Toward a Better Understanding of G4 Evolution in the 3 Living Kingdoms

Anaïs Vannutelli[1,2], Aïda Ouangraoua[2] and Jean-Pierre Perreault[1]

[1]Département de biochimie et de génomique fonctionnelle, faculté de médecine et des sciences de la santé, pavillon de recherche appliquée sur le cancer, Université de Sherbrooke, Sherbrooke, QC, Canada. [2]Département d'informatique, faculté des sciences, Université de Sherbrooke, Sherbrooke, QC, Canada.

Evolutionary Bioinformatics
Volume 19: 1–13

DOI: 10.1177/11769343231212075

Sage

ABSTRACT

BACKGROUND: G-quadruplexes (G4s) are secondary structures in DNA and RNA that impact various cellular processes, such as transcription, splicing, and translation. Due to their numerous functions, G4s are involved in many diseases, making their study important. Yet, G4s evolution remains largely unknown, due to their low sequence similarity and the poor quality of their sequence alignments across several species. To address this, we designed a strategy that avoids direct G4s alignment to study G4s evolution in the 3 species kingdoms. We also explored the coevolution between RBPs and G4s.

METHODS: We retrieved one-to-one orthologous genes from the Ensembl Compara database and computed groups of one-to-one orthologous genes. For each group, we aligned gene sequences and identified G4 families as groups of overlapping G4s in the alignment. We analyzed these G4 families using Count, a tool to infer feature evolution into a gene or a species tree. Additionally, we utilized these G4 families to predict G4s by homology. To establish a control dataset, we performed mono-, di- and tri-nucleotide shuffling.

RESULTS: Only a few conserved G4s occur among all living kingdoms. In eukaryotes, G4s exhibit slight conservation among vertebrates, and few are conserved between plants. In archaea and bacteria, at most, only 2 G4s are common. The G4 homology-based prediction increases the number of conserved G4s in common ancestors. The coevolution between RNA-binding proteins and G4s was investigated and revealed a modest impact of RNA-binding proteins evolution on G4 evolution. However, the details of this relationship remain unclear.

CONCLUSION: Even if G4 evolution still eludes us, the present study provides key information to compute groups of homologous G4 and to reveal the evolution history of G4 families.

KEYWORDS: G-quadruplex, evolution, eukaryotes, archaea, bacteria, co-evolution, RNA-binding proteins

RECEIVED: February 13, 2023. **ACCEPTED:** October 18, 2023.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Natural Sciences and Engineering Research Council of Canada (NSERC) Graduate Scholarship (to J.M.G.); Canada Research Chair in Computational and Biological Complexity (CRC Tier2 Grant 950-230577 to A.O.); Chaire de recherche de l'Université de Sherbrooke en Structure et Génomique de l'ARN (to J.P.P.); Fonds de Recherche du Quebec Nature et Technologies (FRQ-NT); Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-155219-17 to J.P.P., RGPIN-05552-17 to A.O.); Centre de Recherche du CHUS (to J.P.P.); Université de Sherbrooke.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Aïda Ouangraoua, Département d'informatique, faculté des sciences, Université de Sherbrooke, Université de Sherbrooke, 2310 rue bachand, Sherbrooke, QC J1K1T9, Canada. Email: aida.ouangraoua@usherbrooke.ca

Jean-Pierre Perreault, Département de biochimie et de génomique fonctionnelle, faculté de médecine et des sciences de la santé, pavillon de recherche appliquée sur le cancer, Université de Sherbrooke, 2500 Blvd de l'Université, Sherbrooke, QC J1K 2R1, Canada. Email: jean-pierre.perreault@usherbrooke.ca

Introduction

DNA and RNA can fold onto themselves to form secondary structures. Among these structures, G-quadruplexes (G4s) are stable non-canonical structures, made with Hoogsten pairings instead of Watson-Crick pairings. In G4s, the Hoogsten pairings occurs between 4 guanines to form a G-quartet. These G-quartets can stack on top of each other to create a G4 structure.[1,2]

Because they occur in both DNA and RNA,[3] G4 structures impact many biological processes. Indeed, DNA G4s are known to influence telomere homeostasis, epigenetics and transcription.[4-6] RNA G4s (rG4s) have been demonstrated to affect several post-transcriptional regulation mechanisms, such as those in messenger RNA (mRNA) with its impact on splicing, polyadenylation, non-coding RNA like miRNA regulation, translation and RNA transport.[7-12] For more information about G4s and rG4s functions, please refer to Varshney et al.'s review.[13]

G4 structures are involved in biological mechanisms associated with several pathologies, such as cancer and neurodegenerative diseases.[13] For instance, a G4 can fold at the *c-MYC* gene promoter.[14] Usually, this G4 remains unfolded in cancer cells; however, when it adopts a folded state, it activates an apoptotic cascade. Recently, a team designed a peptide to target the *c-MYC* G4, increasing its clinical application.[15] Further, a miRNA increases the production of β-amyloids via the inactivation of *SORL1* translation, which is a cause of Alzheimer's disease.[16] The binding of the miRNA toward its target can be prevented if the rG4 in the pre-miRNA remains folded.[17] Moreover, some G4s are associated with DNA methylation instability, leading to aging and cancer, which increases the potential utility of these structures in clinical studies.[18]

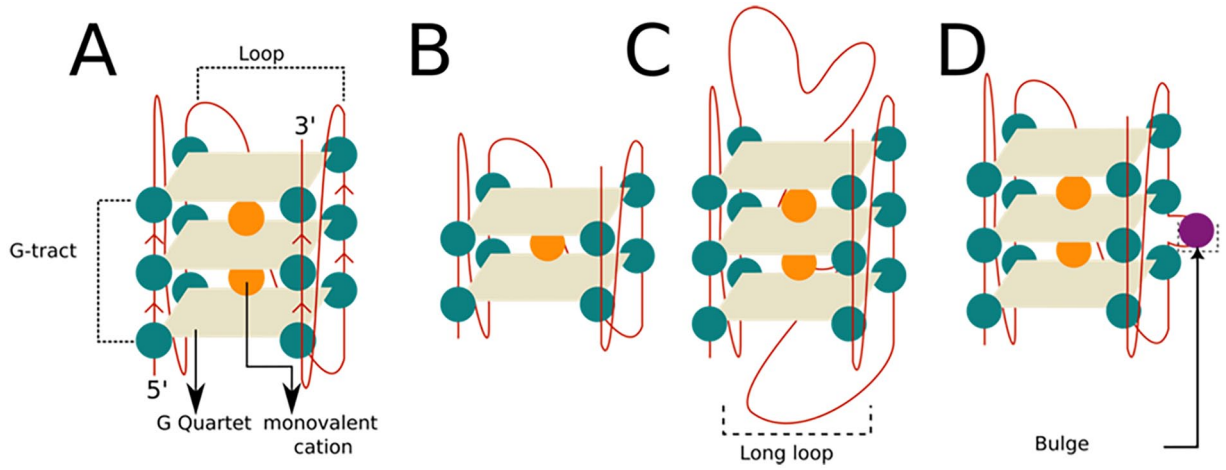

Figure 1. Schematic representation of a canonical G4 (A) and 3 non-canonical G4s: 2 G quartet G4 (B), long loop G4 (C) and bulge (D).

All these discoveries stimulated G4s massive prediction and detection. Initially, G4s were predicted using a canonical motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$[19,20] (see Figure 1). These tools showed that G4 structures are widespread among the human genome, with a particular enrichment in telomere and gene promoters at the DNA level. Over the years, scientists have identified G4s that were not matching the initial motif, thus called non-canonical G4s. These G4s may consist of only two G-quartets, have loops longer than 7 nucleotides or have a bulge inside a G-tract.[21-23] Consequently, the design of new prediction tools consider these non-canonical G4s.[24,25] G4 Hunter and cGcC are 2 tools that were developed using the GC skewness to assess the probability of G4 formation in a sequence, rather than relying on the motif.[26,27] New generation prediction tools use machine learning with experimentally detected G4s or rG4s. These tools include G4RNA screener, DeepG4 and rG4detector.[28-30]

These computational tools are employed to identify predicted G4s (pG4s) within the entire genomes and transcriptomes of various species. G4s are enriched in telomeres and promoters within the human genome. Additionally, the prediction also showed that rG4s are mainly enriched in 5′UTR of coding transcripts, with a modest enrichment in 3′UTR.[31-33] High-throughput G4s and rG4s detection corroborate most of these findings.[34-36] Yet, neither the prediction nor the detection is perfect. Indeed, the former yields some false negatives (ie, failing to predict G4s in sequences that fold into G4s), while the latter yields some false positives (ie, detecting G4s when they should not). Moreover, it has been shown that G4s in transcriptomes seem to be globally unfolded.[37] This study led to the hypothesis that rG4s have co-evolved with RNA-binding proteins (RBPs), which assist rG4s to stay unfolded when they are not required. RNA G4s would be globally unfolded to avoid negative impact on cell transcriptomes and translations, since their DNA counterpart are known as genomic instability marker.[38,39] This hypothesis aligns with the fact that bacteria inhabiting warm environments, where rG4s can freely fold and unfold, possess more rG4s than closely related bacteria in temperate environments.[40]

