Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infection, Genetics and Evolution 4 (2004) 15–19
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Evidence from the evolutionary analysis of nucleotide sequences for a recombinant history of SARS-CoV

Michael J. Stanhope*, James R. Brown, Heather Amrine-Madsen

Bioinformatics Division, Genetics Research, GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, PA 19426, USA

Received 6 October 2003; received in revised form 24 October 2003; accepted 27 October 2003

Abstract

The origins and evolutionary history of the Severe Acute Respiratory Syndrome (SARS) coronavirus (SARS-CoV) remain an issue of uncertainty and debate. Based on evolutionary analyses of coronavirus DNA sequences, encompassing an approximately 13 kb stretch of the SARS-TOR2 genome, we provide evidence that SARS-CoV has a recombinant history with lineages of types I and III coronavirus. We identified a minimum of five recombinant regions ranging from 83 to 863 bp in length and including the polymerase, nsp9, nsp10, and nsp14. Our results are consistent with a hypothesis of viral host jumping events, concomitant with the reassortment of bird and mammalian coronaviruses, a scenario analogous to earlier outbreaks of influenzae.

Keywords: Recombination; Severe acute respiratory syndrome; Coronavirus; Phylogeny; Viral host-shift

1. Introduction

Recently, healthcare institutions around the world, particularly in Asia and Canada, have been forcibly challenged to respond to sudden outbreaks of Severe Acute Respiratory Syndrome (SARS). SARS is a highly communicable, and often lethal, illness thought to be caused by a novel type of coronavirus (Fouchier et al., 2003; Ksiazek et al., 2003; Kuiken et al., 2003), a group of positive, single-stranded RNA viruses known to infect domestic birds and mammals, including humans. The origin of the SARS coronavirus (SARS-CoV) has been the subject of much speculation. One of the leading hypotheses is that SARS-CoV is a hybrid strain (Enserink, 2003), since there are reports of recombination in avian coronaviruses (Lee and Jackwood, 2000), however, until a recent report in this journal (Rest and Mindell, 2003), there was no evidence that SARS-CoV is a recombinant. Our analysis of this question, completed at the time of publication of the Rest and Mindell paper, differs from their work in the choice of methods, the extent of the genome analyzed, taxon sampling, and in the analysis of nucleotides rather than amino acids. Our results act to both corroborate and extend their findings, adding further support to the idea that SARS has had a recombinant history involving different coronavirus lineages and suggest the possibility that the genome could have arisen through a combination of host jumping and recombination events in a manner analogous to previous outbreaks of influenzae (Gregory et al., 2003; Zhou et al., 1999).

2. Materials and methods

2.1. DNA sequence alignments

Many of the molecular evolutionary methods for detection of recombination events involve the analysis of multiple DNA sequence alignments. In choosing coronavirus sequences for our analyses, we made an effort to maximize both genetic diversity of the different coronavirus variants, as well as the length of possible contiguous comparative data (i.e. in excess of 20 kb). We aligned (ClustalW; Thompson et al., 1994) a large portion of the SARS virus TOR2 strain, at the DNA sequence level, between positions 7349–20969, to other coronaviruses from previously designated groups I, II, and III (Ksiazek et al., 2003; Marra et al., 2003; Rota et al., 2003). At the time of manuscript submission, there were 36 complete, or nearly complete, genomes of SARS virus available, all of which were highly similar at the DNA sequence level, thus strain selection does not affect the results of our analyses. The DNA sequence alignments within this region had a few segments which could not be reliably aligned, and thus were excluded from our analyses.

* Corresponding author. Tel.: +1-610-917-6577; fax: +1-610-917-7901.
E-mail address: michael_j_stanhope@gsk.com (M.J. Stanhope).

1567-1348/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2003.10.001

This resulted in 13 separate DNA alignments, which ranged in length from 245 to 3785 bp. Within each of these sub-alignments, any further ambiguous regions were deleted before recombination detection analyses. This was performed in a highly conservative manner, such that not only did we remove any and all remotely ambiguous gaps, but the regions surrounding the gaps were additionally excluded up to areas of clearly anchored sequence alignment (identical or virtually identical stretches of sequence) flanking either side of the gap (alignments available upon request).

