PLOS Computational Biology

RESEARCH ARTICLE

Accounting for Defective Viral Genomes in viral consensus genome reconstruction, application to influenza virus

Kévin Da Silva[1,2,3], **Nadia Naffakh**[4], **Marie-Anne Rameix-Welti**[1,2*], **Frédéric Lemoine**[1,2,3*]

1 Institut Pasteur, Université Versailles St-Quentin en Yvelines, Paris-Saclay INSERM UMR 1173 (2I), Université Paris Cité, Molecular Mechanisms of Multiplication of Pneumovirus, Assistance Publique des Hôpitaux de Paris, Paris, France, **2** Institut Pasteur, Université Paris Cité, National Reference Center for Respiratory Viruses, Paris, France, **3** Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France, **4** Institut Pasteur, Université Paris Cité, CNRS UMR3569, RNA Biology and Influenza Virus, Paris, France

* frederic.lemoine@pasteur.fr (FL); marie-anne.rameix-welti@pasteur.fr (M-AR-W)

OPEN ACCESS

Citation: Da Silva K, Naffakh N, Rameix-Welti M-A, Lemoine F (2026) Accounting for Defective Viral Genomes in viral consensus genome reconstruction, application to influenza virus. PLoS Comput Biol 22(7): e1014115. https://doi.org/10.1371/journal.pcbi.1014115

Editor: Ruslan Kalendar, University of Helsinki: Helsingin Yliopisto, FINLAND

Received: March 9, 2026

Accepted: June 23, 2026

Published: July 16, 2026

Data availability statement: DIPScan workflow is implemented in Nextflow [53], which makes it easily executable on many computing platforms, reproducible, and scalable. Each step uses software containers, which enables a high control of the software environment executed. DIPScan is freely available at https://github.

Abstract

In the context of viral epidemic surveillance, generating accurate consensus viral genomes from sequencing data is critical for tracking the emergence of mutations of concern, evaluating the genomic diversity of circulating viruses, and anticipating which viral strains could become most prevalent. However, this task is made difficult by the presence of Deletion-containing Viral Genomes (DelVGs), which contain truncated (or rearranged) and potentially mutated versions of the full length virus genome. Because these DelVGs can outnumber the full genome in terms of coverage, potential DelVG specific mutations may be erroneously incorporated into the final consensus, thereby compromising its accuracy. Automatic detection of these DelVGs and of the genomic positions that may harbor DelVG specific mutations is therefore crucial. Here, we present DIPScan, a new method able to (i) accurately and efficiently detect DelVGs in short read datasets, and (ii) mask or correct positions in the consensus genome that may be affected by DelVG-specific mutations. DIPScan achieves this through tailored metrics for breakpoint characterization and selection, linear modeling to estimate DelVG relative abundance from well-defined region and junction coverage, and efficient heuristic algorithms for reliable consensus sequence correction. Using several hundreds of simulated and real patient-derived NGS datasets from the National Reference Center (NRC) for respiratory viruses at Institut Pasteur, we demonstrate the capacity of DIPScan to accurately and efficiently detect DelVGs and to correctly adjust the consensus sequences. DIPScan is implemented as a Nextflow workflow, making it highly flexible, scalable, and reproducible, and is now used routinely at the NRC.

com/pasteur-cnrvir/dipscan along with its documentation, and in Zenodo (DOI:10.5281/zenodo.20511342). The code and the datasets used to test DIPScan are available at https://github.com/pasteur-cnrvir/dipscan_analysis/ and Zenodo (DOI:10.5281/zenodo.20511607). Real patient derived datasets analyzed in this study are available in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1426172. Simulated datasets are avaible on Zenodo (DOI:10.5281/zenodo.20511452).

Funding: This work has been partially funded by the European Union's Seq4Epi project (Grant Agreement Project 101113174 - SEQ4EPI EU4H-2022-DGA-MS-IBA-01-02) (to KDS, FL). We also thank the EU4Health Programme (grant number 101102733 DURABLE) for its support (to FL, KDS, MRW, NN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Over the past few years, viral genome sequencing has become a cornerstone of epidemic surveillance, and generating accurate consensus viral genomes is essential for this effort. However, viral samples frequently contain Deletion-containing Viral Genomes (DelVGs). These are truncated and sometimes mutated versions of the full-length genome, that may emerge during replication. Because DelVGs can outnumber the full genome in terms of coverage, they can compromise the quality of the resulting consensus sequences. Moreover, in short-read sequencing data, DelVGs are difficult to detect automatically and to distinguish from the full-length genome. To overcome this challenge, we developed DIPScan, an automated workflow that detects DelVGs in sequencing datasets, and refines the consensus genome by masking or correcting positions potentially affected by DelVG-specific mutations. We benchmarked DIPScan against existing tools and demonstrate its accuracy and efficiency on several hundred simulated and real patient-derived NGS datasets from the National Reference Center (NRC) for respiratory viruses at Institut Pasteur.

Introduction

Routine virological surveillance has been shown to be of utmost importance for monitoring viral epidemics, especially for (i) monitoring the emergence of mutations of concern (e.g., drug resistance mutations, or other advantageous mutations), (ii) monitoring the genomic diversity of circulating viruses and identifying the main circulating lineages, and (iii) predicting the phylogenetic clades that may be most prevalent in the near future.

Regarding these three topics, high throughput sequencing associated with efficient bioinformatics methods have been the cornerstone of virological surveillance in the past 15 years, and many efforts have been made to improve these methods.

Since the COVID-19 pandemics, from the bioinformatics point of view, many data analysis workflows have been developed and used routinely in surveillance laboratories to process virus sequencing data on a large scale and produce consensus sequences for the main circulating respiratory viruses (e.g., SARS-CoV-2, Influenza, RSV) and other non respiratory viruses (e.g., Ebola, Zika, etc.). Among these workflows, each with its unique aspects depending on the type of virus and data, notable examples include Viralrecon [1,2], wf-flu (https://github.com/epi2me-labs/wf-flu), ViReflow [3], ViralFlow [4], SARS2seq (https://github.com/RIVM-bioinformatics/SARS2seq), IRMA [5], ARTIC fieldbioinformatics (https://github.com/artic-network/fieldbioinformatics), to name a few.

These workflows usually work as follows: (i) They map the sequencing reads against a reference genome (the closest to the input samples) using a read mapper (e.g., bwa [6]), (ii) They look for variations between the sequencing reads and the reference genome (SNP calling) using tools such as Samtools [7] and iVAR [8], (iii)

They modify the reference according to the observed differences to obtain a consensus genome (e.g., using iVAR), in which each position consists of the nucleotide present in roughly more than half of the mapped reads at this position. Variants of this simple workflow exist, with additional steps such as reference selection, assembly, filtering, etc. Each step of these workflows is critical to build a reliable consensus sequence [9]. For example, selecting the right reference is of paramount importance, especially for very diverse viruses such as influenza viruses. Choosing the right parameters for SNP calling and consensus building also has a strong influence on the final consensus, such as minimal acceptable coverage, ambiguous nucleotide assignment, regions of the reference genome to mask, etc.

However, one often-overlooked characteristics of the samples can affect the reconstruction of a high-quality consensus sequence is the presence of Deletion-containing Viral Genomes (DelVGs) in the analyzed samples, that can produce defective interfering particles (DIPs) [10].

Definition and types of DIPs

DIPs have been initially discovered several decades ago in Influenza virus as non-replicative incomplete viruses [11] that can interfere with the replication of the non-defective homologous virus [12]. Initially, DIPs were thought to be primarily generated in cultured cells and considered insignificant in natural infections [13]. DIPs have since been detected in various virus families, including DNA viruses, RNA viruses, and retroviruses [14–16].

They consist of "viral particles that contain normal structural proteins but only a part of the viral genome" [14]. DIPs are generated during replication and are incapable of replicating independently, requiring the presence of a full-length (helper) virus to complement the missing sequences and corresponding proteins [15].

According to the definition of Vignuzzi and López [14], defective viral genomes (DVGs) can be classified in three main categories depending on how they were generated: (i) DVGs resulting from mutations in the viral genome (single mutation, frameshifts, or hypermutations); (ii) DVGs resulting from large deletions or rearrangements (i.e. DelVGs, the focus of this study), see Fig 1A, which are thought to form when the viral polymerase detaches from the template strand and resumes replication at a distant point, skipping an large internal genomic region; and (iii) DVGs resulting from snap-back or copy-back, resulting in the duplication of a sequence in reverse complement.

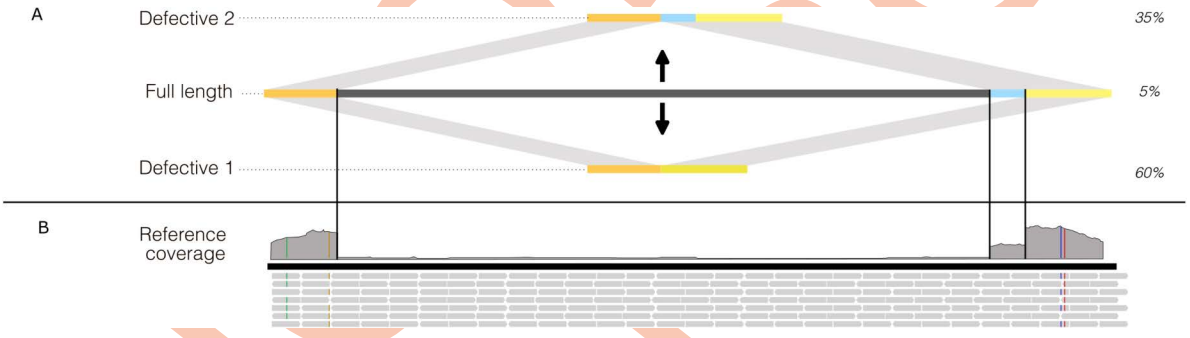


Fig 1. Challenge posed by defective sequences. This schematic illustrates a virtual sample containing two distinct defective sequences, in addition to the full length sequence. **A.** The composition of the two DelVGs, each resulting from the deletion of a unique genomic region. In this virtual scenario, *DelVG 1* constitutes the majority at 60% of the sample sequences, *DelVG 2* accounts for 35%, and the full-length, functional sequence, represents the remaining 5%. **B.** Alignment of sequencing reads to the reference genome, and the resulting coverage profile. Elevated coverage at the termini, accompanied by sharp drops (one near the 3' end and two near the 5' end), indicates the presence of two types of DelVGs. Due to the substantial overrepresentation of DelVGs compared to the full-length genome, mutation frequencies at the extremities must be interpreted with caution, as they likely reflect variants in the defective sequences rather than in the full-length genome.

https://doi.org/10.1371/journal.pcbi.1014115.g001

Mechanisms of production and role of DelVGs

In influenza virus DelVGs, deletions are mainly found in the polymerase segments (PB1, PB2, PA). This was previously thought to be related to the greater length of these segments, but this has not been conclusively confirmed [10]. Moreover, deletion hotspots have been shown to be located at the segments' termini, within the first 400 nucleotides after initiation, for reasons that are not totally elucidated so far, but maybe related to the presence of A/U rich sequences, direct repeats, or to the formation of a t-loop RNA structure, although with no specific sequence motif [10,17–19].

While it has been hypothesised that internal deletions can occur through homologous recombination in positive-strand RNA viruses, such recombination events are rare in negative-strand viruses such as influenza viruses. Recent studies suggest that DelVGs production may be a process regulated by viral factors or host factors. Regarding viral factors, the structure of viral ribonucleoproteins, which could influence break and rejoin sites [15,20], and, possibly, mutations on the influenza polymerase [21], may be linked to deletions. Also, specific sequences in the RSV genome have been shown to regulate the formation of certain DVG types [22]. Regarding host factors, some cell types seem to produce low amount of DIPs [23,24].

DVGs have been proposed to have several key roles. Their primary characteristic is the ability, upon co-infection with a full length virus, to interfere with its replication by competitive inhibition and to accumulate while the proportion of non-defective particles decreases, leading to the term "Defective Interfering Particles" [25]. Beyond this, they have been suggested to promote viral persistence in vitro, to modulate the host innate immune response [26,27], and to contribute to viral evolution by providing a diverse sequence pool for recombination or reassortment with the standard genome [14,16].

