Supplemental Figures & Tables

Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution

Chris Papadopoulos[1], Isabelle Callebaut[2], Jean-Christophe Gelly[3,4,5], Isabelle Hatin[1], Olivier Namy[1], Maxime Renard[1], Olivier Lespinet[1], Anne Lopes[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

[2] Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005, Paris, France

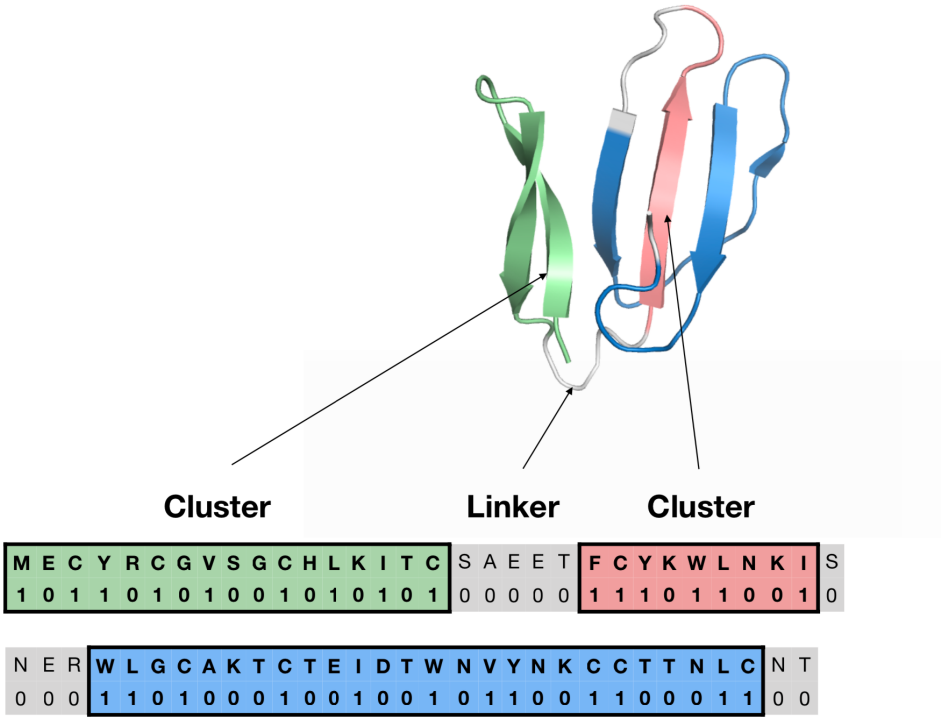
[3] Université de Paris, Biologie Intégrée du Globule Rouge, UMR_S1134, BIGR, INSERM, F-75015, Paris, France.

[4] Laboratoire d'Excellence GR-Ex, Paris, France

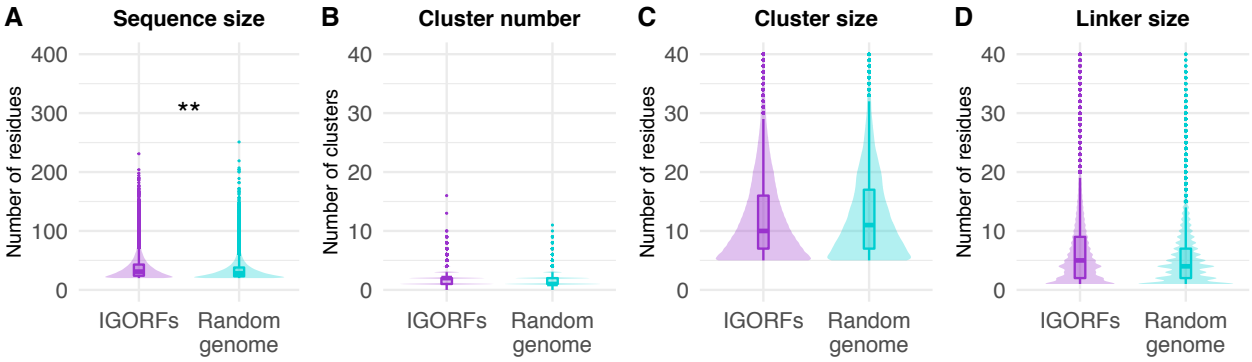
[5] Institut National de la Transfusion Sanguine, F-75015, Paris, France

Content: 17 supplemental figures & 7 supplemental tables

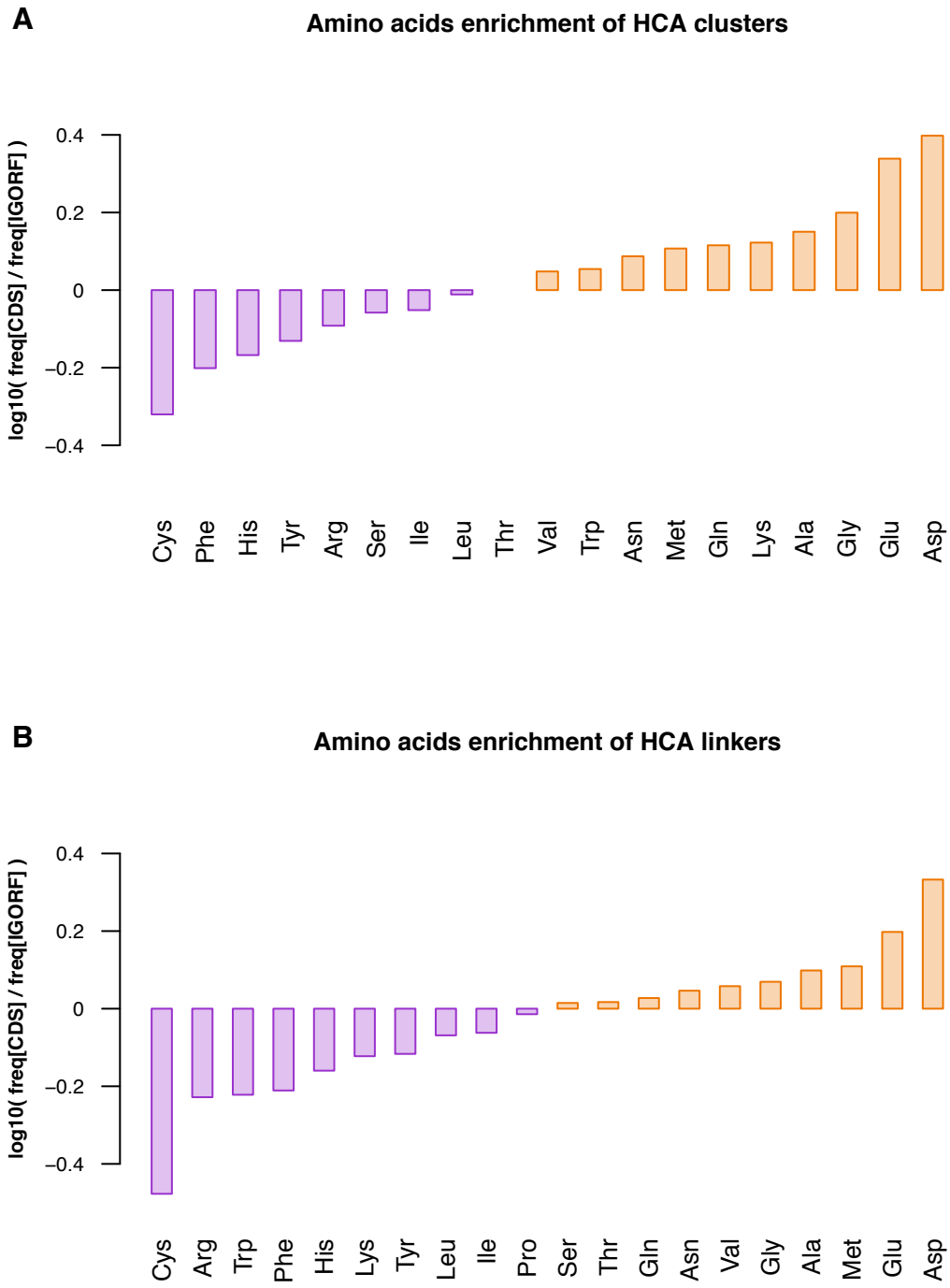
Supplemental Table S2	One-sided Mann-Whitney *U* test p-values for all the ORF categories - Number of clusters
Supplemental Table S3	Two-sided Mann-Whitney *U* test p-values for all the ORF categories – Cluster size
Supplemental Table S4	One-sided Mann-Whitney *U* test p-values for all the ORF categories – Linker size
Supplemental Table S5	Strong hydrophobic residues (V,I,L,F,M,Y,W) frequency per ORF category for the three HCA score categories.
Supplemental Table S6	The 70 de novo genes of *Saccharomyces cerevisiae* used for the ancestral reconstruction.
Supplemental Table S7	Frequencies of the three STOP codons for different ORF categories.



