Arch Virol (2005) 150: 1–20
DOI 10.1007/s00705-004-0413-9

Archives of Virology
Printed in Austria

Testing the hypothesis of a recombinant origin of the SARS-associated coronavirus

X. W. Zhang[1], **Y. L. Yap**[1], and **A. Danchin**[2]

[1]HKU-Pasteur Research Centre, Hong Kong, P.R. China
[2]Pasteur Institute, Unit Genetics of Bacterial Genomes, Paris, France

Received February 27, 2004; accepted August 16, 2004
Published online October 11, 2004

Summary. The origin of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) is still a matter of speculation, although more than one year has passed since the onset of the SARS outbreak. In this study, we implemented a 3-step strategy to test the intriguing hypothesis that SARS-CoV might have been derived from a recombinant virus. First, we blasted the whole SARS-CoV genome against a virus database to search viruses of interest. Second, we employed 7 recombination detection techniques well documented in successfully detecting recombination events to explore the presence of recombination in SARS-CoV genome. Finally, we conducted phylogenetic analyses to further explore whether recombination has indeed occurred in the course of coronaviruses history predating the emergence of SARS-CoV. Surprisingly, we found that 7 putative recombination regions, located in Replicase 1ab and Spike protein, exist between SARS-CoV and other 6 coronaviruses: porcine epidemic diarrhea virus (PEDV), transmissible gastroenteritis virus (TGEV), bovine coronavirus (BCoV), human coronavirus 229E (HCoV), murine hepatitis virus (MHV), and avian infectious bronchitis virus (IBV). Thus, our analyses substantiate the presence of recombination events in history that led to the SARS-CoV genome. Like the other coronaviruses used in the analysis, SARS-CoV is also a mosaic structure.

Introduction

SARS, a new disease characterized by high fever, malaise, rigor, headache and non-productive cough, has spread to over 30 countries with around 8% of mortality rate on average. Sequence analysis of SARS coronavirus (SARS-CoV) [17, 25] showed that it is a novel coronavirus [12]. Anand et al. [1] reported a three-dimensional model of SARS-CoV main proteinase and suggested that

modified rhinovirus $3C^{pro}$ inhibitors could be useful for SARS therapy. Lipsitch et al. [15] developed a mathematical model of SARS transmission to estimate the infectiousness of SARS and the likelihood of an outbreak. Ng et al. [22] suggested that SARS-CoV could have been derived from an innocuous virus or oen causing a mild disease, that would become virulent after some mutational event occuring in some carriers. However, the source of SARS-CoV is not yet exactly known, although it has been reported that a virus highly related to SARS-CoV has infected some wild animals, such as masked palm civet, raccoon dog and badger [7].

Recombination, a key evolutionary process, accounts for a considerable amount of genetic diversity in natural populations. The occurrence of high-frequency homologous RNA recombination is one of the most intriguing aspects of coronavirus replication [14, 27, 31, 34]. The first experimental evidence for IBV recombination was found by Kottier et al. [11], although other studies have concluded that recombination is a feature of IBV evolution [4, 5, 10, 36–38]. Recombination in MHV was also experimentally demonstrated [16]. In particular, Snijder et al. [30] indicated that the recombination occurred between a coronavirus/torovirus-like virus and an influenza C-like virus, resulting in a line of coronaviruses that had a haemagglutinin esterase (HE) gene. This prompted us to explore the possible role of recombination in the emergence of SARS-CoV. A recent report indicated that SARS-CoV has been found in a number of wild animals with 99.8% identity [7]. What would be the role of recombination in the event that created this virus, possibly in a predator animal?

Stavrinides and Guttman [32] have suggested that a possible past recombination event between mammalian-like and avian-like parent viruses is responsible for the evolution of SARS-CoV. In order to further test for the recombination hypothesis, we implemented a 3-step strategy. First, we employed BLAST to determine which viruses (coronaviruses or other viruses) should be included in the sample relevant for recombination detection analysis. Second, we used widely used recombination detection techniques to detect the occurrence of recombination between SARS-CoV and other coronaviruses. Finally, we used phylogenetic tree analysis to confirm the presence of recombination events.

Materials and methods

Sequences

A reference SARS-CoV genome sequence (NC_004718) [17] was downloaded from GenBank. In order to determine which viruses (coronaviruses or other viruses) should be included in the sample relevant for recombination detection analysis, we blasted the whole SARS-CoV sequence against virus database and the result indicated that there are 6 significant hits (at the level of E-value <0.0001. Table 1): Murine hepatitis virus (MHV), Porcine epidemic diarrhea virus (PEDV), Bovine coronavirus (BCoV), Transmissible gastroenteritis virus (TGEV), Avian infectious bronchitis virus (IBV) and Human coronavirus 229E (HCoV). All these sequences were downloaded from GenBank: MHV (AF029248), PEDV (AF353511), BCoV (NC_003045), TGEV (NC_002306), IBV (NC_001451) and HCoV (NC_002645).

Table 1. Search results by BLAST

Virus	Score (bits)	E-value
Murine hepatitis virus	92	2.00E-16
Porcine epidemic diarrhea virus	80	8.00E-13
Bovine coronavirus	72	2.00E-10
Transmissible gastroenteritis virus	58	3.00E-06
Avian infectious bronchitis virus	58	3.00E-06
Human coronavirus 229E	54	4.00E-05
Ovine astrovirus	48	0.003
Streptococcus pyogenes	44	0.043
Saccharomyces cerevisiae chromosome	42	0.17
Saccharomyces cerevisiae chromosome	40	0.67
Equine rhinitis B virus	40	0.67
Equine rhinovirus 3	40	0.67
Callitrichine herpesvirus 3	40	0.67
Turkey astrovirus	40	0.67
Amsacta moorei entomopoxvirus	40	0.67
Salmonella typhimurium bacteriophage	38	2.7
Goatpox virus	38	2.7
Bacteriophage SPBc2	38	2.7
Saccharomyces cerevisiae chromosome	38	2.7
Shrimp white spot syndrome virus	38	2.7
Tupaia paramyxovirus	38	2.7
Rachiplusia ou multiple nucleohedrovirus	38	2.7
Lumpy skin disease virus	38	2.7
Sheeppox virus	38	2.7
Human papillomavirus type 59	38	2.7
Citrus tristeza virus	38	2.7
Pseudomonas phage phiKZ	38	2.7

Recombination detection and phylogenetic analysis

There are a number of methods and software packages that have been developed for detection of recombination events in DNA sequences. The performance of these methods has been extensively evaluated and compared on simulated and real data [23, 24]. In the present study we applied these methods to RNA viruses. SARS-CoV and other 6 coronavirus genomes (SARS-CoV, IBV, BCoV, HCoV, MHV, PEDV, TGEV) were first aligned using CLUSTALW [33]. Sites with gaps were removed and a 25077-nt alignment was generated. Subsequently, seven methods were employed to detect the occurrence of recombination (see corresponding reference in parenthesis for details of each method): BOOTSCAN [26], GENECONV [28], DSS (Difference of Sums of Squares) [20], HMM (Hidden Markov Model) [8], MAXCHI (Maximum Chi-Square method) [19], PDM (Probabilistic Divergence Measures) [9], RDP (Recombination Detection Program) [18].

