The turnover of ribosome-associated transcripts from *de novo* ORFs produces gene-like characteristics available for *de novo* gene emergence in wild yeast populations

Éléonore Durand[1*], Isabelle Gagnon-Arsenault[1,2], Johan Hallin[1,2], Isabelle Hatin[3], Alexandre K Dubé[1,2], Lou Nielly-Thibault[1], Olivier Namy[3] & Christian R Landry[1,2]

[1] Institut de Biologie Intégrative et des Systèmes, Département de Biologie, PROTEO, Centre de Recherche en Données Massives de l'Université Laval, Pavillon Charles-Eugène-Marchand, Université Laval, G1V 0A6 Québec, QC, Canada. [2] Département de biochimie, microbiologie et bio-informatique, Université Laval, G1V 0A6 Québec, QC, Canada. [3] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, 91190 Gif sur Yvette, France.
* Current address : Université de Lille CNRS, UMR 8198-Evo-Eco-Paleo, Lille, France

Supplemental Figures

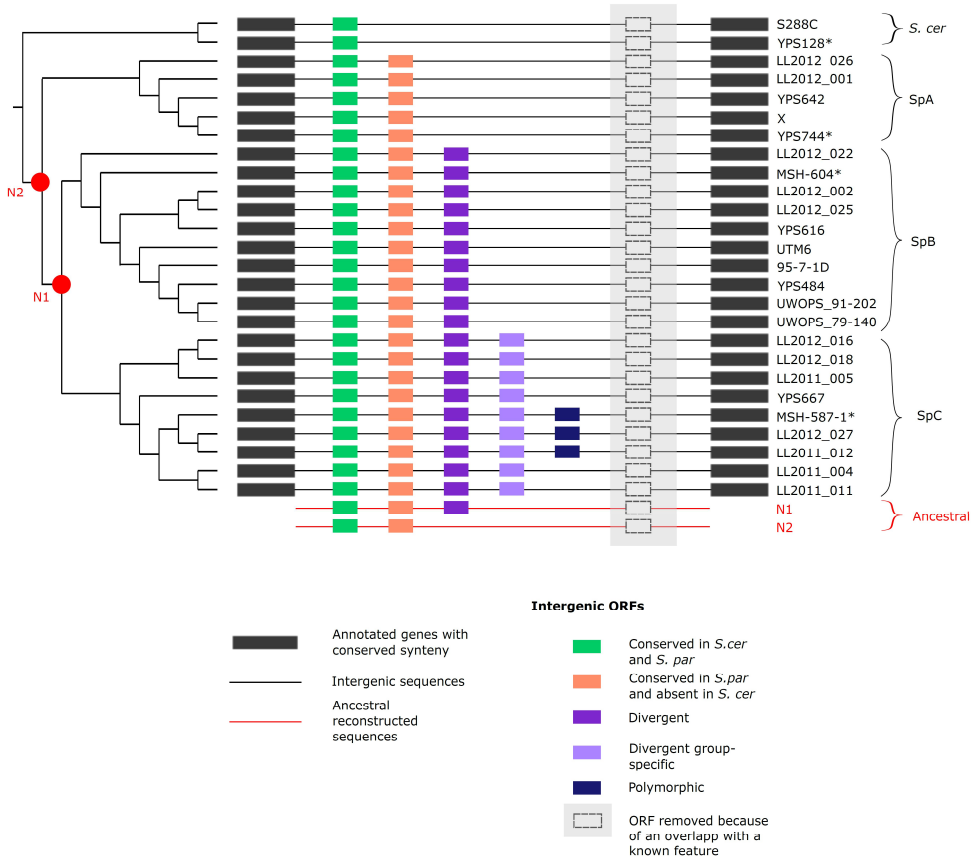


Figure S1. Identification and annotation of iORFs. Genes with conserved synteny were used as anchor to align non-genic sequences for each pair. iORFs were annotated on intergenic aligned sequences and clustered as orthogroups based on the conservation of the positions of their start and stop codons aligned with no disruptive mutation within. iORFs displaying a sequence similarity or an overlap with a known feature were removed from all strains. iORFs are colored according to their conservation group (see Methods and Fig. S1): conserved (cons), *S. paradoxus* (Spar) specific and fixed, divergent (Div), divergent group-specific (DivG) and polymorphic (Pol). Strains used for ribosome profiling experiments are marked with *.

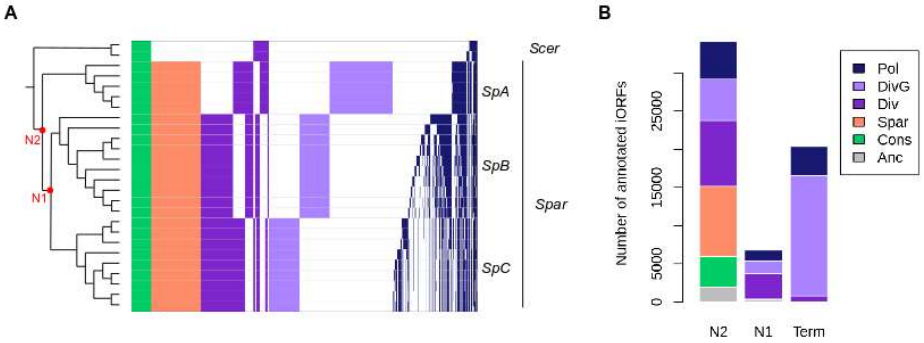


Figure S2. iORF conservation and diversity in wild *Saccharomyces paradoxus* populations. A) Columns represent iORFs sorted according to their conservation levels. iORFs that are absent are shown in white. Others are colored according to their conservation group (see Methods and Supplemental Fig. S1): conserved (cons), *S. paradoxus* (Spar) specific and fixed, divergent (Div), divergent group-specific (DivG) and polymorphic (Pol). **B)** Number of annotated iORFs per age, corresponding to oldest node in which they were detected. 'Term' refers to iORFs appearing on terminal branches and absent in ancestral reconstructions. iORFs detected only in ancestral sequences are plotted in gray.

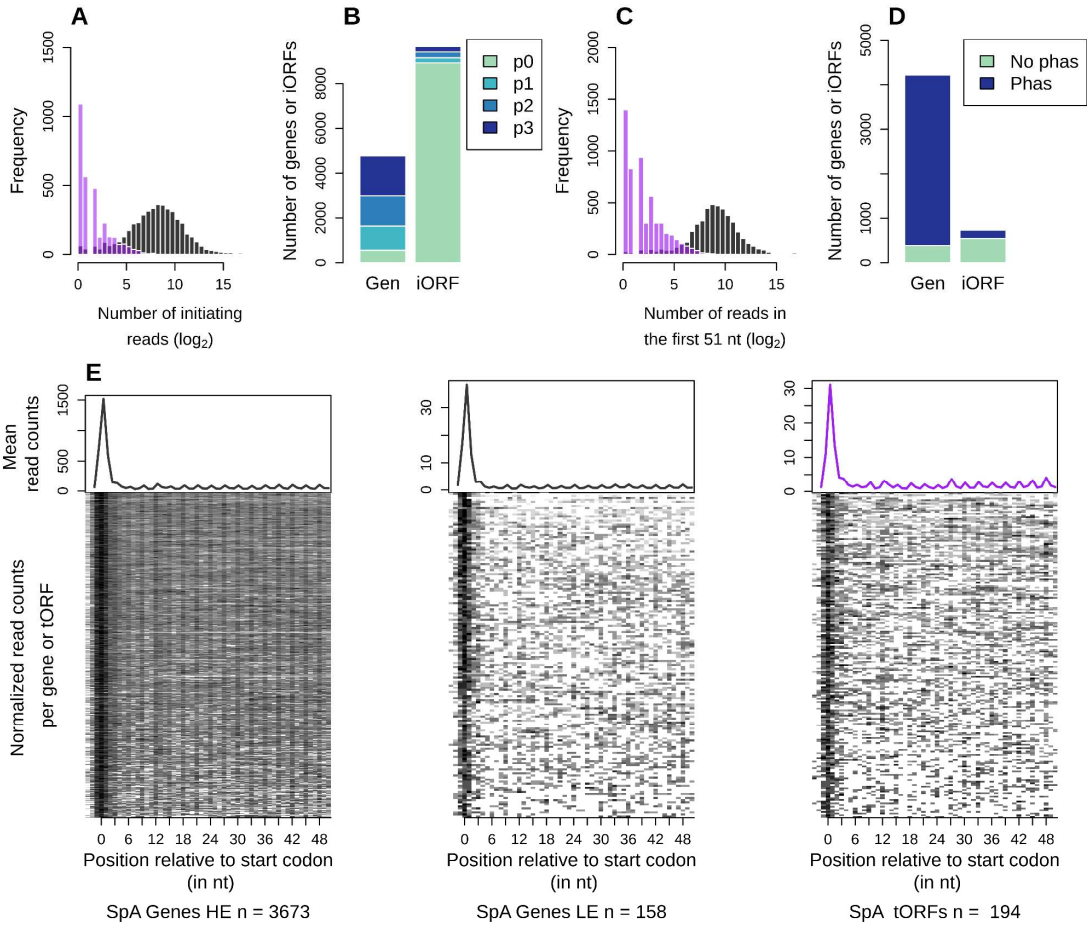


Figure S3. Detection of iORF translation signatures. Results for *SpA* and *SpB* strains (see Fig. 2 for *SpC* results). **A)** Distribution of ribosome profiling read counts for genes (in grey) and iORFs (in purple) at the start codon position. **B)** Proportions of genes (Gen) or iORFs with a detected initiating peak at the start codon position. Peaks are colored according to the precision of the detection (see Methods), from the most precise (p3) to the less precise (p1). No peak detection is in green (p0). **C)** Distribution of ribosome profiling read counts in the first 51 nt of iORFs excluding the start codon. **D)** Proportions of genes or iORFs with a significant codon periodicity (in blue) among genes and iORFs with a detected initiation peak. No peak detection is in green. **E)** Metagene for significantly translated and highly (HE) or lowly (LE) expressed genes in grey, and intergenic tORFs in purple. The mean of 5' read counts is plotted along the position relative to the start codon for significantly translated genes or tORFs. Lines of the matrix indicate the normalize coverage of all genes or tORFs with significant signature of translation.

