# JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.
Volume 2 (2021), Issue 1 ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

## Contents

### Viewpoint

Nonfungible Tokens as a Blockchain Solution to Ethical Challenges for the Secondary Use of Biospecimens: Viewpoint (e29905)
Marielle Gross, Amelia Hood, Robert Miller Jr. . . . 2

### Original Paper

Isolating SARS-CoV-2 Strains From Countries in the Same Meridian: Genome Evolutionary Analysis (e25995)
Emilio Mastriani, Alexey Rakov, Shu-Lin Liu. . . . 8

Viewpoint

# Nonfungible Tokens as a Blockchain Solution to Ethical Challenges for the Secondary Use of Biospecimens: Viewpoint

Marielle S Gross[1,2*], MD; Amelia J Hood[3*], MA; Robert C Miller Jr[4], BA

[1]Department of Obstetrics, Gynecology and Reproductive Services, University of Pittsburgh Medical Center, Pittsburgh, PA, United States

[2]Center for Bioethics and Health Law, University of Pittsburgh, Pittsburgh, PA, United States

[3]Johns Hopkins Berman Institute of Bioethics, Baltimore, MD, United States

[4]Consensys Health, Long Island City, NY, United States

[*]these authors contributed equally

**Corresponding Author:**
Marielle S Gross, MD
Department of Obstetrics, Gynecology and Reproductive Services
University of Pittsburgh Medical Center
300 Halket Street
Pittsburgh, PA, 15213
United States
Phone: 1 412 641 1000
Email: grossms@upmc.edu

## Abstract

Henrietta Lacks' deidentified tissue became HeLa cells (the paradigmatic learning health platform). In this article, we discuss separating research on Ms Lacks' tissue from obligations to promote respect, beneficence, and justice for her as a patient. This case illuminates ethical challenges for the secondary use of biospecimens, which persist in contemporary learning health systems. Deidentification and broad consent seek to maximize the benefits of learning from care by minimizing burdens on patients, but these strategies are insufficient for privacy, transparency, and engagement. The resulting supply chain for human cellular and tissue–based products may therefore recapitulate the harms experienced by the Lacks family. We introduce the potential for blockchain technology to build unprecedented transparency, engagement, and accountability into learning health system architecture without requiring deidentification. The ability of nonfungible tokens to maintain the provenance of inherently unique digital assets may optimize utility, value, and respect for patients who contribute tissue and other clinical data for research. We consider the potential benefits and survey major technical, ethical, socioeconomic, and legal challenges for the successful implementation of the proposed solutions. The potential for nonfungible tokens to promote efficiency, effectiveness, and justice in learning health systems demands further exploration.

***(JMIR Bioinform Biotech 2021;2(1):e29905)*** doi:10.2196/29905

**KEYWORDS**

blockchain; biospecimens; research ethics; nonfungible tokens; research ethics; health platforms; HeLa cells; patient data; deidentification; eHealth; data security; integrity

## Introduction

Deidentifying biospecimens "checks the box" of protecting privacy while permitting unrestricted secondary use of clinical data. This workaround transformed Henrietta Lacks' [1] cervical cancer into death-defying HeLa cells (the paradigmatic learning health platform). According to convention in the then-segregated 1951 Johns Hopkins Hospital, tissue obtained during Henrietta Lacks' cancer treatment was deidentified using the first 2 letters of her first and last names, permitting research on her tissue without her explicit knowledge or consent. Ms Lacks died soon after the procedure that harvested the would-be HeLa cells, leaving her family to mourn their loved one without realizing how the cancer that took her life also enabled her to live on through ongoing replication, distribution, and use of her tissue in perpetuity.

Open access to HeLa cells allowed scientists to ask and answer questions about a wide range of viruses, toxins, drugs, and hormones, while avoiding physical risk for human subjects, accelerating a revolution in biomedical science. Decades later, the provenance of the now ubiquitous HeLa cells came to light [2], when chance reidentification and subsequent efforts to

obtain more samples from her living relatives exposed deficiencies in transparency, accountability, and engagement [3] throughout the learning health lifecycle. Today, Henrietta Lacks is recognized as the mother of modern medicine, the bittersweet result of a system that seeks to maximize the research value of tissue and other data-rich byproducts of clinical care.

Though our technical standards for deidentification have evolved, the spirit of deidentification that disconnected Ms Lacks, a poor, Black, mother of 5, from her legacy remains immortalized in US law and is widely exploited by today's research enterprise. We discuss how deidentification potentiated ethical violations in Ms Lacks' case, drawing parallels to contemporary research practices, and propose that nonfungible tokens (NFTs), an innovation building on blockchain technology, may help create a more ethical system for learning from care. The vastness of digital and genomic data has rendered all matter of biospecimens similarly undeidentifiable [4,5]. This use case focuses on excess surgical biospecimens and derived human cellular and tissue–based products; however, similar arguments may apply to digital data and related proprietary algorithms.

## Ethical/Legal Context for US Biomedical Research and Clinical Care

Physicians' fiduciary duty to maintain patient confidentiality is enforced by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (45 CFR 164). The Belmont Report–inspired [6] Common Rule (45 CFR 46) stipulates that respect for persons and risk of "therapeutic misconception" require research participation to be transparent and voluntary, typically fulfilled by prospective informed consent. In this context, the "safe harbor" for deidentified data is justified by assumptions that privacy is patients' only relevant interest regarding secondary research on the tissue and other data produced during clinical care and that deidentification offers sufficient privacy protection. Thus, preemptive deidentification establishes exemptions from HIPAA and human subject research protections.

Current learning platforms leverage deidentified biospecimens to accelerate scientific breakthroughs, empowering precision medicine [7]. Traditional expectations of research participation as altruistic and primarily beneficial for *other future patients* had implications for the structure of benefit distribution in learning health systems. US law prohibiting reidentification and contact (45 CFR § 164.514) prevents individuals from receiving timely access to relevant benefits that may be produced from research on their data. This is especially problematic for patient-centered outcomes research, which focuses on outcomes that matter to patients, and precision medicine research, which seeks to learn from and improve clinical care in real time [8,9].

Patients' tissue and sensitive data are extracted during clinical care under fiduciary duty to benefit them and minimum necessary standards of Protected Health Information (PHI) (45 CFR 164.502(b), 164.514(d)). Systematic deidentification and repurposing of biospecimens without either explicit permission for research use or an intention to directly benefit that individual wherever possible may betray the sacred trust of the patient-physician relationship, compromising respect for patients as persons. Lack of transparency regarding relevant outcomes may contradict obligations of beneficence and information unblocking mandates set forth in the 21st Century Cures Act [10]. Direct, continuous, and ongoing research feedback is feasible with current bioinformatic technologies and programmable terms of use. Allowing indefinite delays in translation from the bench or the cloud back to the bedside of origin represents a lack of accountability to patients, which disproportionately impacts underserved populations who experience the greatest barriers for accessing cutting-edge care [11].

## Hidden in Plain Sight: Deidentification as Erasure Without Engagement

Henrietta Lacks was eventually reidentified, though she was never truly deidentified. Indeed, the Lacks family was easy to find when their help was needed to clear up HeLa contamination in the 70s [2]. The ability to identify a biospecimen's source remains central for the integrity and value of related research. Likewise, current deidentification techniques are a similarly thin veneer of privacy protection [4,5]. Biospecimens are inherently unique; truly deidentifying them may not be possible given the richness of underlying data, advances in genomics, and maturation of artificial intelligence technology [4,5,12].

Unlike methods of concealing or encrypting personal identity and other data security measures, removing personal details impoverishes data sets, reinforces knowledge silos, and hinders continuous global assessment of the data landscape [13-15]. In addition to compromises on identity protection and scientific progress, deidentification prevents individuals from dynamically controlling or directly benefiting from the use of their biospecimens or understanding how they may have benefited others. Ms Lacks did not live to see the transformative effect she had on the world, but she had access to state-of-the-art care and the physician scientist who granted her cells immortality informed her of their world-changing potential. She "was glad her pain would come to some good for someone" [1].

Further harms experienced by obtaining and distributing the Lacks' health data (ie, harvesting her tissue postmortem, seeking DNA from descendants, and publishing her genome, all without adequate transparency or consent) represented lost opportunities for recognition spanning decades, keeping wounds fresh and distributing their effects. The Lacks family narrative demonstrates an intergenerational harm of erasure that compounds the suffering and loss of illness. Subsequent engagement of the Lacks family on a National Institutes of Health review board overseeing HeLa cell use in government-funded research [16] represents a structural breakthrough in transparency, accountability, and engagement. This approach acknowledges the intergenerational nature of tissue and related learning, but has not yet been extended to the families of other patients whose biospecimens have generated foundational learning health platforms.

Like HeLa cells, our biospecimens are the substrate of learning health platforms, and are similarly deserving of respect. Current paradigms use deidentification to mitigate tradeoffs between privacy and utility of health data, allowing rapid scaling of learning platforms in tandem with digitalization. However, the resulting system effectively imposes the violations Ms Lacks experienced on all patients. How this resource-intensive model of engagement scales is unclear and complicated by uncertainty regarding future utility, the differential value of patients' contributions, and the diversity of preferences [17]. Irrespective of whether the Lacks' current engagement model should be normative rather than exceptional, further technological advancements may be essential for a learning health system that optimizes individual and collective rights and interests.

## Technological Solutions With Blockchain Technology and NFTs

As we strive to fully integrate care and research, advancements in blockchain technology and related privacy and intellectual property–preserving innovations may help embed ethical principles in learning health system architecture [18]. The decentralization of blockchains provides uniquely strong assurances of trust in data security, integrity, and use as the network is surveilled and audited by autonomous "smart contracts." The fundamental transparency of the blockchain could enable individuals to track biospecimen use, and smart contracts could automate translation of potential health or personal benefits. Auditability of the learning health system may be crucial for ensuring that past, present, and future uses of human tissue and other clinically derived data are consistent with communal values.

Blockchains are communities of stakeholders organized around interoperable open-source building blocks with shared standards and information, in which sharing is normative, incentivized, and yields collective benefits. Blockchains' underlying ethos of peer-to-peer engagement and cooperation could serve as the backbone of a learning health system that is designed to engage patients as proper stakeholders in learning, like the Lacks' current oversight of HeLa cell use. Such a system could drive learning and translation by asserting patients' values as contributors and empowering enforceable dynamic consent. Democratic engagement in system governance could dictate learning priorities, informed consent requirements appropriate to specific data use contexts, and operational aspects of participation. Importantly, advanced cryptography of blockchain networks allows transparent public engagement with individuals without compromising private identity.

