**ACKNOWLEDGEMENT**

We are indebted to Rashtreeya Sikshana Samithi Trust, Bengaluru for providing us with all the facilities needed for the successful completion of our Minor Project work at Rashtreeya Vidyalaya College of Engineering (RVCE) during the tenure of our Course.

We would like to thank Dr. K N Subramanya, Principal, for giving us an opportunity to be a part of RVCE and for his timely help and encouragement during the tenure of the Minor Project work.

We are greatly thankful to Dr. Shanta Rangaswamy, Professor and Head, Dept. of CSE for her motivation and constant support during the tenure of our Minor Project Work.

We take this opportunity to convey our sincere gratitude to the internal guide Dr. Hemavathy R, Professor, Dept of CSE for her advice, support and valuable suggestions helped us to accomplish the Minor Project work in time.

Special thanks to Dr. Ramakanthkumar P., Dean CSE Cluster and panel members Dr. Rajashree Shettar, Professor, Dr. Azra Nasreen, Associate Professor for their valuable comments, constructive inputs and feedback during the Phase- I, Phase-II and Phase-III presentations.

We extend our thanks to all who have directly or indirectly extended their constant support for successful completion of our Minor Project work

**Name**

**USN**

**Name**

**USN**

**ABSTRACT**

The COVID-19 pandemic has underscored the need for comprehensive genomic analysis to understand viral evolution, mutation patterns, and their implications on disease severity and persistence. Genomic studies of SARS-CoV-2, particularly in prolonged COVID-19 cases, provide critical insights into the genetic variations that may influence viral transmission, immune escape, and treatment resistance.

This project focuses on the computational analysis of SARS-CoV-2 genomes to detect and characterize genetic variants affecting key viral genes. Utilizing a bioinformatics pipeline, the workflow integrates multiple steps, including sequence alignment, variant calling, annotation, consensus sequence generation, and pathway analysis. The pipeline employs tools such as BWA for sequence alignment, FreeBayes for variant calling and for variant annotation, ensuring accurate identification of mutations. Additionally, BLAST is used for sequence comparison, while BedTools enables the assessment of genomic features affected by mutations.

The findings highlight the most affected genes, suggesting potential implications in viral pathogenicity and persistence. This research provides critical insights into the genomic variations associated with prolonged COVID-19, aiding further investigations into its evolutionary dynamics and functional consequences. The work is carried out sequentially, starting from raw genome sequence preprocessing to variant annotation and functional impact analysis. Variant clustering highlights recurring mutations that may contribute to viral adaptation and immune evasion. The results provide insights into the evolutionary dynamics of SARS-CoV-2, assisting researchers in prioritizing mutations for further experimental validation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TABLE OF CONTENTS** | | | | | |
| **Acknowledgement** | | | | | i |
| **Abstract** | | | | | ii |
| **Table of Contents** | | | | | iii |
| **List of Figures** | | | | | v |
| **Glossary** | | | | | vi |
| **Chapter 1** | | | | |  |
| **Introduction** | | | | |  |
| 1.1 | Overview | | | | 7 |
| 1.2 | Literature Review | | | | 8 |
| 1.3 | Motivation | | | | 17 |
| 1.4 | Problem statement | | | | 18 |
| 1.5 | Objectives | | | | 19 |
| 1.7 | Methodology | | | | 20 |
| 1.8 | Organization of report | | | | 23 |
| **Chapter 2** | | | | |  |
| **Theory and Concepts of Gene Pathway Mapping of SARS-COV-2 Virus** | | | | |  |
| 2.1 | | Biotechnology Role in Genomic Research | | | 24 |
| 2.2 | | DNA and RNA: The Genetic Material | | | 24 |
| 2.3 | | Genome Sequences and Mutation | | | 25 |
| 2.4 | | Bioinformatics | | | 26 |
| 2.5 | | Protein Structure and Function | | | 27 |
| **Chapter 3** | | | | | |
| **Requirement Specification** | | | |  | |
| 3.1 | | | Software Requirements | 29 | |
| 3.2 | | | Hardware Requirements | 30 | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter 4** | | | |
| **Implementation of the Project** | | |  |
| 4.1 System Architecture | | | 31 |
| 4.2 Sequence Alignment and Compression | | | 32 |
| 4.3 Variant Detection and Annotation | | | 33 |
| 4.4 Gene Pathway Mapping and Clustering | | | 34 |
| 4.5 Integration and Visualization | | | 35 |
| 4.6 Applications of Gene Pathway Mapping and Clustering | | | 36 |
| **Chapter 5** | | | |
| **Experiment Analysis and Results of Gene Pathway Mapping** | | | |
| 5.1 | Input to the Pathway Analysis | | 37 |
| 5.2 | Execution of the Pipeline | | 39 |
| 5.3 | Pathway Analysis | | 40 |
| **Chapter 6** | | | |
| **Conclusion and Future Work** | | |  |
| 7.1 | | Conclusion | 46 |
| 7.2 | | Future Work | 47 |
| **REFERENCES** | | | 48 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig No.** | **Figure Name** | **Page No.** |
| 1.1 | Representation of the structure of SARS-COV-2 genome | 12 |
| 1.2 | Cluster View of Syndromes of Long Covid | 14 |
| 1.3 | Methodology of the Project | 24 |
| 2.1 | Structure of DNA and RNA | 27 |
| 2.2 | Structure of Protein | 29 |
| 4.1 | Overview of the Implementation Steps | 33 |
| 4.2 | Flowchart of the Architecture | 34 |
| 4.3 | Burrows Wheel Transform of DNA Sequences | 35 |
| 4.4 | Variant Identification in Genomes | 36 |
| 4.5 | User Interface of the Project | 37 |
| 5.1 | NCBI Website to Obtain Genome Data | 39 |
| 5.2 | Execution of Pipeline | 41 |
| 5.3 | Stage I of the Execution Pipeline | 42 |
| 5.4 | Stage II of the Execution Pipeline | 43 |
| 5.5 | Genes Impacted by Variants | 45 |

|  |  |  |
| --- | --- | --- |
|  |  | **GLOSSARY** |
|  |  |  |
| ATGC | : | Adenine, Thymine, Guanine, Cytosine |
| DNA | : | Deoxyribonucleic Acid |
| RNA | : | Ribonucleic Acid |
| Alignment | : | Process of comparing and matching a sample DNA/RNA sequence to a reference genome to identify similarities and differences |
| BAM | : | Binary Alignment Map |
| Bioinformatics | : | An interdisciplinary field that applies computational techniques to analyze biological and genomic data |
| BLAST | : | Basic Local Alignment Search Tool |
| FASTA | : | File format for storing nucleotide structures |
| GFF | : | General Feature Format. Standardized file format to annotate genomic features such as genes |
| Long COVID | : | Prolonged COVID-19 |
| ORF | : | Open Reading Frame |
| SAM | : | Sequence Alignment Map |
| SARS-COV-2 | : | Severe Acute Respiratory Syndrome Coronavirus 2 |
| SNP | : | Single Nucleotide Polymorphism |
| Variant | : | A version of a virus that has accumulated genetic mutations that  may affect its properties, such as transmissibility or immune escape |
| VCF | : | Variant Call Format |

**INTRODUCTION**

This chapter describes the introduction, background study, motivation, objective, problem statement and methodology of the project. It focuses on the gene pathway mapping of SARS-CoV-2 to explore the complex interplay between viral proteins, their variants affecting the function. In this chapter we have discussed the introduction to the project of building an efficient computational pipeline of the gene pathway mapping.

* 1. **OVERVIEW**

Prolonged COVID-19 has emerged as a significant concern globally [1], with a subset of individuals experiencing persistent symptoms long after the acute phase of the infection. Understanding the genomic basis of prolonged COVID-19 can provide crucial insights into its pathophysiology and aid in the development of targeted therapeutic strategies. The COVID-19 pandemic has led to extensive research efforts to understand the genomic structure of SARS-CoV-2 and its mutations [2], particularly in prolonged cases of infection. This project focuses on the gene pathway mapping and clustering of prolonged COVID-19 cases by analyzing genomic sequences to identify mutations and assess their impact on viral gene functions.

The study employs a computational bioinformatics pipeline to systematically process viral genome sequences, detect genetic variants, and interpret their biological significance. This project outlines a comprehensive approach to decoding prolonged COVID by integrating population genetics and pathway analysis, providing a framework for predicting geographic regions, identifying key genes, and uncovering pathways that can inform targeted treatments.

