Developing an Automated Computational Genomics Pipeline for Breast Cancer Detection Using NGS

Abstract Id : 370

International Conference on
Recent Trends in Materials Science & Devices 2025 (ICRTMD-2025)
24-26 March, 2025

Dilip Sisodia[a,b*], Virendra Kumar Sharma[b*], Rohit Joshi[a], Hemant Arya[c], and Tarun Kumar Bhatt[c]

[a]Engineering College, Ajmer, Rajasthan-305025, India

[b]Bhagwant University, Ajmer, Rajasthan-305004, India

[c]Department of Biotechnology, Central University of Rajasthan, Bandarsindri, Ajmer-305817, India

Introduction

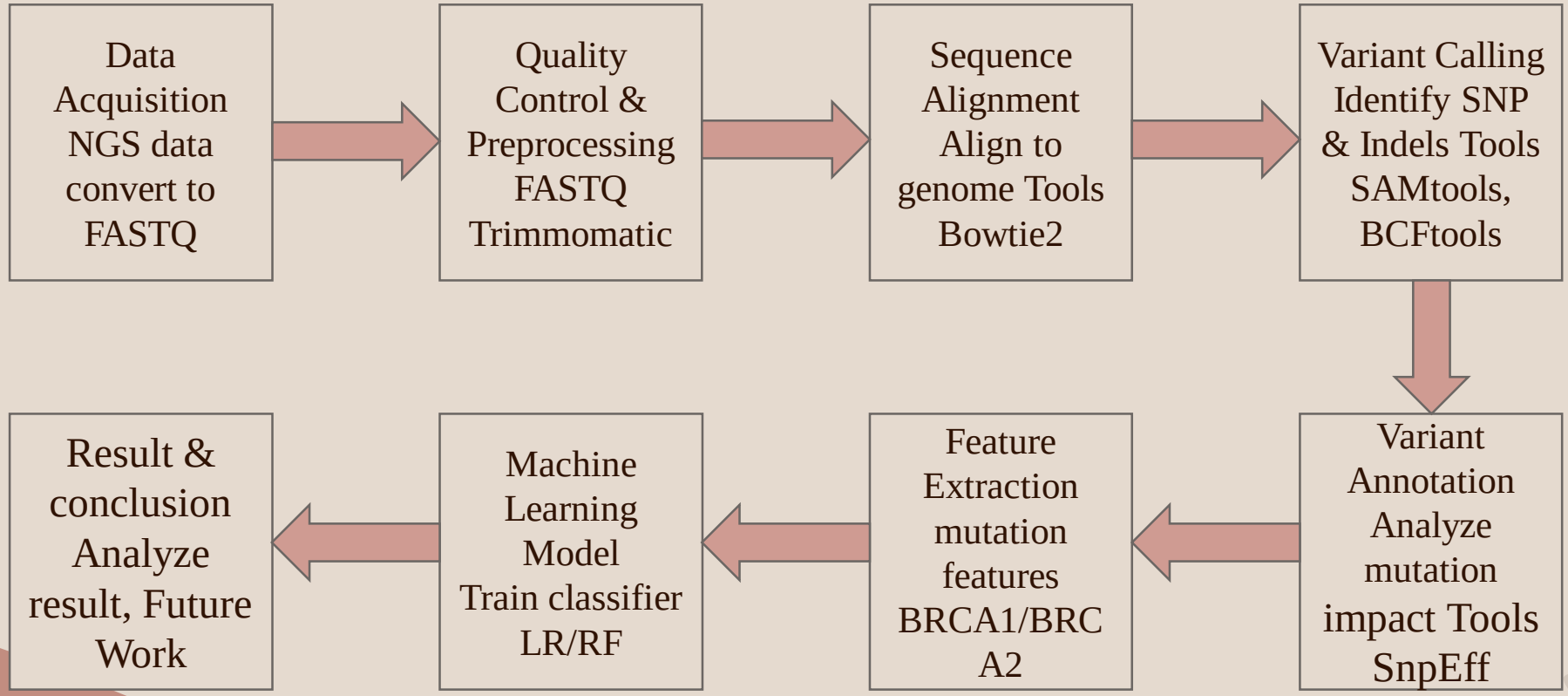
Breast cancer is one of the most common and life-threatening diseases affecting millions of people worldwide. It occurs due to uncontrolled cell growth in the breast tissue, often caused by genetic mutations. According to the World Health Organization (WHO), in 2020, approximately 2.3 million new cases and 685,000 deaths were reported due to breast cancer.

