BGGN-213*

First Year Exam Questions for 2022/2023

Instructions

Save this document to your computer and open it in a PDF viewer such as Preview (available on every mac) or Adobe Acrobat Reader (free for PC and Linux). Be sure to add your name and UC San Diego personal identification number (PID) and email below before answering all questions in the space provided.

Student Name	UCSD PID	UCSD Email
Diwen (Steven) Gan	A59020397	digan@ucsd.edu

Overview:

This open-book, open-notes test consists of 10 required questions and one optional bonus point question (question 11). The number of points for each question is indicated in square parenthesizes at the beginning of each question.

No communication (electronic or otherwise) with your fellow students regarding this test until after the due date.

Please remember to:

- Download the PDF version and open in Preview (Mac) or Acrobat Reader (Windows).
- Type all your answers directly in the space provided below each question.
- Save and upload your completed test to gradescope.

Good luck!

*http://thegrantlab.org/bggn213/

Test Questions:

Visit the following webpage and download your student specific sequences.

> **N.B.** These sequence are unique for you and you must use your sequences to answer the following questions in the space provided.

Q1. [1pt] What protein do these sequences correspond to?

proto-oncogene tyrosine-protein kinase receptor Ret isoform a precursor [Homo sapiens]

Q2. [6pts] What are the tumor specific mutations in this particular case (e.g. A130V)?

K740R, A741Y, L746P, G748V, G751Y, Y752P, T753E, T754E, V778E, N783G, V787R, S795V, G798Y, I803Y, Y806R, A807G, Y809E, R833P, E843Y, D850G, Q857P, I858R, Q863P, Y864G, L865G, N879R, R886G, R908P, G911E, I913P, P914R, S922V, F924P, H926V, T929Y, D933Y, F937P, G938R, W942G, E958V, R959V

Q3. [1pts] Do your mutations cluster to any particular domain and if so give the name and PFAM id of this domain? Alternately note whether your protein is single domain and provide it's PFAM id (e.g. PF02196).

Mutations clustered to: Protein tyrosine and serine/threonine kinase domain (PK_Tyr_Ser-Thr, PF07714)

Q4. [2pts] Using the NCI-GDC list the observed top 2 missense mutations in this protein (amino acid substitutions)?

R133C, M918T

Q5. [2pts] What two TCGA projects have the most cases affected by mutations of this gene?

TCGA-GBM, TCGA-BRCA

Q6. [3pts] List one RCSB PDB identifier with 100% identity to the wt_healthy sequence and detail the percent coverage of your query sequence for this known structure? Alternately, provide the most similar in sequence PDB structure along with it's percent identity, coverage and E-value.

6GL7, 54.49%

Q7. [10pts] Using AlphaFold notebook generate a structural model using the default parameters for your **mutant** sequence.

> Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a "too many amino acids" (i.e. length) error you can focus on the main PFAM domain of interest (your answer to Q3 above).

Once complete save the resulting PDB format file for your records and use Mol-star (or your favorite molecular viewer) to render a molecular figure. In this figure please clearly show your mutant amino acid **side chains as spacefill** and the protein as **cartoon colored by local alpha fold pLDDT quality score**. This score is contained in the B-factor column of your PDB downloaded file. **Upload this image to GradeScope**.

Q8. [2pts] Considering only your mutations in high quality structure regions (with a pLDDT score > 70) are any of the mutations on the surface of the protein and hence have a potential to interfere with protein-protein interaction events? List these mutations below (e.g. A130V)

K740R, A741Y, T754E

Q9. [5pts] Please comment on how useful and/or reliable you think your AlphaFold structural model is for your entire sequence and the main domain where your mutations lie? You may wish to compare your model to the PDB structure you found in Q6.

Q10. [10pts] Obtain the most recently dated COVID-19 Variant Data from the California Health and Human Services (CHHS) open data site.

Upload to gradescope a PDF format report generated from a Quarto or Rmarkdown document that demonstrates reading the above CSV file and generating the below visualization of this data.

> **NB.** You can chose how to make this plot and whether you want to make improvements or stylistic changes. However, you are strongly encouraged to use the ggplot2, lubridate and dplyr packages for this task. Please make sure your name and PID number is on the first page and that your report contains all of your code, text description/narrative text of why you doing a particular task/code chunk and the resulting figure.

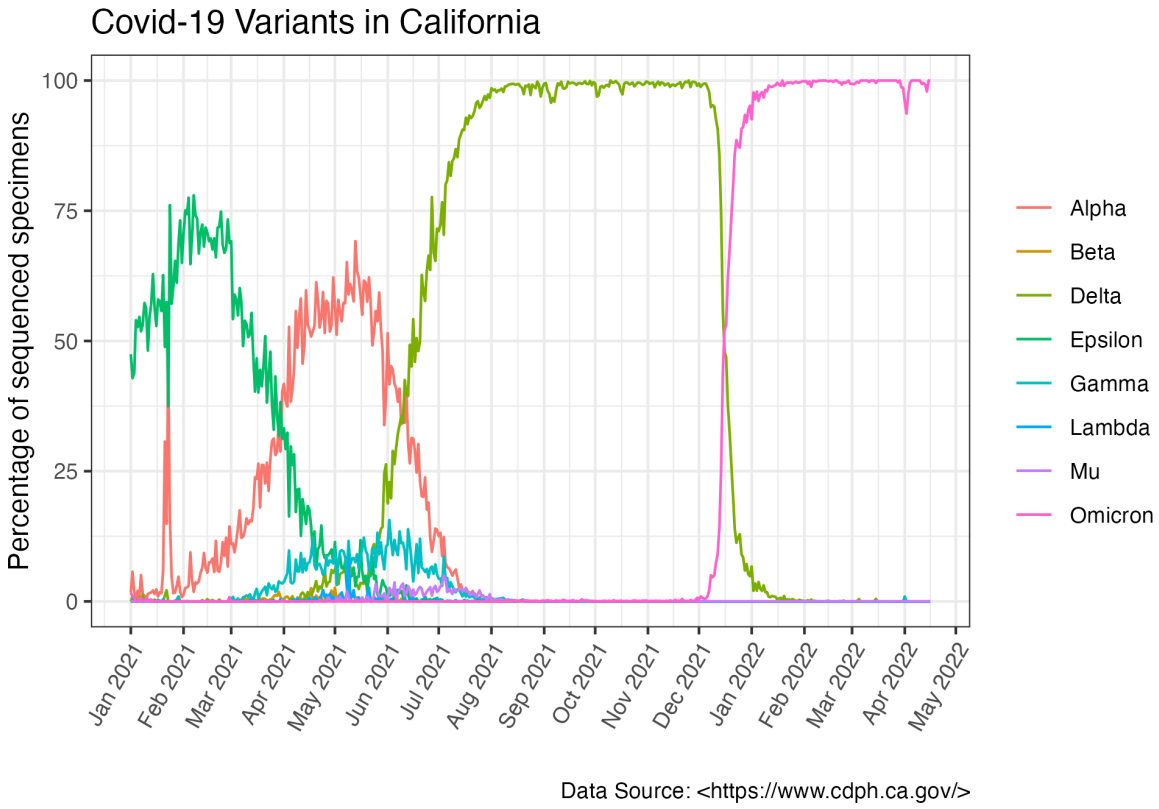

Figure 1: Example plot for Q10.

Q11. [10pts] *Optional: This is not a required question but will yield you 10 extra bonus points.* Using git upload (a.k.a. push) your RStudio project containing your complete work for Q10 to GitHub and provide a link to your project directory here:

https://github.com/StevenVGan/Bioinfo_first_year_final_2023

- End of Test -