Virus-host co-evolution under a modified nuclear genetic code

Among eukaryotes with modified nuclear genetic codes, viruses are unknown. However, here we provide evidence of an RNA virus that infects a fungal host (*Scheffersomyces segobiensis*) with a derived nuclear genetic code where CUG codes for serine. The genomic architecture and phylogeny are consistent with infection by a double-stranded RNA virus of the genus *Totivirus*. We provide evidence of past or present infection with totiviruses in five species of yeasts with modified genetic codes. All but one of the CUG codons in the viral genome have been eliminated, suggesting that avoidance of the modified codon was important to viral adaptation. Our mass spectroscopy analysis indicates that a congener of the host species has co-opted and expresses a capsid gene from totiviruses as a cellular protein. Viral avoidance of the host's modified codon and host co-option of a protein from totiviruses suggest that RNA viruses co-evolved with yeasts that underwent a major evolutionary transition from the standard genetic code.

Derek J. Taylor[1], Matthew J. Ballinger, Shaun M. Bowman, Jeremy A. Bruenn

Department of Biological Sciences, The State University of New York at Buffalo, Buffalo,
NY 14260, USA.

[1]Correspondence and requests for materials should be addressed to D. J. T. (e-mail: *djtaylor@buffalo.edu*; phone: 716-645-2880).

Introduction

Crick (1968) declared the universal genetic code to be nearly immutable because change would cause ‘mistakes’ in so many of the proteins of a cellular life form. However, Crick also implied that viruses are a possible exception to this evolutionary ‘freezing’ process because viruses have but a few protein coding targets. It is now well-established that modified genetic codes have evolved from the universal genetic code at least 34 times (do Céu Santos & Santos 2012). Yet, the consequences of these shifts for virus-host co-evolution remain poorly understood (Holmes 2009; Shackelton & Holmes 2008). Indeed, some authors have proposed that genetic code variants evolved as an antiviral defense (Holmes 2009; Shackelton & Holmes 2008; Taylor & Bruenn 2009). As viruses must use the protein translation machinery of the host, differences in genetic codes could preclude viral transfers among hosts. An evolutionary leap as great as a genetic code change could allow hosts to escape the co-evolutionary struggle with viruses. In agreement with the antiviral hypothesis, viruses are unknown from organisms with modified nuclear genetic codes. Viruses are known to infect the mitochondrial genomes of fungi with alternative mitochondrial genetic codes, but the tremendous divergence of these viruses from known viruses that use the ancestral genetic code obscures their origins (Shackelton & Holmes 2008). Modes of viral adaptation to hosts with non-standard genetic codes remain mysterious.

The CUG codon of the “CTG yeasts” (a diverse monophyletic group that contains human pathogens such as *Candida albicans* and wood-digesting species such as *Scheffersomyces stipitis*) has been reassigned from leucine to serine such that the serine tRNA possesses both derived serine and ancestral leucine sequence motifs (Butler et al. 2009; Santos et al. 2011). This shift in the genetic code, which replaces a hydrophobic residue with a polar residue, interferes with protein folding and can affect surface residue function (Feketová et al. 2010; Rocha et al. 2011). As a consequence, CTG yeasts lack the CUG codon from (>90%) functionally relevant positions in proteins (Rocha et al. 2011) and rarely receive genes via horizontal transfer compared to fungi with the standard code (Fitzpatrick 2011; Richards et al. 2011). Viruses face the same functional barrier posed by this modified code. However, the recent finding of "fossils" of totiviruses in the nuclear genomes of CTG yeast (Taylor & Bruenn 2009) could indicate that exogenous RNA viruses have adapted to a modified nuclear genetic code but remain undetected.

Totiviruses have a double stranded RNA genome (from 4.5-8kb in size in fungi) and are characterized by an overlapping open reading frame between the capsid gene and the RdRp gene with a programmed ribosomal frameshift. The family (Totiviridae) that contains the totiviruses is ancient with a broad eukaryotic host range and extensive co-evolution in the fungi (Liu et al. 2012). Totiviruses 'snatch' the hosts' mRNA caps (modified guanines at the 5' end) with a unique binding mechanism involving at least five proposed residues of the coat protein (Fujimura & Esteban 2011). Frequently, totiviruses in fungi are associated with a satellite killer dsRNA virus that codes for a toxin (Bostian et al. 1984). The fossils of totiviruses are best characterized in the CTG yeast, *S. stipitis* (Frank & Wolfe 2009; Taylor & Bruenn 2009). There are four tandem copies of a capsid-like protein gene in the genome of *S. stipitis*, but it is unknown if these copies are translated into proteins. We explored the existence of RNA viruses in the CTG clade of yeasts and their possible mode of co-evolution by attempting to isolate modern and fossil viral genomes and their products from *S. stipitis* and its relatives *Scheffersomyces segobiensis* and *Scheffersomyces coipomoensis*.

Materials & Methods

Cell Cultures. Freeze-dried culture stock was obtained from the USDA ARS Culture Collection for *Scheffersomyces segobiensis* (Santa Maria & Garcia Aser) Kurtzman & M. Suzuki (2010) NRRL Y-11571 (Type strain) and for *Scheffersomyces coipomoensis* (Ramirez & Gonzalez) Urbina & Blackwell, 2012 comb. nov. NRRL Y-17651 (Type strain). Yeast cultures were grown in 150 ml of YPD broth (yeast extract 1%, peptone 2%, and dextrose 2%) with an inoculum (single colony) of cells from streaked YPD agar plates.

dsRNA assay. Total nucleic acids (depleted of ribosomal RNA) were extracted from whole cells (Bruenn & Keitz 1976). We purified dsRNA using CF-11 chromatography (Franklin 1966). The results (see Fig. S1) were consistent over the two-year period of more than 10 assays, indicating a stable infection.

Viral particle isolation. The tentative totivirus and satellite virus dsRNAs from *S. segobiensis* were isolated from a CsCl gradient fraction with a density of 1.40 g/cc (Fig. S2), which is expected based on the known *Saccharomyces cerevisiae virus L-A*. Note that even though totiviruses and their satellite viruses are separately encapsidated, there is an overlap in their densities, since particles can encapsidate up to four copies of satellite dsRNA (making a total of about 4.8 kbp, essentially the same as the totivirus at 4.6 kbp).

