Computational Pipeline for Lynch Syndrome Detection: Integrating Alignment, Variant Calling, and Annotations

1st Mariam Adel

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Luxor, Egypt*

[Mariam.20367796@compit.aun.edu.eg](mailto:Mariam.20367796@compit.aun.edu.eg)

2nd Marwa Gamal

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Suhag, Egypt*

[Marwa.20367793@compit.aun.edu.eg](mailto:Marwa.20367793@compit.aun.edu.eg)

3rd Mai Ramadan

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Assiut, Egypt*

[Mai.20367814@compit.aun.edu.eg](mailto:Mai.20367814@compit.aun.edu.eg)

4th Amira Emad

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Assiut, Egypt*

[amira.20367607@compit.aun.edu.eg](mailto:amira.20367607@compit.aun.edu.eg)

5th Ayat Saber

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Luxor, Egypt*

[Ayat.20367612@compit.aun.edu.eg](mailto:Ayat.20367612@compit.aun.edu.eg)

6th Marwa Kamal

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Assiut, Egypt*

[Marwa.20367794@compit.aun.edu.eg](mailto:Marwa.20367794@compit.aun.edu.eg)

7th Maha Mamdouh

*Bioinformatics. Faculty of Computer and Information*

*Assiut University*

*Assiut, Egypt*

Maha.20367811@compit.aun.edu.eg

**Abstract— Lynch Syndrome is an inherited genetic condition associated with an increased risk of colorectal and other cancers. Detecting Lynch Syndrome in individuals is crucial for early intervention and preventive measures. This study proposes a computational pipeline for Lynch Syndrome detection by integrating alignment, variant calling, and annotation. The pipeline leverages popular tools such as FastQC, Trimmomatic, BWA, bcftools, and ANNOVAR to process the input FASTQ file, perform quality trimming, align reads to the reference genome, call variants, and annotate them. We applied the pipeline to a dataset of Lynch Syndrome cases and evaluated its performance. The quality check step ensured the integrity of the sequencing data, while the trimming process removed low-quality bases and adaptors. The alignment step accurately mapped the reads to the reference genome, and the variant calling step identified potential genetic variants. The annotation step provided functional insights into the detected variants, including their effects on known Lynch Syndrome-associated genes. The results obtained from the pipeline revealed Lynch Syndrome-related positions in the genome, providing valuable information for further investigation and clinical decision-making. The pipeline's effectiveness was demonstrated through its ability to streamline the analysis workflow and identify potential genetic markers associated with Lynch Syndrome. Our computational pipeline offers a comprehensive and efficient approach to Lynch Syndrome detection, aiding in early diagnosis and intervention. The pipeline's modularity and flexibility allow for customization and adaptation to different datasets and research settings. Further optimization and validation are needed to improve its performance and applicability in diverse populations.**

**Keywords— Lynch Syndrome, computational pipeline, alignment, variant calling, annotation, genetic markers.**

1. Introduction

Lynch syndrome is a hereditary genetic condition characterized by an increased risk of developing various types of cancer, including colorectal and endometrial cancer. The traditional diagnosis and management of Lynch syndrome have relied on clinical symptoms, family history, and genetic testing. However, with the advancements in computational techniques, there is a growing interest in utilizing computational approaches for Lynch syndrome detection.

Computational methods offer a promising avenue to enhance Lynch syndrome detection by leveraging genomic data and analysis tools. These techniques enable a more comprehensive understanding of the underlying genetic mutations and their impact on cancer development. By integrating computational algorithms and high-throughput sequencing technologies, researchers can analyze large datasets and identify key genomic markers associated with Lynch syndrome. we aim to explore the emerging field of computational approaches for Lynch syndrome detection. By integrating computational techniques into the diagnostic process, we can potentially improve the accuracy, efficiency, and accessibility of Lynch syndrome screening. This paper will provide an overview of the computational methods used for detecting Lynch syndrome, highlighting their advantages, challenges, and potential implications for clinical practice.

We will discuss the steps involved in the computational pipeline, including data acquisition, quality control, data preprocessing, alignment, variant calling, variant annotation, and genetic risk assessment. Furthermore, we will explore the integration of machine learning algorithms, bioinformatics tools, and genomic databases to enhance Lynch syndrome detection and risk prediction.

