**Grey – explanation no highlight - code**

This uses Bash’s **extended globbing** to move everything except the Handon directory itself. I had everything on home and I wanted to put it inside a directory, so that I could work hassle free in this project that would generate many new files.

$ shopt -s extglob

$ mv !(Handon) Handon/

I had Trimmomatic and Hisat2 installed, hence got it into my working directory i.e. the home directory in my case.

#### Download the fastq dataset –

Go to SRA Database and select the dataset you want to work with and download its fastq file.

As I am working on WSL1.. so whatever I download on my windows has to be brought to Ubuntu / WSL1.

$ mv /mnt/c/Users/HP/Downloads/SRR32105970.fastq.gz ~/TP53/

$ gunzip SRR32105970.fastq.gz

#### Quality check of the adapters –

$ fastqc SRR32105970.fastq

#### Trimming the adapters –

$ java -jar /home/bidya122/Trimmomatic-0.39/trimmomatic-0.39.jar SE -phred33 SRR32105970.fastq SRR32105970\_trimmed.fastq ILLUMINACLIP:/home/bidya122/Trimmomatic-0.39/adapters/TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36

**Fastqc of trimmed file --**

$ fastqc SRR32105970\_trimmed.fastq

**View html file --**

$ wslview SRR32105970\_trimmed\_fastqc.html

So full human chromosome would not be a choice to align with and very large to work with so I got TP53 sequence:

<https://asia.ensembl.org/Homo_sapiens/Gene/Sequence?g=ENSG00000141510;r=17:7661779-7687546>

put it in WSL1

$ mv /mnt/c/Users/HP/Downloads/Homo\_sapiens\_TP53\_sequence.fa ~/

**Build HISAT2 index --**

* Homo\_sapiens\_TP53\_sequence.fa: The **reference sequence** (TP53 gene)
* TP53\_index: The **prefix name** for the output files
* This command:
* Reads the FASTA file
* Creates 6–8 special .ht2 files that HISAT2 can use to align reads quickly. These files act like a **shortcut map** for fast searching.
* HISAT2 — the "index" is a **pre-processed, searchable version** of your reference FASTA file.

$ hisat2-build Homo\_sapiens\_TP53\_sequence.fa TP53\_index

*Builds 6-8 index files.*

**HISAT2 will align the reads to the reference sequence** that was **previously converted into index files**

$ hisat2 -p 1 -x TP53\_index -U SRR32105970\_trimmed.fastq -S SRR32105970\_TP53.sam

To read a sam file

$ head SRR32105970\_TP53.sam

To search for aligned reads only (not unmapped):

$ grep -v '^@' SRR32105970\_TP53.sam | awk '$3 != "\*"' | head

SAM files are large and uncompressed. Convert to BAM for efficient storage and further processing.

$ samtools view -Sb SRR32105970\_TP53.sam > SRR32105970\_TP53.bam

Alignment = "everyone knows where they belong"  
Sorting = "putting the list in order from start to end of the genome". This makes it much easier for:

🧮 Counting students (featureCounts)

🔍 Looking them up fast (indexing)

🧬 Checking mutations (variant callers)

So this next step is sorting step

$ samtools sort SRR32105970\_TP53.bam -o SRR32105970\_TP53\_sorted.bam

Convert bam to sam

$ samtools view SRR32105970\_TP53\_sorted.bam > SRR32105970\_TP53\_sorted.sam

$ head SRR32105970\_TP53\_sorted.sam

As according to sam file my data aligned to chromosome 17, I downloaded chromosome 17 from the UCSC browser. Gtf file and moved it to WSL1

$ mv /mnt/c/Users/HP/Downloads/chr17.gtf ~/

**Annotation with the chromosome file --**

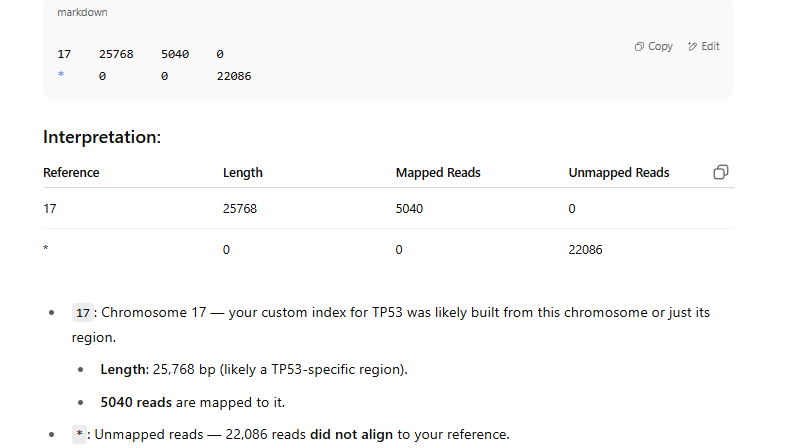
$ featureCounts -a chr17.gtf -o chr17\_counts.txt -t exon -g gene\_id SRR32105970\_TP53\_sorted.bam

The .bai file is the **BAM index** file — required for IGV or any tool that wants to randomly access reads in a BAM file (like to zoom into TP53).

$ samtools index SRR32105970\_TP53\_sorted.bam

This below command gives you read counts per chromosome/contig — and confirms the BAM file is indexed and working properly.

$ samtools idxstats SRR32105970\_TP53\_sorted.bam



### What This Tells You:

* **Only ~18.6% of the reads** (5040 out of 27000+) mapped to chromosome 17 (the TP53 region).
* This is expected for cDNA/RNA-seq targeting TP53, especially if:
  + You trimmed adapters.
  + You used a **partial genome (only chr17)** instead of the whole genome.
  + The sequencing targeted a small region.

Go to the counts.txt check if it has gene id/ transcript id

Accordingly go to <https://www.ensembl.org/biomart/martview>

Ensembl genes 114> Human genes > attributes> check Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version > filters > gene > check Input external references ID list [Max 500 advised] > in it, transcript stable ID > copy and paste your list > then click results.

For Gviz Coverage Plot Visualization done in R for which script and related files are provided