Tutorial

Whole Metagenome Functional Analysis

June 28, 2018

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com · ts-bioinformatics@qiagen.com

Tutorial

Whole Metagenome Functional Analysis

This tutorial will take you through the different tools available in CLC Microbial Genomics Module available for CLC Genomics Workbench and Biomedical Genomics Workbench to perform a whole metagenome functional analysis pipeline.

Introduction The main goal of the tutorial is to demonstrate the assembly of metagenomes derived from two different groups of samples and the subsequent investigation of functional differences. It serves as a template for performing a comparative investigation into the functional composition and diversity of microbial communities. The tools provide a way of looking at different samples in aggregate views and to drill down into differentiating functional categories that result from the comparative analysis.

Prerequisites For this tutorial, you will need either CLC Genomics Workbench or Biomedical Genomics Workbench with MetaGeneMark plugin and CLC Microbial Genomics Module installed.

Overview In this tutorial we will go through a suite of useful components in pipelines for analyzing whole-metagenome NGS data from microbial communities.

- First, we will import "raw" NGS sequencing data into the workbench and prepare the samples for analysis.
- We will then assemble the reads using **De Novo Assemble Metagenome** into contigs.
- With **MetaGeneMark**, we will identify genes and coding DNA sequences (CDS) on the contigs.
- Subsequently, functional annotation of the CDS with Gene Ontology (GO) terms and Pfam domains will be performed with **Annotate CDS with Pfam Domains**.
- Based on the annotations, we will construct a Gene Ontology profile using the **Build Functional Profile** tool for measuring functional diversity.
- We will also create a multi-sample abundance table using **Merge Abundance Tables**.
- Finally, we will set up the data for additional statistical analyses and visualisations.

Downloading and importing the data For this tutorial we will make use of the mock dataset generated by Lindgreen et al., 2016. The dataset contains four samples, divided in two groups (A and B). Group A is enriched in bacteria that perform photosynthesis and nitrogen fixation, while group B is enriched in pathogenic bacteria. The goal of the analysis in this tutorial is to find those functional differences from whole-metagenome sequencing data.

For sake of speed, the dataset used for this tutorial is a small subset of reads that is related to the following functional categories:

- **Photosynthesis**, identified by the GO id 0015979, which defines it as *"the synthesis by organisms of organic chemical compounds, especially carbohydrates, from carbon dioxide (CO_2) using energy obtained from light rather than from the oxidation of chemical compounds."*
- **Nitrogen Fixation**, identified by the GO id 0009399, which defines it as *"the process in which nitrogen is taken from its relatively inert molecular form (N_2) in the atmosphere and converted into nitrogen compounds useful for other chemical processes, such as ammonia, nitrate and nitrogen dioxide."*
- **Pathogenesis**, identified by the GO ids 0009405 and 0009403, which are defined as *"the set of specific processes that generate the ability of an organism to cause disease in another"* and *"the chemical reactions and pathways resulting in the formation of toxin, a poisonous compound (typically a protein) that is produced by cells or organisms and that can cause disease when introduced into the body or tissues of an organism,"* respectively.

The tutorial data consists of the following files:

- **Sequence data**: 4 pairs of fastq files (two for group A and two for group B). The files contain simulated paired-end sequencing reads.
- **Metadata**: "Group metadata.xls". A spreadsheet containing metadata information about the samples and the group they belong to.
- **Pfam database**: "Pfam-A v29 - Tutorial subset.clc". A small subset of Pfam v29 [Bateman et al., 2004, Finn et al., 2016] containing only the Pfam domains relevant to this tutorial.
- **GO database**: "GO database - Tutorial subset.clc". A small subset of the GO (Gene Ontology) database [Ashburner et al., 2000, The Gene Ontology Consortium, 2015] and Pfam2GO mappings (which allow to infer GO terms from Pfam domains). This database contains only terms relevant for this tutorial.

Please note that this tutorial contains only a small subset of the Pfam and GO databases, which should *only* be used for this tutorial. For applying the functional analysis pipeline to real datasets, please download the *complete* databases: the complete versions of the Pfam and GO databases can be easily obtained using the **Download Pfam Database** and **Download GO Database** tools, respectively. If you are not sure where to find these tools in the toolbox, use the Launch button () .

Now that the prerequisites have been described, it is time to start importing the input data.

1. Download the sample data from our website: `http://resources.qiagenbioinformatics.com/testdata/functional_tutorial.zip` and unzip it.

2. Start your CLC Workbench and create a folder for storing input data and results, named for example **Functional analysis**.