PCR, RTPCR, Sanger and Next Generation Sequencing. We extracted total RNA from yeast cells using the Masterpure yeast RNA purification kit (Epicentre) and an RNAse-free DNAse treatment. RNA-seq was used to sequence the putative RNA virus, examine the tRNA expression of the host, test for host expression of known fungal viral sequences, and isolate host protein-coding sequences for bioinformatics analysis. Ribosomal RNA species were removed using the Ribo-zero Magnetic Gold kit (Epicentre). We then prepared an RNA-seq library using the ScriptSeq™ v2 RNA-Seq Library Preparation Kit (Epicentre). Libraries were quantified using the Agilent 2100 Bioanalyzer RNA 6000 Pico Chip and submitted to the University at Buffalo Next Generation sequencing facility for RNA sequencing. The facility carried out RNA-seq using 50-cycle paired-end runs on two flow cells of an Illumina HiSeq 2000. *De novo* RNA sequence assembly was carried out in CLC Assembly Cell 4.06 (http://www.clcbio.com) on an Apple Macintosh Mac Pro Xeon 64-bit workstation. The putative viral contigs were reassembled for mapping purposes using the reference assembly algorithm and the *de novo* contigs as references. A total of 409665 reads were mapped to the totivirus with an average coverage of 4404.80 times.

The putative viral sequence was confirmed by Sanger sequencing using random and specific primers for cDNA library construction. We used the Takara 5'-Full RACE Core Set to expand sequence. Sanger and next generation sequences were compared using Geneious version 5.6.3 created by Biomatters (available from http://www.geneious.com/).

Because endogenous RNA viruses of fungi can be fragmented and differ greatly in their nucleotide sequences from known viruses (Taylor et al. 2009), PCR probes alone are often an ineffective tool for paleoviral discovery. We therefore carried out 454 Life Sciences (http://www.454.com) sequencing with GS FLX Titanium series reagents of a DNA library from *Scheffersomyces coipomoensis*. This form of sequencing also permitted multigene bioinformatics analysis of the host protein coding genes. Strain identity of the assembly was confirmed by BLAST analysis (Altschul et al. 1990) of the nuclear ribosomal RNA sequences that are known for *S. coipomoensis*. We obtained 614185 reads with about 252 Mb of aligned bases. The assembly carried out in Newbler (http://www.454.com), yielded 14.7 Mb of aligned bases (488 contigs) with an average peak depth of 12X.

To establish that the virus was coded by exogenous RNA and not by the DNA of the host, we compared RT-PCR and PCR products. For RNA templates, DNase-treated extracts were exposed to RT-PCR using the Qiagen one step RT-PCR kit. For DNA templates, nucleic acid extracts were exposed to PCR by excluding reverse transcriptase from the RT-PCR protocol. We amplified a fragment of the single copy xylose reductase gene as a positive control for the PCR of DNA. Primers used were: segoxylF CTGTTCTGAACAGATCTACCGTGC (xylose reductase), segoxylR AAGTATGGGTGGTGTTCAACTTGC (xylose reductase), SvLgap3F CGCAATACGACCAGGAGATTG (RdRp of virus from *S. segobiensis*), and segoSvLgap3R: GTACACCAAGGTTAGTAGACAAG (RdRp of virus from *S. segobiensis*). cDNA synthesis was performed at 48° C for 30 minutes, followed by 15 minutes at 94° C for reverse transcriptase deactivation and Taq activation. DNA only reactions were added to the thermal cycler 2 mins before the end of the previous 94° C step to activate the Taq polymerase. PCR amplification was done for 35 cycles of 94° C for 30 s, 48° C for 30 s, and 72° C for 1 min. A final extension at 72° C for 10 min was performed. New sequences from this study have the following Genbank accession numbers: KC610514, KC616419-KC616429.

Bioinformatics and protein mass spectroscopy. We obtained amino acid sequences from totivirids using the BLAST blastp algorithm with the the capsid gene and the RdRp sequences of *Saccharomyces cerevisiae virus L-BC(La)* as queries and E_values <1e-05. We searched the non-redundant (nr) peptide sequence database (National Center for Biotechnology Information, Bethesda, USA) and the Department of Energy Joint Genome Institute (J.G.I.) genome browser for matches. Fossil or paleoviral copies of Totivirus-like genes were identified by significant BLAST tblastx hits (E_values <1e-05) of relevant NCBI databases using the sequences of *Saccharomyces cerevisiae virus L-BC(La)*. Duplicated capsid gene copies adjacent to the complete integrated viral genomes in the assemblies of *S. stipitis* and *D. hansenii* were assumed to be paralogs (Taylor & Bruenn 2009). These duplicated paleoviruses and closely related (i.e. phylogenetic sister viruses) co-infecting viral strains were omitted for the phylogenetic analyses. Sequences were aligned using MAFFT (Katoh et al. 2009) with default settings. We carried out maximum likelihood analyses with PhyML 3.0 as implemented in Seaview 4.3.5 (Anisimova & Gascuel 2006; Gouy et al. 2010). Model optimization in Prottest (Abascal et al. 2005) indicated that the LG + invariable sites parameter (I)+ gamma parameter for among-site rate variation (G) was the best fit under a Bayesian Information Criterion (BIC). For reliability estimates we used SH-like approximate likelihood ratio tests (Anisimova & Gascuel 2006). Searches were comprised of five random starts under the subtree pruning and regrafting (SPR) algorithm and midpoint rooted.

Aguileta et al. (2008) found that concatenation of the two most informative genes from a genomic scale assessment recovered an expected reference fungal phylogeny with strong support. We used the approach of Taylor et al. (2009) who concatenated five of the most phylogenetically informative fungal genes (Aguileta et al. 2008) to estimate fungal relations. Accession numbers for the genes used (Minichromosome Maintenance protein 7[MCM7], Kontroller of Growth[KOG1], Elongator complex subunit[ELP3], NAD-specific glutamate dehydrogenase[GDH2], and acetolactate synthase [ILV2]) in the fungal analysis are presented in Table S1. Data were collected from GenBank and from our newly sequenced cultures of *S. coipomoensis* and *S. segobiensis*. Sequences were aligned in MAFFT and exposed to maximum likelihood analyses in RAxML 7.3.2 (Stamatakis 2006) and in PhyML 3.0. Models were partitioned by gene in RAxML using the best-fit models as indicated by PartitionfinderProtein (Lanfear et al. 2012).

Relative synonymous codon usage (Sharp et al. 1986) and third position base composition for yeasts and viruses was calculated using the CAIcal server (Puigbò et al. 2008). For viruses the entire open reading frame of the genome was used in the calculations. For yeasts we used the representative genes from the phylogenetic analysis. Species used for the yeast and viral codon usage analyses are listed in Table S2. Bivariate plots of RCSU and base composition were graphed using the R statistical programming language (Ihaka & Gentleman 1996).

FSfinder (Moon et al. 2004) was used to locate putative slippery sites and pseudoknots in the totiviral genome. The tRNA$_{CAG}^{Ser}$ for *S. segobiensis* was folded according to the model for *Candida albicans* (Santos et al. 2011) using the VARNA secondary structure visualization program (Darty et al. 2009).

Structural information for *Saccharomyces cerevisiae virus L-A* is from the crystal structure of the ScV L-A capsid protein (Naitow et al. 2002). For ScV L-BC (La) and SsV L, structural information was predicted by the I-TASSER webserver using ScV L-A cap as a template (Roy et al. 2010; Roy et al. 2012).

