**Investigating The Genomes of Different Fish Species and Humans**

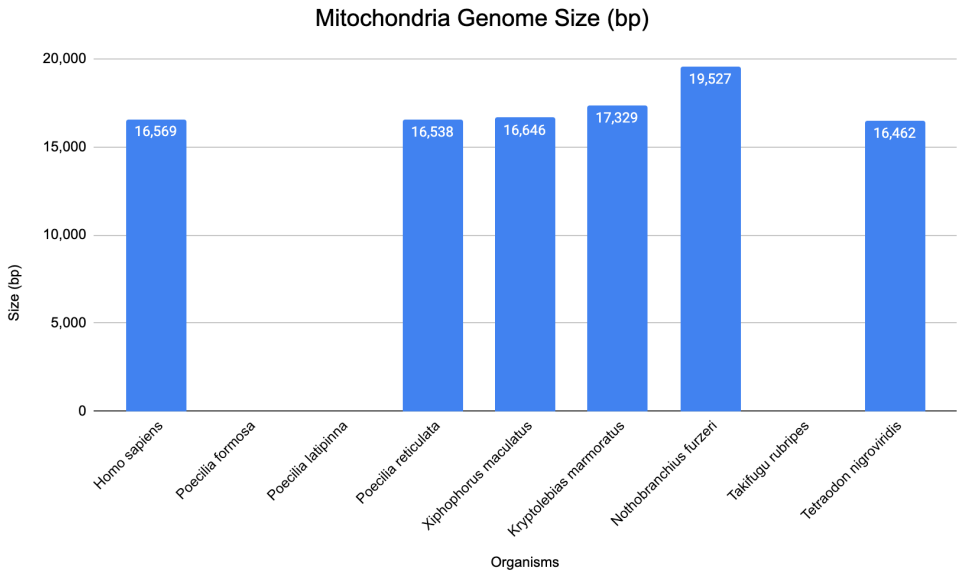
I. Topic_01. Do your Genomes have defined karyotypes?
   A. Galaxy was used to answer this question. The history name is 'Topic_02.'
      1. Search for species karyotype in Ensembl
         a) If not defined in Ensembl, inspect GFF3 file
      2. Examine whether there are specific chromosomes annotations in each genome
         a) If there is not an annotation for chromosomes, it is safe to assume there is no defined karyotype.
   B. Results (see Table 1)
      1. Three organisms did not have defined karyotypes.
      2. Six organisms had defined karyotypes in Ensembl and chromosomes with specific annotations.

**Table 1. Examine Organisms Exhibiting Karyotypes**

| Organisms | Common Name | Defined Karyotype |
|---|---|---|
| *Homo sapiens* | Human | yes |
| *Poecilia formosa* | Amazon molly | no |
| *Poecilia latipinna* | Sailfin molly | no |
| *Poecilia reticulata* | Guppy | yes |
| *Xiphophorus maculatus* | Platyfish | yes |
| *Kryptolebias marmoratus* | Mangrove rivulus | no |
| *Nothobranchius furzeri* | Turquoise killifish | yes |
| *Takifugu rubripes* | Fugu | yes |
| *Tetraodon nigroviridis* | Tetraodon | yes |

II. Topic_02. Do your Genomes have defined Mitochondrial Genomes?
   A. Galaxy was used to answer this question. The history name is 'Topic_02.'
      1. Cut the first column from all of the gff3 files to isolate annotated chromosomes
      2. Isolate unique lines by removing duplicates in the datasets
      3. Search for sequence ID with the expression 'MT' and the genes associated with the mitochondrial genome.
      4. Examine the output from grepping "MT," and look for lines labeled as chromosome/region where it lists the size (bp) of the mitochondrial genome.
      5. Isolate protein coding genes and count the number of lines.
   B. Results (see Figure 1)
      1. Three organisms did not have a defined mitochondrial genome.
      2. Organisms that did exhibit a defined mitochondrial genome had a total of 37 genes and 13 of which are protein coding.

**Figure 1. Organisms With Defined Mitochondrial Genome and Their Size**




III. Topic_03. How many Coding Genes have been annotated?

A. Galaxy was used to answer this question. The history name is 'Topic_03.'

1. For each gff3 file, search in textfiles (grep) utilizing Perl to capture lines containing:

a) "biotype=protein_coding"

(1) Searches for protein coding in different genomic elements (e.g, gene, exon, transcript, mRNA etc.)

b) "ID=gene:"

(1) Searches specifically for genes that have been annotated

B. Results (see Figure 2)

1. *Tetraodon nigroviridis* and *Homo sapiens* have the least amount of protein-coding genes. *Homo sapiens* displays a lower number of protein-coding genes (16,569), which could be due to our cells being very specialized. Number of annotated protein coding genes ranged from 19,602 to 23,753 genes. Xiphophorus maculatus has the most protein coding gene.

