CrossMark

Next-generation sequencing library preparation method for identification of RNA viruses on the Ion Torrent Sequencing Platform

Guiqian Chen[1,2] · Yuan Qiu[1] · Qingye Zhuang[1] · Suchun Wang[1] · Tong Wang[1] · Jiming Chen[1] · Kaicheng Wang[1]

Abstract

Next generation sequencing (NGS) is a powerful tool for the characterization, discovery, and molecular identification of RNA viruses. There were multiple NGS library preparation methods published for strand-specific RNA-seq, but some methods are not suitable for identifying and characterizing RNA viruses. In this study, we report a NGS library preparation method to identify RNA viruses using the Ion Torrent PGM platform. The NGS sequencing adapters were directly inserted into the sequencing library through reverse transcription and polymerase chain reaction, without fragmentation and ligation of nucleic acids. The results show that this method is simple to perform, able to identify multiple species of RNA viruses in clinical samples.

Keywords Next-generation sequencing · Library preparation · RNA virus

Introduction

RNA viruses are the agents of many human, animal, and plant infectious diseases, including influenza, severe acute respiratory syndrome (SARS), and so on [1–3]. Identification and analysis of RNA viruses are important for the diagnosis, treatment, control, and prevention of human and animal infectious diseases [4]. Since the development of next generation sequencing (NGS) technologies, great progress has been made in the rapid identification and characterization of RNA viruses [5–8]. Numerous viruses and variant strains have been identified using NGS approaches. Unlike insensitive traditional virological methods and highly specific reverse transcription-polymerase chain reaction (RT-PCR), NGS methods have the advantage of being able to sequence total or targeted DNA and RNA from samples in an unbiased way, without a priori knowledge of the possible viral agent(s) present, thus making them the ideal tool for novel and divergent viral genome discovery. This facilitates research in virus ecology, novel virus discovery, and the development of larger datasets of complete virus genomes for studies on virus evolution and pandemic prediction.

Four popular second-generation sequencing platforms have been released: Illumina HiSeq, MiSeq and NovaSeq, Ion Torrent PGM, Proton and S5, BGISeq-500, have been commercially available [9]. Among these platforms, Ion Torrent PGM is competitive for detection of viruses and bacteria with respect to instrumental price, sequencing cost, and simplicity of operation, although its sequencing throughput is lower than MiSeq and Proton [10]. Each NGS platform has its own sequencing library preparation procedure. A suitable pipeline of library construction is very essential for virus genome sequencing by NGS. In order to establish the NGS platform in diagnosis and surveillance of viral infection, we developed a NGS library preparation method based on RT-PCR random primers. The effectiveness and practicality to identify viruses and sequence their genomes using this method are discussed in this study.

✉ Kaicheng Wang
    wangkaicheng@cahec.cn

1   China Animal Health and Epidemiology Center, Qingdao, Shandong, China

2   College of Medicine, Jiaxing University, Jiaxing, Zhejiang, China

Materials and methods

Ethics statement

This study was conducted according to the animal welfare guidelines of the World Organization for Animal Health [11], and approved by the Animal Welfare Committee of China Animal Health and Epidemiology Center. The fecal and swab samples were all collected with permission given by the multiple relevant parties, including the Ministry of Agricultural of China, China Animal Health and Epidemiology Center, the relevant veterinary sections in the provincial and county government. Fecal samples were collected from fresh feces in the ground of poultry farms in China. Swab samples were collected by gently taking smears from the trachea and cloacae of domestic fowl in China and then placed in a transport medium.