BOOTSCAN, MAXCHI and RDP are implemented in RDP software package, http://web.uct.ac.za/depts/microbiology/microdescription.htm. GENECONV is implemented in the program, http://www.math.wustl.edu/∼sawyer/geneconv/. DSS, HMM and PDM are implemented in TOPALi software package, http://www.bioss.sari.ac.uk/software.html.

Basically default parameter settings were used in all the programs, except the following values: gscale = 1 (GENECONV), internal and external references (RDP), window size = 300 and step = 10 (DSS, HMM and PDM).

After potential recombination events were identified by at least 3 methods above, separate neighbor joining trees were constructed for each putative recombination region to better evaluate the evidence for conflicting evolutionary histories of different sequence regions. All trees were produced with TOPALi mentioned above.

Results

Recombination detection

Table 2 summarizes the results of BOOTSCAN analysis with 100% bootstrap support and significant P-value (<0.05 for uncorrected and MC corrected P-value). Two regions (13151–13299 and 16051–16449, position in alignment) are identified as putative recombination regions and all 6 coronaviruses are potential parents with SARS-CoV as potential daughter.

GENECONV detected 9 putative recombination events occurred in a wide range of positions 5941–24997 (in alignment) at a significant level $p < 0.05$ for two P-values: simulated P-value (based on 10,000 permutations) and BLAST-like BC KA P-value (Table 3). All 6 coronaviruses are potential parents with SARS-CoV as potential daughter.

MAXCHI identified 15 putative recombination events (Table 4, possible misidentification events are not retained). Most of the breakpoints are significant at about 0.001 level; the position located in alignment spans from 3534 to 22840, but some beginning or ending breakpoints are not determined. Similarly, 6 coronaviruses are potential parents with SARS-CoV as potential daughter.

RDP revealed that 6 putative recombination events occur in the domain of alignment 5910–13334 (Table 5), with the uncorrected and MC corrected p-value at less than 0.002 and 0.05 respectively. In this case, 4 coronaviruses (IBV, BCoV, MHV and PEDV) are potential parents with SARS-CoV as potential daughter.

Figure 1 shows the DSS profiles of putative breakpoints between SARS-CoV and other coronaviruses (Dotted line indicates the 95 percentile under the null hypothesis of no recombination): SARS-CoV, IBV, BCoV and MHV (Fig. 1a), SARS-CoV, MHV, PEDV and TGEV (Fig. 1b), SARS-CoV, IBV, HCoV and TGEV (Fig. 1c). There are about 6 different breakpoints (significant peaks): 13614 and 16085 (Fig. 1a), 11008 and 12850 (Fig. 1b), 12805, 13614 and 16444 (Fig. 1c).

HMM plots for SARS-CoV, IBV, BCoV and HCoV (Fig. 2) revealed that the putative breakpoints are at about position 5500 and 19000. There is a clear transition from state 1 (SARS-CoV grouped with IBV) (Fig. 2a) into state 3 (SARS-CoV grouped with HCoV) (Fig. 2c). The region between 5500 and 19000 is noisy, and at this moment no information can be provided by HMM.

Figure 3 shows the results of PDM analysis performed on SARS-CoV and other coronaviruses (dotted line indicates the 95% critical region for the null

Table 2. Recombination regions identified by BOOTSCAN method

Identified by:	Daughter	Major parent	Minor parent	Beginning in alignment	Ending in alignment	Uncorrected P-Value	MC corrected P-Value	Bootstrap support (%)
Bootscan	SARS	IBV	PEDV	13151	13299	0.001	0.035	100
Bootscan	SARS	IBV	HCoV	16351	16449	0.001	0.035	100
Bootscan	SARS	BCoV	TGEV	16051	16199	0.001	0.035	100

Table 3. Recombination regions identified by GENECONV method

Identified by:	Daughter	Parent	Beginning in alignment	Ending in alignment	Simulated P-Value	BC KA P-Value
GENECONV	SARS	IBV	24970	24997	0.0001	0.00003
GENECONV	SARS	IBV	20708	20727	0.0156	0.0172
GENECONV	SARS	BCoV	12102	12135	0.0329	0.04634
GENECONV	SARS	BCoV	11977	12024	0.0051	0.00509
GENECONV	SARS	BCoV	5941	5965	0.0051	0.00509
GENECONV	SARS	HCoV	10491	10524	0.0033	0.00361
GENECONV	SARS	MHV	12595	12664	0.0185	0.01999
GENECONV	SARS	PEDV	13208	13263	0.0076	0.00827
GENECONV	SARS	TGEV	8399	8425	0.0315	0.02951

Table 4. Recombination regions identified by MAXCHI method

Identified by:	Daughter	Major parent	Minor parent	Beginning in alignment	Ending in alignment	Beginning breakpoint P-Value	Ending breakpoint P-Value
Maxchi	SARS	PEDV	TGEV	9052	9066	0.028108	0.00065
Maxchi	SARS	IBV	HCoV	undetermined	5486	–	0.000336
Maxchi	SARS	HCoV	IBV	14026	undetermined	0.000913	–
Maxchi	SARS	PEDV	TGEV	10668	undetermined	0.000957	–
Maxchi	SARS	Unknown (MHV)	IBV	20676	22840	0.000913	0.000913
Maxchi	SARS	Unknown (MHV)	IBV	undetermined	8996	–	0.000957
Maxchi	SARS	MHV	BCoV	16609	undetermined	0.000913	–
Maxchi	SARS	MHV	BCoV	20514	undetermined	7.75E-06	–
Maxchi	SARS	MHV	HCoV	undetermined	3534	–	0.000336
Maxchi	SARS	PEDV	HCoV	18528	undetermined	0.001015	–
Maxchi	SARS	PEDV	HCoV	undetermined	7281	–	0.00065
Maxchi	SARS	PEDV	HCoV	15742	15763	0.001015	0.009907
Maxchi	SARS	HCoV	PEDV	9137	9156	0.000913	0.010587
Maxchi	SARS	PEDV	HCoV	5474	undetermined	0.000957	–
Maxchi	SARS	HCoV	TGEV	12854	undetermined	0.000253	–

Table 5. Recombination regions identified by RDP method

Identified by:	Daughter	Major parent	Minor parent	Beginning in alignment	Ending in alignment	Uncorrected P-Value	MC corrected P-value
RDP	SARS	IBV	BCoV	5910	6111	5.18E-04	1.81E-02
RDP	SARS	IBV	BCoV	6136	6286	1.56E-05	5.45E-04
RDP	SARS	IBV	MHV	6134	6326	1.28E-03	4.49E-02
RDP	SARS	BCoV	PEDV	13151	13280	3.32E-04	1.16E-02
RDP	SARS	MHV	PEDV	9196	9334	1.72E-05	6.03E-04
RDP	SARS	MHV	PEDV	13152	13334	3.89E-05	1.36E-03

