



**SHIROMANI GURUDWARA PARBHANDHAK COMMITTEE'S**  
**GURU NANAK KHALSA COLLEGE OF ARTS, SCIENCE & COMMERCE**  
(Autonomous)

(Affiliated to University of Mumbai and Recognized by Govt. of Maharashtra)

Accredited by NAAC with 'A' Grade with a CGPA of 3.54

---

**NGS – Precision Medicine Toolkit:  
A Comprehensive User – Friendly GUI for Next -  
Generation Sequencing (NGS) in the context of  
Precision Medicine**

(M. Sc. Bioinformatics – Semester III – Research Project Report)

**Research Project Report Submitted in Partial Fulfilment of The  
Requirements for the Degree of**

**MASTERS OF SCIENCE IN BIOINFORMATICS**

**SUBMITTED BY**

Ms. Prarthi Hrishit Kothari  
(Roll No.: 115)

**UNDER THE GUIDANCE OF**

Prof. Sermarani Nadar  
Dr. Gursimran Kaur Uppal

**SUBMITTED TO**

Department of Bioinformatics,  
Guru Nanak Khalsa College of Arts, Commerce & Science (Autonomous),  
Matunga (East), Mumbai – 400019



**SHIROMANI GURUDWARA PARBHANDHAK COMMITTEE'S**  
**GURU NANAK KHALSA COLLEGE OF ARTS, SCIENCE & COMMERCE**  
(Autonomous)

(Affiliated to University of Mumbai and Recognized by Govt. of Maharashtra)

Accredited by NAAC with 'A' Grade with a CGPA of 3.54

---

## **DEPARTMENT OF BIOINFORMATICS**

### **CERTIFICATE**

This is to certify that Ms. Prarthi Hrishit Kothari ( Roll No.: 115 ), student of M. Sc. Bioinformatics (Semester III), at the Department of Bioinformatics, Guru Nanak Khalsa College of Arts, Science & Commerce (Autonomous), Matunga has submitted the Research Project work titled, 'NGS – Precision Medicine Toolkit: A Comprehensive User – friendly GUI for Next – Generation Sequencing (NGS) in the context of Precision Medicine' for the fulfilment of the Master's degree in Bioinformatics, during the Academic year 2024-2025.

---

Professor – in – Charge  
(Prof. Sermarani Nadar)

---

Head of Department  
(Dr. Gursimaran Kaur Uppal)

---

Examiner

## **ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to all those who have supported me throughout the journey of my research project titled 'NGS – Precision Medicine Toolkit: A Comprehensive User – Friendly GUI for Next – Generation Sequencing (NGS) in the context of Precision Medicine.'

Firstly, I extend my sincere thanks to Dr. Ratna Sharma (The Principal, of Guru Nanak Khalsa College of Arts, Science & Commerce (Autonomous)) and Dr. Gursimran Kaur Uppal (Head of Department of Bioinformatics), for fostering an environment that encourages exploration and creativity. Your leadership and vision have created a space where students can thrive and pursue their academic interests with passion.

Next, I would like to acknowledge the unwavering support of my guide, Prof. Sermarani Nadar, whose invaluable guidance, expertise, and encouragement have been instrumental in shaping this project. Your insights into the complexities of the field have not only enriched my understanding but have also inspired me to explore innovative solutions. Your dedication to teaching and your willingness to share your knowledge have provided me with a solid foundation in both theoretical and practical aspects of this research. I would also like to extend my heartfelt gratitude to Prof. Aparna Patil Kose, whose knowledge and support were instrumental in my academic growth.

I am also grateful to my peers and colleagues who contributed their insights and feedback during various stages of this project. The collaborative spirit within our department has made this research experience enriching and enjoyable.

Lastly, I would like to thank my family for their unwavering support and encouragement throughout this journey. Your belief in my abilities has been a constant source of motivation. This project is a culmination of collective efforts, and I am deeply appreciative of everyone who played a role in its progress.

## **INDEX**

<b>Chapter No.</b>	<b>Title</b>	<b>Page No.</b>
1	Abstract	1
2	Objectives	3
3	Introduction	5
	I. Precision Medicine	6
	II. Next – Generation Sequencing (NGS)	7
	III. Graphical User Interface (GUI)	8
4	Review of Literature	11
	A. The Integration of Next – Generation Sequencing in Precision Medicine	12
	B. Next – Generation Sequencing (NGS) Tools	16
	I. SRA Toolkit	17
	II. FastQC	18
	III. FastP	21
	IV. SAM Tools	23
	V. Variant Calling	24
5	Methodology	26
6	Code	31
7	Results	34
8	Expected Outcome	38
9	Conclusion	40
10	References	43

# **CHAPTER 1: ABSTRACT**

## **ABSTRACT**

This report introduces the ‘NGS - Precision Medicine Toolkit’, a solution designed to streamline next-generation sequencing (NGS) workflows for precision medicine. The toolkit simplifies complex NGS processes, making them accessible to researchers and clinicians with varying levels of expertise through an intuitive graphical user interface (GUI) for efficient data management, visualization, and analysis. Key objectives include integrating essential NGS pipeline steps such as data acquisition, quality control, alignment, base recalibration, and variant calling into a cohesive platform. The toolkit incorporates bioinformatics tools like SRA Toolkit for data retrieval, FastQC and Fastp for quality assessment, and GATK for variant calling. The NGS pipeline emphasizes efficiency and accuracy, with automated processes to minimize user input and maximize throughput. Additionally, the toolkit connects to online databases via APIs, providing insights into genetic variants and therapeutic options. User testing and validation ensure optimal performance and usability, while comprehensive documentation supports users through installation and troubleshooting. In conclusion, the toolkit addresses gaps in current genomic analysis by offering a robust, user-friendly solution, significantly advancing precision medicine and personalized healthcare.

# **CHAPTER 2: OBJECTIVES**

## **OBJECTIVES**

### **1. Develop a User-Friendly GUI for Precision Medicine NGS Workflow**

The goal is to design an intuitive GUI that simplifies NGS workflows for precision medicine, enabling researchers and clinicians to easily navigate and manage processes without technical expertise. Key features include drag-and-drop functionality and accessibility across devices. The focus is on enhancing user experience and efficiency in handling NGS tasks.

### **2. Integrate Key NGS Workflow Steps into the GUI**

This objective involves integrating all major NGS workflow steps into the GUI, including quality control, alignment, variant calling, annotation. Users can monitor progress and access real-time results, centralizing the workflow in one platform to reduce software switching and minimize errors.

### **3. Enable NGS Workflow Customization**

The GUI will allow users to customize their NGS workflows based on specific project needs, enabling modifications like adding/removing steps, adjusting parameters, and integrating custom tools or third-party bioinformatics scripts. This flexibility supports diverse experimental designs and is crucial for advancing personalized precision medicine solutions.

### **4. Standardization and Quality Control Tools**

As bioinformatics tools proliferate, ensuring the quality and reproducibility of analyses becomes crucial. There is a need for standardized tools that can assess the quality of data and methods used in bioinformatics studies, helping to maintain high research standards.

Building a user interface for precision medicine-based NGS workflows is vital for enhancing user experience, streamlining data management, facilitating collaboration, improving data visualization, and integrating advanced technologies. By focusing on usability and accessibility, such interfaces can empower healthcare professionals to harness the full potential of precision medicine, ultimately leading to better patient care and outcomes.

---



# **CHAPTER 3: INTRODUCTION**

# **INTRODUCTION**

## **I. Precision Medicine**

Precision medicine, often referred to as personalized or individualized medicine, represents a transformative approach in healthcare that tailors medical treatment to the unique characteristics of each patient. This methodology integrates a comprehensive understanding of an individual's genomic, environmental, and lifestyle factors to inform medical decisions, aiming to enhance the prevention, diagnosis, and treatment of diseases. The overarching goal is to move away from the traditional 'one-size-fits-all' model of healthcare towards a more nuanced and effective strategy.

### **Key Components of Precision Medicine**

#### **1. Genetic Profiling:**

Genetic profiling involves conducting genetic tests to evaluate how variations in an individual's DNA can affect their response to treatments. For instance, in oncology, molecular testing can identify specific mutations within tumor cells, enabling healthcare providers to select therapies that are more likely to be effective for that particular cancer type. This approach not only optimizes treatment efficacy but also minimizes unnecessary side effects from less effective treatments.



#### **2. Environmental and Lifestyle Factors:**

This aspect considers how external factors such as diet, physical activity, and exposure to environmental toxins influence an individual's health outcomes. Understanding these variables allows for a more comprehensive risk assessment and the development of personalized prevention strategies. For example, dietary modifications can be recommended based on an individual's metabolic profile or genetic predispositions.

#### **3. Data Utilization:**

Advances in genomics and big data analytics play a crucial role in precision medicine. By analyzing large datasets that encompass genetic information, health records, and lifestyle choices, healthcare providers can develop tailored treatment plans that reflect the unique profiles of their patients. This data-driven approach enhances the ability to predict disease susceptibility and treatment responses, ultimately leading to improved health outcomes.

### **The Evolution of Precision Medicine**

The field of precision medicine has gained momentum due to significant advancements in genomic research and technology. The Human Genome Project (HGP), completed in 2003, was a landmark initiative that mapped all human genes, paving the way for understanding genetic contributions to health and disease. Following this project, there has been a surge in research focused on identifying biomarkers and developing targeted therapies that cater specifically to individual patient profiles.

In 2015, President Barack Obama announced the Precision Medicine Initiative, which aimed to accelerate research in this area with substantial federal funding. This initiative highlighted the potential for precision medicine to revolutionize patient care by ensuring that treatments are tailored not only to the disease but also to the individual characteristics of each patient.

## II. Next-Generation Sequencing (NGS)

Next-Generation Sequencing (NGS) is a groundbreaking technology that revolutionizes the field of genomics by enabling the rapid sequencing of DNA and RNA. Unlike traditional sequencing methods, NGS employs a massively parallel approach, allowing for the simultaneous determination of the nucleotide order in entire genomes or targeted regions. This capability results in ultra-high throughput and scalability, making it a powerful tool for both research and clinical applications.

### **Importance of NGS Workflows**

#### **1. Efficiency:**

NGS workflows are designed to optimize the sequencing process, significantly enhancing the speed of sample preparation and analysis. This efficiency is vital in both research settings, where timely results can drive further investigations, and clinical environments, where rapid diagnostics can influence patient care decisions. By minimizing manual steps and automating processes, NGS workflows facilitate quicker turnaround times for obtaining genomic data.

#### **2. Cost-Effectiveness:**

The evolution of NGS has led to a dramatic reduction in sequencing costs over the past decade. As technologies have advanced, the price per base of DNA sequenced has decreased significantly, making high-throughput sequencing accessible to a broader range of researchers and institutions. This affordability empowers smaller labs and universities to engage in cutting-edge genomic research that was previously limited to well-funded facilities.



