# Genesis of a highly pathogenic virus: what can we learn from deep sequencing data?

Alice Fusaro, Gianpiero Zamperin, Adelaide Milani, Alessia Schivo, Annalisa Salviato, Isabella Monne

*Division of Comparative Biomedical Sciences, Istituto Zooprofilattico Sperimentale delle Venezie, Viale dell'Università 10, 35020 Legnaro (PD), Italy*

## The outbreak

Between April 30th and May 16th 2016 Italy reported two highly pathogenic (HP) avian influenza (AI) H7N7 outbreaks in the industrial poultry sector, one in a laying farm (farm-1) and the other in a turkey farm (farm-2). The epidemiologic enquiry indicated a slight increase in mortality rates between the third week of March and mid-April, concurrently with a moderate drop in egg production. This finding, together with the presence of serologically positive birds, led to the hypothesis that the low pathogenic (LP) form of the virus had been introduced into farm-1 through direct contact between wild birds and free-range hens and then rapidly mutated into a HP form within the premises (1).

## The aim

Our objective was to explore the intra-farm evolution of the virus pathogenicity to provide a better understanding of the mechanisms driving the evolution of a LPAI virus into a HP form.


**Fig. 1**. Maximum-likelihood phylogenetic tree of the hemagglutinin (HA) gene of H7 viruses constructed using PhyML 3.1. H7N7 Italian viruses are highlighted in purple (LP) and blue (HP)). Bootstrap supports (100 replicates) >60% are indicated above the nodes.

## Phylogenetic and molecular analyses

- Genetic analysis of the hemagglutinin (HA) gene revealed the co-circulation in farm-1 of both the HP and the LP progenitor virus (**Fig 1**). Three distinct cleavage site motifs have been identified, suggesting a high genetic instability at this site **(Tab. 1**).
- Topologies of the maximum likelihood phylogenetic trees of the eight gene segments showed that the Italian H7N7 LPAI and HPAI viruses clustered together and were separated by long branches from viruses circulating in Eurasia.
- Besides the insertion at the HA cleavage site, we identified at least 10 nucleotide and 6 amino acid differences between the genomes of the HPAI and the LPAI progenitor viruses (**Fig. 2**). These mutations were distributed in the NA, PB2, PB1, PA and M genes, suggesting that in addition to the insertion in the HA cleavage site, mutations in other proteins can play an important role in the evolution of virus virulence.
- Median-joining network analysis of the concatenated HA-NP-NA-NS-M genes of the 2016 Italian H7N7 viruses identified two HPAI and three LPAI genotypes within the first infected farm, confirming the existence of a high genetic variability within the the first infected flock. The high number of median vectors (black dots) identified between the analysed samples revealed missing ancestral sequences from our analyses, which might have been detected with increased sampling (**Fig. 3**).

**Tab. 1**. Amino acid motifs of the HA cleavage sites of the Italian H7N7 LPAI and HPAI samples collected from farm 1 and 2

| Farm | Pathotype | H7N7 cleavage site | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LPAI | P | K | G | - | - | - | R | G | L | F |
| | HPAI | P | K | G | R | **K** | R | R | G | L | F |
| | HPAI | P | K | G | R | E | R | R | G | L | F |
| 2 | HPAI | P | K | **R** | R | E | R | R | G | L | F |

