Bioinformatics Capstone Project

Topic 1: Primer Design and Building an NGS Pipeline

Submitted by - Siya Singh (BSMS - IISER Tirupati)

Introduction

Next-generation sequencing (NGS) has become a core technology in genomics for the detection of genetic variants with high sensitivity and resolution. Accurate variant detection requires not only sequencing, but also careful experimental design, computational processing, and biological interpretation of results.

The KRAS gene is a well-characterized oncogene frequently mutated in human cancers, particularly within exon 2. Variants in this region are clinically relevant and commonly targeted in diagnostic sequencing assays.

This project aims to design primers targeting exon 2 of the human KRAS gene and to construct a complete NGS analysis pipeline, including read simulation, quality control, alignment, variant calling, and visualization. The goal is to demonstrate an end-to-end workflow for targeted NGS analysis and to interpret detected variants at both the nucleotide and protein levels.

Objectives

The specific objectives of this project were:

1. To select and extract the exonic region of interest (KRAS exon 2) from the human genome.
2. To design sequencing primers capable of amplifying the target region.
3. To simulate paired-end NGS reads from the designed amplicon.
4. To perform read quality assessment and alignment to a reference sequence.
5. To identify sequence variants using variant calling tools.
6. To validate detected variants through visualization and codon-level interpretation.

Workflow

The analysis workflow followed a standard targeted NGS pipeline consisting of four major stages:

Data Collection → Analysis → Visualization → Interpretation

Data collection involved the extraction of target genomic regions from Ensembl. Analysis included primer design, sequencing read simulation, alignment, and variant calling. Visualization was performed using quality control and genome browser tools, followed by biological interpretation of detected variants.

Tools Used

Category	**Tool**
Genome Data	Ensembl
Primer Design	Primer3
Read Simulation	ART
Quality Control	FastQC
Alignment	BWA-MEM
Alignment Processing	SAMtools
Variant Calling	bcftools
Visualization	IGV

Data Collection

The canonical KRAS transcript (ENST00000256078; KRAS-201) was selected from Ensembl. Protein-coding exons 2, 3, and 4 were identified as clinically relevant regions due to their known oncogenic hotspot mutations. For this project, exon 2 was selected as the primary target region, as it contains some of the most frequently mutated sites in KRAS

Exon	**Chromosome**	**Start (hg38)**	**End (hg38)**	**Length (bp)**
2	chr12	25,245,395	25,245,274	122

Initial primer design using exon-only sequence was unsuccessful due to insufficient amplicon length and reduced primer stability at exon boundaries. To address this limitation, intronic flanking regions were incorporated, enabling robust primer placement and generation of a

sequencing-compatible amplicon. Primer design was performed using Primer3 with standard NGS-compatible parameters.

Primer Design Results

Parameter	*Left Primer*	*Right Primer*
Sequence (5′→3′)	GATACACGTCTGCAGTCAACT	TCAGTCATTTTCAGCAGGCCT
Length (bp)	21	21
Melting Temperature (°C)	58.05	59.93
GC Content (%)	47.62	47.62
Hairpin / Dimer	Not detected	Not detected

NGS Data Simulation & Quality Control

As experimentally generated sequencing data were not available, paired-end Illumina reads (2 × 150 bp) were simulated using the ART read simulator at approximately 100× coverage. The input template consisted of a genomic fragment encompassing KRAS exon 2 and its intronic flanks, reflecting a realistic targeted amplicon sequencing strategy.

Quality assessment using FastQC indicated high per-base quality scores. Deviations in GC content and overrepresented sequences were observed, consistent with PCR-based targeted sequencing, where all reads originate from a fixed genomic locus.

Link for Reports:

https://drive.google.com/drive/folders/1oiXt_MaII5m_CI4DGEuQApZ59XQasD7c?usp=sharing

Read Alignment

Simulated paired-end reads were aligned to the KRAS exon 2 reference sequence using BWA-MEM. Alignment processing, including sorting and indexing, was performed using SAMtools. Alignment statistics showed that 100% of reads were successfully mapped and

