MCB 149 Problem Set 1
Fall 2021

Name: ……………………………………..
Student ID #: ……………………………...

You have until **Oct 1 at 12pm PST (noon)** to complete the exam. When finished, please upload your file to Gradescope. Late submission of problem sets will be considered on exceptional circumstances, such as an incapacitating illness or accident or a serious illness or death in the immediate family. Prior permission from the instructors must be obtained before the original submission deadline. The exam is open book, meaning that you can use the course materials, textbooks, publications as well as any information online. However, please work on the questions on your own, without getting input from someone else. If you need clarification, please email the GSI or instructor for further details by email. You can either type your answers electronically or use pen/pencil and scan the material afterwards. If you run into any technical difficulties, please let us know as soon as possible. The space allotted is a suggestion only – please feel free to use any extra space you need. To help us give partial credit, show your work and briefly state any assumptions that you make.

Best of luck!

Question 1 (20) …………
Question 2 (25) …………
Question 3 (20) …………
Question 4 (15) …………
Question 5 (20) …………

TOTAL ……….. / 100

MCB 149 Problem Set 1

This semester we discussed a range of ideas and tools in human genetics that can help us understand population history, adaptation and disease. Let's apply this knowledge to understand the evolutionary origin, function and spread of the COVID-19 pandemic in humans. **Please note the parameter values in the problems do not match the real estimates for SARS-COV-2.** Many parameters related to SARS-COV-2 are still unclear and part of ongoing research projects.

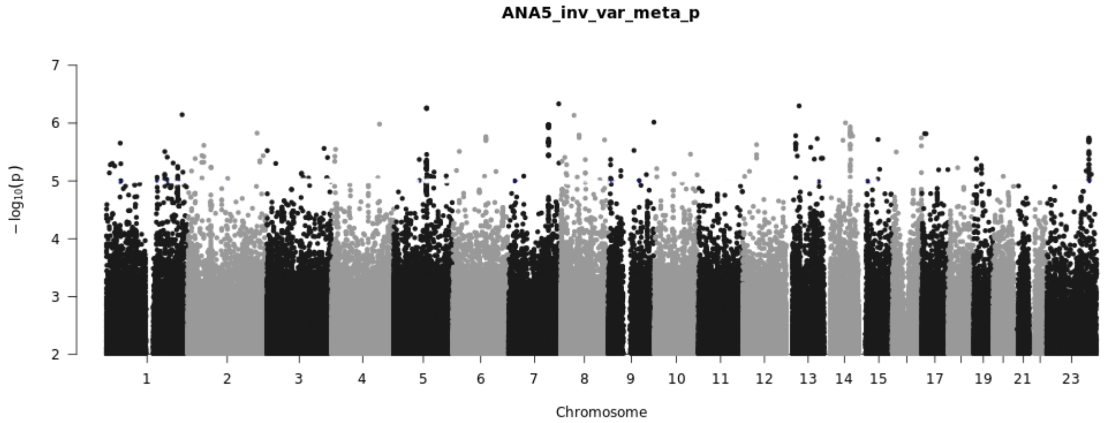
QUESTION 1. Evolutionary forces impacting COVID-19 (20 points)

In order to test if the mutations in coronavirus are neutral, you sequence the Spike (S) protein of SARS-CoV-2 in 5000 New York patients. You focus on one mutation in the S protein that has previously been shown to be important for interaction with the ACE receptor in humans. The putative risk allele (T) arose in Germany and was brought to US by 10 infected non-symptomatic patients. The frequency of the risk allele is 25% in the German strains. In a sample of 7500 New Yorkers, 800 have the TT allele, 3000 have AA and the rest are heterozygous (AT).

a. (10 points) Is this variant neutrally evolving? Suggest two plausible reasons that could lead to deviation from neutrality.

b. (10 points) Assuming the variant is under selection, estimate the selective advantage of the risk genotype over the heterozygotes.

