Conservation of proteoglycan synthesis pathway genes based on organisms' classification

A comparative genomics approach

Flaviu Vadan, Ian McQuillan, Brian Eames

What are proteoglycans?

- Type of sugar-coated proteins
- Variety of sugars presents on surface of proteoglycans “dictate” class
- Influence functions of multiple tissues (cartilage)
- Loss of proteoglycans associated with diseases (osteoarthritis)

Why study this?

- Diseases are not great
- Gene networks are cool
- Evolutionary biology is awesome
- No good understanding of how organisms with different bone structures (or cartilage traits) differ from an evolutionary standpoint

What’s an approach to study this?

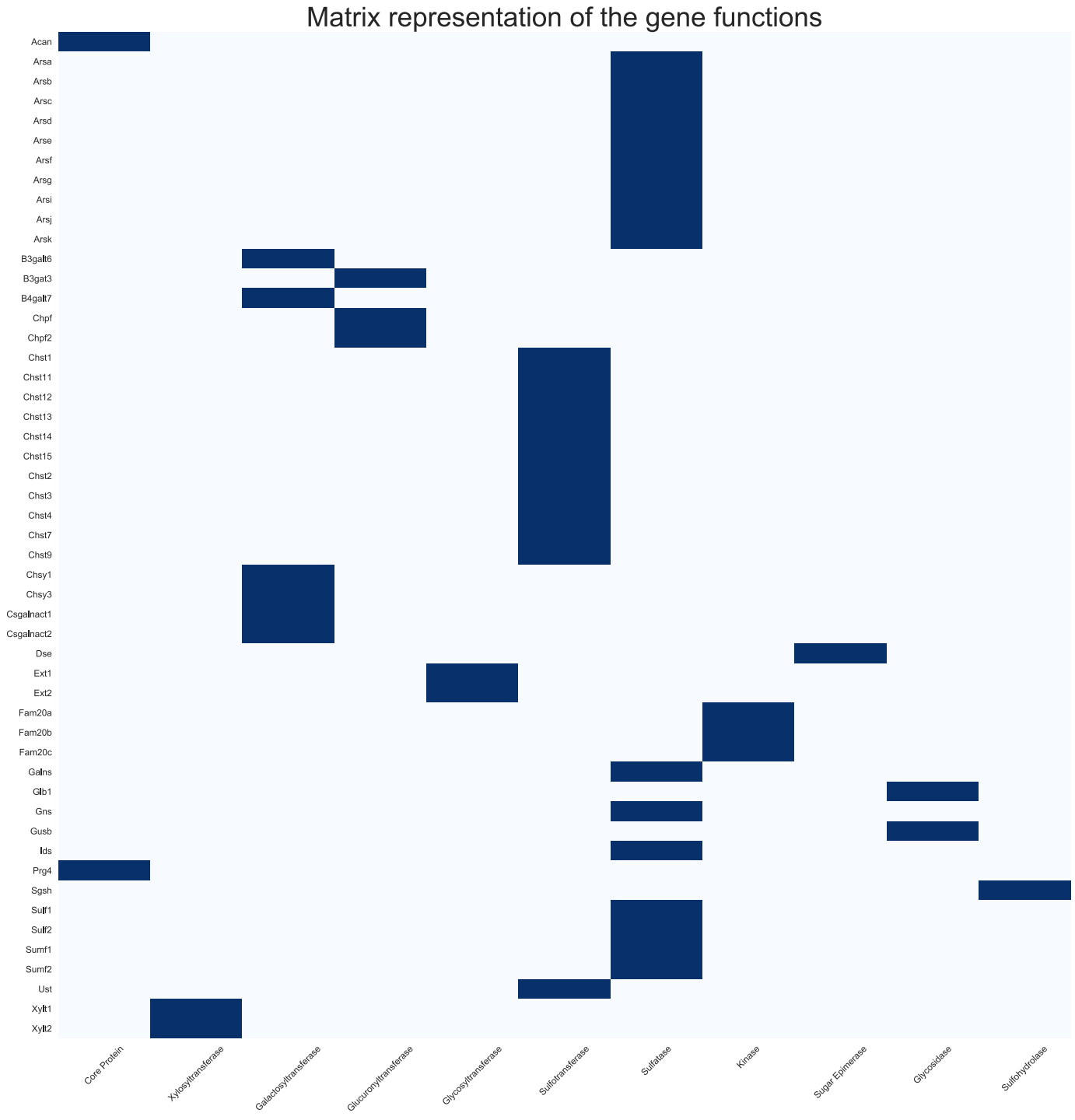
- Look at differences between a small number of organisms
- Study genes individually
- Study organisms individually
- All the above

What can we study?