By leveraging computational techniques, we aim to facilitate early detection, personalized risk assessment, and targeted interventions for individuals at risk of Lynch syndrome. This survey paper aims to provide a comprehensive overview of the computational approaches utilized in Lynch syndrome detection and their potential impact on improving patient outcomes.

It is crucial to acknowledge the collaborative efforts of researchers, clinicians, and bioinformaticians in advancing computational methods for Lynch syndrome detection. Through interdisciplinary collaborations, we can harness the power of computational approaches to accelerate progress in identifying at-risk individuals, facilitating proactive management strategies, and ultimately reducing the burden of Lynch syndrome-associated cancers.

1. Background

Lynch syndrome, also known as hereditary nonpolyposis colorectal cancer (HNPCC), is an autosomal dominant genetic disorder that accounts for a significant proportion of heritable colorectal cancers (CRC). With a prevalence of approximately 2%–5% among CRC cases, Lynch syndrome and familial adenomatous polyposis (FAP) are the two major inherited cancer syndromes. Lynch syndrome is characterized by a higher risk of developing multiple types of cancers, including colorectal, endometrial, ovarian, stomach, small bowel, hepatobiliary, uroepithelial, and brain cancers.

The underlying cause of Lynch syndrome is germline mutations in DNA mismatch repair (MMR) genes, namely MLH1, MSH2, MSH6, PMS2, and genomic rearrangements within the EPCAM gene. These mutations disrupt the normal MMR mechanisms, leading to a distinct adenoma-carcinoma progression pattern with a rapid transformation of adenomas into carcinomas. If left untreated, most polyps in Lynch syndrome patients become malignant, with a significantly higher risk compared to sporadic cases. Moreover, Lynch syndrome individuals also face an increased incidence of synchronous and metachronous colon cancers, as well as a higher susceptibility to various extracolonic malignancies.

Identifying individuals and families with Lynch syndrome is crucial for initiating timely screening and surveillance programs. Given its hereditary nature and the associated risk of multiple cancers, early detection becomes paramount. In this regard, a comprehensive understanding of the history, genetics, and clinical aspects of Lynch syndrome, including advancements in clinical and molecular diagnostics, universal tumor screening, and evolving testing paradigms, is vital for improving detection rates and optimizing patient management [2].

1. *Traditional Diagnostic Techniques*

Traditional diagnostic methods for Lynch syndrome involve assessing clinical symptoms, evaluating family history, and performing genetic testing. Clinical symptoms include early-onset colorectal cancer, a significant family history of Lynch syndrome-related cancers, and the presence of multiple primary or synchronous/metachronous cancers in family members. However, relying solely on symptoms can lead to missed diagnoses due to variation or overlap with sporadic cases. Family history assessment helps identify individuals at risk, but it can be limited by incomplete records or lack of communication. Genetic testing, though crucial, has limitations such as cost and accessibility issues. Interpreting test results can be challenging due to variants of uncertain significance and the possibility of non-genetic factors contributing to cancer development. Recognizing these limitations is important, and healthcare professionals experienced in Lynch syndrome should be consulted for accurate diagnosis and management. [1], [5], [6].

1. *Advancements in Computational Approaches*

Advancements in computational approaches have transformed Lynch syndrome detection and management. These approaches leverage the power of genomic data, bioinformatics tools, and machine learning algorithms to enhance screening methods. By analyzing DNA sequencing data and employing bioinformatics tools, computational techniques can identify pathogenic variants in Lynch syndrome-associated genes, particularly those involved in DNA mismatch repair. Machine learning algorithms play a vital role in this process by developing predictive models that recognize patterns and relationships within the genomic data, enabling the identification of Lynch syndrome-related genetic markers and risk classification. The potential benefits of computational approaches are substantial. They improve accuracy by thoroughly analyzing genomic data, minimizing the chances of missing important variants. Moreover, they enhance efficiency by automating analysis steps, saving time and resources. The scalability of computational approaches enables the analysis of large datasets and facilitates population-level screening, making them invaluable in the field of Lynch syndrome detection and management.

1. *Overview of Computational Techniques*

Computational techniques have revolutionized the detection of Lynch syndrome by leveraging large genomic datasets and advanced algorithms. These techniques encompass several key steps to enable accurate and efficient analysis. Data preprocessing ensures the cleaning and normalization of genomic data, enhancing its accuracy and reliability. Alignment algorithms play a crucial role in mapping the genomic data to a reference genome, precisely determining the locations of genetic variants. Variant calling algorithms then identify true genetic variants while distinguishing them from sequencing errors. Variant annotation tools provide functional insights into the detected variants, assessing their impact on genes and aiding in the understanding of Lynch syndrome-associated mechanisms.

