Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences [version 5; peer review: 2 not approved]

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

Andreas Martin Lisewski

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

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 also be considered.

Results: Here we show that the genomes of SARS-CoV-2 and SARS-CoV-1 significantly diverge from other SARS-like coronaviruses through short chromosomal sequences from the yeast S. cerevisiae at focal positions that are known to be critical for host cell invasion, virus replication, and host immune response. For SARS-CoV-1, we identify two sites: 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; for SARS-CoV-2, one at the start of the viral replicase domain, and the other toward the end of the spike gene past its domain junction. At this junction, we detect a highly specific stretch of yeast origin covering the critical furin cleavage site insert PRRA, which has not been seen in other lineage b betacoronaviruses. As yeast is not a natural host for this virus family, we propose an artificial synthesis model for viral constructs 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 yeast sequence patterns
congruent with chromosomes harboring specific auxotrophic markers further support yeast artificial synthesis.

**Conclusions:** These results provide evidence that the genomes of SARS-CoV-1 and SARS-CoV-2 contain sequence information that points to their artificial synthesis in genetically modified yeast cells. Our data specifically allow the identification of the yeast *S. cerevisiae* as a potential recombination donor for the critical furin cleavage site in SARS-CoV-2.

**Keywords**
SARS related coronavirus, SARS-CoV-2, SARS-CoV-1, COVID-19, virus artificial synthesis, 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:** © 2022 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. Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences [version 5; peer review: 2 not approved] F1000Research 2022, 10:912 https://doi.org/10.12688/f1000research.72956.5

**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 evolution. According to this view, COVID-19 is a zoonosis that probably originated from a species of closely related bat coronaviruses. Prior to a hypothetical spillover event, a recent ancestor to SARS-CoV-2 likely evolved inside bat host cells for many decades. 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 polybasic (PRRAR) furin cleavage site (FCS) at the S1/S2 junction of the SARS-CoV-2 spike (S) glycoprotein; second, the discrepancy between an exponentially suppressed tropism of SARS-CoV-2 in Rhinolophus sinicus bat cells and the high susceptibility of SARS-CoV-2 toward cell entry via Rhinolophus sinicus angiotensin-converting enzyme 2, its primary entry receptor; 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, currently the isolate with the closest evolutionary relationship to the SARS-CoV-2 genome, which is located on the same phylogenetic branch as Rhinolophus sinicus bat coronavirus. Finding the last animal progenitor host of SARS-CoV-2 has been further complicated by the fact that RaTG13 lacks a homologous FCS sequence, and by a continued uncertainty about the origin of RaTG13 itself. Thus even in the third year after the emergence of COVID-19, a more closely related evolutionary progenitor sharing naturally the unusual functional characteristics, like the S1/S2 FCS, with SARS-CoV-2 has yet to be found in China, or outside. In contrast to the natural evolution hypothesis for SARS-CoV-2, the above limitations do not necessarily apply to genetic engineering of viral genomes in laboratory environments. For example, the theory that SARS-CoV-2 could be the product of laboratory manipulation involving a passage through cell culture has been critically discussed. In addition, for 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 protein. Also before 2010, after a period of rapid progress in the understanding the relevant host-virus factors, natural barriers in host range of RNA viruses were rationally extended, leading to artificial genome assembly and directed viral replication in new species including model organisms that originally were not permissive, such as the yeast Saccharomyces cerevisiae. Accordingly, to transform budding yeast into a artificial host for viral synthesis and replication, the general scheme for both positive and negative sense RNA has been to co-express a viral RNA dependent RNA polymerase (RdRp) and, if also necessary for replication, additional factors on plasmids under the control of auxotrophic yeast selectable markers (YSM). For betacoronaviruses, the genus to which all SARS-like coronaviruses belong, the key experimental step had been described already in 2002 by Yount et al., where the essential genomic replicase domain located between nsp3 (non-structural protein 3) and RdRp was cloned and robustly expressed in yeast from a standard pYES2 vector carrying the URA3 (uracil requiring orotidine-5’-phosphate decarboxylase) gene as its only auxotrophic marker. Yeast selectable (auxotrophic) markers, and specifically URA3, have been since described and used experimentally to direct cell lines into stable expression of a large variety of virus derived cDNA constructs and clones, including many human pathogens such as recombinant SARS coronavirus (see columns 14–15 in 20). At the same time, plasmids with YSMs had already been known to function as entry gates for directed insertion of exogenous genetic material into yeast chromosomess. This insertion process, by means of homologous recombination, is a priori independent of both transcription and the optional RdRp driven RNA replication cycle. Following this rationale, Thao et al. have more recently demonstrated that, prior to transcription into infectious RNA, several overlapping genomic domains efficiently assemble into a full-length SARS-CoV-2 coronavirus clone on a yeast artificial chromosome (YAC) using HIS3 (histidine requiring imidazoleglycerolphosphate dehydratase) as YSM. YAC assembly therefore facilitates recombination with endogenous host chromosomess resulting in viral RNA or infectious clones with “traces of yeast genomic DNA.” Our hypothesis is that such artificial synthesis in yeast cells would leave behind traces in the genomic sequences of both the virus construct and the synthetic host.

