nature microbiology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41564-020-0771-4

In the format provided by the authors and unedited.

Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic

Maciej F. Boni[1,8]✉, Philippe Lemey[2,8]✉, Xiaowei Jiang[3], Tommy Tsan-Yuk Lam[4], Blair W. Perry[5], Todd A. Castoe[5], Andrew Rambaut[6]✉ and David L. Robertson[7]✉

[1]Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA, USA. [2]Department of Microbiology, Immunology and Transplantation,KU Leuven, Rega Institute, Leuven, Belgium. [3]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. [4]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China. [5]Department of Biology, University of Texas Arlington, Arlington, TX, USA. [6]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. [7]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [8]These authors contributed equally: Maciej F. Boni, Philippe Lemey.
✉e-mail: mfb9@psu.edu; philippe.lemey@kuleuven.be; a.rambaut@ed.ac.uk; david.l.robertson@glasgow.ac.uk

Supplementary Information for "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic"

Maciej F Boni*, Philippe Lemey*, Xiaowei Jiang, Tommy Tsan-Yuk Lam, Blair Perry, Todd Castoe, Andrew Rambaut and David L Robertson

*Joint First Authors

1 Uncertainty in patristic distance among SARS-CoV-2, RaTG13, and Pangolin 2019 in CTD region and variable loop region of spike protein

The bar graphs at the bottom of Figure 2 of the main text show the ML distances between SARS-CoV-2 and related bat and pangolin viruses, in five different regions of the S protein. As these genetic distance measures are the most basic quantities needed to infer the pattern and direction of recombination in different parts of the S protein, we show in Extended Data Figure 1 the uncertainty in the patristic distances – i.e. distances on the phylogenetic tree, measured in substitutions per site – between SARS-CoV-2 and its nearest bat and pangolin viruses in two key parts of the S protein. Phylogenies were inferred with MrBayes v3.2.6 (Ronquist et al. 2012) for the same six sequences shown in Figure 3, using a GTR+Γ model of evolution. Monte Carlo Markov Chains were run for 1,000,000 iterations, with the first 10% discarded as burn-in, and convergence was assessed visually.

2 Assessing temporal signals using TempEst and BETS

Root-to-tip divergence plots as a function of sampling time indicate no clear pattern of divergence accumulation over the sampling time range for all three data sets; see Extended Data Figure 2. To formally test for temporal signal, we used a recently proposed Bayesian model testing approach that compares the marginal likelihoods estimated for a model that constrains tips in time-measured trees to be proportional to their sampling times and has an estimable evolutionary rate to a model that enforces an ultrametric tree (all sequences sampled at the same time, reflecting that sampling times do not predict tip positions) without a free evolutionary rate parameter (Duchene et al. 2019) . Supplementary Information Table 1 reports the marginal likelihood estimates for these models fitted to both data sets. This results in a log Bayes factor (log BF) support of 3, 10, and 3 in favor of temporal signal for **NRR1**, **NRR2**, and **NRA3**, respectively.

3 Estimating evolutionary rates and divergence dates for NRR1, NRR2, and NRA3

Although formal testing does not reject the presence of temporal signal, the support for this signal remains limited. We therefore aimed to take advantage of prior information on the evolutionary rate in the Bayesian analysis of the three data sets, but in a way that avoids having to make strong assumptions on whether sarbecovirus evolutionary rates should closely match the MERS-CoV or HCoV-OC43 rates. We adopt two evolutionary rate priors that are centred on the mean rates for MERS-CoV and HCoV-OC43, but with standard deviations that are ten times larger than the posterior rate distributions for MERS-CoV and HCoV-OC43. Using these priors, we infer highly similar evolutionary rate posteriors for **NRR1**, **NRR2,** and **NRA3**. In addition, the posterior rate distributions show a considerable reduction in variance compared to their priors, indicating that sampling dates provide information about the rate despite the difficulties in ascertaining the temporal signal based on visual exploration. This is shown in Extended Data Figure 3.

To determine whether the segment or region chosen had an effect on the estimated TMRCA of SARS-CoV-2 and its closest ancestors, we estimated these TMRCAs for all five breakpoint-free regions shown in the bottom set of phylogenetic trees of Figure 1 of the main text. Remember that BFRs **A** through **E** were obtained by joining (concatenating) the original ten BFRs (paragraph 1, Results) when there were no phylogenetic incongruence signals detected between adjacent BFRs. TMRCA estimates for all five BFRs are reasonably close, with estimates ranging from 1963 to 1983; see Supplementary Information Table 3.

4 SARS-CoV-2 synonymous codon usage patterns are not consistent with snake genome

Our analysis of SARS-CoV-2 origins would not be complete without mentioning the early widespread reporting that the ostensibly unique part of the SARS-CoV-2 lineage was recombinant involving a snake virus (Ji et al. 2020). To demonstrate why a snake origin is not supported, a heat map of the relative synonymous codon usage (RSCU) metric for an expanded dataset (see section 4.1) revealed four distinct clusters of species consisting of (1) all coronaviruses, (2) invertebrates (low GC), (3) vertebrates with high GC content, and (4) vertebrates with low GC content (Extended Data Figure 5b). The cluster comprised of coronaviruses was most similar to the cluster of invertebrates, which possess the lowest average GC content of all eukaryotes analyzed. Notably, the cluster of coronaviruses is no more similar to the cluster containing snakes than it is to the cluster containing vertebrates with higher GC content that includes the representative bat species. The first two principal components of a PCA of RSCU values for all species clearly distinguished between coronaviruses and eukaryotes, with no clustering between SARS-CoV-2 and any snake species or the five bat-derived coronaviruses with the representative bat species (Extended Data Figure 5b). Additionally, a PCA of only snake, bat, and coronavirus RSCU measures further illustrates that all sampled coronaviruses generally cluster more closely with snakes than they do with the bat, despite some of them having known bat-derived origins (Extended Data Figure 5c). Lastly, squared Euclidian distance measures of RSCU between SARS-CoV-2 and all other species is linearly correlated with GC content, such that species with a lower GC content similar to that of SARS-CoV-2 exhibit a more similar profile of RSCU than do species with higher GC content (Extended Data Figure 5d), suggesting that Ji et al.'s finding that snakes had the most similar RSCU to SARS-CoV-2 is due simply to the fact that snakes possess the lowest GC content of species that they analyzed, rather than snakes being a likely host reservoir of the coronavirus.

