BGGN-213: FOUNDATIONS OF BIOINFORMATICS
https://bioboot.github.io/bggn213_W22/

Test Questions

Instructions: This open-book, open-notes test consists of **15 required questions** and one optional bonus point question (question 16). The number of points for each question is indicated in green font 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.
- Remember that concise answers are preferable to wordy ones.
- Make sure your name, UCSD email and PID number are on the first page.
- Save and upload your completed test to gradescope.

- **Good luck!**

Name: Mirte Ciz Marieke Kuijpers

UCSD email: mkuijpers@ucsd.edu

PID: A59010989

GitHub username: mcmk3

Q1. [10pt] The following sequences resulted from an analysis of a patients healthy and tumor tissue. You can also download these sequences from https://tinyurl.com/mutantseqs

```
>wt_healthy
MGPWSRSLSALLLLLQVSSWLCQEPEPCHPGFDAESYTFTVPRRHLERGRVLGRVNFEDC
TGRQRTAYFSLDTRFKVGTDGVITVKRPLRFHNPQIHFLVYAWDSTYRKFSTKVTLNTVG
HHHRPPPHQASVSGIQAELLTFPNSSPGLRRQKRDWVIPPISCPENEKGPFPKNLVQIKS
NKDKEGKVFYSITGQGADTPPVGVFIIERETGWLKVTEPLDRERIATYTLFSHAVSSNGN
AVEDPMEILITVTDQNDNKPEFTQEVFKGSVMEGALPGTSVMEVTATDADDDVNTYNAAI
AYTILSQDPELPDKNMFTINRNTGVISVVTTGLDRESFPTYTLVVQAADLQGEGLSTTAT
AVITVTDTNDNPPIFNPTTYKGQVPENEANVVITTLKVTDADAPNTPAWEAVYTILNDDG
GQFVVTTNPVNNDGILKTAKGLDFEAKQQYILHVAVTNVVPFEVSLTTSTATVTVDVLDV
NEAPIFVPPEKRVEVSEDFGVGQEITSYTAQEPDTFMEQKITYRIWRDTANWLEINPDTG
AISTRAELDREDFEHVKNSTYTALIIATDNGSPVATGTGTLLLILSDVNDNAPIPEPRTI
FFCERNPKPQVINIIDADLPPNTSPFTAELTHGASANWTIQYNDPTQESIILKPKMALEV
GDYKINLKLMDNQNKDQVTTLEVSVCDCEGAAGVCRKAQPVEAGLQIPAILGILGGILAL
LILILLLLLFLRRRAVVKEPLLPPEDDTRDNVYYYDEEGGGEEDQDFDLSQLHRGLDARP
EVTRNDVAPTLMSVPRYLPRPANPDEIGNFIDENLKAADTDPTAPPYDSLLVFDYEGSGS
EAASLSSLNSSESDKDQDYDYLNEWGNRFKKLADMYGGGEDD

>mutant_tumor
MGPWSRSLSALLLLLQVSSWLCQEPEPCHPGFDAESYTFTVPRRHLERGRVLGRVNFEDC
TGRQRTAYFSLDTRFKVGTDGVITVKRPLRFHNPQIHFLVYAWDSTYRKFSTKVTLNTVG
HHHRPPPHQASVSGIQAELLTFPNSSPGLRRQKRDWVIPPISCPENEKGPFPKNLVQIKS
NKDKEGKVFYSITGQGADTPPVGVFIIERETGWLKVTEPLDRERIATYTLFSHAVSSNGN
AVEDPMEILITVTVQNDNKPEFTQEVFKGSVMEGALPGTSVMEVTATDADDDVNTYNAAI
AYTILSQDPELPDKNMFTINRNTGVISVVTTGLDRESFPTYTLVVQAADLQGEGLSTTAT
AVITVTDTNDNPPIFNPTTYKGQVPENEANVVITTLKVTDADAPNTPAWEAVYTILNDDG
GQFVVTTNPVNNDGILKTAKGLDFEAKQQYILHVAVTNVVPFEVSLTTSTATVTVDVLDV
NEAPIFVPPEKRVEVSEDFGVGQEITSYTAQEPDTFMEQKITYRIWRDTANWLEINPDTG
AISTRAELDREDFEHVKNSTYTALIIATDNGSPVATGTGTLLLILSDVNDNAPIPEPRTI
FFCERNPKPQVINIIDADLPPNTSPFTAELTHGASANWTIQYNDPTQESIILKPKMALEV
GDYKINLKLMDNQNKDQVTTLEVSVCDCEGAAGVCRKAQPVEAGLQIPAILGILGGILAL
LILILLLLLFLRRRAVVKEPLLPPEDDTRDNVYYYDEEGGGEEDQDFDLSQLHRGLDARP
EVTRNDVAPTLMSVPRYLPRPANPDEIGNFIDENLKAADTDPTAPPYDSLLVFDYEGSGS
EAASLSSLNSSESDKDQDYDYLNEWGNRFKKLADMYGGGEDD
```

