Title

The high 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[1,2,3,4,5], Isabelle Gagnon-Arsenault[1,2,3,4,5], Isabelle Hatin[6], Lou Nielly-Thibaut[1,2,3,4], Olivier Namy[6] & Christian R Landry[1,2,3,4,5]

[1.] Institut de Biologie Intégrative et des Systèmes; [2.] PROTEO; [3.] Centre de Recherche en Données Massives de l'Université Laval; [4.] Département de biologie; [5.] Département de biochimie, microbiologie et bioinformatique. Université Laval, Québec, Québec, G1V 0A6, Canada. [6.] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, 91190 Gif sur Yvette, France.

Correspondence to: christian.landry@bio.ulaval.ca, eleonore.durand11@gmail.com

Running title: High turnover of *de novo* ORFs in yeast populations

Keywords: *De novo* gene birth, wild yeast populations, *Saccharomyces paradoxus*

Abstract

Little is known about the rate of emergence of genes *de novo*, how they spread in populations and what their initial properties are. We examined wild *Saccharomyces paradoxus* populations to characterize the diversity and turnover of intergenic ORFs over short evolutionary time-scales. We identified ~34,000 intergenic ORFs per individual genome for a total of ~64,000 orthogroups, which resulted from an estimated turnover rate relatively smaller than the rate of gene duplication in yeast. Hundreds of intergenic ORFs show translation signatures, similar to canonical genes, but lower translation efficiency, which could reduce their potential production cost or simply reflect a lack of optimization. Translated intergenic ORFs tend to display low expression levels with sequence properties that are on average closer to expectations based on intergenic sequences. However, some predicted *de novo* polypeptides with gene-like properties emerged from ancient as well as recent birth events, illustrating that the raw material for functional innovations may appear even over short evolutionary time-scales. Our results suggest that variation in the mutation rate along the genome impacts the turnover of random polypeptides, which may in turn influence their early evolutionary trajectory. Whereas low mutation rate regions allow more time for random intergenic ORFs to evolve and become functional before being lost, mutation hotspots allow for the rapid exploration of the molecular landscape, thereby increasing the probability to acquire a polypeptide with immediate gene-like properties and thus functional potential.

Introduction

The emergence of new genes is a driving engine for phenotypic evolution. New genes may arise from pre-existing gene structures through genome rearrangements, such as gene duplication followed by neo-functionalization, gene fusion or horizontal gene transfer, or *de novo* from previously non-coding regions (Chen et al. 2013). The mechanism of *de novo* gene birth has long been considered unlikely to occur (Jacob 1977) until the last decade during which comparative genomics approaches shed light on the role of intergenic regions as a regular source of new genes

(Tautz and Domazet-Loso 2011; Landry et al. 2015; Schlotterer 2015; McLysaght and Hurst 2016). Compared to other mechanisms, the *de novo* gene origination is a source of complete innovation because the emerging genes come from mutations alone not from the evolution of preexisting functions (McLysaght and Hurst 2016).

Non-coding regions undergo three major steps to become gene-coding, the first two occurring in any order. First, the acquisition of an Open Reading Frames (ORFs) by mutations conferring a gain of in-frame start and stop codons and second, the acquisition of regulatory sites to allow the ORF transcription and translation and to produce *de novo* polypeptides. The third step would correspond to the retention of this structure by natural selection because of its positive effects on fitness (Schlotterer 2015; Nielly-Thibault and Landry 2018). The subsequent maintenance of the structure by purifying selection will lead to the gene being shared among species, as we see for groups of orthologous canonical genes. There are many ORFs associated with ribosomes in non-annotated regions, supporting their translation and the potential to produce *de novo* polypeptides which are the raw material necessary for *de novo* gene birth (Ingolia et al. 2009; Wilson and Masel 2011; Carvunis et al. 2012; Ruiz-Orera et al. 2014; Lu et al. 2017; Vakirlis et al. 2017; Ruiz-Orera et al. 2018). We distinguish *de novo* polypeptides, encoded by intergenic ORFs, from *de novo* genes because most of *de novo* polypeptides could be non-functional and seem to evolve neutrally (Ruiz-Orera et al. 2018).

Many putative *de novo* genes have been identified (McLysaght and Hurst 2016), but there is generally limited information on their translation and few have been functionally characterized (Begun et al. 2006; Levine et al. 2006; Begun et al. 2007; Cai et al. 2008; Zhou et al. 2008; Knowles and McLysaght 2009; Li et al. 2010; Baalsrud et al. 2017). *De novo* young genes are generally small with a simple intron-exon structure, they are less expressed on average than canonical genes and they may diverge rapidly compared to older genes (Wolf et al. 2009; Tautz and

Domazet-Loso 2011), which makes it more challenging to differentiate *de novo* emerging young genes from non-functional ORFs (McLysaght and Hurst 2016). The absence of sequence similarities of a gene with known genes in other species is not an evidence of a *de novo* origination, and may also be due to rapid divergence between two orthologs. This confusion resulted in spurious *de novo* origin annotations, especially over longer evolutionary time-scale (Gubala et al. 2017). One way to overcome this problem is to compare closely related populations or species and to identify the homologous non-coding sequences through synteny, for instance, which may give access to the causal gene-birth most recent mutation (Begun et al. 2006; Levine et al. 2006; Begun et al. 2007; Cai et al. 2008; Zhou et al. 2008; Knowles and McLysaght 2009; Li et al. 2010).

The process of *de novo* gene birth was framed under hypotheses that consider the role of selection as acting at different time points. The continuum hypothesis involves a gradual change between non-genic to genic characteristics as observed for intergenic ORF sizes for instance (Carvunis et al. 2012). The preadaptation hypothesis predicts extreme levels of gene-like traits in *de novo* young genes, as observed for the intrinsic structural disorder, which is higher in *de novo* young genes compared to ancient ones in some species (Wilson et al. 2017). These two models depend on the distribution of random polypeptide properties and the position of the ones with an adaptive potential in this distribution. Under the continuum hypothesis, selection acts on polypeptides with intergenic-like characteristics on average, which will mature progressively towards gene-like properties. Under the pre-adaptation hypothesis, selection acts on polypeptides with gene-like characteristics located at the extremes of the distribution, and which are more favorable for gene birth (i.e. maintenance by natural selection) than the average of non-coding sequences. If they are preferentially maintained, this creates a gap between the distribution of random polypeptides properties and recently emerging *de novo* genes.

Another question of interest is whether local composition along the genome can accelerate gene birth. The size of intergenic regions, their GC composition and the genomic context (abundant spurious transcription) may affect the birth rate of *de novo* genes (Vakirlis et al. 2017; Nielly-Thibault and Landry 2018). The comparison of yeast species showed that *de novo* genes are preferentially found in GC-rich genomic regions, in recombination hotspot and at the proximity of divergent promoters (Vakirlis et al. 2017). It was also demonstrated that mutation rate varies along chromosomes, for instance it is lower closer to replication origins, especially those that fire early in S phase (Chuang and Li 2004; Stamatoyannopoulos et al. 2009; Lang and Murray 2011; Agier and Fischer 2012). Genomic regions with elevated mutation rate may favor the emergence of *de novo* genes but also their loss in the absence of selection, affecting overall turnover.

It is now accepted that *de novo* genes continuously emerge (Tautz and Domazet-Loso 2011; Neme and Tautz 2013; Palmieri et al. 2014; Vakirlis et al. 2017). They are also frequently lost which explains the constant number of genes observed over time (Palmieri et al. 2014). Because most studies focus on inter-species comparisons, the extent of polymorphism within species in number of *de novo* genes is largely unknown, except in a few cases (Zhao et al. 2014; Li et al. 2016). The rate at which new putative polypeptides appear from non-coding DNA, spread in population and the likelihood that they become functional and are retained by natural selection before they are lost is of strong interest to understand the dynamics of the *de novo* gene birth process. Another pressing question is what are the initial properties of the peptides produced during the neutral exploration period within species. The use of population data may address this issue and allows to precisely monitor the turnover of recently evolving polypeptides over short evolutionary time-scales.

Here we explore the contribution of the intergenic diversity in the emergence and retention of the raw material for the *de novo* gene birth in natural populations of *Saccharomyces paradoxus*. We

characterize the repertoire and turnover of ORFs located in intergenic regions (named hereafter iORFs), as well as the associated putative *de novo* polypeptides using ribosome profiling, and compare how the properties of putative polypeptides covary with their age and expression. We observe a continuous emergence of *de novo* polypeptides that are segregating within *S. paradoxus*. Compared to canonical genes, *de novo* polypeptides are on average smaller and less expressed and show a lower translation efficiency. Translation efficiency tends to decrease in the highly transcribed iORFs, suggesting a regulation acting at the translational level results in a buffering in amount of produced polypeptides, resulting perhaps in a lack of optimization. *De novo* polypeptides display a high variability for various properties, some share gene-like characteristics, suggesting that their functional potential arises directly from non-coding DNA.

Results

A large number of intergenic ORFs segregates in wild *S. paradoxus* populations

We first characterized iORF diversity in wild *S. paradoxus* populations. We used genomes from 24 strains that are structured in three main lineages named *SpA*, *SpB* and *SpC* (Charron et al. 2014; Leducq et al. 2016). Two *S. cerevisiae* strains were included as outgroups: the wild isolate YPS128 (Sniegowski et al. 2002; Peter et al. 2018) and the reference strain S288C. These lineages cover different levels of nucleotide divergence, ranging from ~ 13 % between *S. cerevisiae* and *S. paradoxus* to ~2.27 % between the two closest *SpB* and *SpC* lineages (Kellis et al. 2003; Leducq et al. 2016). We used microsynteny to identify and align homologous non-genic regions between pairs of conserved annotated genes (Fig. S1 and Methods). We identified 3,781 orthologous sets of intergenic sequences representing a total of ~ 2 Mb, with a median size of 381 bp (Fig. S1 and S2). iORFs were annotated on aligned sequences using a method similar to the one employed by Carvunis et al. (2012), that is the first start and stop codons in the same reading frame not overlapping with known features, regardless of the strand, and with no minimum size. We then classified iORFs according to their conservation level among strains (Fig. S1 and

Methods). Because the annotation was performed on aligned sequences, we could precisely detect the presence/absence of orthologous iORFs among strains, based on the conservation of an iORF with the same start and stop positions without disruptive mutations in between. We used *S. cerevisiae* as an outgroup and removed iORFs present only in this species to focus on *S. paradoxus* diversity. However, we conserved iORFs present both in *S. cerevisiae* and in at least one *S. paradoxus* strain to keep the inter-species conservation.

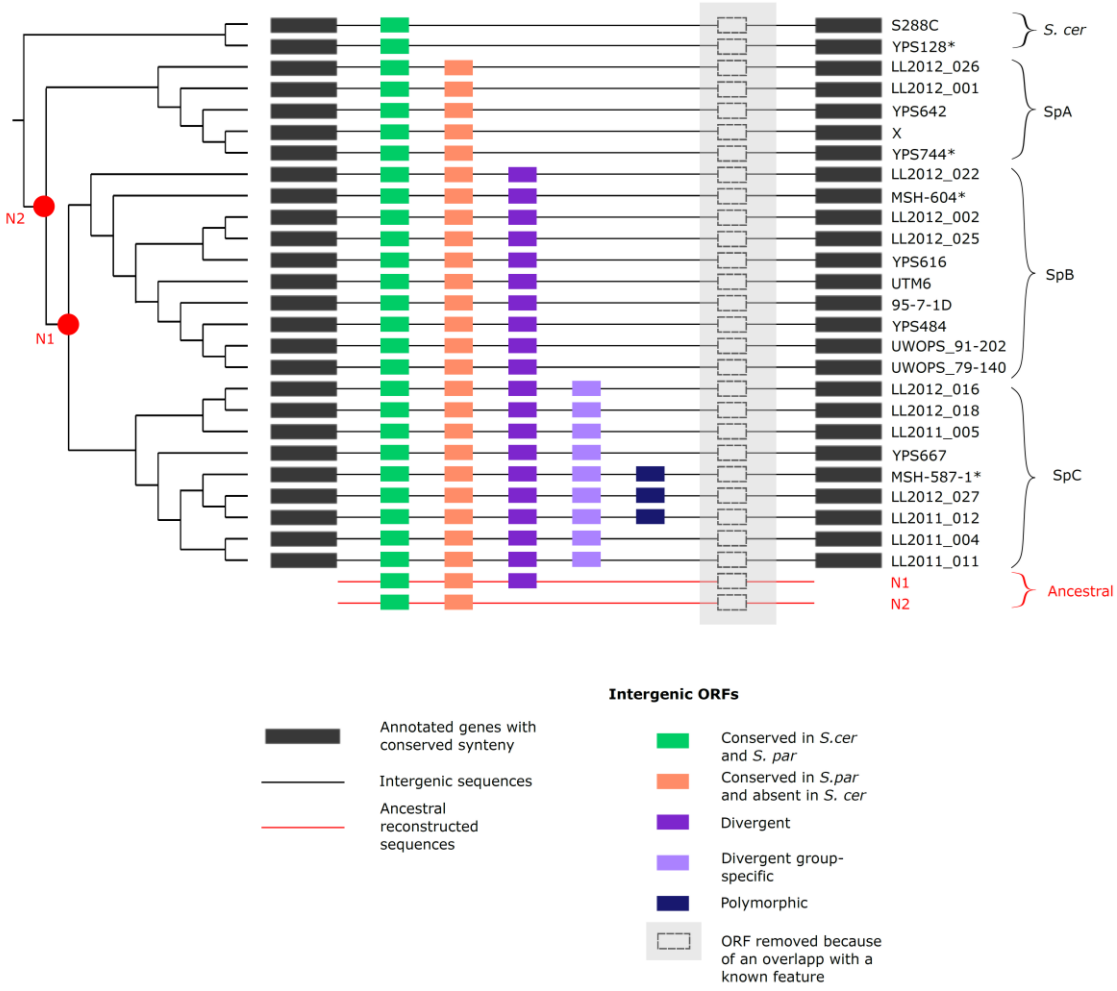

Figure S1. Identification and annotation of iORFs. Genes with conserved synteny were used as anchor to align non-genic sequences beetwen each pair. iORFs were annotated on intergenic aligned sequences and clustered as orthogroups based on the conservation of their start and stop codons aligned positions with no disruptive mutation within. iORFs displaying a sequence similarity or an overlap with a known feature were removed from all strains. Strains used for ribosome profiling experiments are marked with *.

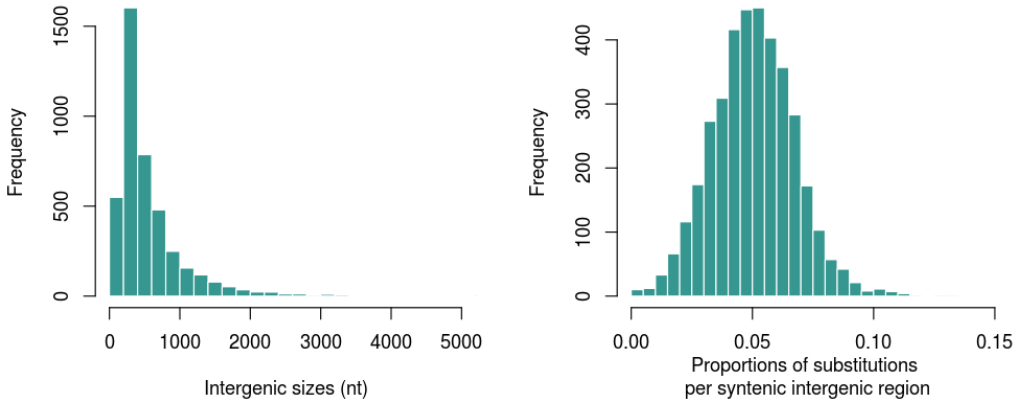

Figure S2. Sizes of intergenic regions and SNP distribution of SNP density.

