Université Paris Cité Year 2024-2025

M2 Bioinformatique

De novo emerged genes in Archaea

Eliott TEMPEZ

Internship supervisor

Anne LOPES

Institute for Integrative Biology of the Cell

Molecular Bioinformatics Team

Contents

1 Introduction

Archaea, with Eukaryotes and Eubacteria, make up one of the three domains of life. They take part in numerous processes on earth, such as the carbon and nitrogen cycles [1], or the animal digestive system [2]. They are also well-known for colonising extreme environments, such as those with high salt concentrations, extreme temperatures, or very acidic or basic pH [3]. The mechanisms by which they adapt to extreme environments are however still poorly understood.

Taxonomically restricted genes (TRGs) and orphan genes, meaning genes that lack homologues outside their own taxonomic group or outside their own species, are involved in important evolutionary processes, and make up 10%-20% of genes in every taxonomic group [4]. They can originate from a few different genetic events, namely gene duplications, fusions, fissions, or horizontal gene transfers (HGTs), after which the TRG would have evolved differently from the ancestor gene, losing its homology signal with it and becoming restrained to the phylogenetic group the event happened in. Another possible TRG origin, though, is de novo gene birth, meaning it evolved from non-coding DNA sequences [5]. Indeed, the de novo genes we are able to detect are those that still have traces of homologue sequences in neighbour genomes, meaning they have emerged recently, and meaning they are generally TRGs or orphan genes.

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. Initially considered unlikely due to the stark differences between coding and non-coding regions, de novo genes have garnered increasing attention in the past decade to elucidate their properties, functions, and mechanisms of emergence. Specifically, it has been shown that these genes display peculiar characteristics: they are associated with unique amino-acid compositions, are expressed in specific tissues or under particular conditions, and are longer than random ORFs, while shorter than canonical genes [5–8]. Furthermore, many were shown to undergo purifying selection and biological functions have been experimentally confirmed for some of them [9, 10]. These observations clearly demonstrate that, despite their distinct features compared to canonical genes, these newly emerged genes can be confidently classified as coding. However, all of these studies

were carried out in eukaryotic species.

To date, no studies have reported de novo genes in Archaea, which are nevertheless associated with many TRGs, including orphan genes, and the question as to whether Archaea can evolve through de novo emergence remains open. Actually, due to the fact that a large fraction of identified de novo genes has been shown to result from intergenic sequences, and that prokaryotes have very compact genomes with very small intergenic regions [11], de novo gene emergence in prokaryotic genomes has long been viewed as unlikely. Yet, recently, a study published by the group of Eugene Koonin reported the existence of thousands of microproteins that likely originated de novo in 5,668 bacterial genomes of the Enterobacteriaceae family [12]. Furthermore, overprinting refers to a special case of de novo gene formation, in which the non-coding origin of the emerged gene isn't intergenic, but comes from alternative reading frames of existing genes. Even if it straddles a gene, this origin is non-coding in the sense that it is not translated in the same way the coding sequence is [5]. This mechanism has been shown to drive gene emergence in viruses and *Escherichia coli* [13, 14], thus it could also be at work in Archaea.

De novo gene identification is often carried out using two methods, most of the time combined. The first method (Fig. 1A) is genomic phylostratigraphy, which aims to detect the TRGs in a genome, to consider as candidates for de novo genes. For each gene in a species of interest, the principle consists of identifying homologues in the tree of life with the use of sequence alignment tools such as Blast [15] and inferring the most likely node of gene emergence, based on its presence/absence profile across lineages. Any gene lacking homologues outside its own taxonomic group will be labelled as a TRG (Fig. 1A).

The second method (Fig. 1B) aims to detect among the identified TRGs, those resulting from an ancestor that was non-coding. It consists of searching homology traces of the de novo candidate in the non-coding regions of the species where the gene is absent. Such evolutionary patterns provide support for a hypothesis of a non-coding ancestor, and the subsequent de novo emergence of the gene family in the ancestor of the species where the gene has been found. Non-coding regions evolve fast, which may affect the detection of homology traces between the candidate and the non-coding regions, that is why methods generally impose the non-coding match to be identified in

synteny (i.e., in relative order of genetic elements enclosing the de novo gene candidate) with the de novo gene candidate, thereby increasing the confidence in the detection of the non-coding match. (Fig. 1B) [5].

Based on these methods, the main constraint in order to identify de novo genes in a not yet studied taxonomic group is to have a high-quality genomic dataset, with neighbours close enough to pick up a homology signal between the candidate gene and its homologue non-coding matches in the neighbour species. As we acquired such a dataset, and with regards to a recently-developed de novo detection tool that allows for the simplification of a once-tedious analysis [16], we offer to answer this question, left unanswered for years: can we identify de novo genes in Archaea?

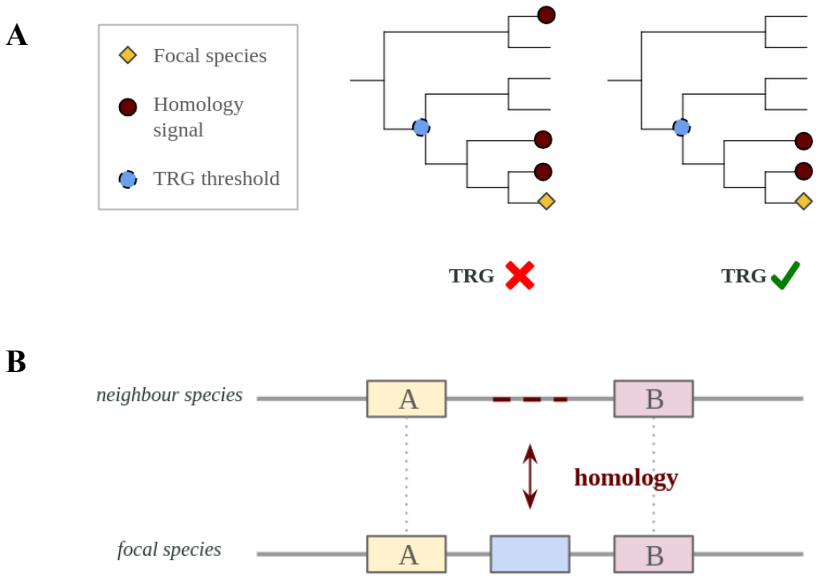


*Figure 1: **A** - Genomic phylostratigraphy: a TRG threshold is chosen, and each gene lacking homologues above this threshold is considered a TRG. **B** - Synteny: The de novo candidate (in blue) is surrounded by the genes A and B, and its non-coding match as well. If it is also a TRG, it will be considered as de novo. Here, the surrounding matches are pictured as in the same direction in both neighbours, but they could also be reversed (ie. B then A in the neighbour). Here, the non-coding match is pictured as intergenic, but it could also be an alternative frame of an existing gene.*

2 Material and methods

2.1 Dataset

The dataset contains the assembled genomes of 116 Thermococcacaea, an Archaea family, and was obtained by our collaborators Violette Da Cunha and Patrick Forterre. It is composed of two different genus, Pyrococcus and Thermococcus. Using the genome average nucleotide identity (ANI) (Supp Fig. 1), and considering that a value above 95 indicates that they belong to the same species, there are 68 different species in our dataset (Supp Fig. 2). Throughout this report though, we will use the word ”species” refering to a single organism or single genome, in an effort to simplify the language.

Gene annotation

Initially, about half of the genomes in the dataset were annotated with the NCBI Prokaryotic genome annotation pipeline [17], and the other half with the MicroScope annotation platform [18]. These two annotation tools having different criteria when considering genes, this heterogeneity would have led to potential homologous genes being considered as genes in certain genomes, and as non-coding in other, creating false positives as well as entirely missing some de novo genes. Thus, we decided to re-annotate all genomes using Prokka [19] version 1.14.6, modifying the hardcoded minimum gene length from 90 nucleotides to 75.

2.2 De novo genes detection

The de novo genes were detected via the DENSE tool, an automated pipeline that relies on both phylostratigraphy and synteny searches in order to infer which genes among a genome have emerged as de novo [16].

The workflow takes as input the genome sequence and annotation of the focal species, as well as the neighbour genomes’. It starts by screening the focal genome’s annotated coding sequences (CDSs) against the nr (non-redundant) database, as well as all the CDSs of all neighbour genomes. This is

done using GenEra [20], a DIAMOND-fuelled [21] genomic phylostratigraphy tool, which returns the estimated ages of each CDS in the focal genome, from which DENSE extracts the TRGs. The TRG threshold (Fig. 1A) given as input was at the family level Thermococcaceae.

DENSE then conducts a homology search of each focal species TRG against all non-coding (intergenic and alternative frames) areas in all neighbour genomes, using tblastn [15]. We chose the strategy n°1 with two outgroups, meaning it mandates at least two outgroup non-coding hits combined with a search of synteny (Fig. 1B) in order to consider a gene as de novo (Fig. 2). Imposing two outgroup species further supports the non-coding status of the ancestor X (Fig. 2) by verifying the non-coding status of an outgroup species (outgroup n°2) with respect to this ancestor.

This step is crucial for de novo detection; the goal is to be certain that our non-coding match can be considered as such, i.e. that it shares a common ancestor sequence with the de novo gene candidate. For this, the query coverage of the match relative to the de novo CDS -returned by tblastn- must be over 50%. As for the genomic context around the detected non-coding match and the candidate, we require that at least two genes among a window of eight genes surrounding the de novo candidate also enclose the non-coding match in synteny.

