Virus Genes
https://doi.org/10.1007/s11262-019-01669-9

ORIGINAL PAPER

Enhanced phylogenetic resolution of Newcastle disease outbreaks using complete viral genome sequences from formalin-fixed paraffin-embedded tissue samples

Salman Latif Butt[1,2] · Kiril M. Dimitrov[1] · Jian Zhang[2] · Abdul Wajid[3] · Tasra Bibi[4] · Asma Basharat[4] · Corrie C. Brown[2] · Shafqat F. Rehmani[4] · James B. Stanton[2] · Claudio L. Afonso[1]

Received: 25 February 2019 / Accepted: 7 May 2019

Abstract

Highly virulent Newcastle disease virus (NDV) causes Newcastle disease (ND), which is a threat to poultry production worldwide. Effective disease management requires approaches to accurately determine sources of infection, which involves tracking of closely related viruses. Next-generation sequencing (NGS) has emerged as a research tool for thorough genetic characterization of infectious organisms. Previously formalin-fixed paraffin-embedded (FFPE) tissues have been used to conduct retrospective epidemiological studies of related but genetically distinct viruses. However, this study extends the applicability of NGS for complete genome analysis of viruses from FFPE tissues to track the evolution of closely related viruses. Total RNA was obtained from FFPE spleens, lungs, brains, and small intestines of chickens in 11 poultry flocks during disease outbreaks in Pakistan. The RNA was randomly sequenced on an Illumina MiSeq instrument and the raw data were analyzed using a custom data analysis pipeline that includes de novo assembly. Genomes of virulent NDV were detected in 10/11 birds: eight nearly complete (> 95% coverage of concatenated coding sequence) and two partial genomes. Phylogeny of the NDV complete genome coding sequences was compared to current methods of analysis based on the full and partial fusion genes and determined that the approach provided a better phylogenetic resolution. Two distinct lineages of sub-genotype VIIi NDV were identified to be simultaneously circulating in Pakistani poultry. Non-targeted NGS of total RNA from FFPE tissues coupled with de novo assembly provided a reliable, safe, and affordable method to conduct epidemiological and evolutionary studies to facilitate management of ND in Pakistan.

Keywords NDV · FFPE · NGS · Chicken · Clinical · Tissue

Introduction

Newcastle disease (ND) is a significant worldwide disease of poultry caused by virulent strains of *Avian avulavirus* 1 (AAvV-1), commonly known as Newcastle disease virus (NDV) [1–3]. Endemicity of this virus in multiple countries is a major challenge to global poultry production and

Edited by William Dundon.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11262-019-01669-9) contains supplementary material, which is available to authorized users.

✉ James B. Stanton
jbs@uga.edu

✉ Claudio L. Afonso
claudio.afonso@ars.usda.gov

[1] Southeast Poultry Research Laboratory, Exotic and Emerging Avian Viral Diseases Research Unit, United States National Poultry Research Center, Agricultural Research Service, United States Department of Agriculture, Athens, GA, USA

[2] Department of Pathology, College of Veterinary Medicine, University of Georgia, Athens, GA, USA

[3] Department of Biotechnology, Virtual University of Pakistan, Lahore, Pakistan

[4] Quality Operations Laboratory (QOL), University of Veterinary and Animal Sciences, Lahore, Pakistan

Published online: 14 May 2019

at least four panzootics have been recognized since it was first identified in the 1920s [4–6]. Reliable and affordable epidemiologic tools that can use tissues transported in a safe and convenient manner across international boundaries are needed to improve the management of ND and other infectious diseases.

Newcastle disease virus is a negative-sense, single-stranded, non-segmented, enveloped RNA virus of the *Paramyxoviridae* family [1]. The genome of NDV is approximately 15 kb and encodes six structural gene products: nucleocapsid protein (NP), phosphoprotein (P), matrix protein (M), fusion protein (F), hemagglutinin-neuraminidase (HN), and RNA-dependent-RNA polymerase (L) [7]. The amino acid sequence of the F protein cleavage site is a major molecular determinant of pathogenicity and virulence of NDV. Virulent NDVs have multiple (3 or more) basic amino acids (^{112}R/K-R-Q-R/K-R↓F^{117}) at the F protein cleavage site and phenylalanine at position 117 [8]. Sequencing of the F gene and other gene segments has been used to predict viral virulence and to study epidemiology and evolution; however, this and other targeted methods have limitations in comparison to random complete genome sequencing.

Genotypic characterization of NDV is commonly done by sequencing the complete F gene [9], whereas complete genome characterization may allow more reliable evolutionary and epidemiological studies. The commonly used Sanger sequencing with overlapping primers for complete genome sequencing has significant limitations such as high cost, reduced output, long turnaround time, and is dependent on pre-existing knowledge of the genetic makeup of the viruses [10–12] making this an impractical method for full-genome sequencing.

Shipping and testing samples with live NDV is problematic because virulent NDV is classified as a select agent in many countries and working with live select agent demands extensive safety precautions (BSL-3 laboratories), expensive transportation, and special permissions. In contrast, formalin fixation inactivates infectious agents; thus, formalin-fixed paraffin-embedded (FFPE) tissues can be easily, safely, and affordably transported across boundaries and processed in a biosafety level 1 laboratory [13]. Additionally, formalin fixation is the gold standard for pathologic preservation and FFPE tissues allow for a full pathologic analysis of the tissue, which will provide a better list of differential diagnoses. As such, FFPE tissues are universally collected clinical samples for routine histopathology [14]. Flinders Technology Associates filter papers (FTA® cards) have been used for hazard-free handling of sample and detection of NDV by RT-PCR [15], and while these show promise for full-genomic characterization of infecting viruses (https://www.ncbi.nlm.nih.gov/pubmed/28684566), FTA cards are limited to genetic analysis only, which prevents broader testing that is often required in a diagnostic setting when samples are collected prior to a definitive diagnosis. Immunohistochemistry (IHC), targeting nucleocapsid protein of NDV as an antigen within FFPE tissues samples, has been used in experimental studies to demonstrate tissue tropism of NDV [16]; however, it does not provide as complete, sensitive, or specific characterization of the NDV as nucleic acid-based methods [13].

