UDS Service des bibliothèques et archives

Savoirs UdeS
DÉPÔT INSTITUTIONNEL

Where are G-quadruplexes located in the human transcriptome?

Anaïs Vannutelli, Sarah Belhamiti, Jean-Michel Garant, Aida Ouangraoua and Jean-Pierre Perreault

Conditions d'utilisation

This is the published version of the following article: Vannutelli A, Belhamiti S, Garant JM, Ouangraoua A, Perreault JP. (2020) Where are G-quadruplexes located in the human transcriptome? NAR Genomics and Bioinformatics, Volume 2, Issue 2, June 2020 which has been published in final form at https://doi.org/10.1093/nargab/lqaa035. It is deposited under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/).

Cet article a été téléchargé à partir du dépôt institutionnel *Savoirs UdeS* de l'Université de Sherbrooke.

Published online 25 May 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 2 **1**
doi: 10.1093/nargab/lqaa035

Where are G-quadruplexes located in the human transcriptome?

Anaïs Vannutelli[1,2], Sarah Belhamiti[1,2], Jean-Michel Garant[2], Aida Ouangraoua[1,*] and Jean-Pierre Perreault[2,*]

[1]Department of Computer Science, Faculté des sciences, Université de Sherbrooke, QC J1K 2R1, Canada and [2]Department of Biochemistry and Functional Genomics, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, QC J1E 4K8, Canada

Received February 03, 2020; Revised April 28, 2020; Editorial Decision May 02, 2020; Accepted May 05, 2020

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

ABSTRACT

It has been demonstrated that RNA G-quadruplexes (G4) are structural motifs present in transcriptomes and play important regulatory roles in several post-transcriptional mechanisms. However, the full picture of RNA G4 locations and the extent of their implication remain elusive. Solely computational prediction analysis of the whole transcriptome may reveal all potential G4, since experimental identifications are always limited to specific conditions or specific cell lines. The present study reports the first in-depth computational prediction of potential G4 region across the complete human transcriptome. Although using a relatively stringent approach based on three prediction scores that accounts for the composition of G4 sequences, the composition of their neighboring sequences, and the various forms of G4, over 1.1 million of potential G4 (pG4) were predicted. The abundance of G4 was computationally confirmed in both 5′ and 3′UTR as well as splicing junction of mRNA, appreciate for the first time in the long ncRNA, while almost absent of most of the small ncRNA families. The present results constitute an important step toward a full understanding of the roles of G4 in post-transcriptional mechanisms.

INTRODUCTION

G-quadruplexes (G4, a list of acronyms is available at the end of the article) are stable, non-canonical secondary structures formed by guanine-rich nucleotide sequences (for reviews see refs (1,2)). Four guanines linked by Hoogsteen base-pairs form each G-tetrad. These tetrads then stack on each other, and the resulting structure is stabilized by a monovalent cation, usually potassium. G4 were first described as sequences corresponding to the canonical motif $G_xN_{1-7}G_xN_{1-7}G_xN_{1-7}G_x$, with the four tracks of guanines forming at least 3 G-tetrads (i.e. $x \geq 3$) that are separated by loops of 1–7 nucleotides (nt) in length. This motif defines the category of canonical G4. Motifs that do not answer to this definition of canonical constitute the category of non-canonical G4.

DNA G4 were shown to be involved in several biological functions by both computational predictions and experimental studies (for reviews see refs (1,2)). These structures were first discovered in telomeres, where it was shown that telomere end-binding proteins can influence the formation of G4 (3) involved in maintaining chromosome stability. It was also shown that G4 located in telomeres are good candidates as targets for anticancer therapeutics because their stabilization can inhibit the telomerase enzyme that is overexpressed in the majority of cancers (4). Some studies have demonstrated a role for G4 in the control of gene expression by acting on transcription (1,5). In particular, Huppert *et al.* (5) demonstrated a 40% enrichment in potential G4 (pG4) sequences in promoter regions as compared to the rest of the genome. Moreover, the use of a structure-specific antibody for the detection of G4 demonstrated that the folding into a G4 structure could be correlated with the cell cycle, and that it could be stabilized by a small-molecule ligand (6).