Sample collection

A swab sample was collected from a duck in a live bird market from Guizhou province, China, in October of 2013. The swab sample was collected through taking smears at both cloacal and oropharyngeal tracts, and stored in 1.5 ml phosphate buffered saline (PBS, pH 7.2) containing 10% glycerol [10, 12]. The sample was negative for Avian influenza virus (AIV) detection, but caused death to specific-pathogen-free (SPF) embryonated chicken eggs in 72 h. The swab sample was clarified by centrifugation at 10,000×g for 5 min, and the supernatants were inoculated in 10-day-old SPF embryonated chicken eggs via the allantoic sac route. The SPF embryonated eggs were purchased from Shandong Healthtec Laboratory Animal Breeding Company (Jinan, China). The inoculated eggs were further incubated for 3 days, and checked twice each day during the incubation period. Dead ones were picked out and stored in a refrigerator. After the incubation period, allantoic fluid was collected to evaluate the unknown virus identification ability of the cDNA library preparation method and the suitable length of reverse transcription time for the first stand cDNA synthesis in the library construction process. Another unknown virus sample was taken from the mixed feces of 52 dead ducks in a poultry farm of Shandong province, China, in June 2014. The fecal sample was collected from approximately 0.5-ml wet and fresh feces, and stored in 3.5 ml PBS (pH 7.2) containing 10% glycerol [10, 12]. The samples were stored at 4 °C and tested in 3 days after collection. The samples were stored at −80 °C after detection.

RNA preparation

Both samples were centrifuged at 12,000×g, 4 °C for 30 min. The supernatant was filtered through a 0.22-μM filter (Millipore, USA) to remove eukaryotic and bacterial particles as much as possible. The 0.22-μM filter (Millipore, USA) could not remove the microorganism of size smaller than 0.22 uM. The filtered solution was precipitated using 1/10 volume of 50% (w/v) polyethylene glycol 6000 (PEG-6000) at 4 °C for 2 h. Then, the solution was centrifuged at 12,000×g for 1 h at 4 °C. Precipitation was suspended into PBS solution. To remove the naked DNA and RNA, the solution was incubated with DNase (Ambion, USA) and RNase (Promega, USA) at 37 °C for 30 min. Viral RNA was extracted with a QiaAmp Viral RNA Kit (Qiagen, Germany). The RNA concentration of the two samples was 187.5 and 27.1 ng/μl determined by a Qubit® 2.0 Fluorometer (Qubit® RNA Assay Kit, Life Technologies), respectively.

NGS library preparation

The method of NGS library preparation is shown in Fig. 1. Briefly, one adaptor was added during the generation of the first strand cDNA by RT-PCR. During the synthesis of the second strand cDNA, the other adaptor was introduced. Primers based on the two adaptors were used to generate the expected cDNA library. The application of random primers in sequencing viral genomes has been reported previously, but reverse transcription time for the first strand cDNA synthesis is variable. To meet the requirements of NGS on a PGM platform, it is better to produce a cDNA library with DNA fragment sizes between 200 and 500 bp. To decide a suitable reverse transcription time for first strand cDNA synthesis in the preparation of NGS library samples, the size distribution and concentration of the first strand cDNA synthesis produced with different reverse transcription times were analyzed by an Agilent 2100 Bioanalyzer, in the NGS library preparation of the first sample. First strand cDNA synthesis produced from reverse transcription times of 10, 20, 25, 30, 40, and 60 min of the first sample were selected for the analysis.