The advancement of massively parallel sequencing (next-generation sequencing; NGS) has significantly improved the ability to sequence full-length genomes using a non-targeted approach [17]. The use of random primers in a non-targeted sequencing approach is ideal for RNA viruses, which rapidly mutate resulting in the problems listed above for PCR and Sanger sequencing. Furthermore, recent studies have shown that the cost of NDV complete genome sequencing by NGS can be reduced approximately tenfold by multiplexing samples in a single sequencing run [10]. Previously, NDV detection and genome characterization by NGS have been done on egg-passaged live viruses [10]. There are limited reports on the application of NGS with FFPE samples for infectious disease studies [18–21]. Recently, a retrospective study describing the molecular evolution of pigeon-adapted NDV in the U.S. using archived FFPE tissue samples from wild pigeons was reported [22]. This approach, however, has not been tested to identify minor evolutionary changes or to trace the epidemiology of closely related isolates, as is often the case in endemic countries involved in intensive production of poultry. For example, recently reports showed that virulent NDVs have been circulating in Iran, a neighboring country of Pakistan, for the last 21 years. A novel virulent strain of NDV has been recently reported and epidemiological information indicates its relatedness to the NDV circulation in China which is also a neighboring country of Pakistan [6, 23, 24]. Due to the endemicity of NDV in Pakistan and the presence of constantly evolving genotypes VIIi in neighboring countries of Pakistan, clinical samples from Pakistan were chosen to test the ability of this method to identify minor evolutionary changes.

The aim of the current study was to sequence and characterize the genomes of NDV directly from clinical FFPE tissue samples by using NGS and to use the obtained data to track the epidemiology of closely related isolates. This method did not include any procedure to target NDV specifically (e.g., sequence-based capture) or to enrich for viral nucleotides generally (e.g., depletion of host ribosomal RNA). This study demonstrates the ability of NGS, assisted by a customized bioinformatics pipeline, to assemble nearly complete NDV genomes from FFPE tissues and that those genomes are suitable for improved phylogenetic differentiation between closely related NDV isolates.

Materials and methods

Sample collection and processing

Thirty-six FFPE field tissue samples (spleen, lung, brain, and small intestine) collected in 2015 from 11 chickens during disease outbreaks in 11 poultry flocks in five different regions of the Pakistani Punjab province were used in this study (Table 1). Collected tissues were fixed for 24 to 48 h in neutral-buffered 10% formalin within 4 h of collection. Fixed tissues were embedded in low-melt paraffin following routine procedures [25]. These paraffin blocks were stored at room temperature and subsequently shipped without refrigeration to the Southeast Poultry Research Laboratory of the USDA, Athens, Georgia, USA, for NDV characterization.

Immunohistochemistry assay

Immunohistochemistry was performed to detect NDV nucleocapsid protein in tissue samples (spleen, lung, brain, and small intestine). Briefly, sections of 3 µm thickness were cut and immunostained with monoclonal antibodies directed against NDV nucleocapsid protein using alkaline phosphatase method as described previously [26]. Immunostained sections were microscopically evaluated and the samples were scored as negative or positive.

RNA extraction

For each sample, six 10-µm-thick tissue sections were cut and collected in one 1.5-ml centrifuge tube and immediately deparaffinized by CitriSolv™ (VWR International, USA). Before and after collecting tissue sections from each sample, the microtome blade was decontaminated with RNase Away (Sigma, USA) to avoid cross-contamination between samples. Total RNA was extracted using the RNeasy FFPE Kit (Qiagen, USA) as per manufacturer's instructions and quantified using the Qubit® RNA HS Assay Kit on a Qubit® fluorometer 3.0 (ThermoFisher Scientific, USA). The purity of RNA (a ratio of absorption at 260 nm and 280 nm wavelength, $A_{260/280}$) was measured on a NanoDrop Spectrophotometer 2000/2000c (ThermoFisher Scientific, USA). Fragment size of RNA was determined using the RNA 6000 Pico kit on an Agilent Bioanalyzer® instrument (Agilent Technologies, USA) as per manufacturer's instructions.

NGS library preparation

DNA Libraries for NGS sequencing were prepared using the KAPA Stranded RNA-Seq Library Preparation Kit for Illumina platforms (Kapa Biosystems, USA) following manufacturer's instructions. Briefly, the protocol involved synthesis of first strand and second strand of DNA from total extracted RNA by using random primers, marking and A-tailing of cDNA fragments, and ligation of unique adapters allowing indexing of each sample for multiplexing purposes. Finally, bead-purified adapter-ligated libraries were amplified with PCR (12 cycles) using library amplification master mix provided with the kit. The Qubit® fluorometer 3.0 was used to quantify dsDNA concentration in libraries using the dsDNA High Sensitivity Assay kits (ThermoFisher Scientific, USA). The average DNA fragment size in each library was determined using the High Sensitivity DNA kit in the Agilent Bioanalyzer®. To facilitate uniform clustering during sequencing process in the MiSeq flow cell, libraries with an average DNA fragment size of 240 to 300 bp and a concentration greater than 3 ng/µl were used. The prepared libraries were diluted to 4 nM. Five microliters of each library were pooled and denatured with NaOH (0.2 N final concentration). After 5 min of incubation at room temperature, the pool was further diluted to 20 pM concentration with chilled HT1 hybridization buffer (Illumina, USA). Using the same buffer, the final concentration of the library

Table 1 Background information of 36 field FFPE tissue samples collected from different regions of Pakistan during disease outbreaks in 2015

Sample ID	Collected organs	No. of samples	Breed of chicken	Location
1162	Spleen, lung, small intestine	3	Broiler chicken	Kassur
1163	Spleen, lung, brain	3	Broiler chicken	Lahore
1164	Spleen, lung, brain, small intestine	4	Desi chicken	Lahore
1165	Spleen, lung, brain	3	Broiler chicken	Kassur
1166	Spleen, lung, brain	3	Broiler chicken	Kamoki
1168	Spleen, lung, brain, small intestine	4	Broiler chicken	Sheikhupura
1170	Spleen, lung, brain, small intestine	4	Broiler chicken	Lahore
1171	Spleen, lung, brain	3	Broiler chicken	Sheikhupura
1172	Spleen, lung, brain	3	Broiler chicken	Lahore
1173	Lung, brain, small intestine	3	Desi chicken	Gujranwala
1174	Lung, brain, small intestine	3	Broiler chicken	Gujranwala

pool was diluted to 10 pM. A control library (3% PhiX174, Illumina, USA) was added and the pool was chilled on ice. The paired-end sequencing was conducted on an Illumina MiSeq instrument using a 300 cycle (2 × 150) MiSeq Reagent Kit v2 (Illumina, USA). After automated cluster generation in MiSeq, the sequencing reads were demultiplexed based on their unique adapter.