RNA G4 (rG4) have been shown to be more stable than DNA G4, a property resulting from the presence of the 2′-hydroxyl group, which contributes to additional stabilizing interactions (2). Moreover, the 2′-hydroxyl group of the ribose locks the RNA in an *anti* conformation, favoring the parallel topology in which all four strands have the same direction. Increasing evidence shows that G4 are abundant in mRNA, and that they play crucial regulatory roles in numerous post-transcriptional mechanisms such as splicing, polyadenylation, miRNA binding, transcription termination and translation (2,7). More recently, G4 structures have been reported in the hairpin of some primary miRNA (pri-

*To whom correspondence should be addressed. Tel: +1 819 821 8000 (Ext. 62014); Fax: +1 819 821 8200; Email: aida.ouangraoua@usherbrooke.ca
Correspondence may also be addressed to Jean-Pierre Perreault. Tel: +1 819 821 8000 (Ext. 75310); Fax: +1 819 820 6831; Email: jean-pierre.perreault@usherbrooke.ca

© The Author(s) 2020. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

miRNA), suggesting a role in miRNA maturation (8). A study on transcriptome performed using a transcriptome-wide G4 profiling method (i.e. rG4-seq; (9)) revealed the presence of thousands of G4 motifs in a specific cell line at a defined time. Thousands of mammalian RNA regions have been demonstrated to fold into G4 structures, but, in contrast to previous beliefs, these regions were found to be overwhelmingly unfolded in cells under the specific conditions used (i.e. the presence of K^+ instead of Na^+) (10). Based on these results, it has been suggested that eukaryotes have a robust machinery that globally unfolds RNA G4; a hypothesis that is currently being under investigation. The current belief is that several RNA binding proteins should be involved in the coordination of the folding and unfolding of the G4 motifs. This would have for effect to allow regulation of the formation of G4 RNA in a transient manner, i.e. only when required for a specific regulation event. Clearly, the increasing number of studies on RNA G4, and their various possible roles, call for more investigation on G4 in transcriptomes. In order to achieve this end, locating and evaluating the abundance of pG4 in whole transcriptomes is a pre-requisite, but it has never been meticulously studied for any species. The present study shows that rG4 are widely present in the transcriptome and in all known classes of transcripts, with 60% of transcripts having at least one pG4.

Several studies have been performed in order to evaluate the number of pG4 sequences in genomes and transcriptomes. For example, computational methods based on the canonical G4 motif for pG4 detection in the human genome led to the identification of around 360 000 pG4 (11). More recently, a high throughput method of DNA G4 detection revealed that the first estimation of the number of pG4 in the human genome was in fact underestimated by about 2-fold. In all, 525 890 pG4 were detected under physiological conditions, and 716 310 in the presence of a G4-specific ligand (12). Under the latter condition, 63% of the observed G4 sequences (OQ) would not have been discovered by the computational method based on the canonical motif. This method has also been applied to numerous species (13). This study showed that, depending on the species, between 60 and 80% OQ do not match to the canonical motif. The higher number of G4 discovered by sequencing methods, and the number of OQ missed by the canonical motif-based method, confirms that the current definition of G4 as canonical motifs is obsolete. For example, it was shown that an increase of the length of the loops in a G4 structure up to a maximum of 15 nt does not influence the ability of the sequence to fold (14). This shows that the upper limit of the loop length at 7 nt is too stringent and should be increased. Together, these results on DNA G4 prediction call for an extension of the definition of G4 and a revision of prediction methods.

RNA pG4 have been identified using the computational and experimental tools developed for their DNA counterparts. However, it was shown that several false positive pG4 located in the 5′UTRs of human mRNAs were detected by computational methods (15). This occurred mainly because the DNA pG4 detection methods do not account for RNA folding. It was hypothesized that the presence of cytosine (C) in the environment around the G4 led to a competition between the Watson–Crick G-C base pairs of the RNA folding structure and the Hoogsten G-G base pairs of the G-tetrads (16). Thus, the folding of a RNA G4 sequence depends on several conditions including the sequence composition, the neighboring sequences composition, and the folding into another secondary structure. The combination of all these features makes the computational prediction of RNA G4 a hard task that may results in several false positives. This observation led to the development of a scoring system called cG/cC for the prediction of rG4 which considers a favorable or an unfavorable environment for the folding of the G4 depending on the number of cytosines in the neighboring regions (15). Based on the same logic, G4Hunter, an algorithm for pG4 detection that takes into account both the G-richness and the C-skewness of a given sequence that produces a quadruplex propensity score, was developed (17).