2.2. Recombination detection

We used the recombination detection program PLATO (Grassly and Holmes, 1997) which employs a maximum likelihood (ML) approach to demarcate the boundaries of anomalous evolving regions in a DNA sequence alignment, with statistical measures of confidence. PLATO has a phylogenetic basis, and such methods have been shown to be somewhat less powerful than substitution distribution methods, in the sense that they are less able to identify more subtle examples of recombination (Posada et al., 2002; Posada and Crandall, 2001). However, this in turn means that such approaches are also more conservative in their overall assessment, and indeed phylogenetic methods can only detect recombination events that change the topology (Posada et al., 2002; Posada and Crandall, 2001). Importantly, the propensity for most recombination detection programs, including PLATO, to detect false positives appears to be low (Posada et al., 2002; Posada and Crandall, 2001). PLATO was used to assess possible recombinant regions for each of the 13 alignments, employing parameters of an HKY model of sequence evolution, five steps for the sliding window, and 1000 replications of Monte Carlo simulation. To add a further level of conservative assessment to our recombination detection, phylogenetic analyses were performed on all partitions identified by PLATO, the putative non-recombinant portions of such alignments, as well as all the remaining alignments. For all of these phylogenetic analyses, the best fitting model of sequence evolution and the corresponding values for the rate matrix, shape of the gamma distribution, and proportion of invariant sites were estimated by the program MODELTEST (Posada and Crandall, 1998). The evolutionary history of each region was compared to the control phylogeny, which was based on a concatenation of the 13 alignments. This control topology was the same as that derived from the concatenated non-recombinant sequence portions. A region was concluded as a SARS-CoV recombinant when all, or at least the majority (for shorter sequences), of phylogenetic methods agreed in their convincing placement of SARS-CoV in an alternative position to that of the control phylogeny. Phylogenies were reconstructed using Bayesian (Huelsenbeck and Ronquist, 2001), maximum likelihood, neighbor joining (NJ, log det distances) and maximum parsimony methods, implemented in PAUP* 4.0b (Swofford, 2002). For ML, starting trees were obtained via neighbor joining and for parsimony analyses addition sequence was employed with 10 random input orders. Tree-bisection reconnection (TBR) was the branch-swapping algorithm used in all analyses. Gaps were coded as missing data in all analyses. Bootstrap support values were obtained with 1000 replicates for maximum parsimony and neighbor joining analyses and 100 replicates for ML. Bayesian analyses were performed using Mr. Bayes (Huelsenbeck and Ronquist, 2001) with 500,000 generations, sampling frequency every 100 generations, four Markov chains, random starting trees, and a burn-in of 100,000 generations.

The PLATO results were corroborated using split decomposition analysis (program SplitsTree; Huson, 1998) and bootscanning (Salminen et al., 1995) (program BOOTSCAN within the SimPlot package). Instances identified by PLATO as possible SARS-CoV recombinants were similarly identified by SplitsTree and bootscanning.

3. Results and discussion

In the unrooted control phylogeny, SARS-CoV branches, with convincing support, along the lineage leading to group II coronaviruses (Fig. 1a), which is in agreement with previous reports (Ksiazek et al., 2003; Marra et al., 2003; Rota et al., 2003). The long branch separating SARS-TOR2 from the group II coronaviruses, in comparison to the branch lengths separating the various group II representatives, is in general agreement with earlier opinions for SARS-CoV as a new, fourth group of coronaviruses (Marra et al., 2003; Rota et al., 2003), and contrary to Snijder et al. (2003) who suggest, based on analysis of replicase ORF1b, that SARS-CoV is more aptly considered a distant member of group II. For the individual alignments the models of sequence evolution identified by MODELTEST were GTR+gamma (alignments corresponding with TOR2 coordinates: 10,645–10,902; 12,613–13,344; 13,725–14,147; 20,100–20,984; and recombinant regions: 15,259–15,342; 19,577–19,862), GTR + invariants (9982–10,125; 13,392–13,610), GTR + gamma + invariants (7366–7710; 10,147–10,626; 11,554–11,973; 11,989–12,516; 18,117–18,980; 14,172–17,936; 19,065–19,871), or HKY + gamma (recombinant region: 15,974–16,108).

Under our recombination criteria, several regions of recombination were evident, involving two alternative positions of SARS-CoV (Fig. 1b and c). These two branching arrangements were SARS-CoV on the branch leading to group III viruses (avian) or as sister lineage to the group I clade (porcine, human, etc.). PLATO identified anomalous regions included 15,259–15,342 (Z value of 5.0666; Z values greater than 3.8896 judged to be significant), 15,974–16,108 (Z value of 4.3997; Z values greater than 3.8896 judged to be significant), and 19,577–19,862 (Z value of 6.1619; Z values greater than 3.6471 judged to be significant). Phylogenetic analysis of 15,259–15,342 supported SARS-CoV with group III (Fig. 1b), whereas 15,974–16,108 supported

