საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 10, №4, 2016
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 10, no. 4, 2016

Molecular Biology

Next Generation Sequencing Method of Avian Influenza Virus, Optimised Research Protocol

Marine Murtskhvaladze* and Adam Kotorashvili*

**L. Sakvarelidze National Center for Disease Control and Public Health, Tbilisi, Georgia*

(Presented by Academy Member Davit Mikeladze)

ABSTRACT. Avian influenza is caused by specified viruses that are members of the family *Orthomyxoviridae* and refers to the genus Influenza Virus A. 16 haemagglutinin (HA) and 9 neuraminidase (NA) subtypes have been isolated from birds. Most avian influenza viruses (AIVs) are of low pathogenicity and cause mild or subclinical infections in aquatic birds. The interface between the natural host reservoir of avian influenza viruses (AIVs), where birds often do not exhibit overt signs of disease (low pathogenic avian influenza (LPAI). Domestic poultry, particularly gallinaceous birds, in which clinical signs may be more obvious, is of key importance, when evaluating the risk of emergence of influenza viruses from the natural host reservoir. The most devastating poultry disease scenario is highly pathogenic avian influenza (HPAI), which is characterised by high morbidity and mortality, and occurs only among H5 and H7. Georgia is important for migration and over-wintering of wild water birds. Thus, it might act as a migratory bridge for influenza virus transmission during migration. In 2009-11 AIV prevalence of 6.3% was observed in ducks and 9% in large gulls during the autumn post-moult aggregations, wintering and migration stop-over period. The molecular characteristics of viruses that exhibit an expanded host range are, to date, poorly understood. Characterization of the virus population in the natural host reservoir, mechanisms of transmissions to other individuals requires full-genome sequencing of each infection cases. *© 2016 Bull. Georg. Natl. Acad. Sci.*

Key words: avian influenza virus, next generation sequencing, avian migration

The influenza A virus genome (family *Orthomyxoviridae),* consists of eight unlinked segments of negative-sense single stranded RNA, which code for 11 proteins. Influenza viruses are classified on the basis of two of these proteins expressed on the surface of virus particles; the haemagglutinin (HA) and neuraminidase (NA) glycoproteins [1, 2]. To date, sixteen haemagglutinin (HA) and nine neuraminidase (NA) subtypes have been isolated from birds [3-5]. Most avian influenza viruses (AIVs) are of low pathogenicity and cause mild or subclinical infections in birds. Only among H5 and H7 occurs highly pathogenic avian influenza (HPAI), which is characterized by high morbidity and mortality. Until recently there has only been limited whole-genome sequence data available for AIVs in Eurasia, Africa, South America and Oceania. The H16 and H13 subtypes have been shown to be mostly gull-spe-

© 2016 Bull. Georg. Natl. Acad. Sci.

cific (order *Charadriiformes*). However H1, H2, H4, H6, H9, and H11 also have been isolated from gulls sporadically [6], moreover, it has been suggested that all H13 viruses have genomes with a mosaic geographic origins [7].

Black sea is an important region for research of influenza virus spreading by aquatic birds during migration, and therefore, wetland habitats in the region are migratory stopover and over-wintering area for tens of thousands of mallards, gulls and other birds. During 2009-2011 surveillance an AIV in Georgia showed prevalence of the virus 6.3% in ducks and 9% in large gulls [8].

This study is the first attempt of performing Next Generation Sequencing (NGS) of Avian influenza viruses in new established Genomic Laboratory at R. Lugar center/NCDC, isolated during AIV surveillance in Georgia. Aim of the research were: To create pipeline of AIV research in Georgia - to adjust and validate protocols for viral genomic study; Describe AIV genomes and deposit in available data bases (Genbank, Influenza research database); This information will then allow us to understand pathogen migration routes, evaluate risk area and risk groups, develop measures for disease prevention and spread in the region.

