IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 March 2025, accepted 12 March 2025, date of publication 17 March 2025, date of current version 28 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3552069

RESEARCH ARTICLE

Cyber-Biosecurity Challenges in Next-Generation Sequencing: A Comprehensive Analysis of Emerging Threat Vectors

NASREEN ANJUM[1], HANI ALSHAHRANI[2,3], (Senior Member, IEEE), ASADULLAH SHAIKH[3,4], (Senior Member, IEEE), MAHREEN-UL-HASSAN[5], MEHREEN KIRAN[6], SHAH RAZ[5], AND ABU ALAM[7]

[1]School of Computing, University of Portsmouth, PO1 2UP Portsmouth, U.K.
[2]Department of Computer Science, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia
[3]Emerging Technologies Research Laboratory (ETRL), College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia
[4]Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia
[5]Department of Microbiology, Shaheed Benazir Bhutto Women University, Peshawar, Khyber Pakhtunkhwa 00384, Pakistan
[6]Department of Computer Science, Anglia Ruskin University, CB1 1PT Cambridge, U.K.
[7]Department of Cyber Security and Computing, University of Gloucestershire, GL50 2RH Cheltenham, U.K.

Corresponding author: Asadullah Shaikh (asshaikh@nu.edu.sa)

This research was supported by the UK-Saudi Challenge Fund (2024-2025), sponsored by the British Council. Additionally, this work was supported by the Quality-Related (QR) Research Grant from the University of Portsmouth (UoP), UK.

ABSTRACT Next-generation sequencing (NGS) has transformed genomic research and healthcare by enabling the rapid and cost-effective sequencing of DNA and RNA, surpassing traditional techniques such as Sanger sequencing. This technological leap has had a profound impact on fields including biomedical research, personalised medicine, cancer genomics, agriculture, and forensic sciences. With its widespread adoption, NGS has made genomic information more accessible, facilitating the sequencing of millions of genomes. However, the growing reliance on NGS has also brought significant challenges related to cyber-biosecurity, particularly the protection of genomic data against cyber threats such as unauthorised access, data breaches, and exploitation. Genomic data is inherently sensitive, and vulnerabilities in NGS technologies, software, data-sharing practices, and open-access databases expose it to risks concerning data confidentiality, integrity, and privacy. While NGS data plays an indispensable role across numerous sectors, research addressing the cyber-biosecurity of these technologies remains fragmented. Most existing studies focus narrowly on specific areas, such as microbial sequencing or system architecture, and fail to provide a holistic perspective on the security challenges that span the entire NGS workflow. Additionally, the lack of interdisciplinary collaboration between the biotechnology and cybersecurity communities further exacerbates these gaps. This paper seeks to bridge these gaps by thoroughly examining cyber-biosecurity threats throughout the NGS workflow. It introduces a tailored taxonomy specifically designed for NGS, aimed at increasing stakeholder awareness of potential vulnerabilities and threats. Key insights include identifying vulnerabilities at various stages of the NGS process—from data generation to analysis and storage—and categorising these threats systematically. The study highlights critical gaps in current research, underscoring the need for interdisciplinary collaboration between experts in biotechnology and cybersecurity. It calls for focused efforts to mitigate risks associated with unauthorised access, data misuse, and exploitation. Failure to address these vulnerabilities could result in severe consequences, such as breaches of medical confidentiality, ethical concerns, and the potential for misuse in malicious applications like genetic warfare or bioterrorism. By providing a comprehensive analysis, this paper advocates for

The associate editor coordinating the review of this manuscript and approving it for publication was Mouquan Shen.

intensified research efforts and collaborative strategies to protect genomic data and ensure its ethical and secure use.

INDEX TERMS Next generation sequencing (NGS), cyber-biosecurity, genomic data security, malware attack, genomic data privacy, cyber attacks on NGS.

I. INTRODUCTION

Next generation sequencing (NGS) technology has revolutionized genomic research and healthcare by enabling high-throughput sequencing of Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), significantly faster and cheaper than traditional methods like Sanger Sequencing [1]. Millions of people now have access to their personal genomics information due to advances in NGS technology and decreasing sequencing costs. According to MIT Technology [2], by the start of 2019, more than 26 million people had taken an at-home ancestry test. By the end of 2025, it is estimated that approximately 60 million people worldwide will have their genome sequenced, indicating the widespread utilization of this technology [3].

In healthcare, NGS has paved the way for groundbreaking discoveries. By enabling a deeper understanding of the genetic basis of diseases, genome sequencing will be routinely used for many fields of research such as biomedical sciences, forensic sciences, cancer genomics, agriculture, personalized medicine, drug development, criminal investigations, environmental monitoring and more [4]. It is further anticipated that this advancing technology will facilitate the generation of sequencing data beyond the traditional laboratory environments. Such data collection is envisioned to occur at investigation sites like crime scenes, as well as directly from consumers or patients, and in other unconventional locations such as space [5]. However, with the continuous growth of NGS, new concerns about cyber-biosecurity are emerging regarding the highly sensitive genomic data produced by these technologies. Cyber-biosecurity refers to the protection of biological information, particularly genomic data, from cyber threats such as unauthorized access, data breaches, theft, and exploitation [6]. Addressing cyber-biosecurity requires transdisciplinary expertise in cybersecurity, biosecurity, and biotechnology [7].

NGS technology faces cyber-biosecurity threats due to the sensitive nature of genomic data, vulnerabilities in NGS technologies and software, risks in data sharing and collaboration, challenges in integrating multiple data sources, and the vulnerability of open access NGS databases. For instance, in the recent Synnovis cyberattack [8], which targeted NHS England's blood test provider, resulted in the theft of sensitive diagnostic data, including blood test results, which were subsequently leaked online. This breach has raised significant concerns about patient privacy and data security. The sensitive nature of such datasets, when combined with genomic information, could facilitate re-identification attacks. For instance, attackers could leverage stolen blood test data alongside genomic data to uncover individuals' identities or exploit genetic predispositions for malicious purposes. In April 2024, Octapharma Plasma, the U.S. arm of the Swiss pharmaceutical company Octapharma, suffered a ransomware attack attributed to the Blacksuit ransomware group [9]. The breach compromised sensitive personal information, prompting the company to notify affected individuals and regulatory authorities. Similarly, in June 2017, Merck was among several organizations hit by the NotPetya ransomware attack, causing extensive disruptions [10], and in June 2024, Japanese pharmaceutical company Eisai experienced a ransomware attack that disrupted logistics and production, delaying product shipments [11].

In the context of NGS, such breaches highlight the potential risks of exposing highly sensitive genomic data stored or analyzed in healthcare and pharmaceutical systems. Attackers could exploit this data for targeted re-identification attacks, genetic profiling, or even unethical research. Furthermore, ransomware and malware attacks, like those experienced by Merck & Co. and Bayer AG, demonstrate how such incidents can cripple critical infrastructures, halting genomic sequencing operations, corrupting sequencing data, and causing significant setbacks in scientific research and clinical diagnostics. These cases underscore the urgent need for robust cybersecurity measures to safeguard genomic data and maintain the integrity of NGS workflows. Unfortunately, these cyber-biosecurity attacks can occur at various stages of the data life cycle, from data generation to sharing and analysis. If genetic data is stored on insecure servers or databases, it becomes vulnerable to unauthorized access. Attackers who gain access to both the genetic data and associated metadata can potentially re-identify individuals. These threats jeopardize data confidentiality, compromise accuracy, and pose risks to individual privacy and medical confidentiality. However, despite the significant investments in genomic data generation and the pivotal role of NGS data in various fields, there remains a conspicuous absence of comprehensive research and discourse on cyber-biosecurity concerning NGS data. This gap is likely due to a lack of cyber security awareness and comprehension within the engineering and computing communities regarding the significance of NGS technology and its data [12]. Furthermore, executing such attacks necessitates extensive interdisciplinary knowledge encompassing computer science, bioinformatics, biotechnology, and microbiology. Nonetheless, the emergence of advanced AI chatbots raises concerns about the potential escalation in the likelihood of such attacks. With AI's capacity to bridge this knowledge gap, there is apprehension that it could facilitate the execution of

these malicious activities. For instance, as mentioned in [13], AI tools like CHATGPT could potentially be leveraged to design and implement biological weapons and attacks.

A. RELATED WORK

Despite the critical importance of NGS technologies and the vast amounts of data they generate, only a handful of studies have been conducted to provide an overview of the security challenges posed by NGS technology. The authors in [14] discuss the cyber security risks associated with the increased use of NGS in public health microbiology. The authors propose policy considerations aimed at mitigating risks, such as implementing robust cybersecurity measures throughout the NGS workflow. The authors in [15] provide an overview of the considerations necessary for setting up NGS analysis architectures, primarily for scientific purposes. It extensively covers the technological, legal, and ethical challenges associated with NGS analysis, including in clinical environments. However, the paper does not delve deeply into security aspects, such as protection against unauthorized data breaches, side-channel leaks, poor authentication, etc., in the context of NGS workflow. The article [16] introduces the concept of the digital DNA lifecycle and addresses privacy attacks on DNA data and proposes countermeasures to safeguard privacy during the storage, querying, and sharing of DNA data. The paper by Peccoud et al. [17] highlights the importance of recognizing and addressing the cyber-biosecurity risks in the biotechnology field. The authors in [18] discuss the cyber-bio security vulnerabilities and challenges associated with pathogen genome databases. The paper, authored by Garrett [19] discusses the security risks associated with genetic information systems. It emphasizes the need for improved security measures in the biotechnology field, especially given the rapid advancements in DNA sequencing and synthesis technologies. With the advancement of genetic technologies, the paper [19] discusses the challenges in creating effective regulatory frameworks to protect genetic information. The paper emphasizes the need for comprehensive policies that balance innovation with security and ethical considerations.

B. RATIONALE AND SIGNIFICANCE OF THE STUDY

Existing research [14], [15], [16], [17], [18] have explored various aspects of NGS security, including microbial sequencing applications, system architecture vulnerabilities, and genomic data protection. However, these works primarily focus on traditional cybersecurity concerns such as data privacy, access control, and encryption, while emerging cyber threats specific to sequencing workflows remain largely unexplored. Moreover, current cybersecurity frameworks fail to account for sequencing-specific attack vectors, including Genomic Inference Attacks, DNA-Encoded Malware Attacks, and Multiplexed DNA Injection Attacks. While existing models prioritize data protection and secure storage, they lack mechanisms to safeguard active sequencing pipelines from targeted cyber threats. Additionally, limited research bridges the gap between cybersecurity, bioinformatics, and sequencing technologies, leading to a lack of interdisciplinary approaches in genomic security. Additionally, synthetic DNA-based malware, adversarial manipulation of sequencing workflows, and AI-driven genomic data breaches remain absent from current cybersecurity taxonomies, leaving a critical gap in biosecurity risk assessment. Addressing these new and evolving threats requires a comprehensive cybersecurity framework that extends beyond traditional data security measures to include sequencing-stage protections and bioinformatics-specific attack mitigation strategies.

Our studies addresses these gaps and advances cyber-biosecurity research by introducing the first structured taxonomy of cyber threats across the entire NGS workflow, systematically mapping vulnerabilities from raw sequencing to bioinformatics analysis and interpretation. Unlike traditional cybersecurity taxonomies—such as MITRE ATT&CK [20] and STRIDE [21], which focus on network security and software vulnerabilities—this study extends threat modeling to biological and computational risks unique to genomic sequencing. This novel classification framework provides a structured approach to identifying, assessing, and mitigating emerging cyber threats in genomics, setting a foundation for real-world risk assessment and security implementation in genomic research.

C. RESEARCH QUESTIONS ADDRESSED IN THIS STUDY

Given these challenges, this study seeks to answer the following research questions:

1) How can a structured cyber-biosecurity threat taxonomy enhance risk assessment and mitigation strategies across the NGS workflow?
2) What are the key vulnerabilities at each stage of the NGS workflow, and how can they be systematically categorized into a comprehensive taxonomy?
3) How do novel cyber threats, such as synthetic DNA-based malware and AI-driven genomic attacks, impact the integrity and security of NGS data?
4) What cybersecurity best practices and countermeasures can be implemented to mitigate emerging cyber-biosecurity risks in NGS?

This study further expands cybersecurity research by introducing and analyzing new types of cyber threats specific to genomic workflows, such as Genomic Inference Attack, DNA-Encoded Malware Attack, Multiplexed DNA Injection Attack, and Genetic Imputation Attack. These threats pose significant risks to the confidentiality, integrity, and availability of genomic data, going beyond traditional cybersecurity concerns of database security and unauthorized access prevention.

In addition, we provide detailed information on the tools and technologies used at each step of the NGS workflow, highlighting potential areas that attackers could exploit to launch cyberattacks. Finally, we propose clear research

directions and effective mitigation techniques to address identified security threats, offering actionable recommendations to guide future research and practical implementations. By integrating these contributions, this study advocates for a resilient cybersecurity framework for NGS, addressing the ethical, technological, and security implications of neglecting cyber-biosecurity. The findings of this study are crucial for researchers, policymakers, and industry professionals seeking to ensure the long-term security and integrity of genomic data against evolving cyber threats.

The overall contributions of this study are as follows:

D. CONTRIBUTIONS

This study pushes the boundaries of genomic cybersecurity by introducing the first structured cyber-biosecurity taxonomy for NGS. The key contributions of this work are as follows:

1) **First Structured Taxonomy of Cybersecurity Threats in NGS:** Unlike traditional cybersecurity models (e.g., MITRE ATT&CK, STRIDE), this new taxonomy systematically categorizes both computational and biological cyber threats across all stages of the NGS workflow, filling a critical gap in genomic risk assessment.
2) **Interdisciplinary Cyber-Biosecurity Framework:** By bridging the knowledge gap between biotechnology and cybersecurity, we offer an integrated approach that facilitates collaboration between genomic researchers, forensic experts, and security professionals.
3) **Identification of Emerging Threat Vectors:** We analyze newly emerging risks, including synthetic DNA malware, adversarial AI genome editing, and sequencing pipeline vulnerabilities, providing the first comprehensive assessment of these attack vectors.
4) **Comprehensive Technological Evaluation**: An in-depth assessment of the tools and technologies utilized across various stages of the NGS workflow is conducted. This includes identifying vulnerabilities prone to cyberattacks and delivering actionable insights absent from prior fragmented studies.
5) **Actionable Security Recommendations**: We propose practical mitigation strategies, including secure sequencing protocols, encryption methodologies, and AI-enhanced anomaly detection, offering a real-world cybersecurity roadmap for genomic institutions.

E. ORGANIZATION OF STUDY

This article is organized as follows: Section II details the research methodology for conducting the study and taxonomy. Section III provides a brief overview of NGS technology and highlights its vulnerabilities to bio-cyber threats. Section IV, Section V, and Section VI explore the vulnerabilities and potential cyber threats in raw sequencing data generation, quality control, and bioinformatics analysis steps, respectively. Section VIII offers clear, practical future directions for mitigating the identified threats. Finally, Section IX summarizes the main points of the article concisely.

II. TAXONOMY, SCOPE, LIMITATIONS, AND RESEARCH METHODOLOGY OF THE STUDY

In this section, the research taxonomy is first discussed, detailing the sequential steps involved in the NGS workflow and emphasizing the importance of securing each stage against potential cyber-attacks. Following this, the scope and limitations of the study are addressed. Finally, the research methodology used to conduct this study is outlined in detail.

A. TAXONOMY

The NGS workflow encompasses several sequential steps. Each step is crucial for transforming raw biological samples into analyzable genetic data. To generate this useful digital genetic information, NGS technology relies on a sophisticated integration of communication systems, bioinformatics tools, software applications, algorithms, and hardware to execute a range of tasks, from the collection of DNA samples to data analysis, processing, and the generation of final decisions and reports. A vulnerability or cyberattack at any phase of the NGS workflow or entry point of its processes can jeopardise the overall integrity and reliability of sensitive data. Such disruptions can lead to data corruption, misinterpretation of results, and ultimately, incorrect conclusions. For instance, a malware attack during the library preparation phase could result in the creation of inaccurate or corrupted DNA fragment libraries [22], [23]. This type of attack can trigger a cascading effect across subsequent stages of the NGS workflow, such as the cluster generation phase, ultimately compromising the accuracy and trustworthiness of the entire sequencing process.

Based on this understanding, we devised a taxonomy to systematically identify vulnerabilities and potential bio-cyber threats at each step of the NGS process. Specifically, it enables a step-by-step evaluation of the workflow, starting from DNA sample collection, library preparation, and cluster generation, through sequencing, quality control, bioinformatics analysis, and final interpretation. By identifying critical entry points for cyber threats at each stage, the taxonomy facilitates the development of targeted security measures. Securing every step, from sample collection to data analysis, ensures the integrity, reliability, and confidentiality of genomic data, protecting against the potentially severe consequences of cyber threats.

Comparison Between Proposed and Traditional Cybersecurity Taxonomies: Traditional cybersecurity taxonomies, such as the MITRE ATT&CK [20] framework and STRIDE threat modeling [21], focus primarily on network security, software vulnerabilities, and insider threats. While these frameworks have been effective in IT-based cybersecurity, they fail to account for the unique attack surfaces present in NGS workflows—such as synthetic DNA-based malware, AI-driven genomic data manipulation, and adversarial attacks on sequencing pipelines.

TABLE 1. Comparison of traditional cybersecurity taxonomies and the proposed NGS cyber-biosecurity taxonomy.

