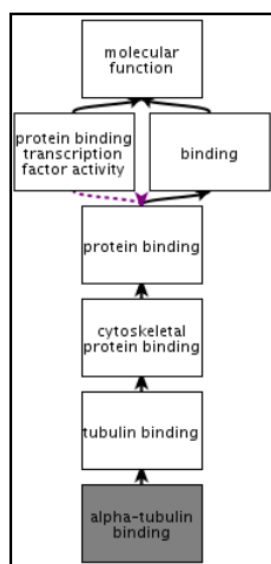


Figure 2. An example of GO terms organised into a hierarchy, with terms becoming less specific as the hierarchy is ascended (e.g., alpha-tubulin binding is a type of cytoskeletal binding, which is a type of protein binding). Note that a GO term can have more than one parent term. The Gene Ontology also 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 project can be found at <http://www.geneontology.org/GO.doc.shtml>

As part of the EMG 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.

Finding functional information about runs on the EMG website

Functional analysis of samples within projects on the EMG website (www.ebi.ac.uk/metagenomics/) can be accessed by clicking on the **Functional Analysis** tab found toward the top of any run page (see Figure 3 below).

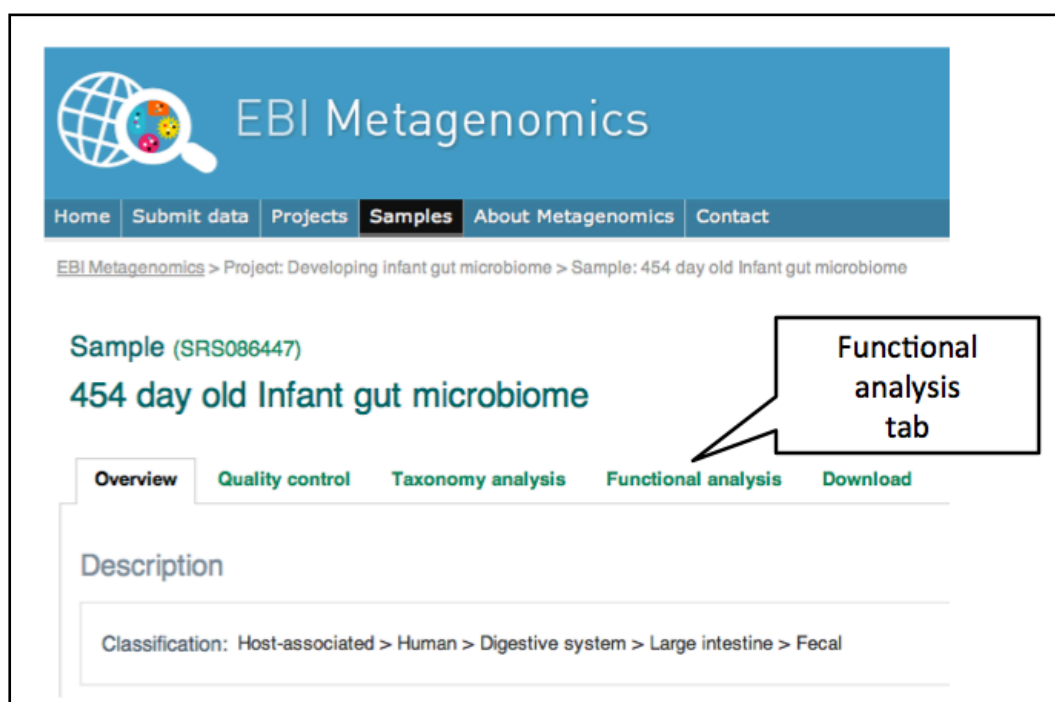


Figure 3. A **Functional analysis** tab can be found towards the top of each run page.

Clicking on this tab brings up a page displaying sequence features (the number of reads with pCDS, the number of pCDS with InterPro matches, etc), InterPro match information and GO term annotation for the run, as shown in Figure 4 below.