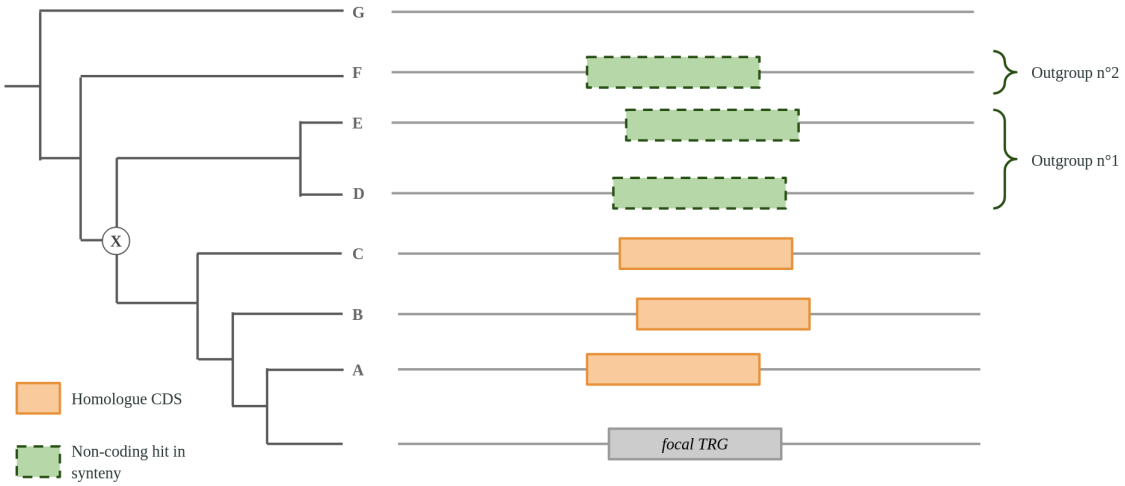


Figure 2: Example or a valid de novo gene for DENSE: the focal TRG does have homologue CDSs in neighbor genomes (A, B, C, in orange), but these are closer to the focal than the non-coding (NC) syntenic matches are (D, E, F, in green). As there are no further-placed CDSs (no orange above F), we conclude that the sequence was also non-coding in the last common ancestor (LCA) of all the organisms for which we found a homology signal (Focal + A through F included). Thus, all three NC matches in synteny (in green) are parts of valid outgroups. Here, there are two distinct outgroups, because the species D and E are at the same phylogenetic distance from the focal species, and as such, are part of the same outgroup.

Integrity analysis

Annotation errors can have major consequences for de novo gene prediction. Specifically, genes that are erroneously annotated as non-coding may lead to the incorrect conclusion that a given genome lacks the gene. This mis-annotation can result in tblastn matches with a presumed non-coding sequence in this species, when in fact the match involves a protein-coding gene. To address this issue, it is critical to verify that regions identified as non-coding truly lack coding potential.

A conservative approach involves considering as non-coding only those matches in which the open reading frame (ORF) of the de novo candidate is disrupted, thereby supporting the absence of a functional coding sequence in the species where the non-coding match was detected. Therefore, to prevent false identification of de novo genes due to annotation errors, we systematically assessed ORF integrity in regions associated with presumed non-coding matches to verify that the presumed non-coding match does involve a non-coding region and not a missed protein-coding gene.

For this, we used the DENSE output; the two furthest outgroups from the genome of interest in the phylogenetic tree with a non-coding syntenic match, that we will simply call outgroups (for an example, see outgroups n°1 and n°2 in Fig. 2), were extracted from the match matrix file. Then, for each species with a syntenic non-coding match in each of these outgroups (in our example Fig. 2, species D, E, and F), we conducted the integrity analysis. Only de novo genes with a disrupted integrity in more than 80% of the considered species in each of the two outgroups were kept for further analysis.

In order to carry out this integrity analysis, the query coverage was extracted from the tblastn output file for each of the species in the two outgroups. For each species, if the query coverage was > 70%, we then computed the length of the longest ORF in the non-coding match, and calculated its query coverage. If it was still > 70%, the de novo gene was discarded as with a preserved integrity, and thus a possibility of annotation error. (Fig. 3).

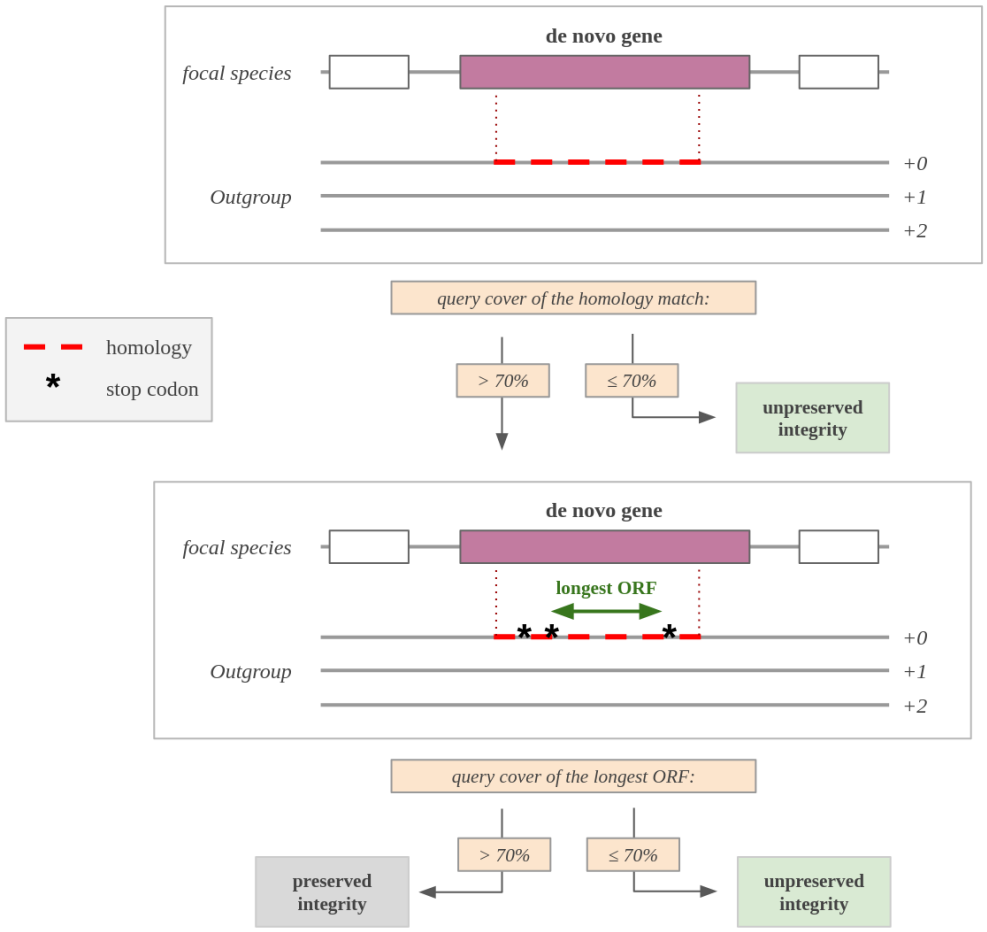


Figure 3: Integrity analysis pipeline. First are selected the de novo genes for which the non-coding match has a query cover $\leq$ 70%. Then, for the other de novo genes, the query coverage of the longest open reading frame (ORF) is computed. The de novo genes with this query coverage $\leq$ 70% are selected too.

Gene clustering

The gene families were computed using MMseqs2 [22] version 17, with the default parameters: the cascaded clustering algorithm, a minimum sequence identity threshold of 50%, and at least 80% coverage of the target sequence. The last common ancestors were extracted with the R library ape [23], version 5.8-1.

2.3 De novo genes characterisation

Descriptors

The intergenic open reading frames (iORFs) were extracted using ORFtrack, a tool developped in the laboratory to extract ORFs within a genome based on its genomic sequence and annotation. iORFs are defined from STOP-to-STOP with at least 60 nucleotides, and they must not cover any other annotated feature (e.g., Coding Sequences (CDSs), tRNA, rRNA, pseudogene etc) by more than 70 percent [24].

All TRGs for each genome were extracted from the DENSE output, and were divided into two subgroups: those that emerged as de novo, and the others. In a similar fashion, only the CDSs that were not TRGs were included in the "CDS" group. In summary, we computed our descriptors for four types of ORFs, and each individual ORF could only be present in one of the four groups, labelled as: iORFs, de novo genes or their associated de novo ORFs when referring to protein properties, TRGs, or CDSs.

A set of descriptors was computed for each of these ORFs: their length, GC rate, amino-acid composition, aromaticity, instability index, mean flexibility, and hydropathy were calculated with Biopython [25] version 1.78. The aromaticity is the relative frequency of Phenylalanine, Tryptophan, and Tyrosine [26]. The instability index is computed depending on the presence of certain dipeptides, associated with unstable proteins [27]. The flexibility of each residue is estimated with a sliding window averaging technique, and the mean for the whole protein is then calculated [28]. The hydropathy level corresponds to the grand average of hydropathy [29].

The foldability of the different ORF categories was assessed with the hydrophobic cluster analysis using pyHCA [30]. Their disorder score and aggregation score were computed using IUPred2A [31], and Tango respectively, [32], via the ORFmine pipeline [24].

Statistical analysis

All statistical analysis were performed in Python version 3.9.19. In order to calculate the p-values with large samples and taking sample imbalance into account, we used an iterative strategy (see

[33]). We randomly sampled n = 100,000 times the CDSs, TRGs, and iORFs with a sample size matching the smallest group, the number of de novo genes.

For each pairwise comparison and each descriptor, we computed the value of the median difference for the two groups D_{obs}. We then divided the pooled sample into two same-sized random subsamples, calculated the difference D_{samp} between sample medians, and derived the p-value as the proportion of D_{samp} values exceeding D_{obs}. The test being one-sided, in order to know which group median to subtract from the other, we pre-emptively calculated the global group medians, and subtracted the group with the lowest median from the one with the highest.

Plots

Plots were created using R version 4.4.3 and ggplot2 [34] version 3.5.1. All plots involving the phylogenetic tree were computed using ggtree [34] version 3.14.0. For better readability, all boxplots and violin plots were created by sampling 5000 values when there were more than 5000, and without representing outliers.

