Challenge Problems (CPs)

Every Tuesday at the start of class, you have the opportunity to submit a new challenge problem, a revision to a previous challenge problem, or both. Remember that by the end of the course, you will need to submit 7 CPs for the A-range, 5 CPs for the B-range, and 4 CPs for the C-range. I recommend completing at least three CPs to satisfaction by the end of Week 8.

I will update this page with challenge problems every few weeks, as we cover material associated with a challenge problem. There should ultimately be ~9 CPs to choose from. Submission instructions will be written with each problem (either an electronic upload or turning in a hard copy). Some additional things to note:

1. CP1 MUST be submitted by Tuesday of Week 3. After that, it is up to you which order you choose to complete CPs.
2. You are NOT allowed to discuss challenge problems with anyone else but your instructor.
3. You are allowed to use your notes, any course materials and external, online resources as is useful.
4. A ✓ is sufficient for a satisfactory submission, but I will comment what is incorrect, and you can still revise for a ✓+. It is recommended you try to get some ✓+'s to help make the argument at the end of the semester for a higher grade than the minimum C-/B-/A-.
5. It is often good to start early, so you can see what you need to do for a CP and schedule office hours if you have questions. Note that occasionally, some questions will require that you message me to get the problem. CPs usually take around 2-3 hours, and can take longer if you are not comfortable with the related content/skill.

Challenge Problem #1 (MUST be completed by Week 3T, or Sep 6, at 12 pm EST)

The goal of this challenge problem is to help you learn how to search for research articles. This will be important as we start looking for primary research for our research project. Here, we will familiarize you with the UR library resources and help you all master finding relevant literature. Successful completion of this problem helps you practice learning objectives D3, D4, and D5.

Submission Instructions: A hard copy of your responses to this worksheet. Worksheet must ALSO be submitted electronically using the 'Assignment Upload' link on BB. Name your file <YourName>_CP1.docx.

Part A: This link is the BIOL199 Library Resources page, the first stop for BIOL199 students. Your first activity is to go through the following resources. After working through all these materials, use the information you learned to answer Part A questions in the worksheet.

- Work through the short video link in the middle box of the 'Welcome' tab.
- Read and explore the information in the 'Annotated Bibliography' link.
- Read and explore the information in the 'Using Citation Linking' tab.

Part B: Find **FOUR** articles satisfying the criteria below and place them in the Part B table with the requested information. Usually, a careful read of the abstract and investigating the site on which the journal article is hosted should be enough to check if the criteria are satisfied, but you may optionally examine more of the article as of interest or use to you. The 'Finding Articles' tab and 'More Assignment Help' on the library webpage may be of use, and generally, I find Google Scholar or Web of Science Core Collection great databases to search through. A recorded lecture (~10 mins) on doing searches for scientific articles is provided here. It starts on the 'Finding Articles' tab of the BIOL199 Library Resources Page.

1. Article 1: Primary research article studying *Drosophila* (fruit flies) using genomic sequencing
2. Article 2: Primary research article studying an infectious disease of your choice using genomic sequencing
3. Article 3: Review article discussing next generation sequencing
4. Article 4: Primary research article on a topic of your choice, so long as it is studying a biological question - in your sentence(s), make sure to explain your chosen topic

Part C: Throughout this course, we will use the Pechenik citation format. Use Chapter 5 of Pechenik to determine how to cite your four articles, placing these citations below the table. Make sure your citations are listed in alphabetical order by first author's last name. For studies with more than 5 co-authors, list the first five co-authors, followed by ellipses, and then the last co-author (see below the table in the worksheet for an example).

Challenge Problem #2 (skills/data introduced by end of Week 2 lab)

A FASTQ file with the name ERR1679631_1.fastq.gz is on the Spydur cluster in the folder /scratch/myang_shared/classes/BDB_F22/data/. Copy this file into your personal folder on Spydur. Then, complete the following questions using Linux commands, the HTML file generated from an analysis using the software *fastp*, and the provided links. Successful completion of this problem helps you practice learning objectives B2, C2, D3, and D5. Steps 3 and 4 are best done after completing CP1.

Submission Instructions: A hard copy of your responses to the following questions. The responses can be hand-written or typed. Submit at the start of a Tuesday class. Submitting later in the week will count towards the following Tuesday's submission, unless you use the appropriate number of tokens.

1. Complete the table below:

Size of Zipped File		**Size of Unzipped File**	
# of lines BEFORE filtering low quality reads		**# of lines AFTER filtering low quality reads**	
# of reads BEFORE filtering low quality reads		**# of reads AFTER filtering low quality reads**	

2. Write one KEY result, describing the difference in read quality before and after filtering out low quality reads using fastp (include numerical values, e.g. percent change). Should be no more than 1-2 sentences.
3. The above data was downloaded from the European Nucleotide Archive (ENA), a database to upload new data from peer-reviewed research articles for access by the greater community. The project from which the FASTQ file came has the accession ID: **PRJEB15857**. Click on the link for the ENA database, and use the top right search bar to search for this accession ID. Find the associated project, and use it to identify the primary research article. Find the primary research article and write a 1-2 sentence summary of the research project, making sure to identify the species whose data you just examined.
4. Cite the primary research article using the Pechenik reference/citation format.

