# NGS - Genomics data analysis

1. **Define Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), and Targeted Region Sequencing (TRS). How do these sequencing methods differ in terms of the genomic regions they cover?**

- **Whole Genome Sequencing (WGS)** is a method used to determine the complete DNA sequence of an organism's genome in a single process. It covers both coding regions (genes) and non-coding regions (intergenic DNA, regulatory sequences, and repeats). WGS provides the most comprehensive view of genetic variation, but it is also the most data-intensive and expensive. It is useful for discovering novel mutations, structural variations, and understanding overall genome organization.
- **Whole Exome Sequencing (WES)** focuses only on the Exome — the part of the genome that codes for proteins (~1–2% of the genome). Since most known disease-causing mutations are found in coding regions, WES allows researchers to detect variants relevant to many genetic disorders while reducing sequencing cost and data complexity compared to WGS. However, it misses mutations in non-coding or regulatory regions.
- **Targeted Region Sequencing (TRS)** is the most focused approach, sequencing only a pre-selected set of genes or genomic regions of interest. The target could be a panel of genes linked to a particular disease, a chromosomal region, or certain variants. TRS offers high coverage of specific areas, making it useful for diagnostics, clinical testing, or research on known pathways. It is faster and cheaper than both WGS and WES, but it cannot detect variants outside the chosen regions.
- Difference in sequencing methods in terms of the genomic regions they cover ;

| **WGS** | **WES** | **TRS** |
|---|---|---|
| It covers the entire genome (coding + non-coding) | It covers all coding regions (exons) | It covers specific pre-selected genes/regions |
| The percentage of genome covered is about 100% | The percentage of genome covered is about 1–2% | The percentage of genome covered is variable (depends on panel size) |
| The data volume is very high | The data volume is very moderate | The data volume is low |
| It is expensive | The expense is lower than WGS | The expense is comparatively low |
| Applications include comprehensive genetic studies, novel variant discovery | Applications include disease-gene studies, rare variant detection | Applications include diagnostics, clinical mutation screening |

2. **Explain the roles of metadata fields such as assay type and library selection in the NCBI SRA database for distinguishing WGS, WES, and TRS datasets. Provide examples of the common terms used for each sequencing type.**

- In the NCBI SRA (Sequence Read Archive), each dataset is described by metadata (assay type and library selection) information that tells us what kind of experiment was performed and how the sequencing library was prepared.

A. Assay type - This field describes the general purpose of the sequencing experiment. It’s basically the label that tells us “what” was sequenced. For example:

- WGS datasets usually have assay type as "WGS" (Whole Genome Sequencing).
- WES datasets are marked as "WXS" (Whole Exome Sequencing).
- TRS datasets might appear as "AMPLICON", "TARGETED", or "GENOMIC" depending on the capture method.

B**.** Library selection - This field indicates how the DNA or RNA fragments were selected or enriched before sequencing. Common terms include:

- WGS as "RANDOM" or "PCR" (random fragmentation without target capture).
- WES as "Hybrid Selection" or "Exome Capture" (using probes to pull down exons).
- TRS as "PCR", "Hybrid Selection", "Oligo-dT", or "RT-PCR" (depending on whether they amplified specific genes, panels, or regions).

When searching SRA, these metadata fields help filter results so that we don’t accidentally mix WGS with WES or TRS. For example, if we only want human exome data, we can search for assay type = "WXS" AND library selection = "Hybrid Selection" which ensures I only download datasets prepared specifically for exome sequencing.

**3. Compare the typical data characteristics of WGS, WES, and TRS in terms of sequencing depth, region size covered, and data file size. How can these parameters help infer the sequencing method when metadata is insufficient?**

- Even without perfect metadata, we can often guess the sequencing method by looking at the dataset’s coverage depth**,** region size**,** and file size. These parameters give clues about how much of the genome was sequenced and at what resolution.

A. Sequencing Depth (Coverage)

- WGS - Moderate coverage (e.g., 20–40× for human genomes). The goal is to cover the entire genome evenly.
- WES - Higher coverage (often 50–100×) in coding regions only, because the target is smaller but needs to be very well covered.
- TRS - Very high coverage (can be 200× or more) in just a few genes or specific regions, since sequencing effort is concentrated on a small target.

B. Region Size

- WGS - Around 3.2 billion base pairs for human, covering coding and non-coding DNA.
- WES – Around 30–50 million base pairs (only protein-coding exons, ~1–2% of the genome).
- TRS - From a few thousand bases to a few million bases, depending on the gene panel.

**C.** Data File Size (FASTQ, paired-end, compressed)

- WGS - Very large (for human ~80–200 GB per sample).
- WES - Medium (~10–20 GB per sample).
- TRS **-** Small (<5 GB, sometimes <1 GB if the panel is tiny).

4. **Discuss the advantages and disadvantages of WGS versus WES for clinical and research applications, focusing on cost, data volume, and variant detection capability.**

- Both Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) are powerful tools for studying genetic variation, but they differ in cost, data size, and the type of variants they can detect. WGS is ideal when you need the full picture of genome and can handle the cost and data load. WES is the budget-friendly option for focusing on known coding regions, especially when exonic mutations are the main suspects in clinical or research applications.

WGS (Whole Genome Sequencing)

Advantages-

- Complete coverage: It includes both coding and non-coding regions, regulatory sequences, and intergenic DNA.
- Detects all variant types: It includes SNPs, indels, structural variants, copy number variations, and variants in non-coding regions.
- Better uniformity: It is less bias compared to WES.

Disadvantages-

- High cost: More expensive to run and store compared to WES.
- Large data volume: Requires significant computational resources for storage and analysis (often >100 GB per sample).
- More complex analysis: A lot of variants with uncertain significance in non-coding regions is recorded.

