Bioinformatics programming assignment

Write a NGS read simulator that randomly picks out reads from a genome and outputs them as a fastq file (use dummy quality values). (Use read length of 50 bp and generate 100,000 reads from the human genome)

Add a uniform error rate of 0.01 (1% of the time a base is randomly replaced with another base) to the fastq file.

Align the resulting fastq file with bwa and find the error rate (read aligned to a part of the genome other than where it originated from).

Validation of Introduced error rate from '*stats.csv*' file generated:

Total number of reads : 100000bp
Read length : 50bp
Total errors introduced : 49954

$$\boldsymbol{Sequencing\ Error\ Rate} = \frac{\boldsymbol{Total\ errors\ introduced}}{\boldsymbol{Total\ no.of\ reads} \times \boldsymbol{Read\ length}}$$

$$\boldsymbol{Sequencing\ Error\ Rate} = \frac{\mathbf{49954}}{\mathbf{100000} \times \mathbf{50}} \times \mathbf{100} \quad = \quad \mathbf{0.99908}$$

Which is approximately 1%

Error Rate due to mapping of reads on genome positions other than where it originated from:

Results from '*calc_error.py*' script:

Reads mapped position other than its origin : 13653
Reads mapped to genome : 99192
Reads unmapped : 808

Results from '*samtools idxstats*' command:

Reads mapped to genome : 99192
Reads unmapped : 808

$$\boldsymbol{Error\ Rate} = \frac{\boldsymbol{Reads\ Mapped\ to\ position\ other\ than\ its\ origin}}{\boldsymbol{Total\ no.of\ mapped\ reads}} \times \mathbf{100}$$

$$\boldsymbol{Error\ Rate} = \frac{\mathbf{13653}}{\mathbf{99192}} \times \mathbf{100} \ = \ \mathbf{13.76}$$

Which is approximately 14%

Note : all analysis files are given in '*Analysis Files*' folder.