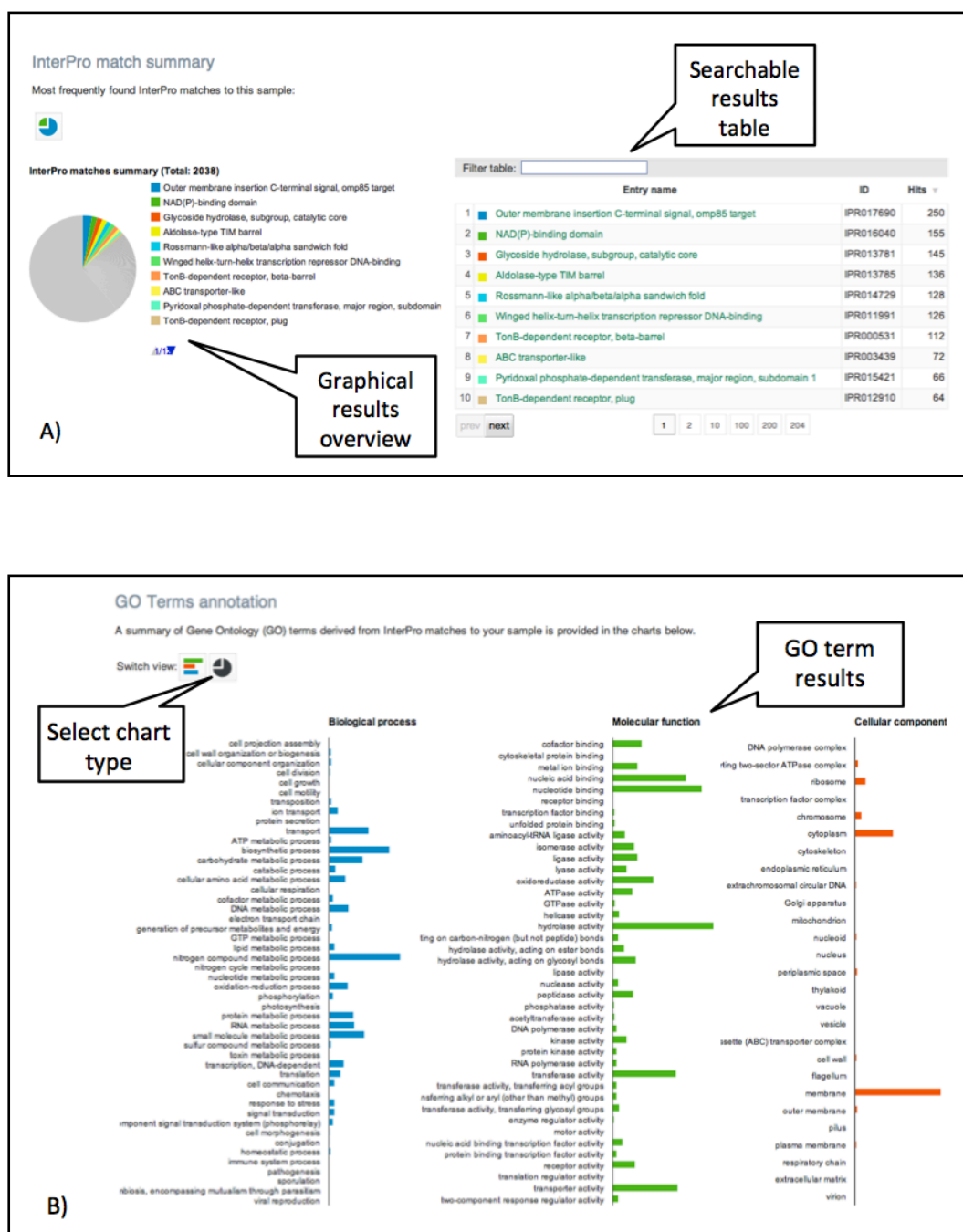


Figure 4. Functional analysis of metagenomics data, as shown on the EMG website. A) InterPro match information for the predicted coding sequences in the run is shown. The number of InterPro matches are displayed graphically, and as a table that has a text search facility. B) The GO terms predicted for 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 (processed reads, pCDS, reads encoding 16S rRNAs, taxonomic analyses, *etc*) can be downloaded for further analysis by clicking the **Download** tab, found towards the top of the page (see Figure 5).