WES (Whole Exome Sequencing)

Advantages-

- Lower cost: Much cheaper than WGS because only ~1–2% of the genome is sequenced.
- Smaller data size: Easier to store and process (~10–20 GB per sample).
- High depth in coding regions: Improves detection of rare variants in exons where most known disease-causing mutations occur.

Disadvantages

- Misses non-coding variants: Can’t detect mutations in promoters, enhancers, or deep intronic regions.
- Capture bias: Some exons may have poor coverage due to uneven hybridization efficiency.
- Limited for structural variants: Large rearrangements or intergenic changes may be missed.

5. **Describe the challenges in identifying targeted sequencing data (TRS) in public repositories and how knowledge of gene panels or capture kits can assist in this process**.

- Identifying targeted sequencing datasets in public repositories such as NCBI SRA can be tricky because the metadata provided by submitters is often incomplete, inconsistent, or uses non-standard terms. Sometimes, the assay type may be labeled simply as "GENOMIC" or "AMPLICON" without clearly stating that the dataset is targeted sequencing. Similarly, library selection fields might list general terms like "PCR" or "Hybrid Selection" that are also used in WES datasets, making it harder to distinguish them. In some cases, researchers may not upload detailed descriptions of the target regions or the gene panel used, leaving only generic experiment titles that do not specify whether the sequencing was targeted.

  Another challenge is that TRS datasets can vary widely in the number of genes or regions covered; some panels target only a handful of genes, while others may include hundreds. This variability means that file size and read depth alone are not always reliable indicators. A TRS dataset with a large panel might appear similar in size to a small exome dataset, causing confusion during filtering.

  Having knowledge of specific gene panels or capture kits used for targeted sequencing can make this process much easier. Many commercial and clinical sequencing providers (e.g., Illumina TruSight, Agilent Sure Select panels) have well-documented lists of the genes they target. If the dataset description or publication mentions one of these panels, you can cross-check the target list to confirm that the data is indeed TRS. Similarly, if you align the reads and see that coverage is concentrated only in a known set of panel genes, this strongly supports that the dataset is targeted. In some cases, the kit name may even appear in the description or title fields in SRA, allowing for direct identification.

  Therefore, identifying TRS in public databases often requires more than just looking at one metadata field .It is more beneficial if we combine available metadata, check associated publications, recognize known panel names, and examine the read mapping pattern. This combined approach helps distinguish TRS from WES or WGS, even when the repository metadata is incomplete or unclear.

6. **What are the typical library preparation methods associated with WGS, WES, and TRS, and how do these influence the sequence data?**

- The choice of library preparation method is a key factor that influences both the quality and characteristics of sequencing data. For Whole Genome Sequencing (WGS), the most common library preparation method is random fragmentation of genomic DNA, followed by end repair, adapter ligation, and PCR amplification. This approach generates libraries that represent the entire genome without bias towards specific regions. As no enrichment step is performed, the coverage is generally uniform across coding and non-coding regions. However, PCR amplification can still introduce some bias and duplicates, especially in low-input DNA samples.

  For Whole Exome Sequencing (WES), library preparation begins similarly with fragmentation and adapter ligation, but includes a target enrichment step. This enrichment is typically done by hybrid capture using biotinylated probes (e.g., Agilent SureSelect, Illumina Nextera Rapid Capture) that are complementary to exonic sequences. These probes bind to the exons, which are then pulled down with magnetic beads, washed, and amplified before sequencing. The enrichment step allows for higher sequencing depth in coding regions, but it can also cause variability in coverage because some exons hybridize more efficiently than others. Poor probe design or low capture efficiency can lead to missing regions in the final dataset.

For Targeted Region Sequencing (TRS), the library preparation method is even more selective. Depending on the design, targeted sequencing may use PCR amplification of specific genes or regions (amplicon sequencing) or custom hybrid capture panels that focus on a defined set of targets, such as cancer-related genes or pharmacogenomics markers. Amplicon-based approaches (e.g., Ion AmpliSeq, multiplex PCR panels) are highly cost-effective and produce extremely high coverage for the targeted regions, but they can introduce amplification bias and may fail if primer binding sites contain variants. Hybrid capture-based TRS works similarly to WES enrichment but uses probes for a much smaller set of regions, allowing for even deeper coverage with less sequencing.

The influence of these methods on the sequence data is significant. WGS libraries tend to produce more uniform coverage but require more sequencing data to achieve the same depth in exonic regions as WES or TRS. WES data shows depth concentrated in coding sequences, but the coverage can be uneven due to capture efficiency. TRS data has extremely high coverage for the targeted regions, making it ideal for detecting low-frequency variants, but offers no information outside the selected panel. Thus, understanding the library preparation method is important when interpreting sequencing results, as it explains the patterns of coverage, potential biases, and the types of variants that can be confidently detected.

**7. Outline an effective workflow to accurately distinguish between WGS, WES, and TRS datasets from public repositories like the NCBI SRA.**

- When retrieving sequencing datasets from public repositories like NCBI's Sequence Read Archive (SRA), correctly identifying whether they are Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), or Targeted Region Sequencing (TRS) is essential before analysis. This process is not always straightforward, because metadata may be incomplete, inconsistent, or use non-standard naming. A robust approach combines metadata inspection, keyword searching, statistical checks, and mapping-based validation. This multi-step approach works as some researchers submit incomplete metadata, and a dataset labelled "GENOMIC" could still be TRS or WES. By cross-checking multiple sources of information and verifying coverage patterns, you reduce the risk of misclassification and ensure that downstream analyses are based on the correct type of sequencing data. An effective workflow to distinguish between datasets is as follows-

A. Metadata Inspection - The first step is to examine the dataset's metadata fields in SRA, especially assay type and library selection. This step often narrows the possibilities, but metadata alone should not be the only criterion.