Blockchains are often accompanied with their own cryptocurrency (tokens) to incentivize disparate parties to organize around a common purpose. Tokenization could incentivize patients to contribute their excess tissue to learning activities by providing transparency, trust, and feedback with a durable digital asset that may accrue in personal and health value over time. This supports the imperative to legitimize and democratize citizen science [19] and could provide suffrage and collective representation for patient advocacy movements [20]. Smart contract infrastructure may enable dynamic patient engagement for specific commercial tissue uses. Tokenization may facilitate development of a system for fairly compensating and maintaining transparency with individuals whose biospecimens become the basis of commercial tissue–based products [21].

NFTs, popularized by application to digital artworks [22], introduce the capacity to protect patient rights and interests regarding use of their inherently unique biospecimens and immortal cell lines by creating a corresponding unique digital asset that retains value even as the products are copied and distributed. NFTs could enable exchange of biospecimens to maximize research utility while retaining the unique signature of their human source. NFTs may allow us to capture and prove the value underlying health data without compromising individual or collective benefits of learning from biospecimens or privacy interests. This presents an opportunity for a paradigm shift in the recognition of nonfungible human beings as the basis of learning health platforms with a potential mechanism for ensuring a more just distribution of benefits. NFTs should be explored for their potential to empower provenance of data and duty in a system of learning from care.

Unsolved technical challenges remain for implementing and scaling NFTs for biospecimens. However, increasing adoption of decentralized ledger technologies in health care and beyond, as well as recent successful scaling operations utilizing privacy-preserving technologies (eg, federated learning [23] and homomorphic encryption [24]), support this further development of an ethical learning health system. Holistically, these innovations are ethically significant for learning health systems given their potential to resolve tensions between data utility and privacy; however, a more detailed discussion is beyond the scope of this article. Blockchain has since evolved to incorporate novel architectures to promote equitable collaborations (eg, Decentralized Autonomous Organizations [DAO]) and scalability of resource-intensive decentralized consensus mechanisms (eg, proof-of-stake algorithms). In the context of research on biospecimens, the use of these technologies must also be accompanied by examination of underlying ethical, legal, and social structures.

## Ethical and Socioeconomic Constraints of Tokenizing Human Tissue

The moral vulnerabilities, legal limitations, and practical constraints of tokenizing biospecimens are significant. Tokenization of tumors could incentivize inappropriate health risks on the part of stakeholders, including patients, physicians, and health systems. Secondary gain could motivate patients to delay surgery to allow for larger more valuable tumor samples, and physicians may be pressured to be less thorough in clinical pathologic examinations or more extensive in surgical interventions to maximize tissue yield. An ecosystem of human tissue tokens must anticipate and guard against potential abuses, including those related to monetization, and further tokenomic research can inform the optimal market design. Blockchain may enhance ethical protections for patients and subjects via an embedded approach to ethical oversight that is continuous, evolving, decentralized, and auditable.

In some settings, return of results raises concerns about the potential harms of disclosing information that may either be unwanted or have clinical consequences, including psychological or physical sequelae of subsequent interventions. While these challenges must be addressed, they do not justify acceptance of the status quo in which patients remain disconnected from research results that may be clinically actionable. NFTs could facilitate incorporation of dynamic consent, allowing patient preferences to guide tissue use and benefit distribution. Additional work is needed to determine appropriate informed consent and engagement of diverse populations. Further innovations and clinical pathways must be advanced to ensure safety, quality, and consent for returning results and delivering health benefits of biospecimen research.

The impact of tokenization on patients' willingness to contribute biospecimens must be further evaluated, as some patients have expressed a preference for no-strings-attached donations, although this perspective may not account for the potential for that individual to receive substantial health, personal, or financial benefits. Tools are needed to socialize the outlook that patients can, and should, directly benefit from research on their tissue, and to communicate the potential for blockchain technology to resolve ethical and technical barriers. Novel economic structures, such as curated markets [25], augmented bonding curves [26], and platform cooperatives [27], could be employed in combination with underlying technical structures to optimally align stakeholder incentives, and research into these methods is ongoing. If tokenization leads to monetization of biospecimens, market design must optimize the use of rare uniquely valuable samples and ensure equitable distribution of benefits, and will require the support of updated legal protections.

## Legal and Practical Barriers for an NFT-Based Learning Health System

Aside from how NFTs could work for prospective biospecimens, retroactive application may be challenged by US regulations prohibiting reidentification and contact of deidentified research subjects. Since reconnecting patients with the knowledge and products from past biospecimen use is a major value proposition, devising strategies that do not violate established legal and ethical obligations will be critical. NFTs may circumvent prohibitions against reidentification, regardless of whether broad consent was initially obtained or not mandated for past specimen uses for which there was no expectation of contact, as they could provide patients an opportunity to opt in for subsequent reidentification, while providing discretionary preferences that could be updated over time. This overcomes the patient-facing concerns about privacy and autonomy, but does not eliminate institutional and researcher concerns about obligations or past lapses in disclosure that may arise.

By comparison with US law, the General Data Protection Regulation (GDPR) seeks to center individuals' data rights as a matter of liberal democracy. A major challenge for the GDPR is the lack of mechanisms for enforcing ethical principles and socioeconomic policies for data use. Despite advantages for consumer autonomy, the GDPR has also been challenged as its limitations on secondary data use may frustrate efforts to maximize the individual and collective value of learning from health data [28]. Endless cookie requests signal consent but may not constitute meaningful control, especially if acceptance of cookies predicates access to needed health services. In practice, GDPR requirements resemble broad consent for the secondary use of biospecimens, aligned with the updated Common Rule [29]. Critically, the GDPR does not address the duty to distribute knowledge and products derived from secondary data and tissue use. Without a means of ensuring a just distribution of benefits, the underlying power asymmetry between individuals and third party data users persists.

Acceptability for current institutional tissue "owners" will be a major barrier to the implementation of NFTs for biospecimens, as fears of adverse press, decreased contributions, and legal repercussions from efforts to tokenize tissue loom large. "Safe harbor" for deidentified data and unsuccessful prior attempts of patients to obtain rights regarding the commercial use of their biosamples may reinforce institutional inertia [30]. Using NFTs for biospecimens recognizes their status as assets and may increase the call for updating definitions regarding rights, ownership, and value distribution.

Maximizing efficiency, effectiveness, and justice will ultimately require global collaboration in learning from care. Novel approaches to data governance must go above and beyond existing policy protections with an eye toward technological evolution and international standards. Transparency, auditability, and smart contract architecture could empower individuals or their representatives to ensure that data use is in accordance with policies and preferences while maximizing collective benefits. Additional measures for enforcing legal compliance, social democratic governance, and ethical oversight are needed to guard against potentially exploitative treatment of vulnerable populations within high income settings and worldwide.

## Conclusion

Henrietta Lacks' story highlights the harms that may occur when deidentification separates research on patient tissue from obligations to promote respect, beneficence, and justice for that individual. Continued reliance on deidentification and broad consent for the "secondary use" of biospecimens may create platforms for learning that recapitulate historically exploitative practices of integrating research and patient care. Blockchain technology promises to build unprecedented transparency, engagement, and accountability into learning health system architecture. NFTs have the potential to embed the primacy of clinical ethics into our clinical research supply chains. HeLa cells are the original "use case" for NFTs, as they demonstrate the imperative of maintaining the provenance of nonfungible human-derived assets and the fiduciary duties to respective patients. Representing biospecimens with NFTs may maximize efficiency, effectiveness, and justice in the future of learning health systems.

## Conflicts of Interest

None declared.

## References


1. Skloot R. The Immortal Life of Henrietta Lacks. New York, NY: Crown Publishing; 2011.
2. Grady D. A Lasting Gift to Medicine That Wasn't Really a Gift. The New York Times. 2010. URL: https://www.nytimes.com/2010/02/02/health/02seco.html [accessed 2021-09-21]
3. Kass NE, Faden RR. Ethics and Learning Health Care: The Essential roles of engagement, transparency, and accountability. Learn Health Syst 2018 Oct 18;2(4):e10066 [FREE Full text] [doi: 10.1002/lrh2.10066] [Medline: 31245590]
4. Na L, Yang C, Lo C, Zhao F, Fukuoka Y, Aswani A. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. JAMA Netw Open 2018 Dec 07;1(8):e186040 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.6040] [Medline: 30646312]
5. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339(6117):321-324. [doi: 10.1126/science.1229566] [Medline: 23329047]
6. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. The U.S. Department of Health & Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html [accessed 2021-04-21]
7. Cohen JK. Mayo Clinic's new data-sharing initiative launches first project. Modern Healthcare. 2020. URL: https://www.modernhealthcare.com/information-technology/mayo-clinics-new-data-sharing-initiative-launches-first-project [accessed 2021-04-21]
8. Precision health: Improving health for each of us and all of us. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/genomics/about/precision_med.htm [accessed 2021-09-21]
9. Our Programs. Patient-Centered Outcomes Research Institute. URL: https://www.pcori.org/about-us/our-programs [accessed 2021-02-25]
10. 21st Century Cures Act. U.S. Food and Drug Administration. URL: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act [accessed 2021-04-21]
11. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec 16;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
12. Hudson K, Collins F. Bringing the Common Rule into the 21st Century. N Engl J Med 2015 Dec 10;373(24):2293-2296 [FREE Full text] [doi: 10.1056/NEJMp1512205] [Medline: 26509903]
13. Mathews D, Jamal L. Revisiting respect for persons in genomic research. Genes (Basel) 2014 Jan 22;5(1):1-12 [FREE Full text] [doi: 10.3390/genes5010001] [Medline: 24705284]
14. Summary Table of Recommendations on the HIPAA Privacy Rule. The U.S. Department of Health & Human Services. 2016. URL: https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2004-september-27-letter-summary/index.html [accessed 2021-04-21]
15. Lund H, Brunnhuber K, Juhl C, Robinson K, Leenaars M, Dorch BF, et al. Towards evidence based research. BMJ 2016 Oct 21;355:i5440 [FREE Full text] [doi: 10.1136/bmj.i5440] [Medline: 27797786]
16. ACD HeLa Genome Data Access Working Group. National Institutes of Health. URL: https://www.acd.od.nih.gov/working-groups/hlgda.html [accessed 2021-04-21]
17. Persad G, Fernandez Lynch H, Largent E. Differential payment to research participants in the same study: an ethical analysis. J Med Ethics 2019 May 07;45(5):318-322. [doi: 10.1136/medethics-2018-105140] [Medline: 30846490]
18. Gross M, Miller R. Ethical Implementation of the Learning Healthcare System with Blockchain Technology. BHTY 2019 Jun 07;2:2 [FREE Full text] [doi: 10.30953/bhty.v2.113]
19. Gross MS, Miller RC. "Paid to Produce Data:" Research Participation as the Labor of Generating Valuable Health Data. Am J Bioeth 2019 Sep 16;19(9):50-52. [doi: 10.1080/15265161.2019.1630500] [Medline: 31419203]
20. Irwin A. No PhDs needed: how citizen science is transforming research. Nature 2018 Oct 23;562(7728):480-482. [doi: 10.1038/d41586-018-07106-5] [Medline: 30353162]
21. Fliesler N. In rare disease, precision medicine meets citizen science. Vector - Boston Children's Hospital. 2015. URL: https://vector.childrenshospital.org/2015/07/power-to-the-people-citizen-science-meets-precision-medicine-for-rare-disease/ [accessed 2021-04-21]
22. What are NFTs and why are some worth millions? BBC. URL: https://www.bbc.com/news/technology-56371912 [accessed 2021-04-21]
23. New Machine Learning Method Allows Hospitals to Share Patient Data Privately. Penn Medicine. 2020. URL: https://tinyurl.com/zmttpbsw [accessed 2021-04-21]
24. Kocabaş Ö, Soyata T. Medical Data Analytics in the Cloud Using Homomorphic Encryption. In: Raj P, Deka GC, editors. Handbook of Research on Cloud Infrastructures for Big Data Analytics. Hershey, PA: IGI Global; 2014.
25. de la Rouviere S. Tokens 2.0: Curved Token Bonding in Curation Markets. Medium. 2017. URL: https://medium.com/@simondlr/tokens-2-0-curved-token-bonding-in-curation-markets-1764a2e0bee5 [accessed 2021-04-21]