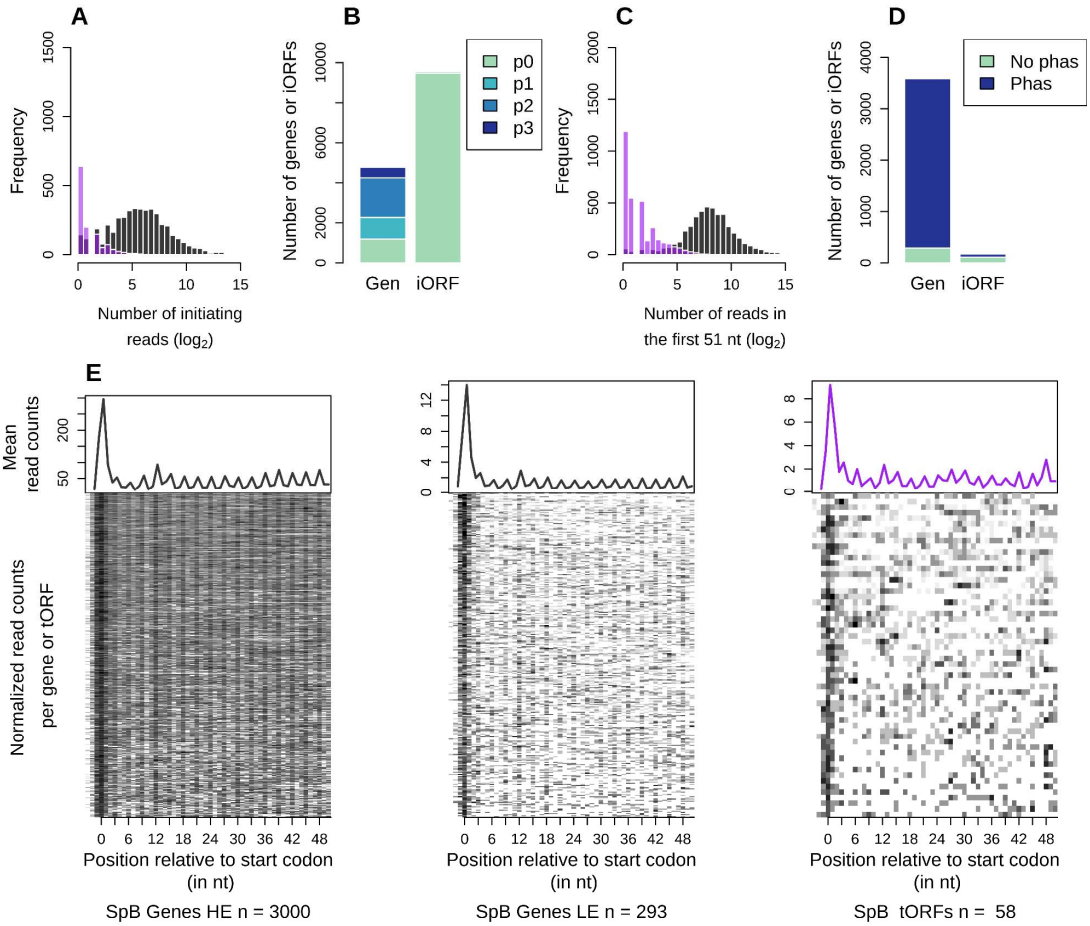


Figure S3. Continued.

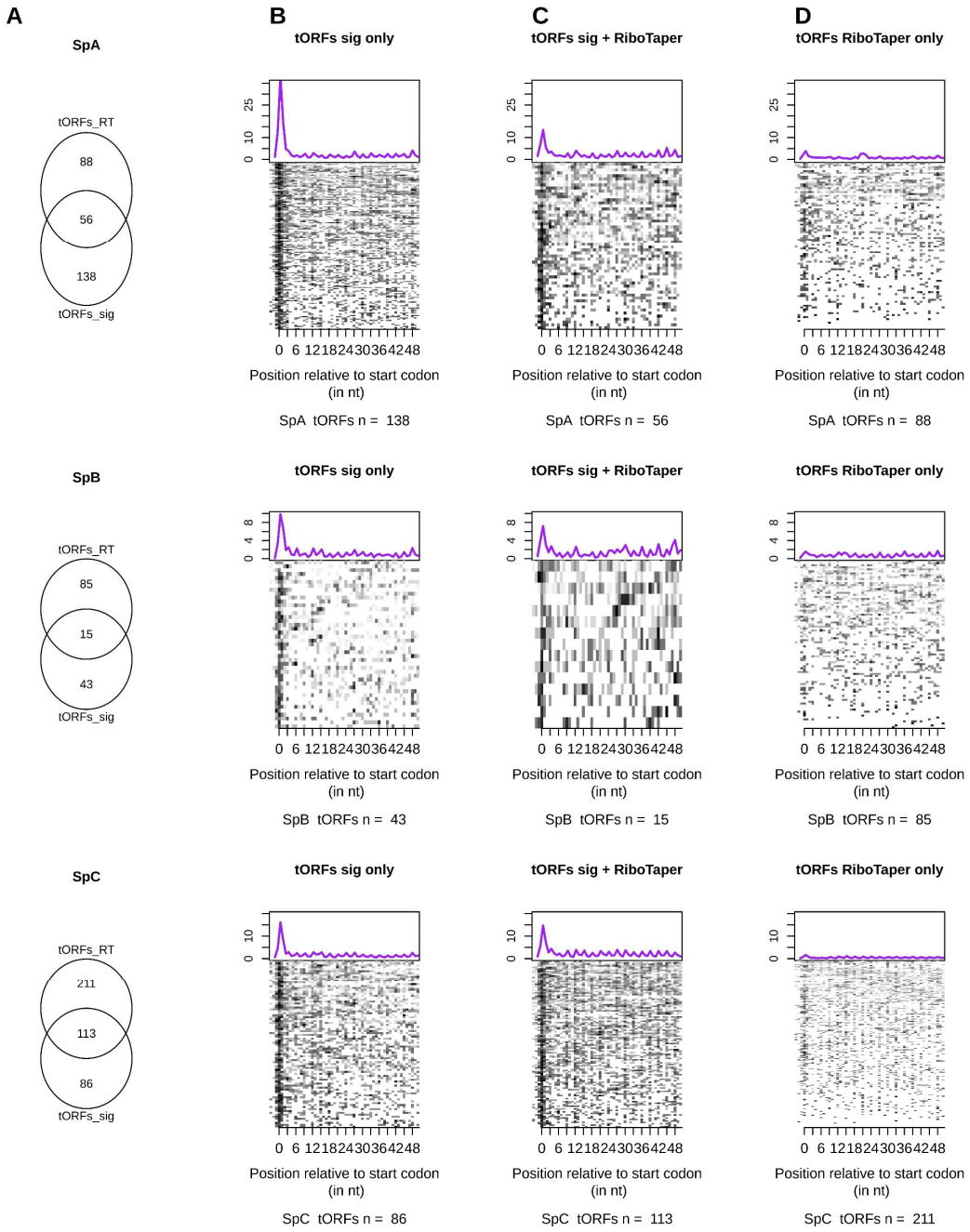


Figure S4. Translation profiles of tORFs with two detection methods in *S. paradoxus* lineages. tORFs_sig and tORFs_RT respectively represents tORFs detected with our custom method and RiboTaper. The figure shows tORFs detected in *Spar* lineages. **A)** Venn diagram showing the number of detected tORFs depending on the method per haplotype. **B-D)** Metagene analysis for tORFs detected **B)** only with our method, **C)** with both methods and **D)** with RiboTaper only. The mean of 5' read counts is plotted along the position relative to the start codon for significantly translated genes or tORFs. The lines of the matrix indicate the normalize coverage of genes or tORFs with significant signature of translation, with one feature per line.

