Title:

Deep selection shapes the intragenomic diversity of rRNA genes

Authors: Daniel Sultanov[1] and Andreas Hochwagen[1,*]

Affiliations: [1]Department of Biology, New York University, New York, NY, USA.

*Corresponding author. e-mail: *ah128@nyu.edu*

Abstract:

Ribosome biogenesis in eukaryotes is supported by hundreds of ribosomal RNA (rRNA) gene copies that are encoded in the ribosomal DNA (rDNA) and are thought to undergo frequent sequence homogenization by concerted evolution. Here, we present species-wide rDNA sequence analysis in *Saccharomyces cerevisiae* that challenges this model of evolution. We show that rDNA copies in yeast are far from homogeneous, both among and within isolates, and that rapid sequence homogenization is unlikely because many variants avoided fixation or elimination over evolutionary time. Instead, the diversity landscape across the rDNA shows clear functional stratification, suggesting that different copy-number thresholds for selection shape rDNA diversity. Notably, variants in the most highly conserved rDNA elements are sufficiently deleterious to exhibit signatures of purifying selection even when present in only 1% of rRNA gene copies. Our results portray a complex selection landscape that shapes rDNA diversity within a single species and reveal unexpectedly deep purifying selection of multi-copy genes.

The ribosomal DNA (rDNA) codes for the universal RNA components of the ribosome and therefore is essential for protein translation and viability in every species. In most eukaryotic genomes, dozens or hundreds of rDNA copies are arranged in one or multiple loci [1]. The current model of concerted rDNA evolution postulates that rDNA copies within a species evolve together and are largely homogeneous [2–4]. Within-genome rDNA sequence polymorphisms are thought to either quickly spread or be quickly eliminated through frequent unequal recombination events (10^{-2} - 10^{-3} per generation) that excise or reshuffle rDNA copies [4–6]. However, previous studies were constrained by very limited sequencing information, and with an increasing number of sequenced species, substantial within-genome polymorphisms have been detected across diverse phyla, including fungi [7–9], invertebrates [10–13], plants [14], and mammals, including humans [15–17].

In an alternative model, rDNA copies are subjected to selection individually rather than collectively, and the frequency of within-genome rDNA polymorphisms directly reflects the level of selection at a given site without needing to invoke the rapid and complete spreading of variants [18]. For example, sequence variation can exist if the mutations are neutral or beneficial, and mutation purging can result from strong purifying selection as long as maladaptive variants are sufficiently deleterious to be detectable against a backdrop of dozens of functional copies. As it were, the sequences of several lower-copy multi-gene families, including the major histocompatibility complex genes and the highly conserved histone and ubiquitin genes are individually under strong purifying selection [19–21]. Whether strong purifying selection similarly acts on the much more numerous copies of the rDNA is unknown. Here, we took advantage of the high sequencing coverage of the 1002 Yeast Genomes Project [22] to quantitatively assess

within- and between-genome sequence heterogeneity of the rDNA in *Saccharomyces cerevisiae*. Together with comparative and structural information of the ribosome, these data allowed us to identify signatures of stringent purifying selection and define regions of constrained and permissive variation of rRNA on a species scale.

Results:

Substantial rDNA sequence diversity

Each rDNA copy of *S. cerevisiae* consists of several structural elements, including the coding sequences of the four rRNAs (*5S*, *5.8S*, *18S*, *25S*), *ETS* and *ITS* elements, which are essential for rRNA maturation by interacting with small nucleolar RNAs and nucleases [23], and non-transcribed sequences (*NTS1* and *NTS2*), which contain the origin of replication and promoters [24–26] (**Fig. 1a**). To investigate the level of sequence diversity across these elements, we analyzed rDNA sequences of the 1002 Yeast Genomes Project, which comprises deep-sequencing data (~200-fold mean coverage) of 918 *S. cerevisiae* isolates from 21 ecological niches [22], including 133 haploid and 694 diploid isolates, as well as 91 isolates with higher ploidies. We used LoFreq [27] to identify rDNA sequences that differed from the rDNA prototype of the S288c reference strains and considered the alternative sequences to be rDNA "variants" if more than 0.4% of the rDNA reads - referred here as "variant frequencies" (VFs) - contained such sequences (**Fig. 1a**). Benchmarking with a test set showed that our pipeline efficiently recovered such low-frequency variants (> 97% recovery; **Supplementary Fig. 1**, also see Methods). As the median rDNA copy number among isolates lies around 180 copies, this approach effectively allows detection of variants that are present in only one or a few rDNA copies within a genome (**Supplementary Fig. 2**).

Across all isolates, we observed a high amount of rDNA sequence variation (**Fig. 1a; Supplementary Fig. 3**). Since clonal populations were sequenced, this diversity reflects sequence heterogeneity of rRNA gene copies within each genome rather than heterogeneity among cells. In total, we detected 2,520 variants across 2,258 polymorphic sites, comprised

primarily of single-nucleotide variants (SNVs) and short indels (< 3 bp; range: 1-24 bp; **Fig. 1b**, **Supplementary Fig. 4**). Almost half (42%) of those variants were shared across multiple isolates (**Supplementary Fig. 5**) but occurred at various VFs (i.e. in different proportions of rDNA copies) and thus represent variant frequency polymorphisms (VFPs; **Fig. 1a**). Among all isolates, we observed a total of 24,183 VFPs with frequencies from 0.4%-100% (**Fig. 1b**), providing a rich dataset for analysis (**Supplementary File 1**).

At least 15% of nucleotide positions in each element were polymorphic (**Fig. 1b**), and the strong correlation between the number of variants and the length of the element (**Supplementary Fig. 6**) is consistent with random mutation accumulation. However, variants were least likely to be shared among isolates if they occurred in the rRNA genes (mean: 2.5 isolates per variant) compared to *ETS+ITS* (10.8) and *NTS* (18.8) elements ($P < 2x10^{-16}$, pairwise Wilcoxon rank sum test; **Supplementary Fig. 7**). As rRNAs fold into complex structures via canonical and non-canonical base pairings, the deleterious effects of base changes most likely prevented their evolutionary persistence, whereas the *NTS1/2* elements contain many spacer sequences and thus are more permissible for variation. This difference in selective pressure was also apparent from the stratification of VFs (**Fig. 1b, c**). Analysis of VFPs separated into 3 categories based on VFs – low (< 5%), mid (from 5% to 95%), and high (≥ 95%) – revealed an uneven distribution across rDNA elements, with predominantly variants with mid- and high VFs in regulatory sequences (*NTS1/2*, *ETS1/2*, *ITS1*) and low VFs in rRNA genes (**Fig. 1b**). Nevertheless, a substantial number of variants in the *25S*, *18S*, and *5.8S* rRNA-encoding sequences had VFs above 5%, indicating regions of increased permissiveness for variation.

When compared to the S288c reference rDNA sequence, VFPs followed a U-shape distribution regardless of the isolates' ploidy (**Fig. 1c**, left, and **Supplementary Fig. 8**), with peaks at very low (below 5%) or very high (above 95%) VFs. For the *NTS* regions, the highest peak corresponded to very high VFs (≥ 99%) and reflects virtually fixed variants among rDNA copies. None of these variants were shared across all 918 isolates (**Supplementary Fig. 9**), indicating that the peak at VF ≥ 99% was not inflated by the choice of S288c as the rDNA reference. However, we did observe five variants, all in the *NTS* regions, that were shared across over 500 isolates and were responsible for 3524/6523 (54%) VFPs with VF ≥ 99%, explaining some of the skew in the distribution. In contrast to the *NTS* regions, the *ETS* and *ITS* regions ("ETS+ITS") exhibited equally distributed peaks at low and high VFs, pointing to greater selection in these elements due to their importance in rRNA processing and maturation [23]. In the rRNA-encoding regions, VFPs with low VFs dominated, presumably as a result of stringent selection.

The U-shaped distribution of VFPs held within each ecological niche, with interesting outliers (**Fig. 1c**, right). For instance, dairy isolates exhibited a greater variation of VFs in the rRNA-encoding sequences, and isolates involved in bioethanol production exhibited a peak with VFs around 50% in the *NTS* regions, perhaps reflecting recent hybridization. Both groups were large (25+ isolates) and consisted primarily of heterozygous diploid isolates (> 92%), arguing against outlier effects from unusual ploidies.

We also observed numerous variant pairs within and between rDNA elements that exhibited consistently similar VFs across multiple isolates (i.e. variants X and Y both had VFs ~ 25% in isolate A and ~ 8% in isolate B; **Supplementary Fig. 10; Supplementary File 2**). The majority

(82%) of those variants had VF ≥ 90%, which unambiguously indicated their co-occurrence in the same set of rDNA copies. Due to the short length of the sequencing reads, co-occurrence of the lower-frequency variants can be directly tested only if they are present within the same sequencing read pair. One notable example included five co-occurring SNVs within a stretch of 38 nucleotides in the *25S* that were captured in the same sequencing reads (**Supplementary Fig. 11**). Taken together, these data reveal pervasive heterogeneity in rDNA sequences within and between genomes of the same species and indicate that functional sequence constraints strongly shape the distribution in VFPs in the rDNA.

Persistence of variants

Genome heterozygosity increases over time by random mutation accumulation and outcrossing, whereas mitotic recombination and inbreeding can lead to partial or complete loss of heterozygosity. To test the characteristics of mutation accumulation in rDNA, we compared rDNA sequence diversity with the level of heterozygosity of non-rDNA sequences for 781 non-haploid isolates (see also Methods). Since most of the natural isolates were diploids while the polyploid isolates were enriched in human-related environments [22], we analyzed these two groups separately.

We found that the number of rDNA variants in each isolate correlated well with the number of heterozygous single nucleotide polymorphisms in the rest of the genome, regardless of ploidy (**Fig. 2a, Supplementary Fig. 12**). Conversely, the number of rDNA variants in diploid isolates negatively correlated with the extent of loss of heterozygosity (**Fig. 2b**). In line with this observation, fully homozygous diploid isolates exhibited lower variant accumulation in the rDNA (median: 8 variants; **Fig. 2c**) compared to heterozygous isolates (median: 20 variants; $P < 2.2\text{x}10^{-16}$, Wilcoxon rank sum test), even though both groups had virtually the same median number of rDNA copies (180 vs 184 copies; $P = 0.36$, Wilcoxon rank sum test; **Supplementary Fig. 13**). Many of the fully homozygous isolates likely resulted from diploidization events after mating-type switching and retained preexisting within-cluster diversity while homozygosing the rest of the genome. Therefore, rDNA variant accumulation correlates with the rest of the genome.

To investigate how long VFs persist across the phylogenetic tree, we determined the distribution of the topmost common variants with VF ≤ 99.5% in each element across the clades defined by the 1002 Yeast Genome Project (**Fig. 3**). In general, the same sequence changes presented

repeatedly within clades and across related clades, while other possible changes of the same nucleotides were largely undetectable. The most parsimonious explanation for these wide-spread variants is that they represent single-origin mutation events that emerged early in the isolates' evolution. For example, A817C, T2227C, and A8934G evolved as early as in the Chinese lineage, which dates back around 15,000 years ago [22]. The intragenomic frequencies of these variants fluctuated widely both within and between clades, implying that these variants are likely neutral or represent standing variation. Importantly, these data demonstrate that genetic diversity in the rDNA can be maintained over extended periods of time, arguing against rapid sequence homogenization of rDNA copies by concerted evolution.

Selection profile of the yeast ribosome

The large number of variants with varying frequencies provided an opportunity to quantitatively investigate the power of selection on the rDNA. The ribosome is one of the most ancient molecular machines that likely appeared first as a catalytic piece of RNA that subsequently grew in complexity [28]. It is comprised of a large subunit (LSU) with the 25S, 5.8S, and 5S rRNAs, and a small subunit (SSU) with the 18S rRNA. Comparative structural studies using high-resolution structures of archaeal and bacterial ribosomes showed that the most conserved sequences are localized in the innermost core of the subunits [29,30]. When analyzed in a series of 10Å-wide concentric shells, the level of rRNA conservation gradually drops from the core outward and plateaus in the outer shells, suggesting that the evolutionarily youngest rRNA elements are located closer to the solvent-exposed surfaces of the ribosome. We adapted this approach and used the rRNA nucleotide distribution data across the shells from [31] to analyze the spatial distribution of the detected variants in the *S. cerevisiae* 80S ribosome (**Fig. 4a** and **Supplementary File 3**).

In line with the evolutionary analyses, VFPs were significantly skewed toward the outer shells of the ribosome (LSU: mid VF $P = 0.009$, high VF $P = 0.008$; SSU: low VF $P = 0.026$, mid VF $P = 0.004$, high VF $P = 0.004$, Wilcoxon rank sum exact test). As expected, variants with VFs ≥ 5% were absent in shell 1 of the LSU, which contains the peptidyltransferase center, the extremely conserved catalytic core of the ribosome. Variants were similarly depleted in shell 1 of the SSU. This is the region of the highly conserved decoding center, which is essential for the fidelity of codon-anticodon interactions and controlling tRNA selection [32]. The depletion of higher-frequency variants in these shells likely reflects mutation-selection balance, whereby newly arising variants are generally highly deleterious and countered by stringent purifying

selection. On the other hand, the outermost regions largely contain rRNA expansion segments (ES), which vary in their sequence and length between eukaryotes [33–35], and thus are more permissible for mutations (**Fig. 4b**). Notably, we observed no significant difference between rRNA variants localized in binding interfaces with ribosomal proteins and variants outside of those interfaces (**Supplementary Fig. 14**), arguing that nucleotides involved in ribosomal protein contacts do not experience significantly different selective forces.

