

ORIGINAL RESEARCH ARTICLE

Identification of conserved SARS-CoV-2 gene regions for diagnostic tool development through whole-genome sequencing

Getnet Hailu^{1,2} , Fisihatsion Tadesse^{3*} , Molalegne Bitew⁴ , Kidist Zealiyas² , Girmay Medhin¹ , Andargachew Gashu² , and Mengistu Legesse¹ 

¹Department of Tropical and Infectious Diseases, Akililu Lemma Institute of Health Research, Addis Ababa University, Addis Ababa, Ethiopia

²Department of Infectious Diseases, Ethiopian Public Health Institute, Addis Ababa, Ethiopia

³Department of Internal Medicine, College of Health Science, Addis Ababa University, Addis Ababa, Ethiopia

⁴Department of Health Biotechnology, Bio and Emerging Technology Institute, Addis Ababa, Ethiopia

Abstract

The high mutation frequency of the SARS-CoV-2 spike (S) gene undermines the reliability of S-based diagnostics. Identifying conserved protein epitopes that remain stable across SARS-CoV-2 variants is paramount in resource-limited settings. This study aims to identify conserved gene regions of SARS-CoV-2 using whole-genome sequencing and immunoinformatics to support the development of monoclonal antibody-based rapid diagnostics. SARS-CoV-2 positive samples ($C_t \leq 30$) collected in Ethiopia (June–August 2022) from a retrospective study ($n = 70$) were sequenced using Illumina NextSeq-550. Conserved regions were identified via Multiple Alignment using Fast Fourier Transform, entropy plots, and mutation profiling. B-cell epitopes were predicted and assessed for antigenicity and cross-reactivity. Structural modeling was conducted using AlphaFold and visualization tools. High-quality whole-genome sequencing of 63 SARS-CoV-2 samples revealed conserved regions in the nucleocapsid gene across major variants. Three surface-exposed, antigenic, and non-allergenic B-cell epitopes were identified with low cross-reactivity to common human coronaviruses. Structural modeling and docking confirmed their accessibility and strong antibody binding, supporting their potential for robust diagnostic applications. In conclusion, conserved N protein epitopes identified in this study are promising targets for developing robust, variant-resistant rapid diagnostic tools.

***Corresponding author:**
Fisihatsion Tadesse
(fishatsion.tadesse@aau.edu.et)

Citation: Hailu G, Tadesse F, Bitew M, *et al.* Identification of conserved SARS-CoV-2 gene regions for diagnostic tool development through whole-genome sequencing. *Microbes & Immunity*. 2026;3(4):025450120.
doi: 10.36922/MI025450120

Received: November 5, 2025

Revised: April 26, 2026

Accepted: May 14, 2026

Published online: June 5, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: SARS-CoV-2; Conserved gene region; Diagnostic tool; Monoclonal antibody-based approach; Ethiopia

1. Introduction

The global emergence and persistence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have highlighted the critical need for reliable, rapid, and widely accessible diagnostic tools. While reverse transcription polymerase chain reaction (RT-PCR) is the current gold standard for SARS-CoV-2 detection, its reliance on specialized equipment

and skilled personnel limits its deployment in low-resource settings and impedes timely diagnosis during outbreaks.^{1,2} Consequently, antigen-based rapid diagnostic tests (Ag-RDTs) have gained prominence as complementary tools due to their ease of use, cost-effectiveness, and potential for point-of-care deployment.³ Ag-RDTs primarily target the viral spike (S) and nucleocapsid (N) proteins. The S protein, which mediates viral entry via the angiotensin-converting enzyme 2 receptor, is highly immunogenic but exhibits considerable genetic variability, particularly in regions subject to immune-driven selection.^{4,5} This variability, observed notably in variants of concern (VOCs) such as Alpha, Beta, Delta, and Omicron, can reduce the sensitivity and specificity of S-targeted diagnostic assays and may also affect vaccine-induced or infection-induced immune recognition.^{6,7}

In contrast, the N protein is highly conserved, abundantly expressed, and elicits a strong early antibody response, making it a promising target for diagnostics.^{1,2} Its structural and functional roles in viral RNA packaging and replication make it less prone to frequent mutations, providing greater stability across circulating variants.⁸ Due to this high degree of conservation, N protein-based Ag-RDTs may maintain diagnostic performance even as new variants emerge.