Methods

SARS and SARS-like betacoronavirus whole genome nucleotide sequences were selected following the comprehensive sequence and phylogenetic analyses by Zhou et al. and from Li et al. 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). This set was extended by 5 additional genomic sequences, BANAL 20-52/20-103/20-236, icSARS-CoV-2,
and rSARS-CoV-2 YAC (see, Repository-hosted data, for he full list). 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)". Each BLAT search from the above set of query sequences against the entire multi-species genome database resulted in a high number of tiles, i.e. perfectly aligned short DNA sequences of length 11, to the yeast \( S. cerevisiae \) (SacCer3/S288c). BLAT identified many homologous regions by aggregating multiple tiles, and to each homologous region it produced an integer score \( S \), which is the number of perfectly matched positions therein. Each of the 18 corresponding BLAT genomic alignments to the yeast \( S. cerevisiae \) (Extended data Tables S2 – S19) produced a profiled BLAT score, \( pS \), which was the genome-wide distribution of \( S \) scores (output table column [SCORE]) weighted by the corresponding length of the homologous genomic region (output table distance between columns [START] and [END]). To remove its shortest-scale fluctuations, these profiles were smoothed by a centered sliding window filter with window size of 200 nucleotides (nt). The cumulative profiled BLAT score, \( cS \), was the total sum over this distribution (excluding matches to mitochondrial DNA). Using \( cS \), a genome-wide measure of yeast homology was generated through the statistical null hypothesis that those profiles, for which no BLAT yeast peaks with \( pS > 20 \) were detected, followed a normal distribution \( N(0,1) \) in their standardized \( cS \) values. This distribution was therefore sampled by shifting \( cS \) values by the sample’s mean and dividing by its standard deviation. The resulting standardized BLAT \( p \)-values, returned by the normal cumulative distribution function and transformed into negative logarithms, became a statistical test of the above null hypothesis and, as such, a measure of sequence homology with \( S. cerevisiae \). A statistical significance (chosen above a level of 0.05) test for pairs of \( p \)-values, \( p_r < p_t \), was calculated with conditional probabilities \( p_r/p_t \). Negative log \( p \)-value for rSARS-CoV-2 YAC was the average of 12 Sanger–sequenced yeast artificial chromosomes with detected mutations (relative to SARS-CoV-2 Y AC was the average of 12 Sanger–sequenced yeast artificial chromosomes). To interrogate the possibility that a similar passage through yeast cells took place within the family of SARS coronaviruses, we initially selected eight reference genomes for further analysis (see Methods): SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank reference NC_045512.2), \( \text{Rhinolophus} \text{affinis} \) bat coronavirus RaTG13 (MN996532.2), \( \text{Rhinolophus} \text{pilus} \) SL-CoV ZXC21 (MG772934.1), \( \text{Rhinolophus} \text{pusillus} \) SL-CoV ZC45 (MG772933.1), \( \text{Rhinolophus} \text{acuminatus} \) bat coronavirus RacCS203 (MW251308.1), \( \text{Rhinolophus} \text{cornutus} \) bat coronavirus Rc-0319 (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). The rationale was that BLAT, a more accurate genome sequence alignment tool than other conventional approaches, would detect such traces of yeast DNA. In line with this hypothesis, a large majority of BLAT matches was on the same two target genomes (see also Extended data Table S1): SARS-CoV-2 (NC_045512.2), a self-match to the only lineage \( \beta \) bat-coronavirus genomic sequence in the BLAT database, and \( S. cerevisiae \) (SacCer3/S288c). To obtain a genome-wide view of this yeast homology pattern 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).

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..7192]max, and a second peak \( P2 \) over ~18,000 bases downstream with a score of 36 in the region [25916..25212]max (see, Figure 1). To put these data into an established gene-function context these two maxima, with half-maximum widths \( w_{\text{1/2}} = 215 \) and \( w_{\text{1/2}} = 219 \), respectively, were annotated with available information from the closest and most specifically annotated genomic region in RefSeq, the NCBI Reference Sequence database. 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..8552]. 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 replication. \( P2 \) 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
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 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 brackets, percent sequence identity of query genome to SARS-CoV-2. Profiles are ordered by decreasing sequence identity to SARS-CoV-2. Of note, detected yeast homology patterns, nucleotide sequence similarity, and geographic location (region, country) do not converge. *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)].
S protein (S\_S1/S2, with interval [23192..25187]). The S\_S1/ S2 domain includes the characteristic furin cleavage site at the S1/S2 junction\(^3\), which has previously been described as unique to SARS-CoV-2 among lineage b betacoronaviruses\(^1\). 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\(^10,34\). 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..9733\_max], and with \(w_{10} = 230\). It intersected with the coding region of the C-terminal domain of nsp4 located at [9770..10046] (Figure 1).

Of special interest in this analysis was a 16 base sequence (TTCTCTCGCGGGGCA) near P2 between position 23599 and 23614, which corresponded to the furin cleavage site and identically aligned with bases [810386..810401] from \(S.\) cerevisiae chromosome XIII. In the forward +1 reading frame this sequence encodes the amino acids SPRRA and thus includes the critical PRRA insert in SARS-CoV-2. This shared sequence could be extended to 17 consecutive nucleotides (TTCTCTCGCGGGGCA), which are identically found in known SARS-CoV-2 variants that emerged after serial passage in cell culture (e.g., GenBank entry MZ995185.1), and—at codon level—are also compatible with the entire ancestral SPRRA motif. As such, TTCTCTCGCGGGGCA represented the longest identical nucleotide sequence between SARS-CoV-2 clade and \(S.\) cerevisiae lineage that covered the furin cleavage site. To test the specificity of TTCTCTCGCGGGGCA across potential eukaryotic host organisms, we performed BLAT and standard blastn sequence searches. For BLAT, no hits were found except for the one in yeast. When restricted to ‘genomic DNA’ sequence records dated before 2020, an extensive blastn search among all GenBank eukaryotic genomic sequences produced no identical sequence hits other than the \(S.\) cerevisiae matches (see, Extended Data File S1). A similar result was obtained when potential host specificity was tested with the shorter TTCTCTCGCGGGCA sequence (Extended Data File S2 and S3), as well as with the entire SARS-CoV-2 genomic sequence (Extended Data File S4 and S5). These data specifically identified the yeast \(S.\) cerevisiae as a potential genomic recombination donor of the critical FCS in the spike protein of SARS-CoV-2.

In the SARS coronavirus Urbani genome (SARS-CoV-1), two additional signals were detected: P4 with a maximum score \(pS = 26\) at position [13486..13497\_max] and \(w_{10} = 222\); and a broader second peak, P5, with \(pS = 41\) at position [22286..22391\_max] and \(w_{10} = 477\). P4 sharply co-localized with the N-terminus of the RdRp domain at [13414..14470]. P5 was annotated with the N-terminal part of the spike gene’s receptor binding domain (Rbd) located in the interval [22443..23199]. In contrast to the five signals identified in these three genomes, an equivalent analysis for the other five (RacCS203, SL-ZC45, SL-ZCX21, 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, \(P\) adaxiosurus hermaphroditus, \(P\) aguma larvata, Asellus stoliczkanus, and \(R\) hinolophus sinicus), which together with SARS-CoV-2 occupy the same phylegenetic branch\(^35\). These data collectively produced a differential yeast homology signature, with only SARS-CoV-1, SARS-CoV-2 and RaTG13 statistically significant, after calculating standardized \(p\)-values (Figure 2) from the entire BLAT profiles to all 13 of the above sequences (Tables S2–S14). This analysis also included the three recently identified bat SARS-like coronavirus genomic sequences from the same clade as RaTG13, i.e., BANAL-20-52, BANAL-20-103, and BANAL-20-236 (Tables S15–S17), none of which yielded statistically significant \(p\)-values. To cross-validate the detected yeast homology signals in P1–P5, we also used an independent sequence alignment method, LALIGN\(^35\), which additionally produced statistics (\(E\)-values) for pairwise alignments. While the peaks P1 and P2, as well as P4 and P5, could be positively 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 S21 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.