We annotated 34,216 to 34,503 iORFs per *S. paradoxus* strain, for a total of 64,225 orthogroups annotated at least in one *S. paradoxus* strain (Table 1). This represents a density of about 17 iORFs per Kb. The iORFs set shows about 6 % conserved among *S. cerevisiae* and *S. paradoxus* strains, and 15 % specific and fixed within *S. paradoxus.* The remaining 79 % are still segregating within *S. paradoxus* (Fig. 1A, 1B and Table 1).

Table 1. Number of iORFs per conservation level

Conservation group	iORF family numbers	Proportion (%)
Conserved	3,961	6
Spar	9,315	15
Div	12,750	20
Spe group	22,740	35
Pol	15,459	24
Total	**64,225**	

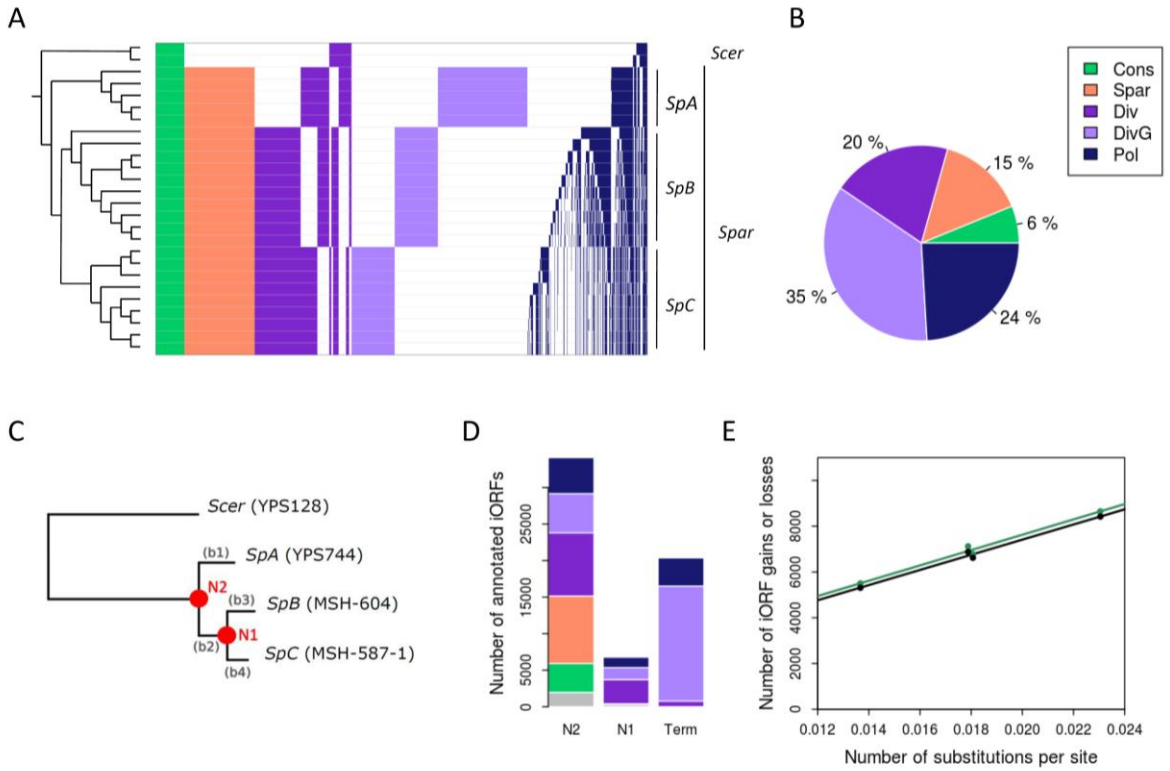

Figure 1. Evolution of iORFs in *Saccharomyces paradoxus* populations. A) Columns represent iORFs sorted according to the conservation. Absent iORFs are white. Others are colored according to their conservation group (see Methods and Fig. S1): conserved (cons), *S. paradoxus* (Spar) specific and fixed, divergent (Div), divergent group-specific (DivG) and polymorphic (Pol). **B)** Percentage of iORFs belonging to each conservation group. **C)** Phylogenetic tree of strains used for the reconstruction of ancestral intergenic sequences. Node and branch names are indicated in orange and grey respectively. **D)** 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. **E)** Number of iORF gains (in green) or losses (in black) as a function of the number of substitutions per site. Points show iORF counts on each phylogenetic branch (b1 to b4) in our dataset.

To understand how iORF diversity changes over a short evolutionary time scale, we estimated iORFs' age and turnover using ancestral sequence reconstruction (Fig. 1C) (see Methods). Because polymorphism within lineages (*SpA*, *SpB* or *SpC*) (Leducq et al. 2016) may affect the topology of the phylogeny (although most diversity in this group is among lineage divergence) we used only one strain per lineage (YPS128 (*S. cerevisiae*), YPS744 (*SpA*), MSH-604 (*SpB*) and MSH-587-1 (*SpC*)) to reconstruct ancestral sequences at two divergence nodes that we labeled

N1 for *SpB-SpC* divergence and N2 for *SpA-SpB/C* divergence. These strains contain 58,952 iORF orthogroups after removing the polymorphic iORFs that are absents in all the four selected strains. Reconstructed sequences were included in intergenic alignments of actual strains and were used to detect the presence or absence ancestral iORFs at each node (Fig. 1C, S1 and Methods).

We estimated the age of the 58,952 iORFs and annotated the 2,291 iORFs detected only in ancestral sequences. 55 % of iORFs were present at N2 (the oldest age category) and are represented in each conservation group depending on iORF loss events occurring after N2 (Fig. 1D and Table 2). We observed a continuous emergence of iORFs with 6,782 iORF gains between N2 and N1 and 5,324 to 8,454 along terminal branches. As expected, the number of iORF gains or losses is correlated and increases with branch length (Fig. 1E). We estimated a rate of emergence and loss at respectively 0.28 +/- 0.01 and 0.27 +/-0.008 ORFs per nucleotide substitution. An ORF is on average gained or lost at every 3.5 substitutions. The *de novo* ORF gain rate, estimated at around 1.1×10^{-3} ORFs per genome per cell division, is about one order of magnitude smaller than the gene duplication rate in *S. cerevisiae* estimated at 1.9×10^{-2} genes per genome per cell division (Lynch et al. 2008).

Table 2. Estimated age of iORFs in *S. paradoxus* lineages

Age (Node or branch)[1]	Numbers	Numbers > or equal to 60nt[2]	Numbers with translation signature[2]
N2	34,092	8,336	221
N1	6,782	2,664	56
b1 (SpA)	8,454	3,608	73
b3 (SpB)	6,860	2,948	13
b4 (SpC)	5,324	2,235	48
Total without redundancy[2]	61,243	19,689	418

[1] N1 and N2 refers to phylogenetic nodes (see Fig. 1-C). b1, b3 and b4 are terminal branches, these categories refer to iORFs absent in ancestral sequences (base on the conservation of the start and stop position in the same reading frame).[2] Some iORF families with no ancestors (so attributed to terminal branches), were found in more than 2 lineages, see results section,

we removed the redundancy in the total counts. [2] The 12 iORFs with significant blastp hits against proteomes (see results and Methods) were removed.

We considered that iORFs with no detected ancestors appeared on terminal branches. Among them, 91 to 93 % are present only in one lineage, which is consistent with the expected conservation pattern for recently emerging iORFs (Table 2). The absence of ancestor for the remaining 7 to 9% iORFs present in more than one lineage can be due to convergence on terminal branches, made possible by the relatively high turnover rate. Convergence events may particularly occur if two lineages acquire independently small indels, not necessarily at the same position but in the same iORF, leading to the same frameshift and resulting in stop codon changes. Finally, regions with a higher rate of evolution may more likely lead to ancestral sequence reconstruction errors and to a small overestimation of the gain rate but this effect should be negligible because of the small number of iORFs with ambiguous age estimations.

As previously observed, iORFs tend to be small with a median value of 43 bp compared to known genes in the reference *S. cerevisiae* (median iORF size of 1,287 bp) (Fig. 2A). Each conservation group of iORFs also contains iORFs longer than the smallest annotated genes in *S. cerevisiae*, revealing an extended set of iORFs with coding potential. In our study, overlapping iORFs between strains, sharing the same start and a different stop position (or the reciprocal) were classified as different orthogroups because of their changed resulting sizes. We investigated the evolution of iORF sizes along the phylogeny, by connecting overlapping iORF orthogroups in actual strains with their ancestors based on the conservation of their start and/or stop positions (Fig. 2B). The majority of iORF orthogroups (65%) were conserved until N2 (Fig. 2B). We identified 19% of iORFs successively connected to N1 and N2 by one or two size changes along the phylogeny (Fig. 2B). Note that a size change is considered as an iORF loss event generally accompanied by the gain of another iORF, which is consistent with the similar iORF gain and loss rates estimated.

iORFs detected only on terminal branches with no 'connected' ancestor tend to display intermediate iORF size values compared to iORFs of conserved size and iORFs resulting from size changes (Fig. 2C).

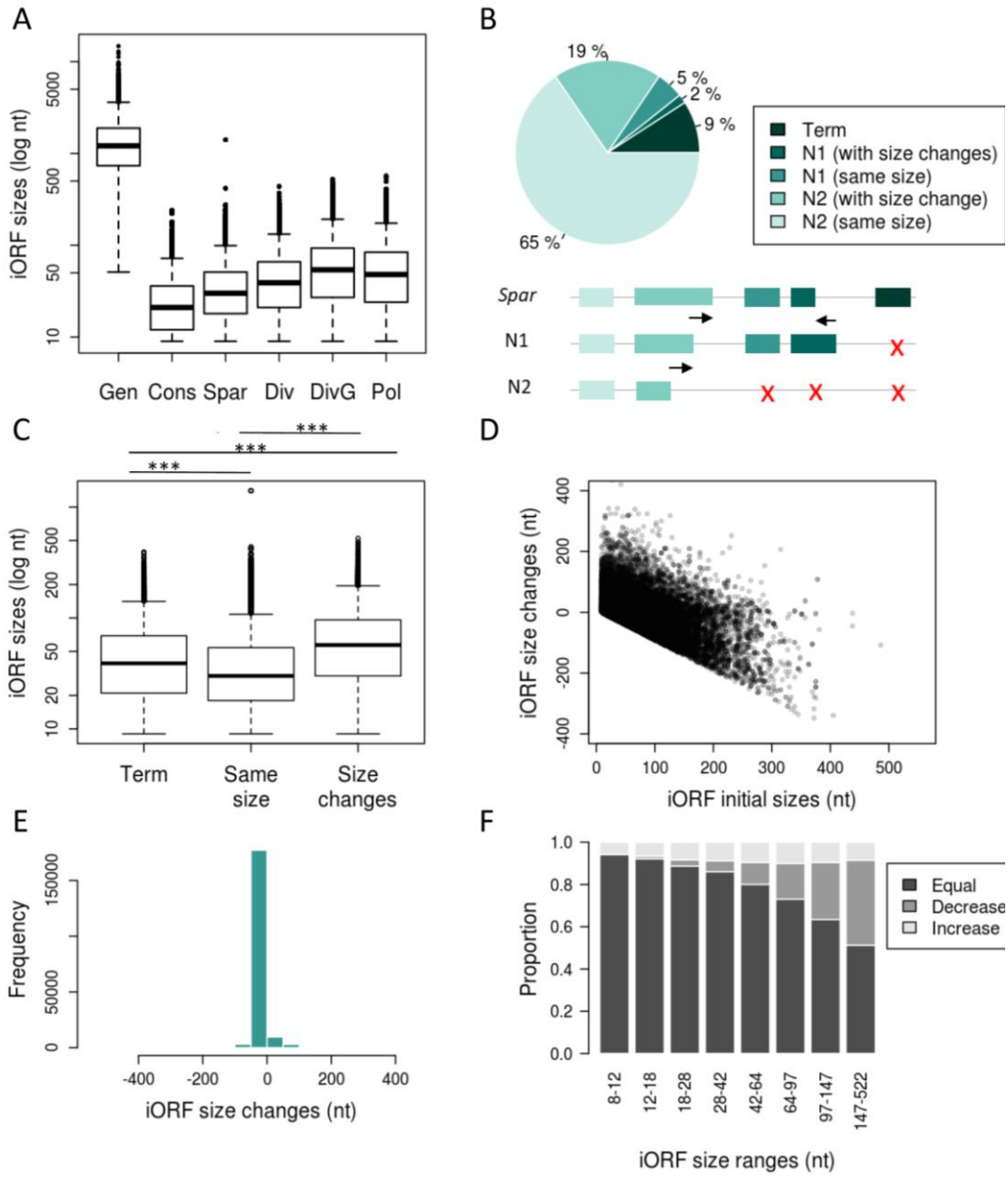

Figure 2. iORFs change their coding potential by frequent size changes. A) Genes (from the reference genome of *S. cerevisiae*) and iORFs sizes (in nucleotides and log scale). **B)** Proportion of iORFs identified in *S. paradoxus* extent sequences (*Spar*) and successively connected to ancestral nodes, or with a connection stopped in N1, or without connection in ancestral sequences (Term). **C)** Distributions of iORFs sizes in modern strains depending on the connection type with an ancestor: 'Term' refers to iORFs appearing along terminal branches and with no connection in any ancestor, 'Same size' refers to iORFs present in ancestral sequences with no size changes

and 'Size changes' refers to iORFs connected to the ancestral sequence by its start or stop position, and submitted to at least one size change. Significant pairwise differences are indicated above each comparison (Student t-test, *** for p-values < 0.001). **D)** Size changes orientation and amplitude depending on the initial iORF size at N2 or N1. No size changes were not displayed for more clarity. **E)** Distribution of iORF size changes relative to ancestral iORFs. **F)** Proportion of terminal iORFs submitted (or not) to size changes (increase or decrease) per range size relative to their ancestor at N1.

Size changes are also mainly small even if some extreme cases are observed (Fig. 2D-E). Compared to smaller iORFs, longer iORFs are less conserved and more submitted to size changes (Chi-square test, p-value < $2.2x10^{-16}$, Fig. 2C and 2F), which might be explained by the higher turnover rate of longer iORFs. This suggests a larger target for mutation accumulation between the start and the stop codon. Longer iORFs also tend to decrease, which might be due to a higher chance to acquire a disruptive mutation resulting in a size decrease, and intergenic size constrains limiting the maximum iORF sizes (Fig. 2E and S2).

Altogether, these analyses show that yeast populations' iORFs repertoire is the result of frequent gain and loss events, and of size changes. 56 % of ancient iORFs detected at N2 are still segregating within *S. paradoxus,* showing the role of wild populations as a reservoir of iORFs that can used to address the dynamics of early *de novo* gene evolution.