- WGS - Assay type is often listed as "WGS", with library selection values such as "RANDOM" or "PCR", indicating random genomic fragmentation.
- WES - Usually "WXS" in assay type, with library selection like "Hybrid Selection" or "Exome Capture".

- TRS - May appear as "AMPLICON", "TARGETED", or "GENOMIC" in assay_type, with library selection such as "PCR" (amplicon sequencing) or "Hybrid Selection" for capture-based panels.

B. Review Experiment Descriptions and Linked Publications - The SRA record's **title** or **description** fields, as well as associated Bio Project or Bio Sample records, often mention kit names or panel types. If a linked publication is available, it can provide exact details about target regions, sequencing depth, and purpose. For example:

- *Agilent SureSelect Human All Exon* - WES
- *Illumina TruSight Cancer Panel* - TRS
- *Whole genome shotgun sequencing* – WGS

C. Assess Region Size and Expected Coverage - Even without reading raw sequence data, you can estimate expected size and coverage from study details:

- WGS - Genome size (e.g., human ~3.2 Gb) with moderate coverage (~20–40×).
- WES – About 30–50 Mb (1–2% of genome) with higher coverage (~50–100×).
- TRS - Ranges from a few Kb to a few Mb, but coverage is often very high (>200×).

The difference in target size directly impacts file size:

- WGS FASTQ files are largest (>80 GB for human).
- WES files are moderate (~10–20 GB).
- TRS files are smallest (<5 GB, sometimes <1 GB).

D. Perform Read Mapping for Confirmation - For absolute certainty, download a small subset of reads and align them to the reference genome using tools like BWA or HISAT2. Then examine coverage with samtools depth or genome browsers. This step is especially useful when metadata labels are vague, such as "GENOMIC".

- WGS - Fairly uniform coverage across all chromosomes.
- WES - Strong peaks at exon coordinates almost no reads in intergenic regions.
- TRS - Sharp, dense peaks at specific panel genes, with zero coverage outside.

E. Match to Known Panels or Capture Kits- If capture kit or gene panel names are provided compare them to known target lists from vendor documentation (e.g., Illumina, Agilent, and Thermo Fisher). For TRS, this can confirm whether the reads match expected targets (e.g., *BRCA1*, *BRCA2*, *TP53* in a cancer panel).

F. Integrate Evidence for Classification- Finally, combine information from metadata, descriptions, coverage patterns, and known panels to classify:

- If genome-wide coverage is confirmed - WGS.
- If coverage matches all exons with enrichment bias - WES.
- If coverage is restricted to a small set of known targets - TRS.

**8. Explain the significance of sequencing depth in targeted sequencing and why TRS generally has higher coverage than WGS or WES.**

- Sequencing depth also called coverage refers to the average number of times a nucleotide in the target region is read during sequencing. For example, 30 × coverage means each base on an average is sequenced 30 times. In TRS, sequencing depth is especially significant because:

A. Purpose of TRS - TRS focuses only on specific genomic regions (e.g., a set of genes linked to cancer or hereditary diseases). Since the sequencer doesn't waste reads on non-relevant regions, it can allocate more reads per target base, naturally increasing the coverage.
**B.** Impact on Accuracy**-** Higher depth signifies greater confidence in variant calling. Low-frequency variants, mosaicism, or subclonal mutations can be detected more reliably with >500× coverage in TRS, compared to ~100× in WES or ~30× in WGS. It also reduces false positives by compensating for sequencing errors.
**C.** Higher Coverage of TRS – The high coverage is due to smaller target region**,** fewer bases to sequence, and the same number of total reads is concentrated in a smaller area. For example:

- WGS (3.2 billion bases, human) with 90 Gb data → ~30× coverage
- WES (~50 Mb target) with same 90 Gb → ~1,800× coverage (though usually not all reads map perfectly)
- TRS (~0.5–5 Mb panel) → can exceed 500–2000× coverage easily.

**D.** Importance in clinical applications – It is used to detect somatic mutations in heterogeneous tumour samples, confirming rare pathogenic variants in inherited diseases, validating variants found in WGS or WES studies and many more.

Therefore**,** TRS achieves much higher coverage than WGS or WES because the sequencing effort is concentrated on a small, predefined set of genes, enabling extremely sensitive and accurate variant detection particularly useful for clinical diagnostics where missing a variant could have serious consequences.

# Assignment-2

1. **Describe the overall workflow for analyzing Whole Exome Sequencing (WES) or Targeted Region Sequencing (TRS) data, starting from raw FASTQ files to annotated variant calls.**

- The overall workflow for analyzing Whole Exome Sequencing (WES) data, from raw FASTQ files to annotated variant calls can be understood as a logical chain of steps, each serving a specific purpose to move from raw sequence data to biologically meaningful insights.

**A. Raw Data Quality Control (QC) -** When sequencing is performed, the output is typically a pair of FASTQ files (one for each read in paired-end sequencing). These files contain not only the DNA sequences but also quality scores for each base. Quality control at this stage is essential to detect problems like low-quality reads, adapter contamination, or sequencing biases. It assess if the sequencing run meets expected standards for depth, quality, and coverage. Tools like FastQC provide visual summaries of these metrics. Poor-quality bases can cause downstream misalignments or false variant calls, so identifying and addressing them early prevents wastage of computational effort.

**B. Read Pre-processing (Trimming and Filtering) -** Once issues are identified, the next step is to clean the reads. Remove sequencing adapters that could interfere with mapping, trim low-quality bases at the ends of reads and filter out reads below a certain length or quality threshold. This step ensures that the data going into the alignment process is as accurate as possible, reducing the risk of mismatches and improving mapping efficiency.