Recent advances in immunoinformatics have enabled efficient *in silico* screening of viral proteins to identify conserved, antigenic, and structurally accessible epitopes that are suitable for monoclonal antibody (mAb) development.⁹ Such epitope-based approaches allow for the rational design of mAb-targeting strategies that can improve the sensitivity and specificity of diagnostic tools. Computational tools such as BepiPred, Immune Epitope Database (IEDB), and structural modeling software (e.g., University of California, San Francisco [UCSF] Chimera, AlphaFold, PyMOL) have facilitated the prediction of linear B-cell epitopes, assessment of antigenicity scores, and analysis of epitope surface accessibility, streamlining the development of diagnostics that are both robust and broad in variant coverage.¹⁰

Although many commercially available AgRDTs are designed to detect the SARSCoV2 N protein, the precise epitopes incorporated into these assays are proprietary, and their diagnostic performance has shown variability as new variants have emerged.¹¹ AgRDTs remain valuable tools because they are inexpensive, provide results quickly, and can be deployed at decentralized testing sites. Nonetheless, their sensitivity may fluctuate depending on factors such as viral load, specimen quality, and epitope changes associated with evolving variants. With Omicron sublineages and their descendants, including XBB-related and BA.2-derived

variants, continuing to circulate globally, there is an urgent need for diagnostic targets that remain robust despite ongoing viral evolution. Systematically identifying N protein epitopes that are evolutionarily conserved, immunogenic, and structurally accessible is therefore critical for informing the development of next-generation AgRDTs that maintain performance across variants. This need is especially pronounced in resource-limited settings, where AgRDTs remain the primary point-of-care testing modality.

2. Materials and methods

2.1. Study design, period, and sample collection

We conducted a retrospective, laboratory-based experimental study. The samples used in this analysis were originally collected during Ethiopia's fifth COVID-19 wave, which occurred between June and August 2022. From the archived repository of the National Virology Reference Laboratory at the Ethiopian Public Health Institute (EPHI), we retrieved 150 confirmed SARSCoV2-positive specimens. All samples were nasopharyngeal swabs obtained from patients living in Addis Ababa, Oromia, Amhara, and the Southern Nations, Nationalities, and Peoples regions (Figure 1).

The EPHI serves as Ethiopia's national reference laboratory for SARSCoV2 testing, and samples included in this study were originally submitted from multiple regions, as well as from diplomatic missions and international travelers. After diagnostic testing, specimens were stored at -80°C for long-term preservation. For sequencing, samples were purposively chosen using their unique identifiers (IDs) from the national electronic database. Only specimens that were real-time quantitative polymerase chain reaction (RTqPCR)-positive with a cycle threshold (Ct) value of 30 or below were eligible. Based on these criteria and available sequencing capacity, 150 SARSCoV2-positive samples were selected for analysis. Deidentified patient metadata, including demographic characteristics, clinical information, and RTqPCR results, were extracted from electronic records. Participants were subsequently categorized by age, sex, clinical presentation, underlying health conditions, travel history, and vaccination status.

2.2. Quality assurance

To improve the likelihood of obtaining high-quality whole-genome sequences, all samples were retested using RTqPCR after thawing at room temperature. RNA was re-extracted using the BGI extraction kit (1000020261, Beijing Genomics Institute, China), and amplification was performed on the real-time polymerase chain reaction detection system (CFX96 Deep Well™, BioRad Laboratories,

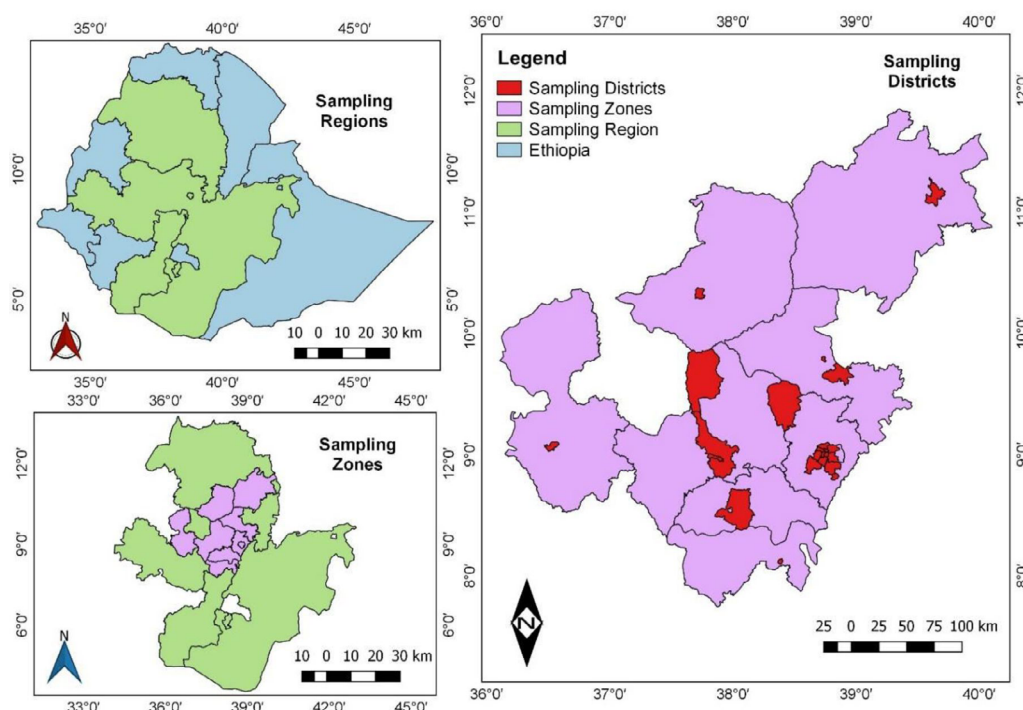


Figure 1. The geographical map of Ethiopia showed the regions and districts where SARS-CoV-2 specimens were collected