As a further validation of our method, we turned to two derived genomic sequences of SARS-CoV-2: the recombinant \(r\)SARS-CoV-2, assembled through reverse genetics into a YAC\(^34\), and the infectious clone icSARS-CoV-2, assembled without yeast through a cell-free \textit{in vitro} ligation method\(^36\). Even though their genomic sequences were identical to SARS-CoV-2 Wuhan-Hu-1 at >99.9% level (see, Table S20), our data (Figure 2) differentiated between \(r\)SARS-CoV-2 YAC, which relative to SARS-CoV-2 Wuhan-Hu-1 yielded a significant (\(p<0.0001\)) increase, and icSARS-CoV-2, which produced no significant difference to this SARS-CoV-2 reference genome. These data suggest that our approach may sensitively and specifically detect traces from a given yeast artificial synthesis history in recombinant SARS-CoV-2.

To explain the observed yeast DNA enrichment pattern in SARS coronavirus genomic sequences, we propose the following artificial synthesis model (Figure 3A): Its starting point is a doubly auxotrophic, synthetic yeast cell line with stable, heterologous expression of a viral replication 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 encoding for viral RNA (nrvRNA1). At this point, nrvRNA1 is any uninterrupted DNA segment
from a SARS coronavirus related genome. Through homologous recombination, the target yeast chromosome is transformed and nrvRNA1 is integrated at the chromosomal site of the auxotrophy conferring allele homologous to YSM2. During 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 conversion, 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 sequences homologous to the YSM1 carrying plasmids, e.g. through ends with overlaps, then the above model implies that higher-order integration events 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.

To test this prediction, we produced the score profile $p_S$, but this time from the yeast sequence hits on each chromosome. For direct comparison, we then transformed each profile into a single number ($c_S$), for all 16 chromosomes (mitochondrial chromosome excluded), by calculating the sum of $p_S$ over the entire chromosome length conditional on the cutoff $p_S > 30$. 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 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 model, we attempted to match the seven most commonly used auxotrophic yeast selectable markers 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 aminoadipate 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 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). For SARS-CoV-1 there was a marked URA3 associated peak (on chromosome V) with a yeast score that exceeded all other observed values by at least 2 orders of magnitude. For SARS-CoV-2, the maximum peak was associated with HIS3 (and ADE2) selectable markers (on chromosome XV). These data imply that for SARS-CoV-2 the two auxotrophic markers (YSM1, YSM2) could be any pair from the triple (HIS3, ADE2, LYS2), and for SARS-CoV-1 the
Figure 3. Yeast artificial synthesis model for SARS coronavirus 1 and 2. (A) First stage assembly and transformation in the artificial host *S. cerevisiae* of a plasmid encoded, non-replicable viral RNA (nrrRNA1) originating from a SARS-CoV related virus. Primary integration of non-homologous nrrRNA1 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 nrrRNA1 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<sub>n</sub>, 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); cS, score profile sum with cutoff pS > 30. 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) Inferred second stage for the synthetic biogenesis of SARS-CoV-2 and SARS-CoV-1. Yeast selectable markers pairings (YSM1, YSM2) matched in (B), chromosomal transformation by three segments nrrRNA1, 2, 3 transcribes into a virus (+)sense RNA, while also recombining with a given yeast artificial chromosome (YAC). Virus-like particle (self-)assembly follows by expression of the structural proteins S, E, M, and N from an enhanced plasmid set Aux*. Rz, self-cleaving ribozyme; YC, yeast chromosome.
pair (URA3, TRP1). Thus SARS-CoV-1 and SARS-CoV-2 both did, but RaTG13 did not completely fit into this artificial yeast model. These results allowed us to infer a scheme for the artificial biogenesis of SARS-CoV-2 and SARS-CoV-1 in transformed yeast cells (Figure 3C). A minimum of three genomic fragments, designed through reverse genetics to assemble into a YAC, provide two outer DNA clone complements of a chosen progenitor SARS viral genome together with the inner segment nrvRNA1. For transformation, integration and assembly, the plasmids carry a YSM2 selectable marker with either the 5′-end (nrvRNA2) or the 3′-end (nrvRNA3) of the target virus genome, each with a specific overlap into both nrvRNA1 ends (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 nrvRNA3. Once these virus genomic RNA encoding segments are integrated into a yeast endogenous chromosome, homologous recombination with the YAC (if concurrently present) and genomic transcription of viral RNA follow. In contrast to the targeted sequence of the YAC, which was designed to not express yeast DNA, the recombinant viral DNA from the transformed chromosome is homologous to the entire YAC while also enriched with yeast genomic DNA. Virus RNA replication then commences upon its further transfection into replication competent host cells, or through additional co-expression of a viral replicase complex (RdRp and Aux, controlled through the auxotrophic marker YSM1, Figure 3C). A final optional step, assembly into a viral particle, may be achieved with a yeast virus-like-particle (VLP) expression system for the structural proteins S, E (envelope), M (membrane), and N (nucleocapsid) that can be expressed from an auxiliary plasmid, Aux* [10,11].

