Gene set completeness with fCAT

Assessing the completeness of sets of proteins, transcripts, but also of genomes is a routine task in genome-scale analyses. The amount of missing gene, i.e. of genes that are expected to be present, but which are not, is used as a standard quality criterion, where more missing genes are interpreted as lower quality data. BUSCO, a standard software for such analyses uses a set of nearly ubiquitously represented single copy genes as the underlying data set. A unidirectional search determines then if *sufficiently similar* sequences exist in the test set such that the gene can be considered present. (Read here for more on BUSCO).

BUSCO leaves open some gaps that we address with the tool fCAT. There are four main advantages of fCAT.

1. users can design their own set of core genes
2. fCAT uses ortholog assignments instead of only a unidirectional search, and because of is not restricted to single copy genes
3. fCAT has four scoring modes (still in beta) that consider next to length differences also differences in the domain architecture of the orthologs
4. the results can be visually interpreted and integrated with the presence/absence pattern of the orthologs in other taxa

fCAT is still in beta testing, so we will use you a bit as guinea pigs to try out the software.

We hope that you like it 🙂

fCAT Core sets

For the course, we have prepared one core set of proteins that are prevalent[1] in **eukaryotes**

Core set eukaryota

This core set uses group identifier from OrthoDB v10

Species	NCBI Taxonomy ID	kingdom	Internal name
Cryptococcus neoformans	215684	fungi	CRYNE@214684@2
Rhizopus delemar	246409	fungi	RHIDE@246409@2
Chlamydomonas reinhardtii	3055	chlorophyta	CHLRE@3055@2
Arabidopis thaliana	3702	streptophyta	ARATH@3702@2
Amphimedon queenslandica	400682	metazoa	AMPQU@400682@2
Nematostella vectensis	45351	metazoa	NEMVE@45351@2
Sorghum bicolor	4558	streptophyta	SORBI@4558@2
Caenorhabditis elegans	6239	metazoa	CAEEL@6239@2

Species	NCBI Taxonomy ID	kingdom	Internal name
Homo sapiens	9606	metazoa	HOMSA@9606@2
Zymoseptoria tritici	336722	fungi	ZYMTR@336722@2
Saccharomyces cerevisiae	559292	fungi	SACCE@559292@2
Drosophila melanogaster	7227	metazoa	DROME@7227@2

Preparing the fCAT run

Prior to your fCAT analysis, you should make sure that the software is installed, and that all necessary files are present and in the correct format.

Checking the software

1. confirm that fCAT is installed on the system
   1. activate the conda environment:

      ```
      conda activate fdog
      ```

   2. type

      ```
      fcat -h
      ```

   3. If fCAT is not installed follow the installation guide lines
2. check for the correct setting of the environmental variable `COILSDIR`

   ```
   echo $COILSDIR
   ```

   - If this does not provide you with a path, then this variable is not set and fdog will not run properly.

     To solve this issue temporarily for the current shell, type

     ```
     export
     COILSDIR=/home/ubuntu/Share/fdog/annotation_tools/COILS2/coils
     ```

     To fix this change, add this line to your bash configuration file `.bashrc`. It will become active upon the next login, or by typing `source ~/.bashrc`. After 'sourcing' the `.bashrc`, you will have to re-activate the conda environment: `conda activate /home/ubuntu/anaconda3/envs/fdog`

Preparing the fCAT run

1. create a working directory for your fCAT analysis:

   ```
   mkdir -p $HOME/Analysis/fcat
   ```

2. change into the new directory:

```
cd $HOME/Analysis/fcat
```

3. soft-link[2)] the protein file you want to analyse into your working directory:

```
ln -sf
/home/ubuntu/Share/Analysis/GeneAnnotation/Results/metaeuk/Crypto_Metae
uk.fas .
```

4. We have already preformed this step for you.
    1. copy the directory harbouring the feature architecture annotations into your Analysis directory

    ```
    cd $HOME/Analysis/fcat
    cp -r /home/ubuntu/Share/ProteinSets/coredir/annotation_dir .
    ```

    2. Annotate the protein domains in the gene set of interest:

    ```
    fdog.addTaxon -f Crypto_Metaeuk.fas -n Crypa_metaeuk -i 5807 -o
    $HOME/Analysis/fCAT --annopath $HOME/Analysis/fcat/annotation_dir/
    --replace
    ```