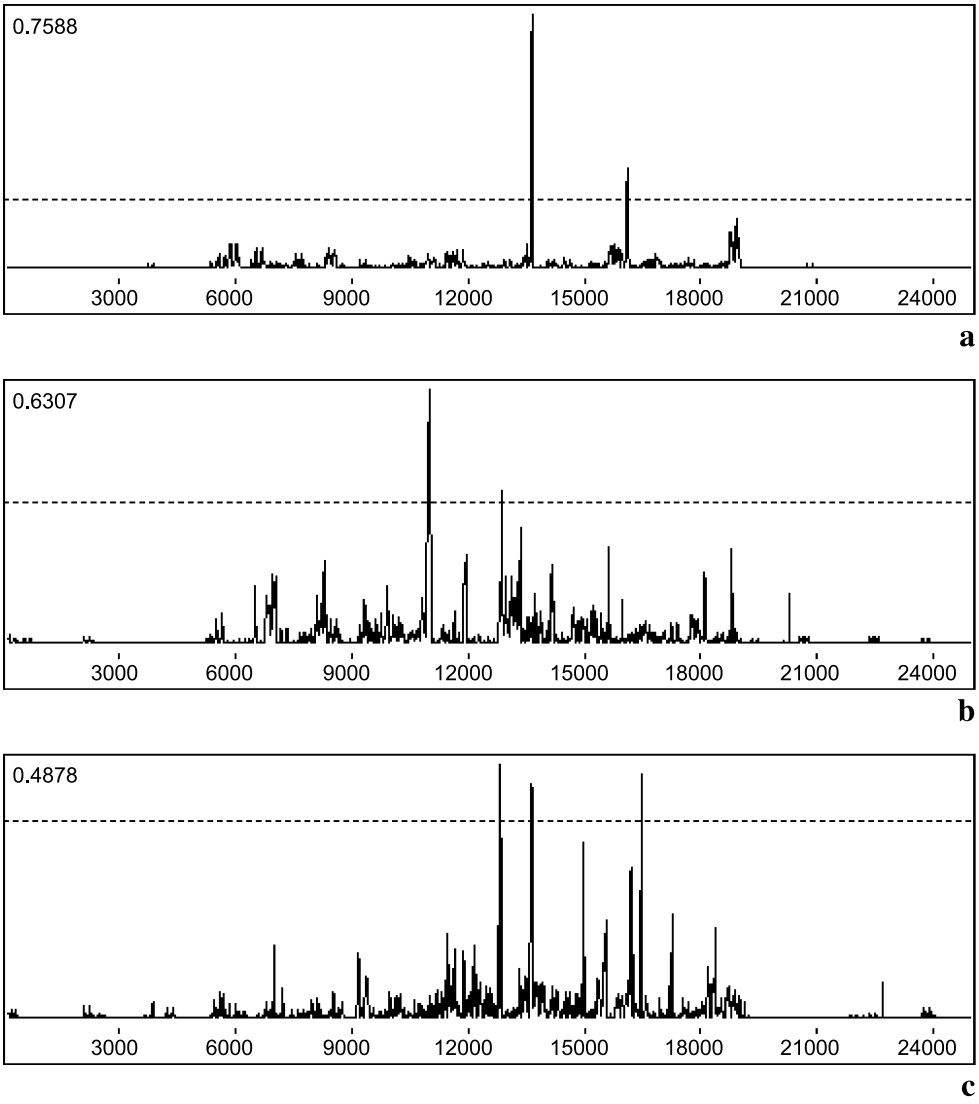


Fig. 1. Predicting recombination regions with DSS (Difference of Sums of Squares) implemented in TOPALi. Default parameter values were used except for the Fitch method, where a window size = 300 and step = 10 were chosen. The horizontal axis represents the site in the alignment, the vertical axis represents the DSS statistic, and the dotted line shows the 95 percentile under the null hypothesis of no recombination. SARS-CoV, IBV, BCoV and MHV for Fig. 1a, SARS-CoV, MHV, PEDV and TGEV for Fig. 1b, and SARS-CoV, IBV, HcoV and TGEV for Fig. 1c, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, PEDV-porcine epidemic diarrhea virus, TGEV-transmissible gastroenteritis virus, BCoV-bovine coronavirus, HCoV-human coronavirus, MHV-murine hepatitis virus, and IBV-avian infectious bronchitis virus

hypothesis of no recombination): SARS-CoV, IBV, BCoV and MHV (Fig. 3a, b), SARS-CoV, MHV, PEDV and TGEV (Fig. 3c, d), SARS-CoV, BCoV, HCoV and MHV (Fig. 3e, f). A number of breakpoints (pronounced peaks) could be concurred: 6380, 13479, 18915 and 20263 (Fig. 3a, b), 1753, 5032, 9256, 10289,

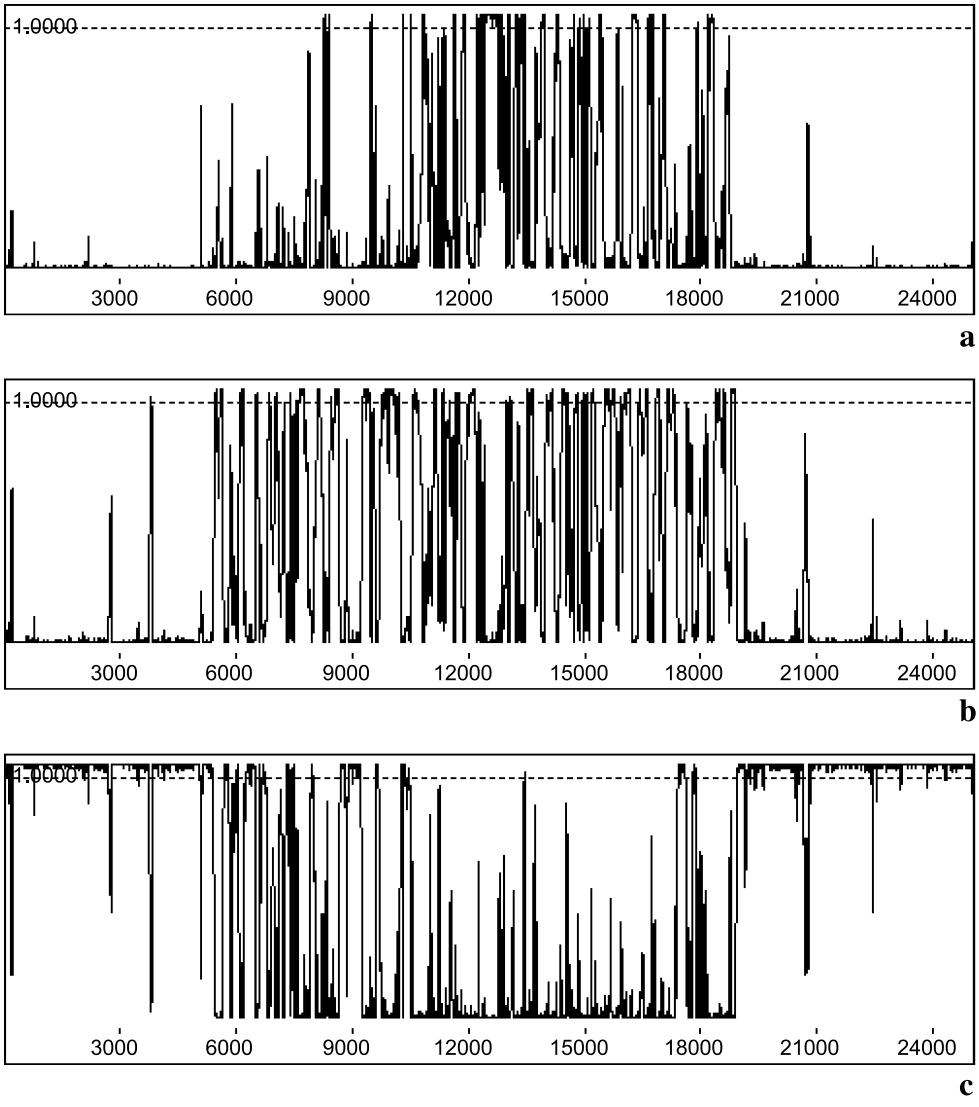


Fig. 2. Predicting recombination regions with HMM (Hidden Markov Model) implemented in TOPALi. Default parameter values were used. The horizontal axis represents the site in the alignment, the vertical axis represents the probability for topology change, and the dotted line shows the 95 percentile under the null hypothesis of no recombination. SARS-CoV, IBV, BCoV and HCoV was used, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, BCoV-bovine coronavirus, HCoV-human coronavirus, and IBV-avian infectious bronchitis virus

15591, 19050 and 22195 (Fig. 3c, d), 1393, 6111, 16624, 19859 and 20802 (Fig. 3e, f).