United States) following standard operating procedures. All laboratory work was conducted in a biosafety level2 facility at the National Virology Reference Laboratory. Of the 150 samples reanalyzed, 70 had Ct values of 30 or lower and were selected for wholegenome sequencing; samples with higher Ct values were excluded due to insufficient viral RNA. Sequencing of the 70 eligible RNA extracts was performed on the next-generation sequencing platform (Illumina NextSeq 550, Illumina, Inc., USA) at EPHI's virology laboratory in March 2023. After bioinformatics processing, 63 genomes met the required quality criteria, defined as >80% genome coverage, absence of clustered mutations, and no unexpected stop codons, and were retained for downstream analyses.

2.3. Comparative genome analysis

Raw sequencing data were obtained in FASTQ format and processed for variant calling and consensus sequence generation. First, the FASTQ sequences' quality was evaluated using the FastQC program (v0.11.9, Babraham Bioinformatics, UK), and then adaptor sequences were trimmed from reads using Cutadapt software (v4.6, TU Dortmund University, Germany)¹² to remove adaptor sequences and poor-quality reads. In addition, a dataset of 118 globally representative SARS-CoV-2 sequences from 62

countries (including the major variants) was downloaded from the Global Initiative on Sharing All Influenza Data (GISAID). Multiple sequence alignment (MSA) of local and global genomes was performed with the Multiple Alignment using Fast Fourier Transform (MAFFT; v7, Kyoto University, Japan). Mutation analysis was performed using Nextclade, CoV-Gene Linked by Underlying Evolution (v3.18.0, University of Edinburgh, UK), and SnpEff (v5.2, Leloir Institute Foundation, Argentina) to identify variant-specific and conserved regions across genomes. All viral genome sequences obtained from this study were submitted to the GISAID with accession ID from EPI_ISL_19147618 to EPI_ISL_19302657¹³ and the National Center for Biotechnology Information (NCBI) repositories with accession numbers from PQ140559 to PP069552.¹⁴

To identify conserved gene regions, MSA was performed between the generated sequences and a diverse dataset of global SARS-CoV-2 sequences downloaded from the GISAID database, representing major VOCs including Alpha, Beta, Delta, and Omicron. Alignment was performed using MAFFT, and sequence conservation was assessed with Jalview (v2.11.3.0, University of Dundee, UK) and entropy plots; regions with low entropy and high

identity were considered highly conserved. Additionally, mutations and variant-specific changes were annotated using Nextclade.

Following the identification of conserved regions, B-cell linear epitope prediction was conducted using BepiPred (v2.0, Technical University of Denmark, Denmark) hosted by the IEDB Analysis Resource. Epitopes were screened for antigenicity using VaxiJen (v2.0, Medical University of Sofia, Bulgaria) with a threshold of 0.5; those scoring above this were further analyzed for allergenicity and toxicity using AllerTOP (v2.0, Medical University of Sofia, Bulgaria) and ToxinPred (v2.0, Council of Scientific and Industrial Research Institute of Microbial Technology, India), respectively. Cross-reactivity with common human coronaviruses (e.g., HCoV-229E, HCoV-OC43, and HCoV-NL63) was assessed via Basic Local Alignment Search Tool (BLASTp) to ensure specificity.

For structural validation, three-dimensional (3D) models of the SARS-CoV-2 N protein were obtained using AlphaFold (v2.3.0, DeepMind Technologies, UK) predictions. The predicted epitopes were mapped onto the 3D structure using PyMOL (v2.3, Schrödinger, LLC, USA) and UCSF Chimera (v1.17, University of California, San Francisco, USA) to evaluate surface accessibility and structural integrity. Solvent accessibility and epitope flexibility were examined using DynaMut. Molecular docking simulations were then performed using ClusPro (v2.0, Boston University, USA) to assess potential binding

interactions between mAb fragments (Fabs) and the predicted epitopes.