Genome assembly

Pre-assembly processing from read quality assessment to digital normalization was conducted as previously described [10] using Fast QC [27], Cutadapt v1.6 (TruSeq LT index sequences were used as reference file the study herein) [28], BWA-MEM v0.2.1 [29], Filter sequencing by mapping v0.0.4 tool (https://github.com/peterjc/pico_galaxy/tree/master/tools/seq_filter_by_mapping), an in-house tool for forward and reverse read re-synchronization (https://github.com/jvolkening/b2b-utils), PEAR v0.9.6.0 [30], and Khmer package v1.1-1 [31]. De novo assembly of the filtered, trimmed, and synchronized reads was performed with MIRA assembler v3.4.1 [32]; however, to account for the relatively low number of NDV reads, as compared to egg-grown NDV samples from the previous study [10], the following parameters for de novo assembly were changed: minimum number of reads required to build a contiguous sequence = 5, minimum overlap = 10, contig length cut off = 60 bp, default values were used for the rest of the settings. The assembled contigs were aligned to the NCBI nt database (accessed on 15 September 2017), using BLASTn (cut off *E* value = 0.001) and the best hit was further used as the reference genome. As previously described [10], the final consensus sequence was obtained by processing the trimmed and un-normalized reads (after read re-synchronization) through BWA-MEM [29] (using the BLASTn-defined genome as the reference) and an in-house tool for parsing (https://github.com/jvolkening/b2b-utils). In addition, when a bird had multiple tissues with NDV, the raw sequence data from those differing tissue samples were merged and re-analyzed again using the same bioinformatics workflow to see if there is any difference in depth and percentage of genome coverage on a per-tissue basis versus a per-bird basis. Tissues were considered NDV positive by NGS if at least one contig hit to NDV (see BLASTn step above).

Phylogenetic analyses

The final consensus sequences were aligned using ClustalW [33] and the concatenated complete genome coding sequences (CDS) were used for nucleotide (nt) distance estimation to closely related NDV isolates obtained from GenBank using MEGA6 [34] and the Maximum Composite Likelihood model [35]. Consensus sequences from six birds in this study (for which complete genome coding sequences were obtained) and 32 sequences of sub-genotypes of genotype VII obtained from GenBank were used for final phylogenetic tree construction (See Table S1). Determination of the best-fit substitution model was performed using MEGA6, and the goodness-of-fit for each model was measured by corrected Akaike information criterion (AICc) and Bayesian information criterion (BIC) [34]. The final tree was constructed using the maximum-likelihood method based on the General Time Reversible model as implemented in MEGA6, with 1000 bootstrap replicates [36]. Additionally, two more datasets were parsed from the initial dataset—one containing the complete fusion gene coding sequences and one with the first 375 nucleotides of the fusion gene coding sequences (denoted as "partial fusion gene sequence"), both regions being commonly used in phylogenetic analyses and epidemiological studies [9, 37]. Nucleotide distance estimation was performed as described above and bootstrap maximum-likelihood phylogenetic trees based on the Tamura 3-parameter (Tamura 1992) were constructed using these smaller datasets utilizing MEGA6.

Accession numbers

The sequences obtained in the current study were submitted to GenBank and are available under the accession numbers from MG200021 to MG200026.

Statistical analyses

Statistical analyses were performed in JMP Pro Version 13.2. A Chi-square test (a contingency Table 2 × 4) was performed to determine whether frequency of NDV identification by NGS is significantly different between different tissue types. In addition, a one-way-ANOVA by ranks Kruskal–Wallis H test was performed to determine if mean percentage genome coverage is significantly different between different types of tissue.

Results

Immunohistochemistry

Thirty-six FFPE tissues (spleen, brain, lung, and small intestine) were examined by IHC with a primary monoclonal antibody to detect NDV nucleoprotein (Table 2, See Fig. S1). At least one tissue per bird was positive by IHC. Of 36 tissue samples, 31 were IHC positive for NDV nucleocapsid protein and 5 samples were IHC negative.

Table 2 Summary of sequencing and IHC data of 31 field FFPE tissue samples collected from different regions of Pakistan during disease outbreaks in 2015

Samples		IHC (±)	Raw read pairs	Filtered read pairs[a]	De novo NDV detection by contig (pos/neg)	NDV reads	NDV reads (% of raw read pairs)	Percent genome coverage
SEPRL ID	Tissue							
1162-L	Lung	+	386,682	31630	Pos	791	0.20	94.92
1162-SI	S. Intestine	–	372,467	1349	Neg	NA	NA	NA
1163-SP	Spleen	+	928	6	Incon.[b]	NA	NA	NA
1163-L	Lung	+	395,061	808	Neg	NA	NA	NA
1163-B	Brain	+	385,336	939	Pos	577	0.15	91.75
1164-SP	Spleen	+	441,166	1194	Pos	346	0.08	77.71
1164-L*	Lung	+	351,767	6158	Pos	5383	1.53	99.72
1164-B[c]	Brain	+	366,257	394	Neg	NA	NA	NA
1165-SP	Spleen	+	407,295	3253	Pos	884	0.22	93.26
1165-L[c]	Lung	+	399,209	1582	Neg	NA	NA	NA
1165-B*	Brain	–	316,718	13426	Pos	6033	1.90	99.72
1166-SP	Spleen	+	4,797,051	47742	Pos	44525	0.93	99.84
1166-L*	Lung	+	1,616,222	14386	Pos	9358	0.58	99.81
1166-B	Brain	–	394,885	5023	Pos	203	0.05	53.67
1168-SP	Spleen	+	445,743	662	Pos	308	0.07	61.56
1168-L	Lung	+	398,836	1676	Pos	255	0.06	55.27
1168-B*	Brain	+	322,405	5281	Pos	3801	1.18	99.77
1168-SI	S. Intestine	+	414,401	3595	Pos	3100	0.75	99.47
1170-SP	Spleen	+	592,168	1087	Pos	768	0.13	94.67
1170-L*	Lung	+	545,772	4250	Pos	2626	0.48	99.72
1170-B	Brain	+	411,037	5430	Pos	761	0.19	96.29
1170-SI	S. Intestine	+	2,006,616	2589	Pos	197	0.01	51.17
1171-SP[c]	Spleen	+	530,855	237	Neg	NA	NA	NA
1171-L	Lung	+	389,248	1347	Pos	179	0.05	42.00
1172-SP[c]	Spleen	+	315,083	380	Neg	NA	NA	NA
1172-L[c]	Lung	+	455,865	1008	Neg	NA	NA	NA
1172-B	Brain	–	442,780	1086	Neg	NA	NA	NA
1173-L	Lung	+	273,342	2236	Pos	1057	0.39	98.06
1173-SI[c]	S. Intestine	+	373,069	2256	Neg	NA	NA	NA
1174-L*	Lung	–	255,408	4670	Pos	2533	0.99	99.80
1174-B	Brain	+	275,619	410	Neg	NA	NA	NA

