PART 1

Please implement the solution to this problem in Perl, Python, or Java.

What you need to submit:

- The scripts/program implementing the solution to this problem with instructions for running your scripts/program.
- Outputs from your scripts/program.
- Test cases for your solutions.

Problem Description:

Given a telephone directory associating names of people and their telephone numbers, output the names of people who share the same phone number.

Expected input format is a 2 column tab-delimited file. In each row, it is expected there exists a:

- Person's name
- 10-digit US phone number, with or without parentheses, hyphens, or dots, and optional 3 digit area code.

Example telephone directory:

Anna Banana	212-867-5309
Bob	212-867-5309
Charlie D	(333)555-1234

Here we can see that Anna Banana and Bob share the same phone number. Charlie D shares the same phone with himself.

Output format should be 2 tab-delimited columns. In each row:

- Column 1: Shared telephone number
- Column 2: Comma-delimited list of people who share the phone number in Column 1.

PART 2

Given gene ADA (RefSeq NM_000022.2), forward translate the DNA substitution mutation c.239A>G to the expected amino acid change and codon position.

c.239A>G denotes that at nucleotide 239 an A is changed to a G.

Nucleotide 1 is the A of the ATG-translation initiation codon.

If you would like to code the solution to this problem, you may, but it is not necessary. Please show your work.

Step 1: Find the mRNA in the GenBank at NCBI at the following website:

https://www.ncbi.nlm.nih.gov/nuccore/NM_000022.2/

Step 2: Save the FASTA file of the sequence of NM_000022.2, the mRNA of Homo sapiens adenosine deaminase (ADA) https://www.ncbi.nlm.nih.gov/nuccore/NM_000022.2/

Step 3: Find out the CDS of NM_000022.2 (129..1220 on the same website of step 1), the start codon of the gene is at 129, which is the ATG-translation initiation codon.

Step 4: Find the c.239 of the whole mRNA. Start from position 129 of gene as the first initial codon, and count 239 base toward the 3'end. The position is 239+129-1= 367

Step 5: Mutate the base c.239A>G at position 367 in the seq NM_000022.2.

Step 6: Find out the codon changes after mutation, c.239A>G. Because 239 mod 3 is 2, meaning the second letter of 3-letter combination of the DNA coding units has been changed by the mutation at c.239A>G. By inspecting mRNA sequence, it should be a codon AAA mutated into AGA.

Step 7: Find out the corresponding change of amino acid, based on the codon changes from step 6. Since the c239A>G will mutate AAA into AGA, by looking up into a DNA codon table, we could conclude a Lysine(K) has been mutated into an Arginine(R) at position math. Floor(239/3) +1 = 80 of protein sequence.

Step 8: Save the protein sequence of ADA. Accordingly, get the protein sequence from the website below: https://www.ncbi.nlm.nih.gov/protein/NP_000013.2?report=fasta

Step 9: Get the mutated protein sequence by changing the amino acid. For ADA protein sequence, change the Lysine(K) at position 80 of protein sequence into the Arginine(R).

PART 3

What information related to a variant in a person's genome would you want to know in order to assess if it is indicative of disease and why? Please name a possible source for each type of information.

Information	**Possible source**
The mapping of variant gene on the chromosome, for example: chr11:13039503-13025677	NCBI nucleotide blast, Ensemble blast, GATK tool
The types of mutations for each match on the chromosome, for example: single nucleotide polymorphism, insertion, deletion, substitutions of multiple nucleotides, gene inversion.	NCBI nblast, Ensemble blast, GATK tool
The changes of copy number of the gene	DNAseq and DNA fingerprinting

The name and the particular component of gene mutated by the variant, for example, the coding region or noncoding region	NCBI nucleotide blast, Ensemble blast, NCBI Gene or Transcripts search
If the coding region is affected by the mutation (the variant impacts the exon): a. The protein sequence of the match gene b. The changes of codon, for example: missense mutation, non sense mutation, silent mutation, frame shift	DNA codon table, NCBI Protein blast, Ensemble blast and NCBI Protein search
If the coding region is not affected by the mutation: 5a. The intron and the splicing impacted by the variant 5b. The regulator influenced by the variant 5c. The downstream gene variation caused by the variant	NCBI blast, Ensemble blast and NCBI Gene search, NCBI transcripts search. The tool of Variant Effect Predictor on Ensemble
If the variant mapped two distinct gene located at distinct region of genome: The name and the particular component of each gene	NCBI nucleotide blast, Ensemble blast, NCBI Gene or Transcripts search
The pathogenic variants at the same region of chromosome. Search the source database on the to find any similar variation which is pathogenic	Variation Viewer, dbSNP, dbVar, and ClinVar
The variant alleles by the variant id.	NCBI variation viewer and Ensemble variant effect predictor, NCBI dbVar, GATK tool.
The downstream gene affected by this variant.	The tool of Variant Effect Predictor on Ensemble, NCBI ClinVar.
Predict potential pathogenic effect of the variants.	The tool of Variant Effect Predictor on Ensemble