Check for updates

BRIEF REPORT

Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis

[version 1; peer review: awaiting peer review]

Andreas Martin Lisewski

Department of Life Sciences and Chemistry, Jacobs University Bremen, Bremen, 28759, Germany

V1 First published: 10 Sep 2021, 10:912
<https://doi.org/10.12688/f1000research.72956.1>
 Second version: 14 Oct 2021, 10:912
<https://doi.org/10.12688/f1000research.72956.2>
 Third version: 19 Jan 2022, 10:912
<https://doi.org/10.12688/f1000research.72956.3>
 Fourth version: 08 Mar 2022, 10:912
<https://doi.org/10.12688/f1000research.72956.4>
 Latest published: 04 Jul 2022, 10:912
<https://doi.org/10.12688/f1000research.72956.5>

Abstract

Background: Knowledge about the origin of SARS-CoV-2 is necessary for both a biological and epidemiological understanding of the COVID-19 pandemic. Evidence suggests that a proximal evolutionary ancestor of SARS-CoV-2 belongs to the bat coronavirus family. However, as further evidence for a direct zoonosis remains limited, alternative modes of SARS-CoV-2 biogenesis should be also considered.

Results: Here we show that genomes from SARS-CoV-2 and from closely related coronaviruses are differentially enriched with short chromosomal sequences from the yeast *S. cerevisiae* at focal positions that are known to be critical for virus replication, host cell invasion, and host immune response. Specifically, for SARS-CoV-2, we identify two sites: one at the start of the viral replicase domain, and the other at the end of the spike gene past its critical domain junction; for SARS-CoV-1, one at the start of the RNA dependent RNA polymerase gene, and the other at the start of the spike protein's receptor binding domain. As yeast is not a natural host for this virus family, we propose a directed passage model for viral constructs, including virus replicase, in yeast cells based on co-transformation of virus DNA plasmids carrying yeast selectable genetic markers followed by intra-chromosomal homologous recombination through gene conversion. Highly differential sequence homology data across yeast chromosomes congruent with chromosomes harboring specific auxotrophic markers further support this passage model. Model and data together allow us to infer a hypothetical tripartite genome assembly scheme for the synthetic biogenesis of SARS-CoV-2 and SARS-CoV-1.

Open Peer Review**Approval Status**

	1	2
version 5 (revision) 04 Jul 2022		
version 4 (revision) 08 Mar 2022		view
version 3 (revision) 19 Jan 2022	view	
version 2 (update) 14 Oct 2021		
version 1 10 Sep 2021		

1. Alexander Y Panchin,

2. Federico Di Lello,

Any reports and responses or comments on the article can be found at the end of the article.

Conclusions: These results provide evidence that the genome sequences of SARS-CoV-1, SARS-CoV-2, but not that of RaTG13 and all other closest SARS coronavirus family members identified, are carriers of distinct homology signals that might point to large-scale genomic editing during a passage of directed replication and chromosomal integration inside genetically modified yeast cells.

Keywords

SARS related coronavirus, SARS-CoV-2, SARS-CoV-1, COVID-19, virus passage, yeast *S. cerevisiae*, directed evolution, genomic transformation, genome editing, synthetic biology

This article is included in the **Cell & Molecular Biology** gateway.

This article is included in the **Coronavirus** collection.

Corresponding author: Andreas Martin Lisewski (a.lisewski@jacobs-university.de)

Author roles: Lisewski AM: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Lisewski AM. This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lisewski AM. **Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis [version 1; peer review: awaiting peer review]** F1000Research 2021, 10:912 <https://doi.org/10.12688/f1000research.72956.1>

First published: 10 Sep 2021, 10:912 <https://doi.org/10.12688/f1000research.72956.1>

