# Automatic identification of recombination events in viruses

Tommaso Alfonsi[1], Anna Bernasconi[*,1], Matteo Chiara[2], and Stefano Ceri[1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. tommaso.alfonsi@polimi.it – https://orcid.org/0000-0002-4707-8425, anna.bernasconi@polimi – 0000-0001-8016-5750, stefano.ceri@polimi.it – 0000-0003-0671-2415
[2] Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy. matteo.chiara@polimi.it, https://orcid.org/0000-0003-3983-4961

[*]corresponding author

*Keywords*: viral genomics, recombination, Influenza A Virus, automatic detection, software.

**Abstract.** Viruses undergo change affecting their genome by several mechanisms, including point mutation and recombination. With the availability of large amounts of genome sequences and the implementation of genomic surveillance systems, the need for light-weight automatic pipelines surges. This work describes our RecombinHunt method, which has been extensively applied to the context of the SARS-CoV-2 pandemic. As a novel contribution, we observe the method applied to Influenza A viruses, where the relevance of recombination is unknown. Given the segmented structure of Influenza A genomes, most of the works on Influenza A concentrate on reassortments, in particular as a mechanism for facilitating inter-species spillovers. However, in this paper, we also show the existence of potential intra-segment recombinations; the relevance of recombination events in Influenza A viruses remains to be assessed, both quantitatively and in terms of impact.

## 1 Introduction

Recombination is an important mechanism of viral evolution; it requires co-circulation and co-infection of two different viral strains in the same host. Recombinant viruses have clinical and epidemiological relevance, as recombinant viruses were associated with enhanced virulence, host immune evasion, and resistance to antivirals [1].

All phylogeny-based approaches assume that the shared history of pathogens, isolated from different hosts, can be described by a branching phylogenetic tree. Recombination breaks this assumption and impacts the application of phylogenetic methods for the reconstruction of chains of contagion, viral evolution, and ultimately genomic surveillance of pathogens [2].

We developed RecombinHunt [3], an approach for effectively detecting recombinations, relying exclusively on data-driven methods. RecombinHunt's conceptual framework stems from a long-lasting tradition of statistical methods for detecting intragenic recombination; it starts from clusters of viral genomes in the form of a list of characterizing mutations.

Every target recombinant sequence is assessed by computing its similarity/dissimilarity with existing lineages/groups of similar genome sequences. RecombinHunt does not reconstruct phylogenies but computes the likelihood of a collection of pre-defined designations/lineages and their combinations (recombinants) based on the mutations in the target sequence. RecombinHunt identifies them as the "most likely candidates" for a recombinant sequence by using an algorithm that explicitly accounts for the frequency of each distinct point mutation.

The method exploits previously existing classifications; we systematically applied it to SARS-CoV-2 (exploiting the Pango lineage classification [4]) and demonstrated its use also in monkeypox (using the classification in [5]). RecombinHunt is technically applicable to any kind

of viral genome, provided that it can be represented as a genome sequence and a classification is present. When these classifications do not exist, they can be fabricated, e.g., using HaploCoV [6], a software framework for clustering viral sequences.

In this article, we show RecombinHunt at work on Influenza A virus; we analyzed data derived from the hemagglutinin segment of genomes assigned to the 6B* clade family of H1N1. Our dataset includes about 74K sequences retrieved from GISAID [7], collected until mid-2023; here, we observed a few cases of recombination.

## 2 The RecombinHunt Method

RecombinHunt accepts a target genome sequence as input in the form of a list of nucleotide mutations. Genome positions with a mutation in the target are denoted as the *target mutations-space*. Candidate "donor" and "acceptor" lineages are defined based on the counts of their mutations in the target mutations-space; we denote as "donor" the lineage with the higher count. For every lineage, the union of the lineage and target mutations-space is denoted as *extended target space*. A cumulative likelihood score is derived according to the following procedure: at each position of the extended target space, we compute the logarithmic ratio (see Figure 1A) between the frequency of the mutation in the lineage and in the complete collection of SARS-CoV-2 genomes (stored in a pre-computed matrix, see Figure 1B). This score is added if the mutation is shared by both the target and the lineage, whereas it is subtracted if the mutation is observed in the lineage but not in the target.

