MICROBIAL GENOMICS

RESEARCH ARTICLE

Van Poelvoorde *et al.*, *Microbial Genomics* 2022;8:000867

DOI 10.1099/mgen.0.000867

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

A general approach to identify low-frequency variants within influenza samples collected during routine surveillance

Laura A. E. Van Poelvoorde[1,2,3,4], Thomas Delcourt[1], Marnik Vuylsteke[5], Sigrid C. J. De Keersmaecker[1], Isabelle Thomas[2], Steven Van Gucht[2], Xavier Saelens[3,4], Nancy Roosens[1]† and Kevin Vanneste[1,*,†]

Abstract

Influenza viruses exhibit considerable diversity between hosts. Additionally, different quasispecies can be found within the same host. High-throughput sequencing technologies can be used to sequence a patient-derived virus population at sufficient depths to identify low-frequency variants (LFV) present in a quasispecies, but many challenges remain for reliable LFV detection because of experimental errors introduced during sample preparation and sequencing. High genomic copy numbers and extensive sequencing depths are required to differentiate false positive from real LFV, especially at low allelic frequencies (AFs). This study proposes a general approach for identifying LFV in patient-derived samples obtained during routine surveillance. Firstly, validated thresholds were determined for LFV detection, whilst balancing both the cost and feasibility of reliable LFV detection in clinical samples. Using a genetically well-defined population of influenza A viruses, thresholds of at least 10^4 genomes per microlitre and AF of ≥5% were established as detection limits. Secondly, a subset of 59 retained influenza A (H3N2) samples from the 2016–2017 Belgian influenza season was composed. Thirdly, as a proof of concept for the added value of LFV for routine influenza monitoring, potential associations between patient data and whole genome sequencing data were investigated. A significant association was found between a high prevalence of LFV and disease severity. This study provides a general methodology for influenza LFV detection, which can also be adopted by other national influenza reference centres and for other viruses such as SARS-CoV-2. Additionally, this study suggests that the current relevance of LFV for routine influenza surveillance programmes might be undervalued.

Received 28 February 2022; Accepted 21 June 2022; Published 28 September 2022

Author affiliations: [1]Transversal activities in Applied Genomics, Sciensano, Juliette Wytsmanstraat 14, Brussels, Belgium; [2]National Influenza Centre, Sciensano, Juliette Wytsmanstraat 14, Brussels, Belgium; [3]Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium; [4]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium; [5]Gnomixx, Ghent University, Melle, Belgium.

***Correspondence:** Kevin Vanneste, kevin.vanneste@sciensano.be

Keywords: Influenza; low-frequency variants; next-generation sequencing; patient data; surveillance.

Abbreviations: AF, Allelic frequency; ARDS, Acute Respiratory Distress Syndrome; DMEM, Dulbecco's modified Eagle medium; ECMO, Extracorporeal membrane oxygenation; FP, False positive; HA, Hemagglutinin; HTS, High-throughput sequencing; Indels, insertions and deletions; IQR, Interquartile range; LFV, Low-frequency variants; MDCK, Madin-Darby canine kidney; NA, Neuraminidase; PFU, Plaque Forming Units; ROC, Receiver operating characteristic; SNV, Single nucleotide variants; SRA, Sequence Read Archive; TP, True positive; WT, Wild-type.

• NCBI Sequence Read Archive (SRA): SRS6396338, SRS6396349, SRS6396411, SRS6396356, SRS6396473, SRS6396318, SRS6396253, SRS6396260, SRS6396269, SRS6396395, SRS6396396, SRS6396403, SRS6396409, SRS6396410, SRS6396437, SRS6396449, SRS6396452, SRS6396453, SRS6396455, SRS6396270, SRS6396276, SRS6396284, SRS6396285, SRS6396293, SRS6396303, SRS6396307, SRS6396228, SRS6396231, SRS6396237, SRS6396238, SRS6396243, SRS6396247, SRS6396248, SRS6396357, SRS6396360, SRS6396370, SRS6396371, SRS6396379, SRS6396383, SRS6396390, SRS6396419, SRS6396422, SRS6396425, SRS6396428, SRS6396429, SRS6396430, SRS6396436, SRS6396457, SRS6396458, SRS6396465, SRS6396468, SRS6396472, SRS6396474, SRS6396312, SRS6396314, SRS6396315, SRS6396316, SRS6396319, SRS6396330 and PRJNA692424 and PRJNA615341.

• GISAID: EPI_ISL_415204, EPI_ISL_415205, EPI_ISL_415207, EPI_ISL_415215, EPI_ISL_415222, EPI_ISL_415223, EPI_ISL_415228, EPI_ISL_415234, EPI_ISL_415241, EPI_ISL_415260, EPI_ISL_415261, EPI_ISL_415266, EPI_ISL_415272, EPI_ISL_415273, EPI_ISL_415278, EPI_ISL_415292, EPI_ISL_415294, EPI_ISL_415296, EPI_ISL_415298, EPI_ISL_415299, EPI_ISL_415305, EPI_ISL_415312, EPI_ISL_415313, EPI_ISL_415320, EPI_ISL_415329, EPI_ISL_415333, EPI_ISL_415338, EPI_ISL_415340, EPI_ISL_415346, EPI_ISL_415347, EPI_ISL_415351, EPI_ISL_415355, EPI_ISL_415356, EPI_ISL_415359, EPI_ISL_415362, EPI_ISL_415371, EPI_ISL_415372, EPI_ISL_415379, EPI_ISL_415383, EPI_ISL_415389, EPI_ISL_415397, EPI_ISL_415399, EPI_ISL_415401, EPI_ISL_415404, EPI_ISL_415405, EPI_ISL_415406, EPI_ISL_415411, EPI_ISL_415413, EPI_ISL_415415, EPI_ISL_415421, EPI_ISL_415424, EPI_ISL_415428, EPI_ISL_415430, EPI_ISL_415433, EPI_ISL_415436, EPI_ISL_415437, EPI_ISL_415438, EPI_ISL_415440, EPI_ISL_415450.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary figure and four supplementary tables are available with the online version of this article.

DATA SUMMARY

All supporting protocols, code and data have been provided within the article, in the supplementary data or in FigShare: https://doi.org/10.6084/m9.figshare.21214256.v1 [1]. All sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA). All generated consensus genome sequences have been deposited in GISAID.

INTRODUCTION

Influenza is a very contagious respiratory tract infection in humans, mainly caused by the Influenza A and B viruses. Both the Influenza A and B genomes consist of eight segments, including the hemagglutinin (HA) and neuraminidase (NA) segments. Due to their location on the viral envelope, the proteins encoded by the HA and NA segments represent key viral antigens and are the principal targets of the humoral immune response of the host [2–4]. A(H1N1) and A(H3N2) are the two principal Influenza A subtypes that circulate in humans [5].

Influenza viruses have a low-fidelity RNA polymerase that lacks proof-reading functionality. This results in a relatively high mutation rate during viral replication [6]. Replicating influenza within a host does therefore not give rise to genetically identical progeny viruses but rather to 'quasispecies', i.e. closely-related viruses that differ by at least one nucleotide from each other. Viral quasispecies are defined as a population of closely-related, non-identical viral genomes in a dynamic host environment that is continuously subjected to competition and selection [7–9]. Although considerable risk exists for producing defective progeny viruses due to the low-fidelity RNA polymerase, this also provides a major opportunity for the virus to rapidly evolve and escape from neutralizing antibodies [10], antiviral drugs [11] and cytotoxic T-cells [12].

The availability and cost-effectiveness of high-throughput sequencing (HTS) technologies have led to their increased use in routine influenza surveillance [13]. HTS allows to determine the sequences of all eight influenza virus segments simultaneously, which offers the opportunity to better understand between- and within-host genetic diversity [14]. Genetic surveillance of influenza virus in biological samples is currently focused on monitoring mutations that are linked to antiviral resistance [15, 16], and antigenic mutations that are relevant for selecting vaccine strains [17]. Studies examining influenza pathogenesis should consequently consider virological and immunological parameters associated to severity as a whole [18]. When investigating viral evolution, transmission, drug and vaccine resistant strains, and pathogenicity, it may not always be sufficient to only examine the consensus genome sequence. Therefore, the current focus is shifting to also include quasispecies while studying genetic diversity [19, 20]. During infection, a particular variant within a quasispecies can by chance obtain a competitive advantage over other variants [21]. This can result in positive selection, and thus an increased frequency of such a variant over time within the patient [22]. However, the spread to other hosts is limited to a small fraction of the quasispecies population and even fewer become fixed in the global viral population [9, 23]. Positive selection of specific quasispecies in hosts has thus far only been observed during long-term infection of immunocompromised patients [24] and in extreme cases of drug resistance [25–27] for the HA and NA genes.

