Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Biophys J.* Author manuscript; available in PMC 2026 October 09.

Published in final edited form as:

*Biophys J.* 2026 November 17; 125(22): 6495–6507. doi:10.1016/j.bpj.2026.10.001.

# VARIANT: Web Server for Decoding and Analyzing Viral Mutations at Genome and Protein Levels

**Rui Wang**[1], **Xuhang Dai**[2], **Xin Cao**[1], **Changchuan Yin**[3,*], **Tamar Schlick**[1,2,4,5,*], **Guo-Wei Wei**[6,7,*]

[1]Simons Center for Computational Physical Chemistry, New York University, New York, NY 10003, USA

[2]Department of Chemistry, New York University, New York, NY 10003, USA

[3]The Yin Research LLC, Hoffman Estates, IL 60192, USA

[4]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

[5]New York University-East China Normal University Center for Computational Chemistry, New York University Shanghai, Shanghai 200122, China

[6]Department of Mathematics, University of Georgia, Athens, GA 30602, USA

[7]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA

## Abstract

A comprehensive analysis of viral mutations is essential for understanding viral evolution, disease epidemiology, diagnosis, drug resistance, and immune escape. However, challenges remain in capturing complex mutation patterns and supporting diverse viral families with varying genome architectures. To address these challenges, we present VARIANT, an openly accessible web server for mutational analysis of RNA viral genomes and associated viral products across both single- and multi-segment virus genomes. The server takes as input a viral reference genome, a reference protein sequence, and/or multiple sequence alignment, and automatically provides full annotation of mutation types, including standard categories such as point mutations (missense, silent, and nonsense), insertions, deletions, and frameshift events in both coding and non-coding regions. In addition, VARIANT detects two biologically significant mutation patterns that are often overlooked by conventional software/packages: "*row mutations*", "*hot mutations*", and

*Correspondence: cyinbox@gmail.com, schlick@nyu.edu, guowei.wei@uga.edu.

AUTHOR CONTRIBUTIONS

Rui Wang (Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing), Xuhang Dai (Software, Validation, Writing - review & editing), Xin Cao (Software, Validation, Writing - review & editing), Changchuan Yin (Conceptualization, Methodology, Software, Validation, Supervision, Writing - review & editing), Tamar Schlick (Conceptualization, Methodology, Supervision, Writing - review & editing), Guo-Wei Wei (Conceptualization, Methodology, Supervision, Writing - review & editing)

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used ChatGPT (GPT-5.5, OpenAI, San Francisco, CA, USA) and Claude Opus 4.8 (Anthropic, San Francisco, CA, USA) in order to improve language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

separately annotates potential *programmed ribosomal frameshifting* regions. The server currently contains automatic analysis of major viral pathogens, including SARS-CoV-2, HIV-1, Influenza H3N2, Ebola virus, and Chikungunya virus. It also allows users to analyze additional viruses by uploading custom genome information. Users can track VARIANT analysis progress in real time, visualize mutation distributions, and download structured results in ZIP format. VARIANT also incorporates dual graph topology analysis to classify frameshifting element structures from dot-bracket notation input. This feature enables systematic comparison of RNA secondary structure motifs across viral families by mapping structures to a comprehensive library of dual graph topologies. The web server is freely available at https://variant.up.railway.app/.

## INTRODUCTION

Rapid and accurate characterization of viral mutations is critical for understanding viral evolution, transmission dynamics, drug resistance, and immune escape mechanisms. For example, the delayed identification of emerging mutations hindered rapid public health responses during the COVID-19 pandemic, resulting in the wide spread of variants with increased transmissibility or immune evasion. Similar challenges arise in chronic viral infections such as HIV-1, where mutation-driven drug resistance continually impacts treatment effectiveness, and in rapidly evolving RNA viruses such as influenza, Ebola, and Chikungunya. Therefore, accessible and advanced computational tools to decode mutational landscapes in viral genomes and associated proteins are needed.

Several mutation analysis tools/web-servers have been developed, each addressing a portion of this need. CoVsurver, integrated within the GISAID platform, annotates mutations in SARS-CoV-2 genomes and marks notable variants. However, it requires GISAID account and is restricted to SARS-CoV-2. Nextclade, part of the Nextstrain system, aligns viral genomes to curated references and identifies nucleotide substitutions and small indels (1). Nextclade is widely used for SARS-CoV-2, influenza, and several other pathogens, but its output focuses primarily on per-site mutations and quality-control (QC) metrics. For instance, clusters of adjacent substitutions are typically flagged as potential sequencing errors (SNP-cluster QC) rather than interpreted as biologically meaningful mutation patterns. Besides web-based resources, variant callers such as V-Phaser 2 (2) and LoFreq (3) excel at identifying intrahost low-frequency variants from deep sequencing reads, yet they do not offer genome- or protein-level functional annotation. Consequence annotation tools such as SnpEff (4) and BCFtools/csq (5) annotate coding consequences from variant calls, and csq additionally uses phase so that adjacent variants sharing a codon are interpreted jointly (5). Both take a VCF together with gene annotations rather than an alignment of consensus genomes, in which co-occurring substitutions are in phase by construction, and both return per-variant consequence records rather than the pattern itself. In addition to mutation-annotation platforms, recent studies have explored geometric, alignment-free, and metric-based approaches for viral variant analysis. For example, Guan et al. constructed a SARS-CoV-2 genome space using natural-vector and k-mer natural-vector representations to compare and classify variants (6), while Yu and Yau proposed an optimal-natural-metric framework for the automatic detection of emerging virus variants (7). These methods provide useful global sequence-space representations for variant classification or detection.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Many issues persist in current computational approaches, although they have their own strengths. First, most platforms prioritize single-nucleotide variants and small indels, leaving limited support for the systematic annotation of coding consequences (missense, silent, and nonsense mutations), non-coding mutations, and frameshift-generating indels across genomic sequences. Second, biologically meaningful combinatorial mutation patterns, such as non-adjacent substitutions or consecutive substitutions, whose joint effect on the encoded residue differs from that of the individual changes, are interpreted jointly by haplotype-aware consequence callers (5) and are not reported as patterns by viral consensus-genome platforms, where clusters of adjacent substitutions are instead marked as sequencing artifacts. These patterns may indicate localized editing or compensatory evolutionary events. Third, many tools are specialized for a single virus (e.g., SARS-CoV-2) or require manual configuration to support new pathogens, limiting their utility for comparative or cross-virus analyses. Critically, existing PRF predictors (8–10) do not have systematic approaches to characterize RNA secondary-structure graph topology in functionally critical regions, such as programmed ribosomal frameshifting elements, which are essential regulatory features in many RNA viruses.

To address these limitations, we present VARIANT, a publicly accessible web server for comprehensive analysis of viral mutations at both the genome and protein levels. VARIANT supports both single-segment and multi-segment RNA viruses and provides automated annotation of all major mutation categories, including point mutations (missense, silent, nonsense), insertions, deletions, frameshift events, and mutations in non-coding regions. A unique feature of VARIANT is its ability to detect two biologically significant but frequently overlooked mutation patterns: row mutations and hot mutations. VARIANT further identifies genomic regions with signatures of programmed ribosomal frameshifting (PRF), which play essential roles in several RNA viruses such as HIV-1 (11) and coronaviruses (12). Additionally, VARIANT integrates dual graph topology analysis to classify RNA secondary structures of frameshifting elements, enabling systematic comparison of structural motifs across viral families.

The VARIANT server currently includes built-in support for major viral pathogens, including SARS-CoV-2, HIV-1, Influenza H3N2, Zaire Ebolavirus, and Chikungunya virus. In addition, VARIANT allows users to analyze other viral species by uploading custom reference genomes and pre-aligned multiple sequence alignments. Users can track processing jobs in real time, visualize mutation distributions, and download results in structured CSV or ZIP format. Together, these capabilities make VARIANT a unified and extensible platform for the viral genomics community, enabling mutation interpretation, evolutionary analysis, and downstream studies in vaccine design, antibody design, antigen design, antiviral drug development, antigenic analysis, and viral surveillance.

## MATERIALS AND METHODS

### The Overview of the VARIANT Workflow

VARIANT is a publicly accessible web server designed to provide comprehensive mutation analysis for viruses, integrating genome- and protein-level annotations, specialized pattern detection, and support for both single- and multi-segment viruses. The platform is developed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

based on previously published techniques (13, 14) and is implemented in Python with a modular architecture that separates user interaction, mutation processing, and result output into distinct components. It is implemented via Railway and uses the FastAPI framework to support asynchronous task execution, real-time progress tracking, and structured file management.

