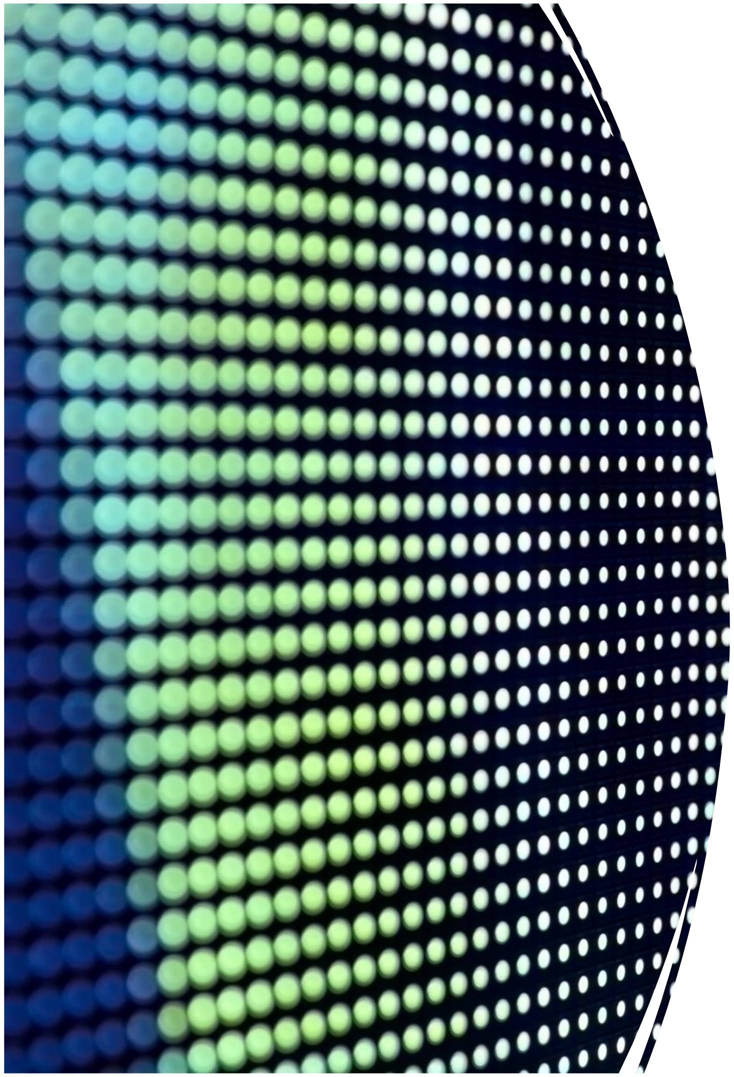
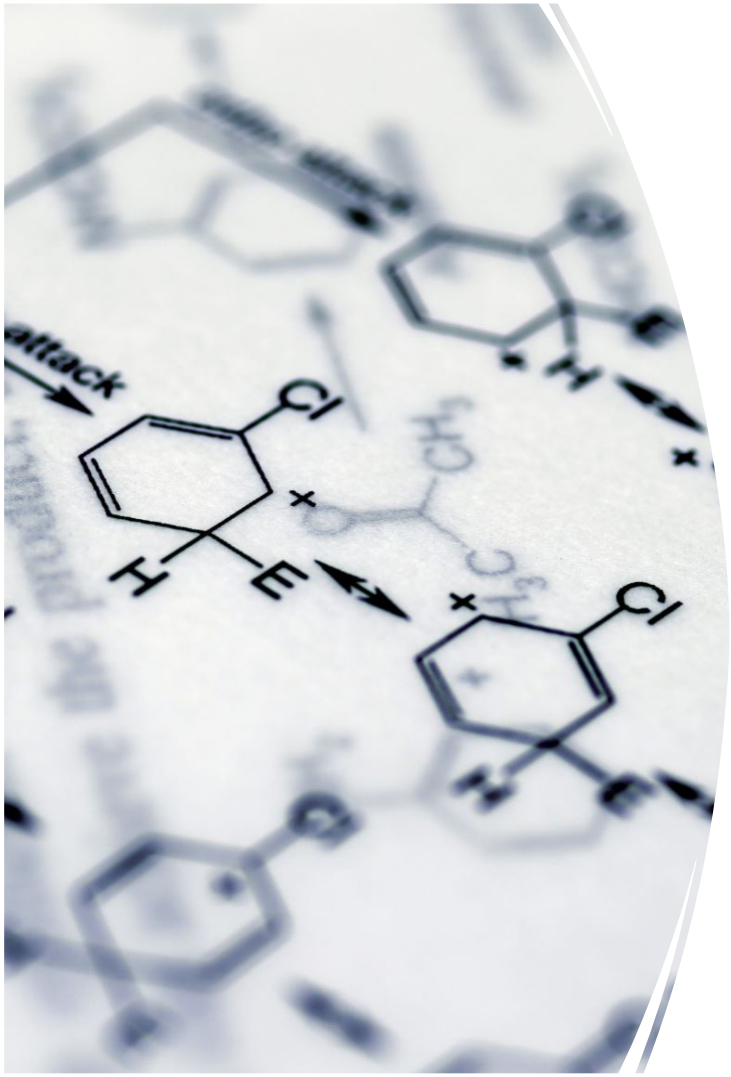


