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, 10003, 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). The multiple copies of rRNA genes are thought to have low sequence diversity within one species. Here, we present species-wide rDNA sequence analysis in *Saccharomyces cerevisiae* that challenges this view. We show that rDNA copies in this yeast are far from homogeneous, both among and within isolates, and that many variants avoided fixation or elimination over evolutionary time. The sequence diversity landscape across the rDNA shows clear functional stratification, suggesting different copy-number thresholds for selection that shape rDNA diversity. Notably, nucleotide variants in the most conserved rDNA regions are sufficiently deleterious to exhibit signatures of purifying selection even when present in only a small fraction of rRNA gene copies. Our results portray a complex evolutionary landscape that shapes rDNA sequence diversity within a single species and reveal unexpectedly deep purifying selection of multi-copy genes.

INTRODUCTION

The ribosomal DNA (rDNA) codes for the ribosomal RNA (rRNA), the universal RNA components of the ribosome, and therefore is essential for protein translation and viability in every species. 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 [1], and non-transcribed sequences (*NTS1* and *NTS2*), which contain promoters and the origin of replication [2–4]. Most eukaryotic genomes harbor dozens or hundreds of rDNA copies arranged in one or multiple loci [5]. It is well established that rDNA copy numbers vary widely even among individuals of a single species, contributing to phenotypic diversity [6–10]. By contrast, how much rDNA sequences vary within a genome remains poorly understood.

rDNA sequence variation has been mostly studied in light of the concerted evolution model, which proposes that rDNA copies share derived changes, both on the level of genomes and species, as a consequence of mechanisms such as gene conversion and repeated unequal crossing over [11–14]. High rates of these mechanisms are expected to result in 1) high homogeneity of rDNA copies within a genome with only a few sequence variants, 2) little difference in the distribution of variation among *rRNA*, *ETS*, *ITS*, and *NTS* elements, and 3) the rapid elimination or fixation of new variants across all rDNA copies [12,13]. However, the assumption of high rates of such mechanisms, and hence these predictions, were based on very limited sequencing information with low sequencing depth from only a few individuals. Indeed, recent studies have detected substantial within-genome rDNA variants across diverse phyla, including fungi [15–17], invertebrates [18–21], plants [22], and mammals, including humans [23–25]. Moreover, several rDNA sequence variants in mammals

and insects appear to have persisted over extended evolutionary times [21,25], conflicting with the postulated rapid fixation of variants.

On the other hand, if different modes of selection contribute to rDNA sequence evolution (with low or moderate rates of concerted evolution), then one expects 1) higher intragenomic rDNA variation, 2) marked differences in variation among different rDNA elements, depending on the phenotypic effect of a given sequence variant, 3) a range of frequencies for different variants proportional to their effects on fitness as a result of directional selection, and 4) randomly fluctuating frequencies of variants that are not under selection. Interestingly, the sequences of several multi-copy gene families, including the major histocompatibility complex genes and the highly conserved histone and ubiquitin genes are individually under strong purifying selection [26–29], demonstrating how selection might act on multiple-copy genes alongside concerted events. Additionally, diversifying selection has maintained multiple variants over long periods of time in the repeated globin genes of mammals [30–32]. How selection acts on the much more numerous rDNA copies has not been determined.

Here, we took advantage of the high sequencing coverage of the 1,002 Yeast Genomes Project [33] to quantitatively assess within- and between-genome sequence heterogeneity and determine the evolutionary landscape 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 on a species scale and define regions of constrained and permissive variation of rRNA in different regions of the ribosome.

RESULTS

Substantial rDNA sequence diversity

To investigate the level of sequence diversity across rDNA elements, we analyzed rDNA sequences of the 1,002 Yeast Genomes Project, which comprises deep-sequencing data (~200-fold mean coverage) of 918 *S. cerevisiae* isolates from 21 ecological niches [33], including 133 haploid and 694 diploid isolates, as well as 91 isolates with higher ploidies. We used LoFreq [34] to identify rDNA sequences that differed from the rDNA prototype of the S288c reference strain and considered the alternative sequences to be "rDNA variants" if more than 0.5% of the rDNA reads contained such sequences (**Fig. 1a**, **Supplementary Fig. 1b**). In addition, because of the variable number of rDNA copies in different isolates (**Fig. 1b**), we used the frequency cutoff corresponding to one rDNA copy for considering variants in each isolate (see Methods). We refer to the relative abundance of a variant among all rDNA reads as "intragenomic variant frequency" (iVF) (**Fig. 1a**, **Supplementary Fig. 1**).

Benchmarking with a test set showed that our pipeline efficiently recovered low-frequency variants down to 0.5% iVF (> 97% recovery; **Supplementary Fig. 2** and Methods), and thus allows detection of variants that are present in only one or a few rDNA copies within a genome. Additional filtering of long homopolymer stretches and low-quality variants further ensured a low number of false positive calls (**Supplementary Fig. 2**).

We observed ubiquitous rDNA sequence variation, both within individual genomes (intragenomic) and between different isolates (intergenomic; **Fig. 1c, Supplementary Fig. 3**). Within genomes, a median of 13 variants per isolate co-existed with S288c reference sequences (**Fig. 1d**). In the

highly conserved rRNA genes, the median was two variants per isolate, including one case with 40 low-frequency variants (isolate name from [33]: "BAM"; **Supplementary Fig. 4**).

Intergenomic variation was similarly common (different colors; **Fig. 1c**), and the number of variants per genome only weakly correlated with the total number of rDNA copies per isolate (**Supplementary Fig. 5a**). Notably, almost half (44%) of the variants were shared across multiple isolates (**Supplementary Fig. 5b**) but occurred at various iVFs (i.e., in different proportions of rDNA copies) and thus represented intragenomic variant frequency polymorphisms (iVFPs; **Fig. 1a**). We note that since clonal populations were sequenced, and every isolate was analyzed separately, every iVF represents the frequency of a variant within a genome of a particular isolate and not the population-wide genomic frequency.

Among all isolates, we detected 2,129 variants across 1,931 polymorphic sites, indicating that most sites exhibited only one type of sequence change. The majority of variants were single-nucleotide variants (SNVs) and short indels (< 3 bp; range: 1-24 bp; **Fig. 1e**, **Supplementary Fig. 6**). Across the entire data set, we observed a total of 21,562 iVFPs with frequencies from 0.5%-100% (**Fig. 1e; Supplementary File 1**), with most polymorphic sites occurring in the *NTS* regions (**Supplementary Fig. 7**, **Supplementary File 2**).

Closer inspection of the functional rDNA elements revealed that at least 13% of nucleotide positions in each element were polymorphic (**Fig. 1e**). The number of variants correlated strongly with the length of each element (**Supplementary Fig. 8**), consistent with random mutation accumulation, but variants were least likely to be shared among isolates if they occurred in the

rRNA genes compared to *ETS+ITS* and *NTS* elements (median: 1 (*rRNA*) vs. 2 (*NTS*; *ETS+ITS*) isolates per variant, $P < 2x10^{-16}$, pairwise Wilcoxon rank sum test with Benjamini-Hochberg correction; **Supplementary Fig. 9**). 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. By contrast, the *NTS1/2* elements contain many spacer sequences and thus are more permissible for variation [26]. The difference in selective pressure was also apparent when iVFPs were empirically separated into three categories based on iVFs – low (< 5%), mid (from 5% to 95%), and high (≥ 95%; **Fig. 1e, f**). We predominantly observed variants with mid- and high iVFs in regulatory sequences (*NTS1/2*, *ETS1/2*, *ITS1*) and low iVFs in the rRNA genes (**Fig. 1e**). Nevertheless, 25% (232) of variants in *25S*, *18S*, and *5.8S* had iVFs above 5%, indicating even these conserved elements contain positions with relaxed sequence constraints for variation.

When compared to the S288c reference rDNA, iVFPs followed a U-shape distribution regardless of the isolates' ploidy (**Fig. 1f**, and **Supplementary Fig. 10**), with peaks at low (below 5%) or high (above 95%) iVFs. For the *NTS* regions, the highest peak corresponded to very high iVFs (≥ 99%) and reflects virtually fixed variants among rDNA copies. None of these variants were shared across all 918 isolates (**Supplementary Fig. 11**), indicating that the peak at iVF ≥ 99% was not inflated by the choice of S288c as the reference. However, we did observe five variants, all in the *NTS* regions, that were shared across over 500 isolates and were responsible for 2989/6507 (46%) iVFPs with iVF ≥ 99% in *NTS*, explaining some of the skew in the distribution. In contrast, the *ETS* and *ITS* regions ("ETS+ITS") exhibited equally distributed peaks at low and high iVFs, pointing to greater selective pressure in these elements due to their importance in rRNA processing

and maturation [1]. In the *rRNA*-encoding elements, iVFPs with low iVFs dominated, presumably as a result of stringent selection.

The U-shaped distribution of iVFPs held within each ecological niche, with interesting outliers (**Fig. 1g**). For instance, dairy isolates exhibited a greater variation of iVFs in the rRNA-encoding sequences, and isolates involved in bioethanol production exhibited a peak with iVFs 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.

Since multiple variants can arise in the same rDNA copy, we asked whether pairs of variants repeatedly co-occurred together. Short-read sequencing data only detects co-occurring variants when they are present in the same short read. However, we can infer that two or more variants may occur in the same rDNA copy if they exhibit consistently similar iVFs across multiple isolates (e.g. variants X and Y both had iVF ~ 25% in isolate A and ~ 8% in isolate B; see Methods). We observed 2,864 such coupled variant pairs within and between rDNA elements (**Supplementary Files 3**, **4**; **Supplementary Figs. 12**, **13**). The majority (71%) of those variants had iVF ≥ 80%, which unambiguously indicated their co-occurrence in the same set of rDNA copies. Occasionally, we also observed variants within the same sequencing read pair. One notable example included five co-occurring SNVs within a stretch of 38 nucleotides in the *25S* rRNA of several industrial strains (**Supplementary Fig. 14**).

Taken together, these data reveal pervasive heterogeneity in rDNA sequences within and between genomes of the same species and show non-uniform functional sequence constraints.

Stratification by environment

Previous work demonstrated that domestication strongly influences mutation accumulation, with wild *S. cerevisiae* isolates accumulating single nucleotide polymorphisms and domesticated strains predominantly utilizing changes in gene copy number to increase fitness [33].

To investigate whether domestication also affected variant accumulation in the rDNA, we took advantage of the recent classification of the yeast isolates based on their habitat – domesticated, wild, human bodyor unassigned [35] (**Fig. 2**). iVFP distribution was unaffected by habitat (**Supplementary Fig. 15**), but we observed a significant enrichment of variants in the *ETS+ITS* elements and *rRNA* of domesticated isolates compared to the wild group (median: 5 vs. 3 variants in *ETS+ITS*, $P = 0.004$; median: 3 vs. 2 variants in *rRNA*, $P = 6\text{x}10^{-4}$; Wilcoxon rank sum test with Benjamini-Hochberg correction; **Fig. 2a**). Copy number estimation further revealed that human body-associated isolates had consistently lower number of total rDNA copies (median: 152 copies) than wild (median: 181 copies, $P = 1.6\text{x}10^{-4}$), domesticated (median: 186 copies, $P = 3.9\text{x}10^{-5}$), and unassigned (median: 188 copies, $P = 0.004$) isolates (Wilcoxon rank sum test with Benjamini-Hochberg correction; **Fig. 2b**). Therefore, the accumulation of rRNA gene variants and rDNA copy numbers in yeast are stratified by growth environment.

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 further investigate 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. Since most of the natural isolates were diploid while polyploid isolates were enriched in human-related environments [33], 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. 3a, Supplementary Fig. 16**). Conversely, the number of rDNA variants in diploid isolates negatively correlated with the extent of loss of heterozygosity (**Fig. 3b**). In line with this observation, fully homozygous diploid isolates exhibited lower variant accumulation in the rDNA (median: 6 variants; **Fig. 3c**) compared to heterozygous isolates (median: 18 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, Wilcoxon rank sum test; **Supplementary Fig. 17**). Therefore, rDNA variant accumulation correlates with the rest of the genome.

To investigate how long iVFs persist across the phylogenetic tree, we determined the distribution of the topmost common variants in each element across the clades defined by the 1,002 Yeast Genome Project (**Fig. 4**). We repeatedly found that the same variants presented within clades and across related clades with different iVFPs. The most parsimonious explanation for these widespread variants is that they represent single-origin mutation events that emerged early in the isolates' evolution and were retained after the groups diverged. For example, the estimate for the domestication time of wine strains is about 1.5 thousand years ago [33] and the wine isolates share

the same variants at different frequencies both within and between clades. Similarly, some variants in the Wine/European clade were stratified by subclades, with a wide distribution of iVFPs (**Supplementary Fig. 18**). For the variants in *rRNA*, the wide range of intragenomic frequencies both within and between clades implies that these variants are likely neutral or represent standing variation, with only modest fitness effects under some stress conditions (**Supplementary Figs. 19**, **20**). These data demonstrate that genetic diversity in the rDNA can be maintained over extended periods of evolutionary time, arguing against rapid sequence homogenization of rDNA copies.

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 [36]. 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 [37,38]. 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 [38]. We adapted this approach and used the rRNA nucleotide distribution data across the shells from [39] to analyze the spatial distribution of the detected variants in the *S. cerevisiae* 80S ribosome (**Fig. 5a** and **Supplementary File 5**).

In line with the evolutionary analyses, iVFPs were significantly skewed toward the outer shells of the ribosome (LSU: low iVF $P = 0.026$, mid iVF $P = 0.013$, high iVF $P = 0.013$; SSU: low iVF $P = 0.015$, mid iVF $P = 0.006$, high iVF $P = 0.006$, Wilcoxon rank sum exact test with Benjamini-Hochberg correction). As expected, variants with iVFs ≥ 5% were absent in shell 1 of the LSU, which contains the peptidyltransferase center, the highly 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 [40]. This pattern is also consistent with a previously reported conservation pattern in the rRNA of other eukaryotic species [25]. The depletion of higher-frequency variants (no mid- and high iVFPs variants in the LSU; no high iVFPs in the SSU and only one variant with iVF

7.1%) in the core shells likely reflects mutation-selection balance, whereby newly arising variants are generally highly deleterious and countered by stringent purifying selection. Several mid- and high frequency variants were localized to the second most inner shell 2. In the LSU, six iVFPs had intragenomic frequencies above 5% in the shell, and three of them interact with ribosomal proteins (RPs) (**Supplementary File 5**). Notably, we observed no significant differences in iVFs between rRNA variants localized in binding interfaces with RPs and variants outside of those interfaces (**Supplementary Fig. 21**), arguing that nucleotides involved in RP contacts do not experience significantly different selective forces. In the SSU, seven iVFPs in shell 2 had intragenomic frequencies above 5%, three of which were located in or near functionally important ribosomal regions (see "*Variants in functional regions*" below), suggesting some of them might be either under neutral or positive selection. On the other hand, the outermost regions largely contain rRNA expansion segments (ES; **Supplementary Fig. 22**), which vary greatly in their sequence and length between eukaryotes [41–43], and thus are more permissible for mutations (**Fig. 5b**).

To estimate the strength of purifying selection, we analyzed variants in the *conserved nucleotide elements* (CNEs), which have virtually the same sequences (> 90% identity) across all available eukaryotic 25S/28S rRNA genes [44]. 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. Mid- and high frequency variants were almost absent in the CNEs, with the majority of identified CNE variants (135/149; 91%) occurring at iVFs below 6% (**Fig. 5c, d; Supplementary File 5**). These low iVFs suggest that mutations in the CNEs are sufficiently deleterious to be selected against even at very low rDNA copy numbers. In fact, even iVFPs with iVFs down to below 1% were significantly underrepresented in CNEs compared to the rest of the *25S* sequence (P = 1.3×10^{-8}, Fisher's exact test with Benjamini-Hochberg

correction; **Fig. 5e**). Such low iVFPs correspond to only one or a few rDNA copies per cell (**Supplementary Fig. 23**). In some cases, the depletion of variants with iVF < 1% in CNEs may be due to the additive presence of other deleterious rRNA mutations in the same genome that shifts the cumulative iVF to > 1% (**Supplementary Fig. 24**; 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 11 CNE variants were observed at higher iVFs (≥ 6%) (**Fig. 5d**). Strikingly, eight of these localized along the GTPase-associated center (GAC; **Fig. 6a, b**), 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 [45]. The eight variants were distributed among nine isolates at various frequencies (**Fig. 6c**). Closely related isolates ("CFH", "CBR", and "CFG") shared the same variant C1248U (25S rRNA nomenclature; iVFs 20-33%), suggesting a single mutational origin, while the same position also exhibited a different variant sequence in isolate "CFA" (C1248A).

Intriguingly, two variants, C1248U and G1237A, were found in the same isolate ("CFH"), 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 *25S* in "CFH" (**Fig. 6d**).

To test whether these variants are expressed, we performed both DNA- and total RNA-sequencing (without poly-A enrichment) of the nine isolates (**Supplementary Figs. 25-28**). With the exception of isolate "CFG", we detected variants at very similar iVFs as reported in [33], indicating that iVFs remained largely stable over the course of culturing. This included the GAC variants, which showed excellent agreement with the data in [33] (**Fig. 6e**; Pearson's $R = 0.98$, $P = 1.4\text{x}10^{-6}$). Importantly, the GAC variants were also detected at matching frequencies in the reads from total RNA-sequencing data (**Fig. 6e**; Pearson's $R = 0.96$, $P = 7.7\text{x}10^{-6}$), indicating that the GAC rRNA variants are expressed. Moreover, the two variants in isolate "CFH" (C1248U and G1237A) also appeared in mutually exclusive transcripts, indicating that they may contribute to different ribosomes.

Secondary structure analysis showed that all but one of the eight variant positions are involved in base-pairing interactions (**Fig. 6b**). 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 [46], these changes could potentially alter the overall geometry of the region and thus affect the interaction with translational GTPases. The relatively high iVFs (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. Interestingly, C1238U, in combination with another variant, was shown to increase translational accuracy in *S. cerevisiae* [47].

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 [48]. We detected 26 variants with mid- and high iVFs within 5Å of intersubunit bridges (**Supplementary Fig. 29; Supplementary File 5**). Among these, variant G1112A in the 18S rRNA (18S rRNA nomenclature; 44.8% iVF in isolate "CDN") is part of helix 27, which dynamically interacts with bridge eB14 during different steps of translation (**Supplementary Fig. 30**) and is near the decoding center. Intriguingly, the same G1112A mutation was previously shown in mutagenesis studies to decrease readthrough of stop codons in *S. cerevisiae* [49], 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. This diversity is non-uniformly distributed across major rDNA elements (*NTS*, *ETS+ITS*, and *rRNA*), and rDNA variants can persist, and be subject to selection, over a wide range of frequencies. The varying frequencies of rDNA polymorphisms likely arose because of new mutations, random duplication, and deletion of rDNA copies, as well as gene conversion. Moreover, deleterious variants are eliminated, and beneficial variants expanded, as long as their relative copy numbers provide sufficient phenotypic consequences, whereas neutral variants can persist and fluctuate in frequency among rDNA copies for extended periods of evolutionary time. These data suggest a lower rate of concerted evolution than previously thought, and a strong role of directional selection that results in a diverse pattern of intra- and intergenomic rDNA sequence variation within one species. As a result, the substantial within-genome polymorphisms seen in this study and observed across diverse phyla [15–25] are consistent with and expected and imply that selection plays an important role in shaping rDNA diversity across eukaryotes.

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 the 28S rRNA coding sequences are disrupted by R1/R2 retrotransposons [50]. We suggest that these differences can be explained by how the respective mutations contribute to ribosome biogenesis and function. The majority of coding variants observed here are SNVs that are expressed as rRNAs and likely incorporated into ribosomes. Given that translation frequently engages multiple ribosomes on the same transcript, the presence of non-functional ribosomes has the potential to cause partially dominant effects and can 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 deleterious rRNA variants are expressed at very low levels [51,52]. We expect that variants in *ETS* and *ITS* that yield defective ribosomal particles may have similar effects [53]. By contrast, mutations that disrupt rRNA transcription altogether, as seen in *Drosophila*, are likely subject to much weaker selection because of the high number of available rDNA copies, which include a substantial number of dormant rRNA genes. Several regulatory mechanisms are known activate these dormant rRNA genes or generally increase rRNA transcription [54]. These mechanisms are expected to provide a substantial buffer against the inactivation of individual rDNA copies and lead to a higher rDNA copy number threshold for selection. We note that the extreme purifying selection seen in our study may also mask disease-causing variants in the human rDNA, especially since rDNA sequences are largely ignored in genotype-phenotype association studies. Based on our data and previous studies in humans [25], we suggest that considering even low-frequency polymorphisms might be important for investigating the role of rDNA variation in health and disease.

The effects of the detected variants on cell fitness will require further functional analyses. It seems likely that many rRNA nucleotide polymorphisms have the potential to impact cellular physiology and adaptation to the environment. rRNA variants exhibit tissue-specific expression and are incorporated into actively translating ribosomes in humans, mice, and *Arabidopsis* [22,25] , and several rRNA nucleotide changes in yeast confer resistance to antibiotics [55] or affect translational fidelity [47,49]. Notably, even variation among rRNA copies can affect fitness, as suggested by work in *E.coli* that demonstrated that intragenomic variation in rRNA sequences confers benefits in response to stress [56]. We uncovered a couple of examples of natural variants that show increased translational fidelity in laboratory experiments [47,49], and may thus be adaptative in certain environments. Our analyses in yeast, along with studies in other organisms [22,25] indicate that rRNA

diversity is likely much more pervasive than previously thought. Therefore, sequencing more individuals of a species could reveal substantial hidden variation.

Finally, our results highlight the importance of considering within-species sequence diversity of rDNA in metagenomic analyses, which frequently use short rDNA segments to estimate species numbers and population complexity in environmental or clinical samples [57] . 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 ID: ERP014555) [33] 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 [58] without annotated rDNA using bowtie2 [59]: -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 [60]. Variants were subsequently called using LoFreq [34]. 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: 1) nucleotide position > 10 and < 9100 (to avoid end effects), 2) iVF > 0.005, 3) homopolymer run < 4, 4) GC content < 0.6 in long indels (> 5 nt) in either reference or alternative sequences, 5) variants were outside of 10-nt stretches of poly A/T/G/C sequences, and 6) the minimal iVF > theoretical one copy iVF computed for each isolate separately (see "*rDNA copy number estimation*"). Additionally, variants with the highest strand bias (≥ 144; top 12%) were removed to further reduce the rate of false positives [61], and replicates of the same isolate were deduplicated such that only the sample with the highest read coverage was considered downstream.