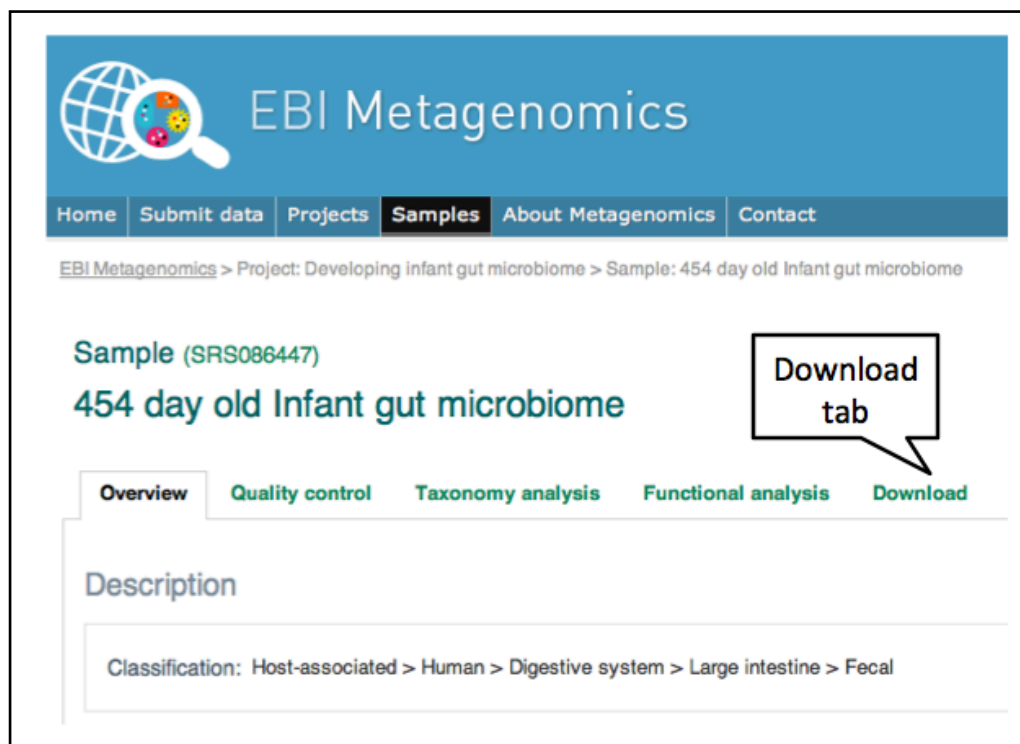




Figure 5. Each run has a download tab, where the full set of sequences, analyses, summaries and raw data can be downloaded.


Exercises


Browsing analysis results on the EMG website


 For this session, we are going to look at the Ocean Sampling Day (OSD) 2014 project, which involved simultaneously sampling from geographically diverse oceanographic sites on Solstice 2014. A map of all of the sampling sites is shown on the project page: <https://www.ebi.ac.uk/metagenomics/projects/ERP009703>


 To get the OSD project page, either follow the above link or open the Metagenomics Portal home page (<https://www.ebi.ac.uk/metagenomics/>), enter 'OSD' in the search box on the top right hand side of the page, and follow the link to project ERP009703.


You should now have a Project overview page, which describes the project, submitter contact details, and links to the samples and runs that the project contains.

 Scroll down to **Associated Runs** and use the 'Filter' search box to find the **OSD80_2014-06-21_0m_NPL022** sample (**ERS667582**). Click on the Sample Name link (not the Run link) to arrive at the overview page, describing various contextual data, such as the geographic location from which the material was isolated, its collection date, and so on.

 **Question 1:** What is the latitude, longitude and depth at which the sample was collected?

 **Question 2:** What geographic location does this correspond to?

 **Question 3:** What environmental ontology (ENVO) identifier has the sample material been annotated with?

 Now scroll down to the 'Associated runs' section of the page. Some samples can have a number of sequencing runs associated with them (for example, corresponding to 16S rRNA amplicon analyses and WGS analyses performed on

the same sample). In this case, there is only 1 associated run: **ERR770971**. Click on the Run ID to go to the Run page.



