We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Application of the New Generation of Sequencing Technologies for Evaluation of Genetic Consistency of Influenza A Vaccine Viruses

Ewan Peter Plant, Tatiana Zagorodnyaya,

Elvira Rodionova, Alin Voskanian-Kordi,

Vahan Simonyan, Zhiping Ye and Majid Laassri

Additional information is available at the end of the chapter

Abstract

For almost half a century, Sanger sequencing has been the conventional method for sequencing DNA. However, its utility for sequencing heterogeneous viral populations is limited because it can only detect mutations that are present in a significant portion of the DNA molecules. Several molecular methods that quantify mutations present at low levels in viral populations were proposed for evaluation of genetic consistency of viral vaccines; however, these methods are only suitable for single site polymorphisms, and cannot be used to screen for unknown mutations.

Next-generation (deep) sequencing methods have enabled the determination of sequences of the entire viral population, including minority components. They enable not only sequencing, but also accurate quantification of mutations. This technique has great value for monitoring the genetic consistency of viral vaccines. Recently, a number of new deep sequencing platforms were introduced (MiSeq, Iron Torrent, etc.) that made such an analysis quite affordable for individual research labs. Here, we review the use of current deep sequencing approaches for influenza virus studies, focusing on the evaluation of the genetic consistency of influenza A vaccine viruses. We also describe a new bioinformatic tool to analyze deep sequencing data and identify artifacts from the true mutants.

Keywords: deep sequencing, DNA and RNA libraries, influenza viruses, mutational profiles, sequence heterogeneity

1. Introductory comments on influenza viruses and vaccines

1.1. Influenza viruses

Influenza A viruses are the causative agents of seasonal epidemics and periodic pandemics. There are many serotypes that infect birds, especially waterfowl, and a few serotypes that infect mammals, including humans. Although some influenza A strains from birds and pigs have jumped the species barrier to infect humans, the majority of human infections are caused by the spread of endemic strains. The endemic strains are continually evolving being one of the reasons that influenza A infections remain a persistent problem. Strains of influenza virus that are used in vaccine production are prone to mutations during the manufacturing process, which can cause breaches in the consistency of vaccine quality. Such mutations can lead to changes in the antigenic structure of the virus and thus affect vaccine effectiveness.

Influenza A viruses belong to the Orthomyxoviridae family of viruses [1]. There are several common features shared among the viruses in this family: they are all negative sense, single-stranded RNA viruses that replicate in the nucleus of the host cell. The influenza A virus genome is comprised of eight RNA segments that encode more than 11 proteins. Two of the segments each encode a major antigenic protein. The fourth largest segment encodes a hemagglutinin (HA) protein and the sixth largest segment encodes a neuraminidase (NA) protein. The different subtypes and strains of influenza A viruses are distinguished by the HA and NA proteins that coat the surface of the virus. There are at least 18 different HA types and 11 NA types [2]. The segmented nature of the viral genome enables two viruses co-infecting the same cell to exchange their segments to produce reassortant progeny. Replication of influenza A viruses also results in mutated viral genomes because of the high error frequency of the RNA polymerase and actions by host defensive elements [3, 4]. Mutations contribute to the emergence of new endemic strains and reassortment may lead to the emergence of epidemic or pandemic strains.

The presence of mutations in influenza A populations has been examined in a variety of contexts. Several groups have isolated clones and used Sanger sequencing to identify mutations. Isolation of a sufficient number of clones has resulted in estimates of the mutation frequencies ranging from 6×10^{-4} to 2×10^{-6} [3, 5, 6]. Although the information about heterogeneity is of great interest, caution must be exercised to ensure that it is accurately reflected in sequence databases. The presence of errors in influenza databases has been noted [7, 8]. To limit discrepancies, some groups have used next-generation sequencing (NGS) to identify sequence heterogeneities [9–12]. Additional technologies such as multisegment reverse-transcription PCR have also been employed [13]. These studies revealed several interesting things. For example, it has been found that differences in viral sequences may occur after a single passage [14], that the same antigenic variants can be detected in different individuals [15], and that oseltamivir resistant and sensitive viruses can be found together as part of heterogeneous viral populations [16].

Most human infections are caused by influenza B viruses and the influenza A serotypes H1N1 and H3N2. In addition to the endemic human influenza transmission, there are cases reported

each year of influenza A infections originating from an animal. Influenza A (H3N2) variant viruses from swine are sometimes transmitted to humans, especially those in close contact with pigs in agricultural settings [17]. Poultry workers are also frequently seropositive for a variety of different avian influenza strains [18–20] suggesting infrequent but detectable transmission. In most cases, there is no person-to-person transmission of the animal viruses. Nevertheless candidate vaccine viruses (CVVs) are prepared each year against some of these viruses to provide a prophylactic option in the event of an outbreak (http://www.who.int/influenza/en/). The CVVs for vaccines against endemic strains and potentially pandemic strains are provided to vaccine manufacturers for use as seed viruses in the manufacturing process.

1.2. Influenza vaccine production

There are several different methods that are used to produce influenza vaccines. Errors may be introduced into the antigenic protein during the replication of the seed virus, no matter which production method is used. The different licensed vaccines are produced from influenza virus grown in eggs or cell culture, or from recombinant viruses expressing the influenza HA from an alternative viral backbone grown in cell culture. Contemporary strains isolated from patients during the current epidemic season normally do not always grow well in cell substrates used for vaccine production. To increase virus yields they are recombined with reference high-growth strains, such that the CVVs have HA and NA-coding RNA segments from the contemporary strain, and RNA segments coding for replicative proteins from the high-growth reference virus. HA is the primary protective antigen and is responsible for binding the cellular receptor. Receptor properties in human and chicken cells differ, forcing the virus HA to adapt to the new receptor, leading to changes in the antigenic specificity potentially affecting vaccine potency.

The most common manufacturing process used in FDA-licensed vaccines is to grow influenza virus in eggs and then inactivate the virus. The inactivated virus is purified and then diluted to the desired potency for filling vials or syringes. Live attenuated influenza vaccines (LAIV) are also grown in eggs but are administered as a nasal spray. Codon deoptimization has also been proposed as a method for creating attenuated viruses [21, 22]. Unlike the current licensed products these may be grown in cell culture.

Some inactivated influenza vaccines are grown in cells. Production of cell-grown vaccines currently uses the same egg isolated seed virus that is used for egg inactivated vaccine production. The cell-grown viruses are harvested, inactivated, and filled into vials or syringes for distribution in a similar manner to the egg grown viruses. The production of recombinant viruses to prepare HA does not require a live seed virus. The HA sequence is cloned into the virus used for production. Although the frequency of errors during replication may differ from that of an influenza virus, the concern still remains. Even if vaccine strains are produced using cloned DNA sequences or synthetic sequences, there is still the possibility of errors arising during amplification of the seed virus. Errors may emerge because of inaccuracies inherent in the replication system or as a response to the host cell defenses.

1.3. Influenza vaccine seed viruses

The seed viruses used to produce influenza vaccines are derived from different sources. Because the influenza virus spreads throughout the world and strains are continually evolving, a network of academic, governmental, and commercial organizations work together to produce new seed viruses. New virus isolates are collected by National Influenza Centers and sent to the World Health Organization Collaborating Centers (WHO CCs). The viruses are typed according to strain and subtype using antigenic and genetic analyses. Viruses are usually isolated in Madin-Darby Canine Kidney Epithelial Cells (MDCK cells) and then amplified in eggs. It is important that the seed virus is as close in sequence and antigenicity to the original isolate as possible. To this end, the egg-amplified virus is used to immunize ferrets for the production of antiserum. The antiserum is then used for antigenic typing of viruses. Such analyses are used to determine how well the strains used in the influenza vaccine are matched to the currently circulating strains.

To produce sufficient quantities of vaccine, the CVVs must have good growth characteristics. Ideally, the viruses should replicate efficiently, have a high antigen (primarily hemagglutinin) to total protein ratio and not have increased pathogenicity. Many viruses do not produce high yields in eggs without adaptation. Some viruses have been propagated for many years and have well-known growth characteristics. They include the influenza A strains Puerto Rico/ 8/1934, cold-adapted A/Ann Arbor/6/1960, A/Leningrad/134/17/1957 and variants of these viruses. Where appropriate these strains have been used as a backbone for influenza A CVVs. Combining the high yield characteristics with the antigenic characteristics of contemporary strains facilitates vaccine production. Reassortant viruses with the desired antigenicity and growth characteristics are produced by two different methods; classical reassortment and genetic reassortment. Dr. Kilbourne of the New York Medical College (NYMC) developed the classical method to create reassortant viruses that expressed the HA and NA from a seasonal strain in the background of a high-growth virus [23]. This method involves co-culture of a contemporary virus and a high yield strain with antibodies to select against the HA and NA of the high yielded strain. Although the resulting viruses have the desired HA and NA genomic segments, the remaining segments may come from either the high-growth strain or the contemporary strain. Another approach based on genetic engineering allows the production of viruses from plasmids expressing the eight influenza virus segments. Genetic engineering also allows the expression of HA and NA proteins in a vector-based system such as a baculo-virus. The desired genetic sequence of HA engineered in this system may be incompatible with baculovirus components, which can result in changes as more efficient mutants displace the parental virus (the so-called gene constellation effect) [24].