For **Fig. 6d** and **Supplementary Fig. 14b**, rDNA reads that fully covered the region of the variants were retrieved from the alignment and visualized in IGV [62]. Random ten percent of GAC RNA reads were sampled using samtools view -b -s 0.1 in.bam > out.bam.

Pipeline performance

To determine the lowest iVF 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 IDs: SRR4074258, SRR4074255) from [58] were mapped using the pipeline, and 10 variants unique to SK1 (not detected in S288c) with iVFs over 95% were tracked (**Supplementary Fig. 2**). Most 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. 23**). To assess the sensitivity and specificity of the pipeline, we simulated artificial datasets of 56 known SNVs (real data – median: 22 SNVs, range: 7-80 SNVs) and 3 indels (real data – median: 3 indels, range: 1-15 indels) at 0.5% iVF distributed randomly across the rDNA prototype using NEAT-genReads [63], with genomic coverages from 20-fold to 400-fold (3,000 – 60,000-fold for rDNA; real data genome coverage: mean 200-fold) (**Supplementary Fig. 2b-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 (SRA ID: SRR4074255), and the read length (the -R parameter) was set to 150. To better recapitulate the real rDNA coverage (**Supplementary Fig. 2b**), 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: 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. 2c, 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% [34]). 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

Total rDNA copy number. First, to calculate how many rDNA copies were present in each isolate, we took copy number estimates of *5.8S, 18S* and *25S* (*RDN5.8, RDN18* and *RND25;* no reported *RDN5*) reported in the "genesMatrix_CopyNumber.tab" table from [33]. Since these sequences are present together in each rDNA copy, we calculated their mean to get an estimate of the rDNA copy number per haploid genome. We then multiplied these values by the isolates' respective ploidies reported in [33] to get the total number of rDNA copies per isolate. We calculated the number of rDNA copies per clade by taking the median number of rDNA copies of isolates in each clade.

One-copy rDNA frequency. For each isolate, we calculated the theoretical one-copy rDNA frequency ("one_copy_iVF") based on the total number of rDNA copies in each isolate:

$$One\ copy\ iVF\ = \frac{1}{total\ number\ of\ rDNA\ copies\ in\ isolate}.$$

Sequence entropy

To examine the degree of sequence diversity at each nucleotide position, we calculated Shannon's entropy *H* (**Supplementary File 2**) for each rDNA nucleotide position *p*:

$$H(p) = -\sum_{i \in \{A,T,G,C,indel\}} \frac{n_i}{N} \times \ln \frac{n_i}{N}$$

, where *N* is the sum of all rDNA copies across the isolates, and n_i is the number of rDNA copies with "A", "T", "G", "C", or "indel" at each position, respectively. We considered total nucleotide composition at each site by pooling all rDNA copies across all isolates (N = 193,069) together, and then calculated how many rDNA copies in the pool contained "A", "T", "G", "C", or "indel" at a given position.

Variant cooccurrence

We estimated pairwise variant co-occurrence by calculating Euclidean distances *D* between iVFs within each isolate:

$$D_{XY} = \sqrt{(X_A - Y_A)^2 + (X_B - Y_B)^2 + \cdots + (X_i - Y_i)^2}$$

, where *X* and *Y* are variants; *A*, *B*, … *i* are isolates; X_A, X_B, … X_i are iVFs (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 ≥ 3 isolates within an ecological niche shared the same pair of variants; 2) $D_{XY} < 0.5$; and 3) the mean of their iVFs ($X_A + \ldots + X_i + Y_A + \ldots + Y_i$)/n was < 0.05. The parameters were empirically chosen to minimize the difference between variant iVFs and maximize the confidence in the computed

variant pairs in the absence of long-read sequencing data, although the confidence would drop for low iVFs. This allowed considering variant pairs that have different iVFs in different isolates but similar iVFs within isolates, e.g. variants *X* and *Y* both have iVF 70% in isolate *A*, iVF 50% in isolate *B* and 7% in isolate *C*, so they always have similar iVFs within an isolate, but different magnitudes in different isolates.

Correlation between rDNA variants and non-rDNA SNPs

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 kilobase of heterozygous regions or the value for loss of heterozygosity (from [33]) were used to compute Pearson's correlation coefficient. For homozygous diploid isolates, the distribution of rDNA variants was plotted in **Fig. 3c**.

rRNA analysis

For the expected distribution of iVFPs in **Fig. 5a**, 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 [39], "Onion" representation) over the total number of nucleotides N in a subunit: $F_i = \frac{n_i}{N}$. Since the number of variants is proportional to the length of a sequence, this suggests that the longer a sequence of interest is, the higher the chance of a variant occurrence in this sequence. We assumed that under a neutral selection model, the likelihood of an iVFP is higher with increased number of nucleotides in a shell ("expected distribution"). For

the LSU (*25S* + *5.8S* + *5S*), N = 3,428 instead of the actual 3,675 (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 = 1,781 (19 nucleotides were missing from the structure). For the observed distribution of iVFPs, F_i* was calculated by dividing the number of iVFPs with a chosen range of iVFs (low iVF - [0.5%;5%); mid iVF - [5%,95%), high iVF - [95%;100%]) in a shell (m_i) by the total number of iVFPs (M) in the chosen range (low, mid- or high iVFs): $F_i^* = \frac{m_i}{M}$. Due to the missing segments in the LSU and SSU structures, some iVFPs detected in **Fig. 1** were not included in the analysis. The excluded number of iVFPs in the LSU: low iVF - 52 out of 927; mid iVF - 60 out of 525; high iVF - 69 out of 241. The excluded number of iVFPs in the SSU: low iVF - 12 out of 379; mid iVF - 3 out of 224; high iVF - 2 out of 36. We then calculated the ratio between the two values: $\log_{10}\frac{F_i^*}{F_i}$ and used it as the metric for constrained (< 0) or relaxed (> 0) sequence variability in the ribosome in **Fig. 5a**, given a significant difference between the expected and observed distributions (Wilcoxon rank sum test; see text). For example, for low iVFs in shell 1 in the LSU, $F_1 = \frac{82}{3428} = 0.024$, $F^*{}_1 = \frac{7}{875} = 0.008$, and $\log_{10}\frac{F^*{}_1}{F_1} = -0.48$. For **Fig. 5b**, **Supplementary Fig. 21**, and **Supplementary Fig. 22**, coordinates of ESs and rRNA:ribosomal protein interactions were retrieved from [64] . The expected distribution of iVFPs in ESs for **Fig. 5b** was derived from the number of nucleotides in ESs (n; LSU: 883, SSU: 390) versus not in ESs (N; LSU: 2,786; SSU:1,410) - "ES": $\frac{n}{n+N}$, "no ES": $\frac{N}{n+N}$. For the observed values, iVFPs were counted for each range (*i* = low, mid-, or high iVFs) within (m_i) or outside (M_i) of ESs, and the fractions ("ES": $\frac{m_i}{m_i+M_i}$, "no ES": $\frac{M_i}{m_i+M_i}$) were calculated. For **Supplementary Fig. 21,** the expected distribution of iVFPs in the rRNA alone or rRNA:protein interaction interfaces and the observed values were

calculated in the same manner. For **Fig. 5c-e**, coordinates of eukaryotic-specific conserved nucleotide elements ("CNEs" in this paper) were retrieved from [44]. Similarly to **Fig. 5b**, the expected distribution of iVFPs in CNEs was derived as the fraction of *25S* nucleotides in CNE (n = 962) and in the rest of the *25S* (N = 2,428; total *25S* length – 3,396) - "CNE": $\frac{n}{n+N}$, "no CNE": $\frac{N}{n+N}$. For the observed values, iVFPs were counted for each range (*i*) indicated in **Fig. 5e** within (m_i) and outside (M_i) of CNEs, and the fractions ("CNE": $\frac{m_i}{m_i+M_i}$, "no CNE": $\frac{M_i}{m_i+M_i}$) were calculated.

DNA sequencing

<u>Preparation of genomic DNA</u>. Yeast isolates were grown in 5ml YPD (20g/l bactopeptone, 10g/l yeast extract, 2% glucose, 0.45g/l L-tryptophan) medium at 30°C overnight. For each isolate, half of each culture was used to prepare total genomic DNA and the other half was used in total RNA-sequencing (see below). The cells were spun down at 1,000 rcf for 5 min at 4°C, and the pellet was washed once in TE buffer (10mM Tris pH 8.0, 1mM EDTA). The pellet was then resuspended in 350µl of breakage buffer (2% Triton X-100, 1% SDS, 100mM NaCl, 10mM Tris pH 8.0, 1mM EDTA) with the subsequent addition of 350µl of phenol:chloroform:isoamyl alcohol (25:24:1) and around 200mg of glass beads. The cells were lysed using a FastPrep homogenizer (MP Biomedicals) for 2x30 sec (power = 4.5) at 4°C and vortexed on an IKA Vibrax VXR for 10 min. The samples were then centrifuged at maximum speed for 15 min at 4°C, and 200µl of the upper aqueous layer was transferred to a new tube and incubated with 2µl RNase A (20-40mg/ml, ThermoFisher Scientific) at 37°C for 1 h. The genomic DNA was precipitated with ethanol and resuspended in 25µl nuclease-free water. The samples were then sonicated using a Bioruptor Pico (Diagenode) for 30 cycles (30 sec "on", 30 sec "off") to yield 250-600bp fragments.

Library preparation and sequencing. Sequencing libraries were prepared using the TruSeq DNA Sample Prep kit v2-Set A (Illumina) for blunting, adding "A" overhangs, and adapter ligation. Excess adapters were removed by using AMPure XP beads (Beckman Coulter) and DNA was eluted in 38µl 10mM Tris-HCl, pH 8.5 (EB from Qiagen). Libraries were amplified with PCR (17 cycles – 30 sec at 98°C, 30 sec at 60°C, 30 sec at 72°C) using Phusion polymerase (ThermoFisher Scientific) and primers complementary to the adapters. The DNA was cleaned up using the MinElute PCR Purification kit (Qiagen) and eluted with 11µl EB. Finally, 300-600 bp fragments were purified from 1.5% agarose gel (1xTAE buffer) using the MinElute Gel Extraction kit (Qiagen). Library concentration was determined by Qubit using the dsDNA HS Assay Kit (ThermoFisher Scientific), and library quality and size range were determined by TapeStation (Agilent) using High Sensitivity D1000 ScreenTape (Agilent). Pooled library (18 total – nine from DNA and 9 from total RNA – see below) concentration was determined using a KAPA library quantification kit (Roche). Libraries were sequenced on a Illumina NextSeq 500 instrument (2x150bp paired-end reads, HighOutput mode v2.5).

Total RNA sequencing

Washed cell pellets were prepared as described in *DNA sequencing.* Total RNA extraction was performed using the RNeasy MinElute Cleanup kit (Qiagen). First, 600µl of RLT buffer with 1% β-mercaptoethanol and 200 mg of glass beads were added to the pellet, and the cells were lysed on an IKA Vibrax VXR for 20 min at 4°C with subsequent centrifugation at maximum speed for 2 min. The supernatant was transferred to a new tube and mixed with an equal volume of 70% ethanol, and transferred to an RNeasy column to proceed with RNA extraction using the manufacturer's protocol. RNA was eluted from the column with 22µl of nuclease-free water. Total

RNA was subsequently fragmented using the Ambion RNA fragmentation reagent and cleaned up with the RNeasy MinElute Cleanup kit (Qiagen). RNA was eluted from the column with 9µl of nuclease-free water.

First-strand cDNA synthesis. Random hexamer priming (8µl total RNA, 1µl random hexamers (Invitrogen), 1µl 10mM dNTPs) was used to synthesize complementary DNA. The samples were incubated at 90°C for 1 min then 65°C for 5 min, quickly chilled on ice for 1 min, and mixed with 2µl 10x RT buffer (ThermoFisher Scientific), 4µl 25mM $MgCl_2$, 2µl 0.1M DDT, 1µl 4U/µl RNase OUT (ThermoFisher Scientific), and 1µl 200U/µl SuperScript III reverse transcriptase (ThermoFisher Scientific). The samples were incubated at 25°C for 10 min followed by 50°C for 50 min. The reaction was heat-inactivated at 75°C for 15 min. dNTPs were removed with ethanol precipitation by adding 80µl water, 1µl glycogen, 10µl 3M sodium acetate pH 5.2 and 200µl 100% ethanol followed by incubation at -80°C for 20 min. The samples were centrifuged at 17,000 rcf for 20 min at 4°C, the supernatant was removed, and the residue was washed with 500µl cold 75% ethanol. The samples were centrifuged at 17,000 rcf for 10 min at 4°C, the supernatant and all the remaining liquid was removed, and the pellet was air-dried for a few minutes. We then resuspended the pellet in 51µl nuclease-free water, 1µl 10x RT buffer (ThermoFisher Scientific), 1µl 0.1M DTT, and 2µl 25mM $MgCl_2$.

Second-strand cDNA synthesis. We further added 15µl 5X second-strand buffer (ThermoFisher Scientific), 2µl 10mM dA/G/C/UTPs, 0.5µl 1U/µl *E. coli* DNA ligase, 2µl 10U/µl DNA Pol I, and 0.5 µl 5U/µl RNase H. The mixture was incubated at 16°C for 2 h and the resulting double-stranded DNA was purified using the MinElute PCR Purification kit (Qiagen) with 20µl EB for elution.

After that, the libraries were prepared and sequenced using the same protocol as described in *DNA sequencing*.

Structural analysis

Ribosome structures PDB IDs: 4v88 (assembly 1) [64] and 6woo [65] were visualized and analyzed in PyMOL. Since ribosomal parts undergo major rearrangements during translation [48,66], we considered variants that are directly in or within 5Å of the intersubunit bridges as potential candidates located in the functionally important regions. Coordinates of the intersubunit bridges were extracted from [64]. Analysis in **Supplementary Fig. 30** was made with local structural alignment of the region (18S:1110-1136 together with eL41) from 4v88 and 6woo in PyMOL. For **Fig. 6b** and **Supplementary Fig. 14c**, the NDB (http://ndbserver.rutgers.edu/) [67] was used to retrieve nucleotide interactions in the 25S rRNA (PDB ID: 4v88).

DATA AVAILABILITY

Sequence data generated in this study have been deposited in the SRA database with accession code **PRJNA867718**. Datasets generated in this study are included in its supplementary information files. **Supplementary File 1** is a tab-delimited table of filtered rDNA variants with their intragenomic frequencies. **Supplementary File 2** is a tab-delimited table of calculated Shannon's entropy values at each position of the rDNA. **Supplementary File 3** is a tab-delimited table of variant pairs with consistently similar frequencies across all isolates. **Supplementary File 4** is a tab-delimited table of variant pairs with consistently similar frequencies calculated by niche. **Supplementary File 5** is a tab-delimited table with annotations of variants in rRNA genes. Source data and additional files are provided with this paper and are also deposited at https://github.com/hochwagenlab: 1) raw_rDNA_var_calls - raw rDNA .vcf files prior to additional filtering; 2) Sequencing_Sultanov_etal – raw rDNA .vcf files (for DNA- and total RNA sequencing) generated in this study and the rDNA coverage for each sample; 3) rDNA_S288c.fsa - the S288c rDNA copy prototype sequence used in this study; 4) rDNA_S288c.bed - annotations associated with the rDNA prototype; 5) rDNA_S288c_benchmark.bed - annotations for benchmarking; 6) positions_in_homopolymers.txt – nucleotide positions in the S288c rDNA prototype that are embedded in the 10-nt poly(A/T/G/C) sequences.

CODE AVAILABILITY

The computer scripts used for processing and analyzing rDNA sequencing reads are available at https://github.com/hochwagenlab.

ACKNOWLEDGEMENTS

We thank Matthew Rockman, David Gresham, Mark Siegal, Loren Williams, David Fitch for helpful discussions, and Joseph Schacherer for sharing strains. 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.

REFERENCES

1. 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 Interdiscip Rev RNA* **6**, 225–242 (2015).
2. 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. *Mol Gen Genet* **233**, 355–362 (1992).
3. Miller, C. A. & Kowalski, D. cis-acting components in the replication origin from ribosomal DNA of Saccharomyces cerevisiae. *Mol Cell Biol* **13**, 5360–5369 (1993).
4. 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 Res* **28**, 3524–3534 (2000).
5. Long, E. O. & Dawid, I. B. Repeated Genes in Eukaryotes. *Annu Rev Biochem* **49**, 727–764 (1980).
6. Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res* **18**, 13–18 (2008).
7. Paredes, S. & Maggert, K. A. Expression of *I-Cre* I Endonuclease Generates Deletions Within the *rDNA* of Drosophila. *Genetics* **181**, 1661–1671 (2009).
8. Ide, S., Miyazaki, T., Maki, H. & Kobayashi, T. Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity. *Science (1979)* **327**, 693–696 (2010).
9. Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S. & Lemos, B. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proceedings of the National Academy of Sciences* **112**, 2485–2490 (2015).
10. Hotz, M. *et al.* rDNA array length is a major determinant of replicative lifespan in budding yeast. *Proceedings of the National Academy of Sciences* **119**, (2022).
11. Dover, G. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).
12. 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 Res* **17**, 184–191 (2007).
13. 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. *Mol Biol Evol* **28**, 2883–2891 (2011).
14. Elder, J. F. & Turner, B. J. Concerted evolution of repetitive DNA sequences in eukaryotes. *Q Rev Biol* **70**, 297–320 (1995).
15. Pawlowska, T. E. & Taylor, J. W. Organization of genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature* **427**, 733–737 (2004).
16. Simon, U. K. & Weiss, M. Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thought. *Mol Biol Evol* **25**, 2251–2254 (2008).
17. 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. *Syst Biol* **63**, 543–554 (2014).
18. Hugall, A., Stanton, J. & Moritz, C. Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic Meloidogyne. *Mol Biol Evol* **16**, 157–164 (1999).

19. Parkin, E. J. Within- and Between-Individual Sequence Variation Among ITS1 Copies in the Meadow Grasshopper Chorthippus parallelus Indicates Frequent Intrachromosomal Gene Conversion. *Mol Biol Evol* **21**, 1595–1601 (2004).
20. McTaggart, S. J. & Crease, T. J. Selection on the Structural Stability of a Ribosomal RNA Expansion Segment in Daphnia obtusa. *Mol Biol Evol* **22**, 1309–1319 (2005).
21. 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).
22. 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. *Nat Commun* **12**, 387 (2021).
23. Tseng, H. *et al.* Mouse Ribosomal RNA Genes Contain Multiple Differentially Regulated Variants. *PLoS One* **3**, e1843 (2008).
24. Holland, M. L. *et al.* Early-life nutrition modulates the epigenetic state of specific rDNA genetic variants in mice. *Science (1979)* **353**, 495–498 (2016).
25. Parks, M. M. *et al.* Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci Adv* **4**, (2018).
26. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu Rev Genet* **39**, 121–152 (2005).
27. 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).
28. 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).
29. Piontkivska, H., Rooney, A. P. & Nei, M. Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. *Mol Biol Evol* **19**, 689–697 (2002).
30. Fitch, D. H., Mainone, C., Goodman, M. & Slightom, J. L. Molecular history of gene conversions in the primate fetal gamma-globin genes. Nucleotide sequences from the common gibbon, Hylobates lar. *J Biol Chem* **265**, 781–93 (1990).
31. Fitch, D. H. *et al.* Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proceedings of the National Academy of Sciences* **88**, 7396–7400 (1991).
32. Storz, J. F. *et al.* Complex Signatures of Selection and Gene Conversion in the Duplicated Globin Genes of House Mice. *Genetics* **177**, 481–500 (2007).
33. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
34. 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 Res* **40**, 11189–11201 (2012).
35. de Chiara, M. *et al.* Domestication reprogrammed the budding yeast life cycle. *Nat Ecol Evol* **6**, 448–460 (2022).
36. Fox, G. E. Origin and Evolution of the Ribosome. *Cold Spring Harb Perspect Biol* **2**, a003483–a003483 (2010).

37. 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 Res* **29**, 5017–5028 (2001).
38. Hsiao, C., Mohan, S., Kalahar, B. K. & Williams, L. D. Peeling the Onion: Ribosomes Are Ancient Molecular Fossils. *Mol Biol Evol* **26**, 2415–2425 (2009).
39. Bernier, C. R. *et al.* RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.* **169**, 195–207 (2014).
40. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu Rev Biochem* **74**, 129–177 (2005).
41. Yokoyama, T. & Suzuki, T. Ribosomal RNAs are tolerant toward genetic insertions: evolutionary origin of the expansion segments. *Nucleic Acids Res* **36**, 3539–3551 (2008).
42. Petrov, A. S. *et al.* Evolution of the ribosome at atomic resolution. *Proceedings of the National Academy of Sciences* **111**, 10251–10256 (2014).
43. 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).
44. Doris, S. M. *et al.* Universal and domain-specific sequences in 23S–28S ribosomal RNA identified by computational phylogenetics. *RNA* **21**, 1719–1730 (2015).
45. Taylor, D. J. *et al.* Structures of modified eEF2·80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J* **26**, 2421–2431 (2007).
46. Gulay, S. P. *et al.* Tracking fluctuation hotspots on the yeast ribosome through the elongation cycle. *Nucleic Acids Res* **45**, 4958–4971 (2017).
47. Velichutina, I. v., Rogers, M. J., McCutchan, T. F. & Liebman, S. W. Chimeric rRNAs containing the GTPase centers of the developmentally regulated ribosomal rRNAs of Plasmodium falciparum are functionally distinct. *RNA* **4**, S1355838298980049 (1998).
48. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nat Struct Mol Biol* **19**, 560–567 (2012).
49. 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).
50. 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).
51. LaRiviere, F. J., Cole, S. E., Ferullo, D. J. & Moore, M. J. A Late-Acting Quality Control Process for Mature Eukaryotic rRNAs. *Mol Cell* **24**, 619–626 (2006).
52. 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. *Mol Cell* **34**, 440–450 (2009).
53. Hage, A. el & Tollervey, D. A Surfeit of Factors: Why is Ribosome Assembly So Much More Complicated in Eukaryotes than Bacteria? *RNA Biol* **1**, 9–14 (2004).
54. 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. *Mol Cell Biol* **23**, 1558–1568 (2003).
55. Chernoff, Y. O., Vincent, A. & Liebman, S. W. Mutations in eukaryotic 18S ribosomal RNA affect translational fidelity and resistance to aminoglycoside antibiotics. *EMBO J* **13**, 906–913 (1994).

