# 1. Do you think that in a phylogenetic tree the parasites that use similar hosts will group together?

Yes, in a phylogenetic tree, organisms that share a closer evolutionary relationship are expected to be grouped together.

# 2. Why?

# Sequences with more than 35% GC content are deleted to remove scaffolds that might come from bird DNA, which typically has a higher GC content. However, even when sequences with high GC content are removed, there may still be contamination or noise from other sources, such as bacterial DNA or sequencing errors

# Q: In SwissProt, the five characters after the underscore is a species abbreviation (what is before the underscore?)

# In the SwissProt database, the part before the underscore (\_) is usually the short or abbreviation of the species, while the five characters after the underscore are the species-specific abbreviation or identifier.

# Typically, the section before the underline is the protein's name or description, which identifies the protein's function, subcellular location, and so on. The five characters following the underscore are usually abbreviations for the corresponding species of the protein and are used to identify the protein

# 3. Insert the missing data in the above table. Use bash, not internet!

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Host | Genome size | Genes | Genomic GC |
| *Plasmodium berghei* | rodents | 17954629 | 7235 | 23.75% |
| *Plasmodium cynomolgi* | macaques | 26181343 | 5787 | 40.38% |
| *Plasmodium falciparum* | humans | 23270305 | 5207 | 19.37% |
| *Plasmodium knowlesi* | lemures | 23462187 | 4952 | 38.83% |
| *Plasmodium vivax* | humans | 27007701 | 5682 | 42.29% |
| *Plasmodium yoelii* | rodents | 22222369 | 4889 | 21.77% |
| *Haemoproteus tartakovskyi* | birds | 27426784 | 15892 | 52.35% |
| *Toxoplasma gondii* | humans | 128105889 | 15048 | 27.52% |

# 4. Compare the genome sizes with other eukaryotes and bacteria. Discuss with your partner (that is student partner) the reason for the observed genome sizes.

# 5. What may cause the biased GC-contents in some of the species?

Biased GC-content in species can be caused by biological selection favoring stability, mutation and genetic drift, horizontal gene transfer, recombination, environmental factors, and genomic architecture. These factors interact to shape the GC-content observed in different organisms.

# 6. What does the curly braces notation stand for?

# The curly braces notation {} in the command proteinortho6.pl {Ht,Pb,Pc,Pf,Pk,Pv,Py,Tg}.faa is used for brace expansion in Bash. Brace expansion allows you to generate arbitrary strings based on comma-separated lists or ranges inside the braces.

# Q: What does the flag -l stands for? and why do we choose apicomplexa?

# The -l flag in the busco command stands for the lineage dataset to be used for assessing the completeness of the input sequences. In this case, -l apicomplexa indicates that the lineage dataset for the Apicomplexa phylum will be used.

# 7. Compare how many BUSCOs (orthologues proteins) that are found in each proteome. Do the investigated parasites have close to complete numbers of BUSCOs?

# Pb: 361 Complete BUSCOs (C) C:80.9%

# Pc: 429 Complete BUSCOs (C) C:96.2%

# Pf: 436 Complete BUSCOs (C) C:97.8%

# Pk: 433 Complete BUSCOs (C) C:97.1%

# Pv: 437 Complete BUSCOs (C) C:97.9%

# Py: 435 Complete BUSCOs (C) C:97.5%

# Tg: 380 Complete BU298 (C) C:85.2%

# Ht: 298 Complete BUSCOs (C) C:66.8%

# Complete BUSCOs account for 80.9% of the total BUSCOs. This means that a fairly high percentage of the genomes of these parasites are conserved genes shared with other species

# 8. Do you think that the assembly of the Haemoproteus tartakowskyi genome is a reasonable approximation of the true genome?

# According to the provided data, the proportion of BUSCOs in the Haemoproteus tartakowskyi genome is not very close to completeness. While there are some complete BUSCOs present, the proportion of missing BUSCOs is relatively high at 28.5%, indicating that there may be some incompleteness in the assembly of this genome. Therefore, it can be said that the assembly of this genome may be a reasonable approximation,

# 9. How many of the BUSCOs are found in all eight organisms?

# Extract all busco ids in complete and duplicated shown once

# Use uniq -c to count the sorted ids and output the number of times each ID appears in all files

awk '/^[^#]/ && ($2 == "Complete" || ($2 == "Duplicated" && !seen[$1]++)) {print $1}' Pb\_busco/Pb/run\_apicomplexa\_odb10/full\_table.tsv > Pb\_buscoIds.txt

awk '/^[^#]/ && ($2 == "Complete" || ($2 == "Duplicated" && !seen[$1]++)) {print $1}' \*\_busco/\*/run\_apicomplexa\_odb10/full\_table.tsv > \*\_buscoIds.txt

awk '{print $1}' \*.tsv | sort | uniq -c | awk -v num=8 '$1 == num {print $2}' | wc -l

# 223

# 10. If Toxoplasma is removed, how many BUSCOs are shared among the remaining seven species. Interpret!

awk '{print $1}' Ht.tsv Pb.tsv Pc.tsv Pf.tsv Pk.tsv Pv.tsv Py.tsv | sort | uniq -c | awk -v num=7 '$1 == num {print $2}' | wc -l

# 252

# 11. Does all protein trees reflect the “true’ ’ species tree?

# No

# 12. What is the phylogenetic position of Plasmodium falciparum?

# First branch

# 13. Do you think that the GC contents have an impact on the tree topology?

# Yes.

# GC content affects the rate and type of mutations, particularly transitions (A <-> G, C <-> T) and transversions (A <-> T, G <-> C). Strong mutation biases towards GC-rich or AT-rich regions may result in differences in substitution rates, thereby influencing tree topology.

# Some organisms exhibit biased DNA mismatch repair favoring GC over AT alleles, known as biased gene conversion. This process can alter the pattern of nucleotide substitutions, thereby affecting tree topology.

# GC content may be influenced by selection pressures related to factors such as codon usage bias, gene expression levels, and environmental adaptation. Genomic regions under different selection pressures may exhibit varying GC content, impacting evolutionary rates and tree topology.

# Differences in GC content between genomic regions (e.g., coding vs. non-coding, introns vs. exons) can affect patterns of molecular evolution, introducing heterogeneity in substitution rates and potentially influencing tree topology.

# Gene flow and hybridization between species with different GC contents can lead to inconsistent gene trees, affecting overall tree topology.

# 14. Do you think that the host range has an impact on the tree topology?

# The host range influences the tree topology of phylogenetic analysis. It may lead to host switching events, co-evolutionary dynamics, horizontal gene transfer, niche conservation, and parasitic speciation, resulting in specific patterns and structures in tree topologies. Therefore, the host range and its effects on related organisms need to be considered when interpreting the results of phylogenetic analyses

# 15. Are the BUSCO proteins also found as orthologs in the proteinortho output?

# Yes

# 16. Make a script that concatenates the alignments for each organism and BUSCO into one fasta file that in the end should contain seven sequences. Alternatively, use bash.

# 17. Make a tree of this “superalignment’ ’. Does it correspond to the consense tree