Research Article

Interpatient Mutational Spectrum of Human Coronavirus-OC43 Revealed by Illumina Sequencing†

Geoffrey J. Gorse[a,b*], Gira B. Patel[a,b], and Xiaofeng Fan[c]

[a] VA St. Louis Health Care System, St. Louis, MO, USA
[b] Division of Infectious Diseases, Allergy and Immunology, Saint Louis University School of Medicine, St. Louis, MO, USA
[c] Division of Gastroenterology and Hepatology, Saint Louis University School of Medicine, St. Louis, MO, USA

* Corresponding Author:
Geoffrey J. Gorse, MD
Division of Infectious Diseases, Allergy and Immunology, Saint Louis University School of Medicine
1100 South Grand Blvd. (DRC 8th Floor)
Saint Louis, MO, USA
Telephone: 314-977-5500
Fax: 314-771-3816
E-Mail Address: gorsegj@slu.edu

Shortened Title/Running Head: Human Coronavirus Genetic Mutations

†This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/jmv.24780]

Additional Supporting Information may be found in the online version of this article.

Received 10 August 2016; Revised 20 December 2016; Accepted 3 January 2017
Journal of Medical Virology

DOI 10.1002/jmv.24780

Abstract

Human coronaviruses (HCoV) are RNA viruses that cause respiratory tract infections with viral replication of limited duration. The host and viral population heterogeneity could influence clinical phenotypes. Employing long RT-PCR with Illumina sequencing, we quantified the gene mutation load at 0.5% mutation frequency for the 4,529 bp-domain spanning the Spike gene (4,086 bp) of HCoV-OC43 in four upper respiratory clinical specimens obtained during acute illness. There were a total of 121 mutations for all four HCoV samples with the average number of mutations at 30.3 ± 10.2, which is significantly higher than that expected from the Illumina sequencing error rate. There were two mutation peaks, one at the 5′ end and the other near position 1550 in the S1 subunit. Two coronavirus samples were genotype B and two were genotype D, clustering with HCoV-OC43 strain AY391777 in neighbor – joining tree phylogenetic analysis. Nonsynonymous mutations were 76.1 ± 14% of mutation load. Although lower than other RNA viruses such as hepatitis C virus, HCoV-OC43 did exhibit quasi-species. The rate of nonsynonymous mutations was higher in the HCoV-OC43 isolates than in hepatitis C virus genotype 1a isolates analyzed for comparison in this study. These characteristics of HCoV-OC43 may affect viral replication dynamics, receptor binding, antigenicity, evolution, transmission, and clinical illness.

Keywords: Coronavirus; Hepatitis C virus; humoral immunity; genetic variability; genetic variation; mutation

1. Introduction

Human coronavirus-OC43 (HCoV-OC43) is within the *Betacoronavirus* genus (family *Coronaviridae*) and is an enveloped, positive sense, single-stranded RNA virus [Belouzard et al., 2012; Bosch et al., 2003]. There are five known genotypes and other betacoronaviruses include severe acute respiratory syndrome (SARS) CoV and Middle East respiratory syndrome (MERS) CoV [Ren et al., 2015]. HCoV-OC43 is prevalent among humans and genotype D has been prominent in recent years [Lau et al., 2011; Ren et al., 2015; Zhang et al., 2015]. Little is known about how HCoV-OC43 genotypes persist in human populations, but continuous adaptation by viral antigenic genes in the Spike protein through genetic drift may be necessary. The Spike protein is the major antigenic protein and is under selection pressure by the host immune response; it is important for host range and tissue tropism. It is cleaved into S1 and S2 subunits for receptor binding and membrane fusion. The N-terminal domain of the S1 subunit is responsible for sugar receptor binding and the S2 subunit is responsible for fusion of viral and host membranes [Li, 2016]. The S1 subunit is more divergent in sequence and the S2 subunit is more conserved [Belouzard et al., 2012; Masters and Perlman 2013; Ren, 2015].

Human coronaviruses cause the common cold and influenza-like illnesses, but can be associated with more severe illnesses such as pneumonia, exacerbations of asthma and chronic obstructive pulmonary disease, croup and bronchiolitis. In patients with chronic obstructive pulmonary disease studied during the 1998 to 1999 influenza season, 13.5% of illnesses were associated with HCoV-229E and HCoV-OC43 infection, while in another study between 2009 and 2013, 19% of acute respiratory illnesses in patients with cardiopulmonary diseases and 21.5% in healthy young adults were associated with HCoV [Gorse et al., 2003, 2006, 2009, 2015]. Coronavirus-associated illness was less severe than influenza but was associated with multiple respiratory and systemic symptoms, and hospitalization [Gorse et al., 2009]. HCoV-229E and HCoV-OC43 infection rates of 2.8% to 26% in healthy young and elderly

adults, high-risk adults, and hospitalized patients were reported during the winters of 1999 to 2003 and they contributed to medical disease burden [Walsh et al., 2013].

Little is known about the degree of heterogeneity of HCoV-OC43 viral quasi-species present in upper respiratory secretions. If present, this may help explain persistent incidence of HCoV-OC43 infections in human populations, if the mutational changes result in antigenic drift. This might allow escape from host immunity and contribute to virus infectivity and pathogenicity.

In the current study, we combined RT-PCR and Illumina sequencing to measure the diversity of HCoV-OC43 Spike gene quasi-species through direct count of the Spike gene mutations, determination of percent nonsynonymous mutation rates and comparison of these rates to hepatitis C virus (HCV), which is in the genus *Hepacivirus*, family *Flaviviridae*. HCV is an RNA virus with a heterogeneous population of quasi-species in chronically infected patients [Fan et al., 2009, 2010; Wang et al., 2014].

