HAYATI Journal of Biosciences xxx (2017) 1–5

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

HAYATI Journal of Biosciences

journal homepage: http://www.journals.elsevier.com/hayati-journal-of-biosciences

Review paper

Next-Generation Sequencing and Influenza Virus: A Short Review of the Published Implementation Attempts

Rasha Ali,[1] Ruth M. Blackburn,[2] Zisis Kozlakidis[1,2*]

[1] Division of Infection and Immunity, University College London, London, United Kingdom.
[2] Farr Institute of Health Informatics Research, University College London, London, United Kingdom.

ARTICLE INFO

Article history:
Received 23 May 2016
Accepted 9 December 2016
Available online xxx

ABSTRACT

Influenza virus represents a major public health concern worldwide after recent pandemics. To aid the understanding and characterization of the virus in ever-increasing sample numbers, new research techniques have been used, such as next-generation sequencing (NGS). The current article review used Ovid MEDLINE and PubMed databases to conduct keyword searches and investigate the extent to which published NGS high-throughput approaches have been implemented to influenza virus research in the last 5 years, during which the increase in research funding for influenza studies has been coincidental with a significant per-base cost reduction of sequencing. Through the current literature review, it is evident that over the last 5 years, NGS techniques have been indeed applied to biological and clinical samples at increasing rates following a wide variety of approaches. The rate of adoption is slower than anticipated by most published studies, with three obstacles identified consistently by authors. These are the lack of suitable downstream analytical capacity, the absence of established quality control comparators, and the higher cost to comparable existing techniques.

1. Introduction

Influenza viruses are well-characterized members of the Orthomyxoviridae family. Genomic subpopulation diversity and new viral mutants emerge constantly because of the continued viral genetic variation and antigenic modification in response to many factors such as host immunity, ecological and environmental factors, resulting in occasional pandemics and annual epidemics (Zhirnov *et al.* 2009). Influenza remains a major threat on the global agricultural and health care systems because of its continued potential to cause pandemics worldwide and because of the increasing number of seasonal infections impacting human and economic health (Fischer *et al.* 2015). The high number of infections and the recurrent seasonality mean that influenza is suitable for a number of high-throughput molecular approaches in addition to the basic virological techniques and clinical expertise to strengthen global pandemic preparedness. In addition, the total and proportional funding for influenza research (£39,139,703, 4.3% of total infection research) increased in 2011–13 compared with 1997–2010 (£126,643,152, 3.4% of all infection research), hence the field is more likely able to afford the use of new and perhaps more expensive technologies than studies of other infectious diseases (Heada *et al.* 2015). Coincidentally, the per-base cost of sequencing in the same period has reduced by 92% from 0.52 to 0.04 USD per DNA Mb (National Human Genome Research Institute, January 2010–January 2015). Hence, according to our working hypothesis, we expected to notice a steady increase in published implementation examples as overall implementation costs were reducing. In this brief report, we review the application of high-throughput next-generation sequencing (NGS) in the study of influenza and present the opportunities and challenges of implementation as reported by the research community.

Currently, there are two major technologies used for influenza genomic sequencing; the NGS and traditional Sanger sequencing (Deng *et al.* 2015). The Sanger sequencing technology referred to as first generation has been used for almost four decades and continues to be the standard reference method used. However, there is a gradual yet notable shift away from this technique and in favor of the use of newer technologies, namely the high-throughput NGS (International Human Genome Consortium 2004). NGS also referred to as deep sequencing or parallel sequencing (massively parallel sequencing) provides high-speed multiplexing capabilities

* Corresponding author.
E-mail address: z.kozlakidis@ucl.ac.uk (Z. Kozlakidis).
Peer review under responsibility of Institut Pertanian Bogor.

http://dx.doi.org/10.1016/j.hjb.2016.12.007

Please cite this article in press as: Ali, R., et al., Next-Generation Sequencing and Influenza Virus: A Short Review of the Published Implementation Attempts, HAYATI J Biosci (2017), http://dx.doi.org/10.1016/j.hjb.2016.12.007

for high-throughput sample sequencing and enormous data volumes of sequencing reads in one run (Barzon *et al.* 2011). Along with the decreasing NGS costs, the applications of NGS techniques within routine diagnostic settings are still evolving because of recent and iterative developments in genome sequencing and bioinformatics analyses (Fischer *et al.* 2015).

