Genome Biology Day 2025

De novo emerged genes in Archaea

Eliott Tempez*, Violette Da Cunha**, Patrick Forterre***, Anne Lopes*

Institute for Integrative Biology of the Cell (I2BC)
Molecular Bio-informatics team

Abstract

De novo gene birth, which refers to the emergence of novel genes from ancestrally non-coding DNA, has been demonstrated to significantly contribute to genome evolution in various eukaryotic species, suggesting a widespread phenomenon in Eukaryotes. To date, no studies have reported de novo genes in Archaea, which are nevertheless associated with many taxonomically restricted genes, and the question as to whether Archaea can evolve through de novo emergence remains open. We acquired the assembled genomes of 116 Thermococcacaea, and we confirmed their status as emerged from a non-coding sequence for 64 of their genes. We further analysed these genes, discussing their evolutionary age, origin, as well as sequence and structure properties. Our findings provide the first evidence that this process is active and potentially underappreciated in archaeal evolution.

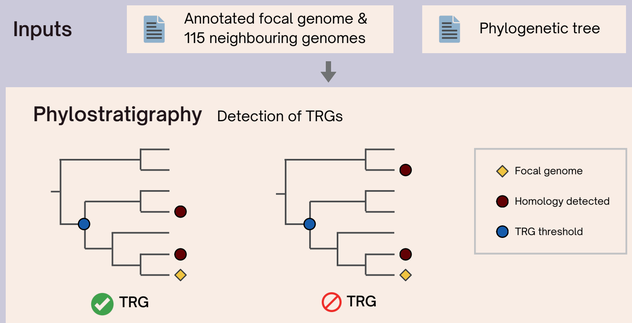
Methods

Data

We obtained the assembled genomes and associated phylogenetic tree of 116 Thermococcacaea, a family of Archaea. Our dataset comprises 2 genus and 67 distinct species (according to the ANI, Average Nucleotide Identity).

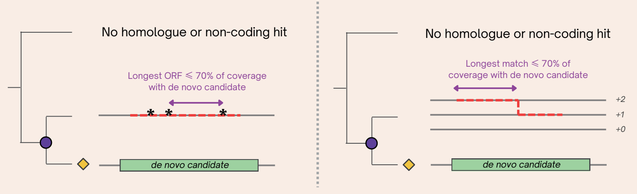
DENSE[7]

DENSE is a Nextflow pipeline developed at the I2BC, that automates the detection of de novo genes. It is based on two distinct steps: detection of taxonomically restricted genes (TRGs) through phylostratigraphy, and filtering of TRGs for de novo emerged genes via genome comparisons and synteny search.



Integrity analysis

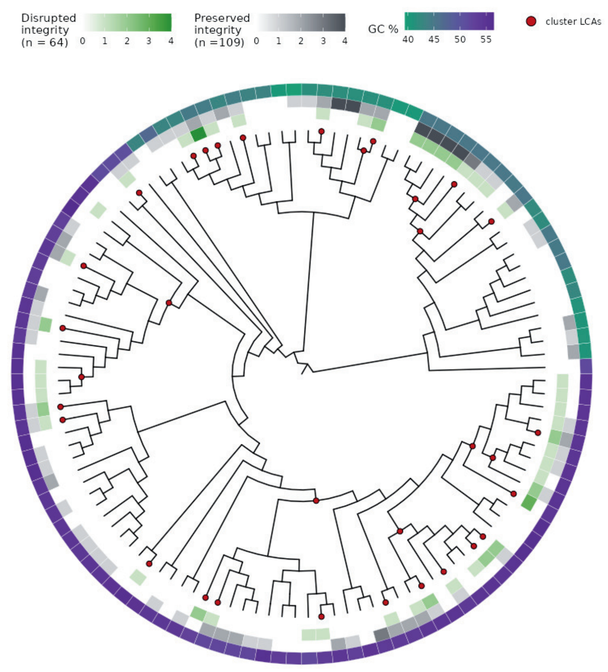
To set aside potential annotation errors that would make us consider canonical genes as non-coding and create false positives, we only took into account de novo genes that had a non-coding match with a disrupted integrity.



Examples of non-coding matches with a disrupted integrity, that would be considered as true positive de novo genes, and selected for the the analysis.