Details of the NGS library preparation method are as follows: 2 μl viral nucleic acids, 1 μl 100 μM primer A15N6 (5′-**GTGTCTCCGACTCAG**NNNNNN-3′), 1 μl dNTP (10 mM), and 6 μl RNase free water were mixed and incubated at 65 °C for 5 min. Then the mixture was placed on ice for at least 1 min. To the RNA/primer mixture was added 10 μl cDNA synthesis mix including 2 μl 10× RT buffer, 4 μl MgCl$_2$ (25 mM), 2 μl DTT (0.1 M), 1 μl RNaseOUT (40 U/μl), and 1 μl SuperScript® III Reverse Transcriptase (200 U/μl, Invitrogen, USA). The first strand cDNA synthesis reaction was performed as 25 °C for 15 min, and 42 °C for 30 min (or 10, 20, 25, 40, and 60 min). The reaction was terminated at 75 °C for 5 min. Then 1 μl RNase H (TaKaRa, Japan) was added to the reaction and incubated at 37 °C for 30 min. After purification using DynaMag™-2 Magnet and Agencourt® AMPure® XP Reagent (Beckman Coulter, USA), the B15N6 primer
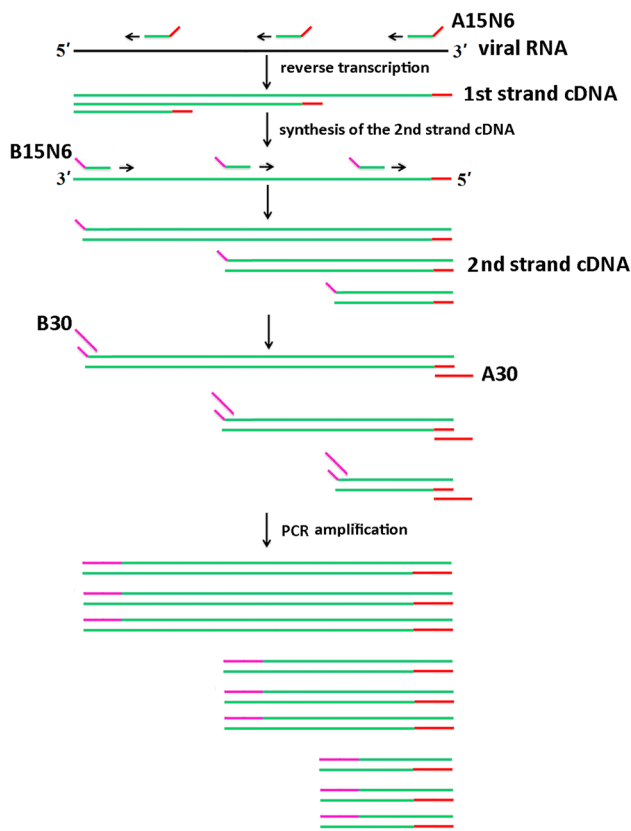
Fig. 1 The method of cDNA library preparation. One adaptor was added during the generation of the first strand cDNA by RT-PCR. During the synthesis of the second strand cDNA, the other adaptor was introduced. Primers based on the two adaptors were used to generate the expected cDNA library

(5′-**TGGGCAGTCGGTGAT**NNNNNNN-3′) was aligned to the purified first strand cDNA and elongated at 37 °C for 1 h with 5 U Klenow fragment (3′→5′ exo-, NEB,USA) and then at 75 °C for 10 min to terminate the reaction. PCR amplification was performed with 5 μl double-stranded DNA template in a final reaction volume of 50 μl, containing 1× Phusion HF buffer, 1 μM primer A30 (5′-CCATCT CATCCCTGC**GTGTCTCCGACTCAG**-3′), B30 (5′-CCG CTTTCCTCTCTA**TGGGCAGTCGGTGAT**-3′), and 0.5 U Phusion High-Fidelity DNA Polymerase (NEB, USA). The library was amplified using the following conditions: 98 °C for 30 s, followed by 14 cycles of 98 °C for 10 s, 55 °C for 30 s, and 72 °C for 1 min, with a final extension at 72 °C for 10 min. DNA fragments between 200 bp and 500 bp were extracted with a Min-Elute gel extraction kit (Qiagen, Germany) to use as the library constructed by the NGS library preparation method. To avoid the contamination of the NGS library, all the materials for NGS library preparation were new, and the operation was performed in the cleaning air-condition laboratory.

Ion Torrent PGM sequencing and analysis

The libraries were sequenced on the Ion Torrent PGM platform with an Ion PGM™ Sequencing 200 Kit. The Ion Torrent PGM singleton reads were compared to the GenBank nucleotide database using the standalone BLAST version 2.2.30 [13]. An E-value of 10^{-5} was used as the cutoff value for significant hits. Reads were further sorted by MetaGenome Analyzer version 5.10.5 (MEGAN,vesion 5.10.5) with default LCA parameters [14] to identify viruses, according to the first hit in the BLAST analysis results. To avoid the false-positive results, all the reads hits of viruses excluding phages were verified manually through online BLAST at NCBI web station. Sorted reads classified into virus categories from uncultured duck fecal sample collected from Shandong were extracted and assembled by De Novo Assembly in the CLC genomics workbench 8.5.1 (Qiagen, Germany). Genome sequencing coverage of the viruses which were hitted with most number of reads was calculated by CLC genomics workbench 8.5.1.

