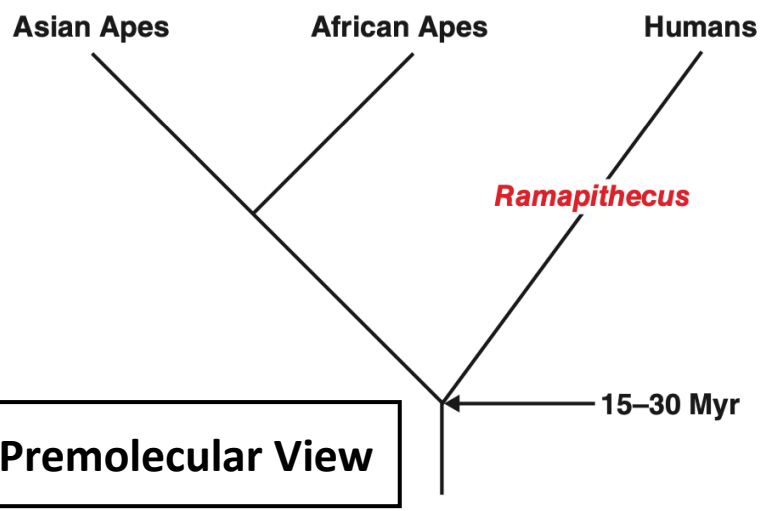
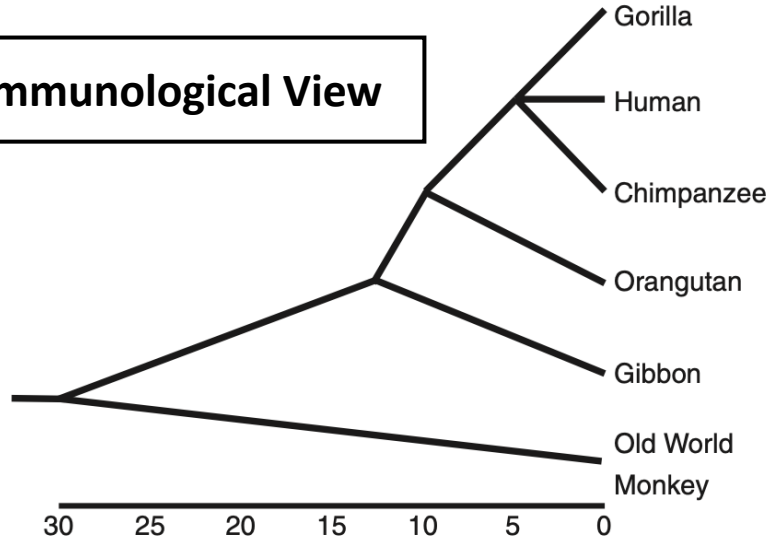
Using Phylogenomic Data to Resolve the Hominoid Trichotomy

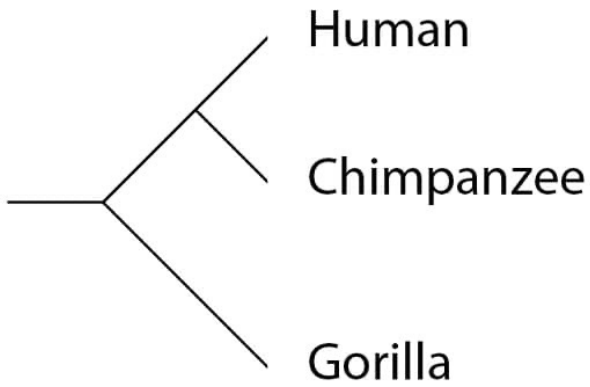
(AKA Joe learns how to convert many file types into a different file type)