The typical VARIANT workflow (Figure 1) begins with three required input files from users: 1) a reference genome in FASTA format, 2) a corresponding reference proteome (amino acid sequence), and 3) a multiple sequence alignment (MSA) file of viral genomes aligned to the reference genome. Users can either select a pre-configured virus (e.g., SARS-CoV-2, HIV-1, Influenza H3N2, Chikungunya, Zaire ebolavirus) or upload custom virus reference files (including reference genome, reference proteome, and MSA) to analyze other viral species. Once inputs are provided, the server automatically performs nucleotide-level variant calling[1] across all aligned genomes in the MSA, maps mutations to their corresponding codons and protein positions, and classifies mutation types, including silent, missense, nonsense, insertion, deletion, frameshift, and non-coding region mutations. For pre-configured viruses, the corresponding reference genome and proteome files are already provided by the server. For custom-virus analysis, VARIANT requires a user-specified reference proteome to ensure accurate and reproducible genome-to-protein mutation mapping. Automatic proteome inference from genome sequence alone can be unreliable for viruses with complex genome architectures, including overlapping ORFs, polyproteins, alternative ORFs, programmed frameshift products, spliced gene products, and multiple genome segments.

In addition to standard mutation annotations, the server performs automated detection of three complex and biologically relevant mutation patterns: (1) row mutations, which are defined as two or more consecutive substitutions; (2) hot mutations, which are defined as two non-consecutive substitutions within a 3-nucleotide window; and (3) potential programmed ribosomal frameshifting (PRF) sites, which are identified by the presence of known slippery sequences and downstream RNA structure motifs.

Upon completion, VARIANT provides users with structured outputs, including genome mutation annotation files, aggregated protein-level summaries, row/hot mutation records, and bundled ZIP archives for streamlined data management. The web interface is fully automated and allows users to monitor the analysis pipeline in real time. Once users click the “Generate All Visualizations”, the interactive visualizations are provided to further interpret the results, such as genome-wide mutation barcode maps, mutation rate statistics on each protein, row/hot mutation spots, and potential PRF sites.

## Mutation Detection and Classification

Following multiple sequence alignment (MSA) input, VARIANT performs comprehensive genome-wide mutation detection by comparing each aligned genome sequence to the specified viral reference genome. This comparison identifies all nucleotide differences, which are categorized into standard (substitution, insertion, deletion) and non-standard

[1]The ‘variant calling’ means the identification of all sequence differences at the nucleotide level relative to the reference genome.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(row and hot mutation) mutation types. The mutation detection pipeline is applied to each genome in the alignment (excluding the reference) and supports both single-segment and multi-segment viral genomes.

## Standard Mutation Types on the Coding Region

Table 1 summarizes three standard mutation types observed in the coding regions of RNA viral genomes, along with their corresponding protein-level mutations. A *substitution* mutation, also known as a point mutation, occurs when a single nucleotide in a codon of the reference genome is replaced by a different nucleotide. If this altered codon encodes the same amino acid, the mutation is *silent* (or *synonymous*). If it encodes a different amino acid, the mutation is *missense*; and if it creates a premature stop codon, the mutation is *nonsense*.

*Insertion* mutations involve the addition of one or more nucleotides to the genome. When the number of inserted nucleotides is a multiple of three, the reading frame remains intact and the protein gains one or more amino acids (an in-frame insertion). However, if the number of inserted nucleotides is not divisible by three, the reading frame shifts result in a *frameshift* mutation that typically alters downstream codons and often produces a nonfunctional or truncated protein.

*Deletion* mutations represent the removal of one or more nucleotides from the genomic sequence. When the number of deleted nucleotides is a multiple of three, the reading frame remains intact, resulting in an in-frame deletion that eliminates one or more amino acids from the protein product. If the number of deleted nucleotides is not divisible by three, it disrupts the reading frame and causes a *frameshift deletion*.

## Non-standard Mutation Types on the Coding Region

A core innovation of VARIANT is its ability to move beyond standard mutation cataloging to detect and characterize the complex and biologically relevant mutation patterns that are often overlooked by existing mutation-parser software. VARIANT includes specialized approaches for detecting two mutation patterns summarized in Table 2: row mutations and hot mutations. These patterns are of particular interest in studies of viral evolution and adaptation, where clustered or coordinated nucleotide substitutions may signal mutational hotspots, (AID/APOBEC)-mediated deamination (15), or co-evolving residues under selective pressure.

A *row mutation* is defined as a set of two or more consecutive nucleotide substitutions. This pattern captures short runs of adjacent mutations that may arise from APOBEC/ADAR-mediated editing, or from compensatory substitutions that preserve structural or functional properties of the encoded protein. Unlike standard point mutation reporting, which treats such changes as independent events, VARIANT identifies and records row mutations as unified mutational blocks.

VARIANT defines *hot mutations* as two non-consecutive substitutions within a sliding 3-nucleotide window, where the middle nucleotide remains unchanged. This pattern is indicative of clustered but independent mutation events localized within a short region of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the genome. Hot mutations may reflect immune pressure, epitope targeting, or error-prone replication hotspots. VARIANT scans each genome sequence using an overlapping window approach and flags all qualifying hot mutation sites.

These two terms are not intended to represent established standard terminology, but are used here to describe two local substitution patterns detected by VARIANT. The purpose of the row/hot mutation module is not to replace single-point mutation annotation, but to preserve local co-occurrence information that would be lost if nearby substitutions were treated only as independent single-nucleotide changes. Multiple substitutions within the same codon, or within immediately adjacent codon positions, can jointly determine whether the resulting protein-level effect is synonymous, missense, or nonsense. This is the same observation that motivates haplotype-aware consequence calling (5). Biologically, row and hot mutations should be interpreted as exploratory indicators of localized clustered substitutions. Such patterns may arise from RNA editing, error-prone replication, compensatory evolution, immune or drug selection, or sequencing and sampling biases. Therefore, VARIANT reports these patterns as candidate local mutation patterns for downstream evolutionary, structural, or experimental analysis, rather than as direct evidence of functional impact.

Both row and hot mutations are reported in dedicated CSV files with detailed annotations including the affected genomic window, substituted bases, and flanking context. These records enable downstream analysis of mutational clustering patterns and their association with functional or structural genomic elements. By systematically detecting and annotating these patterns, VARIANT provides researchers with a higher-resolution view of mutational dynamics beyond isolated single-nucleotide variants.

### Mutations on the non-coding region

VARIANT tracks all mutation types listed in Table 1 and Table 2 within non-coding regions of viral genomes. Although such mutations do not directly alter amino acid sequences, they may have significant functional consequences, such as affecting RNA splicing or gene regulation. For example, substitutions or indels in untranslated regions (UTRs) can disrupt regulatory elements, influence transcript stability, or alter mRNA secondary structures, thereby modulating translation efficiency or RNA degradation pathways.

### Detection of Potential +1/−1 PRF Regions

Programmed ribosomal frameshifting (PRF) is a translational recoding mechanism that allows ribosomes to pause and shift reading frames during protein synthesis, resulting in the production of alternative protein products from a single mRNA transcript. PRF mechanisms include: −1 PRF and +1 PRF, depending on the direction of the frameshift. In −1 PRF, the ribosome shifts one nucleotide backward (in the 5’ direction), while in +1 PRF, it shifts one nucleotide forward (in the 3’ direction).

### −1 Programmed Ribosomal Frameshifting

Many viruses, including retroviruses, alphaviruses, and coronaviruses, exploit −1 PRF to synthesize essential fusion proteins and maintain proper gene expression ratios (11, 12, 16, 17). The canonical −1 PRF signal comprises three elements: (1) a slippery heptanucleotide

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sequence of the form X_XXY_YYZ (where the triplet XXX includes identical nucleotides, YYY is usually AAA or UUU, and Z is generally any base except G) (12, 17); a well-known example is UUUAAAC found in coronaviruses and UUUUUUA in HIV-1 (18, 19); (2) a short spacer (20) (~ 5–9 nucleotides) separating the slippery site from the downstream structure, which ensures proper geometric positioning of the ribosome during the pause; and (3) a stable downstream RNA secondary structure, often a stem-loop or pseudoknot, which stalls the ribosome and increases the probability of simultaneous slippage of the P-site and A-site tRNAs into the −1 frame. In this way, −1 PRF regulates the synthesis of essential fusion proteins such as Gag–Pol in HIV and the ORF1a/ORF1b polyproteins in coronaviruses.