#### **3. Data Management:**

The vast amounts of data generated by NGS pose significant challenges in terms of management, analysis, and interpretation. Effective workflows incorporate robust data management systems that facilitate the handling of large datasets. These systems enable researchers to efficiently store, retrieve, and analyze genomic information, ensuring that meaningful insights can be derived from complex data sets without overwhelming computational resources.

#### **4. Customization:**

One of the key advantages of NGS workflows is their flexibility. Researchers can tailor their sequencing approaches based on specific experimental needs, whether they are focusing on whole-genome sequencing, targeted resequencing, or RNA sequencing. This customization allows for the optimization of protocols to suit different sample types or research questions, enhancing the relevance and applicability of the results obtained.

#### **5. Advancements in Precision Medicine:**

NGS workflows play a crucial role in advancing precision medicine by enabling the identification of genetic variants that inform treatment decisions. By analyzing an individual's genomic information, healthcare providers can tailor therapies to improve patient outcomes based on specific genetic markers associated with diseases. For instance, NGS can uncover mutations linked to cancer susceptibility or drug resistance, guiding oncologists in selecting personalized treatment regimens that enhance efficacy while minimizing adverse effects.

In summary, Next-Generation Sequencing (NGS) represents a transformative advancement in genomics that offers unparalleled efficiency and cost-effectiveness while paving the way for significant strides in precision medicine. As workflows continue to improve and adapt to emerging challenges, NGS will undoubtedly play a pivotal role in shaping the future of genetic research and personalized healthcare.

### **III. Graphical User Interface (GUI)**

A Graphical User Interface (GUI) is a sophisticated type of user interface that enables users to interact with electronic devices through visual elements, such as graphical icons, buttons, and visual indicators. Unlike command-line interfaces that require text-based commands, GUIs provide a more intuitive and user-friendly experience, making technology accessible to a broader audience. GUIs are prevalent in software applications across various platforms, including desktop computers, mobile devices, and web applications, enhancing usability and overall efficiency.

#### **Key Features of GUIs**

##### **1. Visual Elements:**

GUIs utilize icons, buttons, menus, and windows to represent functions and data visually. This design allows users to navigate software applications more easily by clicking or tapping on graphical elements rather than typing commands.

##### **2. Secondary Notation:**

Secondary notation refers to additional visual cues that enhance understanding and usability. For instance, color coding or tooltips can provide contextual information about an icon's function, improving the user experience.

### 3. Interactive Components:

GUIs often include interactive components such as sliders, checkboxes, and dropdown menus that allow users to manipulate data and settings dynamically. This interactivity fosters a more engaging user experience.



## **Benefits of Implementing a GUI in NGS Workflows**

### 1. Enhanced Efficiency:

By providing a visually intuitive interface, GUIs streamline the workflow for Next-Generation Sequencing (NGS) processes. Users can quickly access tools and functionalities without needing extensive training or familiarity with complex commands. This efficiency is particularly beneficial in high-throughput environments where time is critical.

### 2. Improved Data Management:

GUIs facilitate better data management by allowing users to visualize and organize large datasets effectively. Through features like drag-and-drop functionality and customizable dashboards, researchers can manage genomic data more intuitively, making it easier to track samples and results.

### 3. Customization Flexibility:

A well-designed GUI offers customization options that enable researchers to tailor their workflows according to specific experimental needs. Users can adjust settings or modify visual layouts to suit their preferences, enhancing the overall usability of the software.

### 4. Increased Accessibility:

GUIs make complex technologies more accessible to a wider range of users, including those who may not have a strong technical background. By lowering the barrier to entry for using NGS tools, GUIs democratize access to genomic research and analysis.

### 5. Facilitated Collaboration:

Collaborative features integrated into GUIs enable multiple users to work together seamlessly on genomic projects. Shared access to data visualizations and analysis tools fosters teamwork among researchers, enhancing productivity and innovation in precision medicine initiatives.

### **Supporting Advancements in Precision Medicine**

The integration of GUIs into NGS workflows significantly contributes to advancements in precision medicine by streamlining processes that lead to personalized healthcare solutions. By simplifying data analysis and interpretation:

- 1. Rapid Insights:** Researchers can quickly derive insights from genomic data, allowing for timely decision-making in clinical settings.
  - 2. User-Friendly Analysis Tools:** GUIs often include built-in analytical tools that help identify genetic variants associated with diseases, facilitating the development of targeted therapies.
  - 3. Educational Resources:** Many GUI-based applications provide tutorials or help sections that educate users about genomic concepts and analysis techniques, further empowering researchers in the field.
-

# **CHAPTER 4: REVIEW OF LITERATURE**

## **REVIEW OF LITERATURE**

### **A. THE INTEGRATION OF NEXT – GENERATION SEQUENCING IN PRECISION MEDICINE**

#### **Definition and Goals of Precision Medicine**

Precision medicine aims to customize healthcare, with treatments tailored to individual genetic, environmental, and lifestyle factors. It moves away from a “one-size-fits-all” approach to treatment, allowing for more personalized care. The goal of precision medicine is to use patient-specific information, particularly genetic data, to predict disease risk, diagnose illnesses early, and develop treatment plans that are more effective and have fewer side effects. By leveraging detailed biological data, precision medicine enhances the ability to identify the most appropriate therapies for individuals, particularly in complex conditions like cancer, where genetic variability plays a critical role in disease progression and treatment response.

#### **Role of NGS in Precision Medicine**

Next-generation sequencing (NGS) has revolutionized the field of precision medicine by enabling the rapid and accurate sequencing of entire genomes or targeted regions of interest. NGS allows clinicians and researchers to identify genetic mutations that drive diseases, such as cancer, making it possible to match patients with therapies that target these specific mutations. Through NGS, clinicians can identify “actionable mutations,” or genetic changes for which targeted treatments are available.

In oncology, for instance, NGS helps sequence tumor genomes to uncover mutations that can be treated with targeted therapies. Studies like the Genomics Evidence Neoplasia Information Exchange (GENIE) have demonstrated that up to 30% of sequenced cancers reveal actionable mutations that can guide treatment decisions. NGS not only improves diagnosis and therapy selection but also aids in the discovery of new biomarkers and the development of drugs tailored to genetic profiles. By integrating NGS into routine clinical practice, precision medicine offers a future where treatment is more effective, disease risks are mitigated, and healthcare is personalized for better patient outcomes.

#### **Challenges in NGS Workflows for Precision Medicine**

##### **1. Data Management and Integration:**

One of the main challenges in NGS workflows is managing and integrating the massive volume of genomic data generated. NGS produces large, complex datasets that must be organized, stored, and analyzed in a way that makes the data meaningful for precision medicine applications. Aggregating data from different sources, such as clinical records, genomic sequences, and environmental data, requires standardized formats and efficient data processing methods. Without proper data integration frameworks, researchers face difficulties in combining and analyzing heterogeneous datasets, which hinders the discovery of actionable insights for treatment. Moreover, inconsistent use of existing data standards further complicates the ability to share and interpret data across institutions.



## **2. Data Privacy and Security:**

As genomic data is highly personal, privacy and security are critical concerns in precision medicine. Handling sensitive patient data, such as genomic sequences, requires stringent data protection measures to prevent unauthorized access and ensure patient confidentiality. With the potential for sequence-based re-identification, there is a need for secure data-sharing mechanisms that protect privacy while enabling collaboration among researchers. This challenge is compounded when data is shared across institutions or stored in large biorepositories. Initiatives such as the development of machine-readable consent forms and role-based access systems aim to address these concerns, but implementation remains difficult.

## **3. Technological Limitations:**

The computational infrastructure needed to support real-time NGS data analysis poses another major challenge. Analyzing large-scale genomic datasets requires advanced computational resources, including high-performance computing systems, scalable storage solutions, and efficient data processing pipelines. The lack of adequate infrastructure can slow down analysis, making it harder to deliver timely results for clinical decision-making. Developing scalable and cost-effective solutions to support high-throughput sequencing data is essential for realizing the full potential of precision medicine. Additionally, ensuring data accuracy and reliability throughout the process remains a technological hurdle.

## **Barriers to the Clinical Adoption of NGS**

### **1. Interpretation by Physicians:**

One significant barrier to the clinical adoption of next-generation sequencing (NGS) in precision medicine is the difficulty many physicians face in interpreting complex genomic data. While NGS provides vast amounts of actionable information, the ability to understand and apply this data effectively in clinical practice requires specialized knowledge. Many oncologists are not fully trained in genomics, which can lead to discomfort in interpreting sequencing results and translating them into therapeutic decisions. In a survey conducted at the Mayo Clinic, over half of the providers expressed discomfort in interpreting genomic test results. Tumor boards, where genomic experts collaborate with clinicians, have become an essential support system to guide treatment decisions, but reliance on these boards also highlights the knowledge gap among general practitioners.

### **2. Access to Sequencing and Matched Therapies:**

Even when actionable mutations are identified, practical barriers often prevent patients from receiving appropriate sequencing-matched therapies. Access to NGS and its associated treatments can be limited by several factors, including the availability of clinical trials, eligibility restrictions, and geographical barriers. Many patients, especially those with advanced cancer, may not qualify for clinical trials due to previous treatments or other health conditions.

For example, the National Cancer Institute's Molecular Analysis for Therapy Choice (MATCH) trial found that only 33 of 56 patients with targetable mutations met the

eligibility criteria for a matched therapy. Moreover, sequencing may not always be performed early enough in the treatment process, limiting patients' opportunities to benefit from targeted therapies.

### **3. Insurance and Cost Factors:**

The cost of NGS and matched therapies presents another barrier to widespread clinical adoption. Although some insurance companies cover specific companion diagnostic tests, coverage for broader NGS panels, such as whole-exome or genome sequencing, is often limited. These tests are frequently deemed investigational, and many insurers do not provide reimbursement for off-label uses of targeted therapies identified through NGS. Patients often must rely on clinical trials to access sequencing and related treatments, and even when therapies are FDA-approved, off-label use of these drugs can lead to substantial out-of-pocket costs. The variability in insurance coverage means that patient access to these life-saving treatments is inconsistent and often dependent on their financial situation or healthcare provider's network. This lack of coverage remains a significant obstacle to the broader implementation of NGS in oncology.

### **Outcomes and Effectiveness**

Studies have shown that NGS-guided treatments can improve clinical outcomes in cancer patients. For instance, in a study conducted by Tsimberidou et al., advanced cancer patients receiving NGS-matched therapies demonstrated significantly better outcomes, including an overall response rate of 27% compared to 5% in those receiving standard therapies. Similarly, improvements in progression-free survival and overall survival have been reported for patients with sequencing-matched treatments. For example, in one study, patients receiving targeted therapies based on NGS results had a progression-free survival of 86 days, compared to 49 days for those receiving non-matched therapies.

