NGS Variant Calling Pipeline for SRA Accession ERR13985875

This document outlines the complete variant calling pipeline used for processing SRA accession **ERR13985875**. Each section describes the tool, the purpose of the step, and the exact commands (with comments for clarity and reproducibility).

Table of Contents

---

1. Data Acquisition and Conversion

Tools:
- **SRA Toolkit** (`prefetch`, `fastq-dump`)

Purpose:
Download the SRA file and convert it to paired-end FASTQ format.

```
# Download the SRA file for ERR13985875
prefetch ERR13985875 --progress

# Convert the SRA file to paired-end FASTQ files
fastq-dump --split-files ERR13985875/
```

---

2. Quality Control (QC) of Raw Reads

Tool:
- FastQC

Purpose:
Assess the quality of the raw FASTQ files.

```
# Run FastQC on raw FASTQ files to generate quality reports
fastqc ERR13985875_1.fastq ERR13985875_2.fastq
```

---

3. Read Trimming

Tool:
- Trimmomatic

Purpose:
Remove low-quality bases and adapter sequences from the reads.

```
# Create a directory for trimmed reads and switch into it
mkdir -p Trimmed
cd Trimmed

# Run Trimmomatic in paired-end mode with sliding window trimming and minimum
length filtering
trimmomatic PE ../ERR13985875_1.fastq ../ERR13985875_2.fastq \
    ERR13985875_1_paired.fastq ERR13985875_1_unpaired.fastq \
    ERR13985875_2_paired.fastq ERR13985875_2_unpaired.fastq \
    SLIDINGWINDOW:4:20 MINLEN:50

# Run FastQC on the trimmed reads for quality control
fastqc ERR13985875_1_paired.fastq ERR13985875_2_paired.fastq
ERR13985875_1_unpaired.fastq ERR13985875_2_unpaired.fastq

# Return to the parent directory
cd ../
```

---

4. Read Alignment to the Reference Genome

Tools:
- **Bowtie2** and **bowtie2-build**

Purpose:
Align the trimmed paired-end reads to the human reference genome.

```
# Switch to the Trimmed directory
cd Trimmed

# Build index for the reference genome (GRCh38)
bowtie2-build GCF_000001405.40_GRCh38.p14_genomic.fna.gz index

# Align the paired-end reads to the reference genome
bowtie2 --no-unal -p 8 -x ../../Reference_Genome/Human/index \
    -1 ERR13985875_1_paired.fastq -2 ERR13985875_2_paired.fastq -S
alignment.sam

# Move the SAM file to the parent directory for downstream processing
mv alignment.sam ../
cd ../
```

---

5. SAM-to-BAM Conversion, Sorting, and Indexing

Tool:
- **Samtools**

Purpose:
Convert the SAM file to BAM, sort it by coordinates, and create an index.

```
# Convert the SAM file to BAM format
samtools view -Sb -o alignment.bam alignment.sam

# Sort the BAM file by coordinate
samtools sort -O bam -o sorted.bam alignment.bam

# Create an index for the sorted BAM file
samtools index sorted.bam
```

---

6. Adding/Updating Read Groups

Tool:
- GATK (Genome Analysis Toolkit)

Purpose:
Add or update read groups in the sorted BAM file for proper downstream processing.

```
gatk AddOrReplaceReadGroups \
    -I sorted.bam \
    -O sorted_rg.bam \
    --RGID 1 \
    --RGLB lib1 \
    --RGPL ILLUMINA \
    --RGPU unit1 \
    --RGSM SampleName
```

---

7. Marking Duplicates

Tool:
- GATK MarkDuplicates

Purpose:
Identify and mark duplicate reads in the BAM file.

```
gatk MarkDuplicates \
    -I sorted_rg.bam \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -M metrics.txt \
    -O unique_reads.bam
```

---

8. Base Quality Score Recalibration (BQSR)

Tool:
- GATK BaseRecalibrator and **ApplyBQSR**

Purpose:
Generate and apply a recalibration table to adjust base quality scores.

```
# Generate a recalibration table using known variant sites
gatk BaseRecalibrator \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -I unique_reads.bam \
    --known-sites ../Reference_Genome/Human/known_sites.vcf \
    -O recal_data.table
```

```
# Apply the recalibration to adjust base quality scores
gatk ApplyBQSR \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -I unique_reads.bam \
    --bqsr-recal-file recal_data.table \
    -O recalibrated.bam
```

---

9. Variant Calling

Tool:
- GATK HaplotypeCaller

Purpose:
Call variants (SNPs and Indels) from the recalibrated BAM file.

```
gatk HaplotypeCaller \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -I recalibrated.bam \
    -O output.vcf
```

10. Variant Filtering

Tool:
- GATK VariantFiltration

Purpose:
Apply custom filters to flag or remove low-confidence variants.

```
gatk VariantFiltration \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -V output.vcf \
    -O filtered_output.vcf \
    --filter-expression "QD < 2.0" \
    --filter-name "LowQD" \
    --filter-expression "FS > 60.0" \
    --filter-name "HighFS"
```

11 Variant Annotation

Tool:
- GATK VariantAnnotator

Purpose:
Enhance the raw variant calls by adding informative annotations (such as coverage metrics, quality by depth, and mapping quality rank sum) to your VCF file. These annotations support further filtering and prioritization of variants.

```
gatk VariantAnnotator \
    -R ../Reference_Genome/Human/GCF_000001405.40_GRCh38.p14_genomic.fna.gz \
    -V filtered_output.vcf \
    -I recalibrated.bam \
    -O output.annotated.vcf \
    -A Coverage \
    -A QualByDepth \
    -A MappingQualityRankSumTest
```

12. Visualization

Tool:
- **IGV (Integrative Genomics Viewer)**

Purpose:
Load the `sorted.bam` file in IGV to visually inspect alignments and verify variant calls (SNPs and Indels).

This file provides a step-by-step, reproducible pipeline for processing NGS data from raw SRA files to final variant calls. Each section is clearly demarcated with bold and large headings, ensuring clarity and ease of navigation. Enjoy your reproducible workflow!