In recent years, Next-Generation Sequencing (NGS) has emerged as a powerful technique to analyze genetic mutations linked to various diseases, including breast cancer. This project aims to develop an automated computational genomics Pipeline that can analyze NGS data and identify genetics mutations, in BRCA1 and BRCA2 genes.

Objective of the Study

- Process NGS data from breast Cancer patients and healthy individuals.
- Identify genetic mutations in BRCA1 and BRCA2 genes.
- Analyze the correlation between mutations and breast cancer risk.
- Using Machine learning models to classify patients based on genetics variations.
- Support personalized medicine by enabling early detection and risk assessment.

Proposed Methodology



Preprocessing Steps

- SRA Toolkit - Downloads sequencing data from NCBI SRA.
- FASTQC - Checks base Sequence quality, Per Sequence Quality Scores
- Trimmomatic – Removes low-quality reads and adapter sequences.

```
if __name__ == "__main__":
    sra_ids = ["SRR24518778"]  # Add your SRA IDs here
    main_download(sra_ids)
```

```
Installing SRA Toolkit...
SRA Toolkit installed!
SRA Toolkit verified:
/content/sratoolkit.3.0.1-ubuntu64/bin/fastq-dump : 3.0.1


Downloading SRR24518778...
SRA data downloaded successfully! FASTQ files: SRR24518778_1.fastq.gz, SRR24518778_2.fastq.gz
```

Figure 1: Downloading sequencing Data

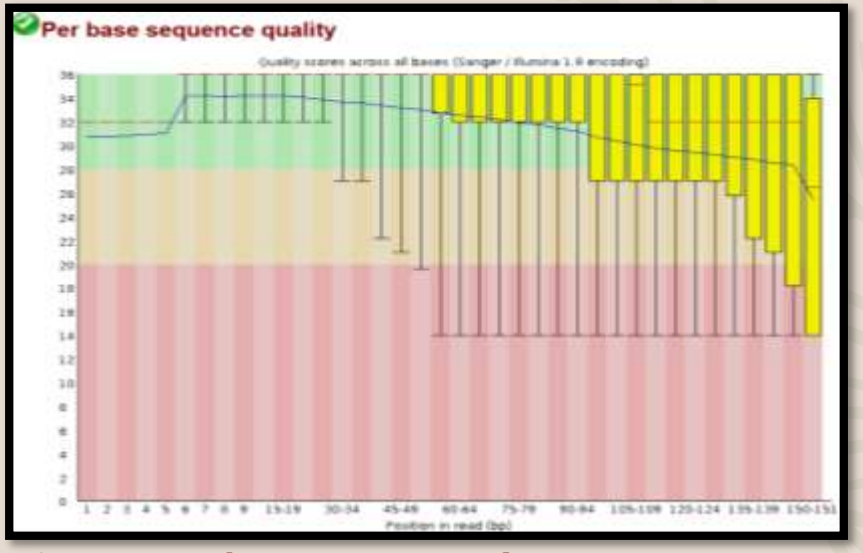


Figure 2: Per base sequence quality

Key Outcome: Clean, high-quality data for accurate mutation detection

Mutation Detection & Annotation

Genomic Basis of BRCA1 & BRCA2:

- ❑ BRCA1 gene is located on chromosome 17.
- ❑ BRCA2 gene is located on chromosome 13.

Mutations in these genes significantly increase breast cancer risk.

Pipeline for Mutation Detection:

1. **Bowtie2** – Aligns sequencing reads to the human genome.
2. **SAMtools & BCFtools** – Identifies genetic mutations.
3. **SnpEff** – Analyzes mutations and their impact.