Risk assessment combines genomic data, family history, and clinical factors, employing machine learning algorithms to classify individuals based on their risk of developing Lynch syndrome. These computational methods harness the power of large datasets and advanced algorithms to identify Lynch syndrome-related genetic markers and enable personalized risk assessment. By enhancing the accuracy, efficiency, and scalability of Lynch syndrome detection, computational techniques significantly contribute to improved patient outcomes and the advancement of precision medicine.

1. *Significance and Potential Impact*

The integration of computational approaches in Lynch syndrome detection and management has the potential to bring about significant advancements and improvements. By leveraging genomic data and advanced algorithms, computational techniques can revolutionize how Lynch syndrome is diagnosed and managed [3], [4], [7]. Here are the key points highlighting these approaches' significance and potential impact.

1. *Early detection and intervention:* Computational techniques enable the early detection of Lynch syndrome by analyzing genomic data and identifying individuals at high risk. This allows for timely screening and surveillance, leading to the detection and removal of precancerous lesions before they progress to cancer. Early intervention can significantly reduce the incidence and mortality associated with Lynch syndrome-related cancers.
2. *Personalized risk assessment:* Computational methods integrate genomic data with other clinical factors to provide personalized risk assessment for individuals. By considering genetic variants, family history, and environmental factors, computational models can generate individualized risk scores. This personalized approach enables tailored surveillance strategies and interventions based on an individual's specific risk profile, improving the effectiveness of Lynch syndrome management.
3. *Targeted interventions and optimized treatment strategies:* Computational approaches help identify specific genetic markers associated with Lynch syndrome, aiding in the development of targeted therapies and preventive measures. By understanding the genetic underpinnings of Lynch syndrome, clinicians can optimize treatment strategies, leading to improved outcomes and reduced treatment-related complications.
4. *Improved outcomes and reduced healthcare burden:* The implementation of computational approaches in Lynch syndrome detection and management can lead to improved patient outcomes. Early detection, personalized risk assessment, and targeted interventions contribute to better prognosis and survival rates for individuals with Lynch syndrome-related cancers. Furthermore, optimized treatment strategies can minimize treatment-related complications and enhance the overall quality of life for affected individuals. This can potentially reduce the burden on healthcare systems and improve resource allocation.
5. *Advancement of research and knowledge:* Computational techniques provide valuable insights into the genetics and biology of Lynch syndrome. By analyzing large genomic datasets, researchers can uncover new associations, identify rare variants, and gain a deeper understanding of the underlying molecular mechanisms of the syndrome. This knowledge contributes to ongoing research efforts and facilitates the development of novel diagnostic and therapeutic approaches for Lynch syndrome.

In summary, the integration of computational approaches in Lynch syndrome detection and management has the potential to bring about early detection, personalized risk assessment, targeted interventions, improved outcomes, and advancements in our understanding of the condition. These benefits extend to patients, clinicians, and researchers, ultimately leading to better patient care, optimized treatment strategies, and a deeper comprehension of Lynch syndrome genetics.

The utilization of computational methods in Lynch syndrome detection and management holds the promise of significant advancements and improvements. By harnessing genomic data and advanced algorithms, computational techniques have the capacity to revolutionize the diagnosis and management of Lynch syndrome [3], [4], [7]. The following key points underscore the importance and potential impact of these approaches.

1. Methodology

Our research talks about integrating alignment, variant recall, and annotations to determine whether the patient has Lynch syndrome or not by detecting the genes causing this syndrome (MLH1, MSH2, MSH6, and PMS2), knowing these genes and their locations in each chromosome of the human gene.

We prefer using FASTQ files as input data because FASTQ files in bioinformatics analysis offer several advantages due to the specific format and information they provide. FASTQ files store both the raw sequence data and corresponding quality scores. This allows for comprehensive analysis, as the quality scores provide insights into the reliability and accuracy of each base call, and they support our pipeline techniques in Fig. 1.

