Acta virologica **64**: 496 – 500, 2020 doi: 10.4149/av_2020_403

Genomic characterization and phylogenetic evolution of the SARS-CoV-2

R.-H. ZHANG[1#], X. AI[2#], Y. G LIU[3], CH.-H. LI[1], H.-L. ZHANG[3*]

[1]Key Laboratory of Preventive Veterinary Medicine, Department of Veterinary Medicine, Animal Science College, HeBei North University, Zhangjiakou, 075131, P. R. China; [2]Collegeof Animal Science and Veterinary Medicine, Tianjin Agricultural University, Tianjin 300384, P. R. China; [3]Department of Veterinary Medicine, Inner Mongolia Agricultural University, Hohhot, 010018, P. R. China

Received July 19, 2020; accepted August 8, 2020

Summary. – The coronavirus disease 2019 (COVID-19) starting on 12 December 2019 in Wuhan, China, caused 7,885,123 cases including 431,835 deaths by 14 Jun 2020 all over the world. Here we report the genomic characterization and phylogenetic evolution of coronavirus SARS-CoV-2 causing COVID-19. The SARS-CoV-2 and other coronavirus genomes were obtained from GISAID and GenBank. The genomes were annotated and potential genetic recombination was investigated. Phylogenetic analysis was conducted and used to determine the evolutionary history of the virus and to elucidate the origin of the virus. The analysis had revealed that SARS-CoV-2 possessed a similar genomic organization to bat-SARS-like-CoV collected in China. The genome sequences of SARS-CoV-2 were very similar, showing 99.6–100% sequence identity. Notably, SARS-CoV-2 was closely related (with 88% identity) to bat-SARS-like coronavirus, but was more distant from SARS-CoV (about 79%) and MERS-CoV (about 50%). Phylogenetic tree of the complete viral genome showed that the virus clustered with bat SARS-like coronavirus. The results of the similarity between SARS-CoV-2 and related viruses did not identify any potential genomic recombination events. Therefore, it seems that the SARS-CoV-2 might be originally hosted by bats, and might have been transmitted to humans via intermediate hosts of currently unknown wild animal(s). Finally, based on the wide spread of SARS-CoV in their natural reservoirs, future studies should focus more on surveillance of coronaviruses, and measures against the domestication and consumption of wild animals should be implemented.

Keywords: coronavirus; SARS coronavirus; SARS-CoV-2; genomic characterization; phylogenetic evolution

Introduction

The coronavirus disease 2019 (COVID-19) emerged on 12 December 2019 in Wuhan, causing a large global outbreak and becoming a major global health concern (China CDC, 2020; Wu *et al*., 2020). The COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, previously provisionally named 2019 novel coronavirus (2019-nCoV) or human coronavirus-19 (hCoV-19)) as designated by the International Committee on Taxonomy of Viruses (Lai *et al*., 2020; Uddin*et al*., 2020). As of June14, 2020, the virus has caused 7,885,123 infections and 431,835 deaths all over the world. Coronaviruses (CoVs) possess a single-strand, positive-sense RNA genome ranging from 26 to 32 kilobases in length and mainly cause respiratory and gastrointestinal tract infections (Su *et al*., 2016). CoVs are genetically classified into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* (Li *et al*., 2016). The *Alphacoronavirus* and *Betacoronavirus* usually infect mammals, whereas the

*Corresponding author. E-mail: zhanghongliang001@126.com; phone: +8615153234981. #These authors contributed equally to this work.

Abbreviations: CoV(s) = coronavirus(es); COVID-19 = coronavirus disease 2019; SARS-CoV = severe acute respiratory syndrome CoV; MERS-CoV = Middle East respiratory syndrome CoV

