

Initial SARS Coronavirus Genome Sequence Analysis Using a Bioinformatics Platform

Hong Luo, Jingchu Luo*

Centre of Bioinformatics, Peking University
Beijing 100871, China

luojc@pku.edu.cn

Abstract

A dedicated anti-SARS bioinformatics web site was setup in April 2003 at the Centre of bioinformatics (CBI), Peking University (<http://antisars.cbi.pku.edu.cn/>). A special bioinformatics platform was constructed to analyse the sequence and structure data of SARS coronavirus and other viruses. A total file of 32 SARS coronavirus genome sequences was retrieved from GenBank and mismatches in 30 sites were revealed from the result of multiple sequence alignment. The SARS coronavirus genome sequences can be divided into three groups based on the phylogenetic analysis using the data set constructed from the sequence mismatches.

Keywords: SARS, virus, genome, multiple sequence alignment, phylogeny

1 Introduction

The outbreak of Severe Acute Respiratory Syndrome (SARS) starting from southern China in November 2002 has a significant influence not only on public health, but also on daily life and social activities around the world. The identification of SARS coronavirus (SARS-CoV) as the major causative factor of the SARS disease (Poutanen et al. 2003, Ksiazek et al. 2003) is the first step in the long route towards the final understanding of the molecular basis of this new emerging virus. The availability of genome sequences of the SARS-CoV makes it possible for bioinformatics analysis to assist biological and medical experiments. In order to provide the user community for the easy access to the variety of resources as well as the analysis results, a dedicated anti-SARS bioinformatics was setup and has been continuously maintained since April 2003. A special platform for the analysis of the coronavirus sequence and structure data has also been setup using the publicly available software tools including the database query system SRS (Etzold et al. 1996), the European Molecular Biology Open Software Suite (EMBOSS) (Rice et al. 2000), the multiple sequence alignment program ClustalW (Thompson et al. 1994), the phylogenetic analysis package Phylip (Felsenstein 1989), etc. Firstly, the SARS-CoVs as well as other virus sequence data were retrieved from the NCBI nucleic acid and protein sequence databases and constructed as to virus specific datasets which can be searched via the Sequence Retrieval System (SRS) installed in the anti-SARS web server.

A table of the SARS genome sequences with links to the individual entries in SRS was created for the easy access. Multiple sequence alignment was then performed using the ClustalW program to reveal the differences in the SARS-CoV genome sequences. The output of the alignment was also reformatted with the ShowAlign program implemented in EMBOSS. The mismatches among different entries of all SARS genome sequences were picked up to generate a mismatch table. The results of mismatches were then analysed with the Phylogenetic package Phylip. All these analysis results were put online and freely accessible from the CBI anti-SARS web site which is also cross linked in the NCBI special SARS genome

page (<http://www.ncbi.nlm.nih.gov/genomes/SARS/SARS.htm>) and the European Bioinformatics Institute (EBI) SARS page (<http://www.ebi.ac.uk/2can/disease/SARS.html>). A mirror site has recently been created at the Asian Pacific Bioinformatics Network (APBioNet) web site hosted at the Bioinformatics Centre at the National University of Singapore (<http://exon.bic.nus.edu.sg:7070/>). In addition to the genome sequences, the protein sequences such as the spike protein and the 3CL enzyme, encoded by the SARS-CoVs were also analysed with various tools installed locally, or through online Web servers.

Although the spread of the SARS disease is currently under control, the origin of the virus, the mechanism of viral infection and transmission remain mysterious. A report of a new SARS case in Singapore was confirmed on 10 September 2003 (www.who.int). It is still a great challenge for biologists to explore the molecular basis of the SARS coronavirus with available genome sequence data which have already been in the public databases.

2 Materials and Methods

2.1 SARS coronavirus genome sequences

The first genome sequence of the SARS coronavirus TOR2 (accession: AY274119) was submitted to the GenBank on 13 April 2003 by the British Columbia Centre for Disease Control and National Microbiology Laboratory Canada (Marra et al. 2003). On 17 April 2003, the US Centers for Disease Control and Prevention sequenced the Urbani strain (Rota et al. 2003) named after Carlo Urbani, the World Health Organization (WHO) officer who identified SARS and died on 29 Mar 2003 (<http://www.who.int/mediacentre/notes/2003/np6/en/>). A dozen of SARS-CoVs were also contributed by the institutions from Hong Kong, Beijing and Singapore in April and released in early May.