Introduction

From the beginning of the COVID-19 pandemic, in March 2020, evidence was put forward that the outbreak of novel coronavirus SARS-CoV-2 within the human population was most likely a product of natural evolution¹. According to this view, COVID-19 is a zoonosis that probably originated from a species of closely related bat coronaviruses². Prior to a hypothetical spillover event, a recent ancestor to SARS-CoV-2 likely evolved inside bat host cells for many decades³. However, the natural evolution hypothesis of SARS-CoV-2 origin is currently not without considerable limitations: first, the difficulty in characterizing the evolutionary origin of the unusual poly-basic (RRAR) furin cleavage site at the S1/S2 junction of the SARS-CoV-2 spike (S) glycoprotein⁴; second, the discrepancy between an exponentially suppressed tropism of SARS-CoV-2 in *Rhinolophus sinicus* bat cells⁵ and the high susceptibility of SARS-CoV-2 toward cell entry via *Rhinolophus sinicus* angiotensin-converting enzyme 2, its primary entry receptor⁶; and third, the persistent inability to identify an intermediate ancestral host between human and the horseshoe bat *Rhinolophus affinis*. This species was reported to be the host of coronavirus RaTG13^{7,8}, currently the isolate with the highest sequence similarity to the SARS-CoV-2 genome, which is located on the same phylogenetic branch as *Rhinolophus sinicus* bat coronavirus⁹. Finding the last animal progenitor host of SARS-CoV-2 has been further complicated by a continued uncertainty about the origin of RaTG13 itself^{10,11}. In contrast to the natural evolution hypothesis for SARS-CoV-2, the above limitations do not necessarily apply to genetic engineering of viral genomes. For example, in the case of SARS coronavirus, it has long been established that introducing a synthetic poly-arginine construct at the furin cleavage site significantly increases the rate of entry into human cells compared with wild-type spike protein¹². Also before 2010, after a period of rapid progress in the understanding the relevant host-virus factors^{13,14}, natural barriers in host range of positive strand RNA viruses were rationally extended, leading to directed viral replication in new species including model organisms that originally were not permissive, such as the yeast *Saccharomyces cerevisiae*¹⁵. Accordingly, to transform yeast into a synthetic host for viral replication, the scheme has been to co-express viral RNA dependent RNA polymerase (RdRp) and, if also necessary for replication, additional viral factors on plasmids under the control of auxotrophic yeast selectable markers (YSM)¹⁶. These selectable markers are primarily there to direct cell lines into stable expression of desired plasmid DNA, but at the same time may function as entry gates for directed insertion of exogenous genetic material into yeast chromosomes¹⁷. In principle, once the RdRp and required auxiliary factors are selectively and functionally expressed, this approach applies to any replication competent SARS coronavirus RNA, including cloned or *de novo* synthesized genomic parts or even entire genomes, thus facilitating their replication as well as integration into the yeast genome. Our hypothesis is that such a passage would leave behind traces in the genomes of both the virus construct and the synthetic host.

Methods

SARS and SARS-like sarbecovirus whole genome nucleotide sequences were taken from the comprehensive sequence and phylogenetic analyses by Zhou *et al.*¹⁸ and from Li *et al.*⁹. In our study, sequences were selected only if they had a valid GenBank accession identifier or an NCBI Reference Sequence (RefSeq) accession identifier, as of 5 June 2021, resulting in the reference set of 13 whole genome virus sequences (see also Extended Data). BLAT whole genome comparative sequence analysis was performed using the BLAT public webserver (BLAT, RRID:SCR_011919) with options set "Genome: Search all" and "All results (no minimum matches)". Given each one of the corresponding 13 BLAT output tables produced from genomic alignments to the yeast *S. cerevisiae* (Extended data Tables S2 – S14), the profiled BLAT score, *pS*, was the genome-wide distribution of BLAT scores (output table column [SCORE]) weighted by the corresponding length of the homologous genomic region (output table distance between columns [START] and [END]). The cumulative profiled BLAT score *cS*, which was used as a genome-wide quantitative indicator of yeast (*S. cerevisiae*) homology, was the total sum over this distribution. After shifting *cS* by the sample's mean and dividing by its standard deviation, the resulting standardized BLAT *z*-score became then a relative indicator of sequence homology with *S. cerevisiae*. Sequence alignments for cross-validation were produced with lalign from the [fasta36-36.3.8/bin/lalign36](https://www.bioinformatics.org/~dave/fasta36-36.3.8/bin/lalign36) software package (version number 36.3.8) with parameter settings: `-f -12 -g 0 -E 1`. This parameter choice followed standard parameters for LALIGN. Sequence identities were calculated using the Clustal Omega public webserver (RRID:SCR_001591) with standard preset parameters.

