OpenRecombinHunt: Automatic detection of recombination in publicly available viral sequences ☆

Tommaso Alfonsi [1,†], **Yavuz Samet Topcuoglu** [1,†], **Matteo Chiara** [2], **and Anna Bernasconi** [1,*]

*1 - **Department of Electronics,** Information, and Bioengineering –Politecnico di Milano, Via Ponzio 34/5, Milan 20133, Italy*
*2 - **Department of Biosciences – Università degli Studi di Milano,** Via Celoria 26, Milan 20133, Italy*

Correspondence to Anna Bernasconi: *tommaso.alfonsi@polimi.it (T. Alfonsi), yavuzsamet.topcuoglu@mail.polimi.it (Y.S. Topcuoglu), matteo.chiara@unimi.it (M. Chiara), anna.bernasconi@polimi.it (A. Bernasconi)*
https://doi.org/10.1016/j.jmb.2026.169811
Editor: Dr. David Mathews

Abstract

Zoonotic transmission and viral spillover events pose severe threats to public health, as underscored by recent pandemics. Mitigating these risks requires robust genomic surveillance systems, supported by the growing availability of openly accessible viral genome sequences through dedicated resources such as NCBI Virus and Nextstrain/Pathogens. This wealth of data highlights the need for lightweight, automated computational tools to monitor viral evolution and spread. OpenRecombinHunt extends our previously published RecombinHunt method, originally developed to identify recombinant SARS-CoV-2 lineages, to prioritize recombination patterns in any virus for which a large corpus of sequences is publicly available. Here, we couple RecombinHunt with HaploCoV, a computational workflow that stratifies viral genomes into distinct groups based on high-frequency genomic variants, without requiring a predefined reference nomenclature. We apply this framework to openly-accessible datasets for SARS-CoV-2, Respiratory Syncytial Virus (RSV) A/B, Monkeypox, Zika, Yellow Fever, and hemagglutinin segments of H5N1 Influenza A, reporting interesting recombination patterns. OpenRecombinHunt monthly updates ensure continuous monitoring, providing temporal snapshots of viral genomes with potential mosaic structure. Our method and Web Server have the potential to unlock large-scale automated support to detection of recombination in viruses, in line with current genomic surveillance interests. The Web Server is freely available at http://gmql.eu/openrecombinhunt/.

Introduction

Recent experience with major viral outbreaks, such as COVID-19, monkeypox, and avian flu, has highlighted the critical importance of genome data analysis for informing health authorities and implementing effective mitigation strategies [14]. As viral genomes constantly evolve through mutation and recombination, the growing body of publicly available genomes deposited in dedicated databases demands the development of powerful and automated computational methods to empower genomic surveillance [16,15,29,18]. Recombination represents a major and rapid driver of evolutionary innovation in viruses, and detecting recombination is critical because it can produce viral variants with increased fitness and pathogenic potential [36]. Indeed, recombinant viral strains have been implicated in changes in host range, virulence, immune

☆ This article is part of a special issue entitled: 'SI: Computation Resources 2026' published in Journal of Molecular Biology.

Please cite tts article as: T. Alfonsi, Y.S. Topcuoglu, M. Chiara, et al., OpenRecombinHunt: Automatic detection of recombination in publicly available viral sequencesJournal of Molecular Biology, https://doi.org/10.1016/j.jmb.2026.169811

escape, and even antiviral resistance [33]. In light of these considerations, monitoring recombination should be considered a top priority for effective genomic surveillance strategies.

However, the sheer scale of publicly available data renders manual inspection for recombination events infeasible, creating a major bottleneck in the rapid identification of potentially significant new variants. Traditional methods focus on detecting recombination based on phylogenetic tree construction and genome-wide scans for mosaic structures [22,19,20,24,35]. Although highly accurate, these approaches are computationally intensive and do not scale efficiently to datasets comprising hundreds or thousands of genomes.

While big data-driven tools for detecting recombination exist [3,37,21], their application is restricted to viruses with well-defined lineage classifications or nomenclature systems, creating a major obstacle for studying recombination in newly emerging viruses or those without an established classification system.

Here, we introduce OpenRecombinHunt, an automated workflow designed to overcome this limitation. Our approach builds upon the previously developed RecombinHunt [3] (recombination detection method) for recombination detection and integrates it with HaploCoV [9], which enables de novo lineage generation through unsupervised clustering of viral genome datasets. OpenRecombinHunt extends recombination surveillance to a wide range of viruses, independent of the availability of an existing classification system. This work aims to provide a scalable, end-to-end solution for automated recombination surveillance that is both powerful and broadly applicable.

OpenRecombinHunt stems from our previous experience in designing computational methods and workflows for genomic surveillance and introduces two major advantages: 1) a holistic framework that integrates HaploCoV with RecombinHunt, and 2) an automatic pipeline and Web reporting system that regularly updates datasets from NCBI Virus [16,32] and Nextstrain [15] sources and identifies novel recombination events. Together, these innovations provide a substantial improvement over previously published approaches. We applied the OpenRecombinHunt workflow to a diverse set of clinically relevant viruses, including SARS-CoV-2, Respiratory Syncytial Virus (RSV) subtypes A and B, Mpox virus, Zika virus, Yellow Fever virus, and the HA segment of Influenza A (H5N1) viruses, providing novel insights into the possible recombination patterns of these viruses and how they are represented in the input datasets.

In the remainder of the manuscript, we describe the considered data sources and summarize the foundational methods on which this work builds; present the characteristics of the Web Server (available at http://gmql.eu/openrecombinhunt/), including a detailed case study on SARS-CoV-2; and conclude with a discussion of OpenRecombinHunt's potential, implications, and limitations.

Materials and methods

Selected viral pathogens and publicly available data

The first release of OpenRecombinHunt focuses on seven viral pathogens of importance for human infection, selected for their public health relevance and the availability of extensive genomic data. A brief overview is provided below; in Table 1A, we report the number of genomes considered in our analysis, along with the IDs of reference sequences [32].

Monkeypox virus (MPXV) is a zoonotic orthopoxvirus with a large (reference sequence 197,209 bp) double-stranded DNA genome. It has historically been classified into broad clades (West African and Central African) [17], and more recently, following a 2022 epidemic in sublineages [12].

Human *respiratory syncytial viruses* (RSV) are negative-sense RNA viruses; we analyze RSV-A and RSV-B (reference sequences are respectively 15,222 bp and 15,225 bp long). Both RSV-A and RSV-B have a reference nomenclature based on G gene variations, but the current classification has a limited granularity, which would hardly allow for capturing recombination patterns.

Yellow Fever virus (YFV) is a mosquito-borne flavivirus that causes hemorrhagic fever. Its genome consists of a single, positive-sense RNA molecule approximately 10,862 nt in length. The current Yellow Fever virus classification includes five genotypes based on geography and genetics.