By identifying key pathways that are altered or hijacked by the virus, we aim to uncover potential targets for therapeutic intervention. The study incorporates bioinformatics tools, databases, and molecular biology techniques to track the gene expression changes and identify significant pathways involved in the viral lifecycle, immune evasion, and inflammation.

Initially, sequences obtained from NCBI are subjected to phylogenetic and clustering analysis using CLC Genomics Workbench, followed by multiple sequence alignment with MAFFT. The reference genome is indexed, and sequences are aligned to produce SAM and BAM files, leading to variant calling through Freebayes. Consensus sequences are generated using tools like bcftools, and subsequent BLAST analysis is performed to compare clusters against the reference genome. Python scripts further analyze SNP variations, generating Excel reports that highlight significant differences.

The workflow offers insights into geographic-specific genetic variations, helping to reveal pathways involved in prolonged COVID, which could ultimately aid in the development of more effective, regionally-tailored treatments.

* 1. **LITERATURE REVIEW**

The phenomenon of Long COVID, characterized by persistent symptoms following SARS-CoV-2 infection, has been extensively studied to understand its impact, pathophysiology, and management strategies. One such comprehensive study is the systematic scoping review conducted by Akbarialiabad et al. (2021), published by Springer Nature, which provides critical insights into the challenges associated with defining and managing Long COVID [1].

The review analyzed 120 publications, categorizing research into three primary areas:

* 49% focused on signs and symptoms, detailing persistent issues such as fatigue, breathlessness, and other long-term health consequences affecting multiple organ systems.
* 23% addressed management strategies, exploring treatment approaches and rehabilitation efforts to mitigate symptoms.
* 11% examined pathophysiology, identifying potential mechanisms driving prolonged illness and the biological factors contributing to its persistence.

One of the key findings of the review is the lack of a universally accepted definition of Long COVID, which has hindered proper recognition, diagnosis, and treatment strategies. The variability in reported symptoms and their duration across studies underscores the complexity of post-COVID conditions.

This project builds upon existing research by applying genomic and bioinformatics approaches to analyze the mutational landscape of SARS-CoV-2 in prolonged COVID cases. While previous studies primarily focused on clinical symptoms and management, our research shifts the focus to genetic factors influencing prolonged infection, using computational tools for variant detection, annotation, and gene pathway analysis. By identifying affected viral genes and their role in prolonged infections, this study aims to bridge the gap between genomic alterations and clinical manifestations of the virus.

The prolonged and complex post-viral condition has emerged as a significant global health challenge. The study by Davis et al. (2023), published in Nature Reviews: Microbiology, provides a detailed overview of the prevalence, symptoms, underlying mechanisms, and gaps in current treatment approaches [2].

The study highlighted that Long COVID affects approximately 10% of individuals infected with SARS-CoV-2, translating to an estimated 65 million cases globally. Over 200 symptoms have been documented, impacting multiple organ systems, including the neurological, cardiovascular, and immune systems, alongside myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)-like manifestations. A significant misconception regarding Long COVID is its classification as a purely respiratory condition, whereas research indicates wide-ranging systemic effects.

Despite the growing recognition of Long COVID, diagnostic tools and treatment options remain insufficient, necessitating urgent prioritization of clinical trials. Current medical interventions fail to adequately address the condition’s complexity, leaving many patients without effective management strategies. However, vaccination has been shown to reduce the risk of Long COVID by 15% to 41%, offering a level of protective benefit.

While previous research has largely focused on clinical symptomatology and epidemiology, this project aims to explore the genomic underpinnings of prolonged COVID-19 cases through variant analysis and gene pathway mapping. By identifying mutations in viral genes that may contribute to prolonged infections, this study provides a complementary perspective to existing research, potentially guiding future therapeutic and diagnostic advancements.

The study by Rahimi et al. (2021), published in Elsevier, explores the genetics and genomics of SARS-CoV-2, providing a foundational understanding of viral genetic diversity and genome detection techniques [3]. Understanding the genetic composition of SARS-CoV-2 is essential for developing effective diagnostic strategies, treatment approaches, and vaccine designs.

The complete SARS-CoV-2 genome sequence (MN908947.3) is available on GenBank, enabling researchers to perform sequence alignment and phylogenetic tree analysis, which has confirmed its classification as a new member of the β-coronavirus (β-CoV) family. The virus is an enveloped, non-segmented, positive-sense single-stranded RNA virus, encoding four key structural proteins:

* Spike (S) – Facilitates viral entry into host cells.
* Envelope (E) – Involved in virus assembly and release.
* Membrane (M) – Maintains the structural integrity of the virus.
* Nucleocapsid (N) – Encapsulates viral RNA for replication.

This research is crucial for gene pathway analysis in prolonged COVID-19 cases, as understanding mutational patterns in key SARS-CoV-2 genes can provide insights into variant-induced changes in viral function and pathogenicity. By analyzing genomic variations and their impact on disease progression, this study contributes to the broader field of viral genomics, vaccine development, and personalized medicine.

Additionally, SARS-CoV-2 contains 12 open reading frames (ORFs), encoding 27 functional proteins, including those responsible for viral replication and host interaction as visualised in the Figure 1.1 below.

A diagram of a virus

AI-generated content may be incorrect.

Figure.1.1 Representation of the structure of SARS-CoV-2 genome [3]

SARS-CoV-2 exhibits a mutation rate of 8 × 10⁻⁴ nucleotides per year, with 17 high-frequency mutations identified in various genes, including ORF1ab, ORF3a, ORF6, S, M, E, and N. The most commonly observed mutations include the changes from 241C > T, 3037C > T, 23403A > G. The D614G in the Spike protein, linked to increased transmissibility.

The study by Niewolik et al. (2024), published in Springer, explores long COVID symptom clustering to understand its syndromic nature and long-term consequences [4]. By assessing 2,371 participants, the study identified distinct symptom clusters, which provide insights into the heterogeneous presentation of long COVID.

Using cluster analysis, the researchers categorized long COVID symptoms into three primary groups:

* Cluster A: Rheumatological symptoms (joint pain, muscle pain, inflammation)
* Cluster B: Neuro-physiological symptoms combined with cardiorespiratory issues (brain fog, memory loss, breathlessness, fatigue, heart palpitations)
* Cluster C: General infection symptoms, dermatological conditions, and otological issues (fever, skin rashes, tinnitus).

This classification highlights the multisystem impact of long COVID, reinforcing the need for personalized treatment approaches. The study relied on self-reported symptom assessments, which may introduce subjective biases. Additionally, the participant group may not be fully representative of the general population, limiting the study’s generalizability. The figure 1.2 depicts the three primary categories based on the symptoms below,

A graph of a number of diseases

AI-generated content may be incorrect.

Figure 1. 2 Cluster View of Syndromes of Long Covid [4]

Future research should focus on linking these symptom clusters to underlying pathophysiological mechanisms, which could help in targeted therapeutic interventions and improve diagnostic accuracy for long COVID.

Understanding symptom clusters is essential in gene pathway mapping, as long COVID symptoms may be linked to persistent genetic or immunological changes. This study provides a valuable framework for correlating genomic mutations with distinct clinical presentations, supporting efforts to analyze the biological basis of the virus.

The study by Smith et al. (2022), published in Molecular and Cellular Probes section under Elsevier, investigates molecular gene signatures mediated by SARS-CoV-2 and provides a comparative transcriptomic analysis of five major respiratory viruses [5].

The five viruses under study were SARS-CoV-2, SARS-CoV-1, Influenza A, RSV, and Rhinovirus. The study aims to identify conserved and unique genetic markers to understand the impact of viral infections on gene expression and immune response.

The study examined 42 transcriptomic datasets comprising over 12,000 genes to identify differentially expressed genes in infected individuals. The analysis found strongest gene regulation similarities between SARS-CoV-2 and RSV, while SARS-CoV-2 and SARS-CoV-1 showed minimal correlation despite both being coronaviruses. Immune Pathway Alterations in SARS-CoV-2 uniquely impacted JAK-STAT immune signaling, cytokine regulation, and DNA replication pathways.

This research provides crucial insights for gene pathway mapping in prolonged COVID-19 cases, as it highlights specific genetic alterations induced by SARS-CoV-2. The identification of differentially regulated genes and immune response pathways is instrumental in understanding how viral mutations influence prolonged infection and disease severity, supporting the objectives of genomic-based SARS-CoV-2 studies.

The systematic scoping review by Akbarialiabad et al. (2021), published in Infection (Springer), provides a comprehensive analysis of Long COVID, its symptoms, pathophysiology, and management. The study reviewed 120 publications to consolidate knowledge on post-COVID-19 syndrome and highlight research gaps [6].

