Phylogenetic Analysis of the Full-Length SARS-CoV Sequences: Evidence for Phylogenetic Discordance in Three Genomic Regions

G. Magiorkinis,[1] E. Magiorkinis,[1] D. Paraskevis,[1,2] A.M. Vandamme,[2] M. Van Ranst,[2]
V. Moulton,[3] and A. Hatzakis[1]*

[1] *National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Athens University Medical School, Athens, Greece*
[2] *Laboratory of Clinical and Epidemiological Virology, Department of Microbiology and Immunology, Rega Institute and University Hospitals, Leuven, Belgium*
[3] *The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden*

The origin of the severe acute respiratory syndrome-coronavirus (SARS-CoV) remains unclear. Evidence based on Bayesian scanning plots and phylogenetic analysis using maximum likelihood (ML) and Bayesian methods indicates that SARS-CoV, for the largest part of the genome (~80%), is more closely related to Group II coronaviruses sequences, whereas in three regions in the *ORF1ab* gene it shows no apparent similarity to any of the previously characterized groups of coronaviruses. There is discordant phylogenetic clustering of SARS-CoV and coronaviruses sequences, throughout the genome, compatible with either ancient recombination events or altered evolutionary rates in different lineages, or a combination of both. ***J. Med. Virol. 74:369–372, 2004.*** © 2004 Wiley-Liss, Inc.

KEY WORDS: SARS-CoV; phylogenetic discordance; recombination; Bayesian scanning

INTRODUCTION

The recent epidemic outbreak of the severe acute respiratory syndrome, known as SARS, was first reported in the Guangdong Province of China late in 2002 and since then several thousand cases have been reported worldwide, mainly in Hong Kong, Vietnam, and Canada [Drosten et al., 2003; Ksiazek et al., 2003; Poutanen et al., 2003]. Analysis of the full-length genomic sequence of the causative agent of SARS [Marra et al., 2003; Rota et al., 2003] established that this agent was a unique coronavirus (SARS-CoV) that displayed no obvious similarity to any other members of the *Coronavirus* genus. Coronaviruses (order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus*) have the longest genome (27,000–31,000 nucleotides) amongst all RNA viruses. They comprise a highly divergent group of RNA-viruses that cause a variety of respiratory and enteric diseases in humans and many animal hosts. Previously available sequences of coronaviruses have been classified according to phylogenetic analysis into three main groups [Marra et al., 2003; Rota et al., 2003]. The newly identified sequence of the SARS-CoV was not found to cluster within any existing coronavirus group, thus suggesting the existence of an additional lineage, fourth group, within the *Coronavirus* genus [Ksiazek et al., 2003; Poutanen et al., 2003; Rota et al., 2003].

The origin of the SARS-CoV still remains unclear, although data from Hong Kong researchers indicate that the SARS-CoV virus is almost identical to viruses infecting civet cats [Guan et al., 2003]. Genetic recombination has been detected between different murine coronaviruses [Fu and Baric, 1992]. Thus, to determine whether recombination has occurred between SARS-CoV and any members of the three main coronaviruses groups, an in-depth analysis was undertaken of the evolutionary relationships between the SARS-CoV genome and representative sequences from all members of the coronaviruses genus available currently.

MATERIALS AND METHODS

Amino acid sequence alignments were obtained using the CLUSTAL W program (v 1.81) [Thompson et al., 1994] and subsequently edited manually for the major gene products ORF1ab, Spike, M and N proteins

*Correspondence to: A. Hatzakis, MD, PhD, Associate Professor, Chief, National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Athens University Medical School, Mikras Asias 75, 11527, Athens, Greece.
E-mail: ahatzak@med.uoa.gr