2. A Number of choices and challenges for NGS platforms

The common process of most NGS technologies is the initial random fragmentation of templates, followed by an amplification process using polymerase chain reaction target-specific primers, resulting in many DNA copies that can be independently sequenced (Metzker 2010). High-throughput sequencing platforms can be divided into two broad groups depending on the template used. The earliest platforms depend on the production of libraries of clonally amplified templates. The recent arrival of single-molecule sequencing platforms determines the sequence of single molecules without amplification. Within these broad categories, there is considerable variation in performance—including in throughput, read length, and error rate—as well as in factors affecting usability, such as cost and run time (Loman *et al.* 2012).

NGS technologies have a unique potential for the *de novo* sequencing of large genomes, genomic markers screening, transcriptome analysis, and several other applications (Bainbridge *et al.* 2006; Cheval *et al.* 2011; Greninger *et al.* 2010; Kuroda *et al.* 2010; Nakamura *et al.* 2009; Pettersson *et al.* 2008; Satkoski *et al.* 2008; Torres *et al.* 2008; Wheeler *et al.* 2008). However, the complexity and large size of the sequencing data constitute one of the main bioinformatics challenges of NGS data interpretation (Nowrousian 2010). The primary approach to NGS data analysis can be accomplished by using either one of three main types of tools, such as general-purpose aligners, *de novo* assemblers, and short-read aligners (Lin *et al.* 2014). NGS methods confer advantages over other techniques such as highly specific reverse transcription-polymerase chain reaction or less-sensitive traditional virological methods for being able to produce unbiased sequencing without prior knowledge of the presence or type of viral agents. This in turn can potentially constitute them into the future gold standard tool for viral genome discovery, especially in the case of recombinogenic viruses, such as influenza (Bialasiewicz *et al.* 2014).

Through the current literature review, it is evident that over the last 5 years, NGS techniques have been indeed applied to clinical samples at increasing rates with some studies concentrating on the detection of novel pathogens or pathogens at low detection levels. Several variant strains and viruses have been successfully identified, such as the PIV4 subtype in late 2013(Alquezar-Planas *et al.* 2013), although it has to be noted that the numbers of unsuccessful attempts are generally not mentioned, unclear, and/or very difficult to even hazard a guess at. Other studies followed the seasonal influenza infections in large population cohorts (Nakamura *et al.* 2009), whereas influenza studies on animals have also used NGS capabilities, such as sequences generated from lung tissues of ferrets experimentally infected with influenza A/California/07/2009 (H1N1) (Lin *et al.* 2014). However, the overall numbers of samples used per study vary widely, and the full implementation of a high-throughput analytical pipeline remains difficult to achieve. The implementation challenges, solutions, and expectations of the authors are also summarized.

3. Methods

Our research based on the Ovid MEDLINE database and the NCBI PubMed databases was conducted with a total of 18 different keywords in different combinations each time (initial concept terms used: Influenza, next generation sequencing, and data not shown). The literature search provided a wide variety of peer-reviewed publications ranging in number from (10–18013). The relevant article abstracts were manually selected corresponding to publications where NGS was actually implemented as opposed to being alluded to for future implementation. Then the exact sequencing techniques used were determined, e.g. Illumina™ MiSeq/HiSeq NGS, Roche™ GS-FLX+ 454-pyrosequencing, and others. Only two inclusion criteria were preselected, that is English language and publication years from 2008 to 2015 inclusive.