Challenge Problem #3: Adding Significance (Statistically Speaking) (needed skills/data introduced by end of Week 3 Lab)

Successful completion of this problem helps you practice learning objectives B4, C4, and D2.

Submission Instructions: A hard copy of your responses to the following questions. The responses can be hand-written or typed. Submit at the start of a Tuesday class. Submitting later in the week will count towards the following Tuesday's submission, unless you use the appropriate number of tokens.

Psoriasis is a skin disease that results in red, itchy, and scaly patches of skin. It runs in families, suggesting a genetic association. There is a region of the genome (8p23.1) associated with a protein called beta-defensin that is known for variable copy numbers - that is, a different number of repeats of a particular DNA sequence, depending on the individual. Hollox et al. (2007) determined the number of repeats (i.e. copy number) in this region for 451 Dutch individuals, dividing the Dutch individuals into those with psoriasis (cases) and those without (controls). They asked whether this region of the genome (8p23.1) is associated with having psoriasis. Answer the following questions related to this study.

1. Construct a hypothesis addressing the question posed by Hollox et al. (2007).
2. After determining copy numbers for each individual in the study, they proceeded to statistically test whether there was a difference in copy number variation at the 8p23.1 region between those with psoriasis (cases) and those without psoriasis (controls). Indicate the null and alternative hypotheses associated with this statistical analysis.
3. You can do this analysis too! I have prepared two TXT files containing the Dutch case data and Dutch control data, where each row indicates the copy number for one individual. Include the following statistics in a nicely formatted table (fit for publication): average mean and standard deviation per dataset, difference in average copy number between two datasets, and the associated t-value, degree of freedom, and p-value. Additionally, upload your Microsoft Excel work onto Box, with the filename 'Name_CP3_Excel.xlsx', so I can check your computations as needed.
4. Write one sentence indicating your key result, making sure to include numerical values to help support your result.
5. Write one sentence interpreting your result, connecting back to the original question asked by the researchers.
6. Cite the primary research article using the Pechenik reference/citation format.

The associated study is linked here, with a PDF copy available here. You should not need this link, but it is provided in case you're interested.

Challenge Problem #4: From Similarities to Inferring Homologs (need skills introduced by end of Week 4 Lab)

Successful completion helps you practice learning objectives B2, B4, C2, C3, D3, and E3.

Submission Instructions: A hard copy of your responses to the following questions. The responses must be typed. Submit at the start of a Tuesday class. Submitting later in the week will count towards the following Tuesday's submission.

On Spydur, I've placed a small FASTA file in /scratch/myang_shared/classes/BDB_F22/data/unknown.fna. Download this file and use BLAST tools to compare this sequence to databases of reference RNA sequences and proteins to help you examine the sequence inside. (You will not use all BLAST tools used in Week 4 Lab.)

1. Describe your objective and the general methodology you used. Make sure you describe enough of your methodology such that a reader lightly familiar with BLAST could replicate your results.
2. Make a table identifying the BLAST search used, name of gene, species scientific name, species common name, query cover, E-value, percent identity, and accession ID for all hits that are a good candidate. Make sure the table is nicely formatted (see description from Paper1 instructions)
3. What do you think is the name of the gene and the species our sequence was most likely found in? Write a results paragraph with your key result, followed by supporting evidence. For any seemingly contradicting evidence (e.g. the other entries in your table), make sure to give a possible explanation for why they showed up in the search.
4. Briefly describe what you learned about the function of the protein associated with this gene. You don't need to do a deep dive, but come up with a 1-sentence general description.
5. Cite any scientific literature or database that you found and used, outside of NCBI BLAST. If you used an online database, see Pechenik Chapter 5 'Website' format for an example, omitting any information unavailable.

Some food for thought:

- You will very likely need to do some research on your own to identify all of the above pieces. Use google to get you started, but make sure to verify against scientific literature or scientific databases so you can refer to supported literature.
- Remember that the two letters starting the accession ID give you information on whether a sequence has been experimentally verified in the lab, or if it is predicted to exist but not yet verified. We want to use verified sequences when possible, if not available, we can consider predicted sequences as well - make it clear what you're using.
- Make sure to include numbers that might be useful, metrics you may want to use are Query Cover, E-value, and Percent Identity

Challenge Problem #5: Practice Annotating Genes (needs skills introduced by end of Week 6 Lab)

Successful completion of this problem helps you practice learning objectives B2, C2, C3, and D3.