Non-coding origin

In order to infer the non-coding origin of the de novo genes, we looked at the non-coding tblastn match of each de novo candidate in the furthest outgroup from the focal genome, to see whether the match was 100% intergenic or if it straddled an existing gene. If the match straddled a single existing gene for less than nine nucleotides though, we discarded it as intergenic. In order to visualise the genes of interest, we used the Integrative Genomic Viewer (IGV) [35].

3 Results

3.1 de novo genes detection

We applied DENSE to the 116 archaeal genomes. This analysis led to the identification of 173 de novo genes, distributed across nearly all genomes. The number of de novo genes per genome ranged

from zero to six, which is significantly lower than the amounts typically reported in eukaryotic species, where dozens of de novo genes are often identified [5, 16] (Fig. 4).

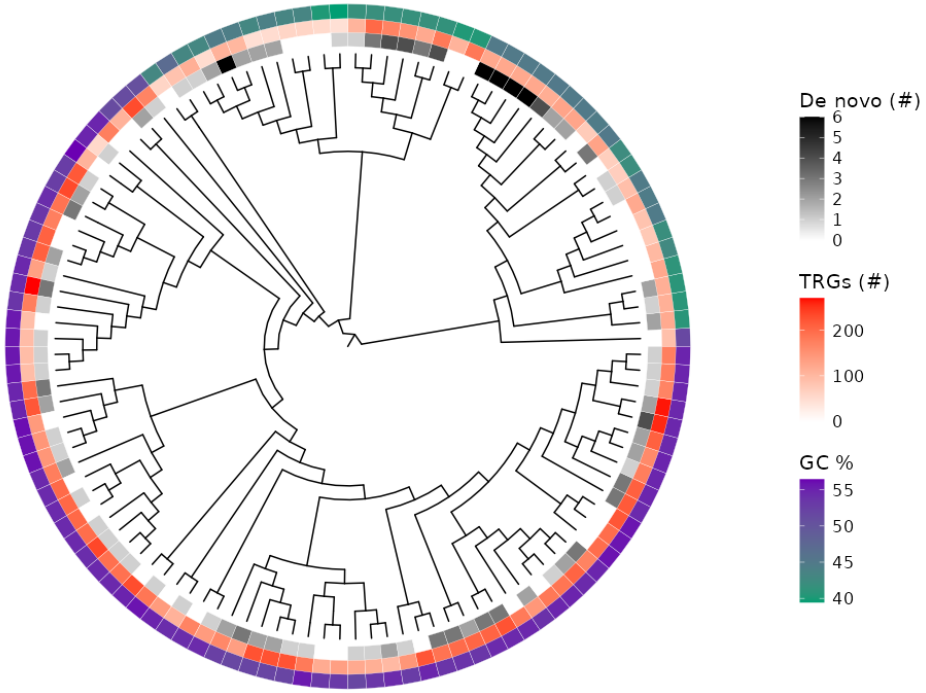


Figure 4: Phylogenetic tree of the 116 genomes (branch lengths not included) with, in order of the heatmaps from the inside to the outside : the number of de novo genes, the number of TRGs, and the GC content per organism.

De novo gene detection being inferred from blast matches in a non-coding portion of the neighbour genomes, in order to make sure these non-coding matches were not the result of an annotation error, we assessed the integrity of the non-coding matches of the two furthest outgroups from the genome of interest (Fig. 3). We -conservatively- decided to keep for further analysis only the de novo candidate genes that had non-coding matches with a disrupted integrity (i.e., either significantly shorter than the candidate de novo ORF due to STOP codons, or disrupted due to insertion/deletion events leading to frameshifts).

For each de novo candidate, we extracted the ones for which at least 80% of the identified non-coding matches in both outgroups (we recall that an outgroup may contain several species - see outgroup n°1 in Fig. 2) had a disrupted integrity. This supports the hypothesis that the matched regions in these outgroups do not correspond to protein-coding genes that were simply missed by the annotation process. Out of the 173, there were 88 de novo genes that met these standards in the

furthest outgroup, and 74 in the penultimate outgroup. As we demanded at least 80% of non-coding matches with a disrupted integrity in both outgroups and not either one or the other, we found that 64 de novo genes out of the 173 fulfilled the criteria (Fig. 5).

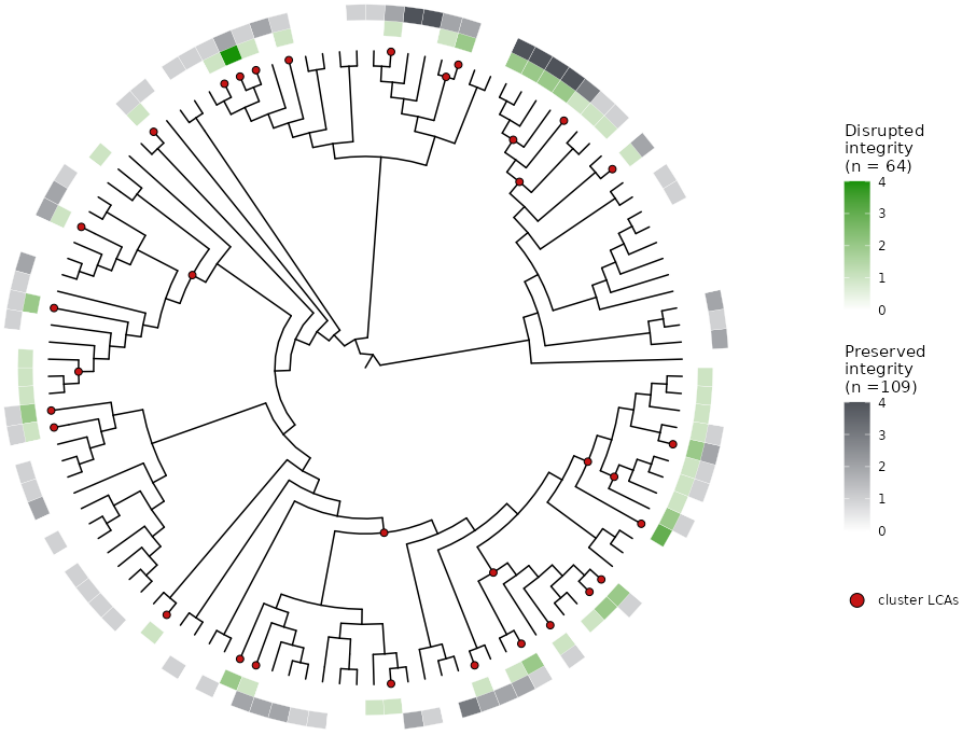


Figure 5: Phylogenetic tree of the 116 genomes (branch lengths not included) with, in order of the heatmaps from the inside to the outside: the number of de novo genes with unpreserved integrity (n = 64) and with preserved integrity (n = 109). The LCAs (last common ancestors) for each cluster are represented with red dots.

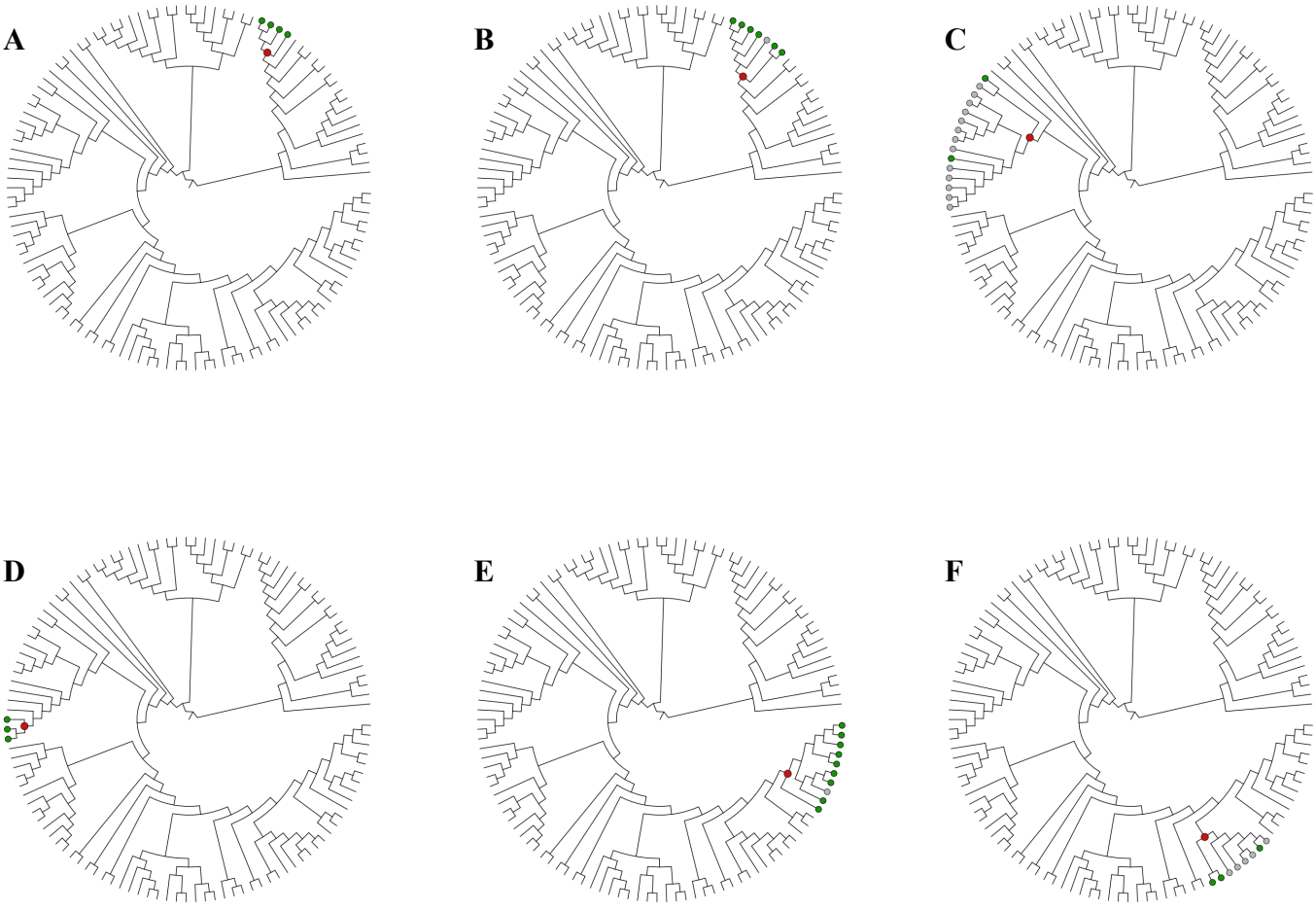
Clustering

The 64 de novo genes were then clustered using MMseqs into 40 different gene families. For each of these clusters, we calculated their associated last common ancestor (LCA) (Fig. 5). 31 out of 40 LCAs were at the leaf level, meaning that there was only one de novo gene making up the cluster. This shows the relatively young age of these 31 genes, and of most de novo genes. This is partly due to the way de novo identification is conducted, i.e. relying on homology detection techniques, meaning that the de novo genes we are able to detect are those that still resemble their ancestor non-coding sequences; as non-coding sequence (and particularly intergenic segments) evolve fast,

that signal is quickly lost. Another explanation is that de novo genes undergo a rapid turnover; as they die quickly, young de novo genes do not necessarily have the opportunity to become older [36].