However, challenges remain, as not all patients with actionable mutations can receive matched therapies due to various barriers, such as access to clinical trials or drug availability. Additionally, despite the success in identifying actionable mutations, only a fraction of patients go on to receive targeted therapies. Nonetheless, the overall trend suggests that NGS-based oncology treatments improve patient outcomes by offering a more personalized, precise approach to cancer care.

### **Future Directions and Innovations in NGS for Precision Medicine**

One of the key challenges in realizing the full potential of NGS in precision medicine is the lack of universal data standards and protocols. Currently, data generated from NGS workflows can vary greatly in format and quality, making it difficult to integrate and analyze across different platforms and institutions. Standardization is crucial to ensuring consistency, data integrity, and interoperability. Without unified standards, the sharing of genomic data between research institutions, hospitals, and healthcare systems becomes complicated, impeding collaborative research and the development of large-scale genomic databases.

Efforts are underway to create these standards. Initiatives like BioSharing, the Global Alliance for Genomics and Health (GA4GH), and the NIH Data Commons aim to address issues related to data security, quality, and standardization. These initiatives focus on creating a framework where genomic data can be securely stored, easily accessed, and efficiently shared. The use of common data models and standardized metadata for genomic datasets will play a crucial role in advancing precision medicine, as they will enable researchers and clinicians to aggregate data from multiple sources and make informed decisions based on comprehensive analyses.

### **Emerging Technologies**

As NGS technologies continue to evolve, several innovations hold great promise for advancing precision medicine. One area of growth is the development of real-time sequencing technologies that can generate faster results, allowing clinicians to make immediate therapeutic decisions. Advances in nanopore sequencing, for example, could reduce the cost and time associated with sequencing, making it more accessible for routine clinical use.

Another promising development is the integration of artificial intelligence (AI) and machine learning into NGS data analysis. AI can help process the vast amounts of genomic data produced by NGS, identifying patterns and mutations that might be missed by traditional methods. Additionally, AI-driven tools can assist in predicting patient responses to targeted therapies based on their genomic profiles, further personalizing treatment strategies.

As these technologies mature, they will likely reduce the cost and complexity of NGS, making precision medicine more accessible to a broader range of patients. With continued innovation in both the technological and computational aspects of NGS, precision medicine is poised to become a routine part of healthcare, offering more tailored and effective treatments.

### **Current Applications of NGS in Oncology**

Next-generation sequencing (NGS) plays a pivotal role in oncology by identifying "actionable mutations," which are specific genetic alterations in tumors that can be targeted with precision therapies. These mutations help clinicians determine which therapies are most likely to be effective for a given patient. For example, mutations in genes like EGFR, BRAF, or KRAS can guide the use of targeted treatments such as kinase inhibitors, immunotherapies, or monoclonal antibodies. Data from initiatives like the Genomics Evidence Neoplasia Information Exchange (GENIE) indicate that about 30% of tumors sequenced have actionable mutations. The identification of these mutations enables clinicians to match patients with therapies that are more likely to be effective, reducing the use of one-size-fits-all treatments and enhancing treatment specificity.

A key advantage of NGS in oncology is its ability to uncover multiple mutations within a single tumor, providing a comprehensive genetic profile that can inform combination therapies or sequential treatment strategies. This personalized approach ensures that patients receive therapies tailored to the molecular drivers of their cancer, increasing the likelihood of treatment success and minimizing adverse effects.

## B. NEXT – GENERATION SEQUENCING (NGS) TOOLS

Bioinformatics tools are essential software applications and algorithms designed to analyze, interpret, and visualize biological data, particularly genomic and proteomic information. Their importance lies in their ability to handle vast amounts of data generated by high-throughput sequencing technologies, enabling researchers to make sense of complex biological systems and contribute to advancements in fields such as genomics, transcriptomics, and proteomics.

### Categories of Bioinformatics Tools

Bioinformatics tools can be classified based on their functionality, which includes:

1. **Data Acquisition:** Tools like the SRA Toolkit facilitate the retrieval of sequencing data from public repositories. They allow researchers to download and manage large datasets efficiently, which is crucial for subsequent analyses.
2. **Quality Control:** Tools such as FastQC are vital for assessing the quality of sequencing data. They provide metrics on various aspects, including sequence quality scores and GC content, helping researchers identify potential issues before proceeding with further analysis.
3. **Pre-processing:** Fastp is an example of a tool that offers comprehensive pre-processing capabilities, including adapter trimming and quality filtering. This step is critical for ensuring that only high-quality reads are used in downstream analyses.
4. **Variant Calling:** The Genome Analysis Toolkit (GATK) is widely used for variant discovery from sequence data. It implements best practices for calling variants, which is essential for understanding genetic variations associated with diseases.
5. **Data Manipulation:** SAM Tools enable researchers to manipulate sequence alignment files (SAM/BAM formats). They provide functionalities such as sorting and indexing, which are necessary for efficient data handling during analysis.

These tools collectively enhance the efficiency and accuracy of bioinformatics analyses, allowing researchers to focus on interpreting results rather than managing data intricacies. The growing number of specialized tools reflects the increasing complexity of biological questions being addressed in modern research.

## I. SRA TOOLKIT

The **SRA Toolkit** is a collection of command-line tools designed to facilitate access to and management of data stored in the Sequence Read Archive (SRA), a public repository for high-throughput sequencing data. This toolkit serves as a vital resource for researchers looking to retrieve and analyze genomic data efficiently.



The SRA Toolkit allows users to download, convert, and manipulate sequencing data from the SRA. It supports various types of sequencing data, making it an essential tool for bioinformaticians and researchers in genomics. The toolkit is particularly advantageous for those who require large

datasets for their analyses, as it streamlines the process of accessing these resources.

### Key Features

1. **Data Access:** The SRA Toolkit provides seamless access to a vast array of datasets available in the SRA. Users can easily search for specific studies or datasets using identifiers such as the Study ID or Run ID.
2. **Downloading Methods:** The toolkit offers multiple downloading options, including direct downloads and the ability to download specific formats like FASTQ. This flexibility allows users to choose the method that best suits their computational environment.
3. **Formats Supported:** The SRA Toolkit supports various file formats, including SRA, FASTQ, and BAM files. This versatility enables researchers to work with different types of data without needing extensive conversions.

### Applications

1. **Genomic Research:** Researchers utilize the toolkit to access large-scale genomic datasets for studies on population genetics, evolutionary biology, and disease association.
2. **Comparative Genomics:** It facilitates comparative analyses by allowing easy retrieval of datasets from different organisms or conditions.
3. **Method Development:** Bioinformaticians often use the toolkit to test new algorithms or methods on existing datasets, contributing to advancements in genomic analysis techniques.

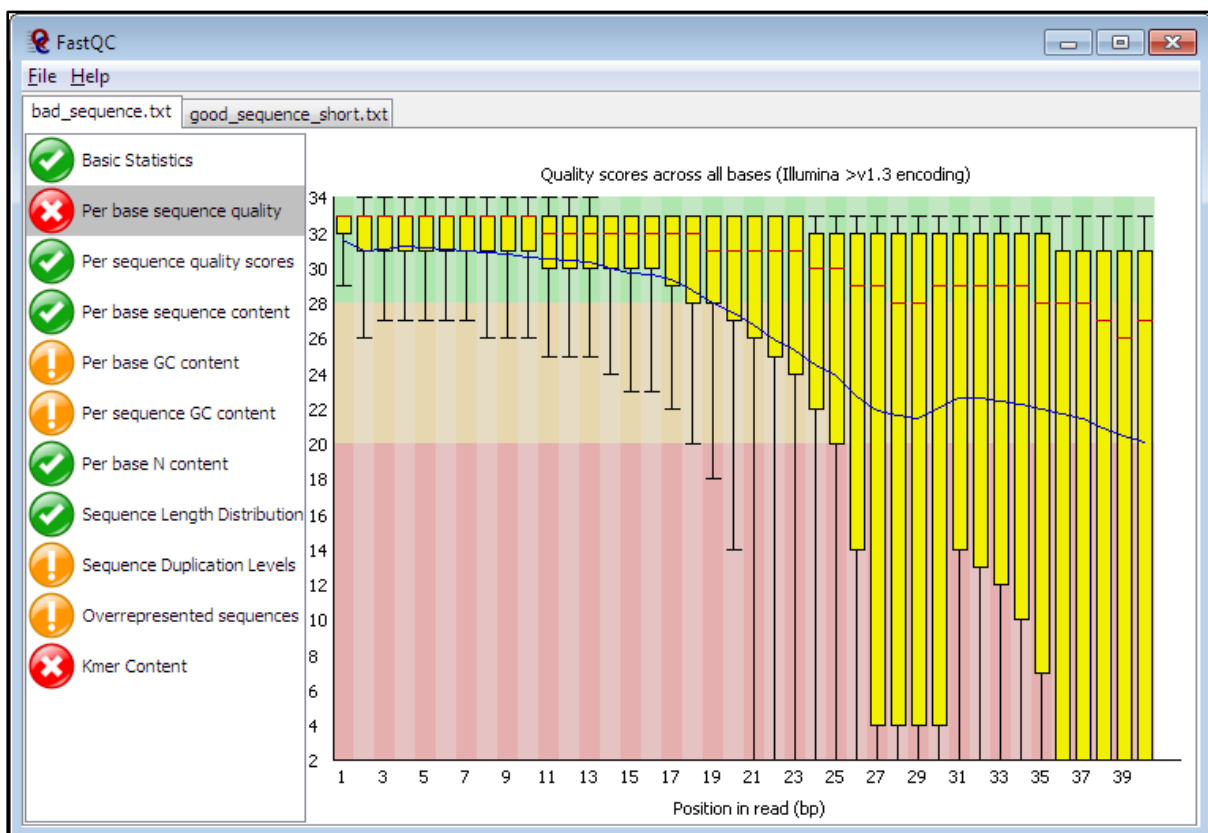
In summary, the SRA Toolkit is an indispensable resource that enhances the accessibility and usability of genomic data, supporting a wide range of research applications in genomics and beyond.

## II. FASTQC

**FastQC** is a widely used bioinformatics tool designed for quality control of high-throughput sequencing data. Its primary purpose is to provide a comprehensive overview of the quality of sequencing reads, helping researchers identify potential issues that may affect downstream analyses.

FastQC serves as an essential first step in the analysis pipeline, offering insights into the quality of raw sequencing data. By evaluating various aspects of the data, it enables researchers to make informed decisions about subsequent processing steps, such as filtering or trimming reads to enhance data quality.

### Key Metrics Analyzed



**Figure 1: Sample FASTQC Report**

### 1. Basic Statistics:

This section provides fundamental information about the input FASTQ file, including:

- File Name:** The name of the analyzed file.
- Quality Score Encoding:** Indicates the encoding method used for quality scores.
- Total Number of Reads:** The total count of sequences in the file.
- Read Length:** The length of the reads, which is crucial for downstream analyses.
- GC Content:** The percentage of guanine (G) and cytosine (C) bases, which can provide insights into the composition of the sequenced material.