- What protein do these sequences correspond to?

E-cadherin isoform 1 preproprotein from Homo Sapiens

- What are the tumor specific mutations in this particular case (e.g. A130V)?

Using the multiple sequence alignment tool CLustal Omega from EMBL-EBI reveals a single mutation: D254V.

- List one RCSB PDB identifier with 100% identity to the *wt_healthy* sequence?

6VEL

- Using the NCI-GDC list the observed top 4 missense mutations in this protein?

The 4 top missense mutations are: chr16:g.68810269G>T D254Y 4 of 380 reported cases; chr16:g.68822190C>T A655V 4 of 380 reported cases; chr16:g.68810290G>A E261K 4 of 380 reported cases; and chr16:g.68822082G>A R619Q 4 of 380 reported cases.
Note that there are four other mutations listed that are more common than these four, but none are missense mutations.

Q2. [6pts] Suggest databases to answer the following questions along with the answer you obtain using the *wt_healthy* sequence in **Q1**. Only one database suggestion per question please. The first one has been done for you for demonstration purposes.

- I want to find the genomic location of some human gene?

NCBI GENE database;
Chromosome 16 (16q22.1)

- I want to determine what protein domains my novel protein contains?

Interpro (of EMBL-EBI)
The protein contains several domains including: Cadherin-pro domain, Cadhrein-pro-2 domain, Cadherin-like domain, Cadherin-2 domain and Cadherin-C/Cadherin-cytoplasmic domain.

- I want to find if an NMR structure has been solved for my protein of interest or it's homologues?

PDB search with filter for NMR. The first result is 1SUH (amino terminal domain of E-cadherin).

- I want to find out about the human mendelian disorders associated with this gene?

OMIM, which reveals that linked/associated phenotypes include:
- Blepharocheilodontic syndrome 1
- Breast cancer, lobular, somatic
- Diffuse gastric and lobular breast cancer syndrome with or without cleft lip and/or palate
- Endometrial carcinoma, somatic
- Ovarian cancer, somatic
- {Prostate cancer, susceptibility to}

Q3. [3pts] What is the name of the major bioinformatics file formats used for storing the following biomolecular data:

- Protein (and other) sequence data that is used by many bioinformatics tools?

.fa

- Protein (and other) atomic structure data?

smaller molecules XYZ format (.xyz), macromolecules PDB format (.pdb) or improved PDBx format

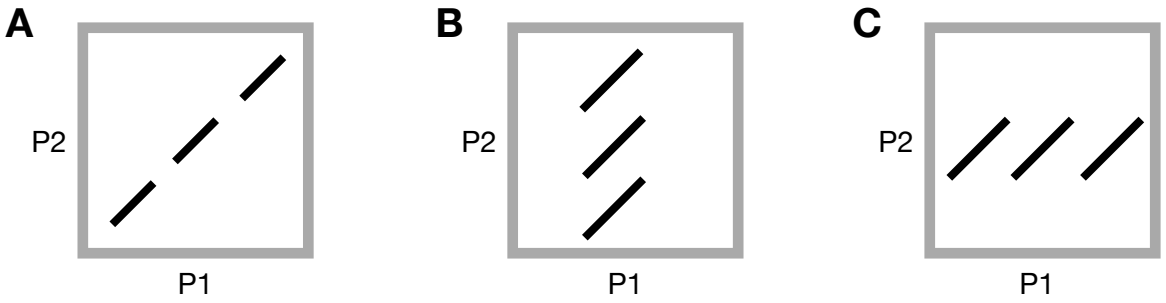
- Sequence reads form NGS?