Considering all these elements, studying the evolution of rG4s and their potential coevolution with RNA binding sites of RBPs might help to understand their distributions and functions. Over the last years, several databases containing information on RNA G4 Binding Proteins (RG4BPs) have been made available. For example, G4IPDB contained over 60 RG4BPs,[41] but is not available anymore. More recently, QUADRatlas (https://rg4db.cibio.unitn.it/) was introduced, featuring data on rG4s overlaid with binding sites of RBPs, presenting information on over a thousand RG4BPs are presented.[42] Among these RG4BPs, many are known to have a meaningful impact on biological processes. One well-known example is DHX36, an RBPs known to bind rG4s, and scientists recently reviewed their interactions.[43] This review delves into many functions of the interaction between the helicase DHX36 and its G4 targets, discussing their implications in diseases like cancer, neurodegenerative diseases, and the aging process. Many studies have started to focus on the interaction between RBPs and G4s, and new RG4BPs continually being discovered, such as G3BP1.[44] This discovery was made by comparing rG4-seq data and eCLIP (enhanced version of the CrossLinking and ImmunoPrecipitation (CLIP) assay) data.[36,45] All these studies show that the comparison between RBPs and G4s still have a lot to reveal.

Until now, and according to our knowledge, limited research has been dedicated to the evolution of G4s. These studies have highlighted their limited conservation.[46-49] rG4s evolution within transcriptomes remains largely unexplored, apart from looking for the conservation at specific rG4s. For instance, a conserved rG4 within ribosomal RNA in mammals has been reported.[50] The complexity of studying the evolution of G4s and rG4s arises from the challenges of working with low-quality sequence alignments. However, their distribution has been examined across various phyla, including mammals, vertebrate, yeast, eukaryotes, bacteria and archaea).[46,51-54]

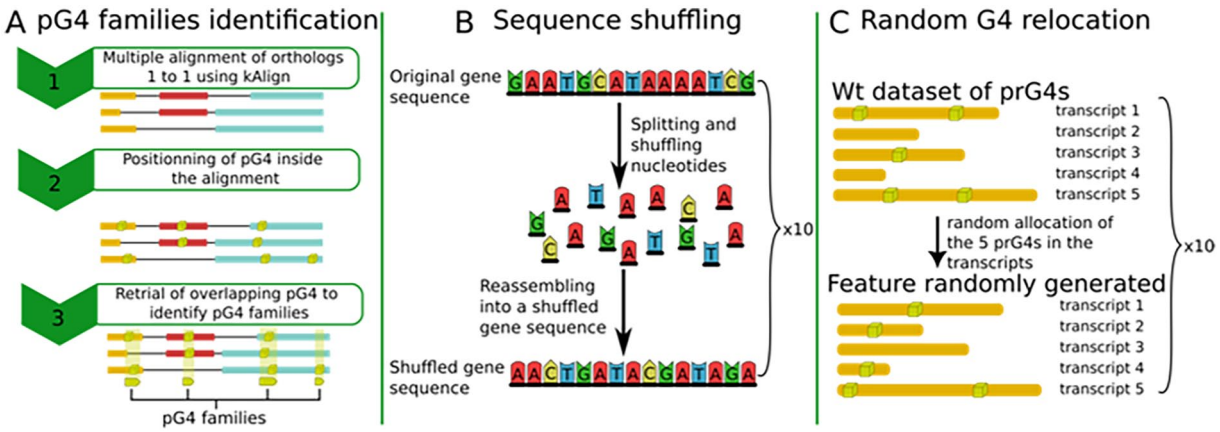

Figure 2. Schematic representation of the main methods. (A) The pG4 family retrial process comprises 3 steps: alignments of orthologous genes, positioning of pG4s, and then identification of pG4 families through detection of overlapping pG4s. The panel (B) demonstrates the gene sequences shuffling used to know how much G4s are predicted by chance. (C) Shows randomly relocated prG4s dataset, which consist in getting the total number of prG4s in a species, and randomly relocate them in transcripts of the species.

The most advanced study on G4s' evolution utilized orthologous genes alignment and suggested that G4s did not seem to be conserved.[47] Nevertheless, recent advances in predicting G4s across entire genomes and transcriptomes have revealed their distribution in every living kingdom and in numerous species.[46,51-54] Since G4s are widespread, a common origin could be possible. We expect to find at least some G4 families shared abroad a spectrum of species, indicating a common evolutionary history.

In the present study, we aim to untie G4s evolution by first using the core principle of the Frees et al[47] method on orthologous genes of eukaryotes, archaea, and bacteria. The principle of this method is to utilize gene alignments of closely related species in order to find related G4s. Secondly, given the apparent importance of RBPs in rG4s folding, we will assess RBPs' propensity and the locations of their binding sites in conjunction with predicted rG4s of different species to explore their interactions. This final step aims to confirm the possible coevolution between rG4s and RBPs.

Material and Methods

pG4 family identification

Data were retrieved from the Ensembl Compara database.[55] This dataset comprises information on genes and transcripts, including their genomic fasta sequences, and details all homology relationships (paralogs and orthologs) among coding genes from 60 species (with 25 eukaryotes, 12 archaea and 23 bacteria, as outlined in Supplemental data Table 1). To ensure the quality of our data, we filtered homology relationships to avoid low sequence conservation among homologous genes. Consequently, only homologous genes with one-to-one ortholog relationships were retained (ie, genes that are derived from a speciation event and present one copy in each species). This stringent criterion enabled us to obtain good quality gene alignments from multiple distantly related species. Orthologous gene families were retrieved via the default homologies file from the pan genome Ensembl release 46. In total, we extracted 9094 genes belonging to 4763 gene families from the Ensembl Compara database.

Subsequently, these gene sequences were aligned using kAlign (https://www.ebi.ac.uk/Tools/msa/kalign/), a tool for semi-global multiple sequence alignments able to align distantly related sequences.[56] Alignments were filtered based on their conservation identity and their ratio of nucleotides (number of nucleotides of a column in the alignment divided by the number of sequences), because some alignments mainly included aligned gaps due to the high phylogenetic distance between species. Thus, only alignments with an average nucleotide ration exceeding 55% were retained. Within these gene alignments, pG4s were positioned. The prediction of G4s was carried out using G4RNA screener (http://scottgroup.med.usherbrooke.ca/G4RNA_screener/), as explained below. Then, overlapping pG4s in the alignment were detected to identify pG4 families, without taking in account alignment identity. The alignments were not manually adjusted to make the pG4s coincide. Figure 2A summarizes the main steps of the pG4 family identification process. We used the G4RNA screener on the sequences of genes to predict G4s.[28] Default parameters were used for the prediction: windows of 60 nucleotides, step of 10 nucleotides between windows, threshold of 0.9, 0.5 and 4.5 for respectively G4 hunter, G4NN and cGcC. The G4s prediction process was previously reported in.[33,54]

Defining pG4 families as groups of overlapping pG4s within gene alignments also allowed to predict additional pG4s by homology, thereby increasing the number of pG4s in pG4 families. In the alignments columns where some gene sequences had pG4s and some other did not, we used GGRS Mapper (https://bioinformatics.ramapo.edu/QGRS/index.php)[57] and G4RNA screener to find pG4s in the genes with missing pG4s. The G4RNA screener scanning results in this step were like the initial one, but the prediction made using QGRS

Mapper with the default parameters predicted more G4s (see Supplemental data Table 1). In total, 36 548 G4s were predicted, with 14 569 identified by G4RNA screener and 22 019 through homology. These computational steps were conducted on a cluster of computers provided by the Digital Research Alliance of Canada.

The pG4 family identification have been limited to DNA pG4s due to disparities in the transcript annotation across different species. This incongruity has made the interpretation RNA pG4 families identification too challenging at present.

pG4s trees computation and comparison with gene trees

pG4 family alignments were computed using Align AI package from the Biopython library, version 1.79.[58] For each pG4 family, we computed a phylogenetic tree with the PhyML option of SeaView (http://pbil.univ-lyon1.fr/software/seaview3) and the default parameters based on pG4s sequences alignment.[59] To facilitate visual comparisons between pG4 family trees and their corresponding gene trees, the branches of pG4 family trees were swapped to closely resemble the gene trees. Next, we used a custom R script to generate mirror trees with the gene trees and pG4 family trees to help the visual comparison. Finally, the python library ete3 was employed to calculate the normalized Robinson-Foulds distance between pG4 family trees and gene trees.[60,61] This metric enables the quantification of the distance between phylogenetic trees, with higher values indicating grater dissimilarity.

In our analysis, species are organized into species trees constructed using super tree methods to combine several species trees from studies.[62-65] These trees were used as relational indicators between the species. Within these trees, we display pG4s densities, which were computed by normalizing the number of pG4s in a species by the length of genes and expressed in kilo base pair (kbp). This normalization process helps mitigate biases of species/genes having longer genes sequences, which might have a higher likelihood of containing pG4s.