Results

To interrogate the possibility that a similar passage through yeast cells took place within the family of SARS coronaviruses, we selected eight reference genomes¹⁸ for further analysis (see Methods): SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank reference NC_045512.2), *Rhinolophus affinis* bat coronavirus RaTG13 (MN996532.2), *Rhinolophus pusillus* SL-CoV ZXC21 (MG772934.1), *Rhinolophus pusillus* SL-CoV ZC45 (MG772933.1), *Rhinolophus acuminatus* bat coronavirus RacCS203 (MW251308.1), *Rhinolophus cornutus* bat coronavirus Rc-o319 (LC556375.1), SARS-CoV Urbani (AY278741.1), and MERS-CoV isolate HCoV-EMC/2012 (NC_019843.3). For comparative genomic sequence analysis we used a standard bioinformatics approach with the BLAST-like Alignment Tool (BLAT) (BLAT, RRID:SCR_011919)¹⁹. Each search from the above set of query sequences against the entire multi-species genome database produced a high number (between 1689 and 5083) of tiles, i.e. perfectly aligned short DNA sequences of length 11. A large majority of these tiles were repeatedly matched on the same two target genomes (out of 107 total; see also Extended data Table S1): SARS-CoV-2 (NC_045512.2), the only coronavirus genome in the database, and *S. cerevisiae* (SacCer3/S288c). In these instances, BLAT identified many

homologous regions by aggregating multiple tiles (Tables S2–S9), and to each homologous region it produced an integer score S , which is the number of perfectly matched positions therein. To obtain a genome-wide view of this homology signal we stacked together all homologous regions weighted by their individual alignment scores S , which resulted in an accumulated homology profile, pS (see Methods and Extended Data Figures S1 and S2). To remove its shortest-scale fluctuations, the profile was smoothed by a centered sliding window filter with window size of 200 nucleotides (nt). The output of eight genomic profiles (Figure 1 and Figure S2) were ordered by decreasing sequence identity to SARS-CoV-2.

For SARS-CoV-2, two prominent ($pS > 20$) peaks indicated highly localized profile scores at levels ~10-fold above the apparent background. A first peak ($P1$) reaching a top alignment score of 47 in the narrow genomic interval $[7191\ nt..7192\ nt]_{\max}$, and a second peak $P2$ over ~18,000 nt downstream with a score of 36 in the region $[25196\ nt..25212\ nt]_{\max}$ (see, Figure 1). To put these data into an established gene-function context these two maxima, with half-maximum widths $w_{1/2} = 215\ nt$ and $w_{1/2} = 219\ nt$, respectively, were annotated with available information from the closest and most specifically annotated genomic region in RefSeq, the NCBI Reference Sequence database²⁰. Thus $P1$ was closest to the start of the C-terminal domain of non-structural protein 3 (designated nsp3C), which extends over the interval $[6962\ nt..8552\ nt]$. The C-terminal domain of nsp3 is known to play a critical role in replication due to its direct interaction with nsp4, thereby facilitating virus-induced membrane rearrangement and replication complex formation; conversely, loss of nsp3C-nsp4 interaction abolishes SARS coronavirus replication²¹. $P2$ was located toward the 3' end of the open reading frame of the spike gene. Here it overlapped with the 3' end of the stretch that covers both the S1/S2 cleavage region and the S2 fusion subunit of the S protein (S_{S1/S2}, with interval $[23192\ nt..25187\ nt]$). The S_{S1/S2} domain includes the characteristic furin cleavage site at the S1/S2 junction²². Cleavage activates the nearby S2 fusion peptide and together they constitute an essential part in SARS-CoV-2 particle-dependent and particle-independent cell entry through fusion of viral and cellular membranes^{23,24}. A similar analysis for the RaTG13 viral genome identified only one isolated peak ($P3$) with a maximum profile score of 50 on the interval $[9713\ nt..9733\ nt]_{\max}$, and with $w_{1/2} = 230\ nt$. It intersected with the coding region of the C-terminal domain of nsp4 located at $[9770\ nt..10046\ nt]$ (Figure 1).

In the SARS coronavirus Urbani genome (SARS-CoV-1), two additional signals were detected: $P4$ with a maximum score $pS = 26$ at position $[13486\ nt..13497\ nt]_{\max}$ and $w_{1/2} = 222\ nt$; and a broader second peak, $P5$, with $pS = 41$ at position $[22286\ nt..22391\ nt]_{\max}$ and $w_{1/2} = 477\ nt$. $P4$ sharply co-localized with the N-terminus of the RdRp domain at $[13414\ nt..14470\ nt]$. $P5$ was annotated with the N-terminal part of the spike gene's receptor binding domain (Rbd) located in the interval $[22443\ nt..23199\ nt]$. In contrast to the five signals identified in these