Methods: Coding sequences were downloaded for 4 additional snake species (*Ophiophagus hannah* (GenBank: GCA_000516915.1) (Vonk et al. 2013), *Python bivittatus* (GenBank: GCA_000186305.2) (Castoe et al. 2013), *Thamnophis sirtalis* (GenBank: GCA_001077635.2) (Perry et al. 2018), and *Deinagkistrodon acutus* (GigaDB: http://dx.doi.org/10.5524/100196) (Yin et al. 2016), 5 coronaviruses with documented origins in bats (bat-SL-CoVZC45 (GenBank: MG772933.1) (Hu et al. 2018), bat-SL-CoVZXC21 (GenBank: MG772934.1) (Hu et al. 2018), BM48-31/BGR/2008 (GenBank: GU190215.1) (Drexler et al. 2010), YNLF_31C (GenBank: KP886808.1), and YNLF_34C (GenBank: KP886809.1)), and 7 additional eukaryotes each known to possess a relatively low genomic GC content (*Lottia gigantea* (GenBank: GCA_000327385.1) (Simakov et al 2013), *Trichoplax adhaerens* (GenBank: GCF_000150275.1) (Srivastava et al. 2008), *Octopus bimaculoides* (GenBank: GCF_001194135.1)

(Albertin et al. 2015), *Sarcophilus harrisii* (GenBank: GCF_902635505.1) (Murchison et al. 2012), *Tetranychus urticae* (GenBank: GCF_000239435.1) (Grbić et al. 2011), *Pediculus humanus* (GenBank: GCF_000006295.1) (Kirkness et al. 2010), and *Apis mellifera* (GenBank: GCF_003254395.2) (Wallberg et al. 2019)). For each species, coding sequence GC content was calculated using statswrapper.sh from BBMap v38.75 (Bushnell et al. 2014), and relative synonymous codon usage (RSCU) was calculated using codonw (Peden, 1997), following Ji et al (2020). Output from codonw was parsed using codonw-parser (https://github.com/juliambrosman/parse-codonw) and combined with RSCU values published in Ji et al. A heatmap of RSCU values was generated using pheatmap (Kolde 2015), with species clustered using the default Euclidean method. Additionally, principle components analysis of RSCU values was conducted on the full dataset as well as a subset of focal species (all snake species, one bat, known bat-derived coronaviruses, and the SARS-CoV-2). Squared Euclidian distance (SED) between RSCU values were calculated between SARS-CoV-2 and all other species, and a simple linear regression was used to test for correlation between SED and relative similarity in coding sequence GC content (measured as the difference in coding sequence GC content between SARS-CoV-2 and a given species).

5 Phylogenetic incongruence among 10 breakpoint-free regions (BFRs)

Shown after the references as Supplementary Figures 1 to 10.

Supplementary Table 1: Log marginal likelihood estimates for a dated tip model versus an ultrametric model for **NRR1**, **NRR2** and **NRA3**.

Data set	Log MLE dated tips	Log MLE ultrametric
NRR1	-91357	-91360
NRR2	-71620	-71630
NRA3	-118921	-118924

Supplementary Table 2. Estimates (posterior means) for the divergence time of RaTG13 bat virus and SARS-CoV-2, using three different non-recombinant regions/alignments (left column).

	HCoV-OC43 rate prior	**MERS-CoV rate prior**
NRR1	1969 (95% HPD: 1930-2000)	1981 (95% HPD: 1951-2007)
NRR2	1982 (95% HPD:1948-2009)	1984 (95% HPD: 1953-2007)
NRA3	1948 (95% HPD:1879-1999)	1966 (95% HPD: 1914-2002)

Supplementary Table 3. Estimates (posterior means) and 95% HPDs for the divergence times of SARS-CoV-2 and the bat virus RaTG13, and SARS-CoV-2 and the 2019 pangolin virus sequence sampled in Guangdong province. We estimated the divergence times for these regions in an analysis that specifies independent phylogenies and clock models for the BFRs, while sharing the same substitution model and coalescent prior among the regions. All other model and prior specifications followed the specifications detailed in the Methods section.

Breakpoint-free region	**TMRCA for SARS-CoV-2 / RaTG13**	**TMRCA for SARS-CoV-2 / pangolin 2019**
BFR1	1967 (1942,2001)	1877 (1812,1934)
BFR2	1975 (1941,2002)	1873 (1781,1951)
BFR3	1963 (1916,2003)	1879 (1767,1966)
BFR4	1983 (1958,2006)	1854 (1729,1951)
BFR5	1975 (1947,2001)	1936 (1885,1974)

Supplementary Table 4. Accession numbers of sequences used in this study.