All CVVs go through several rounds of replication at manufacturers' facilities as they produce their own virus stocks, working seed, and the final product. Some changes to the virus may occur during manufacture so tests to verify the identity, purity, potency and stability of vaccine lots are required. The potency of inactivated influenza vaccines and influenza vaccine pro-duced from recombinant viruses is determined using standardized reagents supplied by national regulatory authorities. The potency of a live-attenuated virus is calculated from the

amount of viable attenuated virus. Genetic characterization of the vaccine viruses is currently achieved by partial genome sequencing or restriction analysis.

It has been suggested that most of the differences between natural isolates and vaccine seed viruses occur during the selection and clonal isolation of the candidate virus prior to manufacture [25]. The fidelity of replication will vary among viruses and es and will depend on other factors such as the host cell line and multiplicity of infection used. There are some limits on the number of times that seed viruses may be passaged so that mutations are less likely to occur. The European Pharmacopoeia monograph 0158 for inactivated influenza vaccines states that the seed virus should not be passaged more than 15 times. However, because regulations tend to lag behind scientific development, there is no universally accepted guideline for influenza vaccine manufacture that covers egg-derived, cell-derived, and synthetic reassortant viruses.

2. Importance of the evaluation of genetic consistency of influenza A vaccine viruses

Virus populations are comprised of genetically variable viruses and this can affect their replication, evolution, attenuation, and pathogenesis [26, 27]. Having an understanding of the mutations present, even at low levels, in a virus population is important for our understanding of how the viruses grow and cause infections. It has recently been shown that an influenza population containing two variants involved in cell exit grows better than populations containing either variant alone [28]. Although good growth properties are a desirable feature in vaccine seed viruses, it is critical that other parts of the genome, such as those causing attenuation or encoding the major antigenic regions, remain stable. Consistency of manufacture is important and having suitable means to assess genetic consistency is valuable. New assays capable of assessing entire viral genomes, and detecting mutations present at a low level, are needed.

The emergence of mutations in the course of vaccine manufacture was shown to contribute to partial reversion to virulence in the oral polio vaccine (OPV). Mutant analysis by PCR and restriction enzyme cleavage (MAPREC) is used to control batches of oral polio vaccine for the presence of neurovirulent mutations [29, 30] and has been expanded to be used for other viruses [31, 32]. Mutations emerging during virus growth may also change antigenic properties and therefore affect protective potency of live and inactivated vaccines. New approaches that can be used not only for monitoring genetic stability of live vaccines, but also for controlling consistency of inactivated vaccines are needed. Influenza vaccines are manufactured in embryonated chicken eggs or cell culture. There is evidence that vaccine seed viruses adapt to grow efficiently in the different substrates and this can lead to changes in the receptor-recognition site of viral hemagglutinin, which is the major protective antigen [33–36]. For this reason, it is important to monitor the changes that may take place in major protective epitopes of the virus. It is also important to know that mutations responsible for attenuated phenotypes are maintained. Knowing which mutations are emerging during virus growth in production

substrates could also be used to optimize genetic structure of vaccine strains. Consistently accumulating mutations have higher fitness, and if they have no deleterious properties, their incorporation into the genome of vaccine virus could increase its yield and improve vaccine potency. Given these concerns it is imperative to screen genomes of viral vaccines for emerging mutations.

3. Methods used for evaluation of genetic consistency of vaccine viruses

As mentioned above viral populations are highly heterogeneous, and even small quantities of mutants in virus stocks may affect their biological properties. Although PCR and restriction analysis of reassortant influenza viruses can demonstrate which parental strain each genomic segment was derived from, it cannot detect new mutations. Even traditional sequencing methods are not sensitive enough to detect small amounts of mutants, and highly sensitive PCR-based methods can only analyze one or few known mutations at a time.

Conventional sequencing approaches are suitable for discovery of mutations that are present in substantial amounts, usually around 20–25% [37]. Determining the actual frequency using conventional sequencing requires the labor- and time-intensive analysis of a large number of virus clones (plaques).

There are indirect approaches based on analysis of electrophoretic mobility in gels [38], which are insufficiently sensitive and do not allow mutations to be located accurately. Mass spectrometry (MALDI-TOF) [39] and hybridization with microarrays of short oligonucleotides [40–42] are sensitive, but are laborious and may require follow-up by direct sequencing. A highly sensitive mutant analysis by PCR and MAPREC [29–31] can detect and quantify mutants at levels as low as 0.1% of the viral population. Recently we developed a quantitative allele-specific PCR (asqPCR) [43] for detection of a low level of mutants in viral vaccines. RT-PCR has been proposed as a method for checking the homogeneity of influenza vaccine seed candidates [44]. However, these methods are only suitable for analysis of one known mutation at a time.

Several versions of high-throughput sequencing technology, known also as deep or massively parallel sequencing (MPS) have been used to assess influenza vaccine viruses. These technologies enable rapid generation of large amounts of sequence information [45]. Three different platforms for deep sequencing are used widely at present time: the Roche/454 FLX [46] (http://www.454.com/) (http://454.com/products/technology.asp), the Illumina/Solexa Genome Analyzer [47]) (http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html), and the Applied Biosystems SOLiD TM System (http://www.thermofisher.com/us/en/home/brands/applied-biosystems.html). Two new sequencing platforms that are improved to sequence long reads have been developed recently: the Pacific Biosciences SMRT Sequencing (http://www.pacb.com/smrt-science/smrt-sequencing/) [48] and Oxford Nanopore Technologies MinION (https://www.nanoporetech.com/technology/the-minion-device-a-miniaturised-sensing-system/the-minion-device-a-miniaturised-sensing-system).

These systems are also called "single molecule" sequencers and do not require any amplification of DNA fragments prior to sequencing.

The deep sequencing technologies were shown to be suitable for analysis of heterogeneities in viral populations [49]. It can produce huge sequencing information in one run. They are used for *de novo* sequencing of large genomes, metagenomics studies (virome, microbiome, etc.), screening for genomic markers, and many other applications [50–58]. Previously, it was demonstrated that deep sequencing can be used to monitor the genetic stability of oral polio vaccines, and could replace the WHO-recommended MAPREC assay for lot release of OPV [59]. Recently we showed that deep sequencing is suitable for evaluation of genetic consistency of influenza vaccine viruses [36, 60].

4. Description of the most used deep sequencing platforms

Deep or massively parallel sequencing refers to several high-throughput methods for DNA sequencing that are often referred to as NGS. They have dramatically improved the ability of biotechnology, scientific, and healthcare researchers to analyze viruses by allowing users to have massive sequencing information for the entire genomes. The high-throughput sequencing field has witnessed the rise of many technologies capable of massive genomic analysis. In the virology field, deep sequencing has made it simple to sequence full viral genomes. Likewise, identification and classification of novel and known viruses, unbiased characterization of viral populations without the need for virus culturing (viromes), molecular epidemiology, viral diversity and evolution, transmission and pathogenesis, and medical virology have greatly benefited from the use of deep sequencing. The cost of deep sequencing has decreased to an affordable price due to the competition between vendors and the ability to analyze multiple samples run in one lane of the sequencing flow cell. This has allowed virologists to study a huge number of viral samples, including mixture of viral populations, and study low-level mutants in a wide range of viruses [36, 59–61].

There are several platforms for deep sequencing. The most widely used sequencing platforms are the Roche/454 FLX [46] (http://www.454.com/), the Illumina/Solexa Genome Analyzer [47] (http://www.illumina.com/technology/next-generation-sequencing/solexa-technology.html), and the Applied Biosystems SOLiD System (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-4-system.html?CID=FL-091411_solid4).

The differences between these platforms include DNA library preparation procedures and chemistry, the sequencing reactions on the amplified strands, the length of reads, the amount of data generated per run, the hardware, the software engineering and the technology used to amplify single strands of a fragment from the library.

In general the DNA libraries of fragment targets are generated, and adaptors containing universal priming sites are ligated to the fragmented target ends, allowing complex genomes to be amplified with PCR primers. After ligation, the DNA is separated into single strands and

attached or immobilized to a solid surface or support. The immobilization of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously.