Introduction

Archaea make up one of the three domains of life and are known for thriving in extreme environments, yet the mechanisms behind their adaptability remain poorly understood[1-2]. De novo genes -genes emerging from previously non-coding DNA- have been shown to contribute to genome evolution in Eukaryotes[3-4], but have not yet been studied in Archaea. Though compact genomes and limited intergenic regions in prokaryotes[5] made this phenomenon seem unlikely, recent findings in Bacteria suggest otherwise[6]. Here, using a newly available archaeal dataset and state-of-the-art de novo detection tool[7], we investigate whether de novo gene emergence occurs in Archaea.



de novo genes do exist in Archaea!

Fig. 1: Phylogenetic tree of the 116 genomes (branch lengths not included) with, in order of the heatmaps from the inside to the outside: the number of de novo genes with unpreserved integrity, i.e. that we took into account for the rest of the analysis (n = 64), with preserved integrity (n = 109), and the GC content. The LCAs (last common ancestors) for each cluster are represented with red dots.

Results

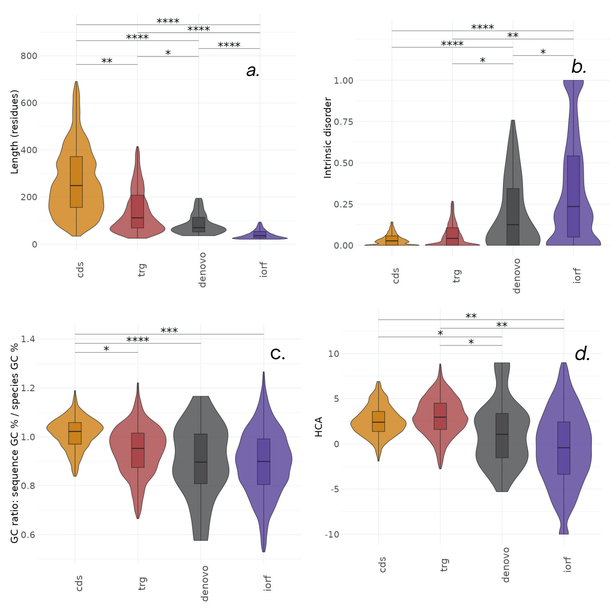
We identified 173 de novo genes for all of the 116 genomes, among which 64 with high confidence (see Methods - integrity analysis). We clustered those genes using MMseqs2, and found 40 clusters, corresponding to 40 independent events of de novo gene birth in Thermococcacaea; fig. 1 represents the last common ancestors (LCAs) of each de novo gene family. Most of these clusters were either comprised of one orphan gene, or had several genes in only one species (according to the ANI) (fig. 2a). There were nonetheless four clusters with several genes across different species (fig. 2b to 2d), suggesting that although de novo genes are known to have a high death rate, some become fixed.

Most de novo genes are recent and associated with high turnover.

Fig. 2 Phylogenetic tree of the 116 genomes (branch lengths not included) for four of the de novo gene clusters. The red dot indicates the LCA of the de novo gene family, the green dots are the genomes in which the de novo gene is present, and the grey dots are the genomes under the LCA for which the gene is absent. Each differently coloured rectangle is a different species, based on the ANI.

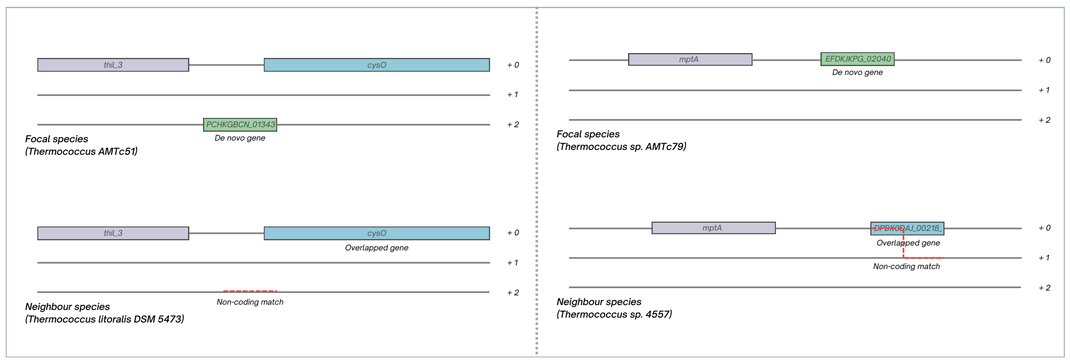
We compared the de novo genes with other TRGs, other CDSs (Coding Sequences), as well as iORFs (Intergenic Open Reading Frames) using sequence and structure properties that have been shown to have significance in de novo gene emergence[8]. For most of those, de novo genes are at the interface between iORFs and TRGs, which may reflect their transitional nature, as they emerged from non-coding sequences and have not yet undergone thesame level of selection and optimisation as older CDSs (fig. 3a-3b). For other descriptors such as the GC content and the hydrophobic cluster analysis, de novo genes were closer to iORFs, denoting their similarity to non-coding sequences (fig. 3c-3d).