Posada [23] suggested that one should not rely too much on a single method for recombination detection. Here we consider the regions identified by at least 3 methods as putative recombination regions. The results are summarized in Table 6. Seven putative recombination regions span a range of positions in SARS-CoV

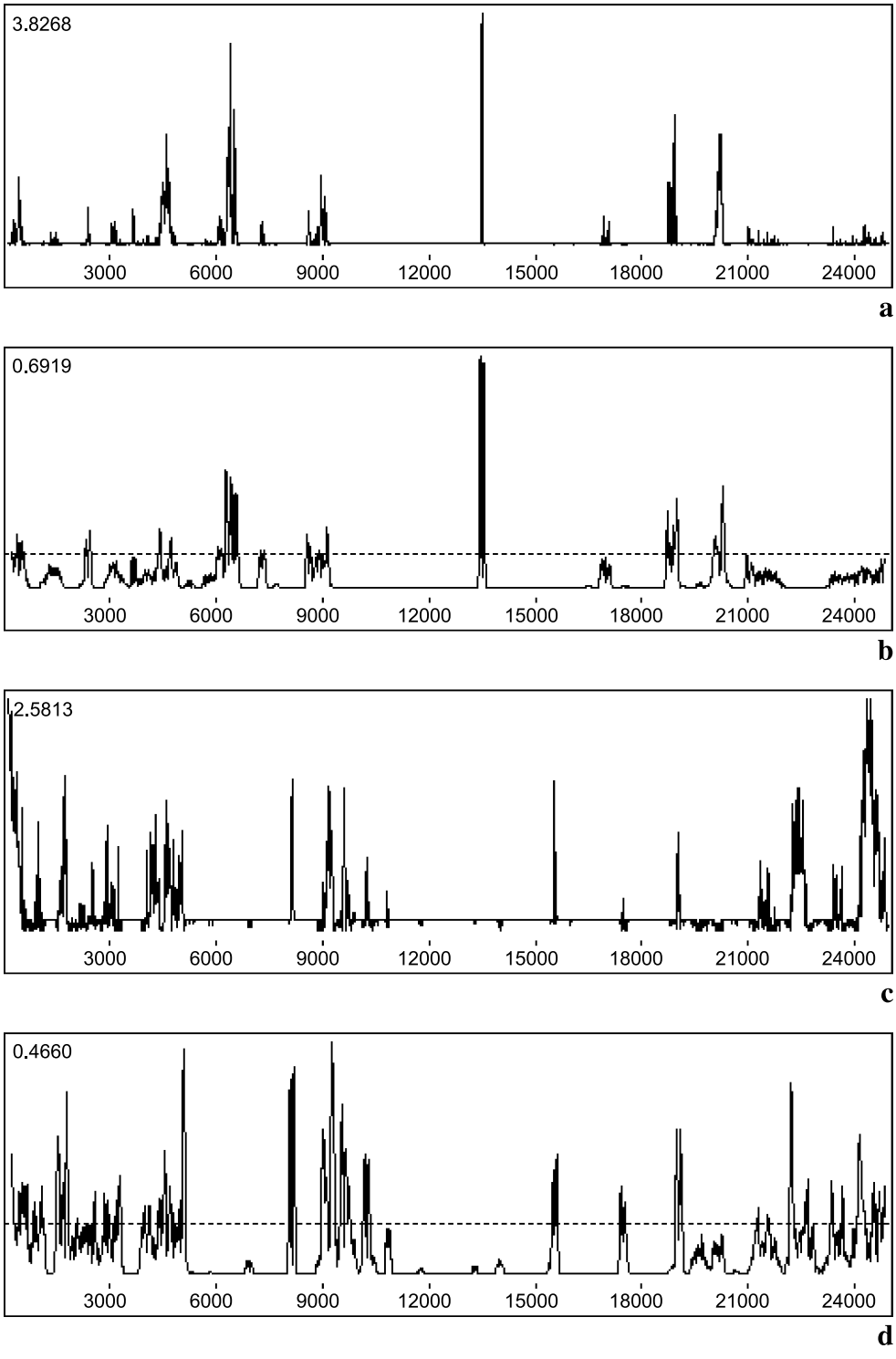


Fig. 3 (*continued*)