As for the nine clusters that contained more than one de novo gene, seven of them held 10 species or less under their LCA, among which the de novo gene was present in most of the species (between 38% and 100%, 38% (Fig. 6F) being the exception with the median being 90%).

The two clusters with the most species under their LCA were also the ones with the less representativeness when considering the de novo presence (Fig. 6C and Fig. 6I): two out of 15 and two out of 37 species respectively held the de novo gene. This once again shows the fast evolution of de novo genes; the more the gene is ancient, the less it can be found in neighbour genomes, highlighting its high death rate.



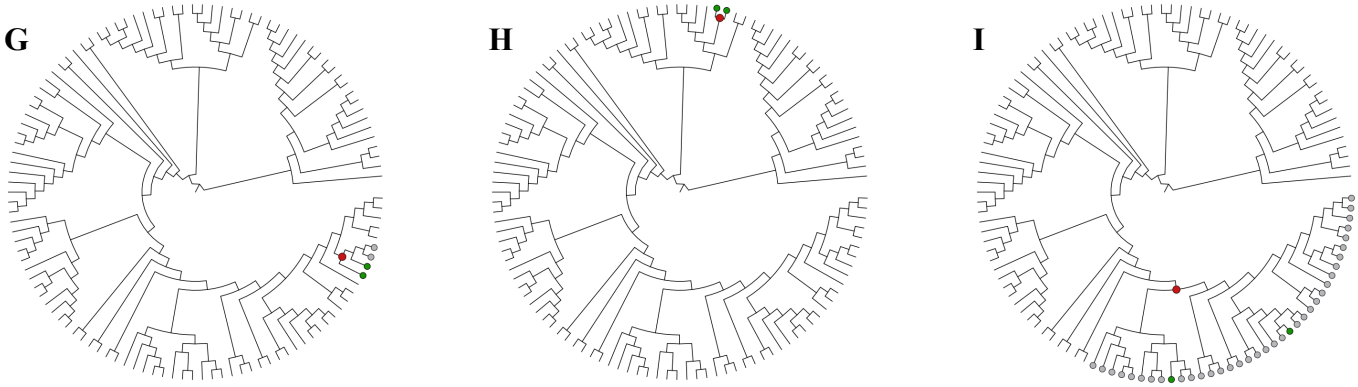


Figure 6: Representation of the clusters containing more than one de novo gene. The phylogenetic tree is represented for each of the nine clusters, with the cluster LCA in red, the species in which the de novo gene is present in green, and the species under the LCA where the gene is absent in grey.

3.2 De novo genes characterisation

De novo genes have been shown to display specific characteristics when considering their sequence and structure properties, that are often at the interface between those of canonical CDSs and intergenic ORFs (iORFs) [5–7]. In order to evaluate if archean de novo genes follow the same patterns, we characterised our 64 de novo candidates by using a set of sequence and structure descriptors.

De novo genes in our dataset are typically shorter than the other TRGs and CDSs (Fig. 7A), which is consistent with the literature on the subject. It is worth noting, though, that the significant size difference between de novo genes and iORFs (as well as between iORFs and other CDSs) here are partly due to their respective detection parameters; all iORFs above 60 nucleotides-long are considered, whereas the length threshold for CDS annotation is 75 nucleotides.

The 116 genomes having different overall GC contents (Fig. 4), we estimated the GC rate: the ratio of the ORF GC content in relation to the global genome GC content. In our dataset, iORFs exhibit significantly lower GC rates than canonical CDSs (Fig. 7B), which was expected as CDSs generally exhibit higher GC rates, reflecting selective pressures associated with canonical translation. De novo GC rates were comparable to those of iORFs and TRGs, which may reflect the transitional

nature of de novo genes, which emerged from non-coding sequences and have not yet undergone the same level of selection and optimisation as older CDSs.

In regards to their amino-acid composition, de novo genes have more polar residues than TRGs but less than iORFs (Fig. 7C), and they have a similar amount of hydrophobic (Fig. 7D), as well as negatively (Fig. 7E) and positively (Fig. 7F) charged residues as iORFs. There however seems to be a slight trend for those last three, where the de novo distributions seem to be situated between the iORFs and TRGs distributions. These patterns further support the idea that de novo genes occupy an intermediate space between non-coding and established coding sequences.

Similar observations can be made regarding the de novo's folding potential; it is lower than the folding potential of CDSs and TRGs, and similar to those of iORFs, but a slight trend can be noticed where it seems to be at the interface of iORFs and TRGs (Fig. 7G). The folding potential refers to how likely a protein is to form a stable 3D structure based on its hydrophobic residue patterns; higher foldability suggests the protein can adopt a functional structure, and it has been shown to play an important role in de novo emergence [7, 37]. For the disorder propensity, we observe a statistical difference between the IUPRED scores of both de novo genes and TRGs, and de novo genes and iORFs (Fig. 7H). For both those descriptors, once again, our observations still place de novo genes as transitional ORFs between iORFs and other CDSs.

For the aggregation score, surprisingly, it is the TRGs that mark a significant differences with de novo genes, as well as iORFs and CDSs, with a higher propensity to aggregate. Those last three bear no statistical difference amongst them. This is surprising because de novo genes and iORFs are typically expected to show higher aggregation propensity than well-established CDSs. This could mean that non-de novo TRGs have different selection constraints that makes them aggregation-prone.

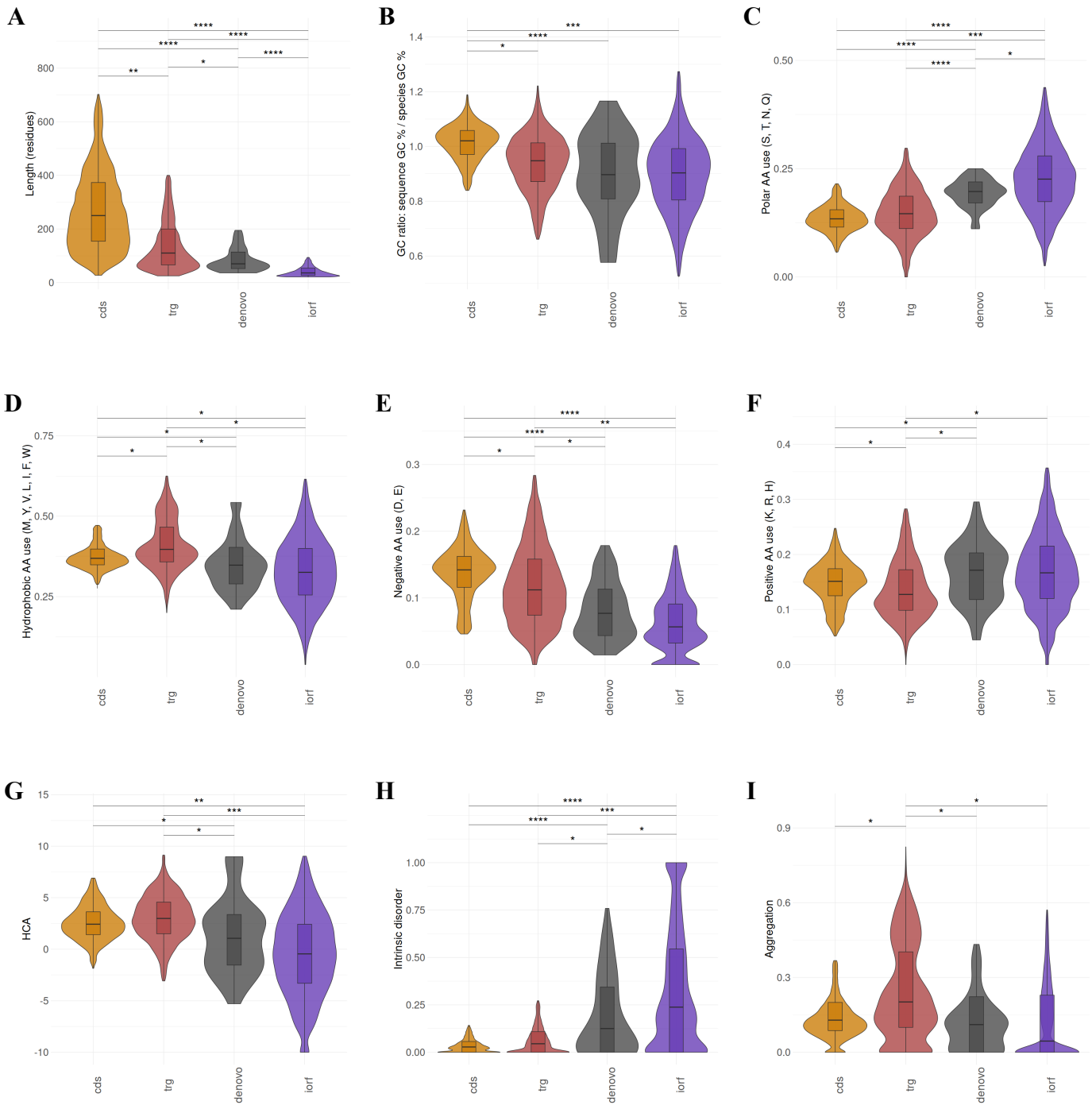


Figure 7: Distribution different descriptors for a sample of 5000 CDSs, 5000 TRGs, all denovo genes (n = 64), and 5000 iORFs. From A to I: ORF length (in residues), GC ratio (ORF GC% / global species GC%), polar residues use, hydrophobic residues use, negatively charged residues use, positively charged residues use, hydrophobic cluster analysis, disorder score, aggregation score. Asterisks denote level of significance: ****: $p \leq 1 \times 10^{-5}$ ***: $p \leq 1 \times 10^{-4}$ **: $p \leq 1 \times 10^{-3}$ *: $p \leq 0.05$

Non-coding origin

We focused on the furthest phylogenetic group from the focal species that had a syntenic non-coding match with the de novo gene, that we will call the furthest outgroup (for the example in Fig. 2, this would have been outgroup n°2 composed of only one species, but the furthest outgroup could also comprise several species). The idea behind this was to study whether this match was intergenic or overlapping an existing gene, in order to infer whether the de novo genes emerged from overprinting or from intergenic sequences.

