Methods

Pre-requirements

To ensure the successful execution of Task 1, the necessary software components were updated and downloaded. Docker was updated to the latest version to ensure compatibility and performance. The GATK Docker container was installed using the following command:

docker pull broadinstitute/gatk

ANNOVAR, a software tool used for variant annotation, was installed according to the guidelines provided in the official documentation.

Additionally, a working directory was set up to organize all relevant files and outputs.

Documentation Review

The documentation for both GATK and ANNOVAR was reviewed to understand the required commands and options for variant selection and annotation.

Task 1a: Variant Selection

The VCF file `Test\_annotate.vcf` was used as the initial input. To filter the SNPs from the VCF file, the `SelectVariants` tool from GATK was utilized. The hg19 reference genome was sourced from Google Cloud Storage as provided by the documentation. The Docker container was then run using the following command:

docker run -v /Users/hm/codeProjects/BioTask/Task1:/mnt/Task1 -it broadinstitute/gatk

The `-v` option was important to load a script and the VCF file into the container to run GATK.

The `SelectVariants` was then executed. This command filtered the SNP variants and produced an output file named `Only\_SNPs.vcf`.

Task 1b: Variant Annotation

Additional Databases

The necessary databases were downloaded using ANNOVAR:

annotate\_variation.pl -buildver hg19 -downdb -webfrom annovar refGene humandb/

annotate\_variation.pl -buildver hg19 -downdb cytoBand humandb/

Running Annovar

A bash script named `vcf\_to\_anno.sh` was written to automate the annotation process. The `convert2annovar.pl` tool was used to convert the VCF file:

convert2annovar.pl -format vcf4 "$vcfFile" > "$avFile"

The `table\_annovar.pl` tool was then used to annotate the variants:

table\_annovar.pl "$avFile" humandb/ -buildver hg19 -out "${outPath}Filtered\_SNPs" -remove -protocol refGene,cytoBand -operation g,r -arg '-splicing 12' -nastring . -csvout

This produced an annotated output file named `Filtered\_SNPs.hg19\_multianno.csv`.

Data Analysis

To analyze the annotated data, a Python script was employed to read the CSV file and extract specific variant information. The script counted the number of unique variants, splicing variants, and start-loss variants using the following command:

python3 /Users/hm/codeProjects/BioTask/Task1/analyzeSNP.py

This analysis provided the required counts and detailed insights into the variant data.

Results:

Number of unique variants: 5746

Number of splicing variants: 221

Number of startloss variants: 4

Time for completion:

Task 1 = ~4 hours

Task 2 = ~1 hour