### +1 Programmed Ribosomal Frameshifting

The +1 PRF mechanism is less common than −1 PRF, but functionally important in specific cellular and viral contexts (17). It is typically stimulated when the ribosome encounters an inefficiently decoded codon in the A-site (for example because of low cognate tRNA abundance or reliance on wobble base pairing), which promotes realignment of the tRNAs in the +1 frame. In many cases, the shift is coupled to an in-frame stop codon in the 0 frame, so that the frameshift allows the ribosome to bypass termination and continue elongation. Unlike −1 PRF, +1 PRF does not always require a strong downstream pseudoknot, although nearby secondary structures or context-specific sequences can contribute to ribosomal pausing.

Since +1 PRF is defined by codon context rather than by a slippery heptamer, VARIANT detects +1 candidates by a codon-pair scan rather than by motif search. Among the viruses supported by VARIANT, functional PRF is predominantly of the −1 type, so +1 predictions are reported as low-frequency candidates for further investigation rather than as confirmed events. The specific codon-context rules used to flag +1 candidates are described in the Supporting Information.

### Workflow for the Detection of −1 and +1 PRFs

The detection of −1 and +1 PRF follows a structured computational workflow that integrates sequence, structural, and translational features with two engines: 1) a structure-driven scanner handles −1 PRF, where the defining feature is a slippery heptamer followed by a downstream fold, and 2) a codon-context scanner handles +1 PRF, where the defining feature is an inefficiently decoded codon pair rather than a structured element. Each engine writes its own annotated output, and the two are combined into a single per-genome table as described in the following 5 stages:

**Stage 1: Slippery-site identification and windowing for −1 PRF.** The structure-driven scanner in `prf_scanner.py` first locates candidate slippery sites by encoding the three-identical-bases geometry of the X_XXY_YYZ grammar as a regular expression with backreferences. This detects any conforming site rather than relying only on a fixed list. The strict grammar requires XXX to be three identical nucleotides, Y to be A or U, and Z to be A, C, or U, and an optional relaxed grammar permits any Y or Z. Every position is evaluated independently, so overlapping sites are all retained. For each site, the scanner then

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

defines the region in which a stimulatory structure could form. The window is anchored at the 3’ end of the heptamer rather than centered on it, because the stimulatory structure lies downstream. For each spacer length from 5 to 9 nucleotides, the window begins after the spacer and extends up to 120 nucleotides toward the 3’ end. This design allows the most favorable spacing to be searched rather than assumed. Sites too close to the 3’ terminus are not reported if they cannot accommodate a sufficient foldable window. Users can also reduce the window size from 120 nt to a smaller number. This downstream foldable-window strategy follows earlier PRF scanners (8, 9).

**Stage 2: Structure prediction and consensus for −1 PRF.** Each candidate window is folded using four complementary tools, each packaged in its own module (Supporting Information Table B). RNAfold from the ViennaRNA package (21) computes a minimum-free-energy for each candidate window. These folds are used to identify stem-loops and rank spacer windows, so that the most stable window represents each site. Since RNAfold cannot represent pseudoknots, three pseudoknot-capable tools are applied to the leading 85 nucleotides of the representative window, namely PKNOTS (22), ProbKnot (23), and NUPACK (24). If all the pseudoknot-capable tools report crossing base pairs, the structure is classified as a high-confidence pseudoknot. If a majority of them report crossing base pairs, the structure is classified as a pseudoknot but with low confidence. Any disagreement among the tools is retained in the output rather than being hidden. The corresponding columns are left blank if the tools are not implemented. We caution that the fold space of frameshifting elements (FSEs) is complex and involves multiple sequence-length dependent conformations (19, 25–27). Therefore, these computational predictions should be treated as starting points for further investigation rather than definitive determinations. Since the pseudoknot-capable tools are computationally costly, this stage is available in the standalone command-line version of the VARIANT GitHub Package and can be run locally. The hosted web server performs only the sequence-level detection of candidate sites.

**Stage 3: Optional translational-context scoring for −1 PRF.** An optional tRNA module in `prf_scanner.py` adds a translational-context layer to each candidate from sequence. Based on a user-supplied host tRNA-abundance table, it returns three metrics: a decoding-supply score (the geometric mean of the tRNA-abundance values for the two codons that re-pair with the P-site and A-site tRNAs in the shifted frame), a limiting-codon pausing proxy from the scarcer of the two supplies, and a flag for codons containing both G and U. These metrics are interpretable prioritization features, not measured tRNA concentrations. They should therefore be recomputed with a validated host-specific table before quantitative interpretation.

**Stage 4: Codon-context scanning for +1 PRF.** In parallel, the codon-context scanner in `frameshift_detector.py` scans for +1 candidates. It reads consecutive P-site and A-site codon pairs in frame and applies the codon-context rules (see detailed information in the Supporting Information Table B), and it additionally re-checks the −1 slippery grammar through an independent implementation that cross-validates the structure-driven scanner. This step captures codon-driven recoding candidates that are outside the design scope of the −1 structure-based scanner.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Stage 5: Output integration.** The −1 scanner in `prf_scanner.py` writes its per-genome tables in CSV and BED format, and a driver script (`run_all_prf.py`) assembles all the outputs. Each row integrates the genomic coordinate, the frameshift type and mechanism, the slippery motif or codon pair, the spacer length, the predicted structures and their consensus, the tRNA metrics (optional), and a flag indicating whether it coincides with a known FSE. We emphasize that the pipeline reports all candidate PRF-like signals, so detection of a site does not guarantee that frameshifting occurs in vivo, and experimental validation remains essential.

## RNA Secondary Structure Topology Classification via RNA-As-Graphs

The efficiency of PRF is believed to depend on the secondary structure formed downstream of the slippery sequence, commonly referred to as the frameshifting element (FSE). The FSE acts as a structural barrier that impedes ribosomal translocation, thereby increasing the probability of frameshifting. Distinct FSE topologies, ranging from simple stem-loops to more complex pseudoknots, are associated with different frameshifting efficiencies (25), and recent surveys of viral RNAs in dual graph space have begun to catalog how these topologies are distributed across viral families (28).

To enable systematic comparison of FSE topologies across viral species and putative PRF candidates, VARIANT incorporates a dual-graph classification pipeline based on the RNA-As-Graphs (RAG) framework developed by the Schlick laboratory (See Existing-Dual-Search on GitHub) (29). In this framework, RNA secondary structures with pseudoknots are represented as two-dimensional dual graphs as follows: each double-stranded helical stem with at least 2 base pairs is encoded as a vertex; single-stranded regions connecting secondary elements such as bulges, loops, and junctions are encoded as edges; and the 5’ and 3’ termini are omitted from explicit representation. Because this graph-based abstraction captures the global fold architecture independently of sequence identity and helix length, it enables the identification of topologically equivalent RNA structures despite substantial sequence divergence. The reduced size of conformational space in such coarse-grained representation makes many analysis, classification (30), and design problems (31) more approachable. A key motivation for working in topology space is that different secondary-structure prediction programs frequently disagree on individual base pairs while describing the same overall shape. By mapping each prediction to its dual graph, the structures can be compared at the level of coarse-grained topology rather than detailed base-pairing, so that two folds sharing a dual graph are treated as topologically equivalent even though their base pairs differ. Such topology classifications have been very useful in analyzing viral frameshifting mechanisms and designing mutants (19, 26, 28, 31–33).

The classification in VARIANT accepts RNA secondary structures in dot-bracket notation as input. For users starting with experimentally determined 3D structures, secondary structures can be extracted from PDB or mmCIF files using the RNApdbee 3.0 web server (34), which annotates secondary structures of knotted and unknotted RNAs and outputs dot-bracket notation. Alternatively, users can directly input pre-computed dot-bracket notation from structure prediction tools (e.g., NUPACK, ProbKnot, PKNOTS).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The RAG dual graph classification progresses in three stages. First, RNA secondary structures in dot-bracket notation (with filename extension .dbn) are normalized to Dissecting the Spatial Structure of RNA (DSSR) records that preserve multi-strand information and pseudoknot annotations using RNApdbee 3.0. Second, these normalized records are converted to connectivity table (CT) format using the `dot2ct` utility from the RNAstructure package (35). Third, CT files are processed using `dualGraphs.py` from the Existing-Dual-Search repository to extract dual graph representations and perform topology matching against the dual graph library (`Dual_Library.txt` in Existing-Dual-Search). The Existing-Dual-Search repository was originally developed by the Schlick laboratory and integrated here into VARIANT for topology classification (36). In this step, dual-graph topologies are characterized using graph invariants, including eigenvalue spectra and degree-based connectivity features, and are then matched against a curated reference library via a two-step procedure consisting of spectral filtering followed by exact adjacency-matrix isomorphism testing. Structures that satisfy both criteria are assigned canonical dual-graph identifiers (graph IDs) and visualized using the matplotlib/plotly library.