56. Kurylo, C. M. *et al.* Endogenous rRNA Sequence Variation Can Regulate Stress Response Gene Expression and Phenotype. *Cell Rep* **25**, 236-248.e6 (2018).
57. Porter, T. M. & Hajibabaei, M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol* **27**, 313–338 (2018).
58. Yue, J.-X. *et al.* Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**, 913–924 (2017).
59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 666 (2012).
62. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
63. Stephens, Z. D. *et al.* Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLoS One* **11**, e0167047 (2016).
64. Ben-Shem, A. *et al.* The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. *Science (1979)* **334**, 1524–1529 (2011).
65. Wang, J. *et al.* Structural basis for the transition from translation initiation to elongation by an 80S-eIF5B complex. *Nat Commun* **11**, 5003 (2020).
66. Behrmann, E. *et al.* Structural Snapshots of Actively Translating Human Ribosomes. *Cell* **161**, 845–857 (2015).
67. Coimbatore Narayanan, B. *et al.* The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res* **42**, D114–D122 (2014).

FIGURES

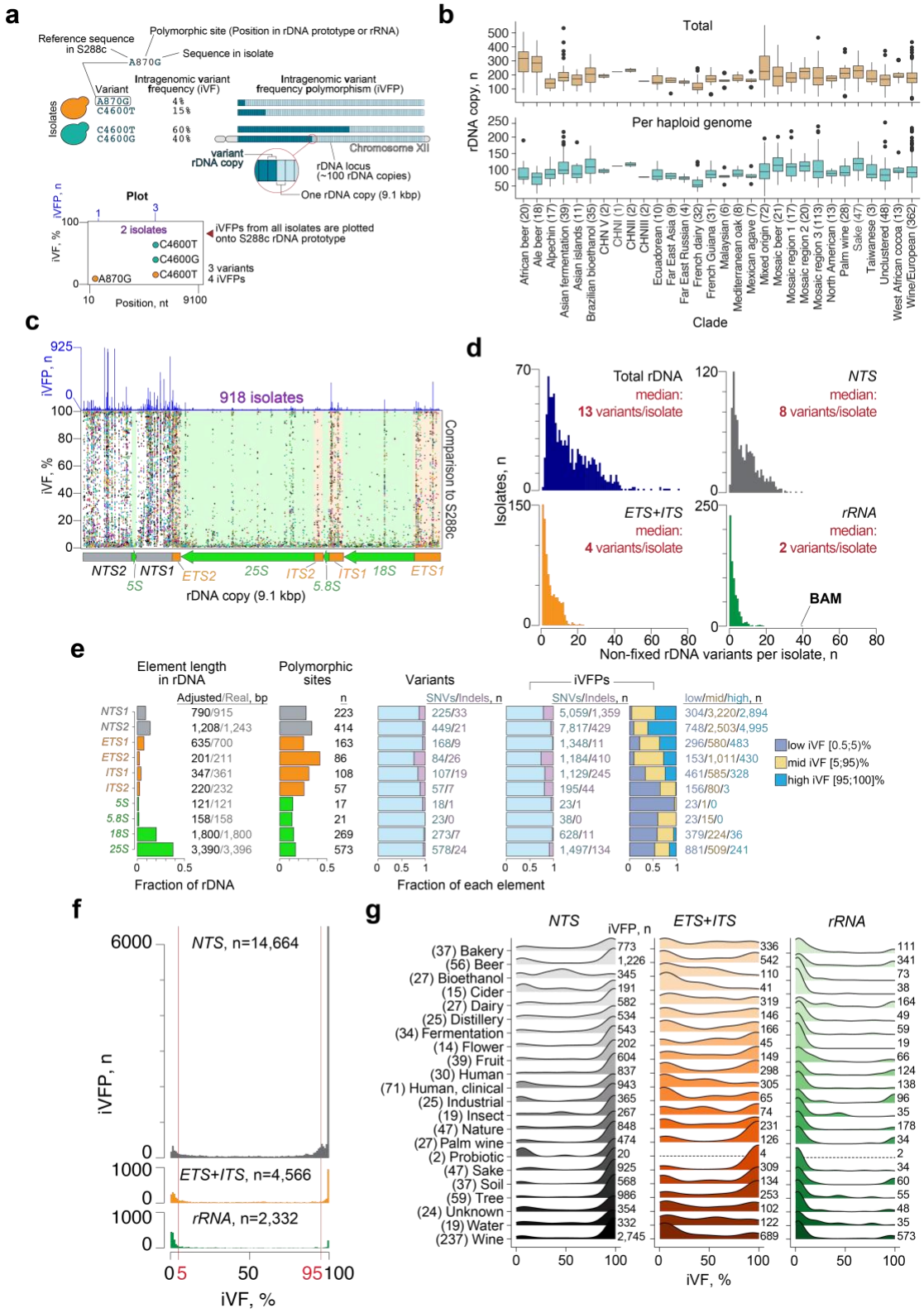

a
Reference sequence in S288c
Polymorphic site (Position in rDNA prototype or rRNA)
Sequence in isolate
A870G
Variant
Intragenomic variant frequency (iVF)
Intragenomic variant frequency polymorphism (iVFP)
Isolates
A870G 4%
C4600T 15%
C4600T 60%
C4600G 40%
Chromosome XII
variant rDNA copy
rDNA locus (~100 rDNA copies)
One rDNA copy (9.1 kbp)
Plot
iVFP, n
iVF, %
2 isolates
C4600T
C4600G
A870G
C4600T
iVFPs from all isolates are plotted onto S288c rDNA prototype
3 variants
4 iVFPs
Position, nt
b
Total
Per haploid genome
rDNA copy, n
Clade
African beer (20)
Ale beer (18)
Alpechin (17)
Asian fermentation (39)
Asian islands (11)
Brazilian bioethanol (35)
CHN V (2)
CHNI (1)
CHNII (2)
CHNIII (2)
Ecuadorean (10)
Far East Asia (9)
Far East Russian (4)
French dairy (32)
French Guiana (31)
Malaysian (6)
Mediterranean oak (8)
Mexican agave (7)
Mixed origin (72)
Mosaic beer (21)
Mosaic region 1 (17)
Mosaic region 2 (20)
Mosaic region 3 (113)
North American (13)
Palm wine (28)
Sake (47)
Taiwanese (3)
Unclustered (48)
West African cocoa (13)
Wine/European (362)
c
iVFP, n
918 isolates
iVF, %
Comparison to S288c
NTS2
NTS1
ETS2
25S
ITS2
ITS1
18S
ETS1
5S
5.8S
rDNA copy (9.1 kbp)
d
Total rDNA
median:
13 variants/isolate
NTS
median:
8 variants/isolate
ETS+ITS
median:
4 variants/isolate
rRNA
median:
2 variants/isolate
BAM
Isolates, n
Non-fixed rDNA variants per isolate, n
e
Element length in rDNA
Adjusted/Real, bp
Polymorphic sites
n
Variants
SNVs/Indels, n
iVFPs
SNVs/Indels, n
low/mid/high, n
NTS1 790/915 223 225/33 5,059/1,359 304/3,220/2,894
NTS2 1,208/1,243 414 449/21 7,817/429 748/2,503/4,995
ETS1 635/700 163 168/9 1,348/11 296/580/483
ETS2 201/211 86 84/26 1,184/410 153/1,011/430
ITS1 347/361 108 107/19 1,129/245 461/585/328
ITS2 220/232 57 57/7 195/44 156/80/3
5S 121/121 17 18/1 23/1 23/1/0
5.8S 158/158 21 23/0 38/0 23/15/0
18S 1,800/1,800 269 273/7 628/11 379/224/36
25S 3,390/3,396 573 578/24 1,497/134 881/509/241
low iVF [0.5;5)%
mid iVF [5;95)%
high iVF [95;100]%
Fraction of rDNA
Fraction of each element
f
NTS, n=14,664
ETS+ITS, n=4,566
rRNA, n=2,332
iVFP, n
iVF, %
g
NTS
ETS+ITS
rRNA
iVFP, n
(37) Bakery 773 336 111
(56) Beer 1,226 542 341
(27) Bioethanol 345 110 73
(15) Cider 191 41 38
(27) Dairy 582 319 164
(25) Distillery 534 146 49
(34) Fermentation 543 166 59
(14) Flower 202 45 19
(39) Fruit 604 149 66
(30) Human 837 298 124
(71) Human, clinical 943 305 138
(25) Industrial 365 65 96
(19) Insect 267 74 35
(47) Nature 848 231 178
(27) Palm wine 474 126 34
(2) Probiotic 20 4 2
(47) Sake 925 309 34
(37) Soil 568 134 60
(59) Tree 986 253 55
(24) Unknown 354 102 48
(19) Water 332 122 35
(237) Wine 2,745 689 573
iVF, %

Fig. 1. rDNA sequence variants across 918 *S. cerevisiae* isolates. a, variant annotation. Variants are defined as nucleotide differences from the S288c prototype (light blue). Intragenomic variant frequency (iVF) is the percentage of rDNA reads that contain a variant sequence within an isolate's genome and is proportional to variant rDNA copies within a genome. Intragenomic variant frequency polymorphism (iVFP) represents a polymorphism of rDNA copy number containing a variant (dark blue), which can differ between isolates. Note that one isolate can have multiple sequences at a given position and thus the sum of the iVFPs can be more than the total number of the sequenced isolates. *Plot* – schematic representation of the iVFPs from the two isolates in one plot. **b**, Total rDNA copy number distribution and per haploid genome across isolates grouped by clade. Numbers in brackets indicate the number of isolates in each clade. **c**, iVFPs 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 iVFPs at a specific position across all isolates. **d**, distribution of the number of non-fixed rDNA variants (those that are present in a maximum of all but one rDNA copy in a genome; see Methods) based on an rDNA element. "BAM" highlights an isolate with an exceptionally high number of rRNA variants. **e**, summary statistics for genic elements in the rDNA. Element length in rDNA - the relative length of each element in the rDNA; Polymorphic sites – the fraction of all variable sites in an element; Variants – the number of observed variants and fraction of single-nucleotide variants (SNVs) and indels; iVFP – the number of iVFPs across all isolates and the fraction SNVs and indels. The numbers indicate absolute values for each distribution. The adjusted element lengths are due to filtering (edge sequences and homopolymer regions; see Mehods). **f**, distribution of iVFPs by iVF in *NTS*, *ETS+ITS*, and *rRNA*. **g**, same analysis as (**f**) but separated by ecological niche. The number of isolates in each niche is indicated in brackets, and the number of iVFPs is shown on the right of each distribution. The missing distributions in the Probiotic niche are due to the low number of iVFPs.

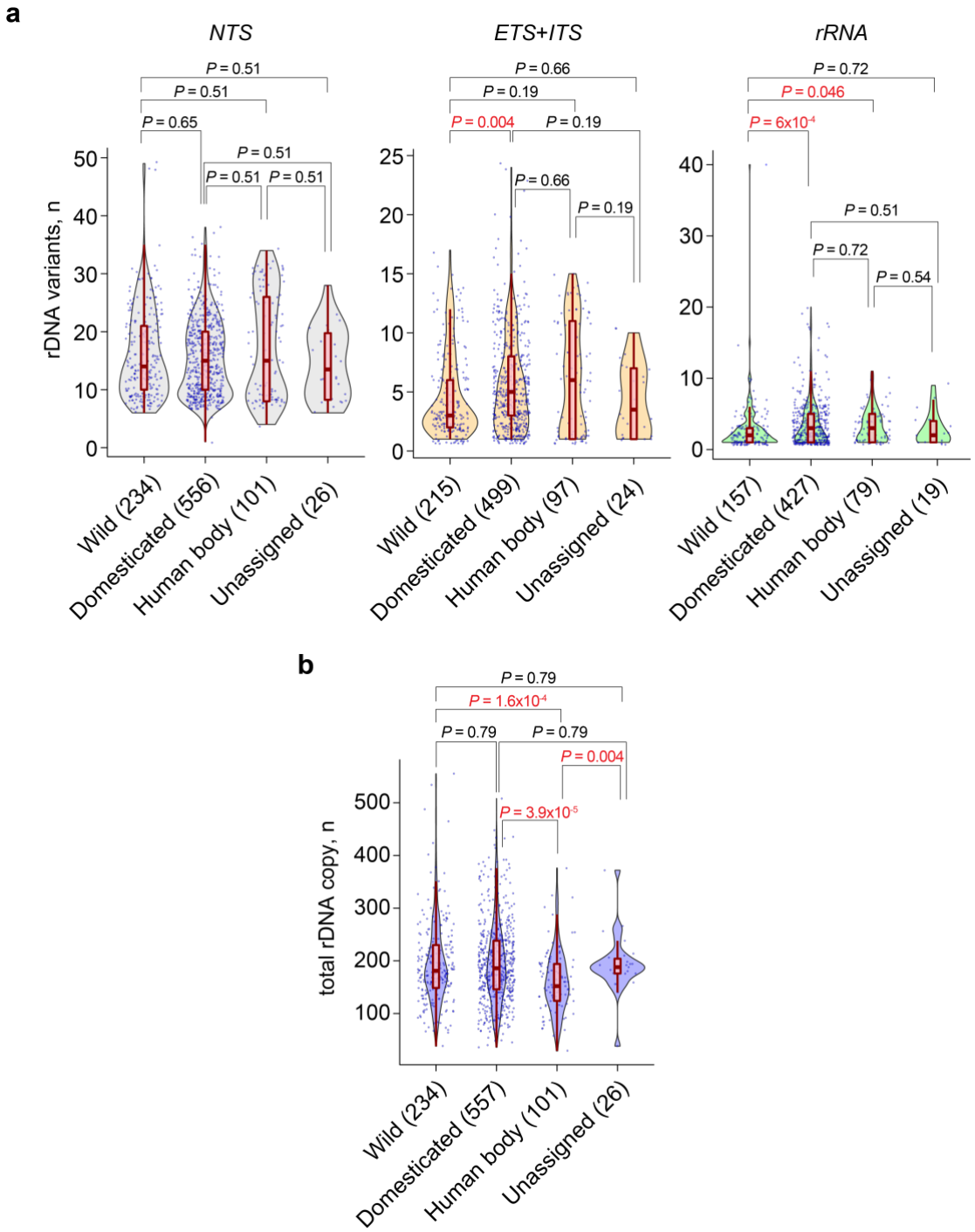


Fig. 2. rDNA variation is stratified by habitat. a, distribution of the number of rDNA variants across different habitats from [35] grouped by rDNA element. **b**, distribution of total rDNA copy numbers across the habitats. The red box plot inside indicates the median. *P*-values were calculated with the Wilcoxon rank sum exact test and adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction.

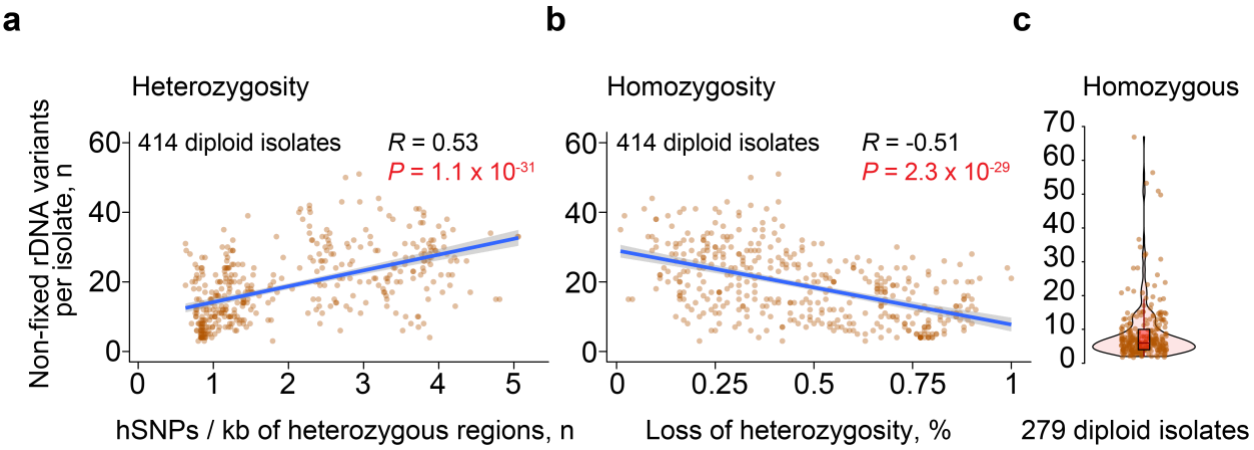


Fig. 3. Relationship between sequence diversity of the rDNA and the rest of the genome in diploid isolates. Correlation between the number of rDNA variants and (**a**) the number of heterozygous SNPs per heterozygous region of the genome (excluding rDNA) [33] or (**b**) the loss of heterozygosity per 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).

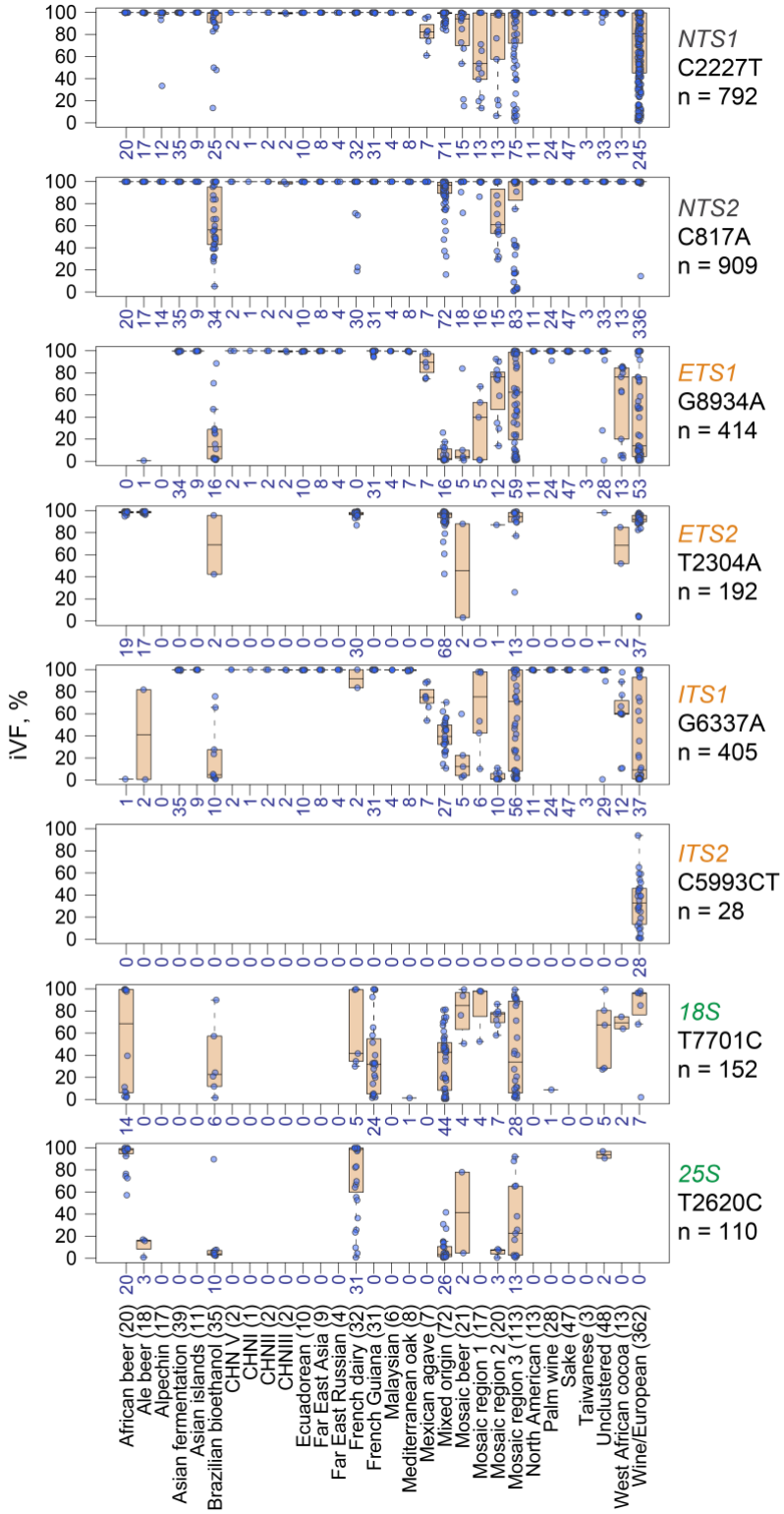


Fig. 4. Boxplot distribution of the topmost abundant rDNA variants across clades. Numbers indicate both the number of isolates with a variant per clade and the number of iVFPs of a specific variant, since one isolate can only have one iVFP of each variant per plot.

