Workflow Report

For the execution of the task provided, I followed the steps outlined below:

Data Retrieval:

Initially, I downloaded 560 FASTQ files (comprising 280 samples related to lymphoma disease) from the SRA database using the SRA toolkit.

Data Merging:

Utilizing the Merge.py script, I merged the downloaded FASTQ files to ensure maximum and minimum overlap, resulting in a single merged FASTQ file. Subsequently, I converted the merged FASTQ file to a FASTA file.

Database Construction:

Next, using the database.py script, I constructed a database where unique sequences contained within specific fixed sequences were extracted and stored in a SQLite database. During this process, a 1% threshold for variation between sequences was considered to determine sequence similarity.

Sequence Comparison and Database Update:

In the query.py script, the objective was to allow users to identify potentially new sequences confined within the fixed sequences. If these sequences were non-repetitive and novel, they were added to the database, and the database was updated accordingly. Additionally, the outputs of this script included a CSV file containing the updated database and a FASTA file containing the new sequences. Furthermore, backups of the database were created before and after the update process.

Testing:

Testing was conducted using two downsized FASTQ files, ensuring repeatability. Seven new sequences were identified and added to the database during this test. The relevant files for this test are available.

Subsequent Testing:

For subsequent testing, the same files can be provided to the script, with the expectation of no new sequences as the database has been previously updated.

Limitations:

It is important to note that alternative approaches such as read assembly to achieve larger contigs and conversion of FASTQ files to SAM and BAM files, followed by sequence searching, were initially considered. However, due to computational constraints and limited access to computing resources, these methods were not pursued.

This workflow encapsulates the steps undertaken to accomplish the task, with the provided details ensuring clarity and completeness.