Results

NGS library construction

In the analysis of suitable reverse transcription time for the RNA extracted from the cultured duck cloacal/oropharyngeal tracts swab sample of Guizhou province, the results showed that reverse transcription times of 20, 25, and 30 min can generate considerably higher concentrations of cDNA fragments between 250 and 500 bp than 10, 40, and 60 min (Fig. 2). Compared to other incubation times, the expected fragment size (250–500 bp) cDNA exhibited the highest percentage (90.77%) of the total cDNA produced by 30-min reverse transcription (Table 1). The concentration of cDNA fragments of the expected size was 20.30 ng/μL determined by a Qubit® 2.0 Fluorometer (Qubit® dsDNA HS Assay Kit, Life Technologies).

Unknown virus identification

The sequence data of the two samples are in the short read archive at GenBank with accession numbers SRR2142090 and SRR5943895, respectively. For the cultured duck cloacal/oropharyngeal tracts swab sample collected from Guizhou, a total of 4,548,888 reads were produced by Ion Torrent PGM NGS. The average read length was 152 bp, and GC content is 54.7%. From these, 2,257,158 (49.62%) reads belong to host cellular organisms, 1472 (0.03%) reads belong to viruses, and 2,134,992 (46.93%) reads belonged to "not assigned" group, which matched the sequence without taxon ids in the GenBank nucleotide database. There were 155,266

Fig. 2 Size distributions when different reverse transcription times were used. The RNA extracted from the cultured duck cloacal/oropharyngeal tracts swab sample of Guizhou province was used for the analysis of suitable reverse transcription time
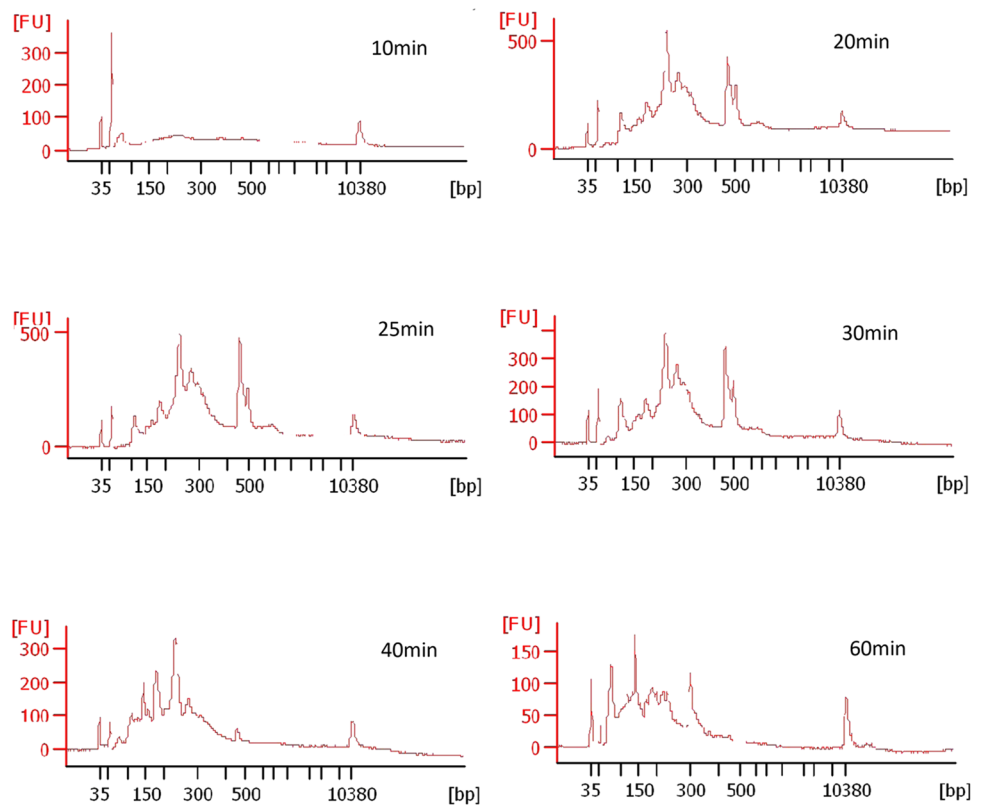

Table 1 The analysis of size distribution and concentration (200–500 bp)

Reverse time (min)	Average size (bp)	Size distribution in CV (%)	Concentration area (pg/µL)	Percentage of the expected fragment size cDNA in the total (%)
10	334	29.02	862.57	61.80
20	356	25.75	3380.31	89.30
25	368	21.87	3596.06	90.17
30	353	22.94	3123.58	90.77
40	342	30.71	2420.33	90.08
60	315	24.95	1427.87	89.51