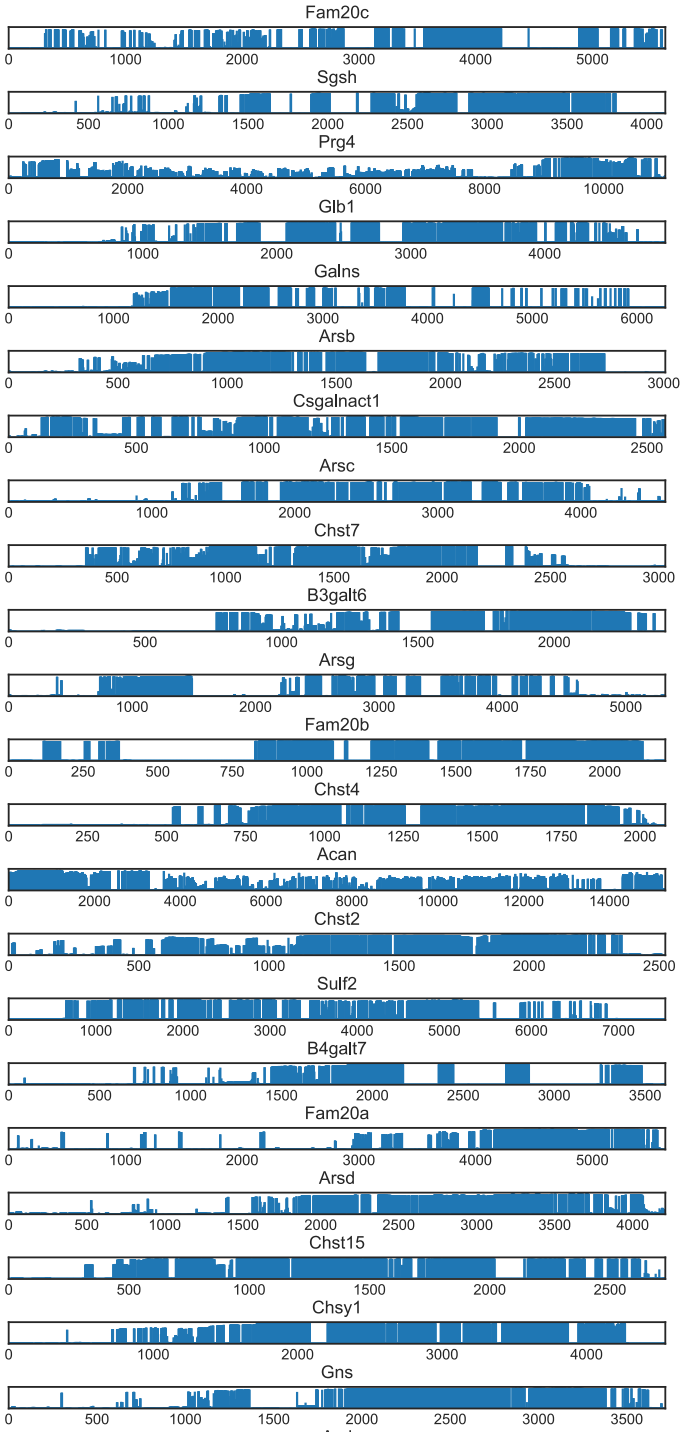
- Evolutionary biology is concerned with how organisms evolve
- Hypothesis: Genes in the proteoglycan synthesis pathway exhibit higher non-synonymous vs. synonymous mutation ratios in organisms that have specific cartilage/bone traits