2. Material and Methods

2.1 Patient Samples

We studied nasal and oropharyngeal swab specimens that were obtained from each of four patients early during symptomatic acute respiratory illness and positive for HCoV-OC43 nucleic acids by multiplex RT-PCR [Gorse et al., 2015]. Serum and nasal wash specimens were obtained at the time of acute illness and 3 to 4 weeks after illness onset. They were assayed by enzyme-linked immunosorbent assay for serum IgG and nasal wash IgA antibodies to tissue culture-adapted HCoV-OC43 (American Type Culture Collection #VR-1558, GenBank: NC_005147.1) that was inactivated by psoralen compound and long-wavelength ultraviolet light, as described [Gorse, 2009, 2010, 2015]. Severity of acute respiratory illness was measured by two scores: a self-reported visual analogue scale of overall illness severity, ranging from 1 (mildest) to 10 (most severe), and a severity of influenza-like symptoms and signs score that was the sum of 16 symptoms and signs that were graded on a scale of 0

(absent) to 15 (most severe) with a maximum score of 240, as described [Arden et al., 1988; Gorse et al., 2009, 2015]. Respiratory and systemic symptoms of the acute illness were recorded. The patients gave written informed consent and the study was approved by the Institutional Review Boards at the VA St. Louis Health Care System and Saint Louis University. Two recombinant clones from a previous study, #1701 and #1709, each containing a 9,022 bp HCV insert, were used to estimate potential errors associated with Illumina sequencing [Fan and DiBisceglie, 2010; Wang et al., 2014]. Also, 19 HCV genotype 1a samples from an earlier report [Ren et al., 2015] were available for re-analysis and comparison in the current study.

2.2 RNA Extraction, RT-PCR and Illumina Sequencing

Total RNA from each nasal and oropharyngeal swab specimen sample was purified using the QIAamp Ultrasens Virus Kit (Qiagen, Valencia, CA) according to the manufacturer's procedures. RT-PCR was then applied to amplify a 4,529 bp amplicon spanning the full-length spike gene (4,086 bp). In brief, 10.6 µL of extracted RNA was mixed with 9.4 µL RT matrix consisting of 1x SuperScript III buffer, 10 mM DTT, 1 µM OC43R1 (reverse primer, 5′-TGC CCC ACA TAC CAC ACA G-3′, position 28164-28182, numbering is according to HCoV-OC43 strain, GenBank accession number: AY391777), 2mM dNTPs, 20 U of RNase OUT recombinant Ribonuclease Inhibitor, and 200 U of SuperScript III Reverse transcriptase (Life Technologies). After 75-min. incubation at 50ºC and subsequent inactivation, an aliquot of 5µL of RT reaction was applied for the first round of PCR that contained 1x GC enhancer (New England Biolabs), 1x Q5 buffer (New England Biolabs), 1.6 mM dNTPs, 0.4 µM OC43F1 (forward primer, 5′-GTA CAG GTT GTT GAT TCG CG-3′, position 23210-23229), 0.4 µM OC43R1 and 1.6 U Q5 High Fidelity DNA Polymerase (New England Biolabs). After initial heating at 94ºC for 1 min., cycle parameters were programmed as the first 10 cycles of 94ºC for 30 sec., 65ºC for 30 sec. and 68ºC for 5 min. followed by 20 cycles of 94ºC for 30 sec., 60ºC for 30 sec. and 68ºC for 5 min. with a 2 sec. autoextension at each cycle. Two µL of the first round of PCR product was used for

the second round amplification with primers OC43F2 (forward primer, 5′-TCT GGC CTC TCT ACC CCT ATG GC-3′, position 23439-23461) and OC43R2 (reverse primer, 5′-CTT GAT TAC GGC ACC AAG CAT GAC-3′, position 27944-27967), under the same cycle parameters as the first round of PCR. Product at expected size was gel-purified using QIAquick PCR purification Kit (Qiagen) and quantitated. About 4 to 5 µg of purified DNA product was subjected to library construction. The fragment library was constructed using Illumina Nextera XT DNA library preparation kit, and followed by Illumina sequencing on NextSeq 500 machine with 1 x 250 bp read output.

2.3 Sequence Data Analysis

We first estimated the error rate associated with Illumina sequencing using two recombinant HCV clones. In doing so, raw sequence reads in fastq format were first filtered in PRINSEQ (v0.19.5) for quality control, including read length ≥70 bp, mean read quality score ≥25, low complexity with DUST score ≤7, ambiguous bases ≤1% and all duplicates [Schmieder and Edwards, 2011]. Filtered reads were mapped onto HCV genotype 1a prototype strain H77 (GenBank accession number AY009606) using a gapped aligner Bowtie 2 [Langmead and Salzberg, 2012]. Mapped files were then converted into binary format (BAM), sorted and indexed in SAMtools [Li et al., 2009] followed by local realignment and base quality recalibration in Genome Analysis Toolkit (GATK) [DePristo et al., 2011]. Next, by converting post-alignment BAM files into mpileup format in SAMtools, the consensus sequence for each clone was called in VarScan (v2.2.3) with the settings of ≥1,000x coverage, ≥25 base quality at a position to count a read and ≥50% mutation frequency [Koboldt et al., 2012; Quinlan and Hall, 2010]. The entire pipeline was repeated using individual consensus sequences. Mutations were called at each position in VarScan under the setting of 0.5% frequency and base quality from 15 to 40, followed by manual check in the Integrative Genomics Viewer [Koboldt, 2012].

Using the value of base quality to define a mutation from above analysis, similar procedures were applied to four patient samples. The HCoV-OC43 strain (GenBank AY391777) was used as the

reference at initial mapping. Over the entire coronavirus Spike gene, the mutation load, the total number of mutations at a given site, was counted through sliding windows, size = 300 bp, overlap = 100 bp. Finally, under the frame of full-length HCoV Spike gene (4,086 bp), the nature of each mutation, either synonymous or nonsynonymous, was determined using a custom script [Van Belleghem et al., 2012].

2.4 Phylogenetic Analysis

The consensus full-length HCoV spike sequences from four patients and reference sequences retrieved from GenBank were used for phylogenetic analysis. The tree was constructed using neighbor-joining approach under nucleotide substitution model of maximum composite likelihood in MEGA program (version 5.2) [Tamura et al., 2011].

2.5 Statistical Analysis

Statistical analyses were done with either two-tailed, unpaired Students test or Chi-square. When applicable, data were expressed as mean value and standard deviation. $P<0.05$ was considered statistically significant.

2.6 Data Availability

Raw sequence data in fastq format from all four patient samples were archived in NCBI Sequence Read Archive (SRA) under SRA accession number SRP071020.