The paper identified the symptom prevalence:

* 49.1% of studies focused on symptoms, identifying fatigue, breathlessness, sleep difficulties, and joint pain as predominant complaints.
* 10.8% of studies examined pathophysiology, noting multi-organ involvement, including neurological, cardiovascular, and renal complications.
* 23.3% of studies addressed management, but effective treatment strategies remain unclear.

The lack of a standardized definition has hindered proper diagnosis and management, leading to under-recognition in clinical settings. The study emphasized that clinical trials are scarce, with only one randomized controlled trial (RCT) identified. Future research should focus on biomarkers, genetic predispositions, and long-term therapeutic strategies to address the prolonged impact of SARS-CoV-2 infections.

This review highlights the long-term effects of COVID-19, which may be linked to underlying genetic factors. By analyzing gene pathways and mutation clusters in SARS-CoV-2, this project aims to bridge the gap between genomic alterations and persistent clinical symptoms, contributing to improved diagnostics and treatment strategies.

The study by Nunes et al. (2022), published in PLOS ONE, investigates the evolutionary dynamics of SARS-CoV-2 variants through deep phylogenetic-based clustering analysis **[7]**. The study focuses on identifying shared and novel mutations in Variants of Concern (VOC) and Variants of Interest (VOI) across multiple geographic regions, including Brazil, India, and the USA.

The study reveals the presence of new and previously unreported mutations in SARS-CoV-2 variants due to directional and convergent evolution. By analyzing over 200,000 whole-genome sequences, researchers identified recombination breakpoints and sites under directional selection in the Spike protein, suggesting their role in immune evasion and increased transmissibility.

Despite genetic similarities, SARS-CoV-2 and SARS-CoV-1 showed minimal correlation in their mutation patterns, reinforcing SARS-CoV-2’s unique evolutionary trajectory. The study found that co-circulating variants and their sub-lineages exhibit similar evolutionary pathways, contributing to regional differences in variant spread and infection rates.

The research highlights the importance of ongoing genomic surveillance to track mutation-driven variant evolution. Additionally, understanding how these mutations affect vaccine efficacy is crucial for future therapeutic and immunization strategies.

This research complements gene pathway mapping in prolonged COVID-19 cases by identifying genomic plasticity and mutational hotspots that influence variant persistence and severity. By integrating variant analysis with pathway mapping, this study aids in understanding the long-term genetic adaptations of SARS-CoV-2.

Longitudinal Host Transcriptional Responses to SARS-CoV-2 Infection study investigates gene expression changes in individuals infected with SARS-CoV-2, focusing on those with extremely high viral loads **[12]**.

Using RNA sequencing of serially collected nasal swabs, the research identifies a dynamic transcriptional response that correlates with viral load. Key findings highlight a strong activation of immune-related genes, including interferon-stimulated genes, cytokines, and stress response markers. The study also contrasts gene expression patterns between high and low viral load cases, revealing that early immune activation is crucial for controlling viral replication.

Additionally, comparisons with in vitro models, such as human nose organoids, suggest distinct host responses based on cellular context. The findings contribute to understanding COVID-19 pathogenesis and may aid in developing targeted therapeutic strategies​.

The study by Taheri & Habibi (2022), published in Applied Soft Computing (Elsevier), presents a comprehensive pathway analysis of COVID-19 using unsupervised machine learning techniques [9]. The research focuses on identifying key biological pathways and molecular mechanisms involved in SARS-CoV-2 infections.

The study employs a two-stage machine learning approach, selecting four key gene sets representing critical COVID-19-related pathways. The research constructs two distinct biological networks for signaling and disease pathways, ranking pathways based on an unsupervised scoring method.

The study emphasizes the need for further validation of identified pathways through experimental and clinical studies. Additionally, integrating multi-omics data could enhance the understanding of COVID-19's systemic effects.

The study by Greenhalgh et al. (2024), published in The Lancet, provides a comprehensive clinical update on Long COVID, addressing its epidemiology, pathophysiology, clinical manifestations, and management strategies [10].

The research consolidates findings from multiple disciplines, integrating virological, immunological, and clinical perspectives to enhance understanding of this persistent post-viral condition. Long COVID is defined as symptoms persisting beyond three months after acute infection, affecting multiple organ systems and leading to severe functional impairment. Possible causes include viral persistence, immune dysregulation, endothelial dysfunction, and microclot formation.

No single biomarker has been identified for diagnosis, complicating clinical assessment. Variability in symptoms and progression makes standardized treatment difficult. Higher risk observed in unvaccinated individuals, women aged 35–50 years, and those with pre-existing conditions like diabetes and asthma.

Vaccination reduces Long COVID risk but does not eliminate it entirely. Despite thousands of studies on Long COVID, clinicians lack standardized diagnostic tools and treatment protocols. The condition places a significant strain on healthcare infrastructure, with many cases going unrecognized. Emerging studies focus on deep molecular profiling and biomarker discovery, which may lead to personalized treatment strategies.

This research reinforces the need for genomic and bioinformatics-based investigations to uncover the underlying molecular pathways involved in Long COVID. By identifying mutational signatures in SARS-CoV-2 that contribute to prolonged infections, this study complements efforts to map gene pathways and develop targeted therapeutic interventions.

The study by Umesh et al. (2022), published in Infection (Springer), presents evidence mapping and review of Long COVID [11], focusing on its pathophysiological mechanisms and systemic complications. The research consolidates data from clinical trials, observational studies, and case reports to highlight the multisystem impact of post-COVID syndrome.

Long COVID symptoms persist for weeks to months post-infection, affecting multiple organ systems. Pulmonary, neuro-psychological, and cardiovascular complications are the most reported.

Potential mechanisms include viral persistence, immune dysregulation, endothelial dysfunction, and microclot formation. A lack of standardized diagnostic criteria leads to under-recognition and misdiagnosis.

Over 50% of industry-sponsored trials focus on pulmonary complications, while cardiovascular research is underrepresented, highlighting gaps in clinical research efforts. Long-term effects on reproductive health and metabolic functions remain poorly understood.

This study underscores the importance of genomic and pathway-based research to decode the molecular underpinnings of Long Covid. By analyzing mutational impacts on viral genes, this project aims to bridge the gap between genetic variations and persistent symptoms, contributing to enhanced diagnostics and therapeutic strategies.

* 1. **MOTIVATION**

The rapid evolution of SARS-CoV-2 variants has significantly impacted public health, with emerging mutations leading to increased transmissibility, immune evasion, and potential resistance to existing vaccines and treatments. While extensive research has been conducted on acute COVID-19 infections, prolonged COVID-19 cases remain poorly understood, posing a challenge for clinicians and researchers alike. Understanding the genetic factors that contribute to prolonged viral persistence and its clinical manifestations is essential for improving disease management and therapeutic interventions.

Advancements in genomic surveillance and computational bioinformatics have enabled large-scale mutation tracking and functional impact analysis, offering deeper insights into variant evolution and host-virus interactions.

However, major challenges remain in accurately mapping gene pathways, identifying high-impact mutations, and correlating them with prolonged infection symptoms. The bioinformatics sector, projected to reach $21.8 billion by 2027, has become a critical field in vaccine development, epidemiology, and personalized medicine, yet efficient computational frameworks for prolonged COVID-19 genomic analysis are still lacking.

This project is driven by the need to bridge this knowledge gap by systematically analyzing SARS-CoV-2 genome mutations in prolonged COVID-19 cases. By employing advanced computational tools and bioinformatics techniques, this study aims to map affected gene pathways, identify critical mutations, and assess their functional consequences. The insights gained from this research will contribute to better diagnostics, targeted therapeutic strategies, and an enhanced understanding of viral evolution in prolonged infections, making it a vital step toward controlling the long-term impact of COVID-19.

* 1. **PROBLEM STATEMENT**

The prolonged effects of COVID-19, commonly referred to as Long COVID, have emerged as a significant public health concern, affecting millions of individuals worldwide. Despite extensive research on the acute phase of SARS-CoV-2 infection, the genomic factors contributing to prolonged viral persistence and post-infection complications remain poorly understood. Patients with prolonged COVID-19 exhibit a wide range of persistent symptoms, including fatigue, respiratory distress, neurological impairments, and cardiovascular issues, with no clear understanding of the underlying molecular mechanisms driving these manifestations.

One of the critical challenges in addressing prolonged COVID-19 is the lack of comprehensive gene pathway mapping and clustering analyses to identify how viral mutations impact host-virus interactions and disease progression.