**C. Alignment to a Reference Genome -** The cleaned reads is then aligned to a high-quality reference genome (e.g., GRCh38 for humans). Alignment is needed because it determines the genomic location of each read. It allows detection of differences (variants) by comparing the sample to the reference. High-quality aligners (like BWA-MEM) produce SAM/BAM files containing not only the mapped positions but also mapping quality scores, which are crucial for distinguishing true variants from sequencing noise.

**D. Post-Alignment Processing -** After alignment, several processing steps refine the mapped reads. Sorting by genomic coordinates ensures that data is organized for efficient analysis. Marking duplicates removes biases introduced during PCR amplification, where the same fragment is sequenced multiple times. Base Quality Score Recalibration (BQSR) adjusts quality scores to correct for systematic errors introduced by the sequencer.These steps improve the reliability of downstream variant calling.

**E. Variant Calling -** At this stage, the processed BAM files are used to identify positions where the sample differs from the reference genome. In WES, this is restricted to the exonic regions targeted by the capture kit. Variant callers assign confidence scores to each detected variant; distinguish between SNPs (single nucleotide polymorphisms) and indels (insertions/deletions). The result is a VCF (Variant Call Format) file listing candidate variants.

**F. Variant Filtering -** Not all detected variants are trustworthy, some are sequencing errors or alignment artefacts. Filtering removes variants with low coverage or low quality scores, as variants in problematic genomic regions prone to misalignment. The goal is to produce a high-confidence set of variants that is suitable for biological interpretation.

**G. Variant Annotation -** Annotation adds meaning to raw variant calls by linking them to biological knowledge. This includes gene names and affected transcripts, predicted impact on protein function (missense, nonsense, frameshift), population allele frequencies to filter out common variants unlikely to cause rare diseases and links to known disease associations (ClinVar, OMIM). This step transforms raw variant lists into interpretable results for researchers or clinicians.

**H. Biological Interpretation -** Finally, the annotated variants are analysed in the context of the biological question. For rare disease studies, rare, high-impact variants in relevant genes are prioritized and for cancer studies, somatic mutations in oncogenes or tumor suppressors are investigated. This step often involves cross-referencing literature, databases, and pathway analyses to assess the variant's relevance.

2. **Explain the purpose of each tool (FASTQC, Trim Galore, BWA, SAM tools, GATK, SnpEff) in the context of WES/TRS data analysis.**

- The purpose of each tool in the context of Whole Exome Sequencing (WES) data analysis is as follows:

  - **FASTQC (Quality Control of Raw Reads)** - Before doing anything else, we need to know whether our raw sequencing reads are good enough for downstream analysis. FASTQC provides a quick, visual summary of the quality of our FASTQ files. It checks base quality scores across the read length (low scores at the ends indicate potential sequencing errors), detects overrepresented sequences (could be adapters, contaminants, or highly repeated sequences), flags GC content anomalies that might indicate contamination and identifies duplication levels (high duplication may suggest PCR bias).In exome sequencing, poor-quality reads or contamination can directly affect variant calling accuracy, leading to false positives. QC helps us decide if trimming or re-sequencing is needed.

  - **Trim Galore (Adapter Removal & Quality Trimming)** – It removes unwanted adapter sequences and trims low-quality bases so that only reliable sequence information is retained for mapping. It detects and removes adapter sequences from both ends of reads, trims bases below a certain quality threshold to avoid mismatches during alignment and works in a strand-specific way if needed. Adapters left in reads can cause misalignment to the genome, while low-quality bases at the ends can introduce false variant calls. WES often has variable coverage at capture boundaries, so trimming improves both alignment accuracy and variant confidence.

- **BWA - Burrows Wheeler Aligner (Read Alignment to the Reference Genome)-** It aligns high-quality reads to a reference genome (e.g., GRCh38) to determine where each fragment came from in the genome. It uses algorithms optimized for short-read mapping, handles mismatches, insertions, and deletions efficiently and produces SAM/BAM files containing mapping coordinates. The exome is small compared to the whole genome, but accurate mapping is critical because even a single base mismatch can affect whether a variant is detected. Misalignment can cause false variant calls, especially in repetitive exonic regions.

- **SAM – Sequence Alignment Map tools (Manipulation of Alignment Files)** – It prepares and manages the aligned data for variant calling. It converts SAM to BAM (compressed format), it sorts reads by genomic position for efficient processing, indexes BAM files for fast retrieval during analysis and generates basic alignment statistics (coverage, mapping rates).Well-organized, sorted, and indexed BAM files make downstream tools (like GATK) run faster and more accurately. Coverage information from SAM tools is crucial for interpreting why certain variants may be missing (low coverage) or unreliable (uneven coverage).

- **GATK - Genome Analysis Toolkit (Variant Calling and Refinement) -** It identifies and refines genomic variants from the aligned reads, including Pre-processing **(**marks duplicate reads, from PCR to avoid counting them as independent evidence), Base Quality Score Recalibration (corrects systematic sequencing errors in base quality scores), variant Calling (detects SNPs and indels from the alignment data) and Variant Filtering **(**removes low-confidence calls using statistical thresholds). GATK's algorithms are tailored for human exome sequencing, where accurate detection of variants in coding regions is essential. Missteps in this stage can lead to missing pathogenic variants or reporting spurious ones.

- **SnpEff – SNP Effect Predictor (Variant Annotation)** – It adds biological meaning to the detected variants by linking them to genes, protein changes, and functional effects. It predicts whether a variant is synonymous, missense, nonsense, frameshift, etc. , links variants to known databases (e.g., ClinVar, dbSNP) and reports which transcripts are affected and how.  Raw variant calls alone are just coordinates and base changes therefore annotation with SnpEff turns them into interpretable biological insights, for example, a mutation that causes a premature stop codon in a cancer-related gene. This is the stage where variants go from "detected" to "biologically meaningful."