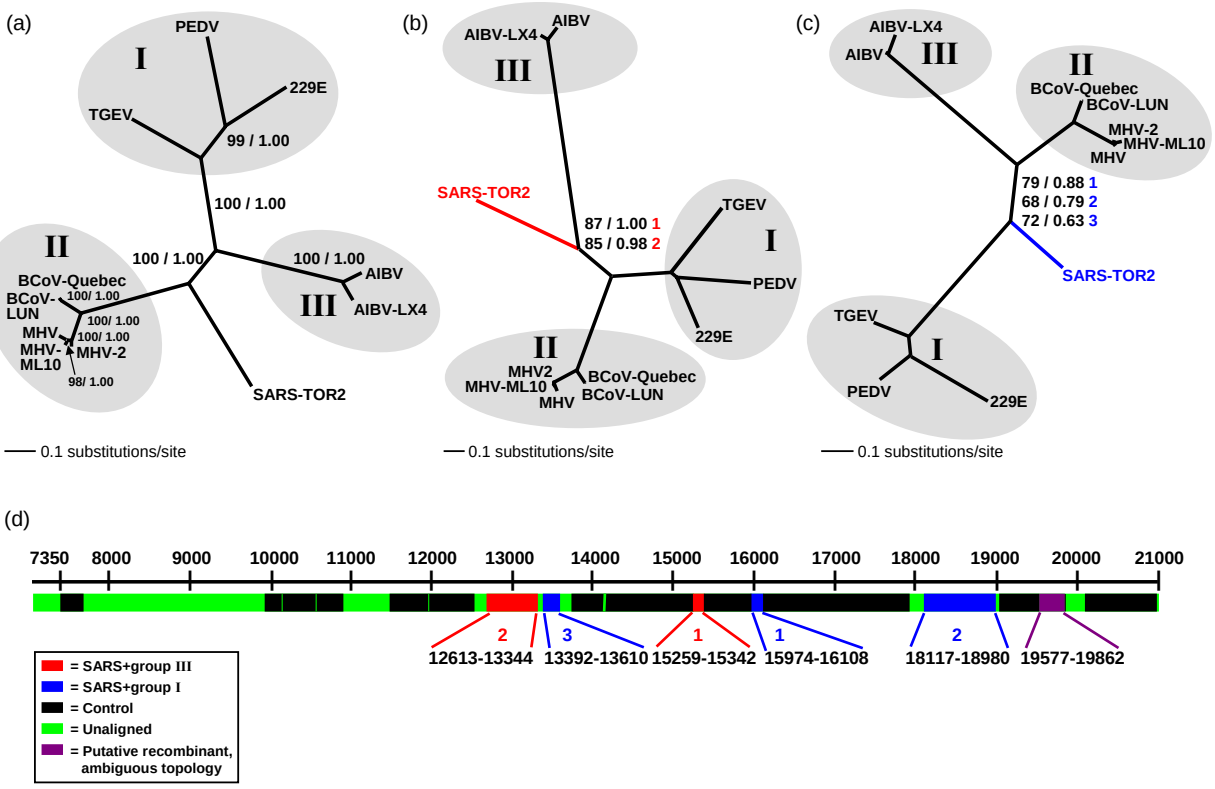

Fig. 1. Examples and summary of recombinational analyses. Sequence identifications are as follows: 229E (human): AF304460; PEDV, porcine epidemic diarrhea virus: AF353511; TGEV, porcine transmissible gastroenteritis virus: AJ271965; AIBV-LX4, Avian infectious bronchitis virus-LX4: AY223860; AIBV, Avian infectious bronchitis virus: M95169; SARS-TOR2: AY274119; MHV-ML10, Murine hepatitis virus-ML-10: AF208067; MHV, Murine hepatitis virus: M55148; MHV-2, Murine hepatitis virus strain 2: AF201929; BCov-Quebec, Bovine coronavirus Quebec: AF220295; BCov-LUN, Bovine coronavirus-LUN: AF391542. (a) the control topology with Bayesian (Huelsenbeck and Ronquist, 2001) posterior probabilities (1.0 for all nodes) and ML (Swofford, 2002) bootstrap values; branch lengths drawn proportional to the amount of sequence change. (b) a tree resulting from one of the PLATO detected anomalous zones, implicating a recombination event involving SARS-CoV and the group III lineage; ML bootstrap and Bayesian posterior probabilities are indicated for both recombination events involving SARS-CoV with the group III lineage, corresponding with the red numbers in d. (c) A tree resulting from a recombinational zone implicating genetic exchange involving SARS-CoV and the group I lineage; ML bootstrap and Bayesian posterior probabilities are indicated for all three recombination events involving SARS-CoV with the group I lineage, corresponding with the blue numbers in (d). (d) A schematic of the recombination and non-recombination events identified in the SARS-TOR2 genome between position 7349 and 20,969.