Intergenic ORFs frequently show signatures of active translation

We performed ribosome profiling to identify iORFs that are translated and that thus putatively produce polypeptides. Only iORFs with a minimum size of 60 bp were considered for this analysis. Among them, 12 iORFs displayed a significant hit when blasted against the proteome of 417 species, including 237 fungi, and were removed for the downstream analysis (see Methods). The set examined consists of 19,689 iORFs. We prepared ribosome profiling sequencing libraries for four strains, one belonging to each lineage or species: YPS128 (*S. cerevisiae*), YPS744 (*SpA*), MSH-604 (*SpB*) and MSH-587-1 (*SpC*), in two biological replicates. All strains were grown in

synthetic oak exudate (SOE) medium (Murphy et al. 2006) to be close to natural conditions in which *de novo* genes could emerge in wild yeast strains.

Typically, a ribosome profiling density pattern is characterized by a strong initiation peak located at the start codon followed by a trinucleotide periodicity at each codon of protein-coding ORFs. We used this feature to identify a set of translated iORFs for which we compared translation intensity with annotated genes. We first detected peaks of initiation sites in the start codon region. As expected, the number of ribosome profiling reads located at the start codon position is lower for iORFs than for annotated genes (Fig. 3A). However, there is a significant overlap between the two read density distributions, illustrating a similar read density between highly expressed iORFs and lowly expressed genes. We observed an initiation peak for 73.9 to 87.9 % of standard annotated genes depending on the haplotype, and for 1.4 to 6.9 % of iORFs (Table 3 and Fig. 3B). This suggests that at least 20% of translated iORFs could be missed using this approach, because of a too low expression levels or condition-specific expression. Detected peaks were classified using three levels of precision and intensity: ‘p1’ for less precise peaks (+/- 1nt relative to the first base of the start codon), ‘p2’ for precise peaks (detected at the exact first base of the start codon) and ‘p3’ for precise peaks with strong initiation signals characterized here by the highest read density in the ORF (see Methods). Among all iORFs with a detected initiation peak, 30, 35 and 34% respectively belong to p1, p2 and p3. A comparable repartition (Chi-square test, p-value= 0.59) was observed for genes with 24, 40 and 36% for each precision group, showing that the precision levels used in our analysis were reliable.

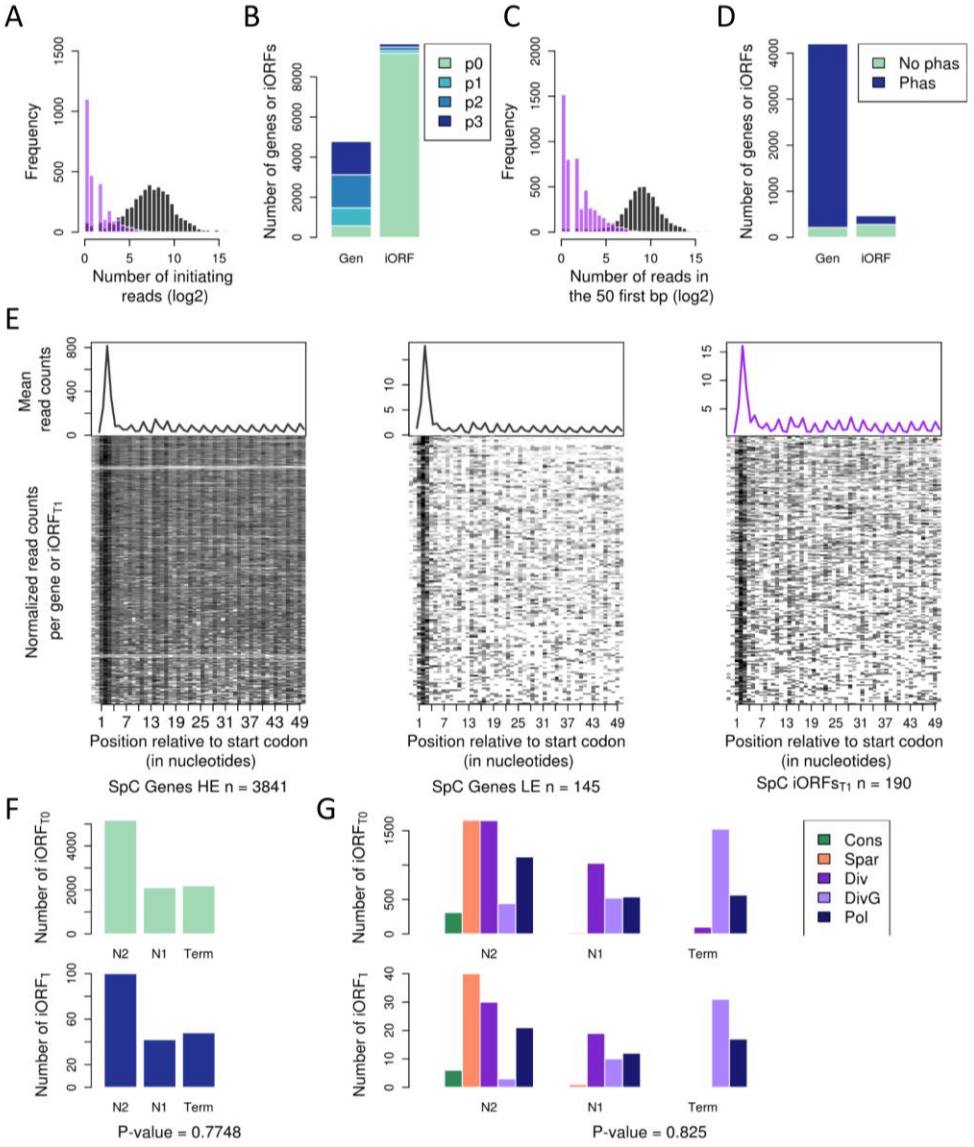

Figure 3. A fraction of iORFs displays translation signatures similar to genes. A) Distribution of the ribosome profiling read counts for genes (in grey) and iORFs (in purple) at the start codon position. **B)** Proportions of genes (Gen) or iORFs with a detected initiating peak at the start codon position. 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 the ribosome profiling read counts in the first 50 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 analysis for significantly translated highly (HE) or lowly (LE) expressed genes in grey, and intergenic iORFs$_{T1}$ in purple. The mean of 5' read counts is plotted along the position relative to the start codon for significantly translated genes or iORFs$_{T1}$. Lines of the matrix indicates the normalize coverage of all Genes or iORFs$_{T1}$ with significant signature of translation. **F-G)** Number of iORFs$_{T0}$ and iORFs$_{T1}$ per age class **(F)** or conservation group per age **(G).** P-values above are for homogeneity chi-square test to compare the proportions of each age or conservation category between iORFs$_{T0}$

and iORFs$_{T1}$. **A-G**) Display results for the *SpC* strain MSH-587-1 (see Fig. S3 for *SpA* and *SpB* results).

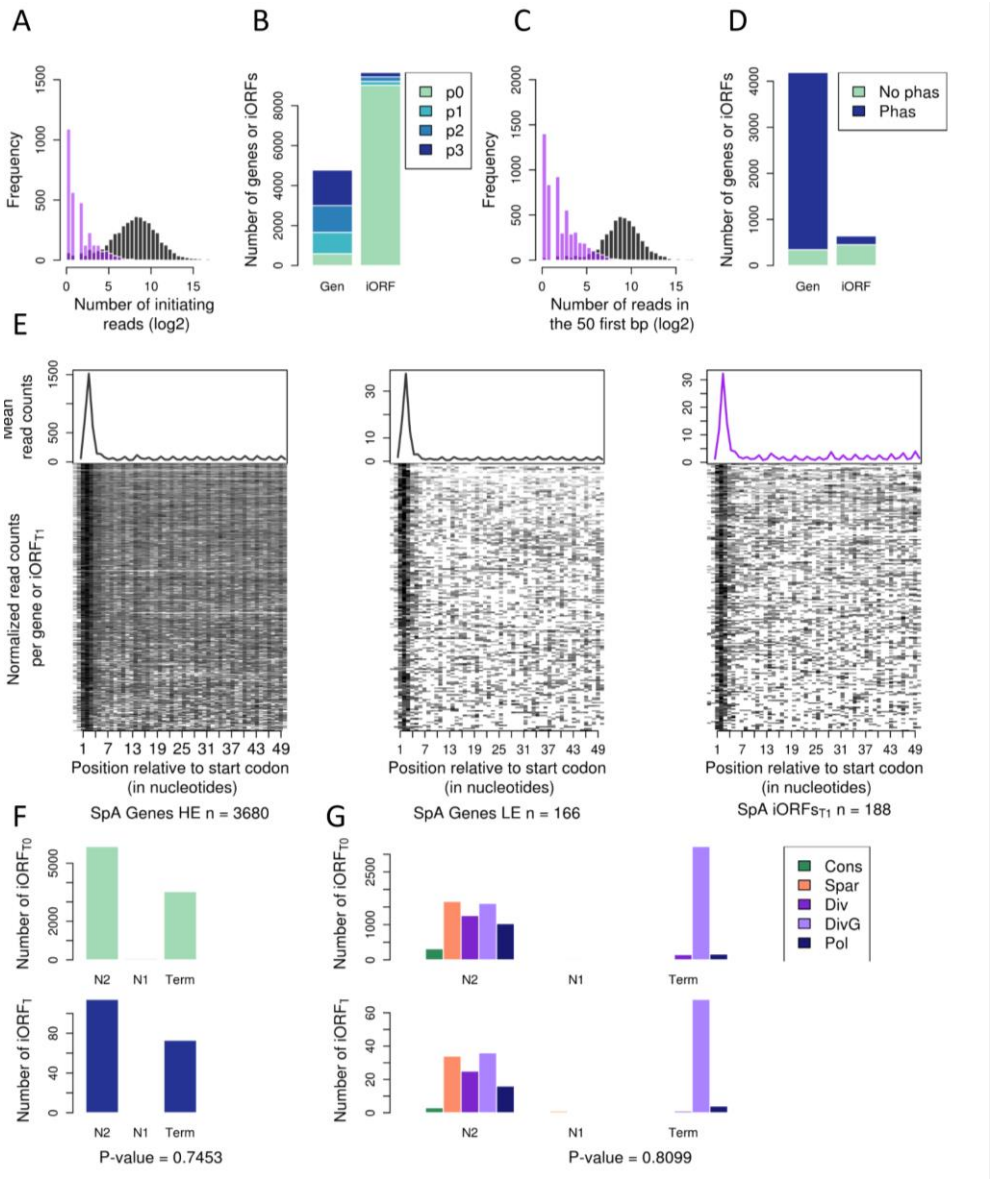

Figure S3. Detection of iORF translation signatures. Results for *SpA* and *SpB* strains (see Fig. 3 for *SpC* results). **A)** Distribution of ribosome profiling read counts for genes (in grey) and iORFs (in purple) at the start codon position. **B)** Proportions of genes (Gen) or iORFs with a detected 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 50 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)** Illustration of metagene analysis results for significantly translated highly (HE) or lowly (LE) expressed genes in grey, and intergenic iORFs$_{T1}$ in purple. The mean of 5' read counts is plotted

along the position relative to the start codon for significantly translated genes or iORFs$_{T1}$. Lines of the matrix indicate the normalize coverage of all genes or iORFs$_{T1}$ with significant signature of translation. **F-G)** Number of iORFs$_{T0}$ and iORFs$_{T1}$ depending on their age **(F)** or conservation group per age **(G)**. P-values above are for homogeneity chi-square test to compare the proportions of each age or conservation category between iORFs$_{T0}$ and iORFs$_{T1}$.

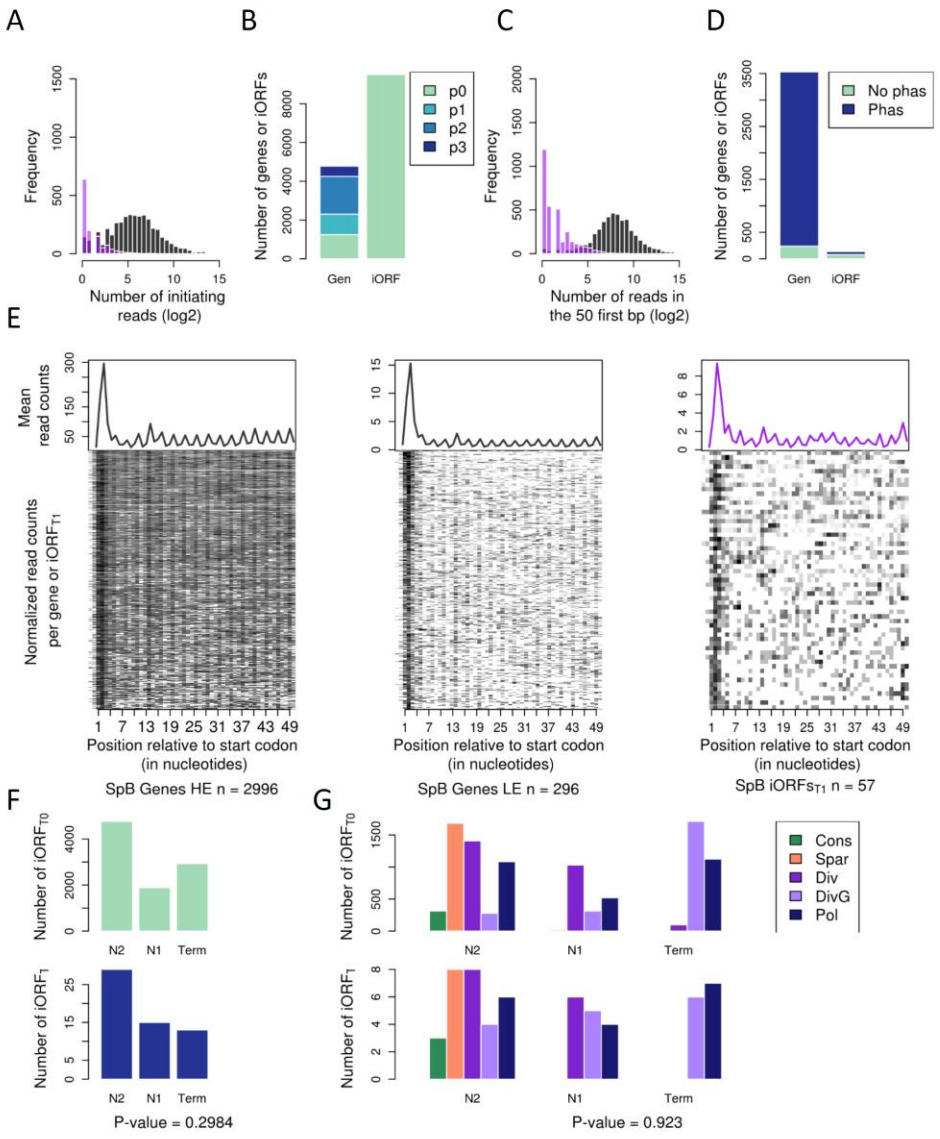

Figure S3. Continued.