Immobilization and separation of the millions of molecules to different surfaces can be achieved by a variety of methods including the Polonator and PicoTiter Plate [47, 62–64]. Attachment of forward and reverse primers to a slide and use of solid-phase amplification also result in the enrichment and amplification of separate template strands [47] (Illumina/Solexa).

Two newer sequencing platforms with longer reads differ from those described above. They are sometimes referred to as "single molecule" sequencers because they sequence molecule by molecule and do not require any amplification of DNA fragments prior to sequencing. The Pacific Biosciences system involves the attachment of a DNA polymerase to the DNA molecule. During the sequencing phase the polymerase adds bases labeled with a fluorophore. The fluorescence unique to each base is recorded and, as each new base is added, the fluorescent label is removed [48]. It generates long sequencing reads (10–15 kb long) from single molecules of DNA, very quickly.

The Oxford Nanopore system runs the sample through very small (1 nm wide) pores. As the DNA passes through these nanopores, the Oxford machine records the electrical charge that is associated with each individual base pair of DNA, like a signature. It produces longer reads (>100 kb long).

A detailed description of the most used two deep sequencing platforms for analysis of influenza viruses and their vaccines is given below.

4.1. Roche/454 FLX pyrosequencer

The Roche/454 FLX sequencing [46] is based on the use of the pyrosequencing technology (http://my454.com/products/technology.asp), in which the incorporation of each nucleotide by DNA polymerase results in the release of pyrophosphate that initiates a cascade of enzymatic reactions that converts the pyrophosphate to a light signal. This light is recorded by CCD camera. This approach, as with most NGS procedures, starts with DNA library preparation; the library DNAs with 454-specific adaptors are denatured into single strands and mixed with agarose beads whose surfaces carry oligonucleotides complementary to the 454-specific adapter sequences on the fragment library, so each bead is associated with a single fragment. The DNA fragments captured by beads are amplified by emulsion PCR (ePCR) [65] to produce approximately one million copies of each DNA fragment on the surface of each bead. These amplified single molecules are then sequenced on a picotiter plate (a fused silica capillary structure) that holds a single bead in each of several hundred thousand single wells, which provides a fixed location at which each sequencing reaction can be monitored.

Individual dNTPs are added to the template in the presence of a DNA polymerase. The sequencing reaction releases pyrophosphate (PPi) after the incorporation of a complementary nucleotide. The released PPi is used by an ATP sulfurylase to release ATP from adenosine 5'-phosphosulfate. The ATP is then used to generate light by converting luciferin into oxyluciferin [66]. Unincorporated dNTPs are degraded by an apyrase, and dATPαS (which is not a

substrate for luciferase) is used instead of dATP. This pyrosequencing reaction is repeated during the sequence of the entire target DNA. This sequencing technology can now produce sequencing reads with up to 1000 bp in length (http://454.com/products/gs-flx-system/).

These raw reads are processed by the 454 analysis software and then filtered to remove poor-quality sequences, mixed sequences, and sequences without the initiating TCGA sequence. Recently, the 454-FLX system was upgraded to reach 99.9% of accuracy after filter and an output of 14 Gb of data per run within 24 h.

4.2. Illumina genome sequencer

The Illumina sequencing [47] method begins with Illumina library preparation flanked with Illumina-specific adapters. Sequencing templates are immobilized on a proprietary flow cell surface that contains immobilized oligos with sequence complementary to those of the adapters, and designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labeled nucleotides. Solid-phase amplification of each single strand DNA from library is performed by bridge amplification, which results in the generation of several million dense clusters of single-stranded DNA in each channel of the flow cell.

The Illumina system sequences DNA in the presence of four reversible terminator-bound dNTPs [47]. At each sequencing step, a fluorescently labeled dNTP is added to the molecule. The fluorescent signal is recorded and then the fluorophore is removed to allow sequencing to continue. The base calls correlate with the signal intensity. Illumina sequencing technology can now produce sequencing reads with up to 600 bp in length (http://www.illumina.com/systems/sequencing.html). The sequencing results are generated in files in which each raw read base has an assigned quality score so that the software can apply a weighting factor in calling differences and generating confidence scores. Illumina data collection software enables users to align sequences to a reference in resequencing applications. This software suite includes the full range of data collection, processing, and analysis modules to streamline collection and analysis of data.

4.3. Deep sequencing data analysis

The massively parallel scale of sequencing implies a similarly massive scale of computational analysis. The conventional pipeline for analysis of next-generation sequencing data includes the following stages: quality control and source data filtering; alignment (mapping); reference profiling (variant-calling, pileup); followed by single-nucleotide polymorphism (SNP) calling (genotyping); and some form of clusterization or classification analysis of samples to discover up or down expression of genes, detect overabundance of SNP positions and correlate those with function and phenotype. Because of the sheer size of the data and amount of calculations needed, such analyses place significant demands on the information technology (IT) infrastructure. Lack of computational power, insufficiency of actively accessible storage facilities in laboratory information management systems (LIMS) and deficiency in network capacity to move data add significantly to the overhead required for high-throughput data production.

The hardware aspect of next-generation sequencing is complicated by the imperfections of current sequence analysis tools, which are suited to shorter sequence read data. There are multiple implementations for all of the stages of the analysis, and some of those are considered to be industry standard tools, running formidable amount of bio-medical analytics. Large-scale analysis of thousands of samples using variety of available tools highlighted important issues with data quality, pre-analytic quality controls, software reproducibility and post-analytic quality controls. Existing data analysis pipelines and algorithms must be modified to accommodate extra-large amounts of short read sequences and combination of shorter and longer read technologies.

To analyze the deep sequencing data for genetic consistency evaluation of influenza vaccine viruses, we have used the corresponding viral reference sequences from NCBI GenBank as a template for alignment of individual sequencing reads. First, sequencing reads with low quality (Phred) score are removed from the data set, and the remaining sequences aligned with reference influenza virus sequence using custom software: The High-performance Integrated Virtual Environment (HIVE, https://hive.biochemistry.gwu.edu/dna.cgi?cmd=main) computer cluster [67, 68].

To create a quantifiable measure comparing the quality of the sequencing and mapping at different positions on a genome, we developed a metrics for assessing positional variant-call quality. To do that, first a histogram is built at every position of a genome where the number
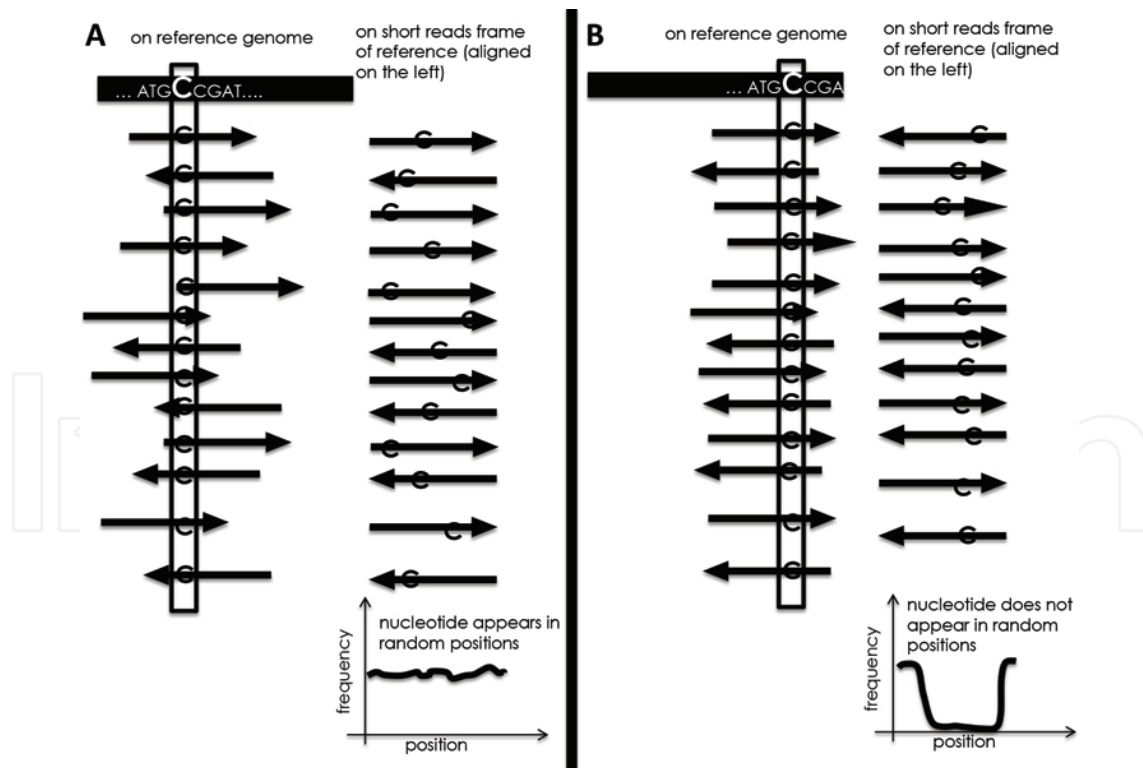

Figure 1. Explanation of entropy computation. For all bases aligned to a particular reference position, the coordinates on a read-frame are accumulated into a positional frequency histogram. For a non-biased position, such distribution is more or less uniform (A). For terminally biased base calls such frequency distribution is skewed towards ends (B) and the Shannon's entropy values drop closer to zero.

of times a base has occurred at a given read-position is accumulated. Additionally, positions of insertion and deletions are also collected in similar histogram. The underlying assumption of the next-generation sequencing method is that the DNA amplification and digestion procedure is random and the short sequences produced by DNA digestion are not strongly biased and not sequence dependent. The default assumption is that a particular variant call should be confirmed by different positions on many reads thus rendering the histogram distribution to be uniform along the entire length of sequence read.