On the VARIANT web server, users can submit RNA secondary structures in dot-bracket notation, and the platform returns the corresponding dual graph identifiers together with graphical representations of the assigned RNA topologies.

## RESULTS AND DISCUSSION

### Case Studies for Genome-Level and Protein-level Mutation Analysis

**Standard Mutation Visualization and Analysis:** VARIANT employs an interactive visualization framework powered by Plotly. As shown in Figure 2 a), the visualization tool presents a mutation barcode aligned with viral gene regions. Each mutation is represented as a vertical line positioned at its exact genomic coordinate on a horizontal gene schematic. 5 mutation types are distinguished by colors: red for missense, green for deletions, purple for insertions, blue for nonsense, and light blue for silent mutations. The *x*-axis indicates genomic position in base pairs, while gene identifiers are displayed along the *y*-axis. Interactive hover functionality enriches the interface. Users can access detailed metadata, including nucleotide changes, corresponding amino acid alterations, and mutation type. This panel enables intuitive identification of mutational hotspots and gene-specific mutation patterns, enabling researchers to pinpoint areas of potential functional importance at a glance. Figure 2 b) and c) are barplots for gene-level and protein-level mutation rates. The gene-level mutation rate is defined as the total number of nucleotide changes normalized by gene length, and the protein-level mutation rate is calculated as the number of total amino acid changes normalized by protein length (silent mutation is not counted). This analysis reveals the position of genomic regions/proteins that accumulate mutations at the highest rate, providing insights into evolutionary pressure and functional consequences. An example of a multi-segment virus, H3N2, is also provided in the Supporting Information Fig. A.

**Non-standard Mutation Visualization and Analysis:** The row mutation (purple) and hot mutation (red) patterns are overlaid as vertical lines at their corresponding genome positions, as illustrated in Figure 3. This analysis enables the detection of complex

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

mutational signatures that may have distinct biological implications beyond those of single-point mutations. For example, there are three consecutive mutations G28881A, G28882A, G28883C, which result in two amino acid mutations on nucleocapsid (R203K and G204R). This triplet of mutations has been identified in more than 60% of SARS-CoV-2 sequences and it creates a transcription regulatory sequence site that yields an additional subgenomic mRNA, N.iORF3, encoding a truncated nucleocapsid fragment that antagonizes type I interferon (37, 38). This example illustrates the motivation for reporting row mutations as local mutation blocks: three adjacent nucleotide substitutions are more informative when interpreted together than as three independent point mutations. An example of a multi-segment virus, H3N2, is also provided in the Supporting Information Fig. B.

**Programmed Ribosomal Frameshifting Detection Analysis:** To evaluate the −1 PRF detection pipeline against independent evidence, we apply it to viral genomes with experimentally documented −1 PRF (SARS-CoV-2 and HIV-1) and to genomes where a −1 PRF signal has been proposed but not experimentally confirmed (Zaire ebolavirus and Chikungunya virus). For the two well-characterized systems, we validate the predicted coordinates, slippery heptamers, and downstream structures against published FSE studies on these same viruses (18, 19, 25–27). For each virus we compare the recovered site with the literature below.

In SARS-CoV-2, VARIANT places the site at the ORF1a/ORF1b junction on the UUUAAAC heptamer at nucleotide 13462, within the published 84-nucleotide FSE spanning nucleotides 13462 to 13545 (25, 27). This is the frameshift that yields the ORF1ab polyprotein. For HIV-1, VARIANT identifies the canonical Gag-Pol junction on the UUUUUUA heptamer at nucleotide 1631. This element is best described as a dynamic ensemble of interconverting folds in which a stable stem-loop dominates while pseudoknotted conformations also appear (18, 27).

Ebola is the most informative test because a recent study reports two frameshifts in the L gene and directly assays one of them (39). VARIANT recovers the −1 signal on the UUUAAAA heptamer at nucleotide 13938, matching the proposed site. It also recovers the +1 signal that the study actually validated. A dual-reporter assay found this +1 site functional at about 10 percent efficiency, and our +1 detector flags it at nucleotide 14028, exactly 90 nucleotides downstream, on a CCU proline codon followed by a UAA stop.

Chikungunya is a prediction without clear experimental evidence. The candidates fall in the structural polyprotein region of the subgenomic RNA, where the characterized alphavirus −1 PRF sits on a UUUUUUA site in the 6K gene and produces the transframe protein (40), an element since modeled in detail by graph-based landscape and kinetic-mutation studies (26, 27). The location is plausible, but the specific site would need the same experimental test applied to the coronavirus and retrovirus elements before any functional claim.

Two main factors suggest these calls represent genuine signal rather than noise. First, the structural filter is selective. Of all the −1 PRF candidate sites detected in a genome, only a minority receive a pseudoknot call from the multi-tool consensus, namely 19 of 72 candidate sites in SARS-CoV-2 and 8 of 39 in HIV-1. Second, nested minimum free energy alone

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

would rank the true frameshifting element well down the list, since the SARS-CoV-2 FSE ranks only 15th of 72 by that measure, which is why the pipeline includes pseudoknot-aware prediction and a transparent multi-tool consensus function. We emphasize that this step flags candidate structured sites for further study rather than resolving the full conformational ensemble, which requires the landscape and kinetic modeling developed for these elements (26, 27). Sites without such support are carried as hypotheses.

## Dual Graph Topology Analysis of Viral Frameshifting Elements

As described above, VARIANT provides a platform to convert RNA secondary structures into their dual graph representations for viral FSEs. Given secondary structures in dot-bracket format obtained from structure prediction tools such as NUPACK, PKNOTS, or extracted from RNApdbee 3.0, VARIANT returns the corresponding dual graph identifier along with representative graph motifs from our updated dual graph library, as shown in the Figure 4.

The web interface streamlines the analysis workflow by automatically matching input structures against our comprehensive dual graph library. For example, when analyzing the SARS-CoV-2 frameshifting element (PDB ID: 7LYJ), VARIANT identifies its dual graph motif and provides the structural classification, enabling rapid comparison with other frameshifting elements. This topological representation reveals conserved architectural features that may be critical for frameshifting function.

We applied this analysis to a curated set of viral frameshifting elements from multiple virus families, including the Chikungunya virus frameshifting element (26), SARS-CoV-2 frameshifting pseudoknot RNA (25), and Rous sarcoma virus frameshifting pseudoknot RNA (41). Figure 5 presents the dual graph representations of representative frameshifting elements. Each panel shows the secondary structure (left) and its corresponding dual graph ID (right). Dot-bracket notations are obtained via the RNApdbee 3.0 website (34). The Chikungunya element was taken from (26). Notably, structures containing pseudoknots exhibited distinct dual graph topologies. Because pseudoknots contain intertwined, non-nested base pairs, they cannot be represented by tree graphs. Dual graphs, which encode stems as vertices and loops as edges and may contain cyclic paths, represent them explicitly (25, 29). This topological classification provides a quantitative framework for comparing frameshifting elements across different viral species and identifying structural determinants of frameshifting efficiency.

We also use dual graph topology analysis to compare how different folding programs describe the same frameshifting element. For the SARS-CoV-2 ORF1a/1b element (NC_045512.2, nt 13462–13550), RNAfold (32 bp) and ProbKnot (34 bp) give different base pairs but *the same* dual graph 3_5, whereas the pseudoknot-aware PKNOTS fold predicts 3_6 and the three-stem H-type pseudoknot topology established for this element by RAG analysis (25) and NUPACK returns a four-stem pseudoknotted graph 4_7. For the HIV-1 *gag–pol* element (NC_001802.1, nt 1638–1690), all four predictors disagree on base-pairing detail (17–19 bp, distinct lower stems) yet map to *the same* 2_2 two-helix motif, reproducing the topology reported for this element (18). In HIV-1 case, the base-pair-level disagreement among tools collapses to a single topological outcome, and the dual-graph

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

identifier directly reports pseudoknot status. In addition, in SARS-CoV-2 case, the base-pair-level disagreement in RNAfold and ProbKnot collapses to the same topological structure (See Supporting Information Figure C).

