**Projekt sommar 2022**

Anteckningar

Start with:

Dunn 2008 – first paper to show contradicting results

Philippe 2009

**Li data organization**

Gene boundaries, in the repository? What are they called?

Taxonomy, how is it unified?

Make database of genes

Compare each gene to all other papers to find where it is used, blast?

Fig. S3, does not show which data overlaps but does show how much data overlaps between studies

**Looking at gene boundaries**

**Find out:**

Does every dataset (publication) have a list? where?

What is the naming scheme used?

Any mapping files?

What scheme does Li et al reanalysis use?

Nexus partition files:

Dunn: Genes are named whatever they were named in the partition files of the raw data, Tribe1006

Philippe: Genes are named by number, Gene:0001

Dunn raw data (C:\Users\matil\Downloads\summer\_proj\data\_raw\Dunn2008\Dunn2008.nex)

Partition names and positions:

CHARSET Tribe1006 = 1 - 99;

CHARSET Tribe1007 = 100 - 336;

CHARSET Tribe1009 = 337 - 547;

CHARSET Tribe1012 = 548 - 704;

CHARSET Tribe1013 = 705 - 897;

CHARSET Tribe1015 = 898 - 1130;

CHARSET Tribe1020 = 1131 - 1213;

…

Can be translated to name and id using

Supplementary Table 3 | Genes selected for phylogenetic analysis in:

<https://static-content.springer.com/esm/art%3A10.1038%2Fnature06614/MediaObjects/41586_2008_BFnature06614_MOESM1_ESM.pdf>

Dunn processed

Partition names: Tribe*xxxx*

Philippe raw data (C:\Users\matil\Downloads\summer\_proj\data\_raw\Philippe2009\Philippe2009.nex)

Partition names: genes = 142:ar21,arc20,arp23,atpsynthalpha-a-mt,cct-A,cct-B, …

Partition positions: partition part = 142:1-136,137-301,302-519,520-1015,1016-1519, …

Supplementary material: Table S1. List of the 128 genes used, names and positions

Philippe processed

Partition names: Gene\_*xxxx*

**Taxonomy, how is it unified?**

C:\Users\matil\Downloads\summer\_proj\reconciliation\reconciliation\scripts\ generate\_taxon\_table

Get names from nexus file, clean up names, search for name in NCBI and return taxa id, fetch taxonomy information from NCBI using taxa id, make taxon table

C:\Users\matil\Downloads\summer\_proj\reconciliation\reconciliation\scripts\prepare\_phylip.py

Converts files to nexus + phy (processed) and adds NCBI scientific names from taxon table

Gene boundaries

**Find out:**

Does every dataset (publication) have a list? where?

What is the naming scheme used?

Any mapping files?

What scheme does Li et al reanalysis use?