Post-alignment quality control includes identification of mutations distributed non-randomly along individual sequencing reads, which may indicate artifacts in PCR amplification or DNA sequencing procedures (**Figure 1A** and **B**). Biased distribution of mutations along sequencing reads was revealed by calculating Shannon entropy values [69]. Low entropy value suggests that a mutation could be an artifact produced in sequencing procedures. This means that there is an abnormal bias in distribution of this mutation. This entropy-based post-alignment quality control value is calculated on the basis of the equation below. It is based on the normalized first order momentum of logarithmic probability distribution for a particular base (*b*) at a particular position (*r*) of reference genome:

$$Entropy(b,r) = -\frac{\sum_{L} p_i(b) \times log(p_i(b))}{log(L^{-1})}$$

where *L* is the length of the longest read, p_i(b) is the frequency distribution of a base *b* in the reference frame of the reads mapped at the location *r*. The index *i* runs over all of the available positions from 1 to *L*. The denominator makes sure the Shannon's entropy is normalized to a unit value of 1 as the maximum value for entirely uniform distribution. In contrast singular value distribution would have a value for entropy equal to zero. This value is computed for all of the reference positions for every base.

Finally, aligned sequencing reads were used to compute SNP profiles for the entire viral genome.

5. The use of deep sequencing for evaluation of genetic consistency of influenza A vaccine viruses

Influenza A viruses are enveloped, single-stranded RNA viruses belonging to the Orthomyx-oviridae family [70], which also contains four other viral species: influenza B virus, influenza C virus, thogotovirus, and isavirus. The segmented genome of influenza A virus is about 13.6 kb in size and encodes for at least 11 proteins. Its genome is highly variable due to the low fidelity of RNA polymerase and reassortment between co-infecting strains [71]. New virus mutants emerge continuously allowing viruses to survive in presence of the host immunity and cause repeated annual epidemics and occasionally pandemics. Because of this frequent change of the antigens, influenza vaccines must be frequently reformulated to include antigens

of the currently circulating strains. Both live and inactivated influenza vaccines are produced mostly by reassortment with high-growth strains for vaccine production [72, 73]. As stated above, adaptation to growth in different cells can lead to changes in viral receptor-binding region, and also in protective epitopes. Therefore, it is desirable to monitor genetic stability of viruses used in vaccine manufacture to ensure that their antigenic structure remains unchanged.

Deep sequencing technology has opened up the possibility for the characterization of viral genomes directly from samples [74, 75]. The viral metagenome or "virome" refers to the collection of viruses found in a particular sample from humans, animals, plants or from a specific environmental sample. Virome studies can lead to the discovery of new viruses and/or to their association with known or novel diseases. Numerous viruses have been identified as part of the virome study, including influenza A viruses [2].

The deep sequencing technologies are a great tool to investigate genetically complex populations of influenza viruses and to detect minority mutant variants with clinical or epidemiological relevance. Deep sequencing-based methods have recently been applied for the assessment of influenza A viruses diversity and their dynamics of evolution [60, 76, 77]. Others have focused on the evolution of avian influenza strains with potential to become pandemic in humans [78–82], as well as the detection of virulence signatures [83] and reassortment patterns [84]. Other studies have investigated the transmission and adaptation of avian influenza viruses to humans, as part of preparedness for a potential influenza pandemic [12, 53, 85–98].

As it is crucial to study transmission and adaptation of avian influenza viruses, and swine strains for epidemics and pandemics in humans, many studies based on the use of deep sequencing techniques have described avian and swine influenza virus evolution [99–103]. Other studies have investigated the predominance and spread of different human influenza viruses in specific geographic areas [16, 104–106].

Study of drug escape variants is an important aspect of epidemiological and clinical virology. Sanger sequencing can only detect mutations present in around 20% of the viral population [37, 107–109], which excludes it for quantitation of low-level viral mutant variants. Using deep sequencing to detect low portion of mutant drug resistant variants at levels as low as 0.1% of the virus population has been demonstrated [110–116]. Other studies have focused on the use of deep sequencing for surveillance of drug resistance-associated mutations for both NA inhibitors and adamantanes [117–127]. Deep sequencing has also been used for the detection and subtyping of human influenza A viruses and reassortants [61, 84].

Recently, deep sequencing-based methods have been proposed for the assessment of influenza A viruses antigenic stability [128] using complete influenza A genomes and exploiting the ability to detect and quantify mutations in heterogeneous viral populations. Deep sequencing was used to study the evolution of influenza A viruses in the vaccinated pigs. The genetic diversity and evolution of the virus at an intra-host level was analyzed directly from nasal swabs collected during infection [129]. The obtained results demonstrated remarkable diversity of influenza A viruses, and rapid change of these viruses during infection of vacci-

nated pigs. These types of complex studies can be done only by high throughput sequence analysis.

To evaluate the genetic stability in influenza vaccine viruses, we have used a deep sequencing approach that was recently qualified for quantitation of all mutants in the entire genome including those that are present at low level in viral populations [59]. Recently, we explored the utility of deep sequencing methods for monitoring the consistency of influenza A vaccines [36, 60]. Also in the same study, we proposed new protocols for simultaneous amplification of all segments of influenza A genomes and new bioinformatic tools to analyze the data and to identify artifacts generated during PCR amplification and deep sequencing procedures. Amplification of the entire genome of influenza viruses presents a challenge because of the difference in size and sequence composition of the eight genomic segments.
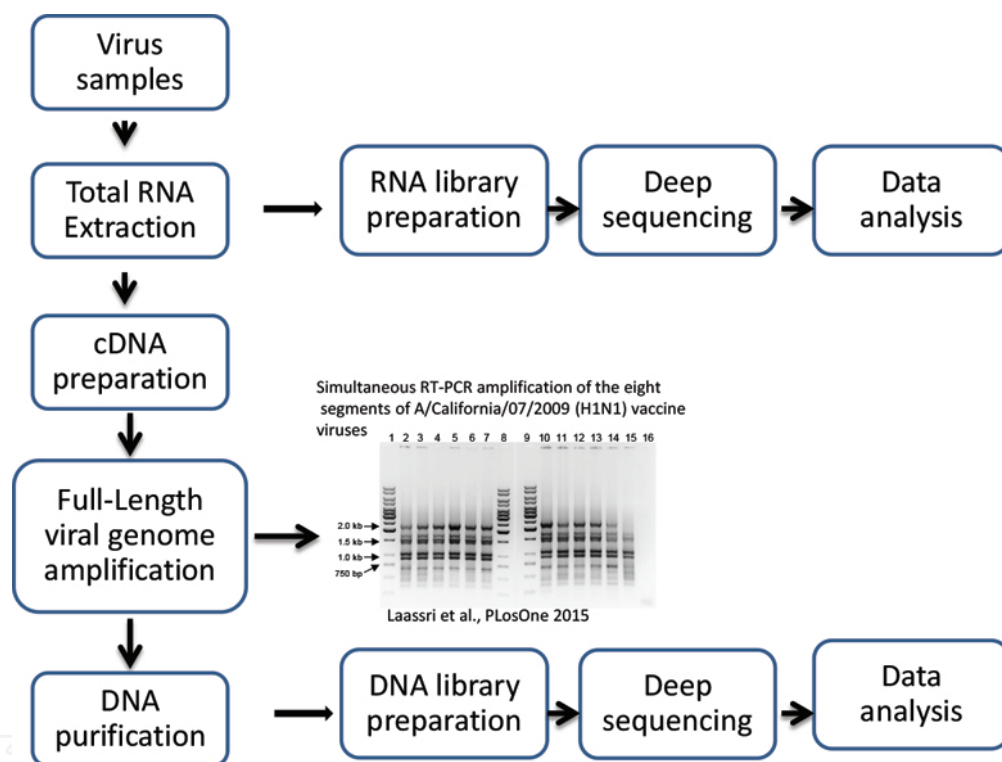

Figure 2. General steps followed for RNA/DNA libraries sequencing of influenza A/California/07/2009 (H1N1) vaccine viruses.