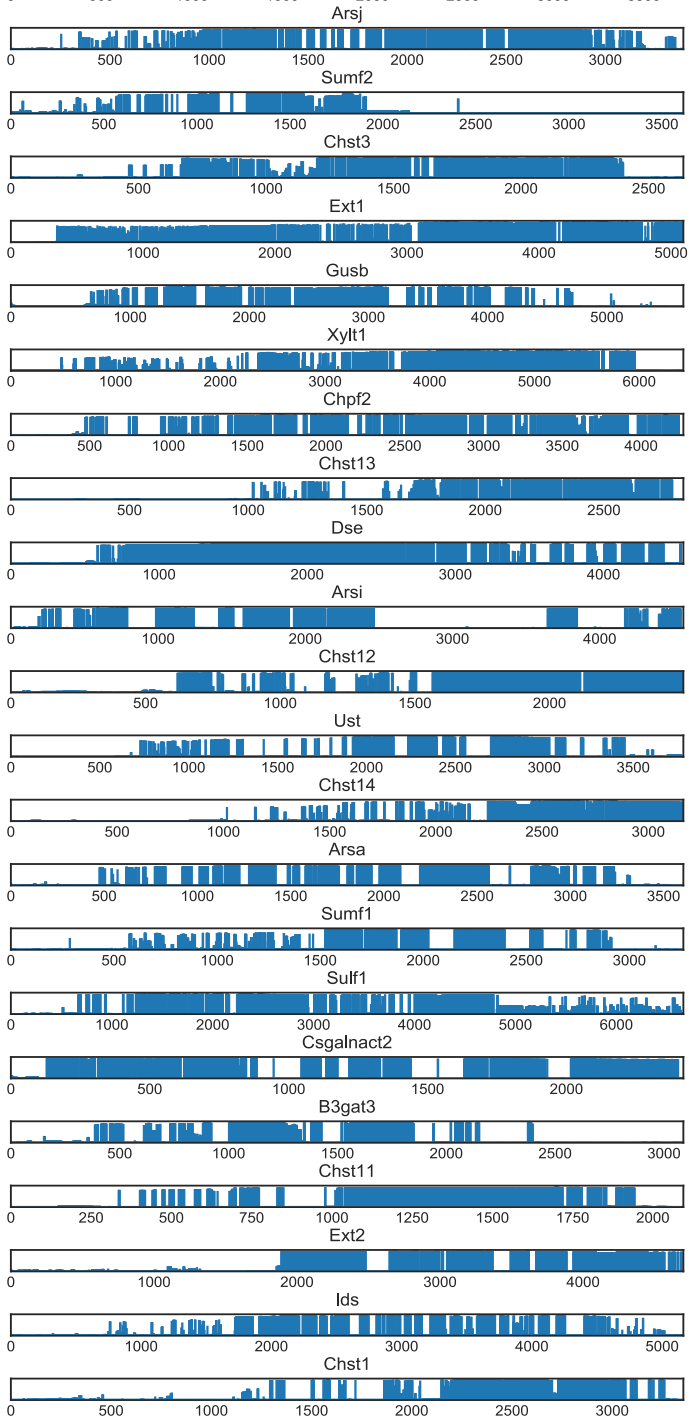
Blah, blah, tell me what you did!

- Collected sequences of all Homo Sapiens proteoglycan pathway gene orthologues (199 organisms, 188 actually) from Ensembl
- Curated, organized, parsed
- Interested in grouping gene functions, multiple sequence alignments, gene frequencies, gene presence, distance between gene presence vectors, phylogenetic relationship between organisms, how everything relates to non-synonymous vs. synonymous mutation ratios