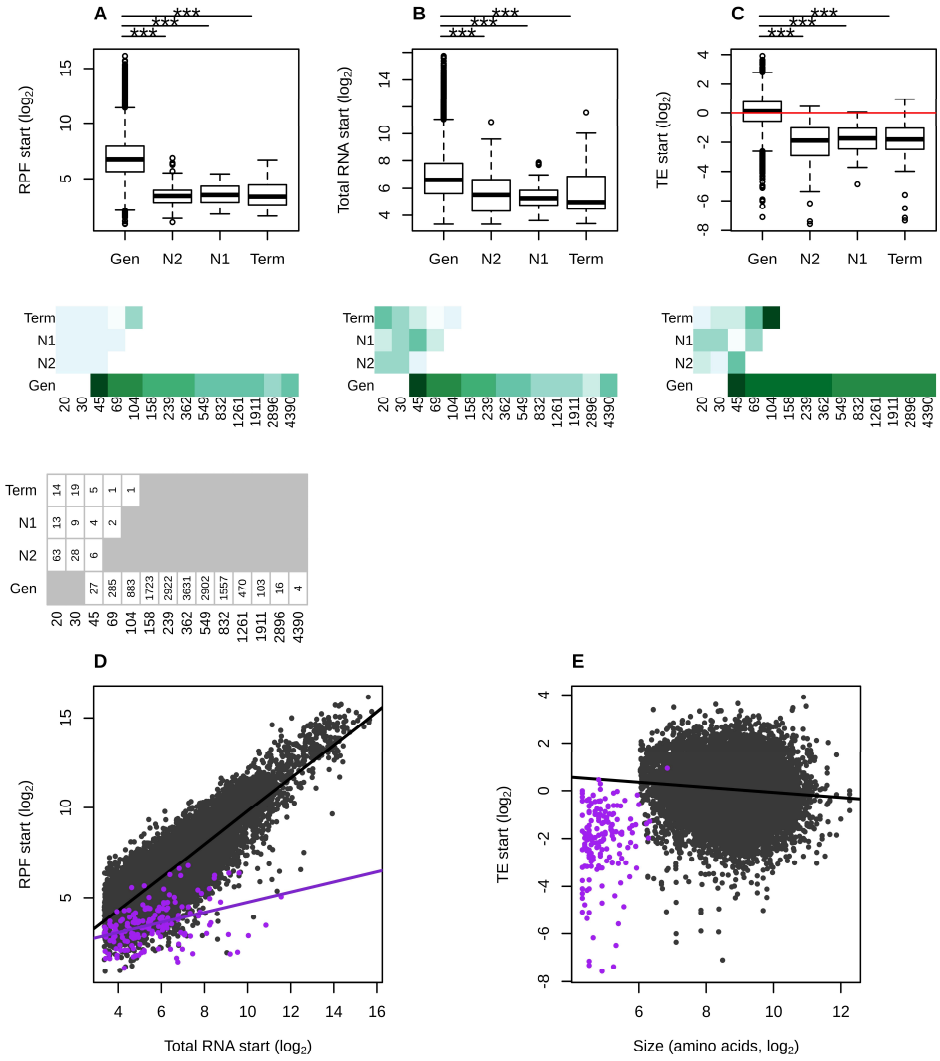


Figure S5. Expression properties of tORFs detected with RiboTaper (see Fig. 3 for tORFs detected with our custom method). A-C) Ribosome profiling (RPF), total RNA and translation efficiency (TE) - read counts in the first 60 nt, normalized to correct for library size differences in $\log_2$ - are displayed for genes (Gen) and tORFs according on their age (N2, N1 and Term). Significant differences in pairwise comparisons are displayed above each plot (Wilcoxon test, *** for p-values < 0.001, ** for p-values < 0.01 and * for p-values < 0.05). Mean estimates per size range are colored in shades of green (from pale for low values to dark green for high values) below. **D)** RPF plotted as a function of total RNA for tORFs in purple, or genes in grey**. E)** TE plotted as a function of tORF or gene sizes (number of amino acid residues in $\log_2$). Regression lines are plotted for significant Spearman correlations (p-values < 0.05). Expression levels were calculated using the mean of the two replicates. We excluded tORFs and genes with less than ten total RNA reads in the first 60 nt (see Methods). Among the 526 tORFs detected with RiboTaper, 165 were supported by at least 10 total RNA reads (see Methods) in the first 60 nt and were used in this analysis.

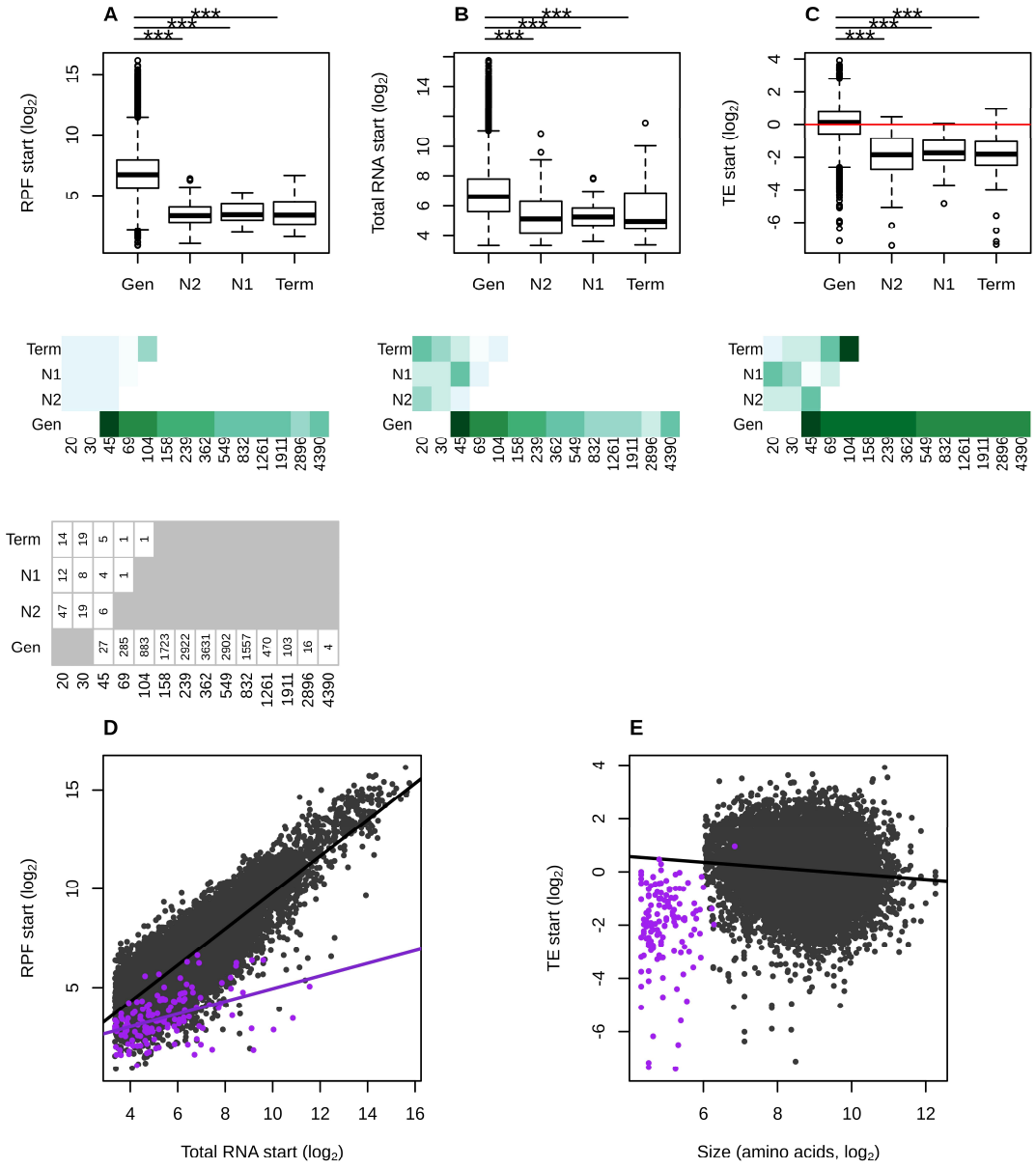


Figure S5. Continued. Expression properties of tORFs detected with both our custom analyses and RiboTaper. Among the 190 tORFs detected with both methods, 132 were supported by at least 10 total RNA reads (see Methods) in the first 60 nt and were used in this analysis.