**Discussion**

Our results reveal a previously unidentified, highly differential sequence pattern in SARS-CoV-2 and SARS-CoV-1 genomes, which—according to our model—points to their history of targeted transformation, integration and recombination in an artificial *S. cerevisiae* host. This orthogonal layer of genomic sequence information significantly deviates from the standard reconstructed natural evolutionary history of lineage b (Sarbecovirus) by indicating a yeast artificial origin of SARS-CoV-1 and SARS-CoV-2. At the same time, our data robustly excludes all other analyzed clade members from this type of yeast artificial 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 evolutionarily, 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 artificial—along with SARS-CoV-2 and SARS-CoV-1 [13], as our results also suggest. As a controversial candidate for a natural ancestral or intermediate SARS coronavirus host, palm civets had in fact never been identified as the original animal reservoir of SARS coronavirus, and a conclusive zoontic host identification or characterization of a natural origin has not been given either. For example, the frequently cited work by Kan et al. concluded that “when SARS-CoV-like virus arrives at an animal market, the majority of palm civets, if not all, will become infected, and that the virus will evolve rapidly in animals to cause disease. Therefore, it is critical to identify the original animal reservoir to remove the continuing threat of SARS.” This conclusion, and further evidence that palm civets were not even an intermediate host, were supported by phylogenetic analysis for the initial stages of the SARS epidemic, where a rooted phylogenetic tree placed the earliest human virus lineage before the first civet infections, and with both viral lineages originating from an unknown reservoir in late 2002. To date, this uncertainty and controversy around the assumed natural origin of SARS-CoV persists, as no close relatives to SARS-CoV-1 or SARS-CoV-2 have been identified in diverse local animals, including palm civets, from relevant Chinese regions [13,14].

If SARS coronavirus had indeed an artificial yeast origin, an important point would be the identification of the putative input progenitor SARS-CoV like nucleotide sequence that went into yeast for assembly. For example, it could be a highly pathogenic virus designed for, or adapted to human cells and subsequently selected for yeast artificial assembly and passage together with some genetic modifications of the virus to attenuate its virulence. Indeed, yeast reverse genetics in the context of stable, genetically easily modifiable and scalable virus vaccine production have been described [10,11]. Then its release back into the human host would likely initiate a rapid succession of complex reversal mutations toward its more pathogenic original structure [11,14]. 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 genome which may have interfered with the yeast homology regions 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.

With regard to the most characteristic sequence signature of SARS-CoV-2, Andersen et al. questioned the possibility that the polybasic cleavage site at the critical S domain junction was acquired during passage in cell culture. However, according to our data, this cleavage site is specifically compatible with a recombination event including chromosome XIII of *S. cerevisiae*, which shares a unique nucleotide sequence that encodes the necessary insert PRRA. From a host viewpoint, our results suggest that an artificial origin of both
SARS-CoV-2 and SARS-CoV-1 should coincide with an emergence of synthetic yeast lineages unnaturally enriched in their chromosomes, due to recombination, with sequences from these coronaviruses. Arguably, such claim would be testable with sequencing data from laboratory and field samples. Collectively, our results offer a new lead for the further understanding of SARS coronavirus origins.

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 Re-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, SARS-CoV-2, complete genome (GenBank: NC_045512.2)
13. Bat coronavirus RaTG13, complete genome (GenBank: MN996532.2)
14. Bat coronavirus isolate BANAL-20-52/Laos/2020, complete genome (GenBank: MZ937000.1)
15. Bat coronavirus isolate BANAL-20-103/Laos/2020, complete genome (GenBank: MZ937001.1)
16. Bat coronavirus isolate BANAL-20-236/Laos/2020, complete genome (GenBank: MZ937003.1)
17. Infectious clone, iCSARS-CoV-2, complete genome (GenBank: MT461669.1)
18. Yeast artificial chromosome (YAC) infections reconstructed genome, rSARS-CoV-2 YAC, complete genome (GenBank: MT108784.1)

Extended data
Harvard Dataverse: Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences. https://doi.org/10.7910/DVN/BK8AL6

This project contains the following extended data files:

- Data_File_S1 : blastn output text file for the input 17 nucleotide sequence TTCCTCCTCGGCGGGGCA. WSIH denotes Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom.
- Data_File_S2 : blastn output text file for the input 16 nucleotide sequence TTCCTCCTCGGCGGGGCA searched against all eukaryotic species records the 'Nucleotide collection (nt)' sequence database (database update 23 February 2022). Output restricted to identical hits of length 16. blastn parameters used standard values except E-threshold (100) and gap cost (6).
- Data_File_S3 : blastn output text file for the input nucleotide sequence NC_045512.2 (SARS-CoV-2 isolate Wuhan-Hu-1 with its poly-A end removed) searched against the 127 different eukaryotic species found in Extended Data File S2. blastn parameters used standard values except E-threshold (100) and gap cost (6).
- Data_File_S4 : blastn output text file for the input nucleotide sequence NC_045512.2 (SARS-CoV-2 isolate Wuhan-Hu-1 with its poly-A end removed) searched against all eukaryotic species records in the 'Nucleotide collection (nt)' sequence database (database update 23 February 2022). blastn parameters used standard values except E-threshold (100) and gap cost (6).
- Data_File_S5 : blastn top hits (E < 0.70) from Extended Data File S4 filtered to 'genomic DNA' sequence records deposited prior to 2020.
- 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 six 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.


12. 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, MERS-CoV and RmYN02 raise multiple questions to be critically addressed by the scientific community. Bioessays. 2021; 43(7): e2100015.


Open Peer Review

Current Peer Review Status: ✗ ✗

Version 4

Reviewer Report 07 June 2022

https://doi.org/10.5256/f1000research.121854.r137743

© 2022 Di Lello F. This is an open access peer review report 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.

Federico Di Lello
Departamento de microbiología, inmunología, biotecnología y genética, Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Instituto de Investigaciones en Bacteriología y Virologia Molecular (IBaViM), Buenos Aires, Argentina

In the article "Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences", the author hypothesizes that among SARS-like coronaviruses, only the genomes of SARS-CoV-1 and SARS-CoV-2 contain information that points to a synthetic passage in genetically modified yeast cells (S. cerevisiae). The work is highly speculative, and the fact that other (+) RNA viruses have been able to replicate in the yeast system does not mean that SARS-CoV-2 can do so. I, therefore, recommend that the author test the possibility of replicating SARS-CoV-2 in yeast in order to reach more reliable conclusions.

Additionally, there is much evidence that supports the natural appearance of SARS-CoV-2.