The significance of defective viral genomes is also well-established in influenza virus research more specifically. First, there is accumulating evidence that these defective genomes influence disease severity, and may contribute to viral persistence [24,28]. Vasilijevic *et al.* correlated a low abundance of DIPs in patients infected with influenza virus with severe disease outcomes [21], while Penn *et al.* suggested that DIP levels during infection influence H5N1 pathogenesis in mice [29].

These properties have prompted the consideration of defective viral genomes as antivirals [30–32] or vaccine adjuvants [14].

DelVGs have been more extensively studied for influenza A than influenza B. However, key differences have been identified between the two types. For instance, type B samples exhibit a greater number of unique junctions in the NA and M segments compared to type A samples [17]. Additionally, recent findings indicate that the length of 5'- and 3'- terminal sequences of DelVGs differ between influenza A and B, with A showing a more pronounced asymmetry between the lengths of the 5'- and 3'-end sequences [33].

How DelVGs impair consensus reconstruction

A standard method for reconstructing consensus viral genomes from sequencing data involves mapping sequencing reads to a reference sequence and adjusting the reference sequence based on observed variations. However, this approach can be problematic when defective viral genomes, especially DelVGs, predominate in the sample over full-length genomes. In such cases, read mapping (as shown in Fig 1B) reveals high coverage at the reference genome's termini (present in both defective and full-length genomes) but low coverage in the central region (present only in the minority full-length genome). Furthermore, defective particles may carry unique mutations that are absent in the full length genome, as well as in the reference genome. Since these mutations appear in most reads, they are often incorrectly incorporated into the final consensus genome.

Overlooking DelVGs during consensus reconstruction can thus lead to a chimeric sequence that misrepresents the true viral population in the sample. The improper integration of mutations may introduce severe errors, such as premature

stop codons, or frameshifts. Fig 1 illustrates this challenge, depicting how DelVGs with distinctive mutations can distort the reconstructed consensus genome when present in sequencing data.

One possible approach to addressing DelVGs in sequencing is to manually review read alignments and exclude affected segments (in segmented viruses like influenza) or discard entire samples (in non-segmented viruses). However, this method is labor-intensive and poses the risk of losing important surveillance data. Some opt to retain the original consensus sequence under the assumption that DelVGs are rare and negligible in clinical samples. Yet, as demonstrated by Saira *et al.* [34], defective genomes frequently appear in Influenza A/H1N1pdm patient samples (13 over the 26 studied), making their proper handling essential.

Existing tools to study DIPs/DelVGs

The generation of accurate consensus sequences necessitates the automated detection of DelVGs in sequencing data. This comes down as identifying and correcting problematic sites (i) where read coverage is higher in the defective genome relative to the full-length genome, and (ii) that have mutations specific to the defective genome.

While existing tools were designed for DelVG detection, DIPScan uniquely integrates detection, quantification, and consensus correction. This positions our method as the first to address the challenges posed by DelVGs in the context of viral genome reconstruction.

Other tools, such as DI-TECTOR [35], ViReMa [36,37], DG-Seq [19], DVG-profiler [38], VODKA [22], or VODKA2 [39] have been developed over the past decade to identify defective genomes in sequencing datasets. However, while these tools can detect various types of defective genomes, (i) they are not suited for large-scale, automated consensus construction in routine settings, and (ii) recent studies have demonstrated low agreement between them when identifying deletion junctions [40].

DI-TECTOR relies on bioinformatics tools that are not adapted to current data production scales, making it impractical for modern dataset sizes. Moreover it is designed primarily for defective genome detection rather than the generation of a clean final consensus genome. ViReMa was developed to detect recombination events in general. While it is designed to map split reads and therefore detect large deletions, it is not designed to clean or correct the full length consensus sequence. DVG-profiler identifies reads that do not align perfectly to the reference. For each read with multiple alignments, it tries to pair alignments that could represent two segments of a single DelVG read. Orientation of the alignment pair determines the type of defective genome, and the number of reads spanning a junction compared to the average full-length coverage determines the defective/full-length ratio. Using only junction-spanning reads may underestimate the proportion of defective genome. Additionally, it is not designed to clean or correct the full length consensus sequence present in the sample. DG-Seq was developed to detect and quantify influenza defective genomes, but not to correct the full length consensus. As for VODKA, it was developed to specifically identify copy back defective genomes, which are not our primary focus, occur less frequently, and make the detection of interesting regions to correct more difficult. Its update, VODKA2, increases its accuracy and speed, and expands its functionality to the detection of delVGs. However, VODKA and VODKA2 are designed to detect and count junction reads, they do not provide DelVG quantification nor full-length consensus correction.

Here, we introduce DIPScan, a bioinformatics workflow specifically designed to identify deletion-containing viral genomes (DelVGs) resulting from large deletions within Illumina sequencing datasets and, when necessary, correct consensus sequences by introducing the likely nucleotide of the full length genome or by masking potentially problematic positions at these positions.

DIPScan requires the following inputs: (i) an Illumina sequencing read file in FASTQ format, (ii) a reference genome in FASTA format, and (iii) an optional precomputed consensus sequence also in FASTA format. The workflow generates two key outputs: (i) a report indicating the presence or absence of DelVGs along with the proportion of internal deletions for each genomic segment, and (ii) a corrected consensus sequence.

Materials and methods

DIPScan workflow

DIPScan features seven primary computational steps shown in Fig 2 and described below: mapping, snp calling, extraction of deletion boundaries, metrics computation, breakpoints selection, DelVGs proportion estimation, and consensus correction.

Mapping split reads. DIPScan begins by mapping the sequencing reads to the user-provided reference genome. This is achieved using a two-step approach that accounts for split reads spanning large deletions. In the first step, the reads are mapped to the reference genome using BWA-MEM2 [41]. From the resulting alignment, only the reads that map with at most 8 clipped nucleotides (unmatched read extremities) are kept. This produces a classical alignment, containing reads that map perfectly, as well as those containing SNPs or short insertions/deletions and less than 8 clipped positions.

In the second step, all unmapped reads and those that map with at least 8 clipped nucleotides are extracted, as they can define split regions in the case of DelVG. These reads then undergo STAR alignment [42] (Spliced Transcripts Alignment to a Reference, v2.7.11a). Although STAR was originally designed for aligning RNA-Seq reads across splicing sites, it is also highly effective at accurately mapping split reads flanking large deletions (with specific run parameters, see Text A in S1 Appendix), with read portions on either side of the deleted sequence, that are considered as "non-canonical splice sites". This makes STAR well-suited to identifying defective genomes of interest, as these split reads often indicate large-scale structural variations within the genome, such as deletions, insertions, duplications, or translocations.

The outputs of these two steps are merged together in a single mapping file (BAM format) containing full reads mappings and split read mappings.

Extraction of deletion boundaries. The mapping file is then used to identify potential deletion boundaries by extracting reads containing "skipped regions" (denoted by 'N' in the CIGAR string of the SAM format) of more than 150 nucleotides. These "skipped reads" each span deleted regions, defining start and end coordinates relative to the reference genome. The resulting position pairs are consolidated into a non-redundant set for downstream analysis.

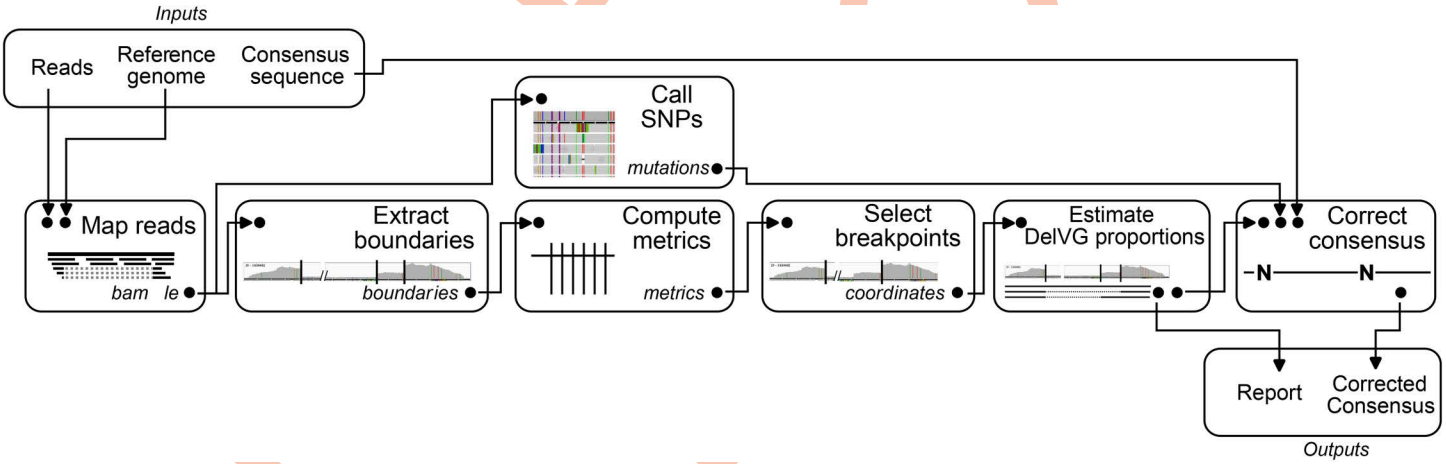


Fig 2. DIPScan workflow. A simplified schematic representation of the DIPScan workflow for identifying DelVGs and correcting the consensus sequence. The pipeline requires an input read file, a reference genome, and an optional user-supplied consensus sequence. Processing consists in seven main steps: (i) read mapping, (ii) SNP calling, (iii) boundary extraction, (iv) metric computation, (v) breakpoint selection, (vi) DelVG proportion estimation, and (vii) consensus sequence correction (if an input consensus sequence is provided). The final output includes two files: a report of identified DelVG boundaries, and the corrected consensus sequence.

https://doi.org/10.1371/journal.pcbi.1014115.g002

Computing defective metrics. Using this set of unique start-end position pairs, various metrics are calculated. For each pair, these metrics include:

- *Supporting reads*: The number of reads spanning the exact deletion coordinates.
- *Split frequency*: The proportion of *supporting reads* relative to the total number of split reads (at any deletion coordinate). This value indicates the relative frequency of this particular deletion compared to other deletions.
- *Total frequency*: The proportion of *supporting reads* relative to the total number of mapped reads. This value indicates the relative frequency of this particular deletion compared to the whole sequencing dataset.
- *Expected minimum frequency*: The expected *Total frequency* if the proportion of DelVG was 2%. This value corresponds to 2% of read length divided by the non-deleted region size (reference genome size minus deletion length).
- *Deletion start ratio*: The local deletion-start prevalence at a breakpoint is quantified as the ratio of mean coverage over the 5 nucleotides downstream to the mean coverage over the 5 nucleotides upstream of the breakpoint start.
- *Deletion end ratio*: The local deletion-end prevalence at a breakpoint is quantified as the ratio of mean coverage over the 5 nucleotides upstream to the mean coverage over the 5 nucleotides downstream of the breakpoint end.

Breakpoint selection. In order to filter out lowly prevalent breakpoints that may add noise in the data, the following default filters (though adjustable depending on the sequencing depth) are applied to ensure the robustness of detected breakpoints:

- *Supporting reads* should be greater than 100 reads. This criterion eliminates background noise.
- *Total frequency* should be greater than the *Expected minimum frequency*. This removes breakpoints that are poorly represented relative to the entire dataset.
- Both *Deletion start ratio* and *Deletion end ratio* must be smaller than 1. This removes meaningless breakpoints having inner-coverage higher than outer-coverage.

The filtered list may include multiple breakpoints that delineate various deleted regions and, notably, several DelVG sequences.

DelVG proportion estimation. The potential risk posed by DelVGs in viral consensus reconstruction is particularly significant when DelVGs are dominant relative to the full-length genome. In such cases, a potential mutation may become dominant and subsequently incorporated into the final consensus sequence. Estimating the proportion of each potential DelVG is therefore important.

For a given segment, the n previously selected unique breakpoint positions (either starts or ends) divide the segment into $n+1$ distinct regions. The median nucleotide coverages, denoted as $cov_1, \ldots, cov_{n+1}$, are computed for all the regions $r_1, \ldots, r_{n+1}$. Because coverage frequently declines at the edges of the first and last regions, the median coverage is calculated after excluding the first (or last) 150 nucleotides of these regions, respectively. If the first (or last) region is shorter than 150 nucleotides, the median coverage of the last (or first) region is used instead. If both regions are shorter than 150 nucleotides, the higher coverage value between the two is selected.