Aspect	Traditional Cybersecurity Taxonomies (e.g., MITRE ATT&CK, STRIDE)	Proposed NGS Cyber-Biosecurity Taxonomy
Primary Focus	Network security, software vulnerabilities, insider threats	Bioinformatics security, sequencing pipeline attacks, genomic data integrity
Threat Landscape	Static digital threats (malware, unauthorized access, phishing)	Dynamic threats evolving throughout the sequencing workflow
Attack Techniques	Phishing, SQL injection, ransomware, DoS/DDoS attacks	Synthetic DNA malware, adversarial AI in sequencing, supply chain attacks in bioinformatics
Risk Assessment Scope	IT infrastructure security (firewalls, endpoint protection)	Comprehensive risk assessment across all NGS stages (sequencing, data processing, analysis, storage)
Vulnerable Assets	Servers, cloud systems, network endpoints	Sequencing instruments, raw sequencing data, bioinformatics pipelines, genomic databases
Security Focus	Protecting static digital information	Ensuring both **digital and biological data security** in real-time sequencing operations
Adaptability to NGS	Not tailored for genomic data workflows	Specifically designed for bioinformatics and sequencing pipeline security

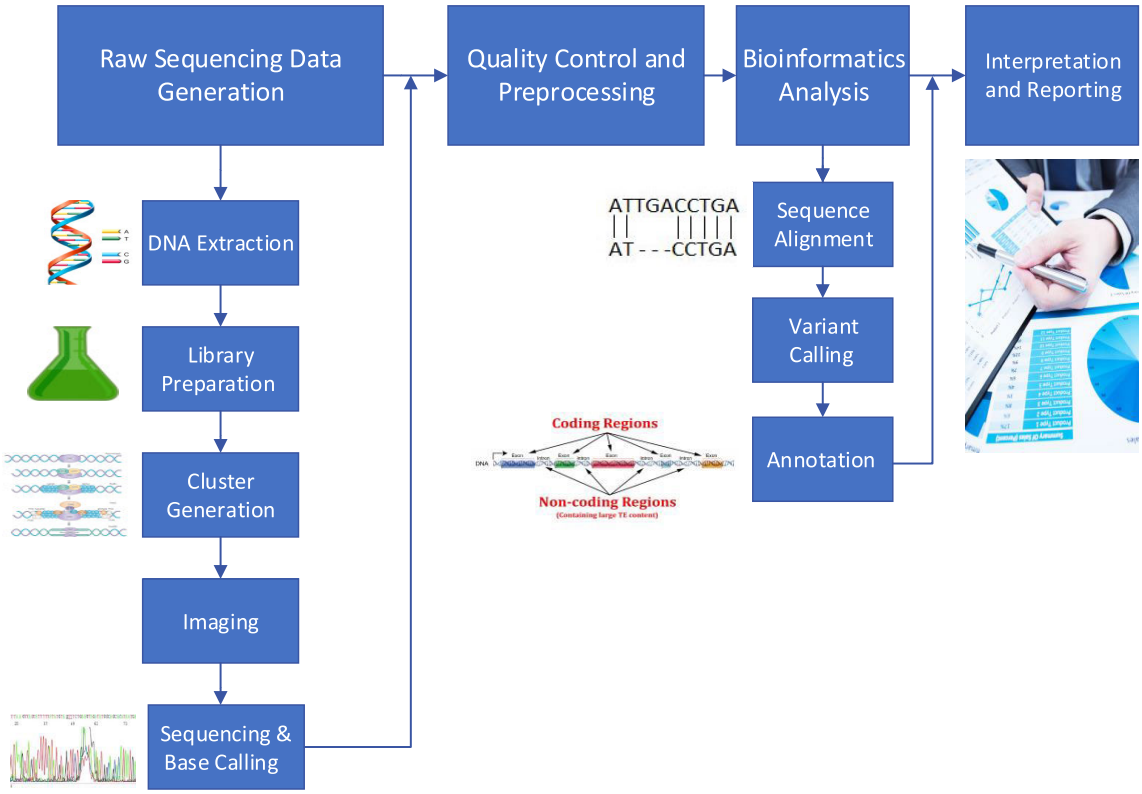


FIGURE 1. NGS workflow and research taxonomy.

This study extends traditional cybersecurity taxonomies by incorporating biological and computational vulnerabilities into a unified cyber-biosecurity framework. Unlike existing models that focus only on protecting static data repositories, this taxonomy captures the dynamic nature of genomic sequencing, where security risks evolve throughout the workflow. Table 1 compares our taxonomy to traditional cybersecurity frameworks, highlighting key gaps that this research addresses.

In our proposed taxonomy (Fig. 1), we have divided the NGS workflow into four major steps: (i) Raw Sequencing Data Generation, (ii) Quality Control and Preprocessing, (iii) Bioinformatics Analysis, and (iv) Interpretation and Analysis. Throughout the workflow, from raw sequencing data generation to interpretation and analysis, various types of files are generated to support quality control, bioinformatics analysis, and final interpretation. The details of these files are summarized in Table 2.

1) **Raw Sequencing Data Generation:** This step involves collecting DNA samples and producing raw sequencing data in digital formats. The data is typically stored as FASTQ or BAM files, which contain essential details

TABLE 2. Summary of files generated across NGS workflow steps.

Step	Description	Generated Files	File Formats	File Description
DNA Extraction	Isolation of DNA from biological samples	N/A	N/A	No files were generated at this step.
Library Preparation	Preparation of DNA for sequencing, including fragmentation, end repair, adapter ligation, and amplification	Library prep reports	PDF, CSV, TXT	Details protocols, reagent usage, and conditions for library preparation steps such as fragmentation, adapter ligation, and amplification efficiency.
Imaging	Capturing images of the sequencing flow cell during the sequencing process	Image files	TIFF, JPEG, PNG	High-resolution images of clusters on the sequencing flow cell, used to determine the presence and quality of DNA fragments.
Cluster Generation	Amplification of DNA fragments on the sequencing flow cell to create clusters	Cluster generation reports	PDF, CSV, TXT	Reports on the density and quality of DNA clusters generated on the flow cell, which are essential for optimal sequencing.
DNA Sequencing and Base Calling	Reading the DNA sequence using sequencing technology and converting raw signals to nucleotide sequences	Raw sequencing data, Base call files	FASTQ, BCL	FASTQ files contain raw sequence reads and their quality scores; BCL files are binary files with raw data from the sequencing machine before base calling.
Quality Control	Assessing the quality of sequencing data and base calling	Quality control reports	PDF, CSV, TXT, HTML	Contains metrics on read quality, adapter contamination, sequence duplication levels, and overall sequencing performance.
Sequence Alignment	Aligning sequencing reads to a reference genome	Alignment files	SAM, BAM	SAM files are text-based formats that contain aligned sequences; BAM files are the binary equivalent of SAM files, storing the same alignment information.
Variant Calling	Identification of genetic variants	Variant call files	VCF	VCF (Variant Call Format) files store information about variants detected in the sample, including SNPs, insertions, deletions, and their respective quality scores.
Annotation	Annotating genomic features, such as genes, exons, and regulatory elements	Annotation files	GFF, GTF	GFF/GTF files contain detailed information about genomic features like gene locations, exons, introns, and regulatory elements, mapping these features to the reference genome.
Interpretation and Reporting	Further analysis such as differential expression, phylogenetics, etc.	Analysis results	Various (TXT, CSV, TSV)	Various formats storing results of additional analyses like gene expression levels, phylogenetic trees, functional annotations, and statistical outputs.

from the workflow, including nucleotide sequences and their associated quality scores. Raw sequencing data generation is performed in controlled environments using NGS machines, such as Illumina HiSeq and MiSeq, along with integrated and supported software applications. These systems are designed to optimize the workflow, covering processes from cluster generation to signal detection and data analysis. They often feature user-friendly interfaces that enable laboratory personnel to easily configure, monitor, and adjust sequencing runs as needed. For instance, Illumina's Real-Time Analysis (RTA) software, embedded within Illumina sequencing platforms, performs real-time analysis of imaging data during sequencing runs to deliver immediate base calling, quality scoring, and error checking [24]. However, these systems are not immune to cyber threats. Attackers could exploit vulnerabilities to alter configuration settings or disrupt the sequencing process, compromising the integrity of the data. The final step of base calling, where raw sequencing data is converted into digital form, is particularly vulnerable to threats such as data manipulation or ransomware, which could render critical genetic data unusable. The key stages within this workflow include DNA extraction, library preparation, cluster generation, imaging, and base calling (Section IV).

2) **Quality Control and Preprocessing:** Before detailed analysis can be conducted, the raw sequencing data must undergo thorough cleaning and formatting. In this process, low-quality data is removed, and undesirable sequences are trimmed [25]. As this step focuses on improving data quality, it serves as a critical foundation for ensuring accurate analysis in the subsequent stages of the NGS workflow. Following this step, the refined data is fed into the bioinformatics analysis stage, where more complex and sophisticated computational processes are performed, including alignment, assembly, and variant calling (Section V).
3) **Bioinformatics Analysis:** After undergoing thorough data cleaning and quality assurance processes, the refined data is advanced to the bioinformatics analysis stage, where meaningful biological insights are derived. This digital biological data is crucial for understanding genetics, disease pathways, and evolutionary biology.

The sub-phases within this analysis stage are highly interdependent, with each relying on the output generated by the previous step. In some instances, analysing the refined dataset necessitates revisiting earlier stages when new data becomes available or additional quality checks are required. Each sub-phase employs sophisticated software, powerful computational tools, and specialised bioinformatics expertise to ensure accurate and reliable results.

This stage is particularly vulnerable to bio-cyber attacks due to its heavy reliance on a diverse array of technologies. Even a minor security breach can compromise the integrity and reliability of research outcomes, potentially stalling or casting doubt on scientific advancements. For instance, in genomic studies aimed at identifying genetic variants associated with diseases, an attacker could alter the data, making benign variants appear cancerous. Such manipulation and unauthorised access could mislead research efforts and yield inaccurate outcomes in clinical settings. Therefore, implementing robust practices to safeguard the security and privacy of bioinformatics workflows is crucial for maintaining the accuracy and reliability of research findings (Section VI).

4) **Interpretation and Analysis**: In this phase, the processed data generated during the bioinformatics analysis stage undergoes a detailed evaluation to extract critical insights. The primary goal of this stage is to address key research objectives and clinical questions. For example, it may involve diagnosing diseases, identifying mutations occurring during biological processes, or utilising genetic markers to predict a patient's response to specific treatments. The insights gained during this phase not only advance scientific discovery but also contribute to the design and development of personalised and precise medical strategies (Section VII).

B. LIMITATION AND SCOPE OF STUDY

The study on bio-cybersecurity threats in the context of NGS relies on theoretical analyses, expert opinions, and a review of available literature. While this approach offers valuable insights into potential vulnerabilities and threat vectors, it has limitations, primarily due to the lack of empirical data on actual cybersecurity incidents affecting NGS in public health. To date, no attacks have been reported on NGS technology. Many cybersecurity incidents, especially in sensitive fields like public health, are underreported due to concerns over reputational damage, regulatory repercussions, or data sensitivity. The evolving nature of genomics and adaptive cyber threats means there is little documented history of cyber attacks specific to this domain. Advanced persistent threats (APTs) and subtle data manipulations can remain undetected for long periods, complicating the collection of concrete empirical data.

Given these challenges, the study likely employs theoretical models to predict potential vulnerabilities and the impact of various cyber threats. Expert opinions are crucial as they combine cybersecurity principles with the specific challenges of genomic data handling.

It is important to note that this study specifically focuses on identifying vulnerabilities and attack vectors present at each stage of the NGS workflow. For readers interested in vulnerabilities specifically related to bioinformatics tools, software, and databases, we refer them to the comprehensive analyses provided in [18] and [26].

C. RESEARCH METHODOLOGY

The process of literature selection was conducted systematically, following PRISMA guidelines, and is detailed in the PRISMA flow diagram (Figure 2). Each step taken to refine the dataset is described below:

1) IDENTIFICATION

Articles were sourced from six comprehensive academic databases including, Web of Science,[1] Google Scholar,[2] IEEE Xplore,[3] Elsevier,[4] Springer Explorer,[5] and Frontiers Journals.[6]

The selection of keywords was guided by expert inputs. Keywords were categorized into primary and secondary groups to ensure a thorough and focused search. The primary keywords included terms such as "Next-Generation Sequencing," 'NGS," "High-throughput Sequencing," "Whole Genome Sequencing," "NGS workflow," "NGS Steps," and "NGS Applications." These primary terms were systematically combined with secondary keywords such as "Cyber Security," "Cyber threats," "Cyber attacks," "CyberbioSecurity," and specific attack types like "Data Breaches," "Malware Attacks," "Ransomware Attacks," "Denial-of-Service (DoS) Attacks," "Phishing Attacks," "Insider Threats," and "Vulnerabilities." Boolean operators ("AND" and "OR") were used to maximize the search's precision and coverage.

The search initially yielded 3,332 articles. These articles were distributed equally among three reviewers, who independently assessed their assigned articles using predefined relevance criteria. Any discrepancies among the reviewers were resolved through group discussions to ensure consistency, objectivity, and minimal bias.

2) SCREENING

During the screening phase, irrelevant records were removed, reducing the dataset significantly. Articles un-related to NGS cybersecurity, those focusing on other technologies such as IoT, blockchain, or network infrastructure, as well as patents, non-English articles, and duplicate records, were excluded.

[1] https://www.webofscience.com/wos/woscc/basic-search
[2] https://scholar.google.com/
[3] https://ieeexplore.ieee.org/Xplore/home.jsp
[4] https://www.elsevier.com/en-gb
[5] https://link.springer.com/
[6] https://www.frontiersin.org/

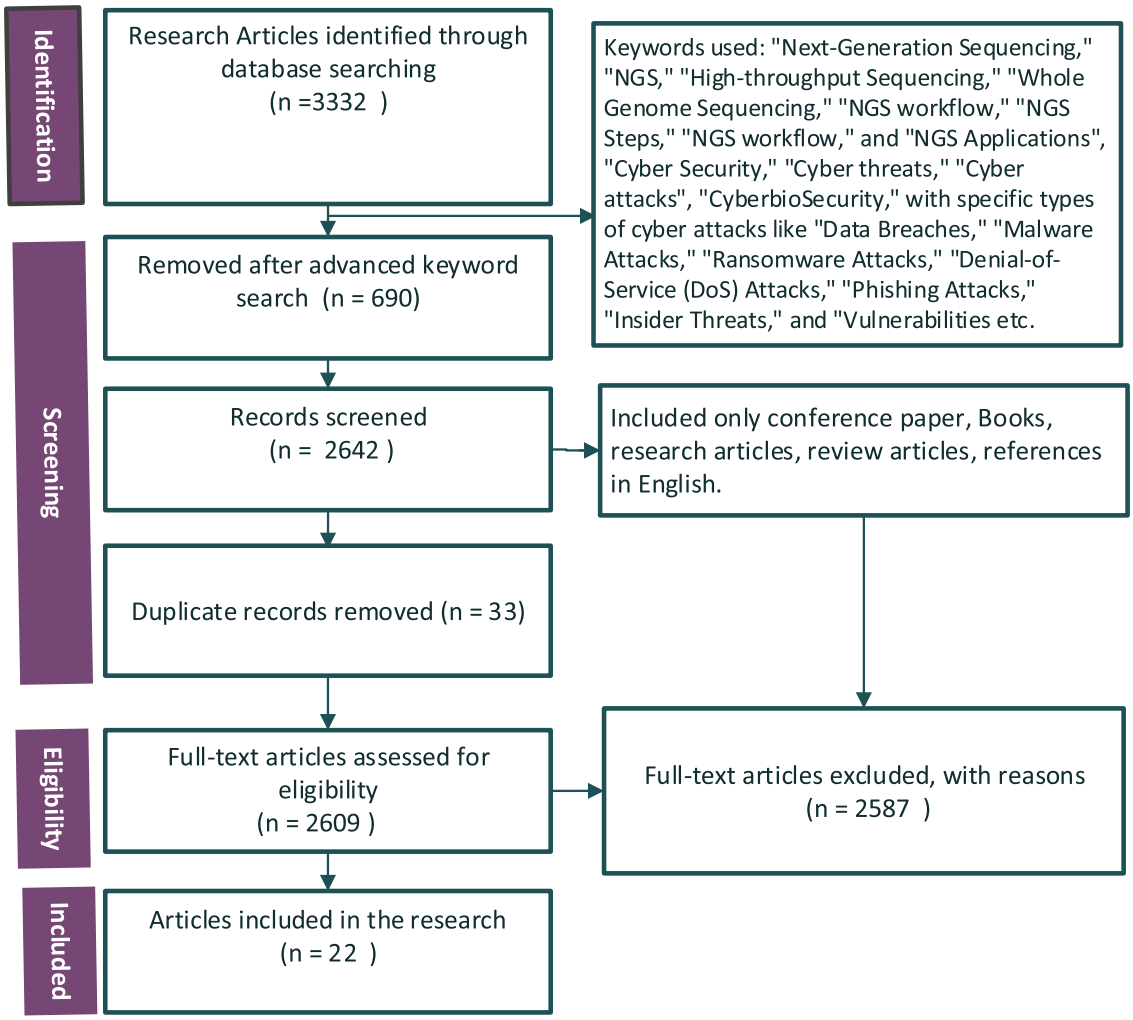


FIGURE 2. PRISMA flow diagram.

The refined results left 690 articles, which were further narrowed down by applying additional filters. These filters excluded 33 duplicate records and 635 irrelevant articles, leaving a total of 22 research articles specifically focused on cyber-biosecurity threats within the NGS workflow.

3) ELIGIBILITY

Following the initial screening, we analyzed the whole content of the remaing 22 articles for their eligibility to include in the study. Each reviewer ensure that each article should address the aim and objectives of the underlined studies i.e., cyber-bio security issues in NGS including mitigation strategies.

4) INCLUSION

Following the eligibility assessment, the final thouroughly analyzed 22 articles were proposed by all reviewrs to include in the study for further analysis due to their storong relevency to the problem domain. These articles formed the basis of the research findings and discussions presented in this study.

III. BACKGROUND ON NGS TECHNOLOGY

This section provides a concise overview of the history and advancements in NGS technology. It then explores the diverse applications of NGS across various fields. Lastly, it highlights the key factors that contribute to the vulnerability of NGS technology to cyber attacks.

A. BRIEF HISTORY AND EVOLUTION

DNA sequencing techniques have undergone significant evolution over the decades, driven by numerous technological breakthroughs. In the late 1970s, Frederick Sanger pioneered the first DNA sequencing method, known as Sanger sequencing, which enabled scientists to identify the precise order of DNA bases [27]. This revolutionary method played a pivotal role in the early stages of molecular biology research, despite its limitations in throughput and cost [28]. Sanger sequencing served as the groundwork for the development of more advanced techniques, including next-generation sequencing (NGS), which addressed these limitations by offering higher throughput and significantly lower costs. For instance, Sanger

sequencing was constrained by a DNA strand read length of approximately 700 to 1,000 base pairs [29], whereas NGS platforms have extended read lengths to an average of 10,000 to 15,000 base pairs [30]. Furthermore, while Sanger sequencing required individual reactions for each DNA fragment, NGS introduced the concept of parallelization, enabling the simultaneous sequencing of millions of DNA fragments in a single run.

NGS platforms can generate varying amounts of data per run, typically measured in terms of several gigabases, where one gigabase refers to one billion bases of DNA sequence or terabases, which are equivalent to thousands of gigabases. DNA sequencing now costs much less per base due to NGS. Large-scale sequencing initiatives have become more feasible because of the high throughput and parallelization of NGS, which allows scientists to sequence genes at a small fraction of the previous cost [31].

