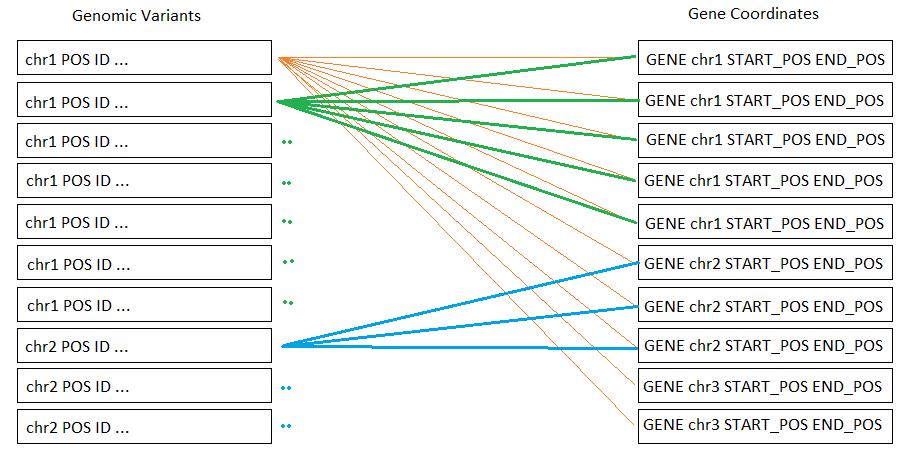
**Junior Software Developer - QMUL31542 - Ibrahim Bashir**

**Question 1**

The code for this question is in test.py.

To find which genomic variants overlapped with which genes, I knew I had to loop through every row in the genomic variants file and compare it with each of the genes in the gene coordinates file to match the CHROM and check if the POS is in the correct range.

To optimise for speed, I used a dictionary so that it would only loop through the entire gene coordinates file once for the first row of the genomic variants. For subsequent genomic variants, it could lookup the matching CHROMs immediately and then only loop through those to check if the POS is in the correct range. I’ve visualised this in the diagram below.



**Question 2**

SELECT v.\*, gc.gene

FROM variants AS v LEFT JOIN gene\_coordinates AS gc

ON v.chrom = gc.chrom

AND v.position >= gc.start\_pos AND v.position <= gc.end\_pos;

**Question 3**

I think SQL is faster because you don’t have to spend time extracting data from the files and evaluating its structure before comparing them. I think Python would be better suited to a much larger dataset or a problem where the comparison is more complex than just comparing chromosomes and positions.

If the number of variants was 10 million, you could:

* use a better CPU/GPU for faster computations.
* split the variants dataset and use multiprocessing to run the computations in parallel.
* split the variants dataset and use multiple computers to run the computations in parallel.