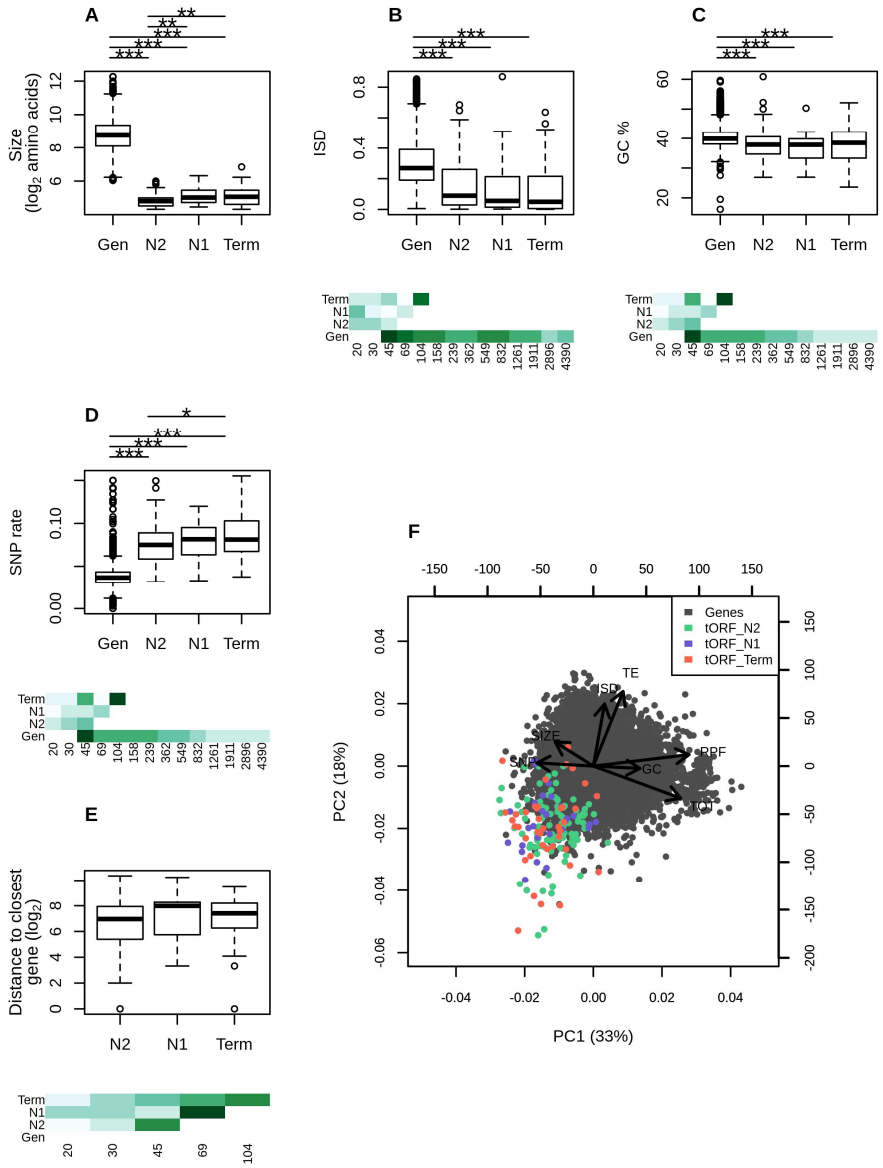


Figure S6. Sequence properties of tORFs detected with RiboTaper (see Fig. 4 for tORFs detected with our custom method). A-E) Sizes (log_2 number of residues), mean disorder (ISD), GC content (%), SNP density and distance to the closest gene are displayed for genes and tORFs as a function of age (N2, N1 and Term). Pairwise significant differences are displayed above each plot (Wilcoxon test, *** for p-values < 0.001, ** for p-values < 0.01 and * for p-values <0.05). Mean estimates per size ranges are colored in shades of green (from pale for low values to dark green for high values) below. **F)** Principal component analysis using the number of residues (SIZE in log_2), ribosome profiling (RPF), total RNA (TOT) and translation efficiency (TE) (as read counts in the first 60 nt normalized to correct for library size differences and in log_2), intrinsic disorder (ISD), the GC content and SNP density (SNP). tORFs are colored according to their age. Among the 526 tORFs detected with RiboTaper, 165 were supported by at least 10 total RNA reads (see Methods) in the first 60 nt and were used in this analysis.

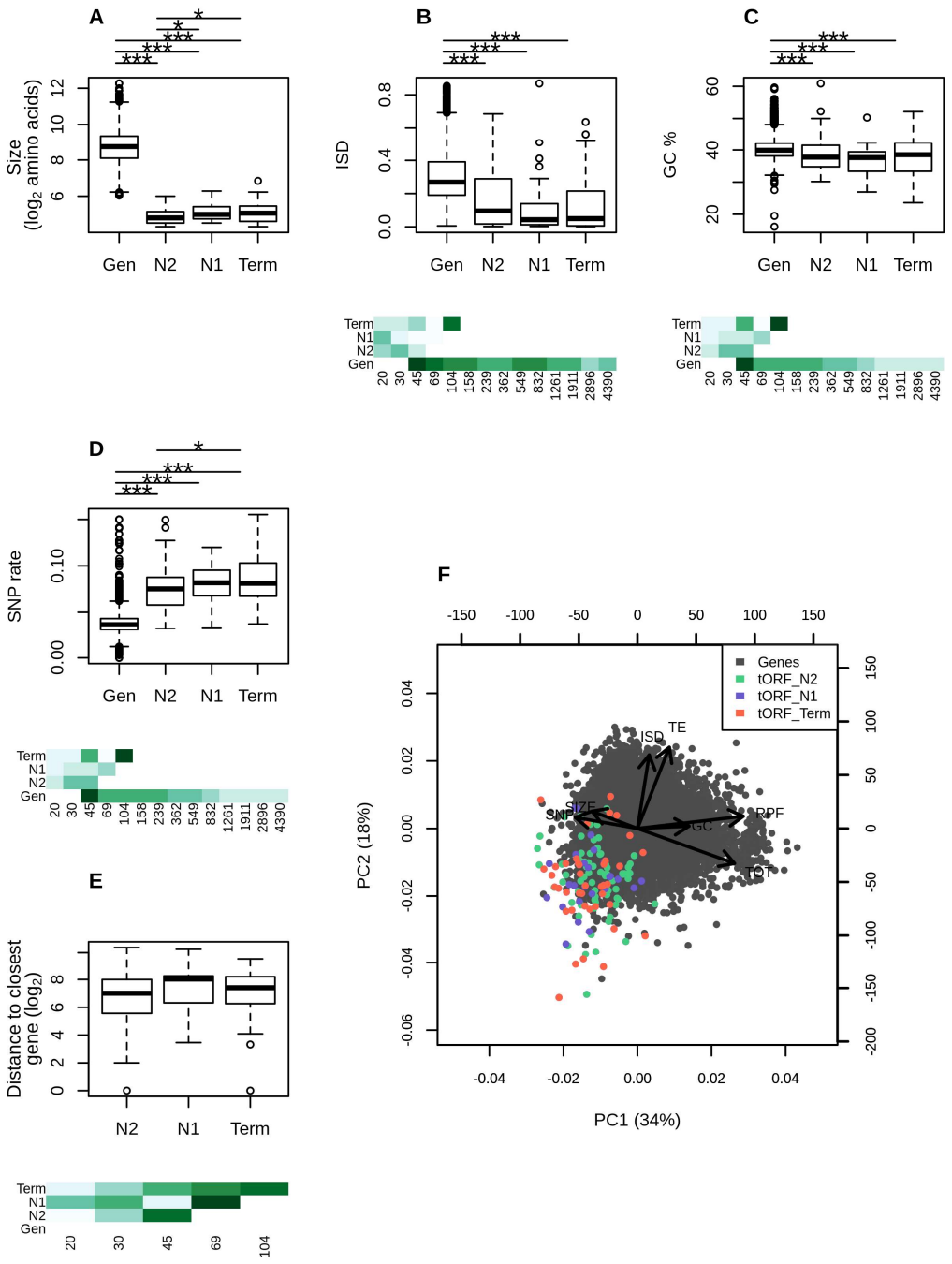


Figure S6. Continued. Sequence properties of tORFs detected with both our custom analyses and RiboTaper. Among the 190 tORFs detected with both methods, 132 were supported by at least 10 total RNA reads (see Methods) in the first 60 nt and were used in this analysis.

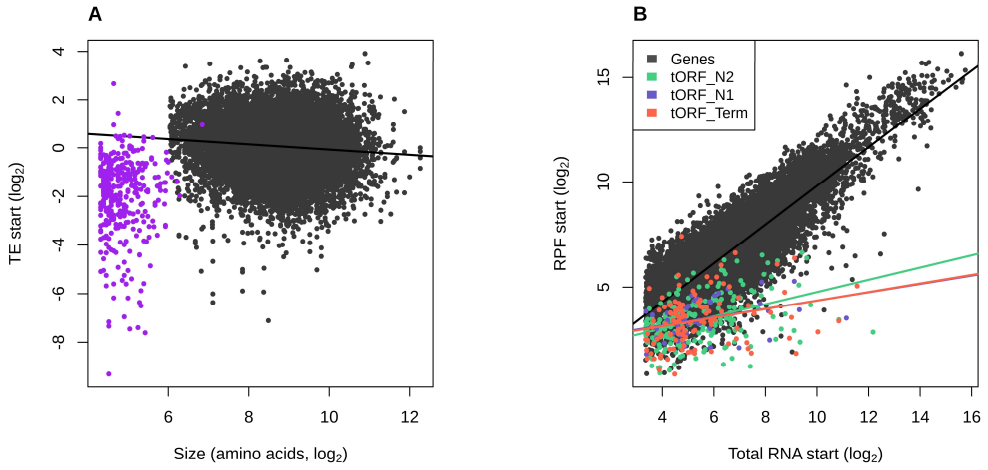


Figure S7. Expression properties of tORFs detected with our custom methods. A) TE plotted as a function of tORF (purple) or gene (grey) sizes (number of amino acid residues in $\log_2$). **B)** RPF plotted as a function of total RNA for tORFs, colored according their ages (Table 1), or genes in grey**.** Regression lines are plotted for significant Spearman correlations (p-values < 0.05). Expression levels were calculated using the mean of the two replicates. We observed no significant pairwise differences between slopes for tORFs of different ages (ANCOVA, p-values = 0.38, 0.29 and 0.96 for N2/N1, N2/Term and N1/Term comparisons respectively).