## **2. Per Base Sequence Quality:**

This module presents a box-and-whisker plot showing quality scores at each base position across all reads. Key features include:

- a. **Mean Quality Score:** Represented by a blue line, indicating average quality.
- b. **Median Quality Score:** Shown as a red line within a yellow box, providing insight into the central tendency of quality scores.
- c. **Quality Score Ranges:** The background color indicates quality levels (green for good, orange for acceptable, red for poor). A common observation is that quality tends to decrease towards the end of reads due to signal decay<sup>234</sup>.

## **3. Per Sequence Quality Scores:**

This plot displays the distribution of average quality scores across all sequences. It helps identify if a significant portion of sequences has low quality, which could indicate issues such as poor imaging or contamination.

## **4. Per Sequence GC Content:**

This module shows the distribution of GC content across all sequences. A normal distribution is expected in whole genome shotgun sequencing, with deviations potentially indicating contamination or over-representation of certain sequences.

## **5. Per Base N Content:**

This metric indicates the percentage of bases called as 'N' (unknown) at each position. A high percentage at any position suggests problems during sequencing, such as instrument errors or insufficient signal.

## **6. Sequence Duplication Levels:**

This analysis assesses how many times each unique sequence appears in the dataset. High levels of duplication can indicate issues with library preparation or PCR amplification bias.

## **7. Overrepresented Sequences:**

FastQC identifies any sequences that are present at unusually high frequencies, which may suggest contamination or bias in library preparation. This module is crucial for detecting potential issues before further analysis.

## **8. Adapter Content:**

This section checks for the presence of adapter sequences that may not have been completely removed during library preparation. High levels of adapter contamination can affect downstream applications and analyses.

### **Interpretation of Results**

Interpreting FastQC output involves examining the provided visualizations and metrics:

1. **Quality Score Graphs:** A consistent high-quality score across all bases indicates good sequencing quality. A significant drop towards the end of reads may suggest the need for trimming.
2. **GC Content Plot:** A GC content that deviates significantly from expected values may indicate contamination or bias in library preparation.
3. **Adapter Content:** If adapter sequences are detected, it signals that additional trimming is necessary before proceeding with further analysis.

### **Interpretation of Flags**

Each module in FastQC is assigned a flag:

1. **Pass (Green Tick):** Indicates no issues detected.
2. **Warn (Orange Exclamation Mark):** Suggests caution; potential issues should be investigated further.
3. **Fail (Red Cross):** Indicates significant problems that need addressing before proceeding with analysis.

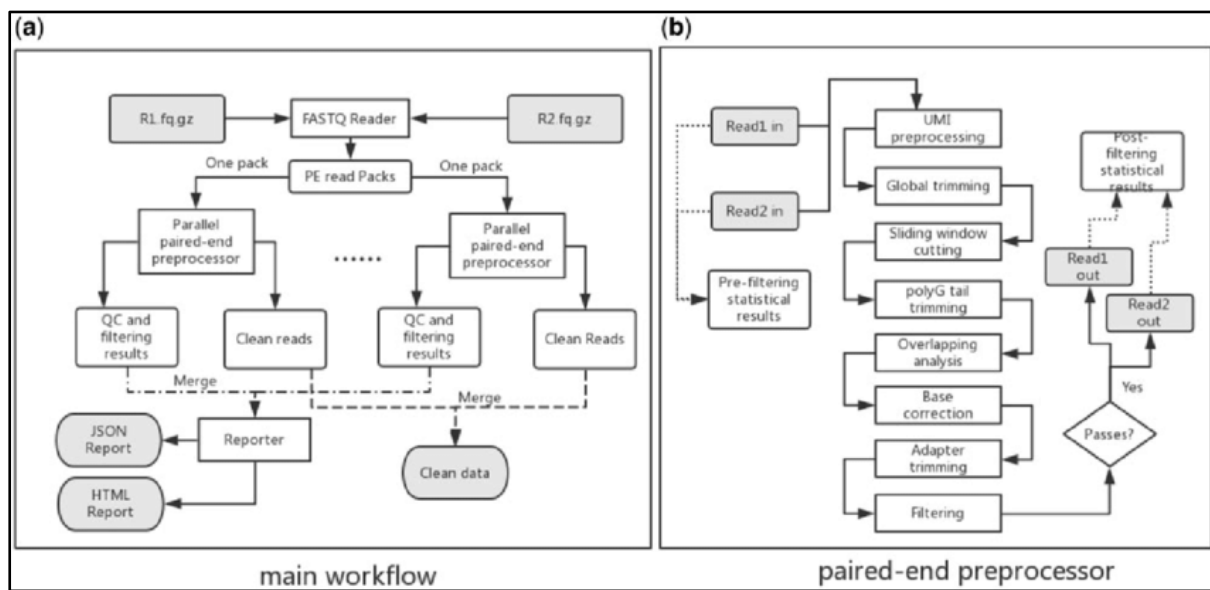
In summary, FastQC is an invaluable tool for ensuring high-quality sequencing data. By analyzing critical metrics and providing clear visualizations, it empowers researchers to make informed decisions regarding data processing and analysis.



### III. FASTP

**Fastp** is an innovative and efficient pre-processing tool designed for high-throughput sequencing data. It streamlines the initial steps of data analysis by providing a comprehensive suite of functionalities that enhance data quality and usability.

Fastp serves as a versatile pre-processing tool that integrates multiple functions into a single application, allowing researchers to handle raw sequencing data efficiently. It is particularly suitable for applications requiring rapid processing of large datasets, making it a popular choice in genomic studies.



**Figure 2: Workflow for the processing of files using the Fastp tool**

#### Features

Fastp offers several key features that distinguish it from other pre-processing tools:

- 1. Quality Filtering:** Fastp employs dynamic quality filtering, which allows users to set thresholds for base quality scores. This ensures that only high-quality reads are retained for downstream analysis, effectively reducing noise and improving overall data integrity.
- 2. Adapter Trimming:** One of Fastp's standout features is its ability to detect and trim adapter sequences automatically. This is crucial because untrimmed adapters can lead to erroneous results in subsequent analyses. Fastp identifies adapter sequences based on user-defined or built-in patterns, ensuring clean reads.
- 3. Read Correction:** Fastp includes a read correction feature that addresses errors in sequencing data. By leveraging overlapping reads, it can correct low-quality bases, enhancing the reliability of the data before further analysis.

### **Performance Comparison**

When compared to other tools like FastQC, Fastp stands out due to its integrated approach. While FastQC primarily focuses on quality assessment rather than pre-processing, Fastp combines quality control with essential pre-processing functions. This dual capability allows researchers to streamline their workflow significantly. In terms of speed and efficiency, Fastp has been shown to outperform many traditional tools by processing large datasets more rapidly without compromising accuracy. Additionally, Fastp generates comprehensive reports that provide insights into the quality of the processed data, similar to what FastQC offers but with the added advantage of immediate corrective actions. In summary, Fastp is a powerful pre-processing tool that enhances the quality of sequencing data through its advanced features, making it an essential component in modern bioinformatics workflows. Its combination of quality filtering, adapter trimming, and read correction positions it as a superior choice for researchers seeking efficient and reliable data preparation methods.

## **IV. SAM TOOLS**

SAM Tools is a suite of utilities designed for manipulating sequencing data in the SAM (Sequence Alignment/Map) and BAM (Binary Alignment/Map) formats. These formats are essential for storing large-scale sequence alignment data generated by high-throughput sequencing technologies. SAM Tools provides researchers with the necessary functionalities to manage and process this data efficiently.

SAM Tools is widely recognized for its ability to handle the complexities of sequencing data. It facilitates various operations that are crucial for preparing data for downstream analysis, making it an indispensable resource in bioinformatics workflows. The toolkit supports a range of tasks, from basic file manipulations to more complex operations, ensuring that researchers can work with their alignment data effectively.

### **Key Functions**

1. **Sorting:** SAM Tools can sort alignment files by genomic coordinates, which is essential for many downstream analyses, including variant calling. Sorted files allow for efficient access and processing of data.
2. **Merging:** The toolkit enables users to merge multiple BAM files into a single file. This is particularly useful when combining data from different sequencing runs or samples, ensuring a cohesive dataset for analysis.
3. **Indexing:** SAM Tools provides indexing capabilities that create index files for BAM files. Indexing allows quick access to specific regions of the alignment data, significantly speeding up data retrieval during analyses.
4. **Viewing:** The toolkit includes utilities for viewing SAM and BAM files in a human-readable format, facilitating easy inspection of alignment data and quality control checks.

### **Applications**

SAM Tools is employed in various practical applications within genomic analyses:

1. **Variant Calling:** Researchers use SAM Tools to prepare alignment files for variant calling pipelines, ensuring that the data is sorted and indexed correctly for efficient processing.
2. **Data Visualization:** The ability to view and manipulate alignment files allows researchers to conduct preliminary analyses and quality assessments before proceeding with more complex analyses.
3. **Comparative Genomics:** SAM Tools aids in merging and sorting alignment files from different species or conditions, enabling comparative studies that explore genetic differences and evolutionary relationships.

In summary, SAM Tools is a vital toolkit for anyone working with sequencing data. Its functionalities—sorting, merging, indexing, and viewing—are essential for preparing high-quality datasets that underpin robust genomic analyses.

## **V. VARIANT CALLING**

Variant Calling is a critical process in genomics that involves identifying variants—such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels)—from high-throughput sequencing data. This process is essential for understanding genetic variations that contribute to diseases, traits, and evolutionary changes.

Variant calling serves as the foundation for many genomic analyses, including population genetics, cancer genomics, and personalized medicine. Accurate variant identification is crucial for discovering genetic markers associated with diseases and for developing targeted therapies. The reliability of variant calling directly impacts the validity of subsequent analyses and conclusions drawn from genomic studies.

### **GATK (Genome Analysis Toolkit)**

GATK is widely regarded as a gold standard for variant calling, particularly in human genomics. It follows best practices that include steps like base quality score recalibration and local realignment, ensuring high accuracy in variant detection .

GATK (Genome Analysis Toolkit) is a powerful software suite developed for variant discovery in high-throughput sequencing data. Its significance lies in its ability to provide researchers with robust tools for detecting and analyzing genetic variants, which are critical for understanding diseases, traits, and evolutionary biology.



**Figure 3: NGS Pipeline involving GATK**

The primary purpose of GATK is to facilitate the accurate identification of variants such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) from sequencing data. As sequencing technologies have advanced, the need for reliable and efficient variant calling has become paramount in both clinical and research settings. GATK addresses this need by offering a comprehensive framework that integrates various computational methods tailored for different types of genomic data.