Supplemental Figure S1 | 3D mapping of HCA hydrophobic clusters and linkers
HCA hydrophobic clusters (colored) and linkers (in grey) delineated for the sequence of Bucandin (pdb code: 1f94). The HCA-based sequence, which consists in translating the protein sequence into a binary pattern, is given under the protein sequence. “1” corresponds to strong hydrophobic amino acids (V, I, L, F, M, Y, W) and “0” to the other amino acids (Methods; Supplemental Methods). HCA clusters and linkers are mapped on the 3D structure of Bucandin with respect to the color code used for the sequence.

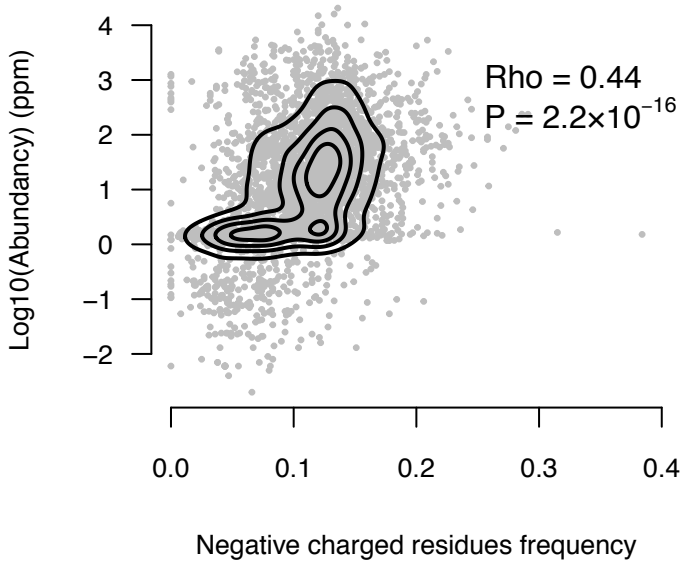


Supplemental Figure S2 | Random IGORFs behave similarly to real IGORFs for most properties
Boxplot distributions of sequence and HCA-based structural properties of real IGORFs and random IGORFs (A) sequence size (B) number of HCA clusters per sequence (C) size of HCA clusters (D) size of linkers. Asterisks denote level of significance: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$

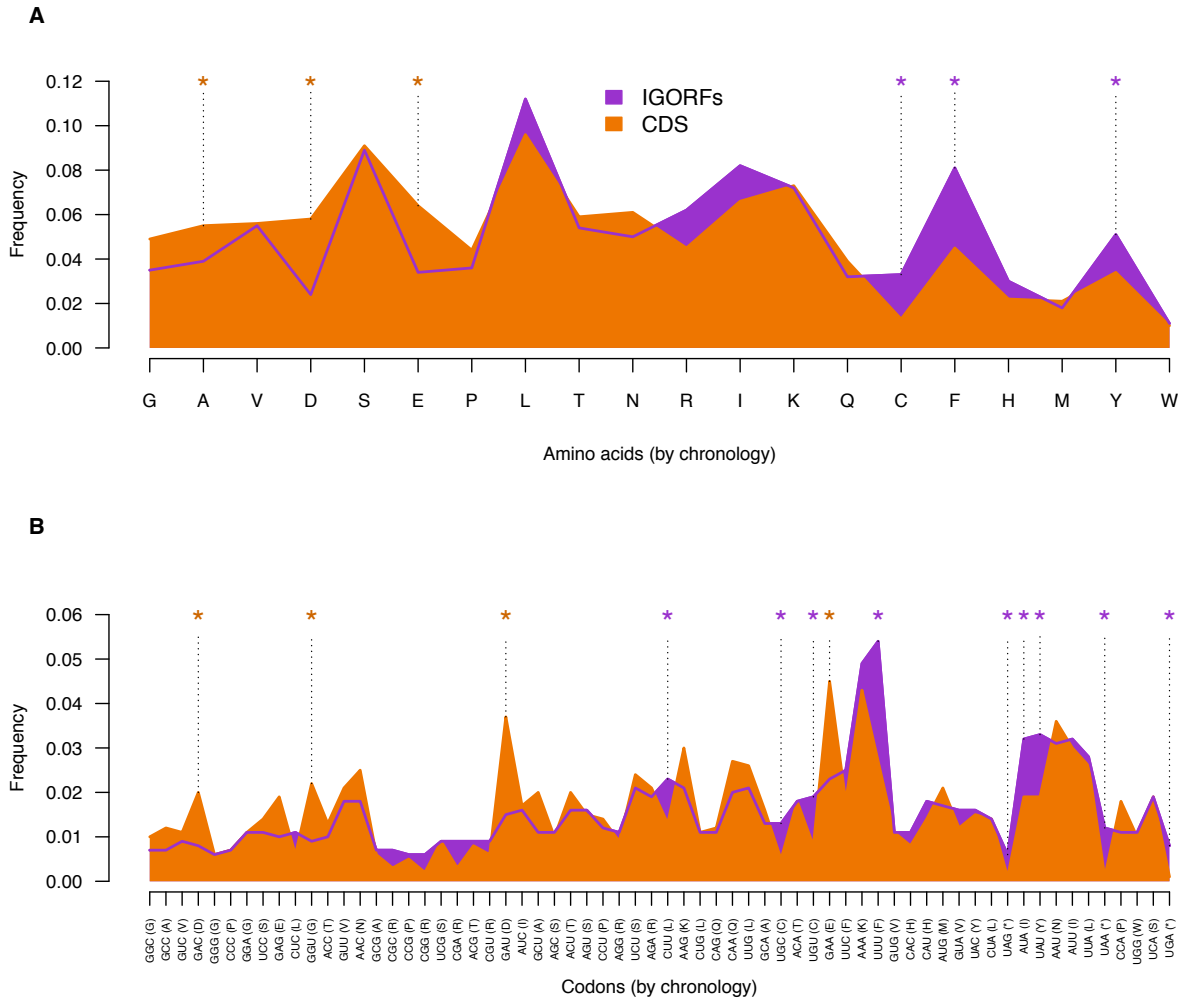


Supplemental Figure S3 | CDS are enriched in hydrophilic residues

(A) Log ratios of amino acid frequencies in HCA clusters of CDS versus HCA clusters of IGORFs. Negative values (purple) correspond to amino acids with higher frequency in IGORF HCA clusters while positive values (orange) correspond to amino acids that are more frequent in CDS HCA linkers. (B) Log ratios of amino acid frequencies in HCA linkers of CDS versus HCA linkers of IGORFs.

