Proto–genes and de novo gene birth

Anne–Ruxandra Carvunis1,2,3, Thomas Rolland1,2, Ilan Wapinski4, Michael A. Calderwood1,2, Muhammed A. Yildirim5, Nicolas Simonis1,2,⁷, Benoît Charloteaux1,2,⁶, César A. Hidalgo7, Justin Barbet1,2,3, Balaji Santhanam1,², Gloria A. Brar⁸, Jonathan S. Weissman⁸, Aviv Regev⁹,¹ⁱ, Nicolas Thierry–Mieg³, Michael E. Cusick¹,² & Marc Vidal¹,²

Novel protein-coding genes can arise either through re-organization of pre-existing genes or de novo⁵. Processes involving re-organization of pre-existing genes, notably after gene duplication, have been extensively described⁶,⁷. In contrast, de novo gene birth remains poorly understood, mainly because translation of sequences devoid of genes, or ‘non-genic’ sequences, is expected to produce insignificant polypeptides rather than proteins with specific biological functions⁸,⁹. Here we formalize an evolutionary model according to which functional genes evolve de novo through transitory proto-genes generated by widespread translational activity in non-genic sequences. Testing this model at the genome scale in Saccharomyces cerevisiae, we detect translation of hundreds of short species-specific open reading frames (ORFs) located in non-genic sequences. These translation events seem to provide adaptive potential, as suggested by their differential regulation upon stress and by signatures of retention by natural selection. In line with our model, we establish that S. cerevisiae ORFs can be placed within an evolutionary continuum ranging from non-genic sequences to genes. We identify ~1,900 candidate proto-genes among S. cerevisiae ORFs and find that de novo gene birth from such a reservoir may be more prevalent than sporadic gene duplication. Our work illustrates that evolution exploits seemingly dispensable sequences to generate adaptive functional innovation.

Both genome-wide surveys and analyses of individual cases have shown that de novo gene birth has occurred throughout the evolution of many lineages, potentially affecting species-specific adaptations and evolutionary radiations¹,²,³,⁶,⁸,⁹. Genes are thought to emerge de novo when non-genic sequences become transcribed, acquire ORFs and the corresponding non-genic transcripts access the translation machinery. However, it is hard to reconcile this proposed mechanism with expectations that non-genic sequences should lack translational activity and, even if translated, should encode insignificant polypeptides. Evidence of associations between non-genic transcripts and ribosomes has suggested that non-genic sequences may occasionally be translated, which could provide raw material for natural selection. It has also been speculated that genes that originate de novo could initially be simple and gradually become more complex over evolutionary time. These ideas are consistent with reports showing that genes that emerged recently are shorter, less expressed and more rapidly diverging than other genes⁴,⁸,¹⁰,¹¹. We developed an integrative evolutionary model whereby de novo gene birth proceeds through intermediate and reversible proto-gene stages, mirroring the well-described pseudo-gene stages of gene death (Fig. 1a)⁴,¹¹.

We investigated this model at genome scale in the context of de novo gene birth in S. cerevisiae⁴,¹¹. In S. cerevisiae, a minimal length threshold of 300 nucleotides was originally used to delineate ORFs likely to be genes from non-genic ORFs occurring by chance in non-genic sequences¹⁵. The resulting gene catalogue has undergone numerous adjustments¹⁶, with currently ~6,000 ORFs annotated as genes and ~261,000 unannotated ORFs containing at least three codons considered to be non-genic ORFs (Supplementary Fig. 1). Non-genic sequences are broadly transcribed in S. cerevisiae¹⁴, their overexpression is mostly non-toxic¹⁸, and the corresponding transcripts can associate with ribosomes, often at AUG start codons¹⁹. We reasoned that translation of non-genic ORFs could be more common than expected. Such translation events would not systematically lead to de novo gene birth, as the corresponding polypeptides would not necessarily have specific biological functions. Instead, upon translation, non-genic ORFs would become proto-genes (Fig. 1b). Proto-genes would provide adaptive potential by exposing genetic variations that are usually hidden in non-genic sequences. A subset of proto-genes could occasionally be retained over evolutionary time, for instance if providing an advantage to the organism under specific environmental conditions. Retained proto-genes could gradually evolve the characteristics of genes, whereas other proto-genes might lose the ability to be translated. Such a reservoir of proto-genes would allow evolutionary innovations to be attempted without affecting existing genes.