Material and Methods

Sample collection. The sampling took place in 2015-2016, AIV's outbreak season, across the wetland in western Georgia. Our filed survey team visited three sampling site near Batumi, Paliastomi Lake and Poti. We collected samples from aquatic wild birds and domestic poultry. Samples were collected according Bioethics Guidelines, without physically injury of the birds. According to Georgian legislation there is no permission requested for sampling water birds as they are not listed in threatened category. Sterile plain cotton swab was inserted into trachea and cloaca, was turned to moisten the swab and then inserted into viral transport media. Capturing and sampling methods of water bird species, as well virus isolation and diagnostics are detailed described [8].

Short-read DNA sequencing using Miseq Illumina at NCDC Tbilisi. Full genome sequencing of AIV was carried out in the newly opened NCDC/CPHR genome center. AIV positive purified RNA was quantified using 2100 Bioanalyzer Laptop Bundle (Agilent Technologies, Santa Clara, USA), and Qubit® 2.0 Fluorometer (Invitrogen, Grand Island, USA). High quality RNA (> 40% of RNA is > 600 nt) was selected for DNA fragmentation and annealing. The first strand cDNA synthesis was conducted using NEBNext Ultra RNA Prep Kit using random primers and amplified at 25 °C for 10 min, 42 °C for 50 min, 70 °C for 15 min. Second Strand Master Mix was added into the original reaction mixture to synthesise the second strand at 16 °C for 60 min. Followed by RNase I Treatment and purification steps. An A-Tailing Mix was added into the purified DNA by incubating at 37 °C for 30 min to adenylate the 32 ends, preventing ligation with one another. Ligation to commercial adapters, a thymidine overhang from the indexed paired-ends adapters, was conducted at 30 °C for 10 min by adding DNA Ligase Mix and adapters to the sample DNA. To stop the ligation, Stop Ligase Buffer was added into the reaction mixture. AMPure XP beads were utilized to purify the complete double-stranded cDNA. Library enrichment was necessary to maximize the amount of DNA fragments that were attached to adapters (adapter-ligated DNA). NEBNext High Fidelity PCR Master Mix, PCR Primer Cocktail and Index PCR Primer were used at 15 min 37 °C 98 °C for 10 s for a total of 15 cycles at 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s and a final extension at 72 °C for 5 min. To maximise the ligation efficiency of the double-stranded DNA to the adapters, the ends of each DNA fragment were repaired by incubating with End Repair Mix at 20 °C for 30 min 65 °C 30 min to generate blunt ends for all DNA fragments. The blunt-ended double-stranded DNA was purified by AMPure XP beads. The enriched library was purified by AMPure XP beads followed by eluting with Resuspension Buffer. To monitor the quality of the reactions, "in-line" or internal controls were added to the sample while conducting the end repair,