All these complete SARS-CoV and other coronavirus genome sequences were timely retrieved from the nucleotide database using the NCBI Entrez query system. The following search strategy was used to reduce redundant entries such as the partial genome sequences:

```
("sars Coronavirus"[Organism] AND
complete genome) AND
(10000[SLLEN]:50000[SLLEN])
```

A table of SARS-CoVs genome sequences with a four-letter code, abbreviation of data source, entry name, accession number, sequence length, date of first deposit and last release was created (Table 1). The four-letter code was defined to distinguish different sources, i.e., CA01 refers to the sequence submitted by a Canadian institute, SG01 indicates that this SARS-CoV was sequenced in Singapore.

No	Code	Source	Name	Accession	Length	First deposit	Last release
1	CA01	GSC	TOR2	NC_004718	29751	13-Apr-2003	13-Aug-2003
2	US01	US-CDC	Urbani	AY278741	29727	17-Apr-2003	12-Aug-2003
3	HK01	HKU	HKU-39849	AY278491	29742	18-Apr-2003	18-Apr-2003
4	HK02	CUHK	CUHK-W1	AY278554	29736	17-Apr-2003	31-Jul-2003
5	HK03	CUHK	CUHK-Su10	AY282752	29736	24-Apr-2003	07-May-2003
6	BJ01	AMMS/BGI	BJ01	AY278488	29725	17-Apr-2003	01-May-2003
7	BJ02	AMMS/BGI	BJ02	AY278487	29745	17-Apr-2003	05-Jun-2003
8	BJ03	AMMS/BGI	BJ03	AY278490	29740	17-Apr-2003	05-Jun-2003
9	BJ04	AMMS/BGI	BJ04	AY279354	29732	17-Apr-2003	05-Jun-2003
10	GD01	AMMS/BGI	GD01	AY278489	29757	17-Apr-2003	18-Aug-2003
11	GD02	GCH	ZMY 1	AY351680	29749	28-Jul-2003	03-Aug-2003
12	ZJ01	ZP-CDC	ZJ01	AY297028	29715	12-May-2003	19-May-2003
13	SG01	GIS	Sin2500	AY283794	29711	27-Apr-2003	12-Aug-2003
14	SG02	GIS	Sin2677	AY283795	29705	27-Apr-2003	12-Aug-2003
15	SG03	GIS	Sin2679	AY283796	29711	27-Apr-2003	12-Aug-2003
16	SG04	GIS	Sin2748	AY283797	29706	27-Apr-2003	12-Aug-2003
17	SG05	GIS	Sin2774	AY283798	29711	27-Apr-2003	12-Aug-2003
18	EU01	WU	Frankfurt 1	AY291315	29727	06-May-2003	11-Jun-2003
19	EU02	WU	FRA	AY310120	29740	29-May-2003	12-Aug-2003
20	EU03	SRISR	HSR 1	AY323977	29751	22-Jul-2003	22-Jul-2003
21	TW01	NTU	TW1	AY291451	29729	06-May-2003	14-May-2003
22	TW02	TW-CDC	TWC	AY321118	29725	11-Jun-2003	26-Jun-2003
23	TW03	TW-CDC	TWC2	AY362698	29727	05-Aug-2003	13-Aug-2003
24	TW04	TW-CDC	TWC3	AY362699	29727	05-Aug-2003	13-Aug-2003
25	TW05	CMUH	TC1	AY338174	29573	08-Jul-2003	28-Jul-2003
26	TW06	CMUH	TC2	AY338175	29573	09-Jul-2003	28-Jul-2003
27	TW07	CMUH	TC3	AY348314	29573	23-Jul-2003	29-Jul-2003
28	TW08	NHRI	TWH	AP006557	29727	30-Jul-2003	02-Aug-2003
29	TW09	NHRI	TWJ	AP006558	29725	30-Jul-2003	02-Aug-2003
30	TW10	NHRI	TWK	AP006559	29727	30-Jul-2003	02-Aug-2003
31	TW11	NHRI	TWS	AP006560	29727	30-Jul-2003	02-Aug-2003
32	TW12	NHRI	TWY	AP006561	29727	30-Jul-2003	02-Aug-2003

Table 1: List of 32 SARS coronavirus genome sequences

Abbreviation of the institutions as data source in Table 1 was used and the full name of each institution was listed in Table 2. Names of the entries were taken from the original authors who deposited the data to GenBank.