To examine protein expression of paleoviral copies we isolated crude protein from *S. stipitis* and *S. cerevisiae* by French press and further isolated proteins migrating between 73 kDa and 92 kDa from 10% SDS-PAGE. Protein mass spectroscopy was carried out at the Seattle Biomedical Research Institute Proteomics Core Facility.

Results and Discussion

Because the fossil viruses in yeast have a similar architecture to dsRNA totiviruses (Taylor & Bruenn 2009), we carried out a specific chromatographic assay for dsRNA. We detected dsRNA products in *S. segobiensis* with approximately the same gel-estimated size to the totivirus (4.5 kb) and satellite virus of *S. cerevisiae* (1.2 kb, Fig. S1). Viral particles containing both sizes of dsRNAs were also isolated by CsCl equilibrium gradient centrifugation (Fig. S2). No such products were detected in *S. stipitis* or in *S. coipomoensis*. A tblastn using the protein sequences of the two known totiviruses from *S. cerevisiae* as queries revealed significant matches to a contig from a database of our RNA sequence assemblies (extracted from *S. segobiensis* cells) of similar length to the dsRNA particles on the gel. Assemblies from Sanger sequencing of the RNA virus agreed with the assembly using Illumina RNA sequencing, but the 5’ UTR was complete only in the Illumina assembly. The assembled virus had the genomic architecture of totiviruses with overlapping capsid and RdRp open reading frames flanked by 5’ and 3’ UTRs (Fig. 1). We identified a putative slippery site for ribosomal frameshifting (GGGTTTT) at position 1981 that was independently identified using the frameshift prediction software fsfinder (Moon et al. 2004). The five sites identified as functionally important for cap-snatching in totiviruses are conserved in the virus from *S. segobiensis* (Fig. S3), suggesting a cap-snatching mechanism similar to those of well-studied totiviruses. These sites show weak conservation in the fossil copies, consistent with the loss of host mRNA decapping in host-coded elements. The successful PCR amplification of a single copy nuclear gene fragment (xylose reductase gene) from the host genome indicates that the DNA template was of sufficient quality to detect endogenous viral genes using our methods (Fig. S4). However, the primers nested within the viral genome failed to PCR amplify a DNA copy from the host (*S. segobiensis*) genome, but RT-PCR did amplify an RNA copy of the viral gene. The results support the existence of an exogenous RNA virus in *S. segobiensis* with the genomic architecture of a totivirus.

Further evidence of affinity to totiviruses comes from sequence analysis of the virus in *S. segobiensis*. A BLAST blastp analysis of the RdRp-like ORF yielded a conserved domain match (E=1.56e-58) to RdRP_4, a viral RNA-directed RNA-polymerase family that includes "RdRPs from Luteovirus, Totivirus and Rotavirus". The best expect value (E= 2e-137) obtained was the totivirus, *Saccharomyces cerevisiae virus L-BC(La)*, with an identity of 37% of residues. The RdRp gene phylogeny (Fig. 2A) positioned the tentative species *Scheffersomyces segobiensis virus L* within the totiviruses and most closely related to *Saccharomyces cerevisiae virus L-BC(La)*. Support values are strong enough to rule out random error as an explanation for the evolutionary position of *Scheffersomyces segobiensis virus L* within the totiviruses. The less conserved capsid gene tree (Fig. 2B) showed a similar association, but with fewer outgroup sequences and more paleoviruses. We detected fossil totiviruses in the genomes of *S. coipomoensis,* and *Pichia membranifaciens* and a putative totivirus in the assembly of *Nadsonia fulvescens*. We deem the sequences from *Nadsonia* to be putative viral sequences because they are present in the RNA-based EST libraries, but not the DNA based genome assembly (see Liu et al. (2012) for a discussion of this approach to dsRNA virus discovery).

As with other totiviruses, we found evidence that *Scheffersomyces segobiensis virus L* has a putative killer satellite virus. We carried out a BLAST search to identify candidate RNA contigs for the dsRNA band observed earlier. A sequence was obtained of the correct size (1.2 kbp) that had a significant match to the killer satellite K2 virus of *S. cerevisiae*. An RTPCR with specific primers confirmed that the contig was not coded in the DNA genome of the yeast. The existence of a satellite virus bolsters the evidence for a totivirus in the CTG clade.

The evolutionary and genomic evidence indicates that *Scheffersomyces segobiensis virus L* originated from exogenous totiviruses that use a standard genetic code. The RdRp permits the deepest assessment of evolution, and it reveals a derived position of the *Scheffersomyces segobiensis virus L* at the tip of the standard code totivirids. The known paleoviruses in the CTG clade are distantly related to *Scheffersomyces segobiensis virus L* and lack the ability to produce the fusion gene product typical of totiviruses (Taylor & Bruenn 2009). Nor is there any evidence of a functioning totivirus genome being coded in the DNA from known genomes of yeasts. Moreover, our RTPCR results indicate that the DNA genome of the host, *S. segobiensis*, lacks coding sequences related to the viral genome. Finally, successful endogenization of a virus that imparts a selective disadvantage by obligately decapping host mRNA seems unlikely. Most of the successful paleoviruses in the CTG clade possess functionally differing residues at the decapping sites. Because the CTG virus we discovered has the conserved residues for such a decapping mechanism, an “escaped” genome hypothesis requires that the unique decapping mechanism was lost in the host genome and then re-evolved in the escaped viral genome – also unlikely. We conclude that adaptation to the genetic code shift of hosts happened in exogenous viruses.

Our viral evolutionary trees are consistent with the jumping of viruses between hosts with different genetic codes. Host products from the CTG clade are phylogenetically interspersed with standard code sequences in both major clades and the virus/paleovirus phylogenies bear little resemblance to the host relationships. For example, a virus from the truffle (*Tuber aestivum*) is most closely related to the fossil virus in the CTG yeast *S. stipitis* rather than the virus of another genus of basidiomycete, *Xanthophyllomyces*. Horizontal transfer of paleoviruses could be a source of some evolutionary noise, but this process appears rare in the CTG clade. The timescale of the totivirus evolution is difficult to estimate because of the dearth of reliable fossil calibrations for fungi (Berbee & Taylor 2010; Rolland & Dujon 2011). However, our observed close sequence and structural similarity for rapidly evolving capsid proteins of RNA viruses likely postdates the ancient split of ascomycetes with basidiomycetes (452 MYA to 1400 MYA) and the origin of the CTG clade (> 150 MYA)(Berbee & Taylor 2010; Massey et al. 2003; Pesole et al. 1995).