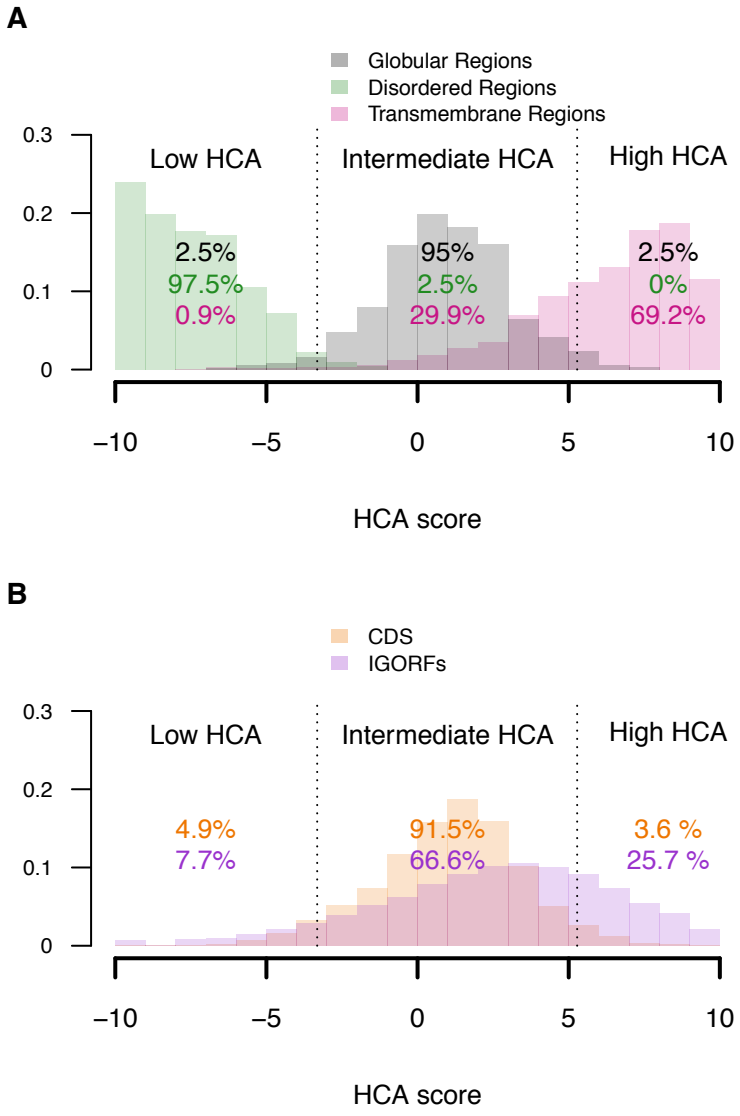


Supplemental Figure S4 | Abundant proteins are enriched in negatively charged amino acids
Protein abundances (in parts per million) of all cytoplasmic proteins are plotted against their corresponding negatively charged residues (Aspartate and Glutamate) frequencies. The Spearman rank correlation coefficient is indicated on the plot (p-value < 2.2×10^{-16}).



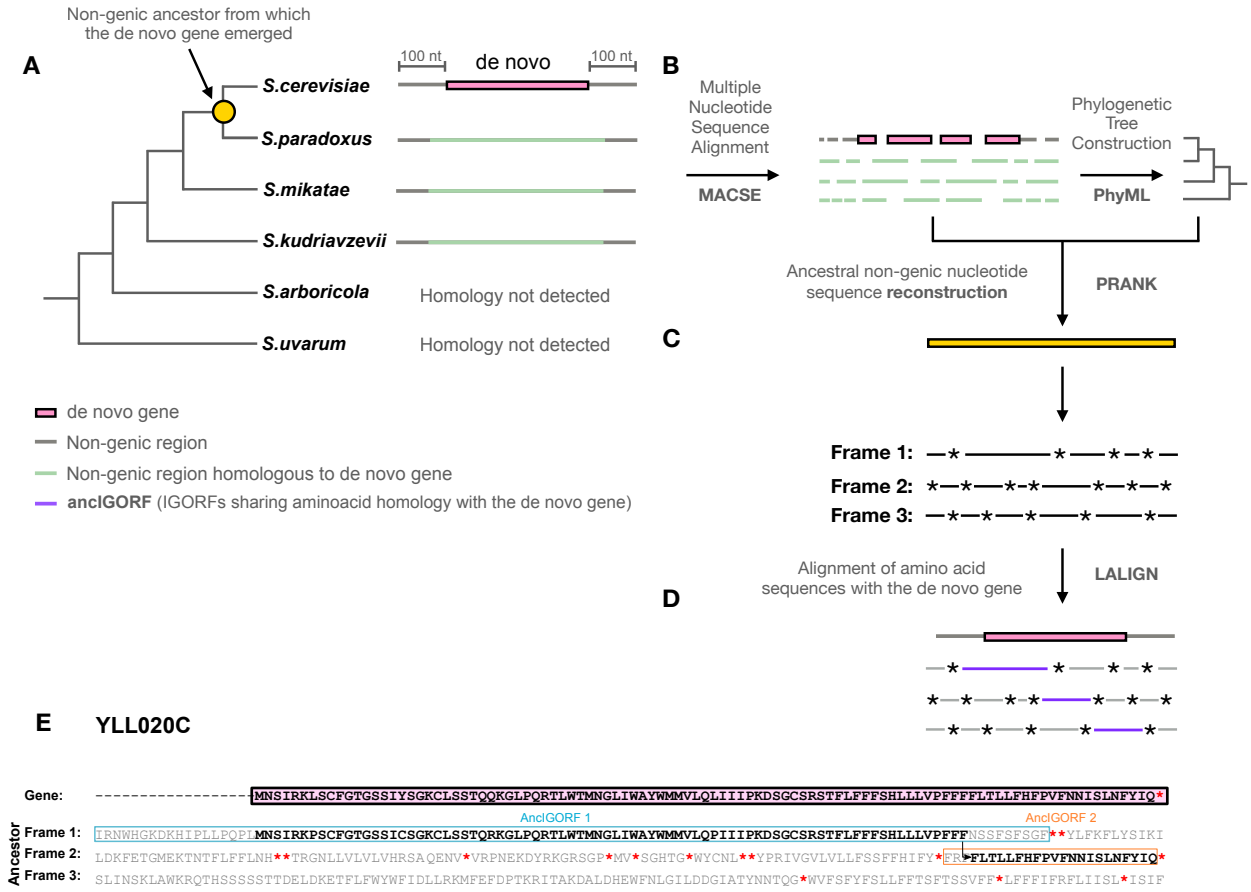
Supplemental Figure S5 | CDS are enriched in ancient amino acids

(A) Frequencies of amino acids of CDS (orange) and IGORFs (purple) ordered according to their chronology of appearance during evolution as defined in Trifonov et al. (2001) (B) Frequencies of codons of CDS (orange) and IGORFs (purple) ordered according to their chronology of appearance during evolution as defined in Trifonov et al. (2001). Amino acids or codons enriched in CDS or IGORFs are indicated by orange or purple stars respectively (z-test, p-values $< 5 \times 10^{-2}$).



Supplemental Figure S6 | IGORFs encompass the large spectrum of fold potential of canonical proteins (raw data)

(A) Histograms of the HCA scores of the three reference datasets (i.e. disordered regions, globular domains and transmembrane regions – green, black and pink histograms respectively). Dotted black lines delineate the boundaries of the low, intermediate and high HCA score categories. The boundaries are defined so that 95% of globular domains fall into the intermediate HCA score category whereas the low and high HCA score categories include all sequences with HCA values that are lower or higher than those of 97.5% of globular domains respectively. (B) Histograms of the HCA scores of CDS and IGORFs. The percentages of sequences in each category are given for all datasets.