We described PCR conditions that allow to amplify all genomic segments of influenza A virus in one reaction [60] that was optimized subsequently during an analysis of the A/California/07/2009 (H1N1) vaccine viruses (derived from X-179A, X-181, 121XP viruses) [36]. We have used both the total RNA (without specific amplification of viral cDNA) and DNA amplicon, for RNA and DNA libraries preparation, respectively, and both protocols were compared for consistency in mutant variants quantitation.

The protocols for deep sequencing of viral DNA libraries and whole-RNA libraries used to determine quantitative profiles of mutations along the entire genome of viruses of influenza

A/California/07/2009 (H1N1) vaccine viruses were described [36]. The steps followed to perform the deep sequencing are presented in **Figure 2** and can be summarized as follows: The PCR product was purified by QIAquick PCR Purification Kit (Qiagen) and fragmented by an ultrasonicator (Covaris) to generate the optimal fragment sizes needed for Illumina sequencing, then the fragmented DNAs were used for library preparation with NEBNext® DNA Sample Prep Reagent Set 1 (New England BioLabs).

For preparation of Illumina sequencing libraries from total RNA, the NEBNext mRNA Sample Prep Master Mix Set 1 (New England BioLabs) was used. Briefly, total RNA (extracted as mentioned above) was fragmented as described above to generate the optimal fragment sizes. Double-stranded cDNA was prepared and ligated to the Illumina paired end adaptors. Finally, the libraries were amplified using 15 cycles of PCR with multiplex indexed primers and purified with magnetic beads using Agencourt Ampure Beads (Beckman Coulter). Deep sequencing was performed at Macrogen (Seoul, Korea) using HiSeq2000 (Illumina) or at our laboratory using MiSeq (Illumina).
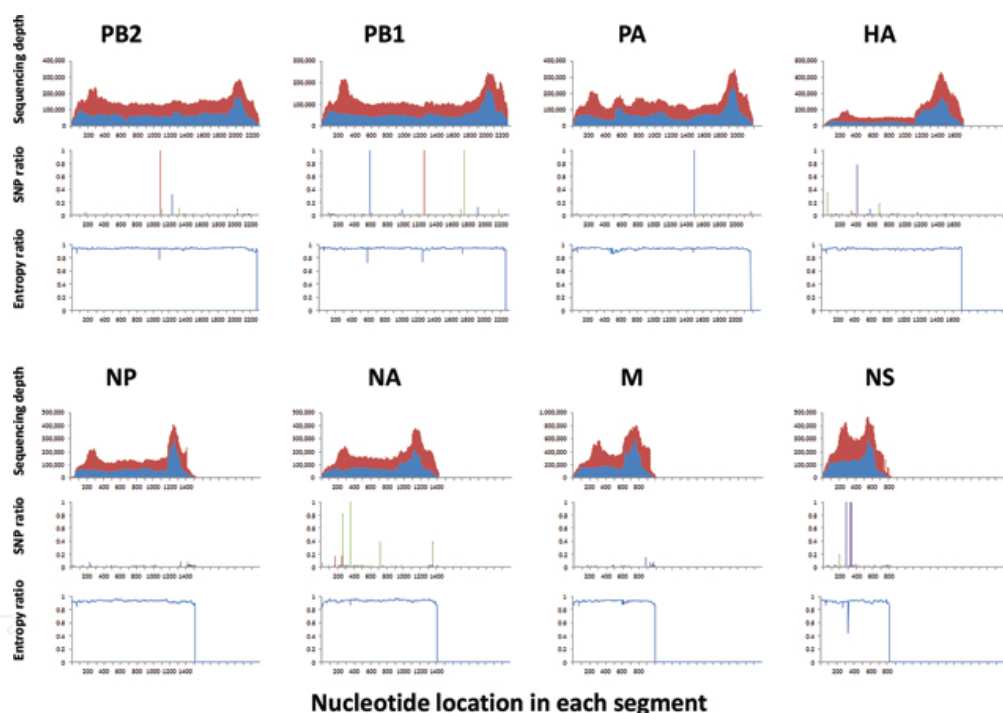

Figure 3. Sequencing analysis of the entire genome of the influenza A/California/07/2009 (H1N1) vaccine virus reassortant tenth passage in eggs sequenced by HiSeq (Macrogen) using 101 bp paired-end Illumina/Solexa sequencing technology. Computations were made by in-house custom software HIVE-align. For each segment, the charts of the depth of sequencing coverage distribution (the number of times every nucleotide was sequenced on ordinate plotted against the position on genome—abscissa), SNP profile (the ratio of mutants on ordinate plotted against the position on genome—abscissa), and entropy were built.

The sequencing data analysis was done using custom software, a highly integrated virtual environment (HIVE) computer cluster (https://hive.biochemistry.gwu.edu/dna.cgi?cmd=main) as described above. The RNA sequences of X-179A, X-181, 121XP, A/California/07/2009 (H1N1), and A/PR/08/34 viruses deposited in NCBI GenBank were used as references

for alignment of the viral sequence reads. We analyzed the depth of sequencing, the single-nucleotide polymorphism profile, and entropy (that allow us to distinct between bias and true mutation) for each segment of influenza virus (see **Figure 3**, for example), the data analysis resulted also on generation of consensus sequences for each segment.

The deep sequencing results revealed several heterogeneities in most genomic segments, and several mutations led to amino acid changes. Deep sequencing of whole-RNA libraries was found to be more reproducible than sequencing of DNA libraries. This may be due to errors introduced during PCR amplification by DNA polymerase and non-specific alignment of primers [36].

The deep sequencing of the X-179A passaged viruses [36] identified several mutations in HA and NA genes. In HA two non-synonymous mutations; $Pro_{314}Gln$ (in 17% of the virus population) and $Asn_{146}Asp$ (in 78% of the virus population) were identified. The $Asn_{146}Asp$ mutation is found in the antigenic site Sa; it was detected at 11% in the A/California/07/2009 (H1N1) strain and as the dominant residue in the X-181 virus [36]. One X-179A stock contained the $Lys_{328}Thr$ mutation at a low level (9%). Viruses derived from X-179A were heterogeneous and contained some complete nucleotide substitutions in comparison to their published sequences in PB2, PB1, NP, and in NS segments [36]. The X-181 virus was developed from the X-179A seed lot by another round of reassortment, and also is subjected to several passages in eggs. Deep sequencing results [36] showed that the $G_{756}T$ ($Glu_{252}Asp$, present at 47%) mutation emerged in HA of the passaged X181 virus, and it is located in the conserved region of the antigenic site Ca [36].
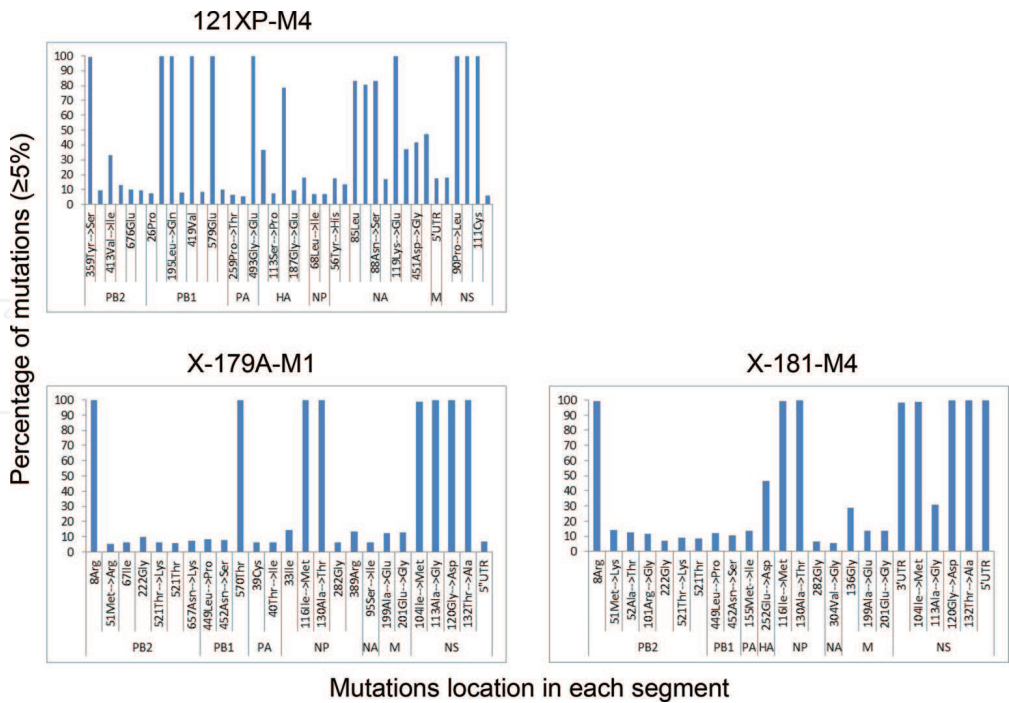

Figure 4. Percent of mutations (≥5%) emerged in X-179A-M1, X-181-M4, and 121XP-M4 viruses (derived from X-179A, X-181, and 121XP viruses, respectively) passaged 10 times in embryonated chicken eggs.