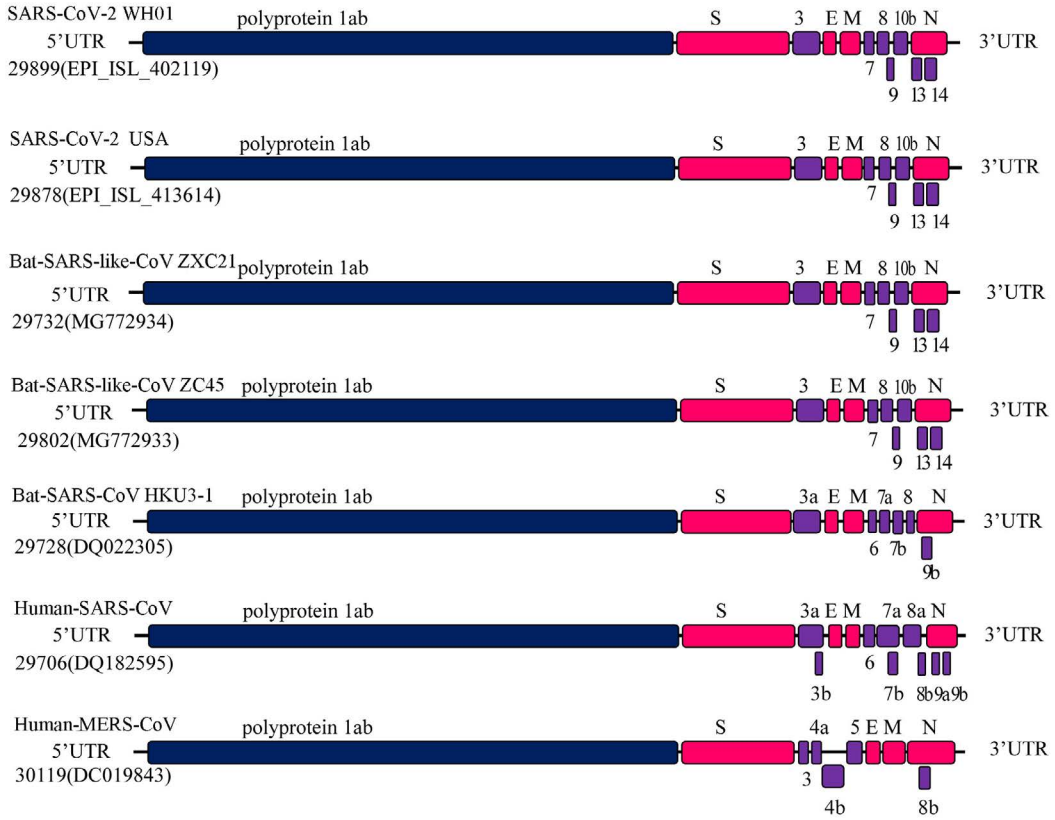

Fig. 1

Genome characterization of SARS-CoV-2

Genomic organization of SARS-CoV-2, bat-SARS-like-CoV, bat-SARS-CoV, human-SARS-CoV, human-MERS-CoV was predicted using Geneious, GeneMarkS and ORF finder with manual check. Only open reading frames of more than 100 nucleotides are shown.

Gammacoronavirus and *Deltacoronavirus* primarily infect birds (Tang *et al*., 2015). Six kinds of human CoVs have been identified previously, however coronaviruses did not attract worldwide attention because most people infected with the virus are associated with mild clinical symptoms. Since 2003, severe acute respiratory syndrome coronavirus (SARS-CoV) was associated with a respiratory illness that emerged in China and resulted in more than 8,000 cases of infections and 700 deaths worldwide (Chinese, S. M. E. C, 2004; Zhong *et al*., 2003). In 2012, Middle East respiratory syndrome coronavirus (MERS-CoV) was responsible for nearly 2,500 cases of infections and 800 deaths. It is considered that SARS-CoV and MERS-CoV are highly pathogenic and were transmitted from bats to palm civets or dromedary camels, and finally to humans (Cui *et al*., 2019; Guan *et al*., 2003; Drosten *et al*., 2014). At present, SARS-CoV-2 is associated with millions of infections and hundreds of thousands of deaths in humans. The infection of the virus has not been effectively controlled worldwide. The molecular epidemiology and phylogenetic evolution of the SARS-CoV-2 needs to be analyzed more closely. This knowledge is important for the prediction of disease progression, drug and vaccine development. Here, we provide genomic characterization, recombination analyses and phylogenetic evolution of the SARS-CoV-2.

Materials and Methods

Virus genome analysis and annotation. Reference virus genomes including complete genomic sequences of bat severe acute respiratory syndrome-like coronavirus (bat-SARS-like-CoV), bat severe acute respiratory syndrome coronavirus (bat-SARS-CoV), SARS-CoV, MERS-CoV, and SARS-CoV-2, were obtained from GISAID and GenBank. The open reading frames (ORF) of the SARS-CoV-2 virus genome sequences were predicted using Geneious (version 11.1.5) (Marchler-Bauer *et al*., 2017). The ORF of the genome sequences were further predicted by GeneMarkS and ORF finder (https://www.ncbi.nlm.nih.gov/orf finder/) with manual check (Wu *et al*., 2020). Pairwise sequence identities were calculated using MegAlign of DNASTAR

program package (MegAlign 5.00, DNASTAR Inc., USA) (Saxena *et al*., 2018; Mo *et al*., 2018).