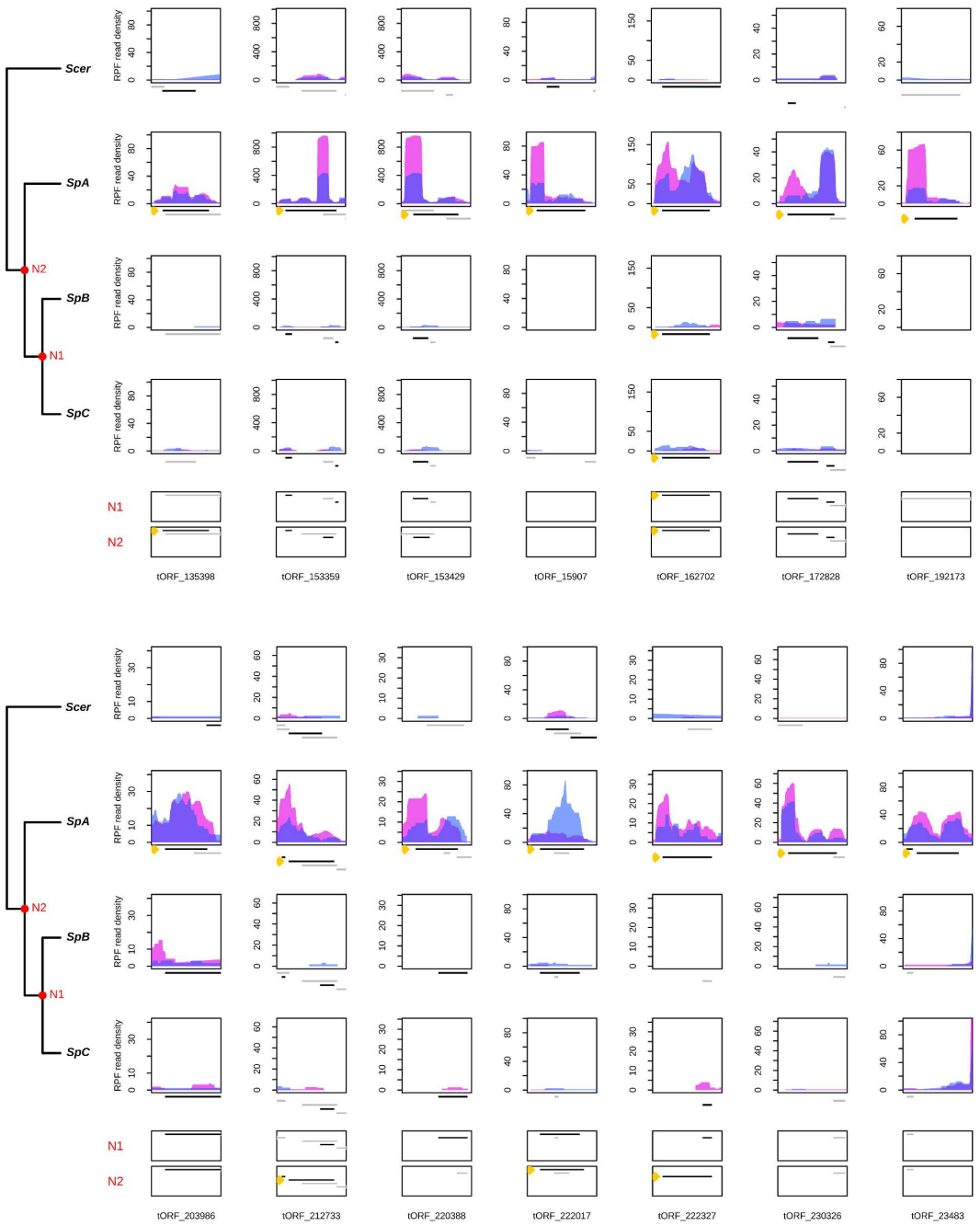


Figure S8. Normalized RPF read coverage for lineage specific (or group specific) tORFs. RPF read coverage is displayed for replicate 1 and 2 with a blue or pink area respectively. The positions of all iORFs (including ntORFs and tORFs) in the genomic area are drawn below each plot. The tORF of interest is labeled with a yellow dot and is plotted in black. iORFs overlapping the iORF of interest are plotted in black when they are in the same reading frame, and in grey when they are in a different reading frame as the selected tORF.

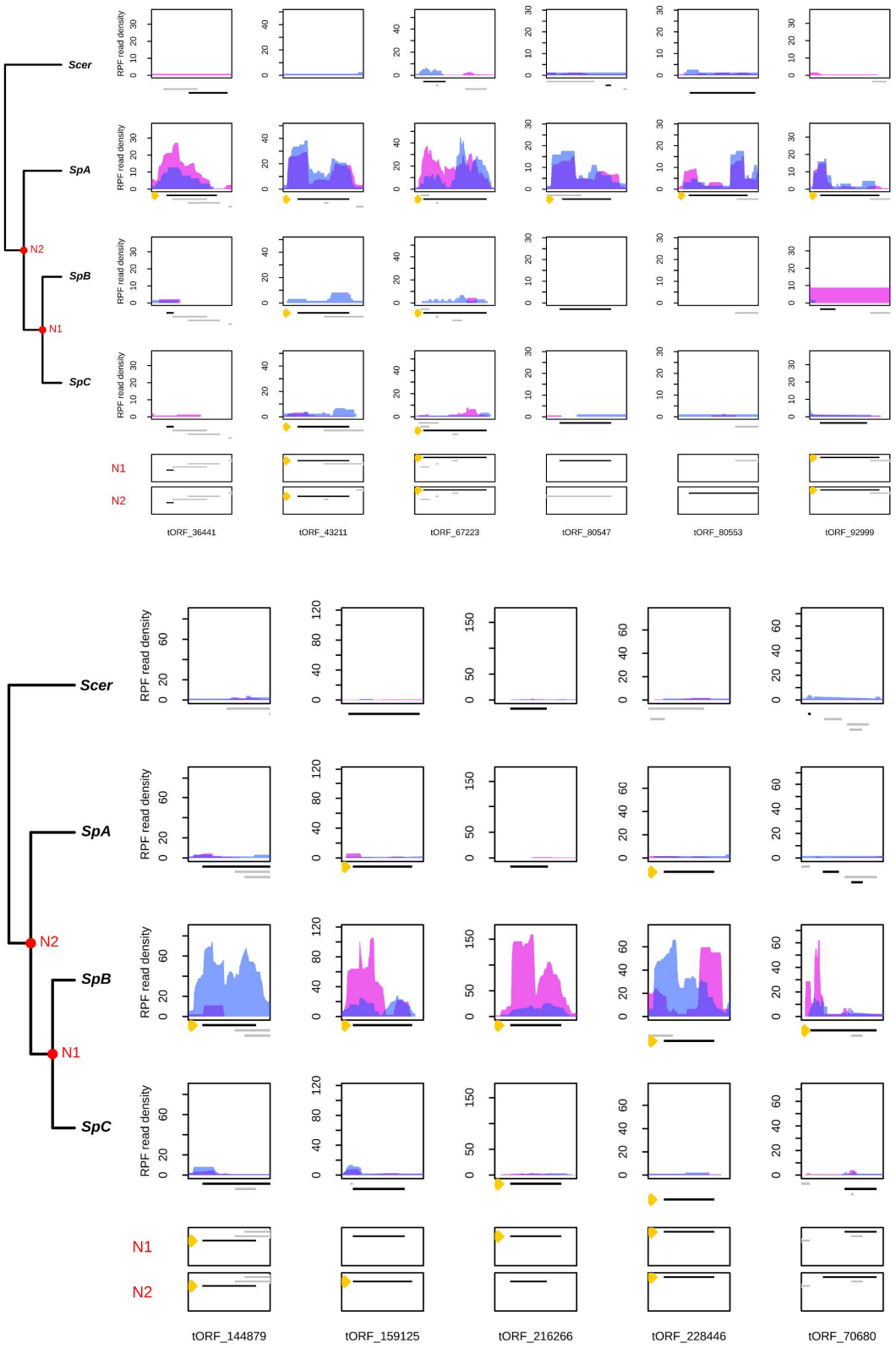
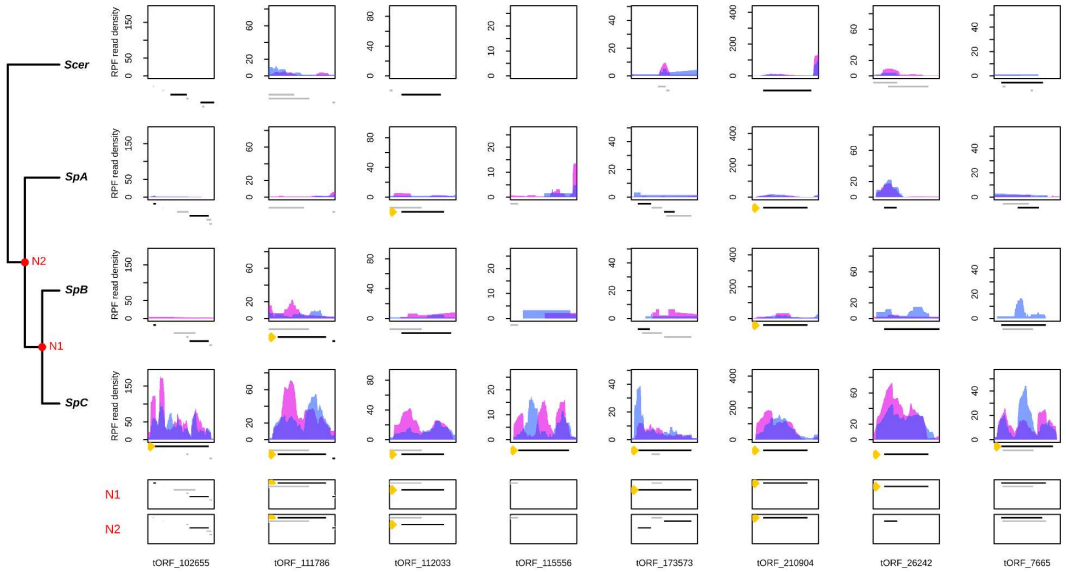


Figure S8. Continued.



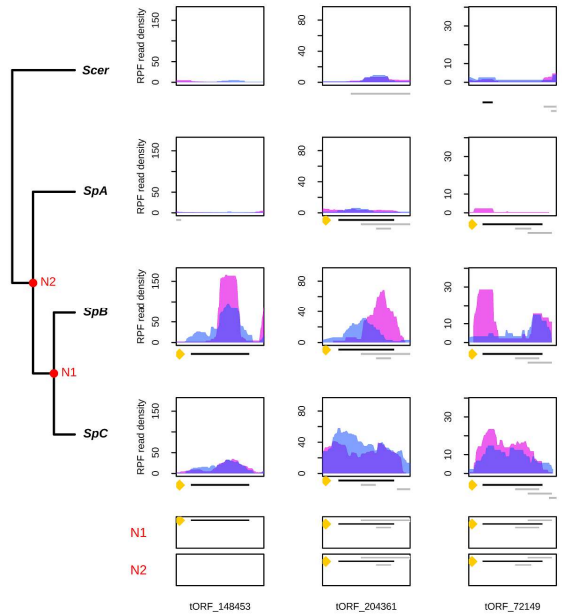


Figure S8. Continued.