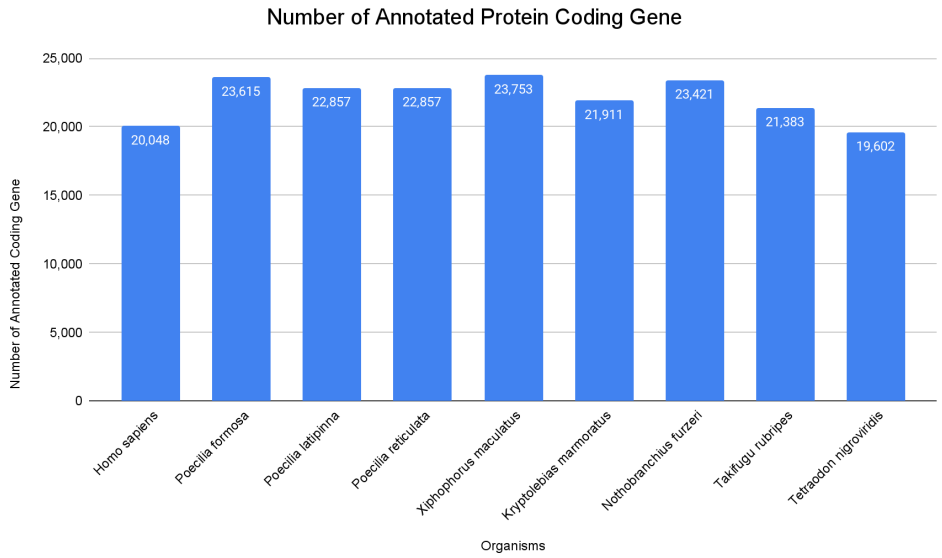



**Figure 2. Number of Annotated Protein Coding Genes**

IV. Topic_04A. How many Proteins have been annotated?

A. Galaxy was used to answer this question. The history names are 'Topic_04A.'

2. For each GFF3 file, search in textfiles (grep) utilizing Perl to capture lines containing:
    a) "\tmRNA\t"
        (1) Capture only lines annotated as mRNA types
    b) "biotype=protein_coding"
        (1) Searches for protein coding from previous output
3. Extract Transcript IDs
    a) Cut Column 9 from the Grep of specific annotated chromosomes
    b) Convert all colons from the output of the previous step to TABS
    c) Utilizing "Text transformation" remove instances of ";Parent=gene//" (s/;Parent=gene//) from the previous step's output
    d) Cut column 2 from the output of the previous step to isolate the Transcript ID
4. Perform Advanced Grep
    a) Utilizing the Tabular pep.all file (input file) and the isolated Transcript ID (pattern file)
    b) Convert Tabular file to Fasta file

V. Topic_04B. How many Canonical proteins have been annotated?

A. Galaxy was used to answer this question. The history names are 'Topic_04B.'

1. To search for canonical proteins
    a) Utilize Topic_04A's Grep of specific annotated genome file to Grep for "Ensembl_canonical"
    b) Repeat the steps to "Extract the Transcript ID"
    c) Perform Advance Grep
    d) Convert Tabular file to Fasta file

B. Results (see Figure 3)

1. *Homo sapiens* has significantly more proteins in their genome than the other organism. Tetraodon nigroviridis had the least number of both proteins and canonical proteins. *Kryptolebias marmoratus* and *Nothobranchius furzeri* have almost double the proteins compared to the other fish species, this could be due to genome duplication events.

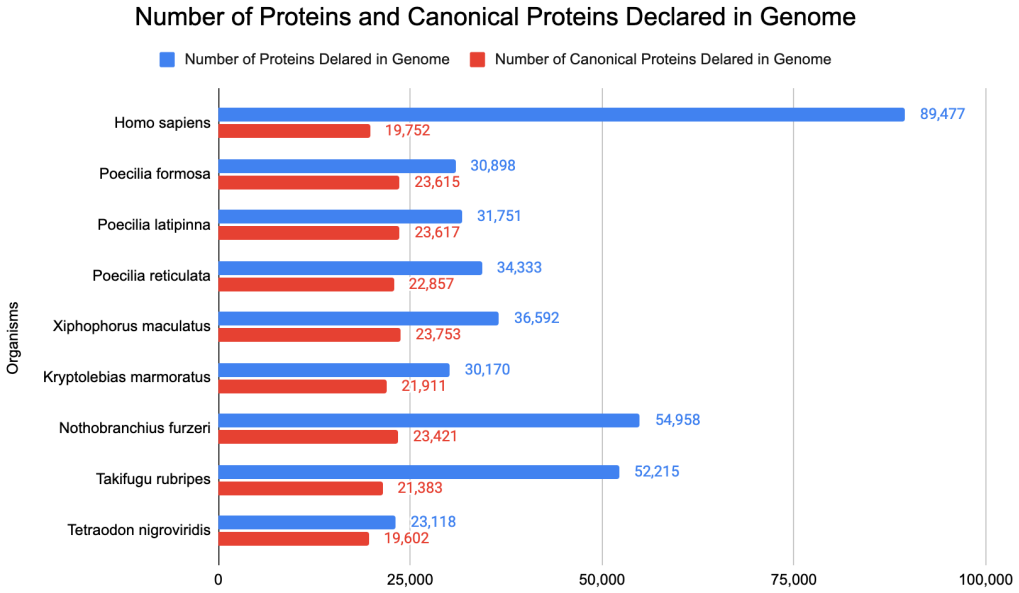



**Figure 3. Evaluating Organism's Protein and Canonical Proteins**

VI. Topic_05A. For each one of the genomes that were assigned to you, does your genome have Chromosome Scaffolds? and if it does how many Proteins are present in those Chromosome Scaffolds?

A. Galaxy was used to answer this question. The history names are 'Topic_05A.'

1. Utilizing the result from Topic_01, we understand that six organisms that displayed specific annotated chromosomes have chromosome scaffolds. Three organisms, *Poecilia formosa, Poecilia latipinna,* and *Kryptolebias marmoratus,* do not have defined karyotypes but exhibit non-chromosomal scaffolds in their gff3 files.
2. Similar to the step from Topic_04A, except after searching for protein-coding mRNA, a search was conducted to isolate specific annotated chromosomes and filter out unplaced/unlocalized scaffolds.
3. Then I continue to follow the steps from Topic_04A to extract the transcript ID to perform Advance Grep.
4. Convert the Tabular Files to Fasta files

B. Results (see Figure 4)

1. Homo sapiens had the most at 89,405 proteins present in their chromosome scaffolds. Tetraodon nigroviridis has the lowest amount at 16,858 proteins present in their scaffolds. *Poecilia reticulata* (32,599) and *Xiphophorus maculatus* (36,432) had similar numbers of proteins in their scaffolds. *Nothobranchius furzeri (48,736)* and *Takifugu rubripes (50,817)* also had similar numbers of proteins. Organisms with similar numbers of proteins might share the same evolutionary history.