				SAMPLING TIME		
Virus_name	**Species**	**Sample_location**	**Accession_no.**	**YEAR**	**MONTH**	**DAY**
RpShaanxi2011	R_pusillus	Shaanxi	JX993987	2011	9	NA
HuB2013	R_sinicus	Hubei	KJ473814	2013	4	NA
279_2005	R_macrotis	Hubei	DQ648857	2004	11	NA
Rm1	R_macrotis	Hubei	DQ412043	2004	11	NA
JL2012	R_ferrumequinum	Jilin	KJ473811	2012	10	NA
JTMC15	R_ferrumequinum	Jilin	KU182964	2013	10	NA
HeB2013	R_ferrumequinum	Hebei	KJ473812	2013	4	NA
SX2013	R_ferrumequinum	Shanxi	KJ473813	2013	11	NA
Jiyuan-84	R_ferrumequinum	Henan-Jiyuan	KY770860	2012	NA	NA
Rf1	R_ferrumequinum	Hubei-Yichang	DQ412042	2004	11	NA
GX2013	R_sinicus	Guangxi	KJ473815	2012	11	NA
Rp3	R_pearsoni	Guangxi-Nanning	DQ071615	2004	12	NA
Rf4092	R_ferrumequinum	Yunnan-Kunming	KY417145	2012	9	18
Rs4231	R_sinicus	Yunnan-Kunming	KY417146	2013	4	17
WIV16	R_sinicus	Yunnan-Kunming	KT444582	2013	7	21
Rs4874	R_sinicus	Yunnan-Kunming	KY417150	2013	7	21
YN2018B	R_affinis	Yunnan	MK211376	2016	9	NA
Rs7327	R_sinicus	Yunnan--Kunming	KY417151	2014	10	24
Rs9401	R_sinicus	Yunnan-Kunming	KY417152	2015	10	16
Rs4084	R_sinicus	Yunnan-Kunming	KY417144	2012	9	18
RsSHC014	R_sinicus	Yunnan-Kunming	KC881005	2011	4	17
Rs3367	R_sinicus	Yunnan-Kunming	KC881006	2012	3	19
WIV1	R_sinicus	Yunnan-Kunming	KF367457	2012	9	NA
YN2018C	R_affinis	Yunnan-Kunming	MK211377	2016	9	NA
As6526	Aselliscus_stoliczkanus	Yunnan-Kunming	KY417142	2014	5	12
YN2018D	R_affinis	Yunnan	MK211378	2016	9	NA
Rs4081	R_sinicus	Yunnan-Kunming	KY417143	2012	9	18
Rs4255	R_sinicus	Yunnan-Kunming	KY417149	2013	4	17
Rs4237	R_sinicus	Yunnan-Kunming	KY417147	2013	4	17
Rs4247	R_sinicus	Yunnan-Kunming	KY417148	2013	4	17
Rs672	R_sinicus	Guizhou	FJ588686	2006	9	NA
YN2018A	R_affinis	Yunnan	MK211375	2016	9	NA
YN2013	R_sinicus	Yunnan	KJ473816	2010	12	NA
Anlong-103	R_sinicus	Guizhou-Anlong	KY770858	2013	NA	NA
Anlong-112	R_sinicus	Guizhou-Anlong	KY770859	2013	NA	NA
HSZ-Cc	SARS-CoV-1	Guangzhou	AY394995	2002	NA	NA
YNLF_31C	R_Ferrumequinum	Yunnan-Lufeng	KP886808	2013	5	23
YNLF_34C	R_Ferrumequinum	Yunnan-Lufeng	KP886809	2013	5	23
F46	R_pusillus	Yunnan	KU973692	2012	NA	NA
SC2018	R_spp	Sichuan	MK211374	2016	10	NA
LYRa11	R_affinis	Yunnan-Baoshan	KF569996	2011	NA	NA
Yunnan2011	Chaerephon_plicata	Yunnan	JX993988	2011	11	NA
Longquan_140	R_monoceros	China	KF294457	2012	NA	NA
HKU3-1	R_sinicus	Hong_Kong	DQ022305	2005	2	17
HKU3-3	R_sinicus	Hong_Kong	DQ084200	2005	3	17
HKU3-2	R_sinicus	Hong_Kong	DQ084199	2005	2	24
HKU3-4	R_sinicus	Hong_Kong	GQ153539	2005	7	20
HKU3-5	R_sinicus	Hong_Kong	GQ153540	2005	9	20
HKU3-6	R_sinicus	Hong_Kong	GQ153541	2005	12	16

HKU3-10	R_sinicus	Hong_Kong	GQ153545	2006	10	28
HKU3-9	R_sinicus	Hong_Kong	GQ153544	2006	10	28
HKU3-11	R_sinicus	Hong_King	GQ153546	2007	3	7
HKU3-13	R_sinicus	Hong_Kong	GQ153548	2007	11	15
HKU3-12	R_sinicus	Hong_Kong	GQ153547	2007	5	15
HKU3-7	R_sinicus	Guangdong	GQ153542	2006	2	15
HKU3-8	R_sinicus	Guangdong	GQ153543	2006	2	15
CoVZC45	R_sinicus	Zhoushan-Dinghai	MG772933	2017	2	NA
CoVZXC21	R_sinicus	Zhoushan-Dinghai	MG772934	2015	7	NA
Wuhan-Hu-1	SARS-CoV-2	Wuhan	MN908947	2019	12	NA
BtKY72	R_spp	Kenya	KY352407	2007	10	NA
BM48-31	R_blasii	Bulgaria	NC_014470	2008	4	NA
RaTG13	R_affinis	Yunnan	EPI_ISL_402131	2013	7	24
P4L	pangolin	Guangxi	EPI_ISL_410538	2017	NA	NA
P5L	pangolin	Guangxi	EPI_ISL_410540	2017	NA	NA
P5E	pangolin	Guangxi	EPI_ISL_410541	2017	NA	NA
P1E	pangolin	Guangxi	EPI_ISL_410539	2017	NA	NA
P2V	pangolin	Guangxi	EPI_ISL_410542	2017	NA	NA
Pangolin-CoV	pangolin	Guangdong	EPI_ISL_410721	2019	3	NA

References

Albertin CB, Simakov O, Mitros Y, et al. "The octopus genome and the evolution of cephalopod neural and morphological novelties." Nature 524(7564):220, 2015.

Bushnell B. BBMap: a fast, accurate, splice-aware aligner. No. LBNL-7065E. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

Castoe TA, de Koning APJ, Hall KT, et al. "The Burmese python genome reveals the molecular basis for extreme adaptation in snakes." Proc Natl Acad Sci USA 110(51): 20645-20650, 2013.

Duchene S, Lemey P, Stadler T, et al. Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. bioRxiv. 2019;10.1101/810697. Available from: https://doi.org/10.1101/810697.

Drexler JF, Gloza-Rausch F, Glende J, et al. "Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences." J Virol 84(21): 11336-11349, 2010.

Grbić M, Van Leeuwen T, Clark RM, et al. "The genome of Tetranychus urticae reveals herbivorous pest adaptations." Nature 479(7374): 487-492, 2011.

Hu D, Zhu C, Ai L, et al. "Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats." Emerg Microb Infect 7(1),1:10, 2018.

Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 92:433–440, 2020.

Kirkness EF, Haas BJ, Sun W, et al. "Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle." Proc Natl Acad Sci USA 107(27):12168-12173, 2010.

Kolde R. "Package 'pheatmap'." R Package 1, no. 7 (2015).

Murchison EP, Schulz-Trieglaff OB, Ning Z, et al. "Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer." Cell 148(4):780-791, 2012.

Peden, J. "CodonW." Trinity College (1997).

Perry BW, Card DC, McGlothlin JW, et al. "Molecular adaptations for sensing and securing prey and insight into amniote genome diversity from the garter snake genome." Genome Biol Evol 10(8): 2110-2129, 2018.

Ronquist F, Teslenko M, van der Mark P, et al. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Syst. Biol. 61:539-542, 2012.

Simakov O, Marletaz F, Cho S-J, et al. "Insights into bilaterian evolution from three spiralian genomes." Nature 493(7433):526-531, 2013.

