# Title of Dataset **Figure 1B**

[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 the proteomic details and quality of our dataset of 37 species. Reference proteomes have been downloaded to perform this analysis.**

## 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 two parts:**

1. **BUSCO analysis of proteomes**
2. **Number of proteins:**
   1. **Proteins**
   2. **Proteins annotated by Interproscanner**
   3. **Stress Annotated Proteins**

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://...)

**Owncloud link https://owncloud.gwdg.de/index.php/s/dH3Y4MAHSfbmhrA**

Data was derived from the following sources:

- []()

**The input files to create figure are as follows:**

1. **BUSCO analysis of proteomes – obtained from Code part 1.1 and Code part 1.2**
2. **Number of proteins:**
   1. **Proteins - from input fasta file**
   2. **Proteins annotated by Interproscanner - from Interproscanner output files**
   3. **Number of Stress Annotated Proteins - data file from Figure 1C approach 3**

## 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.

**Final data files:**

1. **figure1\_BUSCO.csv**
2. **figure1\_table.csv**

**Code part 1.1: busco -m $MODE -i $IN -o $OUT -f -l chlorophyta\_odb10 -o Spinacia\_oleracea\_BUSCO\_out**

1. **IN**: **fasta file of each of 37 species**
2. **MODE**: **proteins**
3. **OUT**: **busco\_out**

**Code part 1.2**: **Concatanate the complete BUSCO scores (C) for every species into the final data file figure1\_BUSCO.csv**

**Code part 1.3**: **Create plot to show completeness of BUSCO scores for dataset of 37 species**

**Code part 2.1**: **interproscan.sh -i $IN -d /output\_directory/ -T /temporary\_directory/ -f tsv IN**: **fasta file of each of 37 species**

**Code part 2.2: annotating stress-relevant proteins**

**Code part 2.3**: **Concatanate for each species the number of:**

1. **Proteins – from input fasta file**
2. **Proteins annotated by Interproscanner – from Interproscanner output files**
3. **Number of Stress Annotated Proteins – data file from Figure 1C approach 3**