Ismael Henarejos Castillo
ihc.ca94@gmail.com

To whom it may concern;

For the skill test nº2 (https://github.com/CERC-Genomic-Medicine/skills_test_2), I wrote a tool in python 3.10, which makes use of multiprocessing Process class to paralelize the search on several cores at the same time (each core works on different vcf files). Python 3.10 is required for specific lines where the bisect module was used. All modules are included with basic python installation and there are no special versions used. High memory usage is avoid with the use of yield when anlyzing variants inside functions. My computer specs are shown in **the figure below**. Details on how the algorithm works are detailed in the comments inside the source code. The general use should be:

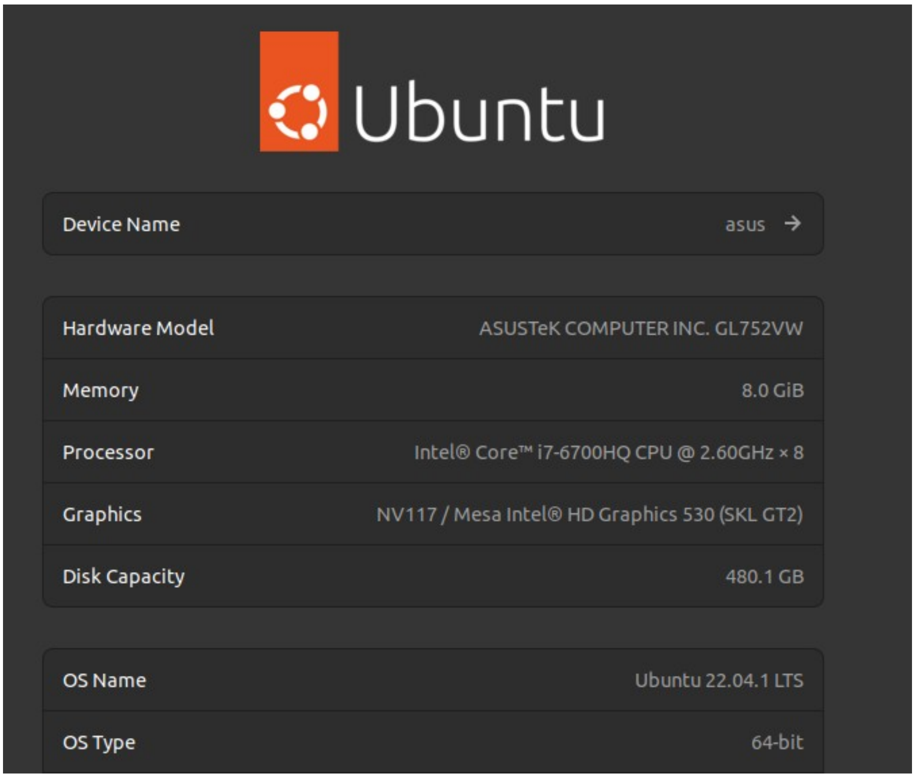
(time) python3 circa_skill_test_ihc.py –vcf (full route to directory) –gtf (full route to gtf) --pb (+- base pairs) --num_process (number of cores to use)

In the shell terminal, help can be used to get info on the parameters. Gtf file was previously simplified using the following line with awk in the shell terminal:
zcat gencode.v38.annotation.gtf.gz | awk -v OFS='\t' '{if($3=="gene"){print $1, $4, $5, $10}}' > gtf_simple.tsv.

Annotated vcf.gz will be created in the same directory along with the original input files. TIme employed in my computer for vcf files with 7 cores was: (real) 48m24.880s

I am open to any criticism as I am still a junior in this field. Thank you for your time and consideration.

Ismael



Device Name	asus →
Hardware Model	ASUSTeK COMPUTER INC. GL752VW
Memory	8.0 GiB
Processor	Intel® Core™ i7-6700HQ CPU @ 2.60GHz × 8
Graphics	NV117 / Mesa Intel® HD Graphics 530 (SKL GT2)
Disk Capacity	480.1 GB
OS Name	Ubuntu 22.04.1 LTS
OS Type	64-bit