To estimate the strength of purifying selection, we analyzed variants in the highly conserved nucleotide elements (CNEs), which have virtually the same sequences (> 90% identity) across all available eukaryotic *25S*/*28S* rRNA genes [36]. This high evolutionary conservation implies that any variants observed in CNEs can be considered highly deleterious, and thus allowed us to probe for the existence of copy-number thresholds for selection, as proposed by the concerted evolution model [3]. Mid- and high frequency variants were almost absent in the CNEs, with the majority of identified CNE variants (135/147; 92%) occurring at VFs below 6% (**Fig. 4c, d; Supplementary File 3**). These low VFs suggest that mutations in the CNEs are sufficiently deleterious to be selected against even at very low rDNA copy numbers. In fact, even VFPs with VFs down to below 1% were significantly underrepresented in CNEs compared to the rest of the *25S* sequence ($P \leq 2\text{x}10^{-15}$, Fisher's exact test; **Fig. 4e**). In some cases, the depletion of variants with VF < 1% in CNEs may be due to the additive presence of other deleterious rRNA mutations in the same genome that shifts the cumulative VF to > 1% (**Supplementary Fig. 15**; median: 3%). Nevertheless, these results indicate that purifying selection acts on variants present at a frequency of as low as 1%, and thus near the level of individual rDNA copies.

Variants in functional regions

A total of 12 CNE variants were observed at higher VFs (≥ 6%) (**Fig. 4d**). Strikingly, eight of them localized along a single loop of the GTPase association center (GAC; **Fig. 5a**), a highly conserved ~ 57 nucleotide-long system of hairpins located in domain II of the 25S rRNA that interacts with translational GTPases during protein synthesis [37]. The eight variants were distributed among nine isolates at various frequencies (**Fig. 5a, b**). Closely related isolates (CFH, CBR, and CFG) shared the same variant C1248U (25S rRNA nomenclature; VF 20-33%), suggesting a single mutational origin, while the same position also exhibited a different variant sequence in another isolate (C1248A). Intriguingly, two variants, C1248U and G1237A, were found in the same isolate (CHF), both at 28% frequency. We retrieved sequence reads spanning both positions in the GAC region of the isolate and found that the variants were present in mutually exclusive gene copies, therefore contributing to three distinct forms of the 25S rRNA gene in CFH (**Fig. 5c**). Secondary structure analysis showed that all but one of the eight variant positions are involved in base-pairing interactions (**Fig. 5d**). For example, U1248 could create a Watson-Crick pair with A1240, thereby extending the stem of the hairpin, whereas C1275U and C1280U could destabilize stem structures by altering the corresponding Watson-Crick base pairs with G1266 and G1229. Since the GAC undergoes major rearrangements upon GTP hydrolysis [38], these changes could potentially alter the overall geometry of the region and thus affect the interaction with translational GTPases. The relatively high VFs (6-33%) in such a highly conserved region, together with their ability to alter base pairing, suggests that these variants are non-deleterious and are either under neutral or adaptive selection.

To search for additional mid- or high frequency variants of possible functional importance, we focused on the intersubunit bridges, which link large and small subunits and mediate major

rearrangements in the ribosome during translation [39,40]. We detected 26 variants with mid- and high VFs within 5Å of intersubunit bridges (**Supplementary Fig. 16; Supplementary File 3**). Among these, variant G1112A in the 18S rRNA (18S rRNA nomenclature; 44.8% VF in isolate CDN) is part of helix 27, which dynamically interacts with bridge eB14 during different steps of translation (**Supplementary Fig. 17**) and is located in the close proximity to the decoding center. Intriguingly, the same G1112A mutation was previously shown in mutagenesis studies to decrease readthrough of stop codons in *S. cerevisiae* [41], thus indicating that this naturally occurring variant can alter cellular fitness and may be under selection.

Discussion:

Our analyses demonstrate remarkable sequence diversity among rDNA copies both within and among isolates. Variants can persist and be subject to selection over a wide range of frequencies. These data argue that quick homogenizing sweeps across the rDNA, as proposed by the concerted evolution model, are likely rare. The varying frequencies of within-genome rDNA polymorphisms likely arise by random duplication or deletion of rDNA copies, as well as gene conversion, and we suggest that these frequency changes are under constant selection. In this model, deleterious variants can be eliminated (and beneficial variants expanded) as long as they provide sufficient phenotypic consequences, whereas neutral variants can persist and fluctuate in frequency among rDNA copies for extended periods of evolutionary time. The substantial within-genome polymorphisms seen in this study and observed across diverse phyla [7–17] is consistent with, and fully expected, in this model.

We note that the extreme purifying selection against deleterious variants observed here seems at odds with a much higher selection threshold seen in *Drosophila*, where up to 60% of 28S rRNA coding sequences are disrupted by R1/R2 retrotransposons [42]. We suggest that these differences can be explained by how the respective mutations contribute to ribosomes biogenesis and function. The majority of coding variants observed here are SNVs that are likely incorporated into ribosomes. The presence of non-functional ribosomes could cause partially dominant effects and thus be sufficiently deleterious to be selected against even at low copy number. Such poisonous effects could explain why defective ribosomes are subject to quality control mechanisms and degraded even if the deleterious rRNA variants are expressed at very low levels [43,44]. We expect that variants in *ETS* and *ITS* that yield defective ribosomal particles may have similar effects [45]. By contrast, mutations that disrupt rRNA transcription altogether, as seen in

Drosophila, are likely subject to much weaker selection because the high rDNA copy number, as well as regulatory mechanisms that activate dormant rRNA genes or upregulate rRNA transcription [46], create a substantial buffer against changes in rRNA expression levels.

How the observed heterogeneity in rDNA sequences might affect cellular phenotypes will require further functional analyses. rRNA variants exhibit tissue-specific expression in humans, mice, and *Arabidopsis*, and are incorporated into actively translating ribosomes [14,17]. However, the effects of these variants on translation remain for the most part to be determined. We uncovered one example of a natural variant (18S:G1112A, VF 44.8%) that shows increased translational fidelity in lab experiments [41], although the added presence of wild-type sequences in the same isolate may buffer any phenotypic effects. Nevertheless, our analyses in yeast, along with studies in *Arabidopsis* [14] and mammals [17], indicate that ribosome heterogeneity is not limited to ribosomal proteins but also involves a diversity of rRNAs that is likely much more pervasive than previously thought. Sequencing more individuals of a species could thus reveal substantial hidden variation.

Finally, our results highlight the importance of considering within-species sequence diversity of rDNA sequences in metagenomic analyses, which frequently use short rDNA segments to estimate species numbers and population complexity in environmental or clinical samples [47]. The extensive diversity seen here within a single species and even within single isolates shows that care must be taken when using exact sequence variants or narrowly defined operational taxonomic units for estimating species numbers, as these approaches may lead to substantial overestimates of true species counts.

Methods:

Read alignment and variant call. Short-read sequencing reads for each isolate (SRA: ERP014555) [22] were used. To remove non-rDNA reads and reduce the rate of false-positive calls downstream, reads of a given isolate were first mapped to an SK1 genome [48] without annotated rDNA using bowtie2 [49]: -5 1 -N 0 -p 8. Unmapped reads were mapped once again to the genome with relaxed settings: -5 1 -N 1 -p 8. The -L parameter was set to 22 by default, and the high divergence between nuclear and mitochondrial rDNA allowed the aligner to discriminate between nuclear and mitochondrial rDNA reads. The resulting nuclear rDNA-enriched reads were mapped with relaxed settings to the annotated copy of the S288c rDNA reference sequence (*Saccharomyces* genome database: Chr XII 451575:460711) whose *NTS* sequences were put before the *35S rRNA* instead of after *ETS1* (see Data availability for the reference .fsa and .bed files). The parameters for bowtie2 were: -5 1 -N 1 -p 8. The output .sam files were converted into .bam files with samtools view -Sbh -F 12 file.sam | samtools sort -@ 8 -o file.sort.bam -O 'bam' file.bam [50]. Variants were subsequently called using LoFreq [27]. Indel qualities were added to the .sort.bam files with indelqual --dindel, and variants were called afterward (call --call-indels). The output .vcf files were filtered using awk. Variant entries were kept if: nucleotide position > 10 and < 9100, VF > 0.004, homopolymer run < 4, and if GC content was < 0.6 in long indels (> 5 nt) in either reference or alternative sequences. Additionally, variants with the highest strand bias (≥ 144; top 12%) were removed to further reduce the rate of false positives [51], and replicates of the same isolate were deduplicated such that only the sample with the highest read coverage was considered downstream. For **Fig. 5c** and **Supplementary Fig. 11b**, rDNA reads that fully covered the region of the variants were retrieved from the alignment and visualized in IGV [52].

Pipeline performance. To determine the lowest VFs that can be confidently detected with this pipeline, we performed a series of *in silico* titrations by "spiking" SK1 reads into S288c reads at different frequencies from 100% to 0.1%. First, paired-end Illumina HiSeq 2500 reads from SK1 and S288c (SRA: SRR4074258, SRR4074255) from [48] were mapped using the pipeline, and 10 variants unique to SK1 (not detected in S288c) that have VFs over 95% were tracked (**Supplementary Fig. 1**). The majority of the variants were detected even at the titration point of 0.5%, which corresponds to a variant in as low as 1 rDNA copy (**Supplementary Fig. 2**). To assess the sensitivity and specificity of the pipeline, we simulated artificial datasets of 56 known SNVs (real data: median is 22 SNVs, range is 7-80 SNVs) and 3 indels (real data: median is 3 indels, range is 1-15 indels) at 0.5% VF distributed randomly across the rDNA prototype using NEAT-genReads [53] , with genomic coverages from 20-fold to 400-fold (3,000 – 60,000-fold for rDNA; real data genome coverage mean is 200-fold) (**Supplementary Fig. 1b-d**). The sequencing error model (genSeqErrorModel.py) as well as G/C content (computeGC.py) and fragment length (computeFraglen.py) models were inferred from the SK1 reads (SRR4074255), and the read length (the -R parameter) was set to 150. To better recapitulate the real rDNA coverage (**Supplementary Fig. 1b**), we modified annotations in the .bed file (see Data availability). Reads with "reference" and "alternative" sequences (SNVs and indels) were generated separately and then combined such that the frequency of variants (alternative sequences) was 0.5%. For example, for 3,000-fold rDNA coverage, reference reads we ran : genReads.py -r rDNA_S288c.fsa -R 150 -o output_file_prefix --bam --pe-model fraglen.p -e seq_error.p --gc-model gcmodel.p -p 1 -M 0 -c 2985 -t rDNA_S288c_benchmark.bed -to 0.4 --rng 456; for alternative reads: python neat-genreads-master/genReads.py -r rDNA_S288c.fsa -R 150 -o output_file_prefix --bam --vcf --pe-model fraglen.p -e seq_error.p --gc-model gcmodel.p -

p 1 -M 0.015 -c 15 -t rDNA_S288c_benchmark.bed -to 0.4 --rng 123. The variants and error model were the same in all the simulations and the coverage patterns best resembled the real data. The datasets were then processed with the bioinformatics pipeline, and the true known prior variants (generated by NEAT-genReads) were compared to those detected by the pipeline (**Supplementary Fig. 1c,d**). We did not detect any false positive calls in the simulated datasets, which is in accordance with the extremely high specificity of the LoFreq caller (False positive rate < 0.2% [27]). Given the actual complexity of the real data, we applied additional filters (see "Read alignment and variant call") to reduce the rate of false positive variants with likely increased rate of false negative calls.

rDNA copy number estimation. The mean of *RDN5.8*, *RDN18* and *RDN25* copy number was calculated from their annotated values and multiplied by the respective ploidies (data from [22]).

Variant cooccurrence. We estimated pairwise variant cooccurrence by calculating Euclidean distances *D* between VFs within each isolate: $D_{XY} = \| (X_A, X_B, \ldots X_i) - (Y_A, Y_B, \ldots Y_i) \| = [(X_A - Y_A)^2 + (X_B - Y_B)^2 + \ldots + (X_i - Y_i)^2]^{1/2}$, where X and Y are variants; A, B, … i are isolates; X_A, X_B, … X_i are VFs (from 0.05 to 1, which corresponds to 5-100%). For each possible pair of variants, only isolates that have both X and Y were considered. Variants X and Y were considered to occur in the same rDNA copies if: 1) $n \geq 3$ isolates within an ecological niche shared the same pair of variants; 2) $D_{XY} < 0.5$; and 3) the mean of their VFs $(X_A + \ldots + X_i + Y_A + \ldots + Y_i)/n$ was < 0.05. This allowed considering variant pairs that have different VFs in different isolates but similar VFs within isolates, i.e. variants X and Y both have VFs 70% in isolate A, VF 50% in isolate B and 7% in isolate C, so they always have similar VFs with an isolate, but different magnitudes in different isolates.

Correlation between rDNA variants and non-rDNA SNPs. Variants with VF [0.5%;99.5%] were considered in this analysis since ~ 0.5% correspond to one rDNA copy. Haploid isolates were excluded from the analysis, and diploid (2n; N = 693) and polyploid (3n-5n; N = 88) isolates were analyzed separately. Four isolates (AEC, AFT, CEN and CFS) were additionally removed from the datasets due to undefined values in ploidy or genome heterozygosity. After that, the number of rDNA variants per isolate (both SNVs and indels) and the respective genome-wide (excluding rDNA) number of heterozygous SNPs per kb of heterozygous regions or the value for loss of heterozygosity (from [22]) were used to compute Pearson's correlation coefficient. For homozygous diploid isolates (those with zero heterozygous SNPs), the distribution of rDNA variants was plotted in **Fig. 2c**.