26. Will bonding curves be researchers' best friends? Blue Steens. URL: https://www.blue-steens.com/token-bonding-curves-curation-markets/ [accessed 2021-04-21]
27. Sandoval M. Entrepreneurial Activism? Platform Cooperativism Between Subversion and Co-optation. Critical Sociology 2019 Nov 13;46(6):801-817. [doi: 10.1177/0896920519870577]
28. Hallinan D. Broad consent under the GDPR: an optimistic perspective on a bright future. Life Sci Soc Policy 2020 Jan 06;16(1):1 [FREE Full text] [doi: 10.1186/s40504-019-0096-3] [Medline: 31903508]
29. Marelli L, Lievevrouw E, Van Hoyweghen I. Fit for purpose? The GDPR and the governance of European digital health. Policy Studies 2020 Feb 10;41(5):447-467. [doi: 10.1080/01442872.2020.1724929]
30. Law School Case Brief. Tort Law. Informed Consent. California Supreme Court Recognizes Patient's Cause of Action for Physician's Nondisclosure of Excised Tissue's Commercial Value. Moore v. Regents of the University of California, 51 Cal. 3d 120, 793 P.2d 479, 271 Cal. Rptr. 146 (1990). Harvard Law Review 1991 Jan;104(3):808. [doi: 10.2307/1341579]


## Abbreviations

**GDPR:** General Data Protection Regulation
**HIPAA:** Health Insurance Portability and Accountability Act
**NFT:** nonfungible token


*Edited by G Eysenbach; submitted 24.04.21; peer-reviewed by V Katavic, M Jungkunz; comments to author 28.06.21; revised version received 30.09.21; accepted 06.10.21; published 22.10.21.*

*Please cite as:*
*Gross MS, Hood AJ, Miller Jr RC*
*Nonfungible Tokens as a Blockchain Solution to Ethical Challenges for the Secondary Use of Biospecimens: Viewpoint*
*JMIR Bioinform Biotech 2021;2(1):e29905*
*URL: https://bioinform.jmir.org/2021/1/e29905*
*doi:10.2196/29905*
*PMID:*

Original Paper

# Isolating SARS-CoV-2 Strains From Countries in the Same Meridian: Genome Evolutionary Analysis

Emilio Mastriani[1,2], PhD; Alexey V Rakov[3], PhD; Shu-Lin Liu[1,2,4], PhD

[1]Systemomics Center, College of Pharmacy, Genomics Research Center, State-Province Key Laboratories of Biomedicine-Pharmaceutics of China, Harbin Medical University, Harbin, China

[2]HMU-UCCSM Centre for Infection and Genomics, Harbin Medical University, Harbin, China

[3]Somov Institute of Epidemiology and Microbiology, Vladivostok, Russian Federation

[4]Department of Microbiology, Immunology and Infectious Diseases, University of Calgary, Calgary, AB, Canada

**Corresponding Author:**
Emilio Mastriani, PhD
HMU-UCCSM Centre for Infection and Genomics
Harbin Medical University
No 157, Baojian Road
Harbin, 150081
China
Phone: 86 13664502721 ext 64502721
Email: emiliomastriani@icloud.com

## Abstract

**Background:** COVID-19, caused by the novel SARS-CoV-2, is considered the most threatening respiratory infection in the world, with over 40 million people infected and over 0.934 million related deaths reported worldwide. It is speculated that epidemiological and clinical features of COVID-19 may differ across countries or continents. Genomic comparison of 48,635 SARS-CoV-2 genomes has shown that the average number of mutations per sample was 7.23, and most SARS-CoV-2 strains belong to one of 3 clades characterized by geographic and genomic specificity: Europe, Asia, and North America.

**Objective:** The aim of this study was to compare the genomes of SARS-CoV-2 strains isolated from Italy, Sweden, and Congo, that is, 3 different countries in the same meridian (longitude) but with different climate conditions, and from Brazil (as an outgroup country), to analyze similarities or differences in patterns of possible evolutionary pressure signatures in their genomes.

**Methods:** We obtained data from the Global Initiative on Sharing All Influenza Data repository by sampling all genomes available on that date. Using HyPhy, we achieved the recombination analysis by genetic algorithm recombination detection method, trimming, removal of the stop codons, and phylogenetic tree and mixed effects model of evolution analyses. We also performed secondary structure prediction analysis for both sequences (mutated and wild-type) and "disorder" and "transmembrane" analyses of the protein. We analyzed both protein structures with an ab initio approach to predict their ontologies and 3D structures.

**Results:** Evolutionary analysis revealed that codon 9628 is under episodic selective pressure for all SARS-CoV-2 strains isolated from the 4 countries, suggesting it is a key site for virus evolution. Codon 9628 encodes the P0DTD3 (Y14_SARS2) uncharacterized protein 14. Further investigation showed that the codon mutation was responsible for helical modification in the secondary structure. The codon was positioned in the more ordered region of the gene (41-59) and near to the area acting as the transmembrane (54-67), suggesting its involvement in the attachment phase of the virus. The predicted protein structures of both wild-type and mutated P0DTD3 confirmed the importance of the codon to define the protein structure. Moreover, ontological analysis of the protein emphasized that the mutation enhances the binding probability.

**Conclusions:** Our results suggest that RNA secondary structure may be affected and, consequently, the protein product changes T (threonine) to G (glycine) in position 50 of the protein. This position is located close to the predicted transmembrane region. Mutation analysis revealed that the change from G (glycine) to D (aspartic acid) may confer a new function to the protein—binding activity, which in turn may be responsible for attaching the virus to human eukaryotic cells. These findings can help design in vitro experiments and possibly facilitate a vaccine design and successful antiviral strategies.

***(JMIR Bioinformatics Biotechnol 2021;2(1):e25995)*** doi:10.2196/25995

**KEYWORDS**

SARS-CoV-2; evolutionary analysis; episodic selective pressure; virus evolution; codon mutation; binding probability; evolution; genome; genetics; COVID-19; virus; strain; codon; pressure; mutation; structure; prediction; protein

## Introduction

The ongoing COVID-19 pandemic caused by the novel SARS-CoV-2 is the most threatening respiratory infection worldwide and has affected almost every country in the world. As of December 30, 2020, over 81 million people were infected with COVID-19, and more than 1.7 million deaths were reported. Many health institutions are attempting to produce effective vaccines against this virus infection, and several are now in the final stages of development before their application to human populations [1,2].

The SARS-CoV-2 genome shares approximately 82% sequence identity with SARS-CoV and MERS-CoV (Middle East respiratory syndrome coronavirus) and more than 90% sequence identity for essential enzymes and structural proteins. This high level of sequence identity suggests a common pathogenesis mechanism and, thus, therapeutic targeting. SARS-CoV-2 contains 4 structural proteins, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins [3]. The structure and the genome of SARS-CoV-2 are being extensively studied, but the results seem to be controversial. For example, a recent study found that the 2 integral membrane proteins (ie, envelope and membrane proteins) tend to evolve slowly by accumulating nucleotide mutations on their corresponding genes, but genes encoding nucleocapsid, viral replicase and spike proteins, which are regarded as important targets for the development of vaccines and antiviral drugs, tend to evolve faster [4]. However, other studies have shown that potential drug targets of SARS-CoV-2 are highly conserved [3].

The genome of SARS-CoV-2 is comprised of a single-stranded positive-sense RNA. A newly sequenced genome of SARS-CoV-2 was submitted to the NCBI genome database (NC_045512.2). The genetic makeup of SARS-CoV-2 is composed of 13-15 (including 12 functional) open reading frames (ORFs) containing ~30,000 nucleotides. The genome contains 38% of GC content and 11 protein-coding genes, together expressing 12 proteins [3].

The genomic characterization of 95 SARS-CoV-2 genomes revealed the 2 most common mutations that might affect the severity and spread of SARS-CoV-2 [5]. Another study highlighted the crucial genomic features that are unique to SARS-CoV-2 and 2 other deadly coronaviruses, SARS-CoV and MERS-CoV. These unique features correlate with the high fatality rate due to infection with these coronaviruses as well as their ability to switch hosts from animals to humans [6]. As a result, it can be speculated that the epidemiological and clinical features of these viruses may differ across countries or continents.

Genomic comparison of 48,635 SARS-CoV-2 genomes has shown that the average number of mutations per sample was 7.23, and most SARS-CoV-2 strains belong to one of the following 3 clades characterized by geographic and genomic specificity: clade G (Europe), clade L (Asia), and G-derived clade (North America) [7]. These results suggest custom-designed antiviral strategies based on the molecular specificities of SARS-CoV-2 in patients from different geographies [7]. Previous studies have also differentiated the 3 variants according to the geographic location (East Asia, Europe, and America) [8]. A more recent genome-wide analysis revealed that the frequency of amino acid mutations was higher in the genome sequences of SARS-CoV-2 strains from Europe (43.07%), followed by strains from Asia (38.09%) and North America (29.64%). However, case fatality rates remained higher in the European temperate countries, such as Italy, Spain, Netherlands, France, England, and Belgium [9].

The aim of this study was to compare the set of SARS-CoV-2 genomes of viral strains isolated from representative countries in the same meridian (longitude), namely, Italy, Sweden, and Congo, which have different climate conditions, to reveal similarities or differences in the patterns of possible evolutionary pressure signatures in their genomes.

