World Health Organization

Next-generation sequencing of influenza viruses

General information for national influenza centres

October 2019

ISBN 978-92-4-000071-1

Suggested citation. Next-generation sequencing of influenza viruses: general information for national influenza centres. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

Contents

Abbreviations

bp	base pairs
DNA	deoxyribonuleic acid
GISRS	Global Influenza Surveillance and Response System
HA	haemagglutinin
HI	haemagglutination inhibition
M gene	gene coding for matrix protein (M1) and transmembrane protein (M2)
NA	neuraminidase
NGS	next-generation sequencing
NIC	national influenza centre
RNA	ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
WHO	World Health Organization

1 Objective

This document is intended for national influenza centres (NICs) of the World Health Organization (WHO) Global Influenza Surveillance and Response System (GISRS) and other influenza laboratories performing virological surveillance. The aim is to provide information on implementing next-generation sequencing (NGS) for genetic characterisation of influenza viruses.

2 Scope

The document provides information on implementing NGS for genetic characterisation of influenza viruses primarily for virological surveillance, but may be used for research or outbreak investigations. It does not cover other scenarios such as diagnosis, and metagenomic detection of other virus pathogens, although some of the general points are applicable to those scenarios.

3 Target audience

The target audience is NICs and other influenza laboratories performing virological surveillance.

4 Introduction to influenza virus gene sequencing

Traditionally, influenza viruses collected for seasonal surveillance to inform influenza vaccine virus recommendations were sequenced, to help to explain and complement results from antigenic characterisation. Initially, full or partial haemagglutinin (HA) gene segments were sequenced because it was thought that much of the information regarding the questions around antigenicity was contained in the HA segments, specifically the HA1 component. However, Sanger-based sequencing technologies did not allow processing of enough specimens to keep pace with the numbers of viruses being subjected to haemagglutination inhibition (HI) testing; hence, only a subset of the HA genes were sequenced. In recent years, with increasing use of antiviral agents targeting influenza gene products, the neuraminidase (NA) and matrix (M) gene segments were also sequenced because they contained sites where encoded amino acid changes pertained to potential antiviral resistance/reduced susceptibility. In this context, Sanger-based sequencing of full-length influenza genes – in particular, the HA, NA and M genes – still provides useful information for genetic surveillance purposes.

Full genome sequencing was performed more frequently for zoonotic viruses than for seasonal viruses because it helped in the assessment of pandemic risk. Mutations associated with mammalian adaptation of influenza type A viruses occur in multiple gene segments of the influenza genome, and identification of these can provide insight into the likelihood of an influenza virus adapting to a new host. Furthermore, other features of the genome can be evaluated; for example, the acquisition or insertion of multiple basic amino acid residues at the protease cleavage site in the HA0 protein that

separates HA1 and HA2, or the presence or absence of other pathogenic determinants in other gene segments which may be species dependent.

The advent of NGS technologies has revolutionized genomic sequencing by increasing throughput, accuracy and cost effectiveness. The high data output (gigabytes) of NGS instruments makes it possible to sequence the genomes of multiple influenza viruses, thereby reducing the cost per genome compared to prior approaches (i.e., Sanger) and, depending on the scheme, improve timeliness for sequencing a large number of samples. However, the cost of sequencing just a few genomes within an institute remains high if using NGS. Additionally, the data produced by NGS allows researchers to explore minor variants in the virus population that may have implications in virus evolution or the development of antiviral resistance or other characteristics.

The sequence data produced by NGS and Sanger sequencing platforms form a vital component in the characterisation of seasonal human epidemic viruses, enzootic viruses and zoonotic viruses that have pandemic potential. Ultimately, sequence data and phenotypic data are used together in the development of vaccines by GISRS. The timely/rapid provision of accurate sequence data that are shared through publicly accessible databases (e.g., GISAID, GenBank, etc.), helps protect the world against influenza.

5 NGS methods – general

NGS sequencing technologies can be separated into two main categories: sequencing by synthesis and direct single molecule sequencing.

Instruments for sequencing by synthesis are available from Illumina, Ion Torrent and Pacific Biosciences. The process involves first isolating molecules of DNA (RT-PCR products for influenza) using a well, bead or array, then monitoring the nucleotide base addition with a reporter, while copying the isolated DNA molecules. In sequencing, some synthesis instruments amplify the isolated DNA on the instrument to amplify the reporter signal (e.g., Ion Torrent and Illumina), whereas others isolate a single DNA strand and report the nucleotide synthesis of the complementary strand as it is copied (e.g., Pacific Biosciences).