The evolutionary position, CUG codon usage, and tRNA expression evidence are consistent with the host, *S. segobiensis,* having a modified CTG genetic code. Our phylogenetic analysis (Fig. 3A) of nuclear protein coding genes revealed strong support for the placement of the host yeast within the CTG clade and as a sister species to the CTG species *S. stipitis*. The monophyly of the CTG clade is well established in evolutionary genomics (Butler et al. 2009; Louis et al. 2012; Wohlbach et al. 2011). The close sister group relationship of *S. segobiensis* and *S. stipitis* has also been independently supported by several studies using the ribosomal rRNA gene family (Cadete et al. 2012; Kurtzman 2010; Urbina & Blackwell 2012). CUG comprises a substantial percentage of codons for leucine residues in standard code yeasts, while being almost absent in the CTG yeasts at the homologous leucine position (Fig. 3B). The usage of CUG in CTG yeast (including *S. segobiensis*) is underrepresented compared to the standard code yeast (Fig. 3C). A BLAST search of RNA contigs for the characteristic modified chimeric serine tRNA of *S. stipitis* that recognizes CUG revealed a significant match in *S. segobiensis*. That is, the sequence has both serine and leucine identity sites (Fig. 3D). We failed to detect a "standard code" leucine tRNA. The pretRNA species (Fig. S5) had unique mutations in the flanking regions from *S. stipitis*, consistent with the modest divergence of a sister species.

As the genetic code shift had a profound functional effect on the proteins of yeasts, we expected the virus to adapt to the shift. We found that *Scheffersomyces segobiensis virus L* had but a single codon of the modified CUG type. This evolutionary loss of CUG codons resulted in the lowest CUG frequency known among related mycoviruses, where the CUG codon is generally overrepresented. The sole CUG in *Scheffersomyces segobiensis virus L* occurs at a position in the capsid protein that appears to be structurally unimportant (Fig. 1). *Scheffersomyces segobiensis virus L* appears to have adapted to the host shift in genetic code by eliminating functionally relevant CUG codons.

In plots of relative synonymous codon usage (RSCU) versus third position base composition (Fig. 4A; Table S2), *Scheffersomyces segobiensis virus L* grouped with CTG yeasts rather than with other totivirids. The same pattern of *Scheffersomyces segobiensis virus L* grouping with CTG yeasts to the exclusion of other viruses was found for relevant leucine codons (Figs. 3B,C). In *S. cerevisiae*, CUN codons are decoded by either tRNA-UAG or by tRNA-GAG. But in the CTG clade yeast, CUN codons are decoded differently. Here a derived tRNA-CAG decodes the reassigned CUG codon, while the remaining CUN family codons are decoded by a single tRNA-IAG. The inosine interacts only weakly with CUA compared to CUC and CUU in *C. albicans* (Massey et al. 2003), and seems the most likely driving force behind reduction in CUA usage in CTG yeasts. We note that *Scheffersomyces segobiensis virus L* lacks a bias in base composition at the third position for the examined codons, suggesting that the observed codon usage bias in *Scheffersomyces segobiensis virus L* is more complicated than base compositional shifts alone.

Our analysis of the genome of *Scheffersomyces segobiensis virus L* is consistent with adaptation to offset the functional effects of the genetic code shift in the host. But, our results also indicate that the endogenization of viral genes by host yeasts of both genetic codes is more common than previously thought. Presently, it is unknown if these endogenous non-retroviral genes function as proteins or are merely transcriptional noise. Here, we show that at least one of these genes derived from a totivirus is expressed as a protein. Isolation of proteins migrating between 73 kDa and 92 kDa from *S. stipitis* yielded approximately 535 proteins, as estimated from the equivalent size range of proteins in *S. cerevisiae*. The mass spectroscopy analysis was able to unambiguously identify the capsid protein (cap) of *Saccharomyces cerevisiae virus L-A* in the *S. cerevisiae* control and 153 proteins in *S. stipitis*. By this method, we were able to detect one of the four NIRV capsid polypeptides from *S. stipitis*. Figures S6A and S6B show the distribution of tryptic peptides detected by mass spectroscopy from *S. stipitis* capsid4 and *Saccharomyces cerevisiae virus L-A* cap. All of the *S. stipitis* virus-like peptides had probabilities of 0.999 or greater. Our inability to detect the remaining cap proteins from *S. stipitis* may be due to the lack of sensitivity of the method used. We were able to detect a single tryptic peptide from *Saccharomyces cerevisiae virus L-BC(La)*cap, which is present at about one tenth of the concentration of *Saccharomyces cerevisiae virus L-A* cap. Our results indicate that the co-option of a non-retroviral RNA viral protein has occurred as with retroviral proteins (e.g. syncytin genes (Feschotte & Gilbert 2012)). Although expression of similar capsid proteins has an antiviral effect (Valle & Wickner 1993; Yao & Bruenn 1995), the current function of the co-opted viral proteins in yeast is unknown.

Conclusions

Our discoveries indicate that a major evolutionary transition involving a change in the genetic code of the fungi failed to result in permanent host escape from viruses. We found evidence of present or past viral infection in five lineages of yeasts with a modified genetic code. Thus, viral infection is likely widespread in the CTG clade of fungi. The mode of viral adaptation recalls the prediction of Crick (1968) that some viruses would be less susceptible to evolutionary “freezing” because they present a reduced protein target. The genomes of totiviruses are among the smallest known for RNA viruses (Holmes, 2009). Still, even with a small viral genome we found evidence that exogenous viral adaptation was associated with the elimination of modified codons from functional positions. Our results also highlight the value of recent paleovirological approaches to understanding virus-host biology (Aswad & Katzourakis 2012; Feschotte & Gilbert 2012; Holmes 2011; Koonin 2010; Patel et al. 2011). Fossil copies in the genomes of yeasts informed us about prior infections and made possible our discovery of a virus adapted to a modified nuclear genetic code. Moreover, we found that at least one of these “fossil” genes is co-opted by the yeast host and expressed as a protein.

Acknowledgements

We thank Thomas W. Jeffries for *Scheffersomyces stipitis* strain CBS 6054. Jennifer Jamison (Next-Generation Sequencing and Expression Analysis Core,

State University of New York at Buffalo) and Yuko Ogato (Proteomics Shared Resource, Fred Hutchinson Cancer Research Center) provided technical advice.

References

Abascal F, Zardoya R, and Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104-2105.

Aguileta G, Marthey S, Chiapello H, Lebrun MH, Rodolphe F, Fournier E, Gendrault-Jacquemard A, and Giraud T. 2008. Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies. *Systematic Biology* 57:613-627.

Altschul SF, Gish W, Miller W, Myers, EW, and Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Anisimova M, and Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology* 55:539-552.

Aswad A, and Katzourakis A. 2012. Paleovirology and virally derived immunity. *Trends in Ecology & Evolution*. 27: 627-636

Berbee ML, and Taylor JW. 2010. Dating the molecular clock in fungi - how close are we? *Fungal Biology Reviews*:1-16.

Bostian KA, Elliott Q, Bussey H, Burn V, Smith A, and Tipper DJ. 1984. Sequence of the preprotoxin dsRNA gene of type 1 killer yeast: multiple processing events produce a two-component toxin. *Cell* 36:741-751.