3. Results

3.1 Clinical Characteristics of HCoV Infections and Antibody Responses.

Samples 3 and 4, both genotype D, were collected within a month of each other in December 2010 and January 2011 from two older patients with significant acute respiratory and systemic symptoms (Table

1). The patients had underlying chronic cardiopulmonary diseases and diabetes mellitus. The two illnesses were associated with greater than a four-fold increase in nasal wash IgA antibody titers but only one with at least a four-fold increase in serum IgG antibody titer to HCoV-OC43, comparing acute illness to convalescent specimens collected 3 to 4 weeks after illness onset (Table 1).

Samples from subjects 2 and 6, both genotype B, were collected about two years apart in January 2010 and March 2012 from a younger patient without underlying chronic illnesses and an older patient with cardiac disease and diabetes mellitus. Both had acute respiratory and systemic symptoms that may have been less severe than those reported by the two patients with genotype D isolates (Table 1). One of the two illnesses with genotype B viruses was associated with a greater than four-fold increase in nasal wash IgA antibody titer to HCoV-OC43, but neither had a four-fold rise in serum IgG antibody titer to HCoV-OC43, comparing acute illness to convalescent specimens collected 3-4 weeks after illness onset (Table 1).

3.2 Quantitation of HCoV-OC43 Mutation Load

The raw data output indicated 70.1% of bases read had a quality score greater than 30. Interpretation of the distribution statistics of base quality scores over read length resulted in trimming the read length at the 3′ end by 6 to 10%. The final results of the quality control are shown in Supplemental Table. The large output gave a very deep base coverage for each HCoV sample, the average was 94,899 ± 21,405 (Supplemental Fig. 1).

To estimate the error rate associated with library construction and Illumina sequencing, mutations were called from two recombinant HCV clones, #1701 and #1709, under a range of base quality settings. Even if a mutation was counted under the base quality as low as 15, there were no differences in the consensus sequences derived either from Illumina or from gene-walking Sanger sequencing (data not shown). However, the number of individual mutations had a sharp drop from the base quality 25 to 30

(Supplemental Fig. 2). Under the conditions of 0.5% mutation frequency and base quality score of ≥30, there were a total of 31 mutations in the two HCV clone samples, suggesting an error rate of about 1.76 mutations per kb.

Applying the same criteria for the coronavirus samples, a total of 121 mutations for all four samples were identified with an average number of mutations of 30.3 ± 10.2 (range: 20 to 40 mutations per sample), which is significantly higher than that expected from the Illumina sequencing error rate (121 vs. 28.76 mutations, $P = 4.2 \times 10^{-14}$). Nonsynonymous mutations accounted for between 61% and 90% of the total mutations (Table 2). No deletions or recombinations were detected. Of the 121 viral mutations, those with frequencies greater than 2% occurred at six positions: Spike gene position numbers 79, 81, 1229, 1859, 2244 and 2858 (Table 3).

Using a sliding window analysis with window size = 300 bp and overlaps = 100 bp, there were two mutation peaks, one at the 5′ end and the other at about Spike gene position number 1550 (Figure 1). Using consensus sequences, a Neighbor-joining tree showed phylogenetically that two of the subjects' samples were genotype D and two were genotype B strains of HCoV-OC43 (Figure 2 A and Figure 2 B). The two genotype B samples both had high mutation frequencies at spike gene nucleotide positions 79 (36.92% and 45.12%) and 81 (7.96% and 7.07%), whereas the two genotype D samples had mutation frequencies of 0.92% and 0.82% at position 81, and less than 0.5% for position 79.

3.3 Comparison of HCoV and HCV.

A previously reported study found that HCV genotype 1a patients experiencing relapse after antiviral treatment (n=19) had a higher average total mutation load measured through 454 sequencing [Ren et al., 2016]. These samples were re-analyzed for the current study using a base quality score ≥ 30 rather than 25 in the earlier report [Ren et al., 2016]. The average mutation load in HCV patient isolates was significantly higher than in HCoV-OC43 patient isolates (296.2 ± 102.2 vs. 30.3 ± 10.2, $P = 7.7 \times 10^{-5}$).

However, nonsynonymous mutations as a percentage of the total mutations were higher among the HCoV-OC43 isolates than among the 19 HCV genotype 1a patient isolates (76.7 ± 14% vs. 26 ± 8%, $P = 3.5 \times 10^{-9}$).

3.4 Consensus Amino Acid Sequences for HCoV-OC43.

The alignment of Spike genes of consensus amino acid sequences show differences between the four clinical strains and the prototype AY391777 strain particularly at the N and C terminal ends of the S1 subunit, and at the S1 and S2 subunit cleavage site, (a.a. 762-766) with a smaller number of amino acid differences in the S2 subunit (Figure 3). The high genetic mutation rates at sites 79 and 81 correspond to several amino acid changes compared to the prototype strain between amino acid positions 20 and 31.

4. Discussion

Through Illumina sequencing, we have introduced high-resolution HCoV-OC43 Spike gene mutational load to quantify viral quasi-species population diversity. The experimental method allows the role of HCoV-OC43 Spike gene heterogeneity to be investigated in a manner not previously reported. The HCoV-OC43 Spike gene does have quasi-species, but the magnitude is almost 10 times lower than the quasi-species found in HCV, while the nonsynonymous mutations take a higher percentage of the total mutation load for HCoV-OC43 than HCV genotype 1a. It should be noted that the Spike gene was amplified with the gene-specific primers. Although these primers are located in the conserved HCoV HE and NS2 domains, the missing of potentially heterogeneous viral variants during the amplification cannot be excluded. As a consequence, quasi-species diversity of HCoV may be underestimated in the current study in comparison to the use of degenerate primers or primer-independent approaches. Viral replicative dynamics, population size, and host immune responses may contribute to this observation, and the finding has implications in terms of HCoV evolution and treatment. High mutation load