Moreover, like for DNA, some RNA sequences were demonstrated to fold into a G4 structure without corresponding to the canonical motif. For instance, human rG4 with a large central loop (10–70 nt in size) and with the other two loops having lengths of 1 nt each were characterized (18). Moreover, some rG4 with the first or the last loop lengths being up to 40 nt were reported (19). In addition, it was shown that some sequences can fold into G4 structures despite the presence of a bulge between the guanosine residues forming the G-track (20). This led to the proposition of bulges as possible binding sites for interaction between the G4 structure and other molecules (21). In addition, several physical evidences supporting the folding of rG4 based only on two tetrads were reported (22–25). The existence of non-canonical RNA G4 illustrates the limitation of the canonical definition of past motif-based computational tools yielding to incomplete detection.

Both the cG/cC and G4Hunter (G4H) scores were proposed as possible solutions for the detection of canonical and non-canonical pG4s, but that they are not sufficient to identify all forms of G4 (26). In order to overcome this limitation, we recently developed a new detection tool that does not rely on a motif definition for RNA sequences (26). An artificial neural network was trained with sequences of experimentally validated G4 obtained from the G4RNA database (27). The resulting prediction score, G4NN, has a predictive power comparable to the reported G richness and G/C skewness evaluations (i.e. the cG/cC scoring system and the G4H score) that are the current state-of-the-art for the identification of RNA pG4. Consequently, the three non-motif-based prediction scores, cG/cC, G4H and G4NN were combined to provide G4RNA screener, a program designed to evaluate sequences in order to identify RNA pG4 (28). Recently, G4-iM, a G4 search engine that use a combination of both G4 motifs and GC content with cGcC and G4H, has been developed (29). G4-iM returns a score accounting for the motif and the GC content by making the average between the G4Hunter score and the PQS-finder score. It also returns the genomic frequency of the predicted G4. Those two features are specific to G4-iM and provide fast results on tested sequences compared to other tools that require specific analyses.

In this study, a complete analysis of the locations and numbers of pG4 in the whole human transcriptome based