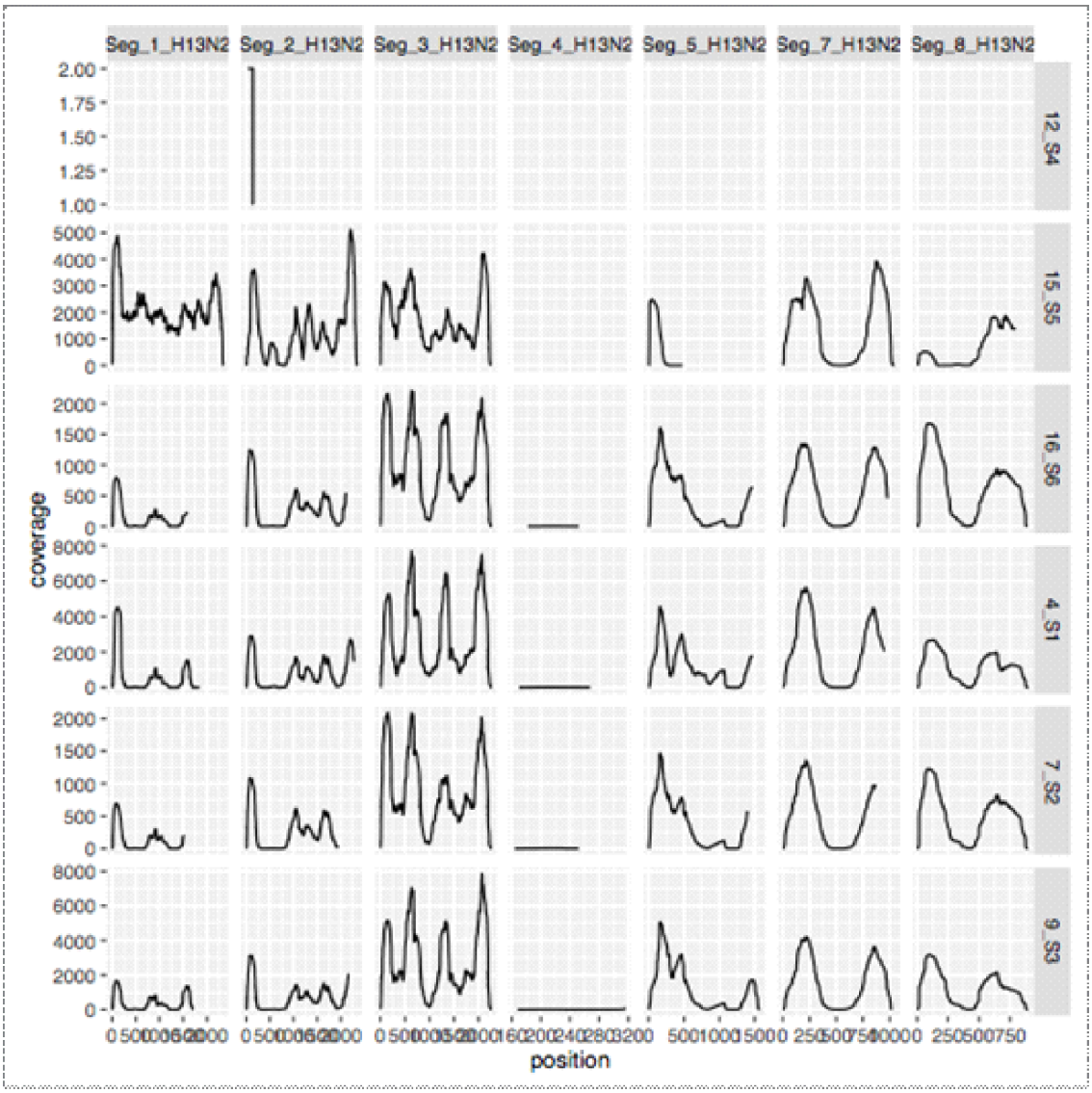

Fig. 1. Depth of coverage plots (DOC) of all five samples. Segments are as following: Seg -1 PB2; Seg 2 PB1; Seg 3 PA; Seg 4 hemagglutinin HA, Seg 5 nucleoprotein NP, Seg 7 matrix (M), and Seg 8 NS. Position from 5′ to 3′ was plotted in *X* axis and DOC was plotted in *Y* axis.

3′ end adenylation and ligation. Ethanol precipitation was performed prior to proceeding to TruSeq LT RNA sample preparation according to the manufacturer's instructions.

DNA quantity and quality control was validated and monitored by 2100 Bioanalyzer Laptop Bundle (Agilent Technologies, Santa Clara, USA), and Qubit® 2.0 Fluorometer (Invitrogen, Grand Island, USA). Concentration of the adapter-ligated DNA was calculated using the following formula: x nM= (Concentration ng/µl * x 10000000)/(660g/mole) (bp Avg. Length) to determine the volume of each sample to pool and make up to 2 nM of library mixtures. All samples were normalized to 2nM, NaOH were added to equal volumes of the normalized DNA libraries for DNA denaturation, followed by dilutions with pre-chilled HT1 buffer to obtain the DNA libraries at 20 pM. The DNA libraries were further diluted to 4 pM and multiplexed with a total volume of 1 ml with pre-chilled HT1 buffer. One percent PhiX, used as internal control, was supplemented into the 4 pM denatured DNA solution to observe the efficiency of DNA

incorporation during DNA sequencing. Libraries were loaded into the cartridge, and sequenced as multiplex two read libraries for 300 cycles (including additional 6 cycles of index reads) according to the manufacturer's protocol. All sequencing runs for 40 mers was performed with GA using the Illumina MiSeq Reagent kit (ver. 3). Fluorescent images were analysed using the Illumina base-calling pipeline 1.4.0 to obtain FASTQ formatted sequence data.