Given that the n breakpoint positions represent d distinct DelVGs (some breakpoints starts or ends may be shared across multiple DelVGs), our objective is to estimate the relative abundances $\{c_1, ..., c_d, c_{d+1}\}$ of each DelVG $(1, \ldots, d)$ and of the full-length segment $(d+1)$ in the sample. These estimates should best explain the observed coverages $cov_1, \ldots, cov_{n+1}$ across the $n+1$ regions.

This estimation is formulated as a system of $n+1$ linear equations, where each equation models the coverage c_r of a region r as a weighted sum of the contributions from all sequences (DelVGs and the full-length segment):

$$cov_r = \sum_{i=1}^{d+1} c_i \times l_{ir}$$

Here, $l_{ir} = 1$ if region r is present in sequence i, and 0 if deleted, while for the full length sequence ($i = d+1$), $l_{ir} = 1$ for all regions.

Beyond coverage of all regions, we also leverage split-read counts at each breakpoint, where each split-read count s_i represents the number of reads spanning the breakpoint present in DelVG i. Using these split-reads counts, we calculate the relative proportion p_i of each DelVG compared to all DelVGs:

$$p_i = \frac{s_i}{\sum_{j=1}^{d} s_j}$$

To ensure consistency, the estimated DelVG abundances $c_1, \ldots, c_d$ should also satisfy:

$$p_i = \frac{c_i}{\sum_{j=1}^{d} c_j}$$

This constraint can easily be reformulated as:

$$(1 - p_i)c_i - \sum_{j \in \{1,\ldots,d\}, j \neq i} p_i c_j = 0$$

yelding d additional equations that are integrated into the system. An illustrative example with 3 DelVGs is given in Fig A in S1 Appendix. Because these equations have no exact solution (no abundance coefficients satisfy all constraints), we estimate the sequence abundances that best match the data using Non-Negative Least Squares (NNLS) [43], as implemented in Scipy (v1.15.3).

After solving, the abundance coefficients are normalized by their sum to obtain the proportions of each DelVG and the full-length segment.

A sample is flagged as containing relevant DelVGs if the sum of DelVG proportions (excluding the full length segment) exceeds 50%.

Consensus correction. Consensus correction relies on a predefined list of candidate mutations. To generate this list, DIPScan first identifies potential variant positions relative to the reference genome using iVar (v1.3) [8], applied to a filtered BAM file from which noisy split reads (those not mapping to selected breakpoints) have been removed. This produces a tabulated output containing all candidate mutation sites, along with comprehensive details for each, such as the mutation type (SNP, INDEL, reference/alternative nucleotides), the number of reads supporting the alternative allele, the total read coverage at the position, the statistical confidence metrics for the mutation, etc. This information then feeds into the consensus correction pipeline, which comprises the following steps, as illustrated in Fig 3:

1) Identification of regions to be corrected The breakpoints identified previously are used to define the maximal region that may harbor dominant DelVG mutations and therefore necessitates correction. This maximal region is composed of two parts:

- The *start region* (orange sequence on Fig 3): spanning from position 0 to the right-most position among the start positions of the breakpoints.
- The *end region* (blue sequence on Fig 3): extending from the left-most position among the end positions of the breakpoints to the end of the reference segment.

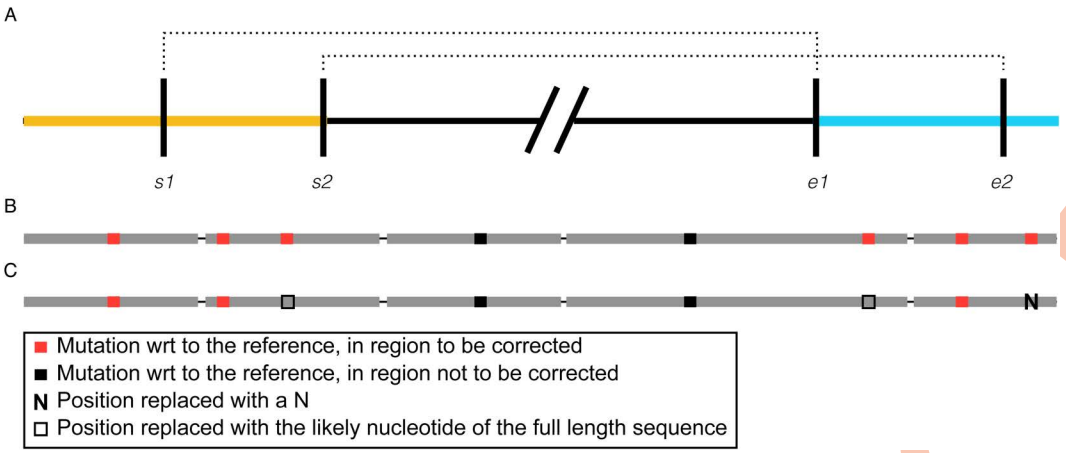


Fig 3. Schematic representation of the consensus sequence correction process. A. Defining regions to be corrected. Based on two previously identified breakpoints (coordinates [*s*1,*e*1] and [*s*2,*e*2]), two regions are flagged for potential correction: the start region (in orange) spans from coordinate 0 to the rightmost breakpoint start (*s*2), and the end region (blue) spans from the leftmost breakpoint end (*e*1) to the end of the reference genome. The central sequence (in black), with coordinates [*s*2,*e*1] is protected from correction as it is presumed to come exclusively from reads originating from the full length genome. **B.** Mutation detection. The initial consensus sequence is aligned against the reference genome. Mutations that fall within the designated correction regions are identified (highlighted in red). **C.** Correction of the identified positions in the consensus. Each identified mutation is resolved through one of the following ways: (i) replacement with an ambiguous character 'N', (ii) retention of the mutation when it is presumed to originate from the full length, or (iii) substitution with the probable full-length nucleotide when the mutation is attributed to a DelVG, the latter determined by comparing its frequency to the coverage of the DelVG sequence.

https://doi.org/10.1371/journal.pcbi.1014115.g003

2) Selection of mutated positions DIPScan selects positions located within either the start or the end region of the genome (in red in Fig 3) that may need correction. To do so, the initial consensus sequence is aligned against the reference genome using MAFFT [44] (v7.525 -auto) in order to match the positions of the user-provided consensus with the positions of the reference genome. Then, using the alignment along with the called mutations from iVAR output, DIPScan selects positions that are covered by at least 10 reads, and specifically for the positions of the consensus with SNPs, also those that are affected by a statistically significant mutation (with a $p.value \leq 0.05$), either major or minor, as computed by iVAR.

3) Choosing the right nucleotide At this final step, for a given mutated position, we aim at associating each observed nucleotide to one or several sequences (DelVGs and full length), and use this information to incorporate the nucleotide associated to the full length sequence in the consensus. This can be done by leveraging the estimated frequency of each sequence (DelVG and full length), and the frequency of each nucleotide at the given position. But first, we can already resolve straightforward cases:

- If the estimated proportion of the full-length genome is below a mutation threshold m_t ($p_{fl} \leq 1 - mt$, by default $m_t = 0.98$), then the genome is considered too low in frequency to confidently determine the correct nucleotide at any position called by iVAR, and an N is incorporated at all the mutated positions in the consensus.
- If the mutation frequency f_{mi} falls within the ambiguous range m_{ar} (by default, $m_{ar} = [0.45, 0.55]$), the mutation cannot be reliably identified, and an N is incorporated in the consensus.
- If $0.55 < f_{mi} < m_t$ (major mutation) and $p_{fl} \geq 0.5$ (the full length sequence is dominant), then the mutation is retained in the consensus.
- If $1 - m_t < f_{mi} < 0.45$ (minor mutation) and $p_{fl} < 0.5$ (the full length sequence in not dominant) and the frequencies of full-length (p_{fl}) and DelVG (p_{dip}) variants do not overlap (*i.e.*, $|p_{fl} - p_{dip}| > \delta$, with $\delta = 0.1$ by default) and the frequency of the

mutation, f_{mi}, is compatible with the full length frequency p_{fl} (*i.e.*, $|p_{fl} - f_{mi}| <= \delta$, with $\delta = 0.1$ by default), then the minor mutation is incorporated in the consensus sequence.

All other cases correspond to more complex scenarios, such as the presence of multiple DelVGs along with the full length sequence, where sequence frequencies must be combined to match to the frequency of a given nucleotide. In this scenario, for a given mutated position, we try to match the detected sequences (DelVGs and full-length) with the possible nucleotides found in the reads at this position, such as the sum of the sequence frequencies associated to a given nucleotide is approximately equal to the frequency of the nucleotide. This corresponds to a variant of a well-characterized problem, known as the "Multiple Subset Sum", and can be formulated as "For each label (nucleotide), identify a subset of objects (sequences) whose combined frequencies match the label's frequency". While this is an NP-hard problem, heuristics (such as simulated annealing or MCMC) can be used to find solutions. Moreover, since the number of sequences and unique nucleotides is very small in our case, a solution is findable very quickly. The solution is then used to assess whether the full length sequence can be associated unambiguously with one nucleotide (the nucleotide is then incorporated in the consensus) or not (a N is incorporated in the consensus).

Additionally, to prevent misaligned reads from being mistaken for mutations, we compare the read depth at each potential mutation site with the expected depth accounting for the full-length genome and all defective genomes spanning the position). If the mutation is in a lowly covered region compared to the expectation (the ratio of the position depth and the expected depth is less than 80%), then the mutation is considered unreliable and an N is incorporated.

Testing datasets

DIPScan was tested on two datasets: (i) A dataset composed of 110 simulated influenza virus sequencing data corresponding to several scenarios, and (ii) a dataset composed of 551 routine sequencing of influenza samples from the NRC for Respiratory Viruses at Institut Pasteur.

Simulated dataset. To evaluate DIPScan in a controlled setting, we simulated sequencing datasets containing DelVGs (see Fig 4). We started with 11 influenza consensus genomes (including all segments) from the real influenza dataset described in the next section, representing various subtypes. These genomes are referred to as *Full Length True Genomes*. We then randomly removed internal regions from one segment (either PB1, PB2, or PA) of these genomes, to create *DelVG True genomes*. Next, we introduced between 1 and 10 new mutations into each *DelVG True genome*, making them distinct from their associated *Full Length True Genome*. These modified genomes are called *DelVG mutated Genomes*. For each *Full Length True Genome* and its corresponding *DelVG Mutated Genome*, we simulated 10 datasets using ReSeq [45] (command reseq illuminaPE –maxFragLen 500) with a full-length/DelVG ratio (based on average coverage per position) ranging from 100% to 10%. This approach allowed us to test both the sensitivity and the precision of DIPScan. This resulted in 110 datasets, further described on Text B and Table A in S1 Appendix.

In parallel, we built two low coverage simulated datasets by sub-sampling each of the 110 simulated samples with a reduced coverage of 10% (500k reads) and 1% (50k reads) of the original sample. These datasets are used to assess the performances of DIPScan on extreme cases (which do not reflect a common occurence, given the short length of the viral genomes, and which may not even be adequate to study viral consensus in the presence of a high coverage DelVG).

Real influenza virus dataset. We applied DIPScan on 551 real Influenza virus sequencing dataset sequenced at Institut Pasteur's NRC (see Text C in S1 Appendix). It comprises 551 samples collected between 24/02/2025 and 04/04/2025, and consisting of: 125 A/H1N1pdm (22.7%), 232 A/H3N2 (42.1%), and 194 B/Victoria (35.2%). All these samples were visually inspected, to search for DelVG patterns (high coverage on the extremities and low coverage in the central region). Among these, 55 A/H1N1pdm (44%), 74 A/H3N2 (31.9%) and 103 B/Victoria (53.1%) showed patterns typical of DelVGs resulting from large deletion.