Zika is positive-sense RNA flavivirus transmitted by Aedes mosquitoes. The genome is 10,808 nt. Since 2007, Zika virus has caused several distinct epidemics linked to neurological disease. The virus also lacks a universally accepted lineage-based classification system.

SARS-CoV-2, the etiological agent of the COVID-19 pandemic, possesses a large positive-sense RNA genome (reference sequence length: 29,903 nt) and is characterized by a comprehensive and widely adopted nomenclature system, including Pango lineages and Nextstrain clades.

Influenza A viruses are segmented, negative-sense RNA viruses. We focused on the H5N1, a highly pathogenic avian influenza virus that has recently drawn worldwide attention due to recurrent spillover in mammals [5,4]. Because the hemagglutinin (HA) segment encodes the protein responsible for binding to host cell receptors and mediating viral entry, we restricted our recombination analysis to this segment (reference sequence length: 1,760 nt). Although a genotype classification

Table 1 (**A**) Overview of datasets downloaded from the sources (date: Oct. 30th, 2025). For each virus, we specify: the source, the number of downloaded vs. analyzed (after applying filters) sequences, the reference sequence, and its length. (**B**) Designation of lineages per virus. When HaploCoV is applied, we indicate `dist` and `size` parameters' values. Then, we provide the number of designations according to HaploCoV, of existing ones, and total ones (used in the following phases). (**C**) Recombination results obtained for each virus, reporting the counts of genomes, total recombined genomes, divided by 1BP and 2BP, their proportion on all the genomes, the number of distinct lineage combinations found, their proportion on all genomes, and the most prevalent pattern.

A)	Virus name	Source	#Download	#Analyse	Reference ID	Ref. length
	Monkeypox	NCBI CL	11,086	8,933	NC_063383.1	197,209 bp
	RSV-A	Nextstrain	34,940	27,054	NC_038235.1	15,222 bp
	RSV-B	Nextstrain	26,336	20,368	NC_001781.1	15,225 bp
	Yellow Fever	NCBI CL	2,288	1,408	NC_002031.1	10,862 bp
	Zika	NCBI CL	2,593	1,266	NC_035889.1	10,808 bp
	Influenza H5N1, HA segm.	NCBI FTP	13,990	13,921	NC_007362.1	1,760 bp
	SARS-CoV-2 (cons.)	Nextstrain	35,182	18,502	NC_045512.2	29,903 bp
	SARS-CoV-2 ($\leqslant$2 m)	Nextstrain	5,486	2,523	NC_045512.2	29,903 bp

B)	Virus name	HaploCoV `dist`	HaploCoV `size`	#HaploCoV design.	#Existing design.	#Total design.
	Monkeypox	3	10	45	54	99
	RSV-A	5	50	16	39	55
	RSV-B	5	50	8	24	32
	Yellow Fever	3	10	3	NA	3
	Zika	3	10	3	NA	3
	Influenza H5N1, HA segm.	5	50	14	NA	14
	SARS-CoV-2 (cons.)	NA	NA	NA	134	134
	SARS-CoV-2 ($\leqslant$2 m)	NA	NA	NA	106	106

C)			Recombinant				Patterns		
	Virus name	#Tot. genomes	#Total Rec.	#1BP Rec.	#2BP Rec.	Rec. Ratio	#Unique Patt.	Patt. Ratio	Most common patt. (#genomes)
	Monkeypox	8,933	966	868	98	0.108	392	0.044	B.1.NmC6 + C.1.NmC2 (97)
	RSV-A	27,054	3,089	3,065	43	0.114	303	0.011	A.D.4.NmC3 + A.3.1.NmC1 (422)
	RSV-B	20,368	3,348	2,301	47	0.164	189	0.009	B.D.E.3 + B.D.NmC1 (367)
	Yellow Fever	1,480	144	101	43	0.097	12	0.008	A.1.NmC2 + A.1.NmC1 (70)
	Zika	1,266	163	130	33	0.129	9	0.007	A.1 + A.1.NmC1 (27)
	Influenza H5N1, HA segm.	13,921	1,263	741	522	0.091	115	0.008	A.1.NmC7 + A.1.NmC13 (154)
	SARS-CoV-2 ($\leqslant$2mos.)	2,523	188	144	44	0.075	142	0.056	KP.3.1.1 + XFG.6 (6)

has been proposed within the GISAID [34] database for the virus type circulating in North America [28], no unified lineage scheme exists for H5N1 HA beyond the WHO clade numbering system.

All datasets used in this manuscript were downloaded on October 30th, 2025. Accordingly, the analyses presented here reflect data available up to that date. However, the OpenRecombinHunt system is automatically updated on a monthly basis, ensuring that future runs incorporate the most recent sequences released in public repositories.

Lineage-based nomenclature

Viral genome classification systems group sequences into biologically or epidemiologically meaningful categories—-most commonly lineages or clades defined through phylogenetic inference [10].

When a reference classification system exists, newly sequenced genomes can be assigned directly to established lineages or clades. For viruses lacking such a framework, unsupervised approaches are required to derive a coherent classification from the genomic data itself. For this task, we selected HaploCoV [9], an automated framework that clusters viral genomes into haplogroups (HGs) based on phenetic clustering of high-frequency genetic variants. Each haplogroup is defined by a distinctive set of shared high frequency variants. HaploCoV operates without a phylogenetic tree, enabling fast and scalable classification. Once HGs are identified, they can be annotated with metadata such as geographic origin, collection date, and correspondence with any existing lineage or clade labels.

We applied HaploCoV to viruses lacking an established nomenclature (Yellow Fever, Zika, Influenza H5N1-HA). For viruses with existing lineages (MPXV, RSV), we compare and integrate the official labels with HaploCoV-defined haplogroups. HaploCoV was not applied to SARS-CoV-2 because this virus already possesses a comprehensive and widely adopted nomenclature.

Recombination detection

Recombination is captured by RecombinHunt [3], a maximum-likelihood–based computational tool that detects recombination by formally comparing recombinant and non-recombinant evolutionary models.

Method summary. RecombinHunt represents genomic sequences to be analysed (target genomes) as a collection of mutations, configuring the *target mutations-space*. Similarly, lineages as defined by a reference nomenclature are represented as sets of characteristic mutations, that is, mutations observed in at least a given proportion of the genomes assigned to that lineage. Target genomes are assigned to one or more candidate parent lineages based on the similarity of mutation patterns by computing a likelihood score. For this computation, the *target mutations-space* is -by turns- extended with candidate lineages' mutations, resulting in the *extended target space*. At each position of this space, RecombinHunt [3] computes cumulative likelihood ratios that quantify how well its mutation pattern aligns with each lineage in the reference nomenclature. The algorithm then evaluates the probability of the observed data under a *non-recombinant model* (single parental lineage) and a *recombinant model* (two parental lineages contributing distinct genome regions).

Model selection is performed using the Akaike Information Criterion (AIC) [1]; target genomes for which the recombinant model provides a statistically significant better fit are classified as putative recombinants.