## DISCUSSION

### Further potential improvement on the PRF detection module

The current PRF detection module focuses on canonical slippery heptamers. Although this strategy provides interpretable candidate PRF/FSE annotations, it may miss frameshifting events driven by imperfect or non-canonical slippery motifs, unusual codon contexts, ribosomal pausing, RNA structural dynamics, or host-specific translational factors. Thus, VARIANT should not be viewed as an exhaustive probabilistic detector of all PRF mechanisms. Probabilistic detection of imperfect heptamers and broader translational-context modeling may be considered in future versions.

### CPU Model and Runtime

The execution time is evaluated using a representative SARS-CoV-2 test case. The benchmark used the SARS-CoV-2 reference genome NC_045512.2 and a query genome from GISAID, EPI_ISL_16327572. The test was performed on an Apple M4 Pro machine with a 12-core CPU, 48 GB RAM, macOS 26.5.1 (Darwin 25.5.0), and Python 3.8.20. VARIANT was executed in single-threaded mode without GPU acceleration or parallelization. Across five independent runs, the complete mutation-decoding pipeline required an average runtime of 0.216 s per query genome. The individual runtimes were 0.305, 0.191, 0.210, 0.185, and 0.190 s. This runtime includes nucleotide- and protein-level mutation annotation, insertion and deletion detection, and row and hot mutation detection. These results indicate that VARIANT can decode a complete SARS-CoV-2 genome in a fraction of a second on a standard desktop workstation.

## CONCLUSION

We developed VARIANT, a comprehensive web platform for analyzing mutations in RNA viral genomes. The server addresses key limitations in existing tools by providing unified support for diverse viral species, detecting biologically meaningful mutation patterns that are typically overlooked, and integrating structural analysis of programmed ribosomal frameshifting elements through “RNA-As-Graph” dual graph identification.

VARIANT automates the complete annotation pipeline from nucleotide changes to protein-level consequences, handling both standard mutations (substitutions, insertions, deletions) and complex patterns such as row mutations, hot mutations, and frameshift events. In addition, the integration of dual-graph topology analysis represents a novel feature in viral mutation analysis platforms. By classifying frameshifting element structures using the RNA-As-Graphs framework, VARIANT enables direct comparison of RNA secondary structure motifs across different viral lineages. This capability bridges sequence-level mutation analysis with structural characterization.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

VARIANT currently supports five major viral pathogens with pre-configured analysis pipelines, while remaining extensible to additional viruses through custom genome uploads. The platform's modular architecture, real-time progress tracking, and interactive visualizations make complex viral genomics analysis accessible to researchers.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

This work was supported in part by NIH Grants R01AI164266 and R35GM148196 and Georgia Research Alliance to G-W. W. Support to T.S. from the National Institutes of Health, National Institute of General Medical Sciences Award R35-GM122562, National Science Foundation Awards (DMS-2151777 and DMS-2330628) from the Division of Mathematical Sciences, and Philip-Morris USA Inc. is gratefully acknowledged. This work was also supported by a grant from the Simons Foundation through the NYU Simons Center for Computational Physical Chemistry Award MPS-T-MPS-00839534 to MET. X.D. acknowledges partial support from a graduate fellowship from the Simons Center for Computational Physical Chemistry (SCCPC) at NYU. R.W. also thanks Soumil Joshi at New York University for the proofreading help during the revision process.

## DATA AND SOFTWARE AVAILABILITY

The VARIANT web server is freely available at https://variant.up.railway.app/ and requires no user registration. The source code of VARIANT is available at https://github.com/wangru25/VARIANT. The original code of dual graph detection and visualization is available at https://github.com/Schlicklab/Existing-Dual-Search. We also plan to deploy/mirror the web server on university-supported academic infrastructure after the first author completes the institutional transitions in Fall 2026. The link to the mirrored web server will be posted on the VARIANT GitHub repository. In the meantime, we will continue to maintain the Railway deployment to ensure uninterrupted public access.

## REFERENCES

1. Aksamentov I, Roemer C, Hodcroft EB, and Neher RA, 2021. Nextclade: Clade Assignment, Mutation Calling and Quality Control for Viral Genomes. Journal of Open Source Software 6:3773.
2. Yang X, Charlebois P, Macalalad A, Henn MR, and Zody MC, 2013. V-Phaser 2: Variant Inference for Viral Populations. BMC Genomics 14:674.
3. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, and Nagarajan N, 2012. LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets. Nucleic Acids Research 40:11189–11201.
4. Cingolani P, 2022. Variant Annotation and Functional Prediction: SnpEff. Methods in Molecular Biology 2493:289–314.
5. Danecek P, and McCarthy SA, 2017. BCFtools/Csq: Haplotype-Aware Variant Consequences. Bioinformatics 33:2037–2039.
6. Guan M, Sun N, and Yau SS-T, 2024. Geometric Analysis of SARS-CoV-2 Variants. Gene 909:148291.
7. Yu H, and Yau SS-T, 2024. New Virus Variant Detection Based on the Optimal Natural Metric. Genes 15:891.
8. Moon S, Byun Y, Kim H-J, Jeong S, and Han K, 2004. Predicting Genes Expressed via −1 and +1 Frameshifts. Nucleic Acids Research 32:4884–4892.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


9. Theis C, Reeder J, and Giegerich R, 2008. KnotInFrame: Prediction of −1 Ribosomal Frameshift Events. Nucleic Acids Research 36:6013–6020.
10. McNair K, Salamon P, Edwards RA, and Segall AM, 2024. PRFect: A Tool to Predict Programmed Ribosomal Frameshifts in Prokaryotic and Viral Genomes. BMC Bioinformatics 25:82.
11. Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, and Varmus HE, 1988. Characterization of Ribosomal Frameshifting in HIV-1 Gag-Pol Expression. Nature 331:280–283.
12. Brierley I, Digard P, and Inglis SC, 1989. Characterization of an Efficient Coronavirus Ribosomal Frameshifting Signal: Requirement for an RNA Pseudoknot. Cell 57:537–547.
13. Yin C, 2020. Genotyping Coronavirus SARS-CoV-2: Methods and Implications. Genomics 112:3588–3596.
14. Wang R, Hozumi Y, Yin C, and Wei G-W, 2020. Mutations on COVID-19 diagnostic targets. Genomics 112:5204–5213.
15. Li Y, Zhu R, Jin J, Guo H, Zhang J, He Z, Liang T, and Guo L, 2024. Exploring the Role of Clustered Mutations in Carcinogenesis and Their Potential Clinical Implications in Cancer. International Journal of Molecular Sciences 25:6744.
16. Jacks T, and Varmus HE, 1985. Expression of the Rous Sarcoma Virus Pol Gene by Ribosomal Frameshifting. Science 230:1237–1242.
17. Atkins JF, Loughran G, Bhatt PR, Firth AE, and Baranov PV, 2016. Ribosomal Frameshifting and Transcriptional Slippage: From Genetic Steganography and Cryptography to Adventitious Use. Nucleic Acids Research 44:7007–7078.
18. Newton K, Yan S, and Schlick T, 2025. Conformational Analysis and Structure-Altering Mutations of the HIV-1 Frameshifting Element. International Journal of Molecular Sciences 26:6297.
19. Yan S, and Schlick T, 2025. Heterogeneous and Multiple Conformational Transition Pathways between Pseudoknots of the SARS-CoV-2 Frameshift Element. Proceedings of the National Academy of Sciences, USA 122:e2417479122.
20. Lin Z, Gilbert RJC, and Brierley I, 2012. Spacer-Length Dependence of Programmed −1 or −2 Ribosomal Frameshifting on a U 6 A Heptamer Supports a Role for Messenger RNA (mRNA) Tension in Frameshifting. Nucleic Acids Research 40:8674–8689.
21. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, and Hofacker IL, 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology 6:26.
22. Rivas E, and Eddy SR, 1999. A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots1. Journal of Molecular Biology 285:2053–2068.
23. Bellaousov S, and Mathews DH, 2010. ProbKnot: Fast Prediction of RNA Secondary Structure Including Pseudoknots. RNA 16:1870–1880.
24. Dirks RM, and Pierce NA, 2004. An Algorithm for Computing Nucleic Acid Base-Pairing Probabilities Including Pseudoknots. Journal of Computational Chemistry 25:1295–1304.
25. Yan S, Zhu Q, Jain S, and Schlick T, 2022. Length-Dependent Motions of SARS-CoV-2 Frameshifting RNA Pseudoknot and Alternative Conformations Suggest Avenues for Frameshifting Suppression. Nature Communications 13:4284.
26. Lee S, and Schlick T, 2026. Kinetic Traps in RNA Folding: Targeted Design of Frameshifting Element Mutants by Thermodynamic and Kinetic Analysis of the Chikungunya Virus. The Journal of Physical Chemistry B 130:1810–1831.
27. Portillo-Ledesma S, Lee S, Laederach A, and Schlick T, 2026. Open Questions on Viral Frameshifting: Exploiting the Structural Plasticity of the Frameshifting Element for Therapeutic Intervention. Biophysical Journal.
28. Zbib J, and Schlick T, 2026. Viral RNAs as Dual Graphs: Extending the Motif Universe of RNAs. RNA NanoMed In press.
29. Gan HH, 2003. Exploring the Repertoire of RNA Secondary Motifs Using Graph Theory; Implications for RNA Design. Nucleic Acids Research 31:2926–2943.
30. Wang R, and Schlick T, 2025. How Large Is the Universe of RNA-like Motifs? A Clustering Analysis of RNA Graph Motifs Using Topological Descriptors. PLOS Computational Biology 21:e1013230.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