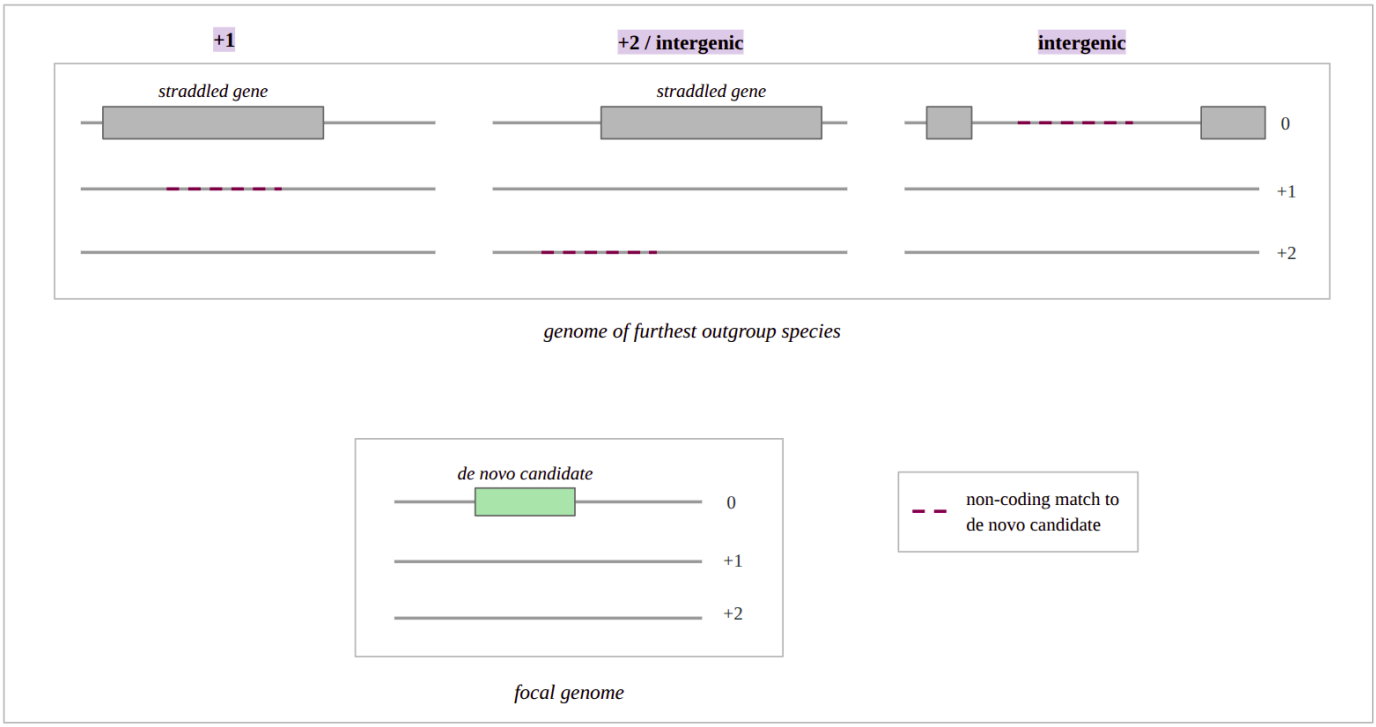


Figure 8: Representation of different possibilities of non-coding origins for a given species in the furthest outgroup. The de novo candidate in the focal genome is in green, at the bottom of the figure. At the top of the figure are three different possibilities (from left to right): the match can be entirely positioned on an alternative frame of an existing gene (here, the +1 frame). It can be overlapping both an alternative frame (here, the +2 frame) and part of an intergenic segment. It can also be 100% intergenic.

Among the 64 de novo genes, only one gene had different origins depending on the species considered in the furthest outgroup (see Fig. 8 to see the different types of origins possible). All other 63 genes had either only one species making up the outgroup, or a homogenous non-coding origin across all species.

In those 63 de novo genes, 49 had a 100% intergenic match across all species in the furthest outgroup, meaning 77% of our de novo genes likely came from ancestral intergenic sequences.

As for the remaining 14 genes, three of them were only overlapping the +1 frame of an existing gene, one was overlapping only the +1 frame as well as an intergenic segment, seven were straddling the +2 frame as well as an intergenic segment, and surprinsingly, we found three genes that were straddling the actual canonical reading frame of an existing gene and an intergenic segment (Table 1). Although this may seem counter-intuitive, these are not considered as coming from a coding sequence because the sequence that makes up the actual reading frame of the existing gene doesn't cover the de novo sequence enough to be considered as a homologue gene; here, the de novo gene has emerged from a combination of intergenic sequence and +0 frame.

		Frame of existing gene		
		+0	+1	+2
Intergenic overlap?	Yes	**3**	**1**	**7**
	No		**3**	**0**

Table 1: Position of the non-coding matches for the 14 de novo genes that were not either 100% intergenic, or that hadn't different origins depending on the species considered in the furthest outgroup.

Overall, those 14 de novo genes were separated into 10 different gene clusters. In order to deepen our analysis, we chose one non-coding match for each of these clusters, and we did a blastp homology search against the focal genome for each of the overlapped genes, in order to see if the overlapped genes were still conserved in the focal genome. For the five non-coding matches that overlapped an alternative frame of an existing gene (either +1 or +2) and an intergenic segment, the overlapped gene had a homologous gene in the focal genome that was right next to the de novo gene (Fig. 9). This shows that these de novo genes truly emerged straddling an existing gene and an intergenic segment, and stayed in the same environment.

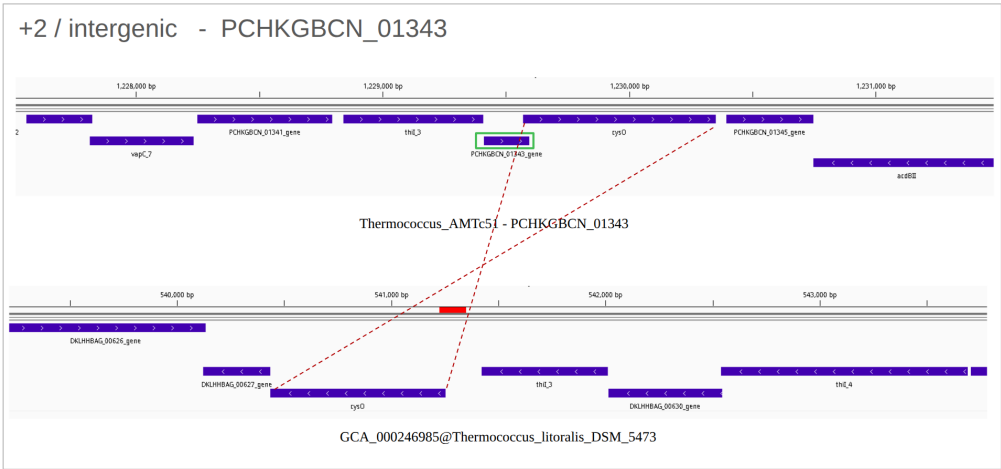


Figure 9: Configuration of the de novo gene PCHKGBCN_01343 and its non-coding match in the furthest outgroup selected for analysis. The representation was done using the IGV browser [35]. At the top, we can see the immediate surroundings of the de novo gene, framed in green. At the bottom, the zone of the non-coding match in the outgroup species is depicted in red. The DKLHHBAG_00628 gene, named cysO on the figure, is the overlapped gene. The red dashed line point to the gene in the focal species that is homologue to DKLHHBAG_00628, and that is right next to the de novo gene.

Among the 10 analysed non-coding matches, there were two that emerged entirely from the +1 frame of an existing gene. When doing a homology search of the overlapped genes, we found two different results; for the first overlapped gene, the search did not identify any homologous gene in the focal species, suggesting that the de novo gene emerged from an existing gene that then disappeared. As for the second result, we found that both the canonical frame and the +1 alternative frame of the overlapped gene shared homology with the de novo gene (Fig. 10). We concluded that the de novo gene emerged from a frameshift event in an existing gene, probably due to an insertion or deletion.

Finally, three of the 10 analysed non-coding matches emerged from the canonical (+0) frame of an existing gene, and an intergenic segment. For two of those, we were not able to detect homologous genes of the overlapped genes in the focal species, suggesting that part of an already existing gene merged with part of the intergenic segment next to it, creating a new gene. The last one was homologous to two genes, as well as the de novo gene, positioned right next to it (Fig. 11). We can conclude that an existing gene probably acquired STOP codons, dividing it into three different genes. In this case, we cannot consider any of the three resulting small ORFs as de novo genes.