Bruenn J, and Keitz B. 1976. The 5' ends of yeast killer factor RNAs are pppGp. *Nucleic Acids Research* 3:2427-2436.

Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL et al. . 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* 459:657-662.

Cadete RM, Melo MA, Dussán KJ, Rodrigues RCLB, Silva SS, Zilli JE, Vital MJS, Gomes FCO, Lachance M-A, and Rosa CA. 2012. Diversity and Physiological

Characterization of D-Xylose-Fermenting Yeasts Isolated from the Brazilian Amazonian Forest. *PLoS One* 7:e43135.

Crick FHC. 1968. The origin of the genetic code. *Journal of Molecular Biology* 38:367-379.

Darty K, Denise A, and Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25:1974.

do Céu Santos M, and Santos M. 2012. Structural and molecular features of non-standard genetic codes. In: Cannarozzi G, and Schneider A, eds. *Codon Evolution: Mechanisms and Models*. Oxford: Oxford University Press, 258-270.

Feketová Z, Mašek T, Vopálenský V, and Pospíšek M. 2010. Ambiguous decoding of the CUG codon alters the functionality of the Candida albicans translation initiation factor 4E. *Fems Yeast Research* 10:558-569.

Feschotte C, and Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics* 13:283-296.

Fitzpatrick DA. 2011. Horizontal gene transfer in fungi. *FEMS Microbiology Letters* 329:1-8.

Frank AC, and Wolfe KH. 2009. Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. *Eukaryotic Cell* 8:1521-1531.

Franklin RM. 1966. Purification and properties of the replicative intermediate of the RNA bacteriophage R17. *Proceedings of the National Academy of Sciences, USA* 55:1504-1511.

Fujimura T, and Esteban R. 2011. Cap-snatching mechanism in yeast LA double-stranded RNA virus. *Proceedings of the National Academy of Sciences, USA* 108:17667-17671.

Gouy M, Guindon S, and Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27:221-224.

Holmes EC. 2009. *The evolution and emergence of RNA viruses*. New York: Oxford University Press.

Holmes EC. 2011. The evolution of endogenous viral elements. *Cell Host & Microbe* 10:368-377.

Ihaka R, and Gentleman R. 1996. R: A language for data analysis and graphics. *Journal of computational and graphical statistics* 5:299-314.

Katoh K, Asimenos G, and Toh H. 2009. Multiple Alignment of DNA Sequences with MAFFT. *Methods Mol Biol* 537:39-64.

Koonin EV. 2010. Taming of the shrewd: novel eukaryotic genes from RNA viruses. *BMC Biol* 8:2.

Kurtzman CP. 2010. Phylogeny of the ascomycetous yeasts and the renaming of Pichia anomala to Wickerhamomyces anomalus. *Antonie van Leeuwenhoek* 99:13-23.

Lanfear R, Calcott B, Ho SYW, and Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29:1695-1701.

Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Yi X, and Jiang D. 2012. Discovery of Novel dsRNA Viral Sequences by In Silico Cloning and Implications for Viral Diversity, Host Range and Evolution. *PLoS One* 7:e42147.

Louis VL, Despons L, Friedrich A, Martin T, Durrens P, Casarégola S, Neuvéglise C, Fairhead C, Marck C, and Cruz JA. 2012. Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization. *G3: Genes| Genomes| Genetics* 2:299-311.

Massey SE, Moura G, Beltrão P, Almeida R, Garey JR, Tuite MF, and Santos MAS. 2003. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. *Genome Research* 13:544-557.

Moon S, Byun Y, Kim HJ, Jeong S, and Han K. 2004. Predicting genes expressed via− 1 and+ 1 frameshifts. *Nucleic Acids Research* 32:4884-4892.

Naitow H, Tang J, Canady M, Wickner RB, and Johnson JE. 2002. L-A virus at 3.4 A resolution reveals particle architecture and mRNA decapping mechanism. *Nat Struct Biol* 9:725-728.

Patel MR, Emerman M, and Malik HS. 2011. Paleovirology--ghosts and gifts of viruses past. *Current Opinion in Virology* 1:304-309.

Pesole G, Lotti M, Alberghina L, and Saccone C. 1995. Evolutionary origin of nonuniversal CUG (Ser) codon in some Candida species as inferred from a molecular phylogeny. *Genetics* 141:903.

Puigbò P, Bravo IG, and Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. *Biology direct* 3:38.

Richards TA, Leonard G, Soanes DM, and Talbot NJ. 2011. Gene transfer into the fungi. *Fungal Biology Reviews* 25:98-110.

Rocha R, Pereira PJB, Santos MAS, and Macedo-Ribeiro S. 2011. Unveiling the structural basis for translational ambiguity tolerance in a human fungal pathogen. *Proceedings of the National Academy of Sciences, USA* 108:14091-14096.

Rolland T, and Dujon B. 2011. Yeasty clocks: Dating genomic changes in yeasts.1-9.

Roy A, Kucukural A, and Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* 5:725-738.

Roy A, Yang J, and Zhang Y. 2012. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research* 40:W471-W477.

Santos MAS, Gomes AC, Santos MC, Carreto LC, and Moura GR. 2011. The genetic code of the fungal CTG clade. *Comptes rendus - Biologies* 334:607-611.

Shackelton LA, and Holmes EC. 2008. The role of alternative genetic codes in viral evolution and emergence. *Journal of Theoretical Biology* 254:128-134.

Sharp PM, Tuohy TMF, and Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* 14:5125-5143.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690.

Taylor DJ, and Bruenn J. 2009. The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biology* 7:88.

Urbina H, and Blackwell M. 2012. Multilocus Phylogenetic Study of the Scheffersomyces Yeast Clade and Characterization of the N-Terminal Region of Xylose Reductase Gene. *PLoS One* 7:e39128.

Valle RP, and Wickner RB. 1993. Elimination of L-A double-stranded RNA virus of *Saccharomyces cerevisiae* by expression of *gag* and *gag-pol* from L-A cDNA clone. *Journal of Virology* 67:2764-2771.

Wohlbach DJ, Kuo A, Sato TK, Potts KM, Salamov AA, Labutti KM, Sun H, Clum A, Pangilinan JL, Lindquist EA et al. . 2011. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proceedings of the National Academy of Sciences, USA*:1-6.

Yao W, and Bruenn JA. 1995. Interference with replication of two double-stranded RNA viruses by production of N-terminal fragments of capsid polypeptides. *Virology* 214:215-221.