3. Go to **Import** | **Illumina** to import the 8 sequence files (ending with "fastq") (figure 1).

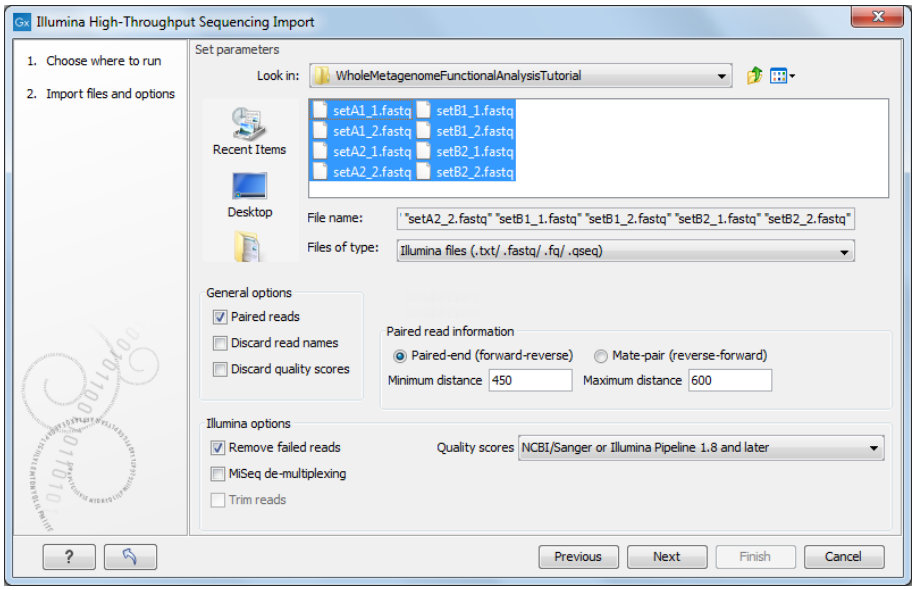


Figure 1: *Import paired-end reads for the four samples.*

- The import type under Options is set to **Paired reads**.
- **Paired-end (forward-reverse)** is selected.
- **Discard read names**, **Discard quality scores** and **MiSeq de-multiplexing** are not checked.
- Set the minimum distance to 450 and the Maximum distance to 600.

4. Click on the button labeled **Next** and select the location where you want to store the imported sequences. You can check that you have now 4 files labeled as "paired"

5. Import the metadata by clicking **Import | Import Metadata** on top of the Navigation Area.

6. A wizard opens (figure 2). Select the spreadsheet Group metadata.xls in the first field. The content fills the Medatata preview table at the bottom of the dialog. Click **Next**.

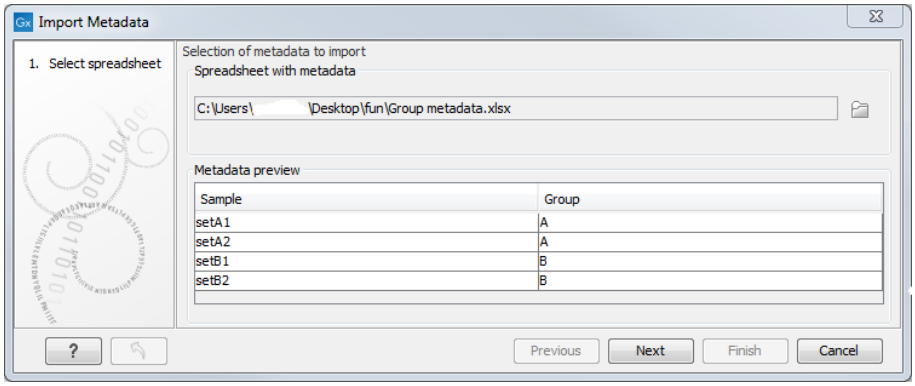


Figure 2: *Import the metadata.*

7. Now select the four reads imported earlier. Check **Partial** for the Data association preview table to fill in and thereby indicate that the association was successful (figure 3). Click **Next**.

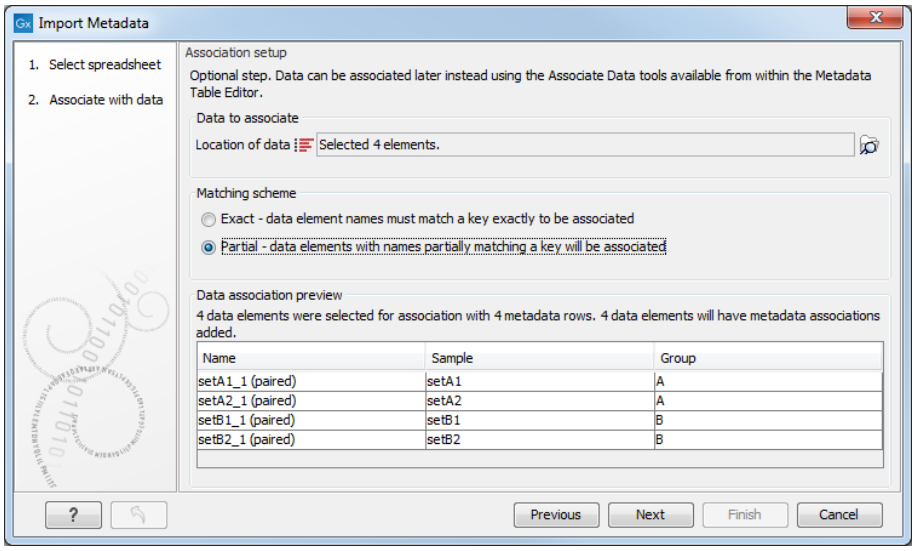


Figure 3: *The metadata and the reads are now associated.*

8. Save the metadata in the folder created earlier. You can rename the metadata table "Group metadata" instead of the default name "Samples".