NA not applicable

*Genome sequences were used to build phylogenetic tree

[a]The number of paired reads after filtering out host (*Gallus gallus* 5.0) and internal control (PhiX) reads

[b]Suboptimal read generation was observed from #1163-spleen and this result was determined to be "inconclusive"

[c]IHC-positive cells ranged from 10 to 200 per studied tissue section

RNA isolation and evaluation

Total RNA from thirty-six tissue samples was extracted with RNA concentrations of 15.8–84 ng/µl. The 260/280 ratios were 1.9–2.03, and the RNA integrity number (RIN) values were 1.7–6.5 (see Table S2 for details).

Next-generation sequencing and genome assembly

Raw read analysis

Five of the prepared thirty-six libraries did not meet the criteria (see materials and methods) set for library quality and

were not submitted for sequencing. Suboptimal read generation was observed for sample 1163-spleen. Only 928 total reads (0.005% of the raw reads) were assigned to this sample and its reads were not included in downstream data analysis. Multiplexed NGS of the remaining 30 tissue samples generated a total of 19,078,363 raw reads (255,408 to 4,797,051 per sample). A total of 166,084 reads (237 to 47,742 per sample) remained after filtering out the host genome and PhiX control reads (Table 2). NDV reads were a small fraction of the total reads obtained (0.01–1.9% per positive sample). Approximately 91–99% of the total raw reads mapped to the chicken genome. However, the chicken reads were filtered out for the purposes of this study. The reads that mapped to bacterial genome were also filtered out. These values are not presented in the results.

NDV contigs: per tissue

Results were initially analyzed on a per-tissue basis. By employing de novo assembly coupled with reference-based consensus re-calling, NDV contigs (NDV positive) were obtained from 64.5% (20/31) of the sequenced tissue samples (55.6% [20/36] of the total extracted), with NGS-positive samples having 179–44,525 NDV reads per positive sample. In 14 of the tissue samples, the NDV genome coverage was greater than 90% (91.75–99.84%) (Table 2). Although NDV was identified by NGS in more lung tissues (72.7%) compared to other tissues (spleen = 62.5%, brain = 62.5%, and small intestines = 50%), these differences were not statistically significant (χ^2 Chi-square test = 0.721; $p = 0.8684$, $\alpha = 0.05$). In terms of genome recovery among different tissues types (Table 2), mean percentage genome coverage was not significantly different among different tissue types (Kruskal–Wallis statistic = 0.9916, $p = 0.8033$, $\alpha = 0.05$). Additionally, no nucleotide differences were identified between sequences obtained from different tissues of a same bird (data not shown).

NDV contigs: per bird

Since identical consensus sequences were identified among different tissues, the raw reads from different tissues of the same bird were merged so that the data could be analyzed on per-bird basis. NDV contigs were assembled from these data in 91% of the birds (10 out of 11 birds) with 191–54,086 NDV reads per bird. When the raw sequence data obtained from different tissues of the same bird were merged, in 9 out of the 11 birds, the genome coverage was greater than 90% (91.87 to 99.83%). The mean depth of each assembled genome per bird ranged from 4 to 513 in the 9 out of 10 NGS-positive birds that had > 90% genome coverage. In one bird (#1171) that was NGS positive, the genome coverage was 45.02% with a max depth of 7× (See Table 3).

Table 3 Summary of genome assembly and sequencing data from each of the 10 NDV-positive birds

SEPRL	Raw read pairs	Filtered read pairs[a]	Mean fragment length	Fragment length SD[b]	Forward read quality[c]	Reverse read quality[c]	Number of reads for consensus[d]	Final coverage depth[c]	Consensus nucleotide length	Percent genome coverage
1162	759,149	43,641	127	42	2|34|37|37|38	2|33|36|37|38	828	0|4|6|9|23	14,469	95.24
1163	780,861	7380	121	43	2|35|37|37|38	2|34|36|37|38	585	0|3|4|6|15	13,958	91.87
1164	1,146,188	30,829	133	48	2|34|37|37|38	2|33|36|37|38	5751	0|29|40|55|191	15,149	99.71
1165	1,123,285	49,586	121	36	2|34|37|37|38	2|33|36|37|38	6925	0|39|53|70|131	15,149	99.71
1166	6,808,158	134,809	148	50	2|35|37|38|38	2|33|36|37|38	54,086	0|394|513|634|1057	15,167	99.83
1168	1,581,385	135,333	130	50	2|34|36|37|38	2|32|36|37|38	7464	0|47|59|78|161	15,159	99.78
1170	3,555,593	42,487	139	48	2|35|37|37|38	2|33|36|37|38	4352	0|30|37|47|96	15,163	99.8
1171	920,103	23,922	106	42	2|34|37|37|38	2|32|36|37|38	191	0|0|1|2|7	6840	45.02
1173	646,411	37,909	126	43	2|34|37|37|38	2|33|36|37|38	1129	0|6|9|12|30	15,013	98.82
1174	531,027	16,168	123	40	2|34|37|37|38	2|33|36|37|38	2536	0|14|20|26|58	15,162	99.8

[a]The number of paired reads after filtering out host (*Gallus gallus* 5.0) and internal control (PhiX) reads

[b]*SD* standard deviation

[c]The numbers represent read quality distribution (minimum | lower quartile | median | upper quartile | maximum)