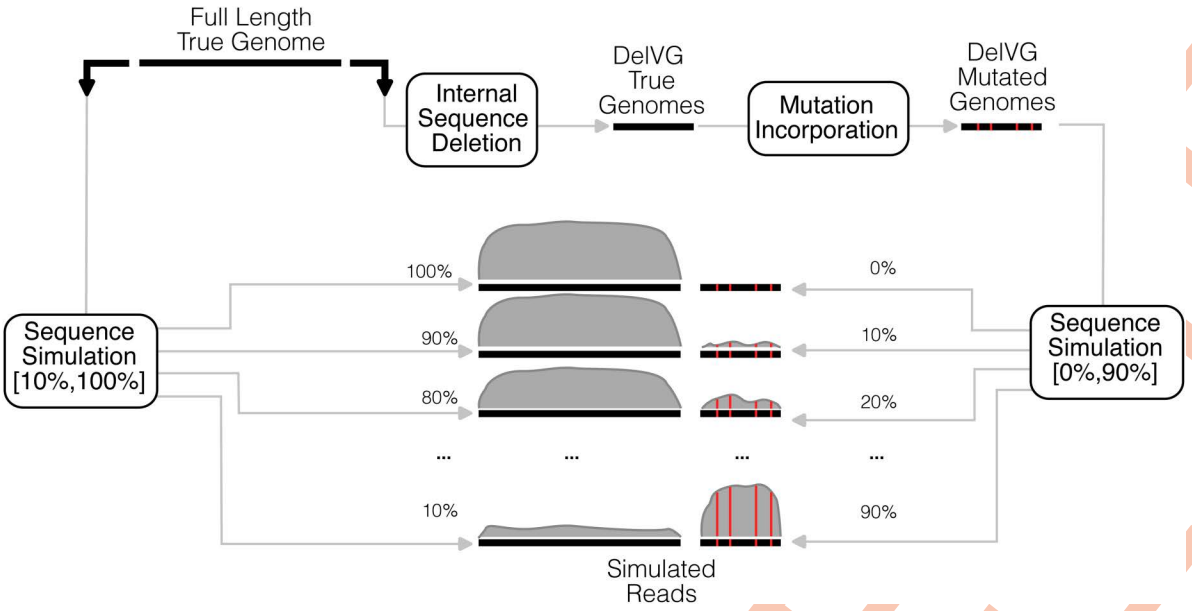


Fig 4. Simulation workflow. Schematic illustration of the simulation workflow for generating mixed full-length and defective genome samples. Starting from a known full-length genome (with 8 segments), an internal sequence of one segment is deleted to create a DelVG sequence, and 1 to 10 random mutations are introduced. Sequencing reads are then simulated from mixtures of the full-length and DelVG genomes at nine distinct ratios(from 100%:0% to 10%-90% full-length:DelVG), yielding ten simulated samples per starting consensus genome.

https://doi.org/10.1371/journal.pcbi.1014115.g004

Results

A. Overview of the prevalence of DelVG sequences in samples

In large-scale sequencing efforts, which often process thousands of samples annually, DelVGs can introduce errors into consensus sequences, leading to their erroneous inclusion in public databases. For instance, during the 2023–2024 season, the National Reference Center (NRC) for Respiratory Viruses at Institut Pasteur Paris sequenced 1,767 influenza samples, of which 540 (30.6%) contained one or more defective segments upon manual review. When restricting the analysis to high-quality sequences eligible for database submission, 471 out of 1,529 samples (30.8%) still exhibited defective segments. A subtype-specific breakdown revealed that defective segments were present in 31.8% of A/H1N1pdm, 30.6% of A/H3N2, and 21.2% of B/Victoria samples, proportions that remained consistent (32.2%, 30.2%, and 21.8%, respectively) when considering only high-quality sequences. Further examination at the segment level demonstrated variability in DelVG prevalence: PB2 had the highest rate (20.1% of high-quality samples), followed by PA (17.7%) and PB1 (14.6%), while HA and NP showed very low rates (0.07% each). No DelVGs were observed in the MP, NA, or NS segments.

These findings underscore the critical need for automated DelVGs detection tools, such as DIPScan, to ensure the accuracy of produced consensus sequences.

B. Simulated datasets

To evaluate the capacity of DIPScan to accurately detect DelVGs and correct the reconstructed consensus genome, we first applied it on the simulated dataset described in the methods section. This dataset consists in 110 virtual influenza virus sequencing datasets, each comprising a mix of reads simulated from one of the 11 original consensus genomes and reads from a corresponding mutated, partially deleted genome segment, with proportions varying from 0 to 90%.

We executed DIPScan on each of the 110 datasets using the default options and assessed its capacity to detect DelVG breakpoints along their proportion, and to correct the consensus sequence. To do so, we computed the following metrics: (i) the detected breakpoints compared to the expected ones, (ii) the estimated DelVG proportion compared to the known proportion in each sample, and (iii) the number of mutations effectively masked or corrected in the final consensus.

Breakpoint detection and tool comparison. The first aspect of DIPScan we assessed was its ability to precisely identify DelVG breakpoints within input datasets. Since ViReMa, DG-Seq, and VODKA2 are specifically designed to detect these breakpoints, rather than estimate DelVG proportions or refine consensus sequences, this presents an ideal opportunity to benchmark DIPScan's performance against these tools. For this comparison, we applied DIPScan, ViReMa, VODKA2, and DG-Seq to each simulated dataset and assessed their performance by matching the identified breakpoints with the ground-truth breakpoints, i.e., those predefined in the simulations. We ran all tools with their default parameters and analyzed their outputs without further processing. Consequently, our analyses incorporated each tool's built-in filtering criteria. For example, DIPScan by default only reports breakpoints separated by more than 150 nucleotides. Hereafter, breakpoints reported by the tools that coincide with simulated breakpoints are considered true positives; reported breakpoints without simulated matches are considered false positives; and simulated breakpoints absent from tool outputs are considered false negatives.

As shown in Fig 5, ViReMa, DG-Seq, and VODKA2 detected 1,060, 582, and 270 breakpoints, respectively, of which only 46 (4.3%), 99 (17%), and 64 (23.7%) were true positives. In contrast, DIPScan identified 93 breakpoints, all of which were correct, while missing 6 true breakpoints. Among these missed breakpoints, four originated from low-frequency DelVGs (10%), one from a 20% DelVG, and one from a 30% DelVG, likely due to inaccurate DelVG proportion estimates or insufficient breakpoint read support, below the detection threshold. These 6 breakpoints do not constitute a problem in terms of consensus correction.

DG-Seq achieved perfect sensitivity (100%), detecting all true junctions, but at the cost of 483 false positives, resulting in low precision (17%), a high false discovery rate (83%), and an F1-score of 29.1%. By comparison, DIPScan demonstrated superior performance across key metrics, with perfect precision (100%), no false discoveries (FDR: 0%), and a global F1-score of 96.9%.

It is worth noting that all methods except DIPScan incorrectly detected many breakpoints in the 0% DelVG simulation, which contained no true breakpoints. While specific filtering criteria could eliminate many false positive breakpoints, optimization of these thresholds falls outside the scope of this study.

DelVG proportion estimation. On the same simulated data, DIPScan accurately estimates the proportion of DelVG sequences. Correlation between the known ratio and the estimated ratio is high (Fig 6A, Pearson correlation of 0.99 between average estimated ratios over the 11 samples at each known ratio and the known ratio).

Except for two samples (see Fig B in S1 Appendix), the estimations were slightly under estimated compared to the known proportion, although close to the expectation.

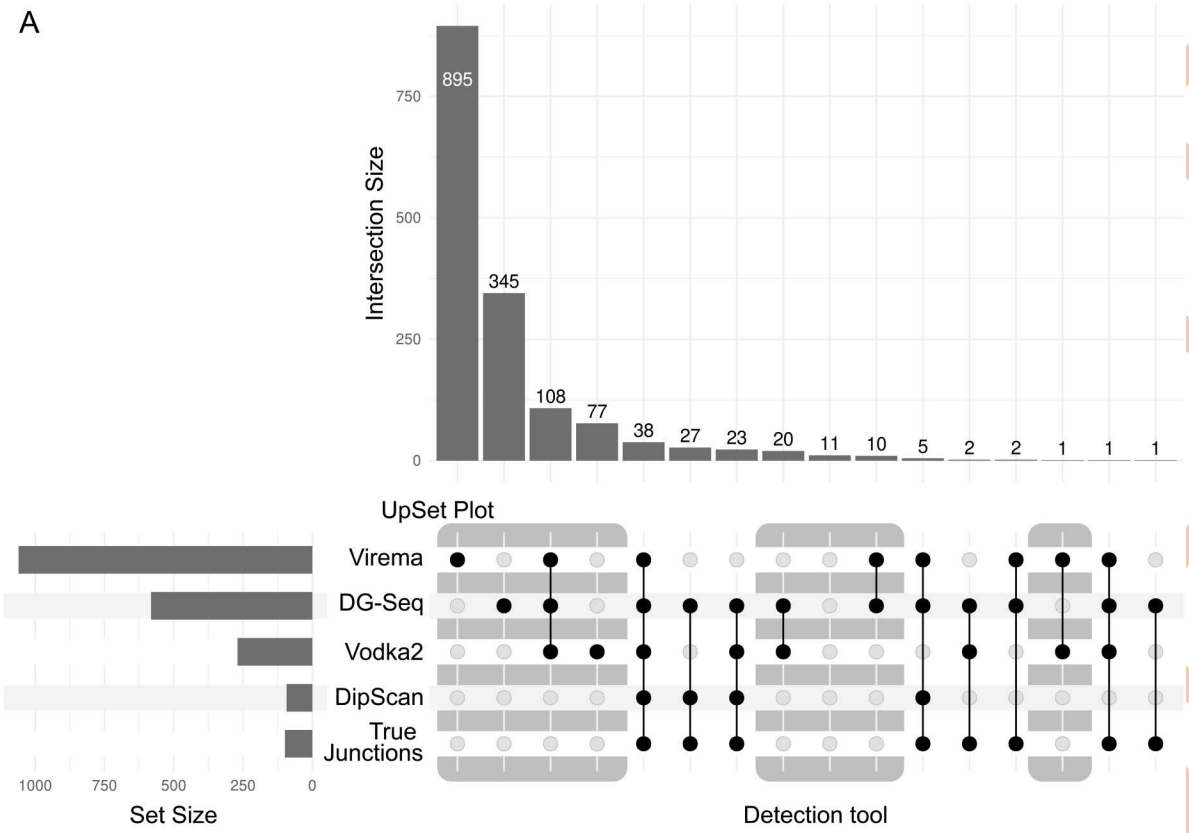
Regarding the sensitivity to low coverage, the Pearson correlation is still good (99%) on the two poorly covered simulated datasets (500k and 50k reads), over the different ratios (see Fig C(a) in S1 Appendix).

Across all combined datasets, DIPScan demonstrated an overall good accuracy of 94.5% (104/110, six incorrectly classified samples were false negative with low DelVG proportions) in classifying samples as containing DelVGs or not.

Mutation detection and consensus correction. In our simulation scenario, three kinds of mutations can in theory be identified relative to the reference genome: (i) Mutations that are shared between the DelVGs and the full length sequence, originating from the original consensus sequence), (ii) mutations that are specific to the DelVGs, randomly introduced, and (iii) mutations that are specific to the full length sequence, if the randomly introduced nucleotide reverts the nucleotide to the reference. The latter case did not happen in our simulations.

As shown in Fig 6B and Fig D in S1 Appendix (percent and absolute values, respectively) and Fig E in S1 Appendix, DIPScan demonstrates a strong performance in handling common and DelVG-specific mutations, respectively. This trend

A



B

	TP	FP	FN	Predictions	Sensitivity	Precision	FDR	F1Score
Virema	46	1014	53	1060	46,5%	4,3%	95,7%	7,9%
DG-Seq	99	483	0	582	**100,0%**	17,0%	83,0%	29,1%
Vodka2	64	206	35	270	64,6%	23,7%	76,3%	34,7%
DIP-Scan	93	0	6	93	93,9%	**100,0%**	**0,0%**	**96,9%**

Fig 5. Evaluation of DIPScan for breakpoint detection. Performance evaluation of DIPScan, ViReMa, DG-Seq, and VODKA2 in detecting DelVG breakpoints. **(A)** UpSet plot illustrating the overlap of breakpoints detected by the three tools and the true ones. For instance, 895 breakpoints are uniquely identified by ViReMa, while 38 breakpoints are correctly detected by all tools. Columns shaded in gray represent false breakpoints. **(B)** Breakdown of detected breakpoints per tool along the true breakpoints, categorized as true positives (TP), false positives (FP), and false negatives (FN), accompanied by key performance metrics: sensitivity, precision, false discovery rate (FDR), and F1-score.

https://doi.org/10.1371/journal.pcbi.1014115.g005

is confirmed in simulated samples with low-coverage, although results show a higher detection difficulty and a greater variability for low-proportion DelVGs (see Fig C(b) in S1 Appendix).