While genomic surveillance has successfully tracked SARS-CoV-2 mutations, current research lacks a systematic approach to link genetic variations with the biological pathways involved in prolonged infections. Moreover, the absence of a standardized computational framework for analyzing mutational effects on viral proteins and their role in immune evasion, persistence, and pathogenicity limits the development of effective diagnostic and therapeutic strategies.

This project aims to address these challenges by performing large-scale mutation analysis, pathway mapping, and clustering of genetic variations observed in prolonged Covid-19 cases. By leveraging bioinformatics tools and computational approaches, this study will identify key viral genes affected by mutations, assess their functional consequences, and establish correlations with disease severity and persistence.

The findings will contribute to improving genomic-based diagnostics, targeted treatment approaches, and a better understanding of the long-term effects of SARS-CoV-2 infections, ultimately aiding in the management and prevention of long covid.

* 1. **OBJECTIVES**

The primary objective of this project is to analyze the genetic variants associated with prolonged COVID-19 and map their impact on biological pathways using computational bioinformatics techniques. The objectives align with the broader goal of understanding the prolonged effects of COVID-19 at a genomic level and developing computational frameworks to facilitate large-scale mutation tracking and functional interpretation. The specific objectives are as follows:

* To collect and process SARS-CoV-2 genomic sequences from affected populations and perform comparative analysis with the reference genome to identify significant variations.
* To identify and characterize genetic mutations associated with prolonged COVID-19 symptoms, enabling a deeper understanding of viral evolution.
* To analyze the functional impact of these mutations on key viral genes and host-virus interactions by mapping them to biological pathways and organ systems.
* To implement computational tools and automation techniques for accurate variant annotation, gene clustering, and pathway analysis, improving efficiency and reproducibility.
* To provide actionable insights into potential health impacts, aiding in the development of targeted diagnostics, therapeutic interventions, and long-term management strategies for prolonged COVID-19.
* To integrate advanced computational techniques, such as SNP analysis, population genetics, and machine learning, alongside a web-based interface for an interactive, comprehensive, and scalable analysis platform.
  1. **METHODOLOGY**

This project follows a structured bioinformatics workflow to analyze genetic variants of SARS-CoV-2 in prolonged COVID-19 cases. The methodology involves multiple phases, including genomic data acquisition, sequence alignment, variant calling, pathway mapping, and computational analysis, ensuring a comprehensive investigation of viral mutations and their impact as outlined in the figure 1.3 below.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 1. 3 Methodology of the Project

**Data Collection and Preprocessing**

This step involves gathering SARS-CoV-2 genome sequences from publicly available databases such as GISAID, NCBI GenBank, and EMBL-EBI. The focus is on acquiring sequences from individuals suffering from prolonged COVID-19 symptoms to detect mutations that may contribute to long-term viral persistence.

The reference genome of SARS-CoV-2 (NC\_045512.2) is used for comparison. Data quality control measures are implemented to filter out low-quality sequences, ensuring the accuracy of subsequent analyses.

**Sequence Alignment and Preprocessing**

The collected viral genome sequences are aligned to the reference genome using computational tools to detect genetic variations.

This step ensures that sequence mismatches corresponding to mutations are accurately captured. The raw alignment data is converted into a structured format, sorted, and indexed for efficient downstream processing. This step is essential for optimizing variant detection, functional analysis, and pathway mapping in the later stages of the project.

**Variant Calling and Annotation**

Genetic variants are identified by comparing the aligned sequences with the reference genome to detect mutations associated with prolonged COVID-19. The identified mutations are then annotated to classify their location, type, and potential functional impact on viral proteins and biological pathways. Special attention is given to mutations in key genes such as ORF1ab, S (Spike), ORF3a, and M (Membrane), which have been implicated in viral replication and immune escape mechanisms.

**Consensus Sequence Generation**

A consensus sequence is generated by integrating the identified mutations into the reference genome, allowing for a comprehensive analysis of mutation patterns in prolonged COVID-19 cases. This process refines the dataset by removing low-confidence mutations and ensuring that only high-confidence, biologically relevant changes are considered for further study. The consensus sequence serves as a baseline for comparative genomic and functional analyses.

**Comparative Analysis and Pathway Mapping**

The consensus sequence is compared to the reference genome to assess the functional consequences of mutations. Identified mutations are mapped to biological pathways using databases such as KEGG and Reactome to understand their role in viral pathogenesis and prolonged COVID-19 symptoms. This step helps in linking genetic variations to potential disruptions in viral replication, host immune response, and disease severity.

**Data Interpretation and Visualization**

The identified mutations are grouped using clustering and classification techniques to detect patterns in mutation frequency and disease severity. Advanced bioinformatics visualization tools are used to generate mutation heatmaps, pathway diagrams, and statistical reports to better understand the genomic alterations observed. This step aids in drawing correlations between genomic variations and clinical outcomes, providing valuable insights into the nature of prolonged infections. The entire analysis pipeline is automated to streamline the process of genomic data analysis, variant annotation, and pathway mapping. A user-friendly web-based interface is developed to allow researchers and healthcare professionals to access and interpret SARS-CoV-2 mutation data efficiently.

* 1. **ORGANIZATION OF REPORT**

Chapter 2 Provides the theory and concepts of gene pathway mapping and clustering of SARS-COV-2 virus and overview of key terminologies.

Chapter 3 Supplies information about the Hardware requirements of processor, storage and other physical hardware required as well as Software requirements of applications.

Chapter 4 Depicts the detailed design, system architecture and flow the project.

Chapter 5 Delivers the description of the execution of gene pathway pipeline, experimental results and analysis about the genes impacted by variants.

Chapter 6 Yields the conclusion of the project and discusses the limitation and future enhancements of the project.

**THEORY AND CONCEPTS OF GENE PATHWAY MAPPING AND CLUSTERING OF SARS-COV-2**

This chapter provides an outline of the steps involved in the process of gene pathway mapping and clustering of the prolonged covid virus. This project falls within the intersection of biotechnology, bioinformatics, and genomic analysis, aiming to understand how genetic variations in SARS-CoV-2 contribute to prolonged covid infections. The following are the key theoretical concepts relevant to this domain.

* 1. **BIOTECHNOLOGY ROLE IN GENOMIC RESEARCH**

Biotechnology is the application of biological systems, organisms, or their derivatives to develop new technologies for healthcare, agriculture, and environmental science. In the context of this project, biotechnology enables genome sequencing, genetic engineering, and bioinformatics-driven data analysis to study mutations in the virus. Techniques such as PCR (Polymerase Chain Reaction), Next-Generation Sequencing (NGS), and CRISPR-based gene editing have revolutionized virology research by allowing rapid genome sequencing, mutation detection, and functional analysis of viral genes.

* 1. **DNA AND RNA: THE GENETIC MATERIAL**

DNA (Deoxyribonucleic Acid) and RNA (Ribonucleic Acid) are the fundamental molecules responsible for storing and transmitting genetic information in living organisms. In the case of SARS-CoV-2, the genetic material is RNA, making it different from human DNA-based genomes.

DNA: A double-stranded molecule composed of nucleotides (A, T, C, G) that carries genetic instructions for the development and functioning of all known living organisms. It is composed of long chains of nucleotides, each containing a sugar molecule sugar molecule, a phosphate group, and one of four nitrogenous bases. The two strands of DNA are held together by hydrogen bonds between the complementary bases, with Adenine pairing with Thymine and Cytosine pairing with Guanine.

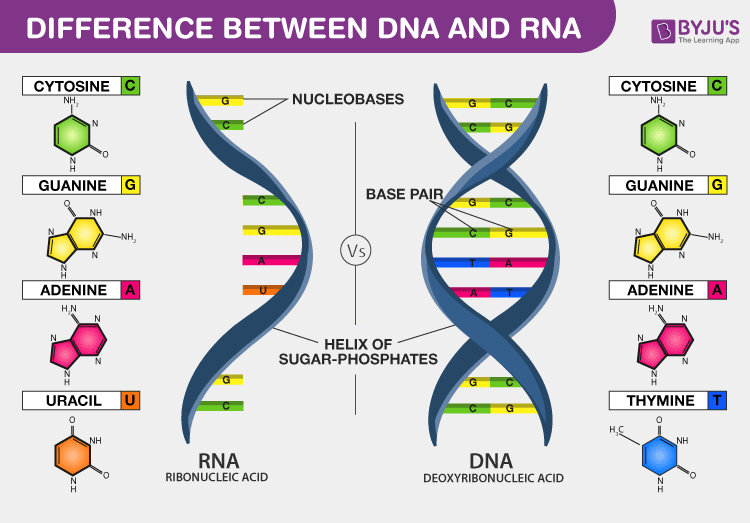
RNA: A single-stranded molecule that plays a crucial role in protein synthesis and viral replication. SARS-CoV-2 is an RNA virus, meaning it relies on host cellular machinery to replicate and propagate within human cells as displayed in figure 2.1 below.