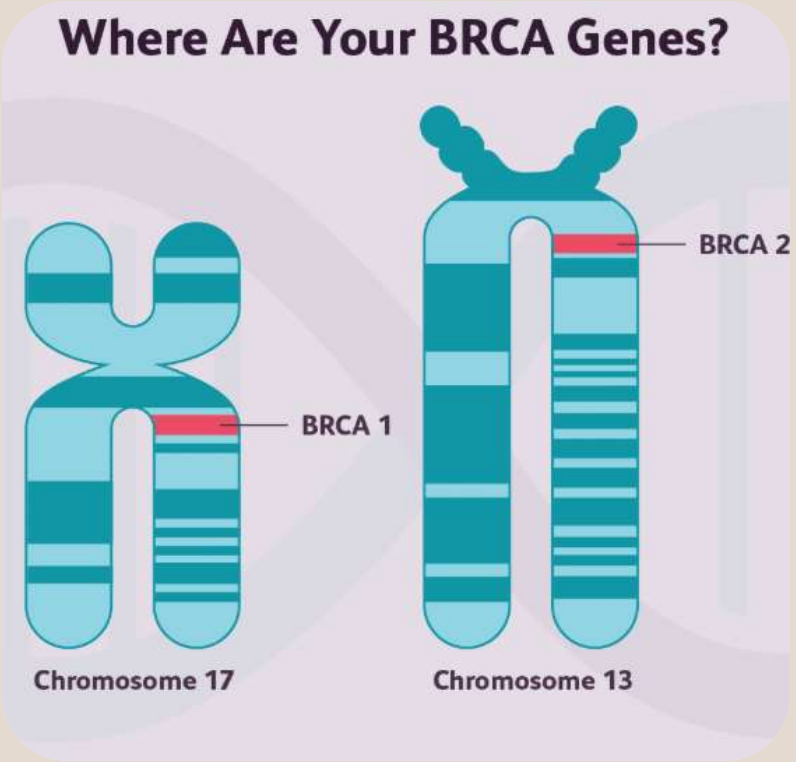


Figure.3: BRCA1 & BRCA2 Gene Locations

Key Outcome:
Efficient detection of genetic mutations in BRCA1 & BRCA2, aiding in early breast cancer diagnosis.

Machine Learning-Based classification

Feature Extraction from Genetic Variants:

- ❑ Extracted key genetic mutation features (BRCA1, BRCA2, missense, synonymous, intron, splice site mutations).

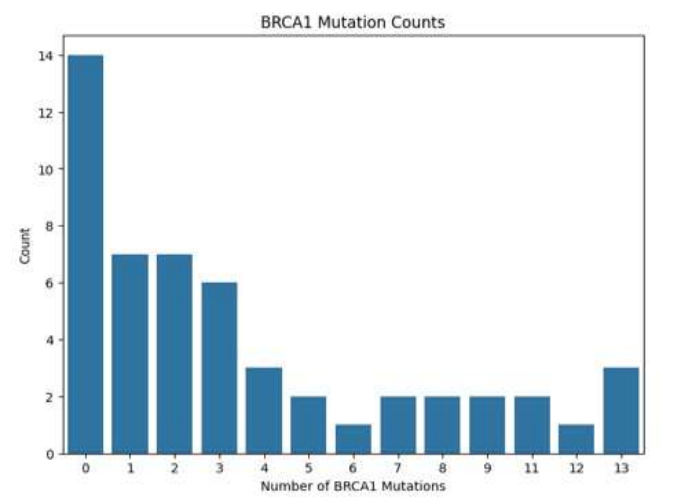


Figure 4: BRCA1 Mutation Count

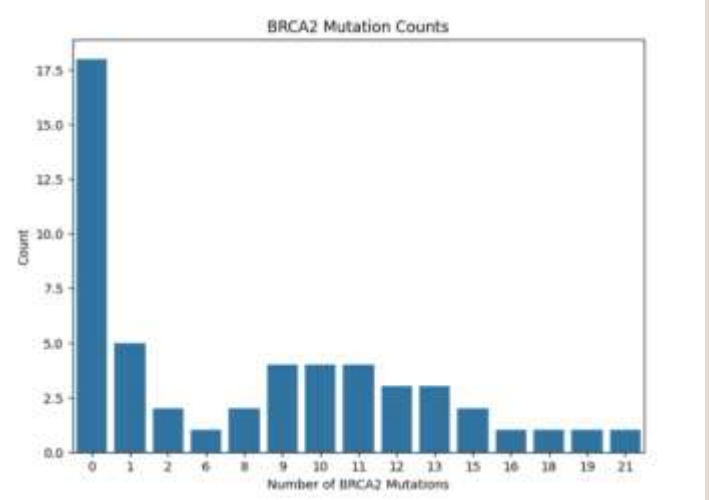


Figure 5: BRCA2 Mutation Count

Created a structured dataset for classification.