First, for shared mutations, where no consensus correction is expected, DIPScan demonstrated high fidelity, correctly retaining 99.6% (3976/3980) of the mutations across all samples. Among the 14 changed positions, 11 were masked (replaced by Ns) as these mutations fell just below the default 98% threshold for fixed mutations. The 3 incorrect changes involved samples where the mutations coincidentally displayed the same full-length/DelVG ratio between two nucleotides. This led DIPScan to misclassify them as DelVG-specific mutations.

Then, DIPScan demonstrated good performance in detecting all DelVG-specific mutations introduced into the test samples. Across all samples, it correctly retained or corrected 73.4% (426/580) of full-length mutations, introduced Ns

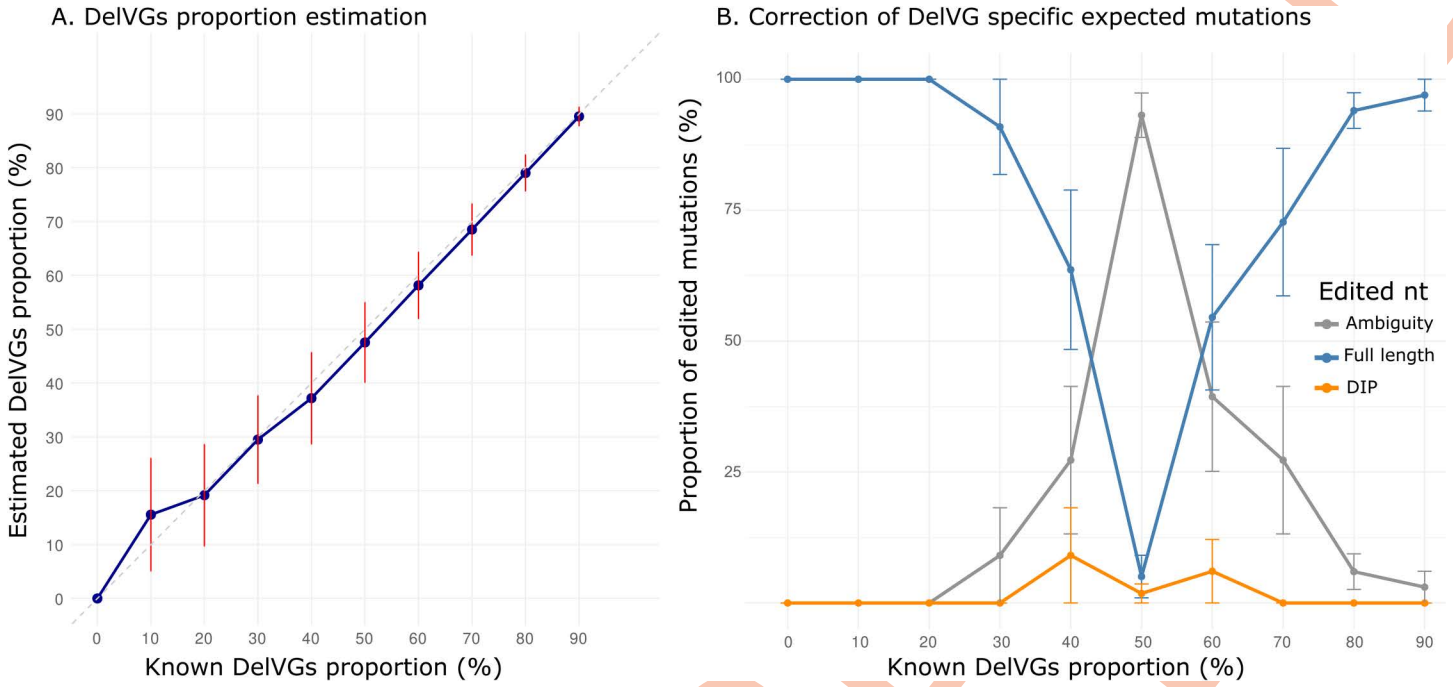


Fig 6. Evaluation of DIPScan for DelVG proportion estimation and consensus correction. A. Accuracy of DelVG proportion estimation. The mean distribution and confidence intervals of DelVG ratios estimated by DIPScan (y-axis) is shown for the 11 samples together across each simulated input proportion (x-axis). **B.** Correction performance for DelVG-specific mutations. The proportion of corrected DelVG-specific mutations (y-axis) as a function of the simulated input DelVG portion (x-axis) is represented for the three following outcomes: replacement with an ambiguous base (gray curve), correct reversion to (or retention of) the full-length genome nucleotide (blue curve), or erroneous retention of the DelVG-specific mutation (yellow curve). As expected, ambiguities peak around a 50% DelVG proportion, while correct retention or reversion peaks at both low DelVG proportion (where the full-length genome is dominant) and high DelVG proportion (where the DelVG is dominant, triggering a reversion to the full-length base).

https://doi.org/10.1371/journal.pcbi.1014115.g006

(ambiguous bases) for 23.8% (138/580), and misassigned 2.8% (16/580). Performance varied across the expected DelVG ratio. In the [50%, 90%] range, where DelVG presence complicates consensus reconstruction, DIPScan correctly retained or corrected 59.3% (172/290) of full-length mutations, assigned Ns for 37.9% (110/290), and incorrectly changed 2.8% (8/290). In contrast, for the [0%, 50%[range, performance improved, with 87.6% (254/290) of mutations correctly retained or corrected, Ns introduced in 9.7% (28/290), and misassignments remaining at 2.8% (8/290).

Notably, all misassigned nucleotides occurred in the same three samples where DelVG proportions were poorly estimated.

Altogether, these results demonstrate the good performances of DIPScan on simulated datasets at detecting DelVGs, their proportion, and correcting the consensus sequences accordingly.

C. Routine influenza virus sequencing datasets

To evaluate DIPScan's performance in a real-world setting, we applied it to a well-characterized influenza virus dataset that had been manually curated for DelVGs. Manual curation was conducted in two distinct phases: i) Initial (Coarse-Grained) Curation. The first phase involved routine manual curation as new samples arrived. It relied on visual inspection of genome coverage plots to detect sharp declines in coverage (and an inner coverage roughly less than half the outer coverage), which suggested the likely presence of DelVGs. However, this process was potentially error-prone due to variability in decision thresholds among multiple curators. ii) Refined Curation. In the second phase, a single curator

re-examined discrepancies between DIPScan results and the initial coarse-grained curation. Using the same criteria but focusing on a smaller subset of segments, this phase ensured a more standardized and careful evaluation.

Based on this, DIPScan was assessed according to three criteria: (i) DelVG detection accuracy, (ii) reliability of DelVG proportion estimates, and (iii) capability for consensus sequence correction. Finally, we analyzed the detected DelVG breakpoints to identify potential genomic hotspots.

DIPScan detection accuracy. To assess DIPScan detection performances, we classified the results into five categories, when compared to the "coarse-grained" manual dataset (Fig 7):

- **Consistently defective**: both manual detection and DIPScan identified DelVGs in the sample;
- **Consistently not defective**: neither manual detection nor DIPScan identified any DelVGs in the genome;
- **Inconsistent, defective in manual**: DelVGs were identified manually but not by DIPScan. These cases are detailed below, as they may represent either a limitation of DIPScan or manual detection errors.
- **Inconsistent, Defective with DIPScan** (<50%): DIPScan identified DelVGs representing less than 50% of the genomes, which manual detection might miss due to overall coverage similarities between and outside the deletion breakpoints. This category is not considered as errors but rather reflects different strategies in defining DelVGs.
- **Inconsistent, Defective with DIPScan** ($\geq$ 50%): DIPScan identified DelVGs representing more than 50% of the genomes. This category is detailed below, as it may indicate DIPScan's oversensitivity or manual detection errors.

Across the 4,408 analyzed segments (551 samples), DIPScan and manual detection agreed on 92.8% of segment classifications (Table 1A). Agreement rates varied by segments: 418 consistent PB1 (75.9%), 438 PB2 (79.5%), 487 PA (88.4%), 549 HA (99.6%), 548 NA (99.5%), 549 NS (99.6%), 551 MP (100%), and 549 NP (99.6%). Discrepancies included 33 segments classified as defective only in the manual dataset and 286 detected solely by DIPScan (Table 1A). Notably, no inconsistencies were found in the HA, NA, NS, NP, or MP segments that could be considered potential errors. Using manual curation as ground truth, DIPScan achieved 91% sensitivity and 54% precision (92.8% accuracy). However, most disagreements involved DIPScan detecting low-frequency DelVGs (<50%) missed during manual inspection. When considering only high-frequency DelVGs (relevant for consensus reconstruction) as true positives (Table 1B), DIPScan's performance improved to 99% sensitivity and 88% precision (98.9% accuracy). Further review showed most remaining discrepancies were either overlooked in manual curation or negligible low-frequency DelVGs, with 49 cases remaining unclassifiable (Table 1B). These unclassifiable cases presented a complex scenario where it was unclear whether manual review or DIPScan was correct.

In particular, among the cases where DIPScan detected DelVGs with a proportion $\geq$ 50% but manual review did not (**Inconsistent, Defective with DIPScan ($\geq$ 50%)**), 12 samples belonged to the A/H1N1pdm subtype. Upon re-evaluation, all instances affecting the PB2, PB1, and/or PA segments ($n = 15$) were found to be potential manual errors, primarily due to the difficulty in visually evaluating the proportions of full-length genomes over DelVGs. For the A/H3N2 subtype, 11 samples were identified, with all cases affecting the PB2, PB1, and/or PA segments ($n = 13$) ultimately attributed to potential manual errors. For the B/Victoria subtype, 29 samples were identified, all affecting the PB2, PB1, and PA segments. In 12 cases, the inconsistency was due a complex scenario likely involving a defective genome with a single breakpoint, resulting in truncation of the entire 5' part. One case was due to a manual report error since a DelVG was unambiguously visible, and 18 cases were found to be also potential manual errors.

In summary, despite using similar detection criteria (identification of breakpoints or sharp coverage declines greater than 50%), key differences between manual curation and DIPScan emerged:

- Sensitivity to low-abundance DelVGs: DIPScan detects minor low-proportion DelVGs much more effectively than manual curation. This limitation led us to focus, in a second step, on DelVGs present at more than 50% abundance.