Breakpoint identification. RecombinHunt partitions the genome into either two regions (1-breakpoint model) or three regions (2-breakpoint model) and evaluates the likelihood of each segment with respect to candidate parental lineages. Breakpoints are first identified between consecutive lineage-informative mutations observed in the extended target space (which may or may not be adjacent). The interval is then mapped to the closest position $[p, p+1]$ in the target mutations-space (ends excluded), and also expressed in the genomic coordinates of the target genome. This yields three possible outcomes: 0BP (non-recombinant), 1BP (one breakpoint), or 2BP (two breakpoints).

Filtering of Putative Recombinants. Putative recombinants are further filtered through additional criteria: (C1) the *p*-value of the likelihood-ratio test is $\leqslant 10^{-5}$); (C2) the breakpoints align (within one mutation-tolerance) with the defining segment of the top-ranked lineage assigned; (C3) if a structured nomenclature is available, the recombinant candidate must belong to the same phylogenetic branch as its inferred parental lineages.

Execution modes. RecombinHunt analyses single genome sequences but also ideal consensus sequences derived from lineages or clades. Lineage/clade consensus genomes sequences are constructed by considering only mutations shared by a certain percentage of high-quality sequences within a cluster (left as user-configurable) and are used only for extensive datasets such as SARS-CoV-2.

OpenRecombinHunt pipeline

The OpenRecombinHunt pipeline comprises a series of modular scripts, orchestrated by a master controller (see Figure S1, Supplementary Materials), responsible for executing each module sequentially, from data acquisition to the final

analysis. Next, we provide an overview of the modules (see Figure S2, Supplementary Materials).

`1) Data Acquisition` activates the automated download of genome sequences and associated metadata for selected viruses from various data sources and formats, producing a standardized set of output files. The pipeline supports three distinct data acquisition workflows. For the virus selection described in Section 2.1, the two primary data sources are the National Center for Biotechnology Information (NCBI) (providing both a Command Line interface and an FTP Server [26,25]) and the Nextstrain platform [27]. We choose a specific source for each virus based on i) the frequency of updates and availability of the most recent genomic sequences; ii) the completeness and reliability of associated metadata (i.e., collection date and assigned lineage/clade); iii) accessibility through programmatic methods and their scalability to accommodate bulk retrieval. The data retrieval mechanism is tolerant with respect to exceptions regarding internet connection; corrupted records are discarded; we assume no duplicates are contained in the data sources. At each run, the database is initiated from scratch (except for the SARS-CoV-2, where only a recent window is considered). Considered files are multi-FASTA for genome sequences and tab-separated for metadata tabular content.

`2) Preprocessing` prepares the collected data for HaploCoV. We apply a series of data quality filters to both genomic sequences and associated metadata, informed by an initial exploratory analysis. These steps include standardizing sequence identifiers and column names, harmonizing date and location formats, retrieving available lineage annotations, and verifying that each record contains complete metadata (collection date, accession, and sequence length). Only genomes with a length greater than 80% of the corresponding reference sequence are retained. Per-genome FASTA files are then retained exclusively for genomes linked to this cleaned metadata table and subsequently compared against the selected reference sequence for each virus. Virus-specific Web Server pages provide further details on filters and reference sequences accession IDs.

`3) HaploCoV` module orchestrates the execution of HaploCoV [9] to generate and/or extend a lineage-based nomenclature and derive genomic variants for every sequence in the dataset. This module is not executed for the SARS-CoV-2 analysis, as the curated Nextstrain metadata for SARS-CoV-2 already includes a fine-grain Pango nomenclature [31] and pre-calculated lists of genomic variants. For each viral genome, HaploCoV performs an alignment against the provided reference sequence using the `nucmer` program from the MUMmer4 suite [23]. The genetic variants (substitutions, insertions, and deletions) identified by this alignment are then extracted. Finally, HaploCoV designates *haplogroups* (HG), as described in Section 2.2 – for ease of reading, we refer to HaploCoV-defined haplogroups (HGs) as "lineages" throughout the manuscript.

We implement two execution modes: i) *De novo clustering:* For viruses that lack a pre-existing classification (such as Yellow Fever and Zika), all sequences are initially assigned a default lineage of "`A.1`". HaploCoV then uses its clustering algorithm to build a new classification system from scratch. ii) *Extension of existing lineages:* For viruses that already have a baseline classification (such as RSV from Nextstrain), HaploCoV uses the pre-assigned lineages as a starting point and applies the same clustering logic to identify potential sub-clusters and refine the existing nomenclature.

HaploCoV requires two parameters. The `dist` parameter defines the maximum number of mutations allowed between two sequences for them to be considered part of the same initial cluster. Choosing smaller values results in finer, more granular clusters. The `size` parameter sets the minimum number of sequences required for a cluster to be considered a new, stable group. A larger value prevents the creation of spurious groups from small, noisy clusters. The choice of these parameters was made using a qualitative approach – heatmaps illustrating the global distribution of lineages under varying configurations were generated (see virus-specific pages, Web Server); we prioritized parameter configurations generating clustering solutions with marked geographical structuring. Table 1B reports details on the first three phases of the pipeline for all the considered viruses.

`4) Postprocessing` performs the standardization of mutation (compound/substitutions, deletions, insertions) notation – this ensures uniformity across all analyses, regardless of the original data source. Depending on the data source, different conversion rules are applied. In the output, for substitutions we use a `POSITION_REF|ALT` format (e.g., `241_C|T`); for deletions we use an underscore-separated format (e.g., `11288_11297`); for insertions we employ the RecombinHunt format (e.g., `28263_.|A`). Compound mutations generated by the HaploCoV output (where a single string represents multiple adjacent events) are reduced to a list of corresponding single-nucleotide mutations. To optimize performance, the script utilizes a caching mechanism. The final output is a new file, which serves as the input for the subsequent pipeline steps.

`5) Prepare for RecombinHunt` module is run to build the specific inputs required by the RecombinHunt tool, specifically per-lineage "sample" files and a single analysis "environment".

Samples contain the `genomeID`, `true_lineage` (name of the assigned lineage), and `nuc_changes` (the list of mutations for each genome).

The environment is a set of pre-computed data files that characterize the genetic landscape of the virus, including: 1) A file that stores the observed frequency of every genetic variant across the entire dataset – this serves as a global reference for mutation frequencies. 2) A Boolean matrix whose rows represent mutations and columns represent lineages; a `True` value indicates that a mutation *m* is characteristic of a lineage *l* (i.e., present in more than a certain % of its genomes)– collectively we refer to the set of 'true' mutations in a row as the *characterization of the lineage* presented in the row. 3) A file that stores the total number of sequences used to calculate its characterization – we set `min_genome_count` = 10 as the minimum number of genomes for a lineage to be included in the environment (preventing statistical noise from very small or singleton clusters from skewing the analysis) and we set `lc_threshold` = 75% (vs 50% for Monkeypox) as the "consensus" lineage characterization threshold, meaning a mutation is considered "characteristic" for the lineage when present in at least 75% of the lineage's sequences. This results in an 'approximation' of the lineages derived from the tree, removing noise. While 75% for SARS-CoV-2 is widely accepted by the community [13], a method to assess this threshold numerically is proposed in [3].