Several recent studies have successfully identified genetic variation in viral quasispecies during clinical influenza infections using deep sequencing with HTS [24, 28–31]. Deep sequencing allows higher genome coverages, and consequently more reliable estimation of the diversity within the quasispecies population present at very low abundances [32]. Apart from the increased experimental costs associated with the use of HTS, many challenges remain to detect low-frequency variants (LFV, i.e. defined as nucleotides differing from the consensus sequence at low allelic frequency at a specific genomic position), including high-quality sequencing reads to ensure that insertions and deletions (indels), and single nucleotide variants (SNVs), can be called confidently. Current variant-calling algorithms for identifying LFV are based on read quality, mapping quality, strand bias, base quality and sequence context [28]. Variants are typically accepted only when their allelic frequency (AF) exceeds the expected sequencing error rate. Several variant-calling methods have been used in multiple HTS-based studies of viral diversity [18, 25, 33]. However, these methods have not always been benchmarked against predefined viral populations, rendering their accuracy for detecting LFV largely unknown. Moreover, not only the bioinformatics approach but also the laboratory process can influence LFV detection. Experimental errors can be introduced during sample preparation, including reverse transcription and PCR amplification, and during sequencing itself [34]. The genome copy number and viral load of samples in particular affect the specificity and sensitivity of variant detection substantially, resulting in more false positive (FP) variant detections for samples with a low concentration due to propagating PCR-amplification errors [28].

In this study, we first established an approach for the quantification of low-frequency variants within influenza samples by using a genetically well-defined population of Influenza A viruses. Thresholds for LFV detection based on HTS with the Illumina technology were validated whilst ensuring that this approach remains powerful enough but also economically feasible in routine surveillance. Secondly, this approach was used to evaluate the prevalence of LFV of influenza A(H3N2) viruses recovered from the Belgian national influenza surveillance network during the 2016–2017 season, demonstrating that several LFV were identified in clinical samples. Finally, potential associations between within-host diversity and patient data were investigated as a proof of concept for the potential relevance of LFV in routine influenza monitoring.

Impact Statement

The influenza virus is prone to mutations and reassortments which leads to a considerable diversity between influenza viruses within different hosts as well as within a host. This results in a population of multiple non-identical viral influenza genomes, or quasispecies, within one patient. Quasispecies may have an impact on the patient by evolving and escaping antiviral drugs, neutralizing antibodies and cytotoxic T-cells. NGS provides the opportunity of not only detecting the majority variant in a sample, but also quasispecies at lower frequencies. This study proposes a general approach to identify low-frequency variants in patient-derived samples obtained during routine surveillance. However, it is quite challenging to distinguish the real low-frequency variants from experimental errors that occur due to PCR and NGS errors. Therefore, validated thresholds were established for low-frequency variant detection, while considering the cost and feasibility of reliable low-frequency variant detection in clinical samples. As a proof of concept for the added value of low-frequency variants for routine influenza monitoring, potential associations between patient data and whole genome sequencing data were investigated.

METHODS

Viruses and cells

A reverse genetics system of Influenza A/Bretagne/7608/2009 (A(H1N1)pdm09) and Influenza A/Centre/1003/2012 (A(H3N2)) in a bidirectional pRF483 plasmid were provided by Institut Pasteur Paris, France. Influenza viruses with a point mutation in the NA segments were obtained by reverse genetics using the QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies) and GeneJET Plasmid Miniprep Kit (Thermo Fischer) according to the manufacturer's instructions. For A/Bretagne/7608/2009, the NA-H275Y mutation (CAC → TAT) was introduced (consisting of two nucleotide mutations). For A/Centre/1003/2012, NA-E119V (GAA → GTA) was introduced (consisting of one nucleotide mutation). The NA plasmids were verified using Sanger sequencing on an Applied Biosystems Genetic Analyzer 3500 using the Big Dye Terminator Kit v3.1 following the manufacturer's instructions using primers described in Table S1.

A co-culture of Madin-Darby canine kidney (MDCK) cells and 293 T cells was maintained in Dulbecco's modified Eagle medium (DMEM) (Gibco) and 1% Penicillin Streptomycin (Gibco). The cells were transfected using FuGene HD Transfection Reagent (Promega) and Opti-MEM (Gibco). The viruses were rescued from transfected cells using an 8-plasmid reverse genetic system containing each a genomic segment. Afterwards these viruses were amplified by two cell passages.

Patient samples

Patient-derived samples were collected from the two main surveillance systems in Belgium, 'influenza-like-illness' (ILI) and 'severe-acute-respiratory-infection' (SARI). ILI cases are defined by a sudden onset of symptoms, including respiratory and systemic symptoms and fever. A SARI case is defined as an acute respiratory illness with onset within the last 10 days of respiratory symptoms, fever, and requiring hospitalization for at least 24 h. These surveillance systems are in place to follow trends of viral spread and changes in circulating influenza viruses. From these two surveillance systems, initially 253 samples were selected [35, 36]. Only samples with a genome copy number above 10^4 genomes per microlitre were retained for the LFV validation (see Results), resulting in 59 retained samples, comprising 44 samples from hospitalized SARI patients and 15 from ILI outpatients, spread over the influenza season (beginning, peak and end of epidemic). The genome copy number of 10^4 genomes per microlitre is based on the Cq values from the routine diagnostic surveillance with qPCR [37] and corresponds with a Cq of 19.53. The samples tested negative using reverse transcription polymerase chain reaction (RT-qPCR) for other respiratory viruses, including respiratory syncytial virus A and B, parainfluenza viruses 1, 2, 3 and 4, enterovirus D68, rhinoviruses, human metapneumoviruses, paraechoviruses, bocaviruses, adenovirus, coronaviruses OC43, NL63, 229 and MERS-CoV [38, 39]. Samples from ILI outpatients were categorized as mild cases (n=15). Samples from hospitalized SARI patients were categorized as moderate (n=34) or severe cases (n=10). Hospital admission (i.e. the SARI case definition) is not a disease severity indicator itself because patients could have been admitted to hospital care for isolation purposes or other medical conditions. A severe case was defined by the presence of at least one severity indicator: death, stay in an intensive care unit, need for invasive respiratory support or extracorporeal membrane oxygenation (ECMO), or the patient having acute respiratory distress syndrome (ARDS). Available patient data are listed in Table 1 with the number of patients exhibiting these characteristics.

Additionally, the median, first quartile and third quartile copy numbers of genomes per microlitre of 1273 A(H3N2) positive influenza samples from the influenza seasons 2015–2019 in Belgium were calculated and plotted with an in-house script (python 3.6) and the matplotlib 3.3.4 library [40] hiding the outliers. The boxplot including the outliers is shown in Fig. S1.

Table 1. Samples stratified according to patient data

Age (years):	<15	15–59	≥60
Beginning of epidemic (<week 4)	4	2	12
Peak of epidemic (week 4–6)	2	3	20
End of epidemic (>week 6)	4	1	11
ILI	15	**SARI**	44
Male*	25	**Female***	32
Vaccinated*	11	**Not vaccinated***	26
Antibiotics administered*	23	**No antibiotics administered***	29
Respiratory diseases	9	**No respiratory disease**	50
Cardiac disease	18	**No cardiac disease**	41
Obesity	6	**No obesity**	53
Renal insufficiency	9	**No renal insufficiency**	50
Diabetes	6	**No Diabetes**	53
Immuno-deficiency	5	**No immuno-deficiency**	54
Neuromuscular disease	7	**No neuromuscular disease**	52
Stay in ICU	5	**No stay in ICU**	54
Resulting in death*	7	**Not resulting in death***	46

*Samples for which certain patient data were unknown, were excluded for analysing that particular characteristic.