Whole exome sequencing (WES) for SNP and short indel discovery

From:

Nirvan Kotha
Leela Sai Krishna Pannem
Sumanth Kumar Lingabathini
Abhishek Arugonda
Aravinth Subramanian



Introduction

- •Whole Exome Sequencing (WES) is a next-generation sequencing (NGS) technique that focuses on sequencing only the protein-coding regions of the genome, known as the exome. The exome comprises about 1-2% of the genome but contains the majority of known disease-related variants.
- •WES is widely used for identifying single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) associated with genetic disorders, cancers, and other diseases.

What is SNP?

A **Single Nucleotide Polymorphism (SNP)** is a genetic variation that occurs when a single nucleotide (A, T, C, or G) in the DNA sequence is altered at a specific position in the genome. SNPs are the most common type of genetic variation among individuals and can influence traits, disease susceptibility, and drug responses.

SNPs help in identifying genetic predispositions to diseases such as cancer, diabetes, and cardiovascular disorders.

Short Indels

Short Indel Discovery refers to the identification of small insertions and deletions (indels) in a genome. Indels are a type of genetic variation where a few base pairs (typically 1–50 bp) are inserted or deleted in the DNA sequence.

Applications of Indel Discovery:

- **Disease Diagnosis**: Identifying pathogenic indels linked to genetic disorders.
- **Cancer Genomics**: Detecting indels that drive tumor formation.
- **Evolutionary Studies**: Understanding genetic variations across populations.

Data set

Raw sequencing data consists of short DNA sequence reads produced by high-throughput sequencing (HTS) platforms like Illumina, Ion Torrent, or PacBio. These reads are unprocessed and contain nucleotide sequences (A, T, C, G) along with quality scores. This data serves as the primary input for downstream genomic analysis, including alignment, variant calling, and identification of SNPs (Single Nucleotide Polymorphisms) and short indels (insertions and deletions).

Format of Raw Sequencing Data

The data is typically stored in **FASTQ format**, which includes:

- **Sequence Reads:** DNA sequences represented by A, T, C, G.
- **Quality Scores:** A Phred score indicating confidence in each base call.
- **Identifiers:** Unique labels for each read.

Why is Raw Sequencing Data Important?

- It provides the **initial genetic information** required for downstream processing.
- Essential for **variant calling**, where SNPs and indels are identified.
- Helps in **genome-wide association studies (GWAS)** and **disease research**.

BIOLOGICAL IMPLICATIONS

Alteration of Protein Function:

Variations such as SNPs or small insertions/deletions can change the amino acid sequence of proteins. This might alter protein structure or function, potentially disrupting normal cellular processes.