4. Results

4.1. Influenza high-throughput DNA sequencing studies

Our research detected 64 research publications within the publication years of 2008–2015. According to their methods, Q4 almost all the studies used one or more of the following NGS platforms (Roche-454 GS Junior/FLX+, Ion Torrent/Proton/Personal Genome Machine sequencing, and Illumina GAIIx/MiSeq/HISeq) accompanied with a diverse and fragmented set of methods for the upstream sample preparation and downstream bioinformatics analyses.

Of the 64 research publications, 35 studies were performed exclusively on human material (Fischer *et al.* 2015; Deng *et al.* 2015; Kuroda *et al.* 2010; Cheval *et al.* 2011; Buggele *et al.* 2013; Depew *et al.* 2013; Baum *et al.* 2010; Rutvisuttinunt *et al.* 2015; Frey *et al.* 2014; Farsani *et al.* 2015; Zhao *et al.* 2015; Rutvisuttinunt *et al.* 2013; Lee *et al.* 2013; Flaherty *et al.* 2012; Téllez-Sosa *et al.* 2013; Borozan *et al.* 2013; Archer *et al.* 2012; Bidzhieva *et al.* 2014; Van den Hoecke *et al.* 2015; Leung *et al.* 2013; Watson *et al.* 2013; Harismendy *et al.* 2009; Zhou *et al.* 2014; Kuroda *et al.* 2015; Burnham *et al.* 2015; Varble *et al.* 2014; Tan *et al.* 2014; Saira *et al.* 2013; Selleri, 2013; Swaminathan *et al.* 2013; Xiao *et al.* 2013; Power *et al.* 2012; Whitehead *et al.* 2012; Yasugi *et al.* 2012), 10 on animal material (Lin *et al.* 2014; Jakhesara *et al.* 2014; Van Borm *et al.* 2012; Dugan *et al.* 2011; Clavijo *et al.* 2013; León *et al.* 2013; Lange *et al.* 2013; Iqbal *et al.* 2014; Peng *et al.* 2011; Wang *et al.* 2012), seven on both animal and human materials (Yu *et al.* 2014; Jonges *et al.* 2014; Kampmann *et al.* 2011; Peng *et al.* 2014; Karlsson *et al.* 2013; Sikora *et al.* 2014; Ren *et al.* 2013), two on plasmid-derived material (Depew *et al.* 2013; Wu *et al.* 2014), and 10 reviewed technical and bioinformatics aspects (Barzon *et al.* 2011; Metzker 2010; Quiñones-Mateu *et al.* 2014; Park *et al.* 2013; Dugan *et al.* 2012; MacLean *et al.* 2009; Radford *et al.* 2012; Ansorge 2009; Shendure and Ji 2008; Tsai and Chen 2011). The number of samples used per study varied widely, with most studies reporting numbers in the low hundreds and less than 10 reporting the use of more than 1000 samples.

4.2. Challenges, opportunities, and solutions of NGS implementation

From the aforementioned, it becomes immediately obvious that the initial NGS applications in the field of influenza research are not reflective of a consistent, universally applied, and true high-throughput approach. Indeed, the picture obtained throughout is one reflecting the initial stages for the adoption of a technical innovation. The challenges mentioned by the various authors are summarized in the Table. The generation of high volumes of data requiring sophisticated downstream bioinformatics analyses is mentioned as the primary challenge for the adoption of the method and interpretation of the NGS outputs. In fact, this single challenge is mentioned in more than two-thirds of all the identified studies. The lack of large-scale validation of NGS outputs with regard to costs and data complexity is challenging and perhaps not feasible for individual research groups to achieve, hence its function as an adoptive impediment. The availability of NGS equipment is a

Please cite this article in press as: Ali, R., et al., Next-Generation Sequencing and Influenza Virus: A Short Review of the Published Implementation Attempts, HAYATI J Biosci (2017), http://dx.doi.org/10.1016/j.hjb.2016.12.007