Creation of mixes of wild-type and mutant viruses

To assess the minimal percentage (i.e. AF) for a LFV to be considered truly present and not constitute a FP observation, mixes were made from the wild-type (WT) and mutant virus, created as described above, for both Influenza A/Bretagne/7608/2009 (A(H1N1)pdm09) and Influenza A/Centre/1003/2012 (A(H3N2)) with eight ratios (0, 0.1, 0.5,1, 5, 10, 20 and 100% mutant virus) (Table S3). Mixes were made in triplicate based on to the plaque forming units (PFU ml^{-1}; concentration of virus) of the infectious virus of the WT and mutant. Constructed mixes were situated mainly in the 0–5% range (Table S4), since previous studies [24, 28–31] have reported most FP being present in this range. RT-ddPCR was used to determine the genome copy numbers of the introduced mutations in the respective mixes (Supplementary Method S2 [1]).

RNA isolation and RT-qPCR

RNA of the A/Bretagne/7608/2009 (A(H1N1)pdm09) and A/Centre/1003/2012 (A(H3N2)) influenza virus mixes was extracted from culture supernatants using the Easy Mag platform (BioMérieux, #280130-#280134 and #280146) according to the manufacturer's instructions. Extraction of nucleic acids of clinical specimens was performed using the Viral RNA/DNA isolation kit (Macherey Nagel, Germany, cat No: MN 740691.4). The RNA extraction was done according to manufacturer's instructions except that the beads were not washed in buffer MV5 but instead left to dry for 10 minutes until the pellet did not appear shiny anymore.

Using 5 µl RNA for each sample, a RT-qPCR was performed using the SuperScriptIII Platinum One-Step Quantitative Kit (Invitrogen) with primers InfA_Forward, InfA_Reverse and InfA_probe. These bind to an influenza M gene section [41]. Each reaction contained 0.5 µl primer/probe, 1 µl SuperScript III RT/Platinum Taq mix, 5 µl nuclease-free water, 12.5 µl PCR Master Mix and 5 µl RNA.

PCR amplification and whole genome sequencing

To amplify RNA extracts, primers designed to target the 3′ and 5′ conserved ends of all eight segment were used as described previously [35]. Concisely, RT-PCR was used to generate sequencing amplicons in a reaction volume of 50 µl. The used protocol is based on Van den Hoecke *et al.* [32] with optimized volumes and RT-PCR conditions. Primers included CommonA-Uni12G (GCCG GAGCTCTGCAGATATCAGCGAAAGCAGG), CommonA-Uni12 (GCCAGAGCTCTGCAGATATCAGCAAAAGCAGG) and CommonA-Uni13G (GCCGGAGCTCTGCAGATATCAGTAGAAACAAGG) [32]. The reaction volumes included 25 µl RT-PCR buffer, 1 µl SuperScript III One-Step RT-PCR Platinum Taq HiFi DNA Polymerase (Invitrogen, USA), 17.375 µl dH_2O, 0.375 µl

of each primer (20 µM), 0.5 µl RnaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, USA) and 5 µl of RNA extract. An error rate (number of misincorporated nucleotides per total number of nucleotides polymerized) of lower than 1×10^{-3} by Invitrogen was estimated for the SuperScript III One-Step RT-PCR Platinum Taq HiFi DNA Polymerase [42]. The following PCR conditions were used: one cycle at 42 °C for 15 min, one cycle at 55 °C for 15 min, one cycle at 60 °C for 5 min, one cycle at 94 °C for 2 min (ramp rate: 2.5 °C s^{-1}); five cycles at 94 °C for 30 s, 45 °C for 30 s (ramp rate: 2.5 °C s^{-1}) and 68 °C for 5 min (ramp rate: 0.5 °C s^{-1}); 37 cycles at 94 °C for 30 s, 55 °C for 30 s and 68 °C for 5 min; and one cycle at 68 °C for 5 min (ramp rate: 2.5 °C s^{-1}). After purifying the generated amplicons with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, Germany) according to the manufacturers' instructions, the concentration of each purification product was quantified with the Qubit 4 Fluorometer (Invitrogen, USA) using the Qubit broad-range assay. Purified products were examined with the Agilent TapeStation (Agilent Technologies, USA) using the Agilent D5000 ScreenTape system.

Sequencing libraries using the Nextera XT DNA Sample Preparation Kit (Illumina, USA) were prepared with the purified RT-PCR products according to the manufacturer's instructions. All libraries were sequenced on an Illumina MiSeq (Illumina, USA) platform using the MiSeq V3 chemistry, as described by the manufacturer's protocol, to produce 2×250 bp paired-end reads. Generated WGS data are available in the NCBI Sequence Read Archive (SRA) [43] under accession number PRJNA692424 for the reverse genetics samples (Table S3) and PRJNA615341 for the patient-derived samples (Table S5).

Consensus genome sequences were obtained as described previously [35]. Concisely, using Trimmomatic v0.32 [44], the raw (paired-end) reads were trimmed with the following settings: 'ILLUMINACLIP:NexteraPE-PE.fa:2:30:10', 'LEADING:10', 'TRAILING:10', 'SLIDINGWINDOW:4:20', and 'MINLEN:40' retaining only paired-end reads. An appropriate reference genome for read mapping was selected from the NCBI viral genomes resource [45] for each sample. Following the GATK 'best practices' protocol [46] using Picard v2.8.3 (https://broadinstitute.github.io/picard/) and GATK v3.7, the consensus sequences for all samples were obtained. First, following best practices in the field [47–50], duplicated reads were marked with PICARD MarkDuplicates in order to remove reads originating from PCR duplicates of the same original DNA molecule which could artificially inflate AF of identified variants. This was followed by indel realignment with GATK and variant calling using GATK UnifiedGenotyper with the following options: '-ploidy 1', '--stand_call_conf 30', and '--genotype_likelihoods_model BOTH'. Subsequently, only high-quality variants with a read depth ≥200 were retained using GATK VariantFilter. Next, GATK FastaAlternateReferenceMaker was used to obtain the consensus sequence based on the called variants and selected reference sequence.

Low-frequency variant identification

Only samples with a viral load $\geq10^4$ genomes μl^{-1} (see above), and a genome median coverage higher than 1000× calculated as described previously [35], were retained. For LFV calling, the consensus genome fasta files were first indexed using Samtools faidx 1.3.1. Bowtie2-build 2.3.0 [51] was then used to generate indexes. Reads were aligned to the consensus sequence using Bowtie2 align 2.3.0 in end-to-end mode for each sample, producing SAM files that were converted into BAM with Samtools view 1.3.1. Reads were then sorted using Picard SortSam 2.8.3 (http://broadinstitute.github.io/picard/) with the option 'SORT ORDER=coordinate'. A dictionary of the reference fasta files was created using Picard CreateSequenceDictionary 2.8.3. Reads originating from PCR duplicates which could bias the observed AF of LFV were removed from read alignments using Picard MarkDuplicates 2.8.3 with the option 'REMOVE_DUPLICATES=true'. The 'LB', 'PL', 'PU' and 'SM' flags are required for downstream analysis by GATK and were set to the placeholder value 'test' using Picard AddOrReplaceReadGroups 2.8.3. The resulting BAM files were indexed by Samtools index 1.3.1 and used as input for GATK RealignerTargetCreator 3.7 [46] followed by GATK IndelRealigner 3.7 for indel realignment. The generated BAM files were then indexed using Samtools index 1.3.1 and LoFreq 2.1.3.1 [52] was used to detect LFV in 'call mode'. LoFreq separates true LFV from erroneous variant calls by using Phred-scores as probability error in a Poisson-binomial distribution. The consensus sequence of each sample was used as its own reference to call LFV, in order to avoid calling high-frequency non-reference bases due to an inadequate choice of a single reference sequence for all samples used by LoFreq to call variants, i.e. nucleotides at low allelic frequency differing from the consensus at a specific genomic position [52]. Average read position values were added to called variants using an in-house script (python 3.6) [53] (Supplementary Methods S2 [1]) based on the one provided by McCrone *et al.* [28]. Only variants with a mean reads location within the central 50% positions (i.e. between bases 62 and 188) were retained for further analysis as advised by McCrone *et al.* [28]. Variants were not further filtered based on Phred-score or mapping quality as was explored in other work, because these metrics are already internally considered by LoFreq for variant calling [28, 52]. An archive containing the code used to call variants and instructions to run it is available as part of the Supplementary Methods S2 [1].