### **Key Features**

1. **Variant Calling Pipeline:** GATK provides a structured pipeline that guides users through the variant calling process. This includes steps such as preprocessing raw sequencing reads, marking duplicates, local realignment, and base quality score recalibration.
2. **Best Practices:** The "GATK Best Practices" guidelines are widely recognized as a standard for variant calling. These recommendations are continually updated to reflect

the latest advancements in sequencing technology and computational methods, ensuring that researchers have access to state-of-the-art techniques.

### **Applications in Research**

1. **Clinical Genomics:** In clinical settings, GATK is employed to identify disease-associated variants in patient genomes. For example, studies have integrated GATK with mapping and assembly approaches to enhance variant detection accuracy in clinical sequencing applications
2. **Population Genomics:** Researchers use GATK to analyze population-level genomic data, allowing them to explore genetic diversity and evolutionary patterns among different species.
3. **Cancer Genomics:** GATK plays a crucial role in cancer research by identifying somatic mutations that drive tumorigenesis. Its ability to handle complex data types makes it suitable for analyzing heterogeneous tumor samples.

In summary, GATK is a cornerstone tool in the field of genomics, providing essential capabilities for variant discovery. Its comprehensive pipeline, adherence to best practices, and broad applicability make it invaluable for researchers aiming to unravel the complexities of genetic variation.

### **Challenges and Considerations**

Despite advancements in variant calling methodologies, several challenges persist:

1. **Data Quality:** High-quality sequencing data is essential for accurate variant detection. Low-quality reads can lead to false positives or missed variants.
2. **Computational Complexity:** Variant calling can be computationally intensive, especially with large datasets or when detecting low-frequency variants. This often necessitates high-performance computing resources.
3. **Reference Bias:** Variant callers rely on reference genomes, which can introduce biases if the reference does not accurately represent the population being studied. This is particularly problematic in diverse populations or when dealing with structural variants.

In summary, variant calling is a fundamental aspect of genomic research that enables the identification of genetic variations critical for understanding biology and disease. While various algorithms exist to facilitate this process, challenges related to data quality and computational demands continue to influence the field.

# **CHAPTER 5: METHODOLOGY**

## **METHODOLOGY**

The methodology for developing an integrated Next-Generation Sequencing (NGS) workflow with a user-friendly Graphical User Interface (GUI) encompasses several systematic phases. These phases ensure a comprehensive approach from problem identification to the final release of the software. The methodology is structured under the following key headers:

### **STEP 1: Problem Statement and Literature Review**

#### **1. Problem Identification:**

The project initiates with the identification of critical challenges in existing NGS workflows, particularly focusing on the complexities associated with data analysis, interpretation, and the lack of intuitive user interfaces that hinder accessibility for clinicians and researchers in precision medicine.

#### **2. Comprehensive Literature Review:**

A thorough review of current literature is conducted to examine existing NGS workflows, precision medicine applications, and GUI design principles. This review aims to identify best practices, uncover gaps in current methodologies, and establish a foundation for developing an enhanced, user-centric NGS pipeline. Key areas of focus include:

- a. NGS Workflow Efficacy:** Evaluation of existing pipelines, their strengths, and limitations in handling large-scale sequencing data.
- b. Precision Medicine Integration:** Analysis of how current workflows support personalized treatment strategies and identify areas for improvement.
- c. GUI Design Principles:** Investigation of user interface designs that enhance usability, accessibility, and functionality for end-users in biomedical research settings.



### **STEP 2: Pipeline Designing and Testing**

#### **1. Workflow Design:**

The NGS workflow is meticulously mapped out to include all essential steps from raw data acquisition to final reporting. The designed workflow encompasses:

- a. Data Acquisition:** Downloading raw NGS data (reads in FASTQ format) using tools like SRA Toolkit.
- b. Quality Control:** Performing quality filtering and adapter trimming using FastQC and Fastp to ensure data integrity.
- c. Alignment and Mapping:** Utilizing alignment tools such as BWT or Bowtie for mapping reads to reference genomes.
- d. Base Quality Recalibration:** Applying GATK for recalibrating base quality scores to improve variant calling accuracy.
- e. Variant Calling:** Identifying mutations through variant calling processes.
- f. Data Integration:** Connecting to online databases via APIs to retrieve additional information on drugs and potential therapeutic targets.

- g. Reporting:** Generating a comprehensive PDF summary of the entire workflow, which can be downloaded by the user.

## 2. Pipeline Testing:

Each step of the pipeline undergoes rigorous testing using sample datasets to validate functionality, ensure data integrity, and optimize performance. Automated scripts and shell scripting are employed to facilitate task automation, manage file operations, and execute command-line tools efficiently.



### STEP 3: Backend Development and Integration of NGS Tools

#### 1. Programming Language Selection:

The backend development leverages a combination of programming languages to harness their respective strengths:

- a. Python:** Utilized for its extensive libraries that facilitate NGS data analysis and GUI development.
- b. Shell Scripting:** Used for automating various tasks within the NGS workflow, enhancing efficiency in file management and tool execution.
- c. Additional Languages:** May be incorporated as needed to support specialized data analysis and processing requirements.

#### 2. Integration of Bioinformatics Tools:

Existing bioinformatics tools and algorithms are seamlessly integrated into the backend to support comprehensive data analysis, visualization, and interpretation of NGS results. This integration ensures that users can perform end-to-end analyses within a unified platform without the need for external tool dependencies.

#### NGS Pipeline Implementation:

##### Downloading Raw NGS Data:

Utilizes SRA Toolkit to download raw sequencing reads in FASTQ format from public repositories.



##### Quality Filtering and Adapter Trimming:

Employs FastQC for initial quality assessment and Fastp for filtering low-quality reads and trimming adapter sequences to ensure high-quality input data.



##### Alignment and Mapping:

Uses alignment tools like BWT or Bowtie to map the filtered reads to a reference genome, producing aligned sequencing data.





**Base Quality Recalibration:**

Applies GATK for recalibrating base quality scores, enhancing the accuracy of subsequent variant calling.



**Variant Calling:**

Identifies genetic variants (mutations) from the recalibrated data, providing insights into genetic differences relevant to precision medicine.



**Connecting to Online Databases:**

Integrates APIs to fetch additional information on identified variants and associated drugs, enriching the analysis with actionable data.



**Generating Summary Report:**

Compiles the entire workflow's results into a comprehensive PDF summary, which includes data analysis outcomes and interpretations. This report is made available for download, facilitating easy sharing and review.



**STEP 4: GUI/Frontend or Software Development**

**1. Prototype Development:**

An initial prototype of the GUI is developed using suitable development tools and frameworks that support graphical interface design. The prototype serves as a foundational model to visualize workflow steps and gather preliminary user feedback.

**2. User Interface Design:**

The GUI is designed following best practices in user interface design to ensure it is intuitive, responsive, and accessible. Key features include:

- a. Workflow Visualization:** Clear mapping of NGS workflow steps, allowing users to navigate through data acquisition, analysis, and reporting seamlessly.
- b. Interactive Elements:** Incorporation of interactive components such as buttons, progress bars, and data visualization tools to enhance user engagement.
- c. Customization Options:** Providing users with the ability to configure analysis parameters and customize reports based on specific research needs.



## STEP 5: Validation and Optimization

### 1. Performance Validation:

The developed GUI and integrated NGS pipeline are subjected to extensive testing using real NGS datasets. This phase assesses the accuracy, reliability, and efficiency of the workflow. Key performance indicators include:

- a. **Data Processing Speed:** Measuring the time taken for each pipeline step to ensure timely analysis.
- b. **Accuracy of Results:** Verifying the correctness of variant calling and data interpretations against known benchmarks.
- c. **System Reliability:** Ensuring the software operates consistently without crashes or data loss.

### 2. Optimization Strategies:

Based on validation results, algorithms and workflow processes are optimized to enhance performance. Optimization efforts focus on:

- a. **Algorithm Refinement:** Improving the efficiency and accuracy of data processing algorithms.
- b. **Resource Management:** Enhancing the software's ability to handle large datasets without compromising performance.
- c. **User Experience Enhancements:** Streamlining processes to reduce user effort and improve overall satisfaction.



## STEP 6: Documentation and Release

### 1. Comprehensive Documentation:

Detailed documentation is developed to cover all aspects of the project, including step-by-step guides on how to use the software, perform analyses, and interpret results.

### 2. Final Release Preparation:

The software undergoes final quality assurance checks to ensure all components function as intended. Packaging includes the executable software, documentation, and necessary dependencies to facilitate easy installation and usage by end-users.

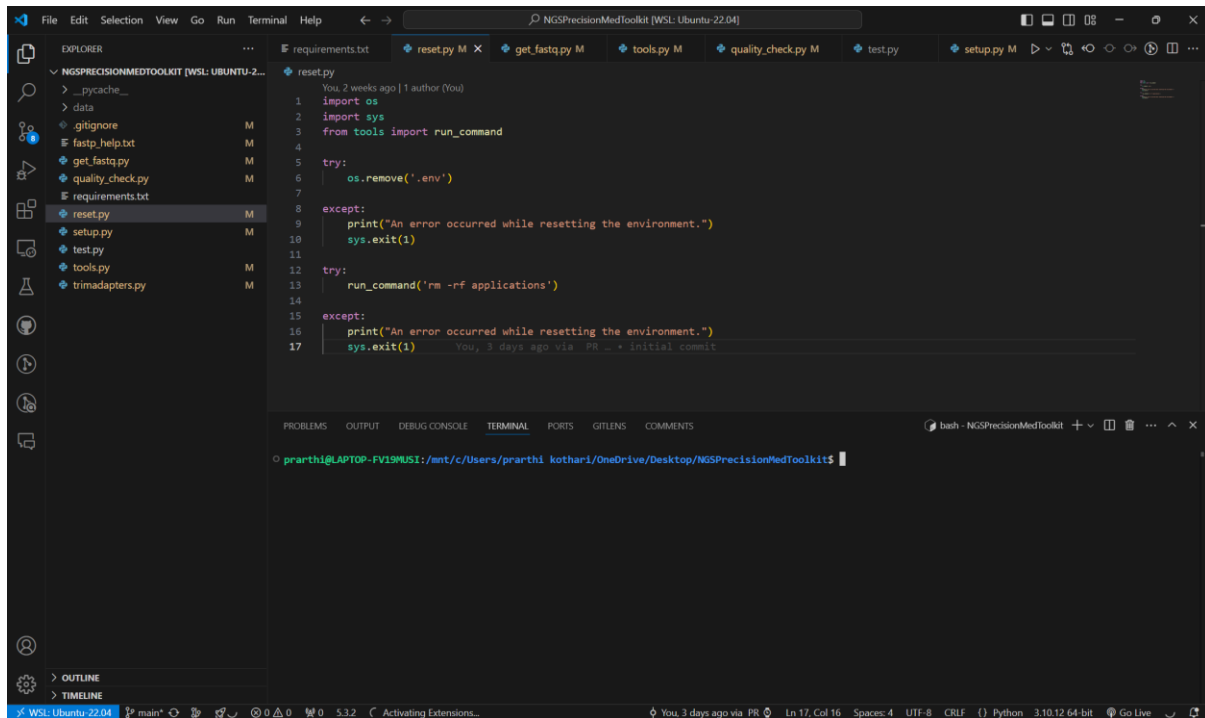
### 3. Release and Deployment:

The finalized software is released to the target audience through appropriate distribution channels. Ongoing support and updates are planned to address any emerging issues, incorporate user feedback, and implement future enhancements.