Unlike the X-179A and X-181 viruses, 121XP was developed by reverse genetics [130]. The deep sequencing of 121XP virus passaged 10 times in eggs (121XP-M4 virus) showed that this virus is more heterogenic than X-179A and X-181 viruses passaged 10 times in eggs (X-179A-M1 and X-181-M4 viruses respectively; **Figure 4**) [36]. In the passaged 121XP virus, the mutation $Lys_{226}Glu$ was emerged at low level (18%) in Ca antigenic site of HA, which is very close to the region that participates in the modulation of HA receptor specificity and that enables H3 influenza viruses to switch specificity from avian to human [131–133]; another mutation $Lys_{136}Asn$ was emerged at a high level (78%) close to the HA antigenic site Sa within the sialic acid-binding pocket [134]. Recently a similar deep sequencing approach was used to study the genetic and potential antigenic diversity of influenza viruses infecting humans, some of whom became infected despite recent vaccination [15].

We found that the deep sequencing approach based on RNA library preparation was effective and reproducible for detection of low quantities of mutants in the entire genome of influenza A vaccine viruses [36]. The deep sequencing approach revealed that the viruses derived from three pandemic A/Ca/07/2009 (H1N1) vaccine viruses have varying levels of sequence heterogeneities some of them in antigenic sites, which may affect their efficacy.

6. Conclusions

In the last few years, the use of deep sequencing has expanded largely to tackle problems in many fields of virology. The greatest benefit of deep sequencing is its ability to detect minor mutant variants, as low as 0.1% of virus population [36, 59, 60, 110, 111, 113, 114]. The deep sequencing approach based on RNA library preparation is effective and reproducible for detection of low quantities of mutants in the entire genome of influenza A vaccine viruses [36], and eliminates the need for full-length amplification. The deep sequencing platforms are improving continuously to combine low error rates with long reads and relatively low cost. It played a key role in the discovery of many new viruses, the characterization of virus populations in humans and the potential of their association with the pathogenesis of several diseases. As described here, there is no doubt that the deep sequencing is facilitating and accelerating the evaluation of the genetic consistency of vaccine viruses. It is an important tool for monitoring vaccine consistency during manufacture and after vaccination. Deep sequencing-based assays are already being implemented for the genetic consistency evaluation of oral polio vaccine and influenza A vaccine viruses [36, 59, 60]. The ability to quantify potentially undesirable mutations in vaccine batches makes this method suitable for quality control to ensure manufacture of safe and effective vaccines.

Acknowledgements

Author details

Ewan Peter Plant, Tatiana Zagorodnyaya, Elvira Rodionova, Alin Voskanian-Kordi, Vahan Simonyan, Zhiping Ye and Majid Laassri[*]

*Address all correspondence to: majid.laassri@fda.hhs.gov

Center for Biologics Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

References

[1] Lamb, R.A. and R.M. Krug, Orthomyxoviridae: the viruses and their replication, in Fields virology, D.M.K. Bernard, N. Fields, P.M. Howley, Editor. 1996, Philadelphia: Lippincott-Raven Publishers, pp. 1353–96.

[2] Tong, S., et al., New world bats harbor diverse influenza A viruses. PLoS Pathog, 2013. 9(10): p. e1003657.

[3] Cheung, P.P., et al., Comparative mutational analyses of influenza A viruses. RNA, 2015. 21(1): pp. 36–47.

[4] Gutierrez, R.A., et al., Biased mutational pattern and quasispecies hypothesis in H5N1 virus. Infect Genet Evol, 2013. 15: pp. 69–76.

[5] Nobusawa, E. and K. Sato, Comparison of the mutation rates of human influenza A and B viruses. J Virol, 2006. 80(7): pp. 3675–8.

[6] Parvin, J.D., et al., Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. J Virol, 1986. 59(2): pp. 377–83.

[7] Krasnitz, M., A.J. Levine, and R. Rabadan, Anomalies in the influenza virus genome database: new biology or laboratory errors? J Virol, 2008. 82(17): pp. 8947–50.

[8] Suarez, D.L., N. Chester, and J. Hatfield, Sequencing artifacts in the type A influenza databases and attempts to correct them. Influenza Other Respir Viruses, 2014. 8(4): pp. 499–505.

[9] Roedig, J.V., et al., Impact of host cell line adaptation on quasispecies composition and glycosylation of influenza A virus hemagglutinin. PLoS One, 2011. 6(12): p. e27989.

[10] Van den Hoecke, S., et al., Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing. BMC Genomics, 2015. 16: p. 79.

[11] Wang, J., et al., MinION nanopore sequencing of an influenza genome. Front Microbiol, 2015. 6: p. 766.

[12] Watson, S.J., et al., Viral population analysis and minority-variant detection using short read next-generation sequencing. Philos Trans R Soc Lond B Biol Sci, 2013. 368(1614): p. 20120205.

[13] Zou, X.H., et al., Evaluation of a single-reaction method for whole genome sequencing of influenza A virus using next generation sequencing. Biomed Environ Sci, 2016. 29(1): pp. 41–6.

[14] Lee, H.K., et al., Comparison of mutation patterns in full-genome A/H3N2 influenza sequences obtained directly from clinical samples and the same samples after a single MDCK passage. PLoS One, 2013. 8(11): p. e79252.

[15] Dinis, J.M., et al., Deep sequencing reveals potential antigenic variants at low frequency in influenza A-infected humans. J Virol, 2016. 90(7): 3355–3365.

[16] Fordyce, S.L., et al., Genetic diversity among pandemic 2009 influenza viruses isolated from a transmission chain. Virol J, 2013. 10: p. 116.

[17] Nelson, M.I., et al., Evolutionary dynamics of influenza A viruses in US Exhibition Swine. J Infect Dis, 2016. 213(2): p. 173–82.

[18] Di Trani, L., et al., Serosurvey against H5 and H7 avian influenza viruses in Italian poultry workers. Avian Dis, 2012. 56(4 Suppl): pp. 1068–71.

[19] Heidari, A., et al., Serological evidence of H9N2 avian influenza virus exposure among poultry workers from Fars province of Iran. Virol J, 2016. 13(1): p. 16.

[20] Huang, S.Y., et al., Serological comparison of antibodies to avian influenza viruses, subtypes H5N2, H6N1, H7N3 and H7N9 between poultry workers and non-poultry workers in Taiwan in 2012. Epidemiol Infect, 2015. 143(14): pp. 2965–74.

[21] Broadbent, A.J., et al., Evaluation of the attenuation, immunogenicity, and efficacy of a live virus vaccine generated by codon-pair bias de-optimization of the 2009 pandemic H1N1 influenza virus, in ferrets. Vaccine, 2016. 34(4): pp. 563–70.

[22] Fan, R.L., et al., Generation of live attenuated influenza virus by using codon usage bias. J Virol, 2015. 89(21): pp. 10762–73.

[23] Kilbourne, E.D., Future influenza vaccines and the use of genetic recombinants. Bull World Health Organ, 1969. 41(3): pp. 643–5.

[24] Plant E.P. and Z. Ye., Gene constellation of influenza vaccine seed viruses in Current issues in molecular virology – viral genetics and biotechnological applications, V. Romanowski, Editor. 2013, InTech, pp 213–237.

[25] Buonagurio, D.A., et al., Genetic stability of live, cold-adapted influenza virus components of the FluMist/CAIV-T vaccine throughout the manufacturing process. Vaccine, 2006. 24(12): pp. 2151–60.

[26] Domingo, E., et al., The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance---a review. Gene, 1985. 40(1): pp. 1–8.

[27] Hansen, H., et al., Recombinant viruses obtained from co-infection in vitro with a live vaccinia-vectored influenza vaccine and a naturally occurring cowpox virus display different plaque phenotypes and loss of the transgene. Vaccine, 2004. 23(4): pp. 499–506.

[28] Xue, K.S., et al., Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. eLife 2016;5: p. e13974.

[29] Chumakov, K.M., Molecular consistency monitoring of oral poliovirus vaccine and other live viral vaccines. Dev Biol Stand, 1999. 100: pp. 67–74.

[30] Chumakov, K.M., et al., Correlation between amount of virus with altered nucleotide sequence and the monkey test for acceptability of oral poliovirus vaccine. Proc Natl Acad Sci U S A, 1991. 88(1): pp. 199–203.