Source	Description
GSC	Genome Sciences Centre, Canada
USCDC	Centers for Disease Control and Prevention, USA
HKU	The University of Hong Kong
CUHK	The Chinese University of Hong Kong
AMMS	Academy of Military Medical Sciences, China
BGI	Beijing Genomics Institute, China
GCH	Guangzhou Children Hospital, China
ZPCDC	Zhejiang Provincial Center for Disease Prevention and Control, China
GIS	Genome Institute of Singapore
WU	University of Wuerzburg, Germany
SRISR	Scientific Research Institute San Raffaele, Italy
NTU	National Taiwan University
NHRI	National Health Research Institutes, Taiwan
CMUH	China Medical University Hospital, Taiwan
TWCDC	Centers for Disease Control, Taiwan

Table 2 Source of data

Currently, this table comprises of 32 entries with all available SARS-CoVs genome sequences released by 18 Aug 2003. We use TOR2 (accession: NC_004718) as the reference entry which has been annotated by NCBI and frequently updated at the Entrez sequence history page: http://www.ncbi.nlm.nih.gov/entrez/sutils/girevhist.cgi?v al=NC_004718. The identical entry (accession: AY274119) was not included.

2.2 Sequence retrieve with SRS

In addition to the SARS-CoVs, datasets of other viruses were also retrieved from NCBI, which may help for comparative analysis. We use the well-known Sequence Retrieval System (SRS) as the database query system. SRS was originally developed by Etzold and colleagues (Etzold et al. 1996) at the European Bioinformatics Institute and now is a commercial package for industrial enterprises. It remains free for academic with a licence agreement. It is one of the most popular molecular biology database query system installed in dozens of bioinformatics centres around the world (Zdobnov et al. 2002, Luo 2000). The latest version is 7.1 is well configured by which the administrator can easily index most of the widely used bioinformatics databases. The configuration files are written with a script language called ICARUS developed by the SRS team.

We installed SRS in the local server with groups of dataset including the coronavirus genome sequences (CVGDB),

coronavirus nucleotide sequences (CVDB), coronavirus protein sequences (CVPDB), as well as other virus sequences and 3D structure of viral proteins.

SRS is a powerful tool to manage, retrieve and analyse the molecular biology data including DNA and protein sequences, three dimensional structure of protein and other molecules, functional information of protein families, literature abstracts, etc. Analysis tools can also be integrated in SRS. ClustalW for multiple sequence alignment and PrositeSearch for sequence pattern detection were installed in our SRS system. The SARS-CoV genome sequence table was created with SRS and has cross-links to the individual entries in SRS for easy access.

2.3 Multiple sequence alignment

Multiple sequence alignment was performed to explore the differences of the SARS-CoVs. We select ClustalW as the multiple sequence alignment program which is commonly used and freely available in the bioinformatics community. The results of multiple sequence alignment of 32 SARS-CoVs genome sequences was reformatted for manual inspection using the ShowAlign program implemented in the EMBOSS package. Both ClustalW and ShowAlign output files are published in the CBI anti-SARS web site.

2.4 Phylogenetic analysis

To explore the mutation rate of the SARS-CoV, the single strand RNA virus, mismatch sites among all the above SARS-CoV genome sequences were inspected based on the results of multiple sequence alignment. Considering the factor of sequencing error, only the mismatch sites that appear at least on two different entries are picked out. The 5' and 3' fragments were not taken into account. A total of 30 mismatch sites were found and a dataset was constructed for phylogeny analysis using the neighbour-joining approach of distant matrix method within the Phylip package. A bootstrap value 100 was chosen and other parameters were set to default.