NGS technology employs a variety of methods that allow for the sequencing of millions to billions of DNA fragments simultaneously. Among these methods are:

1) **Whole-Genome Sequencing (WGS)**: WGS involves sequencing the entire genetic material of an organism. This method provides a thorough analysis of an individual's or organism's complete genetic profile, allowing for the detection of genetic variations such as insertions, deletions, and structural alterations [32].
2) **Whole-Exome Sequencing (WES)**: WES focuses specifically on sequencing the exonic regions of the genome, which are responsible for coding proteins. These exons represent only a small fraction of the entire genome [33].
3) **Chromatin Immunoprecipitation Sequencing:** ChIP-Seq is a technique that combines chromatin immunoprecipitation with next-generation sequencing to investigate protein-DNA interactions. This approach identifies DNA regions bound by specific proteins, such as transcription factors or histones, providing insights into gene regulation mechanisms [15].

B. APPLICATIONS OF NGS IN VARIOUS FIELDS

The advent of NGS technology has revoloutionlized a wide variety of fields including healthcare by facilitating fast, accurate, and parallel sequencing of large number of DNA and RNA. Table 3 provides a summary of the diverse applications of NGS across various disciplines. The detailed descriptions of these applications are elaborated below:

- **Transcriptomics**: NGS has revolutionized research by enabling comprehensive analysis at the gene level. It facilitates the study of genetic variations and the discovery of novel RNA molecules [34]. NGS technology enables the reconstruction of full-length RNA molecules by aligning and assembling short sequencing reads, generating genetic maps that represent the complete RNA content of a cell or tissue. Moreover, NGS has significantly advanced gene analysis at the single-cell level. Techniques such as Single-cell RNA Sequencing have emerged as powerful tools, aiding in the identification of cell types and enhancing our understanding of disease progression [35].
- **Cancer Genomics**: NGS plays a pivotal role in identifying genetic abnormalities associated with various cancers, such as bladder and lung cancer. It allows for the detailed investigation of cancer traits by detecting minor insertions, deletions, single nucleotide alterations, and structural abnormalities in cancer cells. Additionally, NGS supports the identification of chromosomal structural changes, which are critical in cancer development [36].
- **Metagenomics**: NGS enables the detailed analysis of microbial communities in various habitats, including the human gut, soil, and clinical samples. It facilitates the study of microbial diversity, identification, characterization, and taxonomic classification of microbes. By comparing sequencing reads to reference databases, such as 16S rRNA genes, researchers can investigate specific functional capabilities, including genes related to bioremediation and antibiotic resistance [37], [38].
- **Forensic Genomics**: NGS techniques significantly enhance forensic investigations by improving the analysis of DNA evidence for individual identification and relationship determination. It provides high-resolution genetic profiling, increasing the precision of DNA identification. Furthermore, NGS enables the development of databases for DNA, ancestry, and physical traits, supporting the simultaneous analysis of multiple samples and aiding in the identification of missing persons. Additionally, it assists in detecting specific bodily fluids in forensic samples, offering insights into the nature and origin of biological evidence [37].
- **Pharmacogenomics**: NGS aids in identifying genetic variations that influence drug responses and toxicity, helping predict individual reactions to medications and minimizing adverse effects. By analyzing genes involved in drug absorption and excretion, NGS enhances understanding of how genetic differences impact drug concentrations in the body. This knowledge enables clinicians to choose the most effective medications, determine optimal dosages, and design personalized treatment plans tailored to a patient's genetic profile [39].
- **Antimicrobial Resistance Profiling**: Antibiotic resistance, the ability of microorganisms to resist the effects of antibiotics designed to inhibit or kill them, poses a significant public health challenge. NGS streamlines the identification of resistance-related genes by sequencing microbial genomes. Additionally, it supports the study of mobile genetic elements, such as plasmids,[7] which play a crucial role in the spread of resistance. NGS

[7]Plasmids are small, circular DNA molecules that exist independently of chromosomal DNA and often carry antimicrobial resistance genes.

TABLE 3. Summary of applications of NGS technology.

Application of NGS	Description
Transcriptomics	Enables comprehensive gene-level analysis, RNA molecule discovery, reconstructs full-length RNA molecules, aids in single-cell gene analysis for cell type identification and disease progression understanding.
Cancer genomics	Detects genetic abnormalities linked to cancers like bladder and lung cancer, identifies minor alterations and structural abnormalities in cancer cells, including chromosomal changes crucial for cancer development.
Metagenomics	Allows comprehensive analysis of microbial communities, including diversity assessment, microbe identification, and taxonomic classification.
Forensic Genomics	Supports enhanced DNA analysis for individual identification and relationship determination, enables database creation for DNA, ancestry, and physical traits.
Pharmacogenomics	Enables the identification of genetic variations affecting drug responses and toxicity, aiding in predicting individual drug responses and reducing side effects.
Antimicrobial Resistance Profiling	Helps identify genes causing organisms to become resistant to antibiotics by sequencing their genomes. Additionally, it aids in discovering factors affecting gene resistance and clarifying the mechanisms by which resistance is maintained.

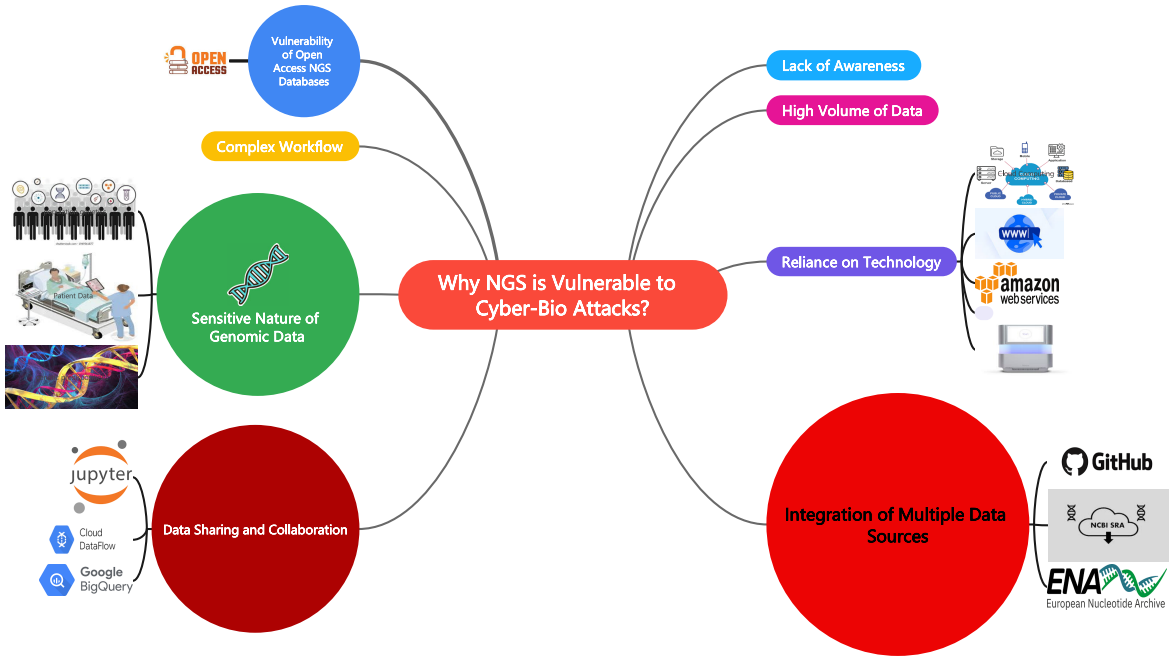


FIGURE 3. Causes of cyber-biosecurity threats in NGS.

also uncovers factors influencing resistance mechanisms and provides insights into the emergence of multidrug-resistant genes, offering valuable information for combating antibiotic resistance [40].

C. WHY NGS TECHNOLOGY IS VULNERABLE TO CYBER-BIOSECURITY THREATS?

NGS technology is at the forefront of genomics research, enabling rapid and detailed analysis of DNA sequences. However, its increasing use also introduces significant vulnerabilities to Cyber-BioSecurity threats, which can compromise data confidentiality and integrity. The key reasons for these vulnerabilities include the following (Figure. 3):

1) **High Volume of Data:** NGS generates massive amounts of genomic data [15], making it a valuable target for cyber attackers. The sheer volume of data also complicates monitoring and security, creating potential entry points for malicious activities.
2) **Sensitive Nature of Genomic Data:** Genomic data contains deeply personal information, such as genetic predispositions, ancestry, hereditary diseases, and population genetics [41]. Unauthorized access to this data can lead to significant privacy violations and ethical concerns. Breaches in bioinformatics can expose sensitive genomic data, risking individual privacy and medical confidentiality.
3) **Complex Workflows**: NGS involves multiple stages, from sample preparation to data analysis. Each stage employs various bioinformatics tools, software, and databases, increasing the attack surface and potential vulnerabilities at each step.

4) **Underreporting and Lack of Awareness:** Many incidents of cyber attacks in the genomics field go unreported due to concerns over reputational damage and regulatory repercussions. Additionally, there may be a lack of awareness and expertise in cybersecurity among genomic researchers and practitioners.
5) **Reliance on Technology:** NGS technology depends on various software tools, algorithms, and computational resources. These tools have complex interdependencies that cyber attackers can exploit through vulnerabilities in bioinformatics software or underlying infrastructure (servers, databases, cloud platforms). Attackers may inject malicious code into bioinformatics tools or target the infrastructure supporting DNA analysis, compromising the integrity of the genomic data.
6) **Data Sharing and Collaboration**: During DNA sequencing, sensitive genomic information is accessed by lab technicians, researchers, and IT personnel. Genomic data is transmitted across potentially insecure networks and stored in shared repositories, creating opportunities for unauthorized access or manipulation by cyber criminals without robust security measures and data transfer protocols.
7) **Integration of Multiple Data Sources**: Bioinformatics analysis often involves integrating data from various sources such as reference genomes and experimental metadata). This complexity broadens the attack surface for cyber threats. Attackers may compromise reference databases, manipulate metadata, or use cross-site scripting attacks, potentially disrupting the entire analysis process.
8) **Vulnerability of Open Access NGS Databases**: NGS databases are crucial for genomic research, with many being openly accessible. For example, platforms like NCBI and EMBL-EBI provide open access to data queries without requiring login credentials and do not enforce strong password policies, such as the use of long phrases, capital letters, symbols, and numbers. Moreover, these platforms do not require two-factor authentication or third-party account login [18]. This lack of security measures makes the databases vulnerable to attacks, where malicious actors could modify or delete data, leading to data loss, corruption, or disruption of NGS operations. If genetic data stored in these databases is linked to personally identifiable information, it could compromise privacy and security. Breaches could expose familial relationships and genetic relatedness, potentially leading to targeted attacks, blackmail, or coercion. Many databases, such as NCBI and EMBL-EBI, allow data queries without login and do not enforce strong password policies, like long phrases, capital letters, symbols, and numbers. Additionally, none require two-factor authentication or third-party account login [18].

IV. CYBER-BIOSECURITY THREATS ON RAW SEQUENCE DATA GENERATION WORKFLOW

In this section, we will explore each step of the experimental workflow process. We will focus on the technologies used in each step and the potential cyber threats associated with them (Fig. 4).

A. DNA EXTRACTION

DNA extraction is the initial step in the NGS workflow (Fig. 5). This process starts with the collection of a biological sample, such as tissue or blood, which contains the DNA of interest. The primary goal during DNA extraction is to isolate the DNA from other cellular components, including proteins and RNA, ensuring that the DNA is not degraded or contaminated during the process [42].

B. TOOLS AND TECHNOLOGIES USED

The DNA extraction process is a manual or physical laboratory procedure involving the handling of biological samples to isolate DNA. This stage does not fundamentally rely on digital systems or networked communications, which are common targets for cyber attacks, thus making the likelihood of direct cyber threats minimal during this phase. However, if an attacker gains physical access to this phase, several potential attacks could be carried out to compromise the integrity and privacy of the genomic data.

1) POTENTIAL CYBER THREATS

i. **Re-identification Attack**: During the sample preparation process, DNA is extracted from biological materials. If an unauthorized individual gains access to these DNA samples, they could potentially analyze short tandem repeats (STRs) present in the raw DNA before sequencing. STRs are small, repeated DNA sequences commonly used in forensic identification, paternity testing, and missing person investigations. This process involves identifying the repeat sequences and counting their occurrences. Once STR profiles are obtained, attackers could query public genetic genealogy databases, such as YSearch and SMGF, to infer potential surnames. These inferred surnames, combined with publicly available demographic information (e.g., age and location), enable attackers to triangulate and re-identify individuals whose genomic data was sequenced. For instance, Gymrek et al. [43] demonstrated that surnames could be inferred from anonymized Y-chromosome data by querying public genealogy databases. Similarly, Erlich and Narayanan [19] showed that combining genetic data with demographic information could uncover individual identities, revealing that genetic data alone is often sufficient for re-identification if adequate safeguards are not in place. Further research by Sweeney [44] revealed that combinations of demographic attributes

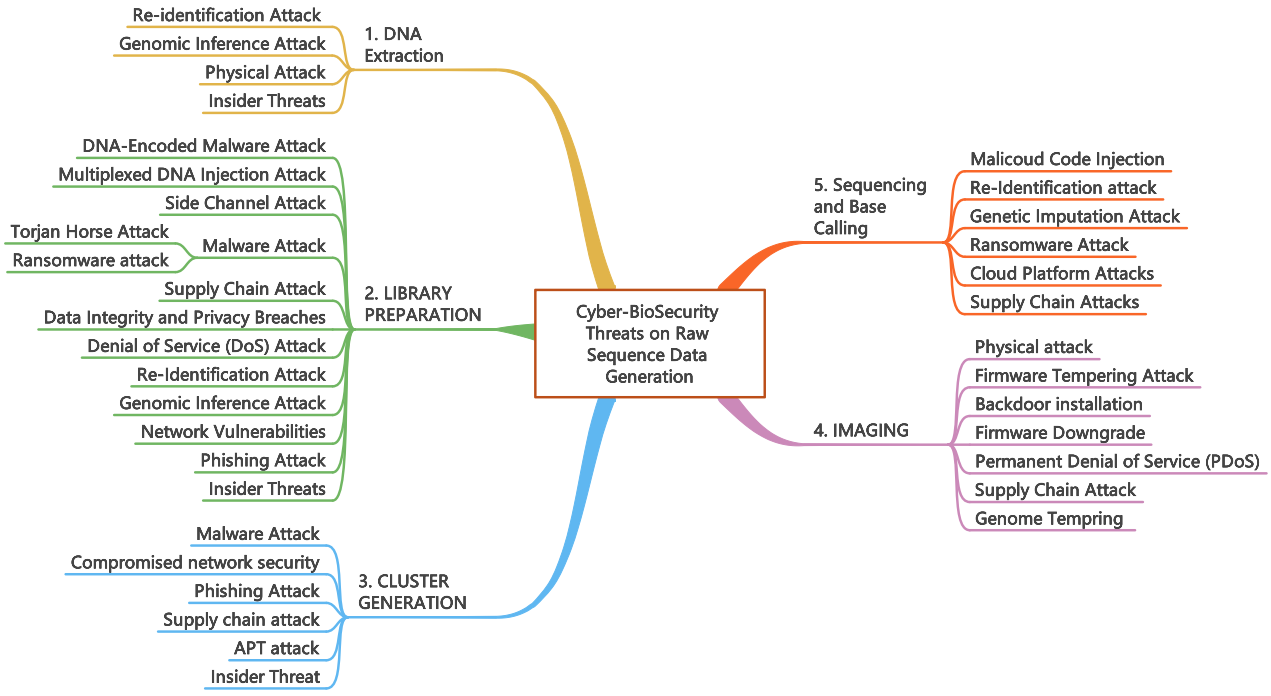


FIGURE 4. Cyber-BioSecurity Threats on raw sequencing data generation Workflow.

such as, ZIP code, gender, and date of birth—could uniquely identify 87% of the U.S. population. This highlights the inadequacy of simple de-identification techniques and the ease with which anonymized data can be re-identified when linked to publicly accessible datasets, such as voter registration lists. In a follow-up study, Sweeney [45] demonstrated that linking demographic data from the Personal Genome Project (PGP) profiles to voter registration lists and other public records enabled re-identification of 84% to 97% of profiles. This re-identification was based solely on demographic data, not DNA information, underscoring the vulnerability of relying on demographic data alone for anonymity. Breaches of genomic data can lead to numerous risks, including the exposure of anonymous paternity through Y chromosome information, which has been used to identify biological fathers, such as sperm donors [46]. For instance, Gymrek et al. [43] demonstrated the re-identification of participants in the 1000 Genomes Project by analyzing their Y-STR profiles and comparing them against public genealogy databases. This approach revealed individuals' identities by linking inferred surnames with publicly available demographic details, such as age and state of residence. Such scenarios pose several significant threats, including:

- **Privacy Violations**: Re-identification from anonymized genomic data infringes on personal privacy and exposes sensitive information without the individual's consent.
- **Identity Theft and Fraud**: Genetic information linked to personal details could be exploited by attackers to access financial and other personal data.
- **Discrimination and Stigmatization**: Misuse of genetic data in employment, insurance, or other contexts may lead to unfair treatment or social stigmatization.
- **Blackmail and Coercion**: Sensitive genetic information could be leveraged by malicious actors to coerce or threaten individuals.

ii. **Genomic Inference Attack:** Partially available genomic data can be used to infer missing genomic information due to the phenomenon known as linkage disequilibrium (LD), which refers to the correlation between different regions of the genome [47]. Adversaries can obtain additional sensitive health information about individuals by analyzing specific disease-related genes, even if only portions of the genome are available. For instance, disease-related genes, such as the APOE gene associated with Alzheimer's disease, can reveal an individual's health status and predisposition to certain diseases [48]. Jim Watson, one of the discoverers of the DNA structure, donated his genome for research purposes but chose to withhold his APOE gene [49]. If an attacker gains access to the extracted DNA, they can obtain sequences that include disease-related genes and their neighboring

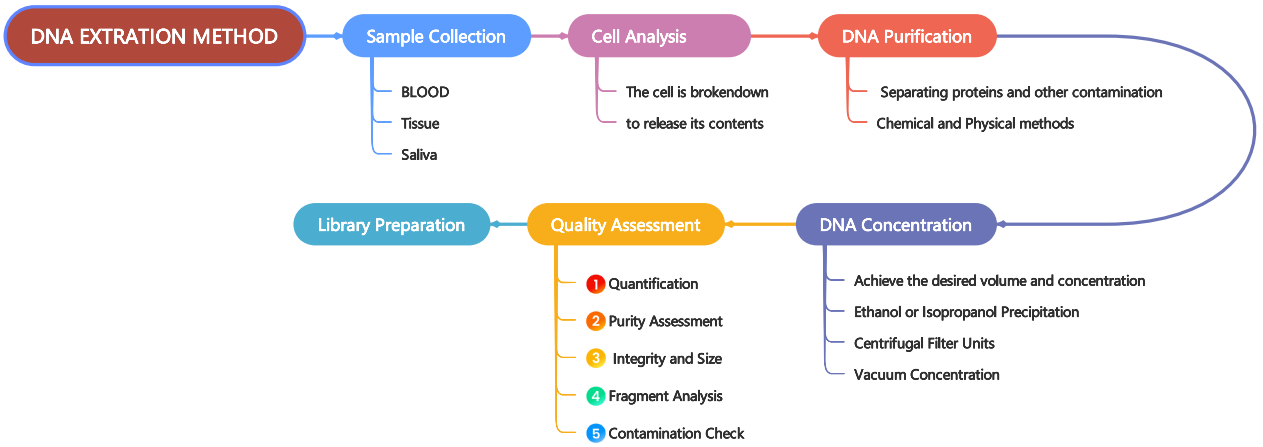


FIGURE 5. DNA extraction method.

genomics variations, which, if misused, could harm the victim's reputation or otherwise disadvantage them.

iii. **Physical Attack:** DNA samples may be physically stolen during the extraction phase, compromising the integrity of the entire dataset. Additionally, attackers could intentionally contaminate samples with extraneous DNA to produce misleading or false results or substitute the original sample with another, leading to erroneous conclusions. Unauthorized individuals might infiltrate the laboratory to tamper with or steal samples. This could be accomplished through social engineering tactics, such as impersonating maintenance staff or new hires, to deceive lab personnel and bypass security protocols.

iv. **Insider Threats:** Insider threats pose a significant risk, as authorized lab personnel with legitimate access to DNA samples or sensitive data may intentionally or unintentionally compromise security. Such breaches can stem from motivations such as financial gain, coercion, personal grievances, or malicious intent. For example, in the case of Yuan Li at Sanofi-Aventis, Li, a researcher at the company, misused her access privileges to download proprietary data, including sensitive research information, onto her company-issued laptop. Her actions, influenced by external affiliations, demonstrated how insiders with legitimate access can exploit their positions to leak confidential data or compromise research integrity [50].