    The second command will annotate protein features, such as PFAM and SMART domains, low complexity regions, transmembrane domains, etc. If you do not want to run the annotation, which will take a couple of minutes using 8 cores, copy it from `/home/ubuntu/Share/ProteinSets/fcat/CRYPA_METAEUK2@5807@240209.json` into $HOME/Analysis/fcat/annotation_dir/

    - fas.doAnno cannot find the COILS program. ! Check the environmental variable $COILSDIR
    - fas.doAnno throws errors about not being able to delete a file. ! Make sure that your fasta header are concise and don't contain any special characters

Running fCAT

To perform the fCAT analysis, perform the following steps

1. run the fCAT analysis on the AWS with the following core set
    1. eukaryota
2. for the **eukaryota core set** and the MetaEuk gene prediction on the CryPa_BCM2021a assembly, invoke the analysis with the following command[3)]:

```
fcat --coreDir $HOME/Share/ProteinSets/coredir/ --coreSet eukaryota --
refspecList "HOMSA@9606@2" --querySpecies Crypto_Metaeuk.fas --taxid
5807 --annoQuery
$HOME/Analysis/fcat/annotation_dir/CRYPA_METAEUK\@5807\@240209.json
```

The analysis will run for about **380 sec** when using 4 cores.

```
##### Generating reports...
Mode 1:
genomeID	similar	dissimilar	duplicated	missing	ignored	total
CRYPA@5807@240206	149	89	0	86	8	332

Mode 2:
genomeID	similar	dissimilar	duplicated	missing	ignored	total
CRYPA@5807@240206	141	97	0	86	8	332

Mode 3:
genomeID	similar	dissimilar	duplicated	missing	ignored	total
CRYPA@5807@240206	215	23	0	86	8	332

Mode 4:
genomeID	complete	fragmented	duplicated	missing	ignored	total
CRYPA@5807@240206	217	21	0	86	8	332
```

fCAT analysis - Output visualization and interpretation

fCAT in combination with PhyloProfile allows to visualize and explore the results of the geneset completeness analysis. Follow the steps below to download the data to your local computer and to open it in PhyloProfile.

You will find all pre-computed fCAT results at `/home/ubuntu/Share/Analysis/fCAT/fcatOutput/eukaryota`. Use these, if your analysis did not complete in time.

Downloading the data

Download the following three files from the fcat output folder, e.g. `$HOME/Analyses/fcat/fcatOutput/eukaryota/CRYPA@5807@250408/phyloprofileOutput` for the *eukaryota* dataset.

1. *.phyloprofile These files contains the information about the presence/absence of

orthologs to the genes in your coreset together with the domain architecture similarity scores. You will find the information for both your taxon of interest **and** the core taxa. **It is the main input file for PhyloProfile**. ❗ Choose the one that is represents the fCAT scoring mode you are interested in.

2. *.mod.fa ❗ This file contains the sequences of the orthologs in FASTA format
3. *.domains ❗ This file contains the feature annotations for the core genes and the orthologs. You will need this for visualization of the feature architectures in PhyloProfile

Opening the data in PhyloProfile

open the results for the *eukaryota* dataset in PhyloProfile. To do so, perform the following steps

1. open a shell on your local computer
2. startup ***R*** by typing R
3. in the R terminal load the package *PhyloProfile*: library (PhyloProfile)
4. start PhyloProfile: runPhyloProfile()
5. now, the start page of PhyloProfile should open in your default browser
6. upload the file *.phyloprofile[4] in the top left of the page. ❗ PhyloProfile will check the input.
7. If PhyloProfile sees a taxon for the first time, it will ask you to fetch the information from the NCBI taxonomy database. Once this is completed, please reload the page and upload the data again
8. upload the *domains file into the field at the lower left
9. specify the origin of group IDs you are using
    1. Dataset *eukaryota*: select **OrthoDB**
10. plot the results by clicking on ‘’Plot’’

Exploring the data

Remember, you are doing this analysis because you want to know

- ❓ which genes are missing
- ❓ how the missing genes are represented in the core species
- ❓ which genes display an unexpected domain architecture