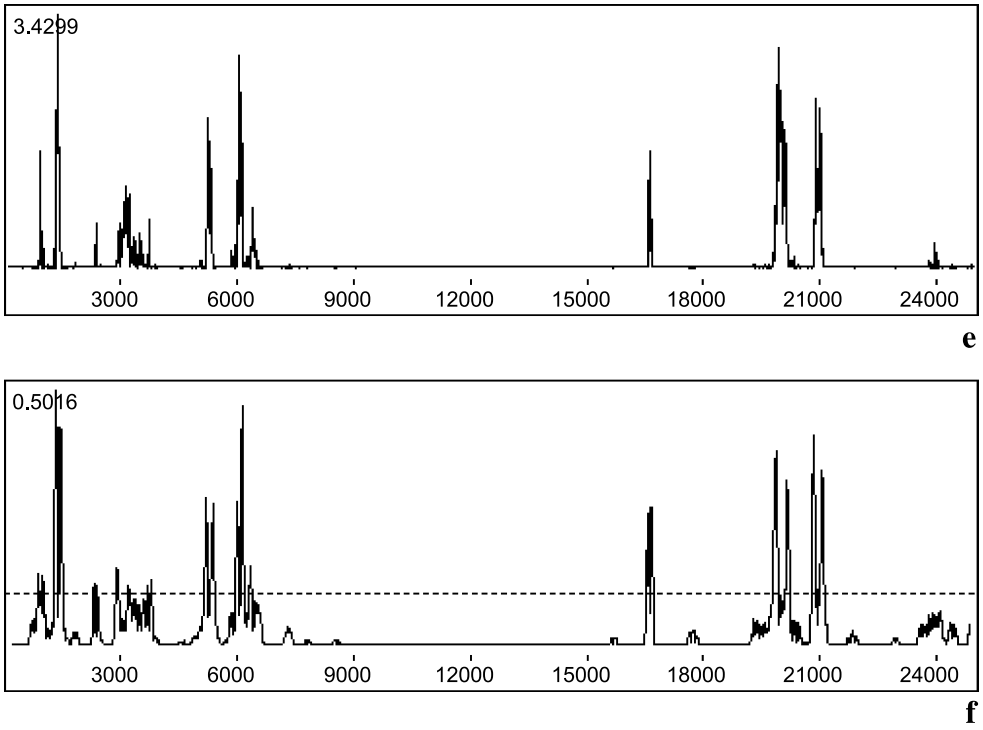


Fig. 3. Predicting recombination regions with PDM (Probabilistic Divergence Measures) implemented in TOPALi. Default parameter values were used with the exception that window size = 300 and step = 10 were used. The horizontal axis represents the site in the alignment, the vertical axis represents the global and local divergence measures, and the dotted line shows the 95% critical region for the null hypothesis of no recombination. SARS-CoV, IBV, BCoV and MHV for Fig. 3**a**, **b**, SARS-CoV, MHV, PEDV and TGEV for Fig. 3**c**, **d**, and SARS-CoV, BCoV, HcoV and MHV for Fig. 3**e**, **f**, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, PEDV-porcine epidemic diarrhea virus, TGEV-transmissible gastroenteritis virus, BCoV-bovine coronavirus, HCoV-human coronavirus, MHV-murine hepatitis virus, and IBV-avian infectious bronchitis virus

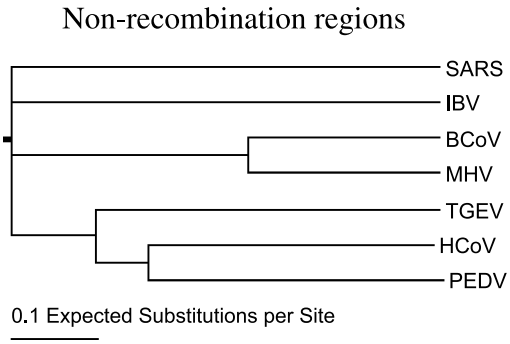
genome from 7475–24133. These regions are separately extracted for phylogenetic analysis.

Phylogenetic analysis

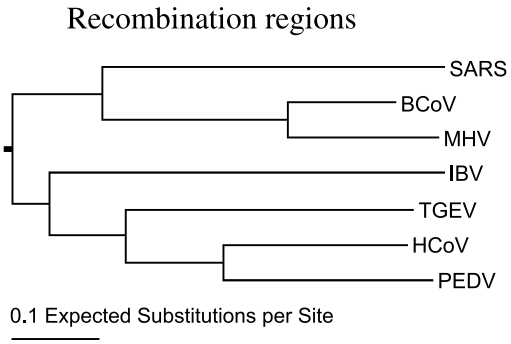
Phylogenetic trees constructed by using putative recombination regions and non-recombination regions identified by above techniques are shown in Figure 4. The left panels stand for non-recombination regions and the right panels for recombination regions. We compared each row of figures and found that the phylogenetic tree in the left panel (non-recombination region) had very different topology when compared to the phylogenetic tree in the right panel (recombination region), which indicates that recombination has occurred. For example, in Fig. 4a, 7 coronaviruses are divided into 4 groups: group 1 for TGEV, HCoV and PEDV, group 2 for BCoV and MHV, group 3 for IBV, and group 4 for SARS-CoV, consistent with Marra et al. [17]; while in Fig. 4b, 7 coronaviruses are divided

Table 6. Recombination regions identified by 7 methods

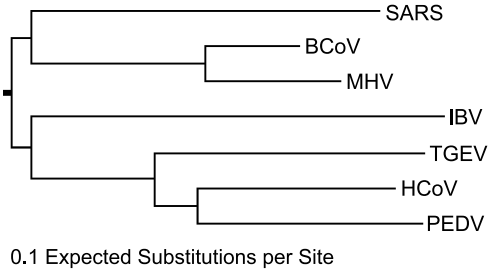
Number of methods	Identified by	Beginning in alignment	Ending in alignment	Beginning in SARS genome	Ending in SARS genome	Protein
5	GENECONV, HMM, MAXCHI, PDM, RDP	5474	6470	7475	8588	replicase 1A
3	MAXCHI, PDM, RDP	9052	9334	11318	11631	replicase 1A
4	DSS, GENECONV, MAXCHI, PDM	10491	10963	12821	13296	replicase 1A
3	DSS, GENECONV, MAXCHI, PDM	12102	12854	14490	15245	replicase 1B
5	BOOTSCAN, DSS, GENECONV, PDM, RDP	13151	13614	15542	16005	replicase 1B
4	BOOTSCAN, DSS, MAXCHI, PDM	16051	16624	18478	19076	replicase 1B
4	GENECONV, HMM, MAXCHI, PDM	19000	20727	21579	24133	spike



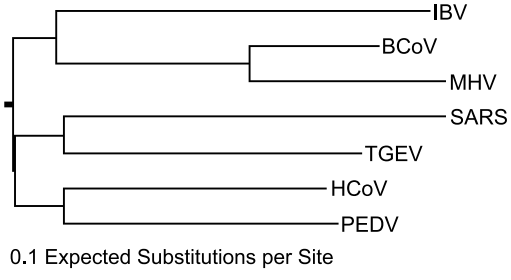
a Region 1-5473



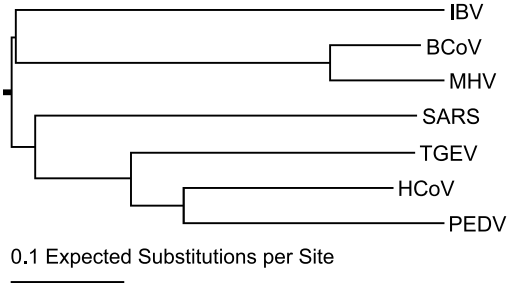
b Region 5474-6470



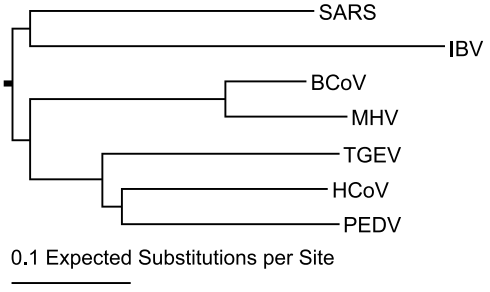
c Region 6471-9051



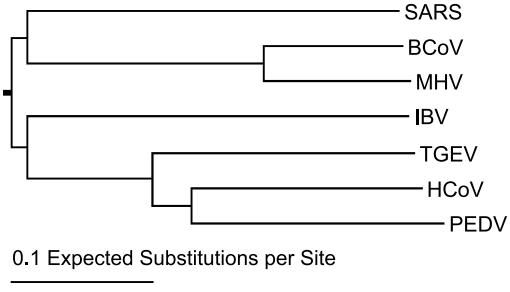
d Region 9052-9334



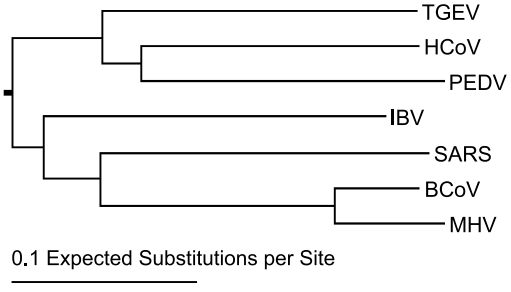
e Region 9335-10490



f Region 10491-10963



g Region 10964-12101



h Region 12102-12854

Fig. 4 (*continued*)

