Variant calling analysis of the human genome.

Background

The objective of this task was to identify genomic variants of the genes on the human chromosomes 12, 15 and 17.

Methodology

Data collection

The reference human genome sequence was given as well as the forward and reverse sequences of our query samples.

Softwares

- Fastp
- Fastqc
- Bwa
- Samtools
- Bcftools

Analysis

First, a directory for the samples was created.

```
mkdir raw_reads
```

In the raw_reads directory, the query sequences were downloaded

```
wget https://zenodo.org/record/2582555/files/SLGFSK-N_231335_r1_chr5_12_17.fastq.gz -O SLGFSK-N_r1_chr5_12_17.fastq.gz
wget https://zenodo.org/record/2582555/files/SLGFSK-N_231335_r2_chr5_12_17.fastq.gz -O SLGFSK-N_r2_chr5_12_17.fastq.gz
wget https://zenodo.org/record/2582555/files/SLGFSK-T_231336_r1_chr5_12_17.fastq.gz -O SLGFSK-T_r1_chr5_12_17.fastq.gz
wget https://zenodo.org/record/2582555/files/SLGFSK-T_231336_r2_chr5_12_17.fastq.g
```

A directory for the reference genome was also created

```
mkdir reference
```

In the reference directory, the reference genome was downloaded

```
wget https://zenodo.org/record/2582555/files/hg19.chr5_12_17.fa.gz > reference
```

Another directory for the quality control reports was created

```
mkdir QC_Reports
```

Fastqc was used to check the quality of our sequence reads

```
fastqc raw_reads/*fastq.gz -o QC_Reports
```

The reads were trimmed of the adapter sequences using Fastp

```
bash trim.sh
```

The query sequences were aligned to the reference genome using the bwa tools. Directories for the results in different file formats were created.

```
mkdir -p results/sam/bam/bcf/vcf
```

Then the reference genome was indexed using the command line:

```
bwa index reference/hg19.chr5_12_17.fa
```

Subsequently, each sequence was aligned to the reference genome with the bwa mem command into a sam file format as follows:

```
bwa mem reference/hg19.chr5_12_17.fa raw_reads/trimmed/SLGFSK-N_r1_chr5_12_17.fastq.gz
raw_reads/trimmed/SLGFSK-N_r2_chr5_12_17.fastq.gz > results/sam/SLGFSK-N.aligned.sam
```

```
bwa mem reference/hg19.chr5_12_17.fa raw_reads/trimmed/SLGFSK-T_r1_chr5_12_17.fastq.gz
raw_reads/trimmed/SLGFSK-T_r2_chr5_12_17.fastq.gz > results/sam/SLGFSK-N.aligned.sam
```

The sam file format was compressed into a bam file using the samtools view command:

```
samtools view -S -b results/sam/SLGFSK-N.aligned.sam > results/bam/SLGFSK-N.aligned.bam
```

```
samtools view -S -b results/sam/SLGFSK-T.aligned.sam > results/bam/SLGFSK-T.aligned.bam
```

The bam files were sorted using the sort command from samtools

```
samtools sort -o results/bam/SLGFSK-N.aligned.sorted.bam results/bam/SLGFSK-N.aligned.bam
```

```
samtools sort -o results/bam/SLGFSK-T.aligned.sorted.bam results/bam/SLGFSK-T.aligned.bam
```

Variant calling was performed using the bcf tools

```
bcftools mpileup -O b -o results/bcf/SLGFSK-N_raw.bcf \-f reference/hg19.chr5_12_17.fa
results/bam/SLGFSK-N.aligned.sorted.bam
```

```
bcftools mpileup -O b -o results/bcf/SLGFSK-T_raw.bcf \-f reference/hg19.chr5_12_17.fa
results/bam/SLGFSK-T.aligned.sorted.bam
```

```
bcftools call --ploidy 1 -m -v -o results/vcf/SLGFSK-N_variants.vcf results/bcf/SLGFSK-
N_raw.bcf
```

```
bcftools call --ploidy 1 -m -v -o results/vcf/SLGFSK-T_variants.vcf results/bcf/SLGFSK-
T_raw.bcf
```

vcf tools was used to filter and report only single nucleotide variants

```
vcfutils.pl varFilter results/vcf/SLGFSK-N_variants.vcf  > results/vcf/SLGFSK-
N_final_variants.vcf
```

```
vcfutils.pl varFilter results/vcf/SLGFSK-T_variants.vcf  > results/vcf/SLGFSK-
T_final_variants.vcf
```

Results and Discussion

After alignment of each query sequence to the reference genome, several variants were found using this command:

```
grep -v "#" results/vcf/SLGFSK-N_final_variants.vcf | wc -l
```

```
grep -v "#" results/vcf/SLGFSK-T_final_variants.vcf | wc -l
```

yielding

70911 variants with the SLGFSK-N sample and 90019 variants for the SLGFSK-T sample.

Visualization of the alignment was performed with `samtools tview`

```
samtools tview results/bam/SLGFSK-N.aligned.sorted.bam reference/hg19.chr5_12_17.fa
```

```
samtools tview results/bam/SLGFSK-T.aligned.sorted.bam reference/hg19.chr5_12_17.fa
```

And this showed the samples mostly matched the reference but with some variations at specific locations.