C. LIBRARY PREPARATION

During the library preparation stage, DNA samples are fragmented into smaller pieces to make them compatible with sequencing. This fragmentation can be carried out through mechanical methods, such as sonication, enzymatic techniques that target specific DNA sequences, or chemical approaches. The choice of method depends on the sequencing technology being used and the desired fragment size. Following fragmentation, adapters—short DNA sequences—are attached to the ends of these fragments. This step functions like labeling sections of a document, ensuring accurate identification and reassembly during sequencing [51].

To ensure sufficient DNA for sequencing, PCR (Polymerase Chain Reaction) is used to amplify the adapter-ligated fragments. Before amplification, a size selection step may be included to ensure that the library contains fragments of the appropriate length. Techniques such as gel electrophoresis or magnetic bead-based methods are commonly employed for this purpose [51]. After amplification, a cleanup process is typically performed to remove unincorporated adapters and primers, leaving only the target DNA fragments for sequencing [52].

1) TOOLS AND TECHNOLOGIES USED

The library preparation step in NGS primarily involves laboratory-based procedures, but it can also integrate network technologies for certain aspects, particularly in high-throughput and automated environments. For instance, barcodes or RFID chips are use to automate the tracking and processing of multiple samples through the library preparation and sequencing workflow [53]. LIMS (Laboratory Information Management Systems) are often networked to manage and track samples throughout the library preparation process. They provide real-time data entry and retrieval, helping to ensure sample integrity and reduce errors. Automated Liquid Handlers (robots) are often networked to central control systems to streamline library preparation steps such as DNA fragmentation, end repair, and adapter ligation. Overall, while the core biochemical steps of library preparation do not inherently depend on network technologies, the management, tracking, and automation aspects of the process increasingly rely on networked systems to enhance efficiency, accuracy, and scalability [54].

2) POTENTIAL CYBER THREAT

i. **DNA-Encoded Malware Attack:** The authors in [22], [23] conducted an experiment where they introduced

a DNA-Encoded Malware during the ''Library Preparation'' step of the NGS workflow. The researchers conceptualized DNA not just as a biological molecule but as a carrier of information—akin to binary code in computing. DNA sequences are composed of four bases (adenine (A), thymine (T), cytosine (C), and guanine (G)), which sequencing machines can interpret in a manner similar to how computers read binary sequences (0s and 1s).

The authors synthesized DNA strands during library preparation that, upon sequencing and processing, generated a digital file. When this file was input into a specially designed, vulnerable program, it enabled the authors to open a socket for remote control, effectively gaining unauthorized access to the computer system, which can compromise the integrity, confidentiality, or availability of the system and its data. This groundbreaking research represents the first documented instance of using biological or synthetic DNA samples to execute a computer system exploit. Researchers demonstrated that it is possible to create DNA that compromises a victim system through sequencing and processing. Their key finding underscores the feasibility of encoding malicious software into DNA strands, introducing a novel security threat vector.

ii. **Multiplexed DNA Injection Attack:** Researchers [22], [23] inadvertently identified an unintended information leakage channel during the multiplexing process, which subsequently exposed vulnerabilities to potential data manipulation attacks. This leakage occurred because standard practice involves multiplexing different DNA samples on the same sequencing machine, which can result in information being shared between samples. The researchers observed that when their exploit-containing synthesized DNA sample was multiplexed and sequenced alongside other samples, the sequencing data erroneously included DNA sequences from those other samples. Such unintended data leakage poses serious risks. For instance, an attacker could exploit this by inserting a harmful DNA sample, leading to manipulation or corruption of genetic data by bioinformatics software, potentially affecting research outcomes or diagnostic results.

iii. **Side Channel Attack:** Multiplexing enables simultaneous sequencing of multiple DNA samples by assigning unique identifiers but presents challenges during the demultiplexing stage. Errors in this stage can result in ''sample bleeding,'' where DNA sequences are incorrectly assigned, potentially affecting over 1% of samples on some platforms. Ney et al. [22] highlight that these vulnerabilities could be exploited through side-channel attacks. An attacker could observe the timing differences in the sequencing process, such as the intervals between signal detections or processing delays specific to different identifiers. By correlating these timing differences with known patterns or sequences, an attacker might infer which sequences correspond to which identifiers. This could lead to sabotaging sequencing runs, influencing outcomes, or compromising sample privacy and could have serious consequences for fields reliant on accurate DNA sequencing, such as medical research, forensics, and personalized medicine.

iv. **Malware Attack:** Accurate control of reagent quantities and mixtures is crucial for successful DNA library preparation, often managed through software overseeing automated liquid handling systems in high-throughput and potentially contaminated lab environments. For instance, the Labcyte Echo Liquid Handler [55] is widely used for precise, non-contact liquid handling in NGS laboratories. However, vulnerabilities in biochips used with DNA sequencers have been highlighted by Ali et al [56]. The authors point out that these biochips are susceptible to malware attacks, including trojans that can masquerade as legitimate software. Once a microfluidic biochip is infected, trojans can leak sensitive sequencing data or manipulate genetic information, severely impacting research integrity and breaching privacy regulations. Such unauthorized access or manipulation can severely impact genetic research integrity, breach privacy regulations, and introduce inaccuracies in crucial genetic data essential for medical diagnostics and research. Additionally, ransomware attacks pose significant threats by encrypting data or system functionalities and demanding ransom payments for decryption keys. Such disruptions can halt laboratory operations, compromising ongoing research and data integrity essential for medical diagnostics and research.

v. **Supply Chain Attack:** Cyber attackers may compromise the software supply chain by injecting malicious code into the software during development, distribution, or maintenance stages. This can lead to the distribution of compromised software to NGS laboratories, potentially exposing them to various security risks.

vi. **Data Integrity and Privacy Breaches:** Hackers may attempt to gain unauthorized access to LIMS or networked systems controlling automated liquid handlers. Weak authentication mechanisms or misconfigured access controls may allow unauthorized individuals to gain access to the software controlling liquid handling in the library preparation systems. Unauthorized access can lead to data manipulation, equipment damage, or sabotage of research activities. This could lead to data theft, manipulation of experimental results, or disruption of operations. Manipulating sample tracking data or experimental parameters can compromise the integrity of sequencing results, leading to incorrect scientific conclusions.

vii. **Denial of Service (DoS) Attacks:** Attackers may launch DoS attacks against LIMS or central control systems, causing downtime or delays in sample processing. This can impact research timelines and productivity. Attackers may launch DoS attacks to disrupt the availability

of the liquid handling systems and associated software. By overwhelming the systems with a flood of requests or traffic, the attackers can render them temporarily or permanently inaccessible, disrupting research activities.

viii. **Re-Identification Attack:** During the library preparation, DNA is fragmented, and certain regions, including STRs, may be amplified. Attackers or laboratory personnel with malicious intent could analyze these regions to create STR profiles. STR profiles can be queried against public genealogy databases to infer personal information associated with the DNA profiles. The laboratory preparation phase often involves recording metadata about the samples (e.g., source, collection date). Attackers could exploit this metadata to link DNA samples back to individuals.

ix. **Genomic Inference Attack:** During the library preparation step, DNA samples are fragmented, and adaptors are attached. If attackers gain access to these fragments, they could analyze the sequences to identify genes and variations linked to diseases. This access would enable them to focus on specific disease-related genes during the preparation process.

x. **Network Vulnerabilities:** The library preparation stage starts with taking patients's blood samples and recording the necessary information on internal or external storage such as could storage. Without adequate network security, the confidential and sensetive data is vulnerable to unauthorised access and data breaches during transmission. Vulnerabilities such as absence of misconfiguration of firewalls, poor access controls, or insufficient encryption can create entry points for malicious actors to exploit critical systems used in library preparation. Such breaches could lead to violations of important regulations like GDPR or HIPAA, which are designed to safeguard personal and health-related data. If attackers gain access, they could manipulate reagent volumes, disrupt sequencing procedures, or compromise the accuracy and reliability of the sequencing data. These risks not only threaten privacy but also pose serious legal and ethical challenges due to the exposure of confidential information.

xi. **Phishing Attacks:** Employees may become targets of phishing campaigns designed to steal login credentials or compromise the security of the networked systems.

xii. **Insider Threats:** Individuals with authorised access to networked systems could misuse their privileges, whether intentionally or unintentionally, potentially jeopardising data integrity or causing disruptions to system operations.

D. CLUSTER GENERATION

During the cluster generation step, clusters are formed by grouping identical DNA molecules on the surface of a sequencing flow cell. This process amplifies the signal, facilitating more accurate detection of the DNA sequence during sequencing [57].

E. TOOLS AND TECHNOLOGIES USED

The DNA library fragments, tagged with adapters during the library preparation step, are then applied to a solid surface, such as a slide or flow cell. On this surface, individual DNA fragments are immobilized and undergo amplification using specialized devices to generate multiple copies of each fragment. During the immobilization process, DNA fragments are attached to the functionalized coating of the slide or flow cell, ensuring they remain fixed throughout the sequencing process. Once immobilized, each DNA fragment undergoes localized amplification facilitated by devices like the Illumina NovaSeq or Oxford Nanopore MinION. This amplification process, often achieved through bridge amplification, duplicates each DNA fragment bound to the surface multiple times in situ. As a result, clusters of DNA are formed. Each cluster contains thousands of copies of the original DNA fragment, all localized to a single spot on the flow cell surface. These clusters are then ready for sequencing. Each cluster provides a strong, localized signal that can be detected and analyzed by the sequencing machine, allowing for high-throughput sequencing of millions of DNA fragments simultaneously [57].

1) POTENTIAL CYBER THREATS

The automation and monitoring software embedded within NGS platforms, such as Illumina, plays a crucial role in ensuring that the cluster generation process is consistent and scalable. It enables high-throughput sequencing with minimal human intervention by precisely controlling environmental conditions within the flow cell, timing chemical reactions accurately, and monitoring cluster development in real-time. However, any tampering with these processes could have catastrophic consequences, resulting in false or unreliable data, which could lead to flawed conclusions and decisions based on erroneous information.

i. **Malware Attack:** Consider a scenario where a cyber criminal expoiltes the vulnerabilities such as a weak password, absence of firewall, or weak access control and injects a malware in the controlling and monitoring software of NGS platform. Once the malware infected the system, it could alter software parameters essential for cluster generation process. For instance, this could involve altering the envoiremental conditions such as temprature and Ph levels. Such undesirable altertion and unauthorized access has the potential to disrupt the timings of critical chemical reactions and processes. This would ultimately results in a cluster that is either too sparse or overly dense. Consequently, this would generate low quality sequencing data or misdiagnoses in clinical applications.

ii. **Insider Threats:**
Individuals with access to NGS platforms may intentionally or accidentally, sabotage the cluster generation process, leading to significant disruptions or a loss of data integrity. Insiders with unauthorised access to the software used for automating and monitoring

cluster generation could tamper with critical parameters, modify amplification protocols, or introduce malicious code into the system. Such intentional or unintentional interference in the NGS plateform can generate inaccurate sequencing data, jeopardising research findings or clinical diagnostics.

F. IMAGING

Following cluster generation, the imaging phase begins, during which DNA synthesis occurs with the incorporation of fluorescently labeled nucleotides. Each nucleotide (A, T, G, C) is tagged with a unique fluorescent marker. As these nucleotides are incorporated into the growing DNA strand, they emit distinct fluorescent signals corresponding to their specific labels. These signals are captured by a camera and subsequently converted into sequence data. This process allows for precise determination of the DNA sequence and is extensively used in various NGS platforms due to its high efficiency and accuracy in sequencing DNA at scale [54].

1) TOOLS AND TECHNOLOGIES USED

This phase relies on sophisticated imaging technologies that play a critical role in accurately capturing the intricate biochemical reactions that occur during DNA sequencing. Cameras used in this process are designed to detect and record the fluorescence signals emitted by nucleotides as they are incorporated into the DNA strand, ensuring the reliable documentation of sequencing data.

2) POTENTIAL SECURITY THREATS

i. **Physical Attack**: Accurate imaging is essential for NGS, as each fluorescent signal must be precisely captured and accurately assigned to a specific nucleotide (A, T, C, G). Physical attacks on the system could involve damaging critical hardware, disrupting environmental controls, or tampering with the imaging components. For example, physical damage to high-resolution cameras or lasers used for fluorescence detection could compromise image quality or render the system non-operational. Additionally, sensitive imaging equipment relies on stable temperature conditions to function properly. Extreme changes in laboratory temperature, whether excessive heat or cold, could disrupt the performance of cameras and other sensitive electronics, jeopardising the sequencing process.

ii. **Hardware Compromise:** The physical components of sequencing machines, including their embedded firmware, are equally at risk. Firmware in imaging cameras controls essential tasks such as image capture and processing, exposure settings, and data transmission to computers for analysis. It ensures the camera operates consistently and with the necessary precision for NGS applications. However, this critical component is vulnerable to various cyber attacks. For instance:

- **Firmware Tampering**: This attack involves modifying the firmware to alter the device's behavior or introduce malicious functionality [58]. An attacker could inject malicious code into the firmware of a sequencing camera, causing it to alter image data or fail at critical moments, thus sabotaging the sequencing process.
- **Backdoor Installation**: A malicious backdoor can be inserted into the firmware, allowing attackers to gain unauthorized access to the device at any time. If a backdoor is installed in the firmware of medical imaging equipment, an attacker could remotely access and manipulate diagnostic tools, potentially leading to misdiagnoses.
- **Firmware Downgrade**: In this attack, devices are forced to downgrade to older, less secure versions of firmware that contain known vulnerabilities [59]. An attacker might force a lab instrument to revert to outdated firmware that has unpatched security flaws, which could then be exploited to gain control over the instrument.
- **Permanent Denial of Service (PDoS)**: known as "bricking," this attack renders a device permanently inoperable by corrupting its firmware [60]. Deliberate corruption of the firmware in critical lab equipment could halt research and cause significant financial loss.
- **Supply Chain Attacks:** Compromising firmware before it even reaches the user, typically during manufacturing or distribution. Malicious actors could tamper with firmware at the source, inserting vulnerabilities or malware before the devices are even shipped to laboratories.

iii. **Genome Tampering Attack:** The software systems embedded within NGS platforms for imaging and signal detection are critical for ensuring the accurate acquisition and analysis of genomic data. Malicious actors may exploit vulnerabilities in these systems or inject malicious code to manipulate algorithms or data. Such tampering could lead to incorrect base calling, compromising scientific research, clinical diagnostics, and other applications dependent on reliable genomic data.

To address this issue, Ma et al. [61] proposed a method utilising fragile watermarking, which is specifically designed to detect unauthorised modifications to DNA sequences. Building on this concept, Fu et al. [62] introduced an approach that combines fragile watermarking with the Merkle Hash Tree. This integration enables the efficient detection of tampered segments in large datasets of DNA sequences, providing an additional layer of security to safeguard genomic data.

G. DNA SEQUENCING AND BASE CALLING

The DNA sequencing and base calling process involves reading DNA sequences through sequencing technology and translating raw signals into nucleotide sequences. This begins with capturing raw fluorescence signals during sequencing,

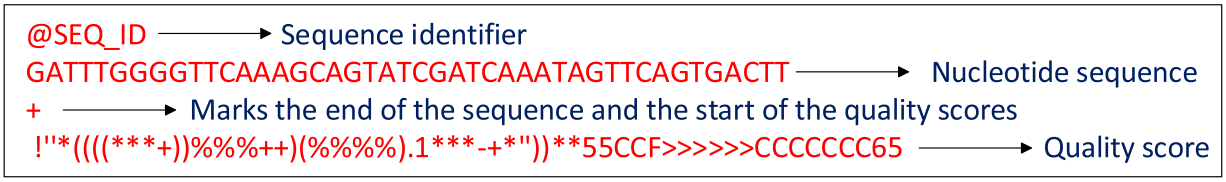


FIGURE 6. A fastq file example.

which are processed by specialised base calling software, such as Illumina Real-Time Analysis (RTA), integrated into sequencing platforms like the HiSeq and NovaSeq series. The software analyses images of the sequencing flow cell to determine the sequence of nucleotides incorporated at each position along the DNA strand.