## Methods

### Sequence Data

We obtained data from the Global Initiative on Sharing All Influenza Data (GISaid) repository and sampled all genomes available therein to that date (May 5, 2020), including the files congo-gisaid_hcov-19_2020_05_05_09.fasta with 75 entries, italy-gisaid_hcov-19_2020_05_05_10.fasta with 69 entries, sweden-gisaid_hcov-19_2020_05_05_10.fasta with 104 entries, and also the outgroup file brazil_gisaid_hcov-19_2020_05_15_04.fasta with 92 entries. The reference genome with accession number NC_045512.2 was downloaded from the GenBank repository.

### Evolution Model Analysis

We used the SARS-CoV-2 Wuhan-Hu-1 genome (RefSeq Acc. No. NC_045512.2) as the reference sequence and the VIRULIGN version 1.0.1 application [10] to perform multiple sequence alignment, with AliView version 1.26 application for visualizing the results of the analyses [11]. HyPhy 2.5.8 (MP) was used to perform recombination analysis by the genetic algorithm recombination detection method and conduct trimming, stop codon removal, and phylogenetic tree and mixed effects model of evolution (MEME) analyses [12]. The MEME web site was used to read JSON output files and generate MEME images and tables.

### RNA Secondary Structure Prediction

We used the RNA_fold web server (part of the Vienna RNA Websuite) to predict secondary structures of both the wild-type and mutated sequences [13], and the Forna package [14] to build the graph diagrams.

### Protein Analysis

Protein disorder analysis was conducted using MFDp2 [15], NetSurfP-2.0 [16], and SPOT-Disorder2 [17] applications.

Transmembrane analysis of the protein was calculated using the TMHMM server v.2.0, MemBrain webserver [18], ProtScale [19], and TMpred [20] (scores normalized for comparison) on the Expasy website [21].

### 3D Protein Structure Prediction and Ontologies

Both protein structures were determined with an ab initio approach by using the Robetta webserver [22], whereas DeeProtein capsule from OCEAN CODE [23] was used to predict ontologies of the predicted proteins. 3D images of protein structures and their ontologies were released using PyMOL 2.4.0 [24].

## Results

### Codon 9628 Evolved Under Episodic Positive Selection

Mixed evolutionary analysis based on the MEME algorithm was conducted on the SARS-CoV-2 data from Italy, Sweden, and Congo (countries from the same geographic meridian) and Brazil (included as an outgroup). The investigation revealed codon 9628 was under episodic positive selective pressure across the countries, as depicted in Table 1.

**Table 1.** Mixed effects model of evolution (MEME_ analysis results showing data obtained from the evolutionary analysis of SARS-CoV-2 from Brazil, Congo, Italy, and Sweden. The top 3 sites for every country are shown, sorted by *P* value.

| Country (ID)/Site | Partition | α | $\beta^-$ | $p^-$ | $\beta^+$ | $p^+$ | LRT | *P* value | Branches under selection | Total branch length | MEME LogL | Fixed effects likelihood LogL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Brazil (BR)** | | | | | | | | | | | | |
| 9628[a] | 1 | 0 | 0 | 0.96 | 10,000 | 0.04 | 16.37 | <.001 | 2 | 0.65 | -27.28 | -20.62 |
| 9928 | 1 | 0 | 0 | 0.82 | 10,000 | 0.18 | 11.12 | <.001 | 4 | 2.71 | -31.03 | -28.53 |
| 81 | 1 | 0 | 0 | 0.04 | 1032.18 | 0.96 | 6.95 | .01 | 5 | 1.49 | -40.77 | -40.77 |
| **Congo (CG)** | | | | | | | | | | | | |
| 9628[a] | 1 | 0 | 0 | 0.97 | 10,000 | 0.03 | 10.89 | <.001 | 1 | 0.25 | -18.18 | -13.54 |
| 2884 | 1 | 0 | 0 | 0.45 | 1273.45 | 0.55 | 3.51 | .08 | 5 | 0.60 | -42.49 | -42.37 |
| 6541 | 1 | 0 | 0 | 0.97 | 10,000 | 0.03 | 2.73 | .12 | 1 | 0.27 | -12.94 | -11.92 |
| **Italy (IT)** | | | | | | | | | | | | |
| 15 | 1 | 0 | 0 | 0.96 | 10,000 | 0.04 | 10.21 | <.001 | 1 | 0.73 | -15.90 | -12.57 |
| 9628[a] | 1 | 0 | 0 | 0.97 | 1,0000 | 0.03 | 11.24 | <.001 | 1 | 0.45 | -17.66 | -12.95 |
| 4 | 1 | 0 | 0 | 0.89 | 10,000 | 0.11 | 7.25 | .01 | 0 | 1.83 | -13.11 | -10.43 |
| **Sweden (SE)** | | | | | | | | | | | | |
| 9628[a] | 1 | 0 | 0 | 0.96 | 9613.52 | 0.04 | 16.03 | <.001 | 2 | 0.51 | -27.43 | -21.10 |
| 4409 | 1 | 0 | 0 | 0.97 | 4356.70 | 0.03 | 7.68 | .01 | 1 | 0.16 | -15.63 | -12.33 |
| 4732 | 1 | 0 | 0 | 0.95 | 10,000 | 0.05 | 3.85 | .07 | 2 | 0.74 | -19.66 | -18.78 |

[a]Indicates site 9628.

In this context, we use the term “site” as a synonym of codon, respecting the HyPhy terminology. The asymptotic *P* value was <.001 for episodic diversification at site 9628. Figure 1 shows the distribution of the *P* value across the sites for all 4 countries.

A deep check of the multiple alignment data of the 4 countries revealed that the episodic positive selective pressure on site 9628 is a consistent mutation of the codon GGG to ACG, as shown in Figure 2.

**Figure 1.** Mixed effects model of evolution site plot. Distribution of the *P* value over the sites in Brazil, Congo, Italy, and Sweden. The purple circle indicates site 9628 that was found to be under episodic selective pressure.

Brazil

| Site | Partition | α | β⁻ | p⁻ | β⁺ | p⁺ | LRT | p-value | # branches under selection | Total branch length | MEME LogL | FEL LogL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9628 | 1 | 0.00 | 0.00 | 0.96 | 10000.00 | 0.04 | 16.37 | 0.00 | 2.00 | 0.65 | -27.28 | -20.62 |
| 9928 | 1 | 0.00 | 0.00 | 0.82 | 10000.00 | 0.18 | 11.12 | 0.00 | 4.00 | 2.71 | -31.03 | -28.53 |
| 81 | 1 | 0.00 | 0.00 | 0.04 | 1032.18 | 0.96 | 6.95 | 0.01 | 5.00 | 1.49 | -40.77 | -40.77 |
| 9929 | 1 | 0.00 | 0.00 | 0.00 | 916.08 | 1.00 | 4.00 | 0.06 | 2.00 | 1.38 | -22.87 | -22.87 |
| 8715 | 1 | 0.00 | 0.00 | 0.00 | 329.35 | 1.00 | 3.80 | 0.07 | 3.00 | 0.50 | -25.49 | -25.49 |
| 3695 | 1 | 0.00 | 0.00 | 0.00 | 526.37 | 1.00 | 3.26 | 0.09 | 7.00 | 0.79 | -35.40 | -35.40 |
| 5469 | 1 | 0.00 | 0.00 | 0.00 | 178.34 | 1.00 | 2.25 | 0.16 | 2.00 | 0.27 | -17.10 | -17.10 |
| 9382 | 1 | 0.00 | 0.00 | 0.00 | 198.57 | 1.00 | 2.18 | 0.17 | 2.00 | 0.30 | -12.96 | -12.96 |
| 4803 | 1 | 0.00 | 0.00 | 0.00 | 176.79 | 1.00 | 1.96 | 0.19 | 1.00 | 0.27 | -19.00 | -19.00 |
| 8521 | 1 | 0.00 | 0.00 | 0.00 | 481.50 | 1.00 | 1.91 | 0.19 | 4.00 | 0.72 | -30.04 | -30.04 |

Congo

| Site | Partition | α | β⁻ | p⁻ | β⁺ | p⁺ | LRT | p-value | # branches under selection | Total branch length | MEME LogL | FEL LogL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9628 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 10.89 | 0.00 | 1.00 | 0.25 | -18.18 | -13.54 |
| 2884 | 1 | 0.00 | 0.00 | 0.45 | 1273.45 | 0.55 | 3.51 | 0.08 | 5.00 | 0.60 | -42.49 | -42.37 |
| 6541 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 2.73 | 0.12 | 1.00 | 0.27 | -12.94 | -11.92 |
| 1754 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 2.60 | 0.13 | 1.00 | 0.28 | -12.24 | -11.40 |
| 9452 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 2.18 | 0.17 | 1.00 | 0.24 | -12.77 | -11.87 |
| 5616 | 1 | 0.00 | 0.00 | 0.51 | 522.55 | 0.49 | 1.88 | 0.19 | 2.00 | 0.22 | -15.60 | -15.59 |
| 8008 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 1.63 | 0.22 | 1.00 | 0.28 | -11.55 | -11.13 |
| 3695 | 1 | 0.00 | 0.00 | 0.00 | 385.65 | 1.00 | 1.46 | 0.24 | 1.00 | 0.33 | -17.03 | -17.03 |
| 9401 | 1 | 0.00 | 0.00 | 0.00 | 192.77 | 1.00 | 1.31 | 0.27 | 2.00 | 0.16 | -15.29 | -15.29 |
| 853 | 1 | 0.00 | 0.00 | 0.00 | 145.84 | 1.00 | 1.22 | 0.28 | 1.00 | 0.12 | -8.80 | -8.80 |

Italy

| Site | Partition | α | β⁻ | p⁻ | β⁺ | p⁺ | LRT | p-value | # branches under selection | Total branch length | MEME LogL | FEL LogL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 1 | 0.00 | 0.00 | 0.96 | 10000.00 | 0.04 | 10.21 | 0.00 | 1.00 | 0.73 | -15.90 | -12.57 |
| 9628 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 11.24 | 0.00 | 1.00 | 0.45 | -17.66 | -12.95 |
| 4 | 1 | 0.00 | 0.00 | 0.89 | 10000.00 | 0.11 | 7.25 | 0.01 | 0.00 | 1.83 | -13.11 | -10.43 |
| 11 | 1 | 0.00 | 0.00 | 0.81 | 4313.16 | 0.19 | 7.57 | 0.01 | 2.00 | 1.39 | -24.95 | -22.23 |
| 10 | 1 | 0.00 | 0.00 | 0.93 | 3851.93 | 0.07 | 3.56 | 0.08 | 1.00 | 0.46 | -15.14 | -12.89 |
| 9965 | 1 | 0.00 | 0.00 | 0.93 | 3420.44 | 0.07 | 3.12 | 0.10 | 1.00 | 0.42 | -15.22 | -12.44 |
| 1419 | 1 | 0.00 | 0.00 | 0.92 | 2815.37 | 0.08 | 2.89 | 0.11 | 2.00 | 0.36 | -17.87 | -17.26 |
| 2290 | 1 | 0.00 | 0.00 | 0.81 | 2108.91 | 0.19 | 1.89 | 0.19 | 3.00 | 0.69 | -20.84 | -20.54 |
| 5486 | 1 | 0.00 | 0.00 | 0.94 | 3954.10 | 0.06 | 1.91 | 0.19 | 1.00 | 0.41 | -11.12 | -10.85 |
| 33 | 1 | 0.00 | 0.00 | 0.97 | 10000.00 | 0.03 | 1.82 | 0.20 | 1.00 | 0.51 | -11.72 | -11.12 |