Figure 1

Comparison of the genomic architecture of the newly discovered *Scheffersomyces segobiensis virus L* [SsV L] (in red) that uses a modified nuclear genetic code with those of related totiviruses (*Saccharyomyces cerevisiae virus L-BC(La)* [ScV L

The figures show overlapping reading frames for capsid (CP) and RNA dependent RNA polymerase genes (RdRp) that are typical of the double-stranded RNA totiviruses. Terminal UTRs (untranslated regions) and central overlapping reading frames are distinguished by dark colored shading. The positions of the ribosomal frameshift sites are indicated in yellow. Coding CUG codons are represented by lollipops. Capsid protein secondary structural domains are shown in purple (α-helices) and orange (β-sheets). BLASTp results for each of the previously known totiviruses to SsV L are shown as grey lines underlying the respective genomes, with darker shades indicating a lower expect value. The scale bar increments represent 50 nucleotides.

Fig. 1

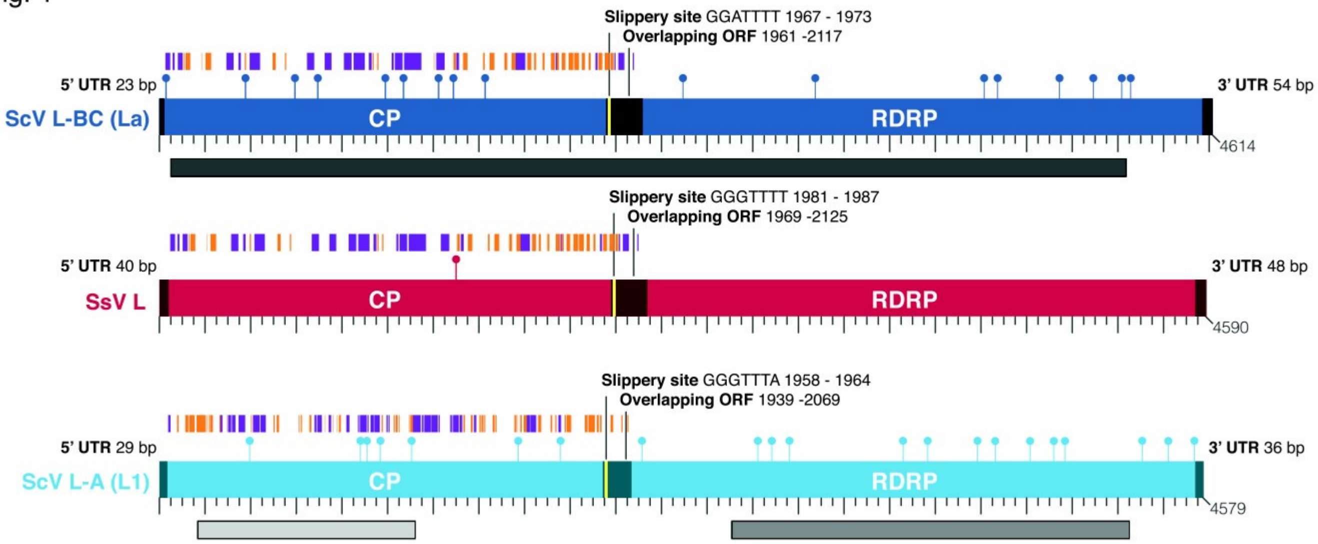
Slippery site GGATTTT 1967 - 1973
Overlapping ORF 1961 -2117
5' UTR 23 bp
3' UTR 54 bp
ScV L-BC (La)
CP
RDRP
4614
Slippery site GGGTTTT 1981 - 1987
Overlapping ORF 1969 -2125
5' UTR 40 bp
3' UTR 48 bp
SsV L
CP
RDRP
4590
Slippery site GGGTTTA 1958 - 1964
Overlapping ORF 1939 -2069
5' UTR 29 bp
3' UTR 36 bp
ScV L-A (L1)
CP
RDRP
4579

Figure 2

Evolutionary relationships of exogenous and endogenous totiviruses showing the derived and non-monophyletic positions of viruses and paleoviruses from CTG yeast.

Asterisks indicate paleoviral sequences. Numbers are support values (SH-like approximate likelihood tests estimated in PhyML 3.0). The phylograms are estimated using the maximum likelihood (ML) optimality criterion from alignments of predicted amino acid residues of (A) the RNA dependent RNA polymerase gene (RdRp) and (B) the capsid gene. Colored dashed boxes indicate the two major clades of totivirus-like sequences. Genbank Accession numbers are provided in parentheses.