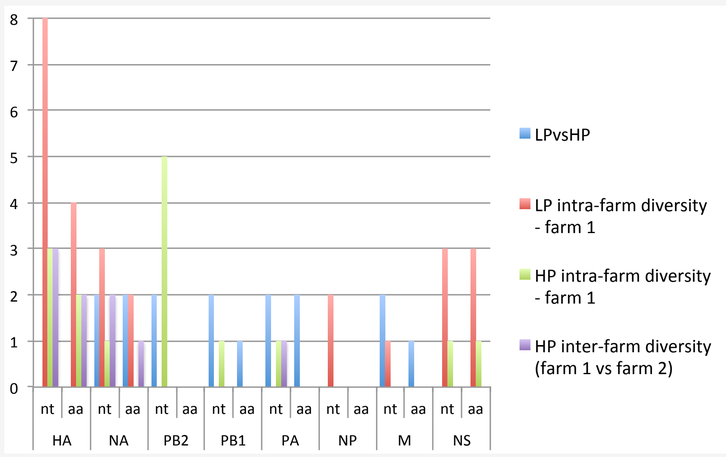


**Fig. 2**. Minimum number of nucleotide (nt) and amino acid (aa) differences identified at the consensus level between i) the LPAI and HPAI viruses (LPvsHP, blue), ii) the LPAI viruses from farm 1 (LP intra-farm diversity – farm 1, red), iii) the HPAI viruses from farm 1 (HP intra-farm diversity – farm 1, green), iv) the HPAI viruses identified in farm 1 and farm 2 (HP inter-farm diversity – farm 1 vs 2, purple)

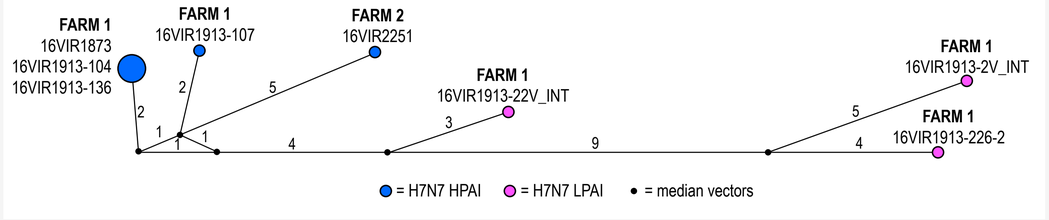


**Fig 3.** Median-joining phylogenetic network constructed from the consensus sequences of the HA-NP-NA-NS-M concatenated gene segments. Each unique sequence genotype is represented by a circle sized relatively to its frequency in the data set. Numbers next to the circles correspond to the samples showing that particular genotype. Branches represent the shortest trees, and numbers above them are the nucleotide mutations that separate each node. Median vectors (lost ancestral sequences) are indicated as black circles.

## Ultra-deep sequencing analyses

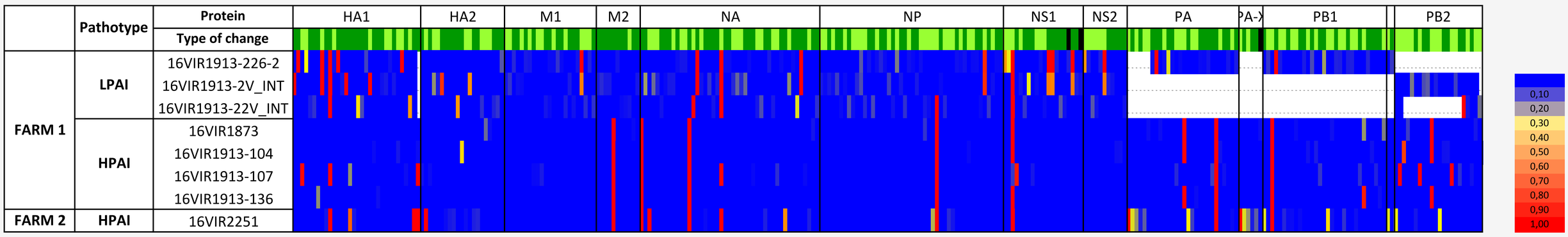


**Fig 4.** Heatmap of the nucleotide frequency. The vertical axis represents the LPAI and HPAI samples collected from farm 1 and 2, while the horizontal axis display only the variable nucleotide positions. The color scale represents the nucleotide frequency according to the scale bar at the right of the figure. White spaces represent positions for which deep-sequencing data were not available (coverage < 1000). The top bar "type of change" indicates the type of nucleotide substitutions: synonimous (light green), nonsynonimous (dark green) and nonsense (black).