Figure 2. Structure of DNA and RNA [8]

SARS-CoV-2 contains a positive-sense single-stranded RNA genome, which means its RNA can be directly translated into proteins by the host's ribosomes. Understanding RNA structure, mutations, and its interactions with host proteins is critical in tracking viral evolution and developing targeted therapies.

* 1. **GENOME SEQUENCES AND MUTATION**

Genome sequencing is the process of determining the complete nucleotide sequence of an organism's DNA or RNA. For SARS-CoV-2, Next-Generation Sequencing (NGS) is widely used to sequence viral genomes and identify genetic variations.

Mutations are changes in the viral genome that can occur due to errors in replication, environmental pressures, or immune system responses. Some common types of mutations include:

* Single Nucleotide Polymorphisms (SNPs): A single base change in the genome.
* Insertions and Deletions (InDels): Addition or removal of nucleotides, thus altering gene the function.
* Missense and Nonsense Mutations: Changes in codons that lead to altered or premature stop codons in protein synthesis.

Understanding these mutations helps in tracking viral evolution, vaccine resistance, and immune escape mechanisms.

* 1. **BIOINFORMATICS**

Bioinformatics is the interdisciplinary field that combines biology, computer science, and mathematics to analyze and interpret biological data, particularly genomic sequences. It is essential for understanding the genetic variations in SARS-CoV-2 by providing computational methods to analyze and interpret genomic data. Sequence alignment allows researchers to compare viral genome sequences, identifying mutations that differentiate variants and influence viral evolution. Variant annotation helps assess the biological significance of these mutations, determining whether they alter protein structures, disrupt viral functions, or enhance infectivity and immune evasion. Phylogenetic analysis traces the evolutionary history by grouping related viral strains based on genetic similarities, enabling scientists to track the emergence and spread of new variants. Pathway mapping links identified mutations to specific biological processes, helping to predict their impact on viral replication, immune system interactions, and potential disease severity. These analyses contribute to a deeper understanding of how genetic changes affect viral behavior, providing critical insights for vaccine design, antiviral drug development, and public health interventions to control the spread of prolonged and severe covid cases.

* 1. **PROTEIN STRUCTURE AND FUNCTION**

Proteins are essential biological molecules that carry out cellular functions, facilitate viral replication, and interact with the host immune system. Proteins are essential biological macromolecules that play a fundamental role in nearly every cellular process, including structural support, enzymatic catalysis, signal transduction, immune defense, and molecular transport. They are composed of amino acids linked together by peptide bonds, forming a polypeptide chain that folds into a specific three-dimensional structure to carry out its function. The structure of a protein is crucial for determining its activity, interactions, and overall biological role as described in the figure 2.2 below.

A diagram of a structure

AI-generated content may be incorrect.

Figure 2. Structure of Protein [8]

Primary Structure: This refers to the linear sequence of amino acids in a polypeptide chain. The specific arrangement of amino acids dictates how the protein will fold and function. Any mutations or substitutions in the primary structure can significantly affect the protein's stability and activity.

Secondary Structure: Local folding patterns within a protein give rise to common structural motifs such as α-helices and β-sheets, which are stabilized by hydrogen bonds. These structures provide the foundation for the overall folding and stability of the protein.

Tertiary Structure: The complete three-dimensional folding of a protein occurs when the secondary structures interact through hydrogen bonds, disulfide bridges, hydrophobic interactions, and ionic bonds. This final shape is essential for the protein’s function, as it determines how the protein interacts with other molecules, including DNA, RNA, and receptors.

Quaternary Structure: Some proteins are composed of multiple polypeptide chains, known as subunits, that assemble into a functional unit. Hemoglobin, for example, consists of four subunits that work together to transport oxygen.

1. :

**REQUIREMENT SPECIFICATION**

This chapter is requirement specification that includes pointers about what softwares are used to carry out the project, The hardware involved in order to successfully implement the proposed ideas. This helps in understanding the tools used to achieve the desired outcome and their roles in each process.

* 1. **SOFTWARE REQUIREMENTS**

The software requirements for this project encompass a combination of bioinformatics tools and computational frameworks, to facilitate genome sequence processing, variant detection, gene pathway mapping. The software components required are:

* **Operating System** – The project requires a Linux-based environment (Ubuntu 20.04 or later) for running bioinformatics tools efficiently. Linux distributions provide native support for genomic data processing, high-performance computing, and compatibility with most open-source bioinformatics software.
* **Sequence Alignment and Preprocessing Tools** – The Burrows-Wheeler Aligner (BWA) is used for mapping sequencing reads to the SARS-CoV-2 reference genome. SAMtools is required for manipulating sequence alignment files, including sorting, indexing, and converting formats. These tools ensure optimized processing of large-scale genomic datasets.
* **Variant Calling and Annotation Software** – FreeBayes is utilized as the primary Bayesian genetic variant caller to identify SNPs, indels, and complex mutations. BCFTools is required for annotating detected variants, classifying their functional impact on viral proteins and pathways.
* **Pathway Mapping and Functional Analysis Tools** – Databases such as Kyoto Encyclopedia of Genes and Genomese, Reactome are integrated for gene pathway analysis. BedTools is used to intersect variant data with genomic annotations, identifying how mutations affect specific genes.
* **Web-Based Interface for Mutation Analysis** – If a web-based interface is developed, Streamlit framework will be used to create an interactive dashboard for genomic data visualization and variant exploration.
  1. **HARDWARE REQUIREMENTS**

Hardware requirements for the implementation and testing of this project are as follows:

* **Processor (CPU):** Multi-core processor (Intel i7 or AMD Ryzen 7 & above) for efficient parallel processing of genomic data.
* **Graphics Processing Unit (GPU):** NVIDIA Tesla T4 (Google Colab) or RTX 3090 for accelerated machine learning tasks and visualization of large datasets.
* **Random Access Memory (RAM):** Minimum 16 GB; recommended 32 GB or higher for handling large-scale genomic datasets and complex computations.
* **Storage:** Solid State Drive (SSD) with at least 1 TB capacity for fast read/write access to large sequencing files.
* **High-Performance Computing (HPC) Support:** Access to a cloud-based or local HPC cluster with multiple CPU cores for parallelized genome analysis.
* **Network Connectivity:** High-speed internet for retrieving genome sequences, accessing online databases, and performing cloud-based computations.
* **Display & Visualization Setup:** Dual-monitor setup or a high-resolution display for simultaneous visualization of genome alignment, mutation heatmaps, and pathway analysis results.

1. **:**

**IMPLEMENTATION OF THE PROJECT**

This chapter of the report outlines the practical steps taken to execute the project, detailing the methods, architecture, models used throughout the process. This section provides a clear description of how the theoretical concepts and designs were translated into a working solution. It includes an overview of the system architecture, the development environment, and application of the project. By following this structure, the section aims to give a comprehensive understanding of the approach taken to achieve the project objectives and the outcomes realized as outlined the figure 4.1.

A diagram of a data flow

AI-generated content may be incorrect.

Figure 4. OVERVIEW OF IMPLEMENTATION STEPS

* 1. **SYSTEM ARCHITECTURE**

The system architecture of this project is designed to analyze SARS-CoV-2 genome sequences, detect genetic variants, and map their effects on gene pathways using a combination of bioinformatics algorithms and computational tools. The framework integrates key methodologies such as sequence alignment, variant calling, annotation, and pathway analysis to derive meaningful biological insights.

The architecture consists of the following primary components. The first step in the framework involves processing and aligning SARS-CoV-2 genome sequences to detect genetic variations efficiently. This begins with organizing raw sequencing data and applying transformation algorithms to optimize storage and retrieval. The Burrows-Wheeler Transform (BWT) algorithm is employed to compress and restructure sequence data, making it easier to handle large-scale genomic datasets. Once the sequences are preprocessed, alignment tools map them against the SARS-CoV-2 reference genome, enabling the identification of mismatches or mutations. This alignment step is crucial for ensuring high accuracy in downstream analyses, as it allows for the precise detection of genetic variations that may contribute. The flow diagram of various modules is depicted in the figure 4.2 below

![A white background with black dots

AI-generated content may be incorrect.](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAEsAAAA4CAYAAAC1+AWFAAAAAXNSR0IArs4c6QAAAARnQU1BAACxjwv8YQUAAAAJcEhZcwAADsMAAA7DAcdvqGQAAACQSURBVHhe7dCxDQAhEMCw5/ff+eiRkEhvlymzZmY+nvxn4M6swKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArMCswKzArGADB/0EbNZfmP8AAAAASUVORK5CYII=)A diagram of a diagram

AI-generated content may be incorrect.