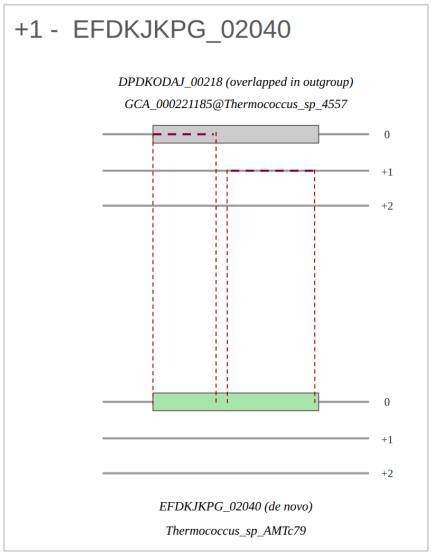


Figure 10: Visualisation of the homology search result between the de novo gene EFDKJKPG_02040 (at the bottom) and the overlapped gene in the furthest outgroup species DPDKODAJ_00218 (at the top). The de novo gene maps both to part of the canonical frame of DPDKODAJ_00218, and to part of its +1 frame.

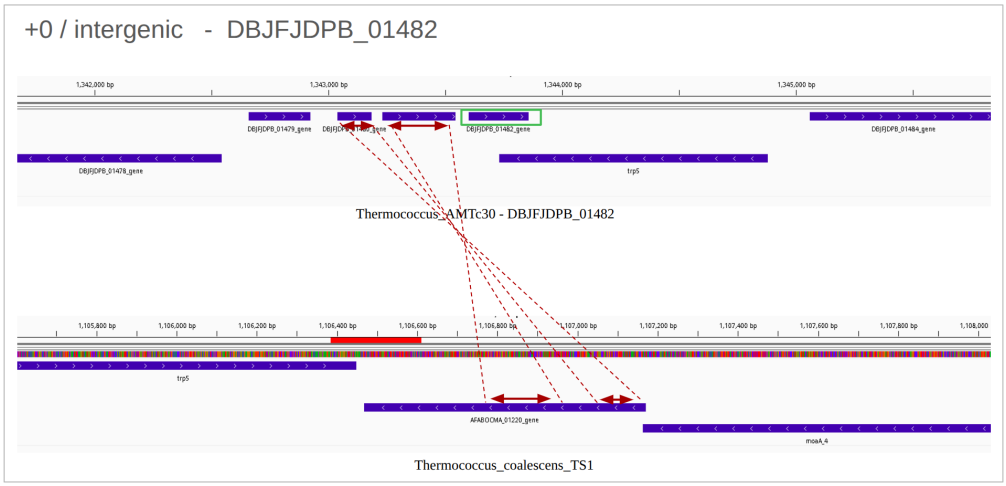


Figure 11: Configuration of the de novo gene DBJFJDPB_01482 and its non-coding match in the furthest outgroup selected for analysis. At the top, we can see the immediate surroundings of the de novo gene, framed in green. At the bottom, the non-coding match in the outgroup species is depicted in red. AFABOCMA_01220 is the overlapped gene. The red dashed line and red arrows show the genes in the focal species that are homologue to AFABOCMA_01220, and that are right next to the de novo gene.

To put it in a nutshell, we identified different events that led to de novo gene emergence in this study including: emergence from entirely ancestral intergenic regions (the majority case); overprinting within alternative reading frames of existing genes (+1 or +2); fusion of an intergenic segment with part of an existing gene, including cases overlapping canonical reading frames (+0); emergence

via frameshift within an existing gene due to likely indels; and gene fragmentation, where a single ancestral gene gained stop codons, splitting into multiple genes including the new de novo one.

3.3 Comparison to other de novo candidates

The 64 de novo genes with unpreserved integrity in the non-coding matches of their two outgroups presented no significant differences to those that had preserved integrity (n = 109), except for their sequence length (Fig. 12). This difference could only be a mechanistic one: longest sequences statistically tend to be more impacted by indels, meaning they have a higher chance of having an unpreserved integrity. The lack of statistical differences for other descriptors, though (Supp Fig. 3), shows that the candidate de novo genes are very similar regardless of their integrity status, suggesting that even the ones with a conserved integrity are actual de novo genes.

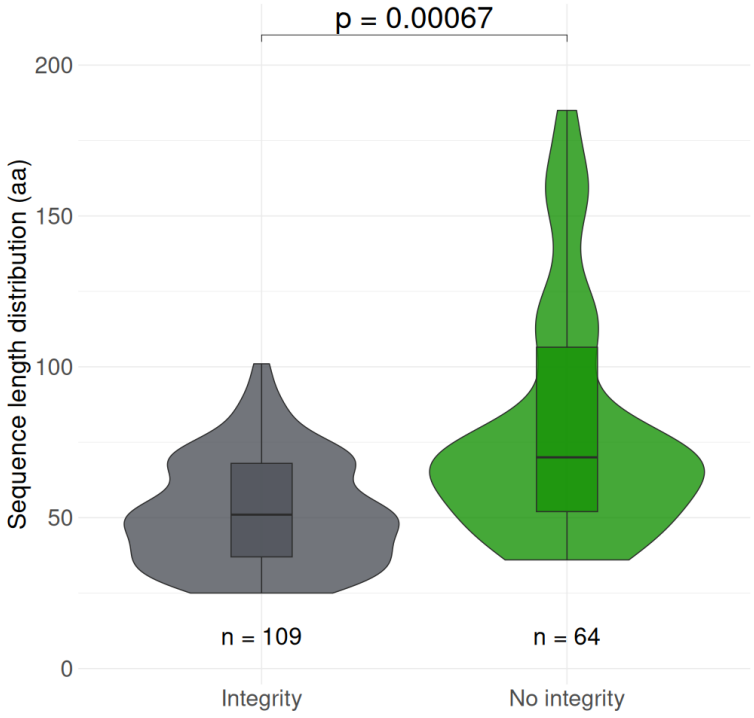


Figure 12: Distribution of the de novo genes sequence lengths (in amino-acids), for the de novo genes with unpreserved integrity (on the right, n = 64) and with preserved integrity (on the left, n = 109).

4 Conclusion and discussion

We detected 173 de novo genes among our 116 species, and we confirmed their status as emerged from a non-coding sequence for 64 of them, for which we conducted further analysis. Although we discarded 109 genes for these investigations, we concluded that they closely resembled the other 64 candidates. Since de novo gene sequences and predicted structures are typically different from other CDSs as well as iORFs, we can hypothesize that the remaining potential de novo genes are also genuine de novo genes whose ORF was still present in the neighbouring species lacking the gene. This recalls the ORF-first model that stipulates that de novo genes emerged from pre-existing long non-coding ORFs that later acquire expression [38].

The clustering analysis revealed that most de novo genes in our dataset are evolutionarily young, with 31 out of 40 gene families containing only one gene and thus having a leaf-level last common ancestor. This reflects both the rapid evolutionary turnover of de novo genes and the methodological bias toward detecting only recent emergence events. We also observed multi-species clusters with more conserved de novo genes across the phylogenetic tree, suggesting that some de novo genes serve an evolutionary purpose and do become fixed. Further analysis of their 3D structure and function prediction could help us understand what sets them apart from more evolutionary unstable de novo genes.

Our de novo genes characterization was consistent with findings in Eukaryotes; de novo genes in our dataset exhibit shorter lengths and lower GC content than TRGs and other CDSs, and their amino acid composition as well as biophysical descriptors (HCA, aggregation, and disorder) were intermediate between iORFs and TRGs. Overall, the characterization of de novo genes confirmed their transitional nature: they consistently display intermediate features between intergenic ORFs and canonical genes, which suggests that de novo genes are still in the early stages of evolutionary optimisation [6, 7]. Their profiles reflect their recent emergence from non-coding regions and partial adaptation to the functional constraints of coding sequences.

Surprisingly, 74% (49 out of 64) of our de novo genes had a 100% intergenic non-coding match in the furthest outgroup from the focal species. Even though archaeal genomes are compact and have small intergenic sequences, they seem to be the privileged origin for de novo emergence. However,

we also identified several cases of more complex origins: some de novo genes emerged through overprinting on alternative frames of existing genes, others from a combination of overprinting and intergenic sequence, and a few appeared as chimeras, combining intergenic segments with the canonical frame of an existing gene. These mixed-origin cases raise the question of whether such genes should still be classified as truly de novo, or whether they represent a distinct category of gene innovation.

De novo gene birth appears to play a more modest role in Archaea compared to Eukaryotes, likely due to the compactness of archaeal genomes and the prevalence of other evolutionary strategies such as horizontal gene transfer and mobile genetic elements. Nonetheless, our results demonstrate that de novo emergence does occur in Archaea, even under genomic constraints that might seem unfavourable. The identification of 173 candidates across 116 species highlights that, although rare, this process is active and potentially underappreciated in prokaryotic evolution.

Importantly, our findings provide the first evidence of de novo gene emergence in Archaea. Given their rapid evolutionary dynamics and the difficulty of detecting old de novo genes, it is likely that more such genes exist but have either diverged beyond recognition or already been lost. Most de novo genes we detected appear transient, supporting the idea that they are generally short-lived unless they provide a selective advantage and become fixed in the population. Future work could explore the genomic context of their emergence —such as proximity to mobile elements, conserved genes, or rapidly evolving regions— and structure prediction tools to infer potential functions or identify novel protein folds. Such insights would help clarify the conditions and mechanisms that favour the birth and retention of new genes.

5 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 (visited on 01/15/2025).

[2] Courtney M Thomas et al. “Factors shaping the abundance and diversity of the gut archaeome across the animal kingdom”. In: *Nature communications* 13.1 (June 1, 2022), p. 3358. ISSN:

2041-1723. DOI: 10.1038/s41467-022-31038-4. URL: https://europepmc.org/articles/PMC9187648 (visited on 01/15/2025).

[3] 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.