[31] Bidzhieva, B., M. Laassri, and K. Chumakov, MAPREC assay for quantitation of mutants in a recombinant flavivirus vaccine strain using near-infrared fluorescent dyes. J Virol Methods, 2011. 175(1): pp. 14–9.

[32] Laassri, M., et al., Microarray hybridization for assessment of the genetic stability of chimeric West Nile/dengue 4 virus. J Med Virol, 2011. 83(5): pp. 910–20.

[33] Gambaryan, A.S., et al., Effects of host-dependent glycosylation of hemagglutinin on receptor-binding properties on H1N1 human influenza A virus grown in MDCK cells and in embryonated eggs. Virology, 1998. 247(2): pp. 170–7.

[34] Hughes, M.T., et al., Adaptation of influenza A viruses to cells expressing low levels of sialic acid leads to loss of neuraminidase activity. J Virol, 2001. 75(8): pp. 3766–70.

[35] Schild, G.C., et al., Evidence for host-cell selection of influenza virus antigenic variants. Nature, 1983. 303(5919): pp. 706–9.

[36] Laassri, M., et al., Deep sequencing for evaluation of genetic stability of influenza A/California/07/2009 (H1N1) vaccine viruses. PLoS One, 2015. 10(9): p. e0138650.

[37] Larder, B.A., et al., Quantitative detection of HIV-1 drug resistance mutations by automated DNA sequencing. Nature, 1993. 365(6447): pp. 671–3.

[38] Orita, M., et al., Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci U S A, 1989. 86(8): pp. 2766–70.

[39] Amexis, G., et al., Quantitative mutant analysis of viral quasispecies by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Natl Acad Sci U S A, 2001. 98(21): pp. 12097–102.

[40] Cherkasova, E., et al., Microarray analysis of evolution of RNA viruses: evidence of circulation of virulent highly divergent vaccine-derived polioviruses. Proc Natl Acad Sci U S A, 2003. 100(16): pp. 9398–403.

[41] Laassri, M., et al., Microarray techniques for evaluation of genetic stability of live viral vaccines, in Viral genomes – Molecular structure, diversity, gene expression mechanisms and host-virus interactions. Maria Laura Garcia Editor. 2012, InTech, pp. 181–94.

[42] Laassri, M., et al., Genomic analysis of vaccine-derived poliovirus strains in stool specimens by combination of full-length PCR and oligonucleotide microarray hybridization. J Clin Microbiol, 2005. 43(6): pp. 2886–94.

[43] Bidzhieva, B., M. Laassri, and K. Chumakov, Allele-specific PCR for quantitative analysis of mutants in live viral vaccines. J Virol Methods, 2014. 201: pp. 86–92.

[44] Shcherbik, S., et al., Application of real time RT-PCR for the genetic homogeneity and stability tests of the seed candidates for live attenuated influenza vaccine production. J Virol Methods, 2014. 195: pp. 18–25.

[45] Rogers, Y.H. and J.C. Venter, Genomics: massively parallel sequencing. Nature, 2005. 437(7057): pp. 326–7.

[46] Margulies, M., et al., Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005. 437(7057): pp. 376–80.

[47] Bentley, D.R., Whole-genome re-sequencing. Curr Opin Genet Dev, 2006. 16(6): pp. 545–52.

[48] Eid, J., et al., Real-time DNA sequencing from single polymerase molecules. Science, 2009. 323(5910): pp. 133–8.

[49] Victoria, J.G., et al., Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol, 2009. 83(9): pp. 4642–51.

[50] Bainbridge, M.N., et al., Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics, 2006. 7: p. 246.

[51] Cheval, J., et al., Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J Clin Microbiol, 2011. 49(9): pp. 3268–75.

[52] Greninger, A.L., et al., A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One, 2010. 5(10): p. e13381.

[53] Kuroda, M., et al., Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PLoS One, 2010. 5(4): p. e10256.

[54] Nakamura, S., et al., Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One, 2009. 4(1): p. e4219.

[55] Pettersson, E., et al., Allelotyping by massively parallel pyrosequencing of SNP-carrying trinucleotide threads. Hum Mutat, 2008. 29(2): pp. 323–9.

[56] Satkoski, J.A., et al., Pyrosequencing as a method for SNP identification in the rhesus macaque (Macaca mulatta). BMC Genomics, 2008. 9: p. 256.

[57] Torres, T.T., et al., Gene expression profiling by massively parallel sequencing. Genome Res, 2008. 18(1): pp. 172–7.

[58] Wheeler, D.A., et al., The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008. 452(7189): pp. 872–6.

[59] Neverov, A. and K. Chumakov, Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. Proc Natl Acad Sci U S A, 2010. 107(46): pp. 20063–8.

[60] Bidzhieva, B., et al., Deep sequencing approach for genetic stability evaluation of influenza A viruses. J Virol Methods, 2014. 199: pp. 68–75.

[61] Seong, M.W., et al., Genotyping influenza virus by next-generation deep sequencing in clinical specimens. Ann Lab Med, 2016. 36(3): pp. 255–8.

[62] Kim, J.B., et al., Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science, 2007. 316(5830): pp. 1481–4.

[63] Leamon, J.H., et al., A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis, 2003. 24(21): pp. 3769–77.

[64] Shendure, J., et al., Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 2005. 309(5741): pp. 1728–32.

[65] Berka, J., et al., Bead emulsion nucleic acid amplification. 2005, Google Patents.

[66] Froehlich, T., D. Heindl, and A. Roesler, Miniaturized, high-throughput nucleic acid analysis. 2010, Google Patents.

[67] Simonyan, V. and R. Mazumder, High-performance integrated virtual environment (HIVE) tools and applications for big data analysis. Genes (Basel), 2014. 5(4): pp. 957–81.

[68] Wilson, C.A. and V. Simonyan, FDA's activities supporting regulatory application of "next gen" sequencing technologies. PDA J Pharm Sci Technol, 2014. 68(6): pp. 626–30.

[69] Shannon, C.E., A mathematical theory of communication. Bell Syst Technical J, 1948. 27(3): pp. 379–423.

[70] Lamb, R.A. and M.A. Krug, Orthomyxoviridae: the viruses and their replication, in Fields Virology, Howley P. M., Knipe D. M., Editor, 4th edition. 2007, Philadelphia: Lippincott Williams & Wilkins, pp. 1487–31.

[71] Steinhauer, D.A., E. Domingo, and J.J. Holland, Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene, 1992. 122(2): pp. 281–8.

[72] Girard, M.P., et al., A review of vaccine research and development: human enteric infections. Vaccine, 2006. 24(15): pp. 2732–50.

[73] McCarthy, M.W. and D.R. Kockler, Trivalent intranasal influenza vaccine, live. Ann Pharmacother, 2004. 38(12): pp. 2086–93.

[74] Julian, T.R. and K.J. Schwab, Challenges in environmental detection of human viral pathogens. Curr Opin Virol, 2012. 2(1): pp. 78–83.

[75] Mokili, J.L., F. Rohwer, and B.E. Dutilh, Metagenomics and future perspectives in virus discovery. Curr Opin Virol, 2012. 2(1): pp. 63–77.

[76] Bhatt, S., E.C. Holmes, and O.G. Pybus, The genomic rate of molecular adaptation of the human influenza A virus. Mol Biol Evol, 2011. 28(9): pp. 2443–51.

[77] Tsai, K.N. and G.W. Chen, Influenza genome diversity and evolution. Microbes Infect, 2011. 13(5): pp. 479–88.

[78] Bourret, V., et al., Whole-genome, deep pyrosequencing analysis of a duck influenza A virus evolution in swine cells. Infect Genet Evol, 2013. 18: pp. 31–41.

[79] Crusat, M., et al., Changes in the hemagglutinin of H5N1 viruses during human infection--influence on receptor binding. Virology, 2013. 447(1–2): pp. 326–37.

[80] Hoper, D., et al., Highly pathogenic avian influenza virus subtype H5N1 escaping neutralization: more than HA variation. J Virol, 2012. 86(3): pp. 1394–404.

[81] Mertens, E., et al., Evaluation of phenotypic markers in full genome sequences of avian influenza isolates from California. Comp Immunol Microbiol Infect Dis, 2013. 36(5): pp. 521–36.

[82] Wilker, P.R., et al., Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun, 2013. 4: p. 2636.

[83] Waybright, N., et al., Detection of human virulence signatures in H5N1. J Virol Methods, 2008. 154(1–2): pp. 200–5.

[84] Deng, Y.M., N. Caldwell, and I.G. Barr, Rapid detection and subtyping of human influenza A viruses and reassortants by pyrosequencing. PLoS One, 2011. 6(8): p. e23400.