3. Results

3.1. Whole-genome sequencing and genome quality

Whole-genome sequencing was successfully performed on SARS-CoV-2-positive clinical specimens from 63 samples collected between June and August 2022. The sequencing yielded high-quality reads with an average depth of $>99\times$ and a genome coverage of over 99%. Quality assessment of raw reads showed minimal adapter contamination and low base-calling error rates. The final assembled genome sequences showed a high degree of similarity to the reference SARS-CoV-2 genome (NC_045512.2) with a consensus sequence identity of $>99.8\%$. The assembled genomes were deposited in GISAID under accession IDs from EPI_ISL_19147618 to EPI_ISL_1930265713, and in NCBI repositories under accession numbers from PQ140559 to PP069552.¹⁴

3.2. Identification of conserved gene regions of SARS-CoV-2

Mutation profiling using Nextclade revealed widespread aminoacid substitutions in the S gene, including D614G, N501Y, P681R, and E484K. To visualize the structural distribution of these highfrequency mutations, the corresponding residues were mapped onto the 3D structure of the S protein (Figure 2), illustrating that many variable

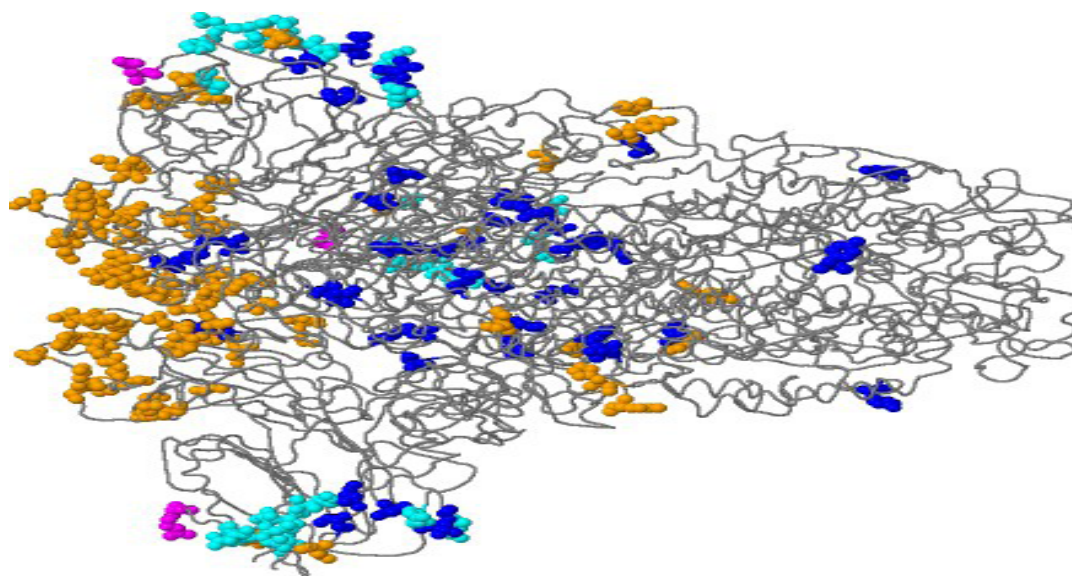


Figure 2. Structure of the SARSCoV2 spike glycoprotein (Protein Data Bank: 6ACC) with the receptorbinding domain in the down conformation. Residues highlighted in color represent highfrequency mutations identified in our dataset and in major variants of concern (e.g., D614G, N501Y, E484K, P681R). Conserved residues are shown in grey. This mapping illustrates the extensive distribution of S-protein mutations, supporting its reduced stability as a diagnostic target. Sources: <https://www.epicov.org/epi3/frontend#1a1eca>

positions occur in surface-exposed and functionally important regions.

To evaluate how strongly each position was conserved, we calculated Shannon entropy from the multiple sequence alignment. This metric reflects the amount of variation at a given amino-acid site: values near zero indicate that a position is highly conserved, whereas higher values indicate greater diversity. Across both the RNA-binding and dimerization regions, the *N* gene showed consistently low entropy (0.0–0.2), demonstrating limited variability. In contrast, the *S* gene displayed substantially higher entropy values (1.1–1.2), aligning with its well-documented pattern of frequent mutations in circulating variants. Together, these results highlight the relative evolutionary stability of the *N* gene compared with other structural genes.

3.3. B-cell epitope prediction and antigenicity screening

Subsequent B-cell epitope prediction using BepiPred identified 17 potential linear epitopes from the conserved regions of the *N* protein. Eight of those exhibited antigenicity scores above 0.5 with VaxiJen, indicating strong potential as diagnostic targets. Among them, three epitopes, RRGPEQTQGNFG (residues 152–163), SSSRGTS (66–72), and DPNFKDQVILL (194–204), were prioritized for downstream analysis due to their high antigenicity scores of 0.81, 0.77, and 0.69, respectively, and their location within or near conserved functional domains (Figure 3).

3.4. Allergenicity, toxicity, and cross-reactivity filtering

All eight antigenic epitopes were evaluated for safety, and five were predicted to be non-allergenic using AllerTOP and non-toxic using ToxinPred. To assess the risk of cross-reactivity with endemic human coronaviruses, BLASTp searches were conducted against HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1. Three epitopes (Section 3.3) showed low sequence identity, i.e., <30%, and E-values of >0.01, which indicates a low risk of cross-reactivity and confirms their diagnostic specificity (Table 1).