.fastq for the sequence data plus read quality information

Q4. [1pts] Represent the following three sequences FGMAKLQGD, FGMGRLQGD and GHAMKLQGD (named **seqA**, **seqB** and **seqC**) in the format you answered for the first part of question **Q3**.

>seqA
FGMAKLQGD
>seqB
FGMGRLQGD
>seqC
GHAMKLQGD

Q5. [1pts] Which schematic figure (**A-C**) below best represents a dotplot for two proteins, **p1** and **p2**, whose only significant similarity is a common domain family, and where the domain is occurring once in **p2** and three times in **p1**.



Your Selection (**A**, **B** or **C**):

C

A shows two proteins with several similar domains along their length. B shows

Q6. [3pts] Consider the following multiple alignment of Transcription Factor Binding site DNA sequences

	1	2	3	4	5
Sequence 1	-	T	A	G	C
Sequence 2	C	T	G	G	A
Sequence 3	C	T	A	-	A
Sequence 4	A	T	A	G	T

Give the (average) profile of the above alignment by filling out the table below (overleaf). The first column (i.e. position in the alignment) has been done for your, now complete the rest. You will use this table for answering questions 7-9 below.

	1	2	3	4	5
A	0.25	0	0.75	0	0.5
C	0.5	0	0	0	0.25
T	0	1	0	0	0.25
G	0	0	0.25	0.75	0
-	0.25	0	0	0.25	0

Q7. [2pts] What is the highest scoring sequence match to your profile above (question **Q6**) and what is it's score?

Sequence: CTAGA or if question is asking for match from 1st table in Q6 then sequence 4

Score: 3.5 or if question is asking for match from 1st table in Q6 then 3

Q8. [2pts] Using your completed profile table above (from question 6) score the following two sequences (**S1** and **S2**):

S1. CTAGC: 3.25

S2. AGAGA 2.25

Q9. [2pts] Following the heuristic threshold for a positive match proposed in Harbison et al. [Nature (2004) 431:99-104.] namely using the threshold for a positive match = 60% x Max Score. Are either of the two sequences in question 12 potential transcript factor binding sites? If so why?

Threshold score = 60% x max score = 60% x 3.5 = 2.1. With a threshold score of 2.1 both the sequences are potential transcription factor (TF) binding sites because their scores are > than the threshold score. However, note that there are many potential binding sites in the genome, yet only some are actually bound, generally those near other TF binding sites that thus allow cooperative DNA binding.

Q10. [3pts] Name two major repositories from which R packages can be obtained and the base R function that can be used to install the package bio3d?

CRAN
Bioconductor

instal.packages("bio3d")

Q11. [2pt] What do the Eigenvalues from principal component analysis (PCA) represent and what PC will have the highest Egenvalue?

The eigenvalue represents how much variation a particular PC explains.

PC1 will have the highest eigenvalue as it explains the most variance in the data.

At a more detailed level, eigenvalues are the values within eigenvectors that determine the vector's direction and magnitude. These eigenvectors are produced from the covariance matrix which measures how data points are associated with each other. The eigenvectors help us understand the directions of spread of the data while eigenvalues are a measure of the relative importance of these directions.

Q12. [1pt] List one major advantage of hierarchical clustering over k-means clustering?

Hierarchical clustering does not require one to specify the number of groups they expect, so there is no requirement for prior knowledge or assumptions about the data in this regard. In comparison, k-means clustering requires, as input, a number of groups into which to cluster the data.

Q13. [2pts] Why do researchers often normalize read counts in gene expression studies and what features/effects are they normalizing for?

