**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?

Bake 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**

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;

…

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, …

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