# **CHAPTER 6:**

# **CODE**

# CODE

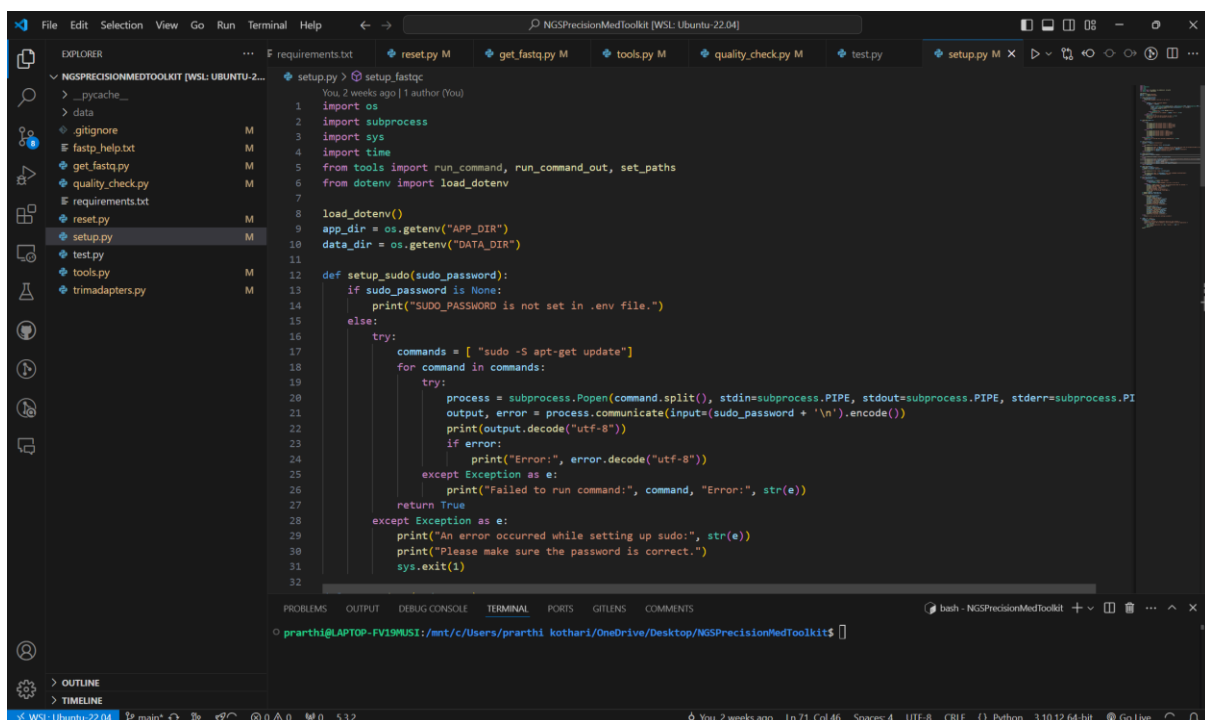


The screenshot shows the Visual Studio Code editor with the 'reset.py' file open. The file is located in the 'NGSPrecisionMedToolkit' project. The code in 'reset.py' is as follows:

```
1 import os
2 import sys
3 from tools import run_command
4
5 try:
6     os.remove('.env')
7
8 except:
9     print("An error occurred while resetting the environment.")
10    sys.exit(1)
11
12 try:
13     run_command('rm -rf applications')
14
15 except:
16     print("An error occurred while resetting the environment.")
17    sys.exit(1)
```

The terminal at the bottom shows the command prompt: `prarthi@LAPTOP-FV19MUS1:/mnt/c/Users/prarthi kothari/OneDrive/Desktop/NGSPrecisionMedToolkit$`.

Figure 4: Code for resetting the environment back to default



The screenshot shows the Visual Studio Code editor with the 'setup.py' file open. The file is located in the 'NGSPrecisionMedToolkit' project. The code in 'setup.py' is as follows:

```
1 import os
2 import subprocess
3 import sys
4 import time
5 from tools import run_command, run_command_out, set_paths
6 from dotenv import load_dotenv
7
8 load_dotenv()
9 app_dir = os.getenv("APP_DIR")
10 data_dir = os.getenv("DATA_DIR")
11
12 def setup_sudo(sudo_password):
13     if sudo_password is None:
14         print("SUDO_PASSWORD is not set in .env file.")
15     else:
16         try:
17             commands = ["sudo -S apt-get update"]
18             for command in commands:
19                 try:
20                     process = subprocess.Popen(command.split(), stdin=subprocess.PIPE, stdout=subprocess.PIPE, stderr=subprocess.PIPE)
21                     output, error = process.communicate(input=(sudo_password + '\n').encode())
22                     print(output.decode("utf-8"))
23                     if error:
24                         print("Error:", error.decode("utf-8"))
25                     except Exception as e:
26                         print("Failed to run command:", command, "Error:", str(e))
27                 return True
28             except Exception as e:
29                 print("An error occurred while setting up sudo:", str(e))
30                 print("Please make sure the password is correct.")
31                 sys.exit(1)
32
33 if __name__ == '__main__':
34     setup_sudo(None)
```

The terminal at the bottom shows the command prompt: `prarthi@LAPTOP-FV19MUS1:/mnt/c/Users/prarthi kothari/OneDrive/Desktop/NGSPrecisionMedToolkit$`.

Figure 5: Code for setting up the environment and downloading dependencies and applications / tools for running the NGS pipeline

```

13  def cmd_to_download(accession: str,
14                      alignment_filter: bool,
15                      compressed: bool,
16                      skip_technical: bool,
17                      remove_adapter: bool,
18                      spot_group: bool,
19                      alignment_filter_type: str = None,
20                      min_reads: int = None,
21                      max_reads: int = None,
22                      ar_specific: str = None,
23                      ar_start: int = None,
24                      ar_end: int = None,
25                      member: str = None):
26
27      command = ["fastq-dump", accession, "--split-3", "--outdir", f"data/{accession}"]
28      if compressed:
29          command.append("--gzip")
30      else:
31          command = command
32      if alignment_filter:
33          if alignment_filter_type == "split-spot":
34              command.remove("--split-3")
35              command.append("--split-spot")
36          elif alignment_filter_type == "aligned":
37              command.remove("--split-3")
38              command.append("--aligned")
39          elif alignment_filter_type == "unaligned":
40              command.remove("--split-3")
41              command.append("--unaligned")
42          elif alignment_filter_type == "aligned-region":
43              command.remove("--split-3")
44              command.append("--aligned-region")
45
46      return command

```

parthi@LAPTOP-FV19WU51: /mnt/c/Users/parthi.kothari/OneDrive/Desktop/NGSPrecisionMedToolkit\$

Figure 6: Code to retrieve FASTQ files using SRA Toolkit

```

1  You 3 days ago | 1 author (You)
2  import os, sys
3  from dotenv import load_dotenv
4  from tools import run_command_out, set_paths
5
6  load_dotenv()
7  fastqc_path = os.getenv("FASTQC_PATH")
8  working_dir = os.getenv("WORKING_DIR")
9
10 def fastqc_check(file, working_dir):
11     file_path = os.path.join(working_dir, file)
12     if os.path.isfile(file_path):
13         if any(ext in file for ext in [".fastq", ".sam", ".bam"]):
14             try:
15                 output_dir = os.path.join(working_dir, 'fastqc_reports')
16                 os.makedirs(output_dir, exist_ok=True)
17                 run_command_out(f"fastqc {file_path} --outdir {output_dir}", dir=fastqc_path)
18             except Exception as e:
19                 print(f"An error occurred while running the command: {e}")
20         else:
21             print(f"The file format of {file} is not supported.")
22
23 def main(working_dir):
24     fastqc_data_dir = os.path.join(working_dir, 'fastqc_reports')
25     set_paths("FASTQC_DATA_DIR", fastqc_data_dir)
26     files = os.listdir(working_dir)
27     for file in files:
28         fastqc_check(file, working_dir)
29     files = os.listdir(fastqc_data_dir)
30     reports = []
31     for file in files:
32         if 'html' in file:

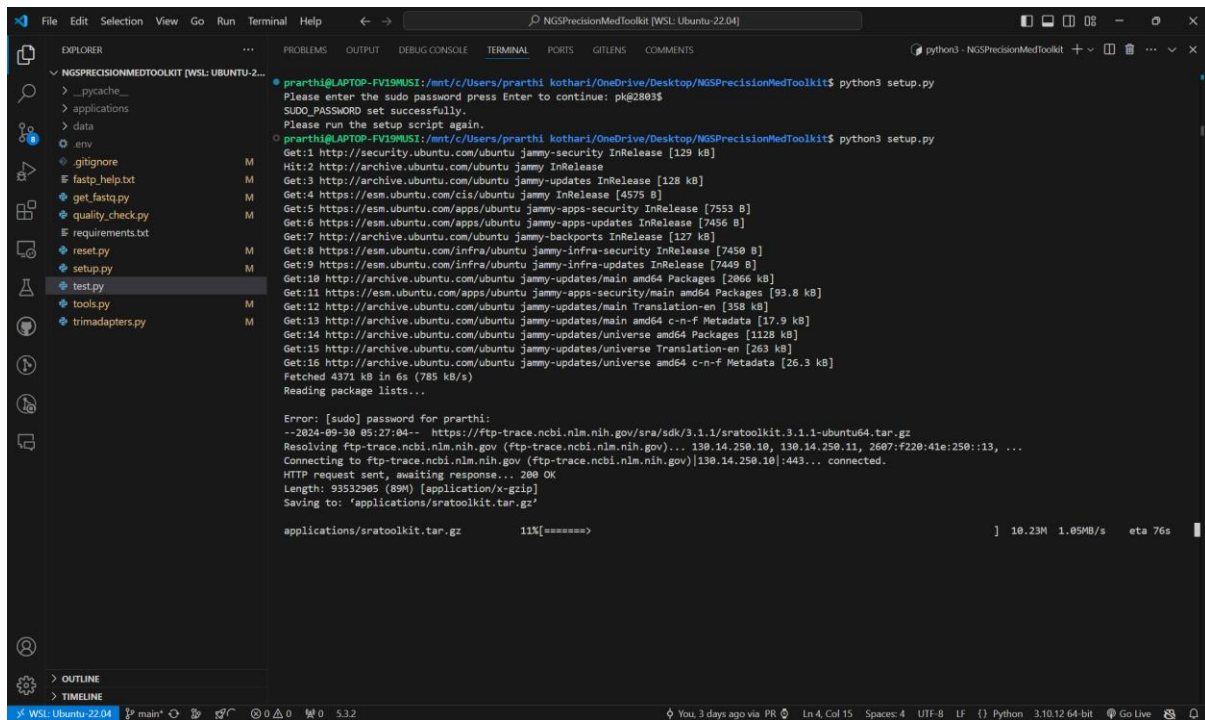
```

parthi@LAPTOP-FV19WU51: /mnt/c/Users/parthi.kothari/OneDrive/Desktop/NGSPrecisionMedToolkit\$