First, Sarbecoviruses circulating in horseshoe bats have a complex recombination history (Boni et al. 2020; Hon et al. 2008; He et al. 2014; Hu et al. 2017; Li et al. 2020; Wang et al. 2020; Zhou et al. 2020) and it should be noted that there are at least 20 different Rhinolophus species across China, leaving many species for which the viruses are unknown. Thus, recombination between unknown coronaviruses cannot be discarded. Additionally, even though the animal that serves as the direct progenitor of SARS-CoV-2 has not been identified, the great diversity of coronaviruses observed in several species may make us think that we are facing a great lack of sampling. Moreover, it was reported that mutations, insertions, and deletions can occur near the S1–S2 junction of coronaviruses, which shows that these changes can occur by evolutionary processes in nature (Zhou et al. 2020). Lastly, Zhou et al. reported that “RmYN02” was characterized by the insertion of multiple amino acids at the junction site of the S1 and S2 subunits of the spike (S) protein. This provides strong evidence that such insertion events can occur naturally in animal betacoronaviruses (Zhou et al 2020).

The results section has a lot of information that belongs to the methods and discussion.
The author stated in the Results: “When restricted to ‘genomic DNA’ sequence records dated before 2020, an extensive blastn search among all GenBank eukaryotic genomic sequences”. - It is not clear why the Blastn search is restricted to eukaryotes and sequences recorded before 2020. The fact that a sequence has been uploaded after 2020 does not mean that it did not exist previously and should be included in the search. Additionally, virus and bacteria sequences must be included too.

There is a lot of missing information about phylogenetic studies and the possible natural origin of SARS-CoV-2 in the discussion and a better review of the literature should be done.

Minor comments:

A large part of the methodology is wrongly included in the results section, for example:

- “For comparative genomic sequence analysis, we used a standard bioinformatics approach with the BLASTlike Alignment Tool (BLAT) (BLAT, RRID:SCR_011919)19”
- “To cross-validate the detected yeast homology signals in P1- P5, we also used an independent sequence alignment method, LALIGN25, which additionally produced statistics (E-values) for pairwise alignments”.

References

Is the work clearly and accurately presented and does it cite the current literature?
Partly

**Is the study design appropriate and is the work technically sound?**
Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Not applicable

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Virology, molecular biology, evolutionary biology, bioinformatics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 26 Jun 2022

**Andreas Martin Lisewski**, Jacobs University Bremen, Bremen, Germany

**Response to Reviewer 2**

The author (Andreas Martin Lisewski) thanks Reviewer 2 (Federico di Lello) for his insightful first review (as published on 7 June 2022) of the manuscript ‘Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences’ (version 4, published 19 January 2022). The points raised by Reviewer 2 were important and have led to several clarifications in the text as well as to new data, which further support the article’s main claims, in the revised version 5 of the manuscript (corresponding copy with tracked and highlighted changes available). The following is a detailed reply to all the points raised by Reviewer 2.

1. "**The work is highly speculative, and the fact that other (+)RNA viruses have been able to replicate in the yeast system does not mean that SARS-CoV-2 can do so. I, therefore, recommend that the author test the possibility of replicating SARS-CoV-2 in yeast in order to reach more reliable conclusions.**"

To this important point, the author responds and stresses that the key genetic interaction between yeast and SARS coronavirus described in the article is not during virus (+)RNA replication but prior to it, i.e. during recombinant virus or virus parts assembly, following
transformation and recombination with plasmids/artificial chromosomes that use yeast selectable markers. The author thanks Reviewer 2 for pointing out what was an insufficient description of the main hypothesis and the proposed model. The text is now revised (version 5, Introduction, and extended Results) to better describe and underscore with new data the difference between recombinant assembly in yeast (which facilitates recombination between genomic yeast DNA and synthetic virus cDNA, leaving “traces of yeast genomic DNA” [1] within the virus genome) and replication (which is independent of a yeast genetic background and which has been realized elsewhere with a number of suitable polymerases).

It is therefore not necessary to do the suggested replication experiment, because a more comprehensive literature review together with an additional analysis of relevant sequencing data indicates that the key experiments with recombination between yeast and SARS coronavirus have already been realized through reverse genetics. Specifically, in the revised Introduction section, the author now provides a more in-depth review of yeast reverse genetics in the context of SARS coronavirus (both SARS-CoV-1 and SARS-CoV-2). Additionally, in the revised and extended Results section, the author provides additional data (new Figure 2) which differentiates significantly between a recombinant rSARS-CoV-2 clone that was assembled on a yeast artificial chromosome (YAC), and a second infectious clone that was assembled without a yeast background. Based on these new and indicative results, and in the direct context of the extended Introduction, the last two revised paragraphs of the Results section now also stronger support the proposed yeast artificial biosynthesis model.

Overall, this revision in response to Reviewer 2 led to the work being more data-based and better embedded into the existing body of relevant literature.

2. "Additionally, there is much evidence that supports the natural appearance of SARS-CoV-2. First, Sarbecoviruses circulating in horseshoe bats have a complex recombination history (Boni et al. 2020; Hon et al. 2008; He et al. 2014; Hu et al. 2017; Li et al. 2020; Wang et al. 2020; Zhou et al. 2020) and it should be noted that there are at least 20 different Rhinolophus species across China, leaving many species for which the viruses are unknown. Thus, recombination between unknown coronaviruses cannot be discarded. Additionally, even though the animal that serves as the direct progenitor of SARS-CoV-2 has not been identified, the great diversity of coronaviruses observed in several species may make us think that we are facing a great lack of sampling. Moreover, it was reported that mutations, insertions, and deletions can occur near the S1–S2 junction of coronaviruses, which shows that these changes can occur by evolutionary processes in nature (Zhou et al. 2020). Lastly, Zhou et al. reported that “RmYN02” was characterized by the insertion of multiple amino acids at the junction site of the S1 and S2 subunits of the spike (S) protein. This provides strong evidence that such insertion events can occur naturally in animal betacoronaviruses (Zhou et al. 2020)."