31. Jain S, Tao Y, and Schlick T, 2020. Inverse Folding with RNA-As-Graphs Produces a Large Pool of Candidate Sequences with Target Topologies. Journal of Structural Biology 209:107438.
32. Schlick T, Zhu Q, Dey A, Jain S, Yan S, and Laederach A, 2021. To Knot or Not to Knot: Multiple Conformations of the SARS-CoV-2 Frameshifting RNA Element. Journal of the American Chemical Society 143:11404–11422.
33. Malekzadeh K, Bhendale M, and Zerze GH, 2026. Enhanced-Sampling Simulations Reveal Distinct Intermediates in SARS-CoV-2 FSE Pseudoknot Interconversion. Biophysical Journal 10.1016/j.bpj.2026.06.039.
34. Pielesiak J, Niznik K, Snioszek P, Wachowski G, Zurawski M, Antczak M, Szachniuk M, and Zok T, 2026. RNApdbee 3.0: A Unified Web Server for Comprehensive RNA Secondary Structure Annotation from 3D Coordinates. Journal of Molecular Biology 169795.
35. Mathews DH, 2006. RNA Secondary Structure Analysis Using RNAstructure. Current Protocols in Bioinformatics 0 12:Unit–12.6.
36. Zhu Q, Petingi L, and Schlick T, 2022. RNA-As-Graphs Motif Atlas—Dual Graph Library of RNA Modules and Viral Frameshifting-Element Applications. International Journal of Molecular Sciences 23:9249.
37. Mears HV, Young GR, Sanderson T, Harvey R, Barrett-Rodger J, Penn R, Cowton V, Furnon W, Lorenzo GD, Crawford M, Snell DM, Fowler AS, Chakrabarti AM, Hussain S, Gilbride C, Emmott E, Finsterbusch K, Luptak J, Peacock TP, Nicod J, Patel AH, Palmarini M, Wall E, Williams B, Gandhi S, Swanton C, and Bauer DLV, 2025. Emergence of SARS-CoV-2 Subgenomic RNAs That Enhance Viral Fitness and Immune Evasion. PLOS Biology 23:e3002982.
38. Mulloy RP, Evseev D, Sharlin N, Bui-Marinos MP, Lacasse É, Dubuc I, Flamand L, and Corcoran JA, 2026. Evolution of a Truncated Nucleocapsid Protein Enhances SARS-CoV-2 Fitness by Suppressing Antiviral Responses. PLOS Biology 24:e3003646.
39. Stubbs DB, Ruzicka JA, and Taylor EW, 2024. Modular Polymerase Synthesis and Internal Protein Domain Swapping via Dual Opposed Frameshifts in the Ebola Virus L Gene. Pathogens 13:829.
40. Firth AE, Chung BY, Fleeton MN, and Atkins JF, 2008. Discovery of Frameshifting in Alphavirus 6K Resolves a 20-Year Enigma. Virology Journal 5:108.
41. Jones J, Christopher P, and Ferré-D'Amaré AR, 2025. Structural Switching Dynamically Controls the Doubly Pseudoknotted Rous Sarcoma Virus–Programmed Ribosomal Frameshifting Element. Proceedings of the National Academy of Sciences, USA 122:e2418418122.
42. Kerpedjiev P, Hammer S, and Hofacker IL, 2015. Forna (Force-Directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams. Bioinformatics 31:3377–3379.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## SIGNIFICANCE

Understanding how viruses mutate is crucial for controlling pandemics, developing effective treatments, and designing vaccines. However, existing computational tools often cannot identify important mutation patterns or analyze the structural features that regulate viral protein production, such as frameshifting elements, through programmed ribosomal frameshifting. VARIANT fills these gaps by providing the first comprehensive web platform that simultaneously detects overlooked mutation patterns, annotates mutations across diverse viral genome architectures, and classifies the RNA structural topologies of frameshifting elements using dual graph approaches. By packing these functions into a web interface, VARIANT empowers researchers to perform sophisticated viral genomic analyses in an intuitive way. This web server can accelerate research in viral evolution, drug resistance mechanisms, and antiviral therapeutic development across multiple viral families.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