Fig. 2

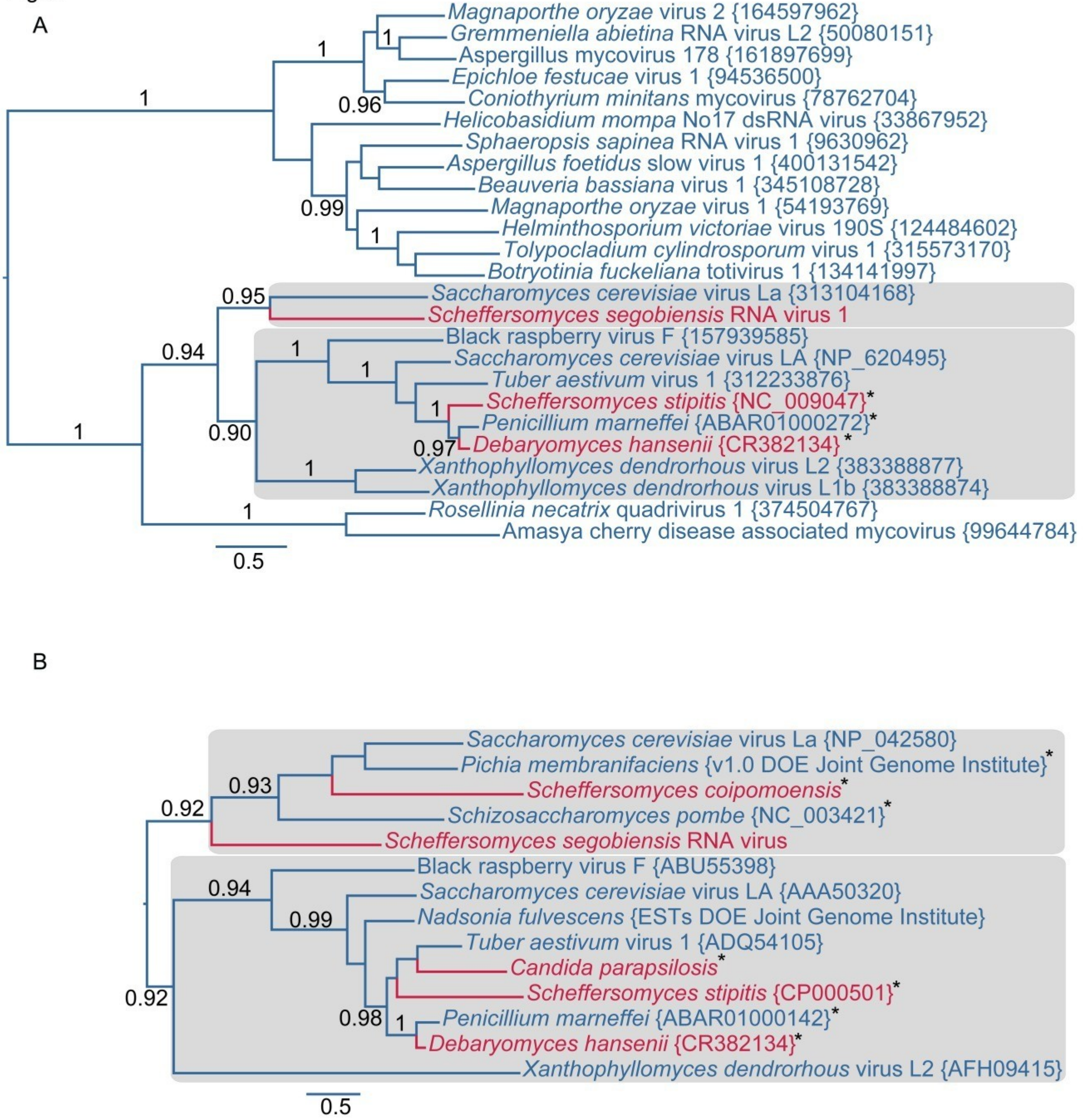
A
Magnaporthe oryzae virus 2 {164597962}
Gremmeniella abietina RNA virus L2 {50080151}
Aspergillus mycovirus 178 {161897699}
Epichloe festucae virus 1 {94536500}
Coniothyrium minitans mycovirus {78762704}
Helicobasidium mompa No17 dsRNA virus {33867952}
Sphaeropsis sapinea RNA virus 1 {9630962}
Aspergillus foetidus slow virus 1 {400131542}
Beauveria bassiana virus 1 {345108728}
Magnaporthe oryzae virus 1 {54193769}
Helminthosporium victoriae virus 190S {124484602}
Tolypocladium cylindrosporum virus 1 {315573170}
Botryotinia fuckeliana totivirus 1 {134141997}
Saccharomyces cerevisiae virus La {313104168}
Scheffersomyces segobiensis RNA virus 1
Black raspberry virus F {157939585}
Saccharomyces cerevisiae virus LA {NP_620495}
Tuber aestivum virus 1 {312233876}
Scheffersomyces stipitis {NC_009047}*
Penicillium marneffei {ABAR01000272}*
Debaryomyces hansenii {CR382134} *
Xanthophyllomyces dendrorhous virus L2 {383388877}
Xanthophyllomyces dendrorhous virus L1b {383388874}
Rosellinia necatrix quadrivirus 1 {374504767}
Amasya cherry disease associated mycovirus {99644784}
1
1
1
0.96
0.99
1
0.95
0.94
1
1
1
0.90
0.97
1
1
1
0.5
B
Saccharomyces cerevisiae virus La {NP_042580}
Pichia membranifaciens {v1.0 DOE Joint Genome Institute}*
Scheffersomyces coipomoensis*
Schizosaccharomyces pombe {NC_003421}*
Scheffersomyces segobiensis RNA virus
Black raspberry virus F {ABU55398}
Saccharomyces cerevisiae virus LA {AAA50320}
Nadsonia fulvescens {ESTs DOE Joint Genome Institute}
Tuber aestivum virus 1 {ADQ54105}
Candida parapsilosis*
Scheffersomyces stipitis {CP000501}*
Penicillium marneffei {ABAR01000142}*
Debaryomyces hansenii {CR382134}*
Xanthophyllomyces dendrorhous virus L2 {AFH09415}
0.93
0.92
0.94
0.99
0.92
0.98
1
0.5

Figure 3

Evidence that the viral host, *Scheffersomyces segobiensis*, uses the modified genetic code of the "CTG" clade.

A) Midpoint-rooted maximum likelihood phylogram of CTG clade yeasts and related yeasts based on the protein sequences of the five most phylogenetically informative genes for fungi. Branches are labeled with support values from approximate likelihood ratio tests and nonparametric bootstrapping. Blue shading indicates standard code yeasts, while red shading indicates CTG code yeasts. Gray spheres indicate lineages with evidence of past or prior infections with totiviruses; B) Comparison of CUG codon positions in each taxon versus homologous amino acid residues in *Saccharyomyces cerevisiae*. Blue bars represent the percent of *S. cerevisiae* leucine residues that are coded by the CUG codon for each taxon, while red bars represent the percent of *S. cerevisiae* serine residues that are coded by CUG for each taxon. C) Boxplots showing relative synonymous codon usage (RSCU) for CUG codon usage by each taxon, the blue plot represents CUG RSCU values for non-CTG clade yeast, while the red plot represents CTG clade yeast including *S. segobiensis*. D) The secondary structure model of the CTG clade type of tRNA^{SER} detected in *S. segobiensis*. The red shading indicates serine identifier sites, while blue shading indicates standard leucine identifier sites. The gray site is a typical guanine residue of the tRNA^{SER} of CTG yeast that lowers the leucine amino-acylation efficiency.