The author appreciates this overview of this relevant line of research. It is here adequate to reply that a recent culmination of this line of research appeared with the May/June 2022 publications of the following three studies (now referenced in the Introduction, v5):

- Wang et al. Coronaviruses in wild animals sampled in and around Wuhan in the

In fact, none of these important studies revealed viruses closely related to either SARS-CoV or SARS-CoV-2, and in particular, no identical or homologous functional S1/S2 furin cleavage site in lineage b coronaviruses was found in any of the animal samples. Thus, the same statement must also be true for the earlier studies pointed out by Reviewer 2. Given this current and comprehensive evidence against it, it may still be possible for some future studies to detect the “missing evolutionary link” (i.e., an animal host to a naturally occurring pre-pandemic lineage b coronavirus with FCS and with significantly higher genomic sequence identity to SARS-CoV-2 than RatG13). But a continued and indefinite referral to potential future discoveries, while potentially disregarding different scientific explanations, may lead to the same logical fallacies that were often brought up against alternatives to the natural origin hypothesis: speculative evidence and selective evidence. Thus, to give a current (as of June 2022) status of evidence for the natural origin of SARS-CoV-1 and SARS-CoV-2, the revised Introduction section now additionally includes this highly relevant result from the above landmark studies.

3. "The results section has a lot of information that belongs to the methods and discussion."

Yes, the parts in question have now been moved to the Methods section, which has also been revised, extended, and provided with more technical details.

4. "The author stated in the Results: “When restricted to ‘genomic DNA’ sequence records dated before 2020, an extensive blastn search among all GenBank eukaryotic genomic sequences”. - It is not clear why the Blastn search is restricted to eukaryotes and sequences recorded before 2020. The fact that a sequence has been uploaded after 2020 does not mean that it did not exist previously and should be included in the search. Additionally, virus and bacteria sequences must be included too."

As a direct reply, the author’s explanation for excluding sequence records after 2019 is now explicitly given in the revised Methods section, i.e.

to records before 2020, the likelihood of assigning such false positive sequence hits to the
pre-pandemic origins of SARS-CoV-2 would be minimized in our study.”

Two additional blastn searches were also performed with the FCS query
‘TTCTCCTCGCGGGCAA” against viridae (taxid:1023, with Sarbecovirus excluded), with no
identical hits; and against bactera (taxid: 2) with no hits to any bacterial systems previously
associated with SARS-CoV reverse genetics.

5. “There is a lot of missing information about phylogenetic studies and the possible
natural origin of SARS-CoV-2 in the discussion and a better review of the literature
should be done.”

Yes, the revised and extended Introduction and Discussion sections now include additional
studies directly relevant to the natural origin hypothesis, including phylogenetic analysis, as
well as most recent landmark studies along this line of research (see also point 2 above).

6. “Minor comments. A large part of the methodology is wrongly included in the
results section, for example:
“For comparative genomic sequence analysis, we used a standard bioinformatics
approach with the BLASTlike Alignment Tool (BLAT) (BLAT, RRID:SCR_011919)19”
Or
“To cross-validate the detected yeast homology signals in P1- P5, we also used an
independent sequence alignment method, LALIGN25, which additionally produced
statistics (E-values) for pairwise alignments”.

Yes, this has now been corrected in v5 (see also point 3 above).

References


Competing Interests: No competing interests.
Alexander Y Panchin  
Sector of molecular evolution, Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation

In the article “Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis” the author claims that “the genomes of SARS-CoV-1 and SARS-CoV-2 contain information that points to a synthetic passage in genetically modified yeast cells”. The claim seems to contradict earlier findings regarding the origin of SARS-CoV-1, which most likely evolved in palm civets and is not considered controversial (Kan et al. (2005)). It is also difficult to reconcile this claim with the fact that the closest relative of the SARS-CoV-2 virus was only recently discovered in North Laos (Temmam et al. (2021)) and that no closer relatives have been identified, specifically among those available to genetic engineers prior to 2019 (the start of the COVID-19 pandemic). Thus I decided to check the validity of the bioinformatics claims provided in the article.

According to the author, the genetic information allegedly acquired by SARS-CoV-2 from yeast includes “a 16 base sequence (TTCTCCTCGGCGGGCA) near P2 between position 23599 and 23614, which corresponded to the furin cleavage site and identically aligned with bases [810386..810401] from S. cerevisiae chromosome XIII”. The article states that “an extensive blastn search among all GenBank eukaryotic genomic sequences produced no identical sequence hits other than the Saccharomyces cerevisiae match above”.

I performed a Megablast search against all mammals in NR database (word size 16), excluding models (XM/XP). I manually checked the submission dates of the sequences to be before 2019 (as in the discussed article). I found other examples of the exact 16bp sequence (TTCTCCTCGGCGGGCA) in Microcebus murinus (GenBank: AB265808.1), Lemur catta (GenBank: AB265807.1) and Homo sapiens (GenBank: AK225546.1).

A similar search against fungi reveals other examples of organisms with the 16 bp exact match such as: Verticillium dahliae (GenBank: CP009079.1), Melanopsichium pennsylvanicum (GenBank: HG529530.1), Thielavia terrestris (GenBank: CP003010.1), Sporisorium reilianum (GenBank: FQ311441.1)

Thus, the author’s claim appears to be false.

This is not surprising, considering that 16bp sequences should occur randomly once per several billion bp. For example, using the complete SARS-CoV-2 reference genome (NC_045512.2) as a query for Megablast (word size 16) against the non-redundant (NR) database restricted to mammals I identified exact matches in Bos mutus (26 bp, GenBank: CP027085.1), a Homo sapiens BAC (26 bp, GenBank: AC117381.5), Ovis canadensis (22 and 25 bp, GenBank: CP011902.1, CP011907.1), Mouse clone (18 bp GenBank: BX005086.9), Sorex araneus (16 bp, GenBank: AC200396.3). All sequences used above were produced prior to 2019.

Furthermore, if the whole SARS-CoV-2 genome is compared using Megablast with all fungi sequences from NR, a number of fungi appear to have more similarity with SARS-CoV-2 than S. cerevisiae (have lower E-values). These include: Saccharomyces jurei (GenBank: LT986464.1), Leptosphaeria maculans (GenBank: FO906009.1), Naumovozyma dairenensis (GenBank: HE580273.2) and others. These sequences were also produced prior to 2019.
Obviously, this does not imply that SARS-CoV-2 acquired sequences from humans, mice or fungi other than *S. cerevisiae* but demonstrates that it is inappropriate to draw conclusions on sequence origins based on short matches from unrelated organisms.