Submission Instructions: Submit a hard copy of your gene model with the information listed below (can be hand- or electronically drawn).

In the UEG module from Weeks 5-6, we learned how to manually annotate the *tra* gene. In this challenge problem, you must manually annotate the **CG32165-RA** isoform, also found on contig1 in the July 2014 (Gene) Assembly for *D. melanogaster*. **Draw a gene model** for CG32165-RA that includes all of the following components (1-6). You may find it useful to make a table to track your phasing/reading frames.

1. CDS, 3'/5' UTR, and intron labels (exon labels not needed because implied with listed labels).
2. The base pairs that indicate the splice donors and splice acceptors for each intron.
3. A bent arrow with TSS label indicating location of TSS and direction of transcription.
4. The coordinates for the beginning and end of all CDS regions.
5. The reading frame for each CDS (perhaps next to the CDS label).
6. The total number of amino acids found in the translated CG32165-RA protein (show me your work! If done in Excel, you can upload onto our 'Assignment Upload' link and note you used Excel in your hard copy).

Challenge Problem #6: Make a meme out of it!

Successful completion of this problem helps you practice many different learning objectives, depending on what you choose to discuss.

Submission Instructions: A hard copy of your summary. The responses must be typed.

I'm not savvy enough to make clever biology memes, but perhaps you are. There are two parts:

1. Create a **meme** that highlights a biological concept/tool/example we learned in class and connect the meme to *at least one learning objective*. Anything from our class and lab topics is fair game. If you have another idea not on that list, feel free to run it by me.

2. In a paragraph, explain how the meme connects to the biological concept/tool/example, making sure to add enough content so a reader who does NOT understand any biology would be able to explain the meme. Remember I am also examining your understanding of that topic, so give me enough information to show your mastery of the topic. **Cite** any sources used using Pechenik format if not from class documents/notes/lecture.

I have provided an example below about human evolutionary genetics, courtesy of a colleague who is more meme-savvy than me. Example paragraph is from me.

Paragraph Description: Denisovans are an archaic human population closely related to Neanderthals. Most of what we know about them is from ancient DNA we sampled from their bones. One interesting finding is that present-day Tibetan populations, who live at very high elevation, possess an allele of the EPAS1 gene that codes for a protein that is adaptive for high altitudes. In a comparison of modern and archaic EPAS1 alleles, it turns out that the adaptive variant in Tibetans is more genetically similar to the Denisovan variant than other variants found in modern humans. This finding suggested to researchers that the high frequency of the adaptive variant in Tibetans is due to gene flow between populations related to Denisovans and Tibetans (since Tibetans are genetically just like other modern humans) and natural selection for the variant in high altitude areas (Huerta-Sanchez et al. 2015). This meme shows the mutation originally occuring in Denisovans, followed by transmission to modern humans (i.e. ancient population that contributed to Tibetans) of the EPAS1 allele, followed by selection on this allele within modern humans, such that no one knew until recently that it originally came from Denisovans. (In fact, we did not even know Denisovans existed until 2010!) This connects to course learning objectives A1 and A4, illustrating an example of evolution that occurred in the past.

Reference: Huerta-Sanchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M.,…, Nielsen, R. 2015. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. 512:194-197.

Challenge Problem #7: PRAs with phylogenies

Successful completion of this problem helps you practice learning objectives A5, B1, D1, D2, E3.

Submission Instructions: A hard copy of your summary. The responses must be typed.

Summarizing primary research articles, with a focus on reading phylogenetic trees. Read one of the following papers (one of Options 1-3) that uses phylogenetic evidence to answer one or more research questions. Write a 2-3 paragraph summary focused on their findings related to their figure with a phylogenetic tree. (Note that this means you do not need to discuss the entire paper in your summary, though I encourage reading the whole paper to decide if there's additional useful information to relay that may help contextualize your summary. The summary should be ~0.5-1 page long, with complete sentences and clear organization. It should be written entirely in your own words. Cite the study at the top of the submitted document using the Pechenik format described in CP1, and don't forget parenthetical citations as appropriate.

Make sure your summary includes the following:

1. The question being asked in the study
2. Background to contextualize the question
3. A brief description of methodology used to make the phylogeny (particularly the data included to analyze the phylogeny and how they made the phylogeny, e.g. software or statistical method).
4. A summary of key results from the phylogeny (While general key results are also important, you must at some point in your summary refer specifically to the figure and tips, accurately describing an evolutionary relationship given in the phylogeny between two or more tips)
5. Interpretation of key results and their importance in broader context of study

Option 1: Baez-Ortega et al. (2019) Somatic evolution and global expansion of an ancient transmissible cancer lineage. Figure 1A (though B-D may prove useful in summary and can be discussed).