Table 3. Detection of translated genes or iORFs

Strain	Genes peak	Genes phasing[1]	iORFs peak	iORFs phasing[1]
YPS128 (*S. cer*)	4,095 (85.7%)	3,874 (94.6%)	83 (6.9%)	29 (34.9%)
YPS744 (*SpA*)	4,190 (87.7%)	3,846 (91.8%)	643 (6.7%)	188 (29.4%)
MSH-604 (*SpB*)	3,531 (73.9%)	3,287(93.1%)	139 (1.4%)	57 (41.0%)
MSH-587-1 (*SpC*)	4,203 (87.9%)	3,985 (94.8%)	472 (4.9%)	190 (40.5%)
Total (without redundancy if shared between strains)	4,573	4,443	1,151	418

[1] Number of iORFs or genes with a significant trinucleotide periodicity in ribosome profiling data among those with an initiation peak

We measured codon periodicity, which is illustrated by an enrichment of reads at the first nucleotide of each codon in the first 50 nt excluding the start codon. As for the start codon region, the number of ribosome profiling reads is lower for iORFs compared to known genes (Fig. 3C). Among the features with a detected initiation peak, 91.8 to 94.8% of genes and 29.4 to 41 % of iORFs show a significant codon periodicity per haplotype (Table 3 and Fig. 3D). The number of detected translation signal is lower in strain MSH-604, which is most likely due to a lower number of reads obtained for this strain and the use of raw read density in this analysis (see Methods). iORFs with an initiation peak and a significant periodicity in at least one strain were considered as significantly translated and labeled iORFs$_{T1}$ whereas iORFs with no significant translation signatures were labeled iORFs$_{T0}$. We performed a metagene analysis on annotated genes and iORFs$_{T1}$, which revealed a similar ribosome profiling read density pattern between low expressed genes and iORFs$_{T1}$, and confirmed a distinct codon periodicity with significant translation signature for iORFs$_{T1}$ (Fig. 3E and S3). The resulting iORFs$_{T1}$ set contains 418 iORF orthogroups with size ranging from 60 to 369 nucleotides. They represent a small fraction (2.12 %) of the 19,689 iORF orthogroups longer than 60 nt. This percentage could be a conservative estimate because the detection depends on the chosen methods and filters and on the ribosome profiling sequencing depth. Also, some iORFs may be expressed under other environmental conditions. Overall, for a

genome of about 5,000 genes, the roughly 400 *de novo* iORFs that show significant translation signatures and which may produce *de novo* polypeptides, could be an important contribution to the proteome diversity of these natural populations.

Translation does not affect intergenic ORFs retention

We looked for an association between translation and iORFs retention, which could a sign that *de novo* polypeptides encoded by $iORFs_{T1}$ contribute to a fitness increase (or decrease) and therefore have beneficial (or deleterious) biochemical activities. We compared the numbers of $iORFs_{T1}$ and $iORFs_{T0}$ with respect to their age and conservation. We observed a similar conservation distribution for $iORFs_{T1}$ and $iORFs_{T0}$ per age category (NS Chi-square test, Fig. 3F-G and S3F-G). This observation suggests that iORFs that become translated are not preferentially conserved (or eliminated) than supposedly neutral $iORFs_{T0}$, suggesting overall weak or no selection acting on them.

We compared $iORFs_{T1}$ and $iORFs_{T0}$ size distributions, as well as the distribution of their size changes relative to their ancestors, to examine if translation may influence iORFs size evolution. More generally, $iORFs_{T1}$ tend to be smaller compared to $iORFs_{T0}$ of the same age, especially for those present at N2 and on terminal branches (Fig. S4C). Note that the absence of effect at N1 may be attributed to a low detection power due to less $iORFs_{T1}$ detected at N1 in our dataset (Table 2). Translation does not influence the distribution of iORF size changes, which are on average similar for $iORFs_{T0}$ and $iORFs_{T1}$ (NS T-test, Fig. S4B). In addition, longer $iORFs_{T1}$ tend to be less submitted to size changes compared to $iORFs_{T0}$ of the same size range (Chi-square test p-value = 0.005, Fig. S4A). By comparison with the fitness effect distribution of new mutations, characterized by a large number of mutations of neutral or small effects and few mutations of large effect (Bataillon and Bailey 2014), we hypothesized that only a small fraction of $iORFs_{T1}$ size

changes may be of strong effects and could influence the retention pattern compared to most nearly neutral iORFs$_{T0}$.

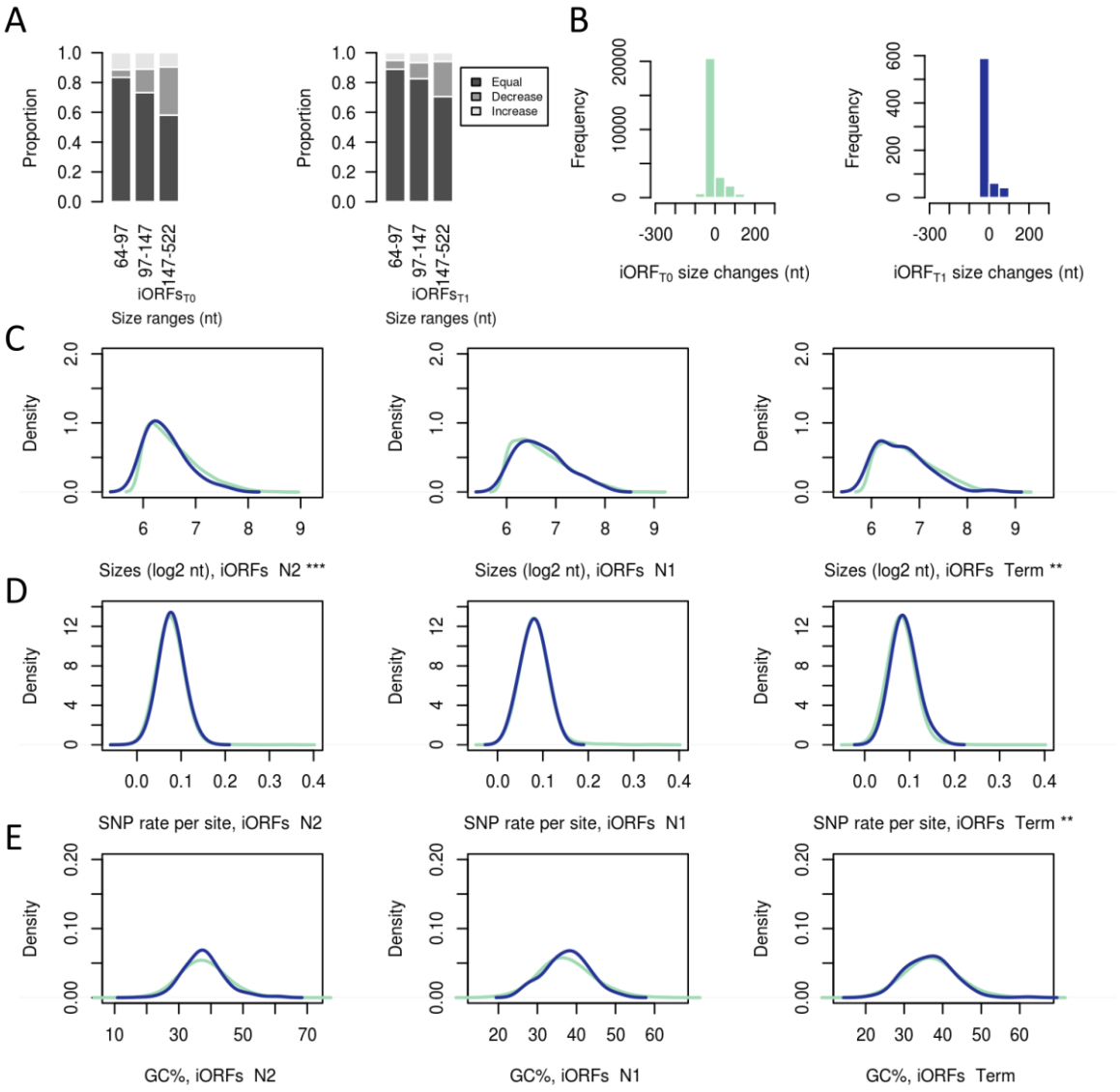

Figure S4. Comparison of iORFs$_{T1}$ and iORFs$_{T0}$ for diverse sequence properties. A) Proportion of terminal iORFs$_{T0}$ or iORFs$_{T1}$ that changed in size (increase or decrease) or conserved per range size relative to their ancestor at N1. **B)** Size change distribution for iORFs$_{T0}$ in green and iORFs$_{T1}$ in blue relative to ancestors. **C-E)** Density distributions for iORF sizes (in nucleotides and log2) **(C)**, SNP density per site in *S. paradoxus* **(D)** and GC % **(E)** per age and between iORFs$_{T0}$ in green and iORFs$_{T1}$ in blue. Significant differences between iORFs$_{T0}$ and iORFs$_{T1}$ per age are indicated above each plot (Wilcoxon test, *** for p-values < 0.001, ** for p-values < 0.01 and * for p-values <0.05). The comparison of iORFs$_{T1}$ general properties with iORFs$_{T0}$ was done on pooled iORFs from the four sequenced strains to increase the number of translated iORFs$_{T1}$.

We assumed that most polymorphisms located in intergenic regions are neutral so we used the polymorphic sites proportion for each syntenic intergenic regions as a measure of the SNP density per genomic intergenic region (see Methods). Recent $iORFs_{T1}$ appearing along terminal branches are located in genomic regions with more polymorphisms compared to $iORFs_{T0}$ (Fig. S4D), suggesting that recent translated iORFs are more likely to occur in regions with higher substitution rates. We tested for an effect of the GC% in the repartition of $iORFs_{T1}$ in the genome. $iORFs_{T1}$ are not preferentially located in GC-rich regions than $ORFs_{T0}$ (Fig. S4E). We removed sequences of low complexity in our filtering methods, so this may biased the average GC content in our data.

Some intergenic translated ORFs display strong expression changes between lineages

$iORFs_{T1}$ came from ancient and recent iORFs gains, showing a regular supply of *de novo* putative polypeptides in intergenic regions (Table 2). We looked for lineage-specific emerging putative polypeptides, among $iORFs_{T1}$, based on significant differences of ribosome profiling coverage between each pair of haplotypes. Note that a translation gain or increase may be due to an iORF gain, or to a transcription/translation increase or both. 33 $iORFs_{T1}$ display a significant lineage-specific expression increase, with 20, 5 and 8 $iORFs_{T1}$ in *SpA*, *SpB* and *SpC* respectively (Fig. 4 and S5). Among them, 24 are accompanied by a lineage-specific presence for the considered $ORFs_{T1}$ within which, 16 were acquired along terminal branches, like the *SpB*-specific iORF_70680 (Fig. 4). Nearly 70 % of strong lineage-specific expression pattern are correlated with the presence of the $iORF_{T1}$ in only one lineage, suggesting that iORF turnover mostly explain translation differences compared to a lineage expression increase in a region already containing a conserved $iORF_{T1}$ for instance. Three $iORFs_{T1}$ are also more expressed in both *SpB* and *SpC* strains compared to *SpA* and *Scer* suggesting an event occurring along branch b2 (Fig. 1C, 4 and S5). We also detected older expression gain/increase events in *S. paradoxus*, specific relative to *S. cerevisiae*, for 9 $iORFs_{T1}$, for instance iORF_69174 (Fig. 4 and S5). This result shows that

ancient iORFs$_{T1}$ may also be conserved over longer evolutionary time-scales, potentially under the action of selection, although there is no evidence for a role of selection.

We observed specific translation patterns resulting from iORFs gains and/or expression increases at different times along the phylogeny. The resulting set of emerging polypeptides of different ages provides key material to examine the properties of *de novo* polypeptides at the onset of gene birth.

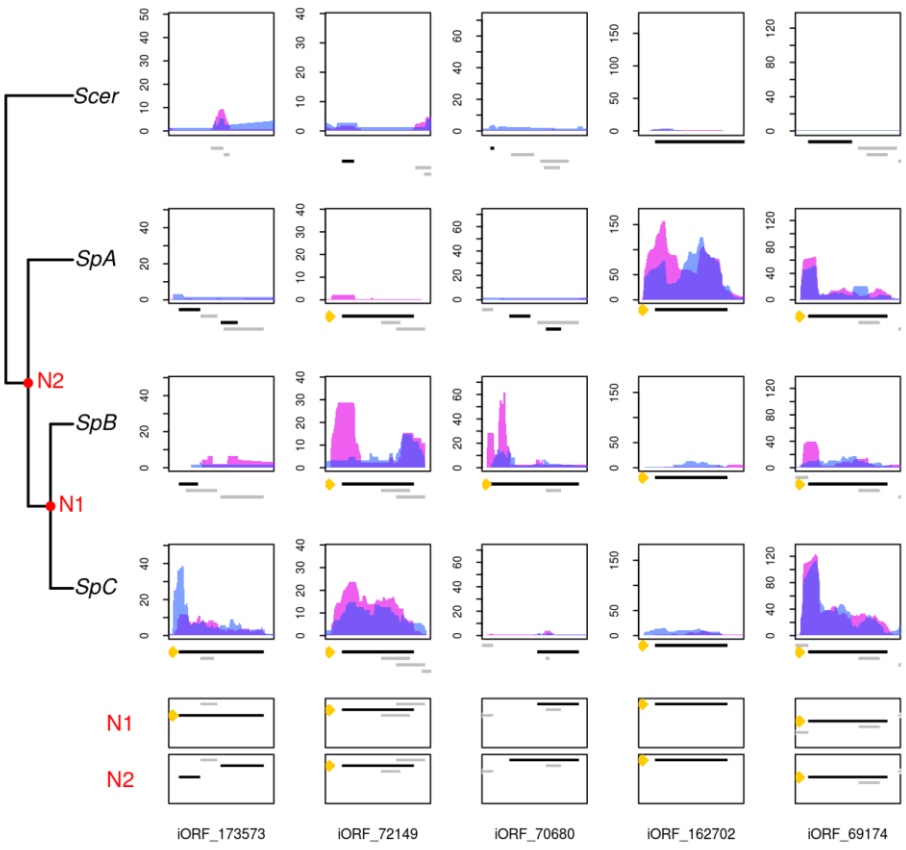

Figure 4. A continuous emergence of putative polypeptides in *S. paradoxus*. Normalized RPF read coverage for a selection of lineage specific (or group specific) iORFs$_{T1}$ per haplotype. Each replicate is displayed with blue and pink area. The positions of all iORFs (including iORFs$_{T0}$) in the genomic area are drown above each plot. The considered iORF$_{T1}$ is tagged by a yellow dot and plotted in black. Overlapping other iORFs are plotted in black when they are in the same reading frame as the selected iORF$_{T1}$, or in grey if different.