facilitates the evolution of viral populations in response to external pressure. Given their slightly deleterious nature [Eyre-Walker et al., 2002]; however, excessive nonsynonymous mutations can be detrimental to such adaptive evolution [Peck, 1994]. Consequently, accumulation of deleterious mutation load during viral infection may contribute to self-limited active infection. In the case of HCoV-OC43, this could contribute to short, self-limited respiratory tract infections compared to HCV, which is a chronic, systemic viral infection with ongoing viral replication. However, genetic diversity manifested by a cloud of quasi-species in some viral systems can be linked to pathogenicity and shown to allow adaptation to new environments such as infection of new hosts [Vignuzzi et al., 2006]. If clonal virus populations are present in human infections with Middle East Respiratory Syndrome (MERS)-CoV [Briese et al., 2014; Cotten et. al., 2013], this could increase the chances of epidemic spread of a particularly virulent strain [Gardner and McIntyre, 2014]. Infections with a cloud of viral strains may enhance spread, for instance, by increasing the chance of a virulent strain being transmitted, or reduce it due to defective and less virulent strains being present. MERS-CoV genome sequences from dromedary camels indicate presence of quasi-species in single samples [Briese et al., 2014]. Recent deep sequencing analyses reported intra-patient viral heterogeneity during a 2015 outbreak of MERS-CoV and the possibility that host immune response provided selection pressure to favor genetic heterogeneity [Park et al., 2016; Kim et al., 2016]. Intra-patient viral heterogeneity may have contributed to transmissibility [Park et al., 2016].

Our four clinical isolates were of the genotypes B and D which have been described as circulating in recent years rather than genotype A [Lau et al., 2011]. Alignment of spike gene consensus amino acid sequences compared to the AY391777 prototype strain sequence (genotype A) indicates the majority of changes were in the S1 subunit. The proteolytic cleavage site of S1 and S2 subunits for the four clinical strains in our study had the RRSRR motif rather than the RRSRG motif of the prototype strain. This G

to R substitution at amino acid 766 may result in increased cleavability and the cleavage process may play a part in fusion activity and viral infectivity [Lau et al., 2011; Vijgen et al., 2005].

There were minimal amino acid changes at Spike gene nucleotide positions 448 to 459 (whole genome positions 24091-24102) in the glomerular part of the Spike gene, encoding the TQDG (a.a. positions 150-153), and low mutation frequencies in that region for our four clinical isolates, although all four had a Y154V substitution. This is in the lectin domain of the S1 subunit involved in attachment of the virus to the cellular receptor which is a derivative of neuraminic acid [Kin et al., 2015]. The TQDG sequence may have evolutionary and functional importance, is not present in the closely related bovine coronaviruses and not inserted in some strains of HCoV-OC43 [Kin et al., 2015]. Four critical sugar-binding residues Y168, E188, W190 and H191 were present in our HCoV-OC43 isolates corresponding to Y162, E182, W184 and H185 in the bovine coronavirus sequence [Peng et al., 2012].

The receptor-binding of SARS-CoV in the S1 carboxy-terminal domain around the receptor binding motif involving amino acids 479 and 487 are areas where nonsynonymous substitutions and genetic diversity occurred for our HCoV-OC43 isolate sequences. These areas are important for binding affinity to human angiotensin-converting enzyme 2 for SARS-CoV, and host immune responses [Li, 2015]. Our isolates had nonsynonymous substitutions at amino acid residues 259 and 260-264 in the N-terminal domain of S1, and, in particular, had nonsynonymous amino acid substitutions within the region between amino acids 471 and 550 at a putative receptor binding domain of HCoV-OC43 (a.a. positions 339-549) [Lau et al., 2011]. In other HCoV strains, for instance HCoV-229E, antibody neutralization of the virus is dependent on the antigenic phenotype of the S1 subunit region [Shirato et al., 2012]. Hence, this may also be a factor for HCoV-OC43 needing further study. Our patients all had acute upper respiratory and systemic signs and symptoms. The small number evaluated here precludes conclusions about pathogenicity and viral mutation rate patterns, but this approach opens up pathways to future study.

In conclusion, quasi-species were present in our HCoV-OC43 strains involving areas of the S1 subunit of the Spike gene that may affect evolution of the viral binding process and antigenicity, as well as host specificity. Further studies are needed to more fully characterize the extent of quasi-species in more clinical isolates and their relation to disease characteristics and host immune responses. Nonsynonymous mutations were more frequent than synonymous ones, contributing to the hypothesis that the mutations are important to viral persistence in the human population over time and to disease pathogenesis.

Acknowledgements: This work was supported by Veterans Affairs (VA) Research, Department of Veterans Affairs Office of Research and Development. The sponsor had no role in the design, conduct, data analysis and reporting of the study. The authors thank Kiana Wilder for secretarial assistance, Mary Margaret Donovan, MSN for clinical nursing assistance, and the Center for Vaccine Development at Saint Louis University.

5. References

Arden NH, Patriarca PA, Fasano MB, Lui K-J, Harmon MW, Kendal AP, Rimland D. 1988. The roles of vaccination and amantadine prophylaxis in controlling an outbreak of influenza A (H3N2) in a nursing home. Arch Intern Med 148:865-868. doi:10.1001/archinte.1988.00380040105016.

Belouzard S, Millet JK, Licitra BN, Whittaker GR. 2012. Mechanisms of coronavirus cell entry mediated by the viral spike protein. Viruses 4:1011-1033. doi:10.3390/v4061011.

Briese T, Mishra N, Jain K, Zalmout IS, Jabado OJ, Karesh WB, Daszak P, Mohammed OB, Alagaili AN, Lipkin WI. 2014. Middle East Respiratory Syndrome coronavirus quasispecies that include homologues of human isolates revealed through whole-genome analysis and virus cultured from dromedary camels in Saudi Arabia. mBio 5:e01146-14. doi: 10.1128/mBio.01146-14.

Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. 2003. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 77:8801-8811. doi:10.1128/JVI.77.16.8801-8811.2003.

Cotten M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P, Pybus OG, Rambaut A, Guan Y, Pillay D, Kellam P, Nastouli E. 2013. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. Emerg Infect Dis 19:736-742. http://dx.doi.org/10.3201/eid1905.130057.

DePristo MA, Banks E, Poplin R. Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, FennelTJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491-498. doi:10.1038/ng.806.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 191:2142-2149.