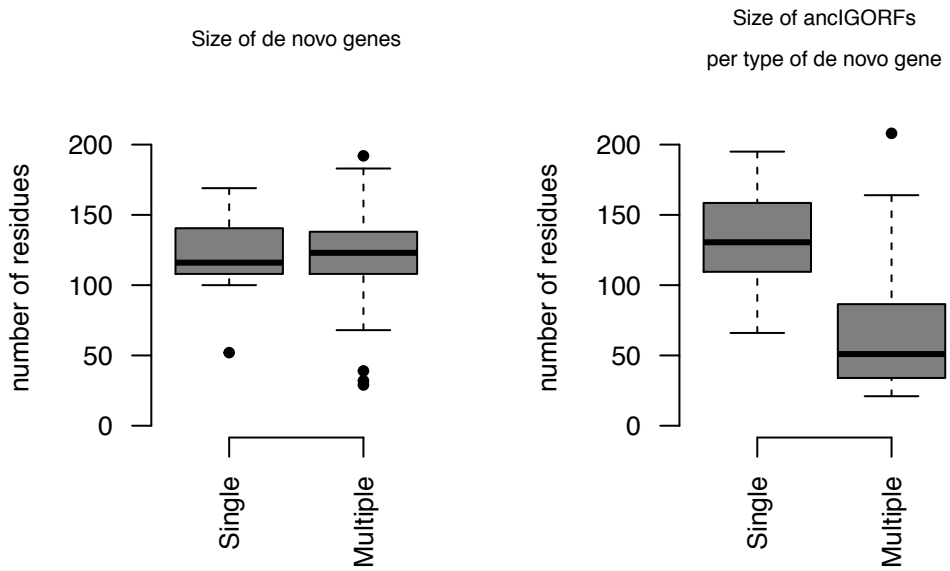


Supplemental Figure S7 | Reconstruction of the ancestral IGORFs (ancIGORFs) which gave birth to known de novo genes

(A) Identification of homologous sequences (that can be an orthologous gene or a homologous noncoding sequence) of the de novo gene of interest in all neighboring species with blast (Altschul et al. 1990) (see Methods for more details) (B) Multiple sequence alignment of the detected homologous nucleotide sequences with MACSE (Ranwez et al. 2011, 2018) and construction of their phylogenetic tree with PhyML (Guindon et al. 2010) (C) reconstruction of the corresponding ancestral nongenic nucleotide sequence (in yellow) with PRANK (Löytynoja and Goldman 2010). The latter is subsequently translated into the three frames. STOP codons are indicated with stars. (D) Alignment of all the reconstructed IGORFs (amino acid sequences) with the de novo gene(s) of interest with LALIGN (Huang and Miller 1991) and detection of the IGORFs sharing a homology with it (i.e. ancIGORFs) (E) Alignment of the *S. cerevisiae* de novo gene YLL020C with the translation products of its corresponding ancestral noncoding sequence as predicted for the ancestor of *S. cerevisiae* and *S. paradoxus*. STOP codons are indicated with red stars. The two IGORFs which gave birth to the YLL020C gene (ancIGORFs) are indicated by blue and orange boxes respectively. The two ancIGORFs are distributed across two frames showing that the current version of YLL020C results from an indel event which induces a frameshift in the original sequence. The sections of the ancIGORFs that participate in the resulting de novo gene are indicated in bold. The HCA scores of the blue and orange IGORFs are 0.48 (foldable) and 7.71 (aggregation-prone) respectively.

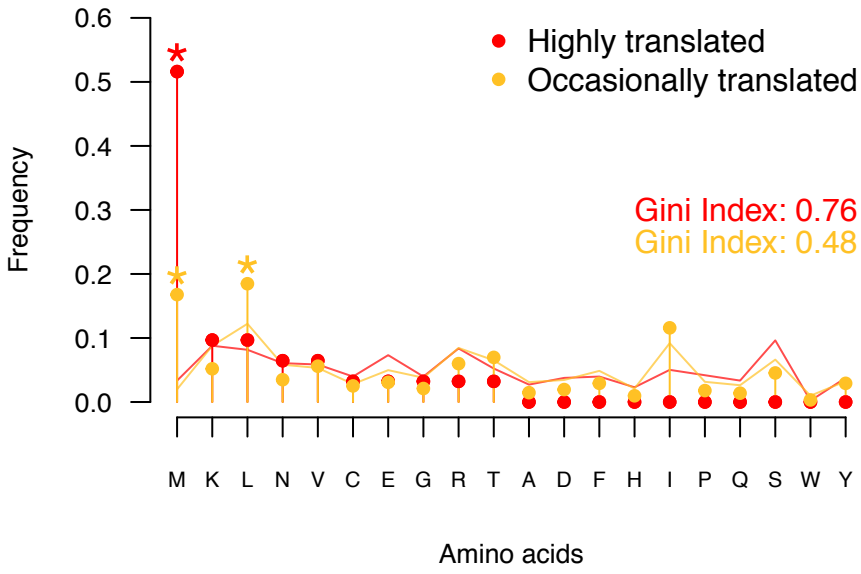
Supplemental Figure S8 | Appearance of a Methionine and fusion of two ancIGORFs in the *S. cerevisiae* lineage

The sequences of the YOR333C de novo gene and its corresponding noncoding regions in the five neighboring species of *S. cerevisiae* are indicated in blue. The ancestral sequences are indicated in yellow. STOP codons are represented with red stars. The appearance of the Methionine in the *S. cerevisiae* lineage is highlighted with a grey box while the STOP codon mutation that led to the fusion of the two ancIGORFs in the *S. cerevisiae* lineage is indicated with a green box.



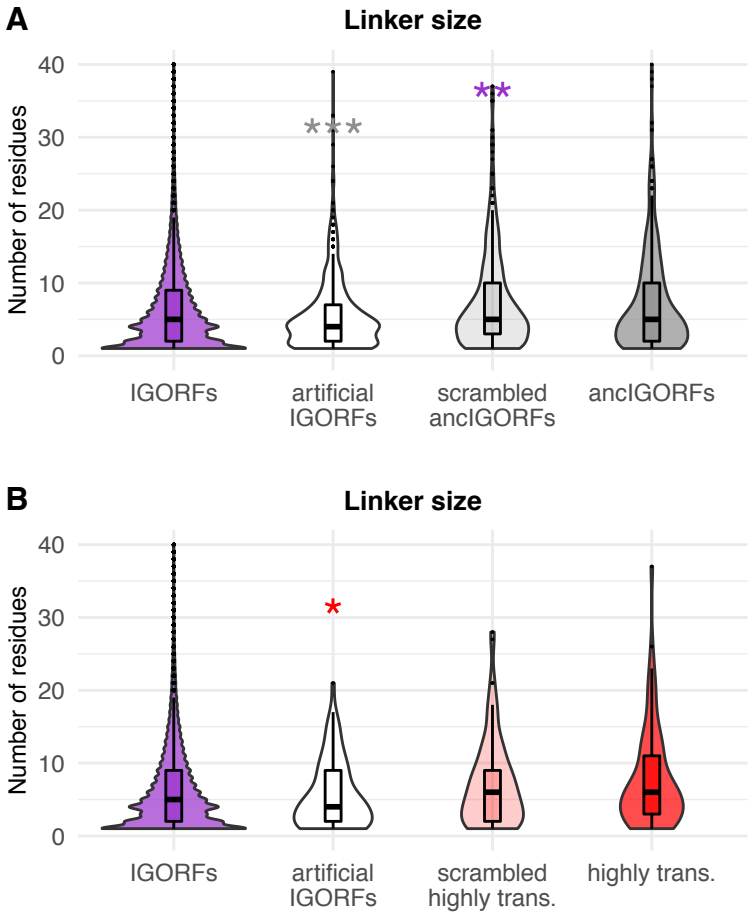
Supplemental Figure S9 | De novo gene categories display similar sizes while their corresponding ancIGORFs exhibit different sizes

(A) Boxplot comparing the sequence size of multiple and single ancIGORF de novo genes. (B) Boxplot comparing the sequence size of ancIGORFs preceding the emergence of single and multiple ancIGORF de novo genes.