2.5 Mismatch table

A table of 30 mismatch sites among 32 SARS-CoV genome sequences was created and inspected manually. Several identical entries which have the same mutation at the same position were found, i.e., the four entries TW09, TW10, TW11 and TW12 from Taiwan National Health Research Institutes, were identical. Therefore, this table was further edited by selecting a representative entry and removing the identical entries (Table 3).

Representative	Identical removed
CA01	SG03, EU03, ZJ01, TW01
SG01	SG02
EU01	EU02
TW03	TW08
TW09	TW10, TW11, TW12

Table 3 List of identical mismatches

3 Results and discussion

3.1 Clustering of SARS-CoVs

Fig. 1 shows the result of phylogeny analysis for the dataset of 30 mismatches among 32 SARS-CoV genome sequences. In addition to the reference sequence CA01 and four identical entries (SG03, ZJ01, TW01 and EU03), the remaining 27 entries can be clustered into three major groups. It seems that the data source has a strong influence on the clustering, as we can see that the 11 entries from Taiwan were in the same group, the 4 entries (BJ01-BJ04) together with GD01 were divided into another group. The samples of these five SARS-CoVs were actually all from Guangdong province, China. The 4 Singapore sequenced entries were also in the same group. However, the three entries from Hong Kong were distributed in three different groups. Other exceptions such as GD02 and TW02 can be also found. This clustering result may give some evidence for the further analysis of the mutation pattern and transmission route of the SARS-CoVs.

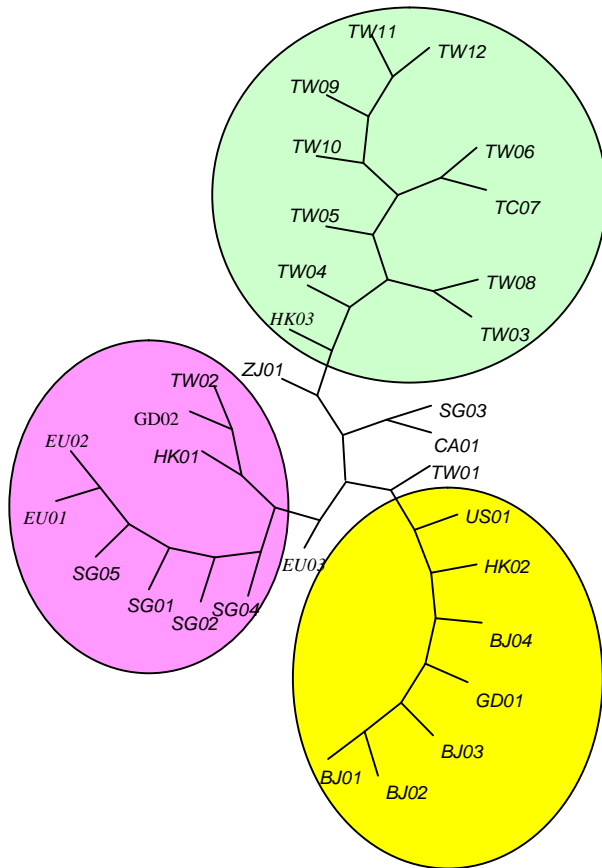


Fig. 1 Clustering of 32 SARS-CoV genome sequences

3.2 Mismatch table

As with other single strand RNA viruses, the SARS-CoV has a high mutation rate which can be seen from the results of multiple sequence alignment. Table 4 shows the 30 mismatch positions among 22 representative SARS-CoV genome sequences. CA01, the first entry deposited to NCBI GenBank by Canadian scientists was taken as a

reference sequence. Three different groups can be clearly distinguished from the positions of the mismatch and the pattern of mutation. Numbers of mutation sites are indicated in the last column. It seems that the group composed mainly of the SARS-CoVs (BJ01-BJ04, GD01) sequenced by the institutions from Beijing has a higher mutation rate than the other two groups. The mismatch sites are shown in the first row of the table. Further analysis of the codon change and co-responding amino acid change which might have some effect on the encoded proteins is underway.

3.3 New released SARS-CoVs

Recently, 10 new SARS-CoV genome sequence from human and animals were released in GenBank (Table 5). All these sequences were deposited by the University of Hong Kong (Guan et al. 2003). The sources of four entries (SZ1, SZ3, SZ16 and SZ43) were live animals.