Fan X, Mao Q, Zhou D, Lu Y, Xing J, Xu Y, Ray SC, Di Bisceglie AM. 2009. High diversity of hepatitis C viral quasispecies is associated with early virological response in patients undergoing antiviral therapy. Hepatology 50:1765-1772. doi:10.1002/hep.23290.

Fan X, DiBisceglie AM. 2010. RT-PCR amplification and cloning of large viral sequences. Methods Mol Biol 630:139-149.

Gardner LM, McIntyre CR. 2014. Unanswered questions about the Middle East respiratory syndrome coronavirus (MERS-CoV). BMC Research Notes 7:358. doi: 10.1186/1756-0500-7-358.

Gorse GJ, Donovan MM, Patel GB, Balasubramanian S, Lusk RH. 2015. Coronavirus and other respiratory illnesses comparing older with younger adults. Am J Med 128:1251.e11-1251.e20.

Gorse GJ, O'Connor TZ, Hall SL, Vitale JN, Nichol KL. 2009. Human coronavirus and acute respiratory illness in older adults with chronic obstructive pulmonary disease. J Infect Dis 199:847-857.

Gorse GJ, O'Connor TZ, Young SL, Habib MP, Wittes J, Neuzil KM, Nichol KL. 2006. Impact of a winter respiratory virus season on patients with COPD and association with influenza vaccination. Chest 130:1109-1116.

Gorse GJ, O'Connor TZ, Young SL, Mendelman PM, Bradle, SF, Nichol KL, Strickland Jr. JH, Paulson DM, Rice KL, Foster RA, Fulambarker AM, Shigeoka JW, Kuschner WG, Goodman RP, Neuzil KM, Wittes J, Boardman KD, Peduzzi PN. 2003. Efficacy trial of live, cold-adapted and inactivated influenza virus vaccines in older adults with chronic obstructive pulmonary disease: a VA cooperative study. Vaccine 21:2133-2144.

Gorse GJ, Patel GB, Vitale JN, O'Connor TZ. 2010. Prevalence of antibodies to four human coronaviruses is lower in nasal secretions than in serum. Clin Vaccine Immunol 17:1875-1880.

Kim D-W, Kim Y-J, Park SH, Yun M-R, Yang J-S, Kang HJ, Han YW, Lee HS, Kim HM, Kim H, Kim A-R, Heo DR, Kim SJ, Jeon JH, Park D, Kim JA, Cheong H-M, Nam J-G, Kim K, Kim SS. 2016. Variations in spike glycoprotein gene of MERS-CoV, South Korea, 2015. Emerg Infect Dis 22:100-104.

Kin N, Miszczak F, Lin W, Gouilh MA, Vabret A, Epicorem Consortium. 2015. Genomic analysis of 15 human coronaviruses OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes. Viruses 7:2358-2377. doi: 10.33901 v 7052358.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568-576.

Langmead B, Salzberg SL. 2012. Fast grapped-read alignment with Bowtie 2. Nat Methods 9:357-359.

Lau SKP, Lee P, Tsang AKL, Yip CCY, Tse H, Lee RA, So L-Y, Lau Y-L, Chan K-H, Woo PCY, Yuen K-Y. 2011. Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. J Virol 85:11325-11337.

Li F. 2015. Receptor recognition mechanisms of coronaviruses: a decade of structural studies. J Virol 89:1954-1964. doi: 10.1128/JVI.02615-14.

Li F. 2016. Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol 3:237-261. Doi: 10.1146/annurev-virology-110615-042301.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079. doi: 10.1093/bioinformatics/btp352.

Masters PS, Perlman S. 2013. Coronaviridae In: Knipe DM, Howley PM, editors. Fields virology 6th Ed., Philadelphia: Lippincott Williams & Wilkins. p. 825-854.

Park D, Huh HJ, Kim YJ, Son D-S, Jeon H-J, Im E-H, Kim J-W, Lee NY, Kang E-S, Kang CI, Chung DR, Ahn J-H, Peck KR, Choi SS, Kim Y-J, Ki C-S, Park W-Y. 2016. Analysis of intra-patient heterogeneity uncovers the microevolution of Middle East respiratory syndrome coronavirus. Cold Spring Harb Mol Case Stud. doi: 10.1101/mcs.a001214.

Peck JR. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137:597-606.

Peng G, Xu L, Lin Y-L, Chen L, Pasquarella JR, Holmes KV, Li F. 2012. Crystal structure of bovine coronavirus spike protein lectin domain. J Biol Chem 287:41931-41938. doi: 10.1074/jbc.M112.418210.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842. doi: 10.1093/bioinformatics/btq033.

Ren Y, Wang W, Zhang X, Xu Y, Di Bisceglie AM, Fan X. 2016. Evidence for deleterious hepatitis C virus quasispecies mutation loads that differentiate the response patterns in interferon-based antiviral therapy. J Gen Virol 97:334-343. doi: 10.1099/jgv.0.000346.

Ren L, Zhang Y, Li J, Xiao Y, Zhang J, Wang Y, Chen L, Paranhos-Baccala G, Wang J. 2015. Genetic drift of human coronavirus OC43 spike gene during adaptive evolution. Sci Rep 5:11451. doi: 10.1038/srep11451.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863-864.

Shirato K, Kawase M, Watanabe O, Hirokawa C, Matsuyama S, Nishimura H, Taguchi F. 2012. Differences in neutralizing antigenicity between laboratory and clinical isolates of HCoV-229E isolated in Japan in 2004-2008 depend on the S1 region sequence of the spike protein. J Gen Virol 93:1908-1917. DOI 10.1099/vir.0.043117-0.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar A. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol Biol Evol 28:2731-2732.

Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F. 2012. De Novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle Pogonus chalceus (Coleoptera, Carabidae). PLoS One 7:e42605.

Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439:344-348.

Vijgen L, Keyaerts E, Leme, P, Moës E, Li S, Vandamme A-M, Van Ranst M. 2005. Circulation of genetically distinct contemporary human coronavirus OC43 strains. Virology 337:85-92.

Walsh EE, Shin JH, Falsey AR. 2013. Clinical impact of human coronavirus 229E and OC43 infection in diverse adult populations. J Infect Dis 208:1634-1642.