(3.41%) reads in the "no hits" group, which did not match any sequence in GenBank nucleotide database. Among the virus reads, 622 belong to Caudovirales (42.26%) and 82 (5.57%) belong to Paramyxoviruses. In the uncultured duck fecal sample collected from Shandong, a total of 2,072,054 reads were produced by Ion Torrent PGM NGS. The average read length was 183 bp, and GC content is 45.93%. From these, 758,547 (36.61%) reads belong to host cellular organisms; 70,430 (3.40%) reads belonged to the "not assigned" group, and 1,220,605 (58.91%) reads belonged to the "no hits" group. Because the sample had not been cultured, most reads were non-hit vial genome sequences. There were 22,472 reads (1.08%) in the "viruses" group, including 18 families (Table 2) and 4190 Phages reads. Most (84.75%) of the reads belonged to Coronaviridae. The main pathogen infecting the ducks was coronavirus.

De novo assembly

From the uncultured duck fecal sample collected from Shandong, 15,494 read sequences showing significant but divergent BLAST hits to Coronaviridae were extracted for assembly analysis. 10,888 reads were mapping to the avian infectious bronchitis (IBV) virus (IBV) genome (Accession NC_001451), covering 71.46% of the reference genome sequence with 29 gaps containing 4423 bases. The mean length of the mapped read is 183 bp, and the total read length is 1,995,756. The average coverage is 61.95 (Min = 0, Max = 2731).

Table 2 The number of hits for each virus species in the uncultured duck fecal sample collected from Shandong

Species	Families	Number of hits
dsRNA viruses		
Human picobirnavirus	Picobirnaviridae	96
Rotavirus A	Reoviridae	2
Rotavirus C	Reoviridae	4
Aquareovirus C	Reoviridae	4
Avian orthoreovirus	Reoviridae	105
Retro-transcribing viruses		
Duck hepatitis B virus	Hepadnaviridae	1
Avian leukosis virus	Retroviridae	17
Rous sarcoma virus	Retroviridae	31
Avian retrovirus	Retroviridae	1
Avian sarcoma virus	Retroviridae	1
Avian endogenous retrovirus EAV-HP	Retroviridae	5
Columba palumbus retrovirus	Retroviridae	1
ssRNA positive-strand viruses		
Avastrovirus 1	Astroviridae	7
Avastrovirus 2	Astroviridae	64
Turkey avastrovirus 3	Astroviridae	1
Mamastrovirus 1	Astroviridae	10
Betacoronavirus 1	Coronaviridae	26
Severe acute respiratory syndrome-related corona-virus	Coronaviridae	4
Avian coronavirus	Coronaviridae	15,464
Circket paralysis virus	Dicistroviridae	3
Drosophila C virus	Dicistroviridae	7
Rhopalosiphum padi virus	Dicistroviridae	4
Foot-and-mouth disease virus	Picornaviridae	1
Encephalomyocarditis virus	Picornaviridae	1
Human enterovirus	Picornaviridae	10
Hepatitis A virus	Picornaviridae	1
Avian encephalomyelitis virus	Picornaviridae	141
Soybean mosaic virus	Potyviridae	11
Watermelon mosaic virus	Potyviridae	27
Zucchini yellow mosaic virus	Potyviridae	179
Sindbis virus	Togaviridae	266
Shallot latnet virus	Betaflexiviridae	5
Cucumber green mottle mosaic virus	Virgaviridae	42
Pepper mild mottle virus	Virgaviridae	84
Tobacco mild green mosaic virus	Virgaviridae	7
Tobacco mosaic virus	Virgaviridae	72
Tomato mosaic virus	Virgaviridae	2
ssRNA negative-strand viruses		
Newcastle disease virus	Paramyxoviridae	125
Influenza A virus	Orthomyxoviridae	1090
dsDNA viruses, no RNA stage		
Cercopithecine herpesvirus 5	Herpesviridae	1
White spot syndrome virus	Nimaviridae	1
ssDNA viruses		
Duck circovirus	Circoviridae	3
Columbid circovirus	Circoviridae	204
Porcine circovirus	Circoviridae	2