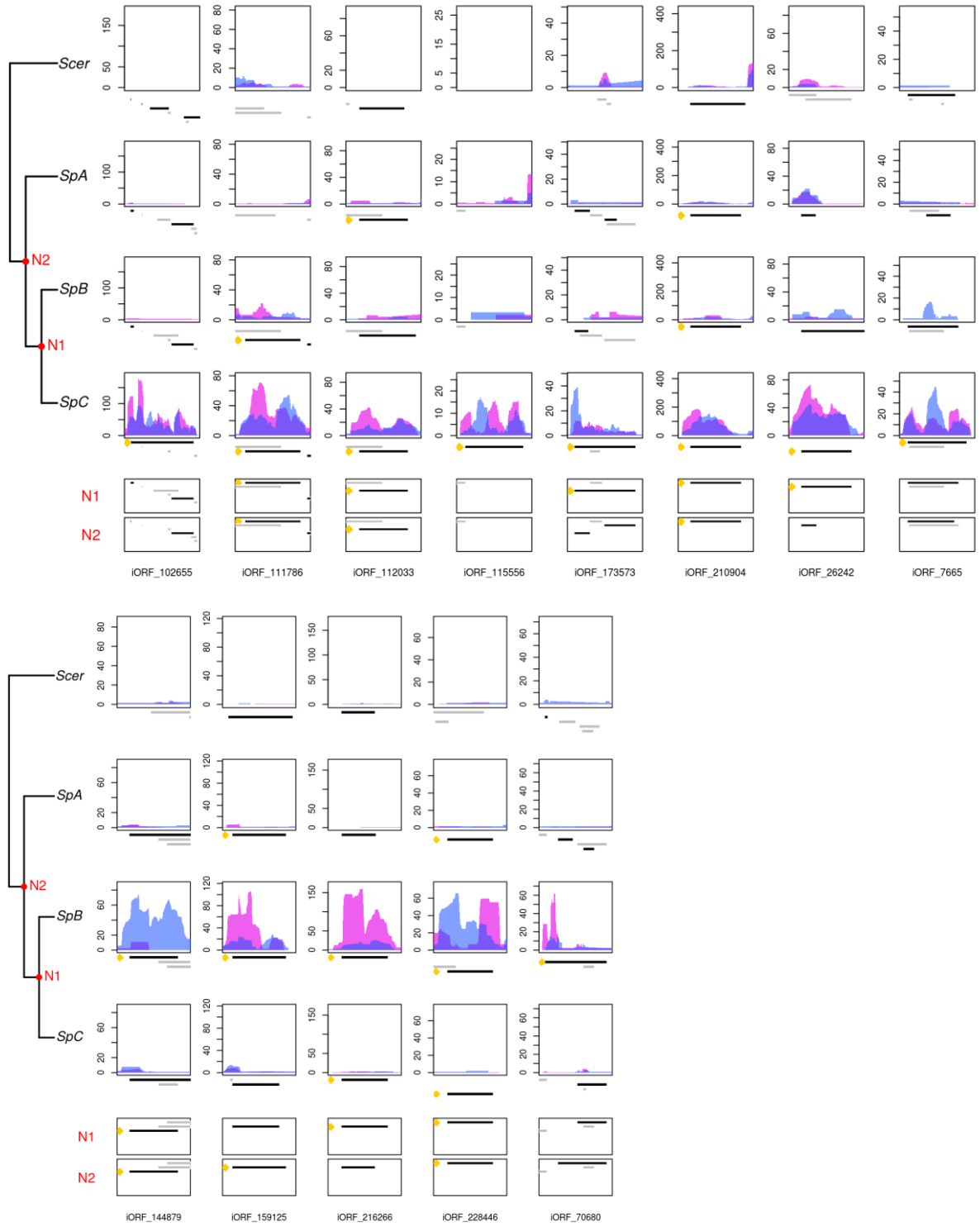

Figure S5. Normalized RPF read coverage for lineage specific (or group specific) iORFs$_{T1}$. Replicates are individually shown with blue and pink area. The positions of all iORFs (including iORFs$_{T0}$) in the genomic area are drown above each plot. The iORF$_{T1}$ shown is tagged with a yellow dot and plotted in black. Overlapping other iORFs are plotted in black when they are in the same reading frame as the selected iORF$_{T1}$, or in grey if different.

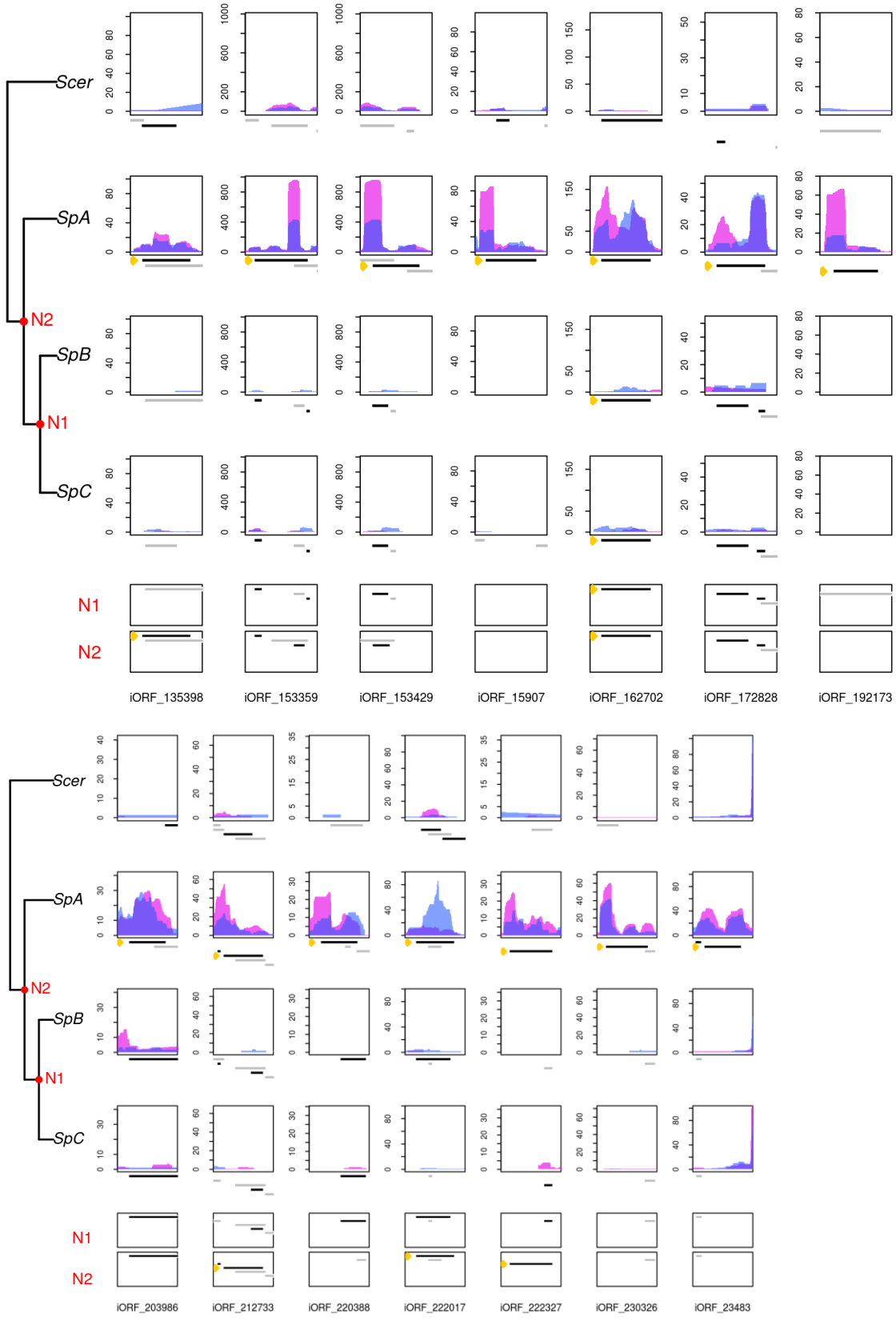

Figure S5. Continued.

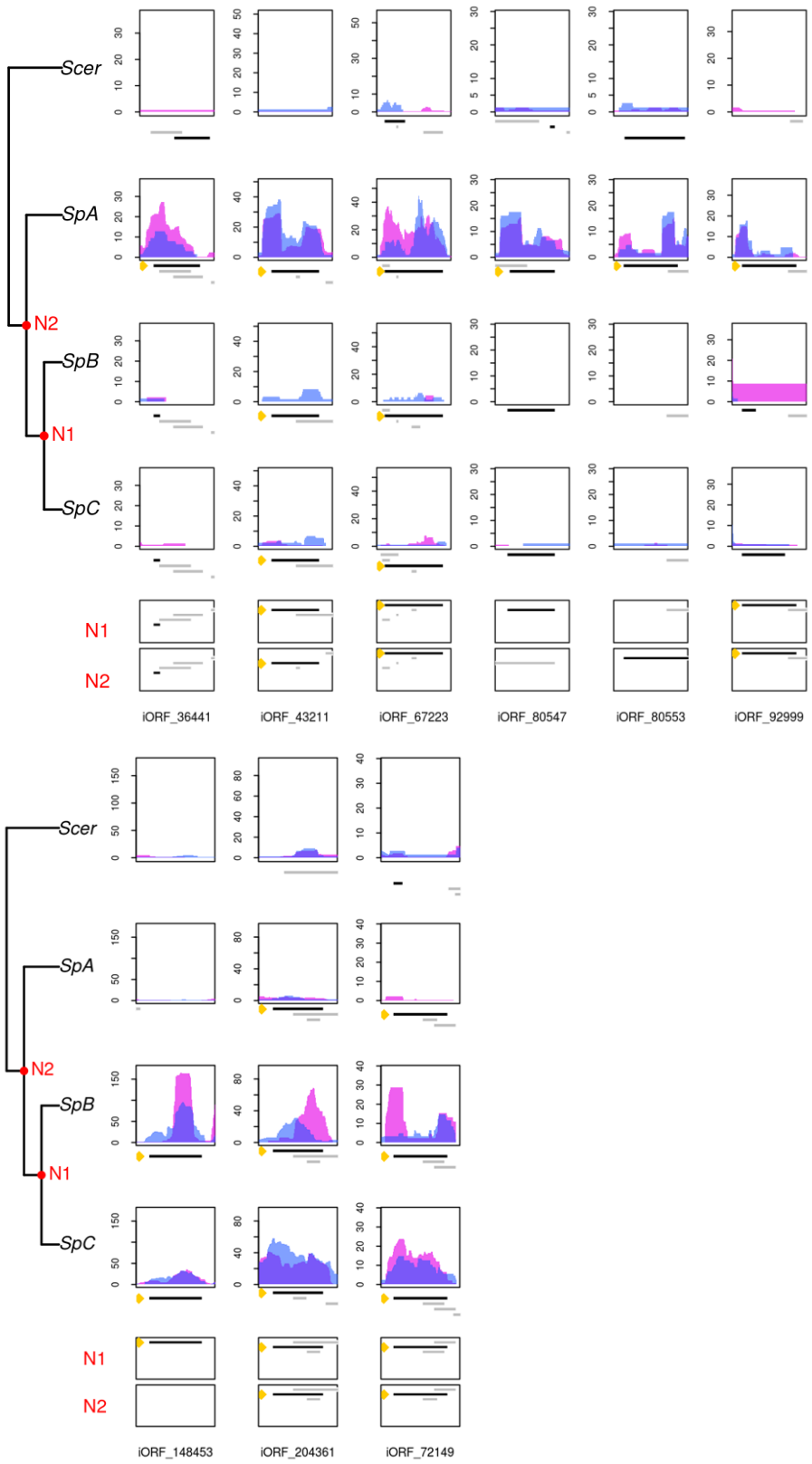

Figure S5. Continued.

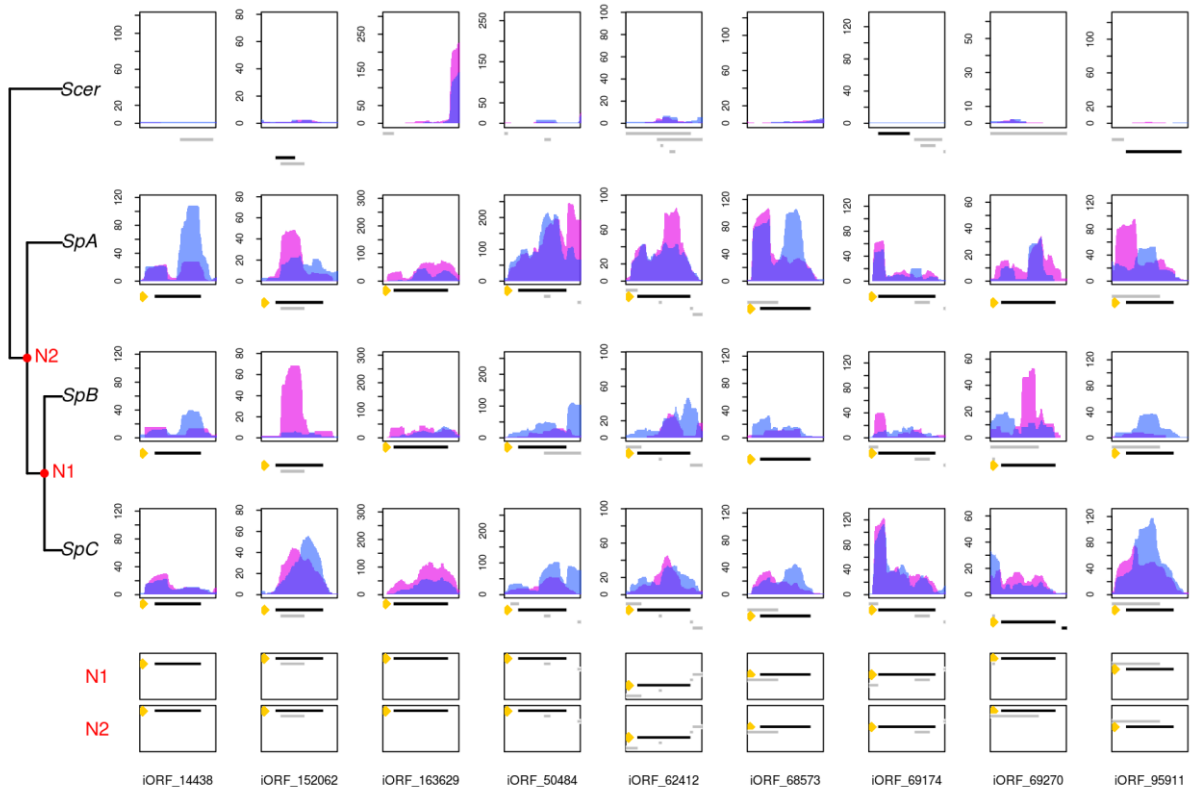

Figure S5. Continued.

Translational buffering acts on intergenic ORFs

We compared the expression of ancient and recent iORFs$_{T1}$ with the one of known genes to examine if *de novo* polypeptides display gene-like expression levels. We looked at the translational and transcriptional levels using ribosome profiling and total RNA sequencing libraries. We estimated translation efficiency (TE) per gene and iORFs$_{T1}$ as the ratio of ribosome profiling reads (named RPFs for ribosome profiling footprints) over total mRNA. This ratio (in log2) is positive when the number of translating ribosomes increases per molecule of mRNA, illustrating a more effective translation per mRNA unit (Ingolia et al. 2009). Note that RPF and total RNA coverages were calculated on the first 60 nt for genes and iORFs$_{T1}$ to reduce the bias introduced by the higher number of reads at the initiation codon, which tends to increase TEs in short iORFs$_{T1}$

compared to longer genes. After this correction, TEs remains significantly correlated with gene size but the effect is small and should not interfere in our analysis (Fig. 5E).