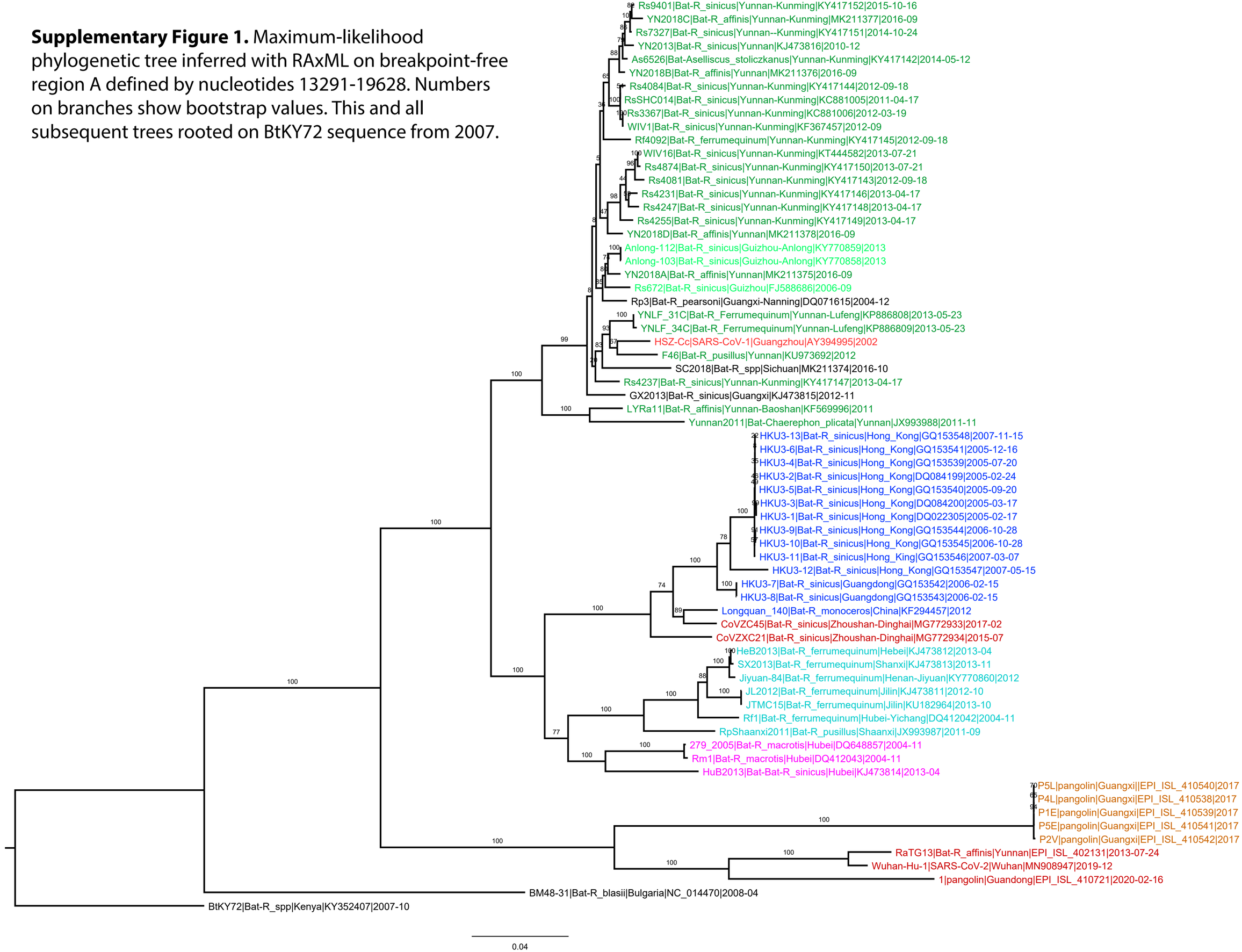
Srivastava M, Begovic E, Chapman J, et al. "The Trichoplax genome and the nature of placozoans." Nature 454(7207): 955-960, 2008.

Vonk FJ, Casewell NR, Henkel CV, et al. "The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system." Proc Natl Acad Sci USA 110(51):20651-20656, 2013.

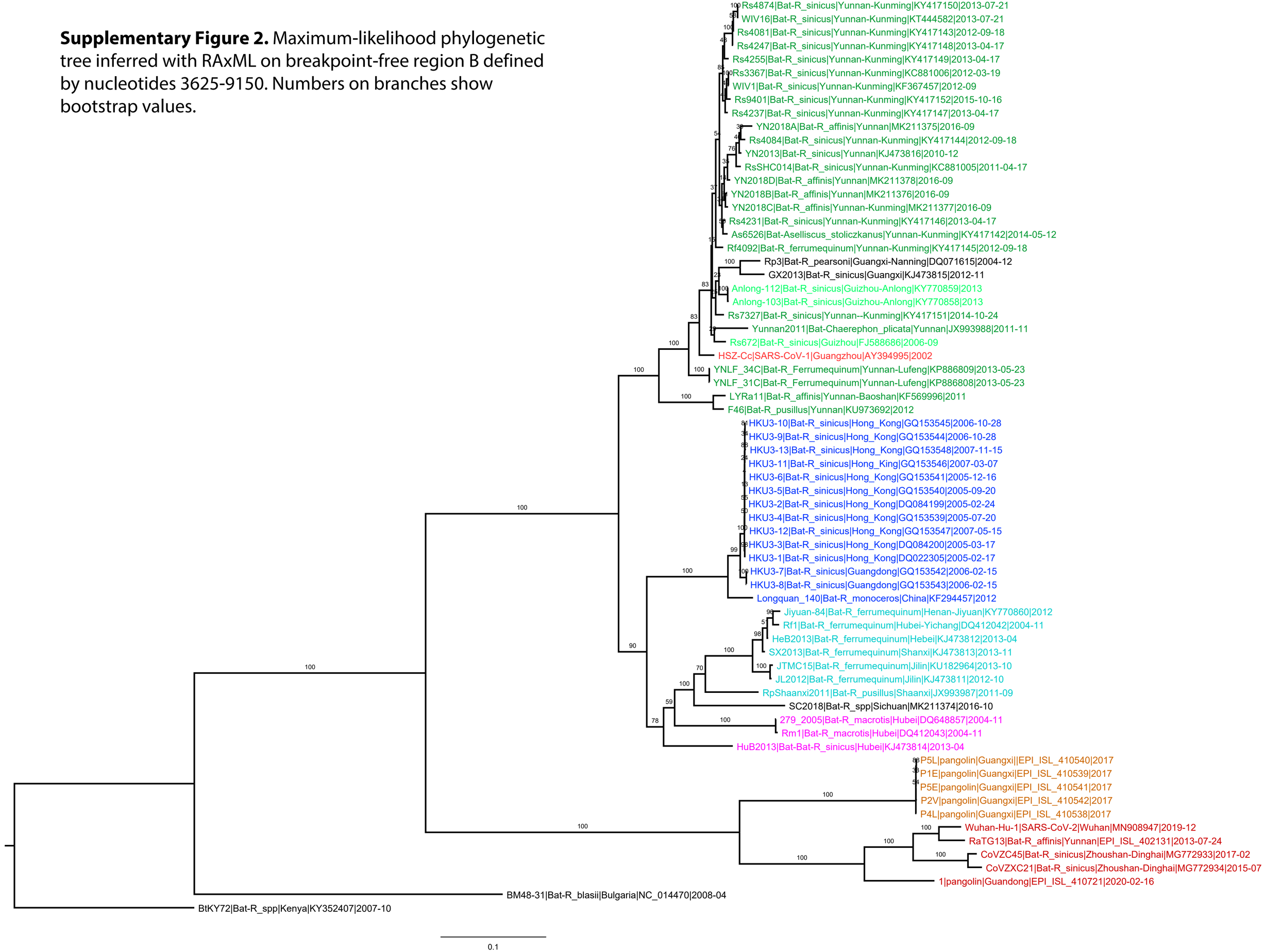
Wallberg A, Bunikis I, Pettersson OV, et al. "A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds." BMC Genomics 20(1):275, 2019.