[d]Number of paired reads used for the final consensus sequence form each bird

Phylogenetic analyses

Complete coding sequences analysis

The complete coding sequences of all six genes were obtained from six of the eleven birds and the complete fusion gene coding sequence was obtained from eight of the eleven birds. The deduced amino acid sequence of the fusion cleavage site for all sequences was specific for virulent viruses ($_{113}$RRQKR↓F$_{117}$, with the exception of sample #1168 that had $_{113}$RRQRR↓F$_{117}$) (See Table S3). The nucleotide distances between the studied sequences and GenBank sequences were estimated. Most of the studied sequences were very closely related to each other (99.7% mean identity within them), except one (#1168) that was 1.5% distant. This first group of sequences was most closely related (99.3–99.4%) to sequences from viruses isolated from a variety of species (chickens, pigeons, and peacock) from Pakistan in 2014 and 2015. The sequence that was more genetically distant (#1168) was closely related (99.2–99.4%) to sequences from viruses isolated from varying species (chickens, parakeets, pigeon, duck) in Pakistan in 2015 and 2016. Also, NDV isolates in this study had nucleotide distance of 1.2% from previously reported sub-genotypes VIIi and (See Table S4). To further confirm the evolutionary relationship between the NDV sequences studied here and sequences available in GenBank, phylogenetic analysis using the complete genome concatenated CDS was performed. In the created phylogenetic tree, the isolates studied here expectedly grouped together within sub-genotype VIIi with the viruses that showed highest nucleotide sequence identity to them (Fig. 1a). Taken together, the results from the distance analysis and the phylogenetic tree demonstrated that all sequences obtained from field FFPE tissue sample from diseased chicken in Pakistan in 2015 belong to NDV class II sub-genotype VIIi. In addition, two separate branches of isolates were identified in the tree. To assess the genetic diversity within sub-genotype VIIi, the nucleotide distance between the sequences in these two branches was estimated and they were found to be less closely related (1.2% nucleotide distance).

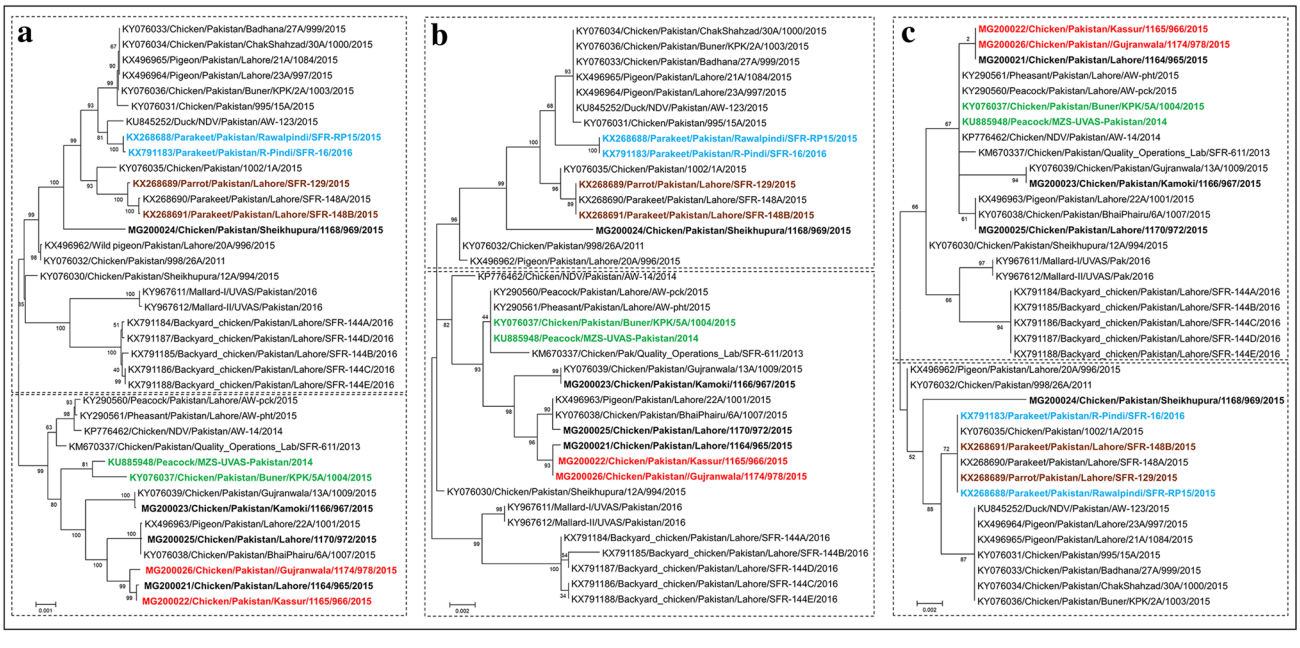

Fig. 1 Phylogenetic analyses based on the complete genome coding sequences (**a**), complete F gene coding sequences (**b**), and partial F gene coding sequences (**c**) of isolates representing class II sub-genotype VIIi Newcastle disease virus. The evolutionary histories were inferred by using the maximum-likelihood method based on General Time Reversible model with 1000 bootstrap replicates as implemented in MEGA 6 [36, 38]. The analyses involved 38 nucleotide sequences with a total of 13,746 (A), 1662 (B), or 375 (C) positions in the final datasets. The sequences obtained in the current study are presented in bold font. Evolutionary analysis was conducted in MEGA6. For all analyses, the codon positions included were 1st + 2nd + 3rd + noncoding, and all positions containing gaps and missing data were eliminated. The GenBank accession numbers are followed by host name, country of isolation, strain designation, and year of isolation. Isolates with bold font in red, brown, green, and blue had 100% nucleotide identity between them based on F (Fig. 1b) and partial F (Fig. 1c) gene trees, were readily differentiated in the complete genome tree, and had nucleotide distances ranging from 0.1 to 0.34%. Isolates in bold black font are from current study