three genomes, an equivalent analysis for the other three (RacCS203, Rc-o319, MERS-CoV) produced only negative results. Their accumulated homology profiles were evenly distributed across the entire genomes consistent with a low random score background from many short spurious matches. As a further specificity control, negative results were obtained (see, Figure S3 and Tables S10–S14) after profiling the five most closely SARS-CoV-1 related betacoronavirus isolates from five wild animals (civet, *Paradoxurus hermaphroditus*, *Paguma larvata*, *Aselliscus stoliczkanus*, and *Rhinolophus sinicus*), which together with SARS-CoV-2 occupy the same phylogenetic branch⁹. These data also lead to a highly differential signal of yeast homology in SARS-CoV-1, SARS-CoV-2 and RaTG13 genomes after calculating standardized z-scores (Figure 2) from the entire BLAT profiles produced to all 13 of the above sequences (Tables S2–S14). To cross-validate the detected yeast homology signals in $P1$ – $P5$, we used an independent sequence alignment method, LALIGN²⁵, which additionally produced statistics (E -values) for pairwise alignments. While the peaks $P1$ and $P2$, as well as $P4$ and $P5$, could be positively cross-validated, the $P3$ signal in RaTG13 detected by BLAT did not yield a statistically significant alignment with LALIGN, with its E -value reaching above 0.01 (see, Table S16 and Figure S4). Taken together, these highly differential data show that, for SARS-CoV-1 and for SARS-CoV-2, genes known to be critical for viral replication and host cell invasion display localized yeast homology at their flanking regions with limited extensions into the corresponding open reading frames.

To explain this yeast DNA enrichment pattern, we propose the following artificial passage model (Figure 3A): Its starting point is a doubly auxotrophic, synthetic yeast cell line with stable, heterologous expression of viral replicase complex (RdRp, optionally together with auxiliary factors for replication, Aux) from a plasmid under the control of a selectable marker YSM1. A second plasmid carries another auxotrophic yeast selectable marker YSM2, which originates from a different chromosome, and regulates the expression of a non-replicative segment of viral RNA (nrvRNA1). At this point, nrvRNA1 is any uninterrupted DNA segment from a SARS-coronavirus related genome prior to passage. Through homologous recombination, the target yeast chromosome is transformed and nrvRNA1 is integrated¹⁷ at the chromosomal site of the auxotrophy conferring allele homologous to YSM2. During passage cell growth double stranded DNA breaks occur, and breaks at both ends of nrvRNA1 ends, their flanking regions, and their homologous extensions into YSM2 are repaired preferably by intra-chromosomal gene conversion²⁶, i.e. through a non-crossover homologous recombination, and with the endogenous site as the homologous repair donor (Figure 3A).

If we assume that nrvRNA1 itself contains a copy of RdRp (and of Aux), then the above model implies that higher-order integration events¹⁷ will occur between the YSM1 plasmid and the primary site of integration. In effect, short segments from its YSM1 region will be also integrated into nrvRNA1. In

Figure 1. Profiled alignment scores (pS) from the alignment output to the query input of six SARS-coronavirus related full genome sequences (for SL-ZC45 and SL-ZXC21 profiles, see Figure S2). Alignment scores retrieved only from hits matching *S. cerevisiae* full genomic sequence assembly SacCer3/S288c. For the corresponding BLAT output, see Table S1, and Table S2–S9. Upper left, in parentheses, percent sequence identity of query genome to SARS-CoV-2. Of note, detected yeast homology signals, nucleotide sequence similarity, and geographic location (*region, country*) of first identified isolates do not converge. Abbreviations: nsp3C, non-structural protein 3 C-terminal domain [YP_009724389.1 (2,232..2,762)]; Rbd, receptor binding domain [SARS-CoV-2: YP_009724390.1 (319..541); SARS-CoV-1: AAP13441.1 (317..569)]; S_S1/S2, spike (S) protein S1/S2 domain cleavage region and the S2 fusion subunit [YP_009724390.1 (543..1,208)]; RdRpN, N-terminal region of the RNA dependent RNA polymerase [AAP13442.1 (4,383..4,735)].

Figure 2. Yeast (*S. cerevisiae*) standardized BLAT z-scores representing the relative homology signal from all alignment scores in thirteen representative SARS-related coronaviruses. BLAT z-scores were calculated using mean and standard deviations from the BLAT outputs (see, Table S2–S14 and Methods section). Evolutionary guide tree represents pairwise sequence identities between genomic sequences (see, Table S15). Grey shaded box indicates $[-1, 1]$ standard deviation interval.

this case the passage model specifically predicts that during *S. cerevisiae* growth nrvRNA1 will accumulate sequences from exactly two yeast chromosomes, i.e. those two which YSM1 and YSM2 originated from.