Distribution of variants across clades. Topmost abundant rDNA variants in each element with VF ≤ 99.5% were considered in this analysis. Since the choice of reference variants was biased because of the use of the S288c rDNA sequence as the reference genome, an ancestral sequence was chosen based on the VF (high) and presence in clades. For example, if all clades contained the variant at high VFs that would mean this variant was an ancestral one and an alternative sequence derived later. Clade data was taken from [22] .

rRNA analysis. For the expected distribution of VFPs in **Fig. 4a**, the fraction of nucleotides in each shell F_i (i indicates shells 1, 2, 3, 4, 5, and 6+7 for the LSU and 6+7+8 for the SSU) was calculated as the number of nucleotides n_i in a shell (from [31] , "Onion" representation) over total number of nucleotides N in a subunit: $F_i = n_i/N$. For the LSU (*25S* + *5.8S* + *5S*), N = 3428 instead of the actual 3675 (S288c reference) because 247 nucleotides were missing from the crystal structure and thus could not be used in the shell annotation (since it relies on the 3D structure of the ribosome). For the SSU (*18S*), N = 1781 (19 nucleotides were missing from the

structure). For the observed distribution of VFPs, F_i* was calculated by dividing the number of VFPs with a chosen range of VFs (low VF - [0.4%;5%); mid VF - [5%,95%), high VF - [95%;100%]) m_i in a shell by the total number of VFPs M in the chosen range (low, mid- or high VFs): $F_i^* = m_i/M$. Due to the missing segments in the LSU and SSU structures, some VFPs detected in **Fig. 1** were not included in the analysis. The excluded number of VFPs in the LSU: low VF - 62 out of 1322; mid VF - 60 out of 527; high VF - 69 out of 241. The excluded number of VFPs in the SSU: low VF - 15 out of 498; mid VF - 3 out of 224; high VF - 2 out of 36. We then calculated the ratio between the two values: $\log_{10}(F_i^*/ F_i)$ and used it as a metric for constrained (< 0) or relaxed (> 0) sequence variability in the ribosome given a significant difference between the expected and observed distributions (Wilcoxon rank sum test; see text). For example, for shell 1 in LSU, F_1 = 82 / 3428 = 0.024, F_1* = 11/1260 = 0.009, and $\log_{10}(F_1^*/ F_1)$ = -0.43. For **Fig. 4b** and **Supplementary Fig. 14**, coordinates of ESs and rRNA:ribosomal protein interactions for were retrieved from [54] . The expected distribution of VFPs in ESs for **Fig. 4b** was derived from the number of nucleotides in ESs (n; LSU: 883, SSU: 390) versus not in ESs (N; LSU: 2,792; SSU:1,410) - "ES": n/(n+N), "no ES": N/(n+N). For the observed values, VFPs were counted for each range (*i* = low, mid-, or high VFs) within (m_i) or outside (M_i) of ESs, and the fractions ("ES": $m_i/(m_i+M_i)$, "no ES": $M_i/(m_i+M_i)$) were calculated. For **Supplementary Fig. 14,** The expected distribution of VFPs in the rRNA alone or rRNA:protein interaction interfaces and the observed values were calculated in the same manner. For **Fig. 4c-d**, coordinates of eukaryotic-specific conserved nucleotide elements (eCNEs; "CNEs" in this paper) were retrieved from [36] . Similarly to **Fig. 4b**, the expected distribution of VFPs in CNEs was derived as the fraction of *25S* nucleotides in CNE (n = 962) and in the rest of *25S* (N = 2,434; total *25S* length – 3,396) - "CNE": n/(n+N), "no CNE": N/(n+N). For the observed values, VFPs

were counted for each range (i) indicated in **Fig. 4e** within (m_i) and outside (M_i) of CNE, and the fractions ("CNE": $m_i/(m_i+M_i)$, "no CNE": $M_i/(m_i+M_i)$) were calculated.

Structural analysis. Ribosome structures PDB_ID: 4v88 (assembly 1) [54] and PDB_ID: 6woo [55] were visualized and analyzed in PyMOL. Since ribosomal parts undergo major rearrangements during translation [39,40], we considered variants that are directly in or within 5 Å of the intersubunit bridges as potential candidates located in the functionally important regions. Coordinates of the intersubunit bridges were extracted from [54]. Analysis in **Supplementary Fig. 17** was made with local structural alignment of the region (18S:1110-1136 together with L41e) from 4v88 and 6woo in PyMOL. For **Fig. 5d** and **Supplementary Fig. 11c**, the NDB (*http://ndbserver.rutgers.edu/*) [56] was used to retrieve nucleotide interactions in the 25S rRNA (PDB_ID: 4v88).

Code and data availability. The computer scripts used for processing and analyzing rDNA sequencing reads are available at https://github.com/hochwagenlab. **Supplementary File 1** is a tab-delimited table of filtered rDNA variants with their frequencies across isolates. **Supplementary File 2** is a tab-delimited table of variant pairs with consistently similar frequencies. **Supplementary File 3** is a tab-delimited table with annotations of variants in rRNA genes. Other files can be found at https://github.com/hochwagenlab: 1) raw_rDNA_var_calls - raw rDNA .vcf files prior to additional filtering; 2) rDNA_S288c.fsa - the S288c rDNA copy prototype sequence used in this study (.fasta format); 3) rDNA_S288c.bed - annotations associated with the rDNA prototype (.bed format); 4) rDNA_S288c_benchmark.bed - annotations for benchmarking.

Acknowledgments:

We thank M. Rockman, D. Gresham, M. Siegal, L. Williams for helpful discussions. This work was supported by the National Institutes of Health [GM111715 and GM123035 to AH].

Author contributions:

Conceptualization, D.S. and A.H.; Investigation and formal analysis, D.S. and A.H.; Coding, D.S.; Writing – Original Draft, Review & Editing, D.S., and A.H.

Competing interests: The authors declare no competing interests.

Figures:

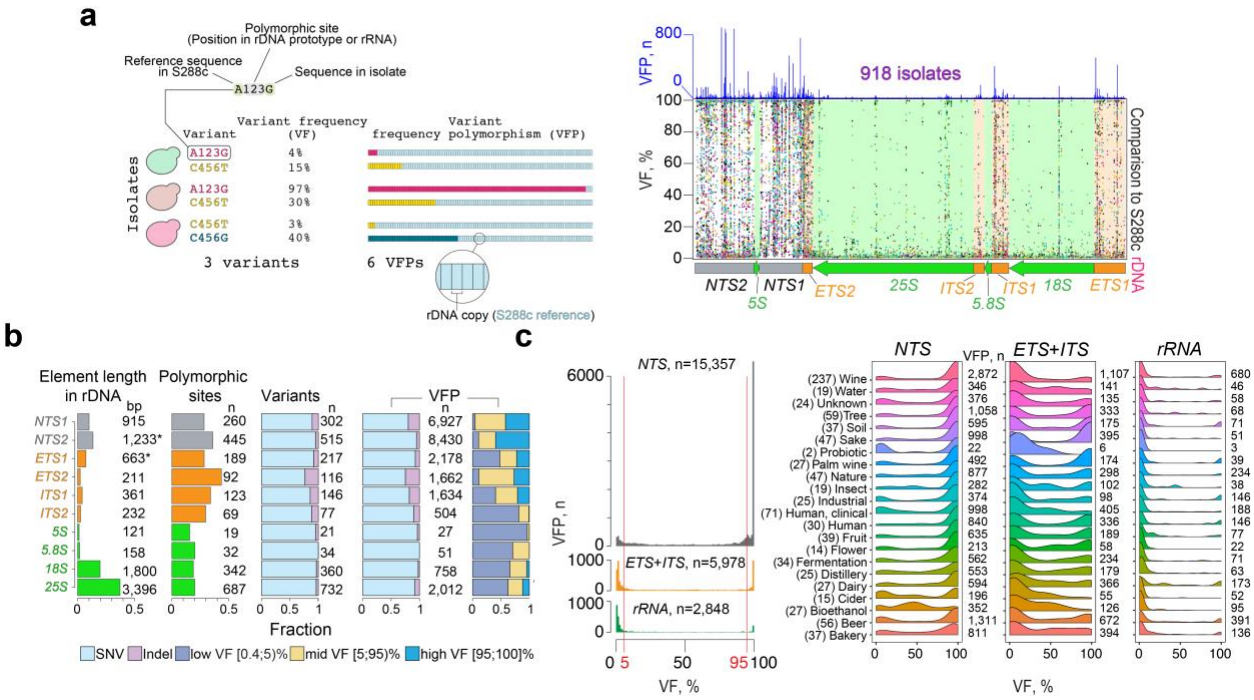

Fig. 1. rDNA sequence variants across 918 *S. cerevisiae* isolates. a, *left* - variant annotation. Variants are defined as nucleotide differences from the from S288c prototype (light blue). Variant frequency (VF) is the percentage of rDNA reads with a particular variant. Variant frequency polymorphism (VFP) represents a polymorphism of rDNA copy number containing a variant, which can differ between isolates. Note that one isolate can have multiple variants and thus the sum of the VFPs can be more than the total number of the sequenced isolates. *Right* - VFs compared to the reference S288c rDNA prototype are plotted against one rDNA copy. Colors reflect different isolates, but some isolates have the same or similar colors because of palette constraints. A schematic of major genic elements within an rDNA copy is shown under the plot. The histogram above shows the number of VFPs at a specific position across all isolates. **b**, summary statistics for genic elements in the rDNA. Element length in rDNA - the relative length each element occupies in an rDNA; Polymorphic sites – the fraction of all

variable sites in an element; Variants – number of observed variants and fraction of single-nucleotide variants (SNVs) and indels; VFP – number of VFPs across all isolates and fraction SNVs and indels. Numbers indicate absolute values for each distribution. The adjusted length for *NTS2* (real: 1,243 bp) and *ETS1* (real: 700 bp) due to filtering pipeline is shown with asterisks. **c**, (*left*) distribution of VFPs by VF in NTS elements, ETS+ITS elements and rRNA sequences; (*right*) same analysis by separated by ecological niche. Numbers of sequenced individuals from each ecological niche are indicated in parentheses, and number of VFPs are shown on the right of each distribution.

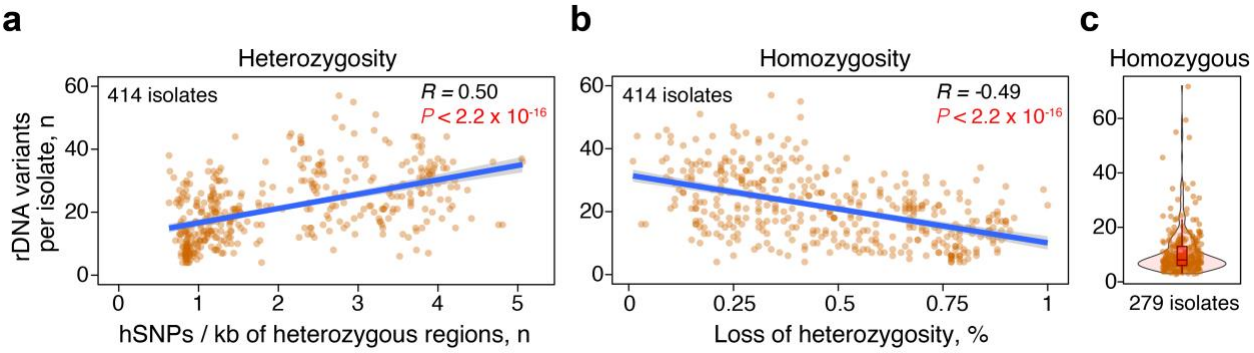

Fig. 2. Relationship between sequence diversity of the rDNA and the rest of the genome in diploid isolates. Correlation between the number of rDNA variants and (**a**) the number of heterozygous SNPs per heterozygous region of the genome (excluding rDNA) [22] or (**b**) the loss of heterozygosity per heterozygous isolate. Pearson correlation coefficient *R* is shown. Each dot represents an isolate, regression line with a 95% confidence interval is shown in blue. **c**, the number of rDNA variants per homozygous isolate (violin plot; red boxplot inside shows the median).

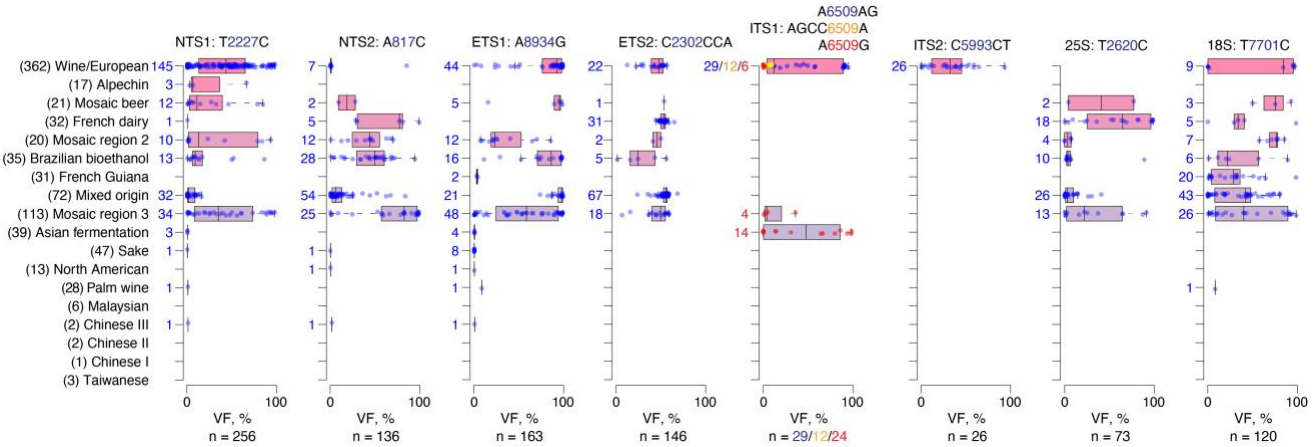

Fig. 3. Boxplot distribution of the topmost abundant variants across clades. The clade information was taken from [22]. Ancestral and derived sequences were determined based on their distribution across clades (i.e.NTS1:2227C is a sequence in the reference rDNA sequence but is shown as the derived T2227C variant here). Numbers in parentheses indicate the total number of sequenced isolates; the number of isolates with a variant per clade is shown in blue, orange, or red. The number of VFPs of a specific variant (n) across isolates is shown below each plot.