[4] Konstantin Khalturin et al. "More than just orphans: are taxonomically-restricted genes important in evolution?" In: *Trends in Genetics* 25.9 (Sept. 1, 2009). Publisher: Elsevier, pp. 404–413. ISSN: 0168-9525. DOI: 10.1016/j.tig.2009.07.006. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(09)00145-0 (visited on 06/13/2025).

[5] 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/ (visited on 01/14/2025).

[6] 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.

[7] 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.

[8] Li Zhao, Nicolas Svetec, and David J. Begun. "De Novo Genes". In: *Annual Review of Genetics* 58.1 (Nov. 2024), pp. 211–232. ISSN: 1545-2948. DOI: 10.1146/annurev-genet-111523-102413.

[9] Li Zhao et al. "Origin and spread of de novo genes in Drosophila melanogaster populations". In: *Science (New York, N.Y.)* 343.6172 (Feb. 14, 2014), pp. 769–772. ISSN: 1095-9203. DOI: 10.1126/science.1248286.

[10] Nikolaos Vakirlis et al. "De novo birth of functional microproteins in the human lineage". In: *Cell Reports* 41.12 (Dec. 20, 2022), p. 111808. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2022.111808.

[11] 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 (visited on 06/13/2025).

[12] 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.

[13] Andrew K Watson, Philippe Lopez, and Eric Bapteste. "Hundreds of Out-of-Frame Remodeled Gene Families in the Escherichia coli Pangenome". In: *Molecular Biology and Evolution* 39.1 (Jan. 1, 2022), msab329. ISSN: 1537-1719. DOI: 10.1093/molbev/msab329. URL: https://doi.org/10.1093/molbev/msab329 (visited on 01/14/2025).

[14] Niv Sabath, Andreas Wagner, and David Karlin. "Evolution of viral proteins originated de novo by overprinting". In: *Molecular Biology and Evolution* 29.12 (Dec. 2012), pp. 3767–3780. ISSN: 1537-1719. DOI: 10.1093/molbev/mss179.

[15] Stephen F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 5, 1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602 (visited on 06/17/2025).

[16] 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 (visited on 01/14/2025).

[17] Tatiana Tatusova et al. “NCBI prokaryotic genome annotation pipeline”. In: *Nucleic Acids Research* 44.14 (Aug. 19, 2016), pp. 6614–6624. ISSN: 1362-4962. DOI: 10.1093/nar/gkw569.

[18] David Vallenet et al. “MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis”. In: *Nucleic Acids Research* 48 (D1 Jan. 8, 2020), pp. D579–D589. ISSN: 0305-1048. DOI: 10.1093/nar/gkz926. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7145621/ (visited on 06/17/2025).

[19] Torsten Seemann. “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14 (July 15, 2014), pp. 2068–2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153. URL: https://doi.org/10.1093/bioinformatics/btu153 (visited on 05/15/2025).

[20] Josué Barrera-Redondo et al. “Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra”. In: *Genome Biology* 24.1 (Mar. 24, 2023), p. 54. ISSN: 1474-760X. DOI: 10.1186/s13059-023-02895-z. URL: https://doi.org/10.1186/s13059-023-02895-z (visited on 01/16/2025).

[21] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. “Sensitive protein alignments at tree-of-life scale using DIAMOND”. In: *Nature Methods* 18.4 (Apr. 2021). Publisher: Nature Publishing Group, pp. 366–368. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01101-x. URL: https://www.nature.com/articles/s41592-021-01101-x (visited on 01/16/2025).

[22] Martin Steinegger and Johannes Söding. “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: *Nature Biotechnology* 35.11 (Nov. 2017). Publisher: Nature Publishing Group, pp. 1026–1028. ISSN: 1546-1696. DOI: 10.1038/nbt.3988. URL: https://www.nature.com/articles/nbt.3988 (visited on 05/21/2025).

[23] Emmanuel Paradis and Klaus Schliep. “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. In: *Bioinformatics* 35.3 (Feb. 1, 2019), pp. 526–528. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty633. URL: https://doi.org/10.1093/bioinformatics/bty633 (visited on 05/22/2025).

[24] Chris Papadopoulos, Nicolas Chevrollier, and Anne Lopes. “Exploring the Peptide Potential of Genomes”. In: *Methods in Molecular Biology (Clifton, N.J.)* 2405 (2022), pp. 63–82. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-1855-4_3.

[25] Peter J. A. Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics (Oxford, England)* 25.11 (June 1, 2009), pp. 1422–1423. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp163.

[26] J. R. Lobry and C. Gautier. “Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes”. In: *Nucleic Acids Research* 22.15 (Aug. 11, 1994), pp. 3174–3180. ISSN: 0305-1048. DOI: 10.1093/nar/22.15.3174.

[27] K. Guruprasad, B. V. Reddy, and M. W. Pandit. “Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence”. In: *Protein Engineering* 4.2 (Dec. 1990), pp. 155–161. ISSN: 0269-2139. DOI: 10.1093/protein/4.2.155.

[28] M. Vihinen, E. Torkkila, and P. Riikonen. “Accuracy of protein flexibility predictions”. In: *Proteins* 19.2 (June 1994), pp. 141–149. ISSN: 0887-3585. DOI: 10.1002/prot.340190207.

[29] J. Kyte and R. F. Doolittle. “A simple method for displaying the hydropathic character of a protein”. In: *Journal of Molecular Biology* 157.1 (May 5, 1982), pp. 105–132. ISSN: 0022-2836. DOI: 10.1016/0022-2836(82)90515-0.

[30] Tristan Bitard-Feildel and Isabelle Callebaut. *HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences*. Pages: 249995 Section: New Results. Jan. 18, 2018. DOI: 10.1101/249995. URL: https://www.biorxiv.org/content/10.1101/249995v1 (visited on 05/26/2025).

[31] Bálint Mészáros, Gábor Erdos, and Zsuzsanna Dosztányi. “IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding”. In: *Nucleic Acids Research* 46 (W1 July 2, 2018), W329–W337. ISSN: 1362-4962. DOI: 10.1093/nar/gky384.

[32] Rune Linding et al. “A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins”. In: *Journal of Molecular Biology* 342.1 (Sept. 3, 2004), pp. 345–353. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2004.06.088.

[33] Chris Papadopoulos et al. “The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation”. In: *Genome Biology* 25.1 (Oct. 14, 2024), p. 268. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03403-7. URL: https://doi.org/10.1186/s13059-024-03403-7 (visited on 05/26/2025).

[34] Hadley Wickham. *ggplot2*. Use R! Cham: Springer International Publishing, 2016. ISBN: 978-3-319-24275-0 978-3-319-24277-4. DOI: 10.1007/978-3-319-24277-4. URL: http://link.springer.com/10.1007/978-3-319-24277-4 (visited on 06/10/2025).

[35] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration”. In: *Briefings in Bioinformatics* 14.2 (Mar. 1, 2013), pp. 178–192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. URL: https://doi.org/10.1093/bib/bbs017 (visited on 06/19/2025).

[36] Anna Grandchamp et al. “Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in Drosophila melanogaster”. In: *Genome Research* 33.6 (June 2023), pp. 872–890. ISSN: 1549-5469. DOI: 10.1101/gr.277482.122.

[37] Paul Roginski, Chris Papadopoulos, and Anne Lopes. “Impact of GC content on de novo gene birth”. In: *In revision* (2025).

[38] Christian Schlötterer. “Genes from scratch–the evolutionary fate of de novo genes”. In: *Trends in genetics: TIG* 31.4 (Apr. 2015), pp. 215–219. ISSN: 0168-9525. DOI: 10.1016/j.tig.2015.02.007.

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 age, origin, as well of sequence and structure conformation. Our findings provide the first evidence that, although rare, this process is active and potentially underappreciated in archaeal evolution.

