Technical Report: Variant Annotation Using SnpEff

The goal of the "Annotate" phase of the project was to utilize SnpEff, a tool for annotating genetic variants from Variant Call Format (VCF) files. Annotation is crucial for understanding the functional significance of genetic variants, which is essential for interpreting genomic data effectively.

The first step involved downloading SnpEff from the official source. This tool is specifically designed for annotating variations in genomic datasets and is available at SnpEff Official Site.

SnpEff runs on Java, so ensuring that a compatible Java version was installed was necessary. Initially, a 64-bit version of Java was required, but a memory error was encountered due to the incompatibility of the Java version with the operating system.

To resolve this, the Java Development Kit (JDK) was updated from JDK 8 to a version above JDK 11, which SnpEff specifically requires for optimal performance.

After setting up Java, the next step was to run SnpEff using a command in the terminal. This step involved integrating Java commands within a Python script, requiring research on executing Java commands from Python.

A Python script was written to automate the running of SnpEff. The script included the necessary commands to execute the annotation process on a test VCF file.

Initially, the output of the annotation was not saved correctly. Adjustments were made to the script to ensure the annotated results were written out to a file. Furthermore, it was critical to use a test file for the initial runs to prevent data corruption or loss in the main dataset.

The annotation of the VCF file was successfully completed after overcoming the initial technical challenges. The output included a detailed annotation of the genomic variants, providing insights into their potential impacts. This output serves as a foundational element for subsequent data transformation and analysis stages.

The ability to annotate genomic data effectively using SnpEff was demonstrated in this phase of the project. The challenges faced were primarily related to software compatibility and output management, which were addressed through updates and script modifications. The successful completion of this stage highlights the importance of proper tool configuration and the potential of genomic data to provide significant insights into genetic research.