Yin W, Wang Z-J, Li Q-Y, et al. "Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper." Nature Commun 7(1):1-11, 2016.

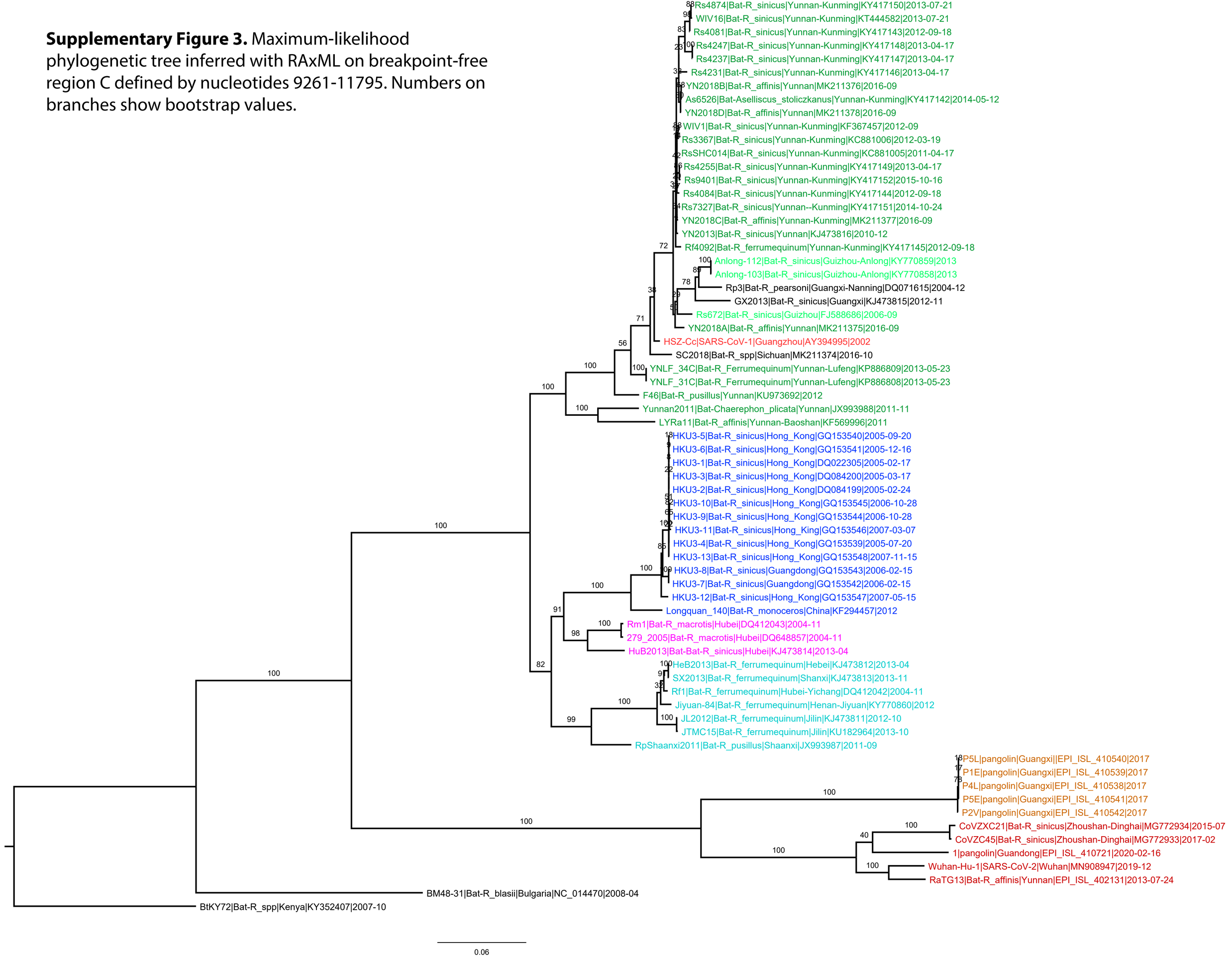
Supplementary Figure 1. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region A defined by nucleotides 13291-19628. Numbers on branches show bootstrap values. This and all subsequent trees rooted on BtKY72 sequence from 2007.