Sweden

| Site | Partition | α | β⁻ | p⁻ | β⁺ | p⁺ | LRT | p-value | # branches under selection | Total branch length | MEME LogL | FEL LogL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9628 | 1 | 0.00 | 0.00 | 0.96 | 9613.52 | 0.04 | 16.03 | 0.00 | 2.00 | 0.51 | -27.43 | -21.10 |
| 4409 | 1 | 0.00 | 0.00 | 0.97 | 4356.70 | 0.03 | 7.68 | 0.01 | 1.00 | 0.16 | -15.63 | -12.33 |
| 4732 | 1 | 0.00 | 0.00 | 0.95 | 10000.00 | 0.05 | 3.85 | 0.07 | 2.00 | 0.64 | -19.66 | -18.78 |
| 8104 | 1 | 0.00 | 0.00 | 0.00 | 124.39 | 1.00 | 1.61 | 0.23 | 1.00 | 0.17 | -9.20 | -9.20 |
| 3695 | 1 | 0.00 | 0.00 | 0.00 | 272.66 | 1.00 | 1.36 | 0.26 | 4.00 | 0.37 | -22.86 | -22.86 |
| 8715 | 1 | 0.00 | 0.00 | 0.00 | 125.60 | 1.00 | 1.35 | 0.26 | 2.00 | 0.17 | -13.95 | -13.95 |
| 853 | 1 | 0.00 | 0.00 | 0.00 | 105.53 | 1.00 | 1.19 | 0.28 | 1.00 | 0.14 | -8.88 | -8.88 |
| 8521 | 1 | 0.00 | 0.00 | 0.51 | 366.97 | 0.49 | 1.23 | 0.28 | 2.00 | 0.24 | -16.90 | -16.89 |
| 571 | 1 | 0.00 | 0.00 | 0.86 | 603.74 | 0.14 | 1.07 | 0.30 | 1.00 | 0.11 | -12.53 | -12.41 |
| 2285 | 1 | 0.00 | 0.00 | 0.50 | 217.46 | 0.50 | 1.10 | 0.30 | 1.00 | 0.15 | -10.07 | -10.07 |

**Figure 2.** Part of the multiple sequence alignment from the Italian data showing the site 9628 under episodic selective pressure. The nucleotides mute from GGG to ACG.


## RNA Secondary Structure Prediction Changes

The prediction of secondary structure before and after mutation shows important differences, as shown by the mutation from GGG to ACG (Figure 3). The comparison between the 2 predicted secondary structures highlighted structural modifications at the top-right ring of the RNA conformation, as depicted in Figure 4, suggesting the GGG to ACG mutation was responsible for a significant modification of the RNA secondary structure.

**Figure 3.** Nucleotide mutation over aligned sequences, illustrating the sequence considered to predict secondary structures in both mutated and wild-type proteins. Site position is indicated in blue, from the start codon (9578) to the open reading frame (9632).

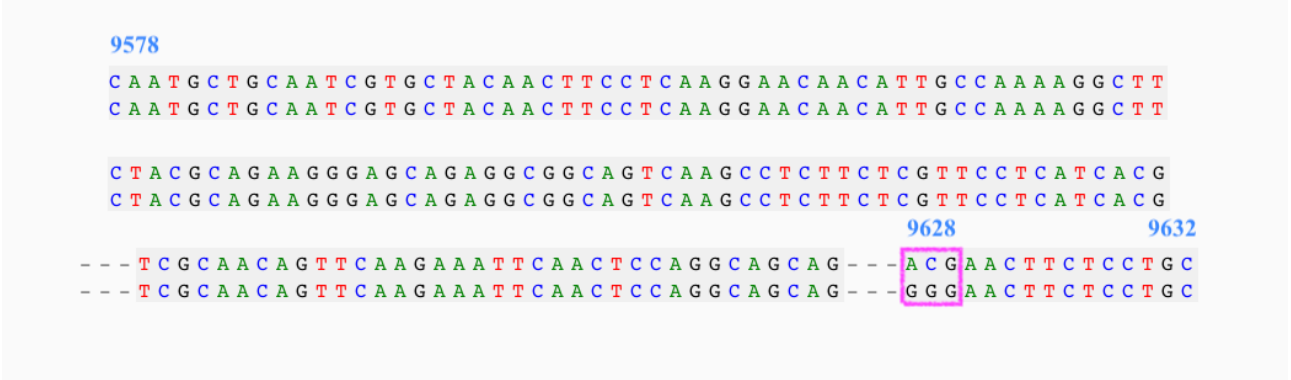
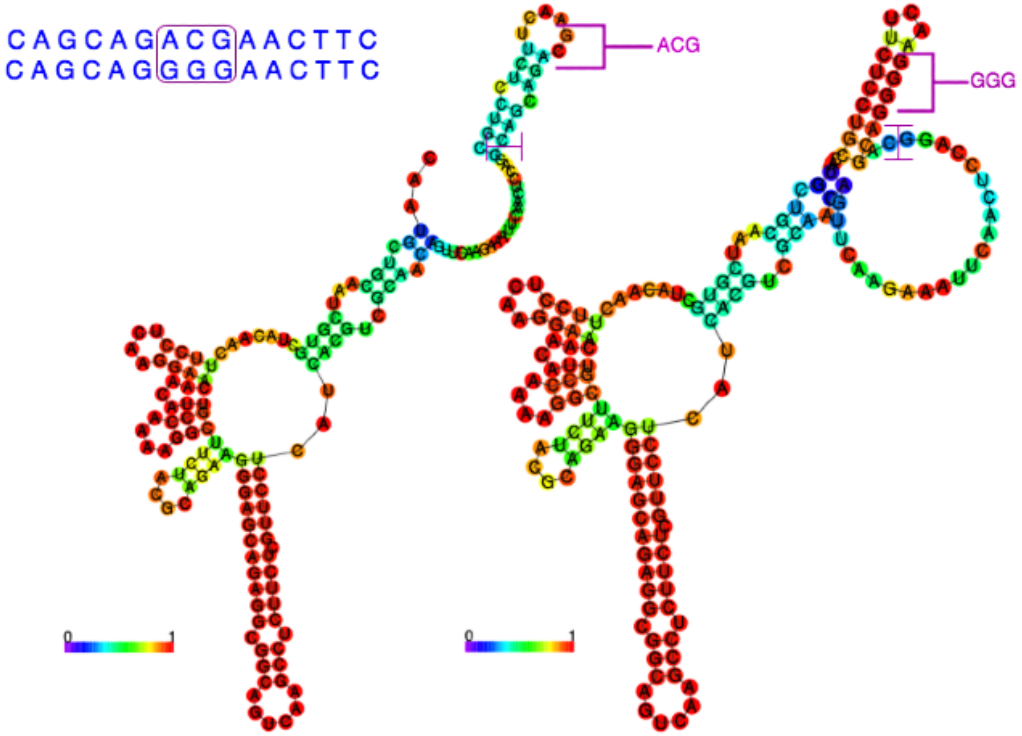


**Figure 4.** Secondary structure prediction. The 2 RNA diagrams exhibit structural modifications affected by the GGG to ACG mutation.




## Protein Analysis

The analysis of the protein conducted for finding its disordered region turned out the positions from 41 to 59 to be more stable with the glycine (G) placed at the 50th position. We obtained results by using 3 different software tools and considering the average value for the probability of disorder, as shown in Figure 5 and reported in Table 2. Further analysis to locate the transmembrane region in the protein revealed that locations 54-67 were associated with this function. The analysis, conducted by using 4 distinct web applications and by evaluating the resultant average values, places the glycine (G) as near the transmembrane region to suppose its involvement. Table 3 reports the data showing the probabilities of each amino acid acting as the transmembrane. The transmembrane topology of the sequence (Figure 6) highlights the amino acid G at location 50 in the middle of the transmembrane region, and the distribution of the probabilities (Figure 7) corroborates this hypothesis.

**Figure 5.** Disorder region analysis. The region 41-59 was found to have the lowest probability to be disordered. The orange lines delimit this region, and the blue dotted line outlines the position of G on the different curves.

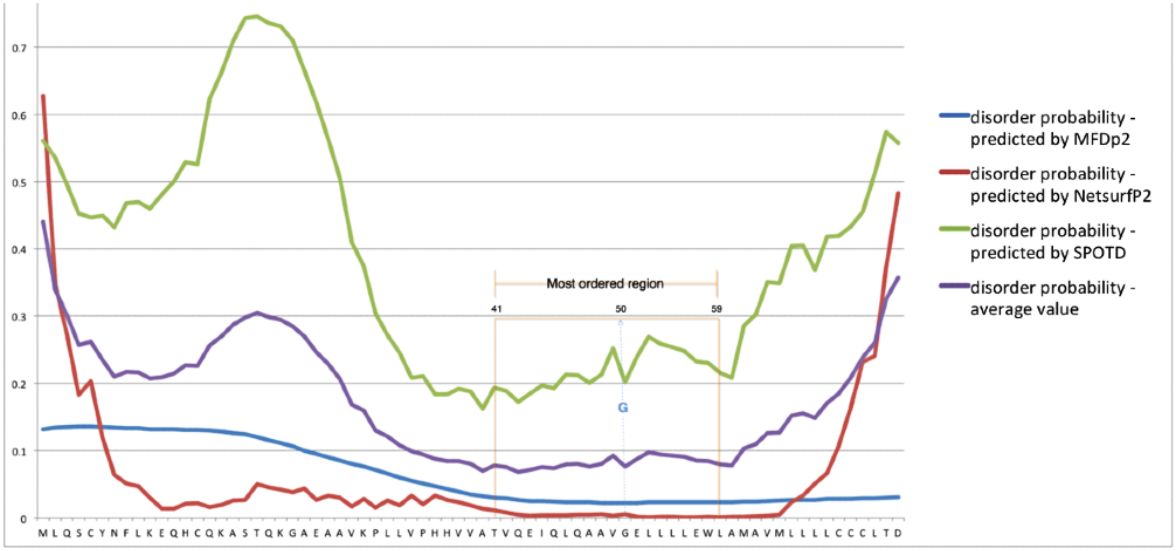


**Table 2.** Protein disorder analysis results showing the probability of disorder for each position of the protein. The probabilities have been calculated using MFDp2, Netsurf, and SPOTD software.