To determine an AF threshold, the workflow described above was used to call variants in the mixes of WT and mutated A(H1N1) pdm09 and A(H3N2) strains. Receiver operating characteristic (ROC) curves for both subtypes were created using an in-house script (python 3.6) and the matplotlib 2.2.2 library [40]. Briefly, called variants were first sorted by decreasing observed AF and then numbers of true and FP variants were calculated at each called AF and plotted as a ROC curve.

Statistical analysis

All statistical analyses were performed using R-software (RStudio 1.0.153; R3.6.1). Sequencing depth and viral concentration were not introduced as covariates, because we assume that the number of amplification and sequencing errors will be limited due to the validated thresholds set up beforehand (viral concentration=10^4 copies μl^{-1}; allelic frequency=5% see Results). Furthermore, any remaining amplification and sequencing errors are expected to be distributed randomly over the genome, and these should consequently not have an influence on the statistical analysis. A glm (link function=quasipoisson) was used to assess the association between number of detected LFV and individual patient data parameters, which included disease severity (classified into mild, moderate and severe), patient age, sampling date, sex, vaccination status, presence of comorbidities and disease severity indicators. Patient data were only evaluated if at least 5% of the retained patient samples met the condition. For example, asthma was not retained because only two out of 59 patients suffered from this condition (3.4%), whereas vaccination status was retained since 11 out of 59 patients were vaccinated (18.6%). Afterwards, all identified significant associations (P <0.05) were fitted simultaneously in a glm with the same link function and only significant associations were retained. In addition to the median, the interquartile range (IQR) and the effect size were calculated.

RESULTS

Validating an AF threshold for LFV calling using an experimental quasispecies population

Sequencing errors affect the frequencies at which variants can reliably be called. At decreasing frequencies, even for high-coverage datasets, the amount of reads containing a certain variant becomes too limited to discriminate real LFV from sequencing errors. Decreasing AF thresholds for accepting LFV will consequently increase sensitivity by identifying more true positive (TP) variants, but also decrease specificity by incorporating more FP variants. It is therefore necessary to establish a validated threshold for the observed AF for accepting LFV. A mutated version of Influenza A/Centre/1003/2012 (A(H3N2)) with high genomic copy number (WT=98 475 genomes μl^{-1}; MUT=312 625 genomes μl^{-1}) was used to create a validation dataset in triplicate, for which the ground truth was known, to determine an AF threshold for accepting called LFV. The mutant included a specific mutation in the NA segment present at 100%, i.e. the well-known A(H3N2) oseltamivir resistance mutation NA-E119V [15], which served as a marker when mixing the WT and mutant virus in different ratios (Table S4). The resulting mixes of the eight ratios (theoretically: 0, 0.1, 0.5,1, 5, 10, 20 and 100% mutant virus), and their triplicates, were then subjected to WGS. High sequencing coverages were obtained for all samples and segments (Fig S2), after which LFV were called with LoFreq. Consequently, 18 TP were expected (i.e. one mutation times six ratios (0.1%, 0.5%,1%, 5%, 10%, 20%) times three replicates). Levels of read deduplication were relatively limited (min=21%, max=61%, average=36%; Table S6), and an additional investigation of variants called with and without read deduplication confirmed that read deduplication did not cause any major bias in the numbers of called variants (Supplementary Information S1). Noteworthy, seven additional variants were detected where the mean of the called frequencies over the triplicates corresponded to expected frequencies based on the TP dilution values, as observed at least in one dilution mix with an AF >5% (Supplementary Information S2). This indicates that during the propagation in cells of both the WT and mutant, other variants emerged even in the absence of external selection pressure. These seven variants were therefore removed from the variant sets used for AF threshold determination as these unexpected variants were not part of the 'ground truth', but showed sufficient evidence for being true variants instead of FP (Supplementary Figure S3). Afterwards, TP variants (i.e. the introduced NA mutation in the different mixes) and FP variants (i.e. any variant called in the different mixes that did not correspond with the WT, excluding the seven aforementioned variants) observed at varying observed AFs were expressed in a ROC curve (Fig. 1), considering triplicate values as independent values. The AFs used in the ROC curve are the observed percentages of the NA mutation as determined with Lofreq. A ROC curve expresses the relationship between sensitivity and specificity for a benchmarked experiment where the ground truth is known by varying a discrimination threshold (here the AF) and plotting the false positive rate (i.e. 1-specificity) and sensitivity on the x- and y-axis, respectively. A perfect assay where all FP are separated from TP is characterized by a ROC curve with a right angle that follows the upper left boundary of the plot (Fig. 1).

For A(H3N2), no FP and 50.00% of TP (n=9/18) were called at an observed AF of 4.82% or higher. This seemingly low sensitivity is explained by the construction of the dataset which aimed at providing a high resolution at low AF to determine the limit of detection and therefore contained half of the variants at an AF lower than 5%. Decreasing the AF threshold increased the sensitivity but impaired a high cost in specificity (Table 2). At an observed AF of 1%, 83.33% of TP (n=15/18) were recovered at a cost of 289 FP. The highest sensitivity was obtained at an observed AF of 0.37%, where 88.89% (n=16/18) of variants were called at a cost of 847 FP. An AF cut-off of 5% was therefore selected as a conservative AF threshold to explicitly minimize the amount of called FP variants to be used for exploring potential associations with host characteristics (see below). Evaluation of the benchmark dataset created for A(H1N1)pdm09 exhibited the same trends, and confirmed 5% to be an adequate threshold to avoid the inclusion of FP observations (Supplementary Information S3).

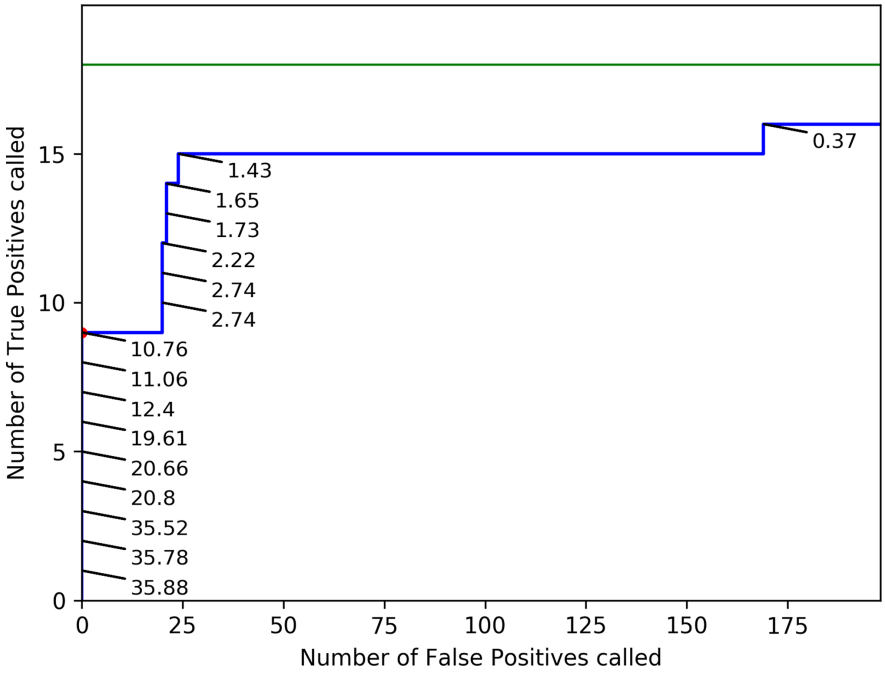

Fig. 1. ROC curve for validating an AF threshold using an A(H3N2) benchmark dataset. The green line represents a theoretical scenario where a perfect variant caller identifies all 18 TP before any FP are called (i.e. perfect sensitivity and specificity). The blue line represents the numbers of observed TP and FP in the benchmark dataset for A(H3N2) at decreasing thresholds for the observed AF of called variants. Observed AF of TP are indicated on the graph. AF thresholds used to create the ROC curve are the numbers plotted in the figure (as percentages). The numbers of FP and TP at the threshold of 5% AF employed for the analysis of patient-derived datasets is depicted by a red dot (no additional TP or FP were observed between 5 and 10.76%). More detailed values are available in Table 2. AF=Allelic Frequency; ROC=Receiver operating characteristic; FP=False Positive; TP=True Positive.