Direct single molecule sequencing is currently only available on sequencers developed by Oxford Nanopore Technologies. By this technique, DNA, RNA or even protein can be sequenced by monitoring the resistance across a molecular membrane as the strand of nucleotides or amino acids moves through a molecular pore. Currently, this instrument is used primarily for DNA sequencing because sequencing of other targets is less well developed. The sequencing device measures resistance in a current; hence, the instrument can be quite small and portable, meaning that it can be used in outbreak investigations in the field. Direct sequencing of RNA is currently unique to nanopore sequencing; it is probably best suited to outbreak or initial investigations of potentially zoonotic viruses, because it does not offer the high throughput of the other NGS platforms.

6 Considerations for implementation of NGS by NICs

6.1 Influenza samples

Laboratories interested in the implementation of NGS for influenza surveillance should assess the average number of samples received by the laboratory within a given time frame, and carefully weigh the costs and benefits of adding sequencing to their testing regimen. If a large number of samples are being sequenced simultaneously, the cost of sequencing an influenza virus genome is much less using NGS instrumentation compared to Sanger-based sequencing. However, the cost of each NGS run is significantly higher than that of a Sanger run; hence, the NGS cost per genome is only cheaper than that for Sanger-based sequencing through the economy of scale. Currently, if a laboratory plans to sequence only a few hundred influenza virus specimens a year, buying an NGS sequencer may be costly and time prohibitive, given the costs of the supporting reagent and the need to develop the required bioinformatics infrastructure. Other alternatives – for example, sequencing service companies or shipping samples to a WHO collaborating centre – may be more financially viable. One option would be for different groups with gene sequencing needs within an institution to liaise with one another, so that economies of scale could be achieved, with various influenza and non-influenza sequences being determined during each NGS run. A guidance figure covering these major factors surrounding NGS adoption is shown in **Annex 1.** Another important consideration is the short shelf life of some of the NGS reagents and whether NICs have easy access to NGS reagent supplies.

Laboratories should also consider their experience with requirements for assay validation and quality control of processes and data (ideally based on sequence analysis and interpretation using data generated through Sanger-based sequencing).

NGS can be performed directly on clinical samples without the need for virus isolation. However, complete genome coverage may not be obtained in samples with low concentrations of virus, usually indicated by high cycle threshold (Ct) values (>30) in the RT-PCR assays used for influenza virus detection.

6.2 NGS technology

Illumina and Ion Torrent are vendors of NGS instruments commonly used for sequencing of influenza viruses. Both companies have instruments available for under US$ 100,000 the data outputs of which allow for significant multiplexing of influenza virus specimens in each run. In general, when determining which instrument to purchase, data output, data quality, run time and running costs are factors to consider. Both platforms have stable sequencing chemistry and sequence the captured DNA through synthesis with roughly equivalent read lengths of up to 600 base pairs (bp).

This document focuses on the implementation of these two technologies as examples of platforms available which have a range of instruments that cost less than US$ 100,000. Laboratories wishing to purchase NGS instruments should also be aware that NGS reagent costs are considerably higher than those for Sanger-based sequencing even though the instrument may be less expensive than a Sanger-

based platform[1] (**Annex 1**; NGS decision tree). It is also important to consider the annual cost for NGS instrument maintenance, which is usually higher than the initial purchase fee over a period of 7–10 years. Table 1 compares the different systems in terms functionality and data output.

Table 1. NGS systems

Platform	Run chemistry	Read length (max.)	Read output
Illumina MiSeq	MiSeq V2 600	600 bp	25 000 000 reads
Illumina MiSeq	MiSeq V2 300	300 bp	25 000 000 reads
Illumina MiniSeq	MiniSeq high output	300 bp	25 000 000 reads
Illumina MiniSeq	MiniSeq mid output	300 bp	8 000 000 reads
Illumina iSeq	I1 chemistry	150 bp	8 000 000 reads
Ion Torrent	PGM 318	400 bp	4 000 000 reads
Ion Torrent	PGM 316	400 bp	2 000 000 reads
Ion Torrent	PGM 314	400 bp	400 000 reads

6.3 NGS library construction

The simplest way to efficiently obtain sequence data of all or part of an influenza virus genome using NGS is through amplicon sequencing. Various methods for enriching or amplifying influenza virus RNA segments or partial segments, and some of the proven strategies are listed in **Annex 2**. These methods make an ideal product on which to base NGS library construction, allowing for efficient sequencing of an influenza genome or partial genome.