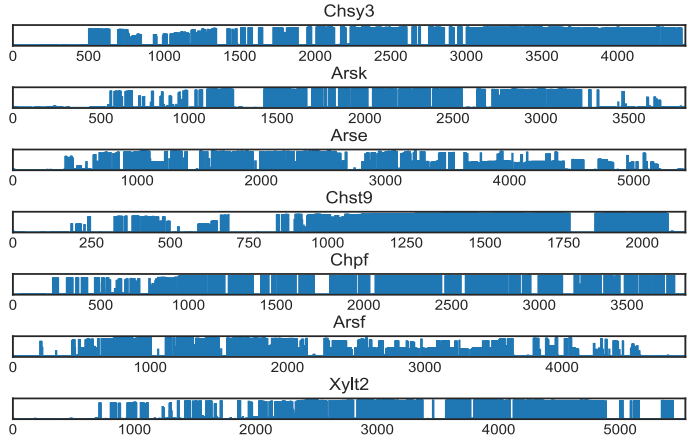


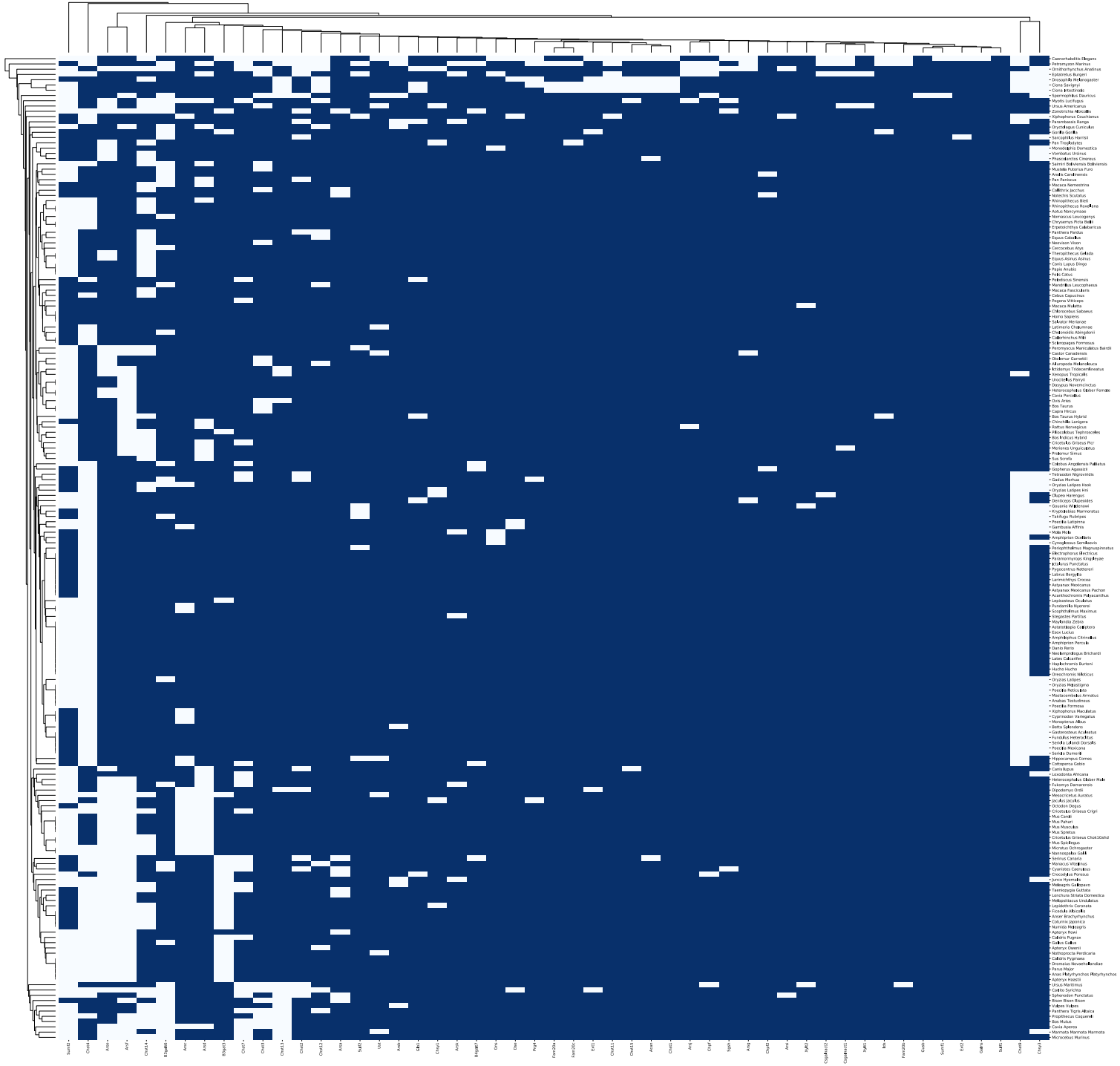
Grouping genes based on function helps us formulate predictions of what patterns might occur in the conservation of genomic sequences