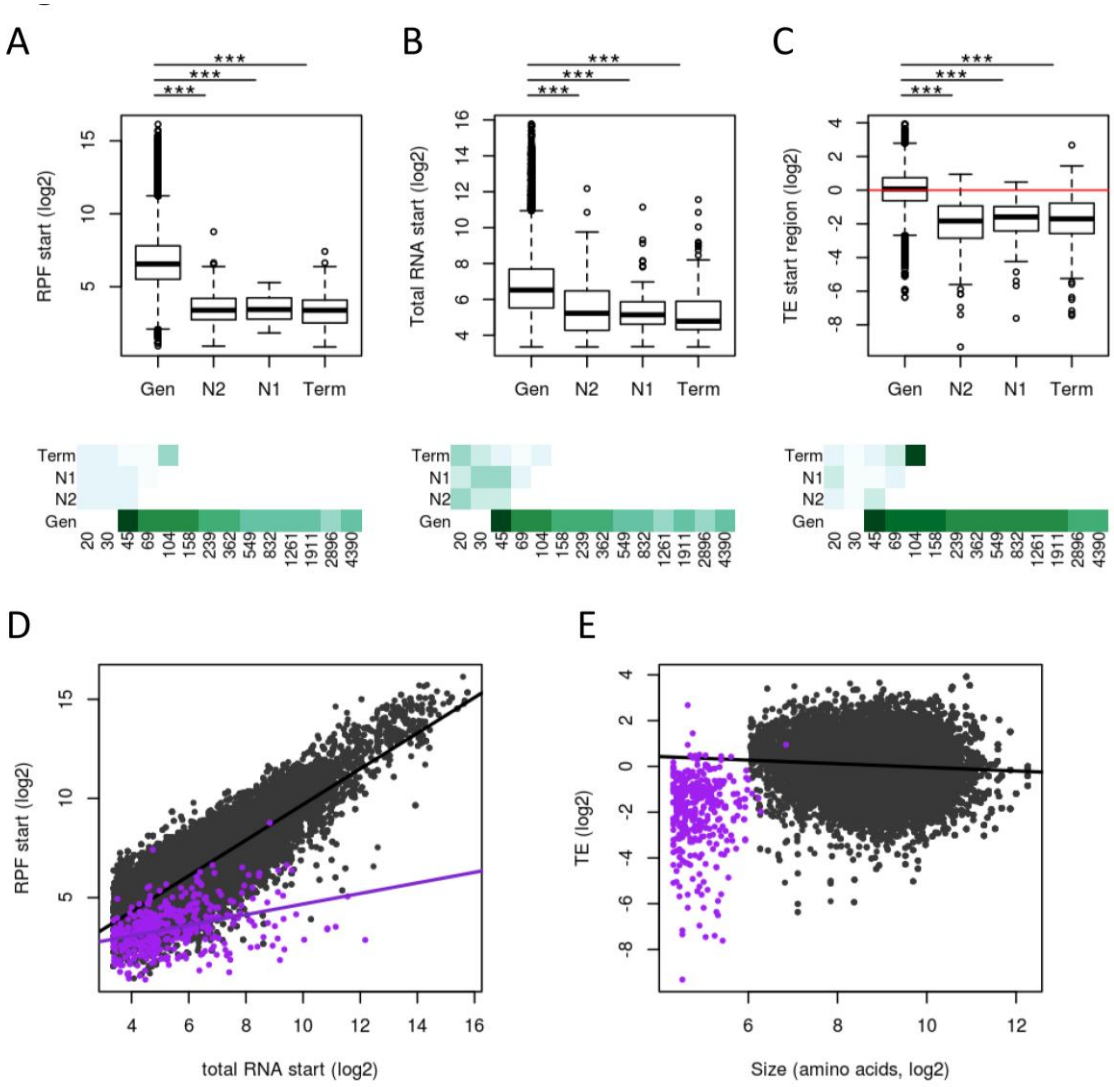

Figure 5. Putative intergenic polypeptides are less effectively translated compared to genes. A-C) Ribosome profiling (RPF), total RNA and translation efficiency (TE) - read counts in the first 60 nt, normalized to correct for library size differences in log2 - are displayed for genes (Gen) and iORFs$_{T1}$ depending on their ages (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 by green intensities (from pale for low values to dark green high values) below. **D)** RPF plotted as a function of total RNA for iORFs$_{T1}$ in purple, or genes in grey**. E)** TE plotted as a function of iORFs$_{T1}$ or gene sizes (number of residues in log2). Regression lines are plotted for significant Spearman correlations (p-values < 0.05). Expression levels were calculated using the mean of the two replicates.

As expected for intergenic regions, iORFs$_{T1}$ were less transcribed and translated than genes (T-test, both p-values < $2.2x10^{-16}$, Fig. 5A-B). We also observed a significant lower TE on average (T-test, p-value = $4.8x10^{-9}$, Fig. 5C) for iORFs$_{T1}$ compared to genes, suggesting that young iORFs$_{T1}$ are less actively transcribed and translated than genes of the same size, excepted for longer iORFs$_{T1}$ appearing on terminal branches which display higher TE levels. More generally, the most transcribed iORFs$_{T1}$ display a more reduced TE compared to genes (Fig. 5D, ANCOVA, p-value < 2.2x10-16). The consequence of this buffering effect acting at the post-transcriptional level is a reduction of polypeptides translated per molecule of mRNA. The buffering of highly transcribed iORFs$_{T1}$ may be due to a rapid selection to reduce the production of toxic polypeptides or may simply be a mechanistic consequence of recent transcription increase without translation optimization. The buffering effect is similar among iORFs$_{T1}$ of different ages, with no significant pairwise differences between buffering slopes (data not shown), which support the mechanistic consequence hypothesis. We also noted a significant overlap between expression levels and TEs in intergenic genes and genes, which means that some iORFs$_{T1}$ have gene-like expression levels.

Translated intergenic polypeptides display a high variability for gene-like traits

A recent study suggested that selection favors pre-adapted *de novo* young genes with a high level of protein disorder (ISD) compared to old genes, whereas random polypeptides in intergenic regions are one average less disordered (Wilson et al. 2017). This would suggest that young polypeptides with an adaptive potential would already be biased in terms of protein structural properties compared to the neutral expectations based on random sequences. We examined the properties of polypeptides as a function of timing of emergence in order to follow their evolution during the time before, or at the early beginning of, the action of selection. We compared the level of intrinsic disorder, GC-content and genetic diversity (based on SNPs density) in iORFs$_{T1}$ as a function of age with that of annotated known genes. On average, protein disorder and GC-content are lower in iORFs$_{T1}$ than in canonical genes regardless of iORFs$_{T1}$ ages (p-values < 0.001, T-

test, Fig. 6B-C). This pattern was confirmed for iORFs$_{T1}$ and genes sharing the same size range of between 45 and around 100 amino acids (Fig. 6B-C). The lower intrinsic disorder for iORFs$_{T1}$ was also observed for random intergenic sequences in Wilson's study (Wilson et al. 2017). However, we observed a subset of iORFs$_{T1}$ with extreme gene-like disorder values that could refer to the subset of non-functional peptides expected to be recruited by natural selection if gene-like characteristics increase their functional potential. iORFs$_{T1}$ are located in more divergent regions compared to genes, which is in agreement with stronger purifying selection on canonical genes (Fig. 6D). We examined if SNP density variation along the genome may influence the iORFs$_{T1}$ turnover. Younger iORFs$_{T1}$, appearing along terminal branches, tend to be in more divergent regions compared to older ones at N2, even when considering the same size ranges (Fig. 6D). This may be due to mutation rate variation or differences in evolutionary constrains acting on iORFs$_{T1}$ age subsets. Older iORFs$_{T1}$ are not preferentially located at the proximity of genes where selection may be stronger (Fig. 6G), suggesting that the lower diversity observed at N2 is mainly due to a lower mutation rate. A correlation between mutation rate variation and replication timing differences along chromosomes has already been observed in yeast, where origins of replication (ARS) activated late show higher mutation rate compared to earlier ones (Lang and Murray 2011; Agier and Fischer 2012). We compared the replication timing in regions of iORFs$_{T1}$ of different ages to examine if the higher diversity observed in younger iORFs$_{T1}$ on average is correlated with late replicating regions. We used estimates from Muller et al. (2014) which are based on the quantification of the amount of DNA during replication by deep-sequencing, which are higher in genomic regions of early replication compared to regions of late replication. We did not see differences for replication timing between genes and iORFs$_{T1}$, neither between iORFs$_{T1}$ ages categories (Fig. 6E). However, we observed that older iORFs$_{T1}$ tend to be closer to replication origins compared to younger iORFs$_{T1}$ (Fig. 6F), which is consistent with the higher genetic diversity observed in recently emerging iORFs$_{T1}$ locations. These observations suggest that younger iORFs$_{T1}$ are more likely to occur in rapidly evolving sequences with higher mutation rates.

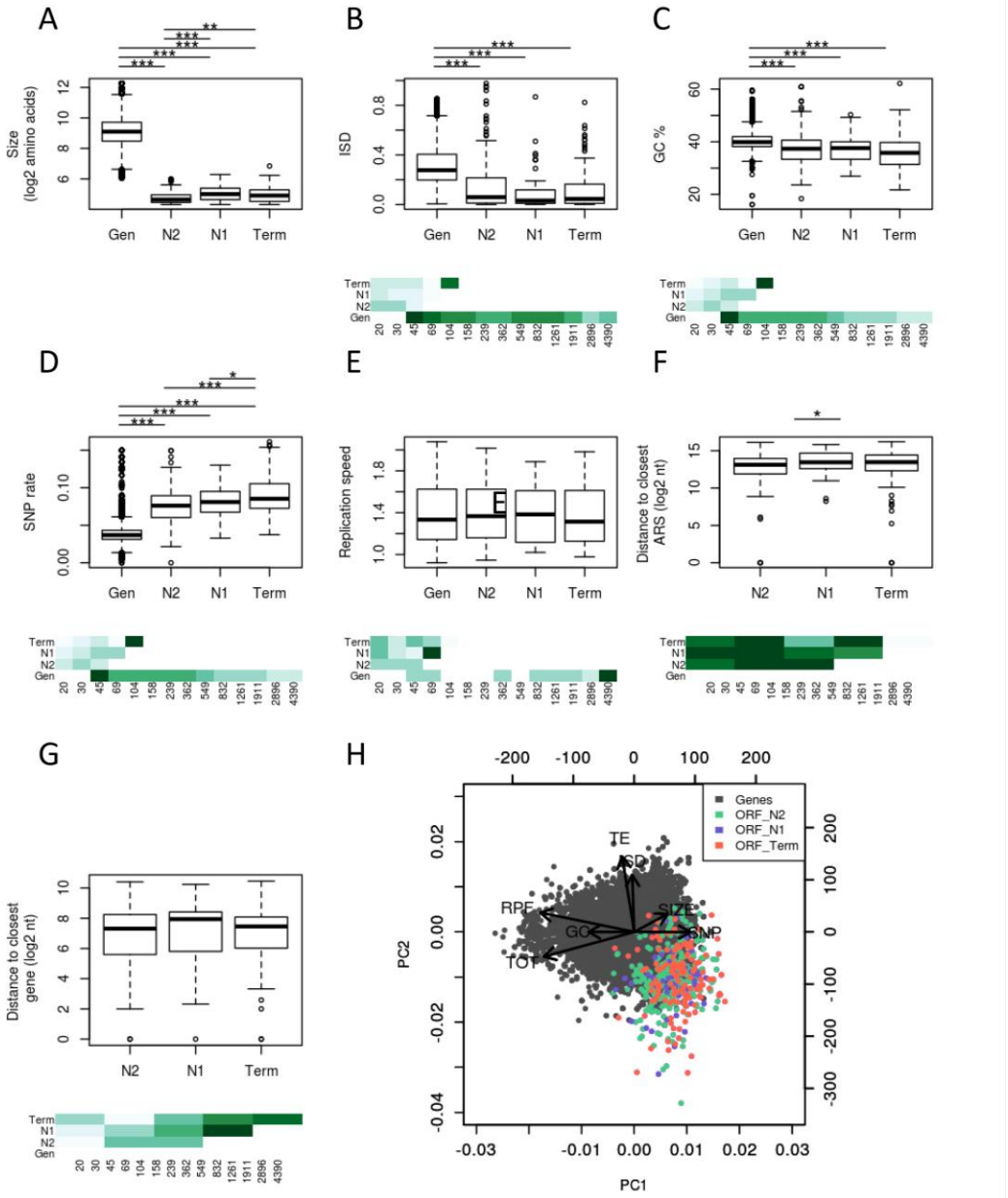

Figure 6. Age-dependent characteristics of intergenic polypeptides. A-G) Sizes (log2 number of residues), mean disorder (ISD), GC %, SNP density, replication speed (based on data from Muller et al. (2014)), and distance to the closest replication origin (ARS) or gene are displayed for genes and iORFs$_{T1}$ as a function of their ages (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 with green intensities (from pale for low values to dark green high values) below. **H)** Principal component analysis using the number of residues (SIZE in log2), 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 log2), intrinsic disorder (ISD), the GC% and SNP density (SNP). iORF$_{T1}$ are colored as a function of their ages. The two first axis explain 32 and 18 % of the variation (total 50 %).

iORFs$_{T1}$ are located in more divergent regions compared to genes, which is in agreement with stronger purifying selection on canonical genes (Fig. 6D). We examined if SNP density variation along the genome may influence the iORFs$_{T1}$ turnover. Younger iORFs$_{T1}$, appearing along terminal branches, tend to be in more divergent regions compared to older ones at N2, even when considering the same size ranges (Fig. 6D). This may be due to mutation rate variation or differences in evolutionary constrains acting on iORFs$_{T1}$ age subsets. Older iORFs$_{T1}$ are not preferentially located at the proximity of genes where selection may be stronger (Fig. 6G), suggesting that the lower diversity observed at N2 is mainly due to a lower mutation rate. A correlation between mutation rate variation and replication timing differences along chromosomes has already been observed in yeast, where origins of replication (ARS) activated late show higher mutation rate compared to earlier ones (Lang and Murray 2011; Agier and Fischer 2012). We compared the replication timing in regions of iORFs$_{T1}$ of different ages to examine if the higher diversity observed in younger iORFs$_{T1}$ on average is correlated with late replicating regions. We used estimates from Muller et al. (2014) which are based on the quantification of the amount of DNA during replication by deep-sequencing, which are higher in genomic regions of early replication compared to regions of late replication. We did not see differences for replication timing between genes and iORFs$_{T1}$, neither between iORFs$_{T1}$ ages categories (Fig. 6E). However, we observed that older iORFs$_{T1}$ tend to be closer to replication origins compared to younger iORFs$_{T1}$ (Fig. 6F), which is consistent with the higher genetic diversity observed in recently emerging iORFs$_{T1}$ locations. These observations suggest that younger iORFs$_{T1}$ are more likely to occur in rapidly evolving sequences with higher mutation rates.