To make relevant comparisons read counts need to be normalised by gene length and total reads in a sample at a minimum.
Genes can have very different lengths, therefore, if raw read counts aren't normalised by transcript length then long genes will have more counts through random sampling alone because a longer gene will have more chances of an ~150-200 bp read along its length than a short gene. For this reason, researchers usually normalize read counts by the length of the gene or its ORF. One common measure using this approach is FPKM which stands for "fragments per kilobase of transcript per million fragments mapped". Similarly, if one sample has far more reads than another it might seem like expression is higher in that sample, therefore it is important to normalise by total number of reads for a sample as well (this is the per million fragments mapped above).
There are also other forms of normalization, for example, it is common to normalise gene-length normalised counts for genes to one or multiple housekeeping genes or to a spike-in of known concentration to allow comparison of datasets from different experiments.

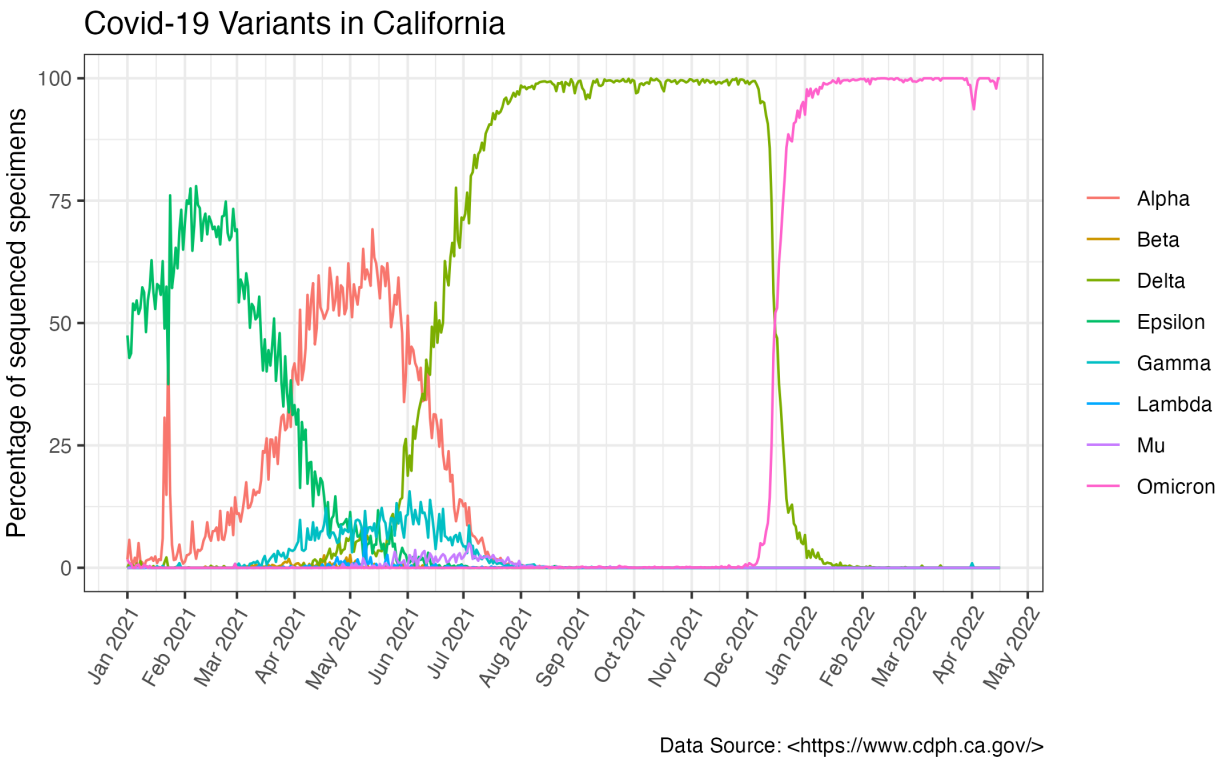
Q14. [2pts] Name two functional databases commonly used for gene set enrichment analysis?

GO database
KEGG database

Q15. [10pt] Obtain the most recently dated **COVID-19 Variant Data** from the California Health and Human Services (CHHS) open data site:
https://data.chhs.ca.gov/dataset/covid-19-variant-data

Upload to gradescope a PDF format report generated from an 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**.



Q16. [10pt] *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 **Q15** to GitHub and provide a link to your project directory here:

https://github.com/mcmk3/BGGN13-First-Year-Exam

—END OF TEST—