Each nucleotide—Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)—is tagged with a unique fluorescent dye, emitting a specific colour: green for Adenine, red for Thymine, blue for Cytosine, and yellow for Guanine. During sequencing, these labeled nucleotides are incorporated sequentially, and their fluorescent signals are captured. The base calling software decodes these light signals, translating them into the DNA sequence, such as ATCA [63].

1) TOOLS AND TECHNOLOGIES USED

During the DNA sequencing and base calling step, two primary types of files are generated: FASTQ or BAM/BAM files and BCL (Binary Base Call) files. FASTQ files contain raw sequence reads along with quality scores for each nucleotide base call. Each entry in a FASTQ file includes a sequence identifier, the nucleotide sequence, a separator, and a quality score string (Fig. 6). These quality scores, typically represented using ASCII characters to encode Phred quality scores, indicate confidence in each base call and are essential for assessing the accuracy and reliability of the sequencing data. BCL files, produced by Illumina sequencing platforms, contain raw data from the sequencing machine, including signal intensities and the base calls derived from those signals. While BCL files are critical for initial data processing and quality assessment directly from the sequencing instrument, they are usually converted to FASTQ format for downstream analysis due to the more accessible and widely used nature of FASTQ in bioinformatics workflows. Accurate base calling is crucial as it ensures high-quality sequencing data necessary for reliable downstream genomic analyses, such as variant detection, genome assembly, and transcriptome analysis.

This step also uses supportive bio-informatics tools such as Bcl2fastq [64]. Bcl2fastq is a conversion tool by Illumina that converts BCL files generated by the sequencer into FASTQ files. Base calling step also utilizes a basespace sequence hub (a cloud-based platform) [65] that allows to hold base calling output.

2) POTENTIAL CYBER THREATS

Consider a scenario where the base calling software of a high-throughput NGS system has been compromised by a malicious entity. The attacker has introduced a subtle bug that shifts the interpretation algorithm, causing a systematic misreading of nucleotides. For example, it might interpret a green signal as Thymine (T) instead of Adenine (A), leading to a sequence output of T instead of A. This alteration can propagate errors throughout the genomic data being analyzed. This could be part of a targeted attack (APT) to skew research results or sabotage a competitor's clinical diagnostic capabilities. Such errors could lead to incorrect genetic screening results, potentially resulting in misdiagnoses or inappropriate treatment plans.

Unlike earlier stages, base calling in NGS demands substantial computational resources and specialized software, making it prone to various cyber threats.

i. **Malicious Code Injection Attack**: FASTQ, or BAM files contain vital information derived from sequencing processes, including nucleotide sequences and their corresponding quality scores (see Figure. 6). In a compromised environment, sophisticated cyber security threats, such as Advanced Persistent Threats (APTs) and zero-day attacks, can target the vulnerabilities inherent in the handling of these data formats. An attacker, for instance, might embed or inject malicious code within these raw data files, exploiting vulnerabilities within the software or indirectly via infected files or updates, causing the software to misinterpret the fluorescent signals. Such actions could lead to unauthorized system access, tampering or corruption of the raw sequencing data, financial losses, and reputational damage.
ii. **Re-Identification attack**: FASTA files contain the actual DNA sequences, which may include sensitive regions such as STRs, disease-related genes, or other identifiable markers. Along with the sequence data, FASTA files or associated records may contain metadata such as patient demographics, clinical data, and sample information that can be used to link sequences back to individuals. Attackers can analyze FASTA files to extract STR regions and create profiles. These profiles can then be queried against public genealogy databases to infer surnames and potentially re-identify individuals. Attackers can use unique genetic markers in the FASTA files and correlate them with data from public databases or previous studies to re-identify individuals.
iii. **Genetic Imputation Attack:** Genetic imputation involves predicting unknown genotypes by using known genetic variants and their linkage disequilibrium (LD) patterns. An example is Nyholt et al.'s [66] attack, where

they inferred a masked gene by interpreting neighboring variations in a genome. During the base calling step, genetic variants and their LD patterns are more clearly defined, making it easier for attackers to perform genetic imputation. The conversion process results in readable sequences that can be exploited to predict unknown genotypes. This attack can recover masked sensitive information, undermining privacy efforts.

iv. **Ransomware Attack:** Generating and processing FASTQ files is resource-intensive, involving substantial computational power for base calling, substantial storage capacity for data management, and significant data transfer bandwidth when these files are moved or shared for further analysis. Ransomware attacks can be launched by exploiting vulnerabilities in the hardware, software or in the broader IT infrastructure, including operating systems and network connections. This could be through security weaknesses in the software, outdated firmware, or through phishing attacks that target users of the sequencing systems. Ransomware attack can encrypt the FASTQ files, making them inaccessible and halting further genomic analysis until a ransom is paid. This could severely disrupt both academic research and clinical diagnostics.

v. **Cloud Platform Attacks**: If the BaseSpace Sequence Hub or similar cloud platforms are used, attackers could target these platforms to access or manipulate sequencing data stored in the cloud.

vi. **Supply Chain Attacks**: Compromising the software updates or third-party tools (like Bcl2fastq) used in the base calling process could introduce vulnerabilities or malicious code into the system.

V. CYBER-BIOSECURITY ATTACKS ON QUALITY CONTROL AND PREPROCESSING

Before analysis, raw data from various sources like DNA sequencing must be cleaned and formatted. This step involves cleaning and filtering raw sequencing reads, trimming low-quality bases, and removing adapter sequences [25]. This is a crucial step, as it directly impacts the quality and accuracy of downstream analyses.

A. TOOLS AND TECHNOLOGIES USED

Several tools and technologies are commonly used to clean and filter raw sequencing reads generated in the base calling step.

- **Quality Control Tools**: Software like FastQC,[8] MultiQC,[9] Qualimap,[10] SeqMonk,[11] Fastp,[12] and NGS QC Toolkit[13] etc., provide quality metrics for raw sequencing data, helping to identify problems such as low-quality bases, adapter contamination, and over represented sequences.
- **Adapter Trimming**: Adapter trimming is a crucial step in the preprocessing of NGS data, involving the removal of adapter sequences ligated to DNA fragments during the preparation of sequencing libraries. These adapters, if not removed, can interfere with the analysis, for example, by affecting sequence alignment and assembly, leading to incorrect interpretation of the sequencing data. Tools such as Trimmomatic[14] and Cutadapt[15] are used to remove adapter sequences that have been ligated to DNA fragments to enable sequencing.
- **Quality Trimming**: Quality trimming focused on removing bases with poor quality scores from the ends of sequencing reads. This process is critical because low-quality bases can introduce errors into the downstream analyses, such as sequence alignment and variant calling, potentially leading to inaccurate conclusions. Tools like Trimmomatic[16] and QIIME[17] (specifically for microbial or metagenomic sequences) are widely used for this purpose.
- **Error Correction**: Error correction in NGS data is a critical step aimed at identifying and correcting errors in the DNA sequence reads. These errors can arise due to various factors inherent in the sequencing process, such as chemical mishaps during sequencing, base-calling inaccuracies, or issues during library preparation. Tools like BayesHammer[18] and Coral[19] are designed to address this need by implementing sophisticated error models.

B. POTENTIAL SECURITY THREATS

The quality control tools are typically used in a local computing environment (e.g., personal computers and institutional servers) and do not inherently require online communication or collaboration technologies for their core functions of analyzing and assessing the quality of sequencing data. The essential input for these tools is a raw or processed sequence. An insecure environment can lead to unauthorized access, manipulation of sequencing data, or even data breaches (Fig. 7)

I. **QualityCompromise Exploit Attack:** This attack targets the integrity of genomic data during its preprocessing stage. The primary goal of the QualityCompromise exploit attack is to degrade the quality of genomic data through malicious alterations within the sequencing workflow. Manipulated quality metrics could lead to the acceptance of poor-quality data, affecting downstream analyses such as variant calling or differential expression studies. Such an attack can be achieved through

[8] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[9] https://multiqc.info/
[10] http://qualimap.conesalab.org/
[11] https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/
[12] https://github.com/OpenGene/fastp
[13] https://ngsqc.org/
[14] https://github.com/usadellab/Trimmomatic
[15] https://github.com/marcelm/cutadapt
[16] https://github.com/usadellab/Trimmomatic
[17] https://qiime2.org/
[18] https://www.biostars.org/p/469538/
[19] https://www.cs.helsinki

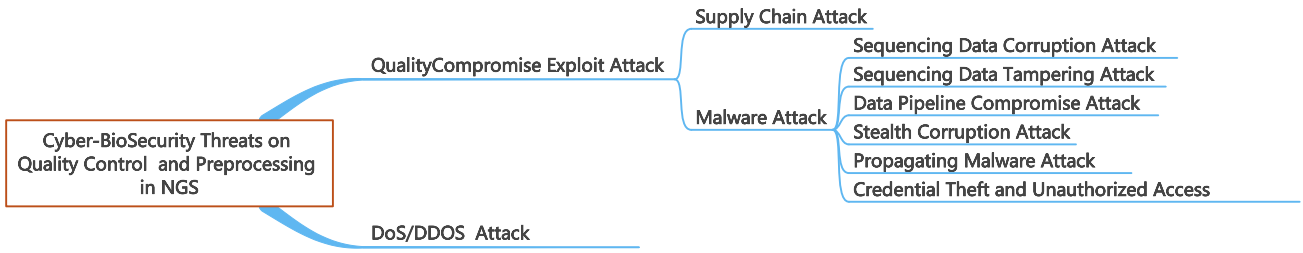


FIGURE 7. Cyber-BioSecurity attacks on quality control and preprocessing in NGS.

manipulating the tools and technologies that ensure data integrity. Here are several methods through which such an attack could be effectively carried out, leveraging vulnerabilities in specific NGS quality control and preprocessing tools:

i. **Supply Chain Attack (Software Libraries and frameworks)**: The software and tools used during the QC and preprocessing steps, such as FastQC, Trimmomatic, Cutadapt, and QIIME, rely on various libraries and frameworks to function. These dependencies are often sourced from external repositories and integrated into the software's ecosystem. Attackers could compromise the software supply chain by injecting malicious code into widely used libraries or frameworks. When NGS preprocessing tools incorporate these compromised libraries, the malicious code becomes part of the legitimate software and could alter the software's functionality or open backdoors for further exploits. For instance, consider a commonly used open-source library within FastQC is tampered with in its public repository. The tampered version includes a hidden backdoor or a subtle bug designed to manipulate the quality metrics reported by FastQC. As laboratories download or update their FastQC installations, they unknowingly integrate the compromised library into their systems. Once active within a laboratory's data processing system, the malicious code could alter the results of the quality control checks, either by failing to report issues with sequencing data or by directly modifying data metrics.

ii. **Malware Attack:** Malware could be programmed to launch variety of attacks on the quality control step in the NGS workflow. For instance:

- **Sequencing Data Corruption Attack**: Malware designed to corrupt .fastq or .bam files can severely disrupt the quality control steps crucial for accurate bioinformatics analysis. By randomizing sequences or making other destructive alterations, this type of malware renders the data useless for accurate analysis. When essential files are corrupted, it becomes impossible to obtain reliable results from the data, leading to wasted time and resources and potentially delaying critical scientific discoveries.
- **Sequencing Data Tampering Attack:** Another insidious form of malware attack involves the subtle alteration of genetic sequences within data files. By inserting or deleting bases, the malware creates or hides genetic variations, skewing analysis results and leading to false conclusions. This type of sequence data tampering is particularly dangerous because it can go unnoticed until significant damage has been done, resulting in incorrect scientific findings and potentially affecting subsequent research built on these flawed results.
- **Data Pipeline Compromise Attack**: A Data Pipeline Compromise Attack targets the software functions responsible for critical preprocessing steps in data analysis pipelines, such as trimming adapters or correcting sequencing errors. The malware introduces subtle errors during these preprocessing stages, compromising the integrity of the entire data analysis pipeline. Researchers might not detect these manipulations until the final stages of their analysis, making it essential to implement redundant verification steps and cross-check results using multiple software tools.
- **Stealth Corruption Attack:** A stealth corruption attack is a highly advanced malware attack that specifically targets detection systems and error correction algorithms within quality control software. The malware introduces subtle errors that evade these protective measures, resulting in undetected data corruption and compromised analyses [67]. This covert form of attack can have a serious impact on the reliability of research outcomes, as the hidden errors can spread throughout the analysis process.
- **Propagating Malware Attack:** Malware that spreads to other systems connected to the initial host, targeting similar data files for corruption, exemplifies a propagating malware attack. This type of malware seeks out and infects additional systems, expanding the scope of data corruption

and magnifying its impact. The widespread disruption caused by such attacks can affect multiple datasets and research projects.

- **Credential Theft and Unauthorized Access Attack:** Malware could be programmed to steal user credentials and use them to gain unauthorized access to systems, posing a serious threat to data integrity. Once inside, attackers can misconfigure settings, leading to intentional errors in data preprocessing and analysis. These errors can compromise the entire pipeline, from raw data generation to final analysis results.

II. **Denial of Service (DoS) attack:** A DoS attack may target the computational resources allocated to quality control processes. For instance, the attacker could flood the server hosting the bioinformatics quality control tools with an overwhelming amount of requests or data, consuming all available processing power and memory resources.

VI. CYBER-BIOSECURITY THREATS ON BIOINFORMATICS ANALYSIS WORKFLOW

In this section, fundamental information regarding the three critical steps involved in bioinformatics analysis—sequence alignment, variant calling, and annotation—is first presented. Following this, the tools and technologies employed in these processes are elaborated upon. Finally, potential cyber attacks and vulnerabilities associated with these technologies are described, with an examination of how they may facilitate such attacks on the bioinformatics analysis workflow.

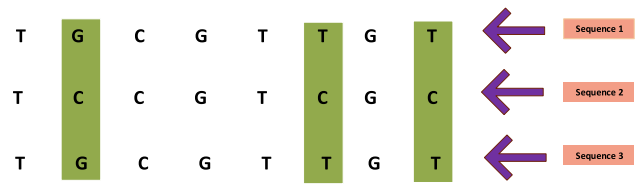


FIGURE 8. Sequence alignment example.

1) **Sequence Alignment:** This process involves the comparison of DNA, RNA, or protein sequences letter by letter to detect regions that correspond to sequences stored in databases of known sequences (Fig.8). This allows for the identification of similarities and discrepancies between the sequences. For instance, when aligning two DNA sequences such as Sequence 1: TGCGTTGT and Sequence 2: TCCGTCGC, we can observe that most of the nucleotides match, except for some differences (e.g., the second nucleotide). These similarities and differences can provide valuable insights into how different organisms are related, how diseases may impact various species or the function of specific genes [68].
2) **Variant Calling:** A variant refers to any deviation or alteration in a DNA sequence when compared to a reference genome [69]. This step involves identifying genetic variants, such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), by comparing the sequencing data to a standard reference (Fig. 9). In medical genetics, variant calling is essential for detecting mutations associated with diseases, understanding genotype-phenotype relationships, and informing personalised treatment plans. In evolutionary biology, it is critical for reconstructing population histories, tracking migration patterns, and studying species adaptations [70]. Advanced variant calling algorithms enable the identification of these genetic differences and generate Variant Call Files (VCFs), which store comprehensive details about SNPs, indels, and structural changes relative to the reference genome [71].

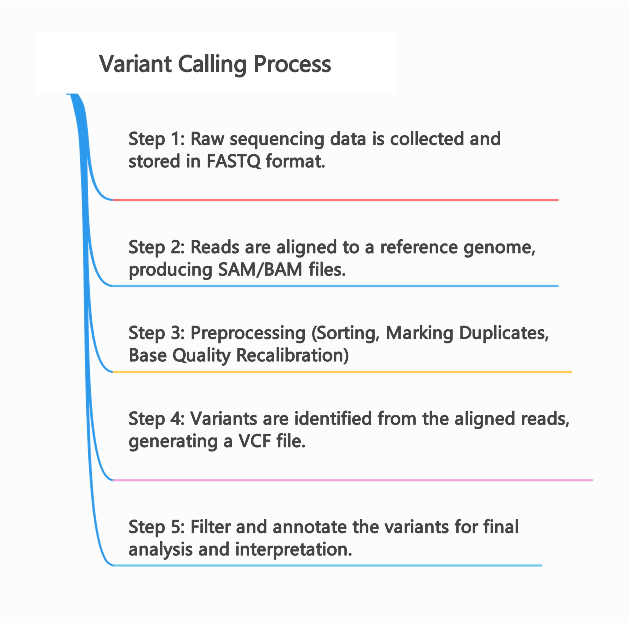


FIGURE 9. Steps in Variant Calling.

3) **Annotation:** Genomic annotation interprets identified variants to determine their biological significance. In cancer research, for example, annotation helps pinpoint mutations in oncogenes or tumour suppressor genes, linking them to disease processes. It is also integral in predicting the pathogenicity of genetic variants and supporting drug discovery efforts [72]. By associating genetic variants with clinical phenotypes, annotation enables researchers to uncover genotype-phenotype relationships and identify genetic risk factors for complex diseases like diabetes or Alzheimer's disease [72]. The output of the annotation process is typically stored in GFF (General Feature Format) or GTF (Gene Transfer Format) files, which contain detailed information about genome features, providing insights into gene function and regulation.