Table 2 (continued)

Species	Families	Number of hits
Diatraea saccharalis densovirus	Parvoviridae	226
Adeno-associated virus	Parvoviridae	1
Rat adeno-associated virus 1	Parvoviridae	5

Discussion

Surveillance and identification of RNA viruses are important to the control and prevention of infectious diseases [4]. NGS is very powerful in the identification of uncharacterized viruses, and will expand the understanding on virus ecology, structure, genome, and pandemic prediction [15]. In this study, our goal was to establish a NGS library preparation method for an Ion Torrent PGM platform, without viral purification and culture to identify novel viruses or obtain genome sequence for known virus species.

It is important to develop a method which would not require prior knowledge of the virus. Identification methods based on culture have disadvantages, such as long turnaround time, increased biohazard risks, and culture bias. Improvements in sequencing and detection technologies over the past 15 years have led to increased detection rates of existing, neglected, and unknown pathogens. To identify unknown viruses by NGS, a shotgun sequencing DNA library or a cDNA library synthesized from RNA with random priming RT-PCR is often used. These methods may result in a huge amount of host cell sequences included in the sequencing data, even in a sample with a very high percentage of viral RNA [16, 17]. Library construction methods based on random primers were reported and applied in viral genome sequencing by NGS platforms [18, 19]. In this method, although host genomic DNA and rRNA was depleted by centrifugation, filtration, and naked DNA/RNA digestion to increase the percentage of viral-specific RNA in the sample, there was also a huge number of host cell and bacterium sequences achieved by NGS. The key to lowering the amount of host contamination is not only the sample pre-processing but also the library preparation method. In order to generate a large number of target size distributions in the NGS library, Agilent 2100 Bioanalyzer was used to characterize size distribution during the random primer reverse transcription over various incubation times. The results showed that a reverse transcription time for 30 min can produce cDNA fragments with an average size of 353 bp. Although the experiment has not been replicated several times, this part of the study was useful in the further research of the relationship between the reverse transcription time and the first strand cDNA fragment sizes, as well as in obtaining a library with suitable fragment sizes and enhancing the quality and quantity of sequencing data. The method has been replicated and compared to the existing standard RNA-seq library preparation protocol. The results showed that more classified viral families and genera were identified using this method than the others [10].

Using the library preparation method, 1 and 18 virus families were identified in the two samples by NGS sequencing, respectively. In the cultured swab sample collected from a healthy duck from Guizhou province, only Paramyxovirinae was detected. In the uncultured fecal sample mixed from 52 dead ducks in a poultry farm of Shandong province, 12 families of animal virus (Picobirnaviridae, Reoviridae, Hepadnaviridae, Retroviridae, Astroviridae, Coronaviridae, Picornaviridae, Paramyxoviridae, Orthomyxoviridae, Herpesviridae, Nimaviridae, Circoviridae), 4 families of plant virus (Potyviridae, Betaflexiviridae, Virgaviridae, Nimaviridae, Circoviridae, Parvoviridae), 2 families of insect virus (Dicistroviridae, Togaviridae,) were identified. The 12 families of animal virus were the main viruses infecting the 52 dead ducks in the farm, which were not the virus infecting one duck. Regarding some virus (Zucchini yellow mosaic virus, Sindbis virus, and Diatraea saccharalis densovirus), they were assumed to be from duck feed sources, as similar viruses had been identified from plants, insects, or shrimps previously. Interestingly, the number of the reads hitting to Avian encephalomyelitis virus was lower than the virus infecting plants and insects. The reason might be that the detected host was not in shedding period of Avian encephalomyelitis virus, which was less than 5 days in adults [20].

Complexity of the library preparation process produced by sequencing is critical in evaluating the NGS library preparation method [21]. The method developed in this study was simple to perform. The NGS library preparation method for RNA virus identification demonstrates its effectiveness in unknown pathogens detection and RNA virus genome sequencing. It also provides a method for rapid pathogen detection and infectious disease investigation, which are important in minimizing morbidity and mortality in viral infectious disease outbreaks. This rapid and low-cost method could be a utility in the routine diagnosis and investigation of viral infections and viral evolution.

Author contributions KW supervised the research, contributed to the design and development of the experimental work, and wrote the paper.

