Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences

[version 5; peer review: 2 ~~not approved~~]

Andreas Martin Lisewski[a1]

Conceptualization

Data Curation

Formal Analysis

Investigation

Methodology

Project Administration

Supervision

Validation

Visualization

Writing – Original Draft Preparation

Writing – Review & Editing

URI: https://orcid.org/0000-0003-3403-6880

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

Author notes:

Correspondence to: [a] a.lisewski@jacobs-university.de

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 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 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 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.

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 [1]. According to this view, COVID-19 is a zoonosis that probably originated from a species of closely related bat coronaviruses [2]. Prior to a hypothetical spillover event, a recent ancestor to SARS-CoV-2 likely evolved inside bat host cells for many decades [3]. 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 (FCS) at the S1/S2 junction of the SARS-COV-2 spike (S) glycoprotein [4]; second, the discrepancy between an exponentially suppressed tropism of SARS-CoV-2 in *Rhinolophus sinicus* bat cells [5] and the high susceptibility of SARS-CoV-2 toward cell entry via *Rhinolophus sinicus* angiotensin-converting enzyme 2, its primary entry receptor [6]; 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 closest evolutionary relationship to the SARS-CoV-2 genome [NEW REF https://pubmed.ncbi.nlm.nih.gov/35172323/], which is located on the same phylogenetic branch as *Rhinolophus sinicus* bat coronavirus [9]. 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 [10,11]. 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 [NEW REF https://doi.org/10.1093/ve/veac046],[NEW REF https://pubmed.ncbi.nlm.nih.gov/35298912/], or outside [NEW REF https://pubmed.ncbi.nlm.nih.gov/35637279/].

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 [1]. 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 [12]. 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 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* [15], [NEW REF 2008 US patent

US9682136B2: https://patentimages.storage.googleapis.com/56/35/7b/353e8edb256ec6/US9682136.pdf]. Accordingly, to transform budding yeast into a artificial host for viral synthesis and replication, the general scheme has been to co-express a 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]. 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 [NEW REF https://pubmed.ncbi.nlm.nih.gov/12368349/]. Yeast selectable (auxotrophic) markers, and specifically URA3, have been since described and used experimentally to direct "humanized yeast" 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 [NEW REF, see columns 14-15 in the 2008 US patent US9682136B2: https://patentimages.storage.googleapis.com/56/35/7b/353e8edb256ec6/US9682136.pdf]. 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 chromosomes [17]. 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*. [NEW REF https://pubmed.ncbi.nlm.nih.gov/32365353/] 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 *imidazoleglycerol-phosphate dehydratase*) as YSM. YAC assembly therefore facilitates recombination with endogenous host chromosomes resulting in viral RNA or infectious clones with "traces of yeast genomic DNA" [NEW REF https://pubmed.ncbi.nlm.nih.gov/32365353/]. Our hypothesis is that such artificial passage through 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*. [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). This set was extended by 5 additional genomic sequnces, BANAL 20-52/20-103/20-236, icSARS-CoV-2 BAC, and rSARS-CoV-2 YAC (see, *Repository-hosted data*, for the 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 produced a high number (between 1689 and 5083) of tiles, i.e. perfectly aligned short DNA sequences of length 11. 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, $p_1 < p_2$, was calculated with conditional probabilities p_1/p_2. 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 Wuhan-Hu-1 reference genome) mapped onto the synthetic genome construct rSARS-CoV-2 with sequence deposited at Genbank MT108784, see Extended Data Table 4 in Thao *et al*. [NEW REF https://pubmed.ncbi.nlm.nih.gov/32365353/]. 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

Andreas 6/19/2022 11:43 AM
Moved (insertion) [1]

Andreas 6/20/2022 10:56 PM
Moved (insertion) [3]

(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 eukaryotic (taxid:2759) 'genomic DNA' sequence records deposited before the year 2020. The reason behind leaving out sequencing data generated after 2019 is growing evidence, since the beginning of the COVID19 pandemic, of exogenous genomic integration in cultured cells and in the infected host [NEW REF https://www.ncbi.nlm.nih.gov/pubmed/33958444] [NEW REF https://pubmed.ncbi.nlm.nih.gov/34702741/], as well as widespread contamination of laboratory environments with SARS-CoV-2 cDNA [NEW REF https://pubmed.ncbi.nlm.nih.gov/34523989/]; this dissemination of SARS-CoV-2 cDNA into the host's natural environment may cause sequencing-based virus testing anomalies [NEW REF https://pubmed.ncbi.nlm.nih.gov/34523989/], and has already resulted in chimeric virus-host sequences in reference databases unseen before 2020 (e.g., https://www.ncbi.nlm.nih.gov/bioproject/PRJNA720932). Therefore, by restricting searches 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. Also, 'Models (XM/XP)', partial, and predicted sequences were excluded. blastn algorithm parameters were set at standard values except for E-value threshold (100 instead of 0.05), and gap cost (6 instead of 5).