Premolecular View

Immunological View





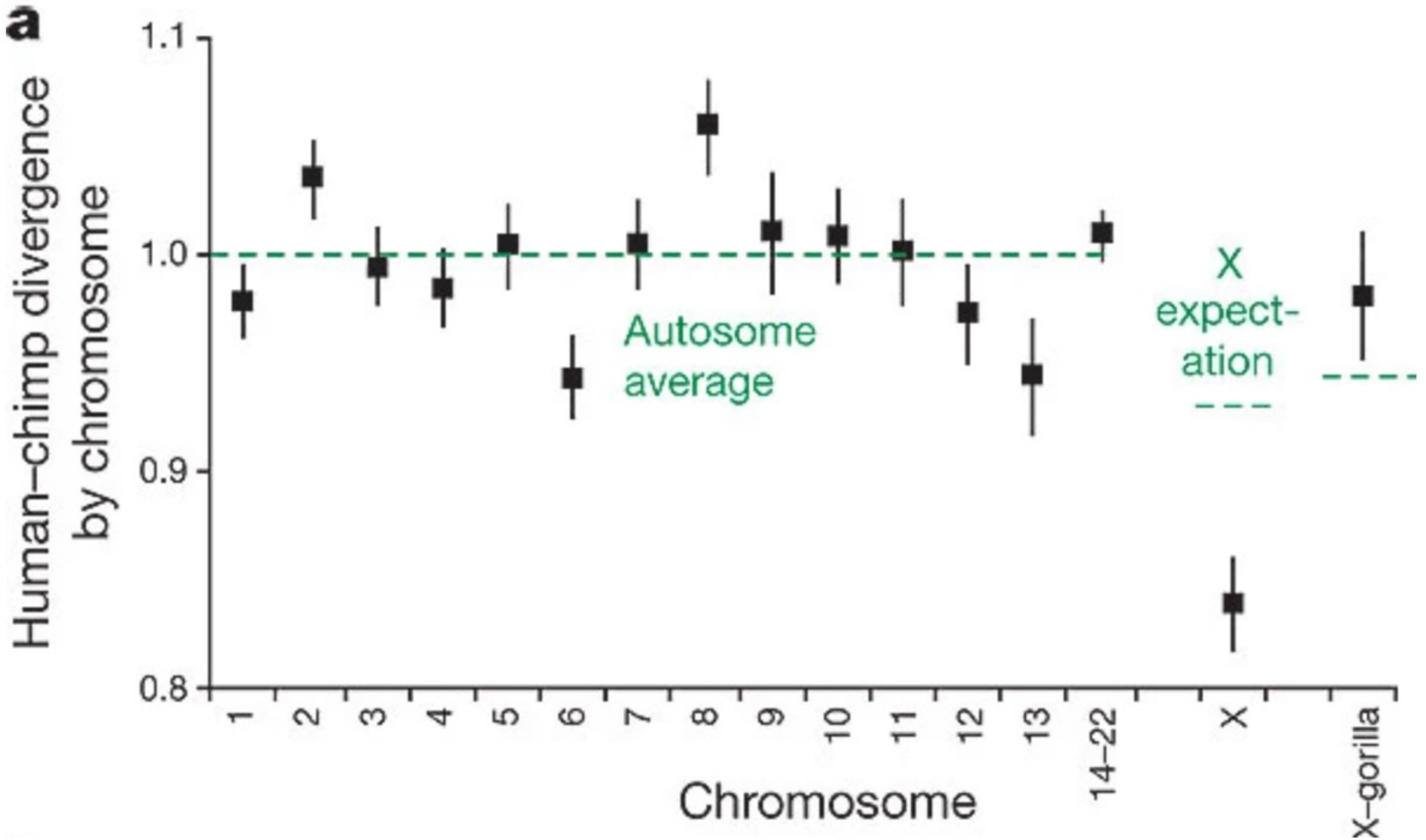
Genetic View

~ 1850s – 1950s

1960s-1990s

1990s - today

Figure 3: Reduced human–chimpanzee time divergence across chromosome X.



Nick Patterson[1], Daniel J. Richter[1], Sante Gnerre[1], Eric S. Lander[1,2] & David Reich[1,3]

Vol 441|29 June 2006|**doi:10.1038/nature04789**

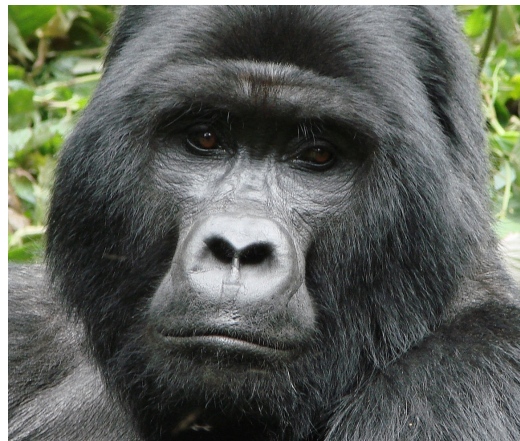
- Contemporary analyses of structural variants in human populations indicate many variants inherited from Denisovans and Neandertals
- These variants are often located in coding regions related to immune function
- Such variants in the genome likely undergo stabilizing selection