3. **Detail the changes or considerations needed—compared to WGS—when applying each analysis step to WES or TRS data, particularly in alignment and variant calling.**

- Whole Exome Sequencing targets only the coding regions (exons) of the genome, which comprise about 1–2% of the total DNA. While the general workflow for WES shares many similarities with WGS including quality control, alignment, variant calling, and annotation, several key considerations are unique to WES due to its targeted nature.

- **Quality Control (QC) -** In both WES and WGS, initial raw FASTQ files undergo quality assessment using tools like **FASTQC** to detect low-quality reads, adapter contamination, and sequence composition biases. For WES, QC interpretation must account for the enrichment method used during library preparation. WES datasets often exhibit variable coverage across targeted exons due to capture efficiency, GC-content bias, and probe design limitations. Unlike WGS, where coverage tends to be more uniform, WES may show overrepresented or underrepresented exonic regions. This means downstream analyses must include checks for coverage uniformity and ensure that enough reads map to targeted regions to avoid false negatives in variant calling.

- **Trimming -** Trimming adapters and low-quality bases is necessary in both workflows. In WES, however, trimming must be conservative, as excessive trimming can disproportionately affect already short exonic reads and reduce the number of reads mapping to targeted regions. Since WES is capture-based, the presence of partial adapter sequences is common if the insert size is small. Trim Galore, which wraps Cutadapt and FASTQC, is often used. Post-trimming QC is especially critical in WES to ensure that the proportion of reads mapping to target intervals remains high.

- **Alignment -** Alignment is a core stage where WES requires special attention compared to WGS. In both, a reference genome (e.g., GRCh38) is used with tools like BWA-MEM to map reads. However, WES reads are only expected to align within exonic regions plus some flanking intronic sequences captured during hybridization. Misalignment risk is higher in WES for two reasons; shorter insert sizes and the higher presence of PCR duplicates can cause reads to map incorrectly to repetitive regions. Also, off-target reads occur due to non-specific capture, meaning a small fraction of reads map outside exonic targets.To optimize alignment for WES, the analysis may use BED files of targeted regions to filter alignments post-mapping, or adjust alignment scoring parameters to better handle shorter reads. Additionally, coverage depth in WES is much higher (often 50–100×) than typical WGS (30–40×), so aligner performance must efficiently handle deep coverage in small genomic intervals.

- **Post-Alignment Processing -** After alignment, SAM/BAM files undergo processing such as sorting, marking/removing PCR duplicates, and indexing. In WES, duplicate removal is more critical because PCR amplification during capture can introduce high duplicate rates. Excess duplicates inflate coverage statistics without providing true sequencing depth, leading to potential false positives. SAM tools handle basic processing, while GATK adds steps like base quality score recalibration (BQSR). For WES, recalibration is particularly valuable because uneven capture efficiency can bias base quality scores, affecting downstream variant calling accuracy.

- **Variant Calling -** Variant calling in WES differs from WGS primarily in its region restriction. Since WES targets exons, variant callers (e.g., GATK Haplotype Caller, Free Bayes) are typically instructed to call variants only within targeted intervals specified in a BED file. This reduces computational load and minimizes false positives in intergenic or poorly covered regions. Key considerations include depth thresholds (WES typically sets a higher minimum depth for calling variants (e.g., 10–20×) compared to WGS, given the higher average coverage in exomes), handling coverage gaps (some exons may have poor coverage due to capture inefficiencies, requiring careful filtering to avoid reporting spurious variants), and indel detection (short indels in WES can be harder to detect if they occur in poorly captured or low-complexity regions). When using GATK for WES, it is explicitly recommended to use target intervals during variant calling and applying exome-optimized hard filters or variant quality score recalibration (VQSR) based on exome datasets.

- **Variant Filtration -** While WGS variant filtering considers genome-wide false positives, WES filtration must focus on exonic-specific issues like edge effects (variants near capture probe edges may have biased coverage), high depth artifacts (extremely high depth in certain targets may be due to PCR over amplification rather than biological signal and off-target variants (some variants detected outside target intervals may still be biologically relevant if they occur in UTRs or splice sites adjacent to exons).

- **Variant Annotation -** Tools like SnpEff (SNP effect predictor) annotate variants with predicted functional impacts. For WES, annotation is more straightforward than in WGS because nearly all variants fall in coding or splice-adjacent regions. However, this also means that functional prediction accuracy is more critical — for example, distinguishing synonymous from missense variants, or identifying stop-gain mutations that might disrupt protein function. Since WES lacks genome-wide data, annotation interpretation must be cautious when predicting regulatory or structural impacts beyond exonic boundaries.

- **Interpretation and Reporting -** In WES, interpretation is enriched for clinically relevant variants because exons harbor the majority of known disease-causing mutations. However, the limitation is that WES misses deep intronic, regulatory, and structural variants outside exonic regions. Compared to WGS, reporting for WES tends to focus on variant pathogenicity (ClinVar, COSMIC), population frequency (gnomAD), and predicted impact on protein structure or function. Functional enrichment or pathway analysis can also be performed, but it’s limited to coding genes.

While WES and WGS share the same fundamental analysis steps, WES workflows are shaped by the targeted, capture-based nature of the data. This affects quality control interpretation (coverage bias), trimming strategy (adapter prevalence), alignment optimization (target restriction, off-target reads), post-alignment processing (duplicate removal), and variant calling (interval-based calling, depth requirements). Alignment and variant calling in WES must be more targeted and parameter-tuned to handle high depth, uneven coverage, and region-specific biases, ensuring that clinically and biologically relevant exonic variants are detected with high confidence.