Figure 7: Code to check the quality of the reads using Fastp tool

# **CHAPTER 7: RESULTS**

# RESULTS

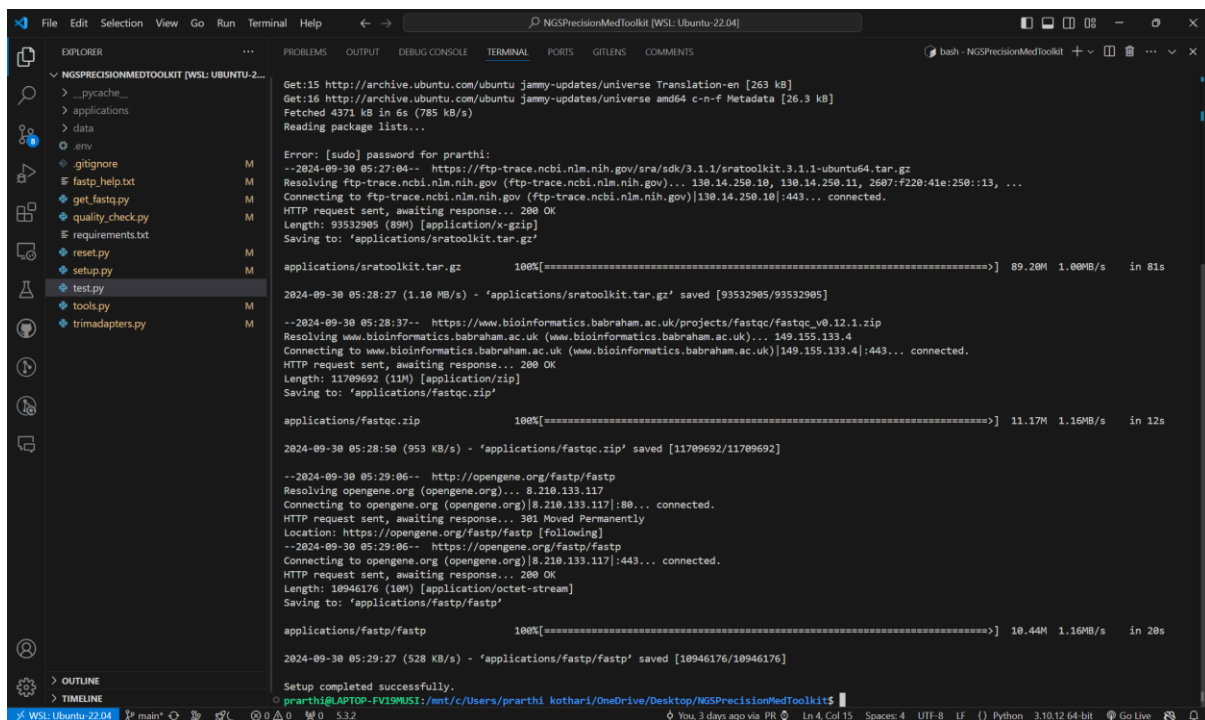


```
prarthi@LAPTOP-FV19WJ51: /mnt/c/Users/prarthi/OneDrive/Desktop/NGSPrecisionMedToolkit$ python3 setup.py
Please enter the sudo password press Enter to continue: pk@2803$
SUDO PASSWORD set successfully.
Please run the setup script again.
prarthi@LAPTOP-FV19WJ51: /mnt/c/Users/prarthi/OneDrive/Desktop/NGSPrecisionMedToolkit$ python3 setup.py
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 https://esm.ubuntu.com/cis/ubuntu jammy InRelease [4575 B]
Get:5 https://esm.ubuntu.com/apps/ubuntu jammy-apps-security InRelease [7553 B]
Get:6 https://esm.ubuntu.com/apps/ubuntu jammy-apps-updates InRelease [7456 B]
Get:7 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:8 https://esm.ubuntu.com/infra/ubuntu jammy-infra-security InRelease [7450 B]
Get:9 https://esm.ubuntu.com/infra/ubuntu jammy-infra-updates InRelease [7449 B]
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2066 kB]
Get:11 https://esm.ubuntu.com/apps/ubuntu jammy-apps-security/main amd64 Packages [93.8 kB]
Get:12 http://archive.ubuntu.com/ubuntu jammy-updates/main Translation-en [358 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 c-n-f Metadata [17.9 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1128 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe Translation-en [263 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 c-n-f Metadata [26.3 kB]
Fetched 4371 kB in 6s (785 kB/s)
Reading package lists...

Error: [sudo] password for prarthi:
--2024-09-30 05:27:04-- https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/3.1.1/sratoolkit.3.1.1-ubuntu64.tar.gz
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.10, 130.14.250.11, 2607:f220:41e:250::13, ...
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|130.14.250.10|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 93532905 (89M) [application/x-gzip]
Saving to: 'applications/sratoolkit.tar.gz'

applications/sratoolkit.tar.gz      11K[=====] 10.23M  1.05MB/s  eta 76s
```

Figure 8: Setting up the environment and downloading required dependencies



```
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe Translation-en [263 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 c-n-f Metadata [26.3 kB]
Fetched 4371 kB in 6s (785 kB/s)
Reading package lists...

Error: [sudo] password for prarthi:
--2024-09-30 05:27:04-- https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/3.1.1/sratoolkit.3.1.1-ubuntu64.tar.gz
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 130.14.250.10, 130.14.250.11, 2607:f220:41e:250::13, ...
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|130.14.250.10|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 93532905 (89M) [application/x-gzip]
Saving to: 'applications/sratoolkit.tar.gz'

applications/sratoolkit.tar.gz      100K[=====] 89.20M  1.00MB/s  in 81s

2024-09-30 05:28:27 (1.10 MB/s) - 'applications/sratoolkit.tar.gz' saved [93532905/93532905]

--2024-09-30 05:28:37-- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip
Resolving www.bioinformatics.babraham.ac.uk (www.bioinformatics.babraham.ac.uk)... 149.155.133.4
Connecting to www.bioinformatics.babraham.ac.uk (www.bioinformatics.babraham.ac.uk)|149.155.133.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 11709692 (11M) [application/zip]
Saving to: 'applications/fastqc.zip'

applications/fastqc.zip            100K[=====] 11.17M  1.16MB/s  in 12s

2024-09-30 05:28:50 (953 KB/s) - 'applications/fastqc.zip' saved [11709692/11709692]

--2024-09-30 05:29:06-- http://opengene.org/fastp/fastp
Resolving opengene.org (opengene.org)... 8.210.133.117
Connecting to opengene.org (opengene.org)|8.210.133.117|:80... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://opengene.org/fastp/fastp [following]
--2024-09-30 05:29:06-- https://opengene.org/fastp/fastp
Connecting to opengene.org (opengene.org)|8.210.133.117|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 10946176 (10M) [application/octet-stream]
Saving to: 'applications/fastp/fastp'

applications/fastp/fastp          100K[=====] 10.44M  1.16MB/s  in 20s

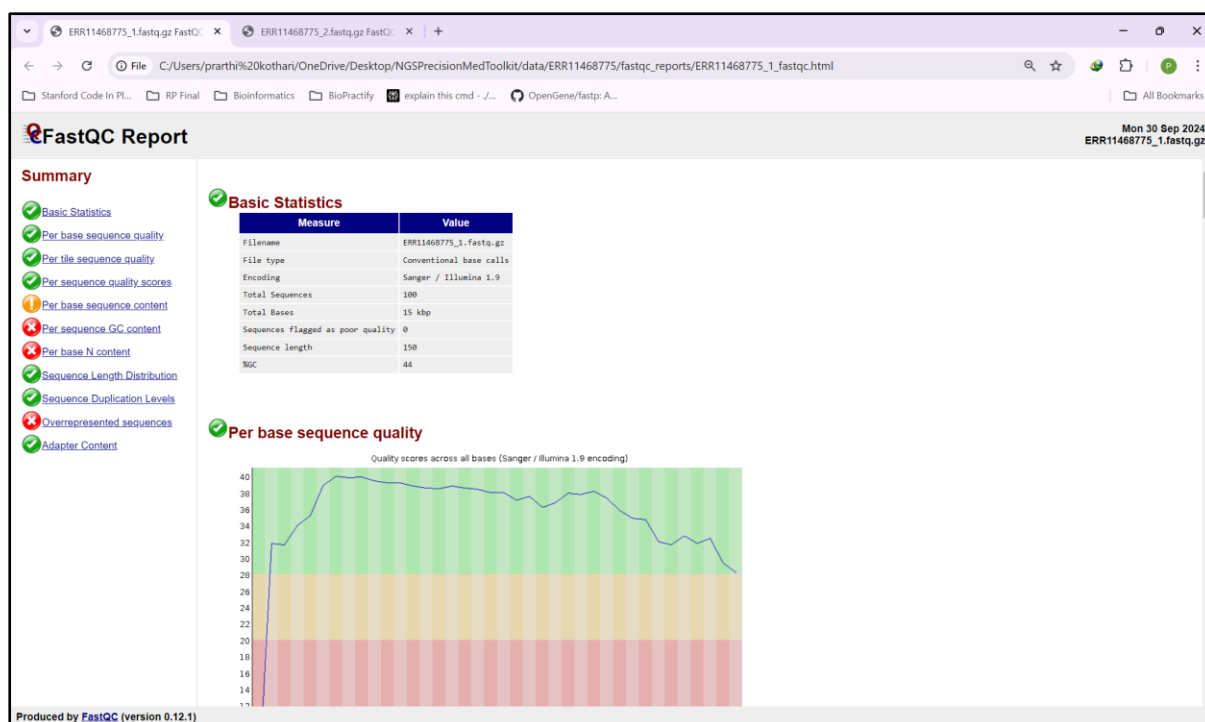
2024-09-30 05:29:27 (528 KB/s) - 'applications/fastp/fastp' saved [10946176/10946176]

Setup completed successfully.
prarthi@LAPTOP-FV19WJ51: /mnt/c/Users/prarthi/OneDrive/Desktop/NGSPrecisionMedToolkit$
```