Trends in Genetics

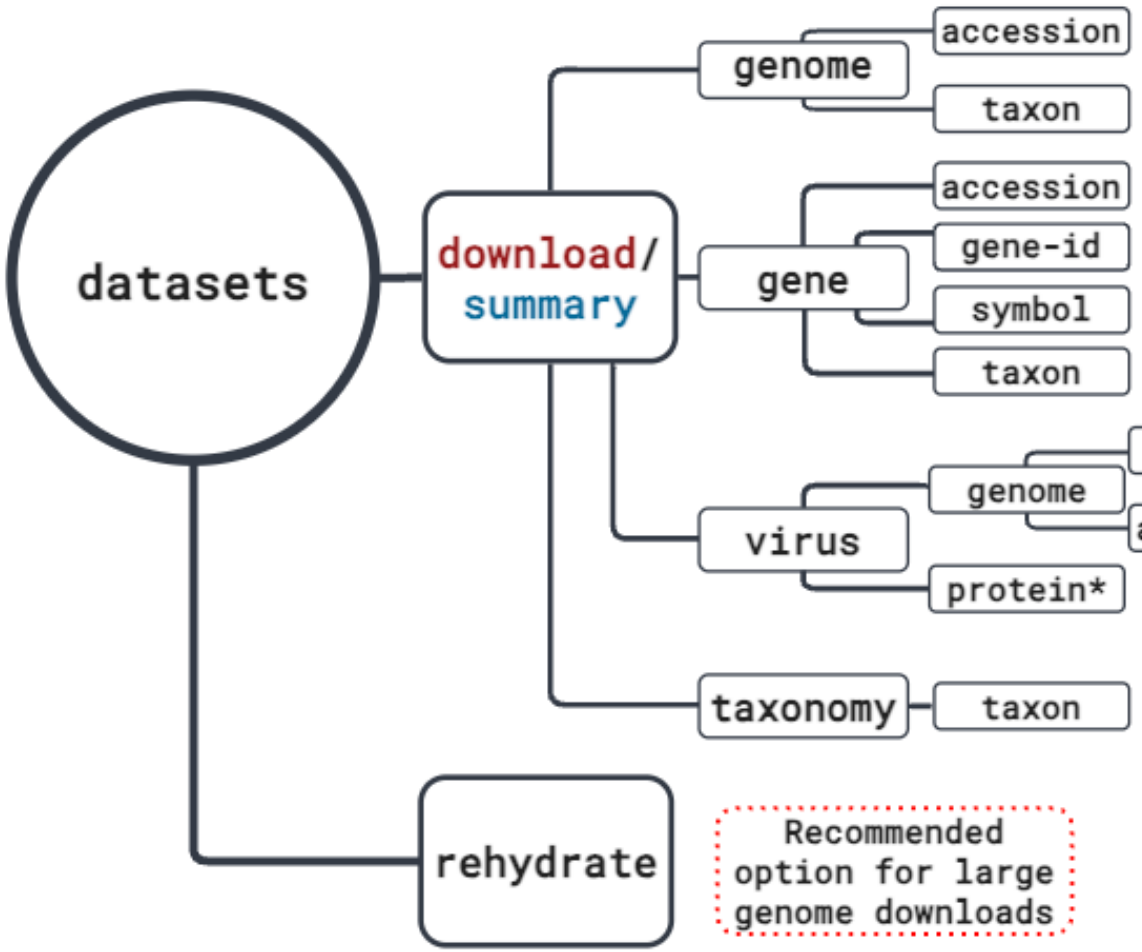
Francesco Montinaro[1,*] and Cristian Capelli[2,*]

Trends in Genetics, October 2020, Vol. 36, No. 10 723

Question: What genes recover different topologies among the four hominids?



COMMAND EXAMPLES




```
datasets download genome accession GCF_000001405.40
datasets summary  genome taxon ailuridae
```

```
datasets download gene accession NP_000483.3
datasets download gene symbol brca1 --ortholog primates
datasets summary  gene gene-id 40650
```

```
datasets download virus genome taxon monkeypox
datasets summary  virus genome accession NC_045521.2
datasets download virus protein ORF10
```

```
datasets download taxonomy taxon "mus musculus"
datasets summary  taxonomy taxon mustelidae --children
```

```
datasets rehydrate --directory my_genomes
```

- First, I downloaded list of all genes available for my relevant taxa from
- For simplicity, used only Ch 12
- This list was then shortened using a custom script

- With a list of 7 genes for analysis, I used the NCBI command line tools to download orthologs
- NCBI HomoloGene orthology is assigned using protein sequence similarity and local synteny information

Alignment

Program: MAFFT (v7.526)