6) `RecombinHunt` module of the pipeline first loads the complete analysis environment, then iterates through each lineage's sample file, processing either all sequences or a single consensus sequence (SARS-CoV-2). For each sample, it runs the RecombinHunt experiment, categorizes the result (0BP, 1BP, or 2BP), and generates a detailed report. The final outputs of this module are structured data files that power the Web-based visualization dashboard, as detailed in the next section.

7) `Streamlit` module builds the visualization and user interaction layer of the pipeline. While the preceding modules focus on computation, lineage clustering, and recombinant detection, Streamlit provides an accessible and interactive interface to explore these results. The Web Server integrates the outputs from all earlier modules—such as lineage definitions, recombinant candidates, and summary statistics—and presents them in a structured dashboard format. The Web Application welcomes the users with a Home page which summarizes the main features of the tool, followed by individual tabs for each virus included in the system; each tab contains three sections. *About the Virus* presents general statistics about the processed dataset, including collection date ranges, number of sequences, and geographical distribution. *Summary Dashboard* offers an overview of lineage dynamics and distribution patterns across time and geography. *Recombinant Explorer* provides an interactive environment to investigate putative recombinant candidates in detail, enabling users to apply filters (e.g., lineage, breakpoint model, continent); for SARS-CoV-2, this section is divided into "last 2 months" (including single genomes' analyses for all sequences collected in the last two months) and "consensus sequences".

Pipeline orchestration

The orchestration of the pipeline is managed by a script that sequentially invokes all the modules. The data processing pipeline, as well as the Streamlit Web Server, has been containerized using Docker (see Figure S3 in the Supplementary Materials). Containerization provides isolation from server dependencies, reproducibility, and simple deployment. The containerization of the frontend decouples the user-facing application from the backend computations, enabling streamlined updates and minimizing dependency issues on the host machine.

The process is parameterized by a global configuration file `config.yaml`, defining critical parameters such as data source locations (URLs or command-line templates), tool-specific settings (for HaploCoV and RecombinHunt), and data quality filtering rules.

The pipeline is designed for automated, unattended execution, as achieved using a `cron` task, a standard Unix-based job scheduler, on a recurring schedule (monthly), enabling the system to update its analyses with the latest available data automatically.

Results

Candidate recombination patterns detected by OpenRecombinHunt are summarized in Table 1C. Signals were also detected in viruses where recombination is generally considered rare or absent, including RSV-A, RSV-B, Zika virus, Yellow fever virus, and the influenza H5N1 HA segment. However, when the number of unique patterns is normalized by the total number of genomes analyzed, these viruses show a very low proportion of distinct recombination patterns despite the relatively large number of candidate recombinant genomes. This indicates that many detections correspond to repeated occurrences of a limited set of patterns, consistent with false positives or other confounding signals. For viruses in which recombination is not expected, the observed proportion of distinct recombination patterns is broadly consistent with the expected type I error rate given the statistical significance threshold used in the analysis (p-value = 0.01). In

contrast, SARS-CoV-2 and Monkeypox display a substantially higher proportion of unique recombination patterns relative to the number of genomes analyzed, a signal more compatible with genuine recombination generating diverse mosaic genomes.

All pipeline results are accessible through the Web Server, which provides an interactive environment for exploring processed datasets and recombination analyses. A navigation bar on the left allows the user to switch between pages dedicated to each virus. Each virus-specific page provides three tabs: *About the Virus*, *Summary Dashboard*, and *Recombinant Explorer*. In the following, we show the distinct sections using the case of Monkeypox for illustration purposes.

The first tab provides background information on the virus, followed by details about the 'Reference Genome' (its identifier, genome length, and an option to download the reference genome sequence in FASTA format); the employed 'Data Source' (direct links or command-line instructions to retrieve the corresponding datasets); and the data 'Quality Filters' applied during preprocessing and 'HaploCoV Parameters' selected for clustering of genomes. The 'Dataset Overview' section reports statistics for the final dataset used in the analysis; the page concludes with 'References and Resources' for further investigation.

The *Summary Dashboard* aggregates results generated by the previous modules. This dashboard supports interactive filtering, either by selecting a date range or by restricting the view to the latest *n* sequences, sorted by their collection date. At the top of the tab, a section named 'Key Metrics' (see Fig. 1A) displays: the total number of analyzed sequences; number of possibly recombinant genomes (split into 1BP and 2BP); top recombinant lineage; and most common parental lineages (with specific combination and cardinality). Additional details are provided in the 'Summary Tables' (see Fig. 1B) section, which includes: (i) a *Lineage Breakdown*, with recombination metrics per lineage; and (ii) *Recombinant Hotspots*, which highlight frequent recombination patterns. The tab concludes with 'Distribution Plots' (see Fig. 1C), both temporal and geographic, implemented as interactive visualizations.

The *Recombinant Explorer* tab enables detailed inspection of possibly recombinant genomes. Users can filter genomes by lineage, breakpoint count, and collection country. The filtered results are displayed in a table. All the data displayed in this tab is retrieved from the structured reports content generated earlier in the pipeline; see Fig. 1D-E for visual representations, respectively showing the selection of a 1BP genome collected