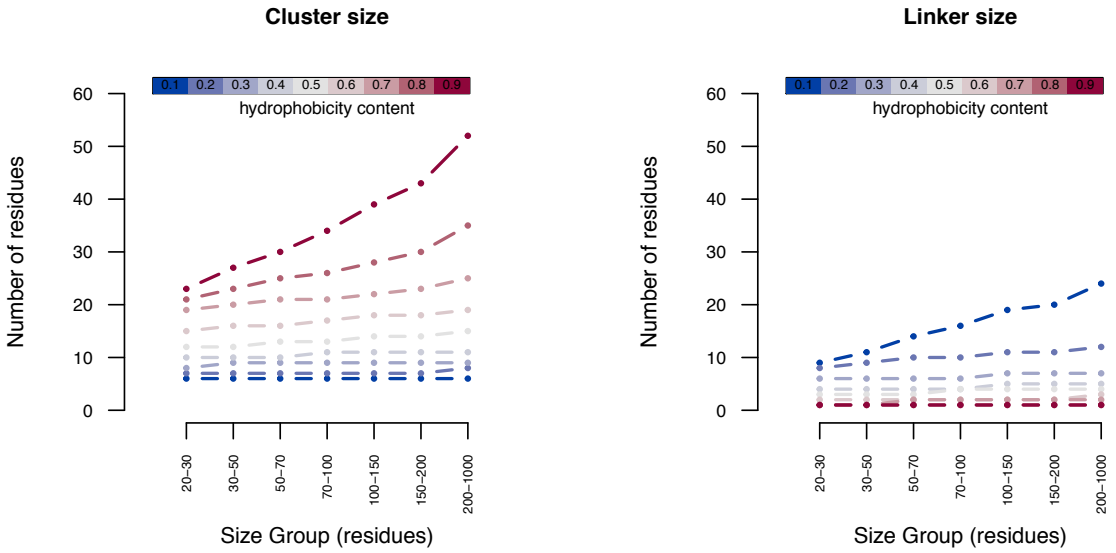
Supplemental Figure S10 | Translated IGORFs are mostly initiated with Methionine

Frequencies of the 20 amino acids at the first translated position for highly translated IGORFs (red) and occasionally translated ones (yellow). Gini indexes which reflect the statistical dispersion of the 20 amino acids at the first translated position are given for highly and occasionally translated IGORFs in red and yellow respectively. Gini index values range from 0 to 1 and high values reflect the fact that the first translated positions are enriched in specific amino acids, particularly, in MET and to a lesser extent in LEU for occasionally translated IGORFs. Amino acids which are significantly observed at the first translated position compared to the other translated positions are indicated with a star (z-test p.value $< 5 \times 10^{-2}$).



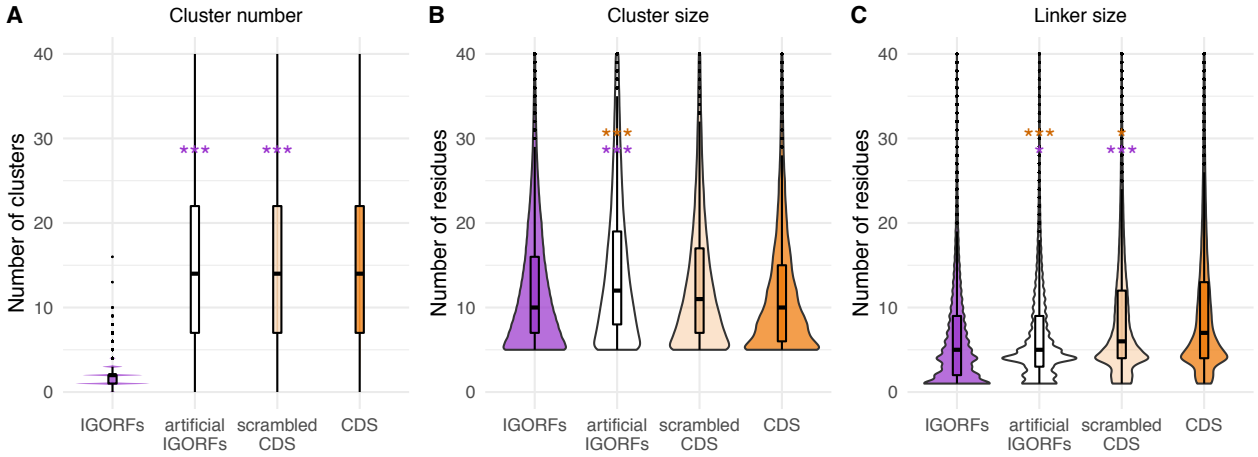
Supplemental Figure S11 | The nucleotide composition of ancestral and highly translated IGORFs seems to play an important role in the linker's size

(A) Linkers' size for real IGORFs (purple), artificial IGORFs (i.e. ORFs with size similar to ancIGORFs but nucleotide composition of IGORFs) (white), ancIGORFs with scrambled nucleotides (light grey) and real ancIGORFs (grey). (B) Linkers' size for real IGORFs (purple), artificial IGORFs (i.e. ORFs with size similar to highly translated IGORFs but nucleotide composition of IGORFs) (white), highly translated IGORFs with scrambled nucleotides (ligh tred) and real highly translated IGORFs (red). The p-values were computed with the Mann-Whitney *U* test (one-sided). Asterisks denote level of significance: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$. The color of the asterisks indicates the ORF category used for the comparison.



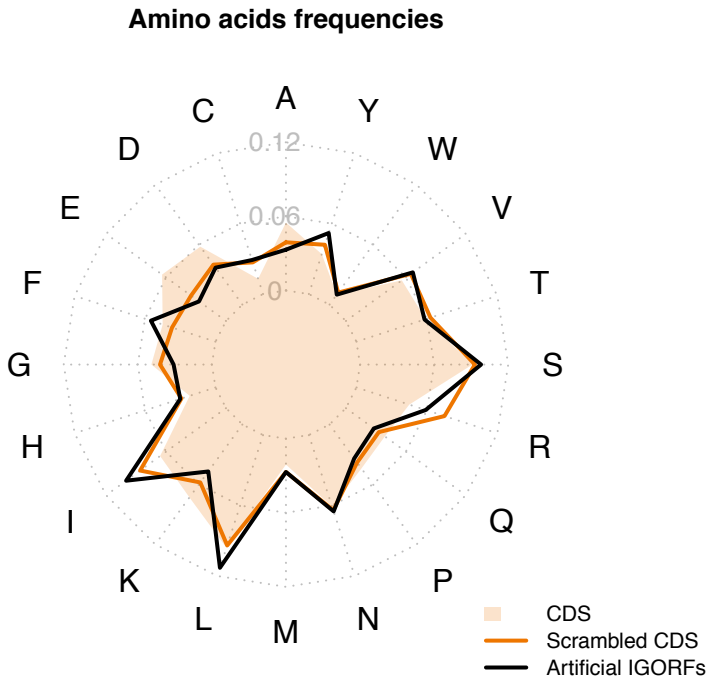
Supplemental Figure S12 | Impact of the hydrophobicity content and sequence length on the size of clusters and linkers

In order to properly decipher the contributions of the amino acid composition and sequence length, we generated artificial sequences with different sizes and different hydrophobic residue contents (1000 sequences per bin of sequence size and hydrophobicity content). (A) The median values of the resulting cluster sizes are subsequently plotted in number of residues. (B) For the same artificial sequences, the median values of the resulting linker sizes are plotted in number of residues. In both plots sequences are colored according to their hydrophobicity content that ranges from 0.1 (i.e. 10% of strong hydrophobic residues according to HCA definition: V, I, L, M, Y, F, W and C) to 0.9. For a given sequence length, hydrophobic and hydrophilic contents have a significant impact on the size of clusters and linkers respectively with an even more important effect on long sequences.