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

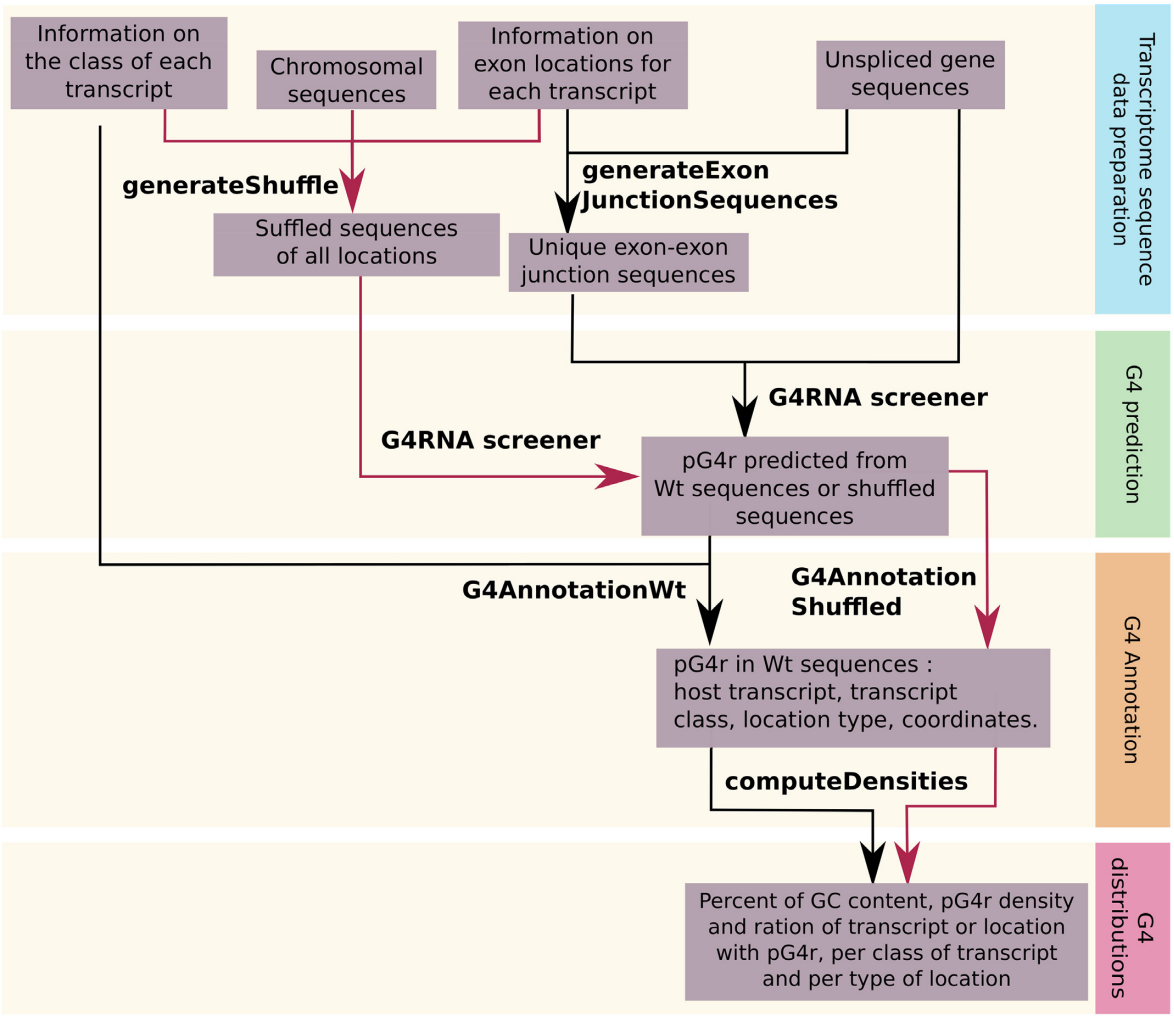

Figure 1. Overview of the methodological pipeline. Black arrows are for the WT dataset and red arrows are for the shuffled data set. Scripts names are in bold characters.

on the computational detection of pG4 using G4RNA screener is presented.

MATERIALS AND METHODS

A methodology was designed for the detection and the analysis of pG4 abundance and location in the human transcriptome. An overview of the four steps of the method is presented in Figure 1. Step 1 consists of the transcriptome sequence data preparation. Step 2 is the pG4 detection phase. Step 3 consists of the annotation of the detected pG4 regions with their location and the class of transcript/gene that contains them. Finally, Step 4 is the analysis of the pG4 distribution across the various classes of transcripts and types of location. The details of each step of the methodology are provided below.

Step 1: Transcriptome sequence data preparation

In order to identify pG4 in the human transcriptome, data were downloaded from the Ensembl genome database for vertebrates and other eukaryotic species hosted by the Wellcome Trust Sanger Institute (WTSI) (30). The data were downloaded for the human genome assembly version GRCh38p12 using the Biomart interface, a database system for flexible querying based on data-agnostic modeling (31) and through the ensembl FTP server.

The first data contains the sequences of all annotated unspliced genes along with their genomic coordinates (chromosome, strand, start and end positions) for all classes of genes. The second data contains information on all human spliced transcripts: their host genes, and the genomic coordinates of their translated/untranslated exons. Note that 1 751 poorly annotated transcripts (for example, non-coding transcripts with UTR) were removed from the analysis. The third data contains information on the classes of human genes and spliced transcripts (coding, non-coding, pseudogene, etc.). The information on the transcript class is required because some transcripts may belong to a class different from their host gene class. The Ensembl transcript/gene class is based on the Vertebrate Genome Annotation (*Vega*) database, a repository for high-quality gene models produced by the manual annotation of vertebrate genomes. Details on the definition of the different