Selection of patient-derived samples based on their genome copy number

For the described validation of an AF threshold of 5% based on the experimentally constructed benchmark dataset, all mixes always contained very high genome copy numbers (≥10^5 genomes µl^{-1}, see above). It has been previously established that the genome copy number and titre of samples can also impact LFV calling. Prior research by McCrone *et al.* indicated that samples with a copy number of ≥10^5 genomes µl^{-1} are acceptable, while samples with a copy number ranging between 10^3–10^5 genomes µl^{-1} should be sequenced in duplicate to reduce FP [28]. In routine surveillance, only a limited number of samples however have a copy number of ≥10^5 genomes µl^{-1}. Only 12 out of 253 sequenced samples of the Belgian influenza season 2016–2017 had a genomic copy number ≥10^5 genomes µl^{-1} (Table S5). This was not due to sample selection bias, since the median of 1273 A(H3N2) positive influenza samples from the influenza seasons 2015–2019 in Belgium was 1168.85 genomes µl^{-1} (IQR: 88.70–8907.89 genomes µl^{-1}) (Fig. 1), with a median associated Cq value of 22.52 (IQR: 19.48–26.68), which corresponds to other observations from the literature [54–56].

Table 2. Number of TP, FP, sensitivity, and specificity at different AF thresholds for the A(H3N2) benchmark dataset. Although the specificity remains high due to the size of the negative class (all positions in the genome that are not positives), the number of FP increases dramatically at lower AF, rapidly exceeding more than ten-fold the number of TP. AF=Allelic Frequency; FP=False Positive; TP=True Positive. *: Sensitivity is considered over the full dataset, and not only variants expected at specific AF; see results for further details

Observed AF (%)	no. of TP	no. of FP	Sensitivity (%)*	Specificity (%)
10.0	8	0	44.44	100.00
5.0	9	0	50.00	100.00
2.0	12	86	66.67	99.97
1.0	15	289	83.33	99.90
0.5	15	678	83.33	99.75

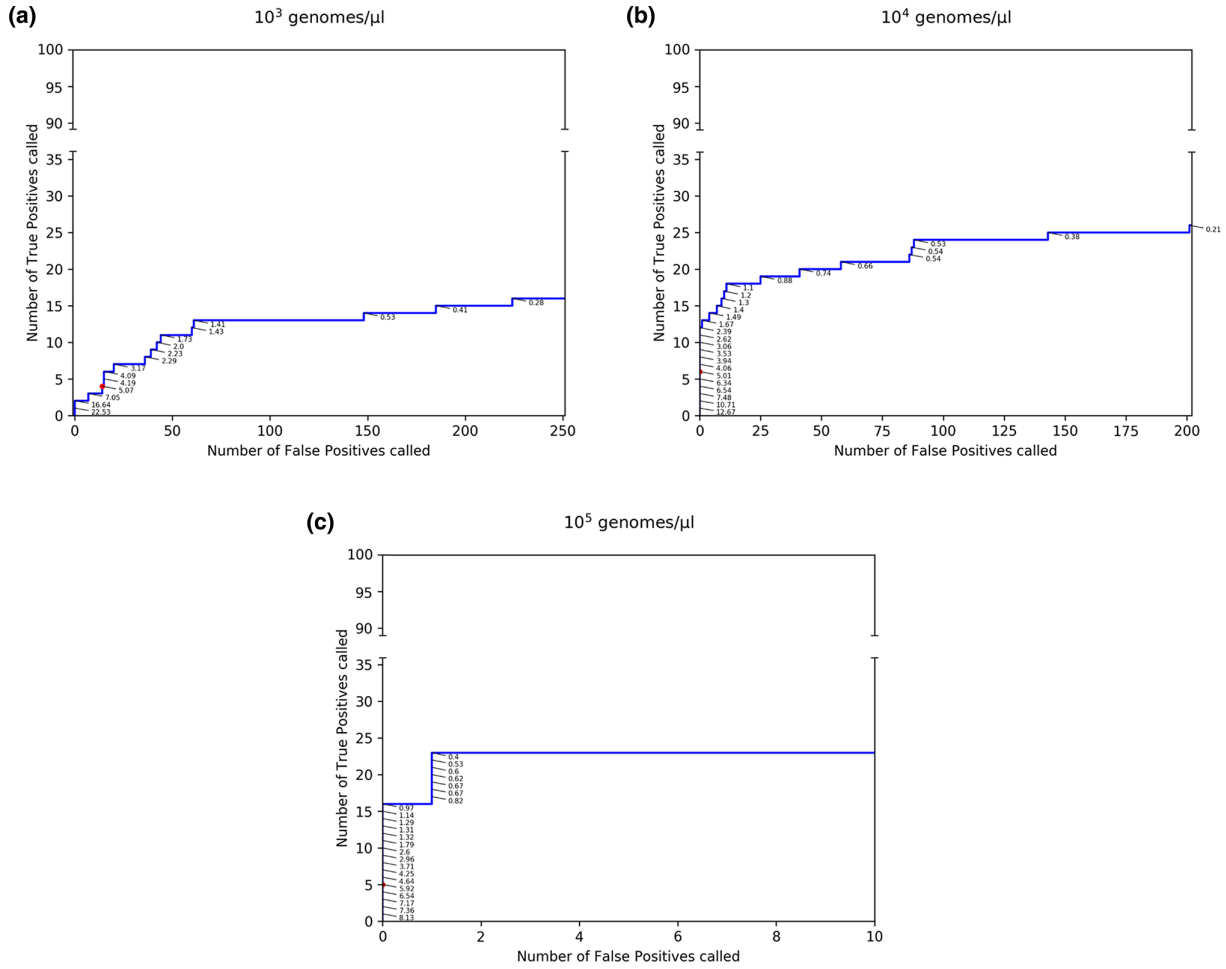

Fig. 2. ROC curves to validate an AF threshold using an A(H3N2) benchmark dataset at different genome copy numbers. Observed TP (out of the 100 expected) and FP counts in the benchmark datasets provided by McCrone *et al*. [28] at variable genome copy numbers. The blue line represents observed TP and FP counts in the benchmark dataset for A(H3N2) at variable thresholds for the AF. Observed AF of called TP are plotted in the figure as percentages. The numbers of observed FP and TP at the threshold of 5% AF employed for the analysis of patient-derived datasets is depicted by a red dot. More detailed values are available in Table 3. Abbreviations: AF=Allelic Frequency; FP=False Positive; TP=True Positive.

To evaluate the impact of adopting a more relaxed genome copy number threshold, we investigated the sensitivity and specificity of the LFV calling workflow on a benchmark dataset containing lower genome copy numbers, for which reference samples of mixes of specific variants at varying targeted AFs and varying initial genomics copy numbers produced and sequenced by McCrone *et al*. [28] were analysed with the same method as described previously. Samples used for this analysis were produced by McCrone *et al*. as an experimental within-host population by inserting 20 mutations in a WSN33 virus genetic background and then diluted to generate five targeted allelic frequencies (5, 2, 1, 0.5 and 0.2%) and three genomic litres (10^3, 10^4 and 10^5 genomes μl^{-1}) [28]. Titres, targeted allelic frequencies and SRA accession numbers of the samples used can be found in Table S2. For samples with 10^3 genomes μl^{-1}, no FP and 2% of TP (n=2/100) were called at an observed AF of ≥16.64%. These particularly low sensitivities are again the result of the dataset encompassing a majority of low allelic frequency variants. The highest sensitivities, 23, 26 and 16% for genomic litres of respectively 10^5, 10^4 and 10^3, were obtained at an observed AF of 0.40, 0.21 and 0.28% at a cost of 1, 201 and 224 called FP, respectively (Fig. 2, Table 3).