Name	Accession	Length	Status
SZ1	AY304489	8439	Partial
SZ3	AY304486	29741	Complete
SZ13	AY304487	8581	Partial
SZ16	AY304488	29731	Complete
GZ43	AY304490	13471	Partial
GZ50	AY304495	29720	Complete
GZ60	AY304491	11006	Partial
HKU-36871	AY304492	13471	Partial
HKU-65806	AY304493	11010	Partial
HKU-66078	AY304494	11010	Partial

Table 5 Newly released SARS genome sequences

A fragment of extra 29 bases near the 3' end was found in four coronaviruses from animals, which exists in the GD01 sequence in human. It is interesting to find out the origin of the virus with the information. Detailed analysis is being carried out.

3.4 Conclusion

In order to provide the biologists and medical scientists, we have setup an anti-SARS bioinformatics web site and constructed a bioinformatics platform for the analysis of SARS-CoV sequences. We have also presented the results of initial analysis of the SARS coronavirus genome sequences including the clustering of different groups and the mismatch sites. The purpose of this paper, however, is not aimed to the accuracy of the analysis results, but focused on the data resources that users can access, the tools we used and the approaches we introduced for the in-depth analysis of sequences of the SARS-CoV as well as other viruses.

Group	Site	1785	2562	3858	7930	8585	9417	9492	9867	10600	11461	11506	16351	17590	17872	18991	19090	19110	19865	21749	22250	22545	24963	26081	26234	26508	26631	27274	27844	27859	28329	No
	TOR2	C	G	T	C	G	T	T	C	A	C	C	A	T	C	T	A	C	A	G	T	A	C	A	C	T	C	C	C	T	C	0
1	HK01	C	G	T	C	G	T	T	C	A	C	C	A	T	C	T	A	C	A	G	T	A	C	A	C	T	T	C	C	T	C	1
	SG01	C	G	T	C	G	T	T	C	A	C	C	A	T	C	T	A	T	A	G	T	A	C	A	C	T	C	C	C	T	C	1
	GD02	C	G	T	C	G	T	T	C	A	C	C	G	T	C	T	A	C	A	G	T	A	T	C	A	C	C	C	C	T	C	1
	TW02	C	G	T	C	G	T	T	C	A	C	C	G	T	C	T	A	C	A	G	T	A	T	C	A	C	T	C	-	T	C	2
	SG04	C	G	T	C	G	T	T	C	A	C	C	A	T	C	T	A	T	A	G	T	A	C	A	C	T	C	C	-	T	C	2
	SG05	C	G	T	C	G	T	T	C	A	C	C	A	T	C	A	A	T	A	G	T	A	C	A	C	T	C	C	C	T	C	2
	EU01	C	A	T	C	G	T	T	C	A	T	C	A	T	C	A	A	T	A	G	T	A	T	A	C	T	T	C	C	T	T	7
2	HK03	C	G	T	C	G	T	T	C	A	C	C	A	T	T	A	C	A	G	T	A	C	A	C	G	C	C	C	T	C	2	
	TW04	C	G	C	C	G	T	T	C	A	C	T	A	T	C	T	A	C	A	G	T	A	C	A	C	G	C	C	C	T	C	3
	TW02	T	G	C	C	G	T	T	C	A	C	T	A	T	C	T	A	C	A	G	T	A	C	A	C	G	C	C	C	T	C	4
	TW05	C	G	C	C	G	T	T	C	A	C	T	A	T	C	T	G	C	A	G	T	A	C	A	C	G	C	C	C	T	C	4
	TW09	C	G	C	C	G	T	T	C	A	C	T	A	T	C	T	A	C	A	G	T	A	C	A	T	G	C	C	T	T	C	5
	TW07	C	G	C	C	G	T	T	C	A	C	T	A	T	C	T	G	C	A	G	T	A	C	A	T	G	C	C	T	T	C	6
	TW06	C	G	C	T	G	T	T	C	A	C	T	A	T	C	T	G	C	A	G	T	A	C	A	T	G	C	C	T	T	C	7
3	US01	C	G	T	T	G	T	T	C	A	C	C	A	T	C	T	G	C	A	G	T	A	C	A	C	T	C	C	C	T	C	2
	BJ04	C	G	T	C	G	T	T	T	A	C	C	A	G	C	T	A	C	G	G	C	A	C	A	C	T	C	T	C	C	C	4
	HK02	C	G	T	C	G	C	C	C	A	C	C	A	G	T	T	G	C	A	A	C	A	C	A	C	T	C	C	C	C	C	6
	BJ03	C	G	T	C	G	C	T	T	A	C	C	A	G	C	T	A	C	G	A	C	A	C	C	C	T	C	C	C	C	C	7
	GD01	C	G	T	C	G	C	C	C	C	C	A	G	C	T	A	C	G	A	C	G	C	A	C	T	C	T	C	C	C	C	8
	BJ02	C	G	T	C	T	C	T	T	A	C	C	A	G	C	T	A	C	G	A	C	G	C	A	C	T	C	T	C	C	C	8
	BJ01	C	G	T	C	T	C	T	T	C	C	A	G	C	T	A	C	G	A	C	A	C	C	C	T	C	T	C	C	C	C	9