Genome assembly. We performed quality control analysis using Fast QC v0.11.5 software [9]. NGS data trimming and adapter clipping was performed with Trimmomatic version 0.33 [10]. We used the quality filtering functionality scanned reads from the 5′ end, and removed bases from the 3′ end with low average quality score. De-novo assembly was performed with Trinity v2.2.0 [11]. Reference nucleotide sequences for each segment of those influenza A strains that have sequences for full genomes available were obtained from the Influenza Research Database (www.fludb.org). A BLAST-database was created from these sequences using BLAST+ v2.3.0 [12]. The de-novo assembled reads were then blasted against this database to identify the most likely present strain in each sample. A representative sequence of the contemptible strain has been identified by clustering a database containing only those strains, separated by segment using CD-Hit v4.6.1 [13]. (These representative references have been used for full seeded iterative assembly of each segment using MITObim v1.8 [14] on the raw reads that have been quality trimmed and adapter clipped beforehand by Trimmomatic v0.36 [10] and merged by FLASh v1.2.11 [15].

The MIRA software [16] was run using default parameters for ''de novo assembly''. The preliminary mitogenome assembly was used as input for the algorithm MITObim [14] which performs MIRA iterations and improves the assembly. The MITO bim assembled sequences were visualised using Tablet [17] to check for the genome coverage.

Results and Discussion

Sample collection and sub -typing. During 2014-2015, in total 648 samples were collected. Among them 603 were from gulls and ducks, and 45 from domestic birds. Within collected samples, 11 Yellow legged Gull samples from Chorokhi delta and five Armenian gull samples from Madatafa Lake were identified as AIV positive.

Next Generation Sequencing at NCDC Tbilisi. Based on MiSeq capability and nature of the samples, the RNA TruSeq LT Kit was chosen for preparation of the eleven samples. After the cDNA was synthesized and dsDNA amplified, the quality of DNA was validated by two different measurements: (1) Aligent 2100 Bioanalyzer Laptop Bundle and (2) Qubit® 2.0 Fluorometer. The final concentrations and average fragment sizes measured by Agilent 2100 Bioanalyzer Laptop Bundle were 216–388 nM and 270–300 bp respectively. Results from Quibit is shown in table 1. After library preparation, The Qubit® 2.0 Fluorometer measured the overall DNA concentration at 2.72 - 32 ng/μl. The best five samples were selected and run on Illumina Miseq. Although the MiSeq highest optimal capacity was 15 pM DNA of template, 5 pM of library mixtures was chosen for sequencing. The sequencing from Illumina MiSeq Platform yielded a total of 8.2 Gbases from the five samples both Read 1 and Read 2, which were generated from a 455 ± 14 K/mm^2 cluster density. The quality of the base calling from images and sequences was determined by quality score (Q). Approximately 96% of the clusters passed QC filters and 93.7% of Read 1 and 83.5% of Read 2 sequences were ≥Q30 (99.9% accuracy of base calling at a particular sequence position). On average, 8.2 Gbases or 88.7% of both reads passed ≥Q30. The incline curve of %≥Q30 as sequencing progressed from Read1 to Read 2. Additionally, the signal to noise ratio was 6–16, which increased with higher quality base calling or %≥Q30. Less than 1% error rate was detected in sequences with 60–100% ≥Q30. A mini-

Table. List of selected samples and number of generated sequences.