Complete and partial fusion gene coding sequences analysis

The evolutionary history was also inferred by phylogenetic analysis using the smaller datasets (same taxa as in the CDS analysis) comprising the complete F gene coding sequences (Fig. 1b) and the partial F gene coding sequence (Fig. 1c) and the results were compared to those of the complete genome sequences (Fig. 1a). The overall phylogenetic topology based on partial or complete F gene coding sequence analyses was consistent with phylogenetic grouping observed in analysis based on complete genome coding sequences. However, when the complete genome CDS were used for the analysis, higher bootstrap values were observed as compared to those in the partial and complete F gene coding sequences. In addition, sequences that appeared identical in the partial and complete fusion gene sequences analyses (MG200022/chicken/Pakistan/Kassur/1165/966/2015 and MG200026//chicken/Pakistan/Gujranwala/1174/978/2015 [bold and red], KX268688//parakeet/Pakistan/Rawalpindi/SFR-RP15/2015 and KX791183/parakeet/Pakistan/R-Pindi/SFR-16/2016 [bold font and blue], KX268689/parrot/Pakistan/Lahore/SFR-129/2015 and KX268691/parakeet/Pakistan/Lahore/SFR-148A/2015 [bold font and brown], KU885948/peacock/Pakistan/MZS-UVAS/2014 and KY076037/chicken/Pakistan/Buner/KPK/5A/1004/2015 [bold font and green]) (Fig. 1b, C) and had 100% nucleotide identity between them were readily differentiated in the complete genome tree and had nucleotide distances ranging from 0.1 to 0.34%. To specify the genomic regions that readily differentiated NDV isolates that were 100% identical in the fusion protein gene-based phylogeny, a pairwise comparison of nucleotide sequences of all six genes of these isolates (MG200022 vs MG200026, KX268691 vs. KX268689 and KY076037 vs. KU885948) was performed (See Table 4). It was observed that in MG200022 vs MG200026 comparison, phosphoprotein gene contributed highest variation (0.4%) followed by Hemagglutinin (HN), polymerase (L), matrix (M), and nucleoprotein (NP). The HN and L genes contributed most of the variation (0.5%) followed by P gene (0.3%) when a comparison of KX268691 vs. KX268689 was made.

Table 4 Evolutionary distance estimated using the complete coding sequences of individual genes of NDV genomes

Gene ID	MG200022 versus MG200026[a]	KX268691 versus KX268689[b]	KY076037 versus KU885948[c]
Fusion (F)	0.000	0.000	0.000
Partial fusion*	0.000	0.000	0.000
Hemagglutinin (HN)	0.002	0.001	0.005
Polymerase (L)	0.001	0.002	0.005
Matrix (M)	0.001	0.000	0.000
Nucleoprotein (NP)	0.001	0.000	0.000
Phosphoprotein (P)	0.004	0.000	0.003

*374-bp-long sequence of Fusion gene that is frequently used for NDV classification and virulence prediction

[a]NDV isolates from current study (#1165 vs. 1174)

[b,c]NDV genotype VIIi sequences from GenBank

Discussion

The feasibility of using FFPE tissues to sequence NDV complete genomes and to conduct evolutionary and epidemiological studies of closely related NDV-infected field tissue samples has been shown. The phylogenetic analyses confirmed that the detected viruses belonged to the virulent sub-genotype VIIi. Furthermore, this study demonstrates that the phylogenetic results from concatenated complete coding sequences obtained from FFPE tissues provide better resolution compared to individual or partial genes and is sufficient to demonstrate viral evolution that would otherwise remain unnoticed because of the limited genomic fragments analyzed. This is also the first use of NGS to characterize NDV genomes from FFPE tissue samples collected during recent disease outbreaks in commercial poultry. Additionally, some variants of NDV, including pigeon paramyxoviruses (PPMVs), are virulent without the polybasic amino acids at the Fusion protein cleavage site, as some of the PPMVs have shown increased virulence to chickens after only a few passages and without any change in the F gene coding region. In addition, replacing the F gene of the avirulent pigeon-adapted NDV with virulent NDV failed to produce virulent chimeric viruses. These observations underscore the importance of other genes in determining the virulence of NDV [39]. Therefore, these observations highlight the need and utility of complete genome analysis of field isolates.

As an example of the utility of this protocol, evolutionary and phylogenetic analyses were conducted. First, it was confirmed that the detected viruses belonged to the virulent sub-genotype VIIi, which is currently circulating in Pakistan [40–42]. Furthermore, two lineages of sub-genotype VIIi NDV were identified. The phylogenetic analysis showed clear separation of sequences from Pakistan isolated between 2014 and 2016 into two independent branches based on the complete coding sequences. While both groups of viruses were simultaneously circulating in geographically close areas, the genetic distance of 1.4% between them suggests at least 15 years of independent evolution of these two groups [43]. These findings demonstrate no direct, or very distant, epidemiological link between the viruses in these

two groups. For example, they may represent two separate introductions events of NDV into Pakistan that evolved elsewhere or that they evolved locally from a common, unidentified ancestor introduced earlier in the region. Wajid and co-authors have recently reported the simultaneous circulation of sub-genotype VIIi virulent NDV in various poultry and non-poultry avian species in Pakistan based on complete F gene analyses [42]. While no particular interdependence among the hosts affected by ND was observed, the role of non-poultry species in the epidemiology and endemicity of NDV in Pakistan was confirmed.

The investigation of the epidemiologic link between highly related NDV viruses is of vital importance especially in closely located geographical regions where ND has acquired endemicity [6, 23, 24]. A phylogenetic analysis of the complete F gene assisted with evolutionary distances has been proposed previously for accurate classification of NDV genotypes [9]. In this study, the general separation of the sequences into major branches was consistent across the three different phylogenetic analyses. However, a better resolution in terms of higher bootstrap values of highly related NDV isolates were observed in the phylogenetic analysis based on complete genome coding sequences of NDV. In addition, the complete genome CDS analysis allowed the differentiation of viruses, which was otherwise impossible using only the complete or partial fusion gene sequences. These findings suggest the application of paraffin-embedded tissues from outbreaks to track epidemiologically very closely related viruses. Although the utility of NGS to sequence complete genomes of NDV from clinical FFPE tissue samples was described here, the utility of this method in diagnostic settings would require further testing to more rigorously establish sensitivity, specificity, and limit of detection, which is beyond the scope of the current study.

In the evaluation of per-tissue percentage genome assembly of NDV, all four tissue types showed different genome assembly percentages. While the sample size for any given tissue was relatively low, no statistically significant difference was identified in the frequency of positivity between the tissue types. This is consistent with Barbezange and Jestin, which reported that RT-PCR showed no differences in the rate of NDV detection between lung, brain, trachea, and spleen [44].