A. TOOLS AND TECHNOLOGIES USED

Sequence alignment, variant calling, and annotation are highly resource-intensive tasks, requiring the comparison and matching of millions of short DNA reads against an extensive reference genome. These steps typically demand significant data exchange, distributed computing resources, and access to various bioinformatics databases and tools. The specifics of the technologies and tools employed in these processes are detailed below:

- **Internet and Networking Protocols**: HTTP and HTTPS are fundamental web protocols used to access various bioinformatics tools and databases available online, such as the widely-used BLAST (Basic Local Alignment Search Tool) from the NCBI (National Center for Biotechnology Information). These protocols enable secure and efficient communication between users and online platforms, allowing researchers to query databases, perform sequence alignments, conduct variant calling, and retrieve critical biological data. Additionally, FTP (File Transfer Protocol) and SFTP (Secure File Transfer Protocol) are frequently used for transferring large sequence datasets to and from remote servers. These protocols are particularly valuable for downloading raw sequencing data from public repositories or uploading data for alignment, especially when leveraging cloud-based bioinformatics services.
- **Distributed Computing and Cloud Services**: Cloud computing platforms like AWS (Amazon Web Services), Google Cloud, and Azure provide scalable computing resources that are essential for handling large datasets, particularly for tasks like sequence alignment that may exceed the capabilities of local systems. These platforms include AWS Genomics,[20] Google Cloud Life Sciences,[21] Azure Genomics,[22] DNAnexus,[23] and the Seven Bridges Genomics Platform.[24] Additionally, Illumina BaseSpace Sequence Hub, a cloud-based solution for managing and analyzing genomic data, offers tools for variant calling and annotation. NVIDIA Parabricks,[25] in collaboration with Google Cloud Platform (GCP), leverages DeepVariant, a deep learning technology developed by Google Brain, to accurately reconstruct genome sequences from high-throughput sequencing data [73].
- Application Programming Interfaces (APIs): RESTful APIs (Representational State Transfer Application Programming Interfaces) [74] are essential in bioinformatics for automating access to tools and databases. They facilitate the integration of sequence alignment into workflows via scripts, adhering to REST principles for standardized internet communication. Bioinformaticians use these APIs to query databases, submit alignment jobs, retrieve results, and perform annotation and variant calling tasks. Ensembl,[26] a popular API, provides access to genomic data, including information on genes, variants, and comparative genomics. BioMart[27] offers a RESTful API for accessing biological data sets, including genomic data across various databases and species.
- **Data Sharing and Collaboration Tools**: Collaborative platforms such as GitHub[28] and Bitbucket[29] serve as centralized hubs for hosting and sharing bioinformatics tools, scripts, and pipelines, including those used for sequence alignment, annotation, and variant calling. Researchers utilize features like issue tracking, pull requests, and code reviews to enhance tools for genomic data analysis collaboratively.
- **High-Performance Computing (HPC)**: HPC interfaces are essential in bioinformatics for compute-intensive tasks such as sequence alignment, variant calling, and annotation [75]. SSH (Secure Shell)[30] is vital for secure access to HPC clusters, enabling researchers to execute commands, transfer files, and run workflows securely. Job scheduling systems such as SLURM (Simple Linux Utility for Resource Management)[31] and PBS (Portable Batch System)[32] efficiently manage tasks like alignment, variant calling, and annotation on HPC clusters, facilitating job submission, monitoring, and resource optimization for bioinformatics analyses.
- **Search Engine and software libraries:**
  The tools and technologies utilized for sequence comparison, such as BLAST,[33] ClustalW,[34] MAFFT,[35] Bowtie & Bowtie2,[36] and HISAT2[37] can be likened to highly specialized search engines tailored for genetic information [77].
  Developed at the Broad Institute, GATK (Genome Analysis Toolkit) is widely recognized as the industry standard for identifying SNPs and indels in germline DNA and RNA-seq data. It offers a variety of tools focused on variant discovery and genotyping [78].

[20] https://aws.amazon.com/health/genomics/
[21] https://www.ocre.cloud.tisparkle.com
[22] https://azure.microsoft.com/en-gb/products/genomics
[23] https://documentation.dnanexus.com/developer/cloud-workstation
[24] https://www.sevenbridges.com/platform/
[25] https://www.nvidia.com/en-gb/clara/parabricks/
[26] https://rest.ensembl.org/
[27] https://www.ensembl.org/info/data/biomart/biomart_restful.html
[28] https://github.com/
[29] https://lasp.colorado.edu/nucleus/login
[30] https://www.ssh.com/
[31] https://slurm.schedmd.com/
[32] https://www.pbspro.org/
[33] A widely used program for comparing an input sequence against a database of sequences. It is available through the National Center for Biotechnology Information (NCBI).
[34] Popular tools for multiple sequence alignment, used to align three or more sequences together [76].
[35] A multiple sequence alignment program that offers a range of algorithms optimized for speed and accuracy.
[36] https://rnnh.github.io/bioinfo-notebook/docs/bowtie2.html
[37] https://daehwankimlab.github.io/hisat2/

Samtools[38] and FreeBayes[39] are used for variant calling and analyzing aligned sequence data for various types of genetic variants.

ANNOVAR[40] is widely used for annotating genetic variants by integrating them with information from various databases, including gene functions, disease associations, and population frequencies. SnpEff annotates genetic variants and predicts their effects on genes, including the impact on protein sequences [79]. VEP[41] provides detailed annotations of genetic variants, including their functional effects on genes and their potential implications in diseases.

- **HGMD:** HGMD (Human Gene Mutation Database)[42] is extensively used in the annotation phase of genomic analysis. During annotation, genetic variants identified through sequencing are compared against known mutation databases to determine their significance. Then, HGMD is utilized in variant calling to interpret the results of the variant calling process. After calling variants from sequencing data, these variants are compared with the HGMD database to identify whether they are previously documented mutations associated with specific diseases. Overall, HGMD integrates with various stages of the genomics workflow to enhance the understanding and interpretation of genetic variants, aiding in clinical diagnostics and research.
- **Deep learning models:** Deep learning models are effectively applied across various genomics subareas, such as variant calling and annotation, disease variant prediction, and gene expression regulation. They leverage the massive datasets generated by NGS technologies to make sophisticated predictions, transforming big biological data into actionable insights [80]. An open-source tool developed by Google, DeepVariant[43] utilizes deep learning techniques to call genetic variants from next-generation DNA sequencing data accurately. This tool showcases how machine learning can enhance the precision of intricate bioinformatics tasks such as variant calling, which is a pivotal step preceding annotation. Developed by DeepMind,[44] AlphaFold uses deep learning to predict the 3D structures of proteins from their amino acid sequences. HMMER[45] is particularly useful for the deep annotation of protein sequences. By employing hidden Markov models, it can identify distant homologs and align sequences to protein families. This is crucial for annotating proteins with uncertain functions by revealing evolutionary relationships and potential functional similarities. PyTorch is another widely used open-source machine learning library that supports various deep learning models. In bioinformatics, PyTorch is utilized for tasks like predicting the effects of genetic variants or understanding gene expression patterns.

[38] https://www.htslib.org/

[39] https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/02_variant-calling.html

[40] https://annovar.openbioinformatics.org/en/latest/

[41] https://www.ensembl.org/info/docs/tools/vep/index.html

[42] https://digitalinsights.qiagen.com/

[43] https://github.com/google/deepvariant

[44] https://github.com/google-deepmind/alphafold

[45] https://www.ebi.ac.uk/Tools/hmmer/

B. POTENTIAL SECURITY THREATS

The bioinformatics analysis workflow can be conducted in diverse environments, such as local (''offline'') setups or cloud-based platforms, based on available resources, data privacy requirements, and specific project needs. However, this critical phase is vulnerable to various cyber-biosecurity threats that can jeopardize both the integrity and confidentiality of genomic data as well as the reliability of the analysis processes. Below is a detailed overview of potential cyber threats associated with bioinformatics analysis:

I. **Man in Middle Man (MitM) Attack:** A study [81] highlighted that out of 10,678 bio-informatics websites analyzed, only 5,278 support HTTPS connections. This indicates that approximately 50% of the communications occur over unsecured connections, while the other 50% use HTTPS. Nonetheless, of those employing HTTPS, only 2,727 websites use valid and trusted HTTPS certificates for authentication and authorization processes, which accounts for merely 7.6% of the total domains reviewed. This poses a significant risk in scenarios where MitM attacks can intercept communications between two parties, potentially allowing attackers to eavesdrop on or modify the exchanged data, potentially leading to unauthorized access and data breaches. Furthermore, cyber criminals may introduce vulnerabilities in FTP implementations or intercept SFTP login credentials to steal sensitive sequence data during transfer.

Consider a scenario in which a bioinformatics research group regularly sends genomic data to a collaborating lab for advanced analysis via FTP and HTTP links through their web application. An attacker exploits the unencrypted protocols to perform a MitM attack, intercepting and modifying genomic data files and analysis results, potentially altering sequence alignments and genetic variants. In a clinical setting, altered genetic information can lead to misdiagnosis or incorrect treatment recommendations. To mitigate the risk of MitM attacks, secure protocols like HTTPS and SFTP should be used to encrypt data, making it harder for attackers to intercept or alter information. However, as reported by [81], the implementation of these secure protocols is often inconsistent across bioinformatics web applications.

II. **Attacks on Genomics Cloud Computing Resources**: Bioinformatics analysis step in NGS often relies on centralized cloud services, which, while offering scalability and computational power, are susceptible to various

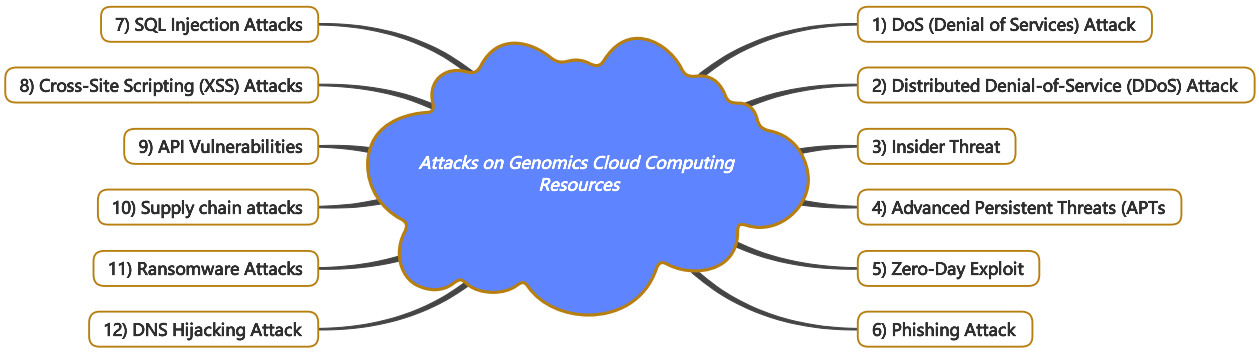


FIGURE 10. Types of Attacks on Genomics Cloud Computing Services.

cyber threats (Fig. 10). Here's an overview of different cyber threats to genomics cloud services could impact bioinformatics analysis workflow:

a) **DoS Attack:** DoS attacks involve flooding servers or network infrastructure with an overwhelming amount of illegitimate traffic or requests. This is done to exhaust resources and bandwidth, thereby rendering the services inaccessible to legitimate users [82]. The primary impact of a DoS attack on NGS services includes delays in genome sequencing processes, potential data loss, and significant disruption of ongoing research and clinical diagnostics. These delays can be critical, especially in time-sensitive situations such as disease outbreak tracking or urgent medical diagnoses.

b) **Distributed Denial-of-Service (DDoS) Attack:** A DDoS attack is a more severe form of DoS that utilizes a network of compromised computers (a botnet) to launch a massive flood of traffic towards the target server or network [83]. In the context of bioinformatics analysis workflows, DDoS attacks can be particularly disruptive, as they can overwhelm the computational infrastructure supporting critical tasks such as sequence alignment, variant calling, and annotation. The scale of these attacks and the distributed nature of the botnets make mitigation challenging, leading to prolonged downtimes. This can severely impact research productivity and compromise the integrity of time-sensitive analyses, overwhelming even robust network defenses.

c) **Insider Threat**: Another critical threat in cloud services is insider threats, which involve malicious insiders with privileged access to cloud resources [84]. These insiders may abuse their credentials to tamper with or exfiltrate sensitive genomic data for malicious purposes. Such actions could lead to data manipulation, deletion, or unauthorized disclosure, posing serious risks to research integrity, patient privacy, and regulatory compliance. Therefore, implementing stringent access controls, monitoring privileged user activities, and conducting regular security audits are essential measures to mitigate the risks associated with insider threats in cloud-based environments.

d) **Advanced Persistent Threats (APTs) Attack:** APTs involve prolonged and targeted cyberattacks where attackers infiltrate a network to steal data or monitor activities over an extended period without being detected [85]. In the context of bioinformatics analysis workflows, APTs can lead to continuous data breaches, intellectual property theft, and long-term espionage. These sophisticated attacks can disrupt critical workflows such as sequence alignment, variant calling, and functional annotation by compromising the integrity and confidentiality of sensitive genomic data. The persistent nature of APTs makes them difficult to detect and mitigate, causing significant strategic damage to research efforts and data security.

e) **Zero-Day Exploit Attack:** This attack takes advantage of previously unknown vulnerabilities in the software. Because it is unknown to the software developers and users at the time of the attack, it is particularly difficult to defend against until patches or updates are released.

f) **Phishing Attack:** Phishing attacks present a significant cybersecurity threat within the context of bioinformatics, particularly when accessing tools or databases online. In these attacks, cybercriminals may impersonate legitimate bioinformatics tools or databases, creating deceptive websites or emails that mimic authentic platforms. Unsuspecting users may be tricked into providing their login credentials or downloading malicious software, compromising the security of their accounts or systems. Once obtained, these credentials can be exploited to gain unauthorized access to sensitive genomic data or to launch further cyberattacks within bioinformatics networks [86]. Therefore, maintaining vigilance,

implementing security awareness training, and adopting robust authentication mechanisms are crucial steps to mitigate the risks associated with phishing attacks in the bioinformatics domain.

g) **SQL Injection Attack:** NGS platforms often provide user interfaces for researchers and clinicians to input data, set parameters for analysis, or retrieve results. These interfaces interact with backend databases that store everything from raw sequencing data to processed results and metadata. If the user inputs are not properly sanitized before being used in SQL queries, attackers can inject malicious SQL code into these queries [87]. The authors in [81] examined various web-based bio-informatics applications and found that a significant majority of them possess outdated versions, rendering them highly susceptible to SQL injection and cross-site scripting (XSS) attack. Attackers could exploit SQL injection vulnerabilities to access sensitive genetic data stored in NGS databases. This could include unauthorized viewing of patient genetic information. Additionally, erroneous mutation data could be inserted, leading researchers or clinicians to make faulty conclusions about genetic predispositions to diseases or the effectiveness of genomics.

h) **Cross-Site Scripting (XSS) Attack:** NGS platforms often provide web-based interfaces for users to upload genetic data, configure analysis parameters, view reports, and share results. These interfaces typically accept input from users, such as metadata, sample descriptions, and custom script parameters. If these inputs are not properly validated or sanitized by the web application, an attacker can inject malicious scripts into these fields. When other users or administrators view pages where these inputs are displayed, the malicious scripts execute within their browsers. XSS can allow attackers to steal session cookies or tokens from users of the NGS platform. This can give attackers unauthorized access to the platform, where they could view or manipulate sensitive genetic data, alter analysis results, or access the personal information of the users.

i) **API Vulnerabilities:** APIs are essential for facilitating communication between client applications and servers via standard HTTP methods. However, their extensive integration into web services and applications makes them prime targets for cyber attacks. Common attacks such as SQL Injection, Command Injection, and Cross-site Scripting (XSS) involve attackers inserting malicious scripts or commands into APIs that interact with databases or backend systems. This can result in unauthorized access to data, data corruption, or even full system control [88]. If not properly secured, these APIs can serve as gateways for further attacks, enabling unauthorized manipulation and access to services.

j) **Supply chain attacks:** Supply chain attacks on cloud services are a growing concern as organizations increasingly rely on complex networks of third-party providers for their cloud infrastructure and services [89]. Attackers might target vendors supplying software or hardware integrated into NGS cloud platforms. By embedding malicious code or backdoors in these components, they can gain unauthorized access to the cloud when these components are deployed.

k) **Ransomware Attacks:** In cloud environments, ransomware can spread rapidly across networked systems. Such attacks can paralyze cloud operations, leading to significant downtime and data loss. Recovery can be costly and time-consuming, and no guarantee paying the ransom will recover the data [90].

l) **DNS Hijacking Attack:** In a DNS hijacking scenario, when researchers or clinicians attempt to access an NGS cloud platform, their DNS query—intended to resolve the domain name of the NGS service to its IP address—gets redirected to a malicious IP address controlled by the attacker. The attacker could take control of the DNS server used by the NGS platform's hosting service. The DNS resolver's cache could be poisoned by inserting false address records, causing users to connect to the wrong server without knowing. The malicious server to which the DNS directs the user could host a spoofed version of the NGS platform. This fake platform might look identical to the legitimate one, tricking users into entering sensitive information. Users might also experience denial of service where they cannot access the legitimate NGS platform due to DNS misdirection, disrupting ongoing research or clinical operations [91].

III. **Algorithm Tampering:** Algorithm tampering involves modifying the algorithms or software used for sequence analysis, variant calling, or annotation to produce incorrect results, introduce biases, or execute unauthorized operations. In bioinformatics analysis workflows, such alterations can lead to incorrect sequence alignments, erroneous variant calls, or flawed annotations. This could lead researchers to draw incorrect conclusions about genetic relationships, functions, or disease associations, thereby compromising the integrity of the research and potentially resulting in flawed scientific findings or misguided clinical decisions.

IV. **Data Poisoning Attack**: In a data poisoning attack, malicious actors inject false or corrupted data into the training datasets used for machine learning models [92]. Within bioinformatics, this can negatively impact the effectiveness of models used for tasks like sequence alignment, variant calling, and annotation. The introduction of poisoned data can result in inaccurate alignments, erroneous variant calls, and flawed annotations, leading

to incorrect conclusions and compromising the reliability of the entire bioinformatics workflow.

V. **Time-of-Check to Time-of-Use (TOCTOU) Attack**: TOCTOU attacks exploit the delay between the validation of a resource and its subsequent use [93]. In bioinformatics workflows, these attacks can be particularly damaging if they target critical stages such as data validation, sequence alignment, or variant calling. For example, an attacker might alter genomic data or computational resources after they have been validated but before they are used, leading to inaccurate results and potentially misleading scientific conclusions. Such attacks compromise the integrity and reliability of bioinformatics analyses.

VI. **Malware Attack**: Malware can infiltrate bioinformatics software or hardware, subtly altering processes such as sequence alignment and introducing errors that may go undetected. These errors could lead to systematic biases in sequencing data, ultimately undermining the reliability and reproducibility of research outcomes.

VII. **Side-Channel Attack**: Side-channel attacks leverage physical characteristics of a sequencing system, such as timing, power consumption, or electromagnetic emissions, to extract sensitive information. This data can potentially reveal critical details, including the sequences being analyzed, compromising the confidentiality of genomic data.

VIII. **Man-in-the-Browser (MitB) Attack**: A variation of the Man-in-the-Middle (MitM) attack, MitB involves malware that infects a user's web browser, enabling real-time modification of web pages, transactions, or content. In bioinformatics, MitB attacks could alter sequence data or parameters entered into web-based alignment tools, leading to flawed analyses or incorrect interpretations.