9. Import the Pfam and GO databases by dragging the ''*GO database - Tutorial subset.clc* ()'' and ''*Pfam-A v29 - Tutorial subset.clc* ()'' files into your destination folder in your workbench, or by using the Standard Import button on top of the Navigation Area.

All of the data needed to get started is now imported and you should have the objects depicted in figure 4. You are now ready to begin the analysis.

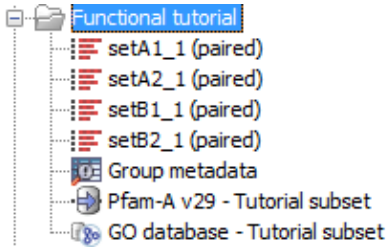


Figure 4: *All files are now imported.*

Assembling and annotating references

In this section, we will assemble the reads into contigs and annotate them with functional information.

1. Create a new folder, for example **Assemblies**, to store the results. We are now ready to assemble the reads into contigs using the De Novo Assemble Metagenome tool:

   Metagenomics ( ) | **De Novo Assemble Metagenome** ( )

2. Since we are processing four samples, select the **Batch** option. Next, select the reads as input (figure 5) and click **Next**.

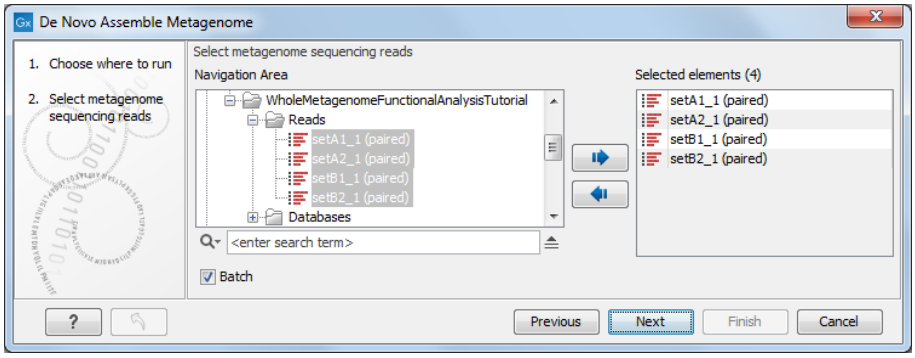


Figure 5: *Batch reads for de novo metagenome assembly.*

3. The next window gives an overview of the batch units. If you do not see this window, it means you forgot to check the batch option at the previous step.

4. In the "De novo options" dialog, make sure **Minimum contig length** is set to 200, choose the **Longer contigs** execution mode and make sure the **Perform scaffolding** checkbox is not checked (figure 6). Click **Next**.

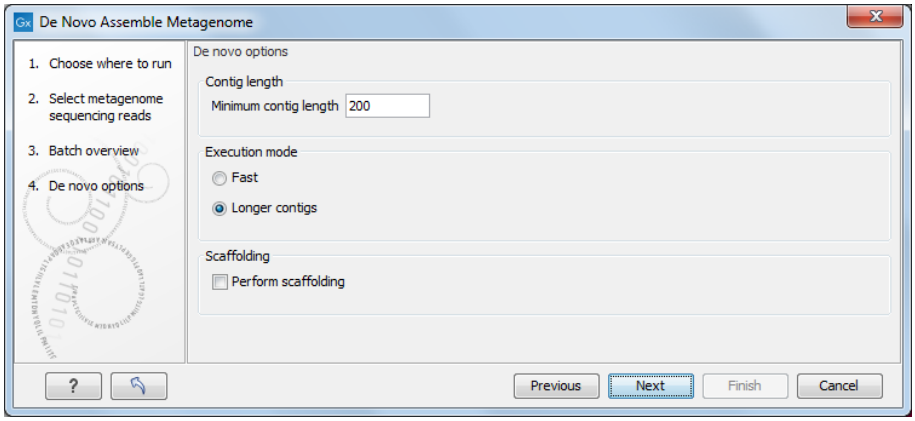


Figure 6: *Select parameters for de novo metagenome assembly.*

5. Choose to save the results in the **Assemblies** folder.

Once the reads have been assembled, we need to functionally annotate the contigs. Before annotation with functional information, we need to identify coding regions in the contigs. We therefore run the MetaGeneMark tool to identify genes and coding DNA sequences (CDS).