In detail, the method works as follows. For a target input sequence, likelihood ratio values are computed for all possible lineages and a ranking is prepared, see Figure 1C); the lineage (termed L1) associated with the maximum value is assigned to the target. If L1 mutations-space is *similar* to the target mutations-space (where similarity is assessed, for each virus, on the basis of the size of the extended target space and on the global likelihood of mutations), then the target is assigned to L1, and designed as non-recombinant; else, the target is designated as recombinant, and L1 is designated as the candidate donor. Note that L1 covers the majority of the mutations of the target, located in the genome segment that starts from one of the two ends (either 5' or 3') – denoted as L1's end – and reaches its maximum value at a position designated as max-L1 - see Figure 1D where the likelihood trend for the range $R_{start:end}$, peaking in max-L1 is depicted.

Upon the identification of a candidate donor, the one-breakpoint model (1BP) and the two-breakpoint model (2BP) are compared. Given the focus of this paper on influenza A viruses, examples of 1BP and 2BP refer to two recombination events for Influenza A, respectively shown in Figure 2, panels A and B. Figure 2A) refers to a viral sequence from Bosnia-Herzegovina sequenced in 2019. Note that the extended mutation space amounts to 247 mutations, and that the candidate donor (lineage 6B.1A.2) is recognized starting from the 5' end. The blue line indicates the likelihood ratio value for the "donor" lineage L1, 6B.1A.7, whose value grows from coordinate 247 (the 3' end) up to coordinate max-L1 = 32, reaching its maximum value. Then the target becomes dissimilar from L1 (it drops), the sequence is designated as recombinant, and the search for an "acceptor" lineage starts.

In the 1BP model, we search for a lineage L2, starting at the opposite end of the genome, and select the L2 lineage with the maximum likelihood ratio value (max-L2), designated as the candidate acceptor. The interval between coordinates max-L1 and max-L2 defines the *breakpoint range*, which is then reduced to a single position. In the example, the 6B.1A.2 lineage is recognized as "acceptor", as the orange likelihood curve starting from the 5' end grows and reaches its maximum value at coordinate max-L2 = 31.

In the 2BP model, the candidate donor L1 lineage is also assigned to the opposite end of the genome; we look for the point where $L1^{opp}$'s likelihood ratio is maximum, denoted as max-$L1^{opp}$.

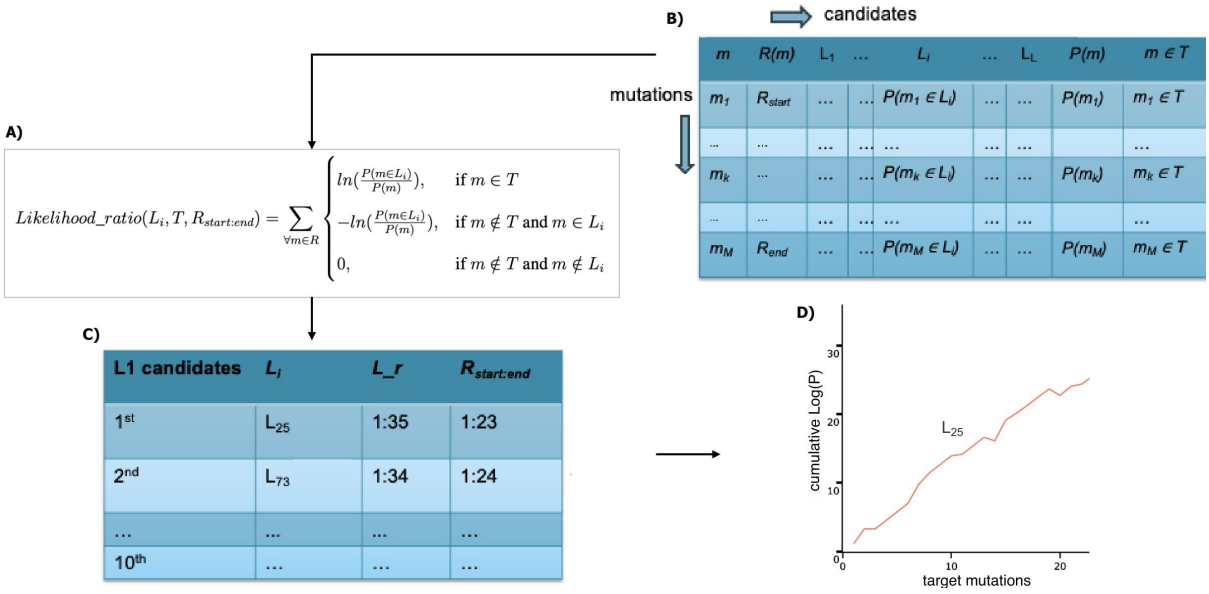


Figure 1: **Overview of the computational structures in the RecombinHunt method, including panels A) target-lineage likelihood formula, B) mutation-lineage probability matrix, C) lineage candidate ranking, and D) graphical likelihood trend**.