Considering that SARS-CoV-2 has higher similarity with some eukaryotes other than *S. cerevisiae*, I see no prior reason to consider yeast as a source of SARS-CoV-2 genetic sequences. And given the analysis provided above, I believe that the sequence similarity between SARS-CoV-2 and *S. cerevisiae* is most likely coincidental. This is supported by the fact that the Megablast e-value of the best SARS-CoV-2 hit against *S. cerevisiae* is 5.3. Thus the author’s hypothesis and conclusions are not supported by sequence similarity analysis.

I would also add a minor comment. The abstract states: “At this junction, we detect a highly specific stretch of yeast DNA encoding for the critical furin cleavage site insert PRRA, which has not been seen in other lineage b betacoronaviruses”. SARS-CoV-2 is an RNA virus and does not contain DNA. I have read this submission. I believe I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

References

**Is the work clearly and accurately presented and does it cite the current literature?**
Yes

**Is the study design appropriate and is the work technically sound?**
No

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, molecular biology, evolutionary biology
I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 25 Feb 2022

Andreas Martin Lisewski, Jacobs University Bremen, Bremen, Germany

Response to Reviewer 1

The author (Andreas Martin Lisewski) thanks Reviewer 1 (Alexander Y. Panchin) for his detailed review (as published on 22 February 2022) of the manuscript “Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis” (version 3, published 19 January 2022). The points raised by Reviewer 1 were directly helpful and have lead to several clarifications in the text as well as to new data in the revised version 4 of the manuscript.

Before responding with a point-by-point reply below, the author puts to record that none of the data in the Results section of the manuscript version 3 have been questioned or refuted by Reviewer 1. Rather, Reviewer 1 has challenged the paper’s main hypothesis by suggesting some additional sequence similarity tests. In the author’s response, these tests have been done, and their outcomes further support the manuscript’s results and conclusions. These additional tests have now been included in the manuscript's new version 4. It is also noted that, in his entire review, Reviewer 1 has focused on only one paragraph (out of seven total) in the Results section.

Point-by-point reply to Reviewer 1:

1. “The author claims that “the genomes of SARS-CoV-1 and SARS-CoV-2 contain information that points to a synthetic passage in genetically modified yeast cells”. The claim seems to contradict earlier findings regarding the origin of SARS-CoV-1, which most likely evolved in palm civets and is not considered controversial (Kan et al. (2005) [1]).”

The molecular evolution analysis of SARS-CoV-1 presented in this 2005 landmark paper by Kan et al (J Virol. 2005;79(18):11892) does not contradict the above claim. Indeed, regarding the origin of SARS-CoV-1, Kan et al conclude that “our observations suggest that when SARS-CoV-like virus arrives at an animal market, the majority of palm civets, if not all, will become infected, and that the virus will evolve rapidly in animals to cause disease. Therefore, it is critical to identify the original animal reservoir to remove the continuing threat of SARS.“ Thus, in this cited work, civets have not been identified as the original animal reservoir, and any further identification or characterization of this hypothetical SARS-CoV-1 origin, whether natural or artificial (synthetic), has not been given either. This conclusion is also supported by other important and high-impact evidence from the same period, e.g. the study by Song et al in 2005 (PNAS. 2005; 102(7): 2430), where a rooted phylogenetic tree analysis for early SARS-CoV-1 isolates placed the earliest human virus lineage before the first civet infection, and with both viral lineages originating from an unknown reservoir in late 2002. To date, this uncertainty and controversy around the origins of SARS-CoV-1 have not vanished but have
been discussed repeatedly in the scientific literature, including the Discussion section of the present manuscript and in the references therein.

2. “It is also difficult to reconcile this claim with the fact that the closest relative of the SARS-CoV-2 virus was only recently discovered in North Laos (Temmam et al. (2021 [2])) and that no closer relatives have been identified, specifically among those available to genetic engineers prior to 2019 (the start of the COVID-19 pandemic).”

As a direct response to this statement, the author points out that – according to the complete genomic sequence analysis given in this recent paper by Temmam et al (doi:10.21203/rs.3.rs-871965/v1) – the North Laos isolates in question (BANAL-20-52, and also BANAL-20-103, BANAL-20-236) are not the closest relatives to SARS-CoV-2. This is demonstrated with the phylogenetic data in their Figure 1B: the maximum likelihood tree robustly places RaTG13, and not BANAL-20-52/103/236, as the nearest neighbor to SARS-CoV-2. This evidence is entirely in line with the arguments made in the manuscript regarding RaTG13, as well as those regarding BANAL-20-52, BANAL-20-103, and BANAL-20-236 (see Results section, 4th paragraph.)

3. “According to the author, the genetic information allegedly acquired by SARS-CoV-2 from yeast includes “a 16 base sequence (TTCTCCTCGGCGGGCA) near P2 between position 23599 and 23614, which corresponded to the furin cleavage site and identically aligned with bases [810386..810401] from S. cerevisiae chromosome XIII”. The article states that “an extensive blastn search among all GenBank eukaryotic genomic sequences produced no identical sequence hits other than the Saccharomyces cerevisiae match above”.

Unfortunately, this paragraph of the Results section in version 3 contains a typographical error when referring to the input 17 bp sequence of the furin cleavage site (FCS), which was identified there as the longest such common sequence between SARS-CoV-2 clade and S. cerevisiae. This error apparently lead to some misunderstanding of the results for which the author apologizes. Specifically, the blastn sequence search as described in this paragraph was based on the extended 17 bp input sequence TTCTCCTCGGCGGGCA and not on the shorter 16 bp TTCTCCTCGGCGGGCA (the typo in line 7 of the paragraph). This fact is documented in the manuscript’s version 3 with the blastn alignment output file Data_File_S1 (Extended Data). As such, the results and conclusions in this paragraph remain valid. The typographical error in line 7 has now been corrected in version 4.

4. “I performed a Megablast search against all mammals in NR database (word size 16), excluding models (XM/XP). I manually checked the submission dates of the sequences to be before 2019 (as in the discussed article). I found other examples of the exact 16bp sequence (TTCTCCTCGGCGGGCA) in Microcebus murinus (GenBank: AB265808.1), Lemur catta (GenBank: AB265807.1) and Homo sapiens (GenBank: AK225546.1).”