As described earlier, there was no difference between consensus sequences from different tissues of the same host; however, low sequencing depth and generation of short reads are limitations that currently prevent this approach from being used to study viral quasispecies diversity within clinical FFPE tissue samples from the same host. As the focus of this study was on consensus-level genome analysis, future studies are required to determine if altered protocols (e.g., targeted sequencing, increased depth of sequencing) could be used for the investigation of tissue effect on quasispecies diversity or similar question involving small genetic differences between sequences.

In this study, without any enrichment procedure for viral RNA, partial to complete AAvV-1 genomes coding sequences were assembled from archived FFPE tissues collected in 2015 from chickens during disease outbreaks. Genome assembly was up to 100% of the protein coding sequences and up to 99.81% of the complete genome of AAvV-1. These data suggest the feasibility of using FFPE tissue samples to sequence complete AAvV-1 genome for epidemiological studies of closely related virus isolates. Using FFPE tissues for direct sequencing of AAvV-1 is advantageous due to the fact that they can easily be transported as pathogens are inactivated making them available to transport across countries for research. As formalin fixation of tissues affect nucleic quality, DV200 values (estimated by Bioanalyzer) represent relative amounts of RNA fragments > 200 bp and may be an adequate predictor of RNA quality from FFPE tissue samples and may be evaluated in the future studies involving FFPE tissue samples. FTA® cards have also been used for hazard-free sample handling and evaluated for successful NDV detection [15], provide a method to inactive samples, and are aimed at preserving nucleic acid integrity; however, the use of FTA cards by practicing clinicians, especially for animal pathogens, is currently limited. Formalin-fixed tissues remain a standard sample as it allows for numerous diagnostic assays, which is valuable especially when there is no clinical diagnosis.

Conclusion

In conclusion, using FFPE tissues for direct sequencing of NDV genome is useful because FFPE tissues can be conveniently and affordably transported, due to pathogen inactivation, and because FFPE tissues are the primary means of preserving tissue for routine diagnostic and pathologic testing and for historical archival. This study demonstrates the capability of full-genome epidemiologic investigations in FFPE samples. The use of random sequencing coupled with the absence of any virus enrichment procedure makes this technique likely to be applicable to sequence virus genomes from clinical FFPE tissues in other viral infections. Additionally, the results demonstrate that sub-genotype VIIi viruses are still circulating and evolving in Pakistan after they were first identified in the country in 2011 and support that active epidemiologic surveillance for NDV is needed.

Acknowledgements The authors would like to thank Poonam Sharma, Jeremy D. Volkening, and Dawn Williams-Coplin for their technical support during the study. The mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement

by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Author contributions SLB prepared NGS libraries, performed NGS data analysis, and drafted the manuscript; KMD contributed in NGS data analysis and helped in phylogenetic analysis; JZ conducted IHC; AW, TS, and AB collected clinical tissue samples during disease outbreak and processed for transportation; CCB aided in experimental design and contributed to the drafting of the manuscript; SFR aided in sample collection during disease outbreaks; JBS and CLA designed and oversaw the tissue-based experiments, and aided with analysis of results and with writing of the manuscript.

Funding This work was supported by the U.S. Department of Agriculture ARS CRIS 6040-32000-072 and the U.S. Department of State BEP/CRDF NDV 31063. Dr. Salman Latif Butt is conducting his PhD research program sponsored by the Fulbright U.S. Student Program.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants and/or animals No human subjects were used in this study. This article does not contain any studies with animals performed by any of the authors. Tissues were collected from dead birds being used for diagnostic purposes during disease outbreaks. Sampling was carried out by veterinarian, who took different samples as part of his routine work and under the permission of the farm owner. Since these are diagnostic specimens, sampling did not require the approval of the Ethics Committee.