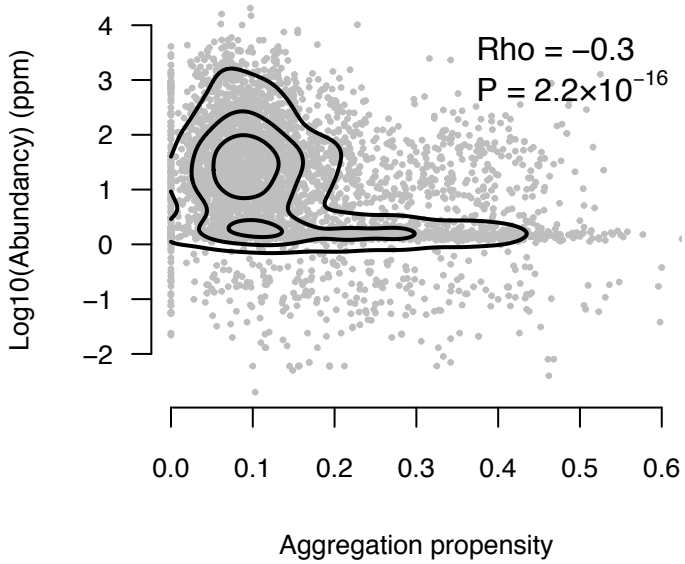


Supplemental Figure S13 | Effect of the sequence length, and GC content on the size of clusters and linkers

Number of HCA clusters (A), size of HCA clusters (B) and size of linkers (C) for real CDS sequences (orange), scrambled CDS sequences (light orange) and artificial IGORFs (i.e. with size similar to CDS but nucleotide compositions of IGORFs (white). The clusters of scrambled CDS are similar to those of CDS while their linkers are slightly shorter (Mann-Whitney *U* test, $P = 4 \times 10^{-2}$) showing that randomly and according to the GC content and size of CDS, long though slightly shorter linkers can be generated. In contrast, the linkers of artificial IGORFs are of comparable size to those of IGORFs though slightly larger, while the artificial clusters are longer (Mann-Whitney *U* test, $P = 4 \times 10^{-2}$ and $P = 6 \times 10^{-4}$ respectively). This reflects that at the IGORF GC content, the sequence length alone has a small impact on cluster size while the effect is marginal on linker size, and overall cannot explain the increase in linker size observed for CDS. Indeed, the artificial linkers are clearly shorter than those of both real and scrambled CDS (Mann-Whitney *U* test, $P = 7.1 \times 10^{-8}$ and 2×10^{-4} respectively) highlighting the impact of the amino acid composition but also of the GC content of the CDS on their linker size. The p-values were computed with the Mann-Whitney *U* test (one-sided). Asterisks denote level of significance: $*p < 5 \times 10^{-2}$, $**p < 1 \times 10^{-2}$, $***p < 1 \times 10^{-3}$. The color of the asterisks indicates the ORF category used for the comparison.



Supplemental Figure S14 | Impact of the GC content on the resulting amino acid compositions
Radar plot reflecting the 20 amino acid frequencies for real CDS (light orange shadow), scrambled CDS (orange line) and artificial IGORFs (i.e. sequences with size similar to CDS but nucleotide compositions of IGORFs (black line)). CDS and artificial IGORFs exhibit slightly different GC contents (GC content of 36.1% and 39.6% for IGORFs and CDS respectively) that lead to slightly different amino acid compositions.

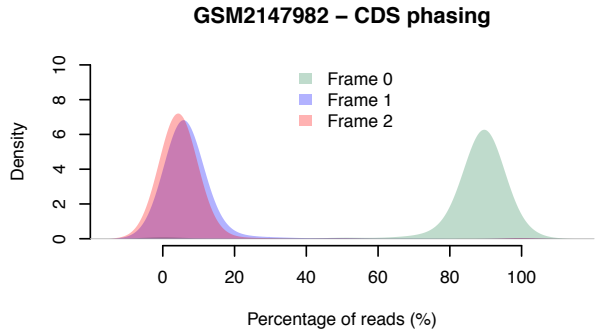


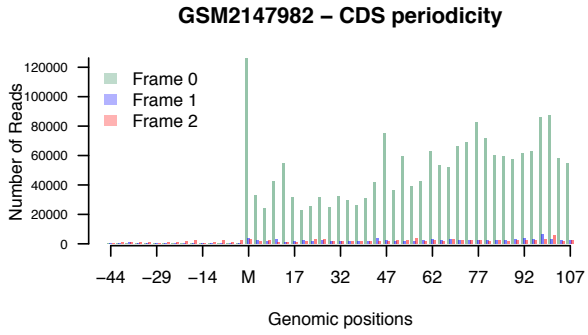
Supplemental Figure S15 | Lowly abundant proteins display a large spectrum of aggregation propensities

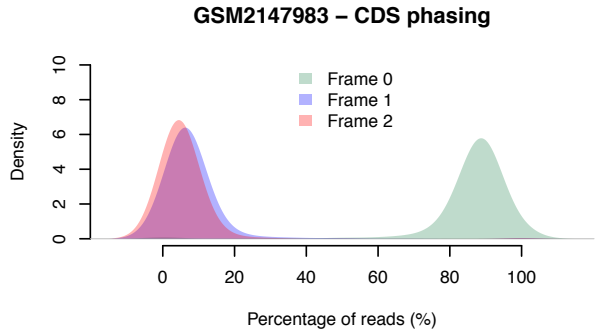
Protein abundances (in parts per million) of all cytoplasmic proteins are plotted against their corresponding aggregation propensity predicted with TANGO (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006). The Spearman rank correlation coefficient is -0.30 with p-value < 2.2×10^{-16}.

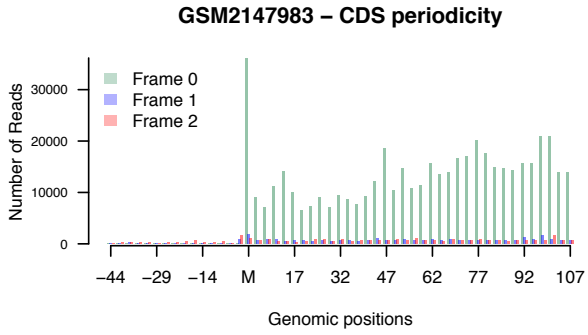
Supplemental Figure S16 | The fusion of IGORFs can lead to longer clusters or linkers

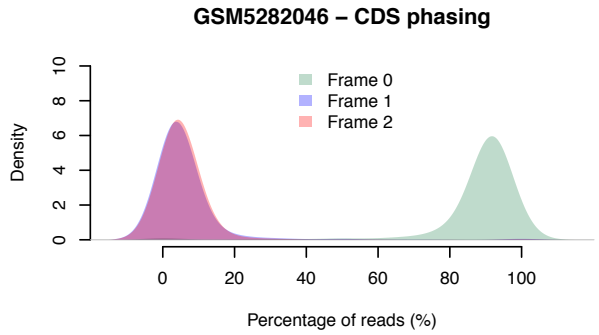
The sequence of the YMR153C-A de novo gene (A) and YPR126C (B) are indicated by the blue boxes while their corresponding ancestral sequences are indicated by the yellow boxes. STOP codons are represented by red stars. HCA clusters are highlighted by red boxes while HCA linkers correspond to the regions connecting two HCA clusters or extremities that are not associated with an HCA cluster.