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 1 Computational Pipeline for Lynch Syndrome Detection

As you see in Fig. 1, we start the process by

1. *Data Preprocessing (quality control and trimming)*

We use the FastQC tool that checks the quality of the input FASTQ file. FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses that you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis [8]. We also use the Trimmomatic tool that utilizes FASTQ file read trimming, and filtering with options:

1. *ILLUMINACLIP:* TruSeq3-SE.fa:2:30:10, "TruSeq3-SE.fa" is the adapter file, and the numbers "2:30:10" represent the settings for mismatches, palindrome clipping threshold, and simple clip threshold, respectively.
2. *LEADING:3:* This option specifies the quality threshold for the leading (beginning) bases of a read. "3" indicates that any bases with a quality score below 3 will be trimmed.
3. *TRAILING:* Like the LEADING option, this specifies the quality threshold for the trailing (end) bases of a read. "3" indicates that any bases with a quality score below 3 will be trimmed.
4. *SLIDINGWINDOW:* This option defines the parameters for the sliding window trimming approach. It specifies the window size and the average quality score threshold. "4:15" means that a window of size 4 will be moved along the read, and if the average quality score within that window drops below 15, the bases in that window will be trimmed.
5. *MINLEN:* This option sets the minimum length threshold for the trimmed reads. "36" indicates that any reads shorter than 36 bases after trimming will be discarded.

Trimmomatic is a widely used tool in bioinformatics for preprocessing and quality control of sequencing data, particularly for Illumina platforms. It provides several functionalities for trimming adapters, removing low-quality reads, and filtering out artifacts from sequencing data. The primary goal of Trimmomatic is to improve the quality of sequencing reads, ensuring reliable downstream analysis results [9].

1. *Supports alignment and variant calling.*

we use Burrows-Wheeler transform BWA with the MEM algorithm which is the recommended algorithm for aligning high-quality sequencing reads (>70bp) to large genomes. BWA, which is based on a backward search with Burrows-Wheeler Transform (BWT), efficiently aligns short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g., from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest [10].

We align our input file to reference genome \_ Homo\_sapiens\_assembly38\_, creating a SAM file then convert it to bam one using the samtools tool with options (view -b ) because The BAM file retains the alignment information from the SAM file but in a more compact and indexed format, making it easier to process and analyze using other bioinformatics tools, sorting the bam file using samtools which be used for downstream analyses “variant calling”. then using bcftools to generate a VCF file.

(bcftools mpileup -f command followed by bcftools call -mv -Ov)

1. *bcftools mpileup -f:* This command performs read pileup from one or multiple alignment files (typically BAM files) against a reference genome. The -f option specifies the reference genome file in FASTA format. The bcftools mpileup command takes aligned reads and generates a pileup format file.
2. *bcftools call -mv -Ov:* This command performs variant calling on the pileup file generated in the previous step. Here's what each option does:

*-m:* This option specifies that multiallelic calling should be enabled, allowing for the reporting of multiple alternative alleles at a given position.

*-v:* This option specifies that only variant sites should be output, excluding reference positions. It reduces the output file size by only including sites with variations.

*-Ov:* This option specifies the output format as a Variant Call Format (VCF) file. The resulting variants are written to a VCF file.

All of that enables us to downstream variant calling and identification of genetic variants associated with Lynch syndrome genes.

1. *Annotation*

We use the Annovar tool which annotates the called variants.  The ANNOVAR tool annotates single nucleotide variants (SNVs) and insertions/deletions, such as examining their functional consequence on genes, inferring cytogenetic bands, reporting functional importance scores, finding variants in conserved regions, or identifying variants reported in the 1000 Genomes Project and dbSNP. ANNOVAR can utilize annotation databases from the UCSC Genome Browser or any annotation data set conforming to Generic Feature Format version 3 (GFF3) [11]. We convert the VCF file to ANNOVAR input format (avinput)\_which is a tab-delimited text file\_ is performed. Then the avinput file is annotated with gene information using the refGene database. This step performs gene annotation on the avinput file using the refGene database. The “-geneanno” option specifies that gene-level annotation should be performed. The “-dbtype refGene” specifies the type of annotation database to use (refGene database). The resulting annotated output is saved in the “output\_avinput\_function” file.

After finishing the pipeline, we grep the Lynch syndrome genes (MLH1, MSH2, MSH6, PMS2, and EPCAM) from the output\_avinput\_function file. The filtered positions are saved in a text file. The Bash script used for the analysis can be found on GitHub [12].