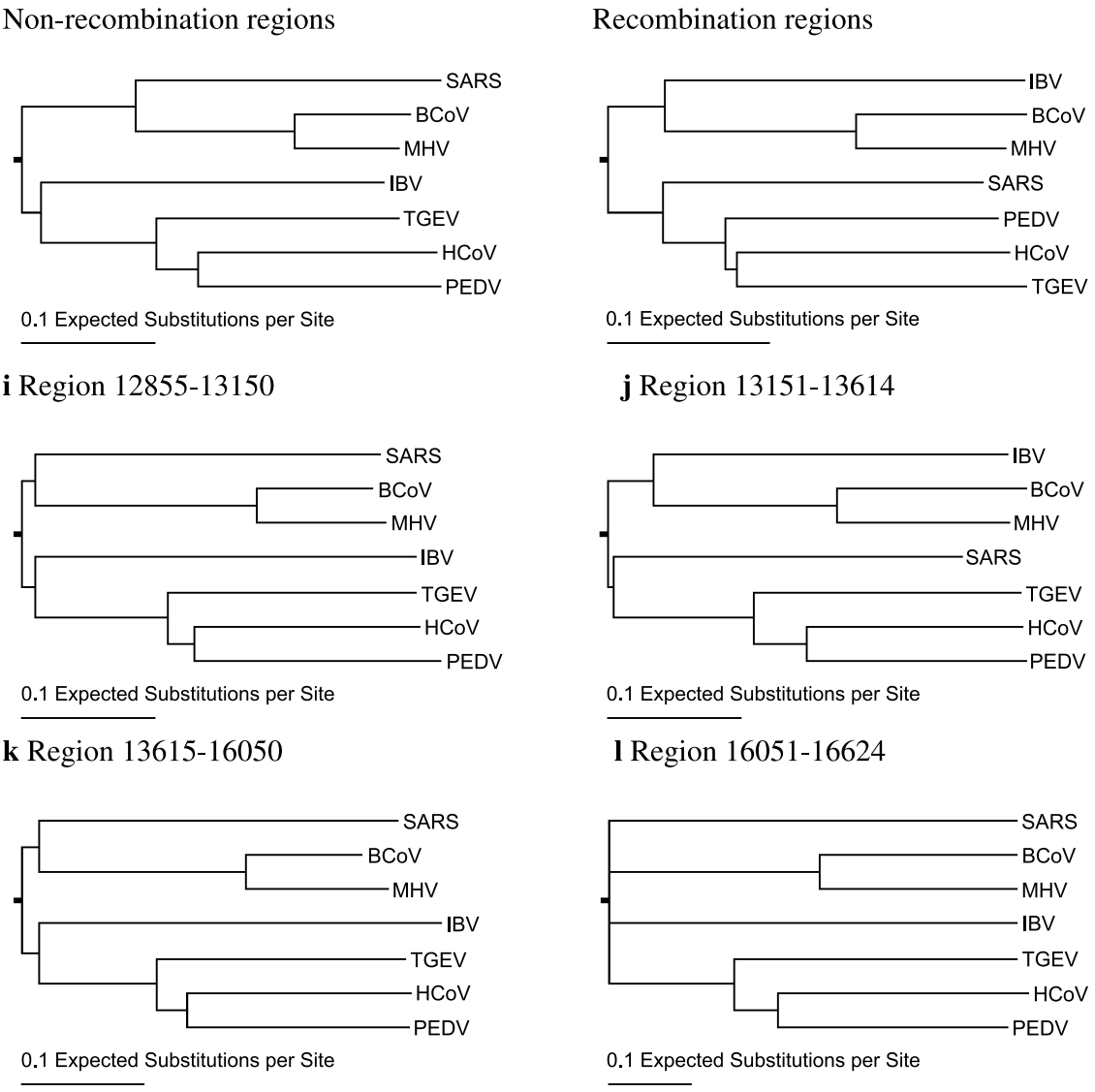


Fig. 4. Phylogenetic analysis of putative recombination regions. Neighbour joining trees were constructed by TOPALi. The sequence region in the alignment used for each tree is written below each figure. The phylogenetic trees in the left panel correspond to non-recombination region and the phylogenetic trees in the right panel correspond to recombination region. All branch lengths are drawn to scale. Six coronaviruses (IBV, BCoV, HCoV, MHV, PEDV and TGEV) are potential parents of SARS-CoV, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, PEDV-porcine epidemic diarrhea virus, TGEV-transmissible gastroenteritis virus, BCoV-bovine coronavirus, HCoV-human coronavirus, MHV-murine hepatitis virus, and IBV-avian infectious bronchitis virus

into 2 groups: group 1 for IBV, TGEV, HCoV and PEDV, group 2 for BCoV, MHV and SARS-CoV, suggests that SARS-CoV is most closely related to BCoV and MHV, which is consistent with a recent report [29]. At the same time, SARS-CoV

is also most closely related to TGEV (Fig. 4d) and IBV (Fig. 4f). Thus, phylogenetic analysis substantiates the presence of recombination events in the history that led to the SARS-CoV genome.

Discussion

In this study, seven recombination detection methods and phylogenetic analyses were performed on SARS-CoV and the six coronaviruses identified by BLAST (IBV, BCoV, HCoV, MHV, PEDV and TGEV). These techniques successfully identified recombination events in bacteria and viruses [2, 3, 6, 21, 26, 39]. Our analysis concurred to suggest the occurrence of recombination events between ancestors of SARS-CoV and these 6 coronaviruses. Indeed, pairwise alignment showed that many segments of high homology with IBV, BCoV, HCoV, MHV, PEDV and TGEV do exist in SARS-CoV genome, Table 7 exhibits the segments with length >20 nt and identiy $>80\%$, and Fig. 5 shows the mosaic structure of the region 14930–15908 in SARS-CoV genome based on the segments with length >50 and identity $>80\%$. Of course, the other coronaviruses used in the analysis are also mosaic structures, for more sequence similarities exist among them than with SARS-CoV.

It is noted that all the sequence comparisons in this study are based on nucleotide sequences. While the protein sequences in SARS-CoV are largely different from those in the known three groups of coronavirus [17], such as, for S protein, the identity is: 25.9% for SARS-CoV and BCoV, 21.7% for SARS-CoV and HCoV, 21.5% for SARS-CoV and IBV, 25.6% for SARS-CoV and MHV, 20.6% for SARS-CoV and PEDV, 19.4% for SARS-CoV and TGEV. Although SARS-CoV is close to BCoV, MHV, TGEV and IBV, the corresponding protein, replicase 1a, is still different: with identity 27.4% for SARS-CoV and BCoV, 24.8% for SARS-CoV and IBV, 32.2% for SARS-CoV and MHV, 25.0% for SARS-CoV and TGEV.