This page has a number of tabs towards the top (**Overview, Quality control, Taxonomy analysis, Functional analysis, and Download**). Click on the 'Download' tab. Click the file labelled 'Predicted CDS' link to save this file to your computer. Find the file (it should be in your Downloads folder), unzip it and examine it using 'less' by typing the following commands in a terminal window:

```
cd ~/Downloads

gzip -d ERR770971_MERGED_FASTQ_pCDS.faa.gz

less ERR770971_MERGED_FASTQ_pCDS.faa
```

Have a look at one or two of the many sequences it contains.



To quit the 'less' view, press 'q'.

You can count the total number of sequences in the file by grepping the number of header lines that start with ">"

```
grep -c ">" ERR770971_MERGED_FASTQ_pCDS.faa
```

In a moment, we will look at the analysis results for this entire batch of sequences, displayed on the EMG website. First, we will attempt to analyse just one of the predicted coding sequences using InterPro (the analysis results on the EMG website summarise these kind of results for hundreds of thousands of sequences).





In a new tab or window, open your web browser and navigate to <http://www.ebi.ac.uk/interpro/> Copy and paste the following sequence, which has been taken from the file you downloaded, into the text box on the InterPro home page where it says 'Analyse your protein sequence':

```
>HWI-M02024:110:000000000-A8H0K:1:1101:23198:21331-
1:N:0:TCAGAGAC_1_267_-
HLLSYRYAYGKFSSTHEATIGGCFLTKDEELDDHIVKYEIWDTAGKNGTIHLPRCTTSKAYXIQVT
WYRNAIAAVVVFVDVTSRDSFEK
```




Press **Search** and wait for your results. Your sequence will be run through the InterProScan analysis software, which attempts to match it against all of the different signatures in the InterProScan database.


 **Question 4:** Which protein family does InterProScan predict your sequence belongs to, and what GO terms are predicted to describe its function?

 Clicking on the InterPro entry names or IPR accession numbers will take you to the InterPro entry pages for your result, where more information can be found.


 Return to the overview page for **ERR770971**.


First, we will find the number of sequences that made it through to the functional analysis section of the pipeline.

 Click on the **Quality control** tab. This page displays a series of charts, showing how many sequences passed each quality control step, how many reads were left after clustering, and so on.


 **Question 5:** After all of the quality filtering steps are complete, how many reads were submitted for analysis by the pipeline?

Next, we will look at the results of the functional predictions for the pCDS. These can be found under the **Functional analysis** tab.


 Click on the **Functional analysis** tab and examine the InterPro match section. The top part of this page shows a sequence feature summary, showing the number of reads with predicted coding sequences (pCDS), the number of pCDS with InterPro matches, etc.

 **Question 6:** How many predicted coding sequences (pCDS) are in the run?

 **Question 7:** How many pCDS have InterProScan hits?

 Scroll down the page to the InterPro match summary section.

 **Question 8:** How many different InterPro entries are matched by the pCDS?

 **Question 9:** Why is this figure different to the number of pCDS that have InterProScan hits?

Next we will examine the GO terms predicted by InterPro for the pCDS in the sample.



Scroll down to the GO term annotation section of the page and examine the 3 bar charts, which correspond to the 3 different components of the Gene Ontology.



Question 10: What are the top 3 biological process terms predicted for the pCDS from the run?



Selecting the pie chart representation of GO terms makes it easier to visualise the data to find the answer.

Now we will look at the taxonomic analysis for this run.



Click on the **Taxonomic Analysis** tab and examine the phylum composition graph and table.



Question 11: How many of the WGS reads are predicted to encode 16S rRNAs?



Question 12: What are the top 3 phyla in the run, according to 16S rRNA analysis?



Examine the Krona chart, which is the default view, or can be selected using



this icon: This brings up an interactive chart that can be used to analyse data at different taxonomic ranks.




Question 13: What is the proportion of *Polaribacter* in the population?