*Phylogenetic analysis of SARS-CoV-2***.** The complete genome sequences were aligned and analyzed using the ClustalX2. Representatives of different coronaviruses were included in the alignment. Phylogenetic analyses based on nucleotide sequences were generated by the neighbor-joining method using the MEGA X software with the neighbor-joining method (Kumar *et al*., 2018; Stecher *et al*., 2020). Estimates of the phylogenic relationships were measured with 1000 bootstrap replicates.

The recombination analysis of SARS-CoV-2 with other coronaviruses. The aligned full sequences were initially scanned for recombination events using SimPlot v.3.5.1. The potential genetic recombination events between SARS-CoVs, bat-SARS-like-CoVs, and bat-SARS-CoVs were investigated using SimPlot v.3.5.1 (Lole *et al*., 1999; Shuai *et al*., 2017).

Results

Sequence comparison and genomic composition of SARS-CoV-2

Based on the determined genomes of the SARS-CoV-2, a genome annotation of this virus was performed with a comparison to related coronaviruses, including bat-SARS-like-CoV, bat-SARS-CoV, human-SARS-CoV, and human-MERS-CoV, whose genomes were obtained from GISAID and NCBI.

Comparison of the predicted coding regions of SARS-CoV-2 showed that they possessed a similar genomic organization to bat-SARS-like-CoV ZC45, and bat-SARS-like-CoV ZXC21. At least 12 coding regions were predicted, including 1ab, S, 3, E, M, 7, 8, 9, 10b, N, 13, and 14. The lengths of most of the proteins encoded by SARS-CoV-2, bat-SARS-like-CoV ZC45, and bat-SARS-like-CoV ZXC21 were similar, with only a slight differences. Compared with the bat-SARS-like-CoV, bat-SARS-CoV, SARS-CoV, and MERS-CoV, an obvious difference was in a longer spike protein encoded by SARS-CoV-2. At the amino acid level, the SARS-CoV-2 is similar to that of bat-SARS-like-CoV, which contains 12 coding regions, including 1ab, S, 3, E, M, 7, 8, 9, 10b, N, 13, and 14. The SARS-CoV-2 is differs from SARS-like-CoV and MERS-CoV. For example, the 4a, 4b and 5 protein is absent in SARS-CoV-2 and SARS-CoV, but present in MERS-CoV. The 6 protein is absent in SARS-CoV-2 and present in SARS-CoV. The 7 protein is present in SARS-CoV-2 and SARS-CoV, and absent in MERS-CoV. The 9a and 9b protein is absent in SARS-CoV-2 and present in SARS-CoV. The 13 and 14 protein is present in SARS-CoV-2, absent in SARS-CoV and MERS-CoV (Fig. 1).

The pairwise sequence identities results showed that the SARS-CoV-2 virus in China, Germany, Italy and USA shared 99.5%–100% genomic sequence identity among themselves. The overall nucleotide sequence identity of the genomes of SARS-CoV-2 with bat-SARS-like-CoV ZC45 was 87.8%–88.2%, which was higher than the sequence identity of the genomes of SARS-CoV-2 with SARS-CoV SZ3 (79.5–79.9%) (Fig. 2).

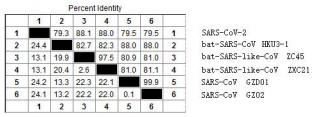

Fig. 2

Sequence comparison of SARS-CoV-2 and other coronaviruses
SARS-CoV-2 have been aligned against SARS-CoVs, bat-SARS-like CoVs, SARS-like-CoV using the MegAlign. GenBank accession numbers are: AY390556 for SARS-CoV GZ02; AY278489 for SARS-CoV GZD01; MG772934 for bat-SARS-like-CoV ZXC21; MG772933 for bat-SARS-like-CoV ZC45; DQ022305 for bat-SARS-CoV HKU3-1; GISAID accession number is EPI_ ISL_402119 for SARS-CoV-2.