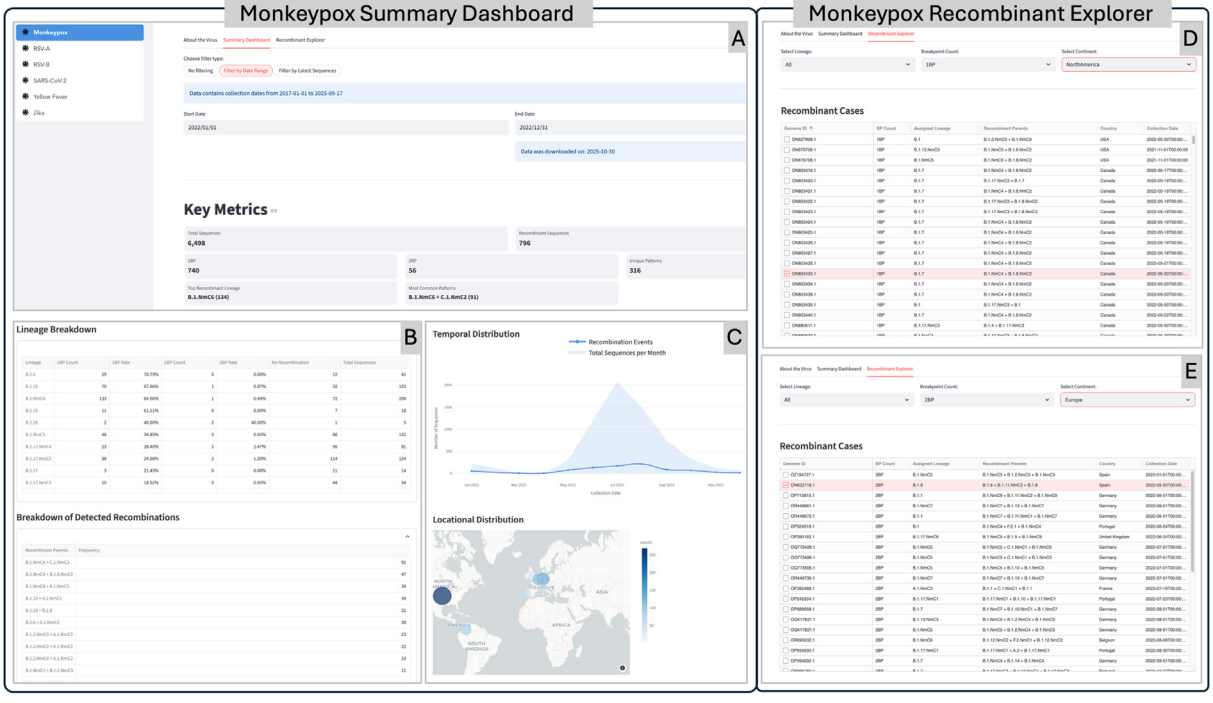

Fig. 1. Overview of the main tabs in Summary Dashboard (A–C), including configurable filters, key metrics, summary tables and plots with temporal/geographical distributions, and in Recombinant Explorer (D–E), including filters and results corresponding to different selections.

in North America and a 2BP genome collected in Europe. Ticking the box next to a genome opens a detailed view consisting of four sections:

- *Case Summary* provides a concise overview of the candidate recombinant, including the source data and overall evaluation metrics.
- *Region Analysis Tables* report analytical details of the RecombinHunt approach. Here, a *region* is defined as a nucleotide sequence between breakpoints. A genome with 1 breakpoint contains 2 regions, while a genome with 2 breakpoints contains 3 regions. Region Tables report the number of sequences, the breakpoint, and the maximum likelihood ratio. The following columns illustrate the comparison of the candidate with the first row of the table: value of one-sided AIC comparison between recombination model and non-recombination model (lower values are when row candidate is similar to the first candidate); p-value of AIC – without multiple comparison corrections; and three conditions: (C1) marks if p-value is $>= \geqslant 10^{-5}$; (C2) marks if row breakpoint is at most one mutation apart from the one of the first candidate; (C3) marks if the candidate belongs to the same phylogenetic branch as the first one. It is possible to prioritize candidate lineages that satisfy all three flag conditions (C1 & C2 & C3).
- *Visualization* plots per-region cumulative $\log(P)$ values on the *y*-axis as a function of target changes on the *x*-axis, showing parental lineage contributions and enabling intuitive interpretation of the statistical evidence for recombination.
- *Target Mutations* include a complete list of mutations involved in the possibly recombinant genome which can also be downloaded as a file via the dedicated button, for further verification and functional/epidemiological analysis.

Fig. 2 presents two such examples of Mpox genomes. Specifically, panel A illustrates the case of *ON803433.1*, collected in Canada and initially assigned to B.1.7 lineage (from the existing nomenclature), with 59 mutations w.r.t. to the reference genome. RecombinHunt recognized this genome as potentially deriving from a single breakpoint recombination pattern, located between the 53rd and 54th mutation (out of the 59 ones in the genome), which is resolved around the 168,204–168,205th nucleotides of the reference genome. We also show the confidence in the test that compares the recombination outcome against both possible non-recombinant cases (B.1.NmC4 vs. B.1.8.NmC2).

The region analysis tables show the candidates for the stretch spanning from the 5'-end and the break point. For completeness, here we show all ten candidates, represented by i) the number of sequences, ii) the mutation at which their likelihood profile reaches the maximum value, and statistical indicators including iii) the Likelihood Ratio, iv) the AIC score, v) the p-value of the test. Only B.1.NmC4 has a satisfactory value in the three criteria we use to filter the candidates. Similarly, we choose B.1.8.NmC2 for the stretch spanning from the breakpoint to the 3'-end. A visual representation of the cumulative likelihood profile is rendered in orange and blue. Below, we see the list of 57 substitutions, one insertion, and one deletion.

Panel B shows the case of *ON622718.1*, collected in Spain and initially assigned to B.1.8 lineage, with 63 mutations w.r.t. to the reference. RecombinHunt recognized its genome as potentially deriving from a double breakpoint recombination pattern, respectively located between the 6-7th and the 20-21st mutations (resolved at the 21,990–21,991st and 64,296–64,297th nucleotides of the reference genome). The region analysis tables show that the strongest candidates are B.1.8 for the external stretches and B.1.11NmC2, meaning that a 14-nucleotide-long insertion of a different strain is likely to have been introduced in a genome originally assigned to B.1.8.

Handling large-scale SARS-CoV-2 data

The SARS-CoV-2 dataset is exceptionally large, containing more than nine million genomes. Because recomputing results for the full dataset is computationally prohibitive, we implemented two complementary analysis modes: 1) *Last2M-mode*, considering only the genomes collected in the last two months (quasi-*real time* surveillance of recombinants) – reporting cases where our tool disagrees with the currently assigned Pango lineage, and 2) *Consensus-mode*, which analyzes one consensus sequence per lineage. Each consensus genome represents the characteristic mutation profile of all high-quality sequences assigned to that lineage. This mode is designed to enable the prioritization of possibly novel recombinant lineages.

For SARS-CoV-2, the Web Server provides a dedicated Summary Dashboard for the *Last2M-mode* and a Recombinant Explorer for both modes: one entry per genome in *Last2M-mode* and one entry per lineage (via its consensus genome) in *Consensus-mode*.

In the *Last2M-mode*, several genomes assigned to recombinant lineages in the Pango system, such as `XFG` (and its descendants), as well as `XFC`, `XFN`, and `XFJ` are also flagged as recombinant by our approach. All other known recombinant lineages show either no representation or only negligible counts in the analyzed dataset.