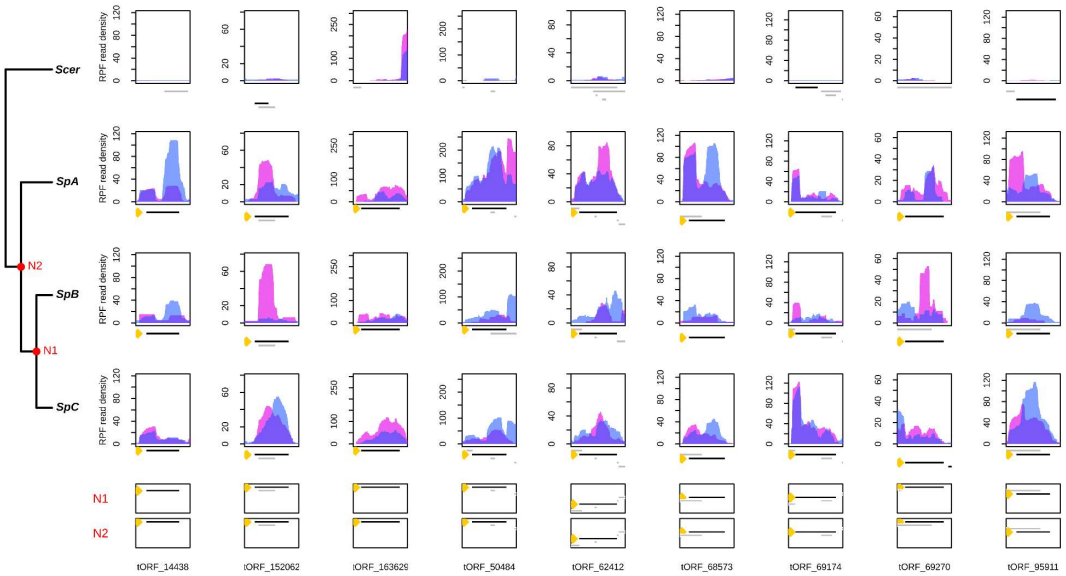


Figure S8. Continued.

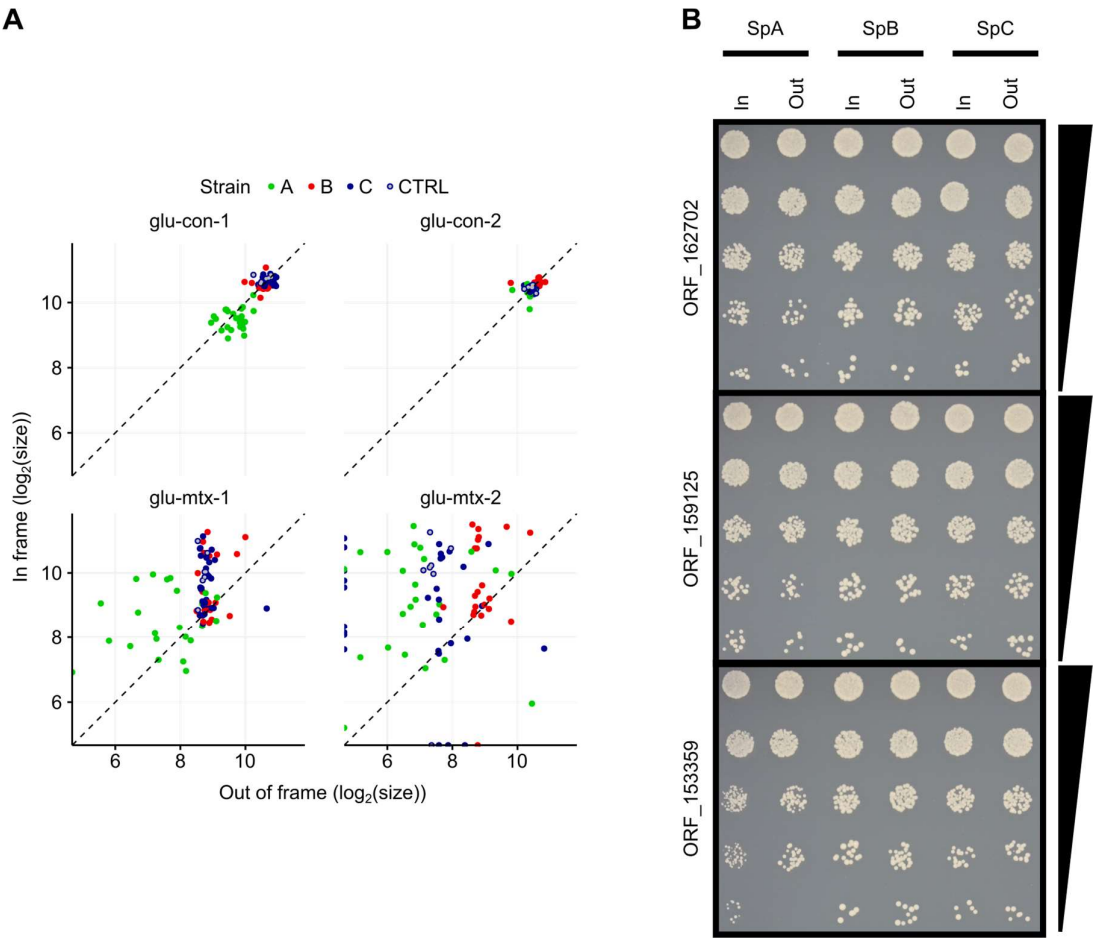


Figure S9. DHFR tagging confirms the translation of tORFs (control experiments). This figure is related to Fig. 6 in the main text. **A)** as in Fig. 6 B, it shows the $\log_2$ colony size of strains tagged either in-frame (y-axis) or out-of-frame (x-axis) with DHFR that confers resistance to methotrexate (MTX). Here, results are shown with (MTX) and without (con) methotrexate added to the media, and the data for the first (1) and second (2) phenotyping run. **B)** The control experiment for the spot dilution assays in Fig. 6C. These are the same strains, spotted at the same time from the same dilutions as in Fig. 6C but in media not supplemented with methotrexate. It shows that strains do not have growth defects.

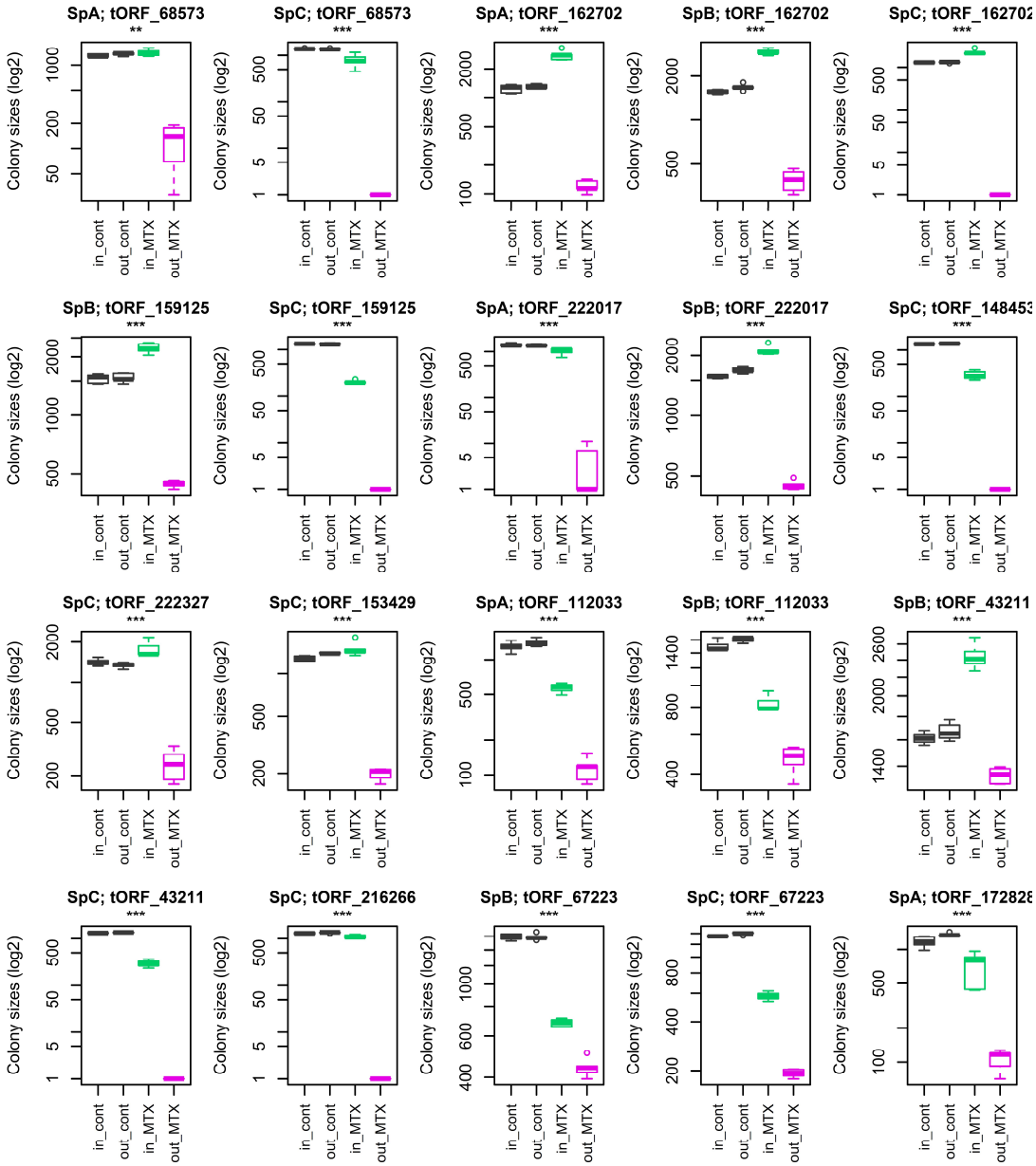


Figure S10. Detection of translated tORFs using the reporter gene DHFR. Each plot displays growth assay results per tORF and lineage (see headers). Note that we did not always get tORFs constructions in the three lineages so some tORFs were tested only in one or two strains (Supplemental Table S5). Positive controls (in frame and out of frame without MTX medium) are in grey, negative controls (out of frame in MTX) are in pink, and the tORFs tested for translation (in frame in MTX) are in green. 6 replicates were performed in each case. Translation was detected i) when we observed $\log_2$ colony size differences between in frame and out of frame constructions on MTX medium with a student t-test, and ii) if both positive controls display colony sizes of more than 1000 and with similar growth for both controls. Significant translated tORFs are indicated by *** for p-values < 0.001, ** for p-values < 0.01 and * for p-values < 0.05. Significant differences indicated between parenthesis correspond to either tORFs with colony size differences between

controls, or to a significant higher colony size on MTX for the out of frame construct, compared with the in frame. Both situations were not considered as a translation signal for the tORF tested, the latest situation is probably due to a translated overlapping iORF.