Supplementary figures

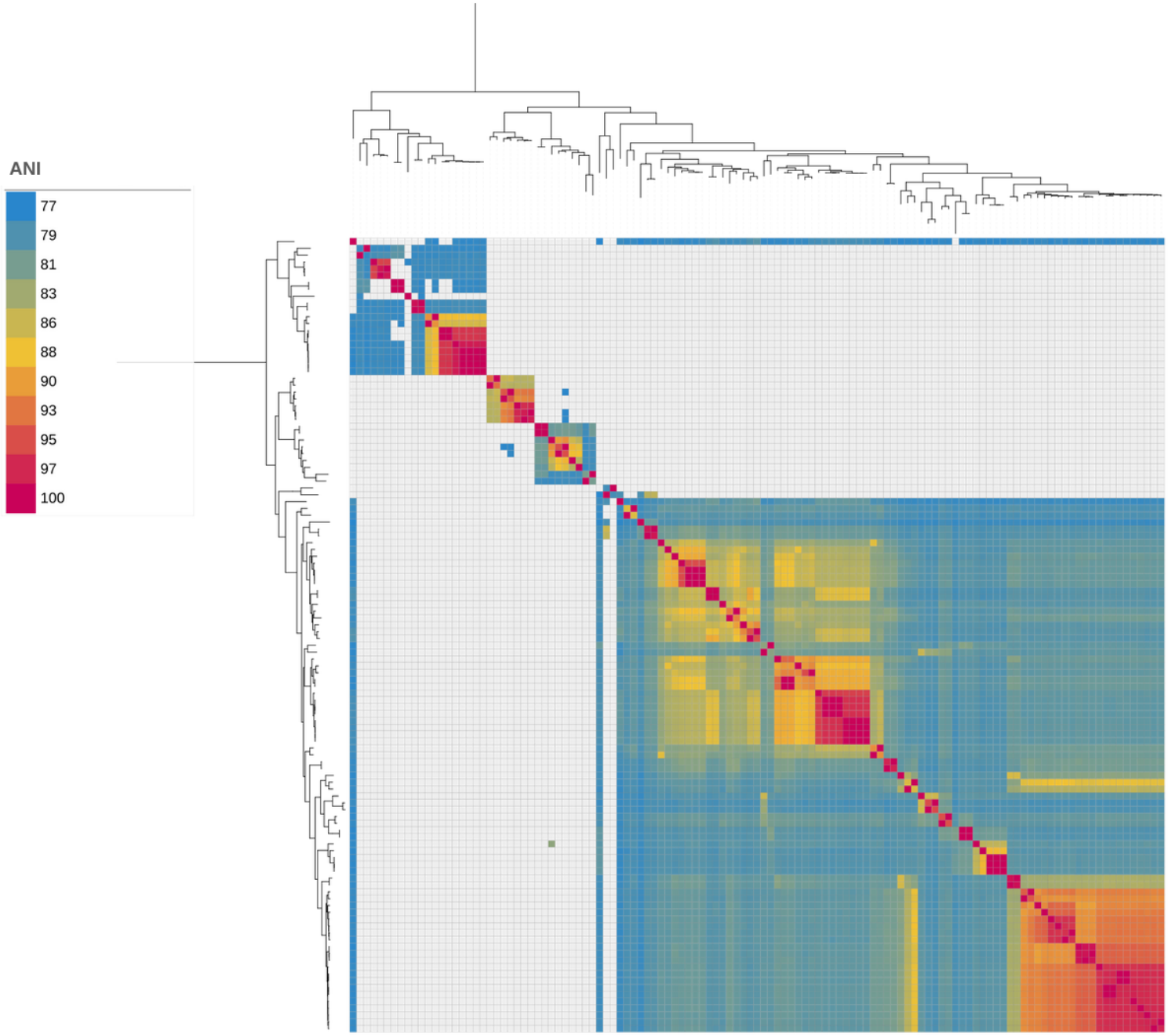


Figure 1: Linear phylogenetic tree of the 116 organisms (branch lengths included) with the Average Nucleotide Identity (ANI) heatmap. ANIs below 75 are not pictured.

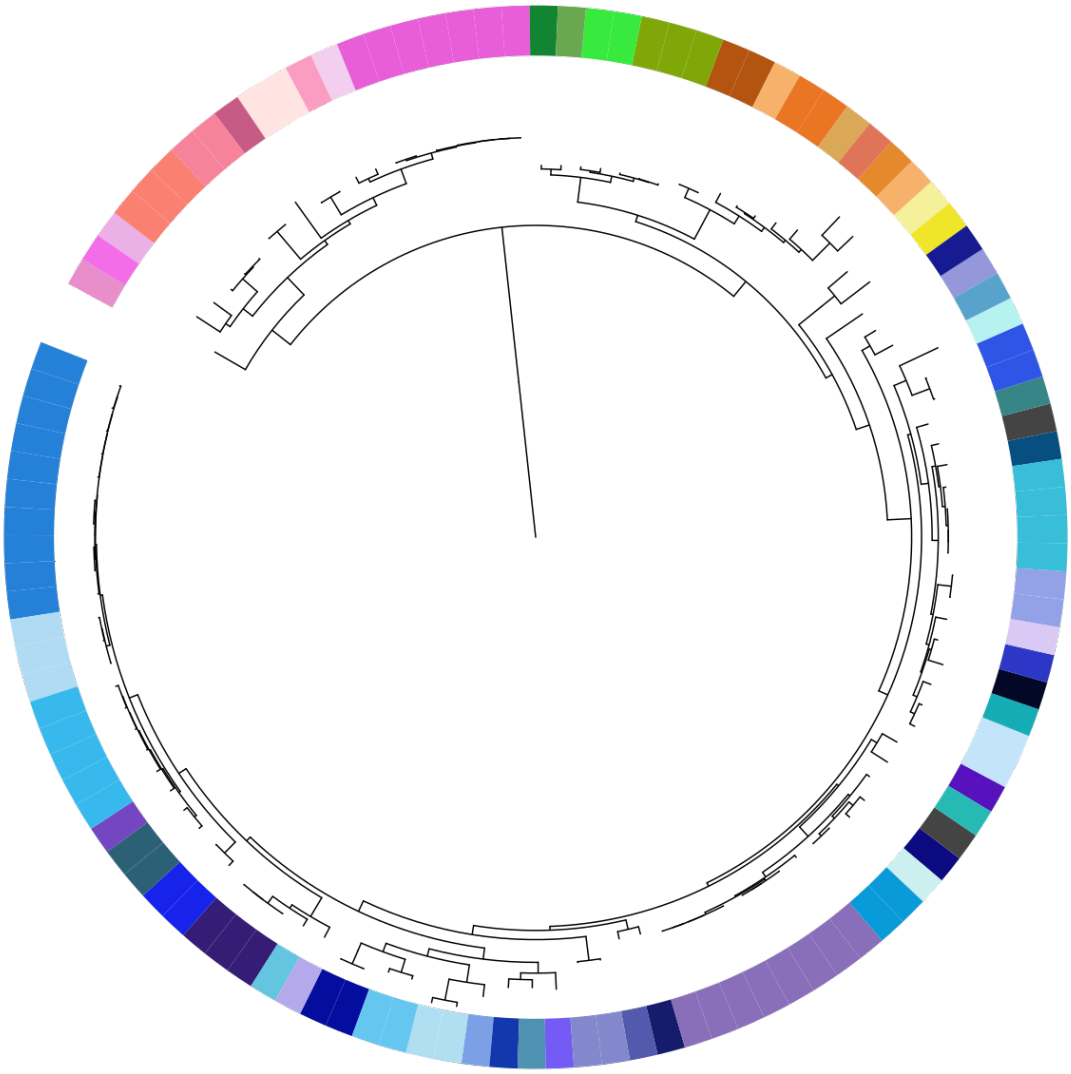


Figure 2: Circular phylogenetic tree of the 116 organisms (branch lengths included). Each separate colour is a distinct species (n = 68).

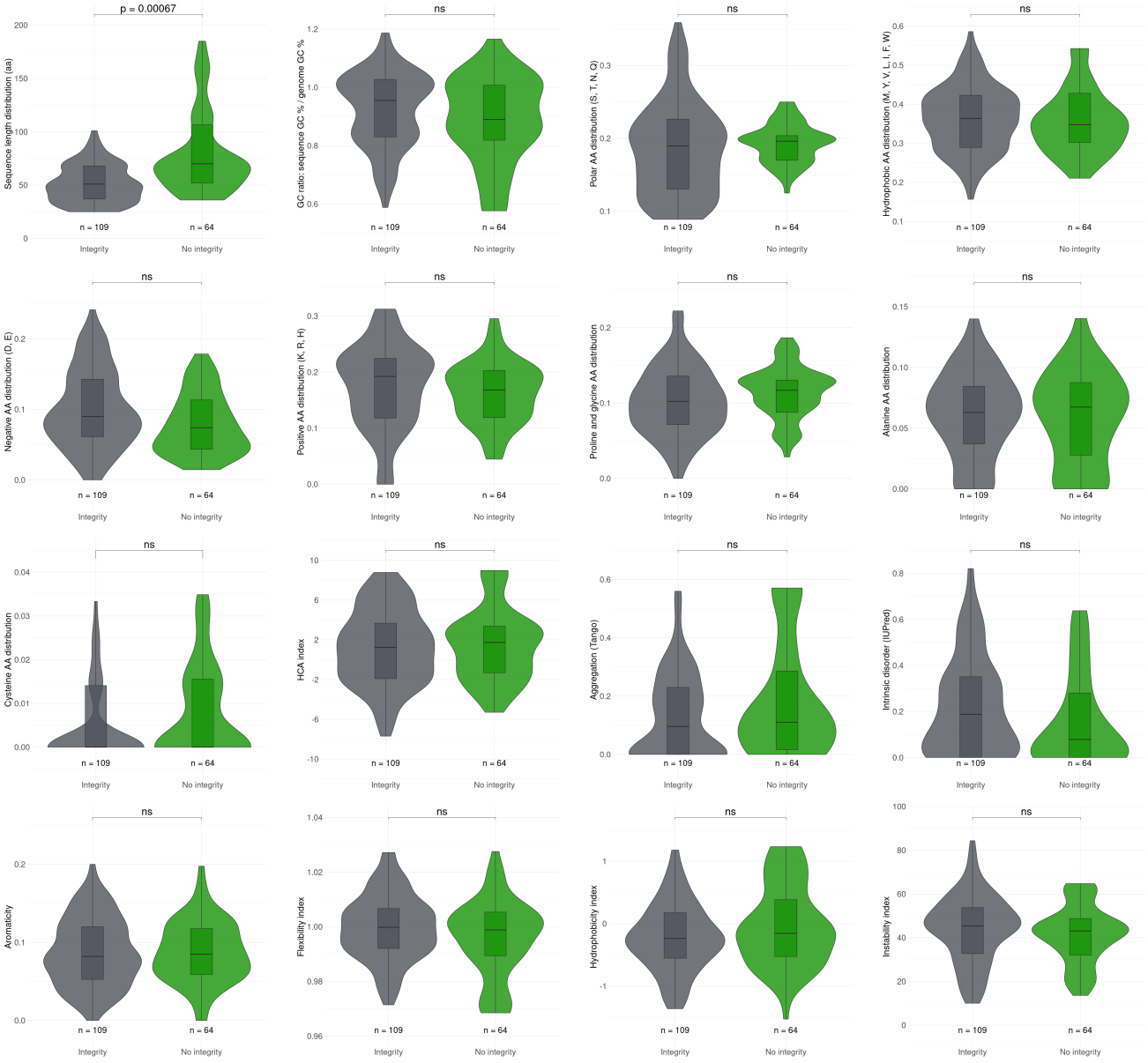


Figure 3: Distribution of different descriptors for the de novo genes with unpreserved integrity (on the right, n = 64) and with preserved integrity (on the left, n = 109). From the left to the right: ORF length (in residues), GC ratio (ORF GC% / global species GC%), polar residues use, hydrophobic residues use, negatively charged residues use, positively charged residues use, prolyne-glycine use, alanine use, cysteine use, hydrophobic cluster analysis, aggregation score, disorder score, aromaticity, flexibility, hydropathy, instability. No statistical difference has been observed, except for the sequence length.