4. **Discuss how to handle targeted region/bed files in exome or panel analysis, and explain their use in coverage analysis and variant filtering.**

- In whole-exome sequencing (WES) analysis, targeted region files, most commonly provided in BED format, play a central role in ensuring that the analysis focuses strictly on the exonic regions or other targeted intervals captured during library preparation. A BED (Browser Extensible Data) file is a simple text file that lists genomic intervals (chromosomes start position and end position) representing the specific targets for sequencing. While WGS data spans the entire genome, WES focuses on ~1–2% of the genome, so handling BED files properly is critical for accurate coverage evaluation, alignment optimization, and variant filtering.

  During the alignment and variant calling process, the BED file serves as a “roadmap” of which genomic regions should be evaluated. In alignment, while most pipelines (e.g., BWA-MEM, Bowtie2) align all reads regardless of location, downstream processes like coverage analysis and variant calling can be restricted to only those intervals in the BED file. Variant calling tools such as GATK Haplotype Caller, bcftools, often include parameters (-L in GATK, --targets in bcftools) to restrict calling to the provided BED intervals, which reduces computational load, minimizes noise from off-target reads, and ensures consistency with the experimental design.

  Coverage analysis is crucial in WES because capture efficiency varies across regions due to hybridization biases, GC content, and probe design. BED files allow tools like bedtools coverage, samtools depth, or mosdepth to calculate per-target coverage metrics. These metrics are used to identify poorly covered exons, which could affect variant detection sensitivity. For example, a minimum threshold (e.g., 20× depth) may be required for confident variant calls; regions falling below this threshold can be flagged for caution or recommended for re-sequencing. In clinical or diagnostic contexts, coverage reports generated from BED files are often required to ensure all clinically relevant exons meet quality standards.

  In WES, variant callers may produce VCF files containing calls from both targeted and incidental off-target regions. To ensure analysis relevance, variants can be filtered to retain only those overlapping BED-defined target regions. This can be done with tools like bcftools view -R, GATK Select Variants, or bedtools intersect. This step is essential for minimizing false positives from low-complexity or poorly captured non-target areas. Additionally, BED files can be used in post-variant-calling annotation workflows to flag variants outside target intervals, ensuring downstream interpretation focus on regions with reliable coverage.

  BED files are essential in WES workflows for defining capture targets, guiding variant calling to relevant regions, ensuring adequate coverage assessment, and filtering variants to maintain analysis accuracy. Without properly incorporating BED files, WES analysis risks including off-target noise, overlooking under-covered exons, and reducing both the sensitivity and specificity of variant detection. Proper handling of targeted BED files thus ensures that WES results are both biologically and technically aligned with the experiment’s intended design.

5. **Based on the output of coverage analysis, how would you determine if your target regions were adequately sequenced?**

- To determine if target regions in an exome sequencing experiment were adequately sequenced, we need to evaluate the coverage analysis output against predefined quality and depth thresholds. The process generally involves analyzing:

  - Depth of coverage (per-base or per-target) - Depth refers to the number of reads mapped to each base within the target region. For exome analysis, adequate sequencing typically means achieving a minimum coverage threshold that ensures reliable variant calling (e.g., ≥20–30× for germline variant detection, and higher, such as ≥100×, for somatic mutations). We should review coverage histograms or summary statistics to see the proportion of bases that met or exceeded this threshold. If many bases in target regions have coverage below the set threshold, they may not be reliable for downstream variant calling.

  - Breadth of coverage (percentage of target bases covered) - Breadth measures what proportion of the targeted bases were sequenced above the depth threshold. For high-quality exome data, you might aim for ≥95% of targeted bases covered at ≥20×. If the breadth is significantly lower, it could indicate problems in library preparation, capture efficiency, or sequencing bias.

  - Uniformity of coverage - Uniformity assesses whether the sequencing depth is evenly distributed across the target regions. Uneven coverage means some targets are overrepresented while others are underrepresented, which can lead to missed variants in poorly covered regions.

  - Identifying under-covered or uncovered targets –After coverage analysis, you can generate a list of low-coverage regions (<threshold) and cross-check them against the BED file to see which targeted genes/exons were poorly sequenced. This is important for clinical or diagnostic applications, where incomplete coverage of critical genes could require re-sequencing or supplemental targeted assays.

  Based on adequacy assessment if most target bases meet the coverage criteria then we can proceed to variant calling confidently. If significant regions fall below the threshold we must consider deeper sequencing, refining capture protocols, or merging data from replicate runs. If poor coverage is localized to specific GC-rich or repetitive regions, bioinformatics adjustments (local realignment, special variant callers) or alternative capture methods might be necessary. Adequate sequencing in exome analysis is confirmed when the majority of target bases exceed a defined coverage depth, the percentage of covered bases meets the required threshold, and coverage is uniform enough to minimize bias, ensuring high sensitivity and accuracy in variant detection.

6. **Interpret a VCF output: How would you prioritize variants identified within targets versus off-target regions in WES/TRS?**

- In Whole Exome Sequencing (WES), the Variant Call Format (VCF) file contains both target-region variants (within the exome capture intervals) and off-target variants (outside the intended regions). Prioritization begins with filtering based on the target region BED file, ensuring that variants inside these regions are analyzed first, as these are most likely to be relevant to the biological or clinical question, given that WES is specifically designed to capture exonic sequences where most protein-altering variants occur. Variants within targets are usually higher in coverage, making them more reliable; therefore, filtering by high sequencing depth (e.g., ≥20–30×) and quality scores helps retain confident calls. Additional annotation using tools like ANNOVAR, VEP, or SnpEff can determine the predicted impact (missense, nonsense, frameshift) and link to databases like ClinVar or gnomAD for population frequency, aiding pathogenicity assessment.