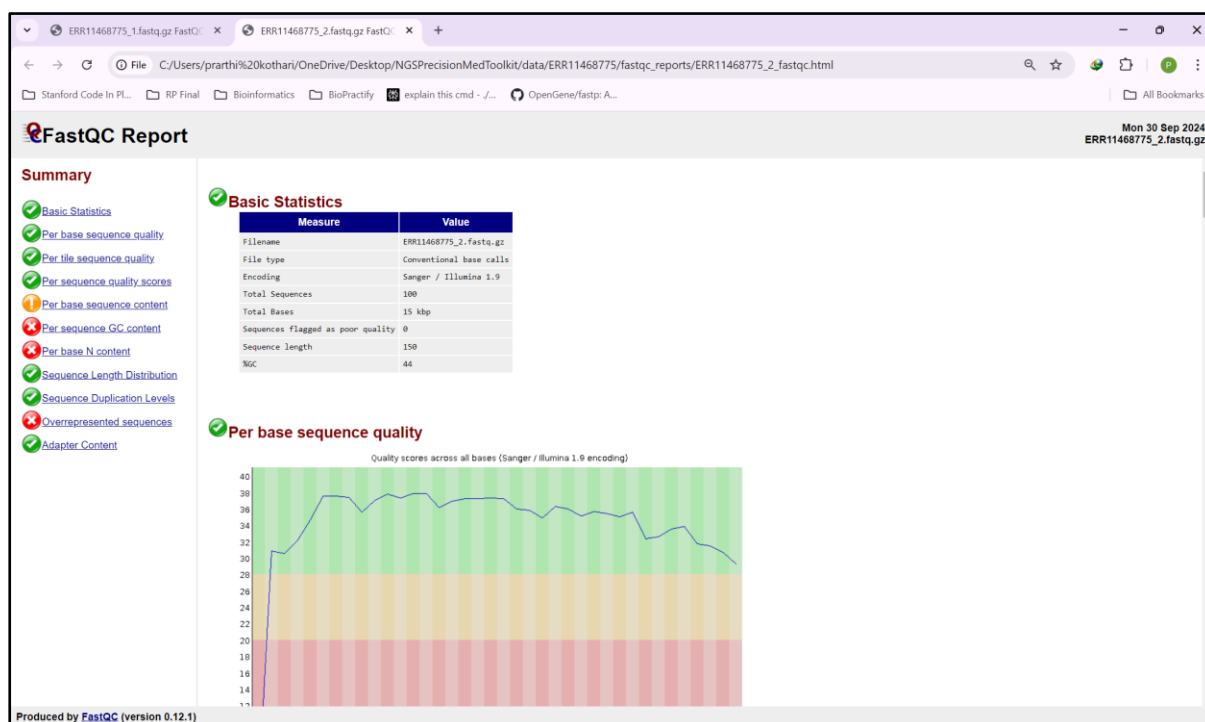
Figure 9: Set – up completed successfully







**Figure 12: FASTQC Report in the form of HTML file for the forward read of the sequence for the SRA ID: ERR11468775**



**Figure 13: FASTQC Report in the form of HTML file for the reverse read of the sequence for the SRA ID: ERR11468775**

# **CHAPTER 8: EXPECTED OUTCOME**

## **EXPECTED OUTCOME**

The development of an integrated Next-Generation Sequencing (NGS) workflow with a user-friendly Graphical User Interface (GUI) is anticipated to yield a comprehensive and efficient tool that significantly enhances the accessibility and usability of genomic data analysis in precision medicine. The expected outcomes are outlined as follows:

- 1. Enhanced Usability:** The GUI will be designed to facilitate intuitive navigation through the NGS workflow, allowing users—particularly clinicians and researchers without extensive computational backgrounds—to easily manage and analyze sequencing data. The incorporation of interactive elements, such as buttons and progress indicators, will further improve user engagement and satisfaction.
- 2. Streamlined Workflow Integration:** By integrating all major steps of the NGS pipeline—from raw data acquisition to variant calling and reporting—the tool will provide a centralized platform for managing complex analyses. This integration is expected to reduce the time and effort required for data processing, enabling users to focus on interpretation rather than technical details.
- 3. Customization Capabilities:** The ability to customize various parameters within the NGS workflow will empower users to tailor analyses according to specific experimental needs. This flexibility is crucial for accommodating diverse research questions and enhancing the relevance of the findings.
- 4. Improved Data Quality and Accuracy:** Rigorous testing and validation phases will ensure that the pipeline maintains high standards of data integrity and accuracy. The application of quality control measures, such as FastQC and Fastp, along with advanced alignment techniques using BWA or Bowtie, is expected to enhance the reliability of variant calling results.
- 5. Comprehensive Reporting:** The generation of detailed PDF reports summarizing the entire workflow will provide users with clear insights into their analyses. These reports will not only facilitate easy sharing among collaborators but also serve as valuable documentation for future reference.
- 6. Robust Documentation:** Comprehensive user manuals, developer documentation, and API guidelines will ensure that users can effectively utilize the software while providing a foundation for future enhancements and maintenance.
- 7. Support for Precision Medicine:** Ultimately, the project aims to contribute to advancements in precision medicine by providing a powerful tool that enables researchers to identify genetic variants relevant to personalized treatment strategies. By connecting to online databases for additional information on drugs and therapeutic targets, the tool will enrich analyses with actionable insights.

In conclusion, this integrated NGS workflow with a user-friendly GUI is expected to significantly advance genomic research capabilities, improve accessibility for end-users, and foster innovations in personalized healthcare through efficient data analysis and interpretation.

# **CHAPTER 9: CONCLUSION**

## CONCLUSION

Despite its promise, precision medicine faces several challenges. Critics argue that many of its anticipated benefits remain unfulfilled, particularly concerning complex diseases where genetic factors alone do not dictate treatment responses. Additionally, the high costs associated with new biotechnologies may exacerbate health inequalities and pose sustainability issues for healthcare systems, especially in low- and middle-income countries.

Moving forward, continued investment in research is essential to overcome these barriers. There is a need for improved diagnostic methods and treatments that are accessible and affordable for diverse populations. As precision medicine evolves, it holds the potential not only for enhanced individual care but also for broader public health improvements through more effective disease management strategies.

As NGS technology continues to evolve, several future directions and challenges emerge:

- 1. Integration with Clinical Practice:** The integration of NGS into routine clinical practice remains a challenge due to regulatory hurdles and the need for standardized protocols. Ensuring that genomic data is interpreted correctly and used effectively in clinical decision-making is crucial for maximizing its benefits.
- 2. Ethical Considerations:** The implications of genetic information raise ethical concerns regarding privacy, consent, and potential discrimination based on genetic predispositions. Addressing these issues will be essential as NGS becomes more prevalent in both research and healthcare settings.
- 3. Data Interpretation:** While data generation has become more efficient, interpreting the vast amounts of genomic data remains complex. Continued advancements in bioinformatics tools and algorithms are necessary to translate genomic findings into actionable insights effectively.

As technology continues to evolve, the role of GUIs in NGS workflows will likely expand further:

- 1. Integration with Artificial Intelligence:** Future GUIs may incorporate AI-driven analytics that assist users in interpreting complex genomic data more effectively.
- 2. Mobile Accessibility:** With the rise of mobile technology, developing mobile-friendly GUIs could allow researchers to access NGS tools on-the-go, increasing flexibility in data analysis.
- 3. Enhanced Visualization Techniques:** Continued advancements in visualization techniques will improve how genomic data is presented within GUIs, making it easier for users to understand intricate patterns and relationships within the data.

The detailed methodology outlines a structured approach to developing an integrated NGS workflow with a user-friendly GUI. By systematically addressing each phase—from problem identification and literature review to pipeline design, backend integration, validation, GUI development, and final release—the project ensures the creation of a robust, efficient, and accessible tool for precision medicine applications. Comprehensive documentation and

iterative user feedback further guarantee that the final product meets the needs of its target users, fostering advancements in genomic research and personalized healthcare.

In summary, Graphical User Interfaces are vital for enhancing the usability and efficiency of NGS workflows. Their implementation not only streamlines processes but also supports advancements in precision medicine by making genomic research more accessible and collaborative. As technology progresses, GUIs will continue to play a crucial role in shaping the future of genomic analysis and personalized healthcare solutions.

---

# **CHAPTER 10: REFERENCES**

## **REFERENCES**

1. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
2. Appiah, Vincent. (2021). Download Fastq Data using SRA Toolkit. [https://www.researchgate.net/publication/355651665\\_Download\\_Fastq\\_Data\\_using\\_SRA\\_Toolkit](https://www.researchgate.net/publication/355651665_Download_Fastq_Data_using_SRA_Toolkit)
3. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
4. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* (Oxford, England), 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
5. Ewels, P., Magnusson, M., Lundin, S., & Källér, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (Oxford, England), 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
6. Jameson, J. L., & Longo, D. L. (2015). Precision medicine--personalized, problematic, and promising. *The New England journal of medicine*, 372(23), 2229–2234. <https://doi.org/10.1056/NEJMsbl503104>
7. Khetani, M. P. R. (2018, September 5). Quality control: Assessing FASTQC results. Introduction to RNA-Seq using high-performance computing - ARCHIVED. [https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc\\_fastqc\\_assessment.html](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html)
8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England), 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
9. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, 17(1), 122. <https://doi.org/10.1186/s13059-016-0974-4>
10. Morash, M., Mitchell, H., Beltran, H., Elemento, O., & Pathak, J. (2018). The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *Journal of personalized medicine*, 8(3), 30. <https://doi.org/10.3390/jpm8030030>
11. Morash, M., Mitchell, H., Beltran, H., Elemento, O., & Pathak, J. (2018). The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *Journal of personalized medicine*, 8(3), 30. <https://doi.org/10.3390/jpm8030030>
12. Ncbi. (n.d.). Home. GitHub. <https://github.com/ncbi/sra-tools/wiki>
13. Ren, S., Ahmed, N., Bertels, K., & Al-Ars, Z. (2019). GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller. *BMC genomics*, 20(Suppl 2), 184. <https://doi.org/10.1186/s12864-019-5468-9>
14. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8), 912–918. <https://doi.org/10.1038/ng.3036>



15. Schmid, S., Jochum, W., Padberg, B., Demmer, I., Mertz, K. D., Joerger, M., Britschgi, C., Matter, M. S., Rothschild, S. I., & Omlin, A. (2022). How to read a next-generation sequencing report-what oncologists need to know. *ESMO open*, 7(5), 100570. <https://doi.org/10.1016/j.esmoop.2022.100570>
  16. Tenenbaum, J. D., Avillach, P., Benham-Hutchins, M., Breitenstein, M. K., Crowgey, E. L., Hoffman, M. A., Jiang, X., Madhavan, S., Mattison, J. E., Nagarajan, R., Ray, B., Shin, D., Visweswaran, S., Zhao, Z., & Freimuth, R. R. (2016). An informatics research agenda to support precision medicine: seven key areas. *Journal of the American Medical Informatics Association : JAMIA*, 23(4), 791–795. <https://doi.org/10.1093/jamia/ocv213>
-