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

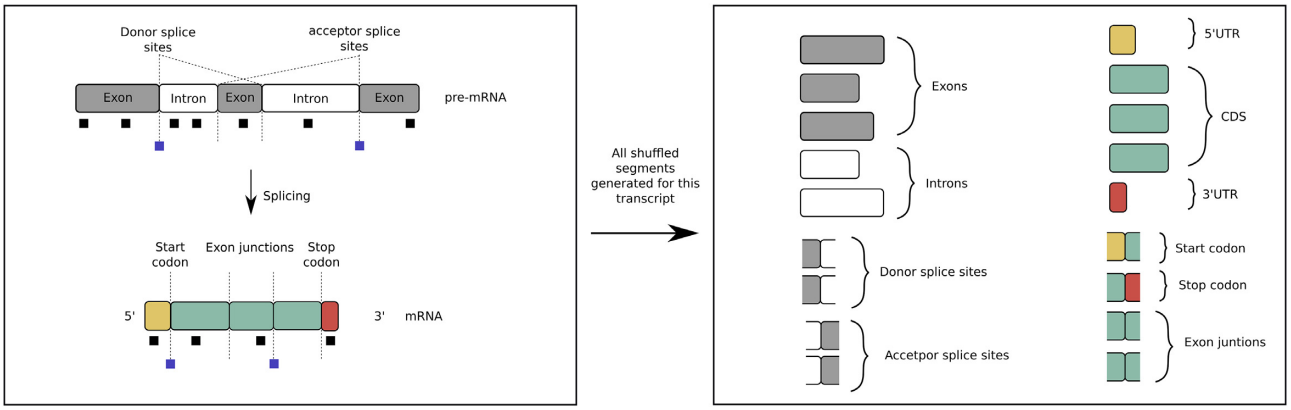

Figure 2. Locations of pG4r in transcripts. The possible locations of pG4r in a transcript are depicted on the left panel while on the right panel all shuffled locations are shown. The segmental locations are represented in gray for exons, yellow for 5′UTR, green for CDS and red for 3′UTR and white for introns. Black boxes are pG4r in segment locations. The point locations are represented in dashed lines at the intersection of two segmental locations and blue boxes are pG4r in point locations. Point locations overlap the start codon, the stop codon, the exon–exon splicing junctions, the donor splice sites and the acceptor splice sites. The G4 represented on the pre-mRNA are not displayed on the mature mRNA.

classes of genes and transcripts can be found in ref. (32). The fourth data contains the sequences of entire chromosomes.

From the unspliced gene sequences and the information on the spliced transcripts exon locations, all sequences of unique exon–exon splicing junctions belonging to the human transcriptome were generated with the genomic coordinates of the intron spliced between the two exons. For each unique exon–exon junction, the sequence of the junction consists of the concatenation of the two 100-nt exon subsequences located upstream and downstream of the junction. The set of sequences of unique exon–exon junctions constitutes the fifth dataset.

Two classes of location types for pG4 in RNA were considered: segmental and point locations. Segmental locations are segments corresponding to 5′UTR, 3′UTR, CDS, Intron or Exon. Point locations are segments overlapping codon start, codon stop, splicing junction, donor splice site or acceptor splice site (see Figure 2 for an illustration). For each segment corresponding to a location, a shuffled counterpart was obtained by generating a segment with the same content in nucleotide but in a random order (see Figure 2 for an illustration). Thus, a location segment and its shuffled counterpart have the same nucleotide content, and thus the same GC content. The shuffled sequence dataset was generated in order to allow the comparison of pG4 prediction between the wild-type (WT) sequence dataset and the shuffled sequence dataset, and to allow investigating the correlation between the frequency of pG4 and the GC content of sequences. The set of shuffled sequences for all locations in the transcriptome constitutes the sixth dataset. Segments were shuffled 10 times to generate 10 shuffled datasets. The pG4 predictions for all datasets were very similar (Supplementary Tables S1–7), so we kept a single shuffled dataset composed of a single shuffled segment for each segment corresponding to a location in the WT dataset.

