Check for updates

BRIEF REPORT

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

[version 3; peer review: 1 not approved]

Andreas Martin Lisewski

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

V3 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 considered.

Results: Here we show that the genomes from SARS-CoV-2 and from SARS-CoV-1 are differentially enriched with 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 critical domain junction. 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. As yeast is not a natural host for this virus family, we propose a passage 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 sequence homology data across yeast chromosomes congruent with chromosomes harboring specific auxotrophic markers further support this passage model.

Conclusions: These results provide evidence that among SARS-like

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**, Russian Academy of Sciences, Moscow, Russian Federation
2. **Federico Di Lello** , Russian Academy of Sciences, Moscow, Russian Federation

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

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. 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 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: © 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. **Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis [version 3; peer review: 1 not approved]** F1000Research 2022, **10**:912 <https://doi.org/10.12688/f1000research.72956.3>

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

UPDATE Updates from Version 2

Furin cleavage site data and discussion included. Identical sequence stretch on *S. cerevisiae* chromosome XIII identified and suggested as potential recombination donor for the furin site insert in the SARS-CoV-2 spike glycoprotein.

Any further responses from the reviewers can be found at the end of the article

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 (PRRAR) 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 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 discussed¹; 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 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 budding 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 betacoronavirus 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` 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. Nucleotide sequence database searches were performed with the NCBI `blastn` webserver (RRID:SCR_001598) against the entire "Nucleotide collection (nr/nt)" restricted to eukaryotes (taxid:2759); herein, "Models (XM/XP)", partial, and predicted sequences were excluded.

Results

To interrogate the possibility that a similar passage through yeast cells took place within the family of SARS coronaviruses, we initially 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..7192]_{max}, and a second peak $P2$ over ~18,000 bases downstream with a score of 36 in the region [25196..25212]_{max} (see, Figure 1). To put these data into an established gene-function context these two maxima, with half-maximum widths $w_{1/2} = 215$ and $w_{1/2} = 219$, 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..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 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..25187]). The S_S1/S2 domain includes the characteristic furin cleavage site at the S1/S2 junction²², which has previously been described as unique to SARS-CoV-2 among lineage b betacoronaviruses⁴. 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..9733]_{max}, and with $w_{1/2} = 230$. 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 (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. 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 (TTCTCCTCGGCGGGCAA), which are identically found in known S^{R685S} variants of SARS-CoV-2 that emerged after serial passage in cell culture (e.g., see GenBank entry MZ995185.1), and—at codon level—are also compatible with the entire ancestral SPRRAR motif. To test the specificity of TCTCCTCGGCGGGCAA across potential 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 GenBank entries dated before August 2020, an extensive blastn search among all GenBank eukaryotic genomic sequences produced no identical sequence hits other than the *Saccharomyces cerevisiae* match above (see, Extended Data File S1). These data specifically identified the yeast *S. cerevisiae* as a potential genomic recombination donor of the critical furin cleavage site 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 $w_{1/2} = 222$; and a broader second peak, $P5$, with $pS = 41$ at position [22286..22391]_{max} and $w_{1/2} = 477$. $P4$ 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, *Paradoxurus hermaphroditus*, *Paguma larvata*, *Aselliscus stoliczkanus*, and *Rhinolophus sinicus*), which together with SARS-CoV-2 occupy the same phylogenetic branch⁹. These data collectively produced a differential yeast homology signature in SARS-CoV-1, SARS-CoV-2 and RaTG13 genomes after calculating standardized z -scores (Figure 2) from the entire BLAT profiles to all 13 of the above sequences (Tables S2–S14). This analysis was also extended by including the three recently identified bat SARS-like coronavirus genomes from the same clade as RaTG13 ($z = 4.72$), i.e., BANAL-20-52 ($z = 0.36$),

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.

BANAL-20-103 ($z = -1.56$) and BANAL-20-236 ($z = 0.20$), which all produced markedly smaller z-scores than SARS-CoV-1 ($z = 9.44$) and SARS-CoV-2 ($z = 5.47$). 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. 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 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

Figure 3. (A) First stage of passage model in the artificial host *S. cerevisiae* of a plasmid encoded, non-replicable viral RNA (nrRNA1) originating from a SARS-CoV related virus. Primary integration of non-homologous nrRNA1 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 nrRNA1 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_n, 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 nrRNA1, 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.

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 sulphydrylase*, 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 *nrvRNA3*¹⁵. 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), 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

Our results reveal a highly differential homology signal in 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, an important point 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 have interfered with the yeast homology 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.*¹ 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. Collectively, these results offer an important new lead for the further understanding of SARS-CoV-2 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 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](#))
14. Bat coronavirus isolate BANAL-20-52/Laos/2020 (GenBank: [MZ937000.1](#))
15. Bat coronavirus isolate BANAL-20-103/Laos/2020 (GenBank: [MZ937001.1](#))
16. Bat coronavirus isolate BANAL-20-236/Laos/2020 (GenBank: [MZ937003.1](#))

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/BK8AL6>³³.

This project contains the following extended data files:

- Data_File_S1. Y8JHH6JM013-Alignment_SHORT.txt: blastn output text file for the input nucleotide sequence TCTCCTCGGCGGGCAA. *WTIH* denotes Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom.
- 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.

Data are available under the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

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; **165**(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, Serviène 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](#)
- 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](#)
- 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](#)
- Kent WJ: **BLAT—The BLAST-Like Alignment Tool.** *Genome Res.* 2002; **12**(4): 656–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988.
[Reference Source](#)
- 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](#)
- 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](#)
- Papa G, Mallery DL, Albecka 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](#)
- 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](#)
- Huang X, Miller W: **A time-efficient, linear-space local similarity algorithm.** *Adv Appl Math.* 1991; **12**(3): 337–57.
[Publisher Full Text](#)
- 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](#)
- 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](#)
- Commonly used auxotrophic markers. SGD-Wiki. [cited 2021 Jun 3].
[Reference Source](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- Lisewski AM: **Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis.** Harvard Dataverse, V1, UNF: 6:BC1twHAK9JEqcRfghK4Dg== [fileUNF]. 2021.
<http://www.doi.org/10.7910/DVN/BK8AL6>

Open Peer Review

Current Peer Review Status:

Version 3

Reviewer Report 22 February 2022

<https://doi.org/10.5256/f1000research.120364.r121768>

© 2022 Panchin A. 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.

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.* (2005¹)). 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.* (2021²)) 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

1. Kan B, Wang M, Jing H, Xu H, et al.: Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J Virol.* 2005; **79** (18): 11892-900 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Temmam S, Vongphayloth K, Salazar E, Munier S, et al.: Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. *Research Square.* 2021. [Publisher Full Text](#)

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

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 led 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 TTCTCCTCGGCGGGCAA and not on the

shorter 16 bp TCTCCTCGGCGGGCAA (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 TTCTCCTCGGCGGGCA, 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

F1000Research