Because sequences are too similar between strains to test for purifying selection individually on each iORFs$_{T1}$, we estimated the likelihood of the global dN/dS ratio for two merged set of iORFs$_{T1}$, containing ancient iORFs$_{T1}$ conserved in all *S. paradoxus* strains (set 1) or iORFs$_{T1}$ appearing at

N1 and conserved between the *SpB* and *SpC* lineages (set 2). Both sets seem to evolve neutrally with no significant purifying selection acting (NS p-values). These results illustrate the continuous emergence of random polypeptides that do not appear to be under significant selection.

The variability observed for expression levels, genetic and structural properties revealed a subset of *de novo* polypeptides with gene-like characteristics. We performed a multivariate analysis to look for polypeptides with extreme values for multiple traits as an indicator of their functional potential. We observed a subset of $iORFs_{T1}$ sharing all considered characteristics with genes and resulting from ancient or recently gained $iORFs_{T1}$ (Fig. 6H). Although iORFs do not appear to be under significant purifying selection, as a neutral pool they provide raw material for selection to act under either the continuum or pre-adaptation models, revealing a rapid potential for molecular innovations.

Discussion

To better understand the early stages of *de novo* gene birth, we characterized the properties and turnover of recently evolving iORFs and their putative peptides over short evolutionary time-scales using closely related wild yeast populations. The number of iORFs identified almost doubles when considering within species diversity, which illustrates the role of intergenic diversity to provide potential molecular innovation. The iORFs presence/absence diversity comes from ancient iORFs that are still segregating within *S. paradoxus* and from a continuous supply of *de novo* iORFs. The turnover and retention of iORFs appear to be mostly guided by mutation rate variation affecting the number of gains and losses, or by size changes with some larger changes, more likely to occur in longer iORFs because of the longer mutational target between start and stop codons. The iORF turnover rate is lower than the rate of gene duplication or loss estimated in yeast (without whole genome duplication, (Lynch et al. 2008)).

Among the ~20,000 iORF orthogroups of 60 nt and longer, only a small fraction (about ~2%, n= 418) shows translation signatures similar to expressed canonical genes. We observed a stronger post-translational buffering in the most transcribed iORFs, reflecting either selection against translation or lack of selection for optimal translation. This mechanism was also observed in interspecies yeast hybrids, especially for genes with transcriptional divergence and was hypothesized to be a result of stabilizing selection on the amount of proteins produced (McManus et al. 2014). The post-translation buffering effect is similar between older and younger iORFs, suggesting a lack of translation optimization which attenuates the amount of *de novo* polypeptides relative to mRNA molecules rather than selection.

Consistent with a model in which most iORFs are neutral, the corresponding *de novo* polypeptides properties are on average close to expectation for random sequences with some having gene-like properties, suggesting a small set of neutrally evolving polypeptides with a potential for molecular innovations. The conservation distribution of iORFs with translation signature ($iORFs_{T1}$) is similar that of non-translated ones, suggesting that iORF retention is mainly guided by random mutations and genetic drift even when translated. The absence of selection signature is also consistent with the neutral evolution of most of intergenic polypeptides observed in rodents (Ruiz-Orera et al. 2018), and with the weak effect of purifying selection acting on younger *de novo* genes in yeast, *Drosophila* and *Arabidopsis* (Carvunis et al. 2012; Palmieri et al. 2014; Zhao et al. 2014; Li et al. 2016; Vakirlis et al. 2017). The resemblance to random sequences does not entirely preclude any potential molecular function and effect on fitness however because a recent study showed that a unneglectable fraction of expressed random sequences confers a positive effect on the fitness (Neme et al. 2017).

Recently emerging $iORF_{T1}$ along terminal branches are more frequent in regions with a higher SNP density, whereas older $iORFs_{T1}$ tend to be located in slowly evolving regions. This

observation suggests variable turnover rates depending of the local mutation rate. Regions with low mutation rates could act as a reservoir of ancient iORFs segregating in population for a longer time before being lost. On the other hand, mutation hotspots may allow to rapidly test many molecular combinations immediately available, which could be advantageous in a changing environment. A small fraction of translated ORFs that recently appears have several gene-like characteristics, suggesting that they are pre-adapted to be biochemically functional, while most have some characteristics but not others, meaning that they would require refinement by natural selection to acquire these traits. These observations could reconcile the two opposing models of *de novo* gene birth (Carvunis et al. 2012; Wilson et al. 2017). Ongoing *de novo* genes would be more likely to progressively acquire gene-like properties in slowly evolving regions (low mutation rate) before being lost, as in the continuum model. Faster evolution in some regions may increase the chance to acquire a polypeptide with an immediate functional potential as in the preadaptation hypothesis.

Material and methods

Characterization of the intergenic ORFs diversity

We investigate intergenic ORF (iORF) diversity in wild *Saccharomyces paradoxus* populations, which are structured in 3 main lineages named *SpA*, *SpB* and *SpC* (Charron et al. 2014; Leducq et al. 2016). The wild *S. cerevisiae* strain YPS128 was used in our experiments and the reference S288C (version R64-2-1) was added in our analysis for the functional annotation.

Genome assemblies

New genomes assemblies were performed using high-coverage sequencing data from 5, 10 and 9 North American strains belonging to lineages *SpA*, *SpB* and *SpC* respectively 1 (Fig. S1) (Leducq et al. 2016) using IDBA_UD (Peng et al. 2012). For strain YPS128, raw reads were kindly provided by J. Schacherer from the 1002 Yeast Genomes project (Peter et al. 2018). We used the

default option for IDBA-UD parameters: a minimum k-mer size of 20 and maximum k-mer size of 100, with 20 increments in each iteration. Scaffolds were then ordered and orientated along a reference genome using ABACAS (Assefa et al. 2009), using the –p nucmer parameter. *S. paradoxus* and *S. cerevisiae* scaffolds were respectively aligned along the reference genome of the CBS432 (Liti et al. 2009) and S288C (version R64-2-1 from the Saccharomyces Genome Database (https://www.yeastgenome.org/)) strains. Unused scaffolds in the ordering and longer than 200 pb were also conserved in the dataset for further analysis.

Identification of homologous intergenic regions

We detected homologous intergenic region using synteny. Genes were predicted using Augustus (Stanke et al. 2008) with the complete gene model for the species parameter "saccharomyces_cerevisiae_S288C". Orthologs were annotated using a reciprocal best hit (RBH) approach implemented in SynChro (Drillon et al. 2014) against the reference S288C (version R64-2-1) using a delta parameter of 3. We used RBH gene pairs provided by SynChro and the Clustering methods implemented in Silixx (Miele et al. 2011) to identify conserved orthologs among the 26 genomes. We selected orthologs conserved among all strains and with a conserved order to extract orthologous microsyntenic genomic regions ≥ 100 nt between each pair of genes (Fig. S1).

Ancestral reconstructions of intergenic sequences

We reconstructed ancestral genomic sequences of internetic regions. Because the divergence between strains belonging to the same lineage is low, we choose one strain per lineage to estimate the ancestral intergenic sequences at each divergence node between lineages (Fig. 1C and S1), that is YPS128 (*S. cerevisiae*), YPS744 (*SpA*), MSH-604 (*SpB*) and MSH-587-1 (*SpC*). The ancestral sequence reconstruction was done using Historian (Holmes 2017), which allows the reconstruction of ancestral indels in addition to nucleotide sequences. Note that indel

reconstruction is essential here to not introduce artefactual frameshifts in ancestral iORFs, see below, which depends on the conservation of the same reading frame between the start and the stop codon. Historian was run with a Jukes-Cantor model and using a phylogenetic tree inferred from aligned intergenic sequences by PhyML version 3.0 (Guindon et al. 2010) with the Smart Model Selection (Lefort et al. 2017) and YPS128 as outgroup.

iORF annotation and conservation level

Orthologous regions identified between each pair of conserved genes in contemporary strains and their ancestral sequence reconstructions were aligned using Muscle (Edgar 2004) with default parameters. Intergenic regions with a global alignment of less than 50% of identity among strains (including gaps) were removed. We annotated iORFs defined as any sequence between canonical start and stop codons, in the same reading frame and with a minimum size of 3 codons, using a custom Python script. Because we are working on homologous aligned regions, the presence-absence pattern does not suffer from limitation alignment bias occurring when we are working with short sequences. We extracted a presence/absence matrix based on the exact conservation of the start and the stop codon in the same reading frame (Fig. S1). iORF aligned coordinates were then converted to genomic coordinates on the respective genomes of each strain, and removed if there was any overlap with a known feature annotation, such as rRNA, a tRNA, a ncRNA, a snoRNA, non-conserved genes and pseudogenes annotated on the reference S288C (version R64-2-1 https://www.yeastgenome.org/). Additional masking was performed by removing iORFs i) located in a region with more than 0.6 % of sequence identity with *S. cerevisiae* ncRNA or gene (including pseudogenes and excluding dubious ORFs) from the reference genome, or *Saccharomyces kudriavzevii* and *Saccharomyces eubayanus* genes (Zerbino et al. 2018), ii) in a low complexity region identified with repeat masker (http://www.repeatmasker.org/) and iii) when local alignments of iORFs +/- 300 bp displayed less than 60% of identity (including gaps). If an

iORF overlapped a masked region detected in only one strain, it was removed for all the other strains in order to not introduce presence-absence patterns due to strain specific masking.

iORFs that do not overlap a known feature were then classified according to the conservation level: 1) conserved in both species, 2) specific and conserved within *S. paradoxus*, 3) fixed within lineages and divergent among, 4) specific and fixed in one lineage, 4) polymorphic in a least one lineage (Fig. S1).

For iORFs with a minimum size of 60 nt, we also performed a sequence similarity search against the proteome of NCBI RefSeq database (O'Leary et al. 2016) for 417 species in the reference RefSeq category and the representative fungi RefSeq category (containing 237 fungi species). iORFs with a significant hit (e-value < 10^{-3}) were removed to exclude any risks of having an ancient pseudogene. Among the 19,701 iORFs tested, only 12 displayed a significant hit, illustrating the stringency of our thresholds for the iORF annotation and filtering above.

Evolutionary history of iORFs

Gain and loss events were inferred by comparing presence/absence pattern between ancestral nodes and actual iORFs. Because the ancestral reconstruction was done using one strain per lineage (see above), polymorphic iORFs absent in all the considered strains have been removed from this analysis. iORFs with no detected ancestors were considered as appearing on terminal branches. We estimated the rate of iORF gain/substitution on each branch as the number of iORF gain/the number of substitution (*i.e* branch length × sequence size) and calculated the mean of the four branches. The iORF gain rate per cell per division was estimated by calculating the number of expected substitution per cell per division (from the substitution rate estimated at 0.33×10^{-9} per site per cell division by Lynch et al. (2008), multiplied by the iORF gain rate per substitution.

The evolution of iORFs sizes was inferred by connecting iORFs with their ancestors along the phylogeny if they shared the same start and/or stop position on aligned intergenic sequences. iORF sizes of two connected iORFs may be conserved if there are no changes, an increase or a decrease if there are connected only by the same start or stop position because the position of the other extremity of the iORFs changed.

Ribosome profiling and mRNA sequencing libraries

Ribosome profiling and mRNA sequencing experiments were conducted with the strains YPS128 (*S. cerevisiae)* (Sniegowski et al. 2002) and YPS744 (*S. paradoxus*), MSH604 (*S. paradoxus*) and MSH587 (*S. paradoxus*) belonging respectively to groups *SpA*, *SpB* and *SpC* according to Leducq et al. (2016). We prepared two replicates per strain and library type. The protocol is described in supplementary methods. Briefly, strains were grown in SOE (Synthetic Oak Exudate) medium (Murphy et al. 2006). Ribosome profiling footprints were purified using the protocol described in Baudin-Baillieu et al. (2016) with modifications (see supplementary methods). The rRNA was depleted in purified ribosome footprints and total mRNA samples using the Ribo-Zero Gold rRNA Removal Kit for yeast (Illumina) according to the manufacturer's instructions. Ribosome profiling and total mRNA libraries were constructed using the TruSeq Ribo Profile kit for yeast (illumina), using manufacturer's instructions starting from fragmentation and end repair step. Libraries were sequenced on Illumina HiSeq 2500 at The Genome Quebec Innovation Center (Montreal, Canada).

Detection of translated iORFs

Both total and ribosome profiling samples were processed using the same procedure. Raw sequences were trimmed of 3' adapters using CUTADAPT (Martin 2011). For RPF data, reads with lengths of 27–33 nucleotides were retained for further analysis as this size is most likely to represent footprinted fragments. For mRNA, reads with lengths of 27–40 nucleotides were

retained. Adapter trimmed reads were aligned to the respective genome of each sample using Bowtie version 1.1.2 (Langmead et al. 2009) with parameters –best –chunkmbs 500.

We used ribosome profiling reads to identify translated iORFs. This analysis was performed on iORFs longer than 60 nucleotides. Annotated iORFs may be overlapping because of the 3 possible reading frames for each strand. Ribosomal speed differences during translation cause an accumulation of ribosome footprints at specific positions within a gene (Ingolia 2016). We used ribosome profiling read density, which is typically characterized by a strong initiation peak located at the start codon followed by a codon periodicity at each codon, to detect the translated iORF among overlapping ones. For each strain, we performed a metagene analysis at the start codon region of iORFs and annotated conserved genes to detect the p-site offset for each read length between 28 and 33 nt. Because the ribosome profiling density pattern is stronger in highly translated regions, metagene analyses were done using the two replicates of each strain pooled in one coverage file. Ribosome footprints were mapped to their 5' ends, and the distance between the largest peak upstream of the start codon and the start codon itself is taken to be the P-site offset per read length. When comparing annotated genes and iORFs, we obtained similar P-site offset estimates per read length, which were used for next analysis. We then extracted the aligned read densities, subtracted by the P-offset estimates, per iORF or genes for next analyses. Metagene analyses were performed using the metagene, psite and get_count_vectors scripts from the Plastid package (Dunn and Weissman 2016), metagene figures were done using R script (R Core Team 2013).

We identified translation initiation signal from ribosome profiling densities, by detecting peaks at the start codon. We defined 3 precision levels of peak initiation: 'p3' if the highest peak is located at the first nucleotide of the start codon, 'p2' there is a peak at the first position of the start codon and 'p1' if there is a peak at the first position of the start codon +/- 1nucleotide because the peak

position is less precise in low expressed feature. A minimum of 5 reads was required for peak detection. Read phasing was estimated by counting the number of aligned reads at the first, second or the third position for all codons of the considered iORF or gene, to test for a significant deviations from expected ratio with no periodicity, that is 1/3 of each, with a binomial test. We applied an FDR correction for multiple testing.

iORF families or genes with an initiation peak and a significant periodicity, *i.e.* a FDR corrected p-value < 0.05, in at least one haplotype were considered as translated.

Differential expression analysis

Reads were strand-specifically mapped to $iORFs_{T1}$ and conserved genes using the coverageBed command from the bedTools package version 2.26.0 (Quinlan and Hall 2010), with parameter -s. We then examined $iORFs_{T1}$ significant expression changes between strains. Differential expression analysis was performed using DESeq2 (Love et al. 2014). Significant differences were identified using 5% FDR and 2-fold magnitude. We identified lineage specific expression increase when the expression of the $iORFs_{T1}$ in the considered lineage was significantly more expressed than the others strains in all pairwise comparisons. For *SpB-SpC* increase, we selected $iORFs_{T1}$ when *SpB* and *SpC* strains were both more expressed than YPS128 and *SpA,* and *S. paradoxus* increase when all *S. paradoxus* lineages were more expressed than YPS128.