This evolutionary model leads to the following predictions: (1) the structural and functional characteristics of S. cerevisiae ORFs (for example, length, expression level or sequence composition) should reflect an evolutionary continuum ranging from non-genic ORFs to genes; (2) many non-genic ORFs should be translated; and (3) ORFs that emerged recently should occasionally have adaptive functions retained by natural selection.

To examine these predictions, we estimated the order of emergence of S. cerevisiae ORFs (Fig. 1c). Annotated ORFs were classified into ten groups based on their conservation throughout the Ascomycota phylogeny (Supplementary Fig. 2). Of ~6,000 annotated ORFs, ~2% are found only in S. cerevisiae (ORFs₅) (Supplementary Fig. 2) and ~12% are found only in the four closely related Saccharomyces sensu stricto species (ORFs₁–₄). The ~88% of annotated ORFs found outside of this group (ORFs₅,₁₀) are well characterized and can confidently be considered genes. ORFs₁–₄ are poorly characterized and their annotation as genes is debatable (Supplementary Fig. 2). The weak conservation of ORFs₁–₄ suggests that they emerged recently, which we corroborated using gene duplication events to control for relative time of emergence (Supplementary Fig. 3). We estimate that over 97% of ORFs₁–₄ originated de novo rather than by cross-species transfer, which could also explain their weak conservation (Supplementary Information). ORFs₅,₁₀ often partially overlap ORFs₅,₁₀ which seems incompatible with cross-species transfer, or preferentially lie within...
subtelomeric regions whose instability may facilitate de novo emergence (Supplementary Fig. 4). In addition to classifying ORFs, we assigned a conservation level of 0 to ~108,000 unannotated ORFs longer than 30 nucleotides and free from overlap with annotated features on the same strand (ORF0) (Supplementary Information). ORFs0 and ORFs1–4 constituted our initial list of candidate proto-genes.

To test the evolutionary continuum prediction, we first verified that ORF conservation level correlates positively with length and expression level (Fig. 2a and Supplementary Fig. 5)10–12. These correlations suggest that genes evolve from non-genic ORFs that lengthen and increase in expression level over evolutionary time. A negative correlation between ORF length and expression level11 was observed among
ORFs, but not among ORFs1-4 (Supplementary Fig. 5). Thus, some ORFs may increase in expression level at different rates than they increase in length over evolutionary time. Lengthening of ORFs could occur by loss of stop codons, possibly following translational readthrough, by shift of start codons or by duplication followed by fusion with other ORFs\textsuperscript{10,22}. Increase in ORF expression level could be mediated by recruitment of existing regulatory elements\textsuperscript{1}. The proportion of ORFs located in the vicinity of transcription factor binding sites increases with conservation level, suggesting that novel regulatory elements could also emerge (Fig. 2a)\textsuperscript{1}.

In line with a study of codon evolution in metazoans\textsuperscript{23}, we observed a positive correlation between codon usage bias and conservation level (Fig. 2b). Relative abundances of amino acids in proteins encoded by ORFs\textsubscript{1-4} show levels intermediate between those in proteins encoded by ORFs\textsubscript{5-10} and in hypothetical translation products of ORFs\textsubscript{0} (Fig. 2c), similar to observations in bacteria\textsuperscript{24}. Probably owing to this biased sequence composition, ORFs\textsubscript{1-4} exhibit a higher hydropathicity, a higher tendency to form transmembrane regions and a lower propensity for intrinsic structural disorder\textsuperscript{10} than ORFs\textsubscript{5-10} (Fig. 2d). Taken together, our observations support the existence of an evolutionary continuum ranging from non-genic ORFs to genes.