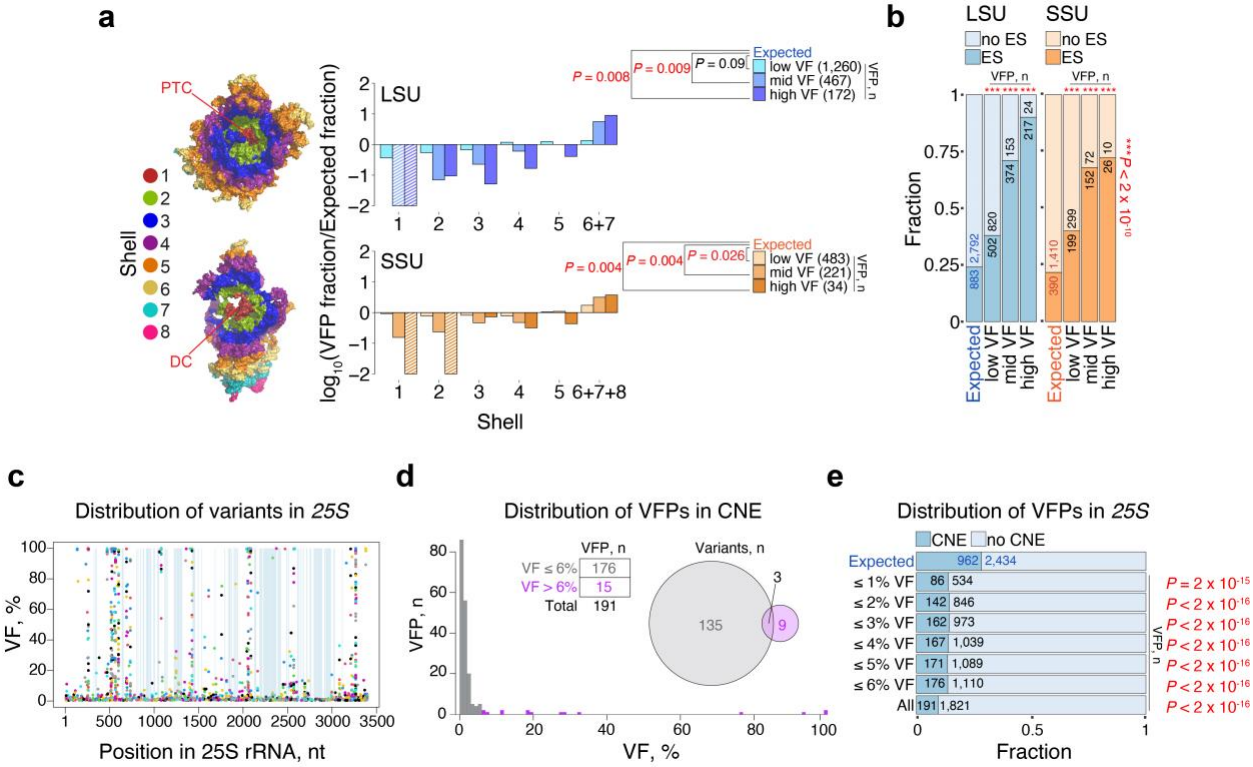

Fig. 4. Selection in the yeast 80S ribosome. a, distribution of the number of low, medium and high VFPs observed (numbers are indicated in brackets) versus expected in both LSU and SSU across their respective concentric shells (see text; depicted to the *left*). The outermost shells in LSU and SSU were pooled together due to the low number of nucleotides. Shells 1 in LSU and 1 and 2 in SSU did not have any variants and their log values were artificially set to -2 (hashed bars). PTC - peptidyltransferase center, DC - decoding center. **b**, distribution of VFPs between expansion segments (ES) and other parts of the rRNA (no ES). **c**, distribution of variants across eukaryotic conserved nucleotide elements (CNEs, highlighted in blue). **d,** distribution of VFPs that are detected in the CNEs. *Inset* – number of VFPs in two groups (below and above 6%; *left*) and the overlap between the variants from the two groups (*right*; three variants have VFPs with VFs both below and above 6% across different isolates). **e**, distribution of VFPs with different VF cutoffs within and outside of the CNEs. For **a**, *P*-values were calculated with Wilcoxon rank sum exact test, for **b** and **e** - with Fisher's exact test. "Expected" distribution was calculated

based on the distribution of total rRNA nucleotides across the shells (**a**) or within and outside of ES (**b**) or CNE (**e**; see Methods).

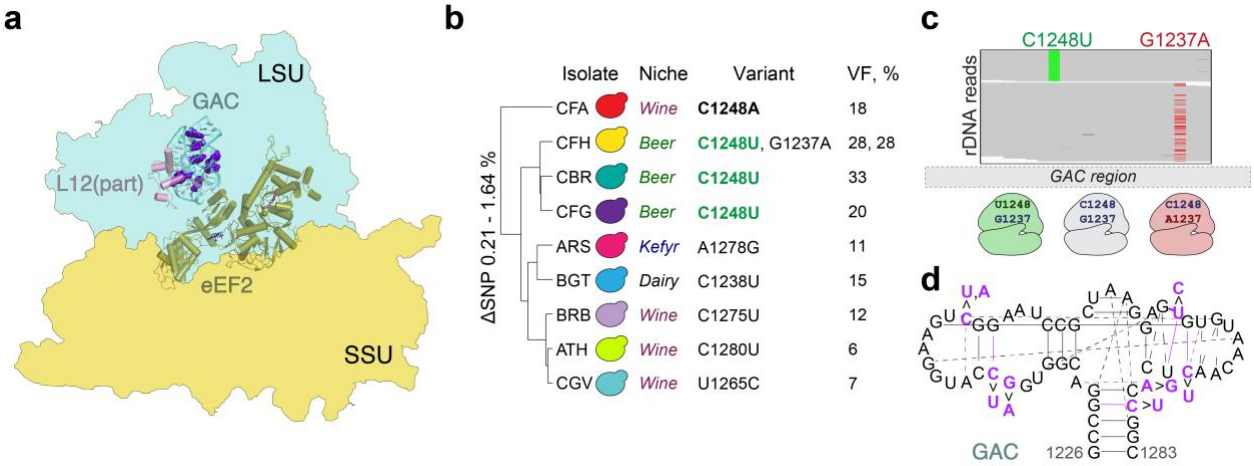

Fig. 5. Variability in the GTPase activation center. a, the GAC (blue) interacts with translational GTPase eEF2 (green) in the 80S yeast ribosome (PDB_ID: 6woo; LSU outline is shown in blue, SSU outline is in yellow). Part of ribosomal protein L12 is shown in pink. Variant positions are shown in purple with the spherical representation. **b**, hierarchical clustering of the isolates with SNVs in the GAC (25S nomenclature) based on genome-wide SNP differences (excluding rDNA variants) from [22]. Bold highlights a position with more than one variant; green indicates a shared variant. **c**, continuous rDNA reads of isolate CFH (horizontal grey lines) mapped to the GAC. Variant nucleotides are indicated in green and red, with 3 possible variant combinations in the 25S rRNA (LSU of the ribosome) shown below. **d**, secondary structure of the GAC. Variant sites and their interactions are denoted in purple, interactions are shown in lines and weaker (or probable) interactions are shown with dashed lines.

References:

1. Long, E. O. & Dawid, I. B. Repeated Genes in Eukaryotes. *Annual Review of Biochemistry* **49**, 727–764 (1980).
2. Dover, G. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).
3. Ganley, A. R. D. & Kobayashi, T. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**, 184–191 (2007).
4. Ganley, A. R. D. & Kobayashi, T. Monitoring the Rate and Dynamics of Concerted Evolution in the Ribosomal DNA Repeats of Saccharomyces cerevisiae Using Experimental Evolution. *Molecular Biology and Evolution* **28**, 2883–2891 (2011).
5. Szostak, J. W. & Wu, R. Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature* **284**, 426–430 (1980).
6. McTaggart, S. J., Dudycha, J. L., Omilian, A. & Crease, T. J. Rates of Recombination in the Ribosomal DNA of Apomictically Propagated *Daphnia obtusa* Lines. *Genetics* **175**, 311–320 (2007).
7. Pawlowska, T. E. & Taylor, J. W. Organization of genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature* **427**, 733–737 (2004).
8. Simon, U. K. & Weiss, M. Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thought. *Molecular Biology and Evolution* **25**, 2251–2254 (2008).
9. West, C., James, S. A., Davey, R. P., Dicks, J. & Roberts, I. N. Ribosomal DNA Sequence Heterogeneity Reflects Intraspecies Phylogenies and Predicts Genome Structure in Two Contrasting Yeast Species. *Systematic Biology* **63**, 543–554 (2014).
10. McTaggart, S. J. & Crease, T. J. Selection on the Structural Stability of a Ribosomal RNA Expansion Segment in Daphnia obtusa. *Molecular Biology and Evolution* **22**, 1309–1319 (2005).
11. Parkin, E. J. Within- and Between-Individual Sequence Variation Among ITS1 Copies in the Meadow Grasshopper Chorthippus parallelus Indicates Frequent Intrachromosomal Gene Conversion. *Molecular Biology and Evolution* **21**, 1595–1601 (2004).
12. Keller, I., Chintauan-Marquier, I. C., Veltsos, P. & Nichols, R. A. Ribosomal DNA in the Grasshopper *Podisma pedestris* : Escape From Concerted Evolution. *Genetics* **174**, 863–874 (2006).
13. Hugall, A., Stanton, J. & Moritz, C. Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic Meloidogyne. *Molecular Biology and Evolution* **16**, 157–164 (1999).
14. Sims, J., Sestini, G., Elgert, C., von Haeseler, A. & Schlögelhofer, P. Sequencing of the Arabidopsis NOR2 reveals its distinct organization and tissue-specific rRNA ribosomal variants. *Nature Communications* **12**, 387 (2021).
15. Holland, M. L. *et al.* Early-life nutrition modulates the epigenetic state of specific rDNA genetic variants in mice. *Science* **353**, 495–498 (2016).
16. Tseng, H. *et al.* Mouse Ribosomal RNA Genes Contain Multiple Differentially Regulated Variants. *PLoS ONE* **3**, e1843 (2008).
17. Parks, M. M. *et al.* Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Science Advances* **4**, (2018).
18. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics* **39**, 121–152 (2005).

19. Nei, M., Gu, X. & Sitnikova, T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences* **94**, 7799–7806 (1997).
20. Nei, M., Rogozin, I. B. & Piontkivska, H. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences* **97**, 10866–10871 (2000).
21. Piontkivska, H., Rooney, A. P. & Nei, M. Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. *Molecular Biology and Evolution* **19**, 689–697 (2002).
22. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
23. Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A. & Gleizes, P.-E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA* **6**, 225–242 (2015).
24. Miller, C. A. & Kowalski, D. cis-acting components in the replication origin from ribosomal DNA of Saccharomyces cerevisiae. *Molecular and Cellular Biology* **13**, 5360–5369 (1993).
25. Wai, H. H. Complete deletion of yeast chromosomal rDNA repeats and integration of a new rDNA repeat: use of rDNA deletion strains for functional analysis of rDNA promoter elements in vivo. *Nucleic Acids Research* **28**, 3524–3534 (2000).
26. Kobayashi, T., Hidaka, M., Nishizawa, M. & Horiuchi, T. Identification of a site required for DNA replication fork blocking activity in the rRNA gene cluster in Saccharomyces cerevisiae. *Molecular and General Genetics MGG* **233**, 355–362 (1992).
27. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40**, 11189–11201 (2012).
28. Fox, G. E. Origin and Evolution of the Ribosome. *Cold Spring Harbor Perspectives in Biology* **2**, a003483–a003483 (2010).
29. Hsiao, C., Mohan, S., Kalahar, B. K. & Williams, L. D. Peeling the Onion: Ribosomes Are Ancient Molecular Fossils. *Molecular Biology and Evolution* **26**, 2415–2425 (2009).
30. Wuyts, J., van de Peer, Y. & de Wachter, R. Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Research* **29**, 5017–5028 (2001).
31. Bernier, C. R. *et al.* RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.* **169**, 195–207 (2014).
32. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annual Review of Biochemistry* **74**, 129–177 (2005).
33. Petrov, A. S. *et al.* Evolution of the ribosome at atomic resolution. *Proceedings of the National Academy of Sciences* **111**, 10251–10256 (2014).
34. Parker, M. S., Sallee, F. R., Park, E. A. & Parker, S. L. Homoiterons and expansion in ribosomal RNAs. *FEBS Open Bio* **5**, 864–876 (2015).
35. Yokoyama, T. & Suzuki, T. Ribosomal RNAs are tolerant toward genetic insertions: evolutionary origin of the expansion segments. *Nucleic Acids Research* **36**, 3539–3551 (2008).
36. Doris, S. M. *et al.* Universal and domain-specific sequences in 23S–28S ribosomal RNA identified by computational phylogenetics. *RNA* **21**, 1719–1730 (2015).
37. Taylor, D. J. *et al.* Structures of modified eEF2·80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *The EMBO Journal* **26**, 2421–2431 (2007).