1. Go to:

   GeneMark Gene Finding | **MetaGeneMark**

2. Enable the **Batch** option and select the four contig list generated in the previous step (figure 7).

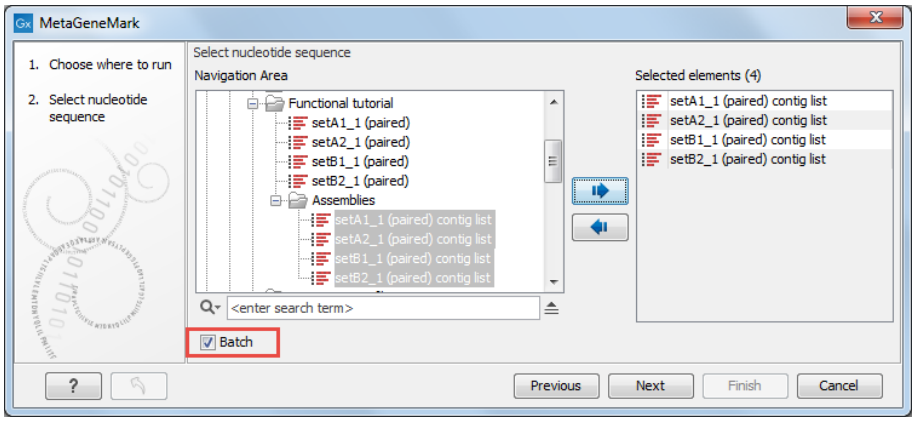


Figure 7: *Remember to select the batch option, right-click on the "Assemblies" folder and choose to "Add folder contents".*

3. The next window gives an overview of the batch units. Click **Next**.

4. In the next dialog, use the default parameters, namely the genetic code "11" (i.e. the genetic code used by bacteria, archea and plant plastids) (figure 8). Click **Next**.

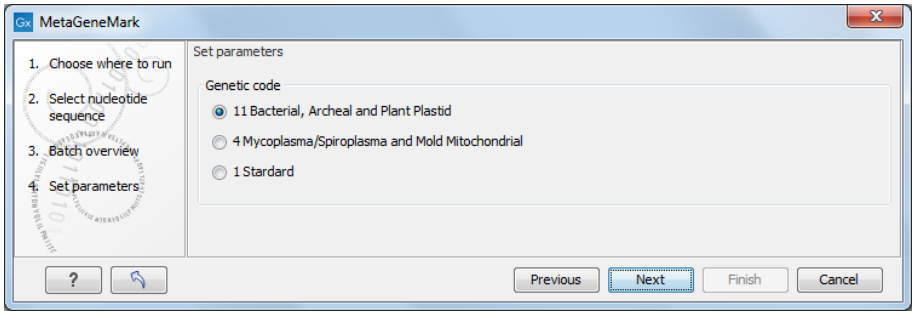


Figure 8: *Default parameters for the De Novo Assemble Metagenome.*

5. Next, keep ".genemark" as output name extension and save the results into a separate folder, for example called "Annotated Assemblies" (figure 9).

In the next step, we will annotate the CDS with Pfam domains and GO terms by using the **Annotate CDS with Pfam Domains** tool.

1. Go to:

   Metagenomics | **Functional Analysis** | **Annotate CDS with Pfam Domains**

2. Enable the **Batch** option again and select the four contig lists generated in the previous step (i.e., the ones ending in ".genemark") (figure 10).

3. Click **Next** at the Batch Overview dialog.

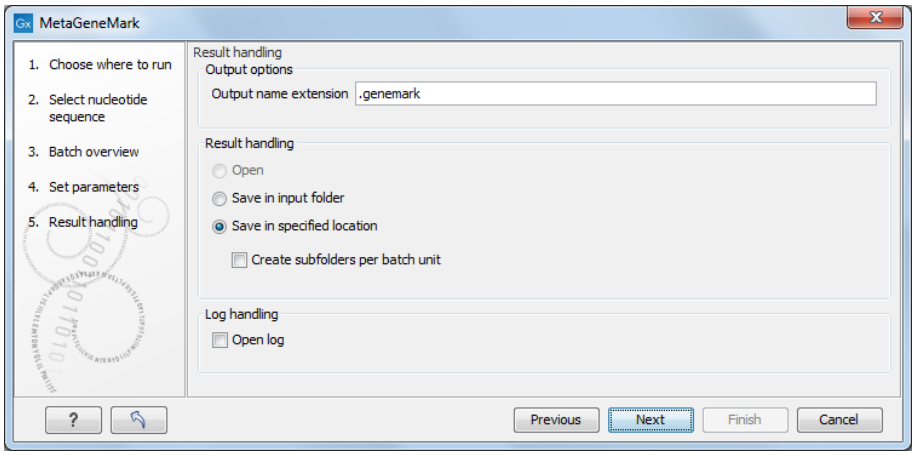


Figure 9: *Default parameters for the MetaGeneMark tool.*

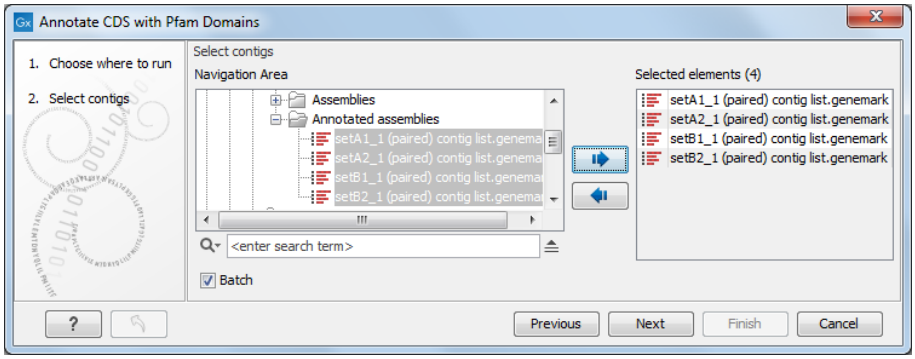


Figure 10: *Remember to select the batch option, right-click on the "Annotated Assemblies" folder and choose to "Add folder contents".*

4. Use the ''*Pfam-A v29 - Tutorial subset.clc* ( )'' as pfam database and ''*GO database - Tutorial subset.clc* ( )'' as GO database, as shown in figure 11. Make sure the genetic code is set to "11 Bacterial, Archeal and Plant Plastid" and that "Use profile's gathering cutoffs" and "Remove overlapping matches from the same clan" are checked. You can keep "Complete GO basic" as GO subset.