Step 2: G4 prediction

G4RNA screener was launched on the full unspliced gene sequences and the sequences of the unique exon–exon splicing junctions constituting the WT sequence dataset, as well as on the shuffled sequence dataset. The screening used a sliding window approach, with a window length of 60 nt and a step of 10 nt (i.e. the recommended default parameters) (26).The threshold score limits for pG4 detection were set to 4.5 for the cG/cC score, 0.9 for G4H and 0.5 for G4NN (as recommended by (26) for stringent detection). G4RNA screener parameters were determined during the development of the tool by testing it on different data and via a battery of tests on a wide range of parameters values (for further details see (25)). The windows length is longer than what is expected for most G4, but it helps to account of the environment of the G4, which has been shown to play an important role in the folding (15,16). This allows to lower the rate of false positive. Successive windows with G4 scores higher than the three thresholds were grouped and returned as pG4 regions (pG4r) that can contain one or more pG4. The two sets of detected pG4r located in the unspliced gene sequences and in the exon–exon junction sequences were merged in order to detect and remove redundancy. The result of this step was a list of unique pG4 regions in human gene sequences and transcript exon–exon junction sequences with their G4 scores and genomic locations, for the WT dataset and the shuffled dataset.

Step 3: G4 annotation

Previous studies have reported an abundance of G4 in specific locations in genes, such as the 5′UTR, the 3′UTR and in the splice sites (5,12,16). In order to study the distribution of pG4r in the whole human transcriptome, in specific classes of transcripts and in the various types of locations in a transcript, an annotation of all pG4r detected in the WT transcriptome dataset was performed. The list of pG4r obtained at the end of Step 2 was mapped on gene sequences and transcript sequences in order to obtain the list of pG4r in genes and the list of pG4r in transcripts with their genomic coordinates. The third data file from Step 1 containing the information on the gene and transcript classes was then used to add the class of gene or transcript to the annotation of each pG4r. Next, based on the status of the tran-

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

scripts, which is protein-coding or non-coding, the pG4r were annotated with their type of location in the transcript. In protein-coding transcripts, the pG4r can be located in the 5′UTR, the 3′UTR, the coding region (CDS), the start codon, the stop codon, the intron, the splicing junctions, the donor splice sites or the acceptor splice sites. In non-coding transcripts, the various types of location in coding transcript exons (i.e. 5′UTR, 3′UTR, CDS, codon start, codon stop) are replaced by a unique type of location which is Exon.

Step 4: Analysis of pG4 distribution

In order to study the pG4r distribution in the transcriptome, the density of pG4r for each type of location in transcripts was evaluated for the WT dataset and the shuffled dataset. The density of pG4r in a type of segmental location S (5′UTR, 3′UTR, CDS, Intron or Exon) was calculated as:

$$\text{Density} = \frac{\text{numberof}\,pG4r}{\text{totallengthoflocation\_}S} \times 1000\,(pG4/kb)$$

The density of pG4r in a type of a point location P was calculated as:

$$\text{Density} = \frac{\text{numberof}\,pG4r}{\text{totalnumberoflocation\_}P}$$

For example, Figure 2 shows the possible locations of G4 in a transcript. In the illustration, the mRNA transcript has seven G4 located in segmental locations, three at point locations on the pre-mRNA form and one more that appears in the mRNA after splicing. The different pG4r densities of segmental locations are: $\frac{1}{length_{5'UTR}}$, $\frac{1}{length_{3'UTR}}$, $\frac{1}{length_{CDS}}$ and $\frac{2}{lenght_{Intron}}$, $\frac{5}{lenght_{Exon}}$. At point locations, the densities are 1 at the codon start, 0.5 at the splicing junction, the donor splice site, the acceptor splice site, and, finally, a null value at the codon stop.

The global density for each type of location was computed within the four classes of transcripts defined in the *Vega* database: (i) the mRNA class (Coding) that contains protein-coding transcripts with an open reading frame (ORF); (ii) the ncRNA class that contains functional RNA that are not translated into proteins, but have roles in the regulation of gene expression at both the transcriptional and the post-transcriptional levels; (iii) the pseudogene class that contains transcripts similar to known protein-coding transcripts, but that contain a frameshift and/or stop codon(s) which disrupt the ORF; and (iv) the predictive class that has been specifically created in the ENCODE project to highlight potential regions that could contain protein-coding genes and require experimental validation. The ncRNA class is further divided in two classes: short ncRNA which are ncRNA shorter than 200 nt, and long ncRNA which are ncRNA longer than 200 nt.