38. Gulay, S. P. *et al.* Tracking fluctuation hotspots on the yeast ribosome through the elongation cycle. *Nucleic Acids Research* **45**, 4958–4971 (2017).
39. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology* **19**, 560–567 (2012).
40. Behrmann, E. *et al.* Structural Snapshots of Actively Translating Human Ribosomes. *Cell* **161**, 845–857 (2015).
41. Velichutina, I. v. *et al.* Mutations in helix 27 of the yeast Saccharomyces cerevisiae 18S rRNA affect the function of the decoding center of the ribosome. *RNA* **6**, S1355838200000637 (2000).
42. Averbeck, K. T. & Eickbush, T. H. Monitoring the Mode and Tempo of Concerted Evolution in the *Drosophila melanogaster* rDNA Locus. *Genetics* **171**, 1837–1846 (2005).
43. LaRiviere, F. J., Cole, S. E., Ferullo, D. J. & Moore, M. J. A Late-Acting Quality Control Process for Mature Eukaryotic rRNAs. *Molecular Cell* **24**, 619–626 (2006).
44. Cole, S. E., LaRiviere, F. J., Merrikh, C. N. & Moore, M. J. A Convergence of rRNA and mRNA Quality Control Pathways Revealed by Mechanistic Analysis of Nonfunctional rRNA Decay. *Molecular Cell* **34**, 440–450 (2009).
45. Hage, A. el & Tollervey, D. A Surfeit of Factors: Why is Ribosome Assembly So Much More Complicated in Eukaryotes than Bacteria? *RNA Biology* **1**, 9–14 (2004).
46. French, S. L., Osheim, Y. N., Cioci, F., Nomura, M. & Beyer, A. L. In Exponentially Growing *Saccharomyces cerevisiae* Cells, rRNA Synthesis Is Determined by the Summed RNA Polymerase I Loading Rate Rather than by the Number of Active Genes. *Molecular and Cellular Biology* **23**, 1558–1568 (2003).
47. Porter, T. M. & Hajibabaei, M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* **27**, 313–338 (2018).
48. Yue, J.-X. *et al.* Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics* **49**, 913–924 (2017).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 666 (2012).
52. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
53. Stephens, Z. D. *et al.* Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLOS ONE* **11**, e0167047 (2016).
54. Ben-Shem, A. *et al.* The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. *Science* **334**, 1524–1529 (2011).
55. Wang, J. *et al.* Structural basis for the transition from translation initiation to elongation by an 80S-eIF5B complex. *Nature Communications* **11**, 5003 (2020).
56. Coimbatore Narayanan, B. *et al.* The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research* **42**, D114–D122 (2014).

Supplementary Figures

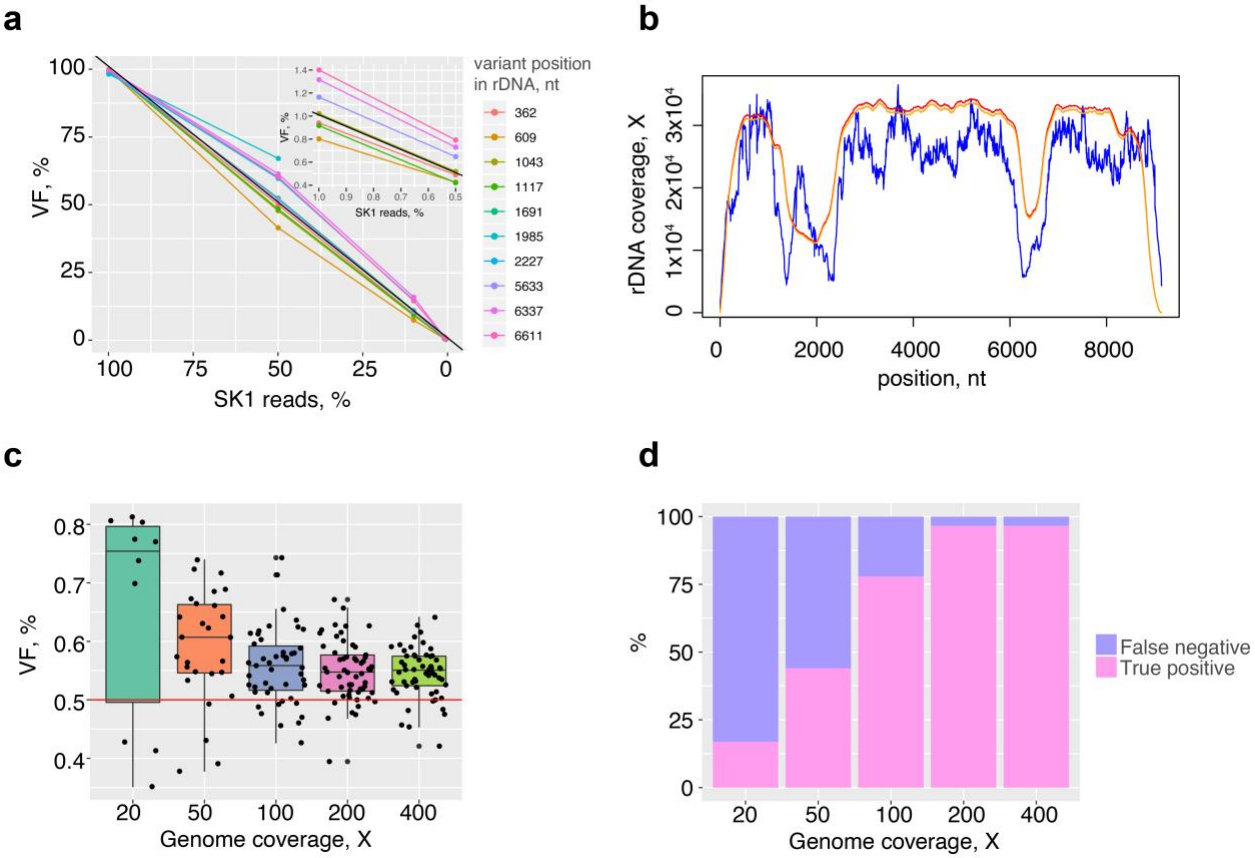

Fig. S1. Pipeline benchmarking. a, *in silico* spiking of S288c read data with 10 highly confident SNVs from another lab strain (SK1) across an rDNA copy (positions are shown in the legend). The black diagonal represents true VF values. *inset* - VFs from 1 to 0.5%. **b**, 200X genomic read coverage across the rDNA copy from S288c experimental data (blue), for a simulated dataset with true priors before (red) or after (orange) running the bioinformatics pipeline. **c**, pipeline performance on simulated datasets of different genomic coverage (20X-400X) but with the same variant alleles (all VF = 0.5%, red line; dots are true positive variants TPs). **d**, pipeline specificity as a function of coverage. No false positives were detected (see Methods).

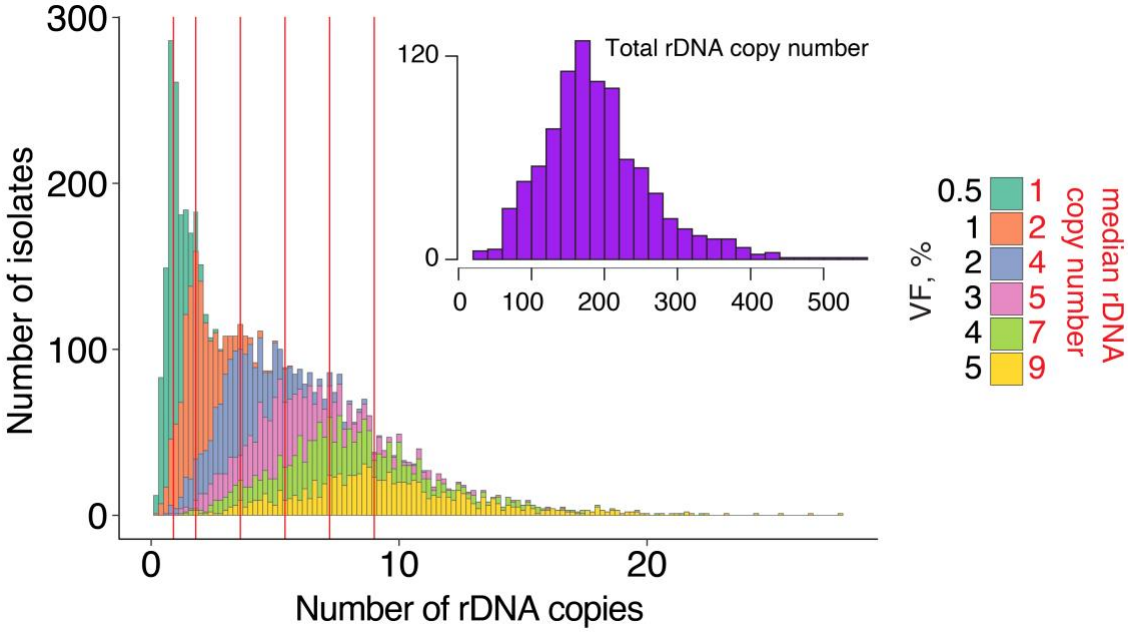

Fig. S2. Relationship between VF and the total number of rDNA copies containing that variant in a given isolate. Profiles show how many copies a VF of 0.5%, 1%, 2%, 3%, 4% or 5% would correspond to based on the number of total rDNA copies across all isolates (inset). The median for each distribution is shown as the red line. Red letters indicate the median copy number that a given VF corresponds to (e.g. a VF of 0.5% corresponds roughly to a copy number of 1 in this data set).

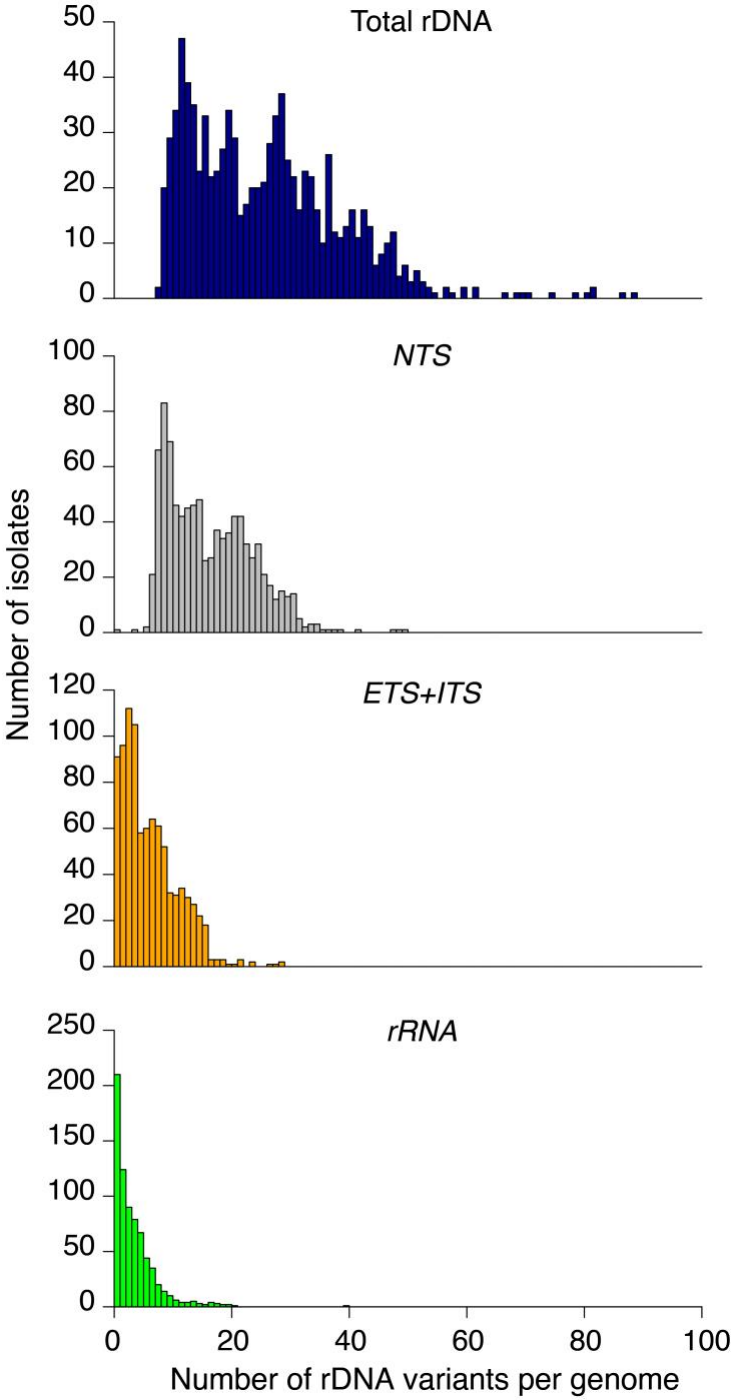

Fig. S3. Distribution of the number of rDNA variants per isolate. NTS – *NTS1* and *NTS2*; ETS+ITS – *ETS1*, *ETS2*, *ITS1* and *ITS2*; rRNA – rRNA genes (*5S*, *5.8S*, *18S* and *25S*).