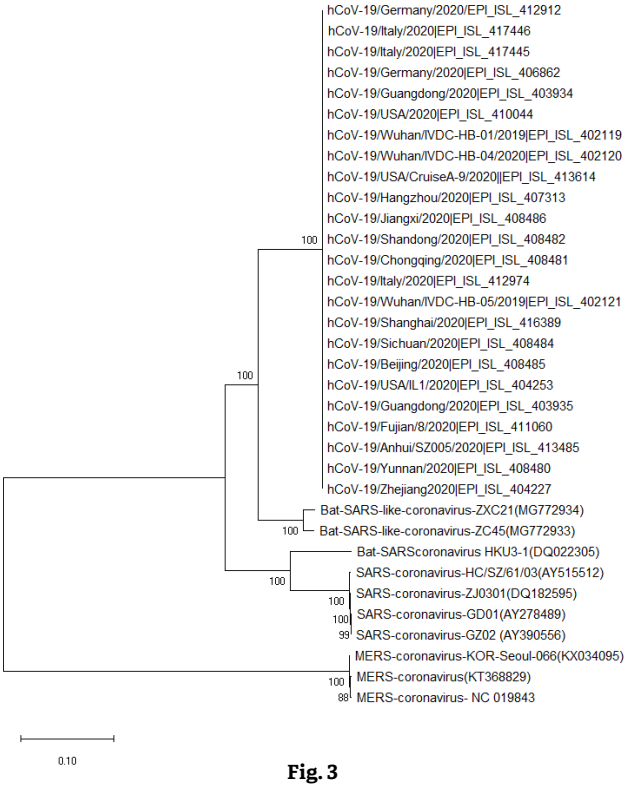

Fig. 3

Phylogenetic analysis for SARS-CoV-2
Phylogenetic analysis was constructed using the neighbor-joining method in MEGA X version with 1,000 bootstrap replicates. All the sequences used for phylogenetic analysis were obtained from GenBank and GISAID.

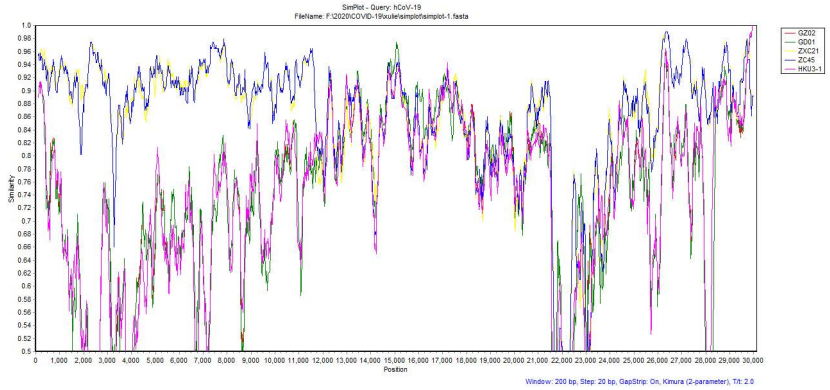

Fig. 4

The recombination analysis of SARS-CoV-2 with other SARS and SARS-like-CoVs.
Similarity plots were conducted with SARS-CoV-2 as the query and bat-SARS-like-CoVs, including ZXC21, ZC45, bat SARS-CoV HKU3-1, and human-SARS-CoVs, including GZ02, GD01, as potential parental sequences. The analysis was performed using the Kimura model, with a window size of 200 base pairs and a step size of 20 base pairs.

Phylogenetic analysis of full-length genomes of SARS-CoV-2

As shown in the phylogenetic tree based on whole genome sequences, the phylogenetic tree falls into two clades (Fig.3). The MERS-CoV constitutes one clade, while the bat-SARS-like-CoV, bat-SARS-CoV, SARS-CoV, and SARS-CoV-2 constitute the other clade. The SARS-CoV-2 is parallel to the bat-SARS-like-CoV, while the SARS-CoV and bat-SARS-CoV are descended from the bat-SARS-like-CoV, indicating that SARS-CoV-2 is closer to the bat-SARS-like-CoV than the SARS-CoVs in terms of the whole genome sequence. All the SARS-CoV-2 from different countries and regions are in the same clade. The phylogenetic analyses for the whole genome clearly shows that the SARS-CoV-2 is most closely related to bat- SARS-like viruses.

The recombination analysis of SARS-CoV-2

To further explore the evolution of SARS-CoV-2, complete genomic sequences of the representative bat-SARS-like-CoV (ZXC21, ZC45), representative bat-SARS-CoV (HKU3-1), and representative SARS-CoV (GZ02, GD01), were obtained from GenBank. The potential genomic recombinant events among the bat-SARS-like-CoV, bat-SARS-CoV, SARS-CoV and SARS-CoV-2 were examined by Simplot analysis (Fig. 4). The results did not identify any potential genomic recombination events.

Therefore, it seems that the SARS-CoV-2 might be originally hosted by bats, and might have been transmitted to humans via intermediate hosts of currently unknown wild animal(s).

Discussion

Phylogenetic analysis showed that the new SARS-CoV-2 clustered with the bat-SARS-like-CoV isolated in 2015 and 2017 in China. This clade is separated from the bat-SARS-CoV and SARS-CoV, suggesting that the SARS-CoV-2 is homologous and genetically more similar to the bat-SARS-like-CoV than to the bat-SARS-CoV and SARS-CoV.