As suggested here by Reviewer 1, the author has repeated this sequence search with the shorter and less specific 16bp sequence TTCTCCTCGGCGGGCA. Due to the reduced specificity of this input sequence, blastn found 473 identical hits across 127 eukaryotic species (of these 473, 189 were on one species, S. cerevisiae), including the 3 hits above as identified by Reviewer 1 (see new Data_File_S2.txt, Extended Data). To sort these hits
according to their similarity to SARS-CoV-2, the author then searched the SARS-CoV-2 full genomic sequence against those 127 species (after removing the poly-A end from the input SARS-CoV-2 sequence.) The corresponding blastn output placed *S. cerevisiae* at second rank (E-value of 0.12), and as the top hit in ‘genomic DNA’ sequence records prior to 2020 (see new Data_File_S3.txt, Extended Data).

5. “For example, using the complete SARS-CoV-2 reference genome (NC_045512.2) as a query for Megablast (word size 16) against the non-redundant (NR) database restricted to mammals I identified exact matches in *Bos mutus* (26 bp. GenBank: CP027085.1), a *Homo sapiens* BAC (26 bp, GenBank: AC117381.5), *Ovis canadensis* (22 and 25 bp, GenBank: CP011902.1, CP011907.1), Mouse clone (18 bp GenBank: BX005086.9), *Sorex araneus* (16 bp, GenBank: AC200396.3). All sequences used above were produced prior to 2019.”

To address this observation, the above specificity search (under point 4) was repeated but this time without the use of TTCTCCTCGCGGGCA, i.e. just by searching the SARS-CoV-2 full genomic sequence against all eukaryotic genomic sequences in the ‘Nucleotide collection (nt)’. In this additional sequence similarity test, blastn again produced *S. cerevisiae* records among the top hits, i.e. at second rank with $E = 0.69$, after the whole output was filtered for ‘genomic DNA’ sequence records submitted prior to 2020 (see new Data_File_S4.txt and filtered Data_File_S5.txt, Extended Data). At first rank ($E = 0.20$) in this filtered output was another budding yeast, *S. yurei*, which is 98% identical to the above $E = 0.69$ *S. cerevisiae* hit from the exact same yeast lineage.

Of note, the correct algorithm for searching such highly dissimilar sequences between species is blastn, and not megablast; the latter is designed primarily for rapid and less sensitive intra-species comparison of highly similar sequences (>95%, see NCBI’s main blastn page at https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome).

Thus, this point 5 and point 4 above further support the manuscript’s original claim that (a) *S. cerevisiae* is a potential synthetic host and potential recombination donor of the genomic region that covers the critical furin cleavage site, and (b) that *S. cerevisiae* and SARS-CoV-2 display a marked level of sequence similarity when compared extensively with other eukaryotic species. These new results are now part of the Results section in the manuscript’s new version 4.

6. “Obviously, this does not imply that SARS-CoV-2 acquired sequences from humans, mice or fungi other than *S. cerevisiae* but demonstrates that it is inappropriate to draw conclusions on sequence origins based on short matches from unrelated organisms.”

As stated in the introduction to this reply, Reviewer 1 has concentrated his review entirely on a single paragraph from the Results section that has 7 paragraphs in version 3. By doing so, the reviewer narrowed down the scientific scope of the manuscript to simple blast sequence similarity searches between random genomic sequences („Thus, the author’s claim appears to be false. This is not surprising, considering that 16bp sequences should occur randomly once per several billion bp.“) However, the paper’s main hypothesis (synthetic origin), its biological model (passage in artificial yeast) as well as the results provided in
support of both hypothesis and model are more than mere similarities between random genomic sequences. For example, the genomic sequences of the FCS and of the RdRp, identified and analyzed in the Results section, are not random or coincidental, but have a specific and critical meaning in the paper’s biological context. It is also noted, on a technical bioinformatics level, that Reviewer 1 has not discussed any of paper’s results produced with bioinformatics methods other than blastn (e.g., BLAT, lalign).

But even when following this narrow bioinformatics approach in this review (sequence similarity testing using blast), the tests suggested by Reviewer 1 – when done systematically and comprehensively – do actually support the paper’s main results and conclusions (see points 4 and 5 above). Thus the reviewer’s main conclusions can be refuted using his own approach.

7. “Considering that SARS-CoV-2 has higher similarity with some eukaryotes other than S. cerevisiae, I see no prior reason to consider yeast as a source of SARS-CoV-2 genetic sequences. And given the analysis provided above, I believe that the sequence similarity between SARS-CoV-2 and S. cerevisiae is most likely coincidental. This is supported by the fact that the Megablast e-value of the best SARS-CoV-2 hit against S. cerevisiae is 5.3. Thus the author’s hypothesis and conclusions are not supported by sequence similarity analysis.”

Please see the replies at points 6, 5, and 4 above.

8. “At this junction, we detect a highly specific stretch of yeast DNA encoding for the critical furin cleavage site insert PRRA, which has not been seen in other lineage b betacoronaviruses’. SARS-CoV-2 is an RNA virus and does not contain DNA.”

Reviewer 1 is of course correct, and the author has rephrased the above sentence. However, even the original sentence does make sense in the given context (passage model with recombination between SARS-CoV-2 synthetic constructs and yeast DNA). In a similar way, for example, one speaks about ‘endogenous retroviruses’ in the human genome even though retroviruses contain no DNA.

New Extended Data files produced for version 4

Extended Data File S2: Lisewski, Andreas Martin, 2021, "Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis", https://doi.org/10.7910/DVN/BK8AL6, Harvard Dataverse, V3; Data_File_S2.txt [fileName]

Extended Data File S3: Lisewski, Andreas Martin, 2021, "Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis", https://doi.org/10.7910/DVN/BK8AL6, Harvard Dataverse, V3; Data_File_S3.txt [fileName]

Extended Data File S4: Lisewski, Andreas Martin, 2021, "Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis", https://doi.org/10.7910/DVN/BK8AL6, Harvard Dataverse, V3; Data_File_S4.txt [fileName]
Extended Data File S5: Lisewski, Andreas Martin, 2021, "Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis", https://doi.org/10.7910/DVN/BK8AL6, Harvard Dataverse, V3; Data_File_S5.txt [fileName]

Andreas Martin Lisewski

Bremen, 25 February 2022

**Competing Interests:** No competing interests disclosed.

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