Figure 4.2 FLOW DIAGRAM OF THE ARCHITECTURE

By leveraging advanced computational techniques, this framework is designed to process vast genomic datasets while maintaining efficiency and reliability in mutation detection, clustering of the pathways. The system architecture of this project combines advanced bioinformatics algorithms, Bayesian statistical models, and machine learning techniques to identify and interpret the mutations efficiently.

* 1. **SEQUENCE ALIGNMENT AND COMPRESSION: BURROWS-WHEELER TRANSFORM ALGORITHM**

One of the critical steps in genomic analysis is sequence alignment, where raw sequencing reads are mapped to the SARS-CoV-2 reference genome. The Burrows-Wheeler Transform (BWT) algorithm plays a key role in this process by optimizing the storage and retrieval of genome sequences.

The BWT algorithm is a text compression technique that reorganizes sequence data into a more structured format, making it easier to search for mutations. It achieves this by transforming the input genome into a matrix of cyclic permutations, extracting the last column, and leveraging the Run-Length Encoding (RLE) and Huffman coding techniques to compress the data. The working of the Burrows Wheel Transform is described in the figure 4.3 below.

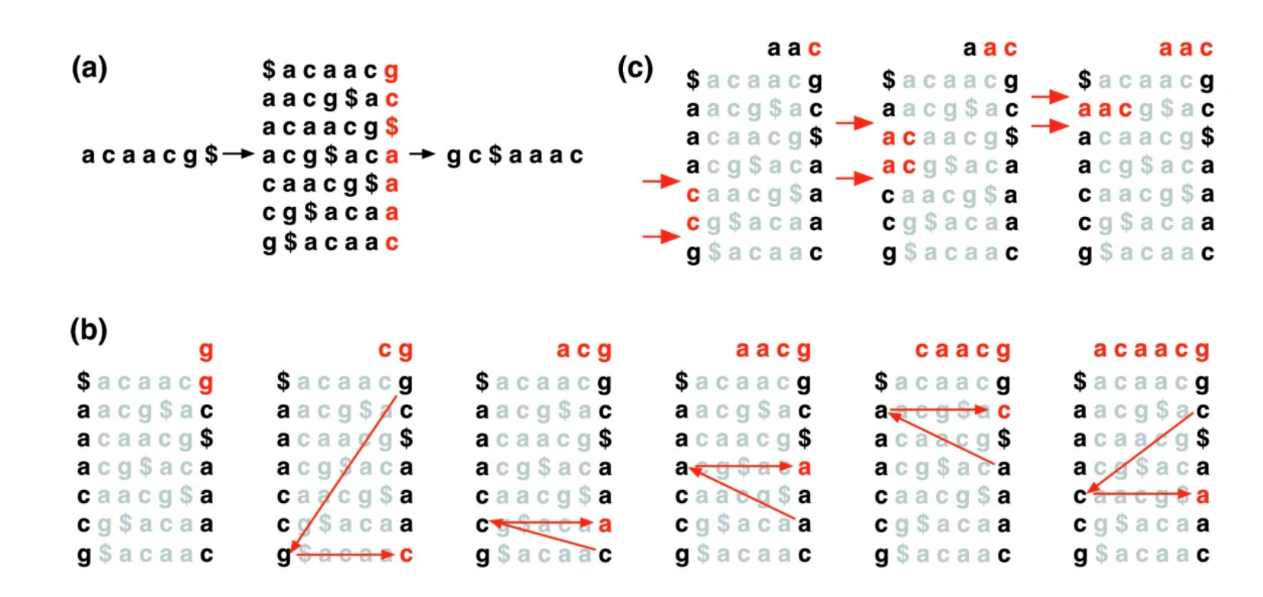


FIGURE 4.3 BURROWS WHEEL TRANSFORM OF DNA SEQUENCES [8]

By applying BWT-based alignment methods, such as those used in BWA (Burrows-Wheeler Aligner), the system efficiently maps short sequencing reads to the reference genome while reducing computational overhead. This enables faster variant detection and more accurate mutation tracking across different SARS-CoV-2 strains.

* 1. **VARIANT DETECTION AND ANNOTATION: FREEBAYES ALGORITHM**

Once the sequencing reads are aligned, the next step is variant calling, which involves detecting genetic mutations such as Single Nucleotide Polymorphisms (SNPs), insertions, deletions, and complex mutations. The FreeBayes algorithm, a Bayesian genetic variant detector, is utilized for this purpose.

FreeBayes employs a haplotype-based approach, meaning it does not just consider individual nucleotide changes but rather examines combinations of alleles (haplotypes) to identify mutations more accurately. This is particularly useful for detecting complex genetic events such as multi-nucleotide polymorphisms (MNPs) and composite insertion-substitution events. By using Bayesian statistical models, FreeBayes assigns probabilities to variant calls, ensuring high accuracy in mutation detection even when analyzing diverse sequencing datasets, including pooled or population-scale genomic samples. This is crucial for studying prolonged COVID-19 cases, where minor genetic variations could have significant functional implications. This step is depicted in the figure 4.4 below.

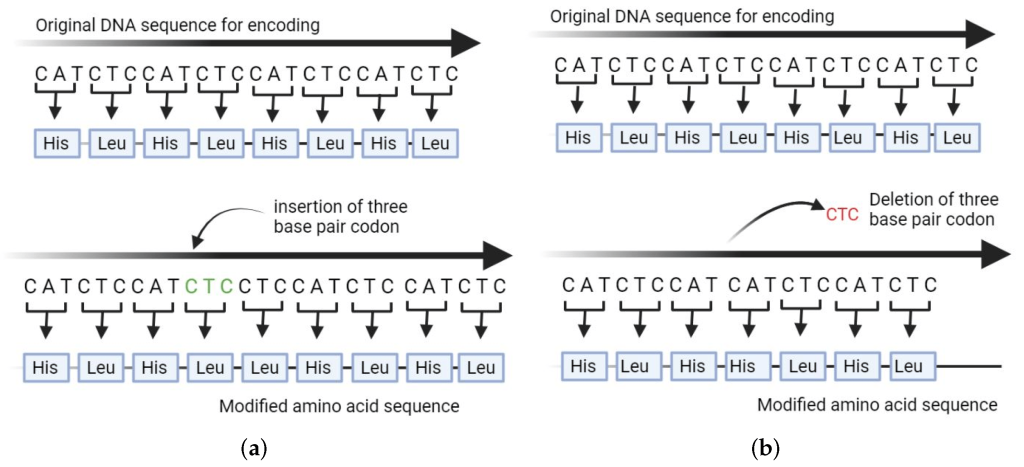


Figure 4.4 VARIANT IDENTIFICATION IN GENOMES [8]

* 1. **GENE PATHWAY MAPPING AND CLUSTERING**

After identifying genetic variants, the next step is to analyze their functional impact on biological pathways. Gene annotation is performed to link detected mutations to specific viral genes, providing insights into how these variations might alter viral function. By leveraging pathway mapping techniques, databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and Reactome are used to identify the biological processes influenced by these mutations. This step is crucial in understanding how genetic changes contribute to prolonged infection, immune system evasion, and altered viral replication mechanisms. Additionally, clustering techniques such as hierarchical clustering and principal component analysis (PCA) are applied to group mutations based on their shared effects on disease progression. These clustering methods help categorize high-impact mutations, revealing patterns that may be linked to long-term COVID-19 symptoms and potential therapeutic targets.

* 1. **INTEGRATION AND VISUALIZATION**

To ensure the findings are accessible and interpretable, the system incorporates advanced data visualization techniques. Interactive tools allow researchers to track variant evolution and frequency across different patient samples, making it easier to observe mutation trends over time. Functional pathway diagrams are generated to illustrate the molecular effects of mutations, helping scientists understand how SARS-CoV-2 mutations alter biological processes.