Each of the five transcript classes was further refined in order to consider various subclasses of transcripts. A summary of the classes and subclasses of transcript considered in the analysis is provided in Supplementary Table S8. The detailed definitions of the subclasses are provided in the 'Results' section. For each transcript class or subclass, and for each type of location in transcripts, the ratio of transcripts in the class containing at least one pG4r in this type of location was reported. Only groups containing enough transcripts, at least 50 transcripts, were considered in the analysis.

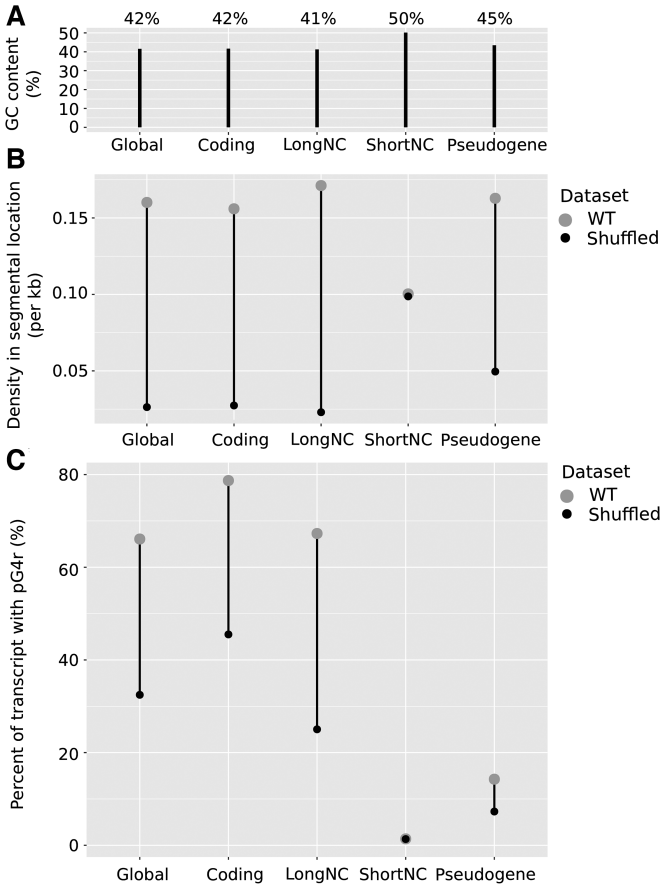

Figure 3. PG4r densities in the human transcriptome and RNA classes. (**A**) GC content, (**B**) pG4r density and, (**C**) ratio of transcripts with pG4r, for the whole transcriptome and each RNA class (coding, long non-coding, short non-coding, pseudogene and predicted). For (B) and (C), the values are shown for the WT dataset (gray points) and the shuffled dataset (black points).

RESULTS AND DISCUSSION

Prevalence of pG4r in the human transcriptome

The application of the method led to the finding of 1 133 712 RNA pG4r, corresponding to 308 745 unique pG4r, in the human transcriptome and in the splicing junctions. Out of the 197 496 transcripts constituting the human transcriptome, 122 845 transcripts contain pG4r. Thus, 60% of human transcripts have at least one pG4r in their sequences (Figure 3C). This high number of pG4r in the human transcriptome agrees with previous studies that have shown the prevalence of potential G4 in the human genome (11,33). The present study reports an even more significant prevalence of G4 in the human transcriptome.

In order to validate this result, the set of pG4r found in the present study was compared to the set of rG4 detected by Kwok *et al.* (9). In the latter, the rG4 were predicted us-

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

ing an experimental approach (rG4-seq) based on a profiling method of rG4 that couples rG4-mediated reverse transcriptase stalling (RTS) with next-generation sequencing. The study was performed on polyadenylated-enriched RNA from the HeLa cell line. Even though the detection was done experimentally, and was limited to a specific cell line, the results of this study are the most suitable for comparison with the predictions made here. In the rG4-seq study, two conditions were used to detect folded G4: the presence of either potassium (K^+) or K^+ with a stabilizer ligand (PDS) (34).