| Position | Amino acid sequence | Disorder probability values | | | |
|---|---|---|---|---|---|
| | | MFDp2 | NetsurfP2 | SPOTD | Average value[a] |
| 1 | M | 0.132 | 0.627823114 | 0.5607 | 0.440174371 |
| 2 | L | 0.134 | 0.347978383 | 0.5358 | 0.339259461 |
| 3 | Q | 0.135 | 0.270706475 | 0.4945 | 0.300068825 |
| … | | | | | |
| 39 | T | 0.03 | 0.010842944 | 0.1936 | 0.078147648 |
| 40 | V | 0.029 | 0.007660664 | 0.189 | 0.075220221 |
| 41 | Q | 0.027 | 0.004478907 | 0.172 | 0.067826302 |
| 42 | E | 0.025 | 0.00340931 | 0.1848 | 0.07106977 |
| 43 | I | 0.025 | 0.003887762 | 0.1968 | 0.075229254 |
| 44 | Q | 0.024 | 0.003997837 | 0.1927 | 0.073565946 |
| 45 | L | 0.023 | 0.00361518 | 0.2129 | 0.079838393 |
| 46 | Q | 0.023 | 0.004551574 | 0.2123 | 0.079950525 |
| 47 | A | 0.023 | 0.004939525 | 0.2011 | 0.076346508 |
| 48 | A | 0.022 | 0.005752307 | 0.2133 | 0.080350769 |
| 49 | V | 0.022 | 0.002826149 | 0.2524 | 0.092408716 |
| 50[b] | G | 0.022 | 0.005828088 | 0.2013 | 0.076376029 |
| 51 | E | 0.022 | 0.001046103 | 0.24 | 0.087682034 |
| 52 | L | 0.023 | 0.000922468 | 0.2694 | 0.097774156 |
| 53 | L | 0.023 | 0.001263275 | 0.2588 | 0.094354425 |
| 54 | L | 0.023 | 0.001187441 | 0.2539 | 0.092695814 |
| 55 | L | 0.023 | 0.000650476 | 0.2483 | 0.090650159 |
| 56 | E | 0.023 | 0.000615434 | 0.2328 | 0.085471811 |
| 57 | W | 0.023 | 0.001080571 | 0.2302 | 0.08476019 |
| 58 | L | 0.023 | 0.000941573 | 0.2154 | 0.079780524 |
| 59 | A | 0.023 | 0.001573079 | 0.208 | 0.07752436 |
| 60 | M | 0.024 | 0.000997698 | 0.2853 | 0.103432566 |
| 61 | A | 0.024 | 0.00227783 | 0.3026 | 0.109625943 |
| 62 | V | 0.025 | 0.003362786 | 0.3503 | 0.126220929 |

[a]Average values of the disorder probability for each position.

[b]Amino acid G placed at position 50, inside the stable region.

**Table 3.** Transmembrane prediction results obtained using TMHMM, MemBrainTHM, ProtScale, and TMpred applications. Results from ProtScale and TMpred have been normalized for comparison with other probabilities.

| Position | Amino acid sequence | TMHMM probability | MemBrain THM propensity | ProtScale normalized score | TMpred normalized score | Transmembrane probability, average value[a] |
|---|---|---|---|---|---|---|
| 1 | M | 0 | 0.000191 | N/A[b] | 0.661425764 | 0.220538921 |
| 2 | L | 0 | 0.002851 | N/A[b] | 0.661425764 | 0.221425588 |
| 3 | Q | 0 | 0.046538 | N/A[b] | 0.661425764 | 0.235987921 |
| … | | | | | | |
| 49 | V | 0.2594 | 0.987914 | 0.646 | 0.603358942 | 0.624168236 |
| 50[c] | G | 0.27719 | 0.987914 | 0.646 | 0.629801679 | 0.63522642 |
| 51 | E | 0.28083 | 0.991702 | 0.736 | 0.660532428 | 0.667266107 |
| 52 | L | 0.32735 | 0.993857 | 0.67 | 0.594246918 | 0.646363479 |
| 53 | L | 0.56651 | 0.993857 | 0.637 | 0.778452743 | 0.743954936 |
| 54 | L | 0.63937 | 0.994522 | 0.632 | 0.73360729 | 0.749874822 |
| 55 | L | 0.64032 | 0.990459 | 0.659 | 0.818831517 | 0.777152629 |
| 56 | E | 0.64052 | 0.96027 | 0.726 | 0.835626228 | 0.790604057 |
| 57 | W | 0.64826 | 0.946819 | 0.701 | 0.822583527 | 0.779665632 |
| 58 | L | 0.6493 | 0.947424 | 0.706 | 0.895122387 | 0.799461597 |
| 59 | A | 0.64928 | 0.947424 | 0.683 | 0.905663748 | 0.796341937 |
| 60 | M | 0.64927 | 0.970735 | 0.683 | 0.947293193 | 0.812574548 |
| 61 | A | 0.64924 | 0.970735 | 0.773 | 0.955511881 | 0.83712172 |
| 62 | V | 0.64903 | 0.937507 | 0.831 | 1 | 0.85438425 |
| 63 | M | 0.64893 | 0.892506 | 0.831 | 0.960871896 | 0.833326974 |
| 64 | L | 0.6482 | 0.846403 | 0.84 | 0.942826514 | 0.819357379 |
| 65 | L | 0.64758 | 0.781733 | 0.847 | 0.924066464 | 0.800094866 |
| 66 | L | 0.63557 | 0.670387 | 0.856 | 0.661425764 | 0.705845691 |
| 67 | L | 0.61835 | 0.539353 | 0.851 | 0.661425764 | 0.667532191 |
| 68 | C | 0.5428 | 0.455615 | 0.819 | 0.661425764 | 0.619710191 |
| 69 | C | 0.51009 | 0.430385 | 0.728 | 0.661425764 | 0.582475191 |
| 70 | C | 0.44702 | 0.380525 | N/A[b] | 0.661425764 | 0.496323588 |

[a]Average values of the probability for each position.

[b]The window size used for the profile computation is 9, so the score is not applicable for positions 1-4 and 70-73.

[c]Amino acid G placed at position 50, inside the stable region.

**Figure 6.** Topology diagram using the MemBrain v3. The illustration depicts the transmembrane topology of the sequence and highlights that the amino acid at position 50 (G) is positioned into the middle of the transmembrane region. Red: transmembrane helix (TMH); blue: loop.

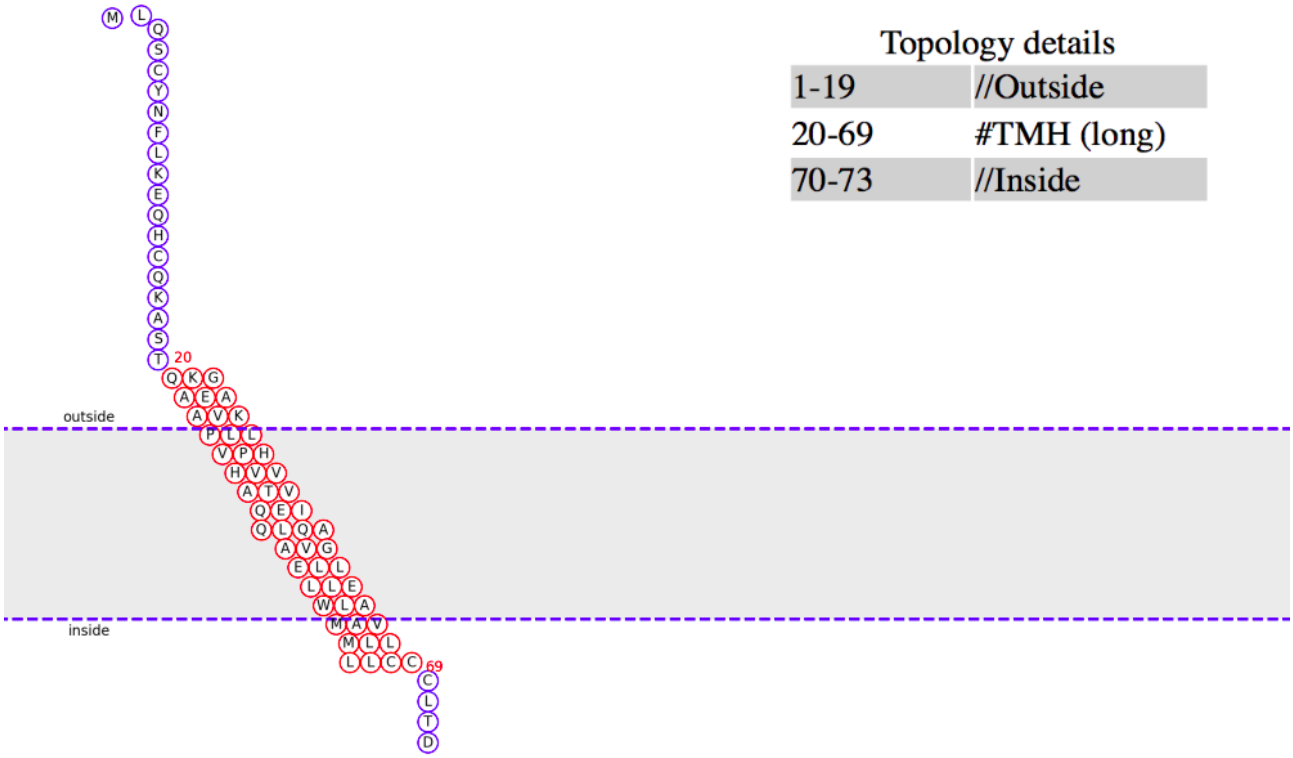
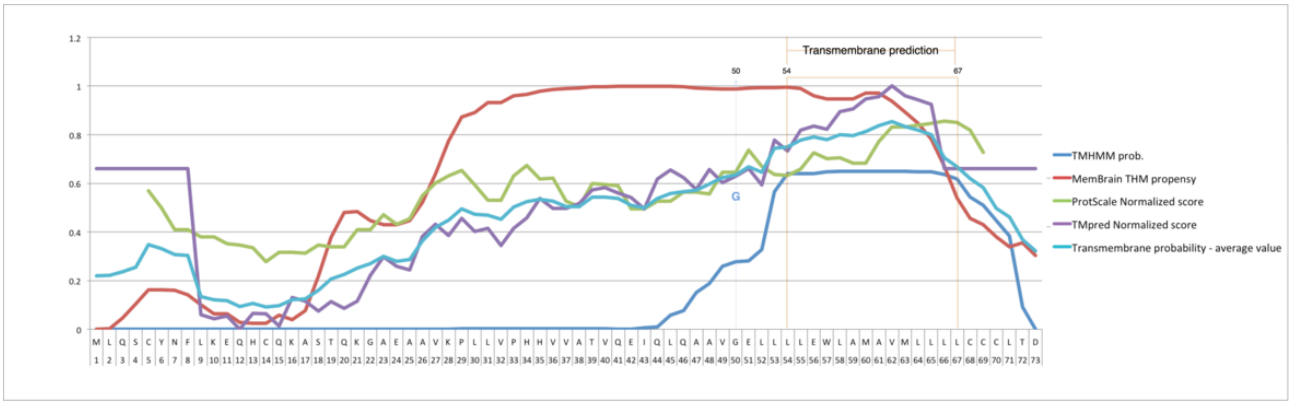



**Figure 7.** Transmembrane prediction. The region 54-67 was found to be the region with the highest probability to code for the transmembrane, and the G amino acid is near enough to suppose its involvement. The orange lines delimit this region, and the blue dotted line outlines the position of G on the different curves.