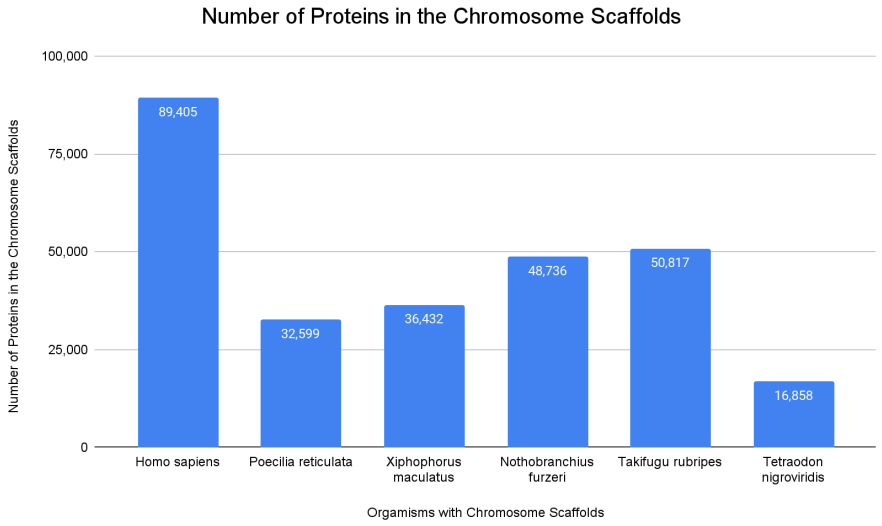



**Figure 4. Organism that Exhibits Chromosome Scaffolds in their Genome and Number of Protein Present**

VII. Topic_05B. If your genome does have Chromosome Scaffolds, what percentage of the Canonical proteins are present in those Chromosome Scaffolds?

a. Galaxy was used to answer this question. The history names are 'Topic_05B.'
   i. Utilize output of Topic_05A grep of annotated chromosomes to search for "Ensembl_canonical"
   ii. Repeat the steps from Topic_04A to extract transcript ID and perform Advance Grep.
   iii. Convert the Tabular Files to Fasta files

b. Results (see Figure 5)
   i. About 83.8% of *Tetraodon nigroviridis's* proteins in their chromosome scaffolds are canonical proteins. *Homo Sapiens* had about 22.1% of canonical proteins in those chromosome scaffolds. The other organisms' percentages of canonical proteins range from 40.7% to 66.2%.