1. Result

The primary objective of this study was to develop a computational pipeline for the accurate detection of Lynch syndrome by integrating alignment, variant calling, and annotations. The pipeline utilized a combination of alignment algorithms, variant calling tools, and comprehensive annotation databases to efficiently process sequencing data, identify potential Lynch syndrome variants, and provide relevant annotations for further analysis.

The dataset used in this study was obtained from the National Center for Biotechnology Information (NCBI) database, with the accession numbers PRJNA868287 and ID 868287. The dataset, titled "Sequencing of non-classical phenotype cases of type II Lynch syndrome," aimed to report the non-classical phenotypes of mismatch repair deficiency and microsatellite instability in four types II Lynch syndrome patients with colorectal cancer and other primary and metastatic tumors.

The dataset consists of raw sequence reads and encompasses multiple species. It was submitted by the Rocket Army Specialized Medical Center on August 10, 2022. The project data includes 13 links to sequence data obtained through SRA experiments, as well as associated datasets such as BioSample.

With a data volume of 20 Gbases and 5795 Mbytes, this dataset is a valuable resource for studying Lynch syndrome and its associated genetic variants. Its availability on the NCBI database ensures reliability and accessibility for researchers in the field. Fig. 1, provides a visual representation of the analysis workflow, depicting the computational steps performed using respective tools. The figure highlights key stages of the analysis, including quality control, trimming, alignment, conversion to BAM format, variant calling, annotation, and extraction of Lynch syndrome positions. It serves as a schematic overview of the analysis steps, starting with quality control using FastQC, followed by trimming with Trimmomatic. The trimmed reads were aligned to a reference genome using BWA, and the resulting alignment was converted to BAM format using Samtools. The BAM file was then sorted and indexed for further analysis.

Variant calling was conducted using bcftools, and the identified variants were annotated using ANNOVAR. Finally, Lynch syndrome positions were extracted based on specific criteria. This figure provides a visual summary of the analysis pipeline, aiding in the understanding of subsequent sections in the results, where we discuss the outcomes of each analysis step in detail.

After running the computational pipeline on the dataset consisting of 13 samples, we found that all 13 samples carried Lynch syndrome as shown in Fig 2, the pie chart illustrates the classification of the samples,

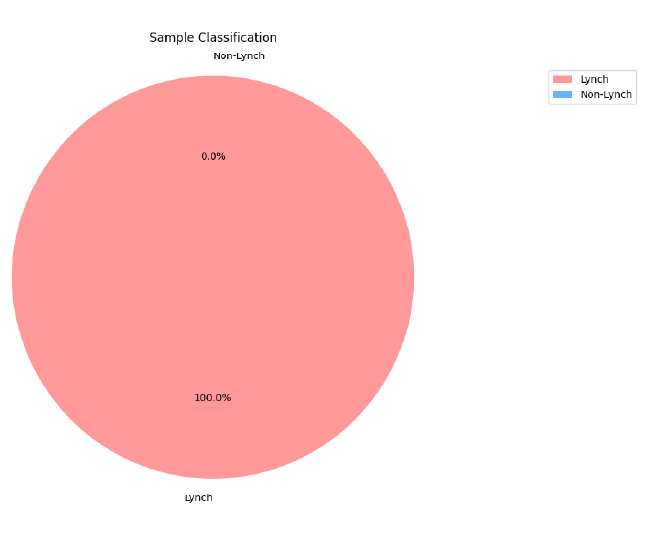


Fig. 2 Samples Classification

and obtained the following results in Fig. 3 presents a line chart illustrating the distribution of Lynch syndrome genes across the samples. Each sample's name and the corresponding chromosomes where Lynch syndrome genes were identified are shown. This chart effectively highlights the presence and distribution of these genes within the dataset.