Computational approaches play a significant role in analyzing and interpreting SARS-CoV-2 mutations. Bioinformatics tools such as BLAST, BWA, FreeBayes, and SnpEff are used for sequence alignment, variant detection, and functional annotation. Machine learning algorithms such as Support Vector Machines (SVMs), Neural Networks, and Bayesian Networks help predict the potential pathogenicity of mutations and their impact on biological pathways. Statistical models are employed to determine mutation frequencies and their correlation with prolonged COVID-19 symptoms, enhancing the predictive power of genomic studies. The web-based interface for the project is depicted in the figure 4.5 below.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4. 5 USER INTERFACE OF THE PROJECT

Additionally, the system is designed to support real-time genomic analysis through a web-based platform, enabling researchers to conduct rapid and large-scale mutation assessments. This integration of computational modeling, pathway mapping, and interactive visualization enhances the ability to interpret genomic variations in prolonged COVID-19 cases, ultimately guiding future studies on viral evolution, treatment development, and personalized medicine approaches.

* 1. **APPLICATIONS OF GENE PATHWAY MAPPING AND CLUSTERING**

Gene pathway mapping and clustering have significant applications in epidemiological surveillance, vaccine development, and genomic-based therapeutics. These techniques help in tracking the emergence of new variants and their impact on disease progression, enabling proactive public health interventions.

In vaccine and drug development, gene pathway analysis identifies critical targets for antiviral strategies and helps evaluate the effectiveness of existing treatments. Personalized medicine applications include identifying genetic susceptibility to prolonged COVID-19 and tailoring treatment approaches based on individual genomic profiles. These insights contribute to improving early detection, risk assessment, and therapeutic planning for long-term SARS-CoV-2 infections.

Gene pathway mapping and clustering provide a data-driven framework for analyzing SARS-CoV-2 mutations and their impact on prolonged COVID-19. By integrating genomic sequencing, bioinformatics tools, and computational models, researchers can track viral evolution, identify high-risk variants, and develop targeted therapeutic strategies. These methodologies contribute to enhancing our understanding of prolonged COVID-19 and its potential long-term health implications, helping shape future research and medical interventions.

1. :

**EXPERIMENT ANALYSIS AND RESULTS OF GENE PATHWAY MAPPING**

This chapter delivers the description of the experimental results and analysis of Gene Pathway Mapping And Clustering Of Prolonged Covid-19 (Sars-Cov-2) Virus. The experiment analysis and results section presents the findings from SARS-CoV-2 genome sequencing, mutation detection, pathway mapping, and clustering. This analysis aims to identify genetic variants associated with prolonged covid and understand their biological impact on viral function and host interactions.

* 1. **INPUT TO THE PATHWAY ANALYSIS**

The dataset used in this project consists of SARS-CoV-2 reference genome, standard genome annotation files (GFF), and genome sequences of individuals affected by the virus. These datasets are obtained from publicly available repositories such as NCBI GenBank ensuring high-quality genomic data for analysis. The dataset obtained from the NCBI website as displayed in the figure 5.1 below

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5. 1 NCBI WEBSITE TO OBTAIN GENOME DATA

The SARS-CoV-2 reference genome serves as a baseline for identifying genetic variations in infected individuals. It is publicly available under the accession number NC\_045512.2 and was originally sequenced from the first known COVID-19 case in Wuhan, China. The genome is approximately 29,903 base pairs (bp) long and belongs to the Coronaviridae family as a positive-sense single-stranded RNA (ssRNA) virus. This reference genome is essential for sequence alignment, variant detection, and functional annotation of viral genes, providing a standard comparison for studying new SARS-CoV-2 variants and their evolution.

The General Feature Format (GFF) file for SARS-CoV-2 contains genome annotations that map genomic positions to their respective features, such as genes, coding sequences (CDS), regulatory regions, and protein domains. The GFF file plays a crucial role in identifying gene locations, linking mutations to their respective protein-coding regions, and performing pathway analysis to understand their impact on viral behavior. A typical SARS-CoV-2 GFF annotation includes key structural proteins such as Spike (S), Membrane (M), Envelope (E), and Nucleocapsid (N), along with non-structural proteins including ORF1ab, ORF3a, ORF6, ORF7a, ORF8, and ORF10. Additionally, regulatory regions such as the 5’ and 3’ untranslated regions (UTRs) are included in the GFF file. The SARS-CoV-2 GFF annotation file (NC\_045512.2.gff) is publicly available on platforms such as NCBI and Ensembl.

To analyze genetic variations in prolonged COVID-19 cases, whole-genome sequences from individuals infected with SARS-CoV-2 are obtained from NCBI Virus, GISAID, and EMBL-EBI databases. These datasets include full-length genome sequences of viral isolates from patients with prolonged COVID-19 symptoms, along with metadata such as patient demographics, sample collection dates, and geographic origins. By comparing patient-derived sequences with the reference genome, researchers can identify mutations, track viral evolution, and determine their functional consequences. The dataset also includes consensus sequences generated from deep sequencing data, which enable high-confidence variant calling and analysis of mutation trends.

* 1. **EXECUTION OF THE PIPELINE**

The execution of the pipeline follows a user-friendly and automated workflow, allowing researchers to analyze SARS-CoV-2 genome sequences from affected individuals efficiently. The pipeline is designed to handle large-scale genomic data and carry out sequence alignment, variant detection, gene annotation, and pathway mapping in a streamlined manner.

The process begins with the user uploading the genome sequence data of individuals affected by prolonged COVID-19. The input files, typically in FASTA or FASTQ format, can be uploaded via a web-based interface or a command-line interface (CLI) for batch processing. Once the files are uploaded, the pipeline automatically initiates the analysis in the backend as shown in figure 5.2.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5. 2 EXECUTION OF PIPELINE

Upon execution, the pipeline first performs quality control and preprocessing of the uploaded genome sequences to filter out low-quality reads. The structured execution of the process to analyze SARS-CoV-2 genome sequences, detect genetic variants, and map their biological impact.

The following steps describe each stage of the pipeline which is designed to ensure accurate sequence alignment, mutation detection, annotation, and gene pathway analysis as shown in the figure 5.3 below,

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 5. 3 STAGE I OF THE EXECUTION PIPELINE

The pipeline begins by indexing the reference genome (NC\_045512.2) to facilitate efficient sequence alignment. Indexing helps in organizing the reference genome into a searchable format, enabling rapid mapping of sequencing reads in subsequent steps.

The uploaded genome sequences of affected individuals are aligned to the reference genome using Burrows-Wheeler Aligner (BWA). This step ensures that sequencing reads are correctly mapped to their corresponding positions in the SARS-CoV-2 genome, allowing for the identification of mismatches and structural variations. These steps are depicted in figure 5.4 below.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 5. 4 STAGE II OF THE EXECUTION PIPELINE

**Conversion of Sequences into Binary Format**

The generated alignment file in SAM (Sequence Alignment Map) format is converted into BAM (Binary Alignment Map) format using SAMtools. BAM files are compressed versions of SAM files, reducing storage space and enhancing processing efficiency.

**Indexing and Sorting BAM Files**

To optimize downstream analysis, the BAM files are sorted and indexed. Sorting arranges the aligned sequences in genomic order, and indexing enables quick retrieval of specific genomic regions for mutation analysis.

**Variant Calling Using FreeBayes**

The variant calling step is performed using FreeBayes, a haplotype-based Bayesian variant detector. This process identifies single nucleotide polymorphisms (SNPs), insertions, deletions (InDels), multi-nucleotide polymorphisms (MNPs), and complex mutations in the uploaded genome sequences by comparing them against the reference genome.

**Variant Annotation**

The identified genetic variants are annotated using BCFTools, which classifies mutations based on their location and functional impact on viral proteins. This step helps in determining whether the mutations affect structural integrity, viral replication, or immune system evasion mechanisms.

**Extracting Chromosome Information**

From the annotated variant file (variants.vcf), the chromosome identifiers associated with the detected mutations are extracted. This information is critical for gene pathway analysis, allowing researchers to focus on specific genomic regions affected by prolonged covid cases.

**Consensus Sequence Generation (Variant Filtering)**

The pipeline generates a consensus sequence incorporating identified mutations, providing a refined representation of the viral genome for each sample. This is achieved by compressing the VCF file and applying Bcftools consensus algorithms to integrate the detected variations into the reference genome.

**BLAST Analysis for Mutation Impact Assessment**

The consensus sequence is compared against the reference genome using BLAST (Basic Local Alignment Search Tool) to assess mutation-induced changes. This step helps in understanding how mutations may affect viral protein structure, function, and antigenicity.

**Gene Pathway Analysis and Functional Impact Assessment**

The impact of identified mutations on biological pathways and viral functions is analyzed using BedTools and pathway mapping databases (KEGG, Reactome). By intersecting variant data with genomic annotations, the pipeline identifies which genes and regulatory elements are affected.