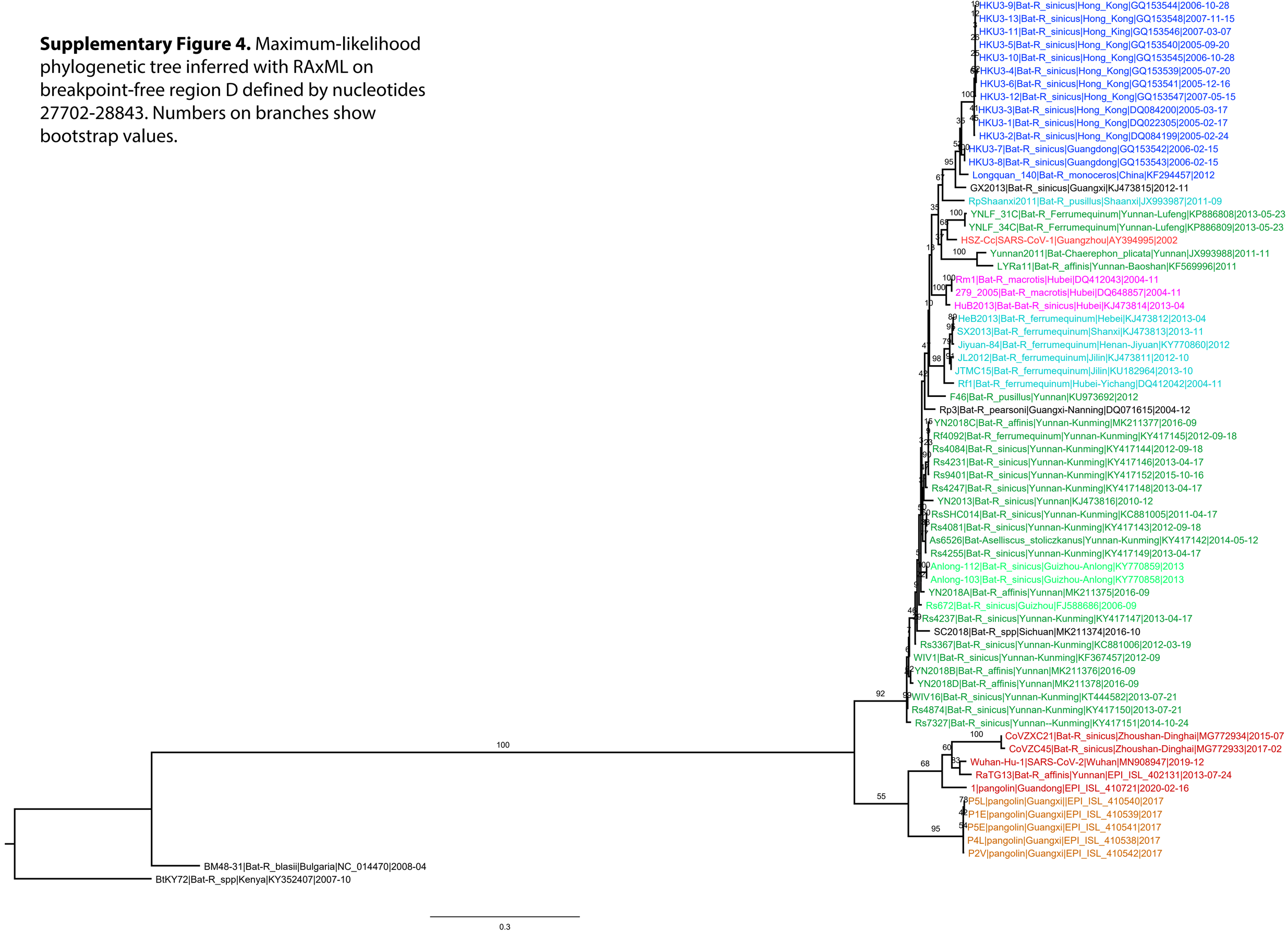
Supplementary Figure 2. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region B defined by nucleotides 3625-9150. Numbers on branches show bootstrap values.



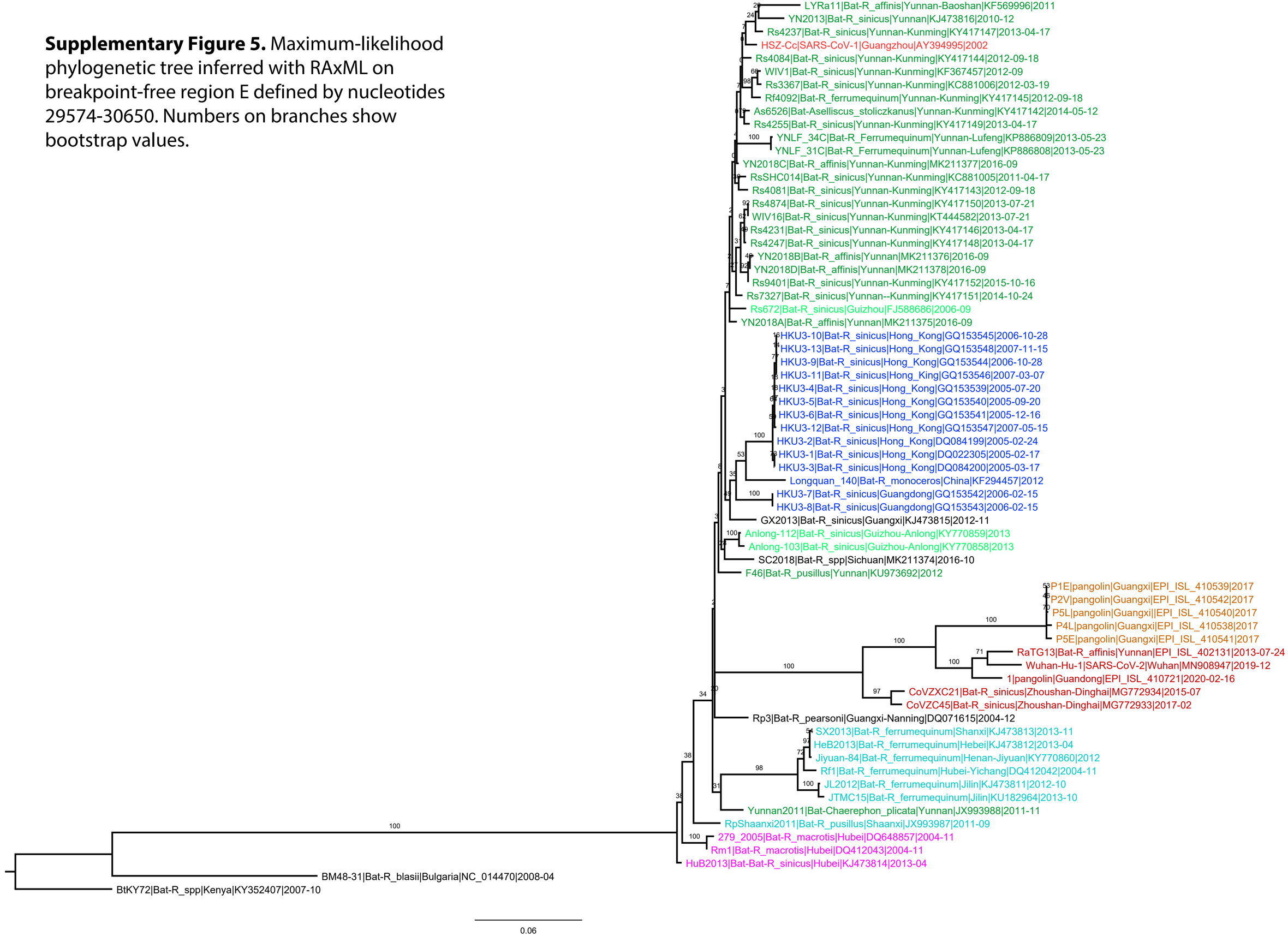
Supplementary Figure 3. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region C defined by nucleotides 9261-11795. Numbers on branches show bootstrap values.