## 3D Protein Analysis

To characterize the deduced protein P0DTD3.1, we predicted the 3D structures for both the wild-type and mutated protein sequences using an ab initio approach. According to the preliminary clue from the secondary structure prediction, the mutated protein presents a slightly different structure when the amino acid residue changed from G to T. Figures 8 and 9 illustrate both the predicted models showing that the mutation would affect the tertiary structure of the protein. The comparison of residues 45-55 between MUT31136 and MOD30336 showed that this portion of the protein with the mutation stretches out with repercussions to the preceding helix. This result suggests that the mutation of the single amino acid from G to T, with consecutive stretching cycles on the 3D structure of the protein, tends to make the protein assume new functions.

**Figure 8.** Prediction of the 3D structure for the mutated protein of SARS-CoV-2. The model MUT31136 represents the predicted 3D model of the protein subject to mutation. (A) Amino acid sequence colored by the spectrum range, with the mutated amino acid indicated in black color at position 50 (T). (B) The protein has been oriented to facilitate the comparison and residue 50 is represented with red dots. (C) Details of the residues 45-55 and their rotation (D) around the Y-axis and (E) around the X-axis with a step of 90˚.

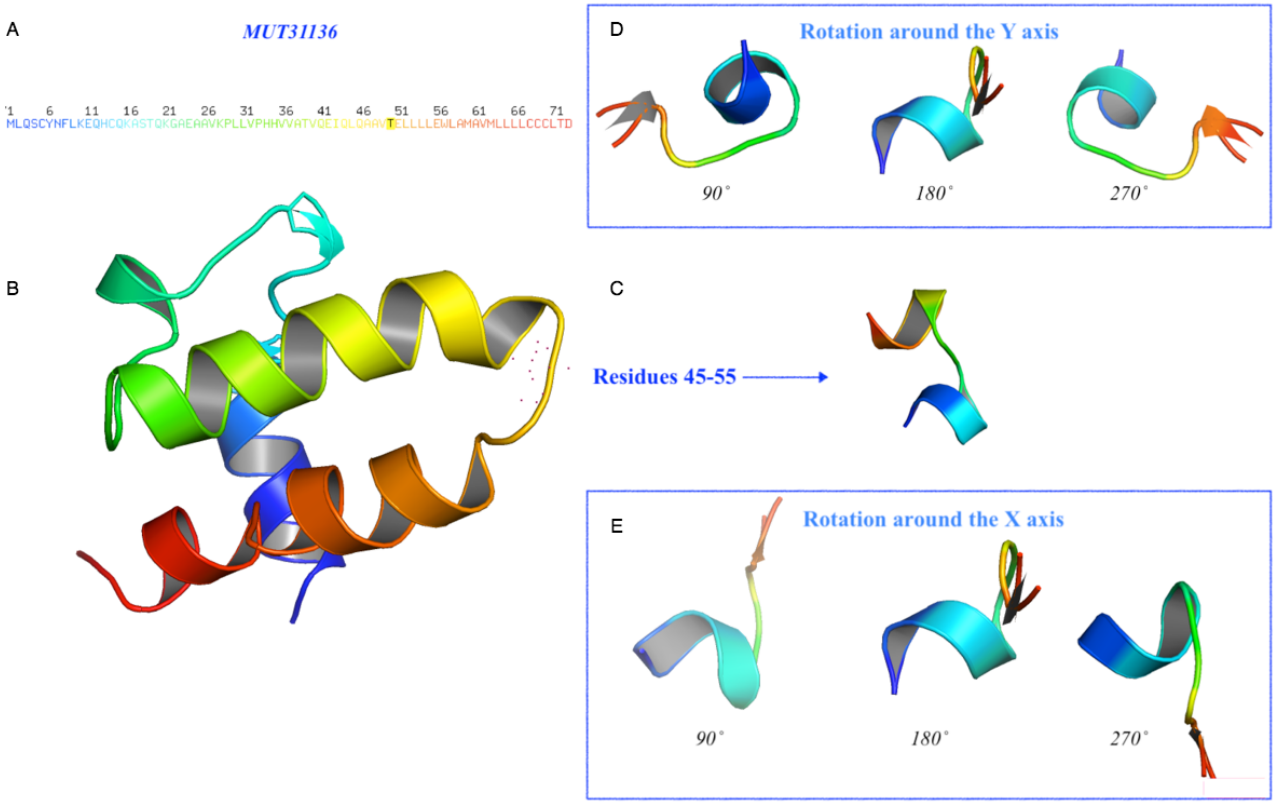
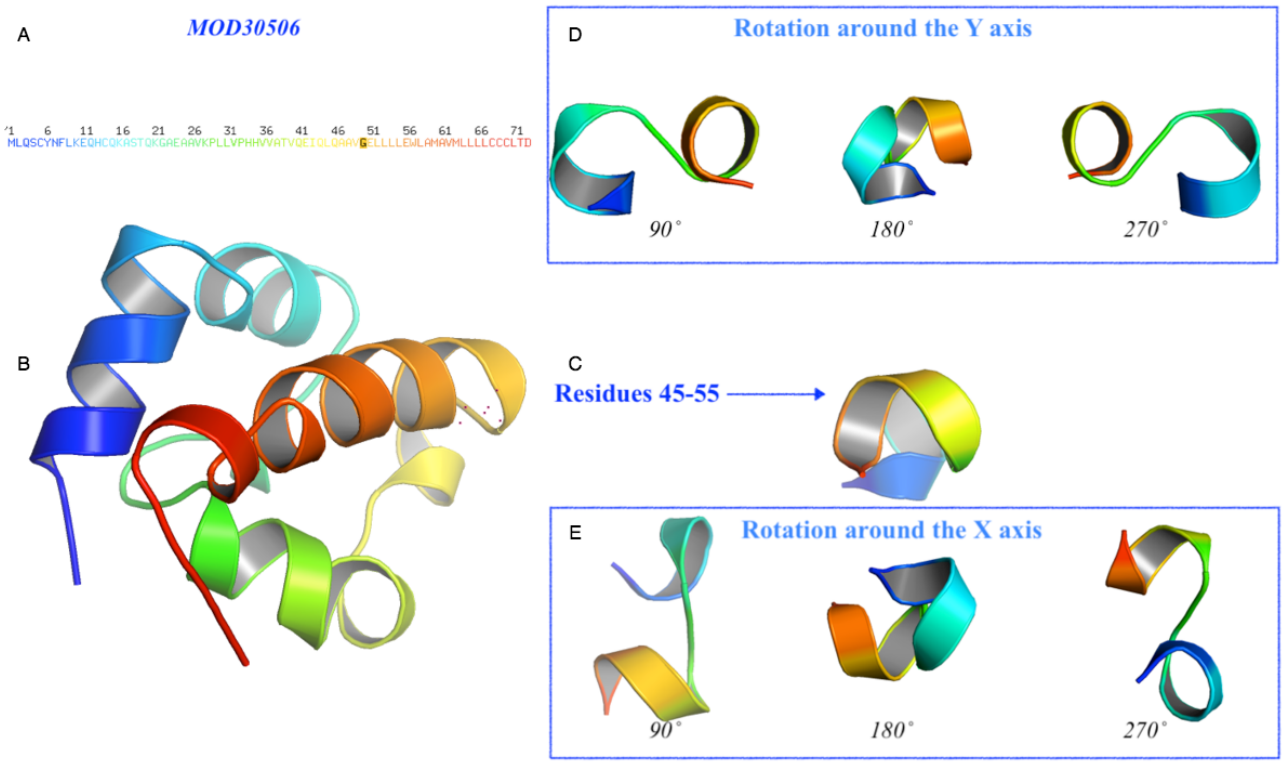



**Figure 9.** Prediction of the 3D structure of the unchanged protein. The model MOD30506 represents the predicted 3D model of the wild-type protein. (A) Amino acid sequence colored by the spectrum range, with the investigated amino acid indicated in black color at position 50 (G). (B) The protein has been oriented to facilitate the comparison and the residue 50 is represented with the red dots. (C) Details of the residues 45-55 and their rotation (D) around the Y-axis and (E) around the X-axis with a step of 90˚.



### Prediction of Protein-Related Ontologies

The analysis of protein ontologies indicates different functions between the wild-type and mutated proteins, owing to their changed structures. As shown in Table 4, the wild-type variant of the protein is linked with a high probability (.978≤*P*≤1) to both catalytic and transferase activities. The mutated variant of the protein presents a remarkable change in its functionality trend: even if usually the scores below 0.5 are interpreted as negative predictions, in an evolutionary context, the decrease in probability of the transferase activity (from 0.98 to 0.375) to favor the binding function (from 0.004 to 0.132) is not regarded as negligible. The contextual inversion of tendency of transferase to binding activity suggests that the episodic evolutionary mutation aims to improve the binding ability of the protein.

**Table 4.** Classification report showing the predicted functions of both (mutated and wild-type) protein sequences and related scores. Only positive scores are reported.

| Gene ontology terms and function | | Score | |
|---|---|---|---|
| | | Wild-type protein sequence | Mutated protein sequence |
| GO:0003674 | Molecular function | 1 | 1 |
| GO:0003824 | Catalytic function | 1 | 0.998 |
| GO:0016740[a] | Transferase activity | 0.978 | 0.375 |
| GO:0016829 | Lyase activity | 0.017 | —[b] |
| GO:0022891 | Transmembrane | 0.07 | —[b] |
| GO:0005488[a] | Binding activity | 0.004 | 0.132 |
| GO:0022892 | Transmembrane transport activity | 0.001 | 0.001 |

[a]Ontological functions subjected to inverted tendency.