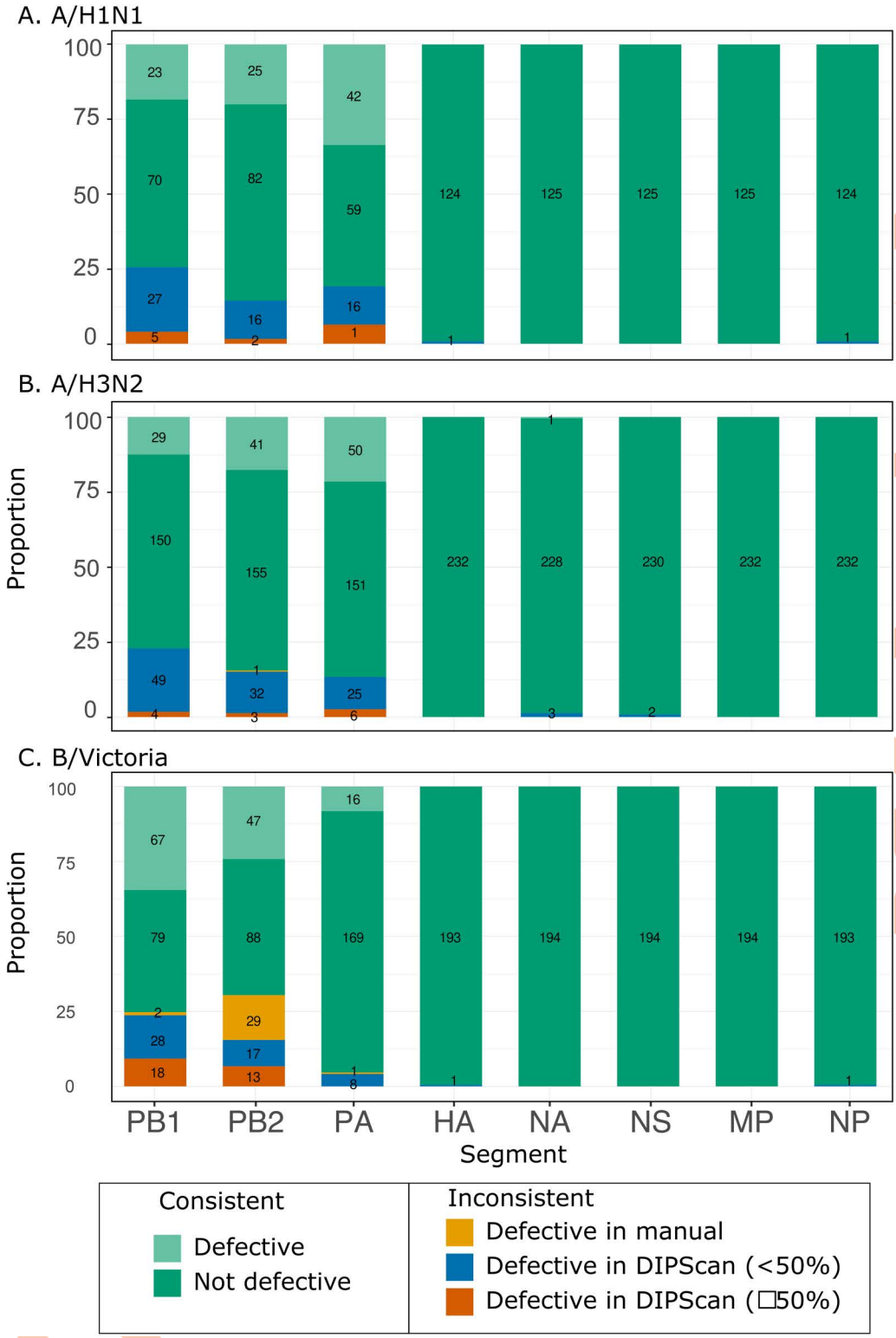


Fig 7. Evaluation of DIPScan on real datasets. Evaluation of DIPScan detection on real Influenza virus sequencing dataset: Manual curation versus DIPScan. Results are categorized based on agreement between methods: Consistent (both methods agree on the presence (light green) or absence

(dark green) of a DelVG) or Inconsistent. Inconsistent results are further broken down into: (i) DelVG detected only in manual curation, (ii) DelVG detected only in DIPScan with an estimated proportion below 50%, and (iii) DelVG detected only by DIPScan with a proportion above 50%. The results are grouped by Influenza virus segment and subtype: **A.** A/H1N1pdm (125 samples), **B.** A/H3N2 (232 samples), and **C.** B/Victoria (194 samples). Overall, we observe a good consistency between DIPScan results and manual curation. DelVGs detected only by DIPScan, at a proportion below 50% were usually missed by manual curation.

https://doi.org/10.1371/journal.pcbi.1014115.g007

Table 1. Confusion matrix comparing DIPScan and manual classification of 4,408 segments from 551 samples.

		DIPScan	
A	**Manual dataset**	Defective	Negative
	Defective	341	33
	Negative	286	3,748
B	**Manual dataset**	Defective ($\geq$ 50%)	Negative (<50%)
	Defective	341	2
	Negative	47	3,969
	Unclassifiable	12	37

(A) All DIPScan-detected sequences are treated as defective. (B) Only DIPScan-detected sequences with a frequency above 50% are considered defective.

https://doi.org/10.1371/journal.pcbi.1014115.t001

- Subjectivity in manual assessment: while both methods use a 50% cutoff to flag problematic DelVGs (to be excluded or corrected), visual inspection is less robust, as the estimated ratio of full-length genomes to DelVGs is less accurate, and the curator bias plays a significant role.
- Complex cases remain challenging: neither method perfectly resolves ambiguous or highly complex cases

DelVG proportion estimation. Direct assessment of DIPScan's accuracy in estimating DelVG proportions in real viral samples is inherently challenging due to the absence of a known ground truth. However, an indirect proxy for the true DelVG proportion can be derived from samples containing DelVG-specific mutations and/or full-length genome-specific mutations. These mutations produce mixed nucleotide signals at specific genomic positions, which can serve as a quantitative indicator of the relative abundance of the corresponding viral genomes within the sample.

We identified candidate mutations by analyzing the real datasets for variable positions (with several possible nucleotides), while systematically excluding regions between breakpoints, and the 150-nucleotide terminal segments of the genome. Using this approach, we identified 233 samples containing a total of 8,117 variable positions, with proportions likely attributable to either DelVGs or full-length genomes. Fig 8 presents the comparison of the DelVG proportions estimated by DIPScan to the observed mixed-position frequencies serving as a proxy for true proportions. The resulting Pearson correlation coefficient ($r = 0.96$, $p < 2.2 \times 10^{-16}$) indicates a high agreement between DIPScan's estimates and the empirical proxy, supporting DIPScan's accuracy in quantifying DelVG abundance.

Consensus correction. Across the whole dataset, 333/627 defective segments (53.1%) were subjected to correction of at least one position (41.4% for PB1, 25.5% for PB2, 32.1% for PA, 0.3% for HA, 0.3% for NA, and 0.3% for NP). Among the corrected positions, DIPScan changed the nucleotide in 17.7% of the cases. In the other cases (82.2%), DIPScan incorporated an ambiguous nucleotide.

To assess DIPScan's accuracy in correcting the consensus sequences, a ground truth dataset was build consisting of 25 positions across 19 samples where nucleotides could be unambiguously (manually) attributed to either the full-length genome or the defective genome.

As depicted in Fig 9, 24/25 positions (96%) in the selected ground truth dataset were correclty handled: 4 positions correctly changed, 13 positions correctly retained (13 unchanged), 7 positions appropriately masked, and only one position

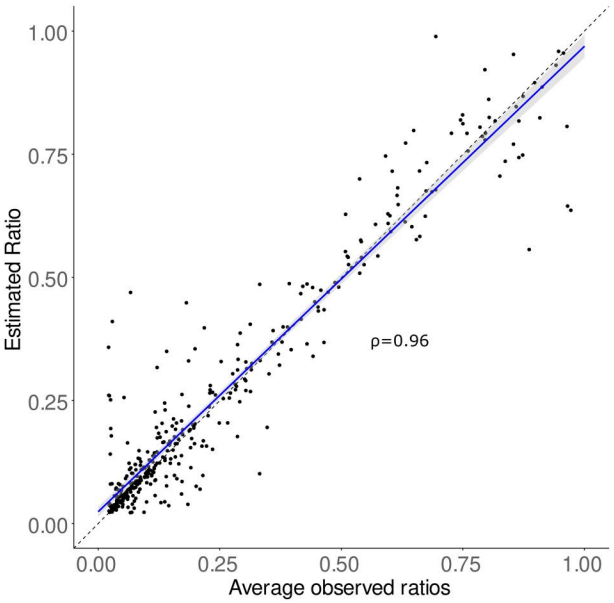


Fig 8. Evaluation of DIPScan proportion estimation on real datasets. Ratios between DelVG-specific coverage and full-length sequence coverage estimated by DIPScan (y-axis) are compared to the ratios derived from DelVG-specific (or full-length-specific) mutation frequencies at highly unambiguous positions (x-axis). Each point corresponds to the average over potentially multiple mutations for a given sample. The high correlation (0.96) between DIPScan estimates and ratios derived from unambiguous positions validates its use for downstream consensus correction.

https://doi.org/10.1371/journal.pcbi.1014115.g008

incorrectly changed, corresponding to a difficult case of DelVG proportion estimation. It is worth noting that DelVGs falling in the 45–55% range were not the most frequent (they represented 5–14% of the PA, PB1 and PB2 DelVGs in A/H1N1 and A/H3N2 samples, and 9–21% in PA, PB1, and PB2 DelVGs in B/Vic samples, see Table B in S1 Appendix).

Masked positions resulted from: (i) Poor alignment (5 positions from 5 samples; detected via comparison of expected depth and coverage of the mutated position), (ii) Inability to match estimated full-length proportions to nucleotide frequencies (1 position from 1 sample), (iii) Estimated full-length proportions below the 2% noise threshold (1 position from 1 sample).

For samples with >50% DelVG content, the unexpected cases of correctly unchanged positions occurred when the initial alignment (used to generate the consensus) was imperfect and resulted in the coincidental introduction of full-length-specific nucleotide into the consensus instead of the dominant DelVG-specific nucleotide.

After correcting the consensus sequences, we assessed the effects of these adjustments on two key aspects: i) Changes in phylogenetic tree topology, ii) Modifications at drug resistance mutation sites. Since viral lineage assignment is predominantly based on the HA segment, a segment that is rarely affected by DelVGs, DIPscan correction is not expected to significantly influence lineage determination.

Phylogenetic tree topology: We inferred phylogenetic trees for the DelVG-affected segments (PB1, PB2, and PA) from H1N1, H3N2, and B viruses, and compared trees generated from uncorrected versus DIPScan-corrected consensus sequences. While the observed topological differences were modest, they were consistent (see Fig F in S1 Appendix), confirming that uncorrected sequences may introduce inaccuracy in phylogenetic reconstructions.

Drug Resistance Mutations: We screened the corrected defective segments for residue known to be associated with resistance to antivirals (oseltamivir, zanamivir, and baloxavir). None of the corrected positions corresponded to established drug resistance markers. Additionally, we examined other samples beyond our dataset for the presence of resistance mutations but we found no overlap with defective sequences. However, it remains a possibility that, in other

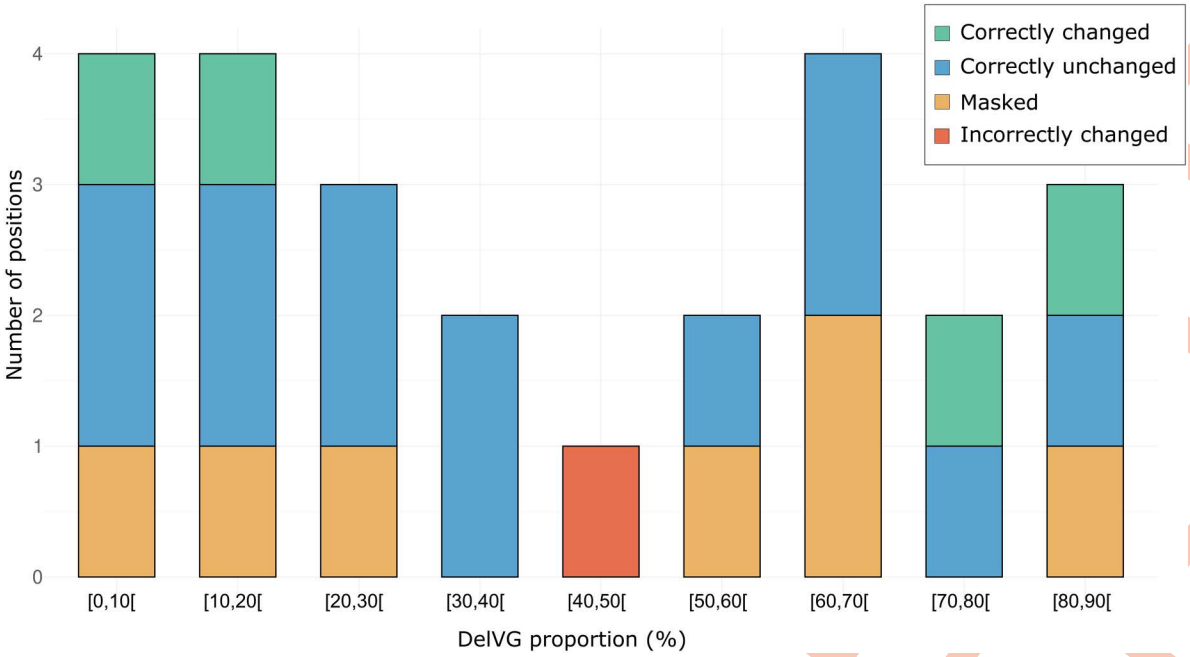


Fig 9. Validation of consensus correction on real datasets. Validation of consensus correction using 25 unambiguous mutations from 19 real Influenza virus datasets. Each of the 25 positions is classified by: (i) The proportion (estimated by DIPScan) of the detected DelVG covering the position (x-axis) and (ii) the outcome of the consensus correction: correctly changed (green), correctly unchanged (blue), masked with an 'N' (orange), or incorrectly changed (red). The y-axis shows the number of position in each category. We observe that globally, most of the positions (96%) are correctly changed, unchanged or masked, and 1 position is incorrectly changed.

https://doi.org/10.1371/journal.pcbi.1014115.g009

datasets or with emerging resistance mutations, DelVG-associated sequence could affect the estimated frequency of antiviral resistance markers, particularly since PB1, PB2, and PA all harbor known or potential resistance-associated sites.