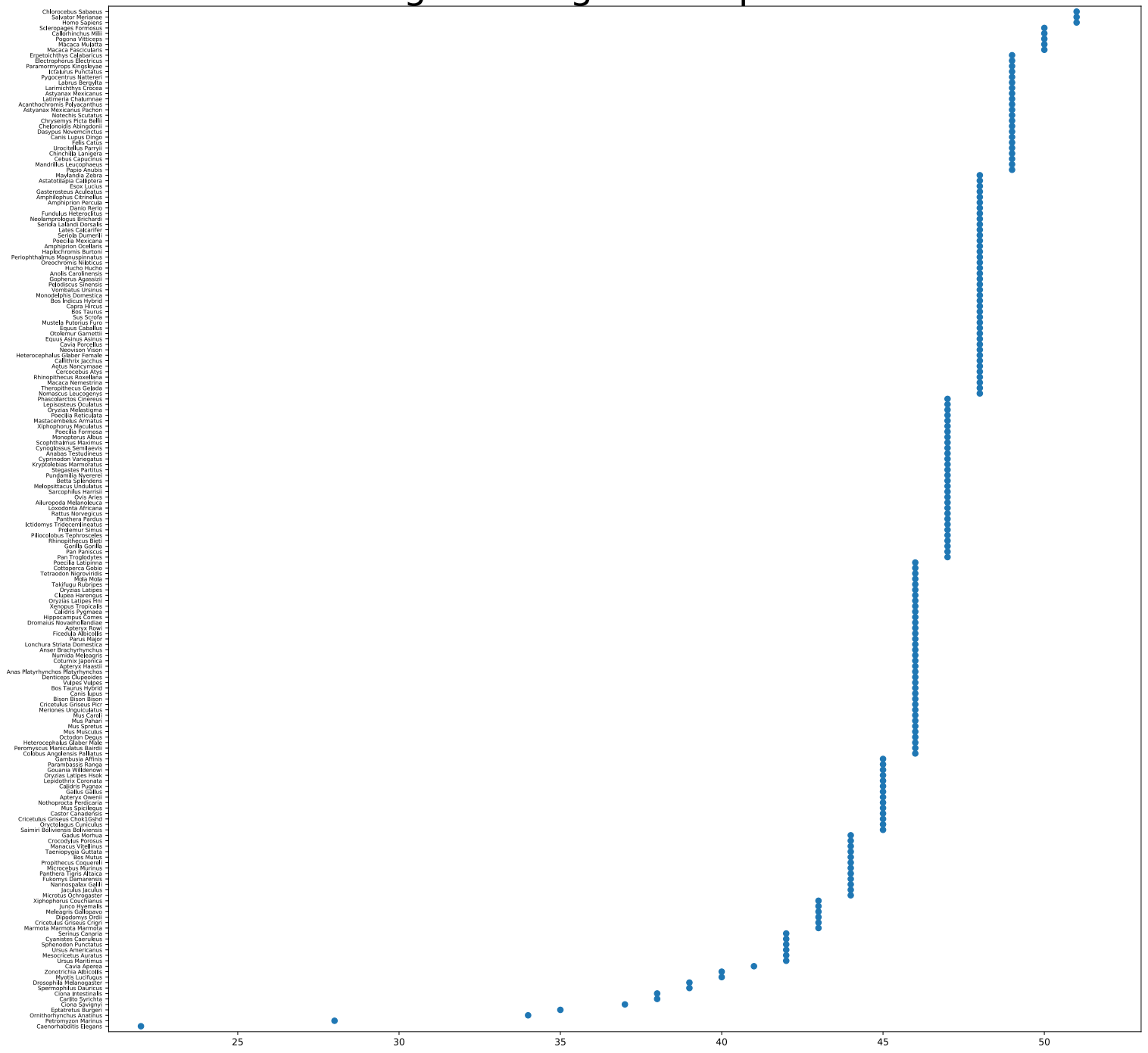
Pretty multiple sequence alignments help with our understanding of sequence preservation