3.5. Structural mapping and solvent accessibility

Structural mapping of the three prioritized epitopes onto the AlphaFold N-protein model showed that all candidates were at least partially solvent-exposed. RRGPEQTQGNFG (152–163) and DPNFKDQVILL (194–204) exhibited high SASA values (>70 Å²), while SSSRGTS (66–72) showed moderate exposure. Two epitopes (RRGPEQTQGNFG and SSSRGTS) were located within flexible coil regions, which are typically more accessible to antibody binding. These structural features support their suitability as diagnostic epitopes (Figures 4 and 5).

3.6. Molecular docking simulations

Finally, molecular docking simulations were conducted using ClusPro to evaluate the binding interaction between the predicted epitopes and Fabs. The docking results

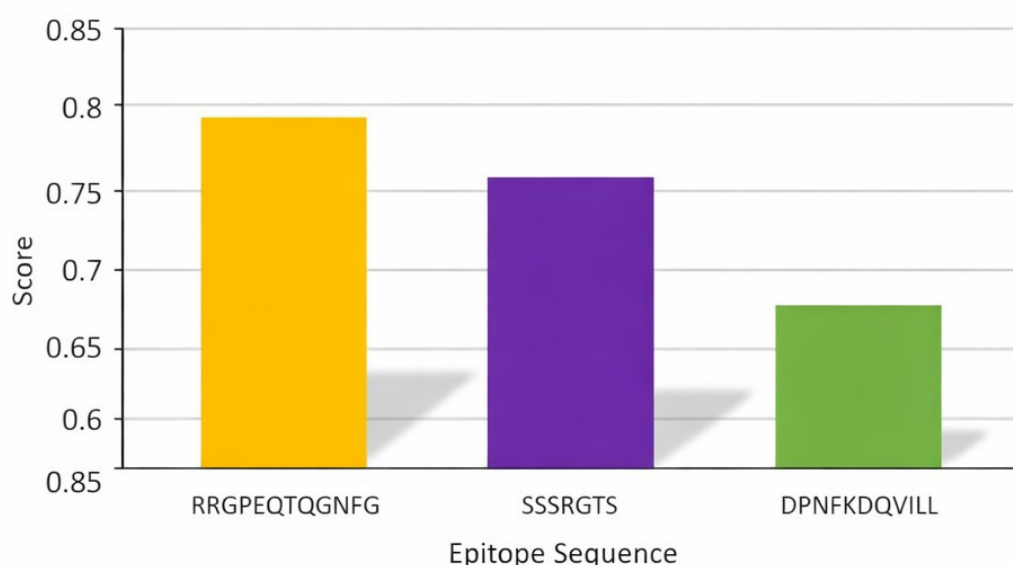


Figure 3. Antigenicity (VaxiJen score) of the sequenced epitopes

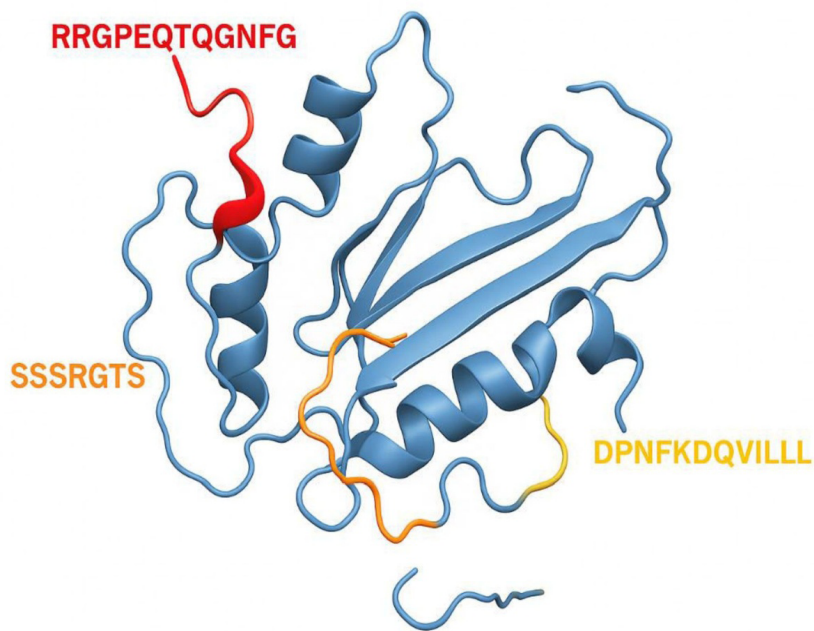


Figure 4. Three-dimensional visualization of SARS-CoV-2 nucleocapsid

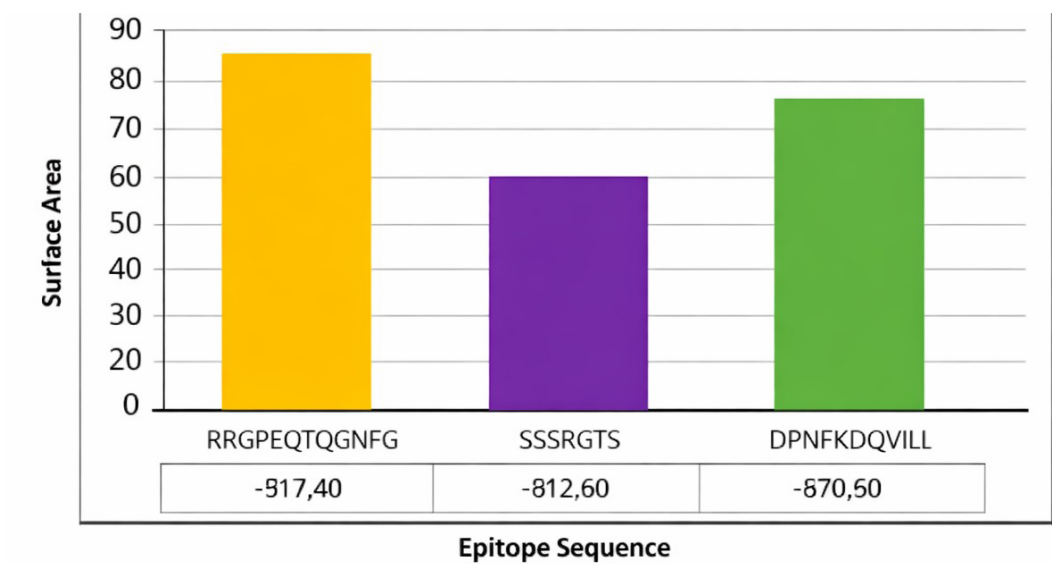


Figure 5. Solvent accessibility of sequenced epitopes

indicated strong affinities, with ClusPro weighted score of -917.4 for RRGPEQTQGNFG, -812.6 for SSSRGTS, and -870.9 for DPNFKDQVILL. The RRGPEQTQGNFG epitope formed stable interactions via hydrogen bonds with key residues (R154, K157, D161), supporting its binding potential and utility in diagnostic design (Figure 6).