We extracted the position of the non-coding match for each de novo gene. Out of the 64, 49 were intergenic; it seems to be the preferred origin of gene emergence, even though prokaryotic genomes are quite compact and their intergenic size small. It is interesting to note though, that the 15 other non-coding matches do overlap an existing gene, and most of the time they overlap an intergenic segment as well. For those overlapped genes, most of them were still present in the focal genome, where the de novo gene was detected (fig. 4). The rest of the overlapped genes showed no homologue gene in the focal genome though. This raises the question of the impact of the type of overlapped gene, as well of its mere presence, on the fixation of the de novo gene.



de novo genes display intermediate properties between intergenic ORFs and established CDSs.

*Fig. 3: distribution of the sequence length, intrinsic disorder score, GC content, , and hydrophobic cluster analysis (HCA) for the different types of ORFs. ****: $p \leqslant 10^{-5}$ ***: $p \leqslant 10^{-4}$ **: $p \leqslant 10^{-3}$ *: $p \leqslant 0.05$*



de novo genes arise from different genomic origins.

Fig. 4: Representation of the configuration of two different de novo genes. At the top, we can see the immediate surroundings of the de novo gene, in green. At the bottom, the match in the neighbour species is shown with red dashes. Genes of the same colour and same name are homologues. The overlapped genes are pictured in blue. On the left subfigure, the de novo gene emerged from a hybrid sequence between an intergenic segment and the +2 frame of an existing gene. On the right, the de novo gene emerged from a frameshift within an existing gene.

References

[1] Brett J. Baker et al. “Diversity, ecology and evolution of Archaea”. In: *Nature Microbiology* 5.7 (July 2020). Publisher: Nature Publishing Group, pp. 887–900. ISSN: 2058-5276. DOI: 10.1038/s41564-020-0715-z. URL: https://www.nature.com/articles/s41564- 020-0715-z

[2] Elena V. Pikuta, Richard B. Hoover, and Jane Tang. “Microbial extremophiles at the limits of life”. In: *Critical Reviews in Microbiology* 33.3 (2007), pp. 183–209. ISSN: 1040-841X. DOI: 10.1080/10408410701451948.

[3] Stephen Branden Van Oss and Anne-Ruxandra Carvunis. “De novo gene birth”. In: *PLoS Genetics* 15.5 (May 23, 2019), e1008160. ISSN: 1553-7390. DOI: 10 . 1371 / journal . pgen.1008160. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6542195/

[4] Anne-Ruxandra Carvunis et al. “Proto-genes and de novo gene birth”. In: *Nature* 487.7407 (July 19, 2012), pp. 370–374. ISSN: 1476-4687. DOI: 10.1038/nature11184

[5] Eugene V. Koonin and Yuri I. Wolf. “Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world”. In: *Nucleic Acids Research* 36.21 (Dec. 1, 2008), pp. 6688–6719. ISSN: 0305-1048. DOI: 10 . 1093 / nar / gkn668. URL: https://doi.org/10.1093/nar/gkn668

[6] Igor Fesenko et al. “The hidden bacterial microproteome”. In: *Molecular Cell* 85.5 (Mar. 6, 2025), 1024–1041.e6. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2025.01.025.

[7] Paul Roginski et al. “De Novo Emerged Gene Search in Eukaryotes with DENSE”. In: *Genome Biology and Evolution* 16.8 (Aug. 1, 2024), evae159. ISSN: 1759-6653. DOI: 10.1093/gbe/evae159. URL: https://doi.org/10.1093/gbe/evae159

[8] Chris Papadopoulos et al. “Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution”. In: Genome Research 31.12 (Dec. 2021), pp. 2303–2315. ISSN: 1549-5469. DOI: 10.1101/gr.275638.121

Poster created for the Genome Biology Day 3rd July 2025

Affiliations

* Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, Gif-sur-Yvette, Cedex, France
** Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
*** Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de Microbiologie, Institut Pasteur, Paris, France, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, Gif-sur-Yvette, Cedex, France