References

1. Miller PJ, Koch G (2013) Newcastle disease. In: Swayne DE, Glisson JR, McDougald LR, Nolan LK, Suarez DL, Nair V (eds) Diseases of poultry, 13th edn. Wiley-Blackwell, Hoboken, pp 89–138
2. Afonso CL, Amarasinghe GK, Banyai K, Bao Y, Basler CF, Bavari S, Bejerman N, Blasdell KR, Briand FX, Briese T et al (2016) Taxonomy of the order Mononegavirales: update 2016. Arch Virol 161(8):2351–2360
3. Amarasinghe GK, Ceballos NGA, Banyard AC, Basler CF, Bavari S, Bennett AJ, Blasdell KR, Briese T, Bukreyev A, Caì Y (2018) Taxonomy of the order Mononegavirales: update. Arch Virol 2018:1–12
4. Dimitrov KM, Ramey AM, Qiu X, Bahl J, Afonso CL (2016) Temporal, geographic, and host distribution of avian paramyxovirus 1 (Newcastle disease virus). Infect Genet Evol 2016(39):22–34
5. Sabouri F, Vasfi Marandi M, Bashashati M (2018) Characterization of a novel VIII sub-genotype of Newcastle disease virus circulating in Iran. Avian Pathol 47(1):90–99
6. Gowthaman V, Singh SD, Dhama K, Desingu PA, Kumar A, Malik YS, Munir M (2016) Isolation and characterization of genotype XIII Newcastle disease virus from Emu in India. Virus-Disease 27(3):315–318
7. Chambers P, Millar NS, Bingham RW, Emmerson PT (1986) Molecular cloning of complementary DNA to Newcastle disease virus, and nucleotide sequence analysis of the junction between the genes encoding the haemmaglutinin-neuraminidase and the large protein. J Gen Virol 67:475–486
8. de Leeuw OS, Koch G, Hartog L, Ravenshorst N, Peeters BPH (2005) Virulence of Newcastle disease virus is determined by the cleavage site of the fusion protein and by both the stem region and globular head of the haemagglutinin-neuraminidase protein. J Gen Virol 86(Pt. 6):1759–1769
9. Diel DG, da Silva LH, Liu H, Wang Z, Miller PJ, Afonso CL (2012) Genetic diversity of avian paramyxovirus type 1: proposal for a unified nomenclature and classification system of Newcastle disease virus genotypes. Infect Genet Evol 12(8):1770–1779
10. Dimitrov KM, Sharma P, Volkening JD, Goraichuk IV, Wajid A, Rehmani SF, Basharat A, Shittu I, Joannis TM, Miller PJ et al (2017) A robust and cost-effective approach to sequence and analyze complete genomes of small RNA viruses. Virol J 14(1):72
11. Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, Subbu V, Spiro DJ, Sitz J, Koo H, Bolotov P et al (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature 437(7062):1162–1166
12. Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25(4):195–203
13. Wakamatsu N, King DJ, Seal BS, Brown CC (2007) Detection of Newcastle disease virus RNA by reverse transcription-polymerase chain reaction using formalin-fixed, paraffin-embedded tissue and comparison with immunohistochemistry and in situ hybridization. J Vet Diagn Invest 19(4):396–400
14. Klopfleisch R, Weiss AT, Gruber AD (2011) Excavation of a buried treasure–DNA, mRNA, miRNA and protein analysis in formalin fixed, paraffin embedded tissues. Histol Histopathol 26(6):797–810
15. Perozo F, Villegas P, Estevez C, Alvarado I, Purvis LB (2006) Use of FTA® filter paper for the molecular detection of Newcastle disease virus. Avian Pathol 35(02):93–98
16. Wakamatsu N, King D, Kapczynski D, Seal B, Brown C (2006) Experimental pathogenesis for chickens, turkeys, and pigeons of exotic Newcastle disease virus from an outbreak in California during 2002-2003. Vet Pathol 43(6):925–933
17. Boheemen S, Graaf M, Lauber C, Bestebroer TM, Raj VS, Zaki AM, Osterhaus AD, Haagmans BL, Gorbalenya AE, Snijder EJ et al (2012) Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. mBio 3(6):e00473
18. Carrick DM, Mehaffey MG, Sachs MC, Altekruse S, Camalier C, Chuaqui R, Cozen W, Das B, Hernandez BY, Lih CJ et al (2015) Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. PLoS ONE 10(7):e0127353
19. Bodewes R, van Run PR, Schurch AC, Koopmans MP, Osterhaus AD, Baumgartner W, Kuiken T, Smits SL (2015) Virus characterization and discovery in formalin-fixed paraffin-embedded tissues. J Virol Methods 214:54–59
20. Mubemba B, Thompson PN, Odendaal L, Coetzee P, Venter EH (2017) Evaluation of positive Rift Valley fever virus formalin-fixed paraffin embedded samples as a source of sequence data for retrospective phylogenetic analysis. J Virol Methods 243:10–14
21. Xiao YL, Kash JC, Beres SB, Sheng ZM, Musser JM, Taubenberger JK (2013) High-throughput RNA sequencing of a formalin-fixed, paraffin-embedded autopsy lung tissue sample from the 1918 influenza pandemic. J Pathol 229(4):535–545
22. He Y, Taylor TL, Dimitrov KM, Butt SL, Stanton JB, Goraichuk IV, Fenton H, Poulson R, Zhang J, Brown CC (2018) Whole-genome sequencing of genotype VI Newcastle disease viruses from formalin-fixed paraffin-embedded tissues from wild pigeons reveals continuous evolution and previously unrecognized genetic diversity in the US. Virol J 15(1):9
23. Mayahi V, Esmaelizad M (2017) Molecular evolution and epidemiological links study of Newcastle disease virus isolates from 1995 to 2016 in Iran. Arch Virol 162(12):3727–3743

24. Esmaelizad M, Mayahi V, Pashaei M, Goudarzi H (2017) Identification of novel Newcastle disease virus sub-genotype VII-(j) based on the fusion protein. Arch Virol 162(4):971–978
25. Bancroft JD, Gamble M (2008) Theory and practice of histological techniques. Elsevier, Amsterdam
26. Susta L, Miller PJ, Afonso CL, Brown CC (2011) Clinicopathological characterization in poultry of three strains of Newcastle disease virus isolated from recent outbreaks. Vet Pathol 48(2):349–360
27. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data
28. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–12
29. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997
30. Zhang J, Kobert K, Flouri T, Stamatakis A (2013) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30(5):614–620
31. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S (2015) The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research 4
32. Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. In: German conference on bioinformatics. Hanover, Germany, pp 45–56
33. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680
34. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729
35. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 101(30):11030–11035
36. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10(3):512–526
37. Courtney SC, Susta L, Gomez D, Hines NL, Pedersen JC, Brown CC, Miller PJ, Afonso CL (2013) Highly divergent virulent isolates of Newcastle disease virus from the Dominican Republic are members of a new genotype that may have evolved unnoticed for over 2 decades. J Clin Microbiol 51(2):508–517
38. Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol Biol Evol 9(4):678–687
39. Dortmans JC, Koch G, Rottier PJ, Peeters BP (2011) Virulence of Newcastle disease virus: what is known so far? Vet Res 42(1):122
40. Rehmani SF, Wajid A, Bibi T, Nazir B, Mukhtar N, Hussain A, Lone NA, Yaqub T, Afonso CL (2015) Presence of virulent Newcastle disease virus in vaccinated chickens in farms in Pakistan. J Clin Microbiol 53(5):1715–1718
41. Miller PJ, Haddas R, Simanov L, Lublin A, Rehmani SF, Wajid A, Bibi T, Khan TA, Yaqub T, Setiyaningsih S et al (2015) Identification of new sub-genotypes of virulent Newcastle disease virus with potential panzootic features. Infect Genet Evol 29:216–229
42. Wajid A, Dimitrov KM, Wasim M, Rehmani SF, Basharat A, Bibi T, Arif S, Yaqub T, Tayyab M, Ababneh M et al (2017) Repeated isolation of virulent Newcastle disease viruses in poultry and captive non-poultry avian species in Pakistan from 2011 to 2016. Prev Vet Med 142:1–6
43. Dimitrov KM, Lee D-H, Williams-Coplin D, Olivier TL, Miller PJ, Afonso CL (2016) Newcastle disease viruses causing recent outbreaks worldwide show unexpectedly high genetic similarity to historical virulent isolates from the 1940s. J Clin Microbiol 54(5):1228–1235
44. Barbezange C, Jestin V (2002) Development of a RT-nested PCR test detecting pigeon Paramyxovirus-1 directly from organs of infected animals. J Virol Methods 106(2):197–207