Accepted 5 July 2004
DOI 10.1002/jmv.20187
Published online in Wiley InterScience
(www.interscience.wiley.com)

to exclude ambiguous parts of the alignment. More specifically, X1, X2, X3, X4, X5, and E genomic regions were excluded since a reliable alignment could not be obtained for these fragments. Amino acid sequences derived from the full-length sequences of avian (NC_001451), bovine (NC_003045), human (NC_002645), murine (NC_001846), porcine (NC_002306, NC_003436) coronaviruses, and the SARS-CoV (AY278488) were used for all the genes. Additional amino acid sequences were used for the *Spike*, *M*, and *N* genes. For the *Spike* gene, additional amino acid sequences were used for avian (NP_040831, AAA70235, AAK27168, AAC54067, AAC42112, CAA28432, P11223), bovine (NP_150077, AAL57308, AAA42908, BAA00557, VGIHQU), canine (CAA01637), feline (CAA56778), human (AAK32190, P15423, CAA71056), murine (AAF68923, AAF69334, P11225, VGIHMJ, BAA23719, JQ1534, Q02385), and porcine (NP_058424, NP_598310, A43573, AAB30949, AAK38656, AAM77000, AAC96004) coronaviruses. For the *M* gene, additional amino acid sequences were used for avian (AAK91808, P04327, P05136, MMIH68), bovine (NP_150082, AAA79959, AAF25524, AAL40405), feline (AAB47501, P25878, CAA74229), human (NP_073555), murine (AAF05705), porcine (NP_058427, AAM77004, AAB71544, AAA47912, P24412, D36607), and rat (AAD33105) coronaviruses. For the *N* gene, additional amino acid sequences were used for avian (AAO46049, AAM82282, AAC54068, AAF06352, CAC39121, VHIHAI), bovine (NP_150083, AAA42758, AAK83362, P10527), human (NP_073556), murine (A45340, P18447), porcine (NP_598314, AAK38660, P04134, CAB91150, S47428, E36607), rat (A45396, BAA01591, Q02915), and turkey (JQ1173) coronaviruses. All the amino acid alignments are available upon request.

After constructing the alignments, the separated genes containing only the common strains in all genes were joined in concatenates and a Bayesian scanning plot was undertaken as described previously [Paraskevis et al., 2003] using the WAG model [Whelan and Goldman, 2001] as implemented in the MrBayes program [Huelsenbeck and Ronquist, 2001]. In particular, within a sliding window of a specified length, Bayesian inference was carried out to determine the support of every clade within a given phylogenetic tree, and these support values were plotted along the alignment as described previously [Paraskevis et al., 2003]. Marginal posterior probabilities between SARS-CoV and sequences of coronaviruses Groups I–III were plotted throughout the concatenated alignment. Posterior probabilities were estimated for a sliding window of 120 amino acids moving in steps of 20 amino acids. Metropolis Coupled Markov Chains Monte Carlo (MCMCMC) analysis was run for 50,000 generations, sampling every 100 steps, and burn-in was set to 5,000 steps. MCMCMC had been checked previously so as to confirm that it had reached equilibrium up to this number of generations in different parts of the alignment, thus suggesting that 5,000 generations could be used as a fixed burnin throughout the alignment.

Phylogenetic trees were then constructed for every piece of the alignment with evidence of phylogenetic discordance, using the maximum likelihood (ML) and Bayesian inference with WAG [Whelan and Goldman, 2001] and MV [Muller and Vingron, 2000] substitution models, which have been constructed by distantly related amino acid sequences, and a Γ-distributed rates heterogeneity among sites as implemented in the Tree-Puzzle (parallel version 5.0) [Schmidt et al., 2002] and MrBayes, respectively. A likelihood mapping analysis was also carried out, as implemented in the Tree-Puzzle program, to investigate the levels of phylogenetic signal in each separate region examined.