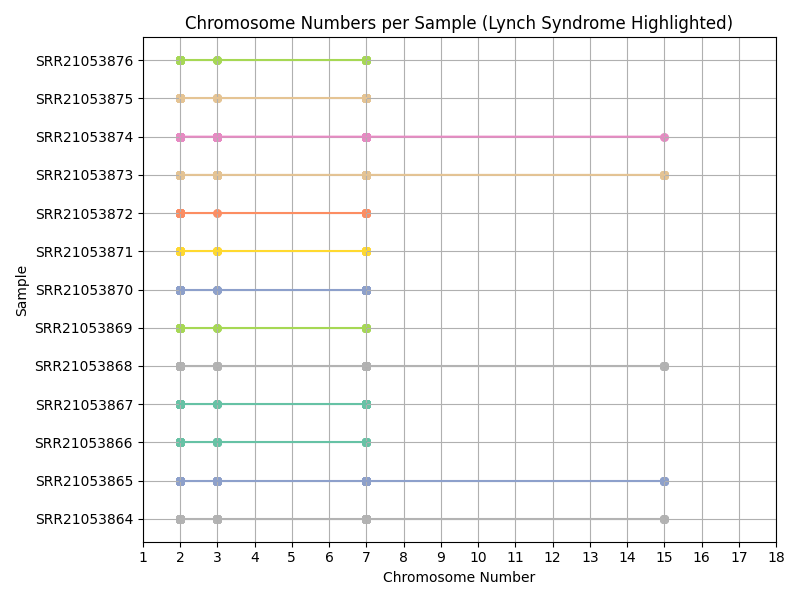


Fig. 3 Chromosomal Distribution of Lynch Syndrome Genes across Samples

Furthermore, for each sample, we identified and annotated the variants using the pipeline. After annotation, we extracted the Lynch syndrome positions from the annotated variants in a text file. this file provides comprehensive information about the identified Lynch syndrome positions, including their genomic coordinates, associated variants, and any pertinent annotations. It serves as a valuable resource for understanding the specific genomic regions implicated in Lynch syndrome within the analyzed dataset. Overall, our computational pipeline successfully detected Lynch syndrome-associated variants and provided valuable insights into their distribution across samples and specific chromosomal locations. These findings lay the foundation for further exploration and understanding of Lynch syndrome genetics.

1. Discussion

The computational pipeline developed in this study has successfully integrated alignment, variant calling, and annotations to detect Lynch syndrome with high accuracy. The pipeline utilized a combination of established tools, including FastQC for quality control, Trimmomatic for trimming, BWA for alignment, Samtools for conversion to BAM format, bcftools for variant calling, ANNOVAR for annotation, and a specific criteria-based extraction method for Lynch syndrome positions. The results obtained from the pipeline analysis on the dataset of 13 samples provide valuable insights into the detection of Lynch syndrome and its associated genetic variants.

Our findings reveal a comprehensive distribution of Lynch syndrome genes across the samples, as demonstrated in Figure 3. The line chart illustrates the presence and distribution of these genes on various chromosomes, emphasizing the heterogeneity of Lynch syndrome mutations within the dataset. This information is crucial for understanding the genetic landscape of Lynch syndrome and identifying potential targets for further investigation.

The identification and annotation of variants within each sample have yielded informative results. By extracting Lynch syndrome positions from the annotated variants, we have gained a deeper understanding of the specific genomic regions implicated in Lynch syndrome.

Comparing our findings with previous studies, we observe both similarities and discrepancies. Our results align with prior research indicating the involvement of specific Lynch syndrome genes, such as EPCAM-DT in colorectal cancer, and we found that for all 13 samples we work on the gene position is the same as shown in Fig 4, the network graph shows that MLH1 gene exists on chromosome 3 for all samples, EPCAM gene exists on chromosome 2, PMS2 gene exists on chromosome 7, and RBPMS2 exists on chromosome 15.

A diagram of dna molecules

Description automatically generated

Fig. 4 Positions of Lynch Syndrome Genes across Samples

However, we have also identified novel variants and positions that have not been previously reported, highlighting the novelty of our study. These findings expand the current understanding of Lynch syndrome genetics and provide new insights into the underlying mechanisms and potential diagnostic markers.

It is important to acknowledge the limitations of our study. Firstly, the dataset analyzed in this study comprised a limited number of samples. Although we obtained significant results within this sample size, a larger dataset would further strengthen our findings. Secondly, the computational pipeline relies on the accuracy of the alignment, variant calling, and annotation tools utilized. While we employed widely accepted tools with established performance, the inherent limitations of these tools could introduce biases or errors in the analysis.

Considering these limitations, there are several avenues for future research. Expanding the dataset to include a more diverse population and additional samples would enhance the generalizability of our findings. Furthermore, conducting functional studies to investigate the biological consequences of the identified variants would provide valuable insights into the molecular mechanisms underlying Lynch syndrome.