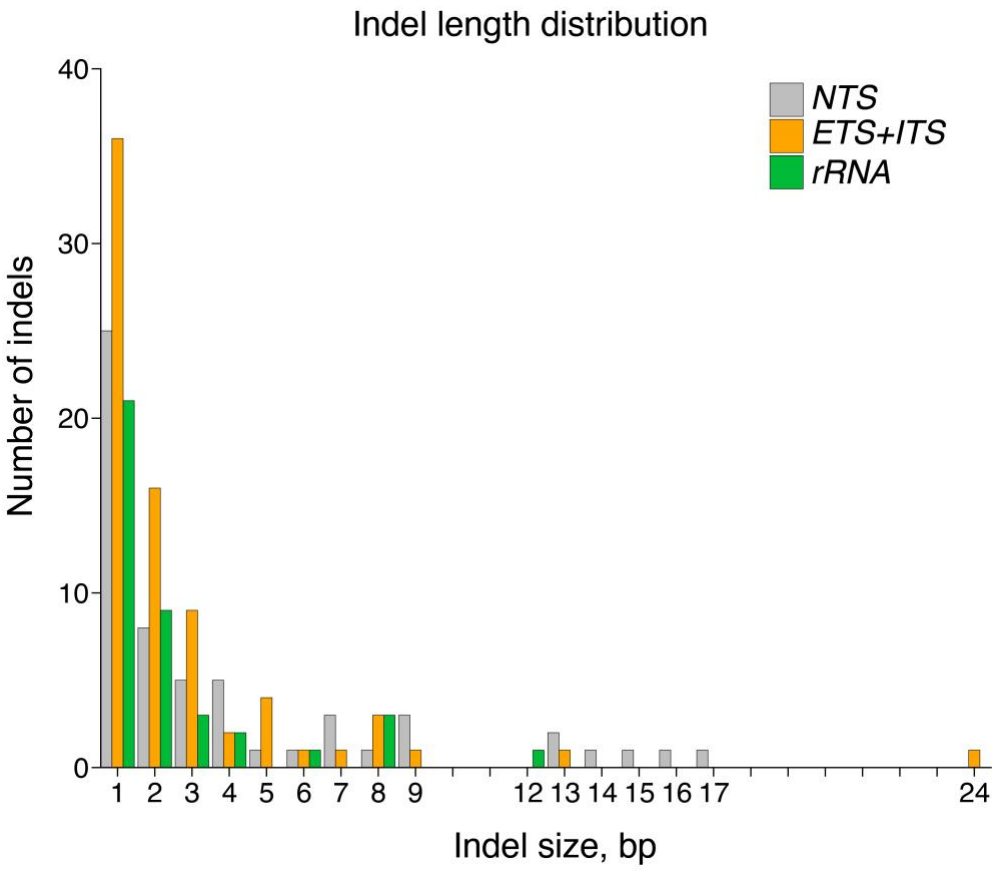

Fig. S4. Indel length distribution across rDNA. The number of indels across all isolates is shown as function of indel size.

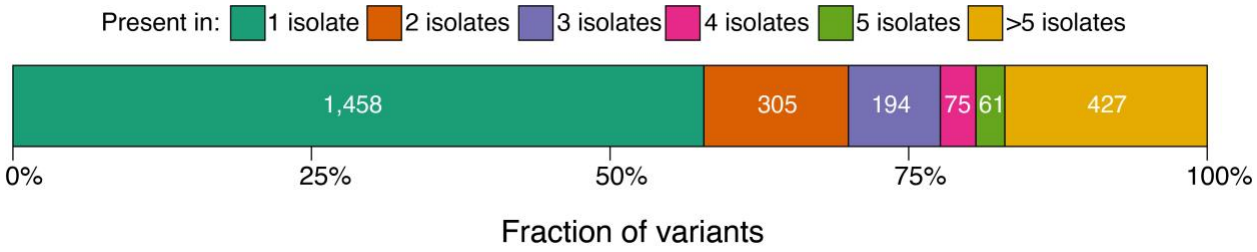

Fig. S5. Sharedness of variants across isolates. Numbers inside the bars indicate the number of variants in each category.

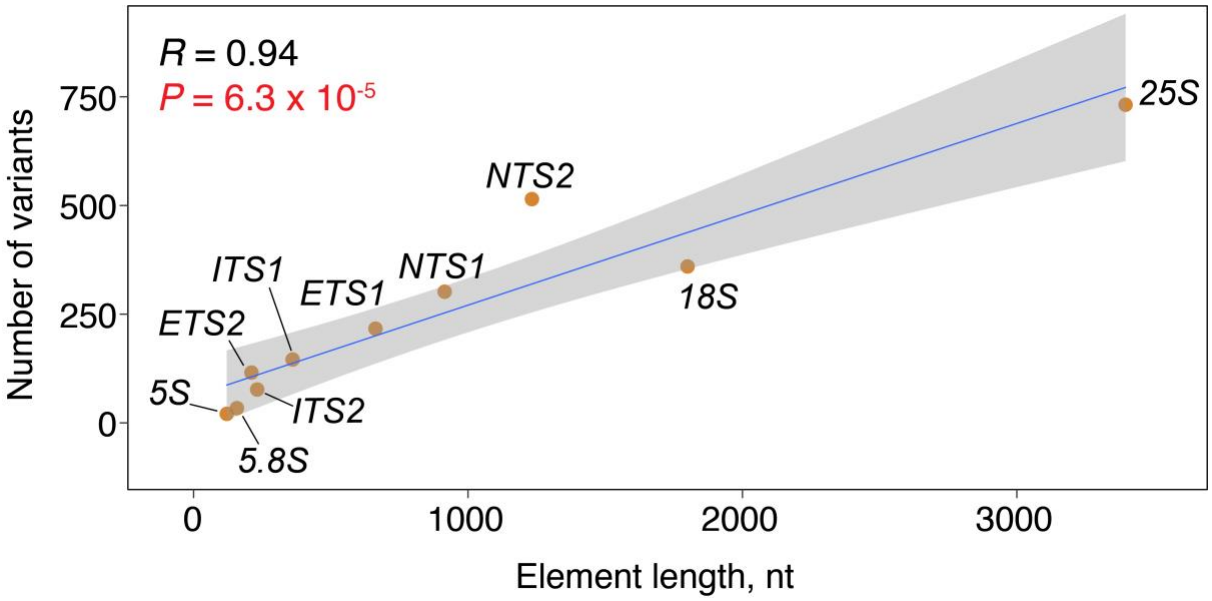

Fig. S6. Correlation between the number of rDNA variants per element and the length of the corresponding element. *R* – Pearson correlation coefficient. Brown dots represent different elements. Regression line with 95% confidence interval is shown in blue. *NTS2* and *ETS1* lengths were adjusted for sequence truncation due to filtering (see Methods).

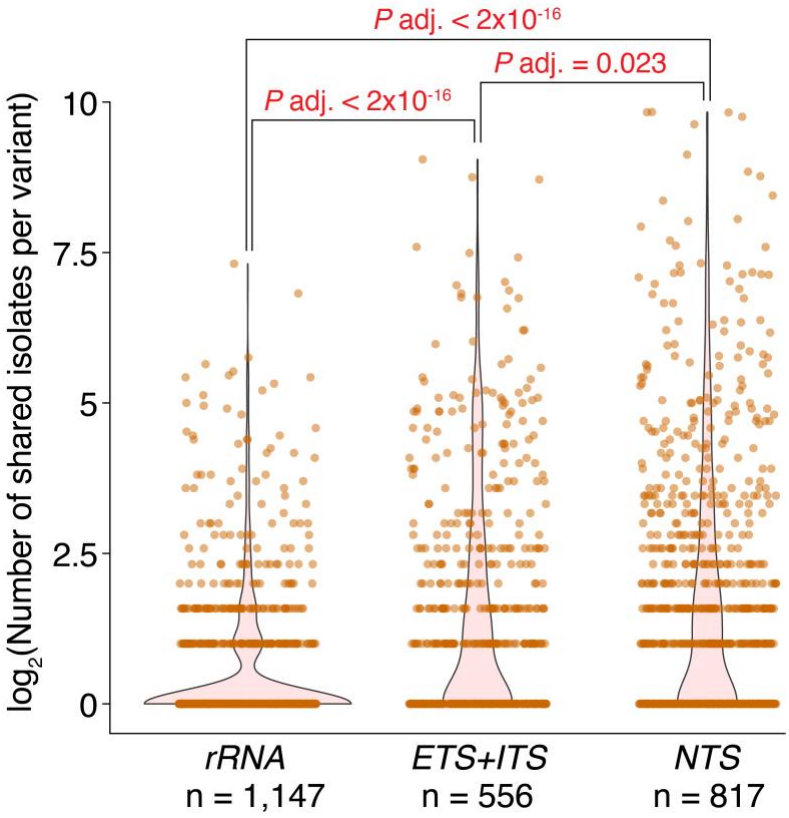

Fig. S7. Sharedness of variants across isolates is stratified by rDNA elements. n - total number of variants in the group. Y-axis is shown in log2 scale. Adjusted p-value (*P* adj) was calculated using Pairwise Wilcoxon rank sum test with Benjamini–Hochberg correction.

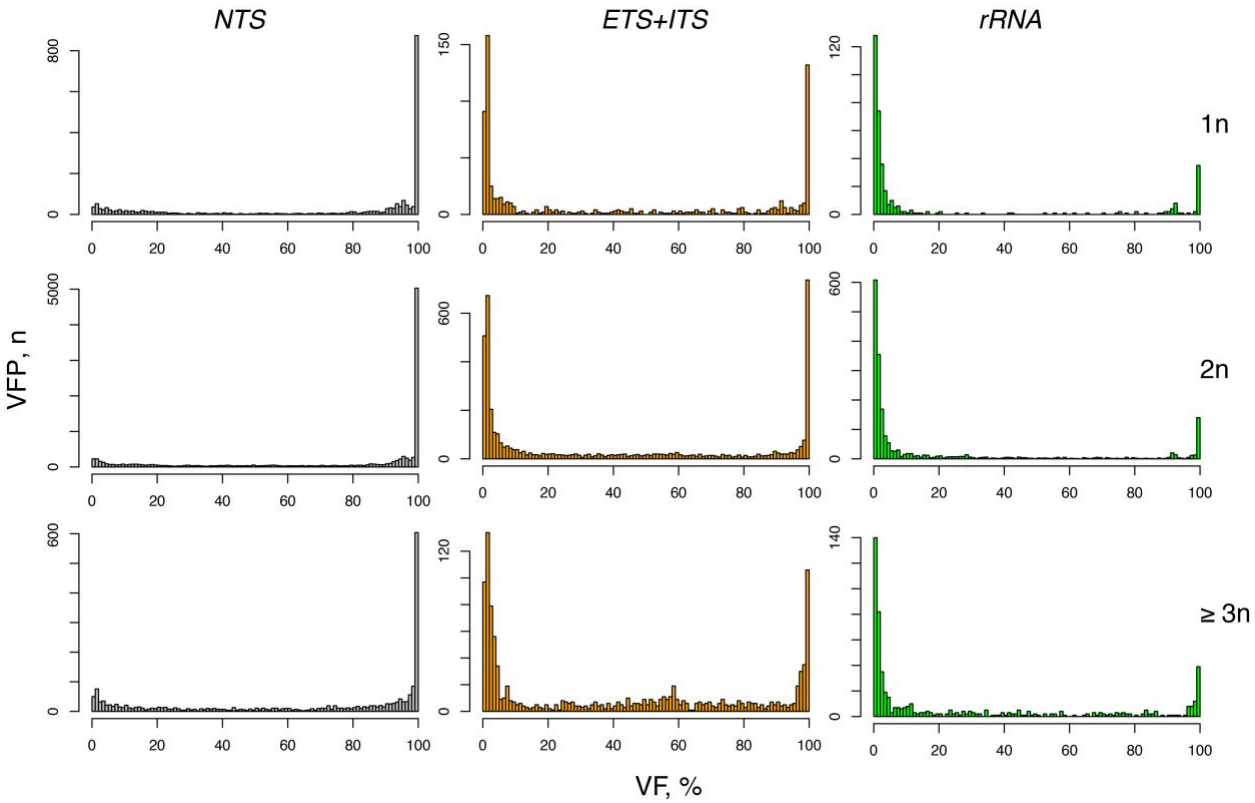

Fig. S8. Distribution of VFPs separated by rDNA elements and isolates' ploidy. 1n - haploid, 2n - diploid, >= 3n - polyploid isolates.