[b]Unpredicted function.

## Discussion

### Principal Findings

SAR-CoV-2, the virus known to cause the COVID-19 pandemic, has many peculiar characteristics, such as rapidly accumulating mutations, compared to other coronaviruses [25]. Specifically, the prevalence of single nucleotide transitions as the major mutational type of SAR-CoV-2 across the world has been shown previously [7]. In this study, we conducted evolutionary analyses on the mutations to determine whether SARS-CoV-2 genomes from different countries in the same meridian might have specific variation patterns. We found that codon 9628 was under episodic selective pressure for all 4 countries in the same meridian. This would affect RNA secondary structure and, consequently, the protein product, with T (threonine) changing to G (glycine) in protein position 50. This position is located close to the predicted transmembrane region. Mutation analysis revealed that a change from G (glycine) to D (aspartic acid) may confer a new function to the protein, that is, binding activity, which in turn may be responsible for attaching the virus to human eukaryotic cells. These bioinformatics findings may help in better designing in vitro (wet lab) and in vivo (animal model) experiments to determine protein variants associated with the virulence of the virus. Therefore, these findings may eventually facilitate vaccine design and successful antiviral strategies. For example, the results of this study suggest the need for site-directed mutagenesis and animal experiments to validate the anticipated effects.

Mercatelli and Georgi [7] demonstrated that clade G, prevalent in Europe, carries a D614G mutation in the spike protein, which is responsible for the initial interaction of the virus with the host human cell. Other studies have also shown different mutation locations among strains isolated from different continents. Mutations at positions 2891, 3036, 14408, 23403, and 28881 are predominantly observed in European strains, whereas those located at positions 17746, 17857, and 18060 are exclusively present in North American strains of SARS-CoV-2 [26]. Their findings suggest that the virus is evolving and that European, North American, and Asian strains of the virus might coexist, with each characterized by different mutation patterns.

Furthermore, a comparison of viral genomes of SARS-CoV-2 strains from 13 countries identified differences in the protein-coding sequences. For example, an Indian strain showed a mutation in the spike glycoprotein at R408I and in the replicase polyprotein at I671T, P2144S, and A2798V, whereas the spike protein of Spain and South Korean strains carried an F797C and a S221W mutation, respectively [27]. Moreover, recently conducted integrative analyses of SARS-CoV-2 genomes of strains from different geographical locations reveal unique features that are potentially consequential to host-virus interaction and pathogenesis [28]. However, the most recent study of genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes indicates that unlike the influenza virus or HIV, SARS-CoV-2 has a low mutation rate, which makes the development of an effective global vaccine very likely [29]. The study determined several hotspot mutations across the whole SARS-CoV-2 genome. In all, 14 nonsynonymous hotspot mutations (whose prevalence of mutations is >10%) have been identified at different locations along the viral genome: 8 in ORF1ab polyprotein (in nsp2, nsp3, transmembrane domain, RdRp, helicase, exonuclease, and endoribonuclease), 3 in nucleocapsid protein, and 1 in each of the 3 proteins spike, ORF3a, and ORF8. Moreover, 36 nonsynonymous mutations

were identified in the receptor-binding domain of the spike protein with a low prevalence (<1%) across all genomes [29].

## Conclusions

All these findings highlight the importance of studying the relationship of geographical locations of SARS-CoV-2 isolates and mutations in their genomes, because the relationship can also be confirmed by phylogenetic tree analyses for elucidation of lineages and clusters based on the geographic locations. In conclusion, this genome evolutionary analysis revealed that codon 9628 is under episodic selective pressure for SARS-CoV-2 strains isolated from all 4 countries (Italy, Sweden, Congo, and Brazil) of the same geographical meridian.

## Acknowledgments

This work was supported by grants of Natural National Science Foundation of China (NSFC81671980, 81871623, 82020108022, Shu-Lin Liu). The funding bodies played no roles in the design of the study; collection, analysis, or interpretation of data; or in writing the manuscript.

## Conflicts of Interest

None declared.

## References

1. Bar-Zeev N, Inglesby T. COVID-19 vaccines: early success and remaining challenges. The Lancet 2020 Sep;396(10255):868-869. [doi: 10.1016/s0140-6736(20)31867-5]
2. Logunov DY, Dolzhikova IV, Zubkova OV, Tukhvatulin AI, Shcheblyakov DV, Dzharullaeva AS, et al. Safety and immunogenicity of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine in two formulations: two open, non-randomised phase 1/2 studies from Russia. The Lancet 2020 Sep;396(10255):887-897. [doi: 10.1016/s0140-6736(20)31866-3]
3. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis 2020 Oct 01;1866(10):165878 [FREE Full text] [doi: 10.1016/j.bbadis.2020.165878] [Medline: 32544429]
4. Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. Codon usage and phenotypic divergences of SARS-CoV-2 genes. Viruses 2020 Apr 30;12(5) [FREE Full text] [doi: 10.3390/v12050498] [Medline: 32366025]
5. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep 2020 Jun;19:100682 [FREE Full text] [doi: 10.1016/j.genrep.2020.100682] [Medline: 32300673]
6. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci U S A 2020 Jun 30;117(26):15193-15199 [FREE Full text] [doi: 10.1073/pnas.2008176117] [Medline: 32522874]
7. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 2020;11:1800 [FREE Full text] [doi: 10.3389/fmicb.2020.01800] [Medline: 32793182]
8. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A 2020 Apr 28;117(17):9241-9243 [FREE Full text] [doi: 10.1073/pnas.2004999117] [Medline: 32269081]
9. Islam MR, Hoque MN, Rahman MS, Alam ASMRU, Akther M, Puspo JA, et al. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci Rep 2020 Aug 19;10(1):14004 [FREE Full text] [doi: 10.1038/s41598-020-70812-6] [Medline: 32814791]
10. Libin PJK, Deforche K, Abecasis AB, Theys K. VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics 2019 May 15;35(10):1763-1765 [FREE Full text] [doi: 10.1093/bioinformatics/bty851] [Medline: 30295730]
11. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 2014 Nov 15;30(22):3276-3278 [FREE Full text] [doi: 10.1093/bioinformatics/btu531] [Medline: 25095880]
12. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 2012;8(7):e1002764 [FREE Full text] [doi: 10.1371/journal.pgen.1002764] [Medline: 22807683]
13. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011 Nov 24;6:26 [FREE Full text] [doi: 10.1186/1748-7188-6-26] [Medline: 22115189]
14. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 2015 Oct 15;31(20):3377-3379 [FREE Full text] [doi: 10.1093/bioinformatics/btv372] [Medline: 26099263]
15. Mizianty MJ, Uversky V, Kurgan L. Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol Biol 2014;1137:147-162. [doi: 10.1007/978-1-4939-0366-5_11] [Medline: 24573480]
16. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins 2019 Jun 09;87(6):520-527. [doi: 10.1002/prot.25674] [Medline: 30785653]

17. Hanson J, Paliwal KK, Litfin T, Zhou Y. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. Genomics Proteomics Bioinformatics 2019 Dec;17(6):645-656 [FREE Full text] [doi: 10.1016/j.gpb.2019.01.004] [Medline: 32173600]
18. Yin X, Yang J, Xiao F, Yang Y, Shen H. MemBrain: an easy-to-use online webserver for transmembrane protein structure prediction. Nanomicro Lett 2018;10(1):2 [FREE Full text] [doi: 10.1007/s40820-017-0156-2] [Medline: 30393651]
19. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. In: Link AJ, editor. 2-D Proteome Analysis Protocols. Methods in Molecular Biology vol. 112. Totowa, NJ: Humana Press; 1999:531-552.
20. Hofmann K, Stoffel W. TMbase-a database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler, 374. Biol. Chem. Hoppe-Seyler 1993;374:166 [FREE Full text]
21. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003 Jul 01;31(13):3784-3788 [FREE Full text] [doi: 10.1093/nar/gkg563] [Medline: 12824418]
22. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004 Jul 01;32(Web Server issue):W526-W531 [FREE Full text] [doi: 10.1093/nar/gkh468] [Medline: 15215442]
23. Upmeier zu Belzen J, Bürgel T, Holderbach S, Bubeck F, Adam L, Gandor C, et al. Leveraging implicit knowledge in neural networks for functional dissection and engineering of proteins. Nat Mach Intell 2019 May 13;1(5):225-235. [doi: 10.1038/s42256-019-0049-9]
24. Rigsby RE, Parker AB. Using the PyMOL application to reinforce visual understanding of protein structure. Biochem Mol Biol Educ 2016 Sep 10;44(5):433-437 [FREE Full text] [doi: 10.1002/bmb.20966] [Medline: 27241834]
25. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 2004 Jun 28;4:21 [FREE Full text] [doi: 10.1186/1471-2148-4-21] [Medline: 15222897]
26. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020 Apr 22;18(1):179 [FREE Full text] [doi: 10.1186/s12967-020-02344-6] [Medline: 32321524]
27. Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk A, Song YG, et al. Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. PLoS One 2020;15(9):e0238344 [FREE Full text] [doi: 10.1371/journal.pone.0238344] [Medline: 32881907]
28. Sardar R, Satish D, Birla S, Gupta D. Integrative analyses of SARS-CoV-2 genomes from different geographical locations reveal unique features potentially consequential to host-virus interaction, pathogenesis and clues for novel therapies. Heliyon 2020 Sep;6(9):e04658 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e04658] [Medline: 32844125]
29. Alouane T, Laamarti M, Essabbar A, Hakmi M, Bouricha EM, Chemao-Elfihri MW, et al. Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: moving toward a universal vaccine for the. Pathogens 2020 Oct 10;9(10) [FREE Full text] [doi: 10.3390/pathogens9100829] [Medline: 33050463]


## Abbreviations

**GISaid:** Global Initiative on Sharing All Influenza Data
**MEME:** mixed effects model of evolution
**MERS-CoV:** Middle East respiratory syndrome coronavirus
**ORF:** open reading frames


*Edited by G Eysenbach; submitted 23.11.20; peer-reviewed by F Pappalardo, S Motta; comments to author 14.12.20; revised version received 30.12.20; accepted 13.01.21; published 22.01.21.*

*Please cite as:*
*Mastriani E, Rakov AV, Liu SL*
*Isolating SARS-CoV-2 Strains From Countries in the Same Meridian: Genome Evolutionary Analysis*
*JMIR Bioinformatics Biotechnol 2021;2(1):e25995*
*URL: http://bioinform.jmir.org/2021/1/e25995/*
*doi:10.2196/25995*
*PMID:33497425*

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org