To assess the extent of non-genic translation, we searched for signatures of translation of ORFs\textsubscript{0} at genome scale in a ribosome footprinting data set generated in both rich and starvation conditions\textsuperscript{26}. In this data set, \( \sim 1\% \) of sequencing reads could not be mapped to ORFs\textsubscript{5-10}. We developed a stringent pipeline to detect unequivocal translation signatures for ORFs\textsubscript{0} located on transcripts associated with ribosomes (Fig. 3a and Supplementary Fig. 6). We found that 1,139 of \( \sim 108,000 \) ORFs\textsubscript{0} show such evidence of translation (ORFs\textsubscript{0}+)\textsuperscript{1}. This number is significantly higher than expected if the ribosome footprinting assay was non-specific, or if the presence of ribosomes on non-genic transcripts was unrelated to the presence of ORFs\textsubscript{0} (Fig. 3b). These ORFs\textsubscript{0}+ are enriched in adenine at position 3 from the start codon, which probably favours translation initiation (Fig. 3c and Supplementary Information). We verified that ORFs\textsubscript{0}+ did not originate from gene duplication or cross-species transfer and are not genes that have failed to be annotated due to their short length (Supplementary Information). The 1,139 ORFs\textsubscript{0}+ therefore appear to be translated non-genic ORFs.

We detected strong differential translation of ORFs\textsubscript{0}+ and ORFs\textsubscript{1-4} in starvation or rich conditions, whereas most ORFs\textsubscript{5-10} are translated in both conditions (Fig. 3d and Supplementary Fig. 6). We found that 1,139 of \( \sim 108,000 \) ORFs\textsubscript{0} show such evidence of translation (ORFs\textsubscript{0}+)\textsuperscript{1}. This number is significantly higher than expected if the ribosome footprinting assay was non-specific, or if the presence of ribosomes on non-genic transcripts was unrelated to the presence of ORFs\textsubscript{0} (Fig. 3b). These ORFs\textsubscript{0}+ are enriched in adenine at position 3 from the start codon, which probably favours translation initiation (Fig. 3c and Supplementary Information). We verified that ORFs\textsubscript{0}+ did not originate from gene duplication or cross-species transfer and are not genes that have failed to be annotated due to their short length (Supplementary Information). The 1,139 ORFs\textsubscript{0}+ therefore appear to be translated non-genic ORFs.

Retention by natural selection was measured by comparing the genome sequences of eight \textit{Saccharomyces} species (\textit{S. cerevisiae}, \textit{S. paradoxus}, \textit{S. mikatae}, \textit{S. bayanus}, \textit{S. cerevisiae Rbi}, \textit{S. paradoxus Rbi}, \textit{S. mikatae Rbi}, \textit{S. bayanus Rbi})\textsuperscript{29,30}. Asterisks mark significant differences between ORFs with and without translation signatures. 5 from the start codon indicates optimum AUG context (Supplementary Information). \( \tau \), Kendall’s correlation statistics. Asterisks mark significant differences between ORFs with and without translation signatures (\( p < 0.05 \), Fisher’s exact test). d, Candidate proto-genomes tend to undergo condition-specific translation. e, Signatures of intra-species purifying selection. The positive correlation (Supplementary Table 4) holds when only considering ORFs that are free from overlap with ORFs\textsubscript{5-10} (Supplementary Fig. 7), and is not entirely driven by the interdependence between strength of purifying selection and expression level (Supplementary Information)\textsuperscript{10,26}. Error bars represent standard error of the proportion in all panels.

nucleotides, show signatures of translation and are under purifying selection, can confidently be considered genes despite being weakly conserved (Fig. 4a and Supplementary Fig. 8). The remaining 1,891
ORFs (1,139 ORFs$^+$ and 752 ORFs$^-\$) present characteristics intermediate between non-genic ORFs and genes, meeting our proto-gene designation (Supplementary Table 3). We propose to place these ORFs in a continuum where strict annotation boundaries no longer have to be set (Fig. 4b).