To test this prediction, we produced the score profile pS , but this time from the yeast sequence hits on each chromosome. For direct comparison, we then transformed each profile into a single number (cS), for all 16 chromosomes (mitochondrial chromosome excluded), by calculating the sum of pS over the entire chromosome length conditional on the cutoff $pS > 30$. In the case of SARS-CoV-2, this procedure resulted in two distinct peaks at chromosome number II and number XV (Figure 3B). For SARS-CoV-1, the highest two peaks were at chromosomes IV and V, followed by a much shallower peak on XVI with only 0.24 the height of IV. One peak was detected for RaTG13, also at XVI, whereas the other three viral genomes produced no signal at the chosen cutoff (see, Figure 3B, also for similar data without a cutoff). To further connect these data to our passage model, we attempted to match the seven most commonly used auxotrophic yeast selectable markers^{27,28} according to their chromosomal origin: ADE2 (adenine requiring *phosphoribosylaminoimidazole carboxylase*, on chromosome XV), HIS3 (histidine requiring *imidazoleglycerol-phosphate dehydratase*, chr. XV), LEU2 (leucine requiring *Beta-isopropylmalate dehydrogenase*, chr. III), LYS2 (lysine requiring *aminoacidipate reductase*, chr. II), MET15 (methionine requiring *O-acetyl homoserine-O-acetyl serine sulfhydrylase*, chr. XII), URA3 (uracil requiring *orotidine-5'-phosphate (OMP) decarboxylase*, chr.V), and TRP1 (tryptophan requiring

phosphoribosylanthranilate isomerase, chr. IV). In agreement with the model prediction, five out of the seven markers could be matched to the four highest of the five chromosome peaks detected in SARS-CoV-2 and SARS-CoV-1 (Figure 3B). This outcome especially implies that for SARS-CoV-2 the two auxotrophic markers (YSM1, YSM2) could be any pair from the triple (ADE2, HIS3, LYS2), and for SARS-CoV-1 either the pair (LEU2, TRP1) or (TRP1, LEU2). Thus SARS-CoV-2 and SARS-CoV-1 both did, but RaTG13 did not fit into this synthetic passage model.

These results further allowed us to infer a specific scheme for the synthetic biogenesis of SARS-CoV-2 and SARS-CoV-1 in transformed yeast cells (Figure 3C). The idea is to stitch together both outer DNA complements of a chosen viral genome with the inner segment nrvRNA1. For co-transformation and integration, two plasmids are designed that carry the YSM2 selectable marker with either the 5'-end (nrvRNA2) or the 3'-end (nrvRNA3) of the target virus genome along with some overlap into nrvRNA1 (regions 1' and 1'', respectively, see Figure 3C). Essential plasmid ingredients are also a transcriptional promoter for nrvRNA2, and a self-cleaving ribozyme (Rz) sequence for the correct 3'-end in nrvRNA3¹⁵. Once these three non-replicable RNA encoding segments are integrated on the chromosome in the correct order, expression of fully replicable virus (+)RNA begins and replication commences upon co-expression of the viral replicase complex (RdRp and Aux, controlled through the auxotrophic marker YSM1). The final step, assembly into a fully infectious viral particle, is conveniently achieved with a yeast virus-like-particle (VLP) expression system for the structural proteins S, E (envelope),

Figure 3. (A) First stage of passage model in the artificial host *S. cerevisiae* of a plasmid encoded, non-replicable viral RNA (nrvRNA1) originating from a SARS-CoV related virus. Primary integration of non-homologous nrvRNA1 sequence occurs through homologous recombination (HR) between the auxotrophic plasmid yeast selectable marker YSM1 (grey box) and its chromosomal homolog (striped grey box); higher-order homologous recombination follows on the flanking regions of nrvRNA1 through intra-chromosomal gene-conversion; co-expression of viral replicase complex (RdRp) and other auxiliary viral genes (Aux). Scheme in parts adapted from Compton *et al.* (1982), and from Alves-Rodrigues *et al.* (2006). P, yeast promoter; A_{tr}, poly-adenosine sequence. (B) Integrated profile scores, cS, from BLAT sequence hits on *S. cerevisiae* by chromosome number from the same six input sequences as in Figure 1 (purple columns); to calculate cS, a score profile cutoff $pS > 30$ was used. Without a cut-off ($pS > 0$), the same order emerged (black horizontal bars, maximum pS score at each chromosome; all other maximum pS scores from the other genomic queries are below, within shaded area). Five common yeast selectable markers are assigned to their chromosomes of origin. (C) Procedural second stage scheme for the synthetic biogenesis of SARS-CoV-2 and SARS-CoV-1. For the possible pairings of yeast selectable markers (YSM1, YSM2) matched in (B), genome editing of the three segments nrvRNA1, 2, 3 leads to a fully replicable virus (+)sense RNA. Virus (self-)assembly follows via expression of the structural proteins S, E, M, and N from an enhanced plasmid set Aux*. Rz, self-cleaving ribozyme.

