Software Design Documentation

Project Name: Fulgent Genetics at home technical test

Date: 6/30/2024

Written By: Khoi Huynh

Introduction:

This script is used to download, align the fastq for NA12878 to hg38 human reference genome. The script is also used to calculate per base coverage for the region of interest of CD4 gene

System Overview:

The script is written to be used on Linux. It can also be adapted to be submitted by SLURM or HPC job manager.

Design Considerations:

Dependencies: samtools, BWA, bedtools, Picard

Strategies:

1. Find a version of NA12878 fastq with highest read quality by reading NA12878.alt_bwamem_GRCh38DH.20150826.CEU.exome.bam.bas. The chosen version is from SRR1518133 study
2. Create hg38 reference genome index by downloading hg38.fa from UCSC genome browser and indexing it with BWA
3. Perform alignment and data quality control using BWA and samtools
4. Generate tab delimited region of interest file for CD4 gene
5. Calculate per base coverage using bedtools coverage with option -d