4. Discussion

In this study, we explored conserved regions within the SARS-CoV-2 genome as potential targets for mAb-based diagnostic tools. Through whole-genome sequencing of locally circulating strains, we found that while the S gene accumulated numerous mutations, the N gene remained remarkably conserved. This observation supports a

growing body of evidence indicating that the N protein, particularly its RNA-binding and dimerization domains, is more genetically stable than the S protein and thus a better target for laboratory diagnosis.¹⁵ Previous studies have consistently highlighted the instability of the S gene due to immune selection pressure, especially following mass vaccination and widespread natural infections.¹⁶ Mutations such as D614G, N501Y, and E484K, commonly observed in our results, have been widely reported as key drivers of immune evasion and increased transmissibility.¹⁷ In contrast, the N gene's lower mutation frequency aligns with earlier genomic surveillance reports from various areas that identified it as a functionally constrained region with limited antigenic drift.^{18,19} These features make the N gene an ideal target for diagnostic development.

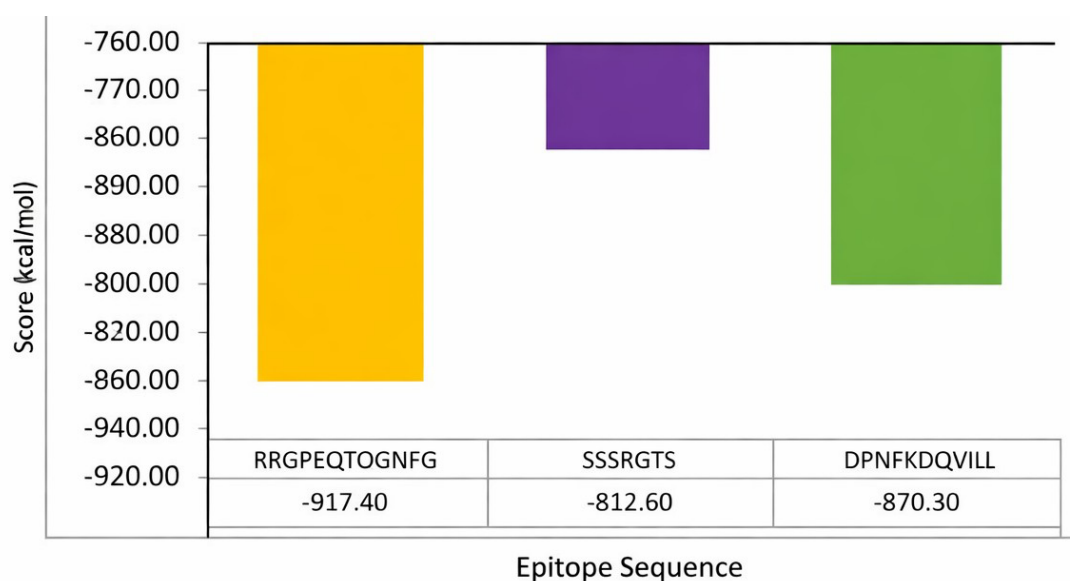


Figure 6. Molecular docking ClusPro score of each epitope. Lower scores indicate stronger affinities.