IX. **Remote Code Execution (RCE)**: RCE attacks take advantage of software vulnerabilities to execute arbitrary code on a remote machine. In the context of NGS workflows, such attacks could allow unauthorized modifications to sequence alignment software, introduce errors into the system, or create backdoors, jeopardizing both the data and the platform's security.

X. **Adversarial Attacks on Deep Learning Models**: These attacks involve injecting misleading data into deep learning models used in genomic analyses, resulting in incorrect predictions or annotations. Additionally, manipulating the training data can cause models to learn inaccurate patterns, reducing their effectiveness and reliability. Such vulnerabilities underscore the need for robust defenses against adversarial inputs, comprehensive validation of training data, and rigorous testing of model outputs to ensure trustworthy genomic analysis.

XI. **Genetic Imputation Attack**: By exploiting correlations between observed and unobserved genetic markers, attackers can infer detailed genetic profiles that were presumed to be anonymized or protected. This poses a significant privacy risk, potentially leading to the unauthorized disclosure of sensitive genetic information.

VII. INTERPRETATION AND REPORTING

Findings from the earlier steps are interpreted and compiled into a report that summarises the genetic results in a clear and concise format. This report typically includes information about the identified variants, their potential biological or clinical significance, and any actionable recommendations. Depending on the preferences of the data owner, the generated genome data can be delivered in various formats, such as FASTQ, SAM, BAM, or Variant Call Format (VCF), either on a physical hard drive or through a cloud-based platform for further bioinformatics analysis.

In clinical settings, the report may also highlight any limitations or uncertainties associated with the findings, ensuring transparency and accuracy. The results are communicated to stakeholders, including healthcare providers, researchers, or study participants, in an understandable and actionable manner. This process often involves discussions to explain the implications of the findings, address any questions, and provide guidance on potential next steps.

VIII. FUTURE RESEARCH DIRECTIONS

The increasing volume and sensitivity of genomic data demand that future research addresses the multifaceted challenges of securing this information. This section outlines key areas where future research should focus to enhance the security, privacy, and integrity of NGS data.

1) **Enhancing Data Security Protocols:** Future research should prioritise the advancement and development of sophisticated and advanced security mechanisms tailored for NGS data. This entails developing resilient encryption techniques that can handle the extensive data generated by NGS while maintaining processing speed and efficiency. Furthermore, it is crucial to examine the possible effects of quantum computing on the security of genomic data. While quantum computing represents a substantial risk to current encryption standards, it simultaneously offers the potential to create novel and robust encryption methods capable of protecting sensitive genomic data.

2) **Implementing AI and ML:** The integration of AI with ML presents significant potentials for the real-time forecasting and prevention of cyber-bio threats. These technologies can analyse trends, discover anomalies in network traffic, monitor user activities, and assess system operations, facilitating the detection of potential security breaches prior to their occurrence [94]. Future research should focus on creating automated AI-driven reaction systems that can rapidly address identified risks. These technologies would provide immediate solutions to counteract cyber-attacks, thereby mitigating damage, decreasing downtime, and ensuring the robustness of essential genetic data processes.

3) **Developing Comprehensive Cyber-Biosecurity Frameworks:** The development and establishment of a comprehensive cyber-biosecurity architecture and framwork are essential for protecting genetic data throughout NGS life cycle, including data collection, processing, and storage. Future research should prioritise the development of standardised protocols and guidelines which could be adobted and impmented by the laboratories and institutions. These frameworks must incorporate optimal techniques for data anonymisation, access control, and secure data sharing to avert unwanted access and data breaches. By using a unified and systematic approach, organizations and instituitions engaged in the NGS process may uphold optimal cyber-biosecurity and therby, effectively mitigating potential threats to sensitive genetic data.
4) **Strengthening Interdisciplinary Collaboration:** In order to address the complex cyber-biosecurity challenges to NGS, it is necessary to strengthen interdisciplinary collaboration among experts in bioinformatics, cybersecurity, computing, and law. Efforts should focus on fostering collaborations across multiple disciplines in research projects. These partnerships aim to close knowledge gaps between different fields, enabling the creation of security solutions that are not only technically strong but also adhere to legal requirements. Activities like conferences, joint workshops, and collaborative research initiatives can significantly aid in promoting this collaborative attitude. Moreover, public-private partnerships that offer funding and support for genetic data security research might enhance the acceleration of technology and information transfer among government, industry, and academic institutions. This will thus foster innovation in this critically significant subject.
5) **Addressing Ethical and Privacy Concerns:** Ethical and privacy considerations in genetic data management should be a primary focus of future research. This entails not just guaranteeing compliance with existing regulations but also proactively confronting growing legal and ethical dilemmas. Research must focus on establishing ethical frameworks that govern the responsible utilisation of NGS data, including informed consent procedures, data ownership rights, and the consequences of prolonged data storage. Such frameworks will foster confidence among stakeholders and guarantee the equitable and responsible use of genetic data.
6) **Exploring Decentralised Data Management:** Future study should examine its potential for decentralised storage and data administration, emphasising on ensuring the data integrity and traceability. The capacity of blockchain to sustain an immutable ledger for monitoring data access and alterations can enhance trust and accountability in NGS operations. Moreover, high-performance centralised ledger databases, as outlined in [95], provide tamper resistance and enhanced performance, rendering them a feasible substitute for blockchain technology. These solutions tackle scalability and compliance issues encountered by decentralised technologies, rendering them essential in the formulation of secure and effective genomic data management methods.
7) **Advancing Data Anonymization Techniques**: Given the importance and ongoing concerns about re-identification security challenges, researchers must prioritise the development of sophisticated anonymisation strategies that safeguard individual identities while preserving data usability for research purposes. This entails examining the use of differential privacy methods in genomic data, which involve introducing controlled noise to the data to avert re-identification while maintaining its research significance. Furthermore, investigating federated learning for genomic data analysis enables institutions to cooperatively train machine learning models while preserving the confidentiality of raw data, thus enhancing privacy and security.
8) **Developing Automated Quality Control Mechanisms:** Mitigating the possibility of malicious users injecting extraneous or detrimental data into public databases requires the establishment of automated quality control systems. These systems must be able to detect and flag suspicious data uploads for examination, hence reducing dependence on manual supervision by database administrators in NGS. This is particularly significant as current quality control procedures frequently do not adequately address this issue. Automated methods will guarantee that extensive dataset uploads are meticulously controlled, averting the unforeseen incorporation of arbitrary or dangerous data into public genomic repositories.
9) **Designing Specialized Intrusion Detection Systems for NGS Data:** Developing specialized intrusion detection systems (IDS) tailored for the unique demands of NGS data is critical to securing its flow and ensuring data integrity. These IDS should leverage advanced anomaly detection algorithms capable of contextualising and understanding the specific characteristics and behaviour of NGS data transactions. By incorporating both supervised and unsupervised machine learning techniques, the accuracy and reliability of anomaly detection can be significantly enhanced.

   Real-time monitoring and stream processing capabilities are essential for enabling immediate threat detection and response. Furthermore, the IDS must adhere to regulatory frameworks such as GDPR and HIPAA, ensuring compliance while maintaining detailed audit trails for forensic analysis. Advanced solutions, such as LedgerDB kchain-based systems [96], offer features like controlled data access and tamper resistance, making them highly effective for maintaining compliance. Similarly, VeDB [97] provides practical solutions for

addressing GDPR-related challenges, combining data mutability with robust security mechanisms. In addition to leveraging behavioural profiling of users and systems to improve detection accuracy, integrating global threat intelligence feeds can bolster the system's ability to counteract emerging threats, ensuring the comprehensive protection of NGS workflows.

10) **Advanced Access Control Mechanisms for Cloud-Based NGS Data**: As the storage of vast amounts of NGS data increasingly shifts to cloud platforms, advanced access control mechanisms are becoming crucial to ensure both security and privacy. Future research should emphasize the development and implementation of privacy-preserving access control schemes designed specifically for the unique requirements of genomic data. Techniques such as spatial range queries and time-controllable keyword search offer precise, restricted access to data subsets such as, specific genomic regions or time-sensitive clinical trial datasets—without exposing the entire dataset [98], [99], [100]. Leveraging technologies like encrypted spatial indexing [101], attribute-based encryption (ABE) [102], and blockchain-based time controls [103] can significantly enhance the confidentiality, integrity, and accessibility of NGS data while maintaining user convenience. Therefore, future studies should focus on integrating these mechanisms into existing cloud-based platforms for NGS data. This includes assessing computational efficiency, scalability, and compatibility with current genomic data sharing systems. By prioritizing such advanced access control methods, the research community can establish a secure, adaptable, and privacy-focused framework for managing and sharing NGS data in the cloud.

IX. CONCLUSION

NGS technology has revolutionized genomic research and healthcare, offering unprecedented opportunities for advancements in personalized medicine, cancer genomics, and forensic investigations. However, the rapid growth and widespread adoption of NGS technologies have introduced significant cyber-biosecurity challenges that threaten the confidentiality, integrity, and availability of genomic data. Existing cybersecurity frameworks remain inadequate in addressing the unique vulnerabilities inherent in the NGS workflow, particularly in the context of AI-driven genomic threats, synthetic DNA-based malware, and adversarial sequencing attacks.

This study makes a significant contribution to cyber-biosecurity by introducing a structured taxonomy of security threats across the entire NGS workflow, an area previously underexplored in genomic cybersecurity. Unlike prior work that focuses primarily on database security and privacy, our study broadens the scope by identifying and categorizing newly emerging threats in both the experimental and bioinformatics analysis phases. By providing a technical framework for assessing and mitigating these risks, this research lays the foundation for developing advanced security frameworks tailored to genomic data protection.

While several mitigation strategies have been proposed, including secure data transmission protocols, role-based access control, and data integrity measures, there remain notable gaps in current security frameworks. Specifically, a more detailed examination of sophisticated attack vectors, such as advanced persistent threats (APTs), supply chain vulnerabilities, and AI-powered cyber threats, is crucial. Future research should focus on developing risk assessment models tailored specifically to NGS, integrating blockchain, quantum cryptography, and AI-driven threat detection into genomic security frameworks, and establishing standardized protocols for cyber-biosecurity in genomic research.

Addressing these challenges will require interdisciplinary collaboration between bioinformaticians, cybersecurity experts, policymakers, and ethicists to ensure that genomic research remains both innovative and secure. By developing comprehensive and adaptive security frameworks, the scientific community can safeguard NGS technologies against evolving cyber threats, ultimately promoting their long-term reliability and impact in research and clinical applications. This work serves as a call to action for strengthening cyber-biosecurity in the genomic era and ensuring that the benefits of NGS technology can be fully realized without compromising data security or ethical integrity.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest to disclose.

DATA AVAILABILITY

All the data is available within the manuscript.

REFERENCES

[1] S. Behjati and P. Tarpey, "What is next generation sequencing?" *Arch. Disease Childhood-Educ. Pract.*, vol. 98, no. 6, pp. 236–238, Aug. 2013.

[2] V. Wadhwa, "The genetic engineering genie is out of the bottle and could unleash the next pandemic," Foreign Policy, Washington, DC, USA, Sep. 2020. [Online]. Available: https://foreignpolicy.com/2020/09/11/crispr-pandemic-gene-editing-virus/

[3] K. A. Phillips, M. P. Douglas, and D. A. Marshall, "Expanding use of clinical genome sequencing and the need for more data on implementation," *Jama*, vol. 324, no. 20, pp. 2029–2030, 2020.

[4] H. Buermans and J. Den Dunnen, "Next generation sequencing technology: Advances and applications," *Biochimica et Biophysica Acta (BBA)-Mol. Basis Disease*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014.

[5] S. L. Castro-Wallace et al., "Nanopore DNA sequencing and genome assembly on the international space station," *Sci. Rep.*, vol. 7, no. 1, p. 18022, Dec. 2017.

[6] S. E. Duncan, R. Reinhard, R. C. Williams, F. Ramsey, W. Thomason, K. Lee, N. Dudek, S. Mostaghimi, E. Colbert, and R. Murch, "Cyber-biosecurity: A new perspective on protecting U.S. food and agricultural system," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 63, Mar. 2019.

[7] M. Bhushan, "Cyber-biosecurity," *J. Defense Stud*, vol. 17, no. 2, pp. 93–119, 2023.

[8] "Cyberattack information centre," Synnovis, London, U.K. [Online]. Available: https://www.synnovis.co.uk/cyberattack-information-centre

[9] S. Alder, "Octapharma plasma notifies individuals affected by April 2024 ransomware attack," HIPAA J., USA, Sep. 2024. [Online]. Available: https://www.hipaajournal.com/octapharma-ransomware-attack/

[10] C. Souza, "What has pharma learned from the Merck cyber attack," Tech. Rep., 2025. Accessed: Jan. 16, 2025.

[11] E. Global. (2025). *Notification of Ransomware Incident*. Accessed: Jan. 16, 2025. [Online]. Available: https://www.eisai.com/news/2023/news202341.html

[12] K. Millett, E. dos Santos, and P. D. Millett, "Cyber-biosecurity risk perceptions in the biotech sector," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 136, Jun. 2019.

[13] A. Mouton, C. Lucas, and E. Guest, "The operational risks of AI in large-scale biological attacks: A red-team approach," RAND Corp., Santa Monica, CA, USA, Res. Rep. RR-A2977-1, 2023. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA2977-1.html

[14] I. Fayans, Y. Motro, L. Rokach, Y. Oren, and J. Moran-Gilad, "Cyber security threats in the microbial genomics era: Implications for public health," *Eurosurveillance*, vol. 25, no. 6, Feb. 2020, Art. no. 1900574.

[15] P. Kulkarni and P. Frommolt, "Challenges in the setup of large-scale next-generation sequencing analysis workflows," *Comput. Structural Biotechnol. J.*, vol. 15, pp. 471–477, Jan. 2017.

[16] M. M. Alsaffar, M. Hasan, G. P. McStay, and M. Sedky, "Digital DNA lifecycle security and privacy: An overview," *Briefings Bioinf.*, vol. 23, no. 2, p. 607, Mar. 2022.

[17] J. Peccoud, J. E. Gallegos, R. Murch, W. G. Buchholz, and S. Raman, "Cyberbiosecurity: From naive trust to risk awareness," *Trends Biotechnol.*, vol. 36, no. 1, pp. 4–7, Jan. 2018.

[18] B. A. Vinatzer, L. S. Heath, H. M. Almohri, M. J. Stulberg, C. Lowe, and S. Li, "Cyberbiosecurity challenges of pathogen genome databases," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 106, May 2019.

[19] G. J. Schumacher, S. Sawaya, D. Nelson, and A. J. Hansen, "Genetic information insecurity as state of the art," *Frontiers Bioeng. Biotechnol.*, vol. 8, Dec. 2020, Art. no. 591980.

[20] MITRE Corp. (2025). *MITRE ATT&CK Framework*. Accessed: Feb. 18, 2025. [Online]. Available: https://attack.mitre.org/

[21] Microsoft Corp. (2025). *The STRIDE Threat Model*. Accessed: Feb. 18, 2025. [Online]. Available: https://docs.microsoft.com/en-us/security/engineering/vstride-threat-model

[22] P. Ney, K. Koscher, L. Organick, L. Ceze, and T. Kohno, "Computer security, privacy, and DNA sequencing: Compromising computers with synthesized DNA," in *Proc. 26th USENIX Secur. Symp.*, 2017, pp. 765–779.

[23] P. M. Ney, "Securing the future of biotechnology: A study of emerging biocyber security threats to DNA-information systems," Ph.D. dissertation, Dept. Comput. Sci. Eng., Univ. Washington, Seattle, WA, USA, 2019. [Online]. Available: https://digital.lib.washington.edu/researchworks/items/d79bdff5-80a3-47dc-b23e-2a474840a03c

[24] (2024). *Sequencing Analysis Viewer Software*. Accessed: Apr. 27, 2024. [Online]. Available: https://emea.support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html

[25] A. S. C. Gaia, M. S. De Oliveira, G. D. S. Moia, V. C. Dos Santos, J. T. C. Alves, P. H. C. G. De Sá, and A. A. D. O. Veras, "E-QA NGS: A user-friendly tool to preprocessing data from next generation sequencing," *Peer Rev.*, vol. 5, no. 3, pp. 91–105, Mar. 2023.

[26] S. Arshad, J. Arshad, M. M. Khan, and S. Parkinson, "Analysis of security and privacy challenges for DNA-genomics applications and databases," *J. Biomed. Informat.*, vol. 119, Jul. 2021, Art. no. 103815.

[27] Oxford Gene Technology, "What lies ahead: The future of next generation sequencing," OGT Blog, Oxford, U.K., Oct. 2022. [Online]. Available: https://www.ogt.com/about-us/ogt-blog/what-lies-ahead-the-future-of-next-generation-sequencing/

[28] M. T. Pervez, S. H. Abbas, M. F. Moustafa, N. Aslam, and S. S. M. Shah, "A comprehensive review of performance of next-generation sequencing platforms," *BioMed Res. Int.*, vol. 2022, no. 1, 2022, Art. no. 3457806.

[29] M. Pop, "DNA sequence assembly algorithms," in *McGraw-Hill 2006 Yearbook of Science and Technology*. New York, NY, USA: McGraw-Hill, 2006.

[30] H. Lee, J. Gurtowski, S. Yoo, M. Nattestad, S. Marcus, S. Goodwin, W. R. McCombie, and M. C. Schatz, "Third-generation sequencing and the future of genomics," *BioRxiv*, vol. 2016, Apr. 2016, Art. no. 048603.

[31] A. Guha Neogi, A. Eltaher, and A. Sargsyan, "NGS data analysis with apache spark," in *Computational Life Sciences: Data Engineering and Data Mining for Life Sciences*. Cham, Switzerland: Springer, 2023, pp. 441–467.

[32] P. C. Ng and E. F. Kirkness, "Whole genome sequencing," *Methods Mol. Biol.*, vol. 628, pp. 215–226, 2010, doi: 10.1007/978-1-60327-367-1_12. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20238084/

[33] K. Retterer et al., "Clinical application of whole-exome sequencing across clinical indications," *Genet. Med.*, vol. 18, no. 7, pp. 696–704, Jul. 2016.

[34] A. Negi, A. Shukla, A. Jaiswar, J. Shrinet, and R. S. Jasrotia, "Applications and challenges of microarray and RNA-sequencing," in *Bioinformatics*. Amsterdam, The Netherlands: Elsevier, Oct. 2022, pp. 91–103.