QUESTION 2. Host Susceptibility Analysis (25 points)
The response of coronavirus has significantly varied across human populations. This has raised questions about the impact of host susceptibility impacting the response to the infection. To answer this question, researchers from around the world ran a genome-wide association for 917 cases and controls from the UK Biobank study containing 1 million markers across the genome. The output of this analysis is shown below as a "Manhattan plot" where the Y-axis = log-scaled *P*-values, a measurement of statistical significance and numbers on the X-axis represents the chromosomes (1-22 = autosomes, 23 = chromosome X). Each point is the output of one variant tested.



a. (1 points) What do the peaks represent?

b. (2 points) How many peaks reach genome-wide significance?

c. (2 points) What can you conclude from this analysis?

You decide to investigate the hypothesis yourself and contact researchers at UCSF who have collected data for 750 cases and 1000 controls. Because it is expensive to genotype a large number of markers, you decide to only investigate 2 markers located close to the ACE gene in humans. You obtain the following counts for your analysis.

Given the following data, test for association between disease and SNP.

SNP 1	AA	AG	GG
COVID-19 positive	337	225	188
COVID-19 negative	375	337	288

SNP 2	CC	CG	GG
COVID-19 positive	132	243	375
COVID-19 negative	250	363	387

d. (15 points) What are the χ^2 and p-values for each SNP? Do either of the SNPs significantly increase the risk of COVID-19 infection? For full credit, show all calculations.

e. (5 points) Estimate the odds ratio for each of the SNPs.

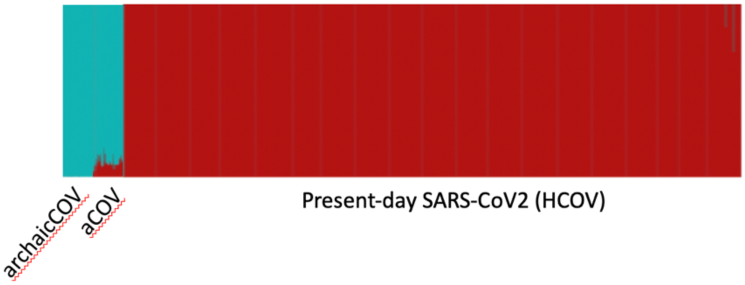
QUESTION 3. Ancient Pathogen Research (20 points)

In looking through the ancient DNA literature, you find a sequence of an ancient bat genome. The study highlights that researchers were able to extract DNA from this sample. In addition to bat DNA, the researchers also found an ancient sequence of coronavirus (aCOV).

a. (5 points) Describe two criteria that you can use to ensure that the ancient coronavirus sequence is authentic (and not modern contamination)?

b. (6 points) You compare the ancient aCOV sequence to modern day strains of SARS-COV2, MERS-COV2 and SARS-CoV1 and estimate the divergence is 0.1%, 0.15% and 0.2% respectively. The haploid genome size of coronavirus is 30,000 nucleotides. Assume the mutation rate in coronaviruses is $2.3x10^{-6}$ per base pair per generation. The mean generation time is 5 days. What is the divergence time of aCOV from the three present-day strains in generations and years?

c. (9 points) From your analysis, you infer that the aCOV is a mixed strain which has ancestry from an archaic extinct lineage of coronavirus and present-day HCOV sequence (see ADMIXTURE plot below). You infer the % of HCOV ancestry is ~10%.



(i) (4 points) Assume two source populations (archaicCOV and HCOV) have allele frequencies of allele A of 0.2 and 0.6 respectively. What is the expected heterozygosity in aCOV tracing 10% of its ancestry to HCOV and 90% of the ancestry to the archaic lineage at that locus?

(ii) (5 points) As you know, mixture between distinct groups creates mosaic chromosomes with ancestry segments from both ancestors, which become progressively smaller as recombination occurs after the gene flow. Assume that on average there is one crossover per generation, and that the length of the genome is ~30,000. What is the typical length of a surviving archaic segment after ~ 2,000 generations?