Breakpoint location hotspots. With DIPScan results in hand, we could further characterize the DelVG sequences. In particular, we used the results from the 551 real dataset samples to examine the breakpoint positions along the genome and identify regions enriched with breakpoints ("hotspots").

To do so, we grouped the results by virus subtype and segment, as they may exhibit distinct patterns of genomic deletions (see Fig 10, Fig G in S1 Appendix).

Our analysis demonstrates that DelVG breakpoints are clearly not uniformly distributed across the viral genome (Fig 10), instead exhibiting distinct "hotspots" that vary by segment and subtype. Moreover, the breakpoint start-end locations seem to be dependent from each other (see Fig G in S1 Appendix). These observations align with previously reported patterns of DelVG breakpoint localization in influenza [33,46,47].

More specifically, breakpoint distribution patterns revealed subtype-specific variations in DelVG generation depending on the segments.

For the PB2 segment (Fig 10B), all three subtypes displayed 5' breakpoints predominantly at around 9% of the genome length, and a second peak was observed around 12% for A/H1N1pdm and A/H3N2, while B/Victoria exhibited a shift to 19%. The 3' breakpoints clustered around 89% for all three subtypes, and a second peak was observed around 85% for A/H1N1pdm and A/H3N2, while B/Victoria exhibited a shift to 80%.

For the PB1 segment (Fig 10A), A/H1N1pdm showed 5' breakpoints at 10%, A/H3N2 at 8% and 13%, and B/Victoria at 15%. A/H1N1pdm and A/H3N2 showed 3' breakpoints at 78% and 86%, B/Victoria also shared a peak at 86% and exhibited a second one with a shift to 88%.

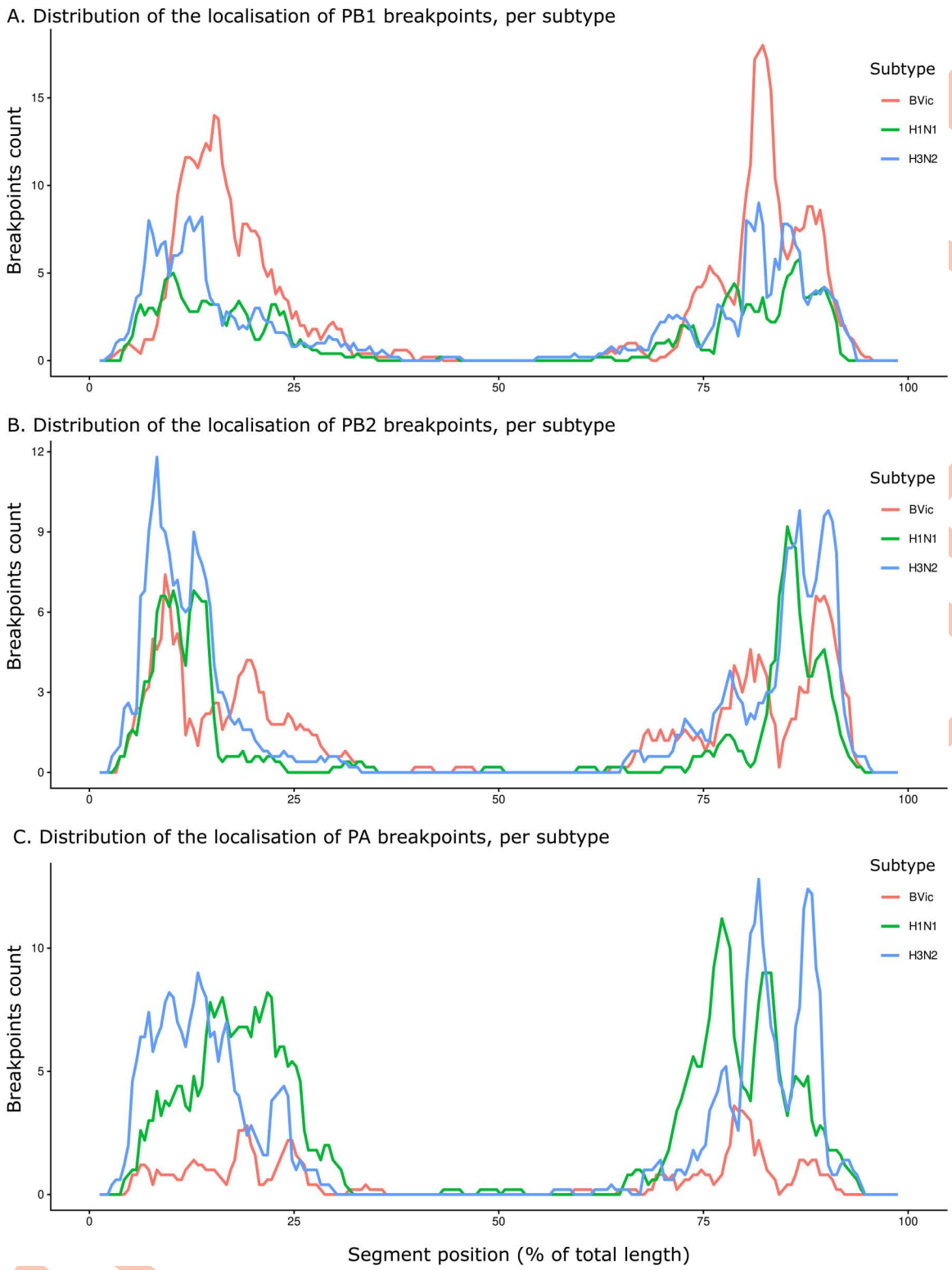


Fig 10. Analysis of potential breakpoint hotspots. The distribution of detected breakpoint ends (y-axis) is shown across the genomic segment, with the x-axis representing the location as a percentage of the segment's total length. This distribution is colored by Influenza virus subtype of the 551

analyzed samples: B/Victoria in red, A/H1N1pdm in green, and A/H3N2 in blue. Only the 3 main affected segments are represented: **A.** PB1, **B.** PB2 and **C.** PA. This confirms that clear breakpoint hotspots exist near the segment ends of the three analyzed subtypes, and the three mainly affected segments (PB1, PB2, and PA).

https://doi.org/10.1371/journal.pcbi.1014115.g010

In the PA segment (Fig 10C), A/H1N1pdm showed 5' breakpoints between 16–22%, A/H3N2 between 7–16%, and B/Victoria exhibited two main peaks at 19% and 24%. All three subtypes displayed 3' breakpoints at 87%, then A/H1N1pdm and A/H3N2 had other peaks at 77% and 82%, and B/Victoria at a second peak at 79%, highlighting subtype-specific variations in DelVG generation.

While confirming these breakpoint hotspots, these results further validate the ability of DIPScan to accurately detect DelVG sequences, and its usage in a routine setting.

Discussion and perspectives

The widespread use of genome sequencing has enabled more precise and large-scale viral epidemic surveillance, for a broad range of pathogens, from EBOV [48], ZIKV [49], to SARS-CoV-2 [50]. While many virus epidemics are now monitored using molecular data, respiratory viruses remain a major concern, and generate the most massive amounts of data [51]. The scale of these datasets, combined with the need for automated consensus generation and curation to ensure accuracy, presents a significant bioinformatics challenge. Consequently, bioinformatics workflows have incorporated specialized functionalities for viruses, including reference selection, segmented genome support, and lineage assignment based on current nomenclatures.

Despite these advances, routine bioinformatics workflows have lacked the capability to account for the specific features of defective genomes when generating the consensus. To address this gap, we developed DIPScan, a Nextflow workflow designed to detect Deletion-containing Viral Genomes (DelVGs) that result from large deletions, and correct the consensus sequence when possible.

DIPScan introduces three key methodological advances. First, it defines dedicated metrics to systematically characterize and select proper breakpoints. Second, it integrates a quantitative estimation of DelVG relative abundance through a linear model integrating regional and junction coverage, which is not provided by existing tools. Third, it implements heuristic consensus correction algorithm to resolve ambiguous genomic positions, which would otherwise be an intractable combinatorial problem.

Using simulated data, we demonstrated that DIPScan accurately detects DelVGs (accuracy of 98.9%) and accurately estimates DelVGs/full length genome proportion (correlation of 0.99 between expected and estimated DelVG proportions). Additionally, DIPScan's correction of consensus sequences is reliable, with 97.2% of the DelVG-specific mutations being either replaced with the correct nucleotide present in full-length genomes or masked, into the final consensus sequence.

Analysis of 551 patient derived sequencing samples revealed a good agreement between DIPScan and manual curation, especially for predominant DelVGs (>50%). Discrepancies were mostly due to manual errors. However, when DelVGs were detected only by DIPScan, they corresponded to genuine internal deletions or to complex cases requiring further review. In all scenarios, DIPScan enabled the rapid identification and correction of DelVGs at a large scale.

In addition, our analysis of real datasets supports prior findings that Influenza A and B DelVGs exhibit distinct characteristics. Specifically, we observed that manual curation identified a greater proportion (29 out of 194) of unique DelVGs in Influenza B, particularly in the PB2 segment, compared to Influenza A (1 out of 357). This discrepancy may stem from a feature present mainly in our Influenza B samples: single-sided breakpoints, which are not explicitly targeted by DIPScan's detection approach. We hypothesize that this may reflect a higher prevalence of copy-back genomes in Influenza B, although this has not been demonstrated specifically between A and B influenza viruses, to our knowledge [52].

Furthermore, the automatic execution of DIPScan on real datasets identified clear preferential genomic breakpoints, "hotspots" positions for defective genomes, which is in accordance with previous reports [10,19], and confirms it on a large number of samples.

In conclusion, DIPScan provides an accurate, efficient and scalable workflow for detecting DelVGs in influenza virus Illumina sequencing datasets, and, importantly, for correcting consensus sequences by accounting for DelVG-specific mutations.

The tool has been integrated into the NRC's routine sequencing pipeline to screen for potential DelVGs. DIPScan is continuously being developed, and future developments will focus on two main areas.

Extension to Other Pathogens: DIPScan is currently implemented and tested more particularly on influenza viruses. We plan to test and benchmark DIPScan on datasets from other viruses known to produce DelVGs, such as RSV and SARS-CoV-2 [16]. This may require parameter optimization to ensure robust performance across different viral genomes. Preliminary tests on one hundred Respiratory Syncytial Virus (RSV) clinical samples identified only a single sample displaying DelVG patterns at more than 50% abundance. However, a more thorough validation on large datasets is needed to assess and confirm the applicability of the approach on RSV, and other viruses beyond influenza viruses.

Detection of diverse types of defective genomes: A key priority is to extend support to other defective genomes, including copy-back and rearrangements. While these are less frequent (although may explain the 29 influenza B samples identified as "defective only in manual") and present a greater challenge for defining genomic regions for correction, we could refine our methodology to identify them. One promising approach is to leverage the chimeric read detection capabilities of mappers like STAR, which could provide the necessary signal to detect these complex structural variants.

Supporting information

S1 Appendix. Supplementary Text, Tables, and Figures.
(PDF)

Acknowledgments

We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work.

Author contributions

Conceptualization: Marie-Anne Rameix-Welti, Frederic Lemoine.

Data curation: Kévin Da Silva, Frederic Lemoine.

Formal analysis: Frederic Lemoine.

Funding acquisition: Marie-Anne Rameix-Welti.

Investigation: Kévin Da Silva, Nadia Naffakh, Marie-Anne Rameix-Welti, Frederic Lemoine.

Methodology: Kévin Da Silva, Frederic Lemoine.

Resources: Marie-Anne Rameix-Welti, Frederic Lemoine.