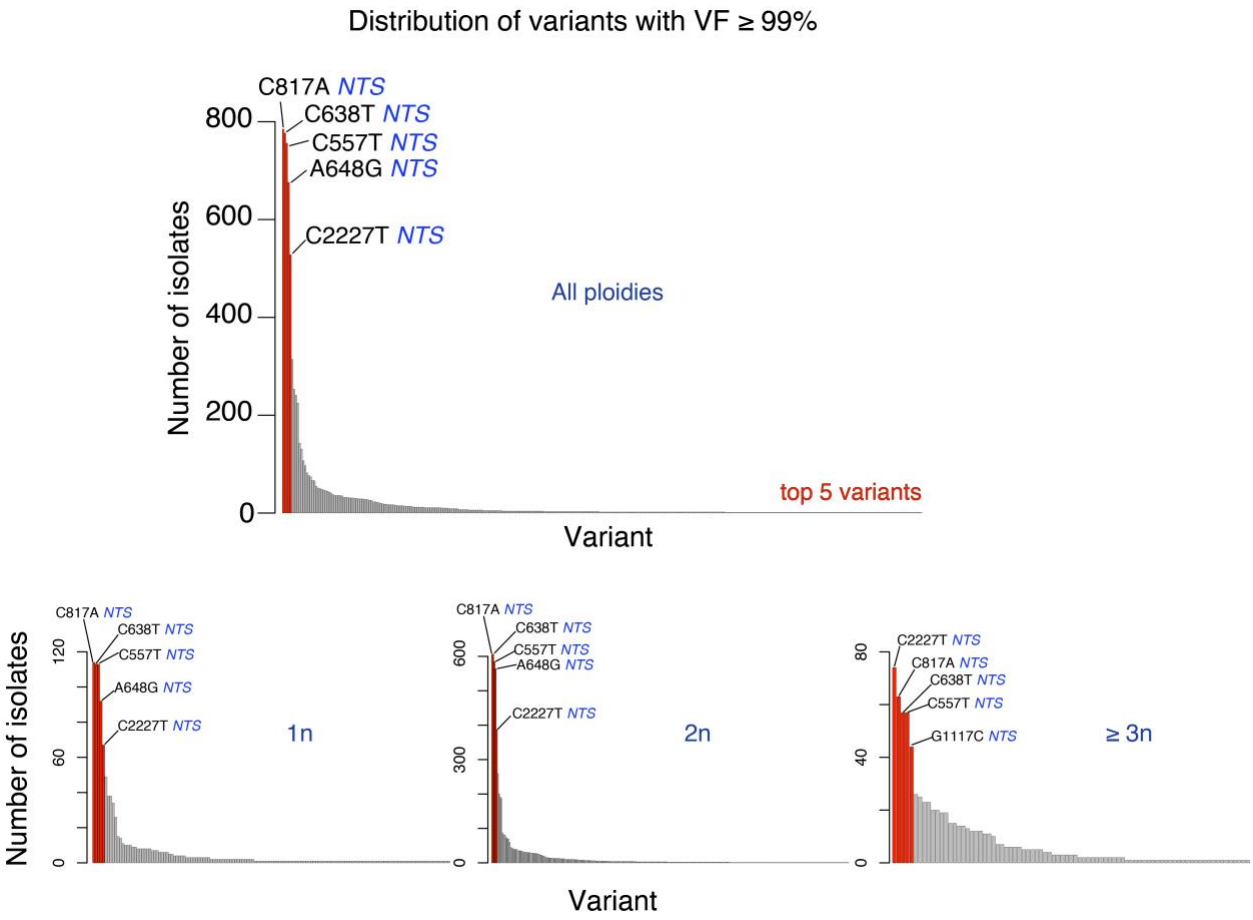

Fig. S9. Distribution of variants with VF ≥ 99% across isolates. Each bar represents a variant, and the height of the bar (y-axis) shows how many isolates contain the variant at VF ≥ 99%. The plots are shown for all isolates ("All ploidies", *top*), and separated by ploidy (*bottom*). The most common variants were all located in *NTS* elements.

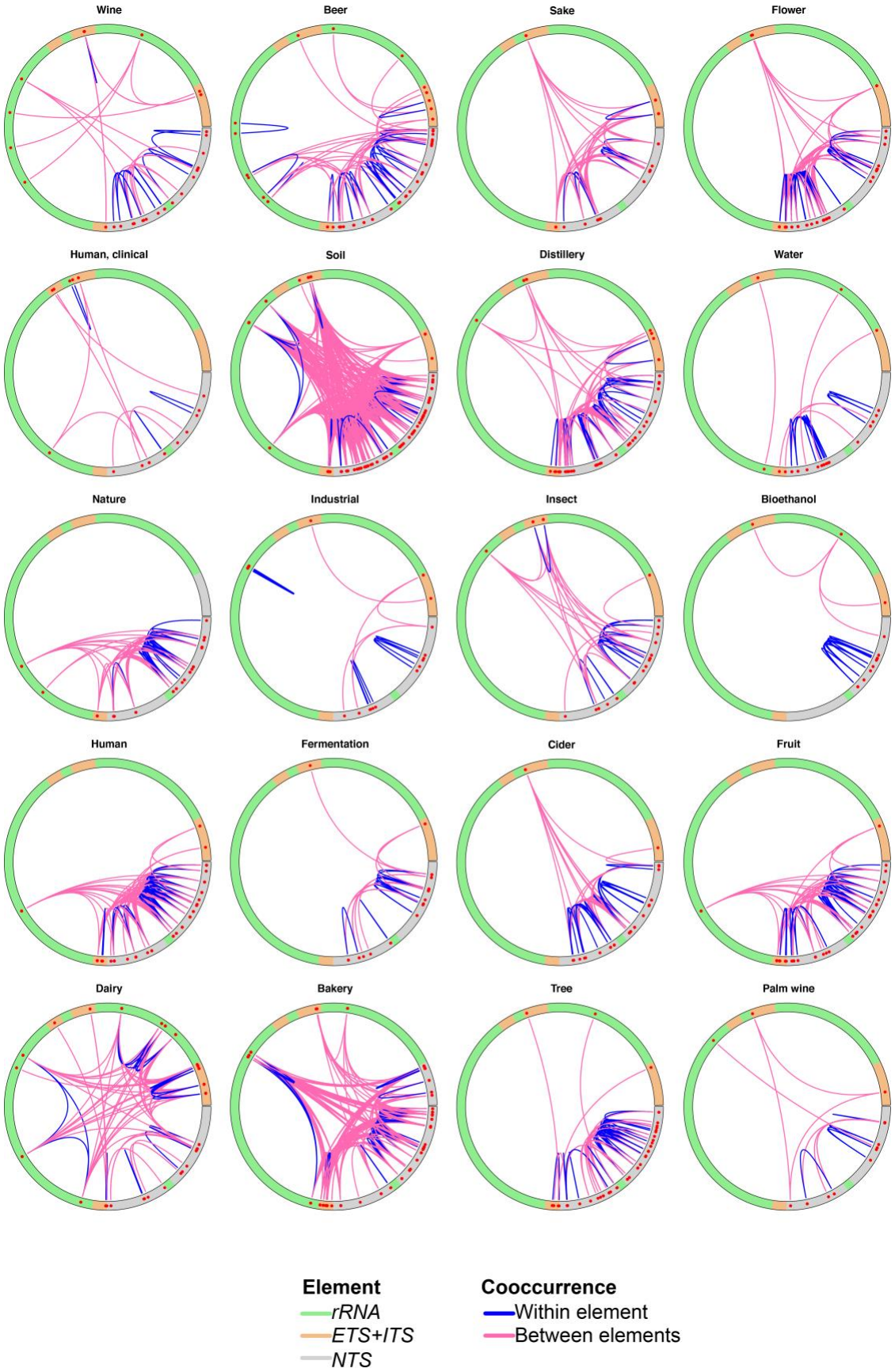
Wine
Beer
Sake
Flower
Human, clinical
Soil
Distillery
Water
Nature
Industrial
Insect
Bioethanol
Human
Fermentation
Cider
Fruit
Dairy
Bakery
Tree
Palm wine
Element
rRNA
ETS+ITS
NTS
Cooccurrence
Within element
Between elements

Fig. S10. Variants with consistently similar VFs in different ecological niches. Cooccurrence was calculated based on Euclidean distances (see Methods) between VFs for every pair of variants within an isolate. An rDNA copy is shown as a circle with colors corresponding to different elements (green – *rRNA;* orange – *ETS1*, *ETS2*, *ITS1* and *ITS2*; grey – *NTS1* and *NTS2*) and variants are shown as red dots. Blue and pink lines indicate variant pairs within or between rDNA elements, respectively. A variant can be connected to multiple other variants.

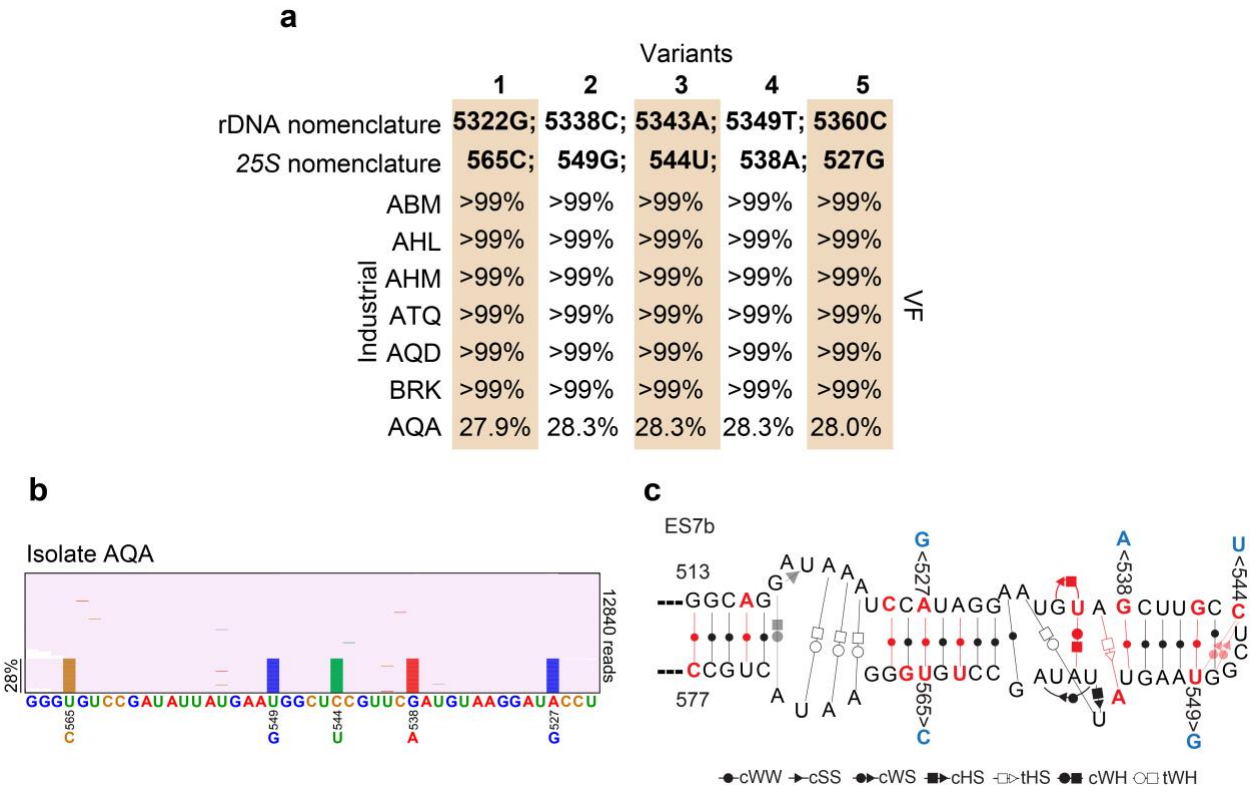

Fig. S11. Five variants cooccur in rDNA copies in several isolates. a, VFs of the five variants in *25S* among seven isolates. **b**, continuous rDNA reads of isolate AQA mapped to *25S*. All the variants are in the same reads and only a subset of rDNA copies (28%) contain the variants in the genome. **c**, localization of the variants (blue) in the two-dimensional context of the hairpin from expansion segment ES7b. Variants that were detected across all the isolates in this region are shown in red. Types of interaction are described below the structure; cWW: cis-Watson-Crick/Watson-Crick, cSS: cis-Sugar/Sugar, cWS: cis-Watson-Crick/Sugar, cHS: cis-Hoogsteen/Sugar, tHS: trans- Hoogsteen/Sugar, cWH: cis-Watson-Crick/Hoogsteen, tWH: trans-Watson-Crick/Hoogsteen.

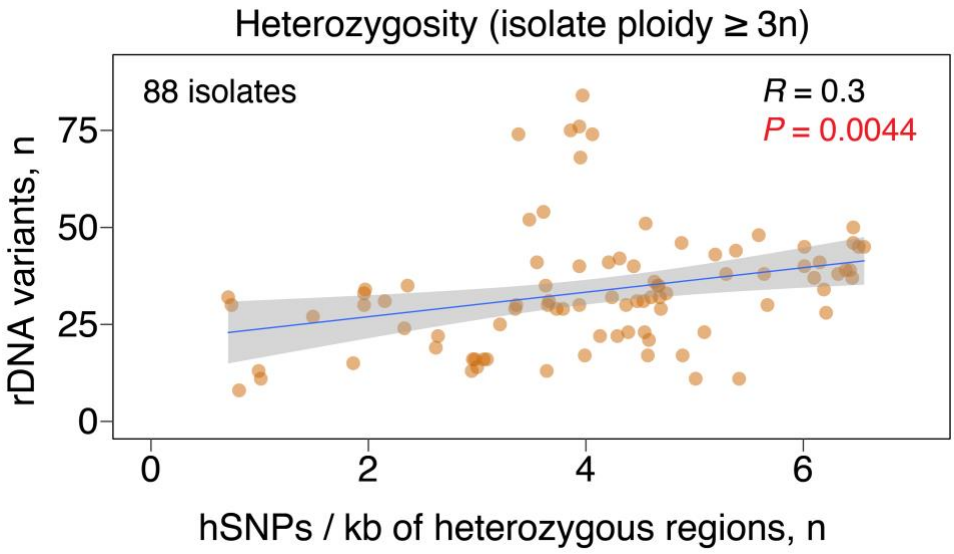

Fig. S12. Relationship between heterozygosity of the rDNA and the rest of the genome in polyploid (3n-5n) isolates. Correlation between the number of rDNA variants and the number of heterozygous SNPs per heterozygous region of the genome (excluding rDNA; from [22]). Pearson correlation coefficient *R* is shown. Each dot represents an isolate, regression line with a 95% confidence interval is shown in blue.