Note: if the cyanobacteria section of the chart looks strange, this is because the version of GreenGenes used for analysis lists chloroplastic organisms under the cyanobacteria category; some of the cyanobacterial counts are, in fact, derived from photosynthetic eukaryotic organisms.


Comparing analyses


Now we will compare these analyses with those for a sample taken at 2 m depth from the same geographical location.

 In a new tab or window, find and open the **Ocean Sampling Day (OSD) 2014** project page again. Find the sample **OSD80_2014-06-21_2m_NPL022** and examine metadata on the Overview page.


 **Question 14:** Other than sampling depth, what are the differences between this sample and **OSD80_2014-06-21_0m_NPL02?**


 Scroll to the Associated runs section, and click on **ERR770970**. Open the **Functional analysis** tab and examine the Sequence feature and InterPro match summary information for this run.

 **Question 15:** How many pCDS were in this run?

 **Question 16:** How many of the pCDSs have an InterPro match?

 **Question 17:** How many different InterPro entries are matched by this run?

 **Question 18:** Are these figures broadly comparable to those for the previous sample?

 Now we are going to look at the differences in slimmed GO terms between the 2 runs. There are two ways to do this. First, you can simply scroll to the bottom of the page and examine the GO term annotation (note - selecting the bar chart representation of GO terms makes it easier to compare different data sets). Alternatively, you can use the **comparison tool**, which allows direct comparison of runs within a project. The tool can be accessed by clicking on the '**Comparison Tool**' tab, illustrated in Figure 6 below. At present, the tool only compares slimmed GO terms, but will be expanded to cover full GO terms, InterPro annotations, and taxonomic profiles as development of the site continues.

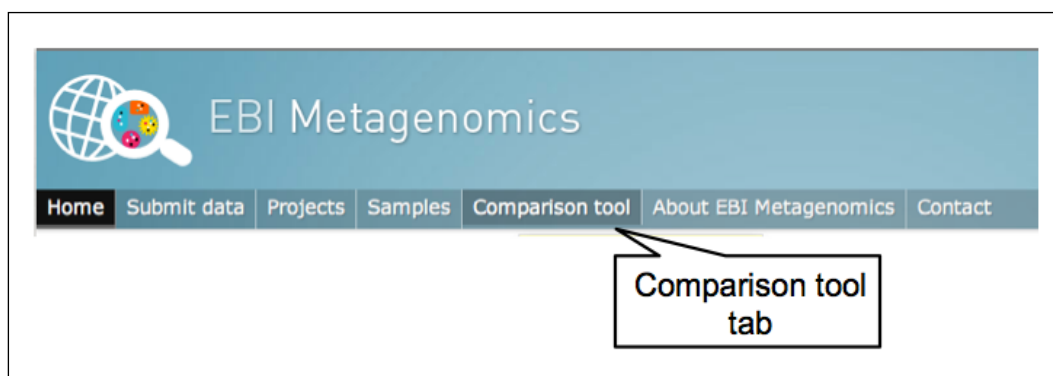





Figure 6. Clicking the appropriate tab takes you to the comparison tool page.

 Click on the Comparison tool tab and choose the **Ocean Sampling Day (OSD) 2014** project from the sample list and select the **OSD80_2014-06-21_0m_NPL02 - ERR770971** and **OSD80_2014-06-21_2m_NPL022 - ERR770970**.


 **Question 19:** Are there visible differences between the GO terms for these runs. Could there be any biological explanation for this?

 Navigate back and open the taxonomic analysis results tab for each run.

 **Question 20:** How does the taxonomic composition differ between runs? Are any trends in the data consistent with your answer to question 19?

Visualising taxonomic data using MEGAN

Next, we are going to look at the taxonomic predictions for all of the runs. To do this, we are going to load them into **MEGAN**.

 MEGAN is a tool suite that provides metagenomic data analysis and visualization. We are going to use only a small subset of its features, relating to taxonomic comparison. Detailed information on MEGAN and the analyses and visualisations it offers can be found here: <http://ab.inf.uni-tuebingen.de/data/software/megan5/download/manual.pdf>

MEGAN can be downloaded from <http://ab.inf.uni-tuebingen.de/software/megan5/> (it requires a licence, but this is free to academic users). However, it should already be installed on your Desktop.