M (membrane), and N (nucleocapsid) that can be used in parallel by an extended set of auxiliary proteins, Aux*²⁹. This hypothetical cellular factory may therefore produce the targeted, fully infectious viral particles without itself being infected by the virus produced.

Discussion

Taken together, our results reveal a highly differential homology signal on SARS-CoV-2 and SARS-CoV-1 genomes, which—according to our model—points to their history of targeted integration, recombination, and directed viral replication through passage in an artificial *S. cerevisiae* host. This genomic pattern suggests similar synthetic origins of SARS-CoV-1 and SARS-CoV-2, but at the same time robustly excludes all other clade members from this type of synthetic origin. A special case is RaTG13, which in our analysis produced both a simpler pattern and a weaker signal of common genetic history with yeast than the two mutually more similar homology signals found in SARS-CoV-1 and SARS-CoV-2. Yet RaTG13 is claimed to be much closer to SARS-CoV-2 evolutionarily⁷, i.e. 96% genomic sequence identity to SARS-CoV-2 against 80% between SARS-CoV-1 and the latter. This divergence suggests that if RaTG13 is assumed to be a product of natural evolution then both the sequences of SARS-CoV-1 and SARS-CoV-2 cannot be. Alternatively, the origin of RaTG13 could be artificial¹¹—along with SARS-CoV-2 and SARS-CoV-1³⁰, as our results also suggest. In this context, a point of uttermost importance would be the identification of the putative input progenitor SARS-CoV nucleotide sequence that went into passage. For example, it could be a highly pathogenic virus designed for, or naturally adapted to human cells and then selected for a transient artificial passage together with some genetic modifications³¹ of the virus to attenuate its virulence. Then its release back into the human host would likely initiate a rapid succession of complex reversal mutations toward its more pathogenic original structure^{30,31}. Intriguingly, during the first months of the SARS-CoV-2 outbreak, the genomic regions of nsp3 and spike protein had the highest mutational rate within the SARS-CoV-2 genome³² which may interfere with the yeast homology signal detected in the present study. During an epidemic, such reversal mutations toward an unidentified artificial genotype would be highly detrimental to most public health countermeasures, including pharmacological interventions and vaccinations. In contrast, through specific guidance of countermeasures such as vaccine development, detailed knowledge about the input progenitor's nucleotide sequence would effectively confer population immunity against the pathogen. Without such specific knowledge, our and further results about the conditions of the yeast artificial passage could offer important directions for SARS-CoV-2 antiviral interventions.

Data availability

Associated or additional data. All data underlying the results are available as part of the article and no additional source data are required.

Repository-hosted data. The following sequence data was retrieved from the NCBI GenBank repository:

1. Middle East respiratory syndrome-related coronavirus isolate HCoV-EMC/2012, complete genome (NCBI Reference Sequence: [NC_019843.3](#))
2. Severe acute respiratory syndrome-related coronavirus Rc-o319 RNA, complete genome (GenBank: [LC556375.1](#))
3. Bat SARS-like coronavirus isolate As6526, complete genome (GenBank: [KY417142.1](#))
4. Bat SARS-like coronavirus isolate Rs4874, complete genome (GenBank: [KY417150.1](#))
5. SARS coronavirus Urbani, complete genome (GenBank: [AY278741.1](#))
6. SARS coronavirus PC4-13, complete genome (GenBank: [AY613948.1](#))
7. SARS coronavirus civet020, complete genome (GenBank: [AY572038.1](#))
8. SARS coronavirus HC/SZ/61/03, complete genome (GenBank: [AY515512.1](#))
9. Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome (GenBank: [MG772933.1](#))
10. Bat SARS-like coronavirus isolate bat-SL-CoVZXC21, complete genome (GenBank: [MG772934.1](#))
11. Bat coronavirus RacCS203, complete genome (GenBank: [MW251308.1](#))
12. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome (GenBank: [NC_045512.2](#))
13. Bat coronavirus RaTG13, complete genome (GenBank: [MN996532.2](#))

Extended data

Harvard Dataverse: Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis. <https://doi.org/10.7910/DVN/BK8AL633>.