RESULTS

Exploratory phylogenetic analyses in the Spike, M, and N genomic regions, including additional partially available coronavirus sequences, revealed that full-length strains are representative and can be used for the exploration of phylogenetic analysis within the coronaviridae family along the full-genome. Bayesian scanning and subsequent phylogenetic analysis revealed that the SARS-CoV sequence was related more closely to Group II than the other two groups in most of its genome (e.g., at the region spanning amino acid positions 4309–5612 in reference to the murine hepatitis virus *ORF1ab* gene) (Fig. 1). This clustering was supported by high quartet puzzling support values and high posterior probabilities under various substitution models, thus suggesting that ~80% of the SARS-CoV genomic sequence is related more closely to coronaviruses Group II than any other members of this family. Likelihood mapping results revealed that all these regions contained enough phylogenetic information (data not shown).

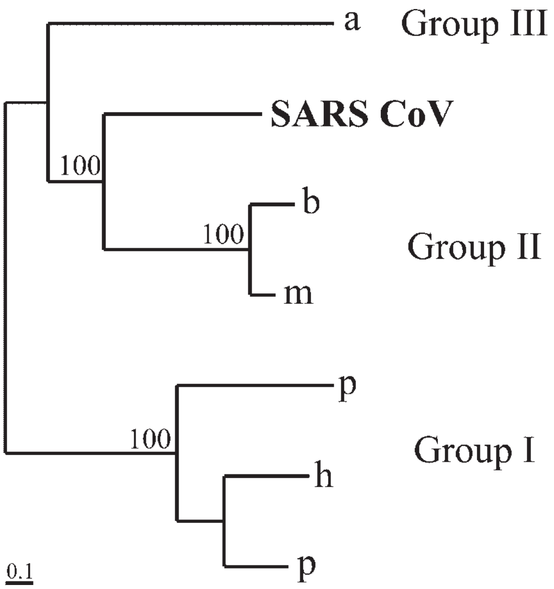

Fig. 1. Representative maximum likelihood tree for a region which the Bayesian scanning plot suggested that SARS CoV clusters with Group II coronaviruses. The viral strains are indicated by the first letter of the species infected by each strain (p, porcine; m, murine; b, bovine; h, human; a, avian).

Intriguingly, clustering of the SARS-CoV within Group II coronaviruses was not supported significantly in regions spanning amino acid positions: 3937−4309 (920−1240), 5612−5734 (2480−2600), 6159−6284 (3000−3120) (positions in reference to the murine hepatitis virus ORF1ab are indicated with their respective positions in reference to the alignment in the parentheses) (Fig. 2). In all cases, SARS-CoV was more closely related to Group I, whereas Group II/Group III−SARS-CoV/Group I nodes were separated by very small distances. Likelihood mapping in the above regions revealed that they contain sufficient phylogenetic signal. Thus, according to the results of phylogenetic analysis, it is suggested that SARS-CoV provides a separate lineage in these three genomic regions (Fig. 2). Moreover, in region spanning amino acid positions 4373−4540 (1300−1440), Bayesian inference suggested a significant cluster between Group III and SARS-CoV, whereas this finding was not supported by the ML method. However, according to likelihood mapping analysis, the phylogenetic signal is very limited in this region; thus we cannot reliably infer the evolutionary relationships in this region. Similarly, in a small fragment at the 5′ of *M* gene (amino acid positions 15−104 (4540−4620) with reference to the M protein of the murine hepatitis virus) showed a similar topology to the previous fragment, whereas given the discrepancy between Bayesian and ML methods, it cannot be deduced that SARS-CoV is related more closely to Group III.

DISCUSSION

According to the Bayesian full-genomic scanning and phylogenetic analyses based on the ML and Bayesian methods, evidence is provided that the SARS-CoV, for the largest part of the genome, is related more closely to Group II than to any other member of the coronaviruses sequences. The clustering between SARS-CoV and Group II is close to the root and probably reflects the evolutionary relationship between the "parental" sequences, from which SARS-CoV originated, and Group II. Moreover, it was found that in three distinct genomic regions (aa positions 3937−4309, 5612−5734, 6159−6284), SARS-CoV shows no apparent similarity with any of the coronavirus groups characterized previously. These findings provide a new insight, in addition to previous studies, where SARS-CoV (using the Neighbor Joining method for phylogenetic tree reconstruction)

