Routine genomic surveillance on samples from COVID-19 patients collected in Poland during summer 2021 revealed the emergence of a SARS-CoV-2 Delta variant with a large 872 nt deletion. This change, confirmed by Sanger and deep sequencing, causes complete loss of ORF7a, ORF7b, and ORF8 genes. The index case carrying the deletion is unknown. The standard pipeline for sequencing may mask this deletion with a long stretch of N's. Effects of this deletion on phenotype or immune evasion needs further study.

Between 1 July and 16 August 2021, 316 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes from 13 of 16 Polish voivodships from isolates obtained during routine surveillance activities, were deposited by our group in the GISAID database [1]. Of these, 31 sequences, all of which belonged to the SARS-CoV-2 variant of concern (VOC) Delta (Phylogenetic Assignment of Named Global Outbreak (Pango) lineage designation B.1.617.2) AY.4 sublineage designation, showed an unusual pattern i.e. a long stretch of N's, indicating a weak or undetermined nucleotide signal, in the region corresponding to the open reading frame (ORF)7a, ORF7b and ORF8 genes in the reference SARS-CoV-2 genome (GenBank NC_045512).

Most of these unique sequences (27/31) originated from coronavirus disease (COVID-19) patient samples collected in the Malopolska voivodship in southern Poland, between 17 July and 11 August 2021. Our aim was to investigate the emergence of the unusual sequence pattern in SARS-CoV-2 sequences from Poland.

Clinical presentation
Clinical data were available for 24 of the 31 patients with atypical Delta VOC sequences. COVID-19 was suspected in all patients, since they showed clinical symptoms compatible with the disease; all were confirmed by a SARS-CoV-2 RT-qPCR diagnostic test, referred by a general practitioner. The median age was 31 years (range: 4 months–84 years), with an equal number of females (n = 12) and males (n = 12). Symptoms included fever (n = 23), cough (n = 15), weakness (n = 10), muscle pain (n = 10), runny nose (n = 7), loss of smell (n = 3), loss of taste (n = 2), sore throat (n = 1), diarrhoea (n = 1), dyspnoea (n = 1) and breathing difficulties (n = 1). Seven patients were hospitalised, and four needed oxygen supplementation. Fifteen of the 24 patients were not vaccinated or partially vaccinated (one dose, various mRNA and non-mRNA vaccines) against SARS-CoV-2; of the seven hospitalised patients, five were not vaccinated. None of the patients travelled abroad before the COVID-19 symptoms developed.
ARTICv3 and Midnight protocols are whole genome sequencing procedures based on amplification of specific SARS-CoV-2 genome regions with the use of sets of primers.

**PCR and sequencing analysis**

For sample collection, nasopharyngeal swab samples were taken according to the WHO guidelines and in line with the protocols approved by the Polish Ministry of Health [2]. RNA was extracted by the various laboratories, accordingly to their validated, routinely performed methods used for SARS-CoV-2 diagnostic. Extracted RNA was selected randomly, was frozen at −80°C and sent from collection to sequencing laboratories within 2 weeks. The sequencing was performed with Oxford Nanopore Technologies (ONT) platform (Oxford, United Kingdom) using the ARTIC amplicon v3 and Midnight whole genome sequencing protocols [3-5]. Sequencing results showed a long stretch of N’s within the region covering the ORF7a, ORF7b and ORF8 genes. Since the technique is based on the targeted amplification, this observed artefact i.e. a long stretch of N’s, was suspected to be the result of a short insertion–deletion mutation (indel) or single nucleotide polymorphism (SNP) within the primer target sites. We selected 14 specimens for further genetic analyses. To exclude an unspecific primer binding event, we tested the primer pair (e.g. 90_Left and 93_Right primers, which was used for the ARTICv3 protocol and overlapped the region of interest) for SARS-CoV-2 genome amplification (Table 1).

| Pair | Sequencing protocol | Primer name | Primer sequence (5’ to 3’) |
|------|---------------------|-------------|---------------------------|
| Pair 1 | ARTICv3 | 90_Left | ACACAGACATTCACTAGGAGT |
| | | 93_Right | AGGTCTTCTTGCCAATGTTGAG |
| Pair 2 | Midnight | 27_Left | TGGATCACCAGTGGAAATGCTA |
| | | 28_Right | GTTTGGCCTTGTGTGTGTGC |

**Rapid barcoding error exclusion**

Such methods of high-throughput whole genome sequencing like ONT of SARS-CoV-2 are usually performed in a 96-well plate format using the barcoding approach, which is time- and cost-effective, but may lead to sporadic misclassification of some reads between the samples. This mistake could be wrongly reported as 'quasi-species'. To exclude artefacts related to the barcoding procedure and to detect SARS-CoV-2 quasi-species, selected samples were run individually on a single flow cell. Moreover, during our analyses, we used a highly accurate ONT Guppy 5.0.11 super high accuracy (SUP) base-calling algorithm supported by detecting a barcode from both ends. None of the reads were mapped to the deleted region, which supported the hypothesis that isolates containing the 872 nt deletion are not mixed with the wild-type virus, thus representing a pure population of these new DeltaAY.4 isolates with the 872 nt deletion.