SARS-CoV with group I (Fig. 1c). Phylogenetic analysis of the third putative recombinant region identified by PLATO (i.e. 19,577–19,862; Fig. 1d), proved inconclusive, with ML and Bayes supporting SARS-CoV with group I, and parsimony and NJ yielding the control topology (bootstrap support under 60%, and Bayesian posterior probability less than 0.50). Three further recombinant regions were identified by phylogenetic analysis, that did not yield significant PLATO results, simply because the entire (or very nearly) alignment appears to represent a recombinant zone (i.e. nothing for PLATO to identify as anomalous; Fig. 1d). Mutational saturation at synonymous positions of codons can be ruled out as a possible explanation for the alternative branching arrangements of these five (possibly six) recombinant zones, because phylogenies for these same regions based on alignments that exclude third codon positions, as well as amino acid sequences, yielded identical topologies. The resulting genomic picture suggests a complex evolutionary history of recombination involving SARS-CoV (Fig. 1d). The placement of SARS-CoV on the branches leading to groups I or III and not nested within these groups indicates that either the recombination events are ancient in nature or the donor species are not present in currently available sequence data. The inclusion of greater host species representation, which is presently possible for a few regions of the genome, such as a 922 bp region of polymerase (for which there are additional GenBank sequences from cat, dog—group I; turkey—group III; human OC43, porcine—group II) (Stephensen et al., 1999), did not allow a more specific identification of the possible species involved, and implicated the same recombination event between positions 15,259–15,342 (Fig. 1d).

Two recent reports regarding the SARS genome suggest, based on analysis of amino acid sequences, that there is either no evidence for recombination (Rota et al., 2003) or no evidence for recent recombination involving other coronaviruses (Marra et al., 2003). Although the methodological details regarding recombination detection are scant in both these reports, we gather that in the one case they came to this conclusion by comparing branching arrangements between

gene trees (Marra et al., 2003), and in the other case by performing an amino acid similarity plot (Rota et al., 2003). In the first case, a comparison of gene trees would not pick up recombination events that crossed gene boundaries, or which involved relatively short stretches of sequence within a gene. In the second instance, similarity plots will only tend to pick up recombination events in comparisons that involved the actual donor, a close relative to the donor, and/or a recent event.

In contrast, our analysis agrees with Rest and Mindell (2003) in identifying recombination in RDRP (RNA dependent RNA polymerase), although our approach tends to suggest more specific break-points, and a larger number of smaller recombinant regions than does their analysis (three regions in RDRP: 13,392–13,610; 15,259–15,342; 15,974–16,108, based on TOR2 coordinates). We also identified several additional recombinant regions in the SARS-CoV genome, encompassing regions not analyzed by Rest and Mindell, including: 12,613–13,344 including all of nsp9 and most of nsp10 and 18,117–18,980 of nsp14.

Analyses of currently available sequences of coronaviruses, yields the conclusion that group III is exclusively composed of avian coronaviruses, while groups I and II have viruses isolated from pig, human, murine rodents, cat, dog and bovine. Our results indicate that SARS-CoV recombined with a member of the group III lineage, suggesting that an avian coronavirus was involved, a further point of general agreement between our results and that of Rest and Mindell (2003). Other recombination events evident from our analysis, involve the branch leading to group I, which encompasses viruses from several mammalian taxa, including two very divergent strains of porcine coronaviruses. Thus, our analyses indicate that human SARS-CoV have a past history of recombination with coronaviruses hosted in distinct animal groups. Mixed animal husbandry practices, in proximity to human populations, could have led to the evolution of the SARS coronavirus and facilitated its progression as an infectious disease in humans. Novel human influenza viruses are thought to have arisen from the reassortment, within porcine hosts, of avian, swine, and human influenza viruses (Gregory et al., 2003; Zhou et al., 1999). We suggest that our recombination results for SARS-CoV implicate a suspiciously analogous history. More specifically, SARS-CoV could have arisen from a combination of host jumping and recombinational events, involving as yet unidentified strains of avian coronavirus group III and mammalian (possibly pig) coronavirus group I. Rest and Mindell (2003) suggested host-species shifts have been relatively common in the diversification of coronavirus lineages, a result consistent with our hypothesis for SARS-CoV. Critical to determination of the evolutionary origin of SARS-CoV are expanded epidemiological surveys of wild and domestic animals, including in particular, additional avian species.