Multiple library chemistries are available from multiple vendors for instruments from both Ion Torrent and Illumina. Library chemistry should be evaluated for ease of use, pairing with a specific NGS platform, consistency, cost and ability to multiplex. The NGS library chemistry used should include a way to fragment the amplicons (e.g., shearing, chemical, enzymatic or transposon) and add specific DNA linkers to index the samples.

6.4 Sequence assembly and data analysis

NGS instruments produce a vast amount of data per run compared with a Sanger-based sequencing instrument. Assembly of NGS data requires a robust assembly program in order to successfully generate a reliable consensus sequence of an influenza virus genome. Free assembly programs and pipelines are available that have been shown to work well with the segmented influenza genome **(Annex 3)**.

Pipelines are the best way to assemble and analyze NGS data for influenza, but they require maintenance by knowledgeable computational staff (bioinformaticians), because they require a command line interface. Several commercially available NGS assembly program packages are available (e.g., Geneious, CLC Bio and DNASTAR Lasergene) but the licenses can be costly and other

[1] For details on Sanger-based sequencing protocols for influenza viruses, see WHO information for molecular diagnosis of influenza virus - update. Geneva: World Health Organization; 2018 (https://www.who.int/influenza/gisrs_laboratory/molecular_diagnosis/en/, accessed 8 June 2019).

freeware options may be suitable for those with more experience. These program packages can be useful for laboratories without bioinformatics support who wish to pursue NGS sequencing by providing analysis tools in a user-friendly graphical user interface (GUI). However, there are limitations to these programs because they are designed to aid in the evaluation of only one genome or single segment at a time; hence, large multiplexed runs of influenza viruses can be challenging and time-consuming. Checking the data quality after the pipeline or commercial program is crucial, because artifacts arise that produce errors in the data; staff should be trained to identify the mistakes caused by programs and should regularly run control samples through the pipeline.

6.5 Data storage and computation

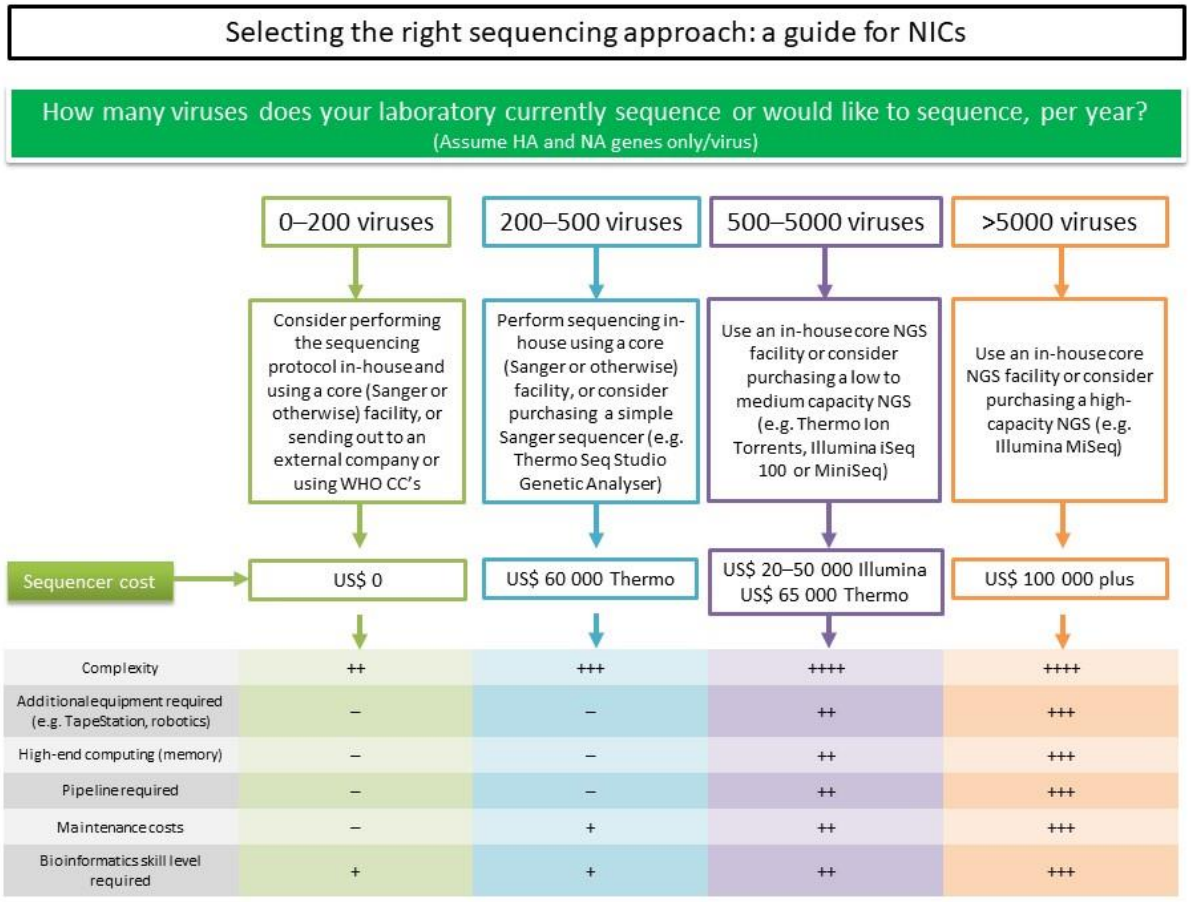
NGS instruments generate large data sets; for example, the largest Illumina instruments can generate hundreds of gigabytes of data in a single run. Hence, before purchasing an NGS instrument, it is important to evaluate the data storage capabilities, network bandwidth and other computational resources of the laboratory and institution. It is also necessary to have a plan that encompasses storage of raw and assembled data from NGS, and the retention of that data in a way that is compliant with the policies of the institution or higher authority, and that takes into account the cost of storage of data. It may be necessary to purchase additional hardware resources to ensure sustainability.