RBP data retrieval and process

To compare RBPs CLIP data and pG4s, RNA pG4s were used. This was achieved by considering transcript locations rather than gene locations. As mentioned previously, G4s are predicted using G4RNA screener, a tool primarily designed for RNA G4s prediction, although it is also capable of predicting DNA G4s.[54] To prevent any confusion, pG4s refer to DNA pG4s, while prG4s denotes RNA pG4s. prG4 seconds data were compared with the RNA binding sites of RBPs obtained from CLIP data. We used the CLIP data from ENCORE (https://www.encodeproject.org) and POSTAR (http://111.198.139.65), which are derived from Cross-Linking Immuno Precipitation experiments focused on the binding of an RBP of interest to RNA[45,66,67] (Supplemental data Tables 2 and 3). In essence, when RBPs are bound to transcripts, a digestion was carried out to remove the unbound RNA. Then, immunoprecipitations were made on the RBPs. The recovered RNA sequences were then sequenced. This procedure led to mapping the RNA binding sites of RBPs onto the transcriptome. For *Homo sapiens*, eCLIP data were retrieved from the ENCORE database for K562 and HepG2 cell lines[45] (Supplemental data Table 4). All RBPs with eCLIP data were retrieved, and subsequently those lacking control or replicate files, or for which a tag was revoked, were excluded. This resulted in a subset of 150 RBPs (103 from HepG2 and 117 from K562 with a common set of 70 to both cell lines). To extract peaks corresponding to binding sites on RNA from the mapped reads available in bam file, we employed MACS3.[68] Then, we processed the files using bedtools and custom python scripts to obtain information on the overlapping binding sites on RNA and prG4 sequences. We computed overlaps between binding sites on RNA of RBPs and prG4s and added 150 nucleotides upstream and downstream. Given that both RBPs and rG4s form 3D structures, even though RBPs do not bind directly the rG4s but a flanking region close to them, this interaction might still be impacted by the rG4s formation or the binding of proteins. The use of flanking regions from both sides allowed us to check for the co-location of RBP binding sites on RNA and prG4s, and not only their direct interaction. For other species (ie, *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae)*, we obtained CLIP data from the POSTAR database where each species contained respectively 46, 2, 7, 6 and 81 RBPs.[67] These datasets were processed using the same procedure as for *Homo sapiens*, with the exception that peaks were already provided.

Shuffle and random datasets

To ensure the validity of our results, we generated different random or shuffled datasets. A first type of dataset was designed to evaluate the construction of pG4 families. In this dataset, we generated multiple shuffled sequences to establish a control for the normal density of pG4s. This comparison allowed to control if there were more or less pG4 than what was expected by chance. Additionally, the shuffled density can be considered as a background that can be subtracted from the normal density. To accomplish this, we used the python library "ushuffle" to shuffle entire genes sequences.[69] We generated 3 types of shuffles: mononucleotide, dinucleotide and trinucleotide (ie, 1-, 2-mer and 3-mer nucleotide shuffling). Each of these shuffling processes was repeated 10 times and only the average appears. This results in most case in decimal numbers.

The second dataset was generated to compare CLIP data and to evaluate if RBPs are binding rG4s more or less than expected by chance. Hence, the number of prG4s of each species was retrieved from the GAIA database (https://gaia.cobius.usherbrooke.ca/) and was randomly relocated in the

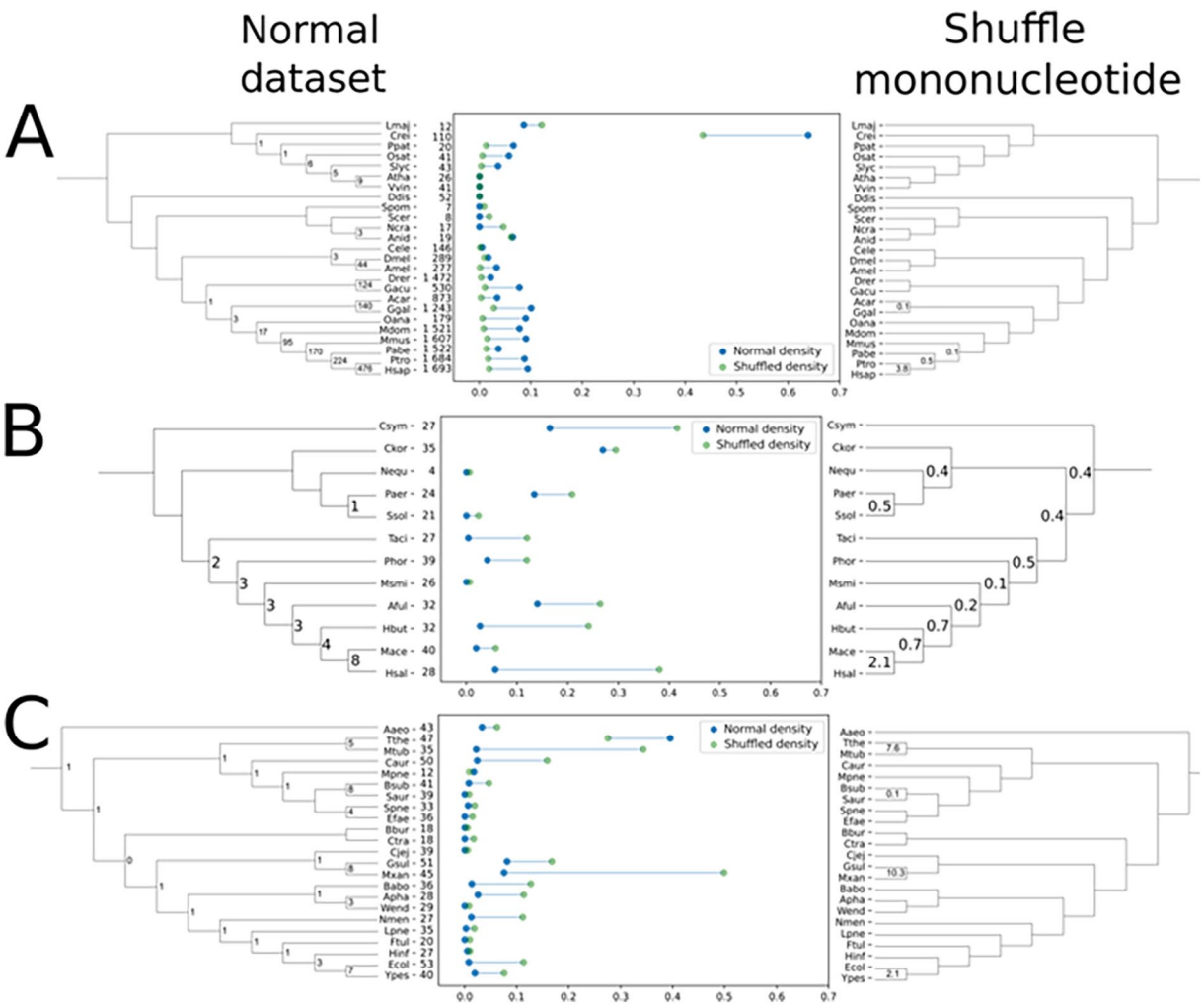

Figure 3. pG4 evolution inferred in a species tree: (A) eukaryote species, (B) archaea species, and (C) bacteria species. Normal densities (pG4/kbp) appear in blue and the average of the mononucleotides shuffled density (pG4/kbp) appears in green. Numbers displayed at the nodes of trees indicate the count of conserved pG4 families, while numbers at the tree leaves indicate the number of predicted G4s within the orthologous groups.

transcripts of the species, creating a "fake prG4" dataset.[70] Then, the fake prG4s were randomly selected, and their locations were randomly allocated throughout the transcriptome. The random data set was generated 10 times, and Figure 2C provides a simplified illustration of the concept.

Results

pG4 evolution inside species trees

The initial objective of our study was to get an overview of pG4 evolution. The evolution of G4s and rG4s remains unresolved. Therefore, we decided to initially focus on gene level since more information is currently available regarding their evolution. Our strategy was to identify families of pG4s within orthologous genes. Therefore, pG4 families were computed through multiple sequence alignments of orthologous genes (refer to Figure 2A for the illustrated method). We identified overlapping pG4s in these alignments as pG4 families, without requiring a minimum overlap. Figure 3 illustrates the overall data. When comparing the prediction of G4s in normal and shuffle datasets, more G4s are predicted than expected by chance in eukaryotes. Conversely, for archaea and bacteria there are as many, or fewer pG4s than expected by chance. These observations were in good agreement with a previous study,[54] where a negative pressure of selection was observed in prokaryotes while positive in eukaryotes. This result is further discussed in the discussion. Also, *Chlamydomonas reinhardtii* (*Crei*) stands out as a unique case among the studied species due to its high pG4 density, exceeding 0.6 pG4/kpb. This phenomenon was also observed in another study in which the link between the high GC content of the species and its high pG4 density is discussed.[54] We noticed that only a few pG4s were conserved in both datasets. To illustrate, let's consider eukaryotes as an example (Figure 3A). Between *Homo sapiens* and *Pan troglodyte* (*Hsap* and *Ptro*) 476 pG4 families were conserved in the normal dataset, whereas in the shuffled dataset, the average number was 3.8 (computed as an average of 10 runs, resulting in decimal numbers). Yet, more than a thousand pG4 are predicted in these species (ie, 1684 for *Ptro* and 1693 for *Hsap*, see Figure 3A). Consequently, less than a third of the

pG4 families are conserved between these closely related species. For eukaryotes, there are more pG4s families in the normal dataset than in the shuffled one, which was expected since the pG4s density in the normal dataset is higher.