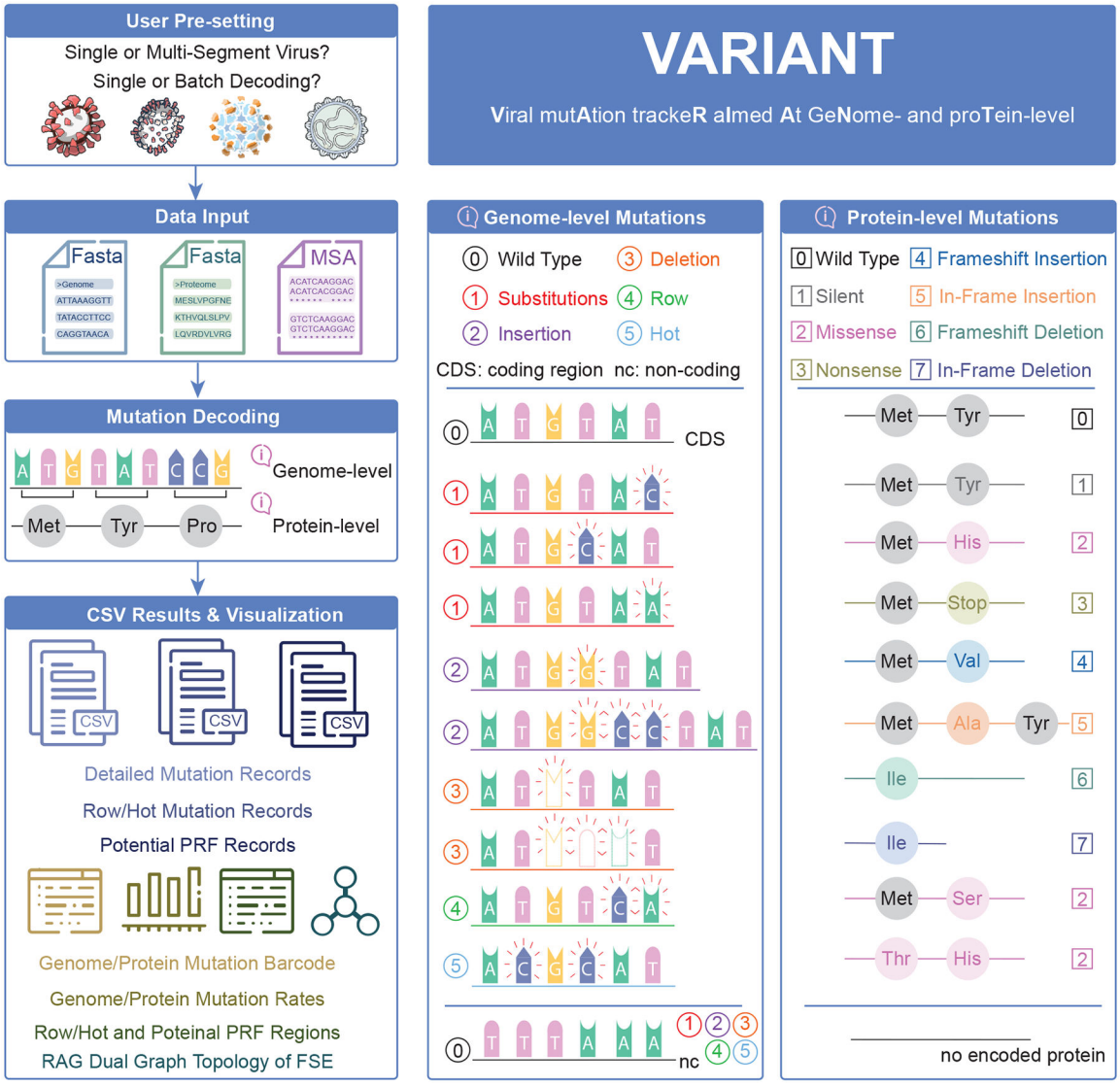


**Figure 1:**

Overview of the VARIANT web server for genome- and protein-level viral mutation analysis. The VARIANT web server supports both single- and multi-segment viral genomes and offers single or batch decoding modes. The workflow begins with user pre-configuration and input of reference genome and proteome sequences (FASTA Format) along with aligned viral genomes (MSA Format). The server performs mutation decoding at both the nucleotide (genome-level) and amino acid (protein-level) resolution, supporting standard mutation types such as substitutions, insertions, and deletions, as well as complex mutation patterns including row mutations, hot mutations, and potential programmed ribosomal frameshift (PRF) signals. Additionally, VARIANT incorporates dual graph topology analysis to classify RNA secondary structures of frameshifting elements using the “RNA-As-Graphs” framework. Output includes structured CSV files of detailed mutation records, row/hot mutation sites, and potential PRF regions. Visualizations of mutation distributions and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

classification summaries are also provided. Mutation classification is shown on the right, including genome-level categories and their corresponding protein-level consequences.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript



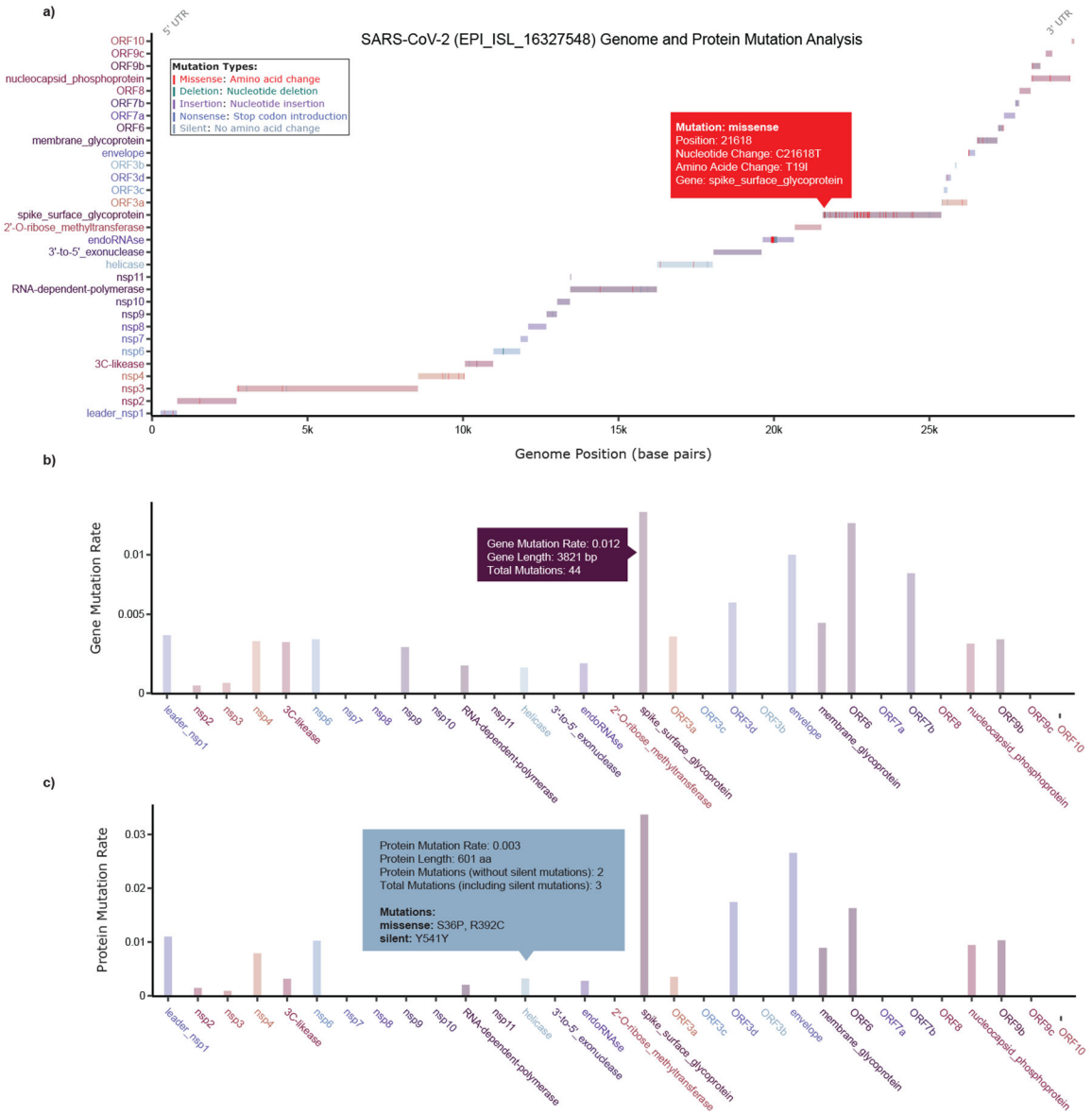


**Figure 2:**

Genome organization and distribution of mutations across SARS-CoV-2 viral genes. **a)** The top panel shows the viral genome as horizontal bars representing individual genes/proteins. Mutations are displayed as vertical lines at their respective genomic positions. Mutation types are represented by different colors (red: missense, green: deletion, purple: insertion, blue: nonsense, light-blue: silent). The *x*-axis indicates genomic position in base pairs, and the *y*-axis lists gene names. Hover interactions reveal detailed mutation information, including nucleotide changes, protein changes, and genomic coordinates. This web visualization enables rapid identification of mutation hotspots and gene-specific mutation patterns across the viral genome. **b)** Gene-level mutation rate, calculated as the total number of nucleotide changes normalized by gene length. Each bar corresponds to a gene and is colored consistently with the top panel. The *y*-axis shows mutation rate (mutations per base pair), and the *x*-axis lists gene names. Hover interactions display

additional details, including mutation counts and gene lengths. **c)** Protein-level mutation rate (silent mutations are not counted), calculated as the number of total amino acid changes normalized by protein length. Each bar represents a protein, with hover interactions revealing protein length and detailed mutation records.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