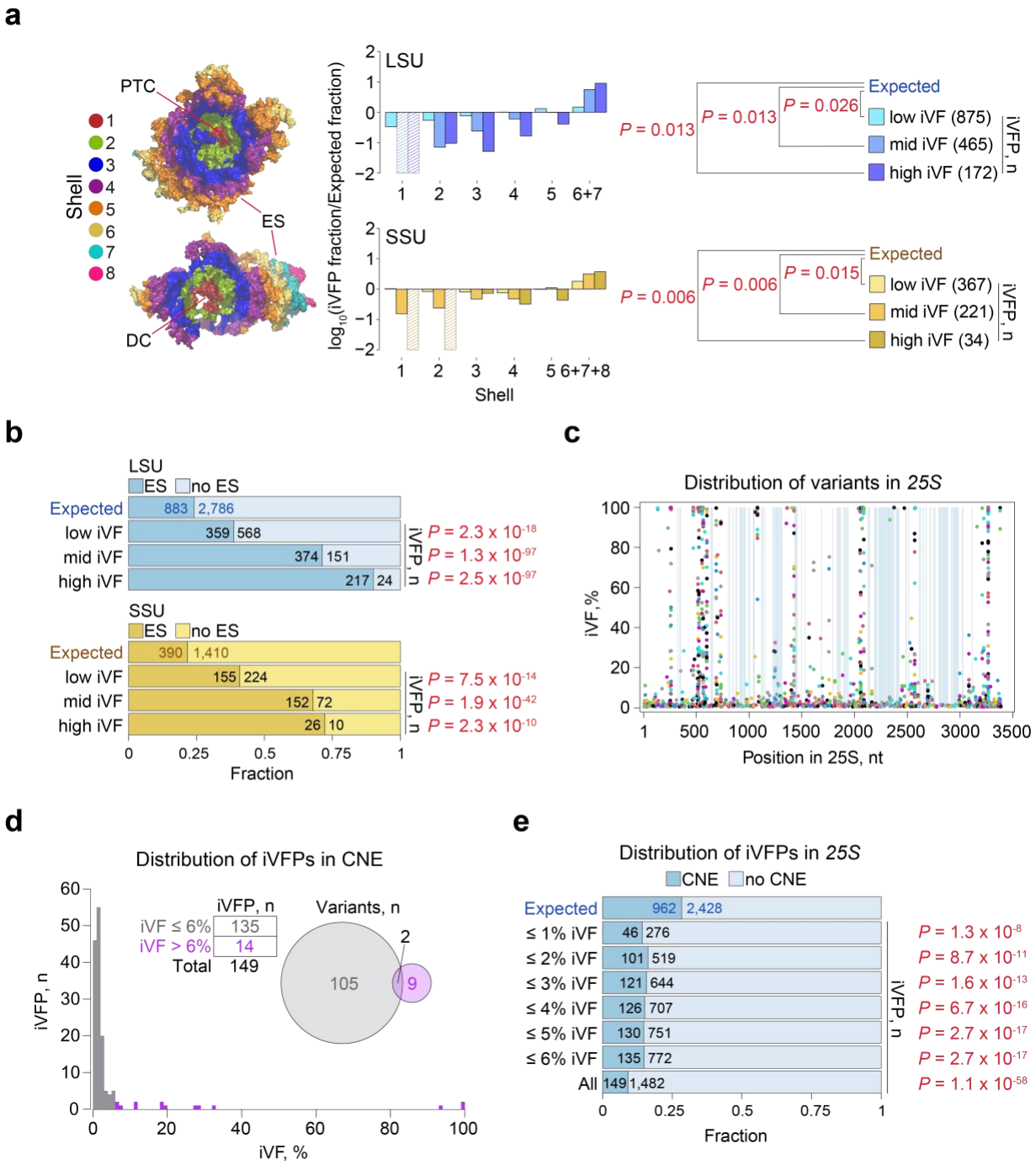

a
PTC
ES
DC
Shell
1
2
3
4
5
6
7
8
LSU
SSU
log10(iVFP fraction/Expected fraction)
Shell
1 2 3 4 5 6+7
1 2 3 4 5 6+7+8
P = 0.013
P = 0.013
P = 0.026
Expected
low iVF (875)
mid iVF (465)
high iVF (172)
iVFP, n
P = 0.006
P = 0.006
P = 0.015
Expected
low iVF (367)
mid iVF (221)
high iVF (34)
iVFP, n
b
LSU
ES no ES
Expected 883 2,786
low iVF 359 568
mid iVF 374 151
high iVF 217 24
iVFP, n
P = 2.3 x 10-18
P = 1.3 x 10-97
P = 2.5 x 10-97
SSU
ES no ES
Expected 390 1,410
low iVF 155 224
mid iVF 152 72
high iVF 26 10
iVFP, n
P = 7.5 x 10-14
P = 1.9 x 10-42
P = 2.3 x 10-10
0 0.25 0.5 0.75 1
Fraction
c
Distribution of variants in 25S
iVF, %
Position in 25S, nt
d
Distribution of iVFPs in CNE
iVFP, n
iVF, %
iVFP, n
iVF ≤ 6% 135
iVF > 6% 14
Total 149
Variants, n
105
2
9
e
Distribution of iVFPs in 25S
CNE no CNE
Expected 962 2,428
≤ 1% iVF 46 276
≤ 2% iVF 101 519
≤ 3% iVF 121 644
≤ 4% iVF 126 707
≤ 5% iVF 130 751
≤ 6% iVF 135 772
All 149 1,482
iVFP, n
P = 1.3 x 10-8
P = 8.7 x 10-11
P = 1.6 x 10-13
P = 6.7 x 10-16
P = 2.7 x 10-17
P = 2.7 x 10-17
P = 1.1 x 10-58
0 0.25 0.5 0.75 1
Fraction

Fig. 5. Selection in the yeast 80S ribosome. a, distribution of the number of low, medium and high iVFPs observed (numbers are indicated in brackets) versus expected in both the LSU and the SSU across their respective concentric shells (see text; depicted to the *left*). The outermost shells in the LSU and the SSU were pooled together due to the low number of nucleotides. Shells 1 in the LSU and 1 and 2 in the SSU did not have any variants and their log values were artificially set to -2 (hashed bars). Red lines point to the location of: PTC – peptidyltransferase center; DC – decoding center; and ES – expansion segments. **b**, distribution of iVFPs 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 iVFPs that are detected in the CNEs. *Inset* – number of iVFPs in two groups (below and above 6%; *left*) and the overlap between the variants from the two groups (*right*; two variants have iVFPs with iVFs both below and above 6% iVF across different isolates). **e**, distribution of iVFPs with different iVF cutoffs within ("CNE") and outside ("no CNE") 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). All reported *P*-values are adjusted for multiple hypothesis testing using Benjamini-Hochberg correction.

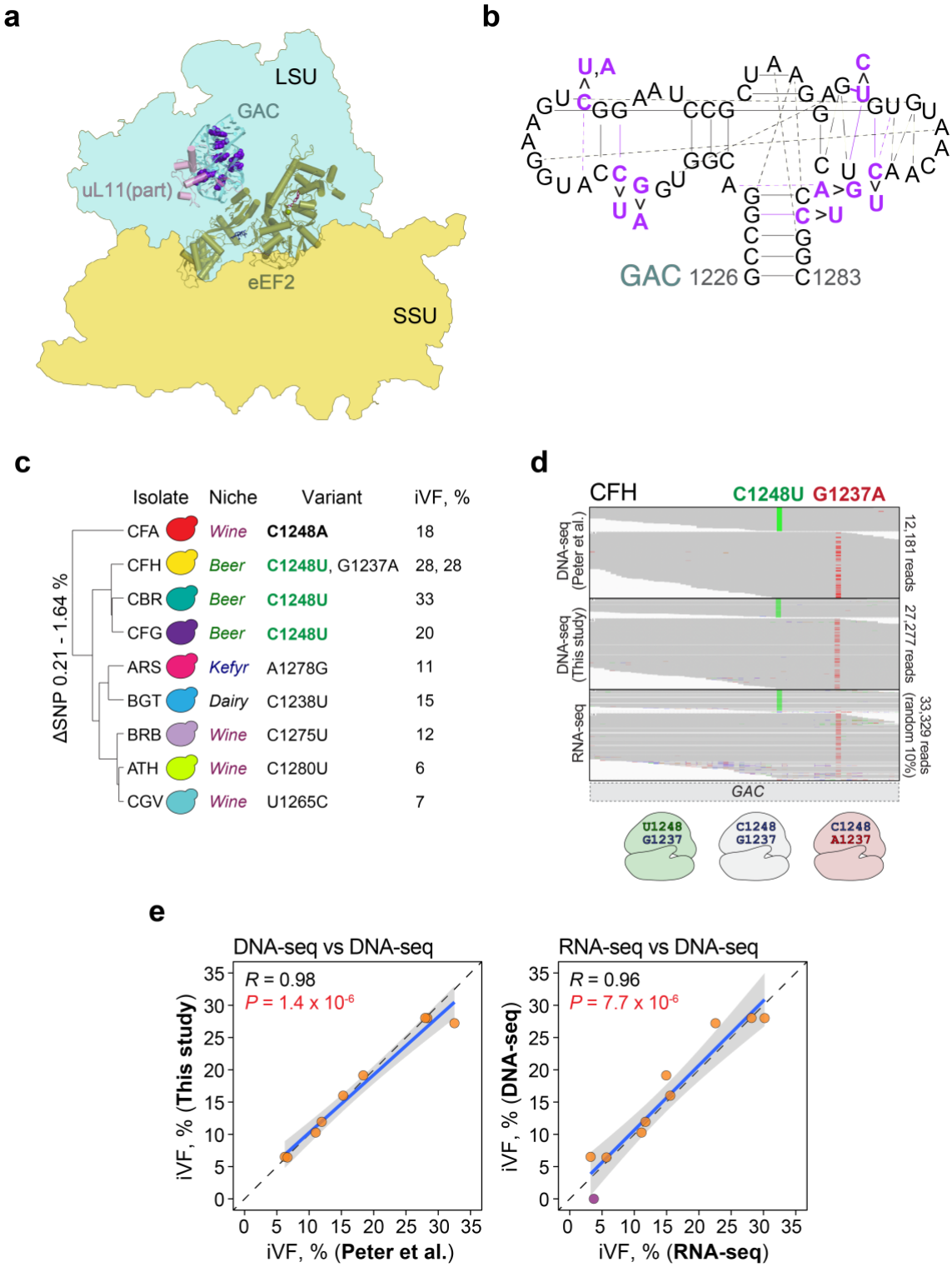
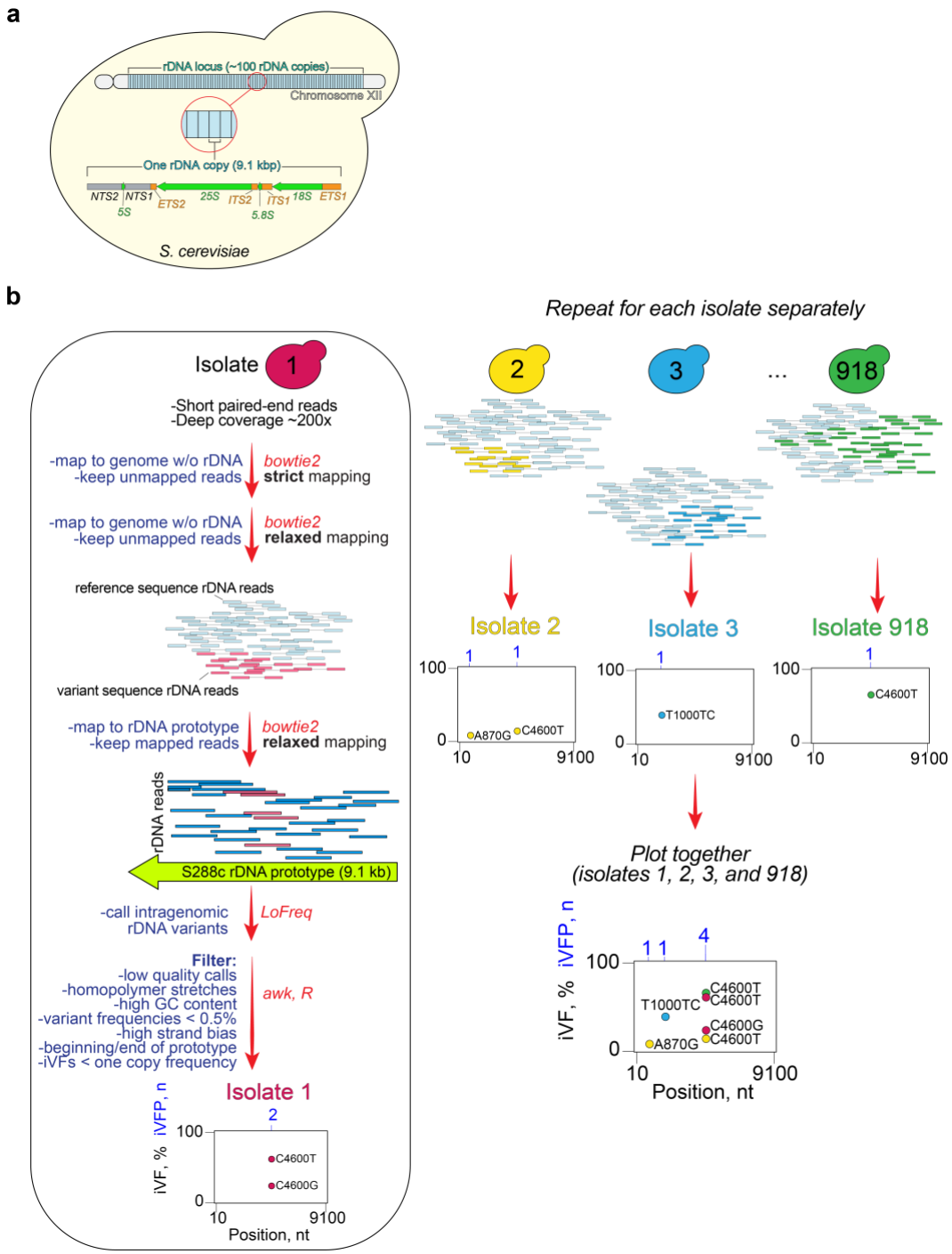
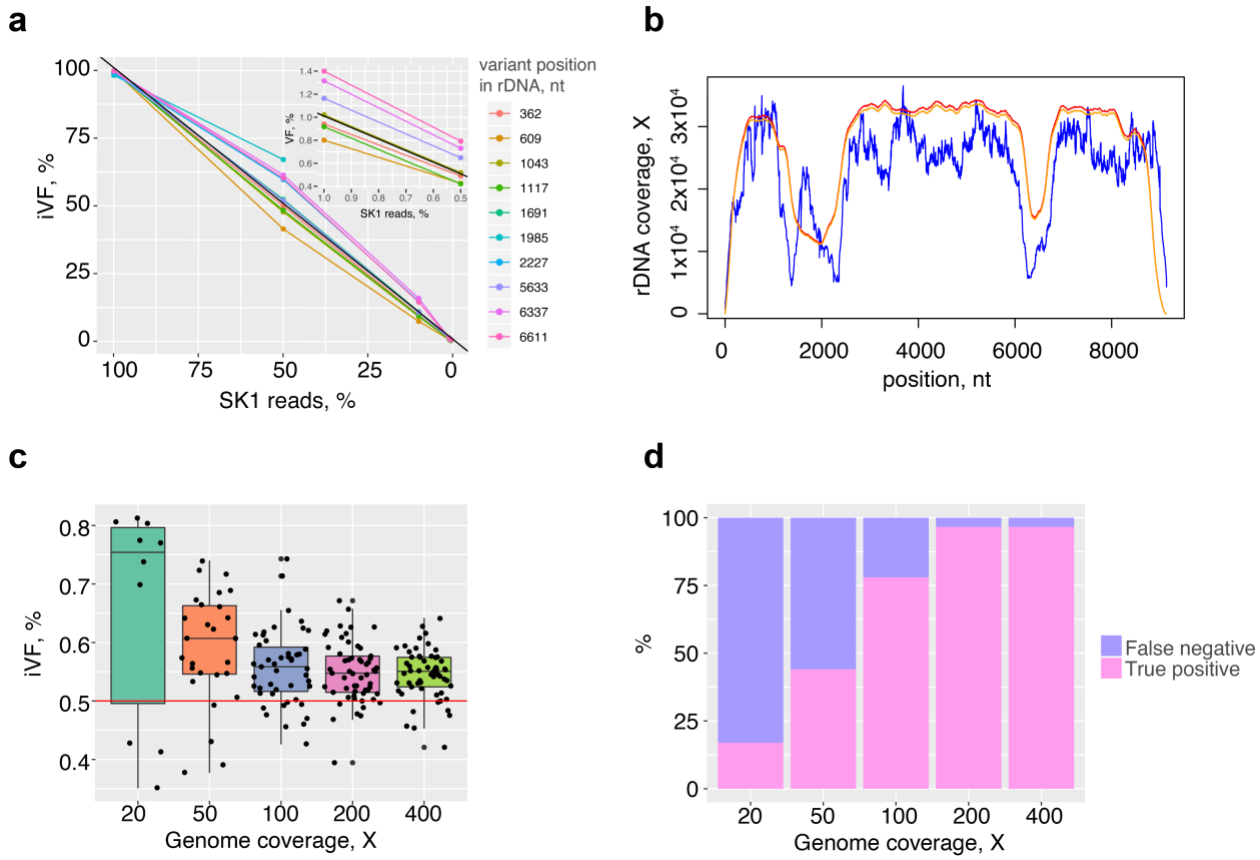

a
LSU
GAC
uL11(part)
eEF2
SSU
b
GAC 1226 G—C1283
c
Isolate Niche Variant iVF, %
CFA Wine C1248A 18
CFH Beer C1248U, G1237A 28, 28
CBR Beer C1248U 33
CFG Beer C1248U 20
ARS Kefyr A1278G 11
BGT Dairy C1238U 15
BRB Wine C1275U 12
ATH Wine C1280U 6
CGV Wine U1265C 7
ΔSNP 0.21 - 1.64 %
d
CFH
C1248U G1237A
DNA-seq (Peter et al.)
12,181 reads
DNA-seq (This study)
27,277 reads
RNA-seq
33,329 reads (random 10%)
GAC
U1248 G1237
C1248 G1237
C1248 A1237
e
DNA-seq vs DNA-seq
R = 0.98
P = 1.4 x 10-6
iVF, % (This study)
iVF, % (Peter et al.)
RNA-seq vs DNA-seq
R = 0.96
P = 7.7 x 10-6
iVF, % (DNA-seq)
iVF, % (RNA-seq)

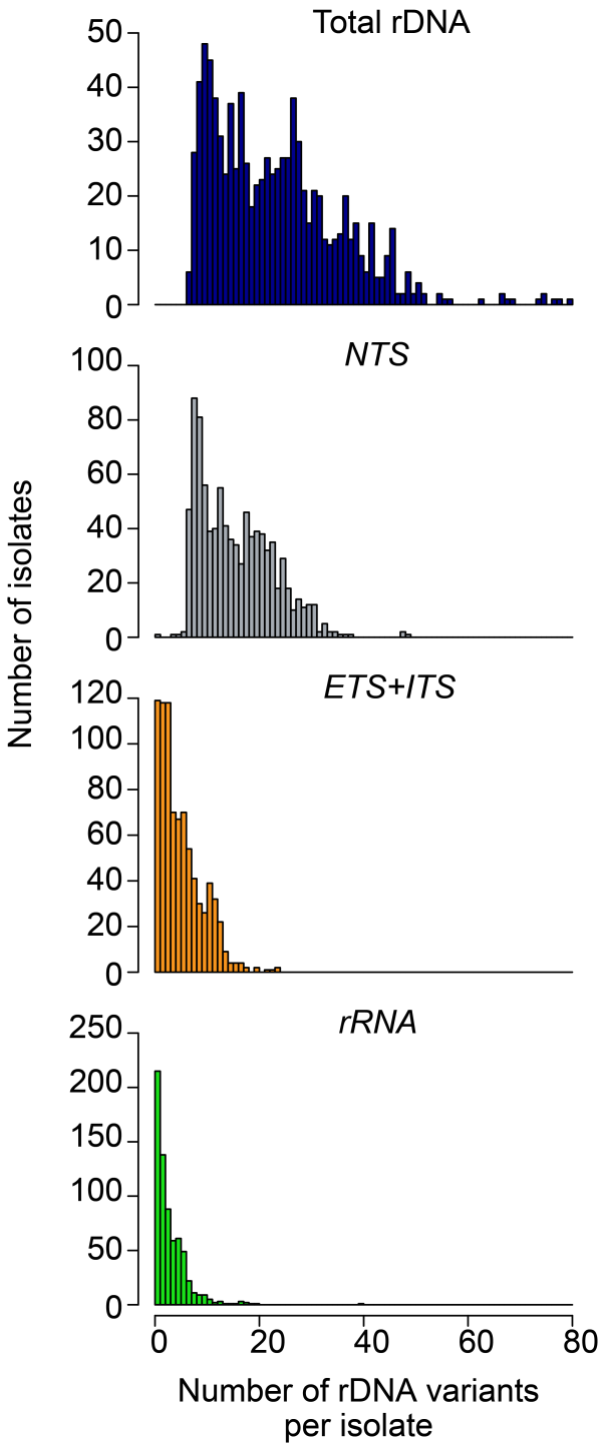
Fig. 6. Variability in the GTPase activation center. a, the GAC (blue) interacts with translational GTPase eEF2 (green) in the 80S yeast ribosome (PDB ID: 6woo; LSU outline is shown in blue, SSU outline is shown in yellow). Part of ribosomal protein uL11 is shown in pink. Variant positions are shown in purple with the spherical representation. **b**, the 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. **c**, hierarchical clustering of the isolates with SNVs in the GAC (25S nomenclature) based on genome-wide single nucleotide polymorphism (SNP) differences (excluding rDNA variants) from [33]. Bold highlights a position with more than one variant; green indicates a shared variant. **d**, continuous rRNA and rDNA reads (from this study and [33]) 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. Due to the large read coverage of the RNA-seq data (over 300,000 continuous GAC reads), only random 10% of the reads were sampled and are shown here. **e**, correlation between iVFs of the GAC variants detected from DNA-sequencing ("DNA-seq") performed in this study compared to the published DNA-sequencing data from [33] (*left*) and correlation between iVFs of the variants in the genome ("DNA-seq") and their expression in RNA ("RNA-seq") performed in this study. Isolate "CFG" is excluded because of inconsistent sequencing results between this study and [33]. Each dot is a variant; the purple dot represents a variant detected in RNA-sequencing but not in DNA-sequencing data. The diagonal shows theoretical one-to-one correspondence between iVFs of the two experiments.



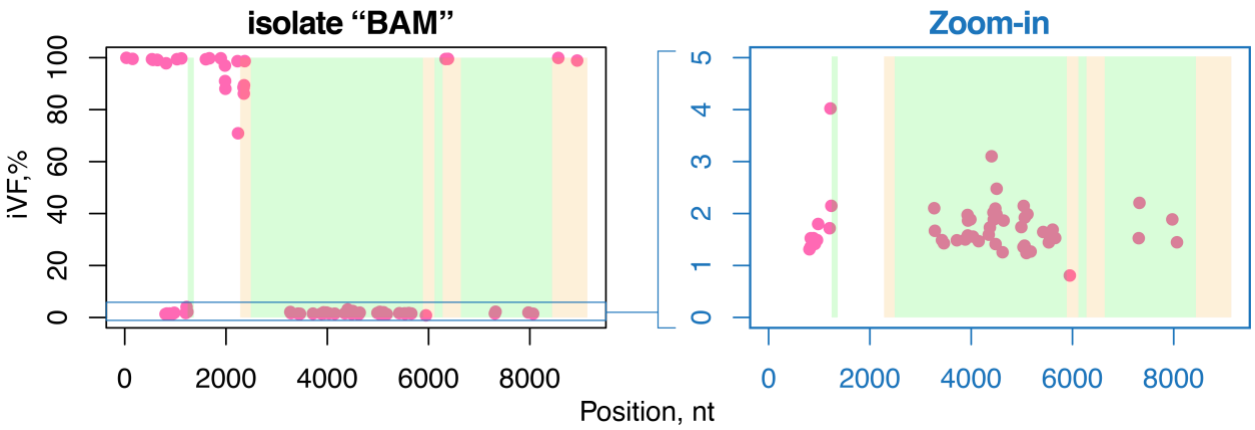
Supplementary Figure 1. Pipeline overview. a, organization of the rDNA locus in *S. cerevisiae*. About 100 rDNA copies are encoded in a head-to-tail fashion on chromosome XII. Each rDNA copy is 9.1 kbp long and contains rRNA genes (*25S*, *18S*, *5.8S*, *5S*; green), gene regulatory elements (*NTS1*, *NTS2*; grey), and transcribed elements that control rRNA processing (*ETS1*, *ETS2*, *ITS1*, *ITS2*; orange). **b**, bioinformatics pipeline described in Methods. The pipeline was run on each isolate separately.