Table 1. Overall measurements/scores of epitope sequences

Epitope sequence	Position (N protein)	VaxiJen score	Allergenicity	Toxicity	Cross reactivity	Solvent accessibility	ClusPro score
RRGPEQTQGNFG	152–163	0.81	Non-allergen	Non-toxic	None	High	–917.4
SSSRGTS	66–72	0.77	Non-allergen	Non-toxic	Low	Moderate	–812.6
DPNFKDQVILL	194–204	0.69	Non-allergen	Non-toxic	None	High	–870.9

Note: These results demonstrate that computational epitope prediction, combined with antigenicity screening, structural validation, and docking studies, can effectively identify conserved, non-allergenic, and surface-accessible B-cell epitopes that may serve as candidates for affordable diagnostic or immunological tools targeting SARS-CoV-2.

The conserved analysis, supported by sequence alignment and entropy profiling, reinforced the N gene's stability across multiple VOCs, including Alpha, Beta, Delta, and Omicron. This finding corroborates previous studies that demonstrated that the N protein retains structural integrity and sequence stability across a wide array of SARS-CoV-2 genomes, even amid global viral evolution.^{19,20}

Epitope prediction within conserved regions of the N protein further validated their diagnostic potential. Predicted B-cell epitopes showed strong antigenicity, surface accessibility, and structural flexibility, key features for effective recognition by host antibodies. These results are consistent with computational and experimental studies from Asia and Europe, which identified similar N-terminal and central N protein regions as immunodominant and suitable for serodiagnostic use.^{21,22} Notably, these studies also emphasized that linear epitopes located in flexible coil regions are more likely to be immunogenic, which aligns with our structural mapping results.

Furthermore, the predicted epitopes demonstrated minimal sequence similarity to common human coronaviruses, suggesting a low risk of cross-reactivity, a major concern in the development of SARS-CoV-2-specific diagnostic tools. This is in agreement with findings from previous studies that underscore the need to carefully filter epitopes for cross-reactivity with endemic coronaviruses, particularly in serological assays.^{8,23}

Molecular docking simulations showed that the predicted epitopes bind stably to Fabs, providing preliminary evidence of their functional relevance. Comparable docking studies have demonstrated similar structural compatibility between predicted N protein epitopes and human immunoglobulins, further validating *in silico* approaches as a valuable screening tool for diagnostic development.^{24,25}

The implications of these findings are significant for countries like Ethiopia, where access to cost-effective, variant-resistant diagnostics remains limited. By targeting stable regions of the SARS-CoV-2 genome and validating their immunological relevance, this study lays the groundwork for the design of lateral flow assays and enzyme-linked immunosorbent assays tailored to the local epidemiological context. Importantly, our approach can also serve as a template for monitoring emerging pathogens beyond COVID-19, especially in settings with limited laboratory infrastructures.

Nevertheless, the study's computational nature is a limitation. While the results are promising, they require experimental validation through peptide synthesis, antibody-binding assays, and clinical performance

evaluation. Such validation would be crucial to confirm the predicted epitopes' diagnostic accuracy and to rule out false positives arising from structural or conformational effects not captured *in silico*.

5. Conclusion

The study's findings support the use of the N protein as a stable, immunogenic target for SARS-CoV-2 diagnostics and highlight specific epitope candidates for further development. This integrative approach, combining genomic surveillance with structural immunoinformatics, can accelerate the identification of reliable diagnostic markers amid evolving viral landscapes.

Acknowledgments

The authors would like to thank the Ethiopian Public Health Institute and Addis Ababa University for their laboratory and logistical support. The team extends gratitude to the laboratory staff and collaborators who contributed to sample processing, sequencing, and data analysis.

Funding

None.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Author contributions

Conceptualization: Getnet Hailu, Fisihatsion Tadesse, Mengistu Legesse

Data curation: Getnet Hailu, Fisihatsion Tadesse, Mengistu Legesse

Formal analysis: Getnet Hailu, Fisihatsion Tadesse, Mengistu Legesse

Investigation: Getnet Hailu, Molalegne Bitew, Kidist Zealiyas, Girmay Medhin, Andargachew Gashu, Mengistu Legesse

Methodology: Getnet Hailu, Molalegne Bitew, Kidist Zealiyas, Girmay Medhin, Andargachew Gashu, Mengistu Legesse

Writing—original draft: Getnet Hailu, Mengistu Legesse

Writing—review & editing: All authors

All authors have read and approved the final manuscript.

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board (IRB) of the Aklilu Lemma Institute of Pathobiology, Addis Ababa University (Protocol No. ALIPB IRB/52/2013/21). Permission to use

archived specimens was granted by the Ethiopian Public Health Institute.

Consent for publication

Not applicable as archived specimens were used in this study.