Disease Pathogenesis:

Many genetic disorders arise from such mutations. For instance, a single nucleotide change or a small deletion can lead to malfunctioning proteins that cause inherited diseases or contribute to cancer development.

Genetic Diversity and Evolution:

SNPs are one of the most common forms of genetic variation and serve as markers for studying genetic diversity. This diversity can provide insights into human evolution, population migration patterns, and adaptation to environmental changes.

MEDICAL IMPLICATIONS

1. Disease Diagnosis & Risk Assessment

- Many inherited diseases are caused by SNPs and indels in protein-coding genes.
- Certain SNPs and short indels are oncogenic mutations that drive tumor growth.

2. Precision & Personalized Medicine

Pharmacogenomics (Drug Response Prediction)

SNP variations influence how individuals metabolize drugs, affecting dosage and effectiveness.

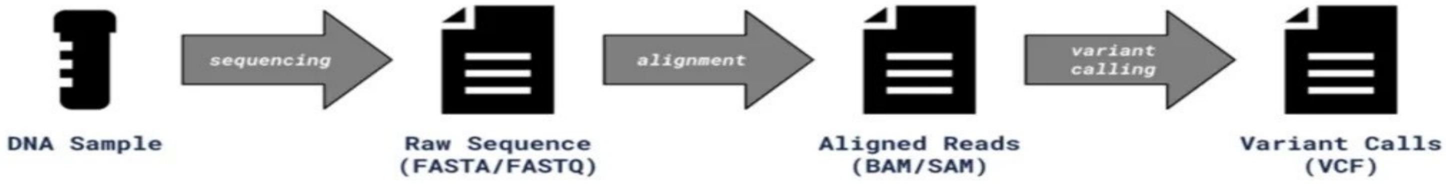
MEDICAL BENEFITS

- Improves diagnostic rates for undiagnosed patients.
- Reduces unnecessary medical tests and procedures by pinpointing the exact cause of disease.
- Enables precision medicine by identifying drug-targetable mutations.
- Supports prenatal and reproductive decision-making through genetic counseling.

Pipeline / Workflow

Aim:

- Start with sequencing reads and perform a series of steps to determine a set of genetic variants.



- Create a file (VCF) with the variants from the data.

What is Genome Analysis Toolkit (GATK) ?

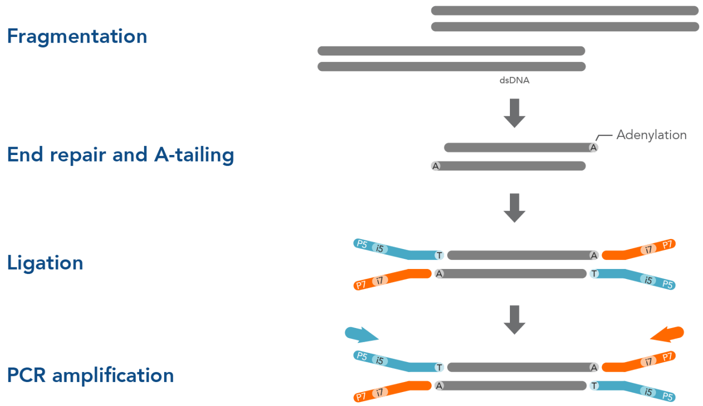
- Package of command-line tools (written in Java)
- GATK Tools vane used individually or chained together into compete workflows
- GATK pipelines reply on another Java Packages, PICARD for processing of alignment files

- Contains multiple tools for
  - NGS data processing
  - Genotyping and variant discovery
  - Variant filtering and evaluation

- Ever evolving and adapting to emerging sequencing technologies
- GATK provide end- to-end workflows, called GATK Best Practices, tailored for specific use cases.
- GATK is designed to run on Linux and other platforms, Which includes MacOS X. windows system are not supported. The major system requirements is java1.8

Library Preparation:

- **Fragmentation:** DNA is sheared into smaller fragments (~150–300 bp) to fit sequencing platform specifications.
- **Adapter Ligation:** Short adapter sequences are added to the DNA fragments to prepare them for sequencing.
- **Target Enrichment:** Use of hybridization probes to selectively capture exonic regions or coding regions for sequencing.
- **Amplification:** Amplify the enriched DNA fragments using PCR to ensure sufficient quantity for sequencing.