Results obtained from consensus genomes were compared with those previously reported in [3], which included an analysis on open data available from Nextstrain at the beginning of 2023. For the present analysis, we retrieved the

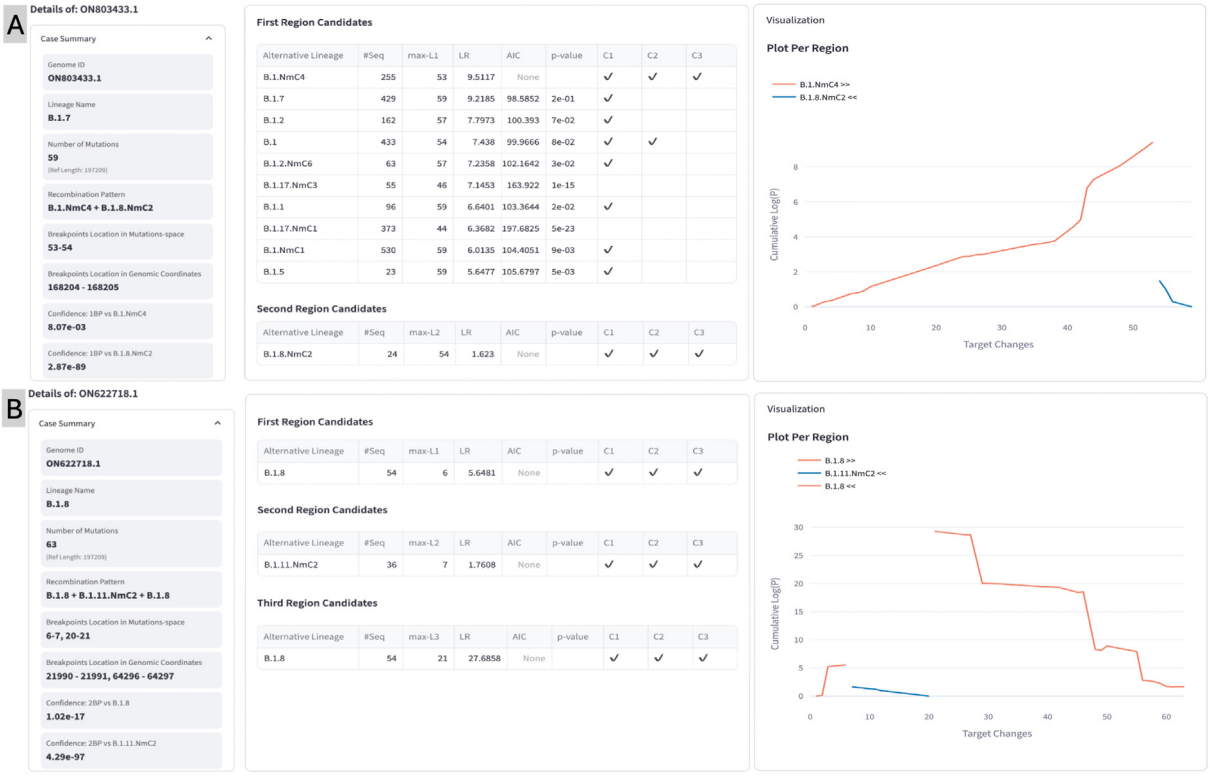

Fig. 2. Genome cards for two sequences identified as possibly recombinant in Mpox. A) illustrates the case of the Canada-collected 1BP genome ON803433.1. B) illustrates the case of the Spain-collected 2BP genome ON622718.1.

corresponding dataset from Nextstrain to ensure consistency of the comparison. In the previous RecombinHunt implementation, only 51 recombinant lineages (i.e., Pango lineages beginning with 'X' and sufficiently populated for consensus construction) could be assessed. In the present analysis, we extend this evaluation substantially: among 134 lineages currently labeled as "recombinant" in the Pango nomenclature, 85 exhibit statistically significant recombinant signals in our framework, while the remaining lineages could not be assessed due to insufficient sequence data. As shown in Table S1 (Supplementary Materials), the large majority of the 51 lineages evaluated previously are confirmed by the updated results. Six lineages (`XBB`, `XBF`, `XBJ`, `XP`, `XAC`, `XBL`) are now classified as non-recombinant, and five lineages show a different combination of inferred parental candidates.

These discrepancies are likely due to updates in the underlying data, including improvements in sequence quality, revised Pango lineage assignments, and substantial changes in lineage size over time. Overall, however, the consistency of the results reinforces the robustness of the approach and supports its continued use for large-scale SARS-CoV-2 recombination surveillance.

Discussion and conclusions

The emergence of novel viral variants, often facilitated by recombination, poses a persistent challenge to global public health. Although genomic surveillance efforts continue to expand, the growing volume of publicly available sequence data–combined with the lack of established lineage classifications for many viruses–creates a major bottleneck in the rapid detection of new recombinant strains.

In this work, we introduced OpenRecombinHunt, a robust and automated pipeline designed to overcome these limitations. By integrating our previously developed maximum-likelihood-based recombination detection method, RecombinHunt, with the de novo lineage creation capability of HaploCoV, we have developed a comprehensive framework for proactive genomic surveillance.

Key contributions and findings. OpenRecombinHunt acts as a large-scale screening framework that flags candidate recombinant genomes for further validation. It is designed to operate on large genomic datasets (tens of thousands of genomes or more), where exhaustive expert-curated analyses would be computationally and conceptually prohibitive. In

this context, the goal of the pipeline is to prioritize genomes showing signals compatible with recombination, which can then be subjected to manual inspection or classical recombination detection approaches. OpenRecombinHunt's main advantage is its broad applicability across diverse viral datasets. Unlike previous tools, which were limited to viruses with established nomenclature (e.g., Pango lineages), our pipeline successfully analyzed diverse viral taxa – including SARS-CoV-2, RSV, Mpox, Zika, Yellow Fever, and the HA segment of H5N1 – many of which lack a granular, unified classification system. The application of HaploCoV's clustering logic to define robust haplogroups or refine existing lineages proved essential for providing the necessary context for RecombinHunt's analysis, enabling the identification of hundreds of potential recombination patterns (Table 1C).

The SARS-CoV-2 analysis served both as a validation and a demonstration of the system's scalability. By utilizing both a quasi-real-time mode (Last2M-mode) for newly collected sequences and a Consensus-mode for established lineages, we were able to confirm the recombinant nature of the great majority of 'X-named' lineages (see Table in the Github repository documentation). This cross-validation, despite differences due to the evolution of the underlying data over 2.5 years, confirms the reliability and continued relevance of the RecombinHunt methodology.

The Streamlit-based Web Server component is a critical aspect of this work, transforming complex genomic data into an accessible, interactive dashboard. This interface empowers virologists and public health officials to explore lineage dynamics, apply filters, and investigate individual recombinant candidates in detail, complete with statistical evidence and cumulative likelihood plots for intuitive interpretation.

Threats to validity. RecombinHunt [3] was originally validated on SARS-CoV-2. OpenRecombinHunt framework extends this method by embedding it within a wider pipeline. Importantly, none of its components rely on coronavirus-specific biological assumptions; rather, the framework depends on the presence of phylogenetically coherent lineage structures and detectable mosaic genetic signals.

Nevertheless, signals detected in other viruses should be interpreted with caution. Based on current biological knowledge, recombination is not expected to be a major evolutionary mechanism in RSV or influenza viruses.