Availability of data

All data supporting the findings of this study are available from the corresponding author upon reasonable request. The generated genome sequences have been submitted to GISAID (Accession IDs: EPI_ISL_19147618–EPI_ISL_19302657) and NCBI GenBank (Accession Nos: PQ140559–PP069552).

References

1. Zhou R, Li F, Chen F, *et al.* Viral dynamics in asymptomatic patients with COVID-19. *Int J Infect Dis.* 2020;96:288–290. doi: 10.1016/j.ijid.2020.05.030
2. Jiang HW, Zhang HN, Meng QF, *et al.* SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. *Cell Mol Immunol.* 2020;17(9):998–1000. doi: 10.1038/s41423-020-0514-8
3. Chen Z, Zhang Z, Zhai X, *et al.* Rapid and sensitive detection of anti-SARS-CoV-2 IgG using lanthanide-doped nanoparticles-based lateral flow immunoassay. *Anal Chem.* 2020;92(10):7226–7231. doi: 10.1021/acs.analchem.0c00784
4. Planas D, Veyer D, Baidaliuk A, *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature.* 2021;596(7871):276–280. doi: 10.1038/s41586-021-03777-9
5. Zahradník J, Marciano S, Shemesh M, *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat Microbiol.* 2021;6(9):1188–1198. doi: 10.1038/s41564-021-00954-4
6. Tegally H, Wilkinson E, Giovanetti M, *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature.* 2021;592(7854):438–443. doi: 10.1038/s41586-021-03402-9
7. Xie C, Jiang L, Huang G, *et al.* Comparison of different samples for 2019 novel coronavirus detection by nucleic acid amplification tests. *Int J Infect Dis.* 2020;93:264–267. doi: 10.1016/j.ijid.2020.02.050
8. Naz A, Shahid F, Butt TT, Awan FM, Ali A, Malik A. Designing multi-epitope vaccines to combat emerging coronavirus disease 2019 (COVID-19) by employing immuno-informatics approach. *Front Immunol.* 2020;11:1663. doi: 10.3389/fimmu.2020.01663
9. Chen J, Wang R, Wang M, Wei GW. Mutations strengthened SARS-CoV-2 infectivity. *J Mol Biol.* 2020;432(19):5212–5226. doi: 10.1016/j.jmb.2020.07.009
10. Cosar B, Karagulleoglu ZY, Unal S, *et al.* SARS-CoV-2 mutations and their viral variants. *Cytokine Growth Factor Rev.* 2022;63:10–22. doi: 10.1016/j.cytogfr.2021.06.001
11. Sisay A, Abera A, Dufera B, *et al.* Diagnostic accuracy of three commercially available one-step RT-PCR assays for the detection of SARS-CoV-2 in resource-limited settings. *PLoS ONE.* 2022;17(1):e0262178. doi: 10.1371/journal.pone.0262178
12. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 2011;17(1):10–12. doi: 10.14806/ej.17.1.200
13. National Center for Biotechnology Information nucleotide database, 2024, PQ140559 to PP069552. Available from: <https://www.ncbi.nlm.nih.gov/nucleotide/> [Last accessed on August 12, 2024].
14. GISAID EpiCoV database, 2024; EPI_ISL_19147618 to EPI_ISL_19302657. Available from: <https://gisaid.org/> [Last accessed on August 12, 2024].
15. Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH. The SARS coronavirus nucleocapsid protein: forms and functions. *Antiviral Res.* 2014;103:39–50. doi: 10.1016/j.antiviral.2013.12.009
16. Harvey WT, Carabelli AM, Jackson B, *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol.* 2021;19(7):409–424. doi: 10.1038/s41579-021-00573-0
17. Korber B, Fischer WM, Gnanakaran S, *et al.* Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell.* 2020;182(4):812–827.e19. doi: 10.1016/j.cell.2020.06.043
18. Wu F, Zhao S, Yu B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature.* 2020;579(7798):265–269. doi: 10.1038/s41586-020-2008-3
19. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol.* 2020;11:1800. doi: 10.3389/fmicb.2020.01800
20. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-

- 19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses*. 2020;12(3):254.
doi: 10.3390/v12030254
21. Grifoni A, Weiskopf D, Ramirez SI, *et al.* Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*. 2020;181(7):1489-1501.e15.
doi: 10.1016/j.cell.2020.05.015
22. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*. 2017;45(W1):W24-W29.
doi: 10.1093/nar/gkx346
23. Sette A, Crotty S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell*. 2021;184(4):861-880.
doi: 10.1016/j.cell.2021.01.007
24. Johnson BA, Zhou Y, Lokugamage KG, *et al.* Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *PLoS Pathog*. 2022;18(10):e1010627.
doi: 10.1371/journal.ppat.1010627
25. Zhang L, Jackson CB, Mou H, *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun*. 2020;11:6013.
doi: 10.1038/s41467-020-19808-4