  Off-target variants those falling outside the capture design should be interpreted cautiously. Although they can sometimes reveal biologically interesting or novel findings (e.g., regulatory variants in UTRs or promoters), they often have lower coverage and higher error rates, making them less reliable. If coverage analysis shows that off-target regions were incidentally well-covered, they can be considered in exploratory analysis but not given the same weightage as in target findings. For clinical diagnostics, off-target calls are typically excluded unless they are in genes of known significance or part of secondary findings reporting guidelines.

  When prioritizing within the target set emphasis should be placed on variants with predicted high functional impact, rare or novel alleles, and those that segregate with the phenotype in family studies. For research studies, broader inclusion criteria may be applied, whereas in clinical settings, stringent criteria based on established guidelines are followed. Therefore, VCF interpretation in WES starts by segmenting variants into in-target and off-target groups using the BED file, giving highest priority to high-confidence, well-covered, in-target variants with functional and clinical relevance. Off-target variants can be retained for exploratory purposes but are secondary in priority. This approach maximizes the reliability and biological relevance of findings while ensuring efficient use of sequencing data aligned to the original capture design.

7. **Evaluate the advantages and limitations of using the given pipeline tools (FASTQC, Trim Galore, BWA, SAMtools, GATK, SnpEff) for WES/TRS versus WGS.**

- When evaluating the use of tools like FASTQC, Trim Galore, BWA, SAMtools, GATK, and SnpEff for Whole Exome Sequencing (WES) compared to Whole Genome Sequencing (WGS), the key difference lies in the data scope, target specificity, and coverage patterns.

- FASTQC is equally essential for both WES and WGS as it provides a quality assessment of raw reads (per base quality, GC content, adapter contamination). In WES, since capture efficiency varies across exons, FASTQC is especially useful for spotting uneven quality profiles caused by hybrid capture bias. In WGS, quality issues are more uniform across the genome.

- Trim Galore performs adapter and quality trimming, which benefits both approaches. However, WES reads often require stricter trimming thresholds because capture-based enrichment may introduce more adapter contamination from short fragments. Over-trimming in WES could risk losing reads mapping to target exons, while in WGS the impact is less critical given the uniform genome-wide coverage.

- BWA (Burrows-Wheeler Aligner) is widely used for aligning short reads to a reference genome. For WES, accuracy in alignment is crucial, especially at exon boundaries, since downstream variant calling focuses on small target regions. Misalignments near exon–intron junctions could cause false positives results. In WGS, the larger and more uniformly covered dataset helps reduce local mapping artifacts, but computational demands are higher.

- SAMtools handles file manipulation, sorting, indexing, and basic variant calling. In WES, indexing BAM files allows rapid access to specific exon intervals, especially when using BED files to subset the data for targeted analyses. In WGS, using SAMtools for indexing is equally important but less likely to involve frequent targeted sub setting.

- GATK is central for variant calling and filtering. In WES, GATK's Haplotype Caller must be used with a targeted BED file to avoid wasting computational effort on off-target regions, and base recalibration must consider the high coverage but uneven distribution across exons. In WGS, recalibration benefits from genome-wide data and is less affected by capture bias but demands significantly more CPU time and memory.

- SnpEff is used to annotate variants for functional impact. In WES, annotation focuses on coding regions, splice sites, and UTRs, making it more relevant for clinical/functional interpretation. In WGS, SnpEff covers both coding and non-coding variants, giving a broader picture but requiring more extensive filtering to focus on biologically relevant changes.

  While the same tools apply to both WES and WGS, WES demands greater emphasis on handling targeted intervals, managing uneven coverage, and minimizing false calls in enriched regions, whereas WGS prioritizes scalability, computational efficiency, and genome-wide interpretation.

8. **Describe potential sources of false positives/negatives in WES/TRS variant calling and how pipeline step modifications (e.g., realignment, base recalibration) can address these issues.**

- In whole-exome sequencing (WES), variant calling accuracy can be influenced by various sources of false positives (variants incorrectly called) and false negatives (true variants missed). False positives often arise from sequencing errors, PCR amplification artifacts, or misalignments, especially in regions with low complexity, high GC content, or repetitive sequences. Poor-quality reads, contamination, or improper adapter trimming can also introduce artificial variants. Inadequate alignment parameters in tools like BWA may lead to mismatches near indels, causing erroneous variant calls. Similarly, incomplete removal of duplicates can result in artificial read depth, leading to false positives.

  False negatives typically occur when the sequencing depth in a target region is insufficient to detect heterozygous variants, often due to capture inefficiency or uneven coverage across the exome. Highly homologous regions (e.g., pseudogenes) can cause reads to misalign or be discarded, hiding true variants. Low base quality, excessive trimming, or overly stringent variant filtering thresholds can also suppress genuine variant calls. In panel-based or exome capture workflows, off-target sequencing or incomplete probe hybridization can leave gaps in the target regions, contributing to missed variants.

Pipeline step modifications can help mitigate these issues. Local realignment around indels corrects misalignments caused by indel-containing reads, reducing false positives in these regions. Base Quality Score Recalibration (BQSR) in GATK adjusts per-base error estimates based on machine cycle and sequence context, allowing more accurate variant quality scoring and reducing false positives from systematic sequencing errors. Marking duplicates (with SAMtools or Picard) eliminates PCR-induced artifacts before variant calling.

Adjusting coverage thresholds during filtering can reduce false negatives e.g., retaining variants with moderate depth but strong allele balance. Incorporating targeted BED files ensures that coverage analysis is focused on intended regions, allowing identification and resequencing of poorly covered targets. Using SnpEff or similar tools for annotation helps prioritize high-impact variants and flag potential sequencing artifacts. Ultimately, combining robust QC (FASTQC, Trim Galore), accurate alignment (BWA), careful duplicate handling, realignment, base recalibration, and region-aware filtering improves both sensitivity and specificity in WES variant calling. The goal is to balance stringent filtering to avoid false positives with sensitivity to retain genuine variants, especially in clinically relevant regions.

