NGS Assignment

# Genome Data Set-Homo sapiens chromosome 22, GRCh38.p12 Primary Assembly

Sanger format can encode a Phred quality score from 0 to 93 using ASCII 33 to 126 (although in raw read data the Phred quality score rarely exceeds 60, higher scores are possible in assemblies or read maps)

readgenerator.py randomly picks out reads from a genome and outputs them as a fastq file(using dummy quality values).

The read length of 50 bp was set as default.

It will generate 100,000 reads from the human genome.

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

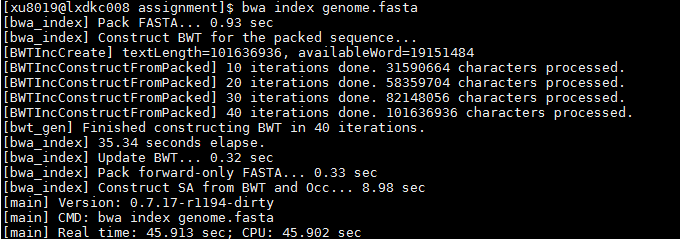
# Run the python script with command :-

readgenerator.py genome.fasta out.fastq

# Aligning the resulting fastq file with bwa. Commands :-

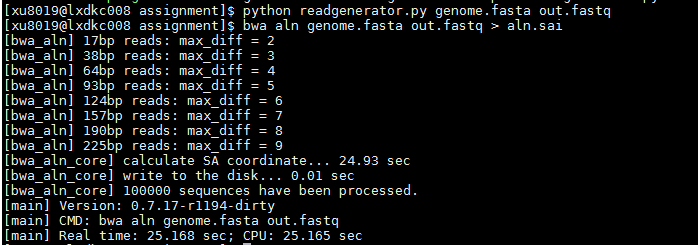
1. Index database sequences in the FASTA format

bwa index genome.fasta



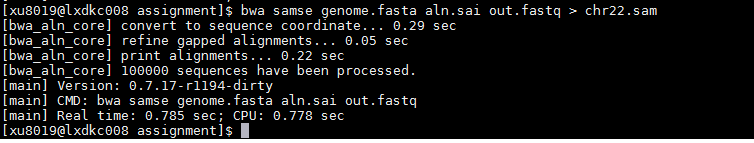
1. gapped/ungapped alignment

bwa aln genome.fasta out.fastq > aln.sai



1. Generate alignments in the SAM format given single-end reads. Repetitive hits will be randomly chosen.

bwa samse genome.fasta aln.sai out.fastq > chr22.sam



Calculating the error rate:-

Error means read aligned to a part of the genome other than where it originated from.

Command:-

python calculaterror.py chr21.sam