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 genomes [18] 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]. The rationale was that BLAT, a more accurate genome sequence alignment tool than other conventional approaches [19], 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 b betacoronavirus 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).

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 [

Andreas 6/19/2022 11:43 AM
Moved up [1]: 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.

Andreas 6/20/2022 10:56 PM
Moved up [3]: To remove its shortest-scale fluctuations, the profile was smoothed by a centered sliding window filter with window size of 200 nucleotides (*nt*).

Andreas 6/19/2022 1:09 PM
Moved down [2]: The output of eight genomic profiles (Figure 1 and Figure S2) were ordered by decreasing sequence identity to SARS-CoV-2.

Andreas 6/19/2022 1:09 PM
Moved (insertion) [2]

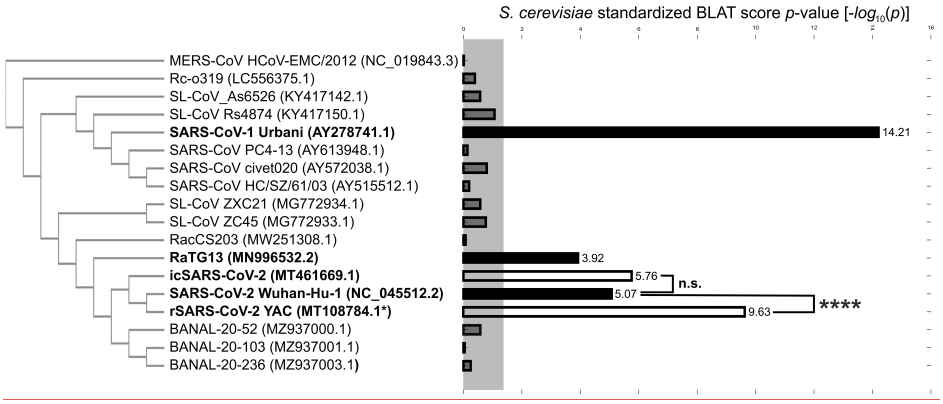
YP_009724390.1 (543..1,208)]; RdRpN, N-terminal region of the RNA dependent RNA polymerase [AAP13442.1 (4,383..4,735)].

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 database [20]. 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 [21]. $P2$ 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 junction [22], which has previously been described as unique to SARS-CoV-2 among lineage b betacoronaviruses [4]. 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 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 SPRRAR motif. As such, TTCTCCTCGGCGGGCAA 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 TTCTCCTCGGCGGGCAA 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 '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 *Saccharomyces cerevisiae* match above (see, Extended Data File S1). A similar result was obtained when potential host specificity was tested with the shorter TTCTCCTCGGCGGGCA 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 $w_{1/2} = 222$; 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 branch [9]. 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 ($z = 0.36$), 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 *P*1- *P*5, we also used an independent sequence alignment method, LALIGN [25], which additionally produced statistics (*E*-values) for pairwise alignments. While the peaks *P*1 and *P*2, as well as *P*4 and *P*5, could be positively validated, the *P*3 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.