In our datasets, genome data quality and reference suitability appear substantially more heterogeneous for RSV and influenza than for SARS-CoV-2, likely reflecting the exceptional scale and consistency of SARS-CoV-2 genomic surveillance. For example, in the RSV dataset, some candidate lineages inferred by HaploCoV were characterized by more than 800 defining mutations (see Figure S4, Supplementary Materials) in a genome of approximately 15 kb, suggesting that the available reference genome may be highly divergent from many circulating strains. In contrast, other candidate lineages were defined by as few as $\sim$20 mutations, indicating that the reference genome may be more appropriate for those sequences. In such heterogeneous settings, the estimation of mutation frequencies becomes less stable, which can in turn affect methods such as RecombinHunt that rely on likelihood models based on mutation frequencies along the genome.

Consistent with this interpretation, although the pipeline flags a relatively large number of candidate recombinant genomes in RSV and influenza, these signals largely collapse into a limited number of recurring mosaic patterns, and their relative proportion is substantially lower than what is observed in SARS-CoV-2. When considering recombination patterns rather than individual genomes, the observed proportion of distinct patterns in viruses where recombination is not expected is broadly compatible with the Type I error rate of the AIC-based statistical test used in the analysis (p-value $= 0.01$).

Thus, the number of sequences flagged by OpenRecombinHunt is larger than expected if interpreted strictly as recombination. However, this outcome should be considered in light of the methodological scope of the framework.

OpenRecombinHunt only flags candidate genomes for further manual verification, and several alternative explanations may generate similar signals, e.g., (i) In short genomes or lineages defined by few mutations, the distinction between recombination and convergent evolution can be subtle within this modeling framework; and (ii) Technical artifacts may produce recombination-like signals. For example, genome mis-assembly or co-infections involving distinct viral lineages may generate chimeric or mixed consensus sequences that mimic recombination.

Limitations and future work. The performance of HaploCoV, and consequently RecombinHunt, is dependent on the quality of the input data and the tuning of clustering parameters (i.e., dist and size). The current parameter choice, based on a qualitative assessment of geographical structuring, may not be universally optimal for all possible viruses. Future work will focus on developing a dynamic, data-driven mechanism for automatic parameter selection to maximize clustering stability and biological relevance.

Furthermore, the recombination detection relies on a maximum-likelihood framework comparing single-parent (non-recombinant) and two-parent (recombinant) models. While highly effective, this approach may struggle with complex, multi-event recombination or when parental lineages are

poorly represented or have diverged significantly over time. In the current implementation, the method is limited to recognizing positions of breakpoints in the mutations target-space, resulting in some degree of uncertainty. In this sense, RecombinHunt aligns with what is also proposed by RDP [24] or RIPPLES/RIVET [36], where breakpoints are provided as 'intervals where the break has likely happened'. Future extension of the proposed pipeline may target multi-segment viruses (for Influenza, we are independently developing reassortment-aware methods [5,4]) or, instead, target recombination across different virus subfamilies or families (by replacing HaploCoV with an appropriate module). In line with previous static analysis [7,2] or online-monitoring [8,6,11,30] of SARS-CoV-2 data in the context of GeCo [38] and SENSIBLE [39] projects, we here proposed the OpenRecombinHunt pipeline. OpenRecombinHunt provides an essential, scalable, and automated layer to the global genomic surveillance infrastructure. By continuously monitoring the evolutionary dynamics of diverse viruses, we aim to accelerate the detection of new recombinant threats, ultimately supporting faster and more informed public health responses worldwide.

Data and code availability statement The code is freely available at https://github.com/DEIB-GECO/open-recombinhunt and the tool is at http://gmql.eu/openrecombinhunt/.

Funding. This work was supported by Ministero dell'Università e della Ricerca (PRIN PNRR 2022 "SENSIBLE" project, n. P2022CNN2J), funded by the European Union, Next Generation EU, within PNRR M4.C2.1.1. Politecnico di Milano, CUP D53D23017400001; Università degli Studi di Milano, CUP G53D23006690001. AB Principal Investigator, MC co-Principal Investigator.

Acknowledgements.

We thank all data contributors to the NCBI Virus and Nextstrain databases. We thank Prof. Manuela Sironi for the useful discussion at the beginning of this project.

CRediT authorship contribution statement

Anna Bernasconi: Writing – original draft, Supervision, Software, Funding acquisition, Conceptualization.**Matteo Chiara**: Writing – review & editing, Supervision, Software, Funding acquisition, Conceptualization.**Yavuz Samet Topcuoglu**: Software.**Tommaso Alfonsi**: Software, Methodology, Investigation.

DATA AVAILABILITY

We describe the process through which public datasets were retrieved. The Web Server updates monthly the datasets thorugh the described URLs/procedure.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.jmb.2026.169811.

Received 9 December 2025;
Accepted 10 April 2026;
Available online xxxx

Keywords:
viral recombination;
genomic surveillance;
bioinformatics pipeline;
HaploCoV;
RecombinHunt

† These authors contributed equally.

References

[1]. Akaike, H., (2003). A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723.

[2]. Al Khalaf, R., Bernasconi, A., Pinoli, P., (2024). Systematic analysis of sars-cov-2 omicron subvariants' impact on b and t cell epitopes. *PLoS One* **19**, e0307873.

[3]. Alfonsi, T., Bernasconi, A., Chiara, M., Ceri, S., (2024). Data-driven recombination detection in viral genomes. *Nat. Commun.* **15**, 3313.

[4]. Alfonsi, T., Bernasconi, A., Chiara, M., Ceri, S., (2025). Lightweight multiscale early warning system for influenza a spillovers. *Sci. Adv.* **11**, eadz7312.

[5]. Alfonsi, T., Chiara, M., Bernasconi, A., (2025). A codon usage-based approach for the stratification of influenza a across recent spillovers. *Comput. Struct. Biotechnol. J.* **27**, 2757–2771.

[6]. Bernasconi, A., Gulino, A., Alfonsi, T., Canakoglu, A., Pinoli, P., Sandionigi, A., Ceri, S., (2021). Virusviz: comparative analysis and effective visualization of viral nucleotide and amino acid variants. *Nucl. Acids Res.* **49** e90–e90.

[7]. Bernasconi, A., Mari, L., Casagrandi, R., Ceri, S., (2021). Data-driven analysis of amino acid change dynamics timely reveals sars-cov-2 variant emergence. *Scient. Rep.* **11**, 21068.

[8]. Canakoglu, A., Pinoli, P., Bernasconi, A., Alfonsi, T., Melidis, D.P., Ceri, S., (2021). Virusurf: an integrated database to investigate viral sequences. *Nucl. Acids Res.* **49**, D817–D824.

[9]. Chiara, M., Horner, D.S., Ferrandi, E., Gissi, C., Pesole, G., (2023). Haplocov: unsupervised classification and rapid detection of novel emerging variants of sars-cov-2. *Commun. Biol.* **6**, 443.

