Gene content protocol

Colored letters correspond to the final datasets as described at the end of the protocol.

A) Create taxa list for analysis. 47 species list created.
B) Subset 47 species list proteomes for analysis.
   Datasets names and taxa number in each:
   - **Opi** (47 taxa proteomes)
   - **Aco** (44 taxa proteomes)
   - **Xen** (41 taxa proteomes)

   - **Hol** (47 taxa proteomes - 6 Fungi taxa proteomes)
   - **HolAco** (44 taxa proteomes - 6 Fungi taxa proteomes)
   - **HolXen** (41 taxa proteomes - 6 Fungi taxa proteomes)

   - **Cho** (47 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)
   - **ChoAco** (44 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)
   - **ChoXen** (41 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)

C) Run DIAMOND with 4 E-values and MCL with 5 I values and create matrices for datasets:
   Three datasets from step A (Opi, Aco and Xen).
   (repeat for homogroups prediction and orthogroups prediction)

D) Run DIAMOND with 1 E-value (1e-3 default) and MCL with 1 I value (1.5 default) and create matrices from datasets:
   All datasets from step A.
   (repeat for homogroups prediction,suffix "homo" and orthogroups prediction, suffix "ortho")

   New matrices are labeled as in step A with additional suffix "disAb" for all and two datasets labeled with additional suffix "Opi":
   - **Opi-homo/ortho-disAb** (47 taxa proteomes)
   - **OpiAco-homo/ortho-disAb** (44 taxa proteomes)
   - **OpiXen-homo-ortho-disAb** (41 taxa proteomes)

   - **Hol-homo/ortho-disAb** (47 taxa proteomes - 6 Fungi taxa proteomes)
   - **HolAco-homo/ortho-disAb** (44 taxa proteomes - 6 Fungi taxa proteomes)
   - **HolXen-homo/ortho-disAb** (41 taxa proteomes - 6 Fungi taxa proteomes)

   - **Cho-homo/ortho-disAb** (47 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)
   - **ChoAco-homo/ortho-disAb** (44 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)

- **ChoXen-homo/ortho-disAb** (41 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)

E) Using TWO matrices created from step D (Opi-homo-disAb and Opi-ortho-disAb) to create new matrices by <u>P</u>runing (method).
New matrices are labeled as in step D but suffix "disAb" changed to suffix "disP":
- No Pruning for the dataset of the 47 taxa proteomes
- **OpiAco-homo/ortho-disP** (44 taxa proteomes)
- **OpiXen-homo/ortho-disP** (41 taxa proteomes)

- **Hol-homo/ortho-disP** (47 taxa proteomes - 6 Fungi taxa proteomes)
- **HolAco-homo/ortho-disP** (44 taxa proteomes - 6 Fungi taxa proteomes)
- **HolXen-homo/ortho-disP** (41 taxa proteomes - 6 Fungi taxa proteomes)

- **Cho-homo/ortho-disP** (47 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)
- **ChoAco-homo/ortho-disP** (44 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)
- **ChoXen-homo/ortho-disP** (41 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes)

F) Reduce from all datasets in A three ingroup taxa.
G) Repeat step D for all new datasets from F.
New matrices are labeled as in step D but suffix "disAb" changed to suffix "neAb":
- **Opi-homo/ortho-neAb** (47 taxa proteomes - 3 ingroups taxa proteomes)
- **OpiAco-homo/ortho-neAb** (44 taxa proteomes - 3 ingroups taxa proteomes)
- **OpiXen-homo-ortho-neAb** (41 taxa proteomes - 3 ingroups taxa proteomes)

- **Hol-homo/ortho-neAb** (47 taxa proteomes   - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)
- **HolAco-homo/ortho-neAb** (44 taxa proteomes - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)
- **HolXen-homo/ortho-neAb** (41 taxa proteomes - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)

- **Cho-homo/ortho-neAb** (47 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)
- **ChoAco-homo/ortho-neAb** (44 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)
- **ChoXen-homo/ortho-neAb** (41 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)

H) Repeat Pruning of three ingroup species (same species as in step F), using TWO matrices created from step E (Opi-homo-disP and Opi-ortho-disP) to create new matrices.
New matrices are labeled as in step E but suffix "disP" changed to suffix "neP":
- **Opi-homo/ortho-neP** (47 taxa proteomes - 3 ingroups taxa proteomes)

- **OpiAco-homo/ortho-neP** (44 taxa proteomes - 3 ingroups taxa proteomes)
- **OpiXen-homo/ortho-neP** (41 taxa proteomes - 3 ingroups taxa proteomes)

- **Hol-homo/ortho-neP** (47 taxa proteomes - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)
- **HolAco-homo/ortho-neP** (44 taxa proteomes - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)
- **HolXen-homo/ortho-neP** (41 taxa proteomes - 6 Fungi taxa proteomes - 3 ingroups taxa proteomes)

- **Cho-homo/ortho-neP** (47 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)
- **ChoAco-homo/ortho-neP** (44 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)
- **ChoXen-homo/ortho-neP** (41 taxa proteomes - 6 Fungi taxa proteomes - 2 Ichthyosporea/Filasterea taxa proteomes - 3 ingroups taxa proteomes)

I) Remove all singletons from datasets.

Results:
Step C results in 120 Phylogenies.
Steps D and G result in 36 Phylogenies.
Steps E and H result in 34 Phylogenies.
A total of 190 Phylogenies.