Comparison of results for a viral load of ≥10^5 genomes μl^{-1} of Table 2 and Table 3, indicates similar trends with increasing AF increasing specificity whilst penalizing sensitivity. The sensitivities of the two benchmark datasets in Table 2 and Table 3 are however not directly comparable because the truth set of mutations is present at different AF, resulting in lower sensitivity values for the McCrone dataset because more real variants were present in the observed AF range of 1–5%. The previously selected AF threshold of 5% was therefore shown to be a conservative value for filtering out FP variants in datasets obtained from samples with low initial genomic copy numbers because despite removing many TP variants, it also effectively safeguards against including

Table 3. Number of TP, FP, sensitivity, and specificity at different AF thresholds using a A(H3N2) benchmark dataset at different genome copy numbers. Although the specificity remains high due to the size of the negative class (all positions in the genome that are not positives), the number of FP increases dramatically at lower observed AF, an effect which is more pronounced at lower genome copy numbers. *: Sensitivity is considered over the full dataset, and not only variants expected at specific AF; see results for further details

Viral load (genomes μl^{-1})	Observed AF (%)	no. of TP	no. of FP	Sensitivity (%)*	Specificity (%)
10^5	10.0	0	0	0.00	100.00
	5.0	5	0	5.00	100.00
	2.0	10	0	10.00	100.00
	1.0	15	0	15.00	100.00
	0.5	22	1	22.00	99.99
10^4	10.0	2	0	2.00	100.00
	5.0	6	0	6.00	100.00
	2.0	12	0	12.00	100.00
	1.0	18	17	18.00	99.97
	0.5	24	67	24.00	99.90
10^3	10.0	2	1	2.00	99.99
	5.0	4	14	4.00	99.98
	2.0	9	41	9.00	99.87
	1.0	13	83	13.00	99.36
	0.5	14	154	14.00	99.76

FPs for genome copy numbers at 10^4–10^5, but not at 10^3 genomes μl^{-1}. A minimal genome copy number of 10^4 genomes μl^{-1} was therefore enforced for the clinical dataset.

Prevalence of LFV in clinical samples

LFV calling was performed on the 59 retained samples with a genome copy number of $\geq 10^4$ genomes μl^{-1} from the Belgian influenza 2016–2017 A(H3N2) season. When the selected threshold of 5% AF was used, at least 20 LFV were detected in seven samples, while for 30 samples between 0 and 20 LFV were detected. Finally, 22 samples did not reveal any LFV (Supplementary Method S2 [1]). Across all samples, LFV at 56 genomic positions were detected in two or more patients, including eight located in PB2, six in PB1, 14 in PA, 12 in HA, six in NP, three in NA, one in MP and six in NS. The majority of these variants were detected at a low observed AF of 5–20%.

Patient data associated with prevalence of LFV

To investigate the potential relevance of LFV for routine influenza monitoring, a proof of concept investigation based on associations of LFV with patient data was performed. The association of patient data with the number of detected LFV was investigated. After an initial glm analysis where all patient data were evaluated individually, disease severity, antibiotics use and age resulted in a significant association. In a second step, a glm was fitted including the three significant patient data simultaneously, which only resulted into a significant result for disease severity. The number of detected LFV was observed to be significantly higher in ILI cases (i.e. mild cases) compared to SARI cases (i.e. moderate and severe cases) (Table 4; Supplementary Method S2 [1]).

Table 4. Statistically significant associations between number of LFV in clinical samples and patient data. Results include the median, first quartile and third quartile of the number of detected LFV across the 59 retained samples, and also *P*-value and effect size. The interpretation of the odds ratio values commonly published in the literature are: <1.68 (small effect), 1.68–3.47 (moderate effect) and >=6.71 (large effect) [70]. ILI cases comprise the mild cases, while the SARI cases include moderate and severe cases. CI=Confidence interval

Patient data	Median	*P*-value	Effect size [CI]
Disease Severity	**Mild**: 19 [3.5–60] **Moderate/Severe**: 1 [0–3]	2.67E-08	26.40 [10.89–83.88]

Additionally, associations between patient data and the proportion of nucleotides at their specific genomic positions, including both LFV and high-frequency variants, were evaluated. Although several associations were identified, these were all below acceptable statistical thresholds. These results are therefore provided in the Supplementary Information S4 for informative purposes only and not further considered below.

DISCUSSION

Since the dynamics of quasispecies can afford influenza a considerable advantage on genetic fitness during within-host evolution, quasispecies information might be relevant for future clinical interventions and epidemiological investigation. HTS renders it nowadays feasible to explore viral quasispecies in patient-derived samples by detecting LFV. However, many challenges remain to obtain reliable results in order to introduce LFV in routine surveillance, in which sampling and funding are often limited. Although HTS enables deep sequencing, it becomes difficult to distinguish sequencing errors from real LFV at low AF. The first goal of this study was to establish an AF threshold for retaining LFV using mixes of a WT and NA-E119V-mutant influenza A(H3N2) virus with different proportions to create a benchmark population that was sequenced followed by LFV calling with LoFreq. While multiple other low-frequency variant callers exist [57–61], LoFreq has been shown to perform particularly well on short read sequencing of virus samples, especially when considering specificity [62, 63]. Other variant callers could alternatively be used as part of the validation approach presented in the current study by other scientists using other software packages. An AF cut-off of 5% was selected as the minimal AF at which no FP variants were called in the experimentally constructed benchmark A(H3N2) population. An additional exploratory analysis with mixes from the A(H1N1) subtype, which included two nucleotide mutations resulting in the NA-H275Y amino acid mutation, confirmed this as being a robust threshold also applicable to other subtypes (Supplementary Information S3). Since the A(H3N2) and A(H1N1) benchmark populations only contained a single and two nucleotide mutations, respectively, publically available data containing more mutations were also considered. The dataset from McCrone *et al*. includes 20 point mutations and also an extra data point at a theoretical AF of 2%, in contrast to our sequenced A(H3N2) population containing a theoretical AF gap between 1 and 5%. Analysis of this dataset with our workflow similarly confirmed 5% to be a robust AF threshold (Fig. 2). This threshold prioritizes specificity over sensitivity, but is context-dependent for three reasons. Firstly, although the established sensitivity of 50% at 5% observed AF (Table 2) may appear low, the benchmark dataset was purposefully constructed to assess the limit of detection of our workflow, and therefore contained half of the inserted variants at frequencies lower than 5%. Conversely, as a result of the choice of thresholds, all variants present at ≥5% in the benchmark dataset were correctly called. Secondly, since our aim was to evaluate associations of LFV with patient data as a proof of concept, we prioritized specificity to minimize potential FP LFV included within the statistical analysis. Depending on the application scope, this AF threshold can be decreased to increase sensitivity if the cost in specificity is deemed acceptable (e.g. approaches that prioritize finding as many LFV as possible). Thirdly, AF thresholds are coverage-dependent once coverage drops below a certain turnkey point [64], with decreasing coverages typically requiring increased AF thresholds. As both the validation dataset and clinical samples consisted of high-coverage data, our established value of 5% should only be applied to high-coverage influenza datasets. Through our emphasis on specificity, the selected AF threshold of 5% is high compared to other AF thresholds reported in other studies in the literature. Gelbart *et al*. [65] investigated the genetic diversity of different viruses, and used a minimum AF threshold of 1% for highly concentrated samples including human immunodeficiency virus, respiratory syncytial virus, and cytomegalovirus. Orton *et al*. [66] focussed on modelling sequencing errors and distinguishing them from real viral variants using foot-and-mouth disease virus as case study. They established a minimum AF threshold of 0.5%, although this was only tested on control samples that were very highly concentrated (10^6 plasmid μl^{-1}). King *et al*. [67] evaluated laboratory and bioinformatic pipelines to accurately identify LFV in viral populations using foot-and-mouth disease as a case study, using an AF threshold of 0.2% for highly concentrated samples (10^7 copies), but observed more errors when a reduced RNA input (10^5 copies) was used and even found consensus-level errors at (very) low RNA inputs (10^3 copies).