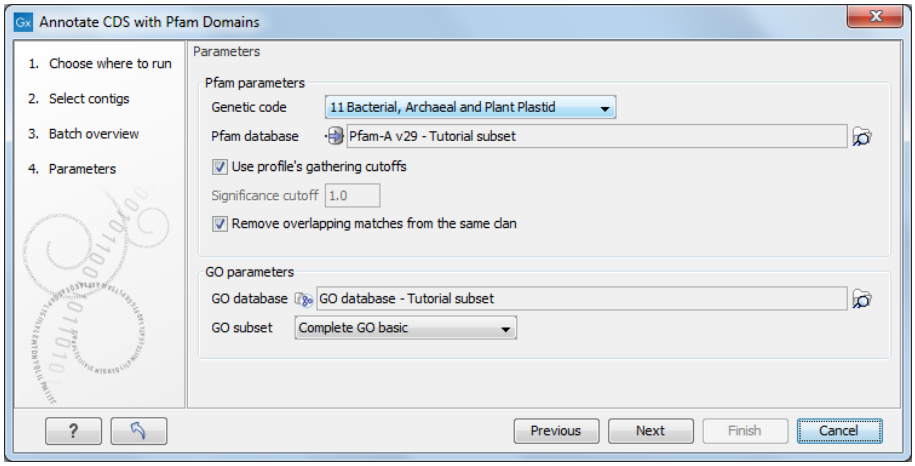


Figure 11: *Parameters for Annotate CDS with Pfam Domains.*

5. Choose where you will save the reports and click **Finish**.

The tool will add annotations to the existing contigs and create a report for each batch unit. You can check that Pfam annotations have been added by opening ‘‘*set_A1_1(paired) contig list.genemark (Pfam)* ( )’’. To see Pfam annotations, open the **Annotation Type** tab on the right panel and click on **Pfam domain**. In the Find tab, select "Annotation" and type in "Pfam" to find all Pfam annotations on each contig. If you hover over a Pfam annotation, you will be able to see the name of the Pfam domain, its description and the score of the match. When the Pfam domain can be matched to a GO term, a GO annotation will also be present, as shown in figure 12.

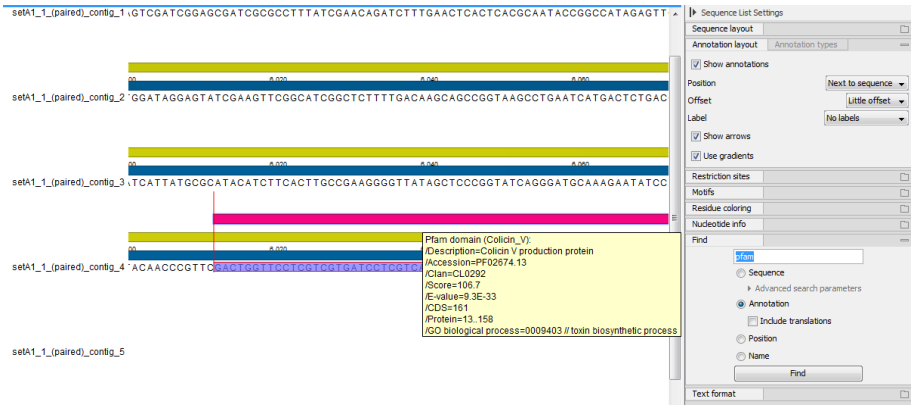


Figure 12: *Contig_4 in set_A1 contains a Colicin V production protein domain, which is related to toxin biosynthesis.*

The tool also generates a table that recapitulates the found Pfam annotations, as well as a description and accession number associated with each annotation (figure 13).

Rows: 8 Pfam results Filter

Sequence	Start	End	Accession	Score	E-value	Description
setA1_1_(paired)_contig_1 CDS 34	485	530	PF08369.7	75.00	3.80E-23	Proto-chlorophyllide reductase 57 kD su...
setA1_1_(paired)_contig_2 CDS ...	469	694	PF00223.16	34.60	6.40E-11	Photosystem I psaA/psaB protein
setA1_1_(paired)_contig_3 CDS ...	416	468	PF08369.7	41.30	1.30E-12	Proto-chlorophyllide reductase 57 kD su...
setA1_1_(paired)_contig_4 CDS ...	11	158	PF02674.13	106.70	9.30E-33	Colicin V production protein
setA1_1_(paired)_contig_5 CDS ...	98	171	PF04319.10	104.30	2.10E-32	NifZ domain
setA1_1_(paired)_contig_6 CDS ...	0	40	PF06988.8	40.00	2.40E-12	NifT/FixU protein
setA1_1_(paired)_contig_7 CDS ...	50	208	PF04891.9	186.60	3.80E-57	NifQ
setA1_1_(paired)_contig_12 CD...	0	76	PF03543.11	87.00	1.40E-26	Yersinia/Haemophilus virulence surface ...

Figure 13: *Table compiling the Pfam results.*

Building functional profiles

The metagenome assemblies are now annotated. We now want to re-map the original reads to estimate the abundance of functional categories in the samples.

First, map the reads to the annotated assemblies. In this case, we will not batch the execution, but we will need to run the read mapping four times against the 4 different references.

1. Create a folder called **Read mappings** to store the results in. Open the **Map Reads to Reference** tool using the Launch button ( ) .
2. Select ''*set_A1_1 (paired)* ( )'' from the Reads folder as input reads and click **Next**.
3. Select ''*set_A1_1 (paired) contig list.genemark* ( )'' from the "Annotated Assemblies" folder as reference as shown in figure 14). Keep "No masking" checked and click **Next**.

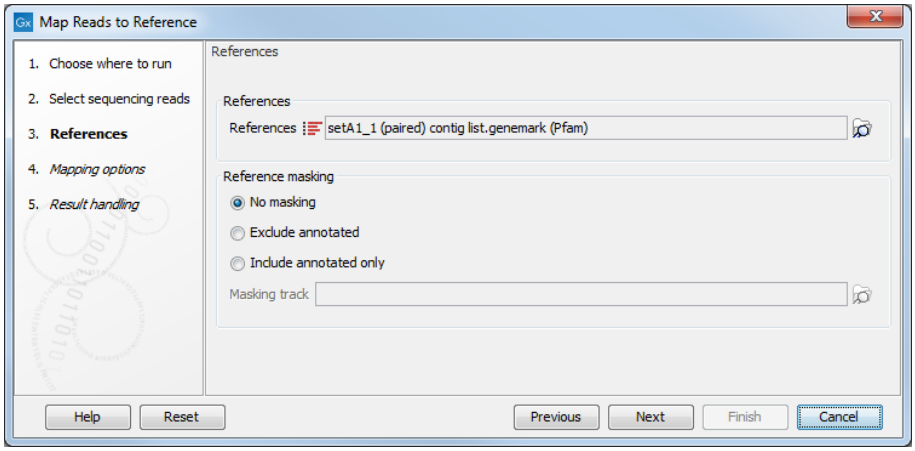


Figure 14: *Select the annotated reference.*

4. Keep the Mapping options at their default values. You can use the **Reset** button if you are not sure whether you have previously changed the parameters for the tool.
5. In the result handling window (figure 15), choose to save the results in the new *Read Mappings* folder you created. If you are working on CLC Genomics Workbench, , choose **Create stand-alone read mappings**. Stand-alone read mappings ( ) are preferable because they allow to run Build Functional Profile without having to specify a reference. If you are using the Biomedical Genomics Workbench, you will have to save the results as reads tracks ( ) and you will have to specify the annotated reference as a parameter in the next step.
6. When the mapping is complete, the read mapping ''*setA1_1 (paired) mapping* ( )'' will be created. Repeat the same procedure for *set_A2*, *set_B1*, and *set_B2*, making sure to use the correct annotated reference for each sample.

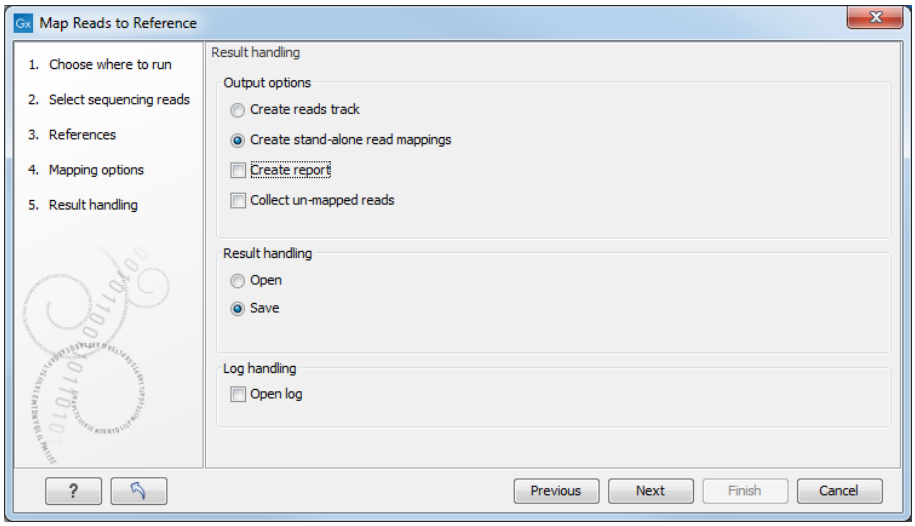


Figure 15: *Save as stand-alone read mappings in CLC Genomics Workbench.*

Next, we build the GO functional profile for each sample using the Build Functional Profile tool.