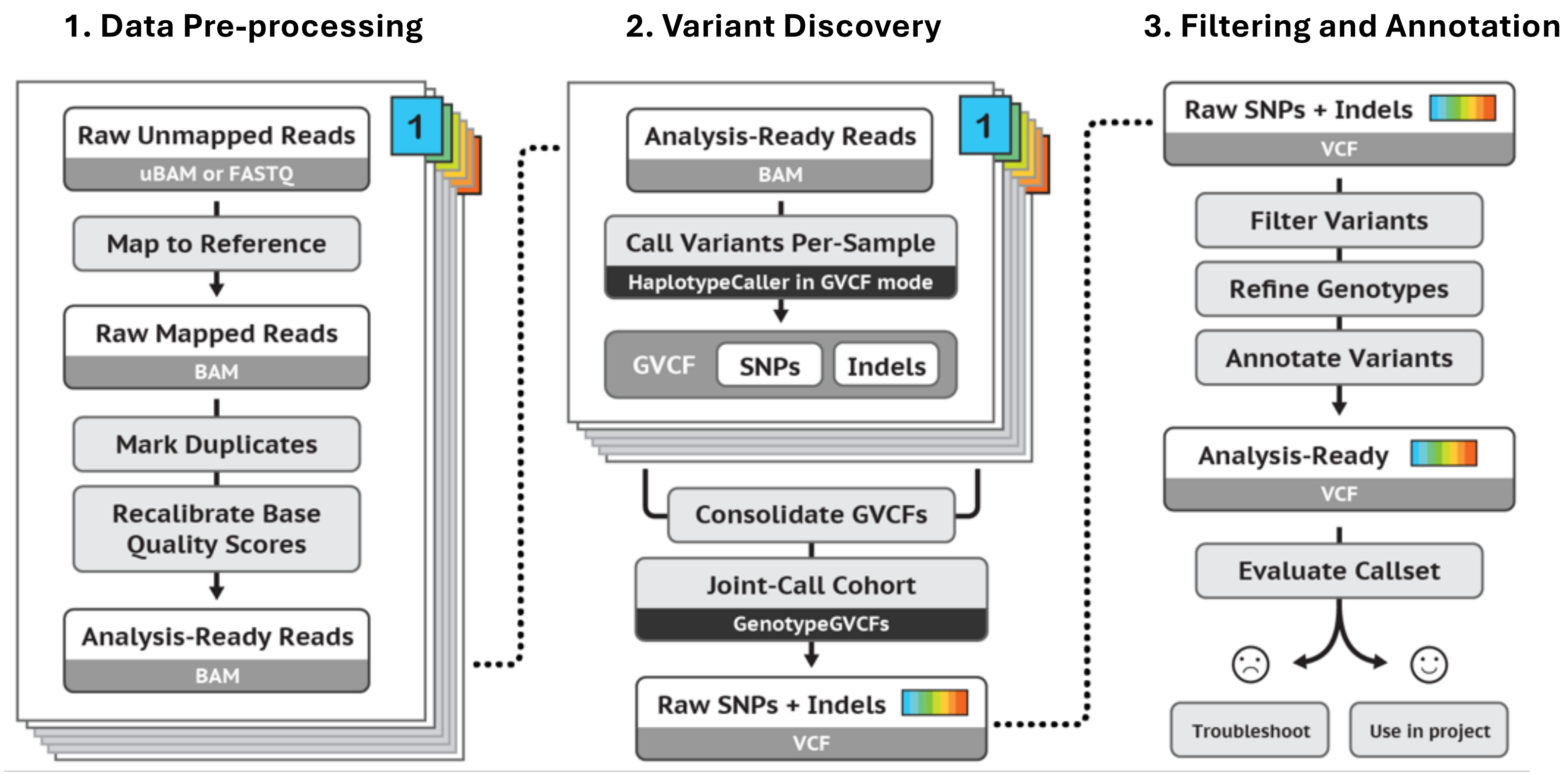
Fragmentation

End repair and A-tailing

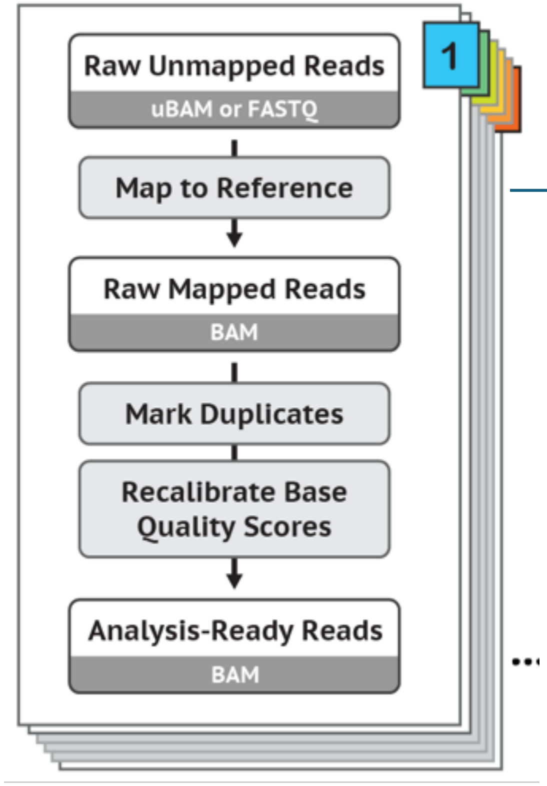
Ligation

PCR amplification

GATK Workflow steps



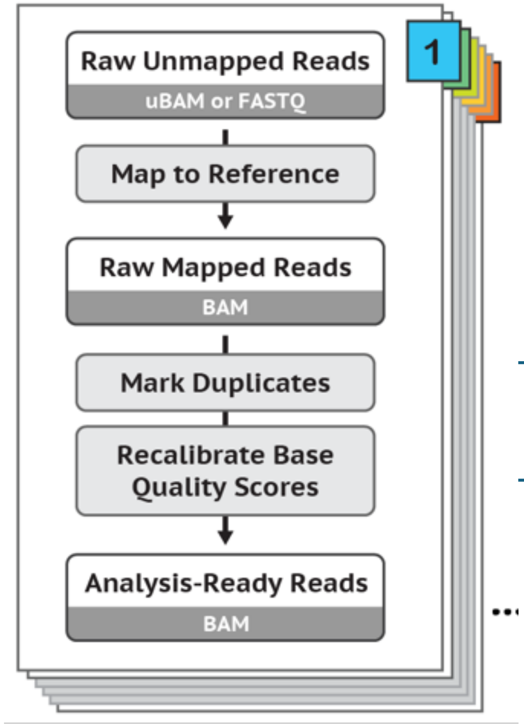
Data Pre-processing Step 1



BWA

- BWA-backtrack = designed for **Illumina sequence reads up to 100bps. (short Range)**
- Mapping is a reference genome is a process of aligning **sequencing reads to a known reference genome** to find exact location.
- Which identify where each read originates in the genome and allows **detection of genetic variations**.
- The best matching location is chosen.
- The **results are stored in a SAM or BAM file**, which have read sequence , position on the reference genome, mapping quality score.