A candidate acceptor L2 lineage is searched in the space between max-L1$^{opp}$ and max-L1; the lineage L2 with the maximum likelihood ratio is selected, then two 'breakpoint ranges' are determined, which are then reduced to a single position. Figure 2B) refers to a viral sequence from Beijing, sequenced in 2018; the extended mutation space amounts to 244 mutations. Also in this case, the "donor" lineage L1 (6B.1A.7) is recognized starting from the 3' end; the likelihood value grows up to coordinate max-L1 = 66. By postulating the 2BP model, L1 is also assigned to the 5' end, and the likelihood value grows up to the max-L1$^{opp}$ = 17 coordinate. The "acceptor" sequence is searched in the 18-65 coordinate interval, and in particular, L2 is assigned to the B6.1A.5 lineage.

## 3 A Summary of SARS-CoV-2 and Monkeypox Results

Our method was executed on 51 of the 57 lineages designated as recombinant by Pango at the end of the COVID-19 pandemic emergency (April 2023). Two lineages were excluded as they had three breakpoints; other four lineages were disregarded since the defining Pango issue was unclear/controversial. The "ground truth", i.e., the description of the recombinant lineage in terms of donor, acceptor, and breakpoints, was reconstructed directly from the corresponding Pango designation issues [8]. RecombinHunt results were in complete agreement with the ground truth for 40 recombinant lineages (37 with one breakpoint and 3 with two breakpoint recombinations).

The remaining 11 lineages – which did not fully agree with the Pango designation – are stratified into three conceptually distinct groups: G1, G2, and G3. Six lineages in group G1 are not flagged as recombinant by RecombinHunt. For all these lineages, recombination is supported only by one or two mutations, over an average of 67 mutations considered in the respective *consensus-genome* (i.e., a lineage's ideal sequence, reconstructed as the mutations shared by $>$75% genomes of the lineage). In two lineages of group G2, RecombinHunt identified the same parent lineages, but additional breakpoints (2BP w.r.t. 1BP) compared with the solutions reported by the ground truth. The remaining three cases of group G3 are controversial; they are discussed in depth in [3]. The small discrepancies observed in our analyses might not necessarily

