Homework 4

Explore using AI for coding. Develop a prompt that addresses a question you find interesting or use one of my suggestions in the class module.

Provide the prompt and/or intended task you want to solve, the code you created and the solution with test examples of data or use of it.

This is the course assignment to accept. This will create an empty repository. Include a README.md where you can document your prompt and how you would use the code and test data. Include your script(s) for your problem, and some test data.

Example of questions you can focus on (pick one or pick your own, if I give you just one problem and everyone uses AI the results are just you feeding my question into the AI prompt, so develop your own twist on the question, test data, etc to think through what would be an interesting component to find).

1. Generate summary statistics for GTF or GFF files on exon size, intron size, frequency (eg number exons per gene on average), length of intergenic space (eg what is the mean size of the distance between genes). Develop this code then test it on some different organisms (get the genomes from NCBI genomes has been shown or other GFF sources). Ideally develop a summary table of these summary statistics for different organisms and compare how bacteria, animals, plants,viruses, fungi, may all have different summary statistics about how long genes are, how close they are in genomes, etc.

NCBI Genome is here https://www.ncbi.nlm.nih.gov/datasets/genome/. - you might also look at other sources for large scale data download if you want to explore https://ensemblgenomes.org/.

2. develop a sequence simulator and test for patterns of sequences frequencies compared to real genome data
3. how far apart are SNPs in a VCF (eg distance between) are the distances random or are they skewed? Or how many SNPs are are found within genes vs not within genes, are the frequencies same or different for those types of genomic regions adjusting for their size in the genome (you need a VCF file and a genome GTF/GFF file)
4. construct a simple set of questions that might relate to data you are working on. These could be super simple like compute the area of rectangles provided based on a set of coordinates in a table/tsv/csv file ... I leave it up to you to challenge yourself to just try out AI coding and report back on something you tried that worked and how you determined if it was useful or not.