Sample #	cDNA concentration (measured by Quibit fluorometer)	Number of sequences per reads	Total number of Nucleotides
4	10.4 ng/μl	3,161,342	1,499,079,253
7	6.4 ng/μl	4,791,037	2,308,783,759
9	9.68 ng/μl	1,926,971	897,273,494
12	32.2 ng/μl	1,261,477	628,345,529
16	7.28 ng/μl	3,281,229	1,594,615,961

mal amount of phasing (delayed sequencing) and pre-phasing (ahead sequencing) were detected (0.0281% and 0.1585%, respectively).

The genomic sequences from the five samples sequenced with MiSeq were assembled into eight genomic fragments consisting of (1) PB2, (2) PB1, (3) PA, (4) HA, (5) NP, (6) NA and (7) NA/NB-additional NB protein [18]. The sequencing from MiSeq showed the majority of viral isolates were seasonal influenza A H13N2 strains. In contrast, sample # 12 had highest cDNA concentration showed less number of sequences per reads and less number of mapped reads to references sequence, therefore was removed from further analysis. Based on multiple alignment analyses of our samples, at least 50x fold DOC on matched references was observed, which was higher than unmatched references. Depth of coverage (DOC) for the samples is shown in figure 1. Overall, the DOC for all five isolates compared to their matched influenza reference ranged from 0 to 8.000. The highest number of reads within all samples was mapped against PA Segment and lower reads against N segment.

Three segments PA, PB2 and M were selected for further analysis. All three segments showed that samples #4; # 7 #9; #15; #16; (Table) are identified as subtypes H13N2, same result were shown by Matrix gene testing. This means that library preparation protocol for next generation sequencing as well as DNA assembly and data analysis protocols are optimized and validated.

To use genomic data for understanding AIV evolution, to reconstruct phylogeny, identify risk groups and risk area, to draw clearer picture it is recommended to continue genomic study of Avian Influenza Virus and generate data from all eight segment, rather than particular segments.

In summary, the NGS MiSeq Platform at Genome Center R. Lugar Center/NCDC can identify and obtain complete sequence information from seasonal Avian Influenza Virus. Research pipeline for viral next generation sequencing, handling avian influenza viruses together with critical data quality-control assessment techniques and genome assembly was created.

Acknowledgement. This work was funded by Shota Rustaveli National Science foundation and Georgian Research and Development Foundation. Nato Kotaria and Ani Machablishvili provided valuable suggestions and were particularly instrumental in laboratory guidance. We are very grateful to Zura Javakhishvili, Nicola Lewis Simon Watson, for their help with analysis and for providing AIV positive samples.

მოლეკულური ბიოლოგია

ფრინველის გრიპის ვირუსის კვლევა ახალი თაობის სეკვენირების მეთოდით -კვლევის ოპტიმიზირებული პროტოკოლი

მ. მურცხვალაძე*, ა. კოტორაშვილი*

** ლ. საყვარელიძის სახ. დაავადებათა კონტროლისა და საზოგადოებრივი ჯანმრთელობის ცენტრი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის დავით მიქელაძის მიერ)