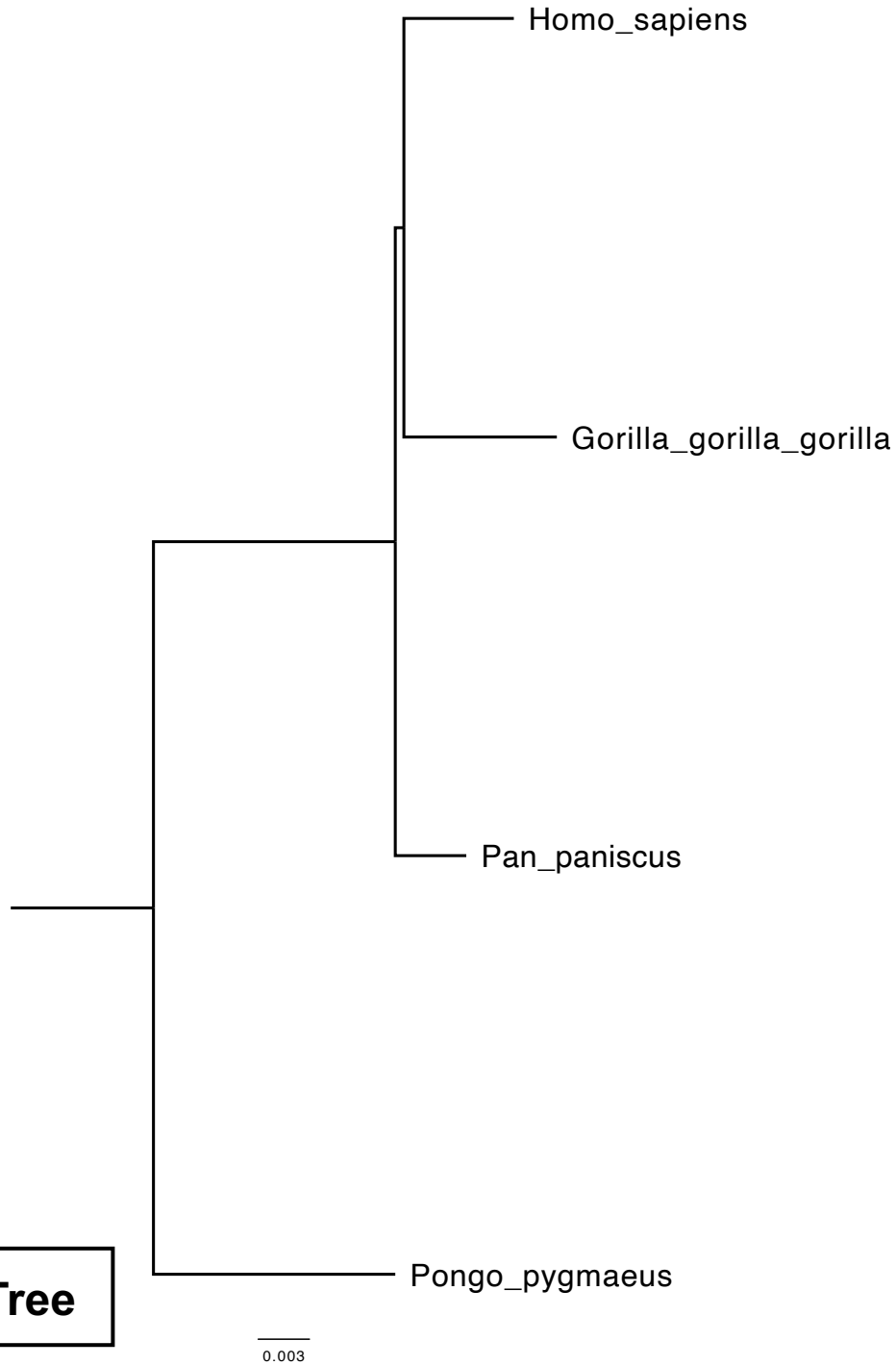
- I aligned each set of orthologous genes with MAFFT
- For simplicity I used the –-auto argument
- Since RAxML did not like the format of the headers used by the NCBI .fna file, I used a custom script to rename headers and remove duplicate samples