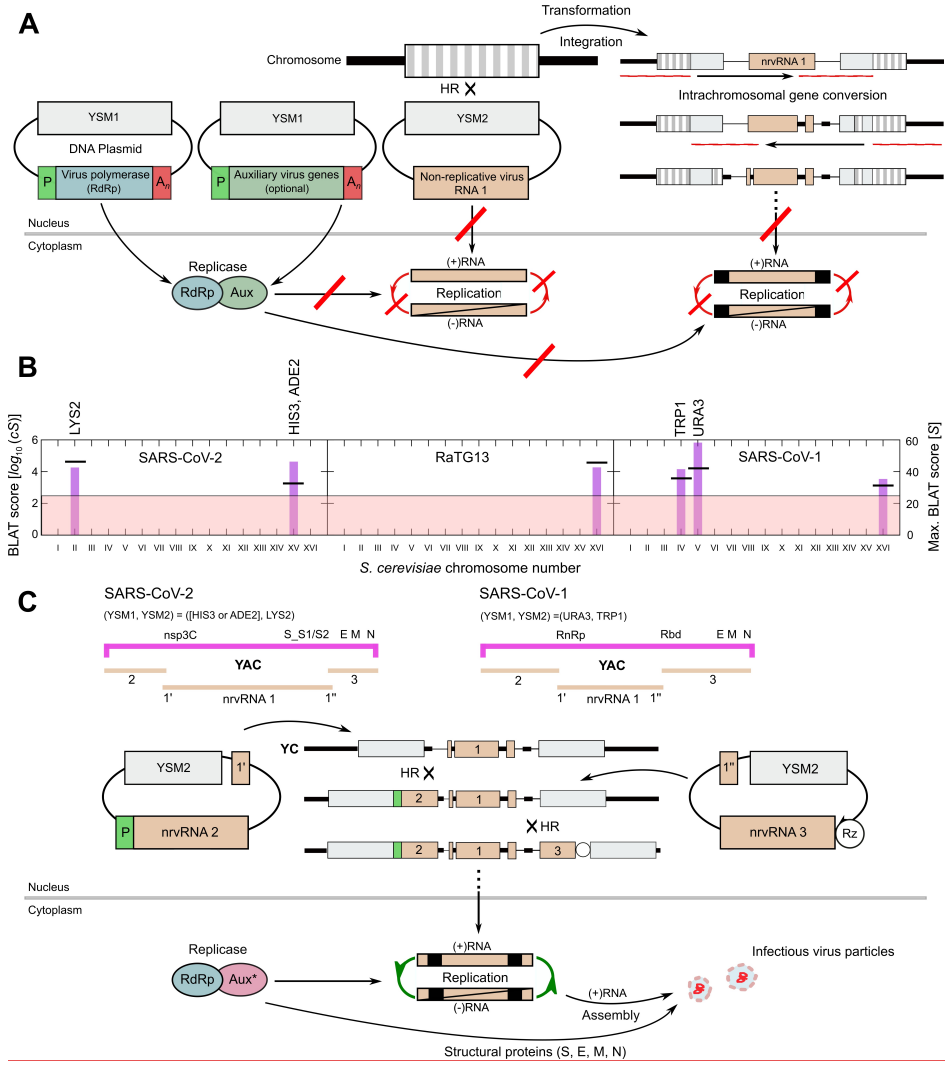
[V5 REVISED Figure 2.] Yeast (*S. cerevisiae*) standardized BLAT *p*-values measuring the relative homology signal from all alignment scores in 18 representative SARS-related coronaviruses.

Individual *p*-values were calculated from sampled means and standard deviations in BLAT outputs (see, Table S2–S19 and Methods). Grey shaded box depicts 0.05 significance level. Pairwise statistical significance test by conditional *p*-values (see, Methods); n.s., not significant. The negative log *p*-value for rSARS-CoV-2 YAC (MT108784.1*) was the average over 12 such values from sequenced YAC clones, see Thao *et al*. 2020 and Methods. Evolutionary guide tree (cladogram) generated by sequence identities between full genomic sequences (see, Table S20).

As a further validation of our method, we turned to two derived genomic sequences of SARS-CoV-2: the recombinant rSARS-CoV-2, assembled through reverse genetics into a YAC https://pubmed.ncbi.nlm.nih.gov/32365353/], and the infectious clone icSARS-CoV-2, assembled without yeast through a cell-free *in vitro* ligation method [NEW REF https://pubmed.ncbi.nlm.nih.gov/32526206/]. Even though their genomic sequences were identical to SARS-CoV-2 Wuhan-Hu-1 at >99.9% level (see, Table S20), our data (Fig. 2) differentiated between rSARS-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 and passage model (Figure 3A): Its starting point is a doubly auxotrophic, synthetic yeast cell line with stable, heterologous expression of a 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 encoding for 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 integrated [17] 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 conversion [26], i.e. through a non-crossover homologous recombination, and with the endogenous site as the homologous repair donor (Figure 3A).



[V5 REVISED Figure 3.] Yeast artificial synthesis model for SARS coronavirus 1 and 2.