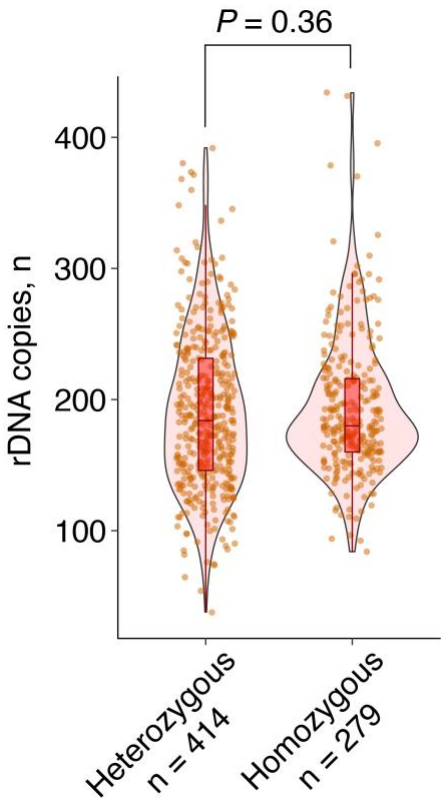

Fig. S13. Number of rDNA copies in heterozygous and homozygous diploid isolates analyzed in **Fig. 2**. n indicates the number of isolates in each group. *P*-value was calculated using Wilcoxon rank sum test.

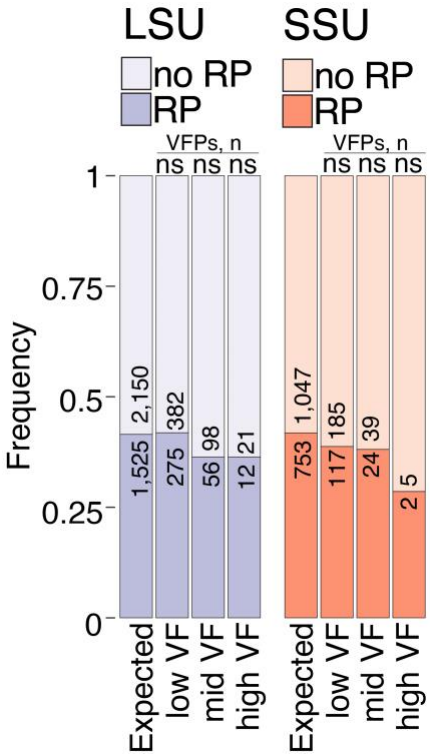

Fig. S14. Distribution of VFPs that are involved in rRNA-ribosomal protein (RP) interactions. ns – non-significant difference (Fisher's test $P > 0.05$)

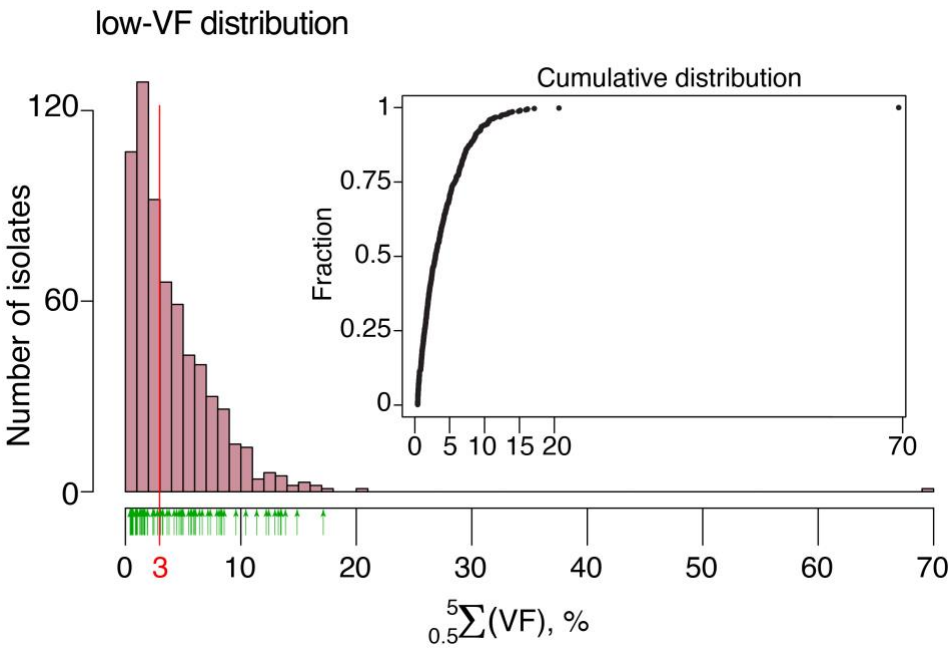

Fig. S15. Distribution of sums of low VFs in coding regions (*5S, 5.8S, 18S, 25S*) across all isolates. Green arrows indicate the presence of isolates with less than 1% of variants in CNEs. The median is shown in red. *inset* – cumulative distribution of the same dataset.

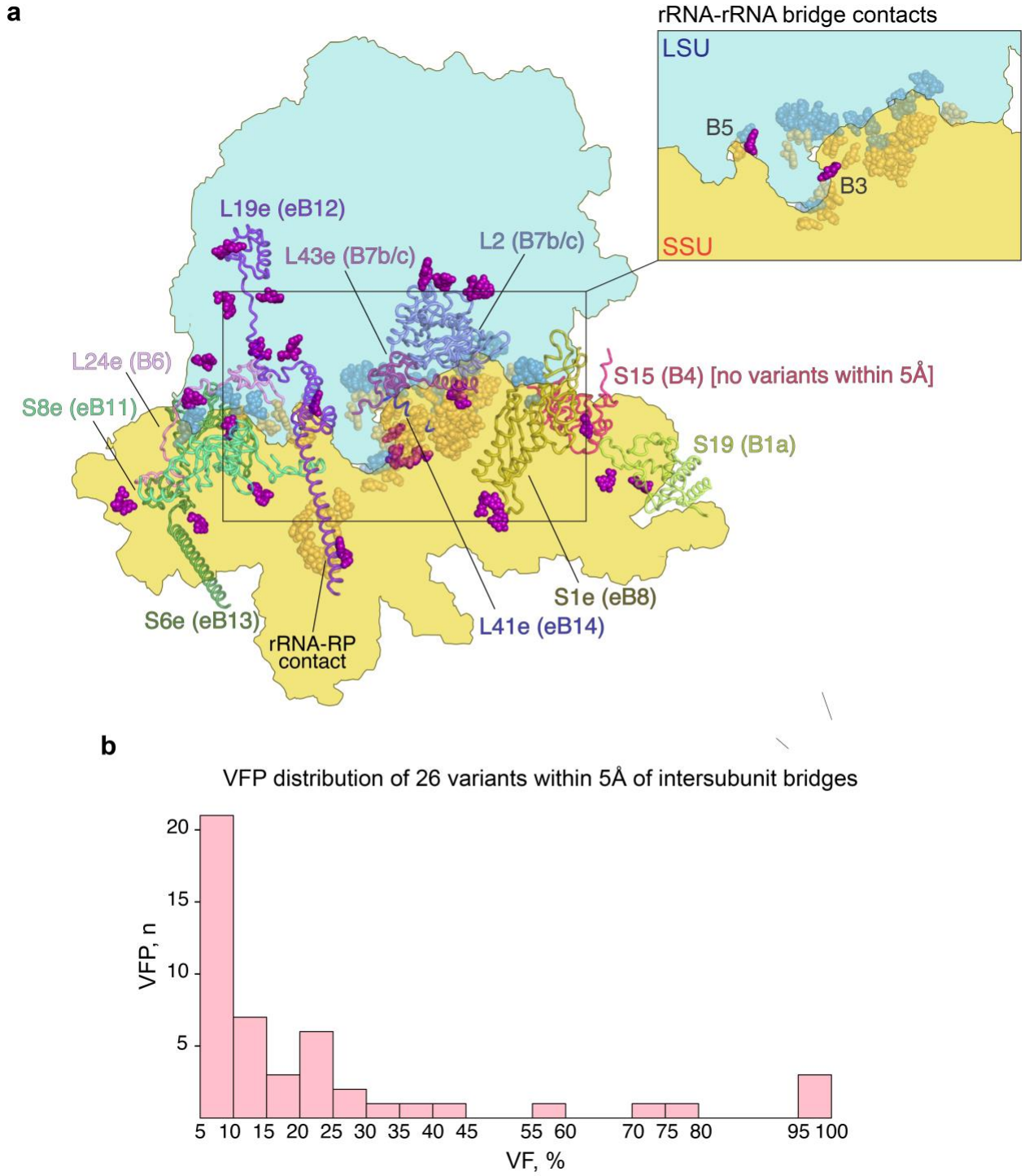

Fig. S16. Variants are localized in the vicinitiy of intersubunit bridges. a, variants (shown in purple with the spherical representation) within 5Å of the intersubunit bridges components

(protein part of a bridge is shown with the backbone representation, rRNA part is show with the transparent spherical representation). Ribosomal protein (RP) names along with the corresponding bridges (in brackets) are indicated. *inset* - bridge contacts mediated by rRNA only. Bridges for which variants were found in their vicinity are indicated in the inset. **b**, VFP distribution of the 26 variants found within 5Å of the intersubunit bridges. Blue outline - large subunit (LSU); yellow outline - small subunit (SSU).

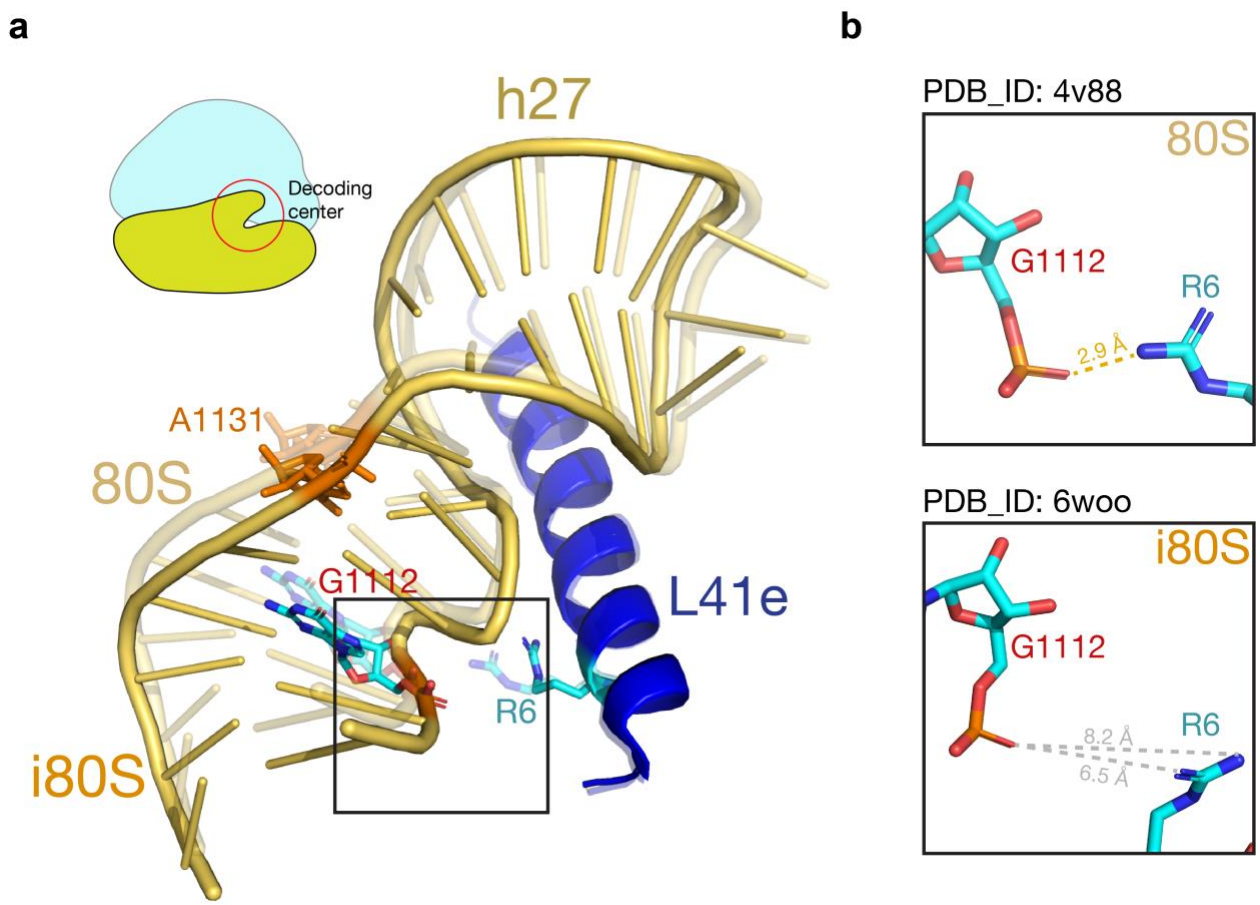

Fig. S17. Structural alignment of a region near the decoding center. a, 80S ribosome structures at two different stages of translation - non-elongating ("80S", transparent; PDB_ID 4v88 at 3.0Å) and initiating-to-elongating state ("i80S", bold; PDB_ID 6woo at 2.9 Å). The region encompasses helix h27 and polypeptide L41e; the latter is part of intersubunit bridge eB14. Positions at which variants were founds are highlighted (G1112 and A1131; 18S nomenclature). **b**, zoom into the region between G1112 and L41e. The phosphate backbone of G1112 forms a hydrogen bond with R6 of L41e in 80S (*top*), while no interaction occurs in i80S (*bottom*) due to conformational change.