Naturally, we should take into account the role of convergent evolution, which would bear its mark on the viral genome. The recombination events that we witnessed in SARS-CoV are present in six different viruses, suggesting sequential horizontal transfers and progressive adaptation to new hosts cells or animals. Indeed because viruses need both receptors to permeate host cells and resist the immune response of the host, their outer layer proteins are submitted to an extremely strong selection pressure that may restrict considerably the possible variations of the corresponding proteins (and accordingly of the corresponding genome pieces of sequences). It is nevertheless remarkable that, despite the inclusion of all possible types of viruses in our sample set (as well as shuffled genomes from the viruses we have identified as relevant) we find a more or less single category of viruses as similar to SARS-CoV. This suggests that even if the contribution of convergent evolution is important, this happened on a more or less common phylogenetic background, suggesting several steps of recombination followed by fine adaptation. In this context, we would like to suggest that ancestors of PEDV, MHV or both are the most plausible origin of SARS-CoV. Guan et al. [7]

Table 7. Mosaic segments in SARS-CoV genome (length >20 nt and identity >80%)

Beginning in SARS	Ending in SARS	Length	Identity	Match percent (%)	Source
10063	10109	47	41	88	MHV
10609	10641	33	30	91	TGEV
12821	12854	34	31	92	HCoV
13844	13879	36	32	89	BCoV
13845	13879	35	31	89	MHV
13986	14011	26	26	100	PEDV
14365	14412	48	45	94	BCoV
14367	14395	29	27	94	MHV
14490	14523	34	32	95	BCoV
14589	14632	44	38	87	BCoV
14685	14729	45	39	87	MHV
14724	14746	23	22	96	IBV
14808	14835	28	26	93	HCoV
14913	14947	35	31	89	HCoV
14933	15070	138	112	82	BCoV
14982	15091	110	89	81	IBV
14986	15055	70	64	92	MHV
15062	15093	32	29	91	HCoV
15123	15173	51	43	85	TGEV
15210	15232	23	22	96	PEDV
15210	15238	29	27	94	BCoV
15210	15253	44	40	91	IBV
15417	15482	66	57	87	BCoV
15417	15457	41	37	91	IBV
15420	15479	63	55	88	MHV
15611	15682	72	64	89	PEDV
15624	15670	47	42	90	HCoV
15633	15672	40	35	88	TGEV
15729	15770	42	40	96	MHV
15765	15817	53	46	87	HCoV
15852	15908	57	49	86	MHV
17088	17125	38	35	93	IBV
17688	17714	27	25	93	TGEV
17757	17800	44	39	89	PEDV
17783	17809	27	25	93	HCoV
18558	18577	20	20	100	PEDV
18771	18847	77	65	85	TGEV
18784	18833	50	44	88	HCoV
19102	19132	31	29	94	IBV
19113	19132	20	20	100	HCoV
19146	19252	107	87	82	MHV
19201	19252	52	45	87	IBV
19206	19253	48	44	92	BCoV
19396	19420	25	24	96	MHV

(*continued*)

Table 7 (*continued*)

Beginning in SARS	Ending in SARS	Length	Identity	Match percent (%)	Source
19396	19420	25	24	96	BCoV
19517	19564	48	42	88	MHV
19548	19588	41	37	91	TGEV
20709	20746	38	34	90	MHV
20712	20747	36	33	92	IBV
20793	20839	47	41	88	HCoV
20797	20827	31	28	91	PEDV
25062	25084	23	22	96	IBV
25068	25088	21	21	100	MHV
25068	25090	23	22	96	TGEV
25068	25090	23	22	96	BCoV
29593	29621	29	29	100	IBV

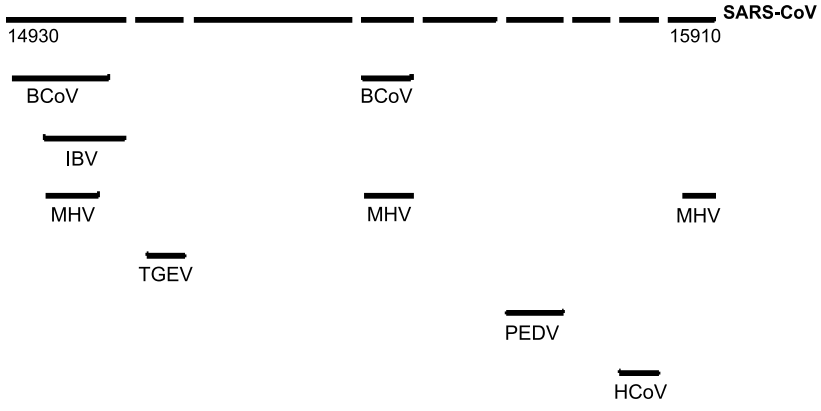


Fig. 5. Mosaic structure of the region 14930–15908 in SARS-CoV genome. Six coronaviruses (IBV, BCoV, HCoV, MHV, PEDV and TGEV) are potential parents of SARS-CoV, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, PEDV-porcine epidemic diarrhea virus, TGEV-transmissible gastroenteritis virus, BCoV-bovine coronavirus, HCoV-human coronavirus, MHV-murine hepatitis virus, and IBV-avian infectious bronchitis virus

indicated that there are 38 nucleotide polymorphisms (26 of them are non-synonymous) in the S genes of human SARS-CoV viruses compared to animal SARS-CoV-like viruses, although the additional 29 nucleotide sequence in the animal viruses exists in ORF10, not in the S protein. These polymorphisms could be responsible for changes in host range and tissue tropism among coronaviruses, for a single nucleotide change can dramatically alter the behaviour of the virus [35].

Based on phylogenetic techniques and BOOTSCAN recombination analysis Stavrinides and Guttman [32] indicated that the replicase of SARS-CoV was a mammalian-like origin, the M and N proteins have an avian-like origin, and the S protein has a mammalian-avian mosaic origin. While in the present study

we used phylogenetic analysis and 7 recombination detection methods, including the powerful methods of MAXCHI and GENECONV among 14 methods studied (SIMPLOT (BOOTSCAN), GENECONV, HOMOPLASY TEST, PIST, MAXCHI, CHIMAERA, PHYPRO, PLATO, RDP, RECPARS, RETICULATE, RUNS TEST, SNEATH TEST, TRIPLE) [23, 24], to conduct whole genome-wide recombination analysis. We identified seven putative recombination regions, which encompass, in terms of proteins involved, replicase 1A, replicase 1B and the spike glycoprotein. Stavrinides and Guttman [32] primarily inferred the occurrence of recombination qualitatively, but did not identify the precise recombination region in the protein involved (the S protein is an exception, they identified a recombination region in S protein, located between nucleotides 2472 and 2694 of the S protein, i.e. between nucleotides 23963 and 24185 of the SARS-CoV genome, basically covered by the last recombination region for S protein (Table 6)). Most importantly, each of our recombination regions is identified by at least 3 methods, because one should not rely too much on a single method, as suggested in [23]. In general, we believe two studies lead to the overall conclusion: the evolution of SARS-CoV has involved recombination.

The recombination event in the replicase is related to the fact that the RNA polymerase of coronaviruses utilize a discontinuous transcription mechanism to synthesize mRNAs. The viral polymerase must jump between different RNA templates regularly during positive- or negative-strand RNA synthesis and depending on the rejoining sites, the resultant RNA recombination will be either homologous or nonhomologous. This is the copy-choice model of recombination in RNA viruses [13, 27, 31, 34]. The recombination event in S protein is certainly important since this allows the virus to alter surface antigenicity and escape immunesurveillance in the animals, thus adapting to a human host.