Click on the MEGAN icon on your desktop to load the s/w.

We now need to download the full taxonomic predictions for all of the runs in the Ocean Sampling Day project.



Navigate to the **Project: Ocean Sampling Day (OSD)**... page, using the breadcrumb link at the top of the page. Click on the **Analysis summary** tab, which will take you to a set of tab separated result matrix files, summarising the taxonomic and functional observations for all runs in the project.



Click on the **Taxonomic assignments (TSV)** link, which will download the corresponding file to your downloads folder.



In order to get the sample names to display properly in MEGAN, we need to tweak a field in the file we have just downloaded. We can do this using a text editor or via the command line.



Open the Terminal, navigate to the Downloads directory and take a look at the file you have just downloaded:

```
cd ~/Downloads
```

```
less ERP009703_taxonomy_abundances_v2.0.tsv
```

You will see it says 'taxonomy' in the first column on the first row.

Type (all on one line):

```
sed -i 's/taxonomy/#SampleID/'  
ERP009703_taxonomy_abundances_v2.0.tsv
```

Look at the file again, using less. 'taxonomy' should now be replaced with '#SampleID'.



From the MEGAN menu, choose 'File' and 'Import'. Select 'CSV Import' and then find the file you have just downloaded and edited.



From the pop up menu, choose the 'Class, Count' option under format, set the separator as 'Tab' and select the Taxonomy classification and press 'Apply' (see **Figure 7** below).

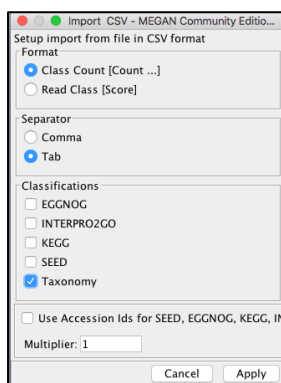


Figure 7. MEGAN settings to import EMG project-level taxonomy results files.



There are many different visualisations and comparisons that can be performed using MEGAN, so feel free to explore the data using the tool.



Question 21: Do any samples contain taxa not found in other samples?



Question 22: Can you discern any patterns in the geographical distribution of certain species (for example, the cluster of samples enriched for *Lactobacillus* or *Maricaulis* sp., compared to other samples)?



The bubble chart visualisation can be useful when comparing a large number of samples. You can access this by clicking the 'show chart' button on the main pane and choosing 'Show bubble chart' from the sub menu.

Searching EBI Metagenomics using indexed metadata



We are now going to take a look at which other datasets in EBI Metagenomics that *Lactobacilli* are found in. Point your browser at <https://www.ebi.ac.uk/metagenomics/search/>



This interface allows you to search the project, sample and run related metadata and analysis results for all of the publicly available datasets in the EBI Metagenomics resource.



Click on the 'Runs' tab. You should now see a number of run-related metadata search facets on the left hand side of the page, including 'Organism'. Click on the 'More...' option under Organism, scroll down to 'Lactobacillus' in the pop-up window and select the check box next to it. Now click 'Filter'. The results page should now show all of the runs that have taxonomic matches to *Lactobacilli*

in their datasets. The 'Biome' facet on the left hand side of the page now shows the number of matching datasets in each biome category (to save space, the 10 biome categories with the most matching datasets are shown by default).



Question 23: Which biome category has the most datasets that contain *Lactobacilli*?



Question 24: How well does this correlate with what's known about these bacteria?



Finally, we will try to use the interface to find functional proteins present under certain environmental conditions. For example, InterPro entry IPR001087 represents a domain found in GDSE esterases and lipases, which are hydrolytic enzymes with multifunctional properties.



Question 25: Using the search interface, can you identify the metagenomics datasets sampled from ocean sites at between 10 and 15 degrees C that contain these enzymes?



Question 26: Can you envisage ways in which this kind of search functionality could be used for target / enzyme discovery?