Wang W, Zhang X, Xu Y, Weinstock GM, Di Bisceglie AM, Fan X. 2014. High resolution quantification of hepatitis C virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha 2a and ribavirin combination therapy. PLoS One 9:e100131. doi: 10.1371/journal.pone.0100131.

Zhang Y, Li J, Xiao Y, Zhang J, Wang Y, Chen L, Paranhos-Baccala G, Ren L, Wang J,. 2015. Genotype shift in human coronavirus and emergence of a novel genotype by natural recombination. J Infect 70:641-650.

Figure Legends

Figure 1. Sliding window analysis of viral mutation loads, calculated as normalized Shannon entropy, over the entire HCoV-OC43 Spike gene (4086 bp) for the four clinical samples.

Figure 2. Neighbor-joining trees constructed with HCoV consensus sequences.

Panel A. All four HCoV-OC43 consensus Spike genes with seven references showing clustering of all four HCoV-OC43 samples with HCoV-OC43 strain (AY391777).

Panel B. Detailed analysis of the four HCoV-OC43 samples showing two clustering in genotype D and two clustering in genotype B.

Figure 3. Tight alignment of consensus amino acid sequences for the Spike genes of the four HCoV-OC43 clinical samples and the prototype strain, AY391777. Dots indicated the identity and the asterisk donated stop codons.

Table 1. Clinical Characteristics of Acute Respiratory Illnesses in Patients Associated with the Four Human Sequenced Coronavirus OC43 (HCoV-OC43) Clinical Samples

Study Subject	Age (Years)	Gender	Date of Illness	Anti-HCoV-OC43 Serum / Nasal Wash Reciprocal Antibody Titers[a] Acute Illness Visit	Anti-HCoV-OC43 Serum / Nasal Wash Reciprocal Antibody Titers[a] Convalescent Visit	Symptoms/Signs of Acute Respiratory Illness	Acute Illness Severity Score[b]	Acute VAS of Severity[c]	Medical History
2	39	Male	1-6-2010	664/<5	1,408/36	Sputum, rhinitis, dyspnea, headache, fatigue, sore throat	45	5	None
3	82	Male	12-17-2010	1,983/59	6,758/1,686	Cough, sputum, rhinitis, dyspnea, chills, headache, myalgia, fatigue, sore throat	57	7	COPD[d], emphysema, ischemic heart disease, diabetes mellitus
4	61	Male	1-10-2011	467/63	7,741/131,410	Cough, sputum, rhinitis, dyspnea, chills, headache, myalgia, body aches and pains, fatigue, sore throat, pharyngitis	80	6	Congestive heart failure, ischemic heart disease, asbestosis, diabetes mellitus
6	62	Male	3-13-2012	1,510/139	1,042/460	Rhinitis, dyspnea, chills, body aches and pains, fatigue, sore throat	27	6	Ischemic heart disease, diabetes mellitus, sinusitis

[a]Serum antibodies were IgG and nasal wash antibodies were IgA binding to UV light and psoralen-inactivated tissue culture-adapted HCoV-OC43 (ATCC#VR-1558) measured by enzyme-linked immunosorbent assay.

[b]Severity of influenza-like symptoms and signs score.

[c]VAS is visual analogue scale score.

[d]COPD is chronic obstructive pulmonary disease.

Table 2. Numbers of Mutations in the Four Human Coronavirus OC43 (HCoV-OC43) Samples at 0.5% Mutation Frequency and Quality Score ≥30

	No. of Mutations (% of total)		
Clinical HCoV-OC43 Samples	**Synonymous**	**Non-Synonymous**	**Total**
2	9 (39%)	14 (61%)	23
3	2 (10%)	18 (90%)	20
4	5 (12.5%)	35 (87.5%)	40
6	12 (32%)	26 (68%)	38
Total	28 (23%)	93 (77%)	121

Table 3. Complete Listing of 121 Viral Mutations and Population Frequencies in Four Human Coronavirus OC43 Clinical Samples

Sample 2 (genotype B)			Sample 3 (genotype D)			Sample 4 (genotype D)			Sample 6 (genotype B)		
Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency
Spike Gene	Complete Genome		Spike Gene	Complete Genome		Spike Gene	Complete Genome		Spike Gene	Complete Genome	
29	23672	1.26%	41	23684	0.50%	29	23672	1.21%	6	23649	0.54%
65	23708	0.99%	65	23708	0.83%	41	23684	0.84%	29	23672	0.98%
79	23722	36.92%	81	23724	0.92%	65	23708	0.95%	65	23708	0.68%
81	23724	7.96%	174	23817	1.25%	81	23724	0.82%	79	23722	45.12%
104	23747	0.56%	259	23902	0.55%	94	23737	0.89%	81	23724	7.07%
734	24377	0.53%	372	24015	0.55%	98	23741	0.56%	104	23747	0.54%
735	24378	0.65%	997	24640	0.71%	104	23747	0.54%	211	23854	0.92%
736	24379	0.67%	1012	24655	0.62%	119	23762	0.51%	284	23927	0.95%
801	24444	0.67%	1675	25318	0.71%	169	23812	0.59%	320	23963	0.51%
894	24537	0.50%	1682	25325	0.53%	242	23885	0.55%	325	23968	0.55%
979	24622	0.97%	1691	25334	0.79%	284	23927	0.75%	541	24184	0.60%
1036	24679	0.51%	1709	25352	0.68%	310	23953	0.53%	672	24315	0.72%
1431	25074	0.52%	1712	25355	0.74%	320	23963	0.81%	766	24409	0.93%
1773	25416	0.57%	1716	25359	0.61%	325	23968	0.59%	820	24463	0.50%
2172	25815	1.05%	1730	25373	0.56%	740	24383	0.55%	929	24572	0.66%
2244	25887	2.44%	1732	25375	0.52%	784	24427	0.58%	1217	24860	0.62%
2383	26026	0.51%	1736	25379	0.60%	1229	24872	2.03%	1437	25080	0.50%
2498	26141	0.57%	1754	25397	0.72%	1457	25100	0.51%	1503	25146	0.81%
2858	26501	2.10%	2000	25643	1.22%	1566	25209	0.68%	1523	25166	0.50%
3133	26776	0.59%	2498	26141	0.52%	1581	25224	0.54%	1544	25187	0.63%
3480	27123	0.50%				1622	25265	0.51%	1553	25196	0.63%
3958	27601	0.96%				1641	25284	0.60%	1596	25239	0.68%
3970	27613	0.68%				1670	25313	0.57%	1600	25243	0.57%
						1675	25318	0.62%	1601	25244	0.65%
						1691	25334	0.60%	1777	25420	0.89%
						1913	25556	1.07%	1841	25484	0.97%
						2494	26137	0.61%	1859	25502	2.27%
						2738	26381	0.50%	2078	25721	0.62%
						2754	26397	1.05%	2283	25926	0.70%
						2775	26418	0.61%	2383	26026	0.56%
						2948	26591	0.95%	2826	26469	0.55%
						3079	26722	0.77%	2862	26505	0.52%
						3095	26738	0.65%	3120	26763	0.51%