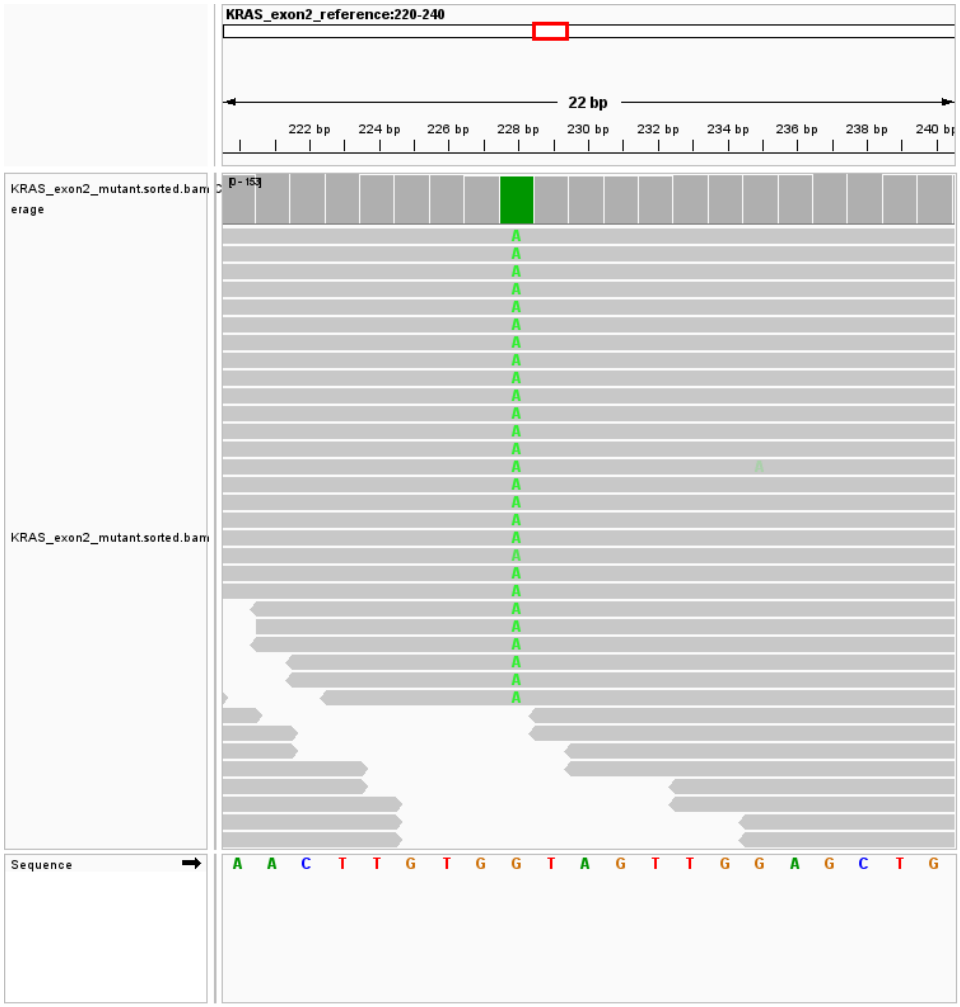
properly paired, with no unmapped reads or duplicates detected, indicating high-quality alignment and specificity of the designed primers.

Variant Calling & Annotation

Variant calling was performed using bcftools mpileup followed by bcftools call. A single high-confidence single-nucleotide variant was identified at position 228 of the amplicon. The reference allele (G) was replaced by an alternate allele (A), supported by 143 reads. The variant was called as homozygous alternate (GT = 1/1) with a Phred-scaled quality score of 225.4, indicating high confidence.

Visualization

The detected variant was visually validated using the Integrative Genomics Viewer (IGV). All aligned reads at position 228 supported the alternate allele, with no evidence of reference allele contamination or strand bias. This visual confirmation corroborated the variant calling results and demonstrated the reliability of the analysis pipeline.



Interpretation

Codon-level analysis revealed that the G→A substitution converts the wild-type codon **TGG** (encoding tryptophan) to **TGA**, a stop codon. This results in a **nonsense mutation**, leading to premature truncation of the KRAS protein.

While activating missense mutations in *KRAS* exon 2 are commonly associated with cancer, truncating mutations are typically not oncogenic and are often filtered out in clinical pipelines. The detection of this mutation in the present study reflects the simulated nature of the dataset and demonstrates the pipeline's ability to accurately identify and annotate coding variants.

Outcomes

- Successful design of NGS primers targeting KRAS exon 2
- Generation of simulated targeted sequencing data
- High-quality read alignment and variant calling
- Visual validation of variants using IGV
- Codon-level annotation and biological interpretation

Scope & Future Work

Future extensions of this work may include primer design for additional *KRAS* exons (exons 3 and 4), incorporation of real sequencing datasets, integration of automated variant annotation tools such as Ensembl VEP, and application of clinical filtering strategies to prioritise actionable mutations.

References

(Ensembl, Primer3, ART, BWA, SAMtools, bcftools, IGV — all standard.)