In prokaryotes, pG4s densities in the shuffled dataset are overall higher than in the normal ones (see Figure 3B and C). This phenomenon may be attributed to the absence of certain mechanisms in prokaryotes, potentially rendering pG4s a non-advantageous feature for them, leading to a negative selection pressure against them. However, in the case of archaea, more pG4s families are conserved in the normal dataset in common ancestor than in the shuffled one. For instance, between *Methanosarcina acetivorans* (*Mace*) and *Halobacterium salinarum* (*Hsal*), there are only 2.1 pG4s families in the shuffled dataset, whereas there are 8 in the normal dataset. In bacteria, even though the pG4 densities in the shuffled dataset are higher than the normal one, there is globally few pG4 families in the shuffled dataset. To illustrate, there are only 4 different common ancestors with conserved pG4 families in the shuffled dataset, while in the normal dataset, there are more than 15 common ancestors with conserved families. Hence, despite the prediction of fewer G4s bacteria than would be expected by chance, it appears that there might be a higher prevalence of shared pG4s in common ancestors than expected by chance.

When comparing species groups, we observe that the number of conserved families are higher in eukaryotes. However, in prokaryotes, pG4s families exhibit conservation in more ancient common ancestors. For example, in eukaryotes, the most ancient conserved families are found in plants or vertebrates, but none are shared by all animals or fungi (Figure 3A). In contrast, in bacteria at least one pG4 family is conserved in the most ancient common ancestors. As a reminder, species trees were built using a super tree method, see Material and Methods for more information.

Globally, these results also appeared when comparing the predictions on the normal dataset to the prediction on dinucleotide and trinucleotide shuffling datasets (see Supplemental Figures 1 and 2). Comparing the results for bacteria (Figure 3C, Supplemental Figures 1C and 2C), the number of conserved pG4s families for the common ancestor between *Geobacter sulfurreducens* (*Gsul*) and *Myxococcus xanthus* (*Mxan*) decreases from 10.3 in the mononucleotide shuffle to 7.8 in the dinucleotide shuffle, and further down to 2.4 in the trinucleotide shuffle. Other species groups yield similar results. This indicates that the closest a shuffled sequence is to its normal sequence (from mononucleotide to dinucleotide and trinucleotide), the fewer conserved pG4 families are found. In summary, although pG4s family conservation is low, it remains higher than what would be expected by chance. The results also demonstrate that pG4s are more conserved in terms of numbers in eukaryotes, but they are not conserved at the root of eukaryotes. While for prokaryotes, fewer pG4s are conserved compared to eukaryotes, but they are present in some ancient common ancestors.

Since the quality of alignments was good for most orthologs gene groups after filterin, we conducted a detailed inspection of the sequence alignments for the 5 most conserved pG4 families in eukaryotes and prokaryotes (for more information on these families, please refer to Supplemental Figure 3). One alignment exhibited numerous insertions and deletions, and this alignment exclusively included eukaryotic genes (Supplemental Figure 3E). This observation can be attributed to the presence of long introns between exons in eukaryotic genes, which are less conserved in sequences.[71] Across all these alignments, we found that the regions of G-tracts were consistently well-aligned and conserved, even in genes where no G4s were initially predicted. As an example, consider Supplemental Figure 3A, where the gene FTT_0154 had a pG4 with the initial prediction, but GSU1819 did not, despite having some conserved G-tracts. However, the regions between the G-tracts, which represent putative loops of pG4s, were less conserved than G-tracts. Surprisingly, G-tracts were found to be most conserved in the gene group with the fewest predicted G4s predicted G4s (Supplemental Figure 3C). Based on this observation, a G4 prediction was made within these genes at the location of a pG4 family using QGRS Mapper with a wide motif.[57] The results revealed that this homology-based prediction approach helped identify more conserved pG4s (refer to Figure 4). For the expanded pG4 families, the gene tree and the pG4 family tree were compared (see Supplemental Figure 4). In some cases, the topology of the trees was very similar, while in others, the trees were very different. This distinction was confirmed by the mirror tree, which represents the relationship between the pG4 family tree and the gene tree, and the normalized Robinson-Foulds measure (RF).[61] The RF metric quantifies the difference between the sets of clades of 2 phylogenetic trees, providing an estimate of the distance between them. For instance, in Supplemental Figure 4A, the trees diverged considerably and exhibited an RF value of 0.79, whereas in Supplemental Figure 4D and 4E, the trees displayed few changes and had an RF value of 0.50. This demonstrates that the evolution of pG4 families and the gene families is not always congruent. Therefore, pG4 families may evolve differently from the genes that contain them.

Subsequently, the evolution of the expanded pG4 families in species trees was investigated using the same strategy employed for the initial prediction. The pG4 family's evolution was inferred in species trees using Count, and the results are presented in Figure 4. As anticipated, pG4 densities in the normal dataset were higher compared to the initial prediction without homology-based prediction. However, the number of G4s predicted in this condition for the shuffled dataset remained similar to the initial prediction. Therefore, the normal densities were higher than the shuffled ones or at similar levels for prokaryotes. This was in contrast to the previous observation where pG4s densities were higher in the shuffled dataset than in the normal dataset. Among eukaryotes, pG4s exhibited higher conservation compared to the initial prediction.

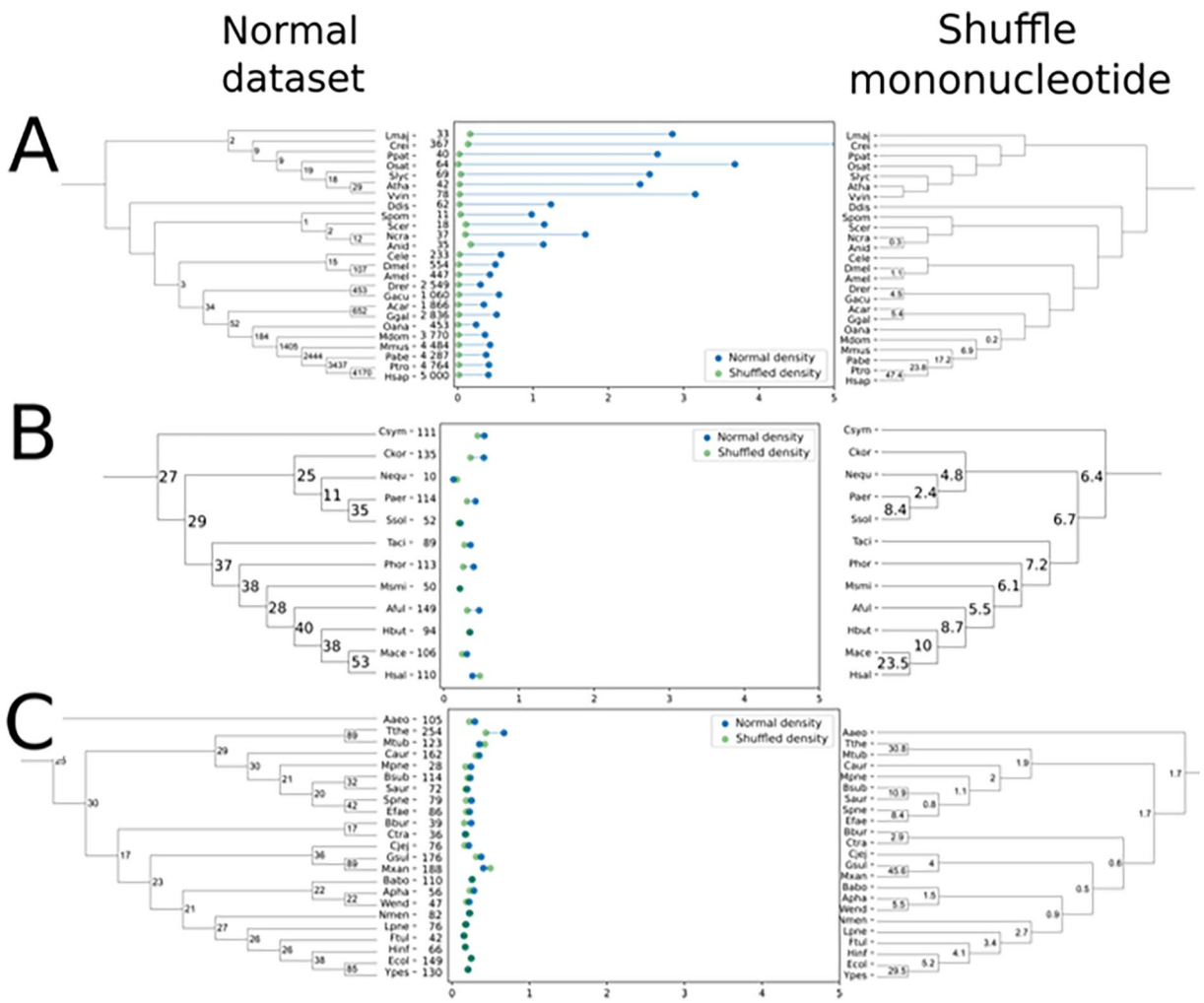

Figure 4. pG4 evolution predicted using homology inferred in species tree: (A) eukaryote species, (B) archaea species, and (C) bacteria species. Normal densities (pG4/kbp) appear in blue and the average of the mononucleotides shuffled densities (pG4/kbp) appears in green. Numbers displayed at the nodes of trees indicate the count of conserved pG4 families, while numbers at the tree leaves indicate the number of predicted G4s within the orthologous groups.