Previous research has indicated that besides correcting for sequencing errors, the viral load and genome copy number of samples also affect LFV calling, independently of sequencing considerations. In this study, the SuperScript III One-Step RT-PCR Platinum Taq HiFi DNA Polymerase with an estimated error rate of less than 1×10^{-3} misincorporated nucleotides per total number of nucleotides polymerized was used to amplify the virus. This error rate will have a larger impact on samples with low viral loads, because they are more likely to propagate PCR-amplification errors that can result in increased FP variant detections [28]. A genome copy number of 10^5 genomes μl^{-1} was recommended by McCrone *et al*. and a copy number of 10^3–10^5 genomes μl^{-1} was considered acceptable if sequenced in duplicate. However, the application of these recommendations to routine surveillance may prove too restrictive as 10^5 genomes μl^{-1} is an extremely high copy number for samples encountered in routine influenza surveillance (Fig. S1), where it is already a considerable challenge to acquire the necessary funds to simply switch from Sanger sequencing the HA and NA segments to WGS. As the genome copy number of our experimental dataset was very high (>10^5 genomes μl^{-1}), we employed the experimental within-host population produced by McCrone *et al*. [28] at a genomic input of 10^3, 10^4 and 10^5 genomes μl^{-1} with our workflow to evaluate FP counts at lower genome copy numbers when enforcing the same 5% AF threshold. We found that also at 10^4 genomes μl^{-1}, no FP were detected, but FP were found at 10^3 genomes μl^{-1} (Table 3). Similar to our experimentally constructed A(H3N2) benchmark dataset, sensitivities were (very) low because the large majority

of LFV were present at AF below 5%. Notwithstanding, a direct comparison of our results with those reported by McCrone *et al*. is not possible for several reasons. Firstly, McCrone *et al*. used *P*-values as a threshold with either deepSNV or LoFreq to determine effects on sensitivity and specificity in samples of varying targeted AF, whereas we used the observed AF as a threshold with LoFreq with default settings (i.e. *P*-value dynamically adapted as part of a Bonferroni multiple test correction) to determine an AF threshold favouring optimal specificity. Secondly, high specificity at low AF could be obtained by McCrone *et al*. by using deepSNV on both mutated samples and control samples containing the same genetic background. This was initially done with LoFreq on our benchmark datasets using the WT samples as controls and resulted in overall higher specificity and lower sensitivity at very low AF (unpublished results), but does not reflect routine influenza monitoring where no control samples are available for clinical samples to begin with. Thirdly, the samples used by McCrone *et al*. were biassed toward very low AF for the TP, which had a large effect on the sensitivity.

The second goal of this study was to evaluate the prevalence of LFV in actual clinical samples collected during routine influenza monitoring, using 59 influenza A(H3N2) samples from the 2016–2017 Belgian Influenza season with a genome copy number $\geq 10^4$ genomes μl^{-1} and retaining only LFV detected at ≥5% AF. It was observed that seven of the 59 samples had at least more than 20 LFV, 30 of the 59 samples had between 0 and 20 LFV, and 22 of the 59 samples did not contain any LFV.

The third goal of this study was to explore potential associations between patient data and the presence and frequency of LFV as a proof of concept for the relevance of LFV analysis in routine influenza surveillance. Statistically significant associations were found between high numbers of LFV and mild cases. It has been suggested in the literature for other viruses that within-host diversity can be driven by host selection pressure [68, 69]. In contrast to our results where more LFV were observed in mild cases, Simon *et al*. observed higher diversity within the PA, HA and NA segments in severe cases compared to mild cases [18]. Additionally, we evaluated potential associations between patient data and the proportion of nucleotides at specific genomic positions. Several associations were found, however, these were below acceptable statistical thresholds (Supplementary Information S4). We are aware, however, of the low statistical power of the association study due to the small sample size of 59 patients and unequal representation of LFV among the patient data groups. More reliable associations will therefore require larger sample sizes in future studies. However, these results show the potential added value to understand viral evolution in relation to the host, but more research is needed.

In conclusion, HTS of clinical influenza samples allows to examine LFV during human infections. Our work provides a general approach for LFV detection by delineating thresholds that balance the number of FP against the feasibility of quasispecies investigation in actual samples collected in the context of routine surveillance programmes. As a proof of concept, several relevant associations with patient data were found while considering LFV, which suggests that the relevance of LFV for influenza monitoring is currently under-valued and could contribute to a better understanding of disease. Although additional validation will be necessary, it could be of great benefit to apply the proposed approach on samples collected during routine influenza monitoring.

Funding information
This study was financed by Sciensano through the Be Ready project.

Acknowledgements
We thank the technicians of the service Transversal activities in Applied Genomics at Sciensano, Belgium for performing the Sanger and Next Generation Sequencing runs.

Author contributions
Conceptualization: N.R., S.D.K., K.V., X.S., S.V.G. Project Administration: N.R. Data Curation: L.V.P., T.D., K.V., I.T. Methodology: L.V.P., T.D., N.R., K.V. Software: T.D., L.V.P. Formal Analysis: L.V.P., T.D., M.V. Validation: T.D., L.V.P., K.V. Investigation: L.V.P., T.D. Visualization: L.V.P., T.D. Writing – original draft preparation: L.V.P., T.D., N.R., K.V. Writing – review & editing: all authors. Funding Acquisition: N.R. Resources: S.D.K., S.V.G., I.T. Supervision: N.R., K.V.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
A central ethical committee and the local ethical committees of each participating hospital approved the SARI surveillance protocol (reference AK/12-02-11/4111; in 2011: Centre Hospitalier Universitaire St-Pierre, Brussels, Belgium; from 2014 onwards: Universitair Ziekenhuis Brussel, Brussels, Belgium). Informed verbal consent was obtained from all participants or parents/guardians.

References

1. Van Poelvoorde L, Delcourt T, Vuylsteke M, De Keersmaecker SCJ, Thomas I, *et al*. A general approach to identify low-frequency variants within influenza samples collected during routine surveillance. *Figshare* 2022.
2. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza a viruses. *Microbiol Rev* 1992;56:152–179.
3. Mosnier A, Caini S, Daviaud I, Nauleau E, Bui TT, *et al*. Clinical characteristics are similar across type A and B influenza virus nfections. *PLoS ONE* 2015;10:1–13.
4. Kosik I, Yewdell JW. Influenza hemagglutinin and neuraminidase: yang proteins coevolving to thwart mmunity. *Viruses* 2019;11:E346.
5. WHO. Influenza (seasonal). Influenza; (n.d.). https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) [accessed 26 March 2018].

6. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. *J Virol* 2010;84:9733–9748.
7. Nowak MA. What is a quasispecies? *Trends Ecol Evol* 1992;7:118–121.
8. Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog* 2010;6:e1001005.
9. Andino R, Domingo E. Viral quasispecies. *Virology* 2015;479–480:46–51.
10. Webster RG, Laver WG, Air GM, Schild GC. Molecular mechanisms of variation in influenza viruses. *Nature* 1982;296:115–121.
11. Hurt AC, Barr IG. Influenza viruses with reduced sensitivity to the neuraminidase inhibitor drugs in untreated young children. *Commun Dis Intell Q Rep* 2008;32:57–62.
12. Forns X, Purcell RH, Bukh J. Quasispecies in viral persistence and pathogenesis of hepatitis C virus. *Trends Microbiol* 1999;7:402–410.
13. Van Poelvoorde LAE, Saelens X, Thomas I, Roosens NH. Next-generation sequencing: an eye-opener for the surveillance of antiviral resistance in nfluenza. *Trends Biotechnol* 2020;38:360–367.
14. Zhou B, Donnelly ME, Scholes DT, St George K, Hatta M, *et al*. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza A viruses. *J Virol* 2009;83:10309–10313.
15. Boivin G. Detection and management of antiviral resistance for influenza viruses. *Influenza Other Respir Viruses* 2013;7 Suppl 3:18–23.
16. Xu Y, Lewandowski K, Downs LO, Kavanagh J, Hender T, *et al*. Nanopore metagenomic sequencing of influenza virus directly from respiratory samples: diagnosis, drug resistance and nosocomial transmission, United Kingdom, 2018/19 influenza season. *Euro Surveill* 2021;26.
17. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GCM, *et al*. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 2013;342:976–979.
18. Simon B, Pichon M, Valette M, Burfin G, Richard M, *et al*. Whole genome sequencing of A(H3N2) influenza viruses reveals variants associated with severity during the 2016. *Viruses* 2019;11:E108.
19. Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, *et al*. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* 2021;37:1673–1680.
20. Tonkin-Hill G, Martincorena I, Amato R, Lawson ARJ, Gerstung M, *et al*. Patterns of within-host genetic diversity in SARS-CoV-2. *Elife* 2021;10:e66857.
21. Luksza M, Lässig M. A predictive fitness model for influenza. *Nature* 2014;507:57–61.
22. Pompei S, Loreto V, Tria F. Phylogenetic properties of RNA viruses. *PLoS ONE* 2012;7:e44849.
23. Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, *et al*. Influenza a virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe* 2014;16:691–700.
24. Xue KS, Stevens-Ayers T, Campbell AP, Englund JA, Pergam SA, *et al*. Parallel evolution of influenza across multiple spatiotemporal scales. *Elife* 2017;6:e26875.
25. Rogers MB, Song T, Sebra R, Greenbaum BD, Hamelin M-E, *et al*. Intrahost dynamics of antiviral resistance in influenza A virus reflect complex patterns of segment linkage, reassortment, and natural selection. *mBio* 2015;6:e02464-14.
26. Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP, *et al*. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *J Infect Dis* 2011;203:168–174.
27. Trebbien R, Pedersen SS, Vorborg K, Franck KT, Fischer TK. Development of oseltamivir and zanamivir resistance in influenza A(H1N1)pdm09 virus, Denmark, 2014. *Euro Surveill* 2017;22:1–8.
28. McCrone JT, Lauring AS. Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant alling. *J Virol* 2016;90:6884–6895.
29. Xue KS, Bloom JD. Linking influenza virus evolution within and between human hosts. *Virus Evol* 2020;6:veaa010.
30. Dinis JM, Florek KR, Fatola OO, Moncla LH, Mutschler JP, *et al*. Deep sequencing reveals potential antigenic variants at low frequencies in influenza a virus-infected humans. *J Virol* 2016;90:3355–3365.
31. Debbink K, McCrone JT, Petrie JG, Truscon R, Johnson E, *et al*. Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLOS Pathog* 2017;13:e1006194.
32. Van den Hoecke S, Verhelst J, Vuylsteke M, Saelens X. Analysis of the genetic diversity of influenza a viruses using next-generation DNA sequencing. *BMC Genomics* 2015;16:1–23.
33. Kundu S, Lockwood J, Depledge DP, Chaudhry Y, Aston A, *et al*. Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin Infect Dis* 2013;57:407–414.
34. Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. *Nat Rev Genet* 2014;15:56–62.
35. Van Poelvoorde LAE, Bogaerts B, Fu Q, De Keersmaecker SCJ, Thomas I, *et al*. Whole-genome-based phylogenomic analysis of the Belgian 2016-2017 influenza A(H3N2) outbreak season allows improved surveillance. *Microb Genom* 2021;7.
36. Van Poelvoorde LAE, Vanneste K, De Keersmaecker SCJ, Thomas I, Van Goethem N, *et al*. (n.d.) whole-genome viral sequence analysis reveals mutations associated with influenza patient data. *Front Microbiol*
37. Thomas I, Barbezange C, Hombrouck A, Gucht SV, Weyckmans J, *et al*. Virological surveillance of influenza in Belgium; season 2016-2017. *Sciensano Influenza Report* 2017:1–33.
38. Brittain-Long R, Nord S, Olofsson S, Westin J, Anderson L-M, *et al*. Multiplex real-time PCR for detection of respiratory tract infections. *J Clin Virol* 2008;41:53–56.
39. Hombrouck A, Sabbe M, Van Casteren V, Wuillaume F, Hue D, *et al*. Viral aetiology of influenza-like illness in Belgium during the influenza A(H1N1)2009 pandemic. *Eur J Clin Microbiol Infect Dis* 2012;31:999–1007.
40. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 2007;9:90–95.
41. WHO. CDC protocol of realtime RTPCR for influenza A(H1N1); (n.d.). https://cdn.who.int/media/docs/default-source/influenza/molecular-detention-of-influenza-viruses/protocols_influenza_virus_detection_feb_2021.pdf?sfvrsn=df7d268a_5 [accessed 28 April 2009].
42. Invitrogen. Advanced analysis with the superscript iv oone-step rt-pcr system; 2018. https://assets.thermofisher.com/TFS-Assets/BID/Reference-Materials/advanced analysis-superscript-iv-one-step-rt-pcr-system-white-paper.pdf [accessed 18 January 2022].
43. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res* 2011;39:D19–21.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
45. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res* 2015;43:D571–7.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
47. Broad Institute. Data pre-processing for variant discovery. GATK; (n.d.). https://gatk.broadinstitute.org/hc/en-us/articles/%20360035535912-Data-pre-processing-for-variant-discovery [accessed 14 January 2022].
48. Marx V. How to deduplicate PCR. *Nat Methods* 2017;14:473–476.

49. Kassahn KS, Holmes O, Nones K, Patch A-M, Miller DK, *et al*. Somatic point mutation calling in low cellularity tumors. *PLOS ONE* 2013;8:e74380.
50. Tian S, Yan H, Kalmbach M, Slager SL. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* 2016;17:403.
51. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
52. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, *et al*. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–11201.
53. van Rossum G. Python tutorial. CWI Report CS-R9526. 1995., pp. 1–65.
54. Kenmoe S, Tchendjou P, Moyo Tetang S, Mossus T, Njankouo Ripa M, *et al*. Evaluating the performance of a rapid antigen test for the detection of influenza virus in clinical specimens from children in Cameroon. *Influenza Other Respir Viruses* 2014;8:131–134.
55. Caselton DL, Arunga G, Emukule G, Muthoka P, Mayieka L, *et al*. Does the length of specimen storage affect influenza testing results by real-time reverse transcription-polymerase chain reaction? An analysis of influenza surveillance specimens, 2008 to 2010. *Euro Surveill* 2014;19.
56. Duchamp MB, Casalegno JS, Gillet Y, Frobert E, Bernard E, *et al*. Pandemic A(H1N1)2009 influenza virus detection by real time RT-PCR: is viral quantification useful? *Clin Microbiol Infect* 2010;16:317–321.
57. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, *et al*. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* 2012;3:811.
58. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, *et al*. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 2016;44:e108.
59. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing; 2012. http://arxiv.org/abs/1207.3907 [accessed 18 January 2022].
60. Verbist BMP, Thys K, Reumers J, Wetzels Y, Van der Borght K, *et al*. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* 2015;31:94–101.
61. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22:568–576.
62. Mohammed KS, Kibinge N, Prins P, Agoti CN, Cotten M, *et al*. Evaluating the performance of tools used to call minority variants from whole genome short-read data. *Wellcome Open Res* 2018;3:21.
63. Deng Z-L, Dhingra A, Fritz A, Götting J, Münch PC, *et al*. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. *Brief Bioinform* 2021;22:bbaa123.
64. Stead LF, Sutton KM, Taylor GR, Quirke P, Rabbitts P. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. *Hum Mutat* 2013;34:1432–1438.
65. Gelbart M, Harari S, Ben-Ari Y, Kustin T, Wolf D, *et al*. Drivers of within-host genetic diversity in acute infections of viruses. *PLOS Pathog* 2020;16:e1009029.
66. Orton RJ, Wright CF, Morelli MJ, King DJ, Paton DJ, *et al*. Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* 2015;16:229.
67. King DJ, Freimanis G, Lasecka-Dykes L, Asfor A, Ribeca P, *et al*. A systematic evaluation of high-throughput sequencing approaches to identify low-frequency single nucleotide variants in viral populations. *Viruses* 2020;12:E1187.
68. Honce R, Schultz-Cherry S. They are what you eat: Shaping of viral populations through nutrition and consequences for virulence. *PLOS Pathog* 2020;16:e1008711.
69. Lin S-R, Yang T-Y, Peng C-Y, Lin Y-Y, Dai C-Y, *et al*. Whole genome deep sequencing analysis of viral quasispecies diversity and evolution in HBeAg seroconverters. *JHEP Rep* 2021;3:100254.
70. Chen H, Cohen P, Chen S. How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation* 2010;39:860–864.