**Viral load**

The viral load, expressed as a mean cycle threshold (CT) value obtained during diagnostic RT-qPCR, was available for 30 samples for three different commonly used gene targets (E: envelope gene, N: nucleoprotein gene, ORF1a: ORF1a gene). To assess the potential influence of the 872 nt deletion on the virus transmissibility, the CT was compared with selected other Delta variantAY.4 lineage samples (n = 34) confirmed in Poland between 15 July and 16 August 2021 (Figure 3). The viral loads did not vary between the groups, and 28,256 of the reference genome. This sizable alteration causes a complete loss of ORF7a, ORF7b, and ORF8 genes (Figure 2). The obtained sequences were deposited in the GISAID database under the following accession numbers: hCoV-19/Poland/PL_P3686/2021; hCoV-19/Poland/PL_P3688/2021; hCoV-19/Poland/PL_P3693/2021; hCoV-19/Poland/PL_P3695/2021; hCoV-19/Poland/PL_P3696/2021; hCoV-19/Poland/PL_P3699/2021; hCoV-19/Poland/PL_P3701/2021; hCoV-19/Poland/PL_P3704/2021; hCoV-19/Poland/PL_P3705/2021; hCoV-19/Poland/PL_P3709/2021; hCoV-19/Poland/PL_P3711/2021; hCoV-19/Poland/PL_P3794/2021; hCoV-19/Poland/PL_P3804/2021; hCoV-19/Poland/PL_P3687/2021; hCoV-19/Poland/PL_P3710/2021.

**SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.**

ARTICv3 [4,5] and Midnight protocols [3] are whole genome sequencing procedures based on amplification of specific SARS-CoV-2 genome regions.
thus it may be assumed that these Delta AY.4 isolates with the 872 nt deletion show similar transmissibility to the parental AY.4 lineage.

**Ethical statement**
No ethical approval was required for this study as samples were collected for routine surveillance.

**Discussion and conclusions**
SARS-CoV-2 VOCs have appeared in Poland in a pattern similar to that observed in other European countries. From July 2021, the Delta VOC has been the dominant circulating variant. Similar to other RNA viruses, the SARS-CoV-2 genome evolves quite rapidly, indicated by a $1.1 \times 10^3$ substitution/site/year mutation rate [6]. Numerous changes within SARS-CoV-2 genomes are continuously reported, from minor SNPs to extensive deletions, which have also been reported for other coronaviruses e.g. Middle East respiratory syndrome coronavirus (MERS-CoV) [7]. Regardless of size, all variations may impair the protein structure and function, which could cause changes in the virus phenotype, infectivity, infectiousness, illness severity, or host immune response. The detected 872 nt deletion leads to complete loss of three accessory proteins encoded by the **ORF7a**, **ORF7b**, and **ORF8** genes.

The gene encoding accessory protein ORF8 shows high variability, and multiple alterations including SNPs, short indels causing frameshifts and partial or complete gene deletions have been reported to date [8-10]. In recombinant SARS-CoV-1, the truncation of ORF8...
led to gradual virus attenuation in vitro [7]. In SARS-CoV-2, the accessory protein encoded by ORF8 has a function related to evasion of the host adaptive immune response via downregulation of MHC-1 (major histocompatibility complex) [11]. Moreover, ORF8 modulates the host’s interferon-mediated antiviral response [12]. However, the importance of ORF8 in vivo remains unresolved. While recombinant ΔORF8 SARS-CoV-2 viruses produced smaller plaques than the wild type (WT) in vitro, the viral growth kinetics remained unchanged [10,13]. Furthermore, human angiotensin converting enzyme 2 (hACE2)-transgenic mice infected with ΔORF8 recombinants showed similar pathological lesions and mortality to the WT strain, suggesting that ORF8 does not contribute to virus pathogenicity [13]. An observational human cohort study revealed that individuals infected with SARS-CoV-2 lacking a functional ORF8 gene was associated with lower odds of developing hypoxia requiring supplemental oxygen [14].