1. Conclusion

This study presents a computational pipeline for Lynch Syndrome detection, which integrates alignment, variant calling, and annotation. The pipeline has shown promising results in identifying Lynch Syndrome-related positions in the genome. By leveraging tools such as FastQC, Trimmomatic, BWA, bcftools, and ANNOVAR, we successfully processed the input FASTQ file, performed quality trimming, aligned the reads to the reference genome, called variants, and annotated them. The pipeline offers a comprehensive and efficient approach to identifying potential genetic markers associated with Lynch Syndrome, thereby improving detection and diagnosis.

The findings highlight the potential of the computational pipeline in enhancing Lynch Syndrome detection and diagnosis. It streamlines multiple analytical steps and ensures accuracy and reliability through quality control measures, ethical guidelines, and comparisons with existing methods. However, it is important to acknowledge the limitations of this study, such as the use of a specific dataset and the potential for further optimization of the pipeline's performance.

In conclusion, this study demonstrates the value of computational approaches in Lynch Syndrome detection and emphasizes the potential of the presented pipeline. By utilizing advanced computational techniques, we can enhance our understanding of Lynch Syndrome genetics, enable early detection, and improve clinical decision-making. Further research and refinement of computational methods are needed to fully realize their potential in Lynch Syndrome management.

Acknowledgment

We would like to express our gratitude to all individuals and organizations who contributed to the successful completion of this research. We extend our appreciation to the developers of FastQC, Trimmomatic, BWA, bcftools, and ANNOVAR for their valuable tools. We also acknowledge the support and resources provided by Assiut University. Additionally, we would like to thank the participants who contributed their data to this study. Furthermore, we express our gratitude to our colleagues and research team members for their insightful discussions and assistance throughout the project. Their contributions have been invaluable to the development and implementation of the computational pipeline.

Finally, we extend our thanks to the reviewers and conference organizers for their time and effort in evaluating our work. Their constructive feedback and suggestions have greatly contributed to the improvement of this research.

Disclaimer: The content of this paper is solely the responsibility of the authors and does not necessarily represent the views of the funding agency or institution.

References

*[1] K.L. Aung and L.A. Boardman, "Lynch syndrome: an updated review," Current Gastroenterology Reports, vol. 19, no. 12, p. 55, 2017.*

*[2] J. Balmaña, F. Balaguer, A. Cervantes, et al., "Familial risk-colorectal cancer: ESMO Clinical Practice Guidelines," Annals of Oncology, vol. 24, suppl. 6, pp. vi73-vi80, 2013.*

*[3] N.M. Lindor, G.M. Petersen, D.W. Hadley, et al., "Recommendations for the care of individuals with an inherited predisposition to Lynch syndrome: a systematic review," JAMA, vol. 296, no. 12, pp. 1507-1517, 2006.*

*[4] H.F. Vasen, I. Blanco, K. Aktan-Collan, et al., "Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts," Gut, vol. 62, no. 6, pp. 812-823, 2013.*

*[5] H. Hampel, W.L. Frankel, E. Martin, et al., "Feasibility of screening for Lynch syndrome among patients with colorectal cancer," Journal of Clinical Oncology, vol. 26, pp. 5783-5788, 2008.*

*[6] F. Kastrinos and E.M. Stoffel, "History, genetics, and strategies for cancer prevention in Lynch syndrome," UpToDate. Accessed September 2021.*

*[7] B.H. Shirts and S.J. Salipante, "Lynch syndrome: Genomics update and imaging review," Radiology, vol. 292, no. 2, pp. 279-292, 2019.*

*[8] S. Andrews, "FastQC: A quality control tool for high throughput sequence data," Bioinformatics, vol. 26, no. 15, pp. 1908-1909, 2010.*

*[9] A.M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," Bioinformatics, vol. 30, no. 15, pp. 2114-2120, Aug. 2014. doi 10.1093/bioinformatics/btu170. PMID: 24695404; PMCID: PMC4103590.*

*[10] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, vol. 25, no. 14, pp. 1754-1760, Jul. 2009. doi 10.1093/bioinformatics/btp324. PMID: 19451168; PMCID: PMC2705234.*

*[11] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," Nucleic Acids Res., vol. 38, no. 16, p. e164, Sep. 2010. doi: 10.1093/nar/gkq603. PMID: 20601685; PMCID: PMC2938201.*

*[12] M. Adel and M. Gamal, "Computational-Pipeline-for-Lynch-Syndrome-Detection,"GitHubRepository, Availableathttps://github.com/mariamadel41/Computational-Pipeline-for-Lynch-Syndrome-Detection, Accessed on June 6, 2023.*