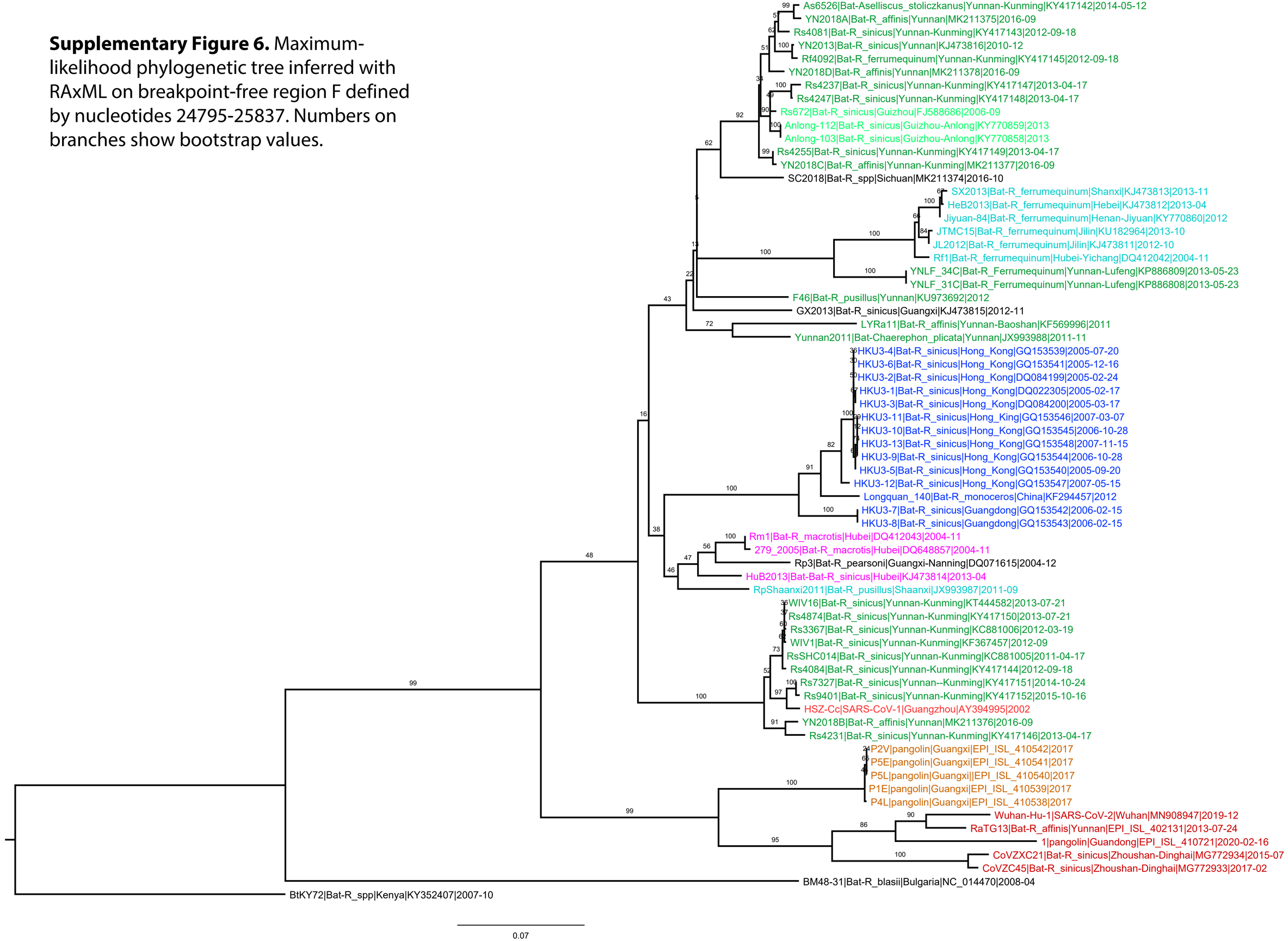
Supplementary Figure 4. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region D defined by nucleotides 27702-28843. Numbers on branches show bootstrap values.



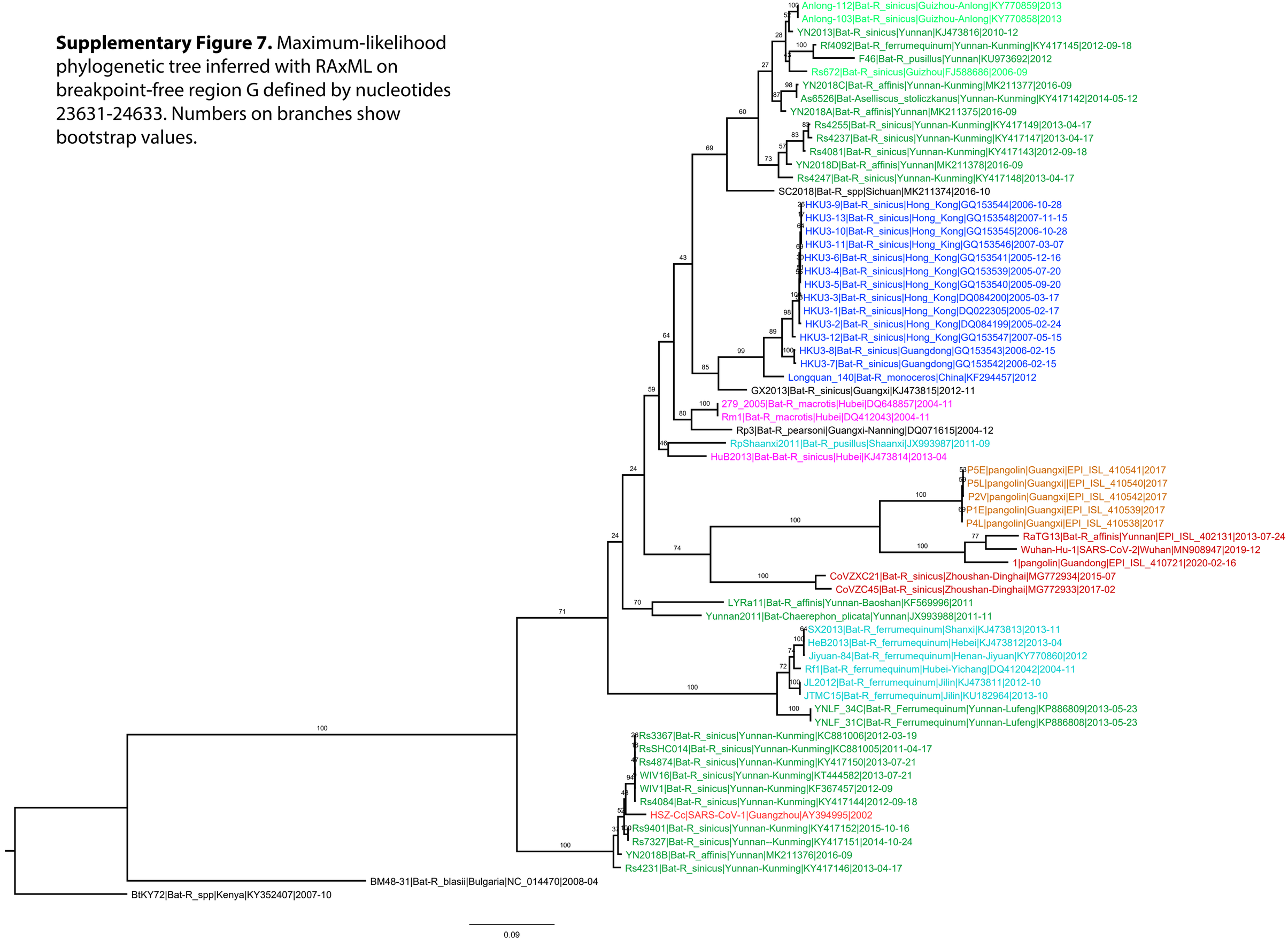
Supplementary Figure 5. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region E defined by nucleotides 29574-30650. Numbers on branches show bootstrap values.