Maximum Likelihood Trees

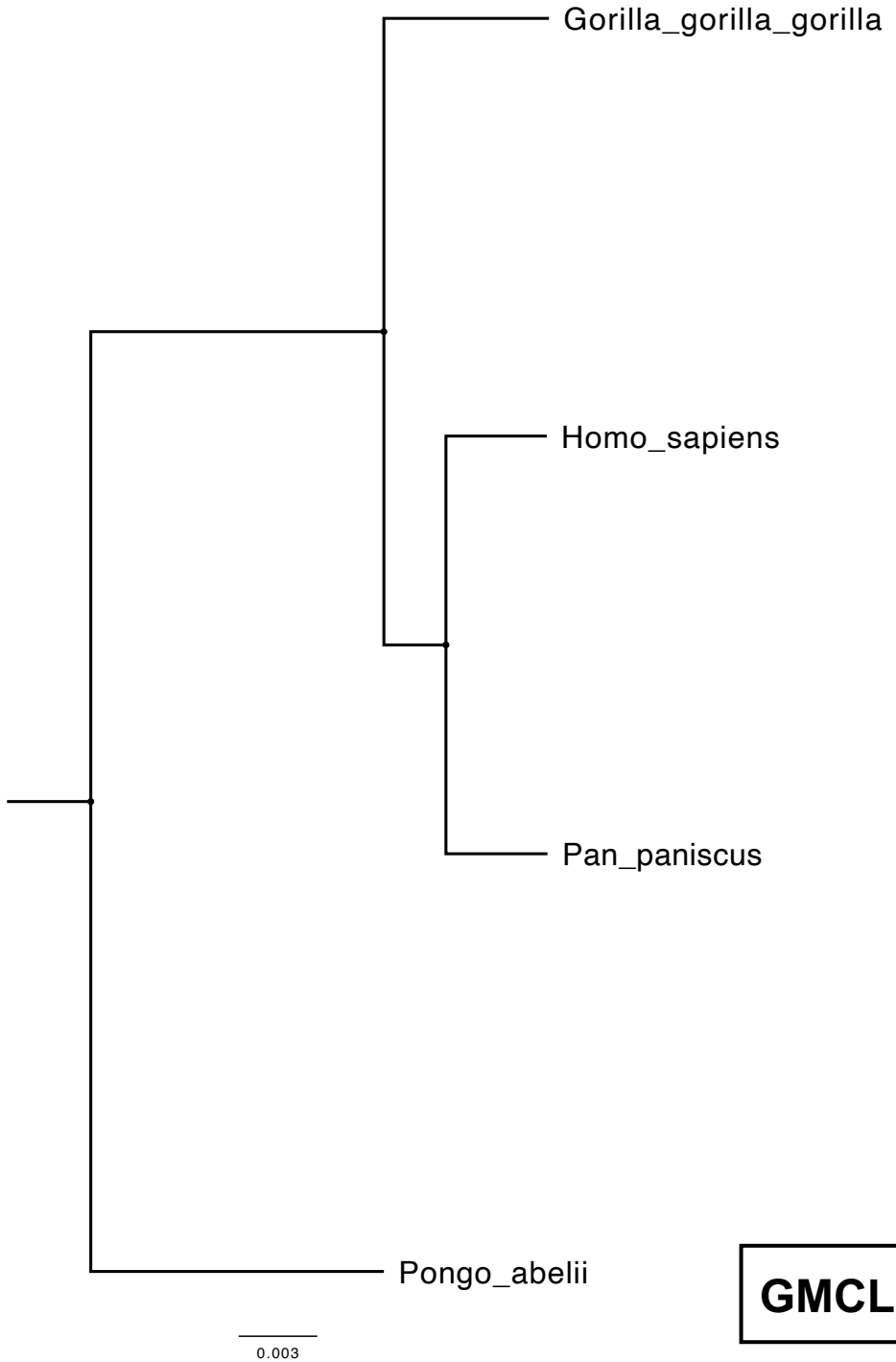
Program: RAxML – NG (v1.2.2)

- I ran RAxML on all edited alignments using this command:

```
#do it for all alignments
for alnp in *.fna
do
    base=$(basename "$alnp" .fna)
    mkdir "$base"
    ./raxml-ng --all \
               --msa "$alnp" \
               --model GTR+G \
               --prefix "$base"/"$base" \
               --threads 4 \
               --bs-trees 100
done
```



CD8A Gene Tree



GMCL1 Gene Tree

Final Note and Future Directions

- I was 3/4^{ths} of the way done with this final project before realizing that I was pruning samples and duplicates AFTER running alignment
- Question: Is this bad practice and will it influence downstream analyses?

- Future Directions:
  - Rectifying previously mentioned issue
  - Running this with a larger sample of orthologous genes
  - Coalescent methods (ASTRAL)
  - Develop a greater understanding of how to compare the different topologies and…
  - How to classify what those genes actually do