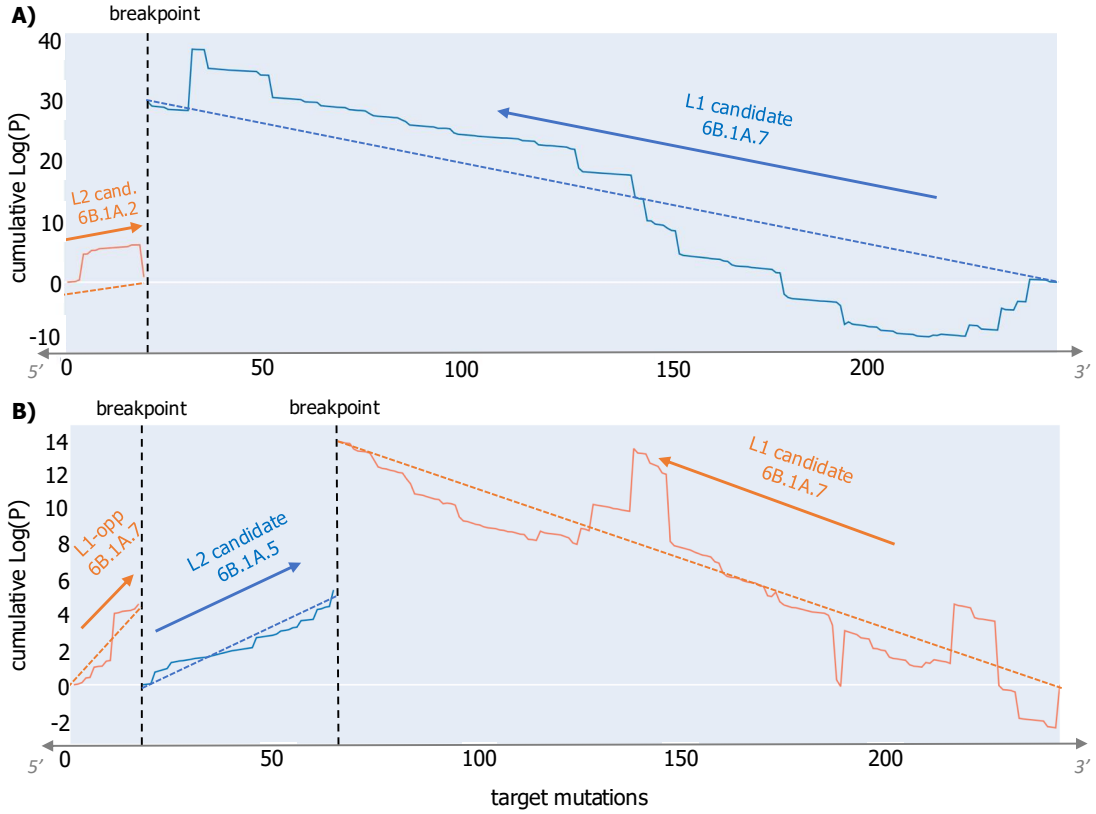


Figure 2: **Recombinant sequences for Influenza A viruses. Panel A) illustrates a 1BP recombination, panel B) illustrates a 2BP recombination**.

reflect errors and could be suggestive of intra-lineage heterogeneity and/or microevolution in some SARS-CoV-2 recombinant lineages.

Once applied to monkeypox, our method was able to replicate the classification of viral sequences indicated as recombinant in [5] by using a sophisticated ad-hoc method based on expert manual annotation. A large number of additional candidate recombinant genomes were also detected, suggesting previously unreported recombination events in monkeypox.

## 4 Application on Influenza Type A Virus

In this article, we apply RecombinHunt to the Influenza Type A Virus, considering the hemagglutinin segment from H1N1 genotypes (analysed in [9]). A classical letter on Nature [10] posed a first baseline in this line of research. Later, Boni et al. [11] suggested that recombination is lacking in human Influenza A Viruses, as other evolutionary mechanisms such as reassortment are typically preferred. However, subsequent results [12] again challenged this position.

We used RecombinHunt to search for evidence. We employed 73,744 sequences from GISAID collected from January 1st, 2007 to June 16th, 2023, divided into 15 clades (from the largely prevalent 6B.1* family and 6B.2) or 'unassigned'. Their sequence length is between 1690 nb and 1780 nb (99% sequences), average length of 1727 nb (standard dev. 34 nb), and mode of 1701 nb. Our dataset exhibited 8,923 unique mutations, with an average of mutations per sequence of 249 (standard dev. 37).

We indeed found some potentially recombinant HA segments. Based on the clades illustrated in Table 1A), the five recombinant sequences indicated in Table 1B) were identified, composed of two candidates, with 1BP or 2BP models. Sequences are identified by the collection location and year.

| A) Clade | Spread interval |
|---|---|
| 6B.1A.5a.2a.1 | 2022-present |
| 6B.1A.5a.1 | 2019-present |
| 6B.1A.5a.2 | 2020-present |
| 6B.1A.5a.2a | 2020-present |
| 6B.1A | 2016-2019 |
| 6B.1A.5a | 2018-2022 |
| 6B.1A.7 | 2018-2020 |
| 6B.1A.2 | 2018-2019 |
| 6B.1A.5b | 2018-2020 |
| 6B.1A.5 | 2018 |
| 6B.1A.1 | 2017-2019 |
| 6B.2 | 2015-2017 |
| 6B.1A.3 | 2017-2018 |
| 6B.1 | 2015-2018 |
| 6B.1A.6 | 2018-2019 |

| B) Accession ID | Location | Year | Recombinant candidates |
|---|---|---|---|
| EPI_ISL_1788918 | Beijing | 2018 | 6B.1A.7 + 6B.1A.5 + 6B.1A.7 |
| EPI_ISL_1255768 | Brisbane | 2018 | 6B.1A.1 + 6B.1A.7 + 6B.1A.1 |
| EPI_ISL_1583010 | Bosnia/Herzegovina | 2019 | 6B.1A.2 + 6B.1A.7 |
| EPI_ISL_1304355 | Illinois | 2018 | 6B.1A.1 + 6B.1A.5a + 6B.1A.1 |
| EPI_ISL_1545804 | St. Kitts | 2019 | 6B.1A.1 + 6B.1A.2 + 6B.1A.1 |

Table 1: **Overview of H1N1 analysis, with A) clades and their period of spread, and B) specific sequences recognized as recombinant by our approach using, respectively, two or three candidates**.

## 5 Conclusion

RecombinHunt is highly computationally efficient: the evaluation of the SARS-CoV-2 recombinant cases takes about 13 minutes on the GISAID dataset (15M sequences) using a laptop. The method is more accurate on large datasets, where classes are represented by a well-defined set of sequences and are well-separated from each other in the mutations-space. However, this does not prevent the application of RecombinHunt also to smaller datasets, with coarse-grained classification (see monkeypox). As a consequence, RecombinHunt is applicable to several viral pathogens, for which curated collections of genome sequences are available within Nextstrain [13] and for which a structured nomenclature has been defined by the respective reference community; these include, for example, dengue, RSV, Enterovirus D68, and West Nile virus.