Option 2: Sjaarda et al. (2021) Phylogenomics reveals viral sources, transmission, and potential superinfection in early-stage COVID-19 patients in Ontario, Canada. Figure 2.

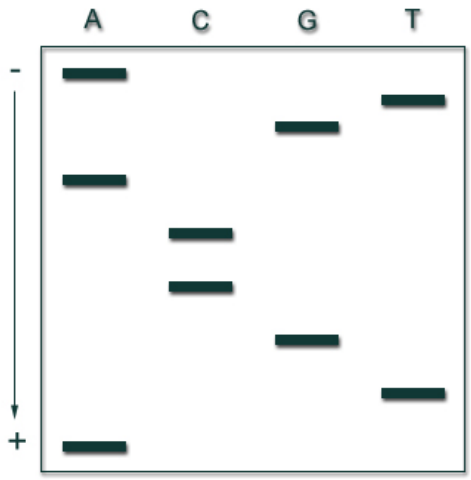
Option 3: West et al. (2017) The Pacific rat race to Easter Island: Tracking the prehistoric dispersal of *Rattus exulans* using ancient mitochondrial genomes. Figure 3A or 4A. (PDF link is to the right, the 'Download Article' button).

Challenge Problem #8: Suddenly, Sanger Sequencing (He purified DNA...Sanger is your man??!)

Successful completion of this problem helps you practice learning objectives B1, B2, B4, D1.

Submission Instructions: A hard copy of your summary. The responses must be typed.

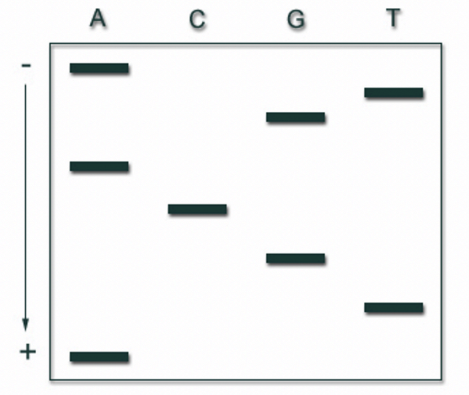
Answer the following FIVE problems thoroughly and accurately to receive credit for this CP. All five problems are related to the image on the right. Kudos to anyone who gets my reference...



1. A DNA fragment in Svante was sequenced using Sanger sequencing. The image to the right shows the resulting gel. **Write a figure title and legend for this image, making sure to clarify all symbols and include all methodology needed to explain Sanger sequencing and the resulting image (it's okay here if the method description runs a bit long).**

2. Indicate the DNA sequence that was analyzed, and clarify <u>which strand was used as the template strand</u>.

3. Assume the sequenced strand shows the beginning of the coding strand of a coding sequence - what are the amino acids observed (place amino acids in order, starting from the first amino acid of the protein sequence.



4. Another DNA fragment was sequenced for the same region in Mel, in the image shown on the left. Examine the difference between this image and the one for Svante - what type of mutation is observed, and what effect do you think it has on the final function of the corresponding protein?

5. Draw a dot plot conveying the difference in **<u>DNA sequence</u>** for Svante and Mel - make sure to label all axes and add units as appropriate. This can be hand-drawn if easy to understand.

Challenge Problem #9 (FINAL CP): Investigating mutations

Successful completion of this problem helps you practice many several LOs, including B2, B3, and D3.

Submission Instructions: A hard copy of your summary. The responses must be typed.

In this challenge problem, you will characterize a DNA sequence from a human patient. To start, **you will need to request a DNA sequence from Dr. Yang by messaging her on Perusall (e.g. 'Please give me a CP9 DNA sequence')**. Your goal is to do a *BLAST* search to identify the protein to which it most likely corresponds in humans (i.e. *Homo sapiens*), using the RefSeq_Protein database. Then, you will examine the alignment for a mutation in your DNA sequence to assess the effect on protein function. You will use UniProt to summarize the protein's function and explore possible effects of the mutation (likely using the 'Function' and 'Disease & Variants' sections). You will write a short summary that contains the following required elements:

A. Methodology you used to identify the protein
B. Identification of protein and your supporting evidence (note that the protein must be tied to experimental rather than predicted data), making sure to indicate the query and subject ranges
C. Identification of amino acid mutation(s) and their position in the subject sequence, along with an explanation of the type of mutation (e.g. from one polar amino acid to another polar amino acid) and your hypothesis of whether the mutation would affect protein function.
D. Summary of protein function, using UniProt and any other resources (make sure to cite those sources) - make sure you are looking at the protein in humans on UniProt (indicate the code or 'Entry name' associated with the UniProt page you examined)
E. Description of genetic disease(s) related to an amino acid mutation in this gene at one of the positions where you identified an amino acid change. Again, any additional sources should be cited.
F. References added should be in the same modified Pechenik format used throughout the course