For example, there were initially 476 pG4 families between *Hsap* and *Ptro*, which increased to 4170 pG4s with the homology-based prediction. Consequently, the conservation of pG4s between these species increased from less than 50% to more than 80%, resulting in a total of 4764 and 5000 predicted G4s in *Ptro* and *Hsap*, respectively. Additionally, pG4 families were more conserved in common ancestors. For instance, in Fungi, one pG4s family was conserved in the most ancient common ancestor, which was not the case with the initial prediction. Furthermore, with the initial prediction, the most ancient conserved pG4 family was in the most ancient common ancestor of vertebrates, but with the homology prediction, there were 3 common pG4s families for all animals used in this study. Nevertheless, even with the discovery of more pG4s families in more ancient common ancestors, no pG4 families were identified in the most ancient common ancestors of all eukaryotes. For prokaryotes, pG4 family conservation highly increases. Specifically, 27 and 25 pG4s families were conserved in all archaea and bacteria respectively, compared to 6.4 and 1.7 in the shuffled dataset. This indicates that more pG4 families are conserved than expected by chance. We also compared the normal predictions to predictions made using dinucleotide and trinucleotide shuffled datasets (see Supplemental Figures 5 and 6), and the results were consistent with the initial prediction. In summary, the initial prediction revealed that only a few pG4 families were common to different species, yet the conservation was more important than expected by chance. With the homology prediction, pG4 conservation increased substantially, especially for prokaryotes and closely related species in eukaryotes.

Relationship between prG4s and RBPs

Previous studies highlighted differences in prG4s densities between eukaryotes and prokaryotes,[37,54] yet the underlying reasons remain unclear. The prevailing hypothesis suggests that RBPs, particularly helicases, might account for this difference by potentially unfolding rG4s in the human transcriptome.[37] Therefore, our next objective was to investigate potential

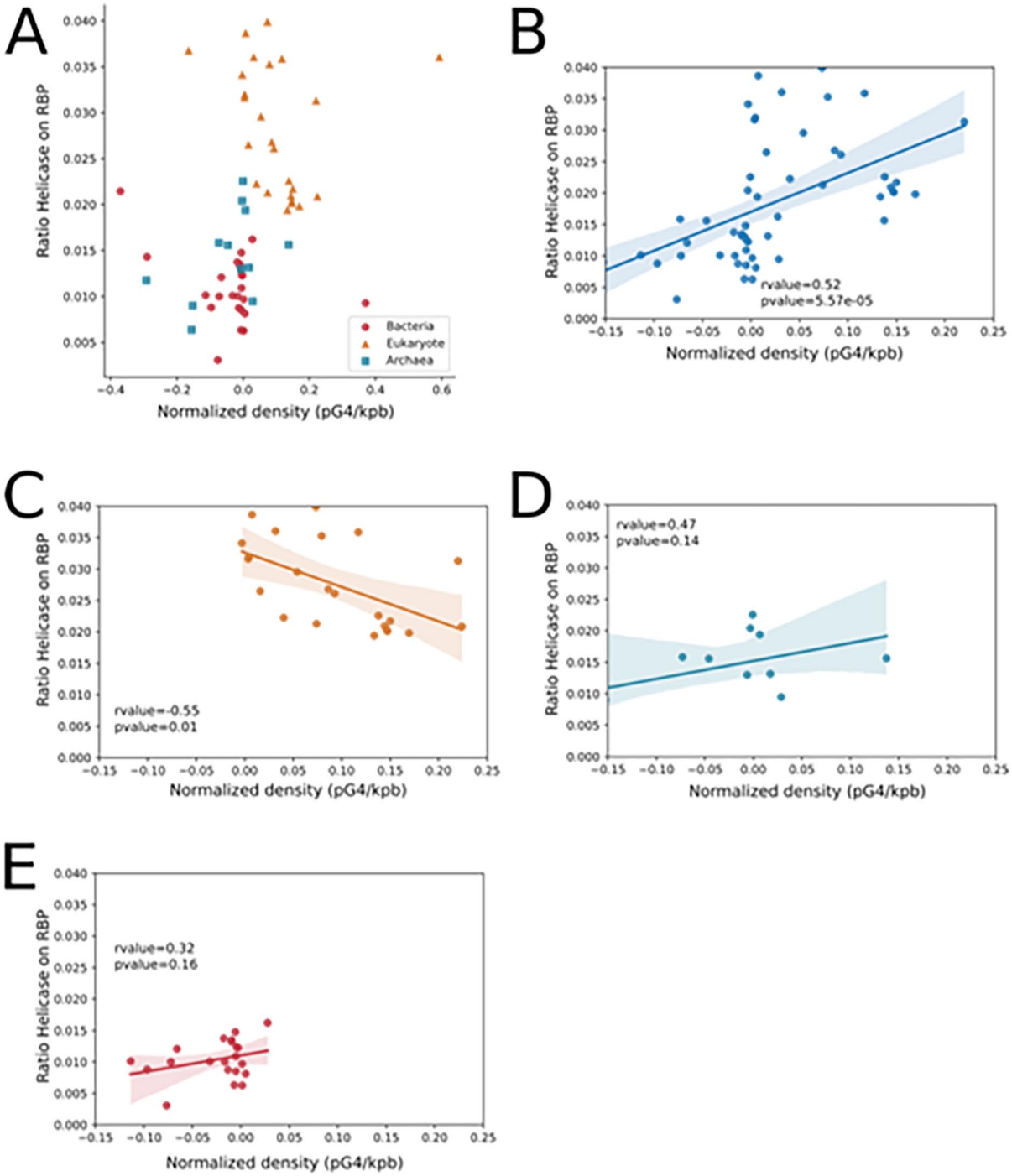

Figure 5. prG4 densities and helicase ratio. (A) Distribution of species based on their helicase ratio and the normalized prG4 density (calculated by subtracting the shuffle density from the normal density); (B–E) depict correlations between the helicase ratio and prG4 densities for respectively all species, eukaryotes, archaea, and bacteria. The r-value represents to the Pearson correlation coefficient.

coevolution between RBPs and prG4s. Our strategy involved analyzing the relationship between prG4s densities obtained here and annotated RBPs from the RBP2GO database (https://rbp2go.dkfz.de).[72] The results suggest a complex relationship between these factors. Figure 5 presents the ratio of the helicases (ie, the number of helicases over the annotated RBPs number) versus the normalized prG4 densities. Since the annotation of RBPs varies among species, normalizing the number of helicases allows for comparisons between species. In Figure 5A, the helicase ratio is plotted against the normalized prG4 density (normal prG4 density minus the shuffled density). Firstly, the ratio of helicases appears higher in eukaryotes than in prokaryotes. Specifically, the helicase ratio ranged from 0.0025 to 0.025 in prokaryotes, while in eukaryotes, it ranged