* 1. **PATHWAY ANALYSIS**

The final step in the pipeline focuses on analyzing the most frequently mutated genes in the SARS-CoV-2 genome within the Indian population. By applying computational filtering techniques, the pipeline identifies high-impact genetic variations, which provide insights into how these mutations influence prolonged covid symptoms.

The analysis result reveals the impact of variant genes observed as displayed in the figure 5.5 below,

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5. 5 GENES IMPACTED BY VARIANTS

The key mutated genes and their biological impact is underlined in the following section,

**ORF1ab (Open Reading Frame 1ab)**

The ORF1ab gene is the largest region in the SARS-CoV-2 genome and encodes non-structural proteins (NSPs) essential for viral replication and transcription. Mutations in ORF1ab may enhance viral replication efficiency, increasing viral load and prolonging infection. These changes can also alter the virus’s ability to evade host immune responses, potentially contributing to persistent viral shedding in prolonged COVID-19 cases.

**Spike (S) Protein**

The Spike (S) protein mediates viral entry into host cells by binding to the ACE2 receptor. Mutations in the Spike gene are commonly associated with increased infectivity, immune escape, and vaccine resistance. Variations such as D614G, N501Y, and E484K have been documented to enhance viral transmissibility and reduce antibody neutralization. Persistent viral presence in prolonged COVID-19 patients may be linked to mutations in the Spike protein, allowing reinfection or sustained immune response activation.

**ORF3a (Accessory Protein ORF3a)**

The ORF3a gene encodes a viral ion channel protein involved in modulating host immune responses and inflammatory pathways. Mutations in ORF3a have been associated with increased cytokine release, apoptosis (cell death), and immune dysregulation, leading to severe inflammation and tissue damage. Studies suggest that variations in this gene contribute to prolonged inflammatory responses in long COVID patients, worsening symptoms such as fatigue and respiratory distress.

The organs affected by these mutated genes that can exhibit syndromes are described below,

**Lungs**: Mutations in the Spike protein and ORF3a contribute to lung inflammation, fibrosis, and long-term respiratory issues, resulting in breathlessness and chronic cough.

**Heart**: ORF1ab and ORF3a mutations have been linked to cardiovascular complications, myocarditis, and irregular heart rhythms due to persistent inflammation and immune response dysregulation.

**Brain and Nervous System**: Mutations in ORF3a have been associated with neuroinflammatory responses, leading to brain fog, headaches, cognitive dysfunction, and fatigue in prolonged COVID-19 patients.

**Kidneys**: Viral persistence and immune dysregulation due to ORF1ab mutations may contribute to kidney injury and dysfunction, which has been observed in some long COVID cases.

**Liver**: ORF1ab-related mutations could impact liver metabolism and enzyme function, leading to abnormal liver function tests in patients with persistent viral infection.

**Immune System**: Spike and ORF3a mutations may dampen the immune response or cause prolonged hyperinflammation, increasing the risk of autoimmune-like symptoms and immune exhaustion.

1. :

**CONCLUSION AND FUTURE WORK**

This chapter presents the conclusion of the project carried out to analyze gene pathway mapping and clustering of prolonged covid and discusses future enhancements to improve the accuracy and efficiency of the analysis.

* 1. **CONCLUSION**

The study successfully identified frequent mutations in genes, which play a crucial role in viral replication, immune evasion, and prolonged infection symptoms. Through sequence alignment, variant detection, annotation, and pathway mapping, the project provided insights into how genetic variations influence disease progression and long-term effects in infected individuals.

The implemented bioinformatics pipeline efficiently processed genomic datasets, performed variant calling using FreeBayes, and mapped mutations to biological pathways. The results demonstrated the impact of genetic mutations on multiple organ systems, including the lungs, heart, brain, kidneys, and immune system, explaining the persistent symptoms observed in prolonged COVID-19 cases.

These mutations contribute to immune evasion, viral persistence, and prolonged inflammatory responses, leading to long-term complications in multiple organ systems. By identifying these mutations and their effects, this study enhances our understanding of prolonged COVID-19 and provides valuable insights for therapeutic development, personalized treatment strategies, and genomic surveillance.

* 1. **FUTURE WORK**

Future enhancements of this project could involve integrating machine learning models for improved mutation classification, expanding the dataset to include a more diverse population, and developing a real-time web-based interface for researchers to analyze SARS-CoV-2 mutations interactively.

Additionally, combining multi-omics data (genomics, proteomics, and transcriptomics) could provide a more comprehensive understanding of how prolonged COVID-19 develops at the molecular level. These advancements will enhance mutation tracking, therapeutic target identification, and public health interventions, contributing to better diagnostics and treatment strategies for long-term SARS-CoV-2 infections.

**REFERENCES**

1. Wolf, S., Zechmeister-Koss, I. & Erdös, J. Possible long COVID healthcare pathways: a scoping review. BMC Health Serv Res 22, 1076 (2022). <https://doi.org/10.1186/s12913-022-08384-6>
2. Davis, H.E., McCorkell, L., Vogel, J.M. et al. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol 21, 133–146 (2023). <https://doi.org/10.1038/s41579-022-00846-2>
3. Rahimi A, Mirzazadeh A, Tavakolpour S. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. Genomics. 2021 Jan;113(1 Pt 2):1221-1232. doi: 10.1016/j.ygeno.2020.09.059. Epub 2020 Sep 30. PMID: 33007398; PMCID: PMC7525243.
4. Niewolik, J., Mikuteit, M., Klawitter, S. et al. Cluster analysis of long COVID symptoms for deciphering a syndrome and its long-term consequence. Immunol Res 72, 605–613 (2024). <https://doi.org/10.1007/s12026-024-09465-w>
5. Smith T, Rohaim MA, Munir M. Mapping molecular gene signatures mediated by SARS-COV-2 and large-scale and genome-wide transcriptomics comparative analysis among respiratory viruses of medical importance. Mol Cell Probes. 2022 Aug;64:101820. doi: 10.1016/j.mcp.2022.101820. Epub 2022 Apr 30. PMID: 35504488; PMCID: PMC9054707.
6. Akbarialiabad H, Taghrir MH, Abdollahi A, Ghahramani N, Kumar M, Paydar S, Razani B, Mwangi J, Asadi-Pooya AA, Malekmakan L, Bastani B. Long COVID, a comprehensive systematic scoping review. Infection. 2021 Dec;49(6):1163-1186. doi: 10.1007/s15010-021-01666-x. Epub 2021 Jul 28. PMID: 34319569; PMCID: PMC8317481.
7. Nunes DR, Braconi CT, Ludwig-Begall LF, Arns CW, Durães-Carvalho R. Deep phylogenetic-based clustering analysis uncovers new and shared mutations in SARS-CoV-2 variants as a result of directional and convergent evolution. PLoS One. 2022 May 24;17(5):e0268389. doi: 10.1371/journal.pone.0268389. PMID: 35609034; PMCID: PMC9129020.
8. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2024 March 22]. Available from: https://www.ncbi.nlm.nih.gov/
9. Taheri, Golnaz & Habibi, Mahnaz. (2022). Comprehensive analysis of pathways in Coronavirus 2019 (COVID-19) using an unsupervised machine learning method. Applied Soft Computing. 128. 109510. 10.1016/j.asoc.2022.109510.
10. Greenhalgh T, Sivan M, Perlowski A, Nikolich JŽ. Long COVID: a clinical update. Lancet. 2024 Aug 17;404(10453):707-724. doi: 10.1016/S0140-6736(24)01136-X. Epub 2024 Jul 31. PMID: 39096925.
11. Umesh A, Pranay K, Pandey RC, Gupta MK. Evidence mapping and review of long-COVID and its underlying pathophysiological mechanism. Infection. 2022 Oct;50(5):1053-1066. doi: 10.1007/s15010-022-01835-6. Epub 2022 Apr 30. PMID: 35489015; PMCID: PMC9055372.
12. Avadhanula V, Creighton CJ, Ferlic-Stark L, Sucgang R, Zhang Y, Nagaraj D, Nicholson EG, Rajan A, Menon VK, Doddapaneni H, Muzny DM, Metcalf G, Cregeen SJJ, Hoffman KL, Gibbs RA, Petrosino J, Piedra PA. Longitudinal host transcriptional responses to SARS-CoV-2 infection in adults with extremely high viral load. bioRxiv [Preprint]. 2023 May 25:2023.05.24.542181. doi: 10.1101/2023.05.24.542181. Update in: PLoS One. 2025 Jan 16;20(1):e0317033. doi: 10.1371/journal.pone.0317033. PMID: 37292999; PMCID: PMC10245966.