GC contributed to the experimental work and the bioinformatics analysis. YQ and QZ contributed to the experimental work. SW performed the revise of the language. TW performed the bioinformatics analysis.

Compliance with ethical standards

References

1. F.A. Al-Ateeg, Saudi Med. J. **24**, 794–798 (2003)
2. W. Li, Z. Shi, M. Yu, W. Ren, C. Smith, J.H. Epstein, H. Wang, G. Crameri, Z. Hu, H. Zhang, J. Zhang, J. McEachern, H. Field, P. Daszak, B.T. Eaton, S. Zhang, L.F. Wang, Science **310**, 676–679 (2005)
3. J. Parry, BMJ **346**, f2568 (2013)
4. J. Chen, B.X. Huang, *Guidelines of Epidemilogical Surveys of Important Animal Infectious Diseases* (Chinese Agricultural Science and Technology Press, Beijing, 2000)
5. K.C. Wang, G.Q. Chen, W.M. Jiang, S. Liu, G.Y. Hou, J.M. Yu, J.P. Li, Q.Y. Zhuang, J.M. Chen, Genome Announc. **1**, e0004513 (2013)
6. N. Fischer, D. Indenbirken, T. Meyer, M. Lutgehetmann, H. Lellek, M. Spohn, M. Aepfelbacher, M. Alawi, A. Grundhoff, J. Clin. Microbiol. **53**, 2238–2250 (2015)
7. A. Routh, P. Ordoukhanian, J.E. Johnson, J. Mol. Biol. **424**, 257–269 (2012)
8. Y. Wang, B. Lupiani, S.M. Reddy, S.J. Lamont, H. Zhou, Poult. Sci. **93**, 485–493 (2014)
9. M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu, BMC Genomics **13**, 341 (2012)
10. Y. Qiu, J.M. Chen, T. Wang, G.Y. Hou, Q.Y. Zhuang, R. Wu, K.C. Wang, Virus Res. **237**, 22–26 (2017)
11. R.C. Edgar, Nucleic Acids Res. **32**, 1792–1797 (2004)
12. G.Q. Chen, Q.Y. Zhuang, K.C. Wang, S. Liu, J.Z. Shao, W.M. Jiang, G.Y. Hou, J.P. Li, J.M. Yu, Y.P. Li, J.M. Chen, PLoS ONE **8**, e72918 (2013)
13. N. Buisine, R. Chalmers, Biotechniques **37**, 987–989 (2004)
14. D.H. Huson, A.F. Auch, J. Qi, S.C. Schuster, Genome Res. **17**, 377–386 (2007)
15. A. Djikeng, R. Halpin, R. Kuzmickas, J. Depasse, J. Feldblyum, N. Sengamalay, C. Afonso, X. Zhang, N.G. Anderson, E. Ghedin, D.J. Spiro, BMC Genomics **9**, 5 (2008)
16. G. Palacios, J. Druce, L. Du, T. Tran, C. Birch, T. Briese, S. Conlan, P.L. Quan, J. Hui, J. Marshall, J.F. Simons, M. Egholm, C.D. Paddock, W.J. Shieh, C.S. Goldsmith, S.R. Zaki, M. Catton, W.I. Lipkin, N. Engl. J. Med. **358**, 991–998 (2008)
17. M. Kuroda, H. Katano, N. Nakajima, M. Tobiume, A. Ainai, T. Sekizuka, H. Hasegawa, M. Tashiro, Y. Sasaki, Y. Arakawa, S. Hata, M. Watanabe, T. Sata, PLoS ONE **5**, e10256 (2010)
18. L.K. McMullan, M. Frace, S.A. Sammons, T. Shoemaker, S. Balinandi, J.F. Wamala, J.J. Lutwama, R.G. Downing, U. Stroeher, A. MacNeil, S.T. Nichol, Virology **422**, 1–5 (2012)
19. C.L. Afonso, Biotechniques **43**, 188, 190, 192 (2007)
20. H.A. Westbury, B. Sinkovic, Aust. Vet. J. **54**, 81–85 (1978)
21. S.R. Head, H.K. Komori, S.A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D.R. Salomon, P. Ordoukhanian, Biotechniques **56**, 61–64, 66, 68, passim (2014)