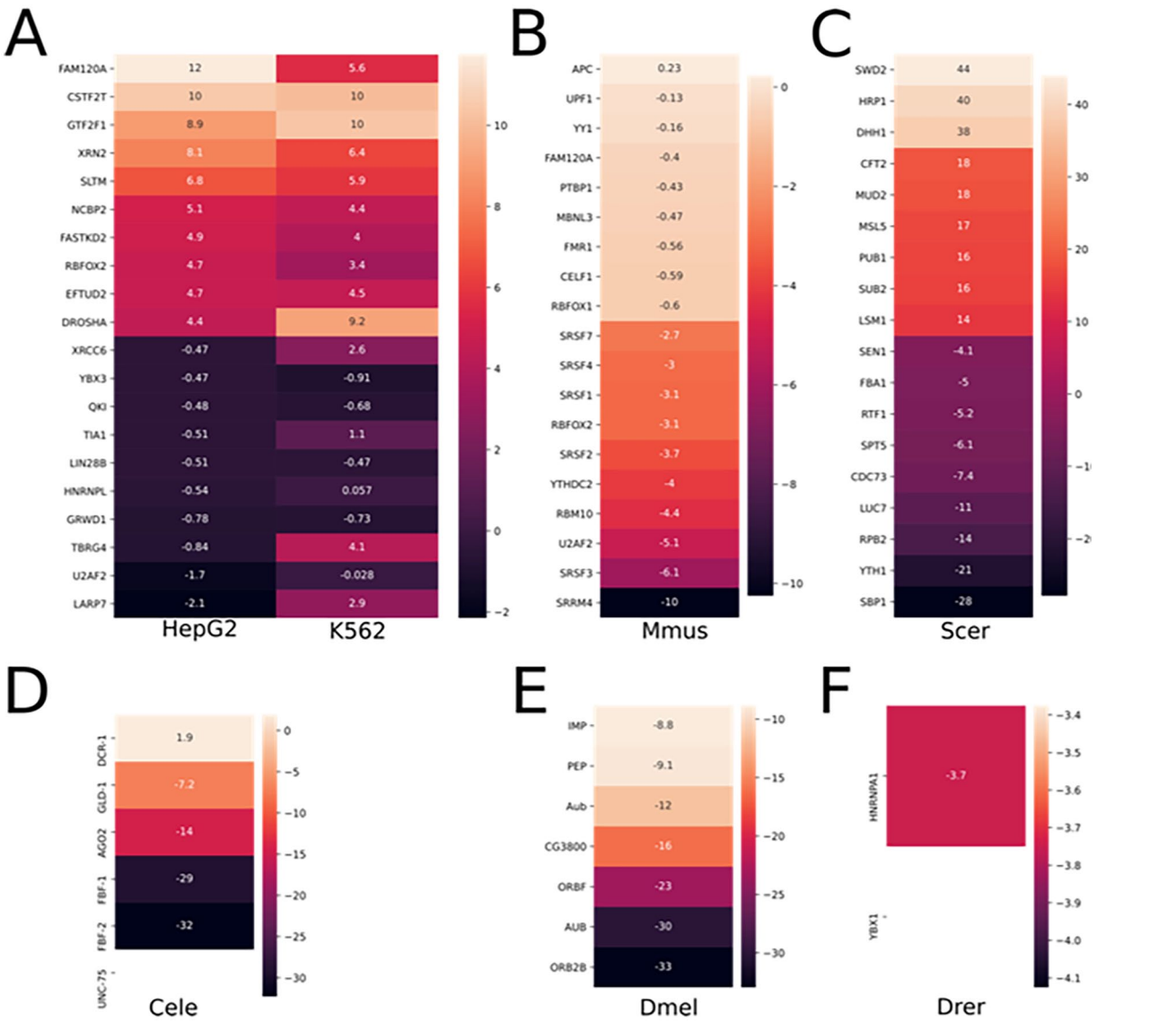

Figure 6. Top and bottom 10 proportions of prG4s near RBP binding sites relative to the total number of prG4s present on the transcripts bound by the RBP. (A) to (F) correspond respectively to *Homo sapiens* (*Hsap*), *Mus musculus* (*Mmus*), *Saccharomyces cerevisiae* (*Scer*), *Caenorhabditis elegans* (*Cele*), *Drosophila melanogaster* (*Dmel*) and *Danio rerio* (*Drer*). For *Hsap*, 2 cell lines are presented: HepG2 on the left column and K562 on the right column. RBPs names are displayed next to heatmaps.

from 0.020 to 0.040. Globally, there seems to be a positive correlation between normalized prG4 densities and the helicase ratio. To confirm this observation, we performed a linear regression on this data, removing outliers using the interquartile range method. The results suggest a significant positive correlation between prG4s densities and helicases ratio (Figure 5B). However, when examining the correlation within each species group separately, the significant positive correlation is not consistent. Specifically, for archaea and bacteria, no significant correlations are found, whereas for eukaryotes, there is a strong significant negative correlation. In conclusion, there is a complex relationship between helicases and prG4 densities with significant correlations, but the nature of this relation varies depending on the species group. Moreover, contrary to our expectations, we found a negative correlation for eukaryotes, which contrasts with the hypothesis. This result is further discussed in the discussion, particularly regarding its interpretation and the use of the helicase ratio over RBPs.

CLIP data *versus* pG4s

To further investigate the relationship between prG4s and RBPs, we examined the relative location of RBPs RNA binding sites and prG4s. This analysis provided an overview of the distribution of these locations across different species. To achieve this, we crossed rG4s prediction data with freely available CLIP data for *Homo sapiens* (*Hsap*), *Mus musculus* (*Mmus*), *Caenorhabditis elegans* (*Cele*), *Danio rerio* (*Drer*), *Drosophila melanogaster* (*Dmel*) and *Saccharomyces cerevisiae* (*Scer*). However, due to variation in the annotation of RBPs

RNA binding sites among different species, the interpretability of these results is somewhat limited. These findings indicate that RBPs interacts prG4s in diverse ways, depending on cell lines and/or species. For the human, we compared the locations of RBPs RNA binding sites to prG4s locations and to a randomly generated prG4 dataset. By subtracting the random count of prG4s located near RBP binding sites from the actual count, we derived a proportion that indicates whether prG4s are more frequently found in close proximity to RBP binding sites than would be expected by chance. In Figure 6A, the top 10 RBPs with the highest positive and negative proportions are presented. Some RBPs exhibit similar binding profiles in the 2 cell lines (eg, HepG2 and K562), while others do not. For instance, GTF2F1 is among the top 10 RBPs with the highest proportion of prG4s in both cell lines, whereas TBRG4 shows a negative prG4 proportion in HepG4 but a positive proportion in K562. Supplemental Figure 7 displays all eCLIP-annotated RBPs from HepG2 and K562, along with additional information, such as whether the RBPs are helicases or known to bind rG4s. Interestingly, the observation applies to the RBP FXR2, which exhibits less co-location than expected in HepG2 but more co-location than expected in K562, despite being an rG4 Binding Protein (RG4BP).[73] Surprisingly, HNRNPL consistently exhibits low co-location with prG4s despite also being an RG4BP. This confirms that while some RBPs are known to bind rG4s, rG4s are not their only or primary target. A general observation is that all proportions are low. The highest proportion in both cell lines does not exceed 20%, indicating that RPBs do not primarily bind sites in close proximity to prG4s. Additionally, over half of the RBPs have one of their proportions under 2%. This is expected since RBPs bind numerous sites independently of prG4s. In most cases, RBPs are binding as much or more randomly relocated features than prG4s. Only 24 out of 103 RBPs HepG2 and 19 out of 117 RBPs in K562 have both their proportions above zero. These findings do not appear to result from the random generation of features, as shown in Supplemental Figure 8, where all 10 runs yield similar results.

We also examined other model organisms using RBPs binding sites data from the POSTAR database for *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (see Figure 6B–F).[67] Our first observation was the variation in annotations between these species. For instance, there were only 2 RBPs' CLIP data for *Danio rerio*, whereas more than 20 RBPs' CLIP data were accessible for *Saccharomyces cerevisiae* and *Mus musculus*. Despite these uneven annotations, most of the proportions are around 0, similar to those observed in humans. The only exception was the results for *Scer*, mainly due to the low number of predicted in this species. The primary explanation for these results is that only 26 rG4s were predicted in *Scer*.[70] With such a small number, the proportion range varied greatly, from −28% to 44%. Importantly, these results do not appear to be the result of poor random generation, as demonstrated in Supplemental Figures 9 and 10.

Overall, the co-location of binding sites and prG4s does appear to support co-evolution. To further investigate the co-evolution between RBPs and prG4s, the effort should concentrate on specific RBPs known to interact with prG4s and study them across many more species.

Discussion

Although pG4s are prevalent in numerous species, only a few pG4 families are conserved. Figure 3 illustrates that few pG4s are conserved (less than half are shared between *Hsap* and *Ptro*), they exhibit a higher level of conservation than expected by chance. In the case of bacteria, even though G4s are rarely predicted, those that are predicted and belong to a pG4 family tend to be conserved across different species. This suggests that pG4s in bacteria might play significant roles, given their conservation in multiple species. In contrast, eukaryotes show lower conservation, a finding consistent with a previous study using a similar methodology.[47] Nevertheless, beyond these 2 methods, there is a need to develop additional techniques for detecting pG4 families. Given that sequence alignment between pG4s is feasible in some cases (see Supplemental Figure 3), the development of a multiple alignment tool designed to maximize the alignment of G-tracts rather than entire sequences could be beneficial. Such an approach could initially use G-tracts to guide the alignment and then focus on aligning of loops, potentially enhancing the alignment quality between G4 sequences. Until now, no specific class of pG4s has been identified apart from motif-based classifications. G4s have traditionally been grouped based on different motifs: G4s with two G-quartets, G4s with two G-quarters and a loop of one nucleotide, G4s with three G-quartets and short loops, G4s with three Gs and long loops, and so on.[36,48,53] While these motif-based grouping provide insights into the distribution of different G4 motif types within datasets, we lack a viable and automated method for grouping and clustering G4s into groups of homologs. Clustering G4s based on metadata like their length, loop, k-mer occurrences in loops, etc., should be possible, especially with recent prediction tools employing machine learning.[29,30] This clustering approach could enhance our ability to study G4 evolution and find properties specific to each G4 cluster, such as potential RBP binding preferences for specific G4s types. Interestingly, in 45 species, more G4s are predicted and conserved through homology than via G4RNA screener (see Supplemental data Table 1). These G4s predicted by homology often correspond to non-canonical G4s with 2 quartets (see Supplemental Figure 3). It appears that non-canonical G4s with 2 quartets exhibit higher conservation compare to more stable G4s with 3 quartets or more. Several factors might explain this result, including the possibility that this motif is less constrained and thus more easily identified. Alternatively, there could be genuine evolutionary selection

favoring 2-tetrad pG4s, or theses 2-tetrad pG4s might be conserved to facilitate the evolutionary selection toward more stable G4s that required by certain species.