- ❑ Algorithms Used:

1. Random Forest – Robust for feature importance analysis.
2. Logistic Regression – Useful for binary classification (Cancer vs. Non-Cancer)

Result & Observations

- Genomic Variant Analysis (SRR24518792 Sample)
  Total Variants Detected: 389
  BRCA1 Variants (Chromosome 17): 13
  BRCA2 Variants (Chromosome 13): 9

- Pipeline Performance:

•Efficient mutation detection in BRCA1 & BRCA2 genes.

•Successfully aligned and annotated sequencing reads.

•Variants identified provide insights into breast cancer susceptibility.

```
Processing SRR24518792...
Checking tools...
All tools already installed!
Indexing Homo_sapiens.GRCh38.dna.chromosome.17.fa...
GRCh38_chr17 already indexed!
Indexing Homo_sapiens.GRCh38.dna.chromosome.13.fa...
GRCh38_chr13 already indexed!
Running FastQC on /content/drive/MyDrive/Cancer_Samples/SRR24518792_1.fastq.gz...
Running Trimmomatic...
Aligning reads to GRCh38_chr17...
Converting SRR24518792_chr17.sam to BAM...
Calling variants for SRR24518792_chr17_sorted.bam...
Annotating SRR24518792_chr17.vcf...
Aligning reads to GRCh38_chr13...
Converting SRR24518792_chr13.sam to BAM...
Calling variants for SRR24518792_chr13_sorted.bam...
Annotating SRR24518792_chr13.vcf...
Summarizing SRR24518792_chr17_annotated.vcf and SRR24518792_chr13_annotated.vcf...
Total variants: 389, BRCA1: 13, BRCA2: 9
```

Figure 6: Variant Analysis for Sample SRR24518792

Key Takeaway:
The automated pipeline effectively detects clinically relevant mutations, supporting genomic-based breast cancer diagnostics.

Conclusion & Future Scope

Conclusion:

- ❑ Strong correlation found between BRCA1/BRCA2 mutations and breast cancer risk.
- ❑ Developed an automated, efficient computational pipeline for genetic mutation detection.
- ❑ Supports early diagnosis and personalized treatment strategies.

Future Scope:

- ❑ Expanding dataset for higher accuracy.
- ❑ Integrating deep learning models for improved mutation classification.
- ❑ Exploring additional genetic markers beyond BRCA1/BRCA2.

References

1. Speiser, Dorothee, and Ulrich Bick. "Primary prevention and early detection of hereditary breast cancer." *Breast Care* 18, no. 6 (2023): 450-456.
2. A. C. Antoniou et al., "BRCA1 and BRCA2 mutations and breast cancer risk," *The American Journal of Human Genetics*, vol. 72, no. 5, pp. 1117-1130, 2003.
3. Parmigiani, Giovanni, Donald A. Berry, and Omar Aguilar. "Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2." *The American Journal of Human Genetics* 62, no. 1 (1998): 145-158.
4. Antoniou, Antonis C., A. P. Cunningham, J. Peto, D. G. Evans, F. Lalloo, S. A. Narod, H. A. Risch et al. "The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions." *British journal of cancer* 98, no. 8 (2008): 1457-1466.
5. Mullis, K., & Faloona, F. (1987). "Specific synthesis of DNA in vitro via a polymerase-chain reaction." *Methods in Enzymology*, 155, 335-350.
6. Metzker, Michael L. "Sequencing technologies—the next generation." *Nature reviews genetics* 11, no. 1 (2010): 31-46
7. Cancer Genome Atlas Network. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012, 490, 61–70.
8. Kleinbaum, D. G., & Klein, M. *Logistic Regression: A Self-Learning Text.* Springer, 2010.
9. Breiman, L. *Random forests.* Machine Learning, 2001, 45(1), 5–32.

Thank You!

Any Queries…..?