QUESTION 4. UCSC Browser (15 points)

Recently, a Neandertal haplotype has been associated with protection against severe COVID-19. This haplotype contains parts or all of the three genes OAS1, OAS2, and OAS3, which encode oligoadenylate synthetases. These enzymes activate ribonuclease L, an enzyme that degrades intracellular double-stranded RNA and activates other antiviral mechanisms in cells infected by viruses. Let’s focus in on the gene OAS1 for now.

a) **(3 points)** What is the genomic location of OAS1?

b) **(2 points)** How many isoforms are shown on the Gencode V36 track? What is the Gencode transcript name (hint: starts with ENST) of the isoform that has the fewest exons?

c) (2 points) One variant (rs10774671) in the Neanderthal haplotype has been described as affecting a splice acceptor site in OAS1. The derived allele at this site, which is the most frequent allele in present-day humans, alters splicing of OAS1 transcript such that several protein isoforms are produced that have lower enzymatic activity compared to the ancestral isoform which is preserved in Neandertals. The reference allele presented in the dpSNP build 151 track for this site is the ancestral allele retained in Neanderthals. What is the frequency of this allele in the 1000Genomes dataset?

d) (3 points) In addition to the splice acceptor site, the Neandertal haplotype contains another variant (rs2660) in OAS1. This variant (or variants in LD with this variant) has been shown to be associated with moderate to strong protection against SARS-CoV.

What are the two alleles have been observed among sequenced individuals at this site? What is the frequency of the ancestral allele for the populations included in the dbSNP build 151 database?

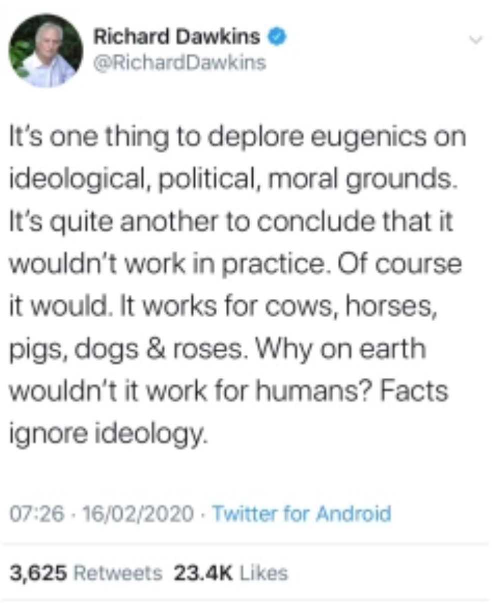
e) (1 point) What type of non-synonymous coding change does this variant (rs2660) result in?

f) (2 points) Let's also explore the SARS-CoV-2 information available in UCSC browser. How large is the SARS-CoV-2 genome?

g) (2 points) From the variant of concern track associated with the SARS-CoV-2 genome, how many variants have been listed for the B.1.490 VOI "California variant"?

QUESTION 5. Ethics and Genetics (20 points)

a. (10 points) Richard Dawkins, an evolutionary biologist, recently suggested that “Eugenics should work in humans”. Discuss some specifics that either support or disagree with this statement. Provide at least two arguments to support your answer.



Richard Dawkins
@RichardDawkins

It’s one thing to deplore eugenics on ideological, political, moral grounds. It’s quite another to conclude that it wouldn’t work in practice. Of course it would. It works for cows, horses, pigs, dogs & roses. Why on earth wouldn’t it work for humans? Facts ignore ideology.

07:26 · 16/02/2020 · Twitter for Android

3,625 Retweets 23.4K Likes

b. (10 points) There are many new technologies that raise concern for eugenics in future. Discuss two technologies that you are concerned about and describe some best practices for future, weighting the ethical and scientific considerations for your recommendations.