The prediction of G4s through homology appears to be a promising approach for discovering additional G4s. However, the viability of this method should first be experimentally confirmed. This method also highlights some limitations of G4RNA screener predictions in some species. For example, the initial prediction identified few G4s in *Scer* or *Ftub* (ie, *Saccharomyces cerevisiae* or *Francisella tuberculosis*), and the number increased with the homology prediction. This might be due either to G4RNA screener yielding false negatives, a known issue with the tool, or from the homology-based prediction being susceptible to false positives.[54]

The comparison of the gene tree with the pG4 family tree within these genes highlights instances where pG4s evolve differently from their host genes. This divergence may be attributed to pG4s experiencing distinct selection pressure compared to the entire gene. It is well-established that different regions within a gene may be under different pressures of selection. In Supplemental Figure 4, some pG4s trees closely resemble gene trees, while others do not. Those pG4s families that closely mirror the gene tree may not have an important role on their own and thus evolved similarly to their host gene. Conversely, pG4s trees with different topologies than their host gene could indicate that these pG4s have important functional roles. These pG4s may be subject to positive or negative pressure of selection depending on how they affect the gene function, resulting in a different evolution than the gene. This divergence could also be influenced by coevolution with other elements such as RBPs. However, it is essential to interpret these results cautiously, as genes with pG4s (both from the initial prediction and homology prediction) and those without pG4s were found as neighbors pG4 family trees. Non-parsimonious gain or loss of pG4s, as well as errors in the tree topology, might explain this phenomenon. Overall, it appears that stable pG4s are less conserved than unstable ones. This observation aligns with the number of G4s predicted through homology, although these predictions require an experimental confirmation. Notably, some specific cases revealed pG4 families with similar trees compared to their host genes, while in other cases, the opposite pattern was observed. The underlying selection pressure driving this evolution remains elusive but is pointing toward coevolution with RBPs. This coevolution could aid in either stabilizing less stable G4s or destabilizing highly stable ones.

Different studies showed have indicated the existence of differences in prG4 densities between eukaryotes and prokaryotes.[37,54] However, the reason remains unknown. The prevailing hypothesis is that prokaryotes possess fewer RBPs and thus, if an rG4 forms in a prokaryotic cell, it is more likely to remain in this state, independently of the cell needs. In contrast, eukaryotes, which have a geater abundance of RBPs, may have more dynamic regulation of rG4s. A study showed that bacteria living in hot environments have more pG4s than other closely related bacteria living in normal temperatures.[40] This suggests that in an environment where prG4s can fold and unfold freely, such as hot environments, more prG4s are present. This aligns with the concept that with more RBPs to help rG4 fold and unfold, there are more prG4s. Figure 5A and B appear to support this hypothesis, but Figure 5C shows a contrary trend in eukaryotes. Several explanations can be considered for these results. First, it is essential to recognize that a correlation does not imply the causality of an event. The correlation between the helicases ratio and prG4s density might occur by chance, and the strength or direction of this correlation could vary among different species groups we selected. Another possibility is that we may be examining at the wrong parameters to evaluate the coevolution of RBPs with prG4s. Our analyses has focused solely on helicases, while chaperone or other RBPs, might have a closer relationship with prG4s. Notably, since there are fewer helicases among RBPs in eukaryotes when the prG4 density becomes high, it might be due to the number of RBPs rising, thus lowering the helicase ratio. RBP2GO was used to retrieve gene ontology associated with RBPs, and only 2 chaperones were identified.[72] This limited availability of chaperone data may have hindered the investigation of the relationship between prG4s RBPs, although it remains an interesting lead for further research. Additionally, there could be other unaccounted-for factors influencing these results. In summary, a clear correlation exists between helicases ratio and prG4s densities across all species, supporting the notion of coevolution between rG4s and RBPs. Yet, for eukaryotes, this correlation is negative, suggesting that higher prG4 densities are associated with a lower ratio of helicases among RBP.

Based on these results, we looked at the co-location BSs and prG4s to determine whether most prG4s interacted with RBPs or not. We used eCLIP data to obtain BSs of 150 RBPs and compared them with normal pG4s and randomly relocated pG4s. Our analysis revealed that RBPs engage with prG4s in different ways, depending on the cell type and on species. This was expected, considering the distinct cellular environments and functions that necessitate different regulatory mechanisms. Surprisingly, even some known RG4BPs appeared to bind fewer prG4s than anticipated.to bind prG4s. For instance, HNRNPL, which has demonstrated interactions with G4s,[74] ranked lower in Figure 6. This outcome may be explained by another study indicating that HNRNPL preferentially binds regions rich in CA repeats,[75] which might explain a higher co-location with false prG4 exists compared to the normal dataset. Additionally, certain RBPs, despite their established prG4-binding capabilities, may bind only specific prG4s and a subset of them. Some other RBPs are expected to appear at the bottom of the list since they are known to interact with short non-coding transcripts where few G4s are predicted (eg, YBX3, LIN28B and LARP7). In Supplemental Figure 7, like some

RG4BPs, helicases exhibit diverse binding profiles contingent on the RBP and cell lines. Yet these results should be mitigated considering many points. Firstly, not all binding sites are comprehensively annotated. While eCLIP data for *Hsap* included 2 cell lines and were relatively comprehensive, data for other model organisms encompassed binding sites from various cell lines through multiple experimental methods (ParCLIP, HITCLIP and others). For instance, *Drer* had data available for only 2 RBPs, with one of them having just one BS. Thus, we limit the presented comparison to the currently available annotation. Also, some RBPs are absent from our analysis, even for *Hsap*, as more than 5000 RBPs are known,[72] but only 150 RBPs eCLIP data were available.

The identification of G4 families yields a complex view of G4 evolution. Most G4s are not grouped into G4 families; however, specific G4 families are found across bacteria or archaea species used in this study, indicating shared ancestral origins for these G4s. Additionally, our study reveals the presence of an intricate relationship exists between RBPs and pG4 density.

Author Contributions

AV, AO, and JPP conceived the study and its design. AV wrote the program, collected the data, ran the experiments, created the figures, and drafted the manuscript. AO and JPP critically revised the manuscript. All authors read and approved the final manuscript.

ORCID iD

Aïda Ouangraoua https://orcid.org/0000-0002-2040-4948

Data Availability

All information to retrieve the data and the scripts used for the analysis are available on the CoBIUS lab GitHub (https://github.com/UdeS-CoBIUS/G4Evolution).