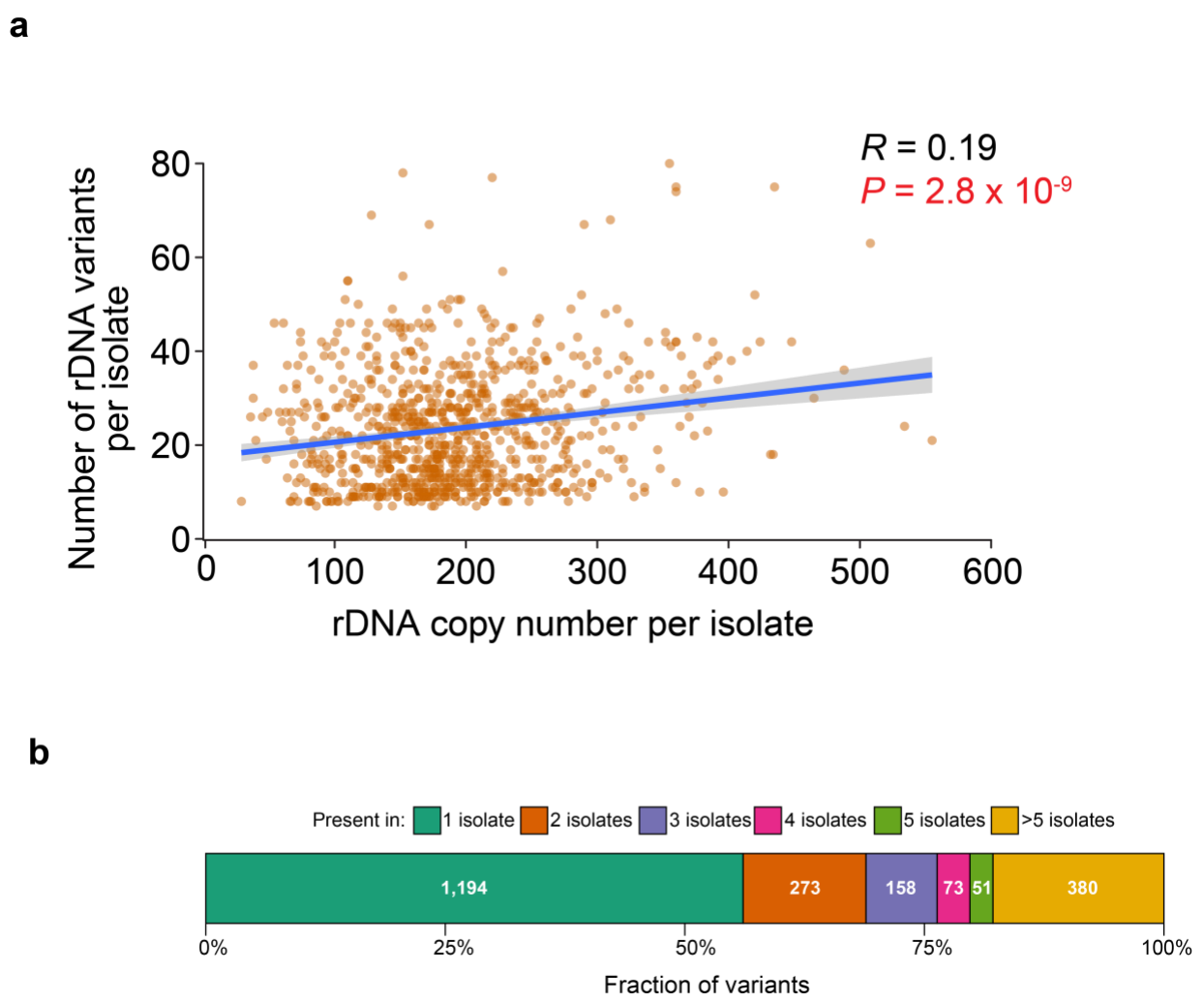
Supplementary Figure 2. Pipeline benchmarking. a, *in silico* spiking of S288c read data with 10 highly confident SNVs from another lab strain (SK1) across the rDNA prototype (positions are shown in the legend). The black diagonal represents true iVF values. *inset* - VFs from 1 to 0.5%. **b**, 200X 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 iVF = 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).



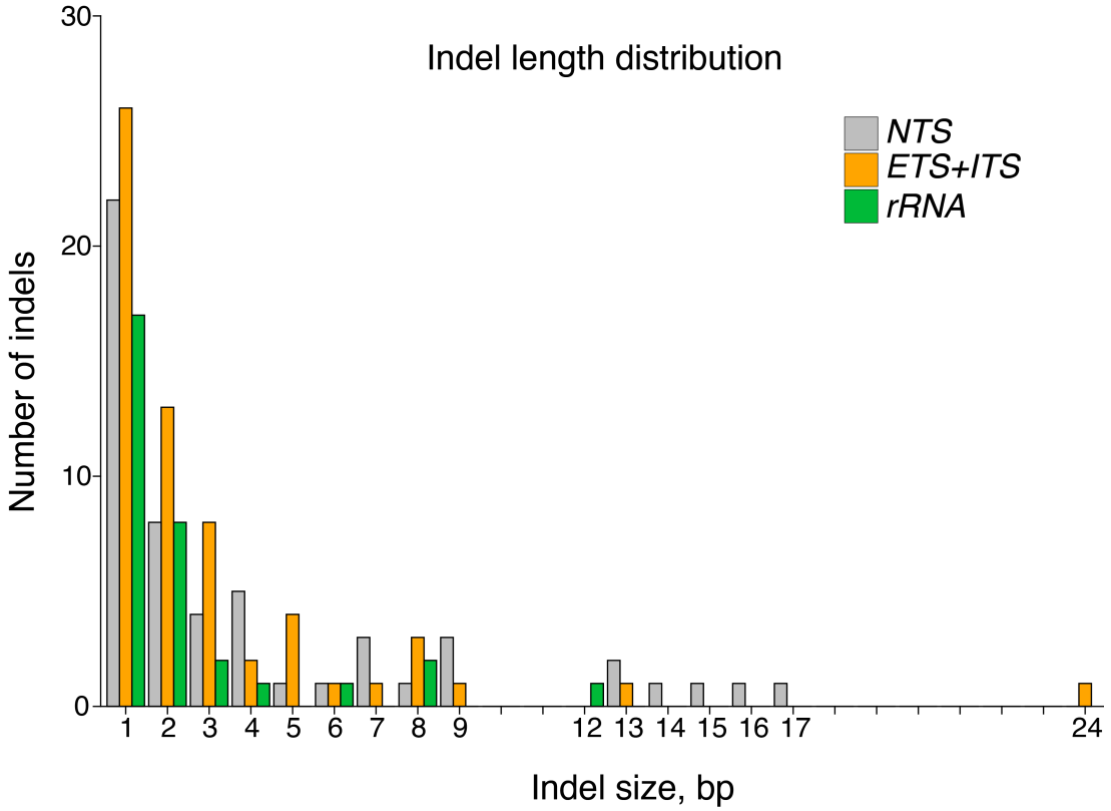
Supplementary Figure 3. Distribution of the number of rDNA variants per isolate. NTS – *NTS1* and *NTS2*; ETS+ITS – *ETS1*, *ETS2*, *ITS1* and *ITS2*; rRNA – rRNA genes (*5S*, *5.8S*, *18S* and *25S*).



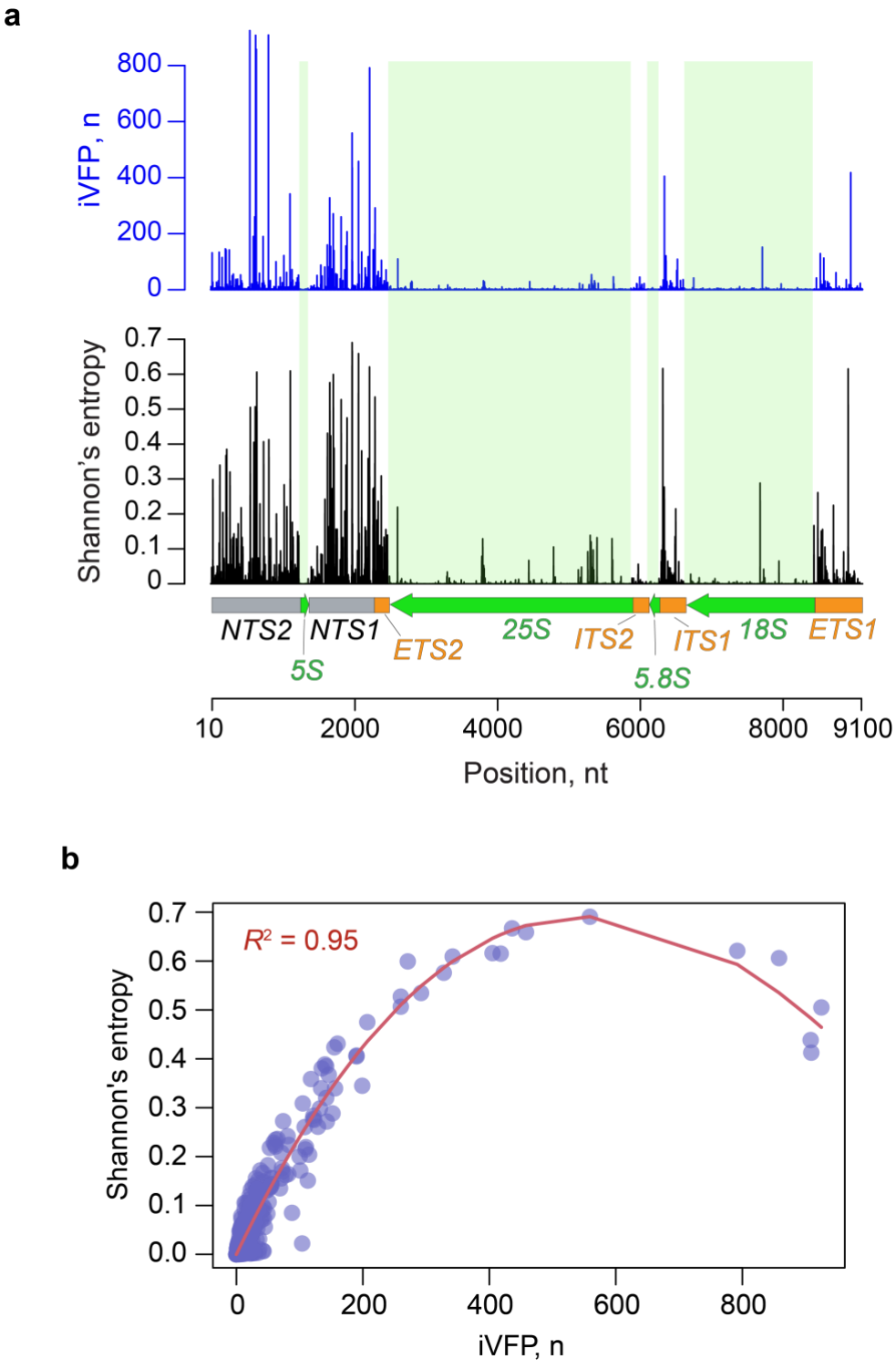
Supplementary Figure 4. Distribution of variants and their iVFs in the "BAM" isolate. Each dot is a variant.



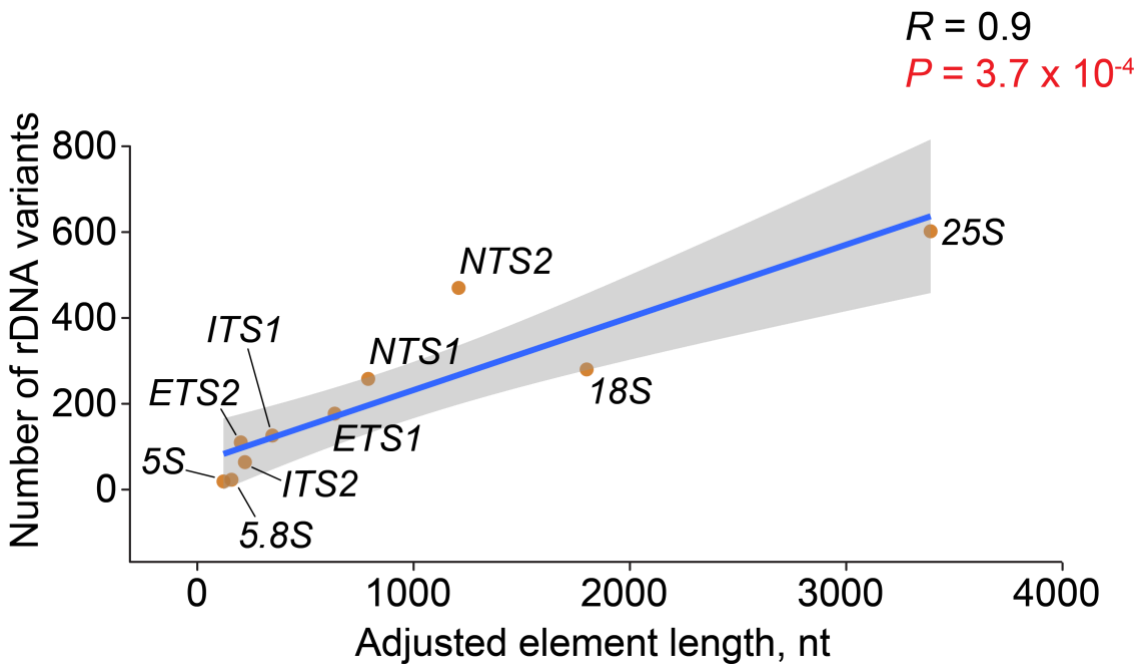
Supplementary Figure 5. **a,** correlation between the number of rDNA variants and the total number of rDNA copies per isolate. *R* – Pearson correlation coefficient. Regression line with 95% confidence interval is shown in blue. **b**, sharedness of variants across isolates. Numbers inside the bars indicate the number of variants in each category.



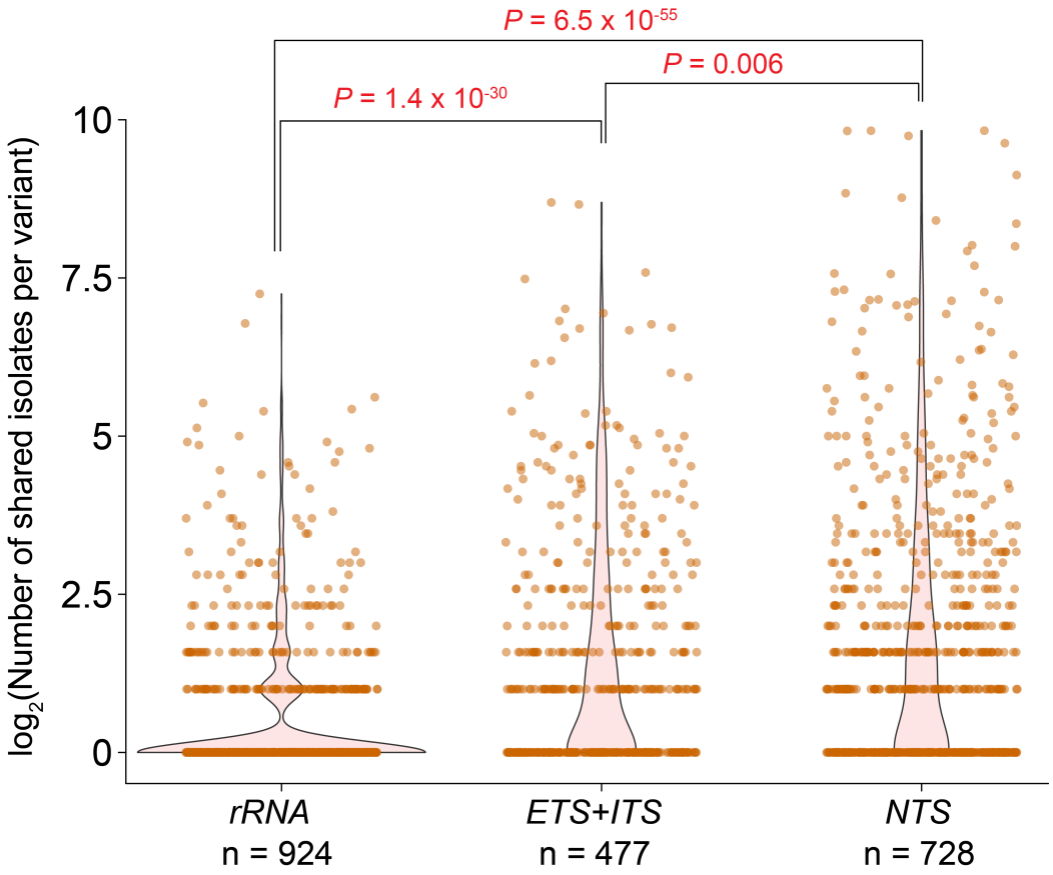
Supplementary Figure 6. Indel length distribution across rDNA. The number of indels across all isolates is shown as a function of indel size.



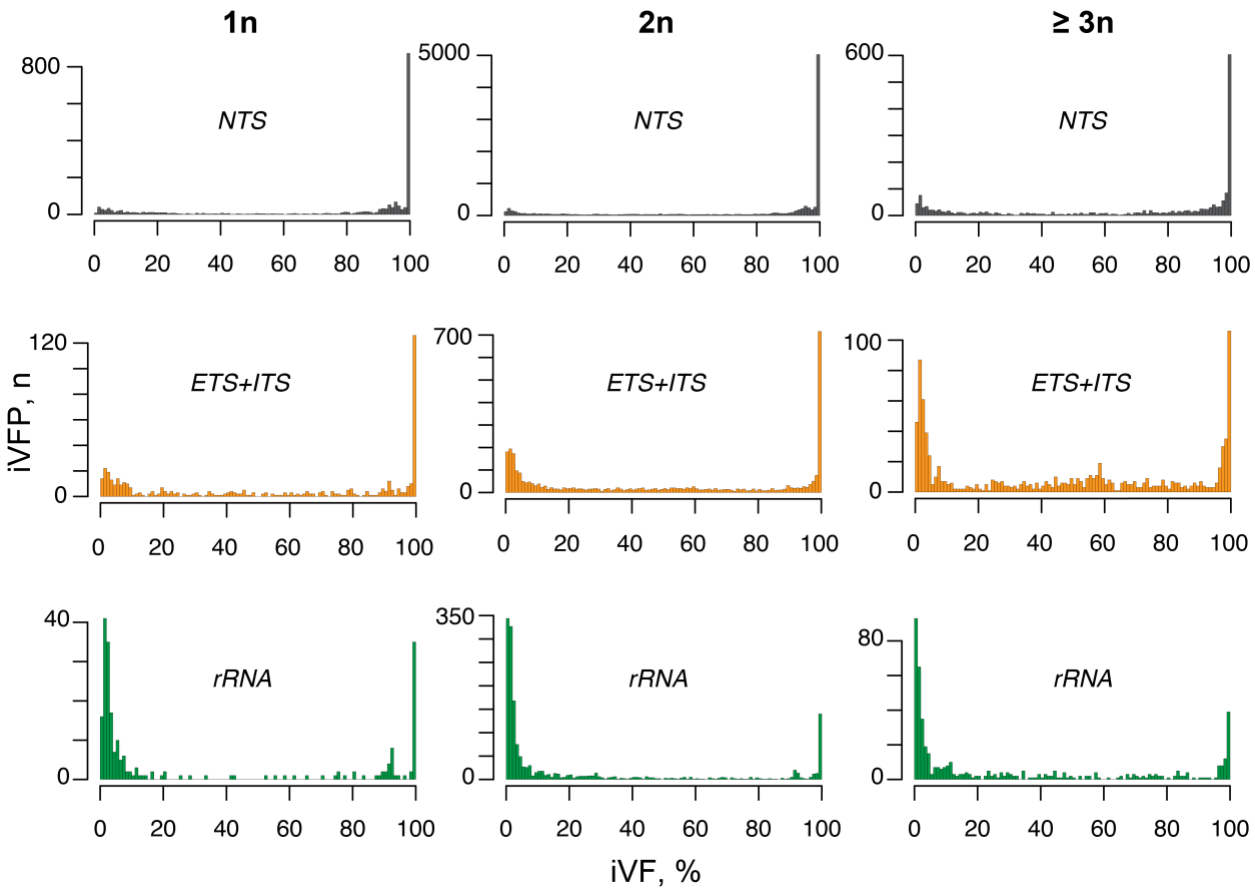
Supplementary Figure 7. a, Sequence diversity at each nucleotide (Shannon's entropy) based on iVFPs calculated for each rDNA position. **b**, correlation between entropy and the number of iVFPs at a given position. Best-fit polynomial is shown in red.