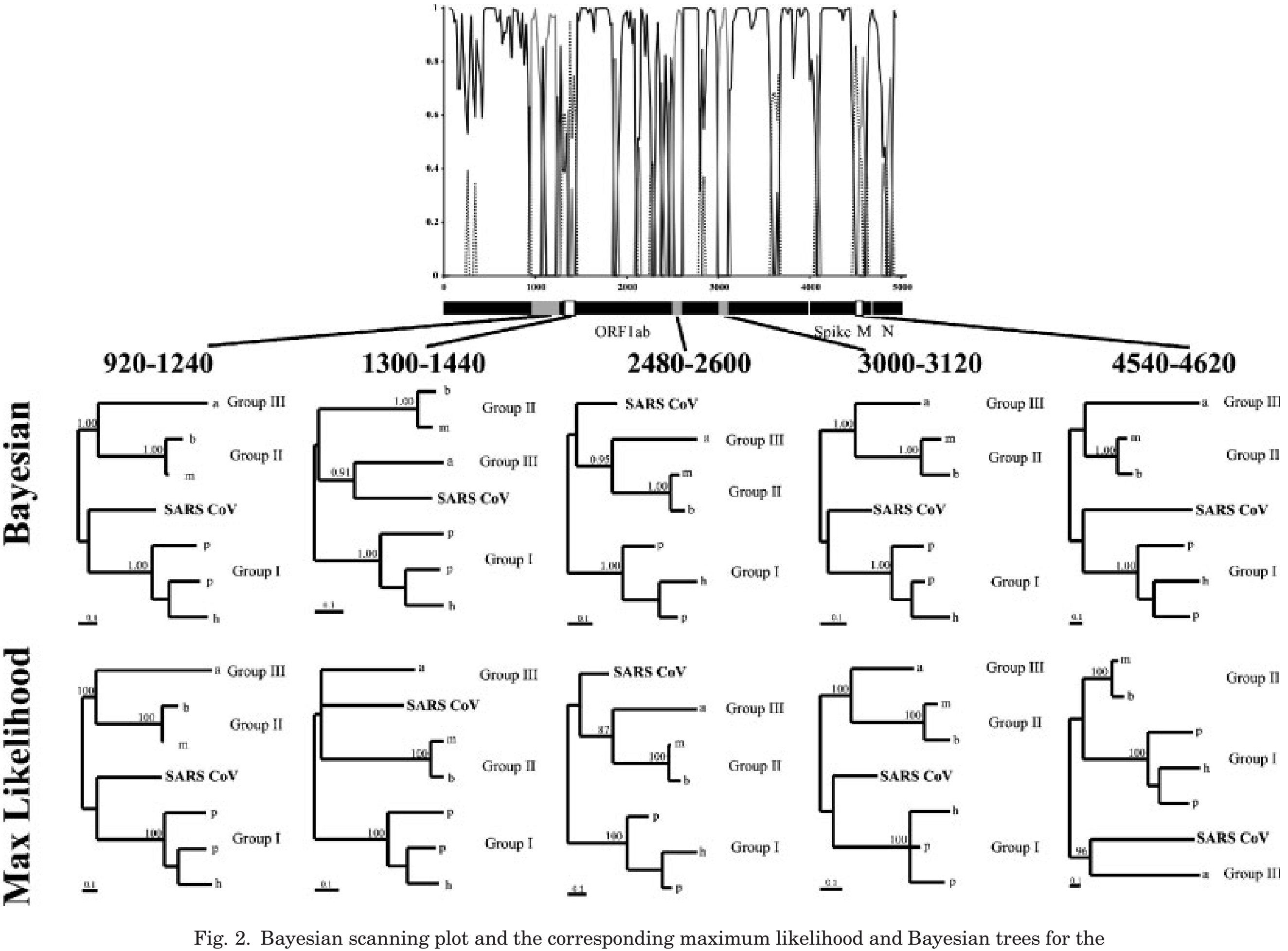

Fig. 2. Bayesian scanning plot and the corresponding maximum likelihood and Bayesian trees for the *ORF1ab*, *Spike*, *M* and *N* genes with the Whelan and Goldman substitution model.