[35] T.-M. Ko, J. V. Beltran, and J.-Y. Huang, "Transcriptomics in Kawasaki disease," in *Kawasaki Disease*. Cham, Switzerland: Springer, 2022, pp. 123–130.

[36] D. Shyr and Q. Liu, "Next generation sequencing in cancer research and clinical application," *Biol. Procedures Online*, vol. 15, no. 1, pp. 1–11, Dec. 2013.

[37] A. Gouello, C. Dunyach-Remy, C. Siatka, and J.-P. Lavigne, "Analysis of microbial communities: An emerging tool in forensic sciences," *Diagnostics*, vol. 12, no. 1, p. 1, Dec. 2021.

[38] H. Kim, S. Kim, and S. Jung, "Instruction of microbiome taxonomic profiling based on 16S rRNA sequencing," *J. Microbiol.*, vol. 58, no. 3, pp. 193–205, Mar. 2020.

[39] M. Wang, "Pharmacogenomics in the clinic," *Nature Rev. Nephrology*, vol. 19, no. 5, p. 277, May 2023.

[40] R. Yee and P. J. Simner, "Next-generation sequencing approaches to predicting antimicrobial susceptibility testing results," *Clinics Lab. Med.*, vol. 42, no. 4, pp. 557–572, Dec. 2022.

[41] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Rev. Genet.*, vol. 15, no. 6, pp. 409–421, Jun. 2014.

[42] A. Abdel-Latif and G. Osman, "Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize," *Plant Methods*, vol. 13, no. 1, pp. 1–9, Dec. 2017.

[43] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, Jan. 2013.

[44] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, Jan. 2000.

[45] L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name (A re-identification experiment)," 2013, *arXiv:1304.7605*.

[46] E. Ayday and J.-P. Hubaux, "Privacy and security in the genomic era," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1863–1865.

[47] E. Halperin and D. A. Stephan, "SNP imputation in association studies," *Nature Biotechnol.*, vol. 27, no. 4, pp. 349–351, Apr. 2009.

[48] D. A. Wheeler et al., "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, no. 7189, pp. 872–876, Apr. 2008.

[49] V. V. Cogo, A. Bessani, F. M. Couto, and P. Verissimo, "A high-throughput method to detect privacy-sensitive human genomic data," in *Proc. 14th ACM Workshop Privacy Electron. Soc.*, Oct. 2015, pp. 101–110.

[50] E. Global. (2025). *Case Study Economic Espionage*. Accessed: Jan. 16, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/2808138.2808139

[51] B. P. Hennig, L. Velten, I. Racke, C. S. Tu, M. Thoms, V. Rybin, H. Besir, K. Remans, and L. M. Steinmetz, "Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol," *G3, Genes, Genomes, Genet.*, vol. 8, no. 1, pp. 79–89, Jan. 2018.

[52] (2024). *The Importance of Sequencing Clean-Up*. Accessed: Jun.25, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29118030/

[53] (2024). *DNA Sizing and Cleanup for Ngs DNA Library Purification*. Accessed: Jun. 25, 2024. [Online]. Available: https://www.thermofisher.com/blog/behindthebench/the-importance-of-sequencing-clean-up-seq-it-out-9/

[54] J. F. Hess, T. A. Kohl, M. Kotrová, K. Rönsch, T. Paprotka, V. Mohr, T. Hutzenlaub, M. Brüggemann, R. Zengerle, S. Niemann, and N. Paust, "Library preparation for next generation sequencing: A review of automation strategies," *Biotechnol. Adv.*, vol. 41, Jul. 2020, Art. no. 107537.

[55] (2024). *Echo Liquid Handlers*. Accessed: Apr. 27, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734975020300343
[56] S. S. Ali, M. Ibrahim, J. Rajendran, O. Sinanoglu, and K. Chakrabarty, "Supply-chain security of digital microfluidic biochips," *Computer*, vol. 49, no. 8, pp. 36–43, Aug. 2016.
[57] A. G. Bracamonte, "Microarrays towards nanoarrays and the future next generation of sequencing methodologies (NGS)," *Sens. Bio-Sensing Res.*, vol. 37, Aug. 2022, Art. no. 100503.
[58] Y. Zhang, Y. Li, and Z. Li, "Aye: A trusted forensic method for firmware tampering attacks," *Symmetry*, vol. 15, no. 1, p. 145, Jan. 2023.
[59] Y. Wu, J. Wang, Y. Wang, S. Zhai, Z. Li, Y. He, K. Sun, Q. Li, and N. Zhang, "Your firmware has arrived: A study of firmware update vulnerabilities," in *Proc. USENIX Secur. Symp.*, 2023, pp. 5627–5644.
[60] S. Abaimov, "Understanding and classifying permanent denial-of-service attacks," *J. Cybersecurity Privacy*, vol. 4, no. 2, pp. 324–339, May 2024.
[61] G. Ma, Q. Tang, W. Zhang, and N. Yu, "Tamper restoration on DNA sequences based on reversible data hiding," in *Proc. 6th Int. Conf. Biomed. Eng. Informat.*, Dec. 2013, pp. 484–489.
[62] J. Fu, W. Zhang, N. Yu, G. Ma, and Q. Tang, "Fast tamper location of batch DNA sequences based on reversible data hiding," in *Proc. 7th Int. Conf. Biomed. Eng. Informat.*, Oct. 2014, pp. 868–872.
[63] C. Ledergerber and C. Dessimoz, "Base-calling for next-generation sequencing platforms," *Briefings Bioinf.*, vol. 12, no. 5, pp. 489–497, Sep. 2011.
[64] (2024). *Bcl2fastq2 Conversion User Guide*. Accessed: May 10, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21245079/
[65] (2024). *Basespace Sequence Hub*. Accessed: Jun. 25, 2024. [Online]. Available: https://bioweb.pasteur.fr/docs/modules/bcl2fastq/2.15.0/bcl2fastq2-user-guide-15051736-b.pdf
[66] D. R. Nyholt, C.-E. Yu, and P. M. Visscher, "On jim Watson's APOE status: Genetic information is hard to hide," *Eur. J. Human Genet.*, vol. 17, no. 2, pp. 147–149, Feb. 2009.
[67] G. Dán and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun.*, Gaithersburg, MD, USA, 2010, pp. 214–219, doi: 10.1109/SMARTGRID.2010.5622046.
[68] T. Warnow, "Revisiting evaluation of multiple sequence alignment methods," in *Multiple Sequence Alignment* (Methods in Molecular Biology), vol. 2231, K. Katoh, Ed., New York, NY, USA: Humana, 2021, doi: 10.1007/978-1-0716-1036-7_17.
[69] R. Sharma, "An overview of variant calling and analysis in NGS data," Tech. Rep., 2024. Accessed: Jun. 30, 2024.
[70] X. Dong, T. Xiao, B. Chen, Y. Lu, and W. Zhou, "Precision medicine via the integration of phenotype-genotype information in neonatal genome project," *Fundam. Res.*, vol. 2, no. 6, pp. 873–884, Nov. 2022.
[71] D. C. Koboldt, "Best practices for variant calling in clinical sequencing," *Genome Med.*, vol. 12, no. 1, p. 91, Dec. 2020.
[72] G. F. Ejigu and J. Jung, "Review on the computational genome annotation of sequences obtained by next-generation sequencing," *Biology*, vol. 9, no. 9, p. 295, Sep. 2020, doi: 10.3390/biology9090295.
[73] M. DePristo and R. Poplin, "Deepvariant: Highly accurate genomes with deep neural networks," *Google AI Blog*, 2017.
[74] (2024). *Restful API*. Accessed: May 10, 2024. [Online]. Available: https://research.google/blog/deepvariant-highly-accurate-genomes-with-deep-neural-networks/
[75] B. Schmidt and A. Hildebrandt, "Next-generation sequencing: Big data meets high performance computing," *Drug Discovery Today*, vol. 22, no. 4, pp. 712–717, Apr. 2017.
[76] (2024). *Multiple Sequence Alignment and NJ / UPGMA Phylogeny*. Accessed: May 10, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28163155/
[77] K. Dang, Y. Zhao, K. Ye, Y. Guo, W. Wang, Q. Ge, and X. Zhao, "Construction of multiplexed transcriptome NGS libraries of microdissected tissue samples based on combinational DNA barcode microbeads," *Biotechnol. J.*, vol. 19, no. 1, 2024, Art. no. 2300294.
[78] (2024). *Genome Analysis Toolkit Variant Discovery in High-throughput Sequencing Data*. Accessed: May 10, 2024. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/biot.202300294?casa_token=HePOMOWHliYAAAAA%3AIIGuI4TwA-dlS5cwJBrTMvPFbCJYxu8KFwlIcnL-4Rluydf9Cmu1beQUTFghI-P5UZfjW4OukfSa8w
[79] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3," *Fly*, vol. 6, no. 2, pp. 80–92, Apr. 2012.
[80] W. S. Alharbi and M. Rashid, "A review of deep learning applications in human genomics using next-generation sequencing data," *Human Genomics*, vol. 16, no. 1, p. 26, Jul. 2022.
[81] T. Tao, Y. Chen, B. Liu, X. Jin, M. Yan, and S. Ji, "Security analysis of bioinformatics Web application," in *Proc. Int. Conf. Secur. Intell. Comput. Big-data Services*. Cham, Switzerland: Springer, 2020, pp. 383–397.
[82] A. B. de Neira, B. Kantarci, and M. Nogueira, "Distributed denial of service attack prediction: Challenges, open issues and opportunities," *Comput. Netw.*, vol. 222, Feb. 2023, Art. no. 109553.
[83] P. Shukla, C. R. Krishna, and N. V. Patil, "Iot traffic-based DDoS attacks detection mechanisms: A comprehensive review," *J. Supercomput.*, vol. 80, no. 7, pp. 9986–10043, May 2024.
[84] A. S. Asha, D. Shanmugapriya, and G. Padmavathi, "Malicious insider threat detection using variation of sampling methods for anomaly detection in cloud environment," *Comput. Electr. Eng.*, vol. 105, Jan. 2023, Art. no. 108519.
[85] A. I. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "A survey on security and privacy issues in modern healthcare systems: Attacks and defenses," *ACM Trans. Comput. Healthcare*, vol. 2, no. 3, pp. 1–44, Jul. 2021.
[86] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers Comput. Sci.*, vol. 3, Mar. 2021, Art. no. 563060.
[87] X. Yang, C. Yue, W. Zhang, Y. Liu, B. C. Ooi, and J. Chen, "SecuDB: An in-enclave privacy-preserving and tamper-resistant relational database," *Proc. VLDB Endowment*, vol. 17, no. 12, pp. 3906–3919, Aug. 2024.
[88] M. A. M. Ariffin, M. F. Ibrahim, and Z. Kasiran, "API vulnerabilities in cloud computing platform: Attack and detection," *Int. J. Eng. Trends Technol.*, vol. 1, pp. 8–14, Oct. 2020.
[89] O. Akinrolabu, S. New, and A. Martin, "Cyber supply chain risks in cloud computing–bridging the risk assessment gap," *Open J. Cloud Comput.*, vol. 5, no. 1, pp. 1–19, 2017.
[90] J. K. Lee, S. Y. Moon, and J. H. Park, "CloudRPS: A cloud analysis based enhanced ransomware prevention system," *J. Supercomput.*, vol. 73, no. 7, pp. 3065–3084, Jul. 2017.
[91] R. Houser, S. Hao, Z. Li, D. Liu, C. Cotton, and H. Wang, "A comprehensive measurement-based investigation of DNS hijacking," in *Proc. 40th Int. Symp. Reliable Distrib. Syst. (SRDS)*, Sep. 2021, pp. 210–221.
[92] X. Zhang, X. Zhu, and L. Lessard, "Online data poisoning attacks," *Learn. Dyn. Control*, vol. 120, pp. 201–210, Jul. 2020.
[93] Y. Y. Zhuang and Y.-N. Tseng, "A novel detection method for the security vulnerability of time-of-check to time-of-use," *J. Inf. Sci. Eng.*, vol. 38, no. 6, p. 1171, 2022.
[94] J.-H. Li, "Cyber security meets artificial intelligence: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 12, pp. 1462–1474, Dec. 2018.
[95] X. Yang, S. Wang, F. Li, Y. Zhang, W. Yan, F. Gai, B. Yu, L. Feng, Q. Gao, and Y. Li, "Ubiquitous verification in centralized ledger database," in *Proc. IEEE 38th Int. Conf. Data Eng. (ICDE)*, May 2022, pp. 1808–1821.
[96] X. Yang, Y. Zhang, S. Wang, B. Yu, F. Li, Y. Li, and W. Yan, "LedgerDB: A centralized ledger database for universal audit and verification," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 3138–3151, Aug. 2020.
[97] X. Yang, R. Zhang, C. Yue, Y. Liu, B. C. Ooi, Q. Gao, Y. Zhang, and H. Yang, "VeDB: A software and hardware enabled trusted relational database," *Proc. ACM Manage. Data*, vol. 1, no. 2, pp. 1–27, Jun. 2023.
[98] Y. Miao, Y. Yang, X. Li, Z. Liu, H. Li, K. R. Choo, and R. H. Deng, "Efficient privacy-preserving spatial range query over outsourced encrypted data," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3921–3933, 2023.
[99] Y. Miao, Y. Yang, X. Li, L. Wei, Z. Liu, and R. H. Deng, "Efficient privacy-preserving spatial data query in cloud computing," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 122–136, Jan. 2023.
[100] Y. Miao, F. Li, X. Li, Z. Liu, J. Ning, H. Li, K. R. Choo, and R. H. Deng, "Time-controllable keyword search scheme with efficient revocation in mobile e-health cloud," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 3650–3665, May 2023.

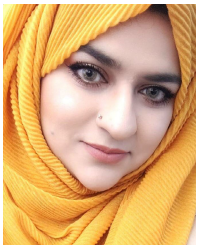
[101] B. Wang, M. Li, and H. Wang, "Geometric range search on encrypted spatial data," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 4, pp. 704–719, Apr. 2016.
[102] Y. Zhang, R. H. Deng, S. Xu, J. Sun, Q. Li, and D. Zheng, "Attribute-based encryption for cloud computing access control: A survey," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–41, 2020.
[103] D. Di Francesco Maesa, P. Mori, and L. Ricci, "Blockchain based access control," in *Proc. 17th IFIP Int. Conf. Distrib. Appl. Interoperable Syst.*, Neuchâtel, Switzerland. Cham, Switzerland: Springer, Jun. 2017, pp. 206–220.

NASREEN ANJUM received the Ph.D. degree from the Department of Informatics, King's College London, U.K. She is currently an Assistant Professor in computer science with the School of Computing, University of Portsmouth, U.K. Her research interests include machine learning algorithms, cybersecurity issues in IoT, and health technologies. She was a recipient of the Future Fellowship Award from the Schlumberger Faculty, from 2014 to 2018.

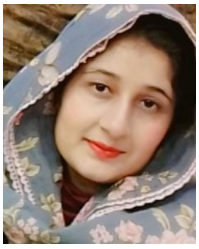
HANI ALSHAHRANI (Senior Member, IEEE) received the bachelor's degree in computer science from King Khalid University, Abha, Saudi Arabia, the master's degree in computer science from California Lutheran University, Thousand Oaks, CA, USA, and the Ph.D. degree from Oakland University, Rochester, MI, USA. He is currently an Associate Professor of computer science and information systems with Najran University, Najran, Saudi Arabia. His current research interests include smartphones, the IoT, crowdsourcing security, and privacy.

ASADULLAH SHAIKH (Senior Member, IEEE) received the B.Sc. degree in software development from the University of Huddersfield, England, U.K., the M.Sc. degree in software engineering and management from the University of Gothenburg, Sweden, and the Ph.D. degree in software engineering from the University of Southern Denmark. He was a Researcher at Universitat Oberta de Catalunya (UOC), Barcelona, Spain. He is currently a Full Professor at the College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia. He has more than 200 publications in the area of software engineering in international journals and conferences. His publications have been cited more than 2500 times (H-index of 26 and i10 index of 81) according to Google Scholar. He has vast experience in teaching and research. His current research interests include software engineering, health informatics, artificial intelligence using machine learning, cloud computing, e-learning, and mobile technologies. He is also an editor and an associate editor of several SCIE journals and an International Advisory Board of several conferences. Further details can be obtained using www.asadshaikh.com.

MAHREEN-UL-HASSAN received the Ph.D. degree from the University of Sheffield, U.K. She is currently a Senior Lecturer at Shaheed Benazir Bhutto Women University, Peshawar, and a dedicated microbiologist and educator with expertise in teaching, research, and academic leadership. Beyond academia, she actively promotes environmental sustainability and community health initiatives, inspiring the next generation of scientists through innovative teaching and mentorship. She has secured multiple research grants, published extensively, and received accolades, such as the Schlumberger Faculty for the Future Fellowship and the Best Teacher Award. Her research interests include protein structure determination, environmental microbiology, and climate resilience, with notable work on peptidoglycan-binding domains in Enterococcus species.



MEHREEN KIRAN received the M.S. degree in computer science from the prestigious Institute of Management Sciences, Peshawar, in 2018. She is currently pursuing the Ph.D. degree with Anglia Ruskin University, with a focus on the fascinating intersection of machine learning and digital twin technology. During her academic tenure, she exhibited exceptional dedication and aptitude, resulting in the honor of receiving the Best Performance Award from the University of Peshawar, a testament to exceptional academic record and dedication to her studies.



SHAH RAZ received the bachelor's degree in biotechnology and genetic engineering from The University of Agriculture, Peshawar, in 2020, and the M.Phil. degree in microbiology from Shaheed Benazir Bhutto Women University. Her research interests include microbial ecology, extremophiles, bacterial characterization, and molecular techniques for microbial analysis. She also participated in an international conference on seed storage techniques hosted by The University of Agriculture.

ABU ALAM currently serves as an Academic Course Leader, the Lead for Collaborative Partnerships, and a Senior Lecturer. With an extensive background in software engineering, he has gained experience across USA, Denmark, and U.K., contributing to the design of various products. He has also held the position of the Project Lead at a major U.K. university. In addition to his academic roles, he holds several esteemed positions, including the Chair of a Professional Computing Society, a Governor, and the QA Lead for an educational institution. He also serves as an External Examiner for multiple universities in U.K. and an Apprenticeship End Point Assessor. His research interests include cyber security and computer science, including machine learning, malware analysis, HCI, secure coding, and automation.