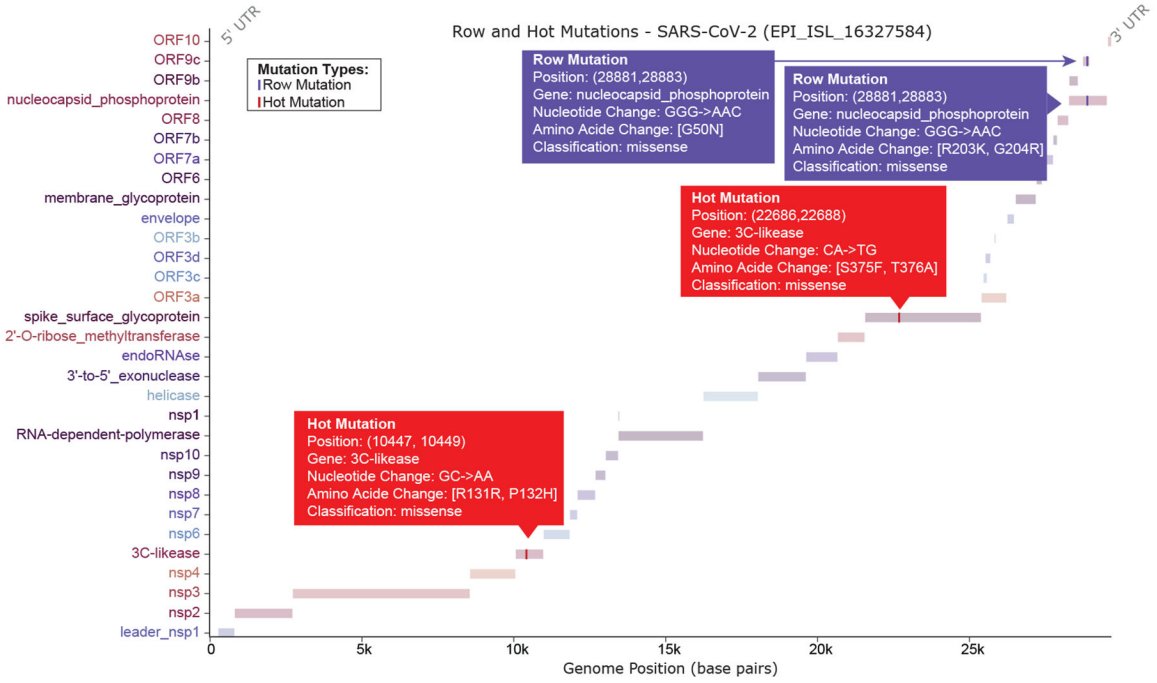


**Figure 3:**

SARS-CoV-2 Genome organization with specialized row and hot mutation pattern detection. The viral genome is displayed as horizontal bars, each representing a distinct gene (for a single-segment virus) or segment (for a multi-segment virus). Row mutation (purple color) and hot mutation (red color) patterns are overlaid as vertical lines at their corresponding genomic positions. The *x*-axis indicates genomic position in base pairs, while the *y*-axis lists gene names. Interactive hover features provide detailed information on each mutation, including mutation type, genomic position, nucleotide and amino acid changes, and biological classification.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

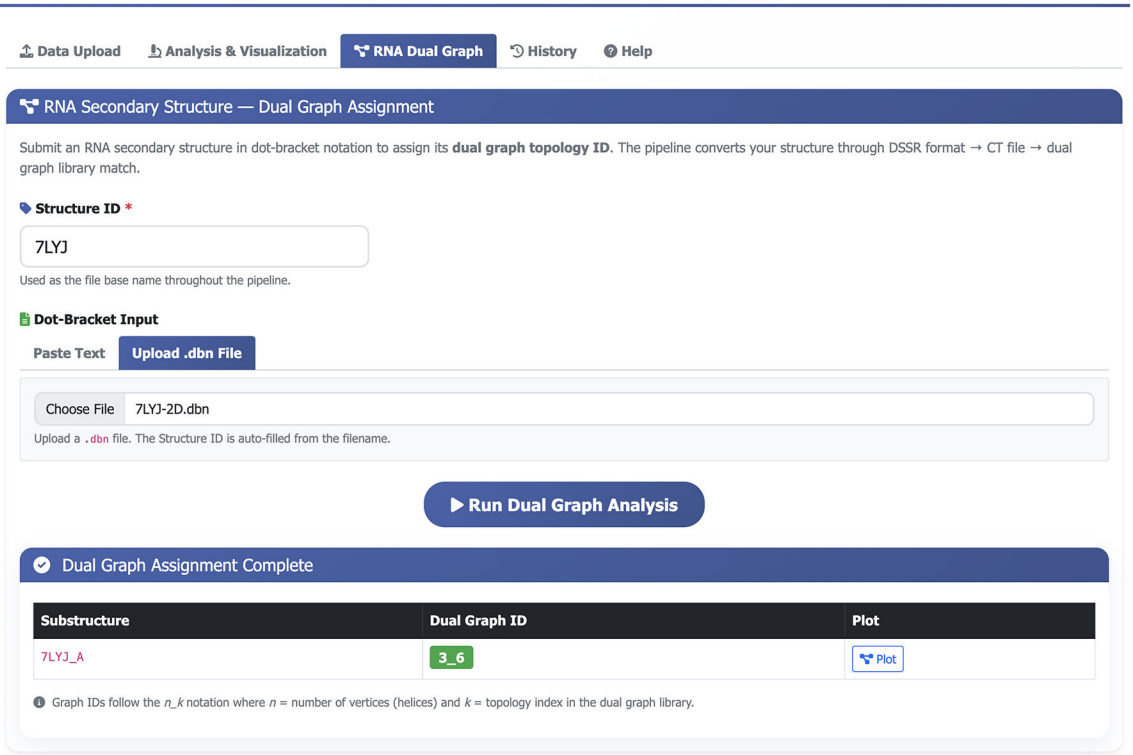


**Figure 4:**

VARIANT web interface for dual graph analysis. The panel accepts dot-bracket notation from structure prediction tools, and then presents the dual graph ID and its corresponding visual representation. The example shows the SARS-CoV-2 frameshifting element (PDB ID: 7LYJ), which exhibits a pseudoknot motif dual graph 3_6.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

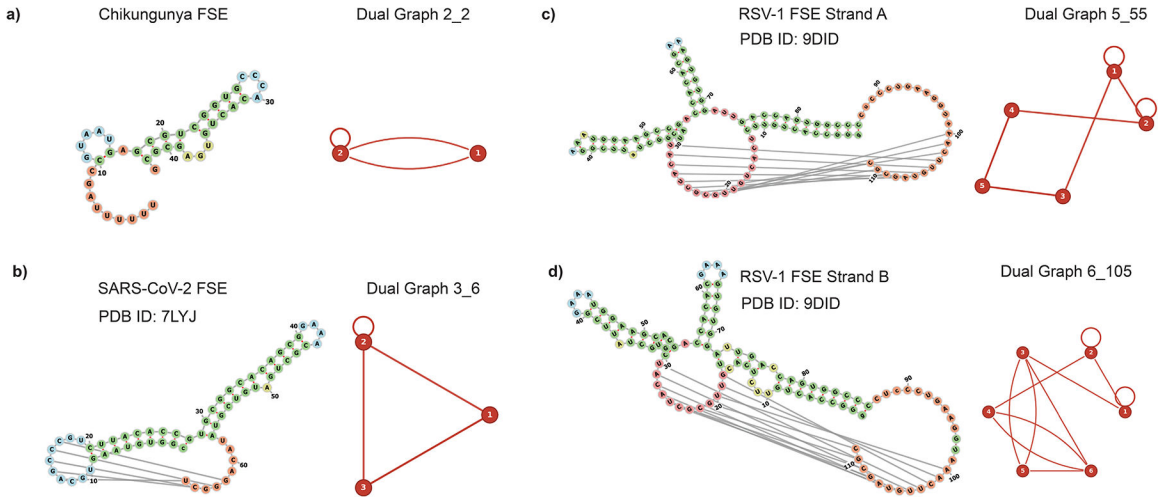


**Figure 5:**

Dual graph representations of viral frameshifting elements. Each row represents a different frameshifting element: **a)** Chikungunya virus frameshifting element (26), **b)** SARS-CoV-2 frameshifting pseudoknot RNA (25), **c)** Rous sarcoma virus frameshifting pseudoknot RNA strand A (41), and **d)** Rous sarcoma virus frameshifting pseudoknot RNA strand B (41). For each structure, the secondary structure diagram (left) produced from Forna (42) and dual graph ID (right) are shown.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1:**

Mapping of standard genome-level mutations (in the coding region) to corresponding protein-level effects.

| Genome-Level Mutation | Definition | Possible Protein-Level Consequences | |
|---|---|---|---|
| **Substitution** | One nucleotide is replaced | **1** | Silent (synonymous) |
| | | **2** | Missense (amino acid (AA) change) |
| | | **3** | Nonsense (introduces stop codon) |
| **Insertion** | Addition of nucleotide(s) | **1** | In-frame insertion (adds AAs) |
| | | **2** | Frameshift insertion (disrupts the reading frame, truncated protein) |
| **Deletion** | Removal of nucleotide(s) | **1** | In-frame deletion (removes AAs) |
| | | **2** | Frameshift deletion (alters AA sequence and likely truncation) |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2:**

Mapping of non-standard genome-level mutations (in the coding region) to corresponding protein-level effects.

| Genome-Level Mutation | Definition | Possible Protein-Level Consequences | |
|---|---|---|---|
| **Row mutation** | Two or more consecutive substitutions | **1** | Silent (synonymous) |
| | | **2** | Missense (AA change) |
| | | **3** | Nonsense (introduces stop codon) |
| **Hot mutation** | Two non-consecutive substitutions within a 3-nt window | **1** | Silent (synonymous) |
| | | **2** | Missense (AA change) |
| | | **3** | Nonsense (introduces stop codon) |