was found not to be related significantly to any of the three coronavirus groups, or was separated by a short internal branch from Group I and III [Rota et al., 2003]. Moreover, three recent analyses indicated that SARS CoV was related more closely to Group II in partial genomic regions [Eickmann et al., 2003; Rest and Mindell, 2003; Snijder et al., 2003]. On the other hand, several analyses have noted that though SARS CoV is close to Group II, it has a recombinant origin [Rest and Mindell, 2003; Stavrinides and Guttman, 2004], but these analyses were based only on partial genomic regions, and did not include an exploratory analysis using a sliding window approach [Rest and Mindell, 2003; Stavrinides and Guttman, 2004].

Evidence is provided that, even though SARS-CoV in the majority of its genomic regions is closer related to coronavirus Group II, phylogenetic discordance in three partial genomic regions suggests a distinct clustering with respect to Group II. This could be a result of ancient recombination events, altered adaptive evolution in different lineages, or a combination of both. We do not want to speculate which of these explanations is correct, since, it is not easy to discriminate with certainty between these two factors.

REFERENCES

Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A, Burguiere AM, Cinatl J, Eickmann M, Escriou N, Grywna K, Kramme S, Manuguerra JC, Muller S, Rickerts V, Sturmer M, Vieth S, Klenk HD, Osterhaus AD, Schmitz H, Doerr HW. 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348:1967–1976.

Eickmann M, Becker S, Klenk HD, Doerr HW, Stadler K, Censini S, Guidotti S, Masignani V, Scarselli M, Mora M, Donati C, Han JH, Song HC, Abrignani S, Covacci A, Rappuoli R. 2003. Phylogeny of the SARS coronavirus. Science 302:1504–1505.

Fu K, Baric RS. 1992. Evidence for variable rates of recombination in the MHV genome. Virology 189:88–102.

Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF, Yuen KY, Peiris JS, Poon LL. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302:276–278.

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

Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W, Rollin PE, Dowell SF, Ling AE, Humphrey CD, Shieh WJ, Guarner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang JY, Cox N, Hughes JM, LeDuc JW, Bellini WJ, Anderson LJ. 2003. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348:1953–1966.

Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, 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 GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL. 2003. The genome sequence of the SARS-associated coronavirus. Science 300:1399–1404.

Muller T, Vingron M. 2000. Modeling amino acid replacement. J Comput Biol 7:761–776.

Paraskevis D, Lemey P, Salemi M, Suchard M, Van der Peer Y, Vandamme A-M. 2003. Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using Bayesian inference: Implications for the origin of HIV-1. Mol Biol Evol 20:1986–1996.

Poutanen SM, Low DE, Henry B, Finkelstein S, Rose D, Green K, Tellier R, Draker R, Adachi D, Ayers M, Chan AK, Skowronski DM, Salit I, Simor AE, Slutsky AS, Doyle PW, Krajden M, Petric M, Brunham RC, McGeer AJ. 2003. Identification of severe acute respiratory syndrome in Canada. N Engl J Med 348:1995–2005.

Rest JS, Mindell DP. 2003. SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. Infect Genet Evol 3:219–225.

Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rassmussen M, Fouchier R, Gunther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ. 2003. Characterization of a novel Coronavirus associated with severe acute respiratory syndrome. Science 300:1394–1399.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE. 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 331:991–1004.

Stavrinides J, Guttman DS. 2004. Mosaic evolution of the severe acute respiratory syndrome coronavirus. J Virol 78:76–82.

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 22:4673–4680.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699.