This project contains the following extended data files:

- Figure_S1.pdf : Profiled alignment scores (pS) without smoothing filter from the BLAT alignment output to the query input of six SARS-coronavirus related full genome nucleotide sequences.
- Figure_S2.pdf : Profiled alignment scores (pS) from the alignment output to the query input of SARS-coronavirus like genome sequences SL-ZC45 and SL-ZXC21.
- Figure_S3.pdf : Smoothed profile yeast BLAT alignment scores of five betacoronavirus isolates from five

wild animals, closely related to SARS-CoV-1 and SARS-CoV-2, after the phylogenetic analysis of Li *et al.* (2020): *Paradoxurus hermaphroditus* (palm civet) SARS coronavirus PC4-13 (GenBank AY613948), Civet SARS coronavirus civet020 (AY572038), *Paguma larvata* SARS coronavirus HC/SZ/61/03 (AY515512), *Rhinolophus sinicus* bat SARS-like coronavirus Rs4874 (KY417150), *Aselliscus stoliczkanus* bat SARS-like coronavirus As6526 (KY417142).

- Figure_S4.pdf : Alignment E-values (inverted, 1/E) as profiles across genomes of SARS-CoV-2, RaTG13, and SARS-CoV-1 calculated with the LALIGN local alignment method by using a sliding window approach with window sizes as given in Table S16.
- Table_S1.tab: Output from the [BLAT web server](#).
- Table_S2.tab: SARS-CoV-2/S. cerevisiae (sacCer3) BLAT results.
- Table_S3.tab: RaTG13/S. cerevisiae (sacCer3) BLAT results.
- Table_S4.tab: RacCS203/S. cerevisiae (sacCer3) BLAT results.
- Table_S5.tab: SL-CoV_ZC45/S. cerevisiae (sacCer3) BLAT results.

- Table_S6.tab: SL-CoV ZXC21/S. cerevisiae (sacCer3) BLAT results.
- Table_S7.tab: Rc-o319/S. cerevisiae (sacCer3) BLAT results.
- Table_S8.tab: SARS-CoV-1 Urbani/S. cerevisiae (sacCer3) BLAT results.
- Table_S9.tab: MERS-CoV/S. cerevisiae (sacCer3) BLAT results.
- Table_S10.tab: SARS coronavirus PC4-13/S. cerevisiae (sacCer3) BLAT results.
- Table_S11.tab: SARS coronavirus civet020/S. cerevisiae (sacCer3) BLAT results.
- Table_S12.tab: SARS coronavirus HC/SZ/61/03/S. cerevisiae (sacCer3) BLAT results.
- Table_S13.tab: SARS-like coronavirus isolate Rs4874 /S. cerevisiae (sacCer3) BLAT results.
- Table_S14.tab: SARS-like coronavirus isolate As6526/S. cerevisiae (sacCer3) BLAT results.
- Table_S15.tab: Percent identity matrix (Clustal 2.1).
- Table_S16.tab: Peak P1-P5 yeast homology signals detected by BLAT, and cross-validated by the LALIGN method.

References

- Andersen KG, Rambaut A, Lipkin WI, *et al.*: **The proximal origin of SARS-CoV-2.** *Nat Med.* 2020; **26**(4): 450–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- MacLean OA, Lytras S, Weaver S, *et al.*: **Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen.** *PLoS Biol.* 2021; **19**(3): e3001115.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Boni MF, Lemey P, Jiang X, *et al.*: **Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.** *Nat Microbiol.* 2020; **5**(11): 1408–17.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gallaher WR: **A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2.** *Arch Virol.* 2020; **65**(10): 2341–2348.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lau SKP, Wong ACP, Luk HKH, *et al.*: **Differential Tropism of SARS-CoV and SARS-CoV-2 in Bat Cells.** *Emerg Infect Dis.* 2020; **26**(12): 2961–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhang HL, Li YM, Sun J, *et al.*: **Evaluating angiotensin-converting enzyme 2-mediated SARS-CoV-2 entry across species.** *J Biol Chem.* 2021; **296**: 100435.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhou P, Yang XL, Wang XG, *et al.*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature.* 2020; **579**(7798): 270–3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhou P, Yang XL, Wang XG, *et al.*: **Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature.* 2020; **588**(7836): E6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Li C, Yang Y, Ren L: **Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species.** *Infect Genet Evol.* 2020; **82**: 104285.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sallard E, Halloy J, Casane D, *et al.*: **Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review.** *Environ Chem Lett.* 2021; 1–17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Deigin Y, Segreto R: **SARS-CoV-2's claimed natural origin is undermined by issues with genome sequences of its relative strains: Coronavirus sequences RaTG13, MP789 and RmYN02 raise multiple questions to be critically addressed by the scientific community.** *Bioessays.* 2021; **43**(7): e2100015.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Belouzard S, Chu VC, Whittaker GR: **Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites.** *Proc Natl Acad Sci U S A.* 2009; **106**(14): 5871–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kushner DB, Lindenbach BD, Grdzelishvili VZ, *et al.*: **Systematic, genome-wide identification of host genes affecting replication of a positive-strand RNA virus.** *Proc Natl Acad Sci U S A.* 2003; **100**(26): 15764–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ahlquist P, Noueiry AO, Lee WM, *et al.*: **Host Factors in Positive-Strand RNA Virus Genome Replication.** *J Virol.* 2003; **77**(15): 8181–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Alves-Rodrigues I, Galão RP, Meyerhans A, *et al.*: **Saccharomyces cerevisiae: A useful model host to study fundamental biology of viral replication.** *Virus Res.* 2006; **120**(1–2): 49–56.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pogany J, Panavas T, Servienu E, *et al.*: **A high-throughput approach for studying virus replication in yeast.** *Curr Protoc Microbiol.* 2010; Chapter 16: Unit16j.1.
[PubMed Abstract](#) | [Publisher Full Text](#)