Gene birth mechanisms involving re-organization of pre-existing genes, notably following gene duplication, have long been regarded as the predominant source of evolutionary innovation$^{29,30}$. Since the split between S. cerevisiae and S. paradoxus, sporadic gene duplications have generated between one and five novel genes$^{31}$. In contrast, 19 of the 143 ORFs that arose de novo during the same evolutionary period were found under purifying selection. Therefore, de novo gene birth seems to be more prevalent than previously supposed$^{30,31}$, in agreement with recent estimations in humans and other primates$^{32}$. The involvement of proto-genes in de novo emergence of protein-coding genes in S. cerevisiae probably holds for other species and may extend to RNA genes and regulatory elements. Examination of translation program remodelling upon stress, in light of our evolutionary model, may further understanding of phenotypic diversity and plasticity of cellular systems$^{29}$.

METHODS SUMMARY

Detection of translation signatures. The mapping of ribosome footprint reads to ORFs does not necessarily indicate full-length, ORF-specific translation events$^{6,25}$. To model the number of ORFs$^+$ expected if the detected presence of ribosomes on non-genic sequences was not related to the presence of ORFs$^+$, we randomized the positions of ORFs$^+$ while maintaining their length distribution and the observed positions of RNA-Seq and footprint reads. To model the number of ORFs$^-$ expected if footprint reads observed outside of annotated ORFs were non-specific, we randomized the positions of footprint reads throughout non-genic sequences while maintaining the length distribution of footprint reads, the positions of RNA-Seq reads and the positions of ORFs$^+$. We optimized three parameters with regard to these two null models: (1) the proportion of ORF length covered in RNA-Seq and footprint reads was fixed at 50% minimum; (2) the factor by which the number of footprint reads per nucleotide in the ORF should be higher than the number of footprint reads per nucleotide in surrounding up- and downstream windows was fixed at a minimum of 5; and (3) the size of these windows was fixed at 300 nucleotides. Any two ORFs that partially overlap on the same strand and show translation signatures in the same experimental conditions were both eliminated from the set of ORFs$^+$ considered to show translation signatures.

Significant purifying selection signatures. We estimated the number of synonymous mutations per synonymous site (dS) and the number of non-synonymous mutations per non-synonymous site (dN) for each ORF present without disruptive mutations in eight S. cerevisiae strains. The likelihood of the dN/dS ratio was determined under two distinct null models: assuming neutral evolution (the rates of synonymous and non-synonymous substitutions are equal) and not assuming neutral evolution. All ORFs with dN/dS < 1 and P < 0.05 (chi-squared distribution of likelihoods with one degree of freedom) were considered to be subject to significant purifying selection.

Received 2 November 2011; accepted 8 May 2012.

Published online 24 June 2012.


Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Duret, E. Levy, J. Vandenhaute, Q. Li, H. Yu, P. Braun, M. Dreze, C. Foo, M. Mann, N. Kulak, J. Cox, C. Maire and S. Jhavery-Schneider as well as members of the Center for Cancer Systems Biology (CCSB), in particular A. Dricot-Zilfier, A. MacWilliams, F. Roth, Y. Jacob and D. Hill for discussions and proofreading. A.R. was supported by a National Institute of Health Pioneer Award, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and the Howard Hughes Medical Institute (HHMI). I.W. is a HHMI fellow of the Damon Runyon Cancer Research Institute. G.A.B. was supported by American Cancer Society Postdoctoral fellowship 117945-PF-09-136-01-RMC. M.V. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, Wallonia-Brussels Federation, Belgium). This work was supported by the grant R01-HG006561 from the National Human Genome Research Institute awarded to M.V.

Author Contributions A.-R.C., I.W., M.E.C. and M.V. conceived the project. A.-R.C. led the project and performed most of the analyses. T.R. evaluated cross-species transfer events, optimized the ribosome footprint analysis pipeline and assisted in other analyses. I.W. designed the conservation level tool and calculated most of the purifying selection statistics. M.A.Y. aligned the sequencing reads. B.S. predicted disordered and transmembrane regions and assisted in the cross-species transfer analyses. N.S. and B.C. assisted in analyses. G.A.B. and J.S.W. shared their expertise in ribosome footprinting data analysis and provided the meiosis ribosome footprinting raw and processed data. A.-R.C., T.R., M.E.C. and M.V. designed the figures. All authors contributed to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.V. (marc_vidal@dfci.harvard.edu).