For the visualization of $iORFs_{T1}$ coverages, we extracted the per base coverage on the same strand using the genomecov command from the bedTools package version 2.26.0 (Quinlan and Hall 2010). The normalization was performed by dividing the perbase coverage of each library with the size factors estimated with DESeq2 (Love et al. 2014).

Expression and sequence properties

Normalized read counts for ribosome profiling and total mRNA samples were extracted with DESeq2 software (Love et al. 2014) and we calculated the mean of the two replicates per library type. Translation efficiency (TE) was calculated as the ratio of RPF over total mRNA normalized read counts on the first 60 nt. We excluded iORFs$_{T1}$ and genes with less than 10 total RNA reads in the first 60 nt for the TE calculation. Slope differences between Genes and iORFs$_{T1}$ were tested with an ANCOVA. We confirmed the buffering effect on iORFs$_{T1}$ annotated on the *S. cerevisiae* reference strain S288C with ribosome profiling and RNA sequencing data obtained in (McManus et al. 2014) study (Fig. S6).

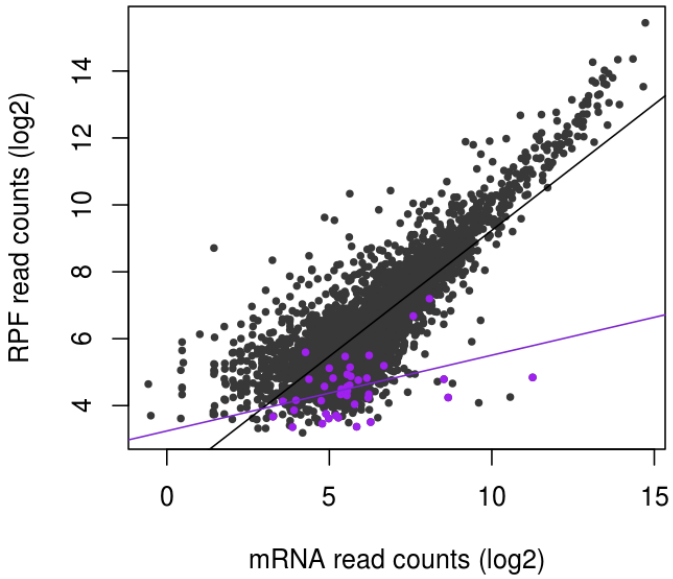

Figure S6. TE buffering in *S. cerevisiae*. Ribosome profiling (RPF) read counts plotted as a function of total RNA read counts for iORFs$_{T1}$ in purple, or genes in dark grey using ribosome profiling and mRNA sequencing from (McManus et al. 2014). Read counts were normalized to correct for library size differences. iORFs$_{T1}$ 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 iORFs$_{T1}$ 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).

The intrinsic disorder was calculated for genes and intergenic $iORFs_{T1}$ using IUPRED (Dosztanyi et al. 2005). The SNP rate was calculated for each syntenic intergenic region by dividing the total number of intergenic SNPs in *S. paradoxus* alignments, by the total number of nucleotides in the region, as in Agier and Fischer (2012) study for intergenic sequences. Replication timing data per 1kb bin comes from Muller et al. (2014) study and were converted to the version R64-2-1 of the reference genome using liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver). We used the *codeml* program from the PAML package version 4.7 (Yang 2007) to estimate the likelihood of the dN/dS ratio, using the same procedure as employed by Carvunis et al. (2012) with codon model 0.

All analysis and figures were conducted with python and R script (R Core Team 2013).

Data access

High-throughput sequencing data have been submitted to the NCBI Sequence ReadArchive (SRA; http://www.ncbi.nlm.nih.gov/sra) and can be accessed under NCBI BioProject number PRJNA400476. *De novo* assemblies and annotations have been submitted to the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide) under NCBI BioProject number PRJNA400476.

Acknowledgments

We thank A. K. Dubé, G. Charron and the IBIS sequencing platform (B. Boyle) for technical help and landry lab members for comments on the manuscript. This project was funded by a FRQNT Team grant to C.R.L and Xavier Roucou and NSERC discovery grant to C.R.L. C.R.L. holds the Canada Research Chair in Evolutionary Cell and Systems Biology.

Author contributions

E.D and C.R.L conceived the project. E.D O.N I.H and I.G.A designed ribosome profiling experiments. E.D I.G.A and I.H. performed the experiments. E.D performed bioinformatics analyses with helpful advices from L.N.T, C.R.L and O.N. E.D wrote the manuscript with revisions from all authors.

Disclosure declaration

The authors have no conflict of interest to declare.

References

Agier N, Fischer G. 2012. The mutational profile of the yeast genome is shaped by replication. *Mol Biol Evol* **29**: 905-913.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**: 1968-1969.

Baalsrud HT, Torresen OK, Hongro Solbakken M, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. 2017. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol Biol Evol* doi:10.1093/molbev/msx311.

Bataillon T, Bailey SF. 2014. Effects of new mutations on fitness: insights from models and data. *Ann N Y Acad Sci* **1320**: 76-92.

Baudin-Baillieu A, Hatin I, Legendre R, Namy O. 2016. Translation Analysis at the Genome Scale by Ribosome Profiling. *Methods Mol Biol* **1361**: 105-124.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics* **176**: 1131-1137.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics* **172**: 1675-1681.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* **179**: 487-496.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370-374.

Charron G, Leducq JB, Landry CR. 2014. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. *Mol Ecol* **23**: 4362-4372.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14**: 645-660.

Chuang JH, Li H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol* **2**: E29.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433-3434.

Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **9**: e92621.

Dunn JG, Weissman JS. 2016. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* **17**: 958.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. 2017. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. *Mol Biol Evol* **34**: 1066-1082.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307-321.

Holmes IH. 2017. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. *Bioinformatics* **33**: 1227-1229.

Ingolia NT. 2016. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**: 22-33.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.

Jacob F. 1977. Evolution and tinkering. *Science* **196**: 1161-1166.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241-254.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752-1759.

Landry CR, Zhong X, Nielly-Thibault L, Roucou X. 2015. Found in translation: functions and evolution of a recently discovered alternative proteome. *Curr Opin Struct Biol* **32**: 74-80.

Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol Evol* **3**: 799-811.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**: 15003.

Lefort V, Longueville JE, Gascuel O. 2017. SMS: Smart Model Selection in PhyML. *Mol Biol Evol* **34**: 2422-2424.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* **103**: 9935-9939.

Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol* **6**: e1000734.

Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biol Evol* **8**: 2190-2202.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337-341.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**: 9272-9277.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**: 10-12.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet* **17**: 567-578.

McManus CJ, May GE, Spealman P, Shteyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* **24**: 422-430.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**: 116.

Muller CA, Hawkins M, Retkute R, Malla S, Wilson R, Blythe MJ, Nakato R, Komata M, Shirahige K, de Moura AP et al. 2014. The dynamics of genome replication using deep sequencing. *Nucleic Acids Res* **42**: e3.

Murphy HA, Kuehne HA, Francis CA, Sniegowski PD. 2006. Mate choice assays and mating propensity differences in natural yeast populations. *Biol Lett* **2**: 553-556.

Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat Ecol Evol* **1**: 0217.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**: 117.

Nielly-Thibault L, Landry CR. 2018. Differences between the de novo proteome and its non-functional precursor can result from neutral constraints on its birth process, not necessarily from natural selection alone. *bioRxiv*: doi: 10.1101/289330.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of Drosophila orphan genes. *Elife* **3**: e01311.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420-1428.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, Sigwalt A, Barre B, Freel K, Llored A et al. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**: 339-344.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team. 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Canas JL, Messeguer X, Alba MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* **2**: 890-896.

Schlotterer C. 2015. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet* **31**: 215-219.

Sniegowski PD, Dombrowski PG, Fingerman E. 2002. Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299-306.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393-395.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692-702.

Vakirlis NN, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2017. A molecular portrait of de novo genes in yeasts. *Mol Biol Evol* doi:10.1093/molbev/msx315.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nat Ecol Evol* **1**: 0146-0146.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A* **106**: 7273-7280.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754-D761.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**: 769-772.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**: 1446-1455.

Supplementary methods

Ribosome profiling and mRNA sequencing libraries

Polysome extract preparation

Ribosome profiling and mRNA sequencing experiments were conducted with the strains YPS128 (*S. cerevisiae)* (Sniegowski et al. 2002), YPS744 (*S. paradoxus*), MSH604 (*S. paradoxus*) and MSH587 (*S. paradoxus*) belonging respectively to groups *SpA*, *SpB* and *SpC* according to Leducq et al. (2016). All strains were diploid. For *S. paradoxus*, we constructed homozygous diploids from haploid heterothallic strains containing a resistance cassette (Nourseothricine or Hygromycine B depending of the mating type) at the *HO* locus. Constructions and crosses were performed according to the protocol described in Leducq et al. (2016). Resulting diploid cells containing the two resistance cassettes were selected on solid YPD (Yeast Peptone Dextrose) medium containing100 ug/ml of Nourseothricine and 250 ug/ml of Hygromycine B.

Strains were grown overnight in 50 mL of SOE (Synthetic Oak Exudate) medium (Murphy et al. 2006), at 30°C with shaking at 250 rpm. These pre-cultures were used to inoculate a 30°C pre-warm 750 mL SOE medium at an initial OD_{600} of ~0.03, and grown to an OD_{600} between 0.6 to 0.7, at 30°C and shaking at 250 rpm. We choose the SOE medium to be closed to natural conditions in which *de novo* genes could emerge in wild yeast strains. Cultures were treated with cycloheximide (50 ug/mL final) for 5 minutes and cells were rapidly collected by vacuum filtration using a 90 mm cellulose nitrate filter with a 0.45 mm pore size and a fritted glass support. Cells were resuspended on ice in 2.5 mL of polysome lysis Buffer (10 mM Tris-HCL pH 7.4, 100 mM NaCl, 30 mM MgCl2, 50 ug/mL cycloheximide). The slurry was then pipetted and frozen by fractions of ~ 20 ul in liquid nitrogen, and stored at -80°C. The resulting cryogenized mix was grinded in a MixerMill 400 (RETSCH) for 15 cycles of 2mn at 30 Hz, with chilling in liquid nitrogen

between each cycle. The powder was gently thawed in the open grinding chamber at room temperature to collect the lysate which was cleared by two rounds of centrifugation for 5 minutes at 3000xg at 4°C, followed by one round of high speed centrifugation for 10 min at 20,000xg at 4°C. The middle layer was quantified by OD_{260} measurement via Nanodrop and samples above 200 OD_{260}/mL were diluted to ~200 OD_{260}/mL in lysis buffer. Cell lysates were divided into 250 ul aliquots of 30 to 50 OD_{260} each, that were flash frozen in liquid nitrogen and stored at -80°C. For each lysate, one aliquot of 250 ul was conserved for direct total mRNA extraction, the other aliquots were pooled for ribosome footprint isolation.

Isolation and purification of ribosome footprints

Cell lysate, corresponding to 50 to 100 OD_{260}, were digested with 15 U of RNAse I (Ambion) per OD_{260} for 60 minutes at 25 °C with shaking. The digestion was stopped by adding 200 U of Superase-in (AMBION). The digested products were loaded on a 24% sucrose cushion (50 mM Tris-acetate pH 7.6, 50 mM NH_4Cl, 12 mM $MgCl_2$, 1 mM DTT) and centrifuged at 4°C 100,000 rpm in a TLa110 rotor for 2h15. Pellets were washed two times with lysis buffer and resuspended in 500 uL of polysome lysis buffer. The extract was treated with DNAse I using the manufacturer's instructions (Truseq Ribo profile illumina kit for yeast). RNA was then extracted using acid-phenol-chloroform extraction protocol, and precipitate overnight at -20 °C with 0.1 Volume of sodium acetate 3M, pH 5.2 and 3 volumes of EtOH 100%. Samples were centrifuged at 4°C for 20 minutes at 10,000 g and pellets were resuspended in 75 ul of RNAse free H_20 supplied with 1 U/ul of Superase-in (AMBION). RNA concentration was measured at 260 nm, resulting in a final amount of ~300 to 1000 ng of digested RNA. RNA fragments were separated by electrophoresis on a denaturing 17% PAGE gel with heating at 60 °C at 200V for ~8 hours. A mix of 28 and 34 nt RNA markers, oNTI199 and oNTI34ARN (Ingolia et al.

2012), was loaded at both extremities of the gel. The gel was stained with SYBR Gold according to the manufacturer's instructions and the region corresponding to the 28 marker was excised. Gel slices were disrupted through needle holes in a 0.5 mL centrifuge tubes nested in a 1.5 mL tube by maximum speed centrifugation. RNA was eluted overnight at 4°C with gentle rotation in an elution buffer (300 mM NaOAc pH 5.5, 1 mM EDTA). The slurry was loaded on SpinX cellulose acetate filter to recover the eluted RNA cleared of gel fragments. The RNA was precipitate overnight at -20°C in ethanol with 0.3 M sodium acetate and 20 ug of glycogen. Samples were centrifuged at 4°C for 30 minutes at maximum speed and pellets were resuspended in 25 ul of nuclease-free water supplemented with 0.1 U/mL of Superase-in (AMBION). These samples contain purified ribosome footprints. Total mRNA was extracted using the same acid-phenol-chloroform extraction protocol as for ribosome footprints samples. Purified ribosome footprints and total RNA were then quantified by fluorescence (Quant-it RNA assay kit, thermofisher) and stored at -20°C.

Library preparation

The rRNA was depleted in purified ribosome footprints and total mRNA samples using the Ribo-Zero Gold rRNA Removal Kit for yeast (Illumina) according to the manufacturer's instructions. Ribo-Zero treated RNAs were then purified by overnight ethanol precipitation. Ribosome profiling and total mRNA libraries were constructed using the TruSeq Ribo Profile kit for yeast (illumina), using manufacturer's instructions starting from Fragmentation and end repair step. Circularized cDNA templates were amplified by 11 cycles of PCR using Phusion-polymerase (New England Biolabs), with primers incorporating barcoded Illumina TruSeq library sequences, according to TruSeq Ribo Profile kit for yeast (illumina). The resulting PCR products were loaded onto a 8% native polyacrylamide gel in TBE and purified using the PCR purification protocol

provided in the TruSeq Ribo Profile kit for yeast (illumine). The quality and size of the purified PCR products were assessed using an Agilent HS bioanalyzer. Libraries were quantified by fluorescence using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher). The 8 total RNA libraries were pooled in one bulk for sequencing. RPF libraries were pooled in 2 bulks, one for the first replicate of the 4 strains, and a bulk for the second replicate of the 4 strains. Libraries were sequenced on an Illumina HiSeq 2500 at The Genome Quebec Innovation Center. The total RNA bulk was loaded onto 5 lanes and RPF bulks were loaded onto 4 lanes each.

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**: 1534-1550.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**: 15003.

Murphy HA, Kuehne HA, Francis CA, Sniegowski PD. 2006. Mate choice assays and mating propensity differences in natural yeast populations. *Biol Lett* **2**: 553-556.

Sniegowski PD, Dombrowski PG, Fingerman E. 2002. Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299-306.