(**A**) First stage of passage model in the artificial host *S. cerevisiae* of a plasmid encoded, non-replicable viral RNA (nrvRNA1) originating from a SARS-CoV related virus. Primary integration of non-homologous nrvRNA1 sequence occurs through homologous recombination (HR) between the auxotrophic plasmid yeast selectable marker YSM1 (grey box) and its chromosomal homolog (striped grey box); higher-order homologous recombination follows on the flanking regions of nrvRNA1 through intra-chromosomal gene-conversion; co-expression of viral replicase complex (RdRp) and other auxiliary viral genes (Aux). Scheme in parts adapted from Compton *et al*. (1982), and from Alves-Rodrigues *et al*. (2006). P, yeast promoter; A $_n$, 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); *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 nrvRNA1, 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.

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 [17] 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 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 viral genomes produced no signal at the chosen cutoff (see, Figure 3B, also for similar data without a cutoff). To further connect these data to our passage model, we attempted to match the seven most commonly used auxotrophic yeast selectable markers [27, 28] according to their chromosomal origin: ADE2 (adenine requiring *phosphoribosylaminoimidazole carboxylase*, on chromosome XV), HIS3 (histidine requiring *imidazoleglycerol-phosphate dehydratase*, chr. XV), LEU2 (leucine requiring *Beta-isopropylmalate dehydrogenase*, chr. III), LYS2 (lysine requiring *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 homology 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 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 [15]. 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* [29].

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) coronaviruses by indicating a common yeast artificial origins 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 [7], 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 [11]—along with SARS-CoV-2 and SARS-CoV-1 [30], as our results also suggest. As a controversial candidate for a natural SARS coronavirus host, palm civets had in fact never been identified as the original animal reservoir of SARS coronavirus, and a conclusive zoonotic host identification or characterization from a natural origin has since not been given either. For eample, 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." [NEW REF https://pubmed.ncbi.nlm.nih.gov/16140765/] 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 [NEW REF https://pubmed.ncbi.nlm.nih.gov/15695582/], 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 [NEW REF https://doi.org/10.1093/ve/veac046][NEW REF https://pubmed.ncbi.nlm.nih.gov/35298912/], [NEW REF https://pubmed.ncbi.nlm.nih.gov/35637279/].

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 a artificial assembly and passage in yeast together with some genetic modifications [31] 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 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 genome [32] 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*. [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. 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 relevant 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 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, 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. Bacterial artificial chromosome (BAC) infectious clone, icSARS-CoV-2 BAC, 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: 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 : blastn output text file for the input 17 nucleotide sequence TTCTCCTCGGCGGGCAA. *WSIH* denotes *Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom*.
- Data_File_S2 : blastn output text file for the input 16 nucleotide sequence TTCTCCTCGGCGGGCA 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 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.txt: BANAL-20-52/Laos/2020/ S. cerevisiae (sacCer3) BLAT results.
- Table_S16.txt: BANAL-20-103/Laos/2020/ S. cerevisiae (sacCer3) BLAT results.
- Table_S17.txt: BANAL-20-236/Laos/2020/ S. cerevisiae (sacCer3) BLAT results.
- Table_S18.txt: SARS coronavirus icSARS-CoV-2 BAC/ S. cerevisiae (sacCer3) BLAT results.
- Table_S19.txt: SARS coronavirus rSARS-CoV-2 YAC/ S. cerevisiae (sacCer3) BLAT results.
- Table_S20.txt: Percent identity matrix (generated with Clustal 2.1).
- Table_S21.xlsx: Peak *P*1-*P*5 yeast homology signals detected by BLAT, and cross-validated by the LALIGN sequence alignment method.

References

1 Andersen KG, Rambaut A, Lipkin WI, et al.: The proximal origin of SARS-CoV-2. *Nat Med*. 2020;26(4):450–2. 32284615 10.1038/s41591-020-0820-9 7095063

2 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. 33711012 10.1371/journal.pbio.3001115 7990310

3 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. 32724171 10.1038/s41564-020-0771-4

4 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. 32737584 10.1007/s00705-020-04750-z 7394270

5 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. 32730733 10.3201/eid2612.202308 7706959

6 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. 33610551 10.1016/j.jbc.2021.100435 7892319

7 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. 32015507 10.1038/s41586-020-2012-7 7095418

8 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. 33199918 10.1038/s41586-020-2951-z

9 Li C, Yang Y, Ren L: Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. *Infect Genet Evol*. 2020;82:104285. 32169673 10.1016/j.meegid.2020.104285 7270525