Sample 2 (genotype B)			Sample 3 (genotype D)			Sample 4 (genotype D)			Sample 6 (genotype B)		
Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency	Nucleotide Position in:		Frequency
Spike Gene	Complete Genome		Spike Gene	Complete Genome		Spike Gene	Complete Genome		Spike Gene	Complete Genome	
						3137	26780	0.51%	3151	26794	0.55%
						3309	26952	0.89%	3277	26920	0.80%
						3462	27105	0.66%	3284	26927	0.83%
						3676	27319	0.59%	3465	27108	0.54%
						3763	27406	1.10%	3620	27263	0.84%
						3929	27572	0.60%			
						4059	27702	0.57%			

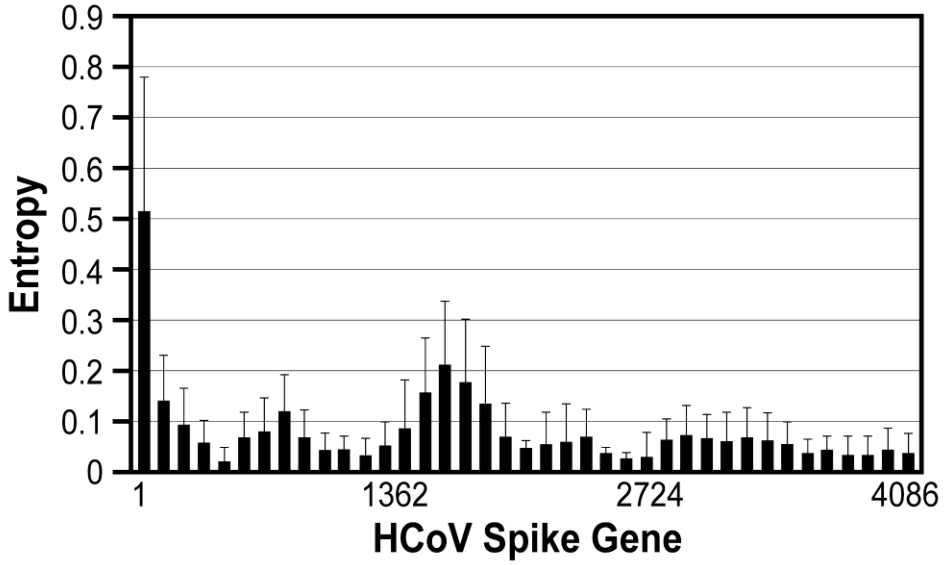

Figure 1

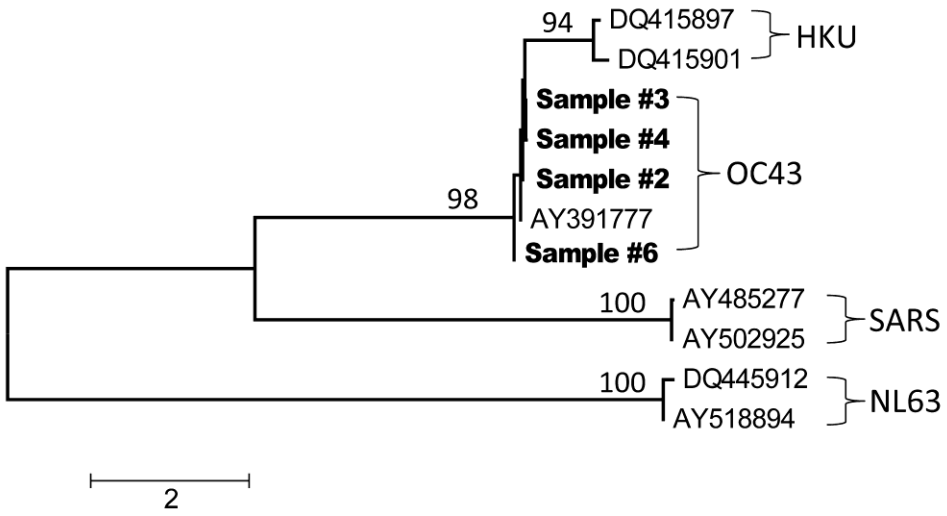

Figure 2a

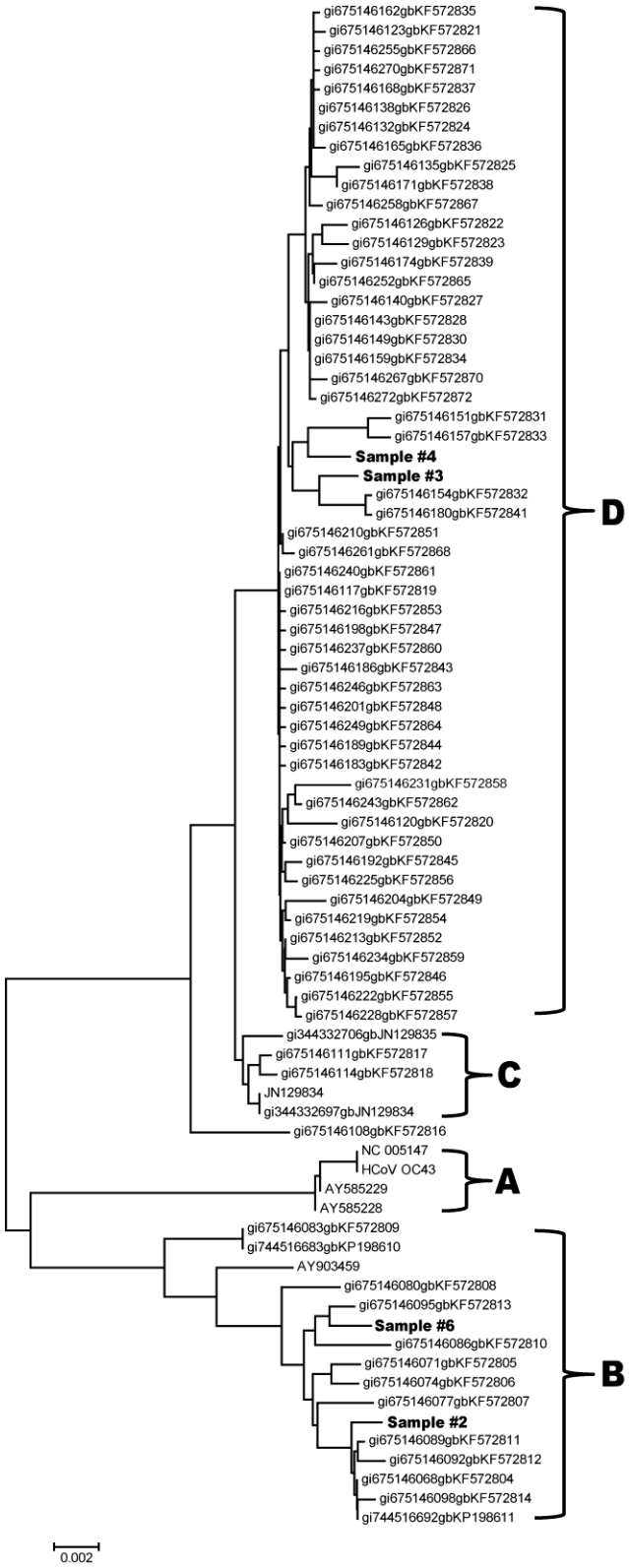

Figure 2b

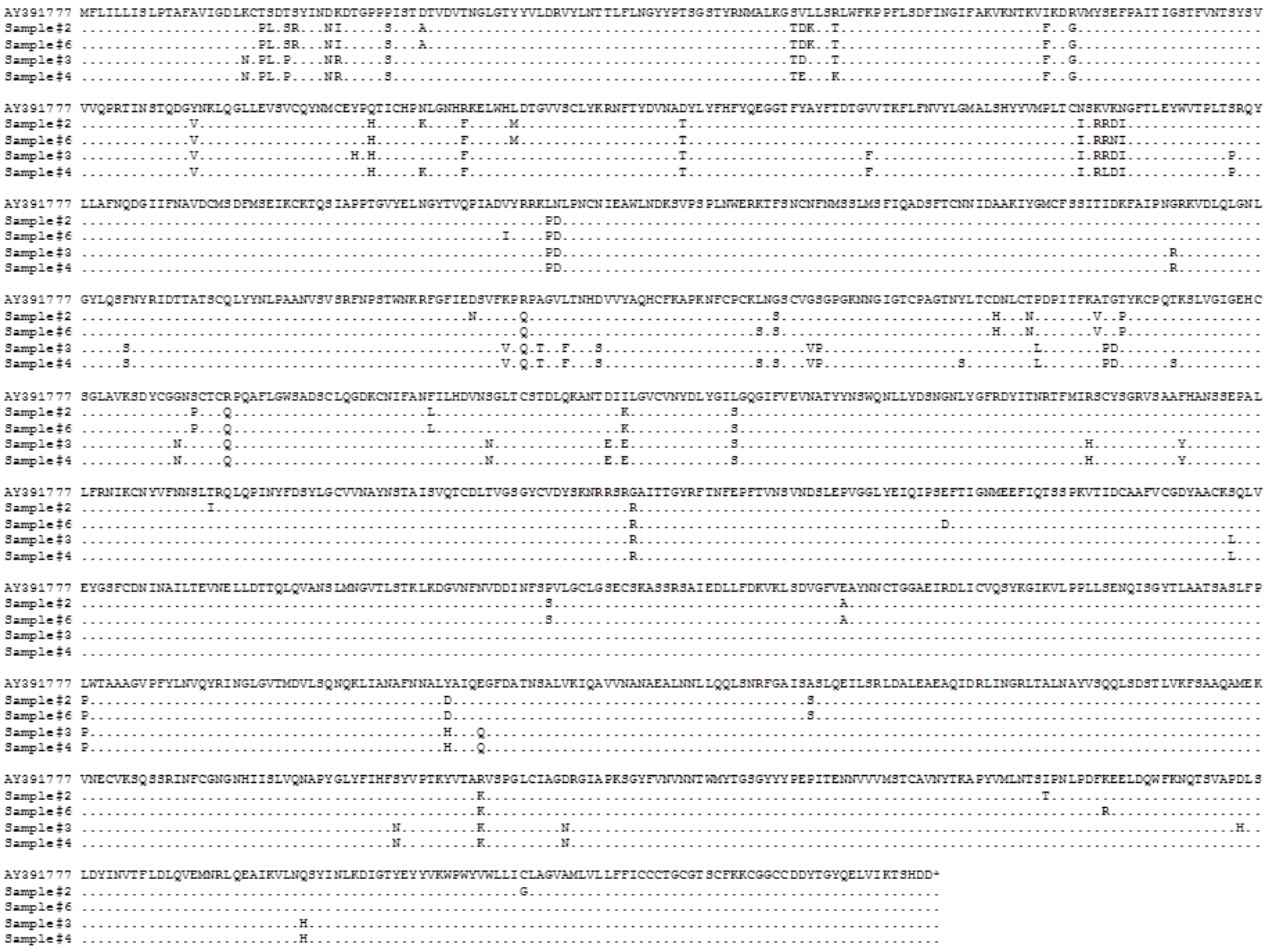

Figure 3