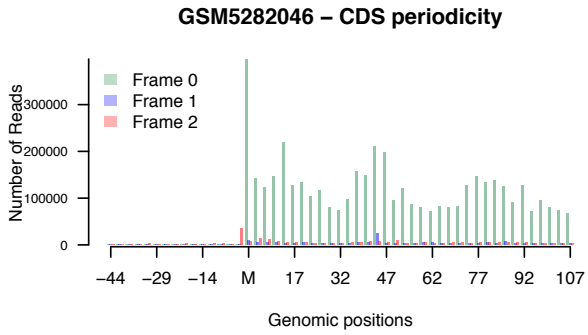

GSM2147982 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
Percentage of reads (%)

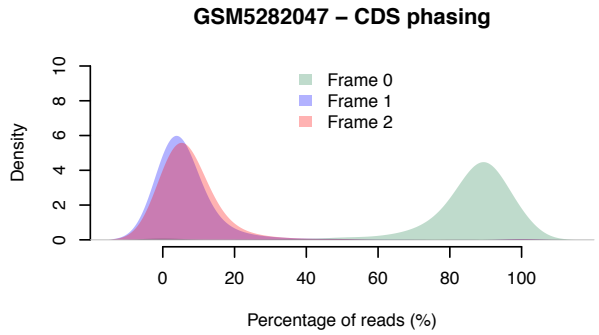

GSM2147982 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
Genomic positions

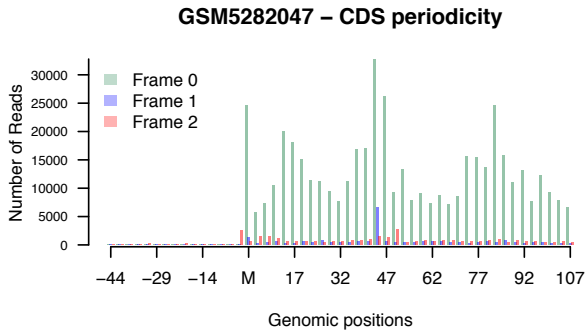

GSM2147983 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
Percentage of reads (%)

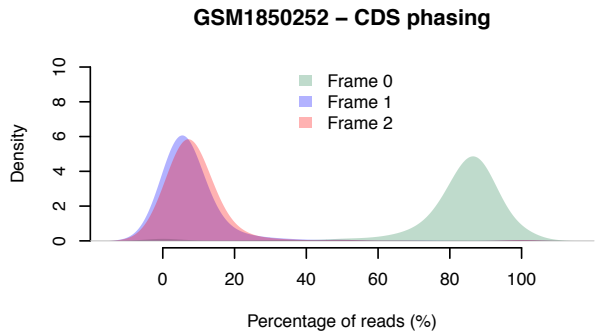

GSM2147983 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
Genomic positions

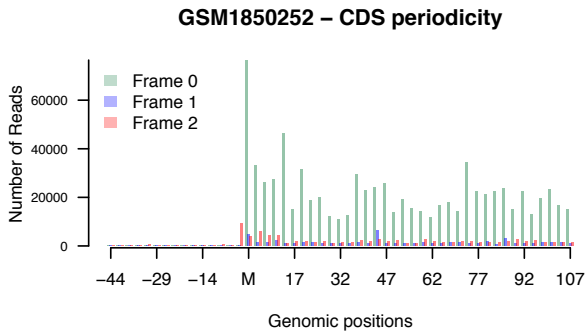

GSM5282046 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
Percentage of reads (%)


GSM5282046 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
Genomic positions


GSM5282047 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
Percentage of reads (%)


GSM5282047 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
Genomic positions


GSM1850252 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
Percentage of reads (%)


GSM1850252 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
Genomic positions

Supplemental Figure S17 | Quality control for the 28-mer Ribosome Protected Fragments (RPFs) used for the detection of occasionally and selectively translated IGORFs for all five experiments

The left panel shows that 90% (on average) of the 28-mer RPFs are in frame with the start codon of the CDS (Frame 0). The right panel presents the number of RPFs at each nucleotide position (determined by the site P of each 28-mer) showing accumulation of signal over the CDS (reads detected only after the start codon), and a nice periodicity (of frame 0) over the 100 first nucleotides. These results inform us about the good quality of the RPF data in all five experiments.

Supplemental Table S1. One-sided Mann-Whitney *U* test p-values for all the ORF categories – Sequence length (in amino acids)

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	3×10^{-4}	3×10^{-2}	2.2×10^{-23}	5.2×10^{-38}	1.4×10^{-153}
Occasionally translated		-	2×10^{-1}	1.3×10^{-15}	1.8×10^{-36}	1.0×10^{-150}
Highly translated			-	1.3×10^{-3}	1.5×10^{-13}	2.2×10^{-19}
Ancestral IGORFs				-	1.1×10^{-16}	1.7×10^{-63}
De novo genes					-	5.2×10^{-20}
CDS						-

Supplemental Table S2. One-sided Mann-Whitney *U* test p-values for all the ORF categories - Number of clusters

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	2×10^{-2}	6×10^{-2}	3.2×10^{-15}	1×10^{-35}	7×10^{-148}
Occasionally translated		-	2×10^{-1}	8×10^{-11}	3.3×10^{-33}	1.9×10^{-142}
Highly translated			-	1×10^{-2}	1.1×10^{-10}	2.2×10^{-18}
Ancestral IGORFs				-	1.7×10^{-13}	1.8×10^{-60}
De novo genes					-	3.3×10^{-20}
CDS						-

Supplemental Table S3. Two-sided Mann-Whitney *U* test p-values for all the ORF categories – Cluster size

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	5×10^{-1}	7×10^{-1}	6×10^{-1}	8×10^{-2}	1×10^{-1}
Occasionally translated		-	6×10^{-1}	6×10^{-1}	6×10^{-2}	1×10^{-1}
Highly translated			-	7×10^{-1}	4×10^{-1}	2×10^{-1}
Ancestral IGORFs				-	5×10^{-2}	1×10^{-1}
De novo genes					-	2×10^{-3}
CDS						-

Supplemental Table S4. One-sided Mann-Whitney *U* test p-values for all the ORF categories – Linker size

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	1×10^{-1}	2×10^{-2}	1×10^{-2}	9×10^{-5}	6.3×10^{-11}
Occasionally translated		-	9×10^{-2}	1×10^{-1}	2×10^{-3}	1.5×10^{-8}
Highly translated			-	7×10^{-1}	3×10^{-1}	8×10^{-3}
Ancestral IGORFs				-	3×10^{-2}	7.9×10^{-7}
De novo genes					-	1.1×10^{-3}
CDS						-

Supplemental Table S5. Strong hydrophobic residues (V,I,L,F,M,Y,W) frequency per ORF category for the three HCA score categories.

	Low HCA	Intermediate HCA	High HCA	Total
IGORFs	0.239	0.391	0.508	0.410
Occasionally translated	0.241	0.384	0.494	0.401
Highly translated	0.251	0.355	0.406	0.353
Ancestral IGORFs	0.241	0.376	0.508	0.392
De novo genes	0.215	0.398	0.479	0.410
CDS	0.219	0.332	0.475	0.328

Supplemental Table S6. ORF names of the 70 de novo genes of *Saccharomyces cerevisiae* used for the ancestral reconstruction. For the last two columns the hydrophobic residues considered are: V,I,L,F,M,Y,W and the hydrophilic ones are: K,R,D,E,Q,N.