Inspect

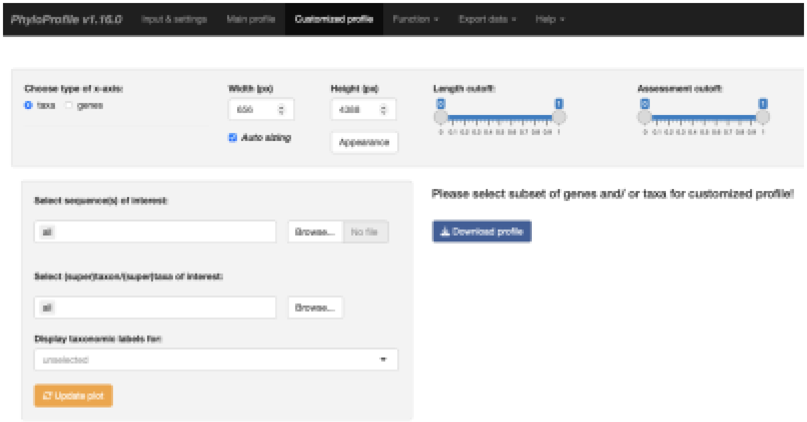
1. the overall pattern of the phylogenetic profiles and reconcile this with the text output of fCAT
2. you can click on individual dots in the profile to gain more information about the detected orthologs. This gives you the option to look up sequence and orthogroup information in the public database[5], and you can expect the domain architectures of the seed protein and the respective ortholog.
3. check out the tab ‘’Functions’’ in the top menu. It gives you, among others, the option to cluster

your phylogenetic profiles based on a variety of distance measures. Try this![6]

4. go back to the clustering function and use the mouse to select a clade in the clustering graph[7]. You will find that the corresponding genes appear in a table to the right. Check the box ''Add to custom plot'' and inspect your selection in tab custom profile
5. redo the selection, this time selecting all genes from the *eukaryota* dataset that are present in all core species but are absent in your *C. parvum* gene set[8]
6. if you do not find a single clade comprising all the genes that are missing in *C. parvum* do the following:
   1. Look for the file

      missing.txt

      in your fCat output folder
   2. go to the tab `Customised profile`
   3. find the button to upload a gene list for selecting a gene set of interest

      Choose type of x-axis: ● taxa ○ genes

      Width (px): 656

      Height (px): 4368

      ☑ Auto sizing

      Length cutoff: 0 – 1

      Assessment cutoff: 0 – 1

      Select sequence(s) of interest: all

      Select (super)taxon/(super)taxa of interest: all

      Display taxonomic labels for: unselected

      Please select subset of genes and/ or taxa for customized profile!
   4. upload the file `missing.txt`
   5. select *Homo sapiens* as the taxon of interest[9]

Download the data for the next analysis step

Download the information about the missing genes. We will need this for the last analysis

1. in the custom plot, select only *Homo sapiens* to be shown
2. go to the ''Download'' tab and download the information in FASTA format
3. inspect the file, and if all is good, upload it to your home directory at the AWS[10]

Navigation

Follow the links below to

- 1) Contamination check - taXaminer
- 2) Gene set completeness - BUSCO
- 3) Work package: Gene set analysis

Use the following links to navigate through the course

- Main page of the Physalia course
- Previous work package: Genome annotation
- Move to the next package: Phylogenetic profiles - fDOG

[1] missing in less than 10% of the core taxa

[2] ln -s

[3] This assumes that you are in $HOME/Analysis/fcat

[4] There are 4 different files, one each for the four modes of fCAT. Try out the **length mode** and **mode3**

[5] of course, this is possible only for groups and sequences for which a public database entry exists. Currently, we support OMA, orthoDB and NCBI

[6] you will have to recheck the box ‘’Sort sequences by ID’’ in the PhyloProfile landing page, though

[7] you may have to increase its height

[8] This requires some experimenting to find the correct clade in the tree, unfortunately

[9] you can play around with the selection of taxa

[10] use the command ‘’scp’’ for that

From:
https://applbio.biologie.uni-frankfurt.de/teaching/wiki/ - **Teaching**

Permanent link:
https://applbio.biologie.uni-frankfurt.de/teaching/wiki/doku.php?id=general:bioseqanalysis:genesetanalysis:fcat

Last update: **2025/04/08 16:34**