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Cheong C, Moore PB. Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. *Biochemistry*. 1992;31(36):8406-8414.
2. Kim J, Cheong C, Moore PB. Tetramerization of an RNA oligonucleotide containing a GGGG sequence. *Nature*. 1991;351(6324):331-332.
3. Joachimi A, Benz A, Hartig JS. A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem*. 2009;17(19):6811-6815.
4. Bryan TM. G-Quadruplexes at Telomeres: Friend or Foe? *Molecules*. 2020;25(16): Article 16.
5. Kim N. The interplay between G-quadruplex and transcription. *Curr Med Chem*. 2019;26(16):2898-2917.
6. Reina C, Cavalieri V. Epigenetic modulation of chromatin states and gene expression by G-Quadruplex structures. *Int J Mol Sci*. 2020;21(11):4172.
7. Beaudoin J-D, Perreault J-P. Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res*. 2013;41(11):5898-5911.
8. Gomez D, Lemarteleur T, Lacroix L, et al. Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res*. 2004;32(1):371-379.
9. Jodoin R, Carrier JC, Rivard N, Bisaillon M, Perreault J-P. G-quadruplex located in the 5'UTR of the BAG-1 mRNA affects both its cap-dependent and cap-independent translation through global secondary structure maintenance. *Nucleic Acids Res*. 2019;47(19):10247-10266.
10. Lammich S, Kamp F, Wagner J, et al. Translational repression of the disintegrin and metalloprotease ADAM10 by a stable G-quadruplex secondary structure in its 5'-untranslated region. *J Biol Chem*. 2011;286(52):45063-45072.
11. Subramanian M, Rage F, Tabet R, et al. G-quadruplex RNA structure as a signal for neurite mRNA targeting. *EMBO Rep*. 2011;12(7):697-704.
12. Tassinari M, Richter SN, Gandellini P. Biological relevance and therapeutic potential of G-quadruplex structures in the human noncoding transcriptome. *Nucleic Acids Res*. 2021;49(7):3617-3633.
13. Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol*. 2020;21(8):459-474.
14. González V, Hurley LH. The c-MYC NHE III1: Function and Regulation. *Annu Rev Pharmacol Toxicol*. 2010;50(1):111-129.
15. Banerjee N, Chatterjee O, Roychowdhury T, et al. Sequence driven interaction of amino acids in de-novo designed peptides determines c-Myc G-quadruplex unfolding inducing apoptosis in cancer cells. *Biochim Biophys Acta Gen Subjects*. 2023;1867(2):130267.
16. Rogaeva E, Meng Y, Lee JH, et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat Genet*. 2007;39(2):168-177.
17. Imperatore JA, Then ML, McDougal KB, Mihailescu MR. Characterization of a G-Quadruplex structure in pre-miRNA-1229 and in its Alzheimer's disease-associated variant rs2291418: implications for miRNA-1229 maturation. *Int J Mol Sci*. 2020;21(3):767.
18. Rauchhaus J, Robinson J, Monti L, Di Antonio M. G-quadruplexes mark sites of methylation instability associated with ageing and cancer. *Genes*. 2022;13(9):1665.
19. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res*. 2006;34(14):3887-3896.
20. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*. 2005;33(9):2908-2916.
21. Bolduc F, Garant J-M, Allard F, Perreault J-P. Irregular G-quadruplexes found in the untranslated regions of human mRNAs influence translation. *J Biol Chem*. 2016;291(41):21751-21760.
22. Lim KW, Alberti P, Guédin A, et al. Sequence variant (CTAGGG)n in the human telomere favors a G·C·G·C tetrad. *Nucleic Acids Res*. 2009;37(18):6239-6248.
23. Mukundan VT, Phan AT. Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J Am Chem Soc*. 2013;135(13):5017-5028.
24. Doluca O. G4Catchall: A G-quadruplex prediction approach considering atypical features. *J Theor Biol*. 2019;463:92-98.
25. Hon J, Martínek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*. 2017;33(21):3373-3379.
26. Beaudoin J-D, Jodoin R, Perreault J-P. New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res*. 2014;42(2):1209-1223.
27. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res*. 2016;44(4):1746-1759.
28. Garant J-M, Perreault J-P, Scott MS. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics*. 2017;33(22):3532-3537.
29. Rocher V, Genais M, Nassereddine E, Mourad R. DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions. *PLoS Comput Biol*. 2021;17(8):e1009308.
30. Turner M, Barshai M, Orenstein Y. (2022). rG4detector: Convolutional neural network to predict RNA G-quadruplex propensity based on rG4-seq data. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–9. doi:10.1145/3535508.3545534
31. Millevoi S, Moine H, Vagner S. G-quadruplexes in RNA biology. *Wiley Interdiscip Rev RNA*. 2012;3(4):495-507.
32. Rouleau S, Jodoin R, Garant JM, Perreault JP. RNA G-Quadruplexes as key motifs of the transcriptome. *Adv Biochem Eng Biotechnol*. 2020;170:1-20.
33. Vannutelli A, Belhamiti S, Garant J-M, Ouangraoua A, Perreault J-P. Where are G-quadruplexes located in the human transcriptome? *Nar Genom Bioinform*. 2020;2(2):lqaa035.
34. Chambers VS, Marsico G, Boutell JM, et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol*. 2015;33(8):877-881.
35. Fernando H, Sewitz S, Darot J, et al. Genome-wide analysis of a G-quadruplex-specific single-chain antibody that regulates gene expression. *Nucleic Acids Res*. 2009;37(20):6716-6722.
36. Kwok CK, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. RG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods*. 2016;13(10):841-844.

37. Guo JU, Bartel DP. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*. 2016;353(6306):aaf5371.
38. Magis AD, Manzo SG, Russo M, et al. DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc Natl Acad Sci*. 2019;116(3):816-825.
39. Rider SD, Gadgil RY, Hitch DC, et al. Stable G-quadruplex DNA structures promote replication-dependent genome instability. *J Biol Chem*. 2022;298(6):101947.
40. Ding Y, Fleming AM, Burrows CJ. Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes. *Sci Rep*. 2018;8(1):15679.
41. Mishra SK, Tawani A, Mishra A, Kumar A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep*. 2016; 6(1):38144.
42. Bourdon S, Herviou P, Dumas L, et al. QUADRatlas: the RNA G-quadruplex and RG4-binding proteins database. *Nucleic Acids Res*. 2023;51(D1):D240-D247.
43. Antcliff A, McCullough LD, Tsvetkov AS. G-Quadruplexes and the DNA/RNA helicase DHX36 in health, disease, and aging. *Aging*. 2021;13(23):25578-25587.
44. He X, Yuan J, Wang Y. G3BP1 binds to guanine quadruplexes in mRNAs to modulate their stabilities. *Nucleic Acids Res*. 2021;49(19):11323-11336.
45. Sloan CA, Chan ET, Davidson JM, et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res*. 2016;44(D1):D726-D732.
46. Capra JA, Paeschke K, Singh M, Zakian VA. G-Quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS Comput Biol*. 2010;6(7):e1000861.
47. Frees S, Menendez C, Crum M, Bagga PS. QGRS-Conserve: a computational method for discovering evolutionarily conserved G-quadruplex motifs. *Hum Genomics*. 2014;8(1):8.
48. Wu F, Niu K, Cui Y, et al. Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Commun Biol*. 2021;4(1): 98-11.
49. Zhao Y, Du Z, Li N. Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett*. 2007; 581(10):1951-1956.
50. Mestre-Fos S, Penev PI, Richards JC, et al. Profusion of G-quadruplexes on both subunits of metazoan ribosomes. *PLoS One*. 2019;14(12):e0226177.
51. Bartas M, Čutová M, Brázda V, et al. The presence and localization of G-Quadruplex forming sequences in the domain of bacteria. *Molecules*. 2019; 24(9):1711.
52. Eddy J, Maizels N. Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res*. 2008;36(4):1321-1333.
53. Puig Lombardi E, Holmes A, Verga D, et al. Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species. *Nucleic Acids Res*. 2019;47(12):6098-6113.
54. Vannutelli A, Perreault J-P, Ouangraoua A. G-quadruplex occurrence and conservation: more than just a question of guanine-cytosine content. *Nar Genom Bioinform*. 2022;4(1):lqac010.
55. Vilella AJ, Severin J, Ureta-Vidal A, et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 2009; 19(2):327-335.
56. Lassmann T, Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. 2005;6(1):298.
57. Kikin O, D'Antonio L, Bagga PS. QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*. 2006;34(l_2): W676-W682.
58. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009; 25(11):1422-1423.
59. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and Phylogenetic Tree Building. *Mol Biol Evol*. 2010;27(2):221-224.
60. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33(6):1635-1638.
61. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981; 53(1):131-147.
62. Battistuzzi FU, Hedges SB. A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol*. 2009;26(2):335-343.
63. Cavalier-Smith T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*. 2002;52(Pt 2):297-354.
64. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256-W259.
65. Munoz R, Yarza P, Ludwig W, et al. Release LTPs104 of the all-species living tree. *Syst Appl Microbiol*. 2011;34(3):169-170.
66. Ule J, Jensen K, Mele A, Darnell RB. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*. 2005;37(4):376-386.
67. Weihao Z, Shang Z, Yumin Z, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res*. 2022;50(D1):D287-D294.
68. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of chip-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
69. Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*. 2008;9(1):192.
70. Vannutelli A, Schell L, Perreault J-P, Ouangraoua A. GAIA: G-quadruplexes in alive creature database. *Nucleic Acids Res*. 2023;51(D1):D135-D140.
71. Frigola J, Sabarinathan R, Mularoni L, et al. Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet*. 2017;49(12):1684-1692.
72. Caudron-Herger M, Jansen RE, Wassmer E, Diederichs S. RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. *Nucleic Acids Res*. 2021;49(D1):D425-D436.
73. Busa VF, Favorov AV, Fertig EJ, Leung AKL. Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding. *Cell Rep Methods*. 2021;1(6):100088.
74. Mori K, Lammich S, Mackenzie IR, et al. HnRNP A3 binds to GGGGCC repeats and is a constituent of p62-positive/TDP43-negative inclusions in the hippocampus of patients with C9orf72 mutations. *Acta Neuropathol*. 2013; 125(3):413-423.
75. Blatter M, Dunin-Horkawicz S, Grishina I, et al. The signature of the five-stranded vRRM fold defined by functional, structural and computational analysis of the hnRNP L protein. *J Mol Biol*. 2015;427(19):3001-3022.