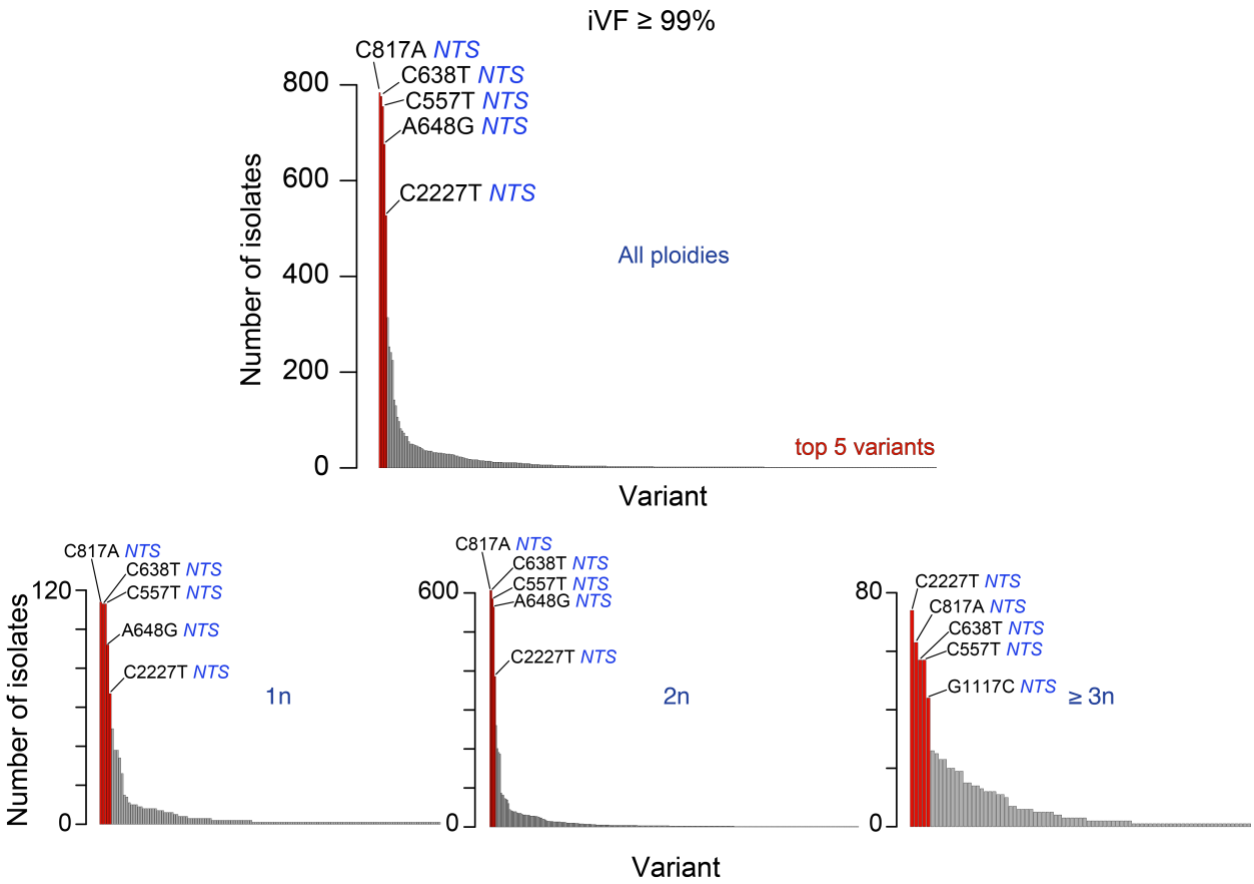
Supplementary Figure 8. Correlation between the number of rDNA variants per element and the length of the corresponding element. *R* – Pearson correlation coefficient. Brown dots represent different elements. Regression line with 95% confidence interval is shown in blue. Adjusted element lengths are due to filtering (see Methods).



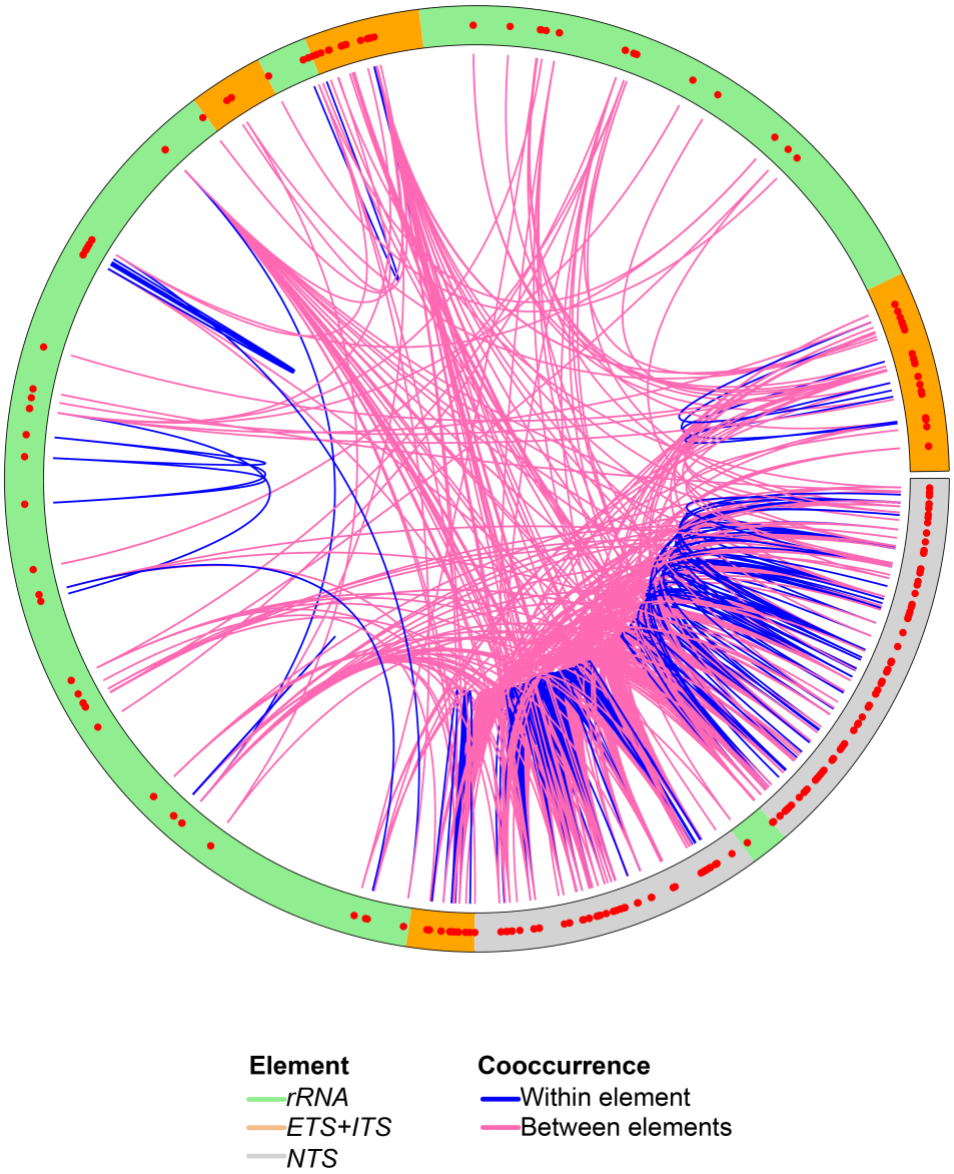
Supplementary Figure 9. Sharedness of variants across isolates is stratified by rDNA elements. n - total number of variants in the group. Y-axis is shown in log2 scale. Adjusted p-value (*P* adj) was calculated using Pairwise Wilcoxon rank sum test with Benjamini–Hochberg correction.



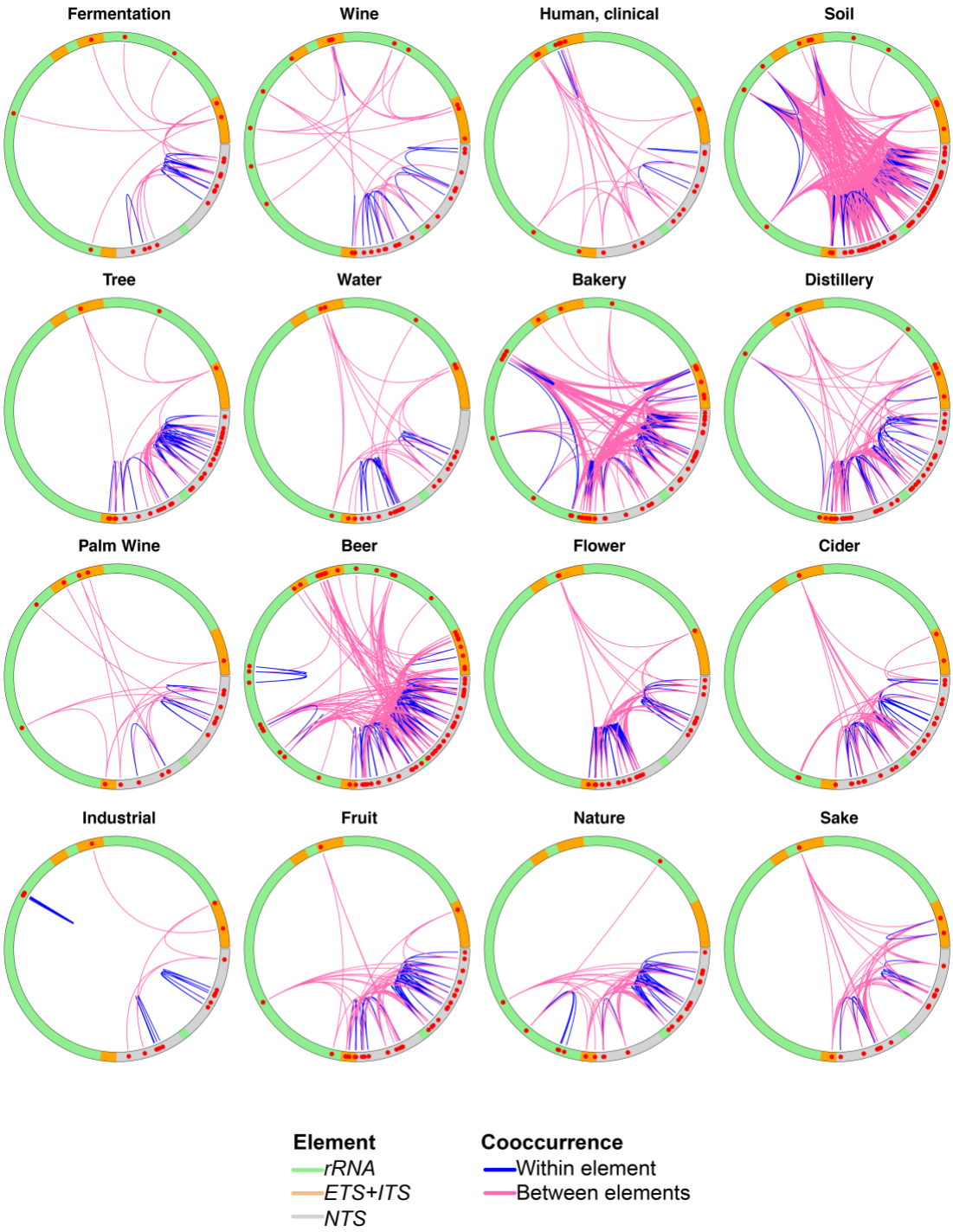
Supplementary Figure 10. Distribution of iVFPs separated by rDNA elements and isolates' ploidy. 1n - haploid, 2n - diploid, >= 3n - polyploid isolates.



Supplementary Figure 11. Distribution of variants with iVF ≥ 99% across isolates. Each bar represents a variant, and the height of the bar (y-axis) shows how many isolates contain the variant at iVF ≥ 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*.



Supplementary Figure 12. Variants with consistently similar iVFs. Cooccurrence was calculated on all isolates based on Euclidean distances (see Methods) between iVFs 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.



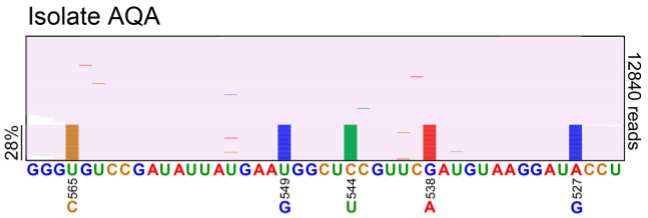
Supplementary Figure 13. Variants with consistently similar iVFs in different ecological niches. Same as **Supplementary Figure 12** but performed within each ecological niche individually.

a

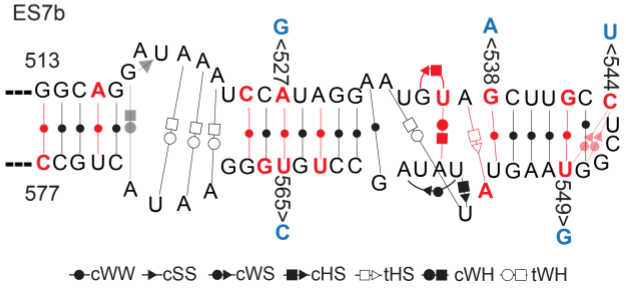
	Variants				
	1	**2**	**3**	**4**	**5**
rDNA nomenclature	**5322G;**	**5338C;**	**5343A;**	**5349T;**	**5360C**
25S nomenclature	**565C;**	**549G;**	**544U;**	**538A;**	**527G**
Industrial ABM	>99%	>99%	>99%	>99%	>99%
Industrial AHL	>99%	>99%	>99%	>99%	>99%
Industrial AHM	>99%	>99%	>99%	>99%	>99%
Industrial ATQ	>99%	>99%	>99%	>99%	>99%
Industrial AQD	>99%	>99%	>99%	>99%	>99%
Industrial BRK	>99%	>99%	>99%	>99%	>99%
AQA	27.9%	28.3%	28.3%	28.3%	28.0%

iVF

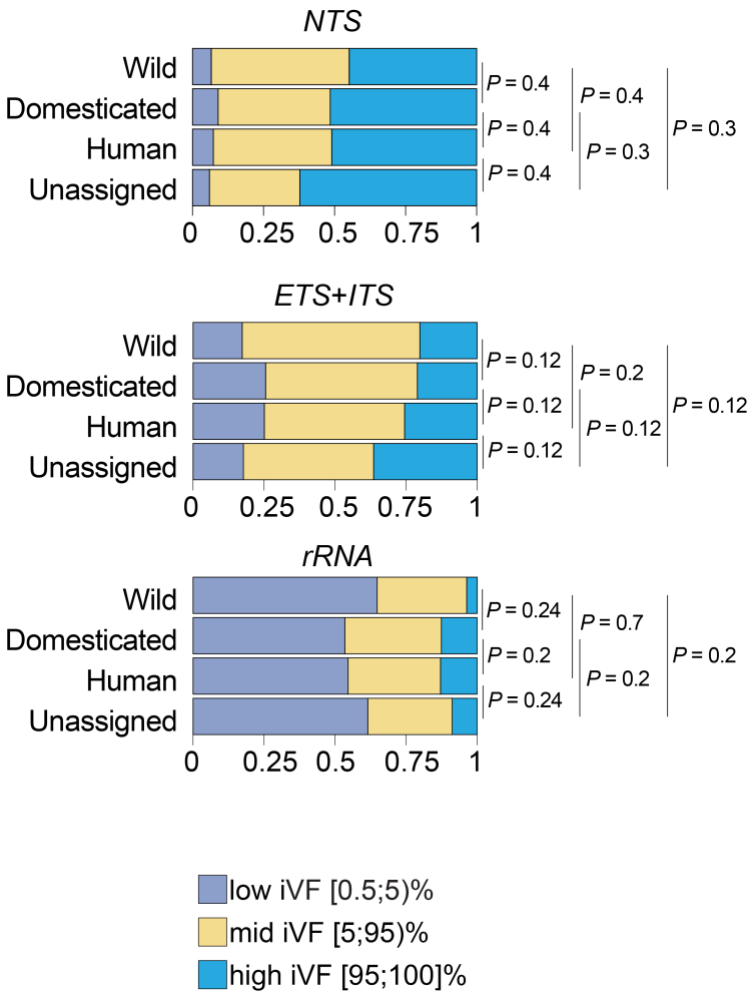
b



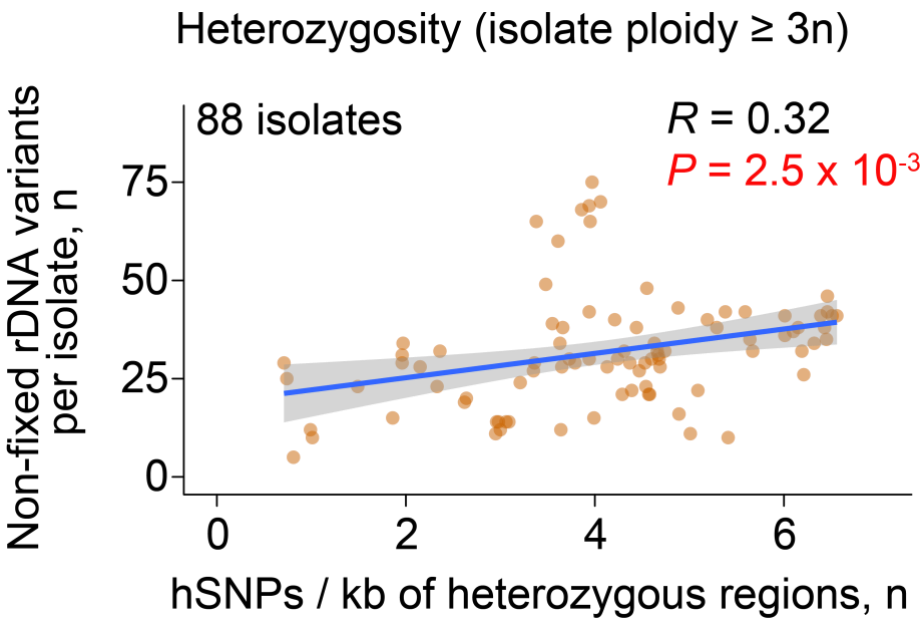
c



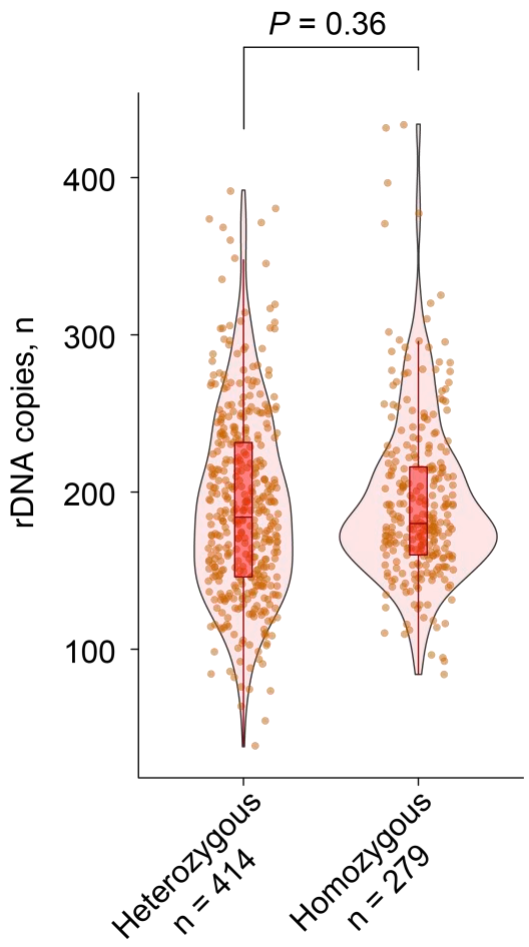
Supplementary Figure 14. Five variants cooccur in rDNA copies in several isolates. a, iVFs of the five variants in *25S* among seven isolates. **b**, continuous rDNA reads of isolate "AQA" mapped to *25S*. All the variants are in the same reads and only a subset of rDNA copies (28%) contain the variants 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.



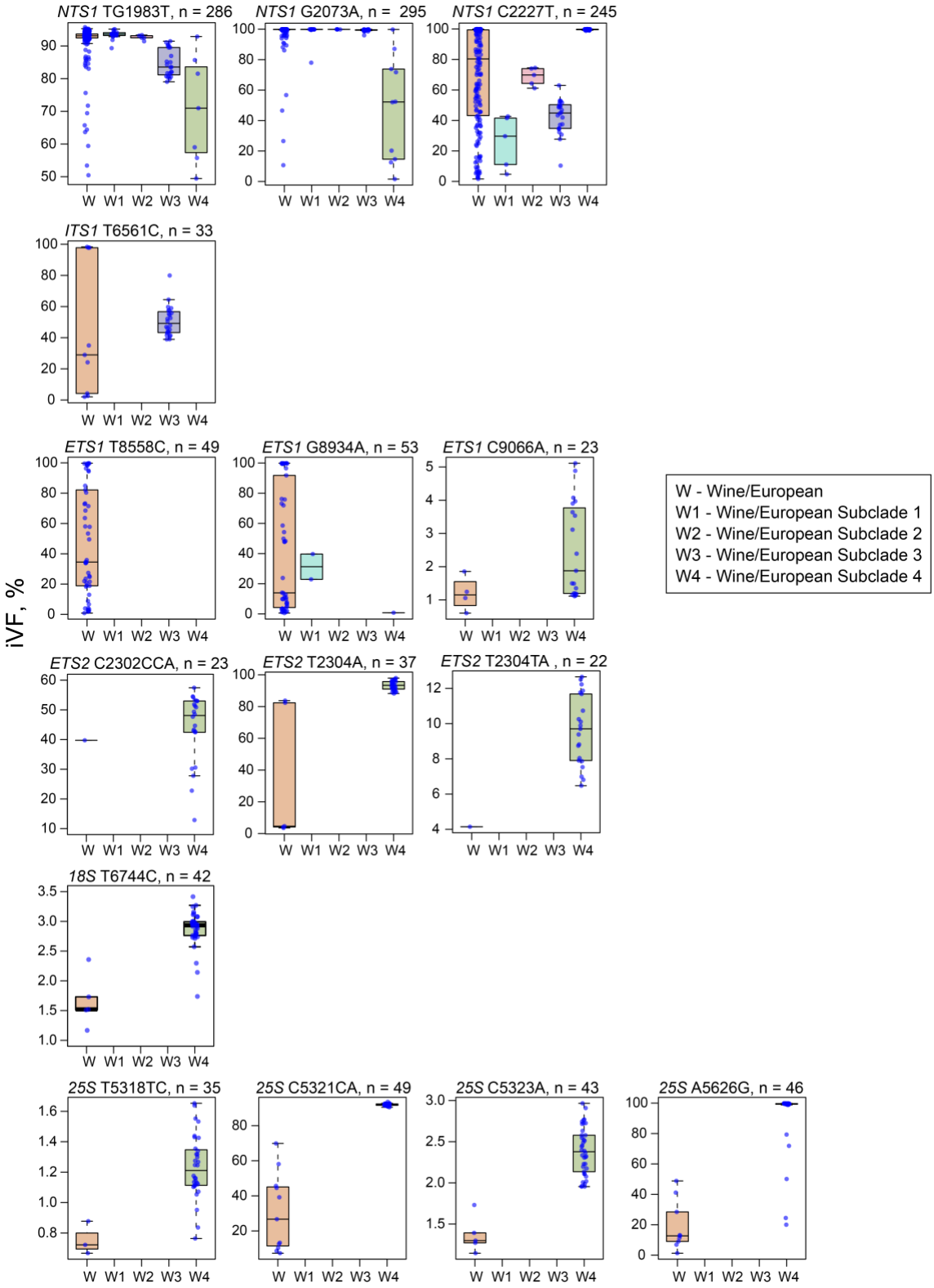
Supplementary Figure 15. iVFP distribution based on rDNA genic elements in different ecological groups from [1]. *P* – values were calculated by using Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple hypothesis testing.