Software: Kévin Da Silva, Frederic Lemoine.

Supervision: Marie-Anne Rameix-Welti, Frederic Lemoine.

Validation: Kévin Da Silva, Nadia Naffakh, Marie-Anne Rameix-Welti, Frederic Lemoine.

Visualization: Kévin Da Silva, Frederic Lemoine.

Writing – original draft: Kévin Da Silva, Frederic Lemoine.

Writing – review & editing: Kévin Da Silva, Nadia Naffakh, Marie-Anne Rameix-Welti, Frederic Lemoine.

References

1. Patel H, Monzón S, Varona S, Espinosa-Carrasco J, Garcia MU, Bot NC. nf-core/viralrecon: nf-core/viralrecon v2.6.0 - Rhodium Raccoon. Zenodo; 2023. Available from: https://doi.org/10.5281/ZENODO.3901628
2. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8. https://doi.org/10.1038/s41587-020-0439-x PMID: 32055031
3. Moshiri N, Fisch KM, Birmingham A, DeHoff P, Yeo GW, Jepsen K, et al. The ViReflow pipeline enables user friendly large scale viral consensus genome reconstruction. Sci Rep. 2022;12(1):5077. https://doi.org/10.1038/s41598-022-09035-w PMID: 35332213
4. da Silva AF, da Silva Neto AM, Aksenen C, Jeronimo P, Dezordi F, Almeida S, et al. ViralFlow v1.0—a computational workflow for streamlining viral genomic surveillance. NAR Genom Bioinform. 2024;6(2):lqae056. https://doi.org/10.1093/nargab/lqae056
5. Shepard SS, Meno S, Bahl J, Wilson MM, Barnes J, Neuhaus E. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics. 2016;17(1):708. https://doi.org/10.1186/s12864-016-3030-6 PMID: 27595578
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352
8. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019;20(1):8. https://doi.org/10.1186/s13059-018-1618-7 PMID: 30621750
9. European Centre for Disease Prevention and Control. European SARS-CoV-2 and influenza Bioinformatics External Quality Assessment (ESIB-EQA): 2023. LU: Publications Office; 2024.
10. Alnaji FG, Reiser WK, Rivera-Cardona J, Te Velthuis AJW, Brooke CB. Influenza A virus defective viral genomes are inefficiently packaged into virions relative to wild-type genomic RNAs. mBio. 2021;12(6):e02959-21. https://doi.org/10.1128/mBio.02959-21
11. von Magnus P. Studies on interference in experimental influenza: Biological observations. No. vol. 1 in Arkiv för kemi, mineralogi och geologi. Almqvist & Wiksell; 1947. Available from: https://books.google.fr/books?id=f-IEHQAACAAJ
12. Huang AS, Baltimore D. Defective viral particles and viral disease processes. Nature. 1970;226(5243):325–7. https://doi.org/10.1038/226325a0 PMID: 5439728
13. Alnaji FG, Brooke CB. Influenza virus DI particles: defective interfering or delightfully interesting? PLOS Pathogens. 2020;16(5):e1008436. https://doi.org/10.1371/journal.ppat.1008436
14. Vignuzzi M, López CB. Defective viral genomes are key drivers of the virus-host interaction. Nat Microbiol. 2019;4(7):1075–87. https://doi.org/10.1038/s41564-019-0465-y PMID: 31160826
15. Genoyer E, López CB. The impact of defective viruses on infection and immunity. Annu Rev Virol. 2019;6(1):547–66. https://doi.org/10.1146/annurev-virology-092818-015652 PMID: 31082310
16. Brennan JW, Sun Y. Defective viral genomes: advances in understanding their generation, function, and impact on infection outcomes. mBio. 2024;15(5):e0069224. https://doi.org/10.1128/mbio.00692-24 PMID: 38567955
17. Alnaji FG, Holmes JR, Rendon G, Vera JC, Fields CJ, Martin BE, et al. Sequencing framework for the sensitive detection and precise mapping of defective interfering particle-associated deletions across influenza A and B viruses. J Virol. 2019;93(11):e00354-19. https://doi.org/10.1128/JVI.00354-19 PMID: 30867305
18. Oade MS, Te Velthuis AJW. Molecular insights into noncanonical influenza virus replication and transcription. Annu Rev Virol. 2025. https://doi.org/10.1146/annurev-virology-092623-101331
19. Boussier J, Munier S, Achouri E, Meyer B, Crescenzo-Chaigne B, Behillil S, et al. RNA-seq accuracy and reproducibility for the mapping and quantification of influenza defective viral genomes. RNA. 2020;26(12):1905–18. https://doi.org/10.1261/rna.077529.120 PMID: 32929001
20. Jennings PA, Finch JT, Winter G, Robertson JS. Does the higher order structure of the influenza virus ribonucleoprotein guide sequence rearrangements in influenza viral RNA? Cell. 1983;34(2):619–27. https://doi.org/10.1016/0092-8674(83)90394-X
21. Vasilijevic J, Zamarreño N, Oliveros JC, Rodriguez-Frandsen A, Gómez G, Rodriguez G, et al. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. PLoS Pathog. 2017;13(10):e1006650. https://doi.org/10.1371/journal.ppat.1006650 PMID: 29023600
22. Sun Y, Kim EJ, Felt SA, Taylor LJ, Agarwal D, Grant GR, et al. A specific sequence in the genome of respiratory syncytial virus regulates the generation of copy-back defective viral genomes. PLoS Pathog. 2019;15(4):e1007707. https://doi.org/10.1371/journal.ppat.1007707 PMID: 30995283
23. Holland JJ, Kennedy SIT, Semler BL, Jones CL, Roux L, Grabau EA. Defective interfering RNA viruses and the host-cell response. In: Comprehensive virology: vol. 16: virus-host interactions: viral invasion, persistence, and disease. Springer; 1980. p. 137–92.
24. Ziegler CM, Botten JW. Defective interfering particles of negative-strand RNA viruses. Trends Microbiol. 2020;28(7):554–65. https://doi.org/10.1016/j.tim.2020.02.006 PMID: 32544442
25. Huang AS. Defective interfering viruses. Annu Rev Microbiol. 1973;27(1):101–18. https://doi.org/10.1146/annurev.mi.27.100173.000533
26. Marcus PI, Sekellick MJ. Defective interfering particles with covalently linked [+/-]RNA induce interferon. Nature. 1977;266(5605):815–9. https://doi.org/10.1038/266815a0 PMID: 194158

27. Tapia K, Kim W-K, Sun Y, Mercado-López X, Dunay E, Wise M, et al. Defective viral genomes arising in vivo provide critical danger signals for the triggering of lung antiviral immunity. PLoS Pathog. 2013;9(10):e1003703. https://doi.org/10.1371/journal.ppat.1003703 PMID: 24204261
28. Barrett AT, Dimmock NJ. Defective interfering viruses and infections of animals. Curr Top Microbiol Immunol. 1986:55–84.
29. Penn R, Tregoning JS, Flight KE, Baillon L, Frise R, Goldhill DH, et al. Levels of influenza A virus defective viral genomes determine pathogenesis in the BALB/c mouse model. J Virol. 2022;96(21):e0117822. https://doi.org/10.1128/jvi.01178-22 PMID: 36226985
30. Dimmock NJ, Easton AJ. Defective interfering influenza virus RNAs: time to reevaluate their clinical potential as broad-spectrum antivirals? J Virol. 2014;88(10):5217–27. https://doi.org/10.1128/JVI.03193-13 PMID: 24574404
31. Pelz L, Rüdiger D, Dogra T, Alnaji FG, Genzel Y, Brooke CB, et al. Semi-continuous propagation of influenza A virus and its defective interfering particles: analyzing the dynamic competition to select candidates for antiviral therapy. J Virol. 2021;95(24):e0117421. https://doi.org/10.1128/JVI.01174-21 PMID: 34550771
32. Dogra T, Pelz L, Boehme JD, Kuechler J, Kershaw O, Marichal-Gallardo P, et al. Generation of "OP7 chimera" defective interfering influenza A particle preparations free of infectious virus that show antiviral efficacy in mice. Sci Rep. 2023;13(1):20936. https://doi.org/10.1038/s41598-023-47547-1 PMID: 38017026
33. Lohmann JJG, Le M, Alnaji FG, Zolotareva O, Baumbach J, Laske T. Meta-analysis of genomic characteristics for antiviral influenza defective interfering particle prioritization. NAR Genom Bioinform. 2025;7(2):lqaf031. https://doi.org/10.1093/nargab/lqaf031 PMID: 40191586
34. Saira K, Lin X, DePasse JV, Halpin R, Twaddle A, Stockwell T, et al. Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. J Virol. 2013;87(14):8064–74. https://doi.org/10.1128/JVI.00240-13 PMID: 23678180
35. Beauclair G, Mura M, Combredet C, Tangy F, Jouvenet N, Komarova AV. DI-tector: defective interfering viral genomes' detector for next-generation sequencing data. RNA. 2018;24(10):1285–96. https://doi.org/10.1261/rna.066910.118 PMID: 30012569
36. Jaworski E, Routh A. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. PLoS Pathog. 2017;13(5):e1006365. https://doi.org/10.1371/journal.ppat.1006365 PMID: 28475646
37. Routh A, Johnson JE. Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data. Nucleic Acids Res. 2014;42(2):e11. https://doi.org/10.1093/nar/gkt916 PMID: 24137010
38. Bosma TJ, Karagiannis K, Santana-Quintero L, Ilyushina N, Zagorodnyaya T, Petrovskaya S, et al. Identification and quantification of defective virus genomes in high throughput sequencing data using DVG-profiler, a novel post-sequence alignment processing algorithm. PLoS One. 2019;14(5):e0216944. https://doi.org/10.1371/journal.pone.0216944 PMID: 31100083
39. Achouri E, Felt SA, Hackbart M, Rivera-Espinal NS, López CB. VODKA2: a fast and accurate method to detect non-standard viral genomes from large RNA-seq data sets. RNA. 2023;30(1):16–25. https://doi.org/10.1261/rna.079747.123 PMID: 37891004
40. Taylor A, Rosa C, Archetti M. Defective but promising: evaluating the utility of currently available bioinformatic pipelines for detecting defective viral genomes in RNA-Seq data. J Gen Virol. 2025;106(11):002176. https://doi.org/10.1099/jgv.0.002176 PMID: 41247266
41. Vasimuddin M, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE; 2019. p. 314–24.
42. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886
43. Bro R, De Jong S. A fast non-negativity-constrained least squares algorithm. J Chemom. 1997;11(5):393–401. https://doi.org/10.1002/(sici)1099-128x(199709/10)11:5<393::aid-cem483>3.0.co;2-l
44. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690
45. Schmeing S, Robinson MD. ReSeq simulates realistic Illumina high-throughput sequencing data. Genome Biol. 2021;22(1):67. https://doi.org/10.1186/s13059-021-02265-7 PMID: 33608040
46. Sheng Z, Liu R, Yu J, Ran Z, Newkirk SJ, An W, et al. Identification and characterization of viral defective RNA genomes in influenza B virus. J Gen Virol. 2018;99(4):475–88. https://doi.org/10.1099/jgv.0.001018 PMID: 29458654
47. Martin MA, Berg N, Koelle K. Influenza A genomic diversity during human infections underscores the strength of genetic drift and the existence of tight transmission bottlenecks. 2023. https://doi.org/10.1101/2023.06.18.545481
48. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228–32. https://doi.org/10.1038/nature16996 PMID: 26840485
49. Faye O, de Lourdes Monteiro M, Vrancken B, Prot M, Lequime S, Diarra M, et al. Genomic epidemiology of 2015-2016 Zika virus outbreak in Cape Verde. Emerg Infect Dis. 2020;26(6):1084–90. https://doi.org/10.3201/eid2606.190928 PMID: 32441631
50. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet. 2022;54(4):499–507. https://doi.org/10.1038/s41588-022-01033-y PMID: 35347305
51. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Eurosurveillance. 2017;22(13):30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMID: 28382917
52. Li X, Ye Z, Plant EP. 5' copyback defective viral genomes are major component in clinical and non-clinical influenza samples. Virus Res. 2024;339:199274. https://doi.org/10.1016/j.virusres.2023.199274 PMID: 37981214
53. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9. https://doi.org/10.1038/nbt.3820 PMID: 28398311