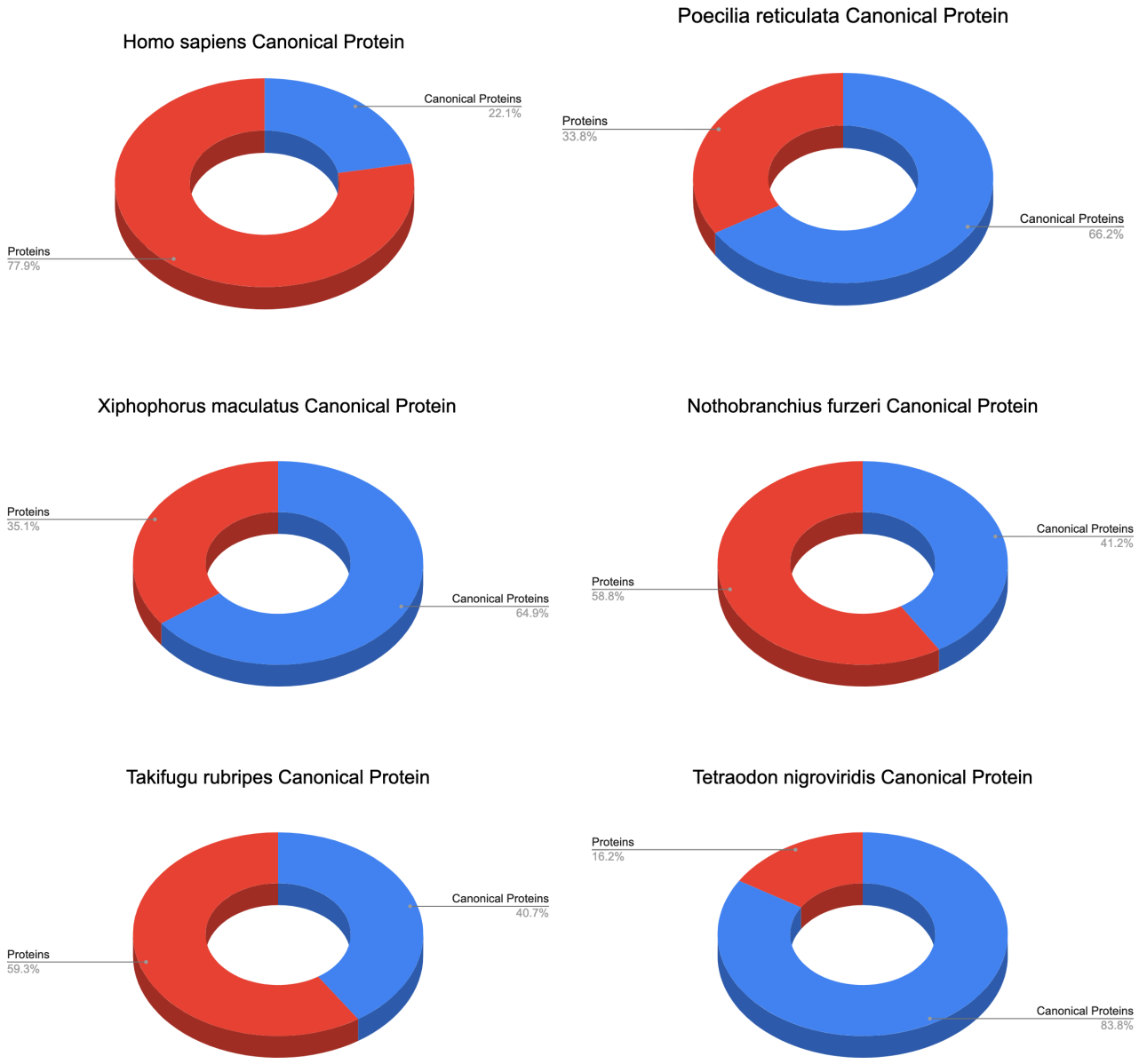



**Figure 5. Percentages of Canonical Proteins Present in each Organisms' chromosome Scaffolds**

VIII. Topic_06A. If your genome does have Non-Chromosome Scaffolds, how many proteins are present in those Non-Chromosome Scaffolds?

A. Galaxy was used to answer this question. The history names are 'Topic_06A.'

1. Utilizing the result from Topic_01, we understand that all organisms exhibit non-chromosomal scaffolds in their gff3 files.
2. Follow the step from Topic_05A, but instead of searching for annotated chromosome scaffolds we are grepping non-chromosomal scaffolds.

3. Then follow the rest of the steps from Topic_04A to extract transcript ID and perform Advance Grep.
4. Convert Tabular files to Fasta files

IX. Topic_06B. If your genome does have Non-Chromosome Scaffolds, how many Canonical proteins are present in those Non-Chromosome Scaffolds?

A. Galaxy was used to answer this question. The history names are 'Topic_06B.'

