

# Assignments - Level I

Attempt to complete the following questions. Section-I is mandatory. Solutions for Section-II will be counted as bonus. Choose a programming language of your choice. **All code should be checked into github. Send the completed Excel and link to code with instructions for running to [ramani@indna.in](mailto:ramani@indna.in)**

## SECTION - I

1. Find attached a file "`mutect_immediate.vcf`". It has the following format.

```
Line 1 ## Comment
Line 2 ## Comment
Line 3 ## ...
Line 4 ## Comment
Line 5 ## Comment
Line 6 #CHROM  POS      ID      REF      ALT      QUAL      FILTER  INFO
FORMAT SAMPLE
Line 7 chr1    115258741      .      A      AA      45      PASS
DP=35404;TI=NM_002524;GI=NRAS;FC=Frameshift      GT:GQ:AD:VF
0/1:45:34968,436:0.0123
Line 8 chr2    209113192      rs11554137      G      A      1000
PASS      DP=93129;TI=NM_005896;GI=IDH1;FC=Synonymous_G105G
GT:GQ:AD:VF      0/1:1000:50037,43060:0.4624
Line 9 ...
```

Each row that does **not** start with a '#' in the "`mutect_immediate.vcf`" file corresponds to a "**mutation**" in the human genome. Its absolute position on the human DNA strand is given under the column "pos" and the chromosome on which it lies is given under the column "chrom".

Your job is to find out if the "mutation" is present in a given "**panel**". A panel is a set of regions on the DNA strand which are described by rows of the following form : [ start position, end position, and the chromosome-number on which the region lies ]. These "**regions**" are provided to you in the "`truseq.bed`" file.

Write a program to check if every mutation [ row that does **not** start with a '#' ] from the "`mutect_immediate.vcf`" file **falls in the given regions of the panel**. You can do that by comparing the mutation's position listed under 'pos' with the 'start' and the 'end' positions of the regions from the bed file. If they fall in between the start and end positions of **at least one** region, then consider that mutation to be an

**"accepted mutation".** Write these "accepted" mutations/rows in a new **Excel** file **"accepted.xlsx/accepted.xls"**.

**Note:** Rows in **"mutect\_immediate.vcf"** file are tab separated.

Example of bed file:

```
chr1 43815005 43815137
chr1 115256525 115256653
chr2 209113002 209113413
chr10 43609073 43609201
chr10 43609929 43610049
```

For the above input, output **"accepted.xls"** would look like the following:

CHROM	POS	ID	R E F	A L T	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr2	209113192	rs11554137	G	A	1000	PASS	DP=93 129;TI =NM_0 05896; GI=IDH 1;FC=S ynony mous_ G105G	GT:GQ:AD: VF	0/1:1000:5 0037,4306 0:0.4624
...	...	...	...	...	...	...	...	...	...

**How:** 209113192 falls in between 209113002 and 209113413 and both mutations have the same chromosome, **"chr2"**

## SECTION - II

2. Install the latest version of MySQL Server on your machine. Test if you're able to create a database, a table and are able to perform basic table related operations: INSERT, UPDATE, DELETE ROWS, etc.
3.
  - a. Create a table in MySQL database called **"my\_vcf"** with the below-mentioned columns. **Refer the "mutect\_immediate.vcf" to decide the data-types of these columns.**
    - i. Chr
    - ii. Pos
    - iii. Id

- iv. Ref
  - v. Alt
  - vi. Qual
  - vii. Filter
  - viii. Info
  - ix. Sample
- b. Write a program to insert the values from mutations [ rows ***not*** beginning with a '#' ] from "**mutect\_immediate.vcf**" into table "**my\_vcf**". [ You will have to establish connectivity between MySQL and the programming language of your choice ].
- c. Fire a query on "my\_vcf" in the same program to fetch mutations with quality > 50 and < 500 and write these mutations to a **.txt** file. [ 'quality' corresponds to the column Qual ].