10 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. 33558807 10.1007/s10311-020-01151-1 7859469

11 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. 34046923 10.1002/bies.202100015 8209872

12 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. 19321428 10.1073/pnas.0809524106 2660061

13 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. 14671320 10.1073/pnas.2536857100 307642

14 Ahlquist P, Noueiry AO, Lee WM, et al.: Host Factors in Positive-Strand RNA Virus Genome Replication. *J Virol*. 2003;77(15):8181–6. 12857886 10.1128/jvi.77.15.8181-8186.2003 165243

15 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. 16698107 10.1016/j.virusres.2005.11.018 7114155

16 Pogany J, Panavas T, Serviene E, et al.: A high-throughput approach for studying virus replication in yeast. *Curr Protoc Microbiol*. 2010;Chapter 16: Unit16J.1. 21053256 10.1002/9780471729259.mc16j01s19

17 Compton JL, Zamir A, Szalay AA: Insertion of nonhomologous DNA into the yeast genome mediated by homologous recombination with a cotransforming plasmid. *Mol Gen Genet*. 1982;188(1):44–50. 6294480 10.1007/BF00332994

18 Zhou H, Ji J, Chen X, et al.: Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*. 2021;184(17):4380–4391.e14. 34147139 10.1016/j.cell.2021.06.008 8188299

19 Kent WJ: BLAT--The BLAST-Like Alignment Tool. *Genome Res*. 2002;12(4):656–64. 11932250 10.1101/gr.229202 187518

20 National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.1988. Reference Source

21 Sakai Y, Kawachi K, Terada Y, et al.: Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication. *Virology*. 2017;510:165–74. 28738245 10.1016/j.virol.2017.07.019 7111695

22 Coutard B, Valle C, de Lamballerie X, et al.: The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res*. 2020;176:104742. 32057769 10.1016/j.antiviral.2020.104742 7114094

23 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. 33493182 10.1371/journal.ppat.1009246 7861537

24 Theuerkauf SA, Michels A, Riechert V, et al.: Quantitative assays reveal cell fusion at minimal levels of SARS-CoV-2 spike protein and fusion from without. *iScience*. 2021;24(3):102170. 33585805 10.1016/j.isci.2021.102170 7871100

25 Huang X, Miller W: A time-efficient, linear-space local similarity algorithm. *Adv Appl Math*. 1991;12(3):337–57. 10.1016/0196-8858(91)90017-D

26 Agmon N, Pur S, Liefshitz B, et al.: Analysis of repair mechanism choice during homologous recombination. *Nucleic Acids Res*. 2009;37(15):5081–92. 19553188 10.1093/nar/gkp495 2731894

27 Pronk JT: Auxotrophic yeast strains in fundamental and applied research. *Appl Environ Microbiol*. 2002;68(5):2095–100. 11976076 10.1128/AEM.68.5.2095-2100.2002 127579

28 Commonly used auxotrophic markers. SGD-Wiki. [cited 2021 Jun 3]. Reference Source

29 Nooraei S, Bahrulolum H, Hoseini ZS, et al.: Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers. *J Nanobiotechnology*. 2021;19(1):59. 33632278 10.1186/s12951-021-00806-7 7905985

30 Xu D, Sun H, Su H, et al.: SARS coronavirus without reservoir originated from an unnatural evolution, experienced the reverse evolution, and finally disappeared in the world. *Chin Med J (Engl)*. 2014;127(13):2537–42. 24985597

31 Jimenez-Guardeño JM, Regla-Nava JA, Nieto-Torres JL, et al.: Identification of the Mechanisms Causing Reversion to Virulence in an Attenuated SARS-CoV for the Design of a Genetically Stable Vaccine. *PLoS Pathog*. 2015;11(10):e1005215. 26513244 10.1371/journal.ppat.1005215 4626112

32 Pereson MJ, Mojsiejczuk L, Martínez AP, et al.: Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence. *J Med Virol*. 2021;93(3):1722–31. 32966646 10.1002/jmv.26545 7537150

33 Lisewski AM: Differential enrichment of yeast DNA in SARS-CoV-2 and related genomes supports synthetic origin hypothesis. Harvard Dataverse, V1, UNF: 6:BC1twHAk9jEwqcRfghK4Dg== [fileUNF].2021. http://www.doi.org/10.7910/DVN/BK8AL6