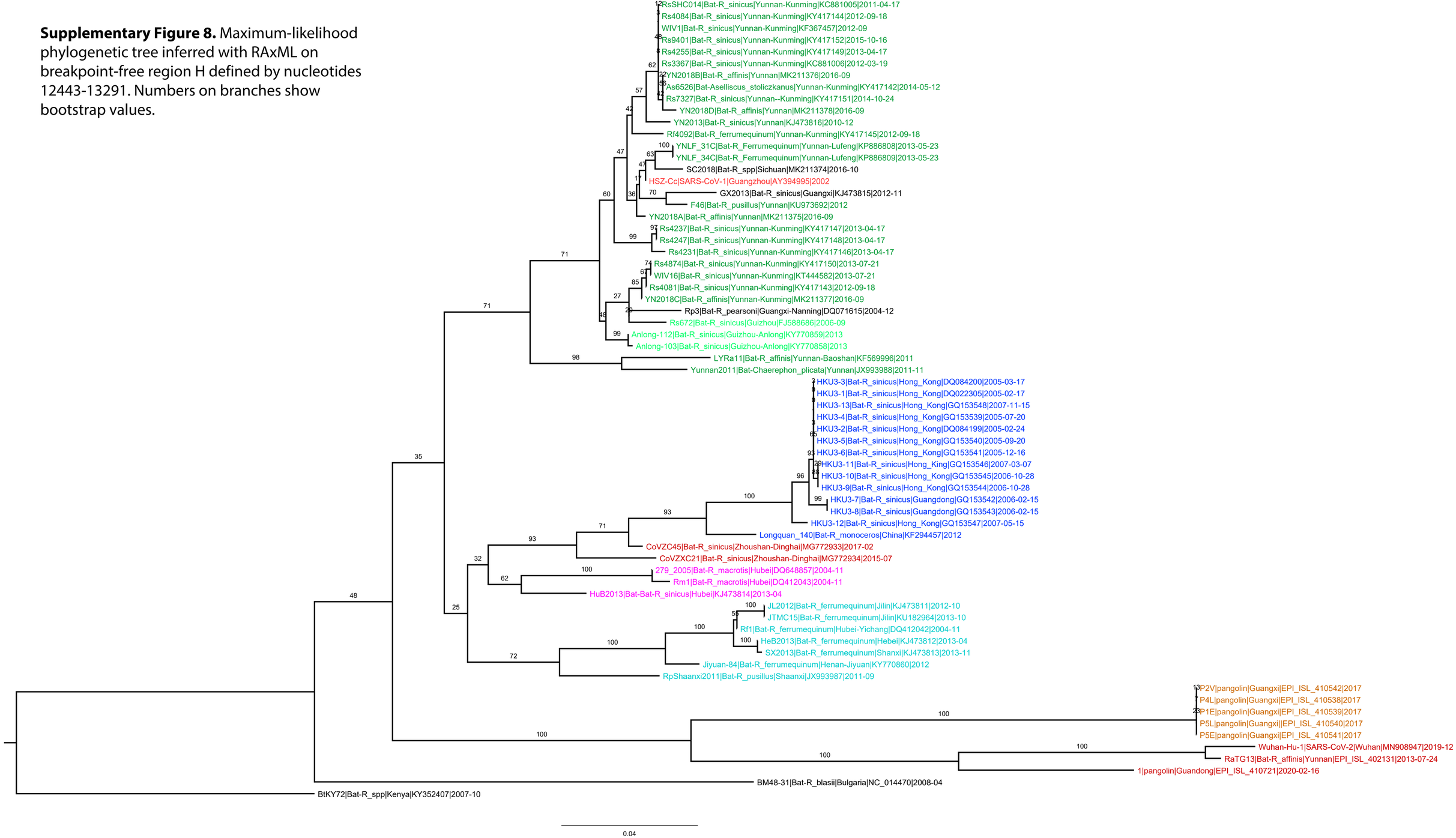
Supplementary Figure 6. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region F defined by nucleotides 24795-25837. Numbers on branches show bootstrap values.



Supplementary Figure 7. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region G defined by nucleotides 23631-24633. Numbers on branches show bootstrap values.



Supplementary Figure 8. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region H defined by nucleotides 12443-13291. Numbers on branches show bootstrap values.



Supplementary Figure 9. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region I defined by nucleotides 962-1686. Numbers on branches show bootstrap values.

Supplementary Figure 10. Maximum-likelihood phylogenetic tree inferred with RAxML on breakpoint-free region J defined by nucleotides 147-695. Numbers on branches show bootstrap values.