Lab 19*

Cancer Mutation Mini-Project

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

Background:

To identify somatic mutations in a tumor, DNA from the tumor is sequenced and compared to DNA from normal tissue in the same individual using variant calling algorithms.

Comparison of tumor sequences to those from normal tissue (rather than 'the human genome') is important to ensure that the detected differences are not germline mutations.

To identify which of the somatic mutations leads to the production of aberrant proteins, the location of the mutation in the genome is inspected to identify non-synonymous mutations (i.e. those that fall into protein coding regions and change the encoded amino acid).

As you go through this mini-project please remember to:

- Download the PDF version of this lab sheet (as noted above).
- Type all your answers directly in the space provided below each question.
- Save and upload your completed PDF to gradescope.

Good luck!

*http://thegrantlab.org/teaching/

Questions:

Visit the following webpage and download your student specific sequences. These sequences resulted from an NGS analysis of patient healthy and tumor tissue.

> **N.B.** Note that these sequence are unique for you and you must download your sequences and use them to answer the following questions in the space provided.

Q1. [1pt] What protein do these sequences correspond to? (Give both full gene/protein name and official symbol).

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

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/accession and name (e.g. PF00613 and PI3Ka).

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

Q5. [2pts] What two TCGA projects have the most cases affected by mutations of this gene? (Give the TCGA "code" and "Project Name" for example "TCGA-BRCA" and "Breast Invasive Carcinoma").

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. Does this structure "cover" (i.e. include or span the amino acid residue positions) of your previously identified tumor specific mutations?

Optional Extension:

The following 3 questions are not required for this lab session but are here for motivated students to see how "druggable hot-spots" near these mutation sites might be identified:

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. **Send this image to Barry for some bonus points**.

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)

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] Visit the FTMap online server and sign-up for an account. Once logged in upload your PDB structure model, provide a job name and submit. Note that that these jobs can take some time (over a day) to run if there are multiple runs scheduled ahead of you.

FTMap aims to identify potential **binding hot spots** on the surface of your protein that can bind with high affinity to ligands, drugs, or other proteins. The server does this by docking

a set of small chemical probes onto the protein surface. The server then clusters the probes based on their spatial proximity. Each cluster represents a potential binding site, with larger clusters generally indicating stronger or more likely binding sites.

> **NB.** Note that in advanced options you can also run in **PPI Mode** with the goal of detecting binding hot spots for protein-protein interactions using an alternative set of parameters. You might wish to try this as a second calculation.

Are any of the identified "hot spots" near your cancer specific mutation sites or the most commonly mutated sites from the NCI-GDC? If so which mutation site(s)?

- End of Lab -