1. Utilize output of Topic_06A when grepping for non-chromosomal scaffolds to grep for "Ensembl_canonical"
2. Repeat steps from Topic_04A to extract transcript ID to run Advance Grep.
3. Convert Tabular files to Fasta files

B. Results for Topic_06A&B (see Figure 6)

1. *Homo sapiens* have significantly fewer proteins in their non-chromosomal scaffolds compared to other organisms, and most of the proteins found are canonical. *Poecilia formosa*, *Poecilia latipinna,* and *Kryptolebias marmoratus* genomes' did not contain specific annotated chromosomes but had the most proteins in non-chromosome scaffolds compared to organisms that exhibited annotated chromosomes.

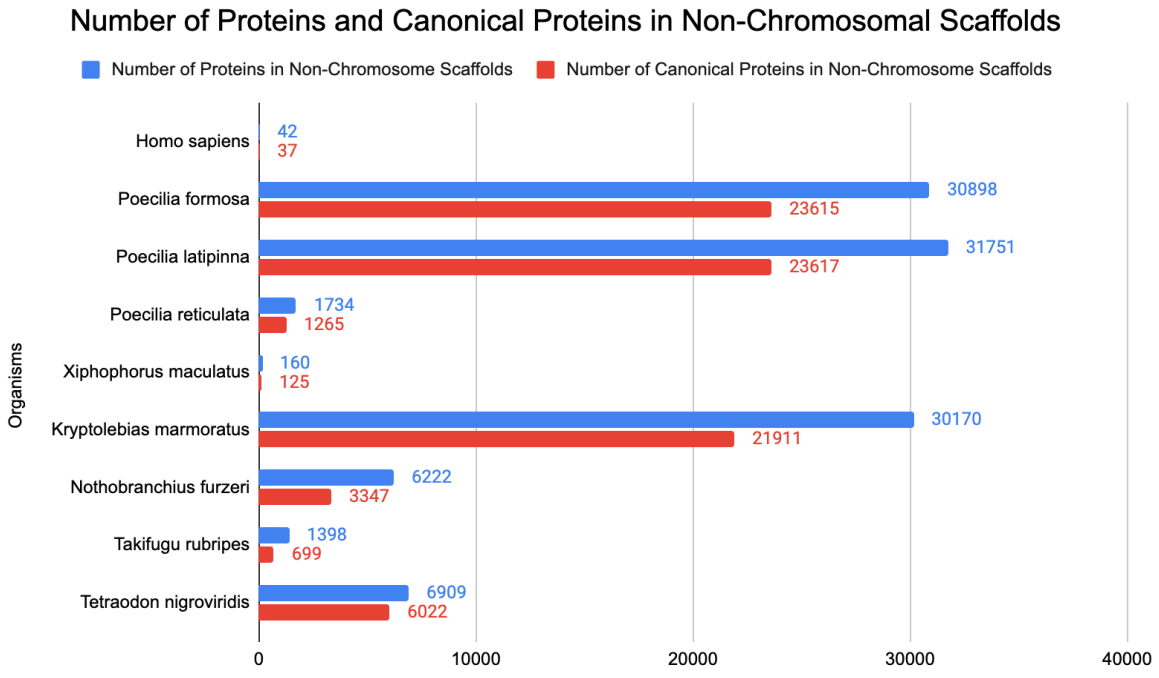



**Figure 6. Comparison of The Number of Proteins and Canonical Proteins in each Organisms Respective Genomes**

X. Topic_07A. Can you find Alternative (Alt) Scaffolds associated with the genomes that were assigned to you, and if you do, how many proteins are associated with those Scaffolds?