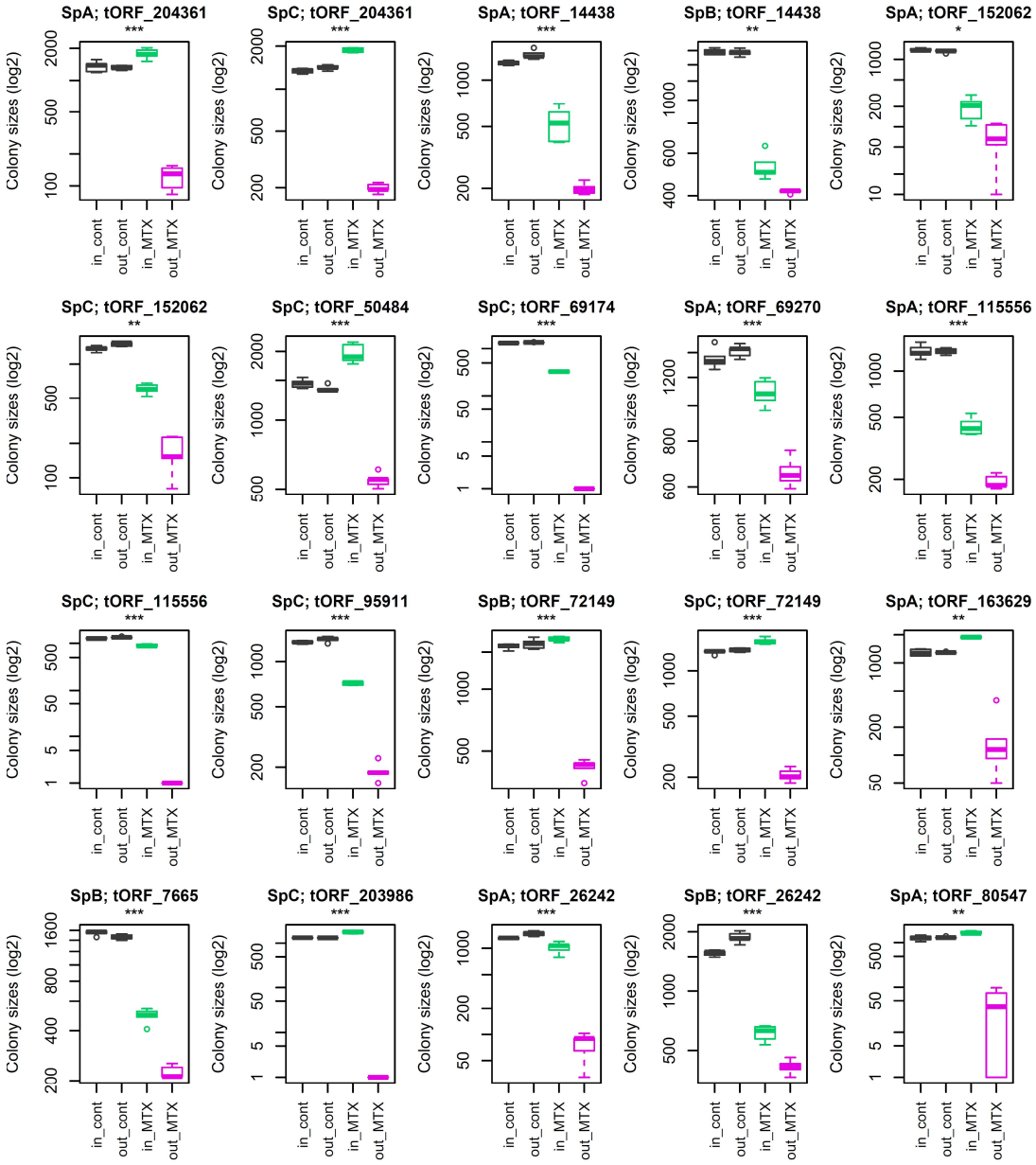


Figure S10. Continued. Detection of translated tORFs using the reporter gene DHFR.

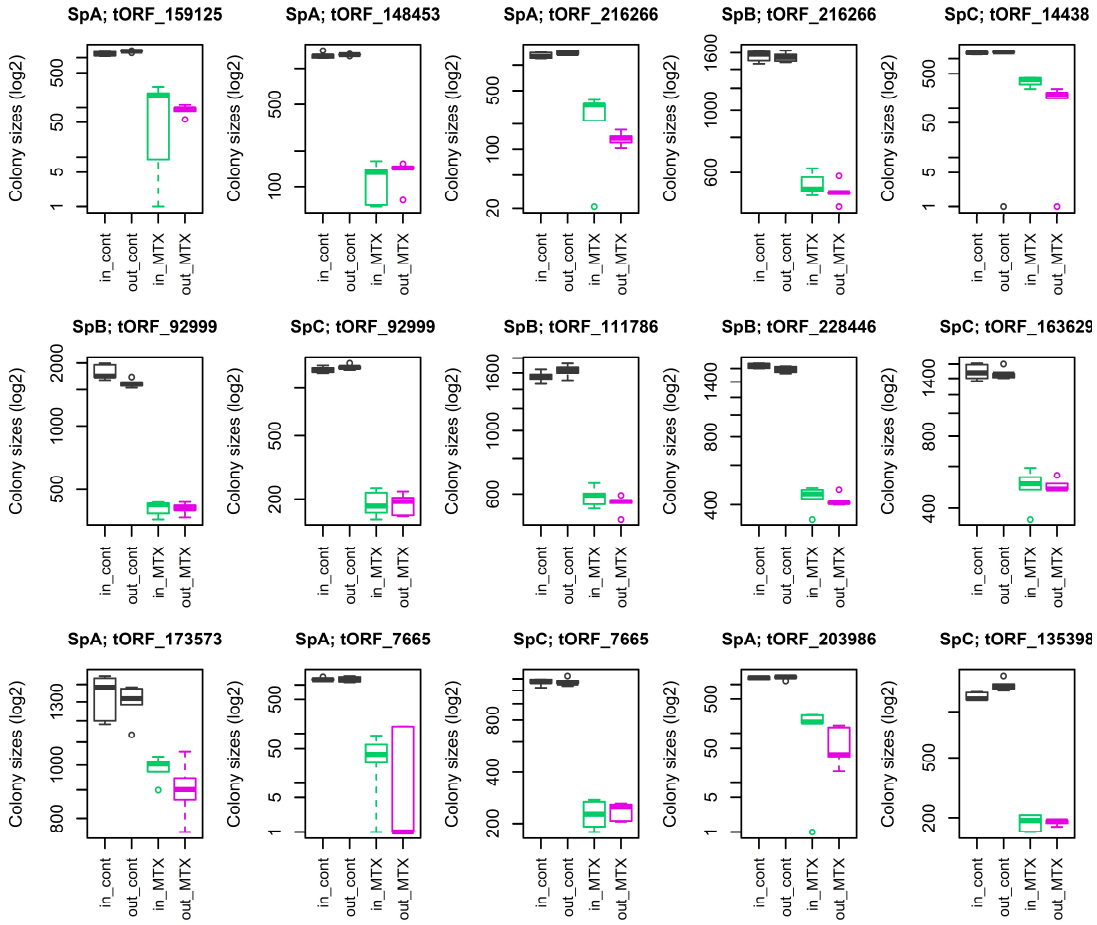


Figure S10. Continued. Detection of translated tORFs using the reporter gene DHFR.