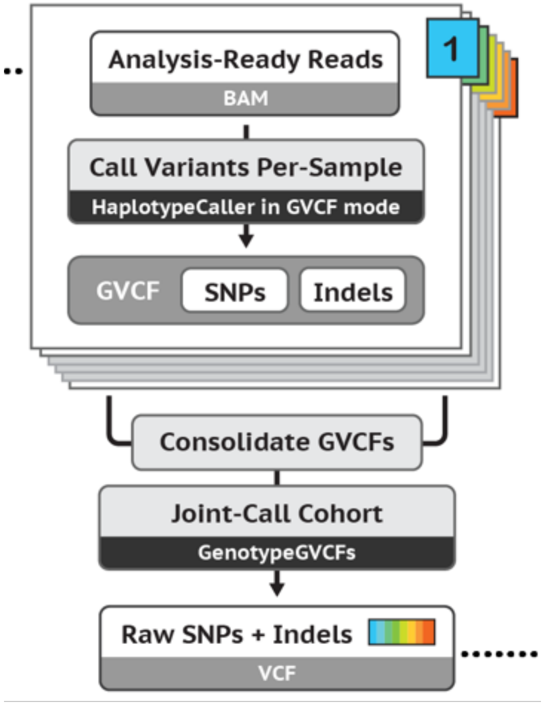
Data Pre-Processing Step 1



- PCR duplicates are identical DNA sequences generated during PCR amplification in sequencing, essentially creating copies of DNA sequence.
- These duplicates can bias variant calling results, so their removal ensures accurate downstream analysis.
- Identified and removed using Tools Picard, SAM tools.

- Recalibrating base quality scores corrects biases in sequencing data by adjusting confidence values to reflect true errors rates, improve scores. This ensures better variant calling.
- Using GATK Base Recalibrator Tools, the base quality scores are adjusted.

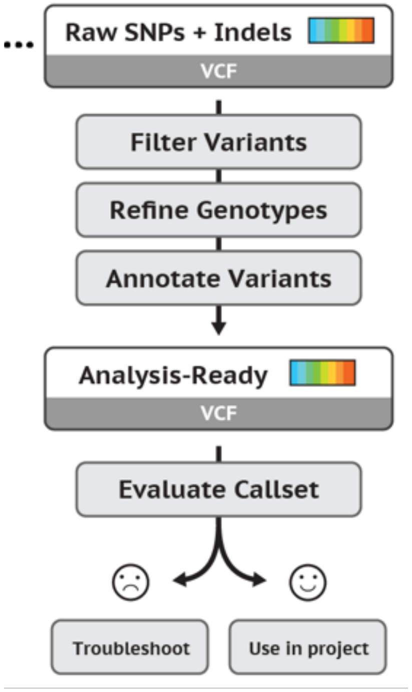
Data Pre-Processing Step 2



HaplotypeCaller(GATK)

- Which identifies regions of genome that shows signs of variation.
- Detects the SNP and Indels in high sequencing data.
- This examines the read data and compares it to the reference genome to detect difference.
- Generates a VCF file with the variant annotations for downstream analysis.

Data Pre-Processing Step 3



- The **raw variants** are filtered based on quality metrics such as **Depth(DP)**, **Strand bias**, and **Genotype Quality(GQ)** to remove low-quality or false-positive variants, ensuring high-confidence variants for analysis.

- Next, the filtered variants are **annotated** using tools like **SnpEff**, **VEP**, or **ANNOVAR** to add functional information, such as gene effects and protein impact, which helps interpret their biological and clinical significance.

- Analyzing a VCF file involves identifying variants by filtering for rare, high impact, or pathogenic variants and performing population genetics.

Tools and Software required

- **Quality Control** : FastQC
- **Alignment reference to Genome:** BWA back track
- **Post alignment :** SAMtools , Picard Markduplicates, GATK
- **Variant Calling** : GATK Hapotypecaller, bcftool
- **Variant filtering :** GATK Variant Filtering, bcftool
- **Variant Annotation:** SnpEff, ANNOVAR
- **Variant Evaluation:** hap.py, vacfeval, IGV
- **Analysis:** PLINK, CADD, MultiQC
- **Visualization :** IGV, R and python.

Thank you