There are several laboratory information management systems (LIMS) that provide NGS sample workflow management. These systems help in organizing the sample identifiers, multiplex barcodes and data that are common features of NGS runs. Although these systems are not required for NGS, they are desirable because they are useful in helping laboratories to track samples through an NGS pipeline.

6.6 Concluding remarks

Overall there are many options availble and will continue to be developed, it is important for those considering installing NGS platforms to consider the number of specimens needed to be sequence to meet your goals, computational requiremnets of the various NGS platformes, and/or training that will be required. It is advisable to consult with laboratories that actively preforming NGS to further understand requirments that may be best suited for your needs.

Annex 1. NGS decision tree



CC: collaborating centre; HA: haemagglutinin; NA: neuraminidase; NGS: next-generation sequencing; NIC: national influenza centre; US: United States; WHO: World Health Organization.

Annex 2. NGS library construction methods

1. Yamauchi Y (ed.). Influenza virus: methods and protocols. New York: Humana Press; 2018.

2. Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ et al. Direct RNA sequencing of the coding complete influenza A virus genome. Sci Rep. 2018;8(1):14408.

3. Zhou B, Deng YM, Barnes JR, Sessions OM, Chou TW, Wilson M et al. Multiplex reverse transcription-PCR for simultaneous surveillance of influenza A and B viruses. J Clin Microbiol. 2017;55(12):3492–3501.

4. Zhou B, Lin X, Wang W, Halpin RA, Bera J, Stockwell TB et al. Universal influenza B virus genomic amplification facilitates sequencing, diagnostics, and reverse genetics. J Clin Microbiol. 2014;52(5):1330–7.

5. Zhou B, Donnelly ME, Scholes DT, St George K, Hatta M, Kawaoka Y et al. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. J Virol. 2009;83(19):10309–13.

6. Zhou B, Wentworth DE. Influenza A virus molecular virology techniques. Methods Mol Biol. 2012;865:175–92. doi: 10.1007/978–1-61779–621–0_11.

Annex 3. Assembly and data analyses methods

1. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP. INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. Genome Med. 2018;10(1):46.

2. 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;17:708.

3. Zhou B, Deng YM, Barnes JR, Sessions OM, Chou TW, Wilson M et al. Multiplex reverse transcription-PCR for simultaneous surveillance of influenza A and B viruses. J Clin Microbiol. 2017;55(12):3492–3501.