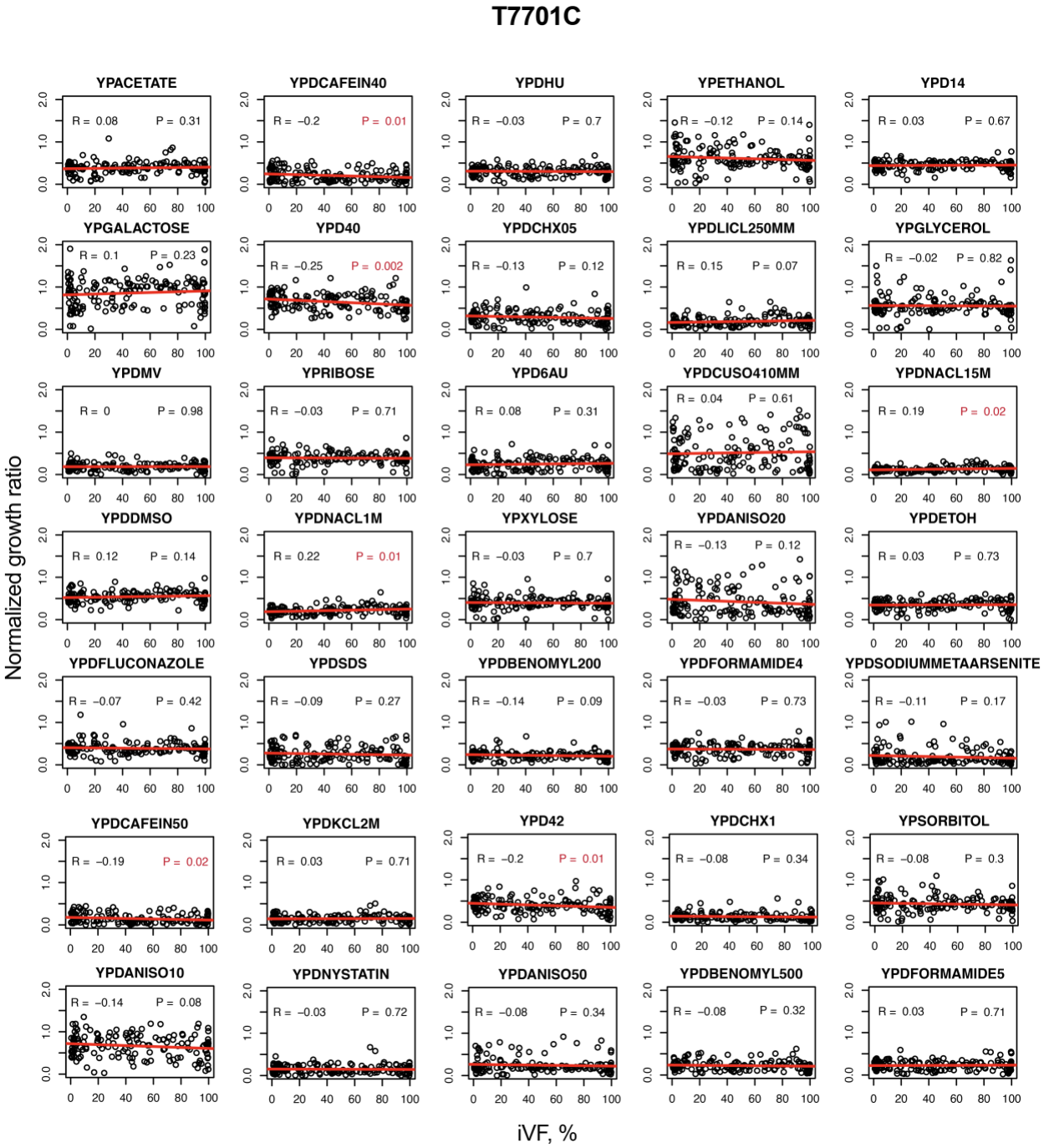
Supplementary Figure 16. Relationship between heterozygosity of the rDNA and the rest of the genome in polyploid (3n-5n) isolates. Correlation between the number of rDNA variants and the number of heterozygous SNPs per heterozygous region of the genome (excluding rDNA; from [2]). Pearson correlation coefficient *R* is shown. Each dot represents an isolate, regression line with a 95% confidence interval is shown in blue.



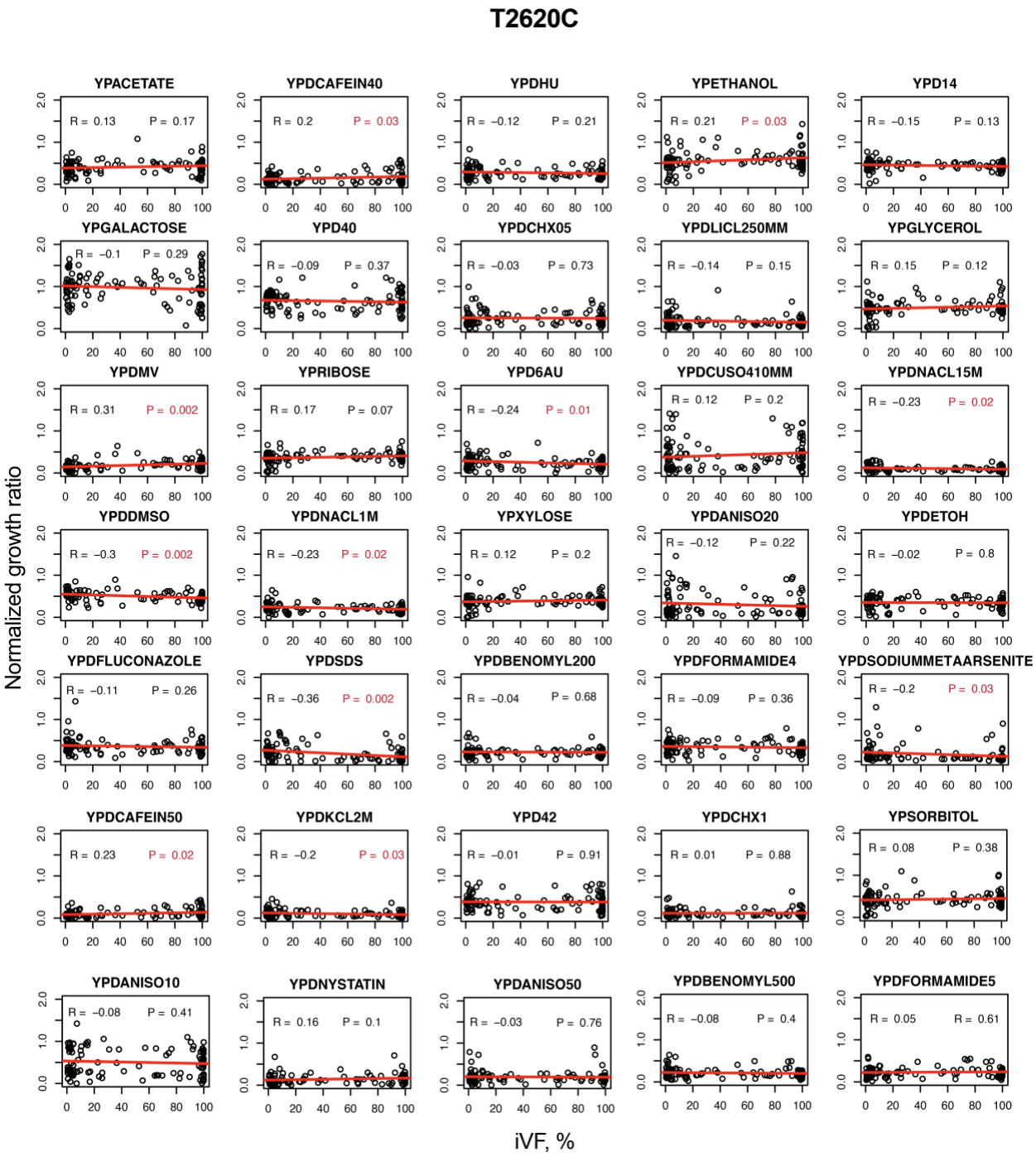
Supplementary Figure 17. Number of rDNA copies in heterozygous and homozygous diploid isolates analyzed in **Fig. 3** (main text). n indicates the number of isolates in each group. *P*-value was calculated using Wilcoxon rank sum test.



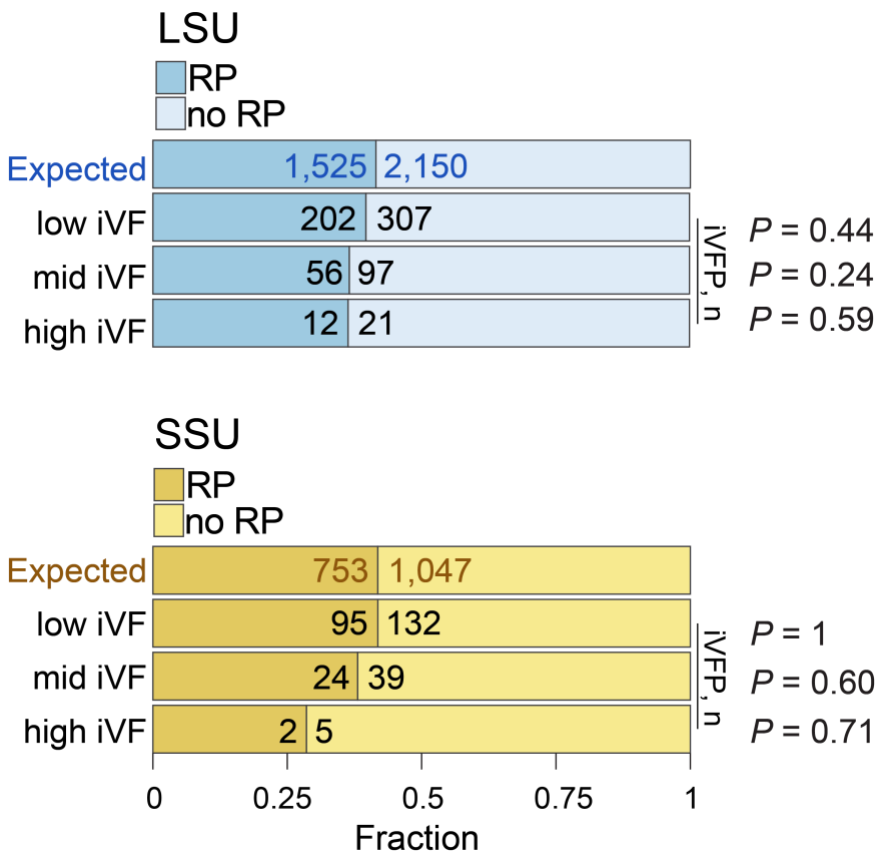
Supplementary Figure 18. Distribution of some of the most represented iVFPs in the Wine/European subclades. n indicates the total number of iVFPs for each variant across subclades.



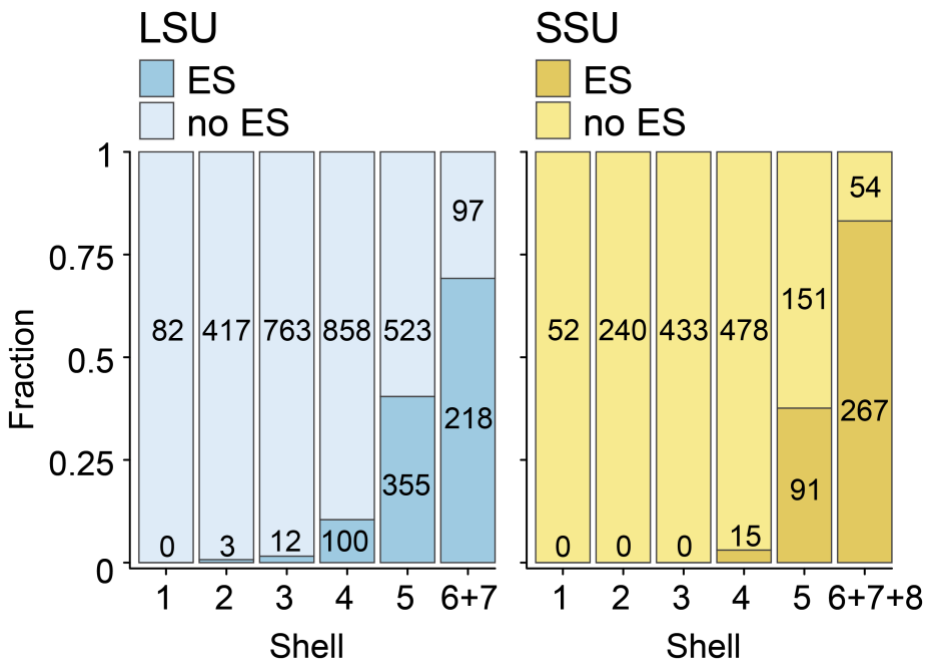
Supplementary Figure 19. Correlation between fitness and variant's iVFs for variant T7701C. Stress conditions and their fitness values (growth ratio normalized to a non-stressed growth condition) are taken from [2]. Pearson correlation coefficient R and a P-value are shown.



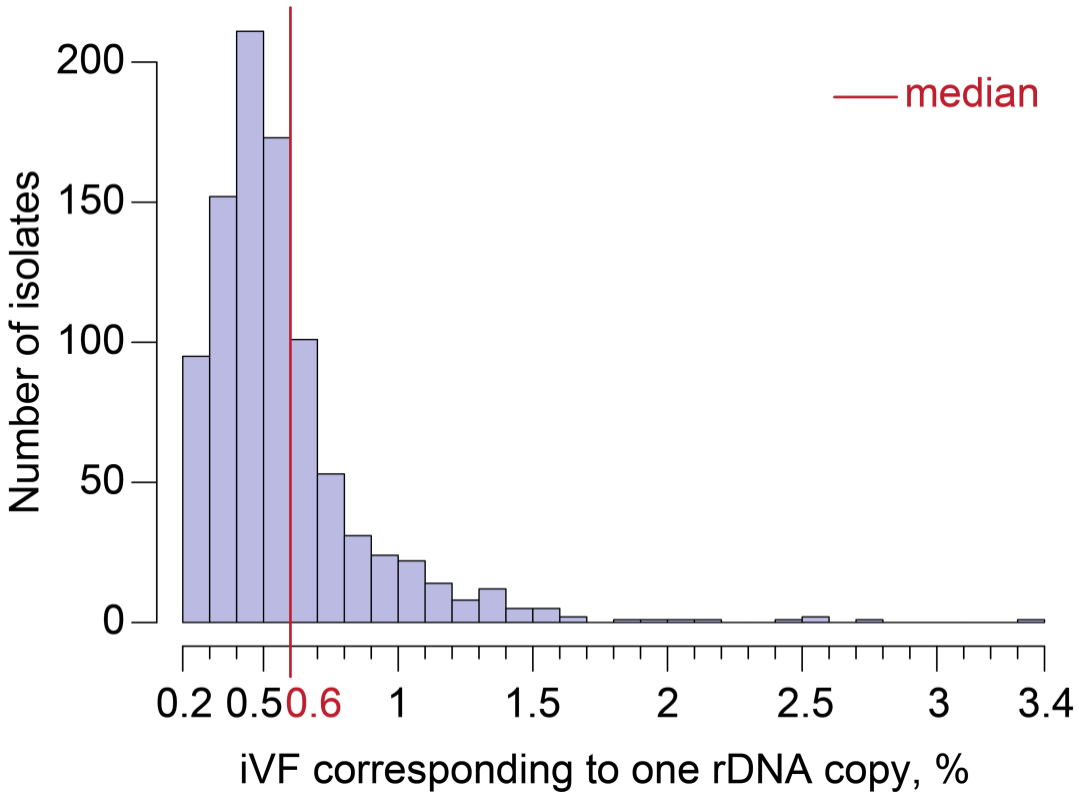
Supplementary Figure 20. Correlation between fitness and variant's iVFs for variant T2620C. Stress conditions and their fitness values (growth ratio normalized to a non-stressed growth condition) are taken from [2]. Pearson correlation coefficient R and a P-value are shown.



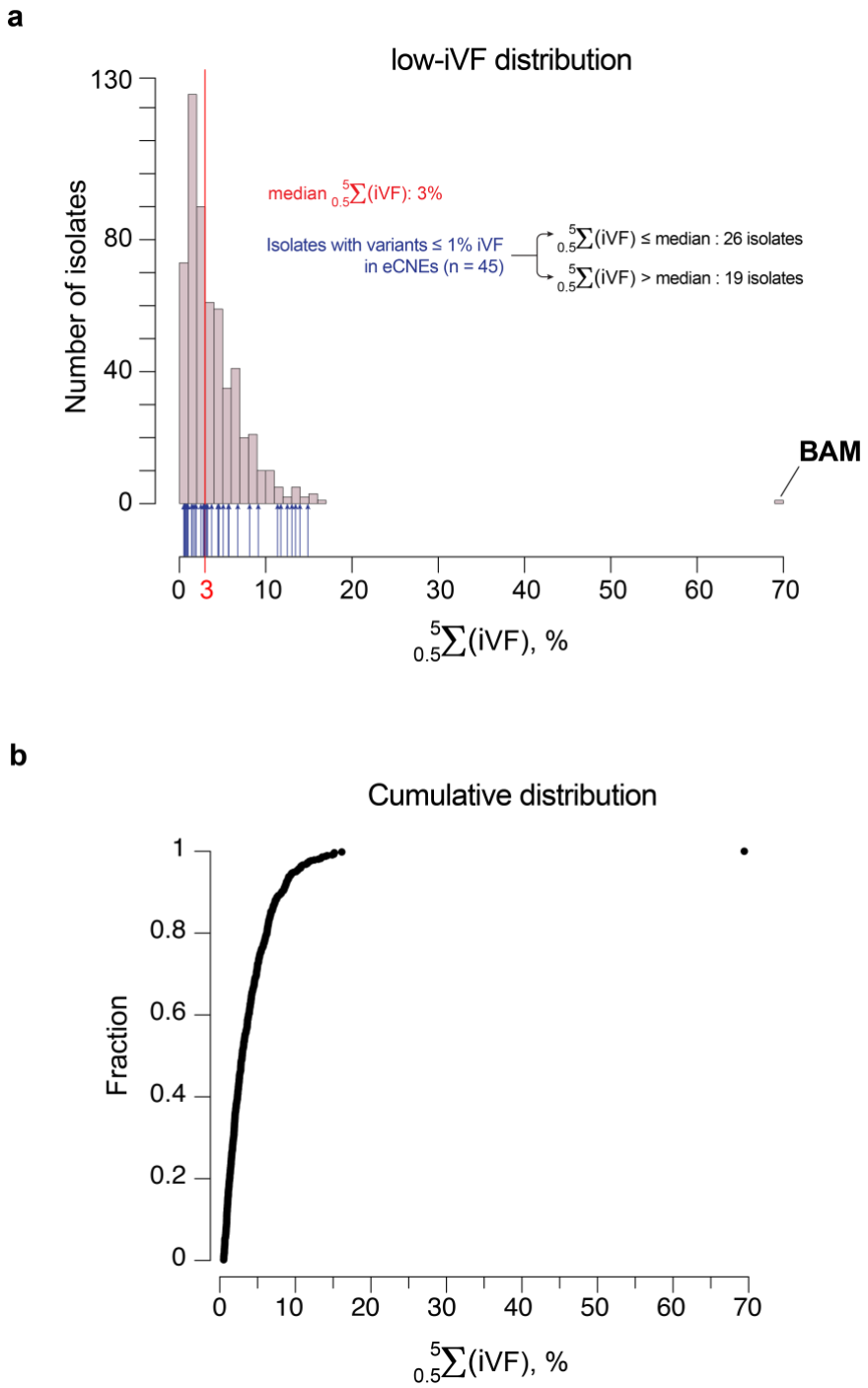
Supplementary Figure 21. Distribution of VFPs that are involved in rRNA-ribosomal protein (RP) interactions. LSU – large subunit, SSU – small subunit. P-values are calculated using Fisher's exact test.



Supplementary Figure 22. Distribution of nucleotides from expansion segments (ES) across shells. "no ES" – nucleotides outside ES.

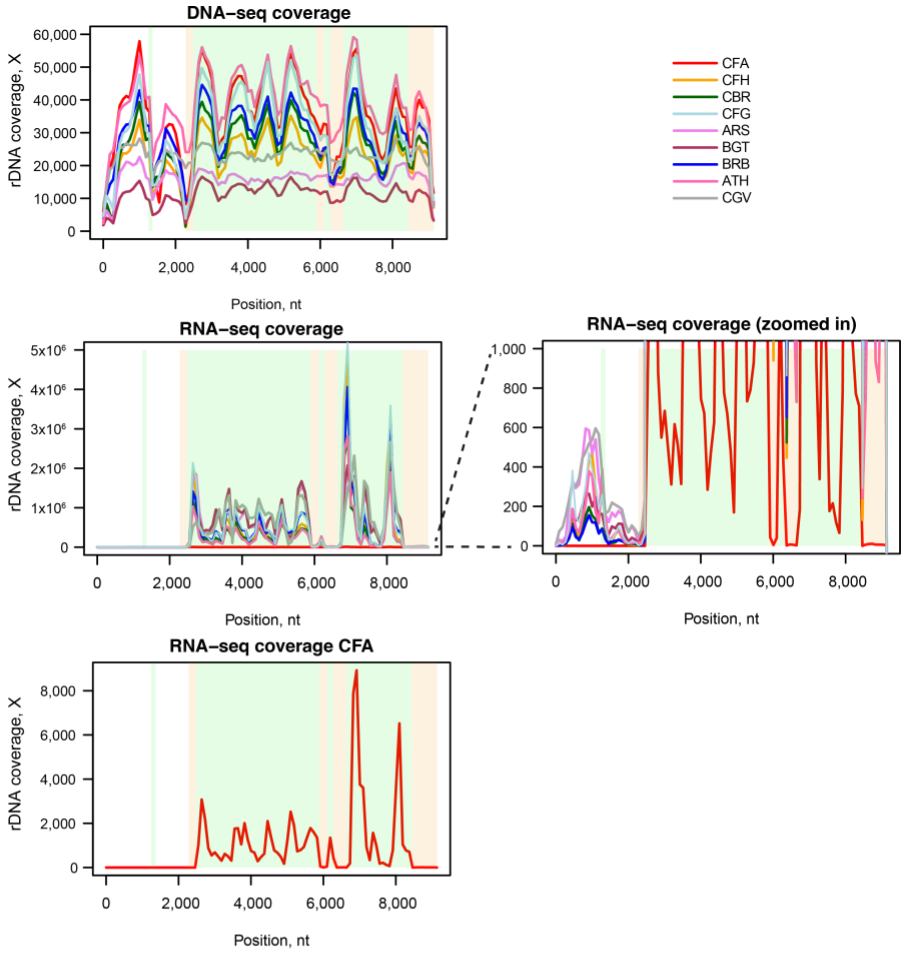


Supplementary Figure 23. Distribution of iVFs that correspond to one rDNA copy. Each iVF was calculated for every isolate from its total rDNA copy number. The red line indicates the median.