The comparison of pG4r and rG4 datasets was done using blastn (v2.6.0+). Blastn with minimum word length 11, and no mismatch allowed, was used two times, one with rG4 as query and pG4r as target, and conversely. Both datasets (rG4 and pG4r) were filtered in order to get the more comparable datasets possible. It is important to recall that the rG4-seq detection was performed experimentally only on mRNA from a specific cell line, i.e. polyadenylated-enriched RNA from the HeLa cell line. Thus, the first filter kept only rG4 and pG4r that are in transcripts detected in the rG4-seq study. This allows to get a common set of transcripts. The second filter is the removal of all locations not contained in rG4-seq study, i.e. introns or exon–intron junction sites. Surprisingly, rG4 were detected experimentally using RTS but we found that 12% of those rG4 did not have enough guanine residues to form a G-track. A last filter was then applied to remove RTS sites that did not match a minimal G4 motif. This motif was defined to be the less restrictive possible, i.e. $G_2X_nG_2X_nG_2X_nG_2$. It matches any two-tetrad G4 (22–25) with no restriction on loop lengths. In order to infer a match between a pG4r and a rG4, only hits within the first 5 hits of blastn were kept. Then, hits on sequences with different strand or chromosome were removed.

We found 52.8% of rG4 matching pG4r, but only 22.4% of pG4r matching rG4 (Table 1) for the K^+ condition. It is not surprising to find only 52.8% of rG4 corresponding to pG4r as described in (26). Yet, 22.4% of pG4r corresponding to rG4 was not expected. This could be due to several reasons: (i) some G4 are just folded transiently and not very stable so they may not be detected in K^+ but in K^+PDS; and (ii) RTS sites are detected by comparing K^+/K^+PDS condition against Li^+ condition, which do not allow the detection of rG4 able to fold in the Li^+ condition, but they could be predicted by G4RNAScreener. Further drawbacks of the rG4-seq detection method are discussed in (35). This justifies the increase from 22.4 to 44.8% in the K^+PDS condition of the percent of pG4r matching to rG4. Moreover, in the rG4-seq study, they report that only 18% of G4 predicted using motifs (canonical and non-canonical) correspond to rG4 predicted in K^+ condition (9). The 22.4% proportion reported in our study is comparable to their 18%, which increase to 26% in K^+PDS condition.

Comparison of the distribution of G4 in RNA classes

The distributions of pG4r in the five transcript classes (mRNA, long ncRNA, short ncRNA, pseudogene and predicted transcripts) were compared without accounting for the type of location in the transcript. The aim was to evaluate if pG4r were under- or over-represented in specific classes (Figure 3). Comparable pG4r densities are observed for mRNA, long ncRNA and pseudogenes. The percentage of transcripts containing at least one pG4r is lower for the pseudogene class as compared to those three other classes. This low rate means that although the density of pG4r is similar in the classes, pG4r are carried by a smaller fraction of transcripts in the pseudogene class. Interestingly, there are a low density of pG4r predicted in short ncRNA (0.1/kb) while it is the class of transcripts with the higher GC content (50.2%). We also observe that the pG4r density for the shuffled dataset is always lower than the density for the WT dataset, except for short ncRNA. The same observation can be made for the ratio of transcripts containing pG4r. It means that we find more pG4r in mRNA, long ncRNA and pseudogenes than would be expected randomly, and same pG4r in short ncRNA than would be expected randomly. We observe that pG4r density in shuffled segments of short ncRNA is higher than in shuffled segments of other RNA classes. This can be due to the high GC content in short ncRNA and the small length of segments (<200 nt).

G4 enrichment in 5′UTR and at donor splice sites in the mRNA class

Inside the mRNA transcript class, the distribution of pG4r in the various types of locations was evaluated. The mRNA class can be divided into three subclasses of transcripts based on the transcript type (see also Supplