**FAQ for Comparative\_db**

**1) How the pipeline works**

*-What is* *Comparative\_db?*The Comparative\_db is a customizable comparative genomics database which enables the user to compare and analyse genomes. Comparative\_db integrates multiple tools for comparative genomic analyses and facilitates the access to comprehensive annotations from multiple databases including UniProt (curated and automated protein annotations), KEGG (annotation of pathways) and COG (orthology). Next to this, phylogenetic trees and phylogenetic relationships of genes, conservation of gene neighbourhood and taxonomic profiles can be visualized using dynamically generated graphs, available for download.

*-How does Comparative\_db work?*  
Comparative\_db consists of two parts:  
1. generation of a SQL database through a Nextflow–based pipeline applied on two or more input .gbk files. (see question ‘*-Which analysis does the pipeline include?’)*  
2. launch of an overview and interacting webpage that displays the results retrieved in step 1.

*-What is the input format file?*  
Standard Genbank file that contains genome sequences and annotations.  
Only the extention “.gbk” is used, consequently Genbank files downloaded from NCBI, with the extention “.gbff” must be renamed as “.gbk”.

You can annotate a genome and generate the corresponding “.gbk” file using Prokka (see Prokka documentation – put citation).

*-Which input files do I need to provide?*

1. STEP 1 – generation of SQL database:

-Input tabel (.csv/.tsv ):  
it contains an identifier for each input file and the local or absolute path of the gbk files.   
Download an example of input table here ( link to an example table)  
For a detailed explanation go here (link to the documentation).  
-Config file  
it let you control which analyses among those classified as ‘optional’ in question X (How does Comparative\_db work?) will be run.  
Download an example of Config file here (link).  
Rely on the Documentation to set up this file in the proper way. (link Documentation)

1. STEP 2 – launch of the web app  
   settings file (not sure if the user will need to modify it or it will be automatically linked to the newly generated db)

*-Which tools do I need to install to run the pipeline?*  
You need to install Nextflow (see Nextflow documentation – install – put link) in your environment for running the annotation analysis.  
For launching the webapp you need to create an environment based on the chlamdb.yml file provided.  
Please see the documentation

---- What about providing a single environment with Nextflow and everything? ----

*-How many genomes can I compare?*There is no limits on the number of genomes given as input

-How long does it take to run the pipeline?  
Test with x number of genomes, size of genomes, cpu to run it, etc (to do)

*-Which analyses does the pipeline include?*  
The pipeline performs the following analyses:  
-Orthologous proteins identified with [OrthoFinder](https://github.com/davidemms/OrthoFinder) (optional)  
-Detailed annotation of proteins based on   
 [KEGG](https://www.genome.jp/kegg/) (Kegg orthologs) - (optional)  
 [COG](https://www.ncbi.nlm.nih.gov/COG/) (clusters of orthologous groups) - (optional)  
 [SwissProt](https://www.uniprot.org/) (manually annotated proteins) (NOT SURE).  
-Precomputed homology searches with [RefSeq](http://155.105.138.104:8084/) and [SwissProt](http://155.105.138.104:8084/) databases for each protein   
-Precomputed phylogenetic reconstructions of orthologous groups  
-COG phylogenetic profiles  
-Generation of customized databases for blast search (optional)

For a detailed description link (Bastian’s documentation)

*-Can I customize the analyses I want to perform?*The user can select which analyses she/he wants to perform turning on or off some of the parameters of the configuration file. See the list of ‘optional’ parameters in Q-‘Which analyses does the pipeline include?’and refer to documentation (link)

**2) How to use the Comparative\_db web**

*-How do I identify which proteins are part of a specific orthogroup?  
-How do I use the search bar? – list of terms (cog, kegg ,accession)*

-Other analysis ...

**3) Biological aspects**

*-Are plasmids identified?*  
Genes belonging to plasmids are reported only if this feature is annotated in the input file, otherwise no distinction between chromosome and plasmids is done.

-Orthogroups accurate (?)