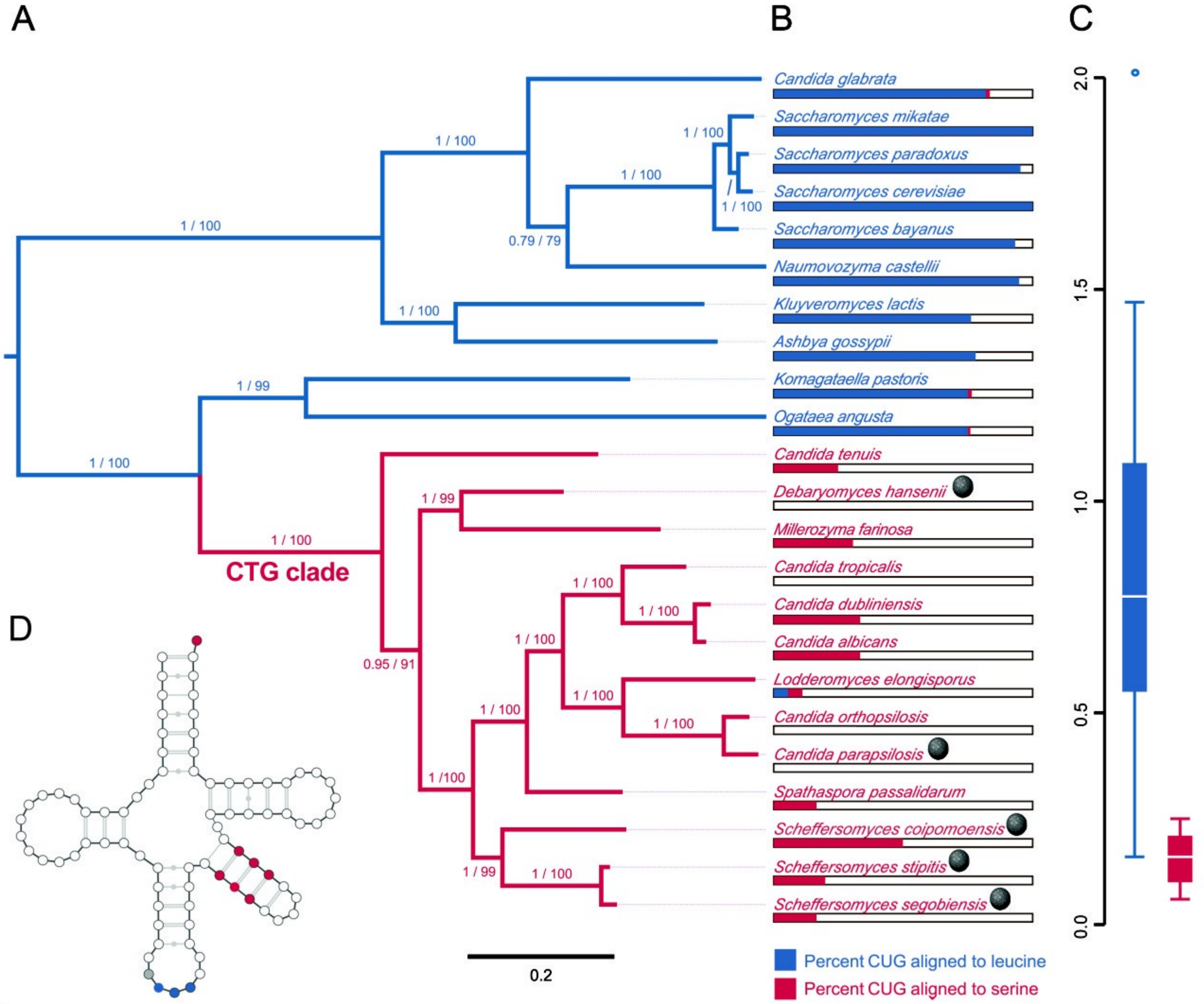
A
B
C
D
Candida glabrata
Saccharomyces mikatae
Saccharomyces paradoxus
Saccharomyces cerevisiae
Saccharomyces bayanus
Naumovozyma castellii
Kluyveromyces lactis
Ashbya gossypii
Komagataella pastoris
Ogataea angusta
Candida tenuis
Debaryomyces hansenii
Millerozyma farinosa
Candida tropicalis
Candida dubliniensis
Candida albicans
Lodderomyces elongisporus
Candida orthopsilosis
Candida parapsilosis
Spathaspora passalidarum
Scheffersomyces coipomoensis
Scheffersomyces stipitis
Scheffersomyces segobiensis
1 / 100
1 / 100
1 / 100
1 / 100
1 / 100
0.79 / 79
1 / 100
1 / 99
1 / 100
1 / 100
CTG clade
1 / 99
0.95 / 91
1 / 100
1 / 100
1 / 100
1 / 100
1 / 100
1 / 100
1 /100
1 / 99
1 / 100
0.2
2.0
1.5
1.0
0.5
0.0
Percent CUG aligned to leucine
Percent CUG aligned to serine

Fig. 3

Figure 4

Bivariate plots of relative synonymous codon usage (RSCU) for serine and leucine versus third position base composition in yeast and their dsRNA viruses. CTG clade yeasts are shown as solid red points, and their viruses as hollow red points.

Standard code yeasts are shown as solid blue points and their viruses as hollow blue points. A) CUG is used by standard code yeasts and their viruses but avoided by CTG clade yeasts and *Scheffersomyces segobiensis virus L* . B) Leucine codon UUG is overused by CTG clade yeasts and *S. segobiensis virus L* relative to standard code yeasts and their viruses. C) Leucine codon CUA is used by standard code yeasts and their viruses, but avoided by CTG clade yeasts and *Scheffersomyces segobiensis virus L* . Comparison with %N3 suggests that these codon preferences are not attributable to nucleotide composition alone.

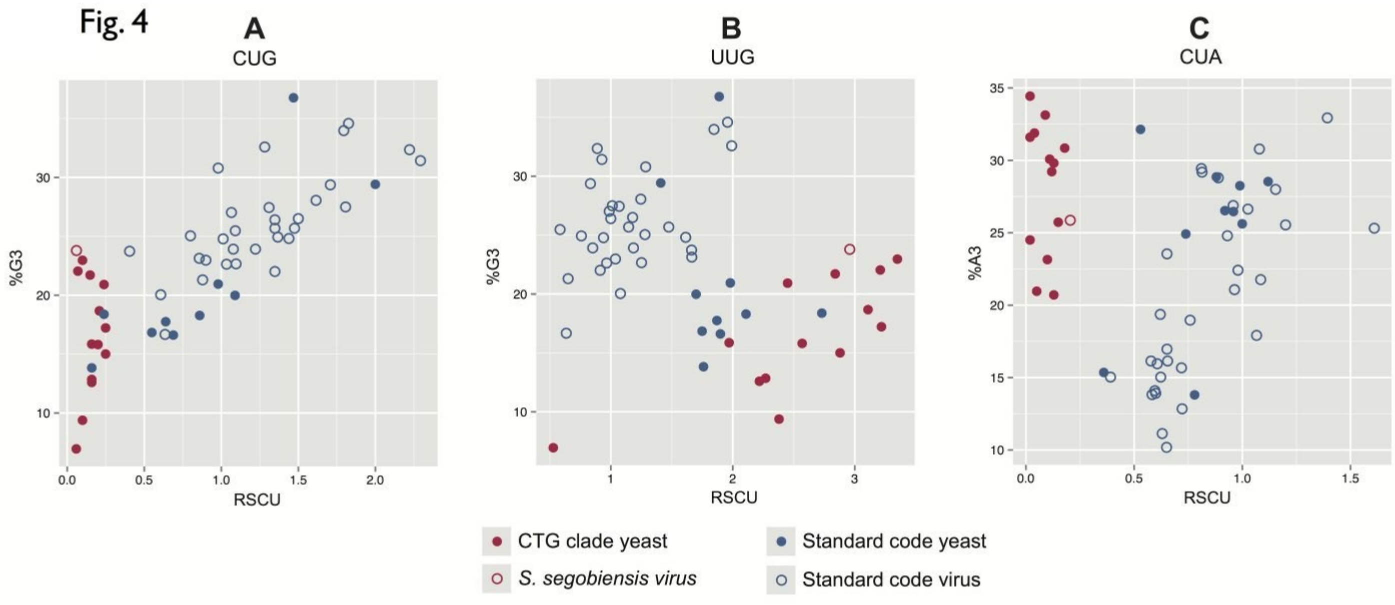
Fig. 4
A
CUG
%G3
RSCU
B
UUG
%G3
RSCU
C
CUA
%A3
RSCU
● CTG clade yeast
○ S. segobiensis virus
● Standard code yeast
○ Standard code virus