Gene name	Ancestral type	Protein size	HCA score	HCA bin	Clusters count	Disorder propensity	Aggregation propensity	Hydrophobic percentage	Hydrophilic percentage
YAL026C-A	multiple	145	5.54	high	5	0	0.407	0.475	0.31
YAL031W-A	multiple	102	3.87	intermediate	4	0	0.284	0.431	0.225
YAL047W-A	single	109	4.06	intermediate	6	0.055	0.312	0.451	0.203
YBL100W-C	multiple	39	5.97	high	3	0.205	0.256	0.386	0.361
YBR056C-B	single	52	-5.44	low	1	0.5	0.096	0.211	0.308
YBR206W	multiple	107	1.68	intermediate	4	0.056	0.121	0.299	0.243
YCL058C	single	152	6.28	high	6	0	0.546	0.52	0.107
YCR085W	multiple	117	4.33	intermediate	6	0	0.385	0.512	0.196
YDL016C	single	100	0.78	intermediate	3	0	0.05	0.37	0.25
YDL158C	multiple	102	7.32	high	2	0	0.598	0.509	0.256
YDR024W	multiple	161	0.01	intermediate	7	0.062	0.174	0.336	0.273
YDR154C	multiple	116	0.34	intermediate	3	0.086	0.19	0.379	0.198
YDR327W	multiple	108	1.98	intermediate	5	0.046	0.093	0.381	0.27
YDR396W	multiple	166	4.39	intermediate	8	0	0.373	0.385	0.222
YDR426C	multiple	125	6.5	high	4	0	0.392	0.504	0.232
YER014C-A	multiple	153	4.46	intermediate	10	0.039	0.301	0.399	0.248
YER046W-A	multiple	109	2.96	intermediate	7	0.073	0.165	0.404	0.212
YER076W-A	single	115	3.82	intermediate	3	0.087	0.252	0.409	0.26
YER087C-A	multiple	183	2.54	intermediate	9	0.055	0.213	0.382	0.131
YER133W-A	multiple	113	2.3	intermediate	4	0.071	0.124	0.39	0.239
YFR026C	single	169	1.56	intermediate	6	0.101	0.213	0.314	0.32
YGL152C	multiple	225	5.4	high	9	0	0.409	0.422	0.129
YGL165C	multiple	192	3.51	intermediate	7	0.026	0.349	0.421	0.218
YGL214W	single	161	0.34	intermediate	7	0.05	0.081	0.324	0.267
YGR011W	multiple	108	3.06	intermediate	5	0	0.352	0.407	0.24
YGR050C	multiple	118	1.79	intermediate	5	0.153	0.042	0.372	0.287
YGR064W	multiple	122	5.04	intermediate	5	0	0.115	0.353	0.229
YGR137W	multiple	124	2.85	intermediate	5	0.04	0.298	0.434	0.242
YGR151C	single	111	0.26	intermediate	7	0.045	0.135	0.369	0.387
YHL006W-A	single	117	2.59	intermediate	8	0	0.051	0.352	0.155
YHR022C-A	multiple	29	4.86	intermediate	3	0	0	0.447	0.274
YHR071C-A	single	106	7.51	high	3	0	0.179	0.452	0.264
YHR180W	single	163	1.77	intermediate	9	0.037	0.227	0.404	0.227
YIL028W	multiple	132	4.84	intermediate	6	0	0.333	0.448	0.219
YIL030W-A	multiple	112	5.07	intermediate	4	0	0.384	0.482	0.242
YIL066W-A	multiple	147	2.14	intermediate	6	0.088	0.259	0.341	0.205
YIL071W-A	multiple	158	4.16	intermediate	8	0.082	0.399	0.444	0.196
YIL086C	multiple	102	-0.97	intermediate	5	0.157	0.118	0.332	0.314
YJL077W-B	multiple	32	4.09	intermediate	2	0	0.156	0.436	0.374
YJL119C	single	107	0.29	intermediate	4	0.047	0.224	0.328	0.326

YJL142C	multiple	130	4.83	intermediate	7	0.054	0.523	0.483	0.215
YJL211C	multiple	147	0.02	intermediate	7	0.095	0.136	0.328	0.273
YJR018W	multiple	120	4.38	intermediate	4	0	0.25	0.39	0.184
YJR020W	single	110	3.38	intermediate	5	0	0.245	0.391	0.227
YJR087W	multiple	116	3.87	intermediate	5	0	0.267	0.432	0.207
YKL036C	multiple	130	1.35	intermediate	3	0.131	0.362	0.37	0.261
YKL053W	multiple	124	2.3	intermediate	6	0.056	0.452	0.54	0.217
YKL076C	multiple	127	3.34	intermediate	6	0	0.299	0.447	0.259
YKL123W	multiple	126	4.13	intermediate	5	0	0.31	0.428	0.317
YKL136W	multiple	132	-0.14	intermediate	6	0	0.311	0.355	0.182
YKL153W	multiple	169	3.54	intermediate	6	0.036	0.249	0.414	0.308
YLL020C	multiple	101	3.85	intermediate	6	0.059	0.436	0.487	0.18
YLR041W	multiple	106	4.71	intermediate	3	0.094	0	0.34	0.378
YLR171W	single	129	4.02	intermediate	6	0.047	0.504	0.473	0.165
YLR255C	multiple	117	-1.42	intermediate	3	0.197	0.154	0.308	0.342
YLR412C-A	multiple	68	-4.51	low	2	0.676	0	0.221	0.515
YLR434C	multiple	127	5.39	high	4	0	0.181	0.393	0.329
YMR052C-A	single	121	7.22	high	4	0	0.479	0.562	0.207
YMR103C	multiple	120	-2.13	intermediate	3	0.108	0.158	0.342	0.251
YMR119W-A	single	124	8.48	high	2	0	0.5	0.557	0.216
YMR153C-A	multiple	111	4.12	intermediate	3	0.045	0.252	0.432	0.252
YMR173W-A	multiple	394	2.82	intermediate	21	0.018	0.312	0.445	0.111
YNL150W	multiple	135	1.89	intermediate	6	0.104	0.222	0.348	0.23
YNL226W	multiple	136	2.43	intermediate	5	0.037	0.228	0.434	0.236
YNL269W	multiple	131	1.98	intermediate	7	0.122	0.168	0.375	0.305
YOR316C-A	multiple	69	-1.2	intermediate	3	0.217	0	0.318	0.274
YOR333C	multiple	138	5.46	high	6	0	0.159	0.413	0.326
YPL056C	multiple	101	0.53	intermediate	6	0	0.218	0.367	0.209
YPR126C	multiple	102	7.54	high	4	0	0.461	0.568	0.186
YPR150W	multiple	173	5.63	high	5	0	0.416	0.462	0.169

Supplemental Table S7. Frequencies of the three STOP codons for different ORF categories.

	UAA	UAG	UGA
IGORFs	0.45	0.24	0.31
Occasionally translated	0.48	0.23	0.29
Highly translated	0.48	0.32	0.20
CDS	0.47	0.23	0.30

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**: 403–410.

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology* **22**: 1302–1306.

Guindon S, Dufayard J-F, 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. *Systematic biology* **59**: 307–321.

Huang X, Miller W. 1991. A time-efficient, linear-space local similarity algorithm. *Advances in applied mathematics* **12**: 337–357.

Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *Journal of molecular biology* **342**: 345–353.

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC bioinformatics* **11**: 579.

Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular biology and evolution* **35**: 2582–2584.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS one* **6**: e22594.

Rousseau F, Schymkowitz J, Serrano L. 2006. Protein aggregation and amyloidosis: confusion of the kinds? *Current opinion in structural biology* **16**: 118–126.

Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN. 2001. Distinct stages of protein evolution as suggested by protein sequence analysis. *Journal of molecular evolution* **53**: 394–401.