Exercise 2: Interpreting a human genome

Hello! You are part of a CSI team who tries to determine as much as possible genetic information on the genome at hand.

Prerequisite:

1. Download the genome file from your group table
2. Note that I use the word loci and SNPs interchangeably.

Technology (18 points)

1. What technology has been used to produce this genome (note: we are asking about a technology not a company)?

2. Please count the number of loci per chromosome and produce a bar graph

3. How many loci does the file have?

4. What's the cost per locus (SNP) if it costs the customer $39 to get this file?

5. Find the lengths of each chromosome of the human genome (hint: use google).
    a. Draw a scatter plot with X-axis as the length of each chromosome and the y-axis as the number of SNPs in the file.
    b. What's the r^2 between chromosome lengths and number of SNPs?
    c. Are there over represented or under represented chromosomes?

6. Draw the interval distribution of the gap between each two loci in 1Kb (1000nt) and determine the mean and median.

Data (12 points)

7. What's the percentage of "no call" loci?

8. Only for the autosome, create a 6x6 matrix with columns as "A", "C", "G" and "T", "I", "D" and rows along the same values. For each called genotype, place the call in the table based on the first and the second allele. For example: "AA" will go to the first row and first column. "AC" will go to the second column and first row, and so on. Report the distribution of the table. What's the most prevalent genotype?

9. Looking only on the autosome, what's the percentage of loci with homozygous SNPs?

10. Is this person a male or a female? Explain why.

Health (20 points)

11. ApoE is a gene that has three types of alleles: ApoE2, ApoE3, and ApoE4. Read the SNPedia page on ApoE.
    a. Create a function that takes the genome file of a person and returns the ApoE status.
    b. For the person of interest, what are the two ApoE alleles?
    c. What can you tell this person about her/his risk to Alzheimer?

12. Delta F508 is the most prevalent mutation in Cystic Fibrosis, a recessive disease. For the sake of the exercise, let's assume it is the only mutation that causes Cystic Fibrosis. Read the SNPedia page on Cystic Fibrosis.
    a. Create a function that takes the genome file of a person and returns the Delta F508.
    b. For the person of interest, what's her/his Cystic Fibrosis status?
    c. This person considers marrying someone who is a carrier. What do you recommend for this person?

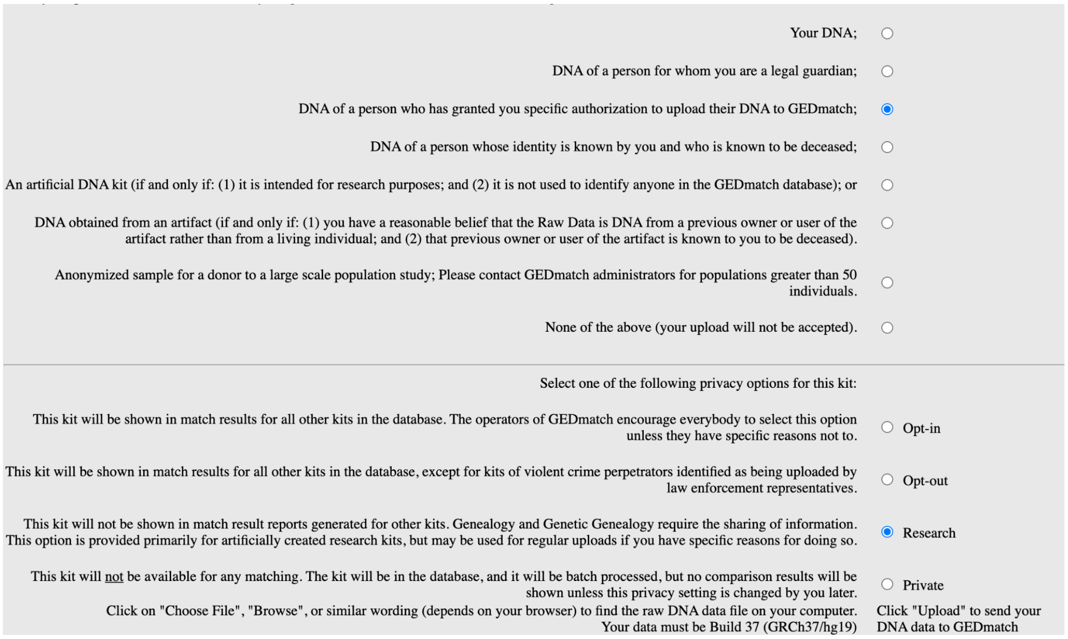
Scale up (25 points)

13. Go to OpenSNP.org and download the *latest* 100 genomes from 23andMe. For each genome, call the ApoE status and Delta F508. Report the results in an Excel table with openSNP id, ApoE status, and Delta F508.

14. Report table with the frequency of the three ApoE alleles in the population and the delta F508.

GEDmatch (25 points)

15. Upload the original genome to GEDmatch. Use the following setting and report back the kid ID.



Your DNA; ○

DNA of a person for whom you are a legal guardian; ○

DNA of a person who has granted you specific authorization to upload their DNA to GEDmatch; ●

DNA of a person whose identity is known by you and who is known to be deceased; ○

An artificial DNA kit (if and only if: (1) it is intended for research purposes; and (2) it is not used to identify anyone in the GEDmatch database); or ○

DNA obtained from an artifact (if and only if: (1) you have a reasonable belief that the Raw Data is DNA from a previous owner or user of the artifact rather than from a living individual; and (2) that previous owner or user of the artifact is known to you to be deceased). ○

Anonymized sample for a donor to a large scale population study; Please contact GEDmatch administrators for populations greater than 50 individuals. ○

None of the above (your upload will not be accepted). ○

Select one of the following privacy options for this kit:

This kit will be shown in match results for all other kits in the database. The operators of GEDmatch encourage everybody to select this option unless they have specific reasons not to. ○ Opt-in

This kit will be shown in match results for all other kits in the database, except for kits of violent crime perpetrators identified as being uploaded by law enforcement representatives. ○ Opt-out

This kit will not be shown in match result reports generated for other kits. Genealogy and Genetic Genealogy require the sharing of information. This option is provided primarily for artificially created research kits, but may be used for regular uploads if you have specific reasons for doing so. ● Research

This kit will not be available for any matching. The kit will be in the database, and it will be batch processed, but no comparison results will be shown unless this privacy setting is changed by you later. ○ Private

Click on "Choose File", "Browse", or similar wording (depends on your browser) to find the raw DNA data file on your computer. Your data must be Build 37 (GRCh37/hg19)

Click "Upload" to send your DNA data to GEDmatch

16. Use the one-to-many tool, what's the closest match to the person? How many cM do they share? What do you think is their familial relationship?

17. Explore GEDmatch a bit. Are the parents of this person cousins?

18. What do you think is the ethnic background of this person based on matches?

Bonus (10pt):

1. Consider that the sensitivity of calling correctly a rare disease variant is 100% and the specificity is 99.5%. For example, consider that the disease variant is “A” and the healthy variant is “C”. Now, for 1000 people who are “CC”, the 995 will be reported as “CC” and 5 as “AC”.
    a. Based on internet searches find the allele frequency of rs386833395 for Europeans
    b. What disease is associated with this variant
    c. You find that a person is a carrier for this variant. What are the chances that the person is truly positive?
    d. Suggest at least one method to confirm the status of the variant
    e. Will you use Promethase after knowing the answer for this question?