[10]. Chiara, M., Horner, D.S., Gissi, C., Pesole, G., (2021). Comparative genomics reveals early emergence and biased spatiotemporal distribution of sars-cov-2. *Mol. Biol. Evol.* **38**, 2547–2565.

[11]. Cilibrasi, L., Pinoli, P., Bernasconi, A., Canakoglu, A., Chiara, M., Ceri, S., (2022). Viruclust: direct comparison of sars-cov-2 genomes and genetic variants in space and time. *Bioinformatics* **38**, 1988–1994.

[12]. De Pascali, A.M., Ingletto, L., Brandolini, M., Rocchi, E., Tarozzi, M., Turba, M.E., Casadio, R., Gentilini, F., Gatti, G., Dionisi, L., et al., (2025). Understanding the evolutionary dynamics of monkeypox virus through less explored pathways. *Scient. Rep* **15**, 25849.

[13]. Gangavarapu, K., Latif, A.A., Mullen, J.L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Zeller, M., Aceves, C. M., Zaiets, K., et al., (2023). Outbreak. info genomic reports: scalable and dynamic surveillance of sars-cov-2 variants and mutations. *Nat. Methods* **20**, 512–522.

[14]. Gottdenker, N.L., Streicker, D.G., Faust, C.L., Carroll, C. R., (2014). Anthropogenic land use change and infectious diseases: a review of the evidence. *Trends Parasito.* **30**, 580–593.

[15]. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123.

[16]. Hatcher, E.L., Zhdanov, S.A., Bao, Y., et al., (2017). Virus variation resource: a comprehensive viral genome resource. *Nucl. Acids Res.* **45**, D482–D490.

[17]. Karagoz, A., Tombuloglu, H., Alsaeed, M., Tombuloglu, G., AlRubaish, A.A., Mahmoud, A., Smajlović, S., Ćordić, S., Rabaan, A.A., Alsuhaimi, E., (2023). Monkeypox (mpox) virus: classification, origin, transmission, genome organization, antiviral drugs, and molecular diagnosis. *J. Infect. Public Health* **16**, 531–541.

[18]. Knyazev, S. et al, (2021). Accurate and fast tree placement using UShER. *Nat. Commun.* **12**, 6413.

[19]. Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., (2006). Gard: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098.

[20]. Lam, H.M., Ratmann, O., Boni, M.F., (2018). Improved algorithmic complexity for the 3seq recombination detection algorithm. *Mol. Biol. Evol.* **35**, 247–251.

[21]. Li, J.Y., Wang, H.Y., Cheng, Y.X., Ji, C., Weng, S., Han, N., Yang, R., Zhou, H.Y., Wu, A., (2024). Comprehensive detection and dissection of interlineage recombination events in the SARS-CoV-2 pandemic. *Virus Evol.* **10**, veae074.

[22]. Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H. W., Ray, S.C., (1999). Full-length human immunodeficiency virus type 1 genomes from subtype c-infected seroconverters in india, with evidence of intersubtype recombination. *J. Virol.* **73**, 152–160.

[23]. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., (2018). MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944.

[24]. Martin, D.P., Varsani, A., Roumagnac, P., Botha, G., Maslamoney, S., Schwab, T., Kelz, Z., Kumar, V., Murrell, B., (2021). Rdp5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087.

[25]. National Center for Biotechnology Information, NCBI FTP website. URL: https://ftp.ncbi.nlm.nih.gov/. Last accessed Dec 8th 2025.

[26]. National Center for Biotechnology Information, 2004. NCBI Virus. URL: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/. Last accessed Dec 8th,2025.

[27]. Nextstrain, Nextstrain-maintained pathogen analyses. URL: https://nextstrain.org/pathogens#list. Last accessed Dec 8th 2025.

[28]. Nguyen, T.Q., Hutter, C.R., Markin, A., Thomas, M., Lantz, K., Killian, M.L., Janzen, G.M., Vijendran, S., Wagle, S., Inderski, B., et al., (2025). Emergence and interstate spread of highly pathogenic avian influenza a (h5n1) in dairy cattle in the united states. *Science* **388**, eadq0900.

[29]. O'Toole, A., Scher, E., Underwood, A., et al., (2021). Assignment of epidemiological lineages in an emerging pandemic using the pango nomenclature. *Nat. Microbiol.* **6**, 1403–1407.

[30]. P. Pinoli, A. Canakoglu, S. Ceri, M. Chiara, E. Ferrandi, L. Minotti, A. Bernasconi, 2023. Varianthunter: a method and tool for fast detection of emerging sars-cov-2 variants. Database 2023, baad044.

[31]. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., Du Plessis, L., Pybus, O.G., (2020). A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407.

[32]. Sayers, E.W., Cavanaugh, M., Frisse, L., Pruitt, K.D., Schneider, V.A., Underwood, B.A., Yankie, L., Karsch-Mizrachi, I., (2025). Genbank 2025 update. *Nucl. Acids Res.* **53**, D56–D61.

[33]. Shiraz, R., Tripathi, S., (2023). Enhanced recombination among omicron subvariants of sars-cov-2 contributes to viral immune escape. *J. Med. Virol.* **95**, e28519.

[34]. Shu, Y., McCauley, J., (2017). Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22**, 30494.

[35]. Song, H., Giorgi, E.E., Ganusov, V.V., Cai, F., Athreya, G., Yoon, H., Carja, O., Hora, B., Hraber, P., Romero-Severson, E., et al., (2018). Tracking hiv-1 recombination to resolve its contribution to hiv-1 evolution in natural infection. *Nat. Commun.* **9**, 1928.

[36]. Turakhia, Y., Thornlow, B., Hinrichs, A., McBroome, J., Ayala, N., Ye, C., Smith, K., De Maio, N., Haussler, D., Lanfear, R., et al., (2022). Pandemic-scale phylogenomics reveals the sars-cov-2 recombination landscape. *Nature* **609**, 994–997.

[37]. Z.J. Zhou, C.H. Yang, S.B. Ye, X.W. Yu, Y. Qiu, X.Y. Ge, VirusRecom: an information-theory-based method for recombination detection of viral lineages and its application on SARS-CoV-2. Briefings in Bioinformatics 24, 2023.

[38]. Ceri, S., Bernasconi, A., Canakoglu, A., Gulino, A., Kaitoua, A., Masseroli, M., Nanni, L., Pinoli, P., (2017). Overview of GeCo: a project for exploring and integrating signals from the genome. *In: International Conference on Data Analytics and Management in Data Intensive Domains*. Springer International Publishing, Cham, pp. 46–57. https://doi.org/10.1007/978-3-319-96553-6_4.

[39]. Bernasconi, A., Chiara, M., Alfonsi, T. and Ceri, S., 2024. SENSIBLE: implementing data-driven early warning systems for future viral epidemics. In CEUR WORKSHOP PROCEEDINGS (No. 3692, pp. 18-25). CEUR-WS. https://ceur-ws.org/Vol-3692/paper3.pdf.