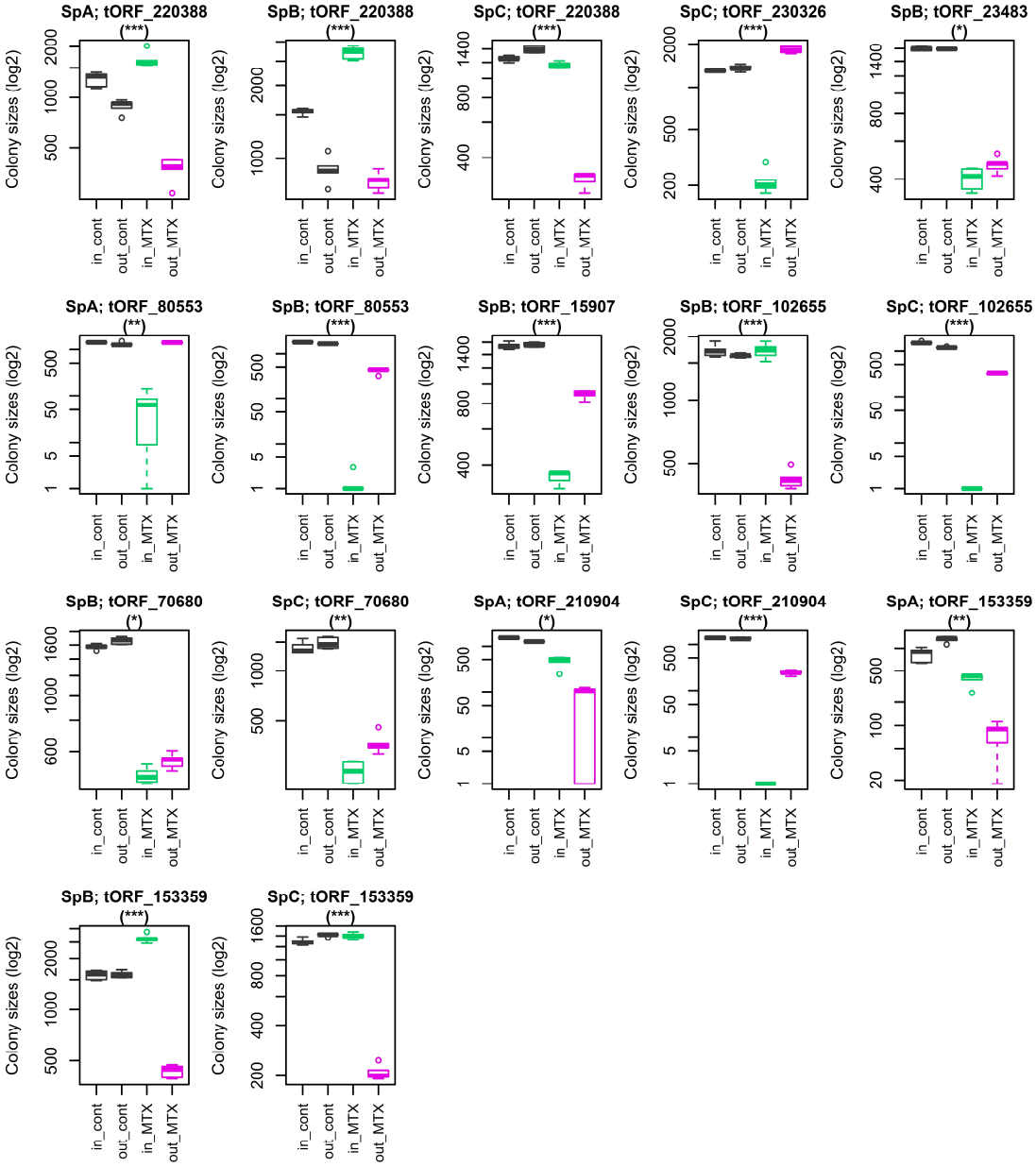


Figure S10. Continued. Detection of translated tORFs using the reporter gene DHFR.

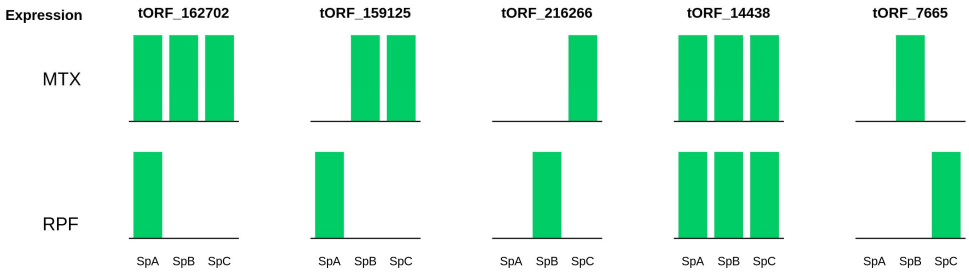


Figure S11. Comparison of translation signals detected with the growth assay (MTX, above) or by RPF sequencing (below) in the three *S. paradoxus* lineages. Each bar represents the detection of a translation signal per strain and per tORF, no bar is an absence. Here, we see that some tORFs are translated only under specific conditions (i.e. tORF_216266), others are more broadly translated (i.e. tORF_14438). Growth conditions therefore appear to have a strong effect on the expression of tORFs. In addition, the DHFR fusion approach could be more sensitive that the RPF approach for detecting translation.

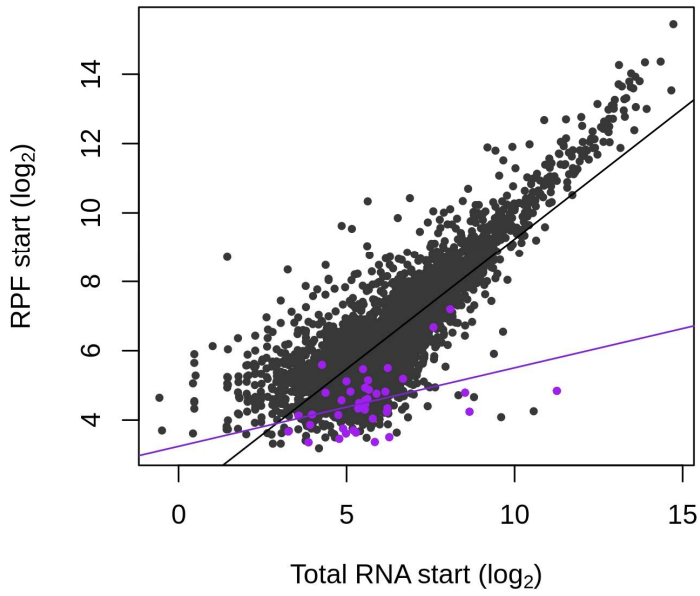


Figure S12. TE buffering in *S. cerevisiae*. Read counts from ribosome profiling (RPF) plotted as a function of read counts total RNA for total RNA for tORFs in purple, or genes in dark grey using ribosome profiling and mRNA sequencing from (McManus et al. 2014). Read counts were normalized to correct for library size differences. tORFs were identified based on iORFs annotations on the S228C reference strain (including *Scer* specific annotations), using the same procedure as in our analyses. We detected 40 tORFs in this dataset, which is smaller compared to our data probably due to a lower RPF coverage here. Regression lines are plotted for significant Spearman correlations (p-values < 0.05).

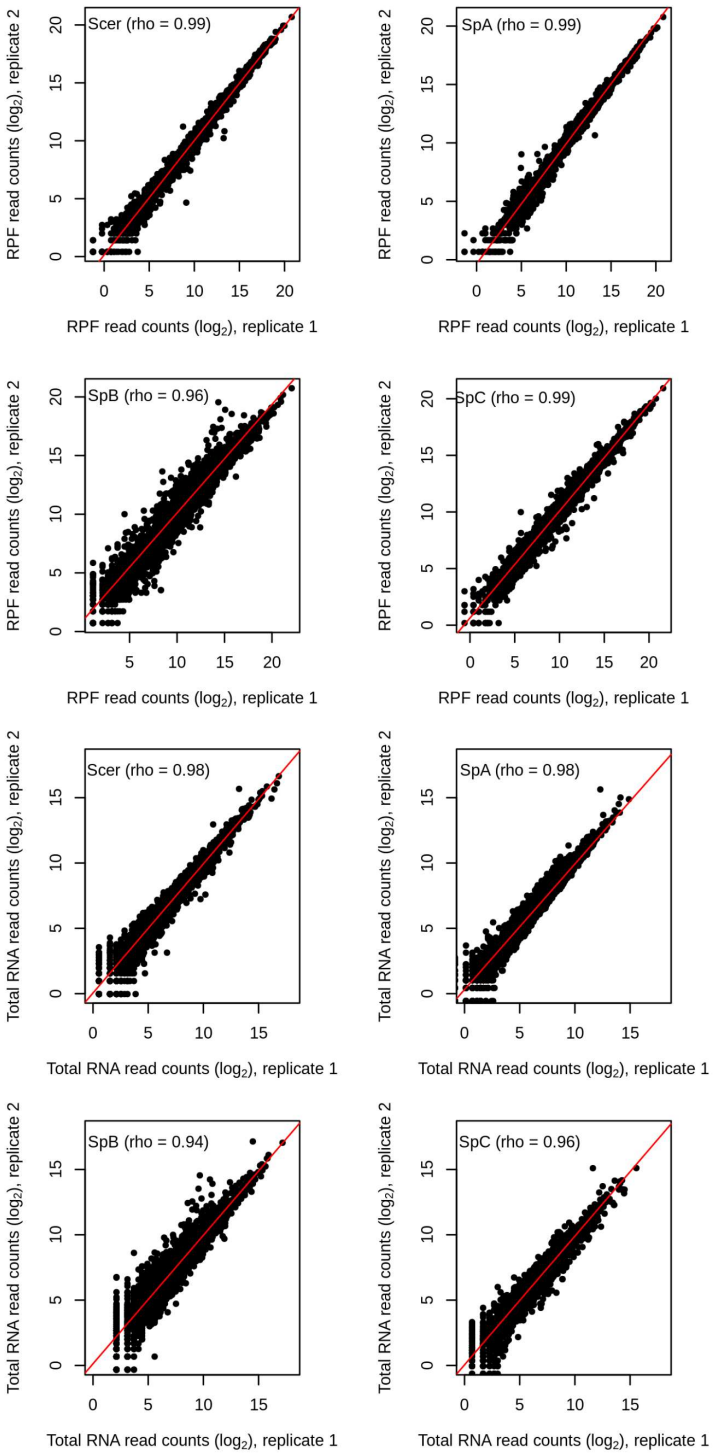


Figure S13. Comparison of RPF or Total RNA replicate sequencing experiments. The number of normalized read counts for biological replicate 1 is plotted against biological replicate 2 for each gene or tORF, per library type (RPF or Total RNA) and per strain. The Spearman's rank correlation coefficient (rho) is shown for each plot.