# Docker Installation & Ewing Sarcoma Dataset Analysis Workflow

## 1. Docker Installation

Docker is a platform that enables the creation and management of lightweight, portable containers. These containers package software and dependencies, ensuring consistent execution across environments.

### 1.1 Installing Docker on Ubuntu

Update the package list:

sudo apt update

Install required packages:

sudo apt install apt-transport-https ca-certificates curl software-properties-common

Add Docker’s official GPG key:

curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg

Set up the stable repository:

echo "deb [arch=$(dpkg --print-architecture) signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] https://download.docker.com/linux/ubuntu $(lsb\_release -cs) stable" | sudo tee /etc/apt/sources.list.d/docker.list > /dev/null

Install Docker Engine:

sudo apt update && sudo apt install docker-ce docker-ce-cli containerd.io

Verify installation:

docker --version

## 2. Using Docker for GATK

The Genome Analysis Toolkit (GATK) is used for variant discovery in high-throughput sequencing data. Using Docker ensures that all dependencies are pre-packaged.

Pull the latest GATK image from Docker Hub:

sudo docker pull broadinstitute/gatk:latest

Run the GATK container with mounted data directory:

sudo docker run -v /home/shobita/EwingSarcoma:/gatk/data -it broadinstitute/gatk:latest

## 3. Ewing Sarcoma Dataset Analysis Workflow

This workflow processes paired-end FASTQ files from an Ewing Sarcoma dataset, performing quality control, trimming, alignment, variant calling, and annotation.

### Step 1: Quality Control

Run FastQC to assess read quality before and after trimming:

fastqc SRR25029932\_1.fastq.gz SRR25029932\_2.fastq.gz -o fastqc\_before\_trim/

fastqc SRR25029932\_1\_trimmed.fastq.gz SRR25029932\_2\_trimmed.fastq.gz -o fastqc\_after\_trim/

### Step 2: Adapter Trimming

Use fastp for adapter trimming and quality filtering:

fastp -i SRR25029932\_1.fastq.gz -I SRR25029932\_2.fastq.gz -o SRR25029932\_1\_trimmed.fastq.gz -O SRR25029932\_2\_trimmed.fastq.gz --detect\_adapter\_for\_pe

### Step 3: Alignment to Reference Genome

Align trimmed reads to the human genome reference (Homo\_sapiens\_assembly38.fasta) using BWA-MEM:

bwa mem Homo\_sapiens\_assembly38.fasta SRR25029932\_1\_trimmed.fastq.gz SRR25029932\_2\_trimmed.fastq.gz > EwingSarcoma.sam

### Step 4: Variant Filtering

Filter variants to retain only those that passed quality checks:

bcftools filter -i '%FILTER="PASS"' hardfiltered\_recalibrated.vcf > hardfiltered\_recalibrated\_PASS.vcf

### Step 5: Variant Annotation

Annotate variants using SnpEff:

java -Xmx4g -jar snpEff.jar -v hg38 hardfiltered\_recalibrated\_PASS.vcf > Final\_annotated.vcf