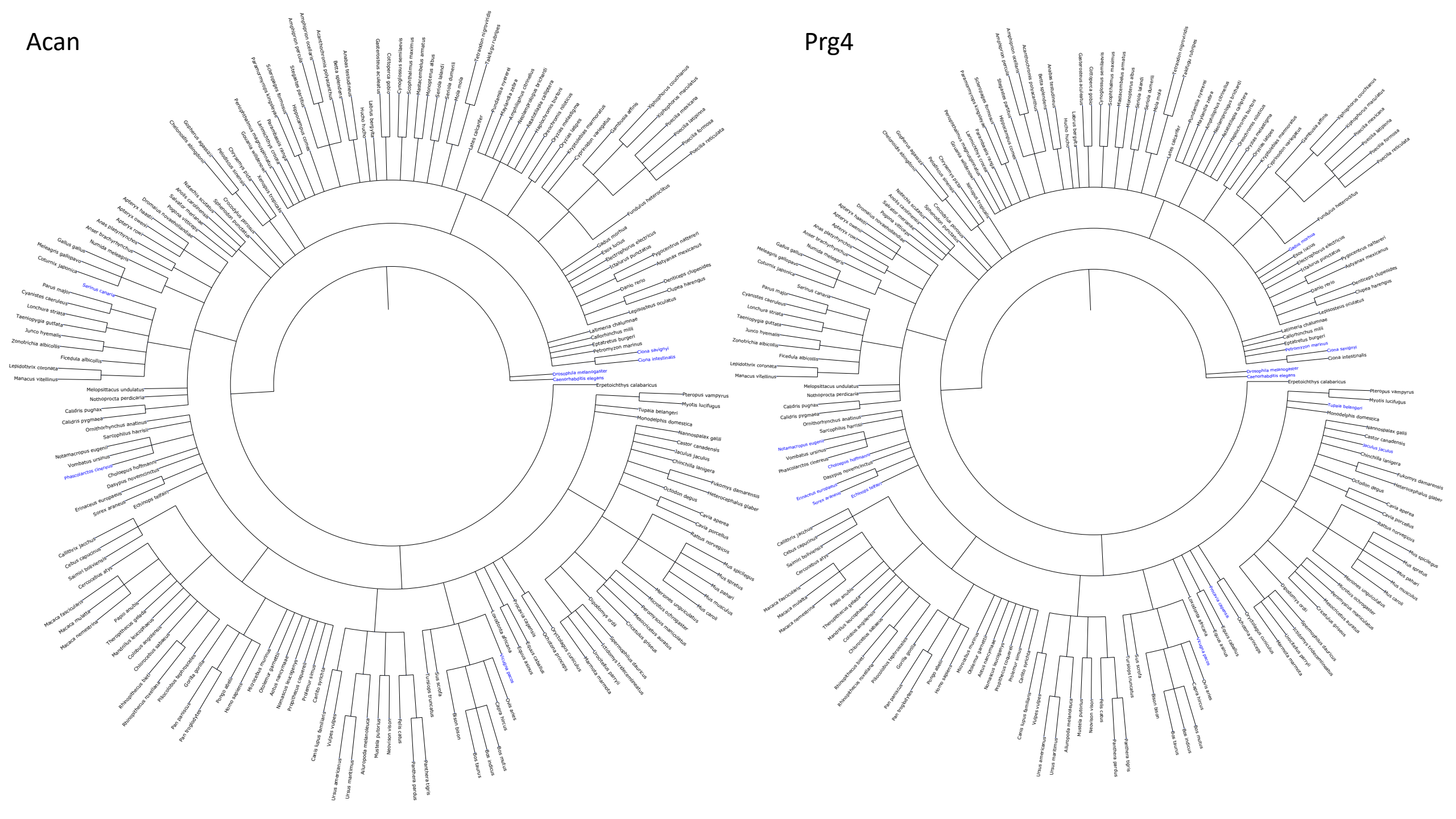


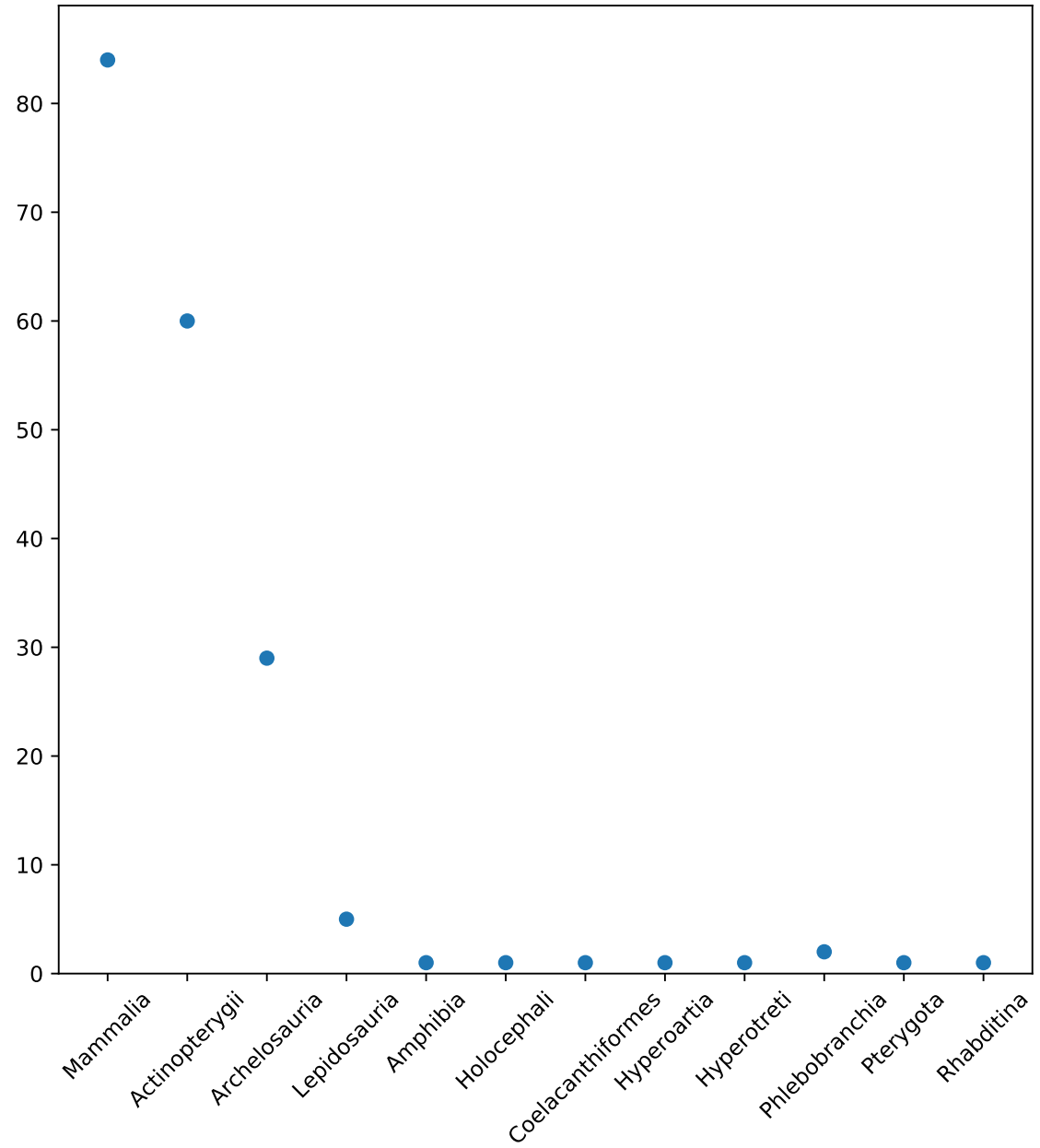
Clustering organisms based on gene presence similarity can provide insight into the proteoglycan pathway genomic consistency between organisms

Organisms gene frequencies



Gene frequencies allow us to focus on specific organisms that may be interesting from a gene presence perspective (e.g outliers such as *C. Elegans*)


Acan
Prg4
Serinus canaria
Drosophila melanogaster
Caenorhabditis elegans
Vicugna pacos



Taxonomic frequency of the analyzed dataset. The presented classes are:
Mammalia - mammals; Actinopterygii - ray-finned fish; Archelosauria - turtles and archosaurs;
Lepidosauria - scaly reptiles; Amphibia - amphibians; Holocephali - cartilaginous fish;
Coelacanthiformes - bony fish (ancient); Hyperoartia - jawless bony fish;
Hyperotreti - hagfish; Phlebobranchia - sea squirts; Pterygota - winged insects;
Rhabditina - nematodes.

Grouping organisms based on taxonomic information is challenging and suggests a data problem*

*the only diagram that can stand by itself

Where are you going with this?

- Need to further analyze clustering of organisms and genes from an evolutionary perspective
- Need to analyze phylogenetic relationships and gene presence
- Must compute ratios of non-synonymous vs. synonymous mutations and check whether they're consistent with current results' direction*

* after all, they're mentioned in the hypothesis, so they must be important

Who helped?

- My supervisors!
- StackOverflow and Python libraries (ete3)
- EBI (MSAs)*

- D.S. Brown, B. F. Eames. Emerging tools to study proteoglycan function during skeletal development. Book chapter. 2016
- D. Graur. Molecular and Genome Evolution. Book. 2016.
- A. P. Hendry et al. Evolutionary principles and their applications. Evolutionary Applications. 2011

*can we have kalign installed on tuxworld? Way faster than emma!