Table 4 Mismatches of 30 sites among 20 SARS-CoV genome sequences

Acknowledgements

Thanks to Wu JM and Sun Y for the initial design of the CBI anti-SARS web site, to Chen YJ, Gao G and Ye ZQ for the analysis of SARS coronavirus sequence. We are grateful to Lopez R at EBI and Bao YM at NCBI for various help providing the resource. Thanks to Bhuvana and Tan TW to setup the anti-SARS mirror site on the APBioNet web server.

References

- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003): GenBank. *Nucleic Acids Res.* 31(1):23-7.
- Etzold, T., Ulyanov, A. and Argos, P. (1996): SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* 266:114-128.
- Felsenstein, J. (1989): PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics.* 5, 164-166.
- Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., Butt, K.M., Wong, K.L., Chan, K.W., Lim, W., Shortridge, K.F., Yuen, K.Y., Peiris, J.S. and Poon, L.L. Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China, Published online 4 September 2003, 10.1126/science.1087139.
- 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. and Anderson, L.J. (2003): A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N Engl J Med* 348(20):1953-66. (published online 10 Apr 2003)
- Kumar, S., Tamura, K., Jakobsen, I.B. and Nei M (2001): MEGA2: Molecular Evolutionary Genetics Analysis software, Bioinformatics (submitted).

- Luo, J. (2000): A PC/Linux based SRS platform. *Science Bulletin*, 45(9):1006-1008. (in Chinese)
- 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. and Roper, R.L. (2003): The Genome sequence of the SARS-associated coronavirus, *Science* 300(5624): 1399-1404. (published online 1 May 2003)
- Poutanen, S.M., Low, D.E., Henry, B., Finkelstein, S., Rose, D., Green, K., Tellier, R., Draker, R., Adachi, D., Ayers, M., Chan, A.K., Skowronski, D.M., Salit, I., Simor, A.E., Slutsky, A.S., Doyle, P.W., Krajden, M., Petric, M., Brunham, R.C., McGeer, A.J. and National Microbiology Laboratory, Canada; Canadian Severe Acute Respiratory Syndrome Study Team. (2003): Identification of severe acute respiratory syndrome in Canada. *N Engl J Med* 348(20):1995-2005. (published online 31 Mar 2003)
- Rice P., Longden I., Bleasby A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276-7.
- 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. and Bellini, W.J. (2003): Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300(5624):1394-9. (published online 1 May 2003)
- Ruan, Y.J., Wei, C.L., Ee, L.A., Vega, V.B, Thoreau, H., Yun, S.T.S., Chia, J.M., Ng, P., Chiu, K.P., Lim, L., Tao, Z., Peng, C.K., Ean, L.O.L., Lee, N.M., Sin, L.Y., Ng, L.F.P., Chee, R.E., Stanton, L.W., Long, P. M. and Liu, E.T. (2003): Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *The Lancet* 361(9371):1779-1785. (published online 9 May 2003)
- Thompson, J.D., Higgins, D.G and 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. *Nucl Acids Res* 22:4673-4680.
- Zdobnov, E.M., Lopez, R., Apweiler, R. and Etzold T. (2002): The EBI SRS server--recent developments. *Bioinformatics* 18(2):368-73.