1. Create a folder called **Functional profiles** to store results and go to:

   Metagenomics ( ) | **Functional Analysis** ( ) | **Build Functional Profile** ( )

2. On Biomedical Genomics Workbench you cannot perform this step in batch, so select just one mapping and click **Next**. Working with CLC Genomics Workbench, enable the **Batch** option and select the four read mappings (figure 16). Click **Next** twice to pass the Batch overview window.

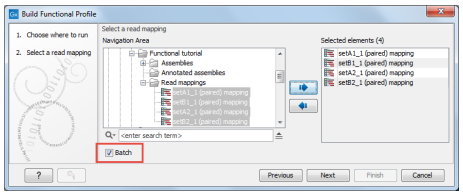


Figure 16: *Select the four read mappings and analyze them in batch.*

3. If you are running Biomedical Genomics Workbench, specify the annotated reference (e.g. ''*set_A1_1 (paired) contig list.genemark* ( )'' for set_A1) in the **Reference** parameter, and repeat the analysis with the appropriate reference for each read mapping. If you are working with CLC Genomics Workbench, you do not need to specify a reference in the "Parameters" dialog.

   In both cases, use ''*GO database - Tutorial subset.clc* ( )'' as GO database, as shown in figure 17.

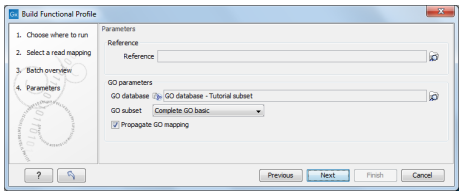


Figure 17: *Parameters for Build Functional Profile.*

4. Finally, choose **Create GO functional profile** only and save to the "Functional profiles" folder (figure 18).

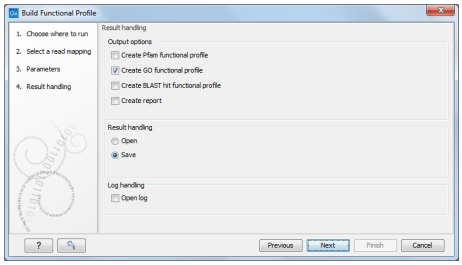


Figure 18: *Create only a GO functional profile.*

You have now built a functional profile for each sample (figure 19).

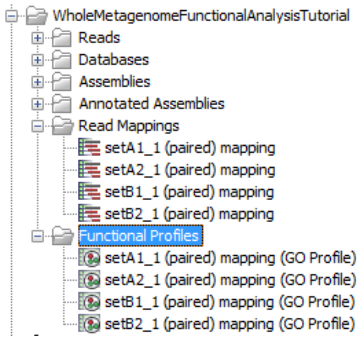


Figure 19: *GO functional profiles have been created.*

We now want to merge them using the Merge Abundance Tables tools.

1. Create a folder to store the results (e.g. **Statistical analyses**) and go to:
   Metagenomics ( ) | **Functional Analysis** ( ) | **Build Functional Profile** ( )
2. Select the four GO profiles as input. In this case, we do not batch the analysis, as the tools should be run only once using all the four functional profiles as input (figure 20).

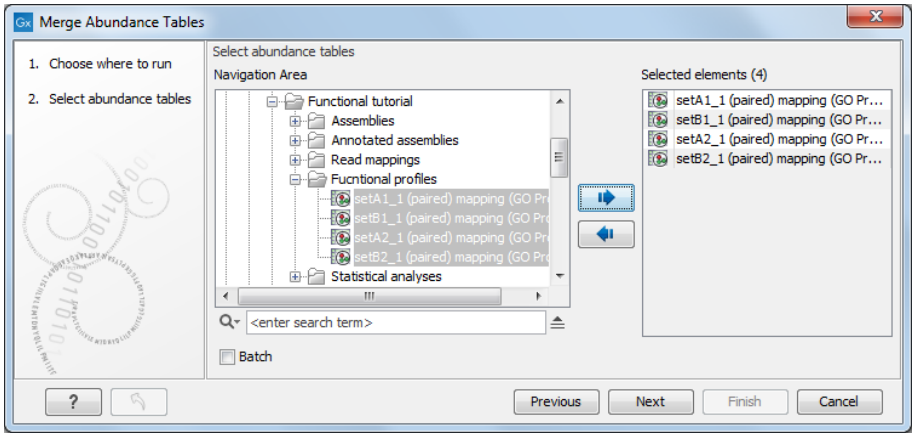
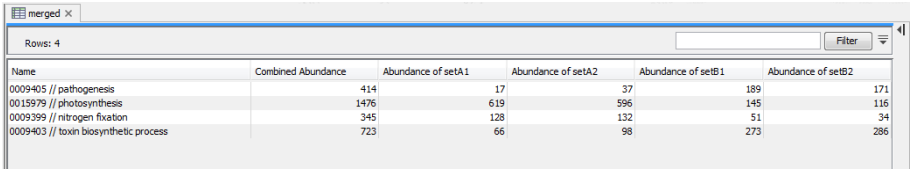


Figure 20: *Merge the GO functional profiles.*

3. Save the merged profile in the "Statistical analyses" folder.

The tool will create an abundance table called ‘‘*merged* ( )’’ containing functional abundances for the four samples. You can now open the table to explore the results of the functional analysis (figure 21). Observe the functional abundance values for the four GO terms. As expected, abundance values for "pathogenesis" and "toxins" are higher in group B, whereas group A is enriched in "photosynthesis" and "nitrogen fixation".



