**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 here: <https://www.dropbox.com/s/atdw13kd8nlexe3/genome.txt?dl=0>
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).
   1. 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.
   2. What’s the r^2 between chromosome lengths and number of SNPs?
   3. 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)**

1. What’s the percentage of “no call” loci?
2. 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?
3. Looking only on the autosome, what’s the percentage of loci with homozygous SNPs?
4. Is this person a male or a female? Explain why.

**Health (20 points)**

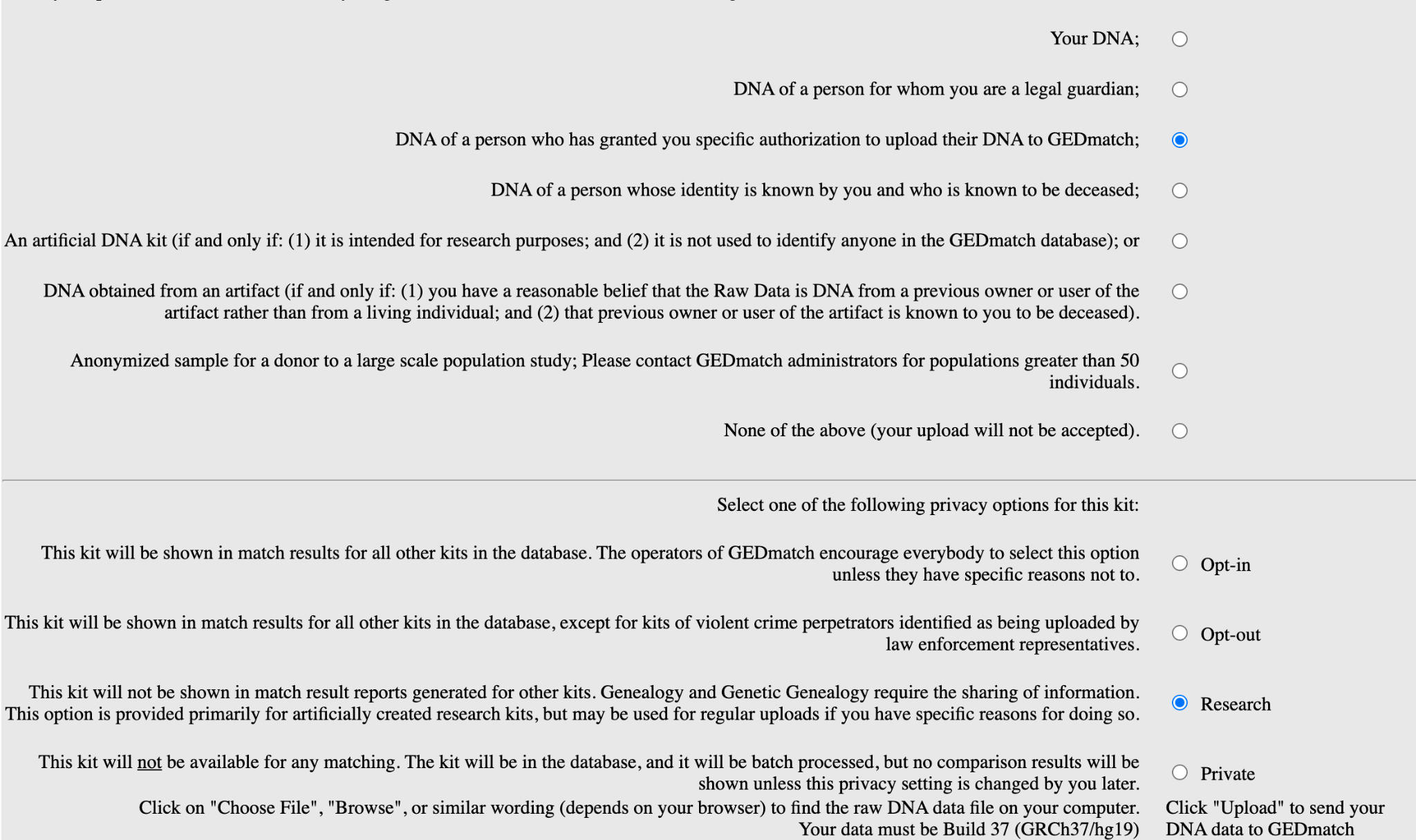
1. ApoE is a gene that has three types of alleles: ApoE2, ApoE3, and ApoE4. Read the SNPedia page on [ApoE](https://www.snpedia.com/index.php/APOE).
   1. Create a function that takes the genome file of a person and returns the ApoE status.
   2. For the person of interest, what are the two ApoE alleles?
   3. What can you tell this person about her/his risk to Alzheimer?
2. 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](https://www.snpedia.com/index.php/Cystic_Fibrosis).
   1. Create a function that takes the genome file of a person and returns the Delta F508.
   2. For the person of interest, what’s her/his Cystic Fibrosis status?
   3. This person considers marrying someone who is a carrier. What do you recommend for this person?

**Scale up (25 points)**

1. 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.
2. Report table with the frequency of the three ApoE alleles in the population and the delta F508.

**GEDmatch (25 points)**

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



1. 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?
2. Explore GEDmatch a bit. Are the parents of this person cousins?
3. 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”.
   1. Based on internet searches find the allele frequency of rs386833395 for Europeans
   2. What disease is associated with this variant
   3. You find that a person is a carrier for this variant. What are the chances that the person is truly positive?
   4. Suggest at least one method to confirm the status of the variant
   5. Will you use Promethase after knowing the answer for this question?