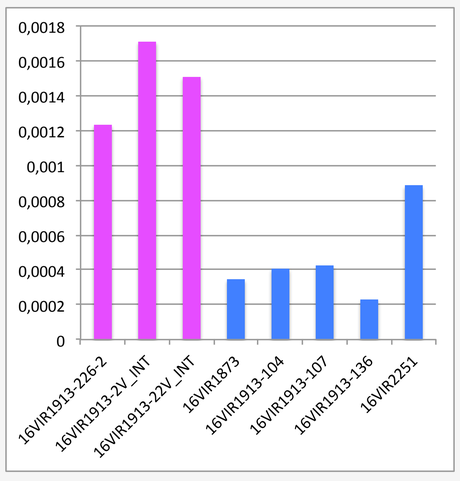


**Fig 5.** Comparison of the Shannon entropy of the complete genome of each sample (purple for LPAI and blue for HPAI viruses). We used one-way ANOVA to determine significant differences between the entropies of each gene for each sample. A *P* value of0.05 was considered significant.

- Overall, we observed 298 sites showing genetic variability distributed among the eight gene segments, of which 164 were non-synonymous, 131 synonymous and 3 nonsense (**Fig. 4**).
- A total of 51 consensus-level nucleotide substitutions (24 non-synonymous and 27 synonymous) were recovered along the entire genome, defining 8 different genomes, 7 of which identified within the first infected farm (**Fig 4**).
- A total of 247 sites showed minority variants in at least one sample with a frequency ranging from 1 to 40%. Only few minority variants (3) were shared among at least two samples.
- The complexity of the viral population (Shannon entropy) of the HP viruses from farm-1 was lower compared to the LP viruses ($p<0.05$), which suggests that the HP viruses had recently experienced a narrow bottleneck (**Fig 5**). This hypothesis is further supported by the lower number of minority variants (47) observed among the HPAI viruses compared to the LPAI (164).

## Conclusion

We demonstrated that a high level of virus genetic heterogeneity within a farm, resulting from the prolonged circulation of a virus in multiple hosts, combined with the occurrence of population bottlenecks, may have caused the emergence of viruses with a high virulence. Hence, the prompt detection of H7 LPAI infection and application of depopulation measures are essential to prevent the emergence and spread of a HPAI strain.

### Methods

***Sequencing:*** Using a deep sequencing approach (Illumina MiSeq), we sequenced 7 H7N7 positive samples collected from symptomatic and asymptomatic animals at farm-1 and one sample from farm 2. We obtained the complete genome of the 5 HPAI viruses (4 HPAI and 1 LPAI), while only partial or no sequences of the gene segments codifying for the polymerase proteins (PB2, PB1 and PA) were obtained for the 3 LPAI viruses.

***Assembly:*** Raw data were filtered by removing: a) reads with more than 10% of undetermined bases; b) reads with more than 50 bases with Q score below 7; c) duplicated paired-end reads. Remaining reads were clipped from Illumina adaptors Nextera XT with scythe v0.991 (2) and trimmed with sickle v1.33 (3). ***Reads*** > 50 bases or unpaired were discarded. High quality reads were aligned against the reference genome using BWA v0.7.5a (4). Alignments were processed with SAMtools v0.1.19 (5) to convert them in BAM format e sort them by position. SNPs were called using LoFreq v0.6.1 (6). From the final set of variants, SNPs called in a position with less than 10X fold coverage or with strand-bias were excluded.

***Phylogenetic analysis:*** phylogenetical analyses of the eight gene segments were performed using the Maximum likelihood method implemented in PhyML 3.1.

***Median-Joining network***: The HA-NP-NA-M-NS gene segments of the influenza virus genome were manually concatenated, and the alignment was used to construct a phylogenetic network using the median joining method implemented in the program NETWORK 4.5 (20). The parameter "epsilon" was set to 0, and the transition/transversion ratio was set to 3:1.

**REFERENCES**:
**1**) Mulatti P, et al.. H7N7 Highly Pathogenic Avian Influenza in Poultry Farms in Italy in 2016. Avian Dis. 2017 Jun;61(2):261-266. **2**) https://github.com/vsbuffalo/scythe; **3**) https://github.com/najoshi/sickle; **4**) Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595; **5**) Li H., et al.(2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9; **6**) Wilm et al. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012; 40(22): 11189-201.

**ACKNOWLEDGEMENTS**: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727922