ფრინველის გრიპს იწვევს ვირუსი, რომელიც ეკუთვნის ოჯახ Orthomyxo-viridae და გვარ influenzavirus A. A ტიპის გრიპის ვირუსების ზედაპირზე ორი სპეციფიკური პროტეინი არსებობს: ჰემაგლუტინინი (H) და ნეირამინიდაზა (N). "H" და "N" პროტეინული შემადგენლობა განსაზღვრავს ვირუსის შტამის კლასიფიკაციას. დღესდღეობით, 16 ჰემაგლუტინინი და 9 ნეიროამინიდაზა არის იდენტიფიცირებული. ამ ვირუსების უმრავლესობა წარმოადგენს დაბალპათოგენურ ფორმას და იწვევს უმნიშვნელო ან სუბკლინიკურ ინფექციებს ფრინველების უმრავლესობა სახეობებში. უნდა აღინიშნოს ისიც, რომ ვირუსით გამოწვეული სიკვდილიანობა განსხვავდება სახეობების მიხედვით და საკმაოდ მაღალია ქათმისნაირ შინაურ ფრინველებში. ფრინველის გრიპის მაღალპათოგენური ფორმა, რომელიც ხასიათდება მაღალი სიკვდილიანობით, გვხდება მხოლოდ H5 და H7 სუბტიპებში. საქართველო წარმოადგენს სამიგრაციო დერეფანს მოზამთრე და მიგრირებადი წყალმცურავი ფრინველებისთვის. რაც შესაძლოა ხელს უწყობდეს ფრინველის გრიპის ვირუსის გავრცელებას ფრინველთა მიგრაციის დროს. 2009-2011 წლებში, წყალმცურავი ფრინველების სამიგრაციო და საზამთრო ადგილებში ჩატარებულმა კვლევამ გამოავლინა ფრინველის გრიპის ვირუსის შემთხვევათა 6,3% გარეული იხვის და 9% თოლიების სახეობაში. დღეისათვის, საქართველოში გავრცელებული AIV ვირუსის ბიოლოგია სუსტადაა შესწავლილი და არც ვირუსის სხვა მასპინძლებზე ტრანსმისიის მექანიზმებია ნათელი. ამ ფენომენის შესწავლის ერთ-ერთი გზა გენომური კვლევებია.

REFERENCES:

1. *Webster R.G., Bean W.J., Gorman O.T., Chambers T.M., Kawaoka Y*. (1992) Microbiological reviews. **56**, 1: 152-79.
2. *Fujii K., Fujii Y., Noda T., Muramoto Y., Watanabe T., Takada A., Goto H., Horimoto T., Kawaoka Y*. (2005) Journal of virology. **79**, 6: 3766-3774.
3. *Alexander Dennis J.* (2000) Veterinary microbiology. **74**, 1: 3-13.
4. *Elfidasari D., Solihin D.D., Soejoedono R.D., Murtini S., Noor Y.R.* (2011) Makara Journal of science. **15,** 2: 179-185.
5. *Jackwood, Daral J., and Susan Sommer-Wagner* (2007) Virology. **365,** 2: 369-375.
6. *Kawaoka Y., Chambers T.M., Sladen W.L., Gwebster R*. (1988) Virology. **163,** 1: 247-250.
7. *Olsen B., Munster V.J., Wallensten A., Waldenstrom J.M., Osterhaus A.D., Fouchier R.A*. (2006). Science. **312**, 384–388.
8. *Lewis N.S., Javakhishvili Z., Russell C.A., Machablishvili A., Lexmond P., Verhagen J.H.* (2013) PLoS ONE. **8** (3): 58534.
9. http://www.bioinformatics.babraham.ac.uk/projects/download.html
10. *Bolger A., Giorgi F.* (2014) Trimmomatic: a flexible read trimming tool for illumina NGS data. URL http:// www. usadellab. org/cms/index. php.
11. *Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J, Couger M.B., Eccles D., Li B., Lieber M., MacManes M.D. D.* (2013) Nature protocols. **8,** 8: 1494-512.
12. *Weizhong Li and Godzik Adam* (2006) Bioinformatics. **22**, 13: 1658-1659.
13. *Hahn, Christoph, Lutz Bachmann, and Bastien Chevreux* (2013) Nucleic acids research. **1**, 3: 371-378.
14. *Magoè, Tanja, and Steven L. Salzberg.* (2011) Bioinformatics. **27,** 21: 2957-2963.
15. *Chevreux B., Wetter T., Suhai S.* (1999) Computer Science and Biology: Proc. German Conf Bioinform. **99,** 1: 45–56.
16. *Milne I., Stephen G., Bayer M., Cock P.J.A, Pritchard L., Cardle L., Shaw P.D., Marshall D.* (2013) Brief Bioinformatics. **14,** 1: 193–202.
17. *Hatta M., Kawaoka Y.* (2003) Journal of Virology. **77,** 10: 6050-6054.

Received September, 2016