9. **How would you incorporate gene panel/target information in the downstream analysis and interpretation of results?**

- In downstream analysis of Whole Exome Sequencing (WES) or targeted resequencing (TRS) data, incorporating gene panel or target information is essential for focusing interpretation on clinically or biologically relevant regions. After raw reads are processed through alignment, variant calling, and annotation, the next step is to integrate BED files or curated gene panel lists to restrict attention to variants in regions of known interest (e.g., disease-associated genes, pharmacogenomics markers, or specific signaling pathways). This is typically done by intersecting the final annotated VCF with the target regions file using tools like BEDTools intersect, bcftools view –R, or GATK's Select Variants. This filtering ensures that variants outside of the intended target regions often of lower reliability or clinical relevance are excluded from priority consideration.

Gene panel information also guides variant prioritization based on the biological role of the genes in question. For example, in a cancer panel, variants in tumor suppressor or oncogene hotspots would be ranked higher for follow-up validation. Annotation tools like SnpEff or ANNOVAR can be configured to annotate variants with gene-specific details such as transcript consequences, known pathogenicity (ClinVar), and disease association databases (OMIM, COSMIC). These annotations become more meaningful when cross-referenced with the curated gene panel to provide a context-specific interpretation.

Coverage metrics derived from the BED-defined targets are also incorporated at this stage to evaluate sequencing adequacy. Poorly covered target regions (e.g., <20× depth) are flagged, as variants in these regions may be false negatives. This allows for interpretation caution or even re-sequencing of low-coverage targets. Off-target variants, if retained, can be assessed separately but are generally deprioritized unless they have known high clinical relevance. In clinical diagnostics, gene panel integration supports the creation of structured reports where variants are categorized according to AMP (Association for Molecular Pathology) guidelines, with emphasis on those in target genes. For research applications, it allows more focused statistical analyses (e.g., variant burden tests limited to panel genes). In both cases, targeted information enhances the specificity of results, reduces noise from genome-wide background variation, and streamlines biological interpretation. Overall, incorporating gene panel/target region information in downstream WES/TRS analysis ensures that results remain clinically actionable, biologically relevant, and methodologically robust, while also aligning the interpretation process with the initial design goals of the sequencing experiment.

**10. Outline an approach for reporting clinically relevant variants using the workflow and tools provided.**

- Raw Data Quality Assessment (FASTQC) - To ensure the sequencing reads are of sufficient quality before proceeding. Use FASTQC to generate per-base quality scores, GC content plots, adapter content analysis, and sequence duplication levels. Then identify potential issues like low-quality bases, presence of adapters, overrepresented sequences, or unexpected GC profiles. The importance is high-quality raw reads reduce the likelihood of false variant calls, which is crucial for clinical decision-making.

- Pre-processing (Trim Galore) – To improve the accuracy of alignments and variant calls. Use Trim Galore to remove sequencing adapters and trim low-quality bases (based on Phred scores).Then re-run FASTQC post-trimming to verify improvement. It reduces false positives caused by poor-quality bases or adapter contamination.

- Alignment to Reference Genome (BWA) – It maps reads accurately to the human reference genome (e.g., GRCh38).Use BWA-MEM for aligning trimmed reads. The generated output SAM/BAM files are further used for downstream analysis. Accurate mapping ensures correct genomic positions for variant identification.

- Post-alignment Processing (SAMtools) – It prepares alignment files for variant calling. First sort and index BAM files using SAMtools. Then generate alignment statistics (e.g., coverage depth, mapping quality). This ensures efficient processing and easy access to genomic coordinates during variant calling.

- Variant Calling (GATK)-To identify single nucleotide variants (SNVs) and insertions/deletions (indels).Following GATK best practices like marking duplicates (remove PCR bias), Base Quality Score Recalibration, and using Haplotype Caller for variant calling. Then in output we get a raw VCF file containing all potential variants. It helps minimizes technical artifacts and maximizes accuracy in variant detection.

- Variant Annotation (SnpEff) – It adds biological and clinical meaning to the raw variant list. Annotate variants with SnpEff to determine genomic location (exonic, intronic, intergenic), functional consequence (missense, nonsense, frameshift), and predict impact (high, moderate, low). It includes gene name, amino acid change, and functional class. It converts raw variant data into interpretable biological information.

- Filtering Clinically Relevant Variants – It narrows down to variants of potential clinical impact. Remove variants with low read depth, low quality scores, or high frequency in general population databases, prioritize variants in clinically relevant genes or target regions, then use AMP classification guidelines to categorize as pathogenic, likely pathogenic, likely benign, or benign.

- Cross referencing with Clinical Databases – It validates and interprets variants against established knowledge. Check ClinVar for existing clinical interpretations, then use dbSNP for reference IDs, search OMIM for gene-disease associations and review literature for newly reported pathogenic variants. It ensures that the reported variant has supporting evidence and context.

- Clinical Report Preparation – To present findings in a clear, concise, and actionable format. Include patient/sample information (ID, test date, reference genome version).For each clinically relevant variant provide genomic coordinates (chromosome, position, reference/alternate allele), gene name and transcript reference, zygosity (heterozygous/homozygous), variant consequence (e.g., nonsense mutation), clinical classification (ACMG category), associated conditions with references and evidence strength (database entries, functional studies, segregation analysis). Also state any limitations (e.g., structural variants not covered). It allows clinicians to make informed decisions based on scientifically validated data.