# Title of Dataset **Figure 1C**

[Access this dataset on Dryad](Dataset DOI link)

**Github link:** [**https://github.com/ADhabaliaAshok/EvolutionaryAssemblyFromProteinDomains/tree/main**](https://github.com/ADhabaliaAshok/EvolutionaryAssemblyFromProteinDomains/tree/main)

**Owncloud link:** [**https://owncloud.gwdg.de/index.php/s/dH3Y4MAHSfbmhrA**](https://owncloud.gwdg.de/index.php/s/dH3Y4MAHSfbmhrA)

Give a brief summary of dataset contents, contextualized in experimental procedures and results.

**Figure 1B shows a comparison among three approaches to annotate proteins and orthologs as stress-relevant**

## Description of the data and file structure

This is a freeform section for you to describe how the data are structured and how a potential consumer might use them. Be as descriptive as necessary. Keep in mind that users of your data might be new to the field and unfamiliar with common terminology, metrics, etc.

**This figure is divided into three approaches of annotating proteins and orthologs with:**

1. **Approach 1: *A.thaliana* and *P.patens* model organisms open source database**
2. **Approach 2: Gene Ontology open source tool known as Eggnogmapper**
3. **Approach 3: Intersection of Approach 1 and Approach 2**

Describe relationships between data files, missing data codes, other abbreviations used. Be as descriptive as possible.

## Sharing/Access information

This is a section for linking to other ways to access the data, and for linking to sources the data is derived from, if any.

Links to other publicly accessible locations of the data:

- [http://...](http://...)

**DatabaseFiles (Owncloud Link):** [**https://owncloud.gwdg.de/index.php/s/dH3Y4MAHSfbmhrA**](https://owncloud.gwdg.de/index.php/s/dH3Y4MAHSfbmhrA)

**Model organisms**

1. ***A.thaliana* database:** [**https://www.arabidopsis.org/download/list?dir=GO\_and\_PO\_Annotations%2FGene\_Ontology\_Annotations**](https://www.arabidopsis.org/download/list?dir=GO_and_PO_Annotations%2FGene_Ontology_Annotations)
2. ***P.patens* database:** [**https://peatmoss.plantcode.cup.uni-freiburg.de/ppatens\_db/downloads.php**](https://peatmoss.plantcode.cup.uni-freiburg.de/ppatens_db/downloads.php)

Data was derived from the following sources:

- []()

**The base input file needed is the 37 species clustered into orthogroups – obtained from Code part 1. The Orthogroups file is used to annotate the orthogroups and proteins as follows:**

1. **Approach 1 – from model organisms database, and orthogroup file**
2. **Approach 2 – fasta files of 37 species as input for Eggnogmapper, and orthogroup file**
3. **Approach 3 – orthogroup IDs from Approach 1 and Approach 2**

## Code/Software

This is an optional, freeform section for describing any code in your submission and the software used to run it.

Describe any scripts, code, or notebooks (e.g., R, Python, Mathematica, MatLab) as well as the software versions (including loaded packages) that you used to run those files. If your repository contains more than one file whose relationship to other scripts is not obvious, provide information about the workflow that you used to run those scripts and notebooks.

**Code part 1: Create orthogroups using OrthoFinder**

**orthofinder.py -f /directory\_with\_37species\_fasta\_files/**

**Code part 2.1: python3 make\_input\_3\_1.py**

**INPUT FILES REQUIRED (all files can be downloaded from DatabaseFiles):**

* **Ppatens\_GO - download file from *P.patens* database**
* **ATH\_GO\_GOSLIM\_2021\_removed4lines.txt – modified file from *A.thaliana***
* **PEATmossTAIR\_GOBP\_GORTS - download file from *P.patens* database**
* **PEATmossTAIR\_stress\_proteins – modified from Ppatens\_GO**
* **concat\_all\_files\_len13\_ipr – interproscanner files**
* **Orthogroups.tsv - from Code part 1**
* **species\_ordered\_list\_final\_m – input 37 species list**

**OUTPUT FILES:**

* **3\_1\_input\_file\_m**

**Code part 2.2: python3 make\_output\_3\_1.py**

**INPUT FILES REQUIRED:**

* **3\_1\_input\_file\_m**

**OUTPUT FILES:**

* **py\_plot\_AtPpOG\_m.png – distribution of stress-annotated orthogroups from Approach 1**

**Code part 3.1: python3 make\_input\_3\_2.py**

**INPUT FILES REQUIRED (all files can be downloaded from DatabaseFiles):**

* **database\_GOID\_GORTS – output from python3 create\_database\_GO\_GORTS.py**
* **Orthogroups.tsv – from Code part 1**
* **species\_ordered\_list\_final\_m - input 37 species list**

**OUTPUT FILES:**

* **3\_2\_input\_file**

**Code part 3.2: python3 make\_output\_3\_2.py**

**INPUT FILES REQUIRED:**

* **3\_2\_input\_file**

**OUTPUT FILES:**

* **py\_plot\_EggOG.png - distribution of stress-annotated orthogroups from Approach 2**

**Code part 4.1: python3 compare\_index.py**

**INPUT FILES REQUIRED:**

* **3\_1\_index\_final\_m - obtained from header of 3\_1\_input\_file\_m**
* **3\_2\_index - obtained from header of 3\_2\_input\_file**

**OUTPUT FILES:**

* **3\_3\_index\_overlaps**

**Code part 4.2: python3 make\_input\_3\_3.py**

**INPUT FILES REQUIRED:**

* **3\_1\_input\_file\_m – from Code part 2**
* **species\_ordered\_list\_final\_m - input 37 species list**
* **3\_3\_index\_overlaps – from Code part 4.1**

**OUTPUT FILES:**

* **3\_3\_input\_file\_m**

**Code part 4.3: python3 make\_output\_3\_3.py**

**INPUT FILES REQUIRED:**

* **3\_3\_input\_file\_m**

**OUTPUT FILES:**

* **py\_plot\_AtPpOG\_EggOG\_m.png - distribution of stress-annotated orthogroups from Approach 3**