Name	Combined Abundance	Abundance of setA1	Abundance of setA2	Abundance of setB1	Abundance of setB2
0009405 // pathogenesis	414	17	37	189	171
0015979 // photosynthesis	1476	619	596	145	116
0009399 // nitrogen fixation	345	128	132	51	34
0009403 // toxin biosynthetic process	723	66	98	273	286

Figure 21: *Result of the GO functional analysis.*

Performing statistical analyses

A heat map and dendrogram help assessing similarity between samples.

1. Open the **Microbial Genomics Module ( )** | **Metagenomics ( )** | **Abundance Analysis ( )** | **Create Heat Map for Abundance Table ( )**

   and choose the "merged" table as input.

2. Leave the parameters as set by default, i.e., the distance to Euclidean and clusters to Complete linkage. Click **Next**.

3. In the next wizard window, do not set any particular filter by selecting the option "No filtering" and click **Next**.

4. Save the result in the **Statistical analyses** folder.

Display the heat map by double-clicking on it in the Navigation Area (figure 22).

As we would have expected from the description of the data-set by Lindgreen et al., 2016 in the beginning of this tutorial, the normalized values for toxin biosynthesis and pathogenesis are over-expressed in group B, while the normalized values for photosynthesis and nitrogen fixation are enriched in group A. Furthermore, the samples from each group cluster together, as shown in the dendrogram at the bottom of the figure.

It is also possible to use as additional statistical analyses the Differential Abundance Analysis tool, although the interest of a Venn diagram is quite limited when the data set is only made of two distinctive groups as it is for this tutorial.

Although the results are hardly surprising, it is always re-assuring and good scientific practice to first apply a method to a problem with a known solution in order to verify everything works out exactly as expected before moving on to harder problems. Well done! At this point, we’d like to point out again that it is important to download the full versions of the Pfam and GO databases (by using the tools provided in the workbench) prior to using the functional annotation pipeline for a complete functional analysis of your own datasets.

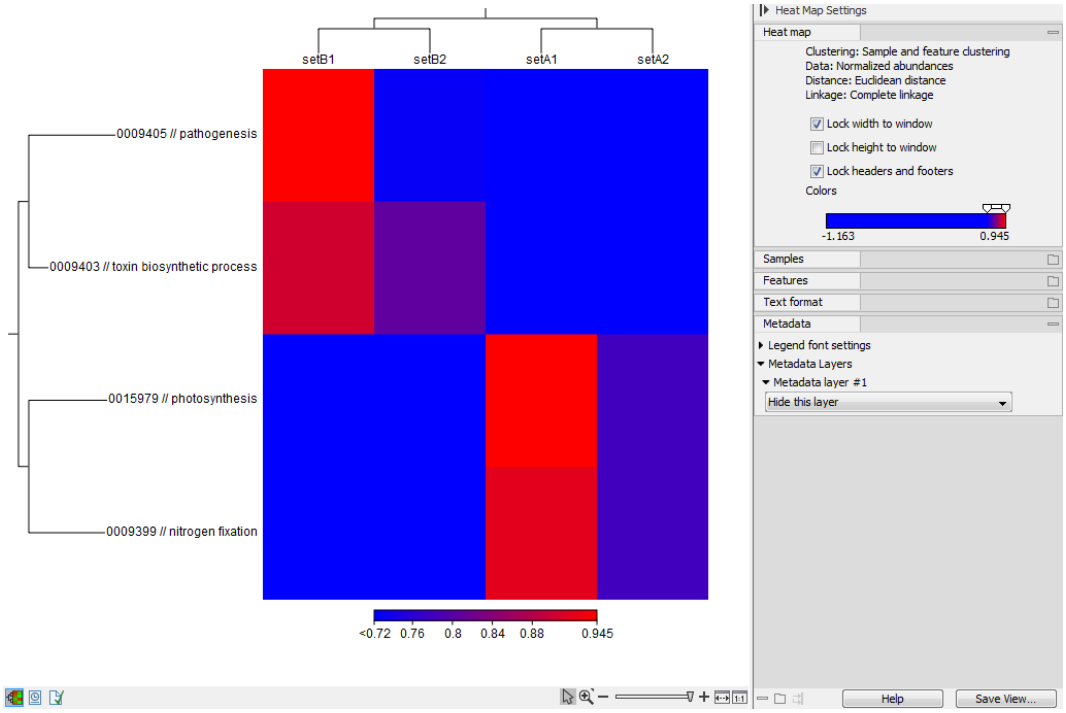


Figure 22: *Heat map from the abundance table.*

Bibliography

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29.

[Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–D141.

[Finn et al., 2016] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

[Lindgreen et al., 2016] Lindgreen, S., Adair, K. L., and Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6:19233–.

[The Gene Ontology Consortium, 2015] The Gene Ontology Consortium (2015). Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056.