The deletions within the ORF7a and ORF7b region have also previously been reported [15-18]. ORF7a activates the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kappa B) pathway and induces proinflammatory cytokine expression [19,20]. The function of ORF7b is largely unknown, but a recent study has shown that the protein activates the type-I interferon (IFN) signalling pathway and promotes expression of IFN-beta, interleukin (IL)-6, and tumor necrosis factor (TNF)-alpha, which induce apoptosis [21]. Viruses lacking these genes also produced smaller plaques than WT viruses in vitro, but their growth kinetics were similar to the parental strain [13]. Based on studies in animal models, the proteins have a minor impact on pathological lesions and disease outcomes [13]. This is the first study, to our knowledge, confirming the emergence of such a large deletion within the SARS-CoV-2 genome, causing the complete loss of three consecutive ORF sequences, e.g. ORF7a, ORF7b, and ORF8. The clinical data obtained from 25 infected individuals suggest a typical course of COVID-19, with mild to moderate symptoms. The origin of these Delta AY.4 isolates with the 872 nt deletion remains unknown, but their detection in a short period of time with high frequency in the Malopolska voivodship suggests that they may be maintained in the population. Based on GISAID data, during recent weeks (September 2021), more than 70 similar sequences containing a stretch of N’s within the region corresponding to ORF7a, ORF7b and ORF 8 were confirmed in other areas in Poland, including Mazowieckie, Pomorskie and Swietokrzyskie voivodships. These sequences, similar to those presented here, could also contain the 872 nt deletion and thus, detection could be missed because of the bioinformatic algorithm leading to a long stretch of N’s at the position of the deleted genes. However, a re-analysis of raw sequencing data or Sanger sequencing is needed to confirm the deletion in these samples. Considering the potential spread of a virus strain similar to these Delta AY.4 isolates with the 872 nt deletion to the other regions in Poland, it may be hypothesised that its transmissibility has not been compromised. Unfortunately, the original swab samples were unavailable and the virus isolation was impossible at this point, but will be a subject for the future studies. The monitoring of this 872 nt deletion in clinical samples and the publicly available databases should be continued.
Figure 3
Comparison of viral load between the SARS-CoV-2 Delta AY.4 lineage (n = 34) and the Delta AY.4 isolates with the 872 nt deletion (n = 30), Poland, 17 July–11 August 2021

CT: cycle threshold; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

Each point refers to single sample, for which the CT value was calculated as a mean CT obtained from three target genes (E, N, ORF1a) obtained during routine diagnostic RT-qPCR. CT values are compared between SARS-CoV-2 AY.4 samples (n = 34) and Delta AY.4 isolates with the 872 nt deletion (n = 30). Error bars indicate standard deviation of the mean. The difference is statistically insignificant (p > 0.05, unpaired t-test).

Acknowledgements
We would like to thank Provincial Sanitary and Epidemiological Station in Krakow for their excellent cooperation and clinical data disclosure.

Funding statement: This work was supported by the subsidy from the Polish Ministry of Science and Higher Education for the research on the SARS-CoV-2 and the grant 2020/ABM/104525. This work was supported by the subsidy from the Polish Ministry of Science and Higher Education for the research on the SARS-CoV-2 and the grant 2020/ABM/104525.

Conflict of interest
None declared.

Authors’ contributions
Study design: NMP, LR, KP. Data analysis: NMP, LR, KP. Experimental procedures: NMP, LR, TG, NG, AP, KM, KS, KS, JG, PL. Manuscript preparation: NMP, LR, KP. Manuscript revision: all authors.

References
1. Global Initiative on Sharing All Influenza Data (GISAID). Munich: GISAID. [Accessed: 15 Sep 2021]. Available from: https://https://www.eurosurveillance.org

www.eurosurveillance.org
for human CD14+ monocytes. iScience. 2021;24(3):102187. https://doi.org/10.1016/j.isci.2021.102187 PMID: 33615195

21. Yang R, Zhao Q, Rao J, Zeng F, Yuan S, Li M, et al. SARS-CoV-2 Accessory protein ORF7b mediates Tumor Necrosis Factor-α-induced apoptosis in cells. Front Microbiol. 2021;12:654709. https://doi.org/10.3389/fmicb.2021.654709 PMID: 34484133

License, supplementary material and copyright

This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) Licence. You may share and adapt the material, but must give appropriate credit to the source, provide a link to the licence and indicate if changes were made.

Any supplementary material referenced in the article can be found in the online version.

This article is copyright of the authors or their affiliated institutions, 2021.