[85] Archer, J., et al., Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics, 2012. 13: p. 47.

[86] Bartolini, B., et al., Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. New Microbiol, 2011. 34(4): pp. 391–7.

[87] Deyde, V.M. and L.V. Gubareva, Influenza genome analysis using pyrosequencing method: current applications for a moving target. Expert Rev Mol Diagn, 2009. 9(5): pp. 493–509.

[88] Flaherty, P., et al., Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res, 2012. 40(1): p. e2.

[89] Hoper, D., B. Hoffmann, and M. Beer, Simple, sensitive, and swift sequencing of complete H5N1 avian influenza virus genomes. J Clin Microbiol, 2009. 47(3): pp. 674–9.

[90] Hoper, D., B. Hoffmann, and M. Beer, A comprehensive deep sequencing strategy for full-length genomes of influenza A. PLoS One, 2011. 6(4): p. e19075.

[91] Kampmann, M.L., et al., A simple method for the parallel deep sequencing of full influenza A genomes. J Virol Methods, 2011. 178(1–2): pp. 243–8.

[92] Levine, M., et al., Detection of hemagglutinin variants of the pandemic influenza A (H1N1) 2009 virus by pyrosequencing. J Clin Microbiol, 2011. 49(4): pp. 1307–12.

[93] Ren, X., et al., Full genome of influenza A (H7N9) virus derived by direct sequencing without culture. Emerg Infect Dis, 2013. 19(11): pp. 1881–4.

[94] Rutvisuttinunt, W., et al., Simultaneous and complete genome sequencing of influenza A and B with high coverage by Illumina MiSeq Platform. J Virol Methods, 2013. 193(2): pp. 394–404.

[95] Taubenberger, J.K., The virulence of the 1918 pandemic influenza virus: unraveling the enigma. Arch Virol Suppl, 2005(19): pp. 101–15.

[96] Taubenberger, J.K., et al., Characterization of the 1918 influenza virus polymerase genes. Nature, 2005. 437(7060): pp. 889–93.

[97] Tumpey, T.M., et al., Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science, 2005. 310(5745): pp. 77–80.

[98] Xiao, Y.L., et al., High-throughput RNA sequencing of a formalin-fixed, paraffin-embedded autopsy lung tissue sample from the 1918 influenza pandemic. J Pathol, 2013. 229(4): pp. 535–45.

[99] Clavijo, A., et al., Identification and analysis of the first 2009 pandemic H1N1 influenza virus from U.S. feral swine. Zoonoses Public Health, 2013. 60(5): pp. 327–35.

[100] Croville, G., et al., Field monitoring of avian influenza viruses: whole-genome sequencing and tracking of neuraminidase evolution using 454 pyrosequencing. J Clin Microbiol, 2012. 50(9): pp. 2881–7.

[101] Marchenko, V.Y., et al., Ecology of influenza virus in wild bird populations in Central Asia. Avian Dis, 2012. 56(1): pp. 234–7.

[102] Van Borm, S., et al., Phylogeographic analysis of avian influenza viruses isolated from Charadriiformes in Belgium confirms intercontinental reassortment in gulls. Arch Virol, 2012. 157(8): pp. 1509–22.

[103] Yu, X., et al., Influenza H7N9 and H9N2 viruses: coexistence in poultry linked to human H7N9 infection and genome characteristics. J Virol, 2014. 88(6): pp. 3423–31.

[104] Barrero, P.R., et al., Genetic and phylogenetic analyses of influenza A H1N1pdm virus in Buenos Aires, Argentina. J Virol, 2011. 85(2): pp. 1058–66.

[105] de la Rosa-Zamboni, D., et al., Molecular characterization of the predominant influenza A(H1N1)pdm09 virus in Mexico, December 2011-February 2012. PLoS One, 2012. 7(11): p. e50116.

[106] Lin, J., et al., Influenza seasonality and predominant subtypes of influenza virus in Guangdong, China, 2004-2012. J Thorac Dis, 2013. 5(Suppl 2): pp. S109–17.

[107] Church, J.D., et al., Sensitivity of the ViroSeq HIV-1 genotyping system for detection of the K103N resistance mutation in HIV-1 subtypes A, C, and D. J Mol Diagn, 2006. 8(4): pp. 430–2; quiz 527.

[108] Halvas, E.K., et al., Blinded, multicenter comparison of methods to detect a drug-resistant mutant of human immunodeficiency virus type 1 at low frequency. J Clin Microbiol, 2006. 44(7): pp. 2612–4.

[109] Leitner, T., et al., Analysis of heterogeneous viral populations by direct DNA sequencing. Biotechniques, 1993. 15(1): pp. 120–7.

[110] Archer, J., et al., Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. PLoS One, 2012. 7(11): p. e49602.

[111] Dudley, D.M., et al., Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. PLoS One, 2012. 7(5): p. e36494.

[112] Fisher, R., et al., Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. J Virol, 2012. 86(11): pp. 6231–7.

[113] Gibson, R.M., et al., Sensitive deep-sequencing-based HIV-1 genotyping assay to simultaneously determine susceptibility to protease, reverse transcriptase, integrase, and maturation inhibitors, as well as HIV-1 coreceptor tropism. Antimicrob Agents Chemother, 2014. 58(4): pp. 2167–85.

[114] Shao, W., et al., Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. Retrovirology, 2013. 10: p. 18.

[115] Simen, B.B., et al., Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J Infect Dis, 2009. 199(5): pp. 693–701.

[116] Wang, C., et al., Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res, 2007. 17(8): pp. 1195–201.

[117] Arvia, R., et al., Monitoring the susceptibility to oseltamivir of Influenza A(H1N1) 2009 virus by nested-PCR and pyrosequencing during the pandemic and in the season 2010-2011. J Virol Methods, 2012. 184(1–2): pp. 113–6.

[118] Chen, L.F., et al., Cluster of oseltamivir-resistant 2009 pandemic influenza A (H1N1) virus infections on a hospital ward among immunocompromised patients--North Carolina, 2009. J Infect Dis, 2011. 203(6): pp. 838–46.

[119] Correia, V., et al., Antiviral drug profile of seasonal influenza viruses circulating in Portugal from 2004/2005 to 2008/2009 winter seasons. Antiviral Res, 2010. 86(2): pp. 128–36.

[120] Dharan, N.J., et al., Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. JAMA, 2009. 301(10): pp. 1034–41.

[121] Duwe, S.C., et al., Genotypic and phenotypic resistance of pandemic A/H1N1 influenza viruses circulating in Germany. Antiviral Res, 2011. 89(1): pp. 115–8.

[122] Pontoriero, A., et al., Virological surveillance and antiviral resistance of human influenza virus in Argentina, 2005-2008. Rev Panam Salud Publica, 2011. 30(6): pp. 634–40.

[123] Tellez-Sosa, J., et al., Using high-throughput sequencing to leverage surveillance of genetic diversity and oseltamivir resistance: a pilot study during the 2009 influenza A(H1N1) pandemic. PLoS One, 2013. 8(7): p. e67010.

[124] Bright, R.A., et al., Incidence of adamantane resistance among influenza A (H3N2) viruses isolated worldwide from 1994 to 2005: a cause for concern. Lancet, 2005. 366(9492): pp. 1175–81.

[125] Bright, R.A., et al., Adamantane resistance among influenza A viruses isolated early during the 2005-2006 influenza season in the United States. JAMA, 2006. 295(8): pp. 891–4.

[126] Deyde, V.M., et al., Surveillance of resistance to adamantanes among influenza A(H3N2) and A(H1N1) viruses isolated worldwide. J Infect Dis, 2007. 196(2): pp. 249–57.

[127] Higgins, R.R., et al., Differential patterns of amantadine-resistance in influenza A (H3N2) and (H1N1) isolates in Toronto, Canada. J Clin Virol, 2009. 44(1): pp. 91–3.

[128] Warren, S., et al., Extreme evolutionary conservation of functionally important regions in H1N1 influenza proteome. PLoS One, 2013. 8(11): p. e81027.

[129] Diaz, A., et al., Genome plasticity of triple-reassortant H1N1 influenza A virus during infection of vaccinated pigs. J Gen Virol, 2015. 96(10): pp. 2982–93.

[130] Robertson, J.S., et al., The development of vaccine viruses against pandemic A(H1N1) influenza. Vaccine, 2011. 29(9): pp. 1836–43.

[131] Connor, R.J., et al., Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. Virology, 1994. 205(1): pp. 17–23.

[132] Matrosovich, M., et al., Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. J Virol, 2000. 74(18): pp. 8502–12.

[133] Stevens, J., et al., Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science, 2006. 312(5772): pp. 404–10.

[134] Varghese, J.N., et al., The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. Proteins, 1992. 14(3): pp. 327–32.