The existence of SARS-CoV-like viruses (99.8% homology to human SARS-CoV) in several wild animals in a live animal market in Guangdong [7] indicated that interspecies transmission among the human and animal SARS-CoV-like viruses had occurred. The mutation analysis of sequence variations among these isolates will help identify the genetic signature of SARS virus strains when a sufficient amount of sequence data is available.

The very fact that several species of animals are affected does not allow one to trace directly the origin of the virus as endemic in one of these species, but, rather, might be indicative that animals and men might have been contaminated by a virus from a common origin, presumably located in animal food present in local markets in the Guangdong province. Investigating a wide variety of animal coronaviruses, especially in relation to rodents, birds, snakes and farm animals, would be interesting with regard to the origin of the SARS-CoV that caused disease in humans.

Finally, a challenging question arises. What is the molecular basis of recombination in SARS-CoV? Many requirements are needed for recombination to occur: (1) Two coronaviruses can infect a host simultaneously and continue to replicate without interference with each other; (2) Sufficient nucleotide identity between these genomes is essential for genome-switching to occur during RNA replication; (3) The proteins arising from recombination must be functional; (4) The recombinant virus must have some selective advantage for its survival. That

is, the recombination that creates a successful "new" coronavirus is probably a rare event. So, we must stress that the potential recombination events in SARS-CoV, identified in the present study, are most likely "old" events, which may represent the events that occurred thousands of years ago. Although the recent findings indicated that SARS-CoV did exist in a number of wild animals [7], we have not yet determined where these SARS-CoV-like virus strains come from.

Acknowledgement

We wish to thank the Hong Kong Innovation and Technology Fund for supporting the present research.

References

1. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R (2003) Coronavirus main proteinase ($3CL^{pro}$) structure: basis for design of anti-SARS drugs. Science 300: 1763–1767
2. Anderson JP, Rodrigo AG, Learn GH, Madan A, Delahunty C, Coon M, Girard M, Osmanov S, Hood L, Mullins JI (2000) Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. J Virol 74: 10752–10765
3. Carr JK, Salminen MO, Koch C, Gotte D, Artenstein AW, Hegerich PA, Louis DSt, Burke DS, McCutchan FE (1996) Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. J Virol 70: 5935–5943
4. Cavanagh D, Davis PJ (1988) Evolution of avian coronavirus IBV: sequence of the matrix glycoprotein gene and intergenic region of several serotypes. J Gen Virol 69: 621–629
5. Cavanagh D, Davis PJ, Cook JKA (1992) Infectious bronchitis virus: evidence for recombination within the Massachusetts serotype. Avian Pathol 21: 401–408
6. Gao F, Robertson DL, Morrison SG, Hui H, Craig S, Decker J, Fultz PN, Girard M, Shaw GM, Hahn BH, Sharp PM (1996) The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J Virol 70: 7013–7029
7. 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 JSM, Poon LLM (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302(5643): 276–278
8. Husmeier D, McGuire G (2002) Detecting recombination with MCMC. Bioinformatics 18: S345–S353
9. Husmeier D, Wright F (2001) Probabilistic divergence measures for detecting interspecies recombination. Bioinformatics 17: 1–8
10. Jia W, Karaca K, Parrish CR, Naqi SA (1995) A novel variant of avian infectious bronchitis virus resulting from recombination among three different strains. Arch Virol 140(2): 259–271
11. Kottier SA, Cavanagh D, Britton P (1995) Experimental evidence of recombination in coronavirus infectious bronchitis virus. Virology 213(2): 569–580
12. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W et al. (2003) A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348: 1953–1966
13. Lai MMC (1992) RNA recombination in animal and plant viruses. Microbiol Rev 56: 61–79

14. Lai MMC (1996) Recombination in large RNA viruses: coronaviruses. Semin Virol 7: 381–388
15. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH et al. (2003) Transmission dynamics and control of severe acute respiratory syndrome. Science 300(5627): 1966–1970
16. Markino S, Keck JG, Stohlman SA, Lai MMC (1986) High-frequency RNA recombination of murine coronaviruses. J Virol 57: 729–737
17. Marra MA, Jones SJM, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YSN, Khattra J, Asano JK, Barber SA, Chan SY et al. (2003) The genome sequence of the SARS-associated coronavirus. Science 300: 1399–1404
18. Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16: 562–563
19. Maynard Smith J (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126–129
20. McGuire G, Wright F, Prentice M (1997) A graphical method for detecting recombination in phylogenetic data sets. Mol Biol Evol 14: 1125–1131
21. Millman KL, Tavare S, Dean D (2001) Recombination in the ompA gene but not the omcB gene of Chlamydia contributes to serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism. J Bacteriol 183: 5997–6008
22. Ng TW, Turinici G, Danchin A (2003) A double epidemic model for the SARS propagation. BMC Infect Dis 3: 19
23. Posada D (2002) Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Mol Biol Evol 19: 708–717
24. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci USA 98: 13757–13762
25. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH et al. (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300: 1394–1399
26. Salminen MO, Carr JK, Burke DS, McCutchan FE (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by Bootscanning. AIDS Res Hum Retrovir 11: 1423–1425
27. Sawicki SG, Sawicki DL (1998) A new model for coronavirus transcription. Adv Exp Med Biol 440: 215–219
28. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526–538
29. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WM, 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
30. Snijder EJ, den Boon JA, Horzinek MC, Spaan WJ (1991) Comparison of the genome organization of toro- and coronaviruses: evidence for two nonhomologous RNA recombination events during Berne virus evolution. Virology 180(1): 448–452
31. Spaan W, Delius H, Skinner MA, Armstrong J, Rottier P, Smeekens S, Siddell SG, van der Zeijst B (1984) Transcription strategy of coronaviruses: fusion of non-contiguous sequences during mRNA synthesis. Adv Exp Med Biol 173: 173–186
32. Stavrinides J, Guttman DS (2004) Mosaic evolution of the severe acute respiratory syndrome coronavirus. J Virol 78(1): 76–82
33. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
34. van Marle G, van der Most RG, van der Straaten T, Luytjes W, Spaan WJ (1995) Regulation of transcription of coronaviruses. Adv Exp Med Biol 380: 507–510

35. Vogel G (2003) Flood of sequence data yields clues but few answers. Science 300: 1062–1063
36. Wang L, Junker D, Collisson EW (1993) Evidence of natural recombination within the S1 gene of infectious bronchitis virus. Virology 192(2): 710–716
37. Wang L, Junker D, Hock L, Ebiary E, Collisson EW (1994) Evolutionary implications of genetic variations in the S1 gene of infectious bronchitis virus. Virus Res 34(3): 327–338
38. Wang L, Xu Y, Collisson EW (1997) Experimental confirmation of recombination upstream of the S1 hypervariable region of infectious bronchitis virus. Virus Res 49: 139–145
39. Worobey M, Rambaut A, Holmes EC (1999) Widespread intra-serotype recombination in natural populations of dengue virus. Proc Natl Acad Sci USA 96: 7352–7357

Author's address: Dr. Xue Wu Zhang, HKU-Pasteur Research Centre Ltd., Dexter H.C. Man Building, 8 Sassoon Road, Pokfulam, Hong Kong, P.R. China; e-mail: xwzhang@hkucc.hku.hk