Supplementary Figure 24. a, distribution of sums of low iVFs 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. The "BAM" isolate is indicated in the plot. **b** – cumulative distribution of the same dataset.

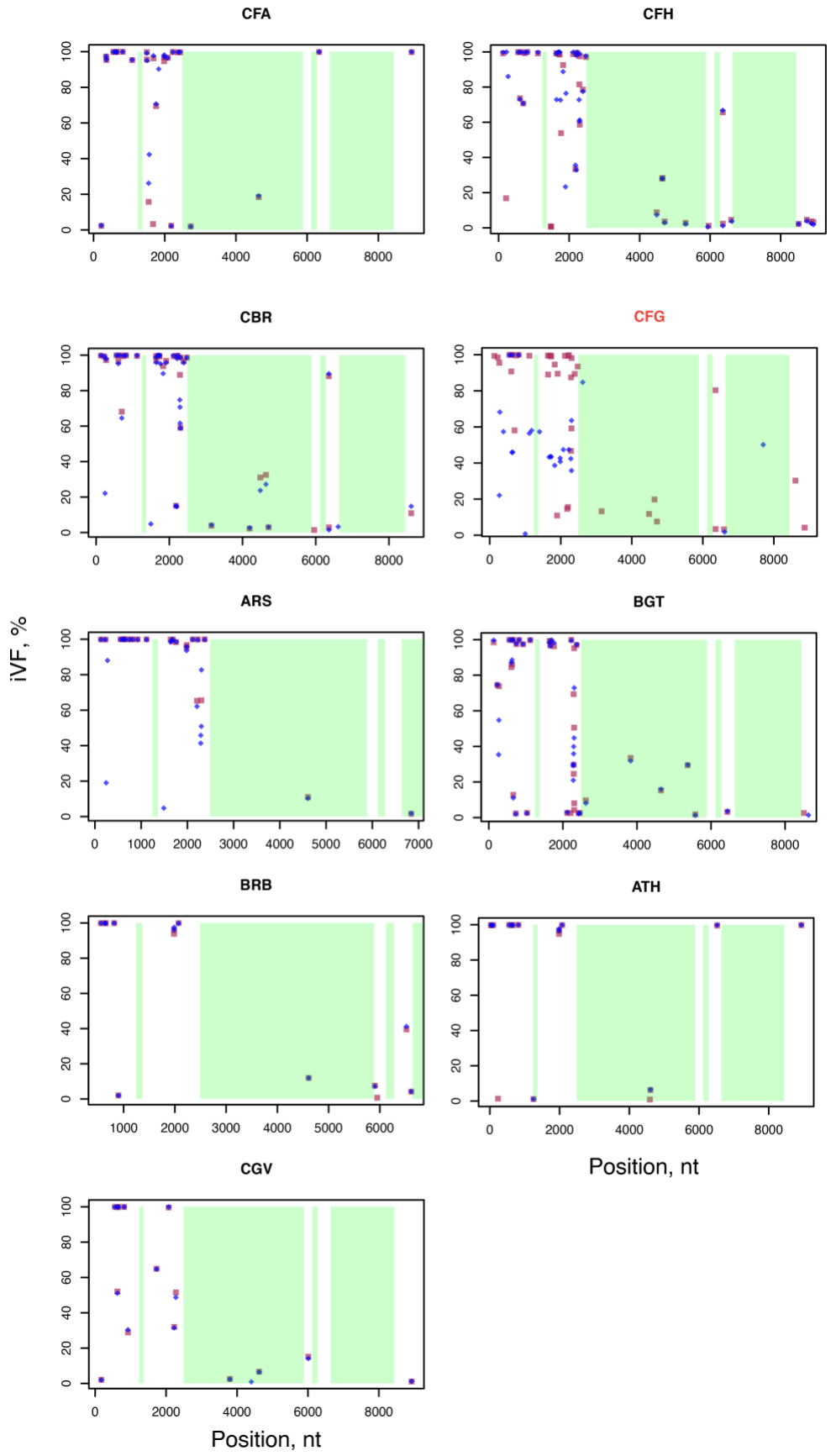
a



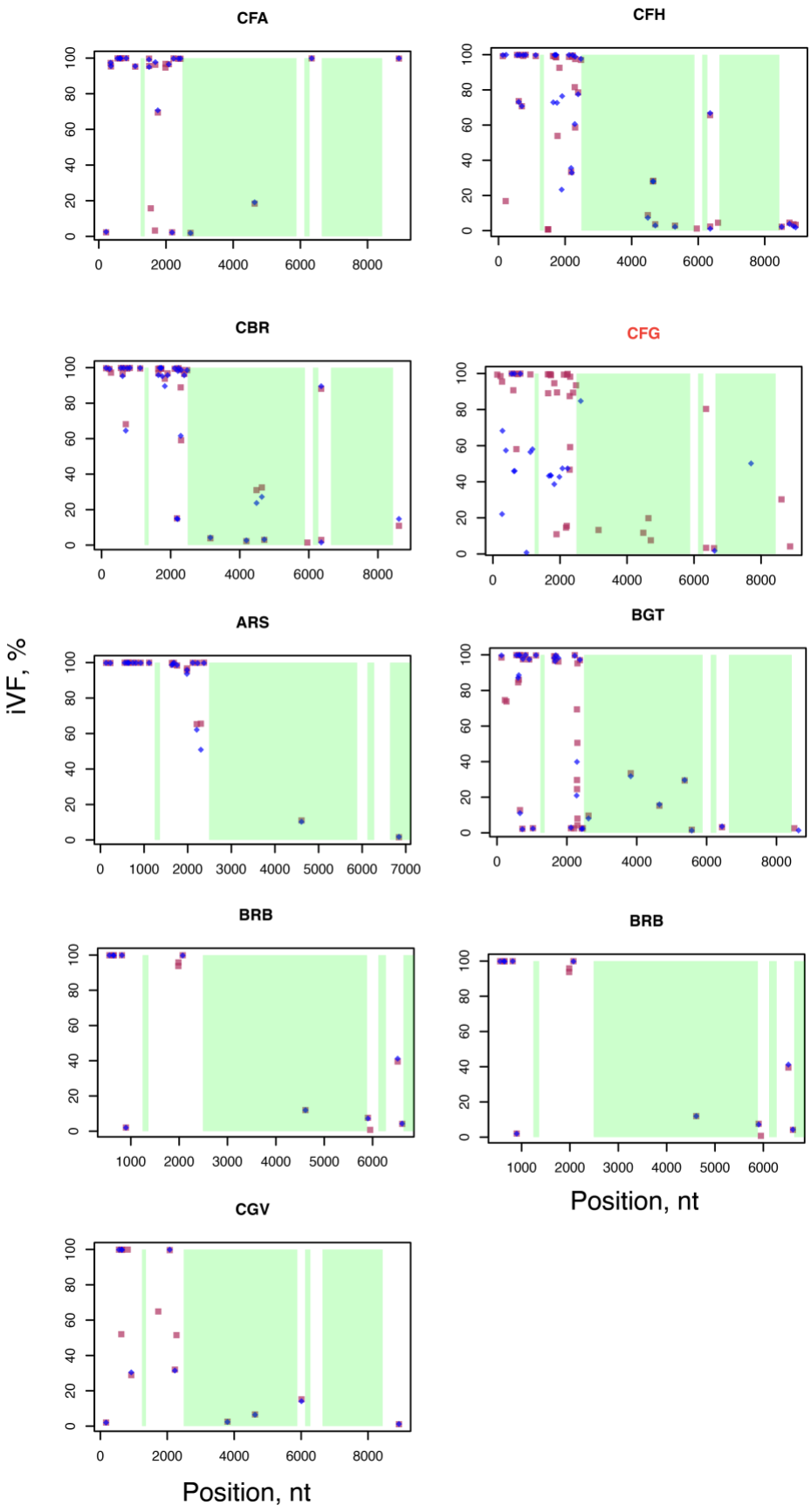
b

Library	Total Read Count
undetermined_library	23 208 839
12267_DNA	11 927 317
12268_DNA	12 051 221
12269_DNA	14 410 096
12270_DNA	12 055 410
12271_DNA	14 671 076
12272_DNA	12 585 743
12273_DNA	11 859 724
12274_DNA	12 797 641
12275_DNA	11 675 009
12267_totalRNA CFA RNA-seq	167 807
12268_totalRNA	23 112 037
12269_totalRNA	29 918 763
12270_totalRNA	30 632 545
12271_totalRNA	17 787 776
12272_totalRNA	21 149 883
12273_totalRNA	25 329 291
12274_totalRNA	36 546 306
12275_totalRNA	18 912 060

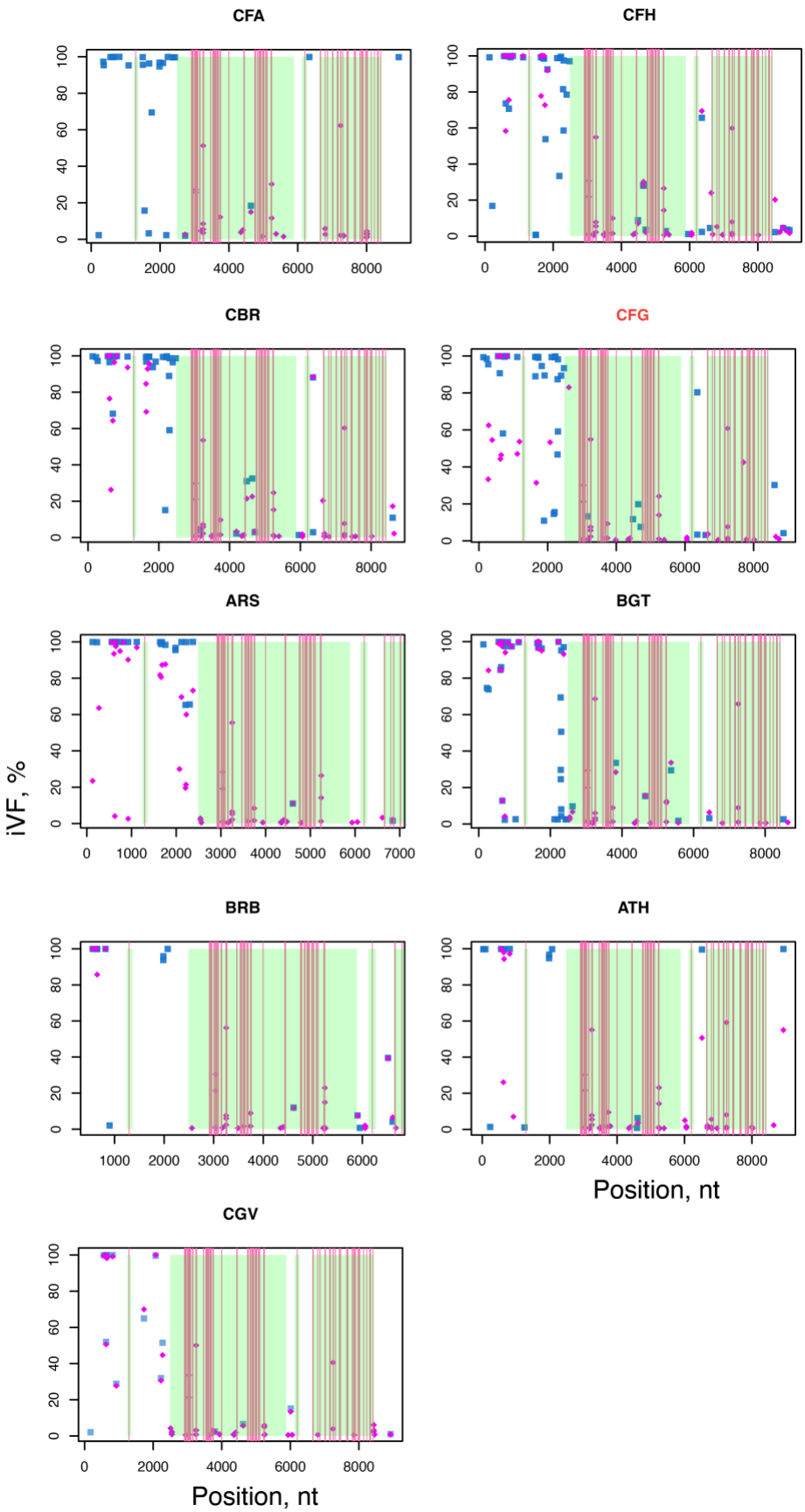
Supplementary Figure 25. Sequencing coverage of the nine isolates with variants in the GAC. a, DNA- and total RNA-seq coverage of the isolates plotted against the rDNA prototype. Each line is a separate isolate. Note the extremely high coverage of the RNA-sequencing data, which indicates that the transcripts are coming from true ribosomal RNA and are expressed in the cell. The lower coverage of the isolate CFA is due to very a low concentration of the sample library in the pool (**b**). Additionally, note that some reads in the RNA-seq data are from *NTS* (non-transcribed sequences), and their very minor fraction suggest they are coming either from the genomic DNA or leaky transcription.


CFA
CFH
CBR
CFG
ARS
BGT
BRB
ATH
CGV
iVF, %
Position, nt
Position, nt

Supplementary Figure 26. Variants detected from the DNA-sequencing data. Red dots indicate variants detected in DNA sequencing data from [2], and blue dots indicate variants detected in DNA sequencing data from the current study (without filtering for high strand bias). The discordant "CFG" isolate is highlighted in red.

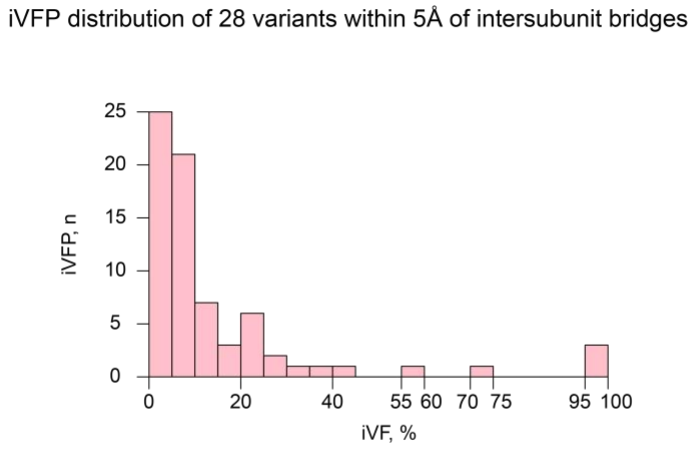

CFA
CFH
CBR
CFG
ARS
BGT
BRB
BRB
CGV
iVF, %
Position, nt
Position, nt

Supplementary Figure 27. Same as Supplementary Figure 26 but with additional filtering for high strand bias. Top 10% of variants with the highest strand bias were removed in each isolate. Some variants in *NTS*, *ETS* and *ITS* were affected by the implicated strand bias filter, suggesting a trade-off between the decreased false positive rate and increase true positive calls. The discordant "CFG" isolate is highlighted in red.

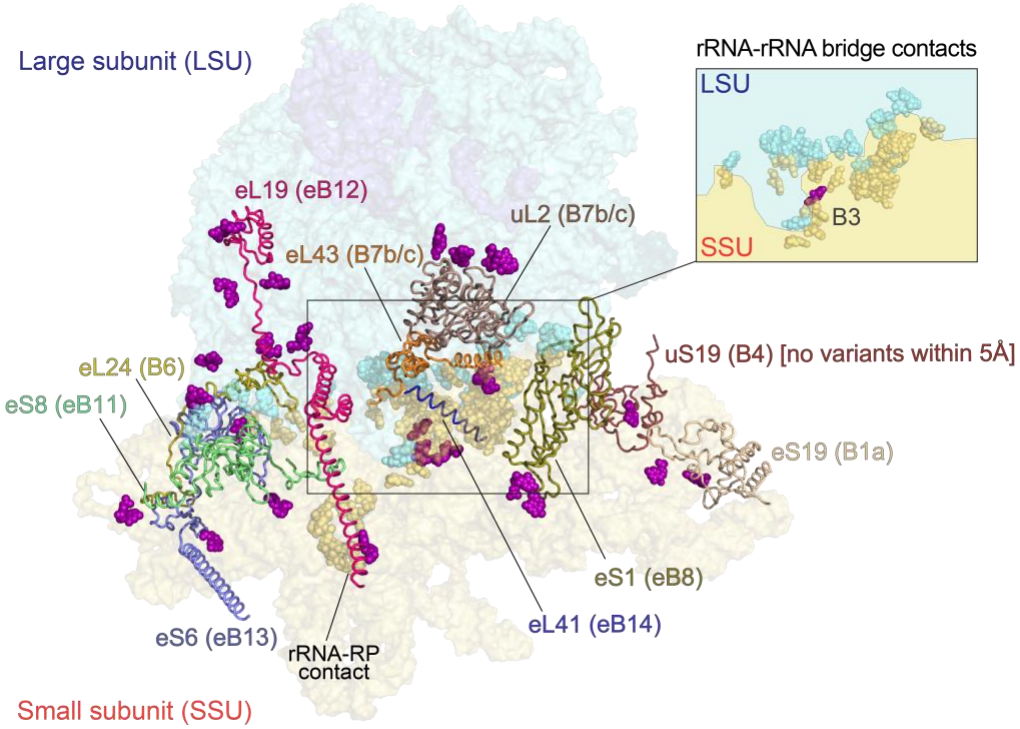

CFA
CFH
CBR
CFG
ARS
BGT
BRB
ATH
CGV
iVF, %
Position, nt
Position, nt

Supplementary Figure 28. Variants detected in the current study. Blue dots are variants detected in the DNA-sequencing data from this study, and purple dots are variants detected in the RNA-sequencing data. Red lines indicate positions of rRNA post-transcriptional modification from [3], which align with consistent false-positive calls from RNA-sequencing data. The presence of variants in *NTS* the in RNA-sequencing data is due to possible DNA contamination or leaky transcription (see **Supplementary Figure 25**).

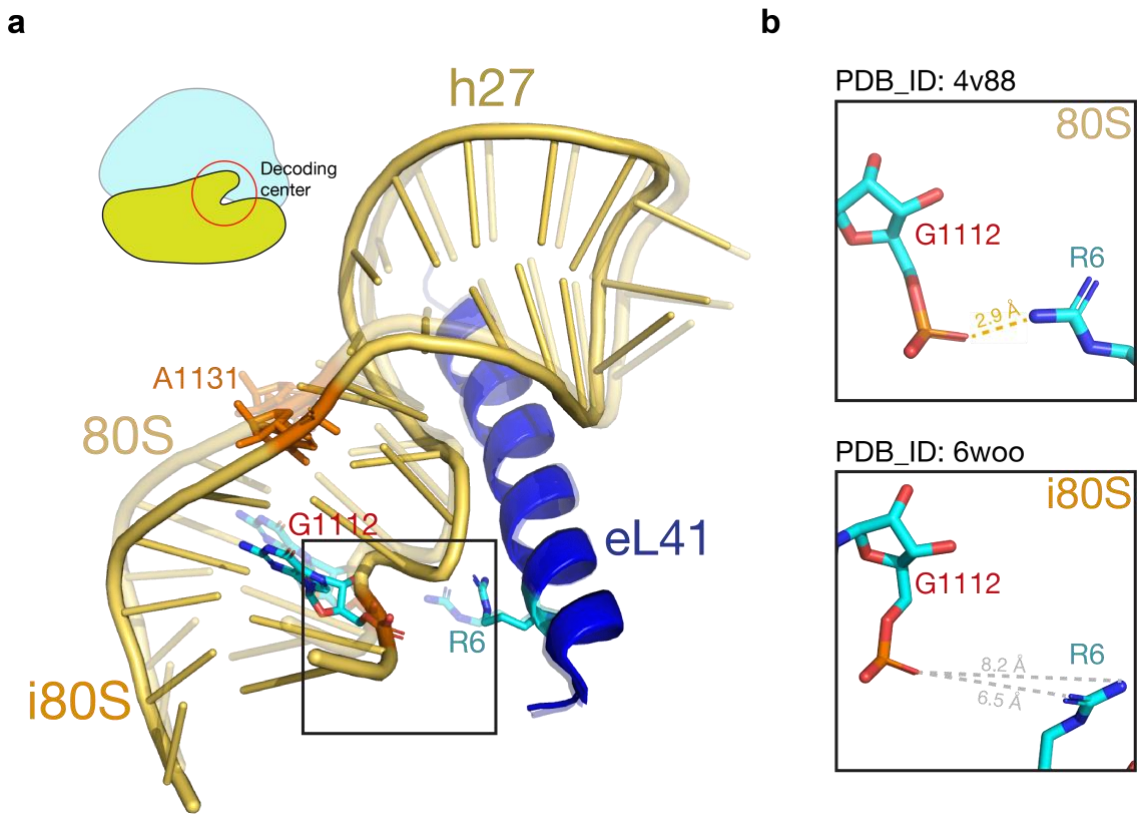
a



b



Supplementary Figure 29. Variants are localized in the vicinitiy of intersubunit bridges. a, iVFP distribution of the 26 variants found within 5Å of the intersubunit bridges. **b**, 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). Blue outline - large subunit (LSU); yellow outline - small subunit (SSU). Ribosomal protein (RP) names along with the corresponding bridges (in brackets) are indicated. *inset* - bridge contacts mediated by rRNA only. A variant position near bridge B3 is indicated in the inset.



Supplementary Figure 30. 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 eL41; 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 eL41. The phosphate backbone of G1112 forms a hydrogen bond with R6 of eL41 in the 80S (*top*), while no interaction occurs in the i80S (*bottom*) due to conformational change.

SUPPLEMENTARY REFERENCES

1. de Chiara, M. *et al.* Domestication reprogrammed the budding yeast life cycle. *Nat Ecol Evol* **6**, 448–460 (2022).
2. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
3. Sloan, K. E. *et al.* Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol* **14**, 1138–1152 (2017).