These data support the hypothesis that a bat indeed is a reservoir for SARS-CoV-2 in particular. However, the sequence identity between SARS-CoV-2 and its close relatives, bat-SARS-like-CoV ZC45 and bat-SARS-like CoV ZXC21 was less than 90%, conforming the long branches between them. Hence, there are other animals acting as an intermediate host between bats and humans. Bat-SARS-like-CoV ZC45 and bat-SARS-like-CoV ZXC21 are not direct ancestors of SARS-CoV-2.

Therefore, it seems that the SARS-CoV-2 might be originally hosted by bats, and might have been transmitted to humans via intermediate hosts of currently unknown wild animal(s). Finally, based on the wide spread of SARS-CoV in their natural reservoirs, future studies should focus more on surveillance of coronavirus, and measures against the domestication and consumption of wild animals should be implemented.

References

China CDC. Tracking the Epidemic. http://weekly.chinacdc.cn/news/ Tracking the Epidemic.htm?from=timeline#Beijing%20Municipality%20Update.2020.

Chinese SARS Molecular Epidemiology Consortium (S.M.E.C) (2004): Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 303, 1666–1669. https://doi.org/10.1126/science.1092002

Cui J, Li F, Shi ZL (2019): Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181–192. https://doi.org/10.1038/s41579-018-0118-9

Drosten C, Kellam P, Memish ZA (2014): Evidence for camel-to-human transmission of MERS coronavirus. N. Engl. J. Med. 371, 1359–1360. https://doi.org/10.1056/NEJMc1409847

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, 276–278. https://doi.org/10.1126/science.1087139

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018): MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549. https://doi.org/10.1093/molbev/msy096

Lai C-C, Shih T-P , Ko W-C , Tang H-J , Hsueh P-R (2020): Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int. J. Antimicrob. Agents 55, 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924

Li F (2016): Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. 3, 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC (1999): Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152–160. https://doi.org/10.1128/JVI.73.1.152-160.1999

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH (2017): CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203. https://doi.org/10.1093/nar/gkw1129

Mo L, Shi J, Guo X, Zeng Z, Hu N, Sun J, Wu M, Zhou H, Hu Y (2018): Molecular characterization and phylogenetic analysis of a dengue virus serotype 3 isolated from a Chinese traveler returned from Laos.Virol. J. 15, 113. https://doi.org/10.1186/s12985-018-1016-5

Saxena A, Biswas SK, Chand K, Naskar J, Chauhan A, Mohd G, Tewari N, Kurat-Ul-Ain, Ramakrishnan MA, Pandey AB (2018): Genetic and phylogenetic analysis of the outer capsid protein genes of Indian isolates of bluetongue virus serotype-16. Vet. World 11, 1025–1029. https://doi.org/10.14202/vetworld.2018.1025-1029

Shuai L, Yanqun W, Yingzhu C, Bingjie W, Kun Q, Zhao J, Lou Y, Tan W (2017): Discovery of a novel canine respiratory coronavirus support genetic recombination among betacoronavirus. Virus Res. 237, 7–13. https://doi.org/10.1016/j.virusres.2017.05.006

Stecher G., Tamura K, Kumar S (2020): Molecular evolutionary genetics analysis (MEGA) for macOS. Mol. Biol. Evol. 37, 1237–1239. https://doi.org/10.1093/molbev/msz312

Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF (2016): Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24, 490–502. https://doi.org/10.1016/j.tim.2016.03.003

Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia X-Q (2015): Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci. Rep. 5, 17155. https://doi.org/10.1038/srep17155

Uddin M, Mustafa F, Rizvi TA, Loney T, Suwaidi HA, Al-Marzouqi AHH, Eldin AK, Alsabeeha N, Adrian TE, Stefanini C, Nowotny N, Alsheikh-Ali A, Senok AC (2020):SARS-CoV-2/COVID-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions. Viruses 12, 526. https://doi.org/10.3390/v12050526

Wu A, Peng Y, Huan B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T (2020): Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27, 325–328. https://doi.org/10.1016/j.chom.2020.02.001

Zhong NS, Zheng BJ, Li YM, Poon LLM, Xie ZH, Chan KH, Li PH, Tan SY, Chang Q, Xie JP, Liu XQ, Xu J, Li DX, Yuen KY, Peiris JSM, Guan Y (2003): Epidemiology and cause of severe acute respiratory syndrome (SARS)in Guangdong, People's Republic of China, in February, 2003. Lancet 362, 1353–1358. https://doi.org/10.1016/S0140-6736(03)14630-2