17. Compton JL, Zamir A, Szalay AA: **Insertion of nonhomologous DNA into the yeast genome mediated by homologous recombination with a cotransforming plasmid.** *Mol Gen Genet.* 1982; **188**(1): 44–50.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Zhou H, Ji J, Chen X, *et al.*: **Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses.** *Cell.* 2021; **184**(17): 4380–4391.e14.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Kent WJ: **BLAT—The BLAST-Like Alignment Tool.** *Genome Res.* 2002; **12**(4): 656–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. **National Center for Biotechnology Information (NCBI).** Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988.
[Reference Source](#)
21. Sakai Y, Kawachi K, Terada Y, *et al.*: **Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication.** *Virology.* 2017; **510**: 165–74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Coutard B, Valle C, de Lamballerie X, *et al.*: **The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade.** *Antiviral Res.* 2020; **176**: 104742.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Papa G, Mallery DL, Albeck A, *et al.*: **Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion.** *PLoS Pathog.* 2021; **17**(1): e1009246.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Theuerkauf SA, Michels A, Riechert V, *et al.*: **Quantitative assays reveal cell fusion at minimal levels of SARS-CoV-2 spike protein and fusion from without.** *iScience.* 2021; **24**(3): 102170.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Huang X, Miller W: **A time-efficient, linear-space local similarity algorithm.** *Adv Appl Math.* 1991; **12**(3): 337–57.
[Publisher Full Text](#)
26. Agmon N, Pur S, Liefshitz B, *et al.*: **Analysis of repair mechanism choice during homologous recombination.** *Nucleic Acids Res.* 2009; **37**(15): 5081–92.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Pronk JT: **Auxotrophic yeast strains in fundamental and applied research.** *Appl Environ Microbiol.* 2002; **68**(5): 2095–100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. **Commonly used auxotrophic markers.** SGD-Wiki. [cited 2021 Jun 3].
[Reference Source](#)
29. Nooraei S, Bahrulolum H, Hoseini ZS, *et al.*: **Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers.** *J Nanobiotechnology.* 2021; **19**(1): 59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Xu D, Sun H, Su H, *et al.*: **SARS coronavirus without reservoir originated from an unnatural evolution, experienced the reverse evolution, and finally disappeared in the world.** *Chin Med J (Engl).* 2014; **127**(13): 2537–42.
[PubMed Abstract](#)
31. Jimenez-Guardeño JM, Regla-Nava JA, Nieto-Torres JL, *et al.*: **Identification of the Mechanisms Causing Reversion to Virulence in an Attenuated SARS-CoV for the Design of a Genetically Stable Vaccine.** *PLoS Pathog.* 2015; **11**(10): e1005215.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Pereson MJ, Mojsiejczuk L, Martínez AP, *et al.*: **Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence.** *J Med Virol.* 2021; **93**(3): 1722–31.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Lisewski AM: **Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis.** Harvard Dataverse, V1, UNF: 6:BC1twHAK9jEwqcRfghk4Dg== [fileUNF]. 2021.
<http://www.doi.org/10.7910/DVN/BK8AL6>

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research