Understanding the origin and evolutionary history of SARS-CoV is important to proper vaccine development as well as the epidemiological modeling of future outbreaks. Current perception of the SARS-CoV genome is one of relative genetic stability (Brown and Tetro, 2003; Ruan et al., 2003), however, our analyses indicate that SARS-CoV has a complex history of recombination, suggesting that the genome may not be as stable as previously thought. We propose that future epidemiological modeling efforts and vaccine development take this new evidence into account.

References

Brown, E.G., Tetro, J.A., 2003. Comparative analysis of the SARS coronavirus genome: a good start to a long journey. Lancet 361, 1756–1757.

Enserink, M., 2003. Infectious diseases. Calling all coronavirologists. Science 300, 413–414.

Fouchier, R.A., Kuiken, T., Schutten, M., van Amerongen, G., van Doornum, G.J., van den Hoogen, B.G., Peiris, M., Lim, W., Stohr, K., Osterhaus, A.D., 2003. Aetiology: Koch's postulates fulfilled for SARS virus. Nature 423, 240.

Grassly, N.C., Holmes, E.C., 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. 14, 239–247.

Gregory, V., Bennett, M., Thomas, Y., Kaiser, L., Wunderli, W., Matter, H., Hay, A., Lin, Y.P., 2003. Human infection by a swine influenza A (H1N1) virus in Switzerland. Arch. Virol. 148, 793–802.

Huelsenbeck, J.P., Ronquist, F.R., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73.

Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J.A., Lim, W., Rollin, P.E., Dowell, S.F., Ling, A.E., Humphrey, C.D., Shieh, W.J., Guarner, J., Paddock, C.D., Rota, P., Fields, B., DeRisi, J., Yang, J.Y., Cox, N., Hughes, J.M., LeDuc, J.W., Bellini, W.J., Anderson, L.J.,. SARS Working Group 2003. A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. 348, 1947–1958.

Kuiken, T., Fouchier, R.A., Schutten, M., Rimmelzwaan, G.F., van Amerongen, G., van Riel, D., Laman, J.D., de Jong, T., van Doornum, G., Lim, W., Ling, A.E., Chan, P.K., Tam, J.S., Zambon, M.C., Gopal, R., Drosten, C., van der Werf, S., Escriou, N., Manuguerra, J.C., Stohr, K., Peiris, J.S., Osterhaus, A.D., 2003. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. Lancet 362, 263–270.

Lee, C.-W., Jackwood, M.W., 2000. Evidence of genetic diversity generated by recombination among avian coronavirus IBV. Arch. Virol. 145, 2135–2148.

Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C., Roper, R.L., 2003. The genome sequence of the SARS-associated coronavirus. Science 300, 1399–1404.

Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19, 708–717.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U.S.A. 98, 13757–13762.

Rest, J.S., Mindell, D.P., 2003. SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. Infect. Genet. Evol. 3, 219–225.

Rota, P.A., Oberste, M.S., Monroe, S.S., Nix, W.A., Campagnoli, R., Icenogle, J.P., Penaranda, S., Bankamp, B., Maher, K., Chen, M.H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J.L., Chen, Q., Wang, D., Erdman, D.D., Peret, T.C., Burns, C., Ksiazek, T.G., Rollin, P.E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Gunther, S., Osterhaus, A.D., Drosten, C., Pallansch, M.A., Anderson, L.J., Bellini, W.J., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Ruan, Y.J., Wei, C.L., Ee, A.L., Vega, V.B., Thoreau, H., Su, S.T., Chia, J.M., Ng, P., Chiu, K.P., Lim, L., Zhang, T., Peng, C.K., Lin, E.O., Lee, N.M., Yee, S.L., Ng, L.F., Chee, R.E., Stanton, L.W., Long, P.M., Liu, E.T., 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 361, 1779–1785.

Salminen, M.O., Carr, J.K., Burke, D.S., McCutchan, F.E., 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res. Hum. Retroviruses 11, 1423–1425.

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L., Guan, Y., Rozanov, M., Spaan, W.J., Gorbalenya, A.E., 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 29, 991–1004.

Stephensen, C.B., Casebolt, D.B., Gangopadhyay, N.N., 1999. Phylogenetic analysis of a highly conserved region of the polymerase gene from 11 coronaviruses and development of a consensus polymerase chain reaction assay. Virus Res. 60, 181–189.

Swofford, D.L., 2002. PAUP* Version 4.0b10. Sinauer Associates, Sunderland, MS.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 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, 4673–4680.

Zhou, N.N., Senne, D.A., Landgraf, J.S., Swenson, S.L., Erickson, G., Rossow, K., Liu, L., Yoon, K., Krauss, S., Webster, R.G., 1999. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. J. Virol. 73, 8851–8856.