In this work, we showed preliminary results of applying RecombinHunt to a dataset of H1N1 Influenza Type A viruses, useful to retarget an open research problem on whether this kind of virus considers intra-segment recombination in addition to reassortment. In future developments, we will test its use in integration with HaploCoV for viruses whose genome is expected to recombine (e.g., Respiratory syncytial virus and Zika). We are also addressing the problem of capturing viral reassortments with a data-driven approach, in particular by detecting macro-changes in the viral Influenza A genome, including spillovers. Our systems contribute to genomic surveillance at large, with automatic detection of potentially dangerous and sudden changes in human health-threatening viruses.

### Conflict of interests

The authors declare no conflicting interests.

### Acknowledgments

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We also thank Ilaria Capua for the initial discussion regarding our research on Influenza A viruses.

**Funding**

The work was supported by Ministero dell'Università e della Ricerca (PRIN PNRR 2022 "SENSIBLE" project, n. P2022CNN2J [14]), 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. PI A.B., co-PI M.C.


**Availability of data and software code**

The RecombinHunt updated software code is available at `https://zenodo.org/records/13349272`. The analyzed Influenza Type A dataset is at `https://doi.org/10.55876/gis8.250512ny`.

**References**

[1] Daniele Focosi, Fabrizio Maggi, Massimo Franchini, Scott McConnell, and Arturo Casadevall. Analysis of immune escape variants from antibody-based therapeutics against COVID-19: a systematic review. *International journal of molecular sciences*, 23(1):29, 2021.

[2] Yatish Turakhia, Bryan Thornlow, Angie Hinrichs, Jakob McBroome, Nicolas Ayala, Cheng Ye, Kyle Smith, Nicola De Maio, David Haussler, Robert Lanfear, et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature*, 609(7929):994–997, 2022.

[3] Tommaso Alfonsi, Anna Bernasconi, Matteo Chiara, and Stefano Ceri. Data-driven recombination detection in viral genomes. *Nature Communications*, 15(1):3313, 2024.

[4] Andrew Rambaut, Edward C Holmes, Áine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis Du Plessis, and Oliver G Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11):1403–1407, 2020.

[5] Ting-Yu Yeh, Zih-Yu Hsieh, Michael C Feehley, Patrick J Feehley, Gregory P Contreras, Ying-Chieh Su, Shang-Lin Hsieh, and Dylan A Lewis. Recombination shapes the 2022 monkeypox (mpox) outbreak. *Med*, 3(12):824–826, 2022.

[6] Matteo Chiara, David S Horner, Erika Ferrandi, Carmela Gissi, and Graziano Pesole. HaploCoV: unsupervised classification and rapid detection of novel emerging variants of SARS-CoV-2. *Communications Biology*, 6(1):443, 2023.

[7] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13), 2017.

[8] Pango. Designation GitHub Repository – Issues. https://github.com/cov-lineages/pango-designation/issues.

[9] Tommaso Alfonsi, Matteo Chiara, and Anna Bernasconi. A codon usage-based approach for the stratification of influenza a across recent spillovers. *Computational and Structural Biotechnology Journal*, 2025.

[10] William J Bean Jr, Nancy J Cox, and Alan P Kendal. Recombination of human influenza A viruses in nature. *Nature*, 284(5757):638–640, 1980.

[11] Maciej F Boni, Yang Zhou, Jeffery K Taubenberger, and Edward C Holmes. Homologous recombination is very rare or absent in human influenza A virus. *Journal of virology*, 82(10):4807–4811, 2008.

[12] Tommy Tsan-Yuk Lam, Yee Ling Chong, Mang Shi, Chung-Chau Hon, Jun Li, Darren P Martin, Julian Wei-Tze Tang, Chee-Keng Mok, Shin-Ru Shih, Chi-Wai Yip, et al. Systematic phylogenetic analysis of influenza A virus reveals many novel mosaic genome segments. *Infection, Genetics and Evolution*, 18:367–378, 2013.

[13] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[14] Anna Bernasconi, Matteo Chiara, Tommaso Alfonsi, and Stefano Ceri. SENSIBLE: Implementing Data-Driven Early Warning Systems for Future Viral Epidemics. In *CEUR PROCEEDINGS*, volume 3692, pages 18–25. CEUR-WS, 2024.