A. Galaxy was used to answer this question. The history names are 'Topic_07A.'

1. Upload the gff3 file and pep.all file of Homo sapiens

2. Convert the Fasta pep.all file to a Tabular file
3. In the GFF3 file, search in textfiles (grep) utilizing Perl to capture lines containing:
    a) “\tmRNA\t”
        (1) Capture only lines annotated as mRNA types
    b) “biotype=protein_coding”
        (1) Searches for protein coding from the previous step’s output
4. Extract Transcript IDs
    a) Cut Column 9 from the Grep of “biotype=protein_coding”
    b) Convert all colons from the output of the previous step to TABS
    c) Utilizing “Text transformation” remove instances of “;Parent=gene//” (s/;Parent=gene//) from the previous step’s output
    d) Cut column 2 from the output of the previous step to isolate the Transcript ID
5. Perform Advance Grep
    a) Use the Tabular pep.all file (input file) and isolated Transcript ID from the previous step (pattern file)
6. Compare two Datasets
    a) Compare the tabular pep.all file against Advance Grep file using column 1. To find non-matching lines in the pep.all file.
    b) Convert the Tabular file to a Fasta file

Topic_07B. Can you find Alternative (Alt) Scaffolds associated with the genomes that were assigned to you, and if you do, how many Canonical proteins are associated with those Scaffolds?

B. Galaxy was not used to answer this question.
C. Results for Topic_07A&B
    1. *Homo sapiens* had 32,319 proteins associated with Alt scaffolds. Alternative scaffolds are not present in any of the other genomes. Canonical proteins are not present in Alt scaffolds. However, this might have been due to limitations in the annotation of the genome. Canonical is one representative transcript for each locus that can code for multiple polypeptides. Alt scaffolds are not part of the reference assembly to be considered if it is canonical.

XI. Topic_08A. How many tRNAs can you detect in the Scaffolds that have been declared for each one of the genomes that have been assigned to you?

A. Galaxy was used to answer this question. The history names are ‘Topic_08A.’
    1. Download toplevel.fa file for each organism
        a) Top level file contains all chromosome sequences and unplaced/unlocalized Scaffolds. (Primary assembly only contains placed/annotated chromosomes)
    2. Run tRNA predictions
        a) Search for Eukaryotic tRNA
    3. Search in Textfiles (grep)
        a) Search for data lines that contains (match) “tPseudo” (\tPseudo\t) in the output Tabular file
        b) Search for data lines that does not contain (Don’t Match) “tPseudo”

4. Group Data
    a) Group dataset from the search for lines not containing “tPseudo.”
        (1) Use “Count” operations to group column 1 (chromosome name)
        (2) Use “Count” operation to group column 5 (amino acids)
5. Sort data
    a) Sort dataset output from grouping of column 1 in descending order
    b) Sort data set from count of grouping of column 5 in descending order
6. Examine dataset
    a) Are all 20 amino acids present
    b) which chromosome is associated with the highest number of tRNA
    c) Which amino acids is associated with highest number of tRNAs

B. Results (see Figure 7 and Table 2)

1. All 20 amino acids were present in every genome.
2. *Homo sapiens* (861) and *Tetraodon nigroviridis* (705) have the fewest tRNA compared to the other organisms. *Poecilia reticulata, Xiphophorus maculatus,* and *Nothobranchius furzeri* had the most tRNA, around 4605-4734 tRNAs.
3. For *Homo sapiens,* Ala was associated with the highest number of tRNA. Also chromosomal Scaffold 6 was associated with the highest number of tRNA. The other organisms associated the amino acids Val, Ile, Gly, Arg, Lys, and Leu with the highest number of tRNA in their respective scaffolds.

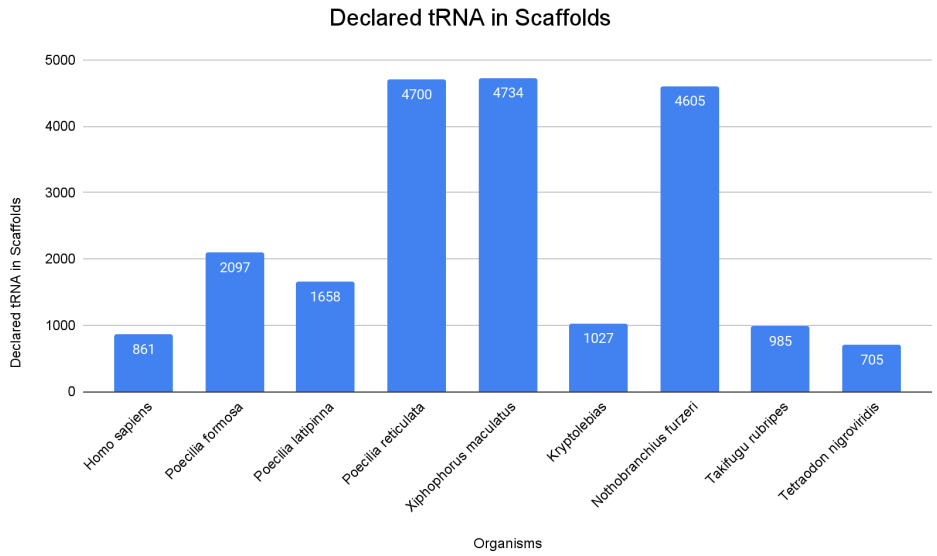



**Figure 7. Number of tRNAs Detected in the Scaffolds for each Genome**

**Table 2. Analysis of tRNA Prediction Results**

| Organisms | Amino Acid with the Highest Number of tRNA | Scaffolds with the Highest Number of tRNA |
|---|---|---|
| *Homo sapiens* | Ala | 6 |
| *Poecilia formosa* | Val | KI520089.1 |
| *Poecilia latipinna* | Val | KQ549592.1 |
| *Poecilia reticulata* | Ile | LG17 |
| *Xiphophorus maculatus* | Gly | 23 |
| *Kryptolebias marmoratus* | Arg | LWHD01000011.1 |
| *Nothobranchius furzeri* | Lys | sgr04 and sgr06 |
| *Takifugu rubripes* | Val | 15 |
| *Tetraodon nigroviridis* | Leu | 2 and 7 |

XII. Topic_08B. How complete are your Canonical proteomes? How many complete BUSCOS can you find? How many incomplete BUSCOS can you find? How many duplicated BUSCOS can you find? Does the number of duplicated BUSCOS correlates with the known genome duplication history of the organism in question?

A. Galaxy was used to answer this question. The history names are 'Topic_08B.'
   1. Uploaded gff3 files was used as reference only
   2. Upload Topic_04B's output Advance Grep Fasta file
   3. Run Busco on annotated gene set (protein) on that Fasta file
      a) Run both on lineage Eukaryota and Metazoa
   4. Use BUSCO plot tool and plot both short summary data

B. Results (Figure 8 and 9)
   1. In the BUSCO Assessment of the Eukarytoa lineage, *Poecilia latipinna* and *Takifugu rubripes* have almost double the duplicates that *Homo sapiens* have, 12 and 7, respectively. This could be evidence that genome duplication events took place. Only *Homo Sapiens* has complete BUSCO results without fragments or missing BUSCOs in both lineages. This matches the known genome duplication history of the organisms.
   2. In the BUSCO Assessment of the Metazoa lineage, *Poecilia latipinna* and *Poecilia reticulata* had close to double *Homo sapiens* duplicates. We could assume that genome duplication events could have occurred.This matches the known genome duplication history of the organisms. This matches the known genome duplication history of the organisms.

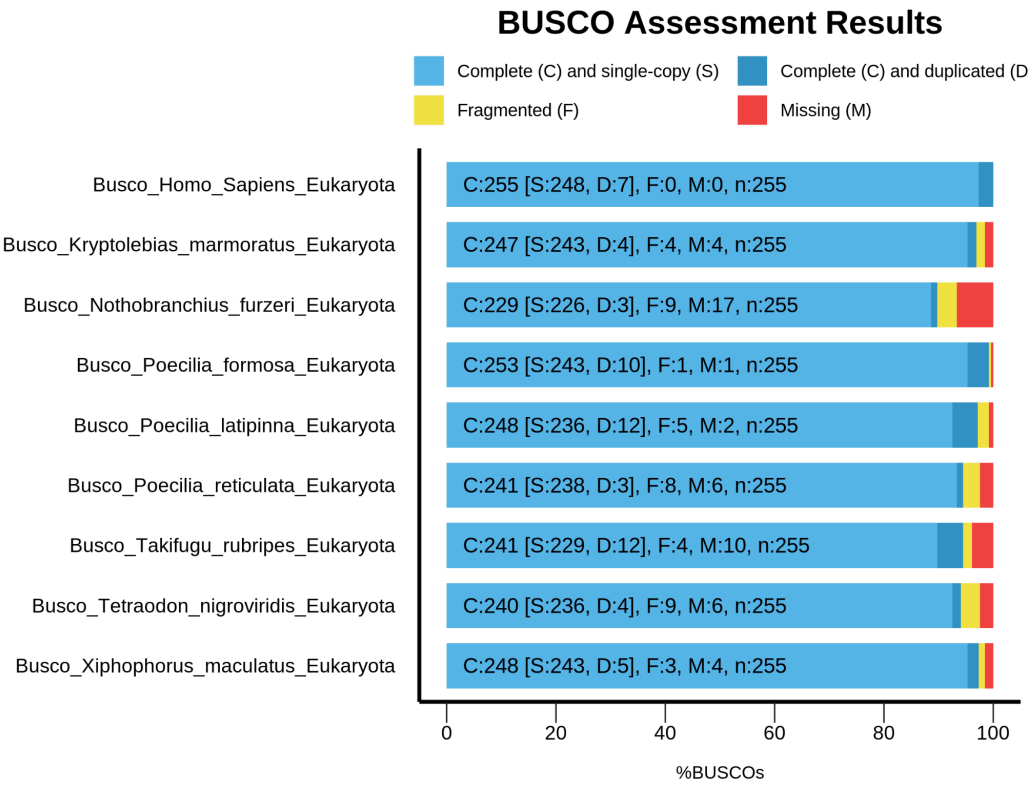



**Figure 8. Busco Assessment Results From Eukaryota Lineage**

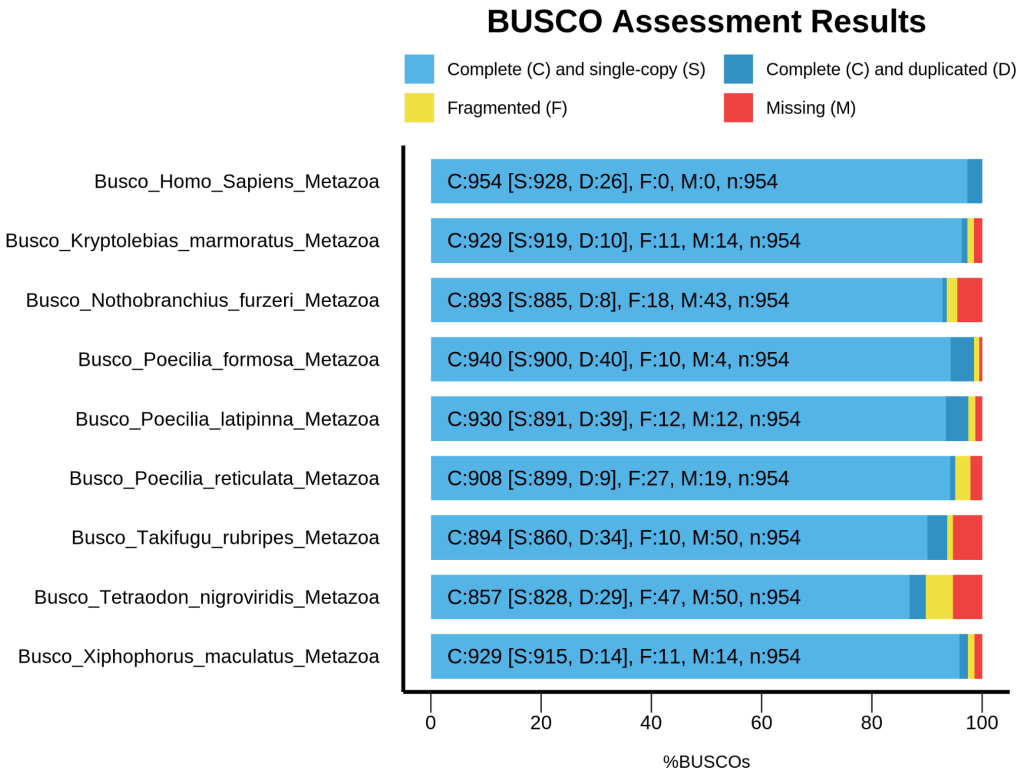



**Figure 9. Busco Assessment Results From Metazoa Lineage**

XIII. Topic_09A. What is the degree of duplication/redundancy present in your proteomes? Using the Reverse-Blast-Hit Galaxy tool (100%Identity and 100% length), or the CD-HIT-Protein Galaxy tool (100%Identity and 100% length), can you reduce the complexity of your proteomes? What do you get, in terms of proteome reduction?

A. Galaxy was used to answer this question. The history names are 'Topic_09A.'
   1. Upload Advance Grep fasta file from Topic_04A
   2. Run CD-HIT PROTEIN tool with 100% identity and 100% length
      a) Similar threshold: 1.0
      b) Length difference cutoff: 1.0
      c) Yes on print alignment overlap in . clstr file and slow cluster
      d) All other setting are default

XIV. Topic_09B. What is the degree of duplication/redundancy present in your Canonical proteomes? Using the Reverse-Blast-Hit Galaxy tool (100%Identity and 100% length), or the CD-HIT-Protein Galaxy tool (100%Identity and 100% length),, can you reduce the complexity of your Canonical proteomes? What do you get in terms of proteome reduction?

A. Galaxy was used to answer this question. The history names are 'Topic_09B.'
   1. Upload Advance Grep fasta file from Topic_04B
   2. Run CD-HIT PROTEIN tool with 100% identity and 100% length
      a) With the same settings as in Topic_09A

B. Results for Topic_09A&B ( Figure 10 & 11 and Table 3 & 4):
   1. *Homo sapiens* had the highest percentage of duplication in their proteome at 15.5% compared to the other organisms that ranged from 0.30% to 3.65%. However, *Homo sapiens* has a significantly lower percentage of duplication in canonical proteome at 1.19%. *Tetraodon nigroviridis* has one of the lowest degrees of duplication in both the proteome and canonical proteome. Comparing the degree of duplication in both proteome and canonical proteome, the percent of reduction is significantly lower for canonical proteins.

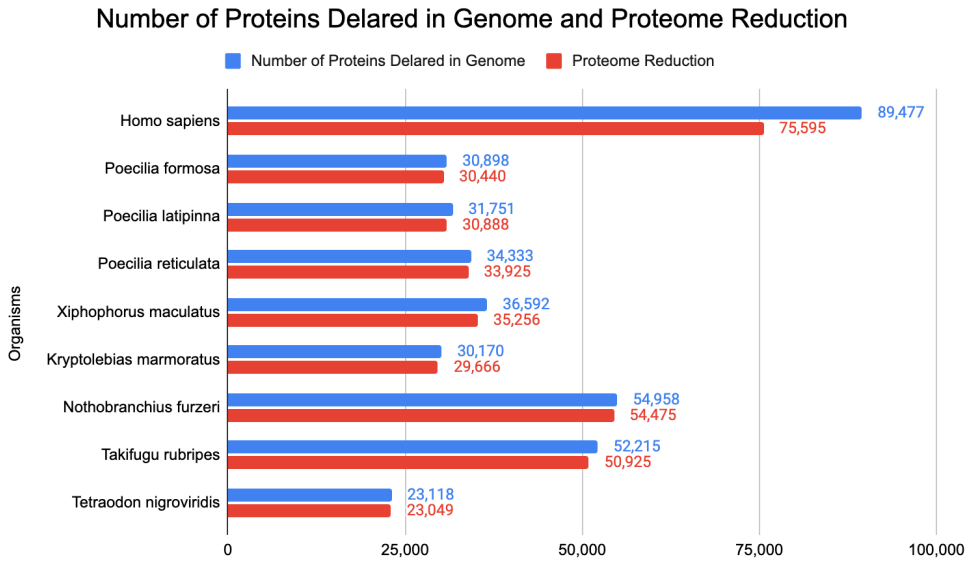


**Figure 10. Number of Protein in Genome and Reduction of Duplicates in Proteome**

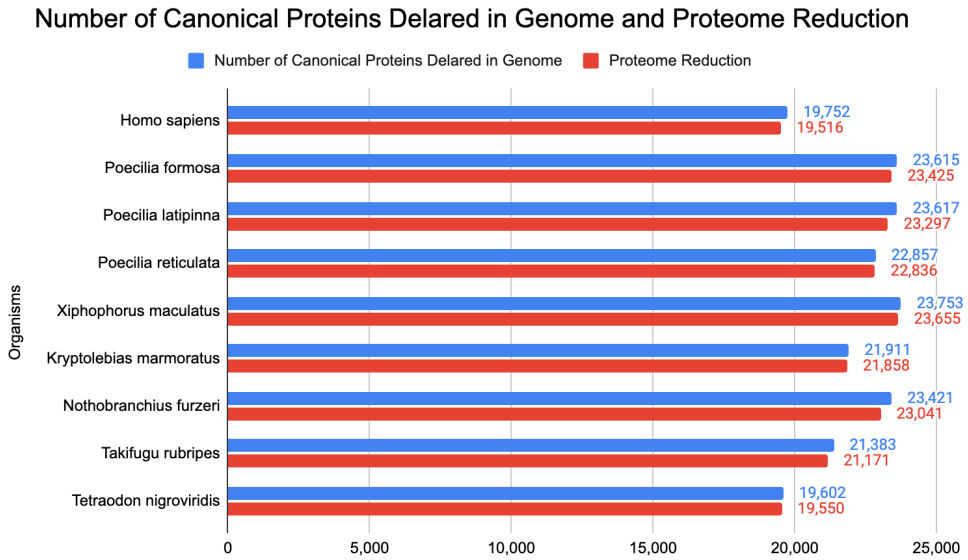



**Figure 11. Number of Protein in Genome and Reduction of Duplicates in Canonical Proteome**

**Table 3. Degree of Duplication/Reduction in Proteome**

| Organisms | Degree of Duplication in Proteome |
|---|---|
| *Homo sapiens* | 15.50% |
| *Poecilia formosa* | 1.48% |
| *Poecilia latipinna* | 2.71% |
| *Poecilia reticulata* | 1.19% |
| *Xiphophorus maculatus* | 3.65% |
| *Kryptolebias marmoratus* | 1.67% |
| *Nothobranchius furzeri* | 0.88% |
| *Takifugu rubripes* | 2.47% |
| *Tetraodon nigroviridis* | 0.30% |

**Table 4. Degree of Duplication/Reduction in Canonical Proteome**

| Organisms | Degree of Duplication in Canonical Proteome |
|---|---|
| *Homo sapiens* | 1.19% |
| *Poecilia formosa* | 0.80% |
| *Poecilia latipinna* | 1.35% |
| *Poecilia reticulata* | 0.09% |
| *Xiphophorus maculatus* | 0.41% |
| *Kryptolebias marmoratus* | 0.24% |
| *Nothobranchius furzeri* | 1.62% |
| *Takifugu rubripes* | 0.99% |
| *Tetraodon nigroviridis* | 0.27% |

XV. Topic_10. How related to each other are your different Canonical proteomes? Can you identify a set of proteins that are common to all your proteomes? Can you identify a set of proteins that are unique to each proteome?

A. Galaxy was used to answer this question. The history names are 'Topic_10.'

1. Upload all representative.fasta files from Topic_09B
2. Concatenate all the files
3. Run CD-HIT PROTEIN on concatenate dataset
    a) Similar threshold: 1.0
    b) Length difference cutoff: 1.0
    c) Yes on print alignment overlap in . clstr file and slow cluster
    d) All other setting are default
4. Search in textfiles (or grep) CD HIT PROTEIN .clstr file with lines that contains:
    a) "^>Cluster" or "^0"
        (1) Representative sequence mix of clustered and unclustered sequence
    b) "^1"
        (1) Clustered that contain two or more sequence
    c) "^2"
        (1) Clusters that contains three or more sequence
    d) "^3"
        (1) Clusters that contains four or more sequence
    e) "^4"
        (1) Clusters that contains five or more sequence
    f) "^5"
        (1) Clusters that contains six or more sequence
    g) "^6"
        (1) Clusters that contains seven or more sequence
    h) "^7"
        (1) Clusters that contains eight or more sequence
    i) "^8"
        (1) Clusters that contains nine or more sequence

j) “^9”

(1) Since we started with 9 proteome so there cannot be more than 9 proteins per cluster

B. Results: (Table 5)

1. There were 194,728 representative sequences. About 1.44% of clustered proteins have clusters with 2 or more proteins. Only 0.0005% of clustered protein has 9 proteins, so it is safe to assume that this one protein is unique to *Homo sapiens*. As *Homo sapiens* is distantly related to the different fish species. Since we started with 9 proteomes, we cannot have more than 9 proteins per cluster, so the output of grepping ‘^9’ is zero. Proteins that did not cluster are considered unique to each proteome. Sets of proteins that cluster together are considered common or conserved in the different proteomes.

**Table 5. Clusters Percentages and Composition from CD-HIT PROTEIN**

| Cluster Composition | Percentage of Clustered Proteins |
|---|---|
| 2 or more proteins | 1.44 |
| 3 or more proteins | 0.27 |
| 4 or more proteins | 0.095 |
| 5 or more proteins | 0.028 |
| 6 or more proteins | 0.014 |
| 7 or more proteins | 0.0046 |
| 8 or more proteins | 0.0026 |
| 9 proteins | 0.0005 |

**References**

Watson, C. T., Gray, S. M., Hoffmann, M., Lubieniecki, K. P., Joy, J. B., Sandkam, B. A., Weigel, D., Loew, E., Dreyer, C., Davidson, W. S., & Breden, F. (2011). Gene duplication and divergence of long wavelength-sensitive opsin genes in the guppy, Poecilia reticulata. *Journal of molecular evolution*, *72*(2), 240–252. https://doi.org/10.1007/s00239-010-9426-z

Kai, W., Kikuchi, K., Fujita, M., Suetake, H., Fujiwara, A., Yoshiura, Y., Ototake, M., Venkatesh, B., Miyaki, K., & Suzuki, Y. (2005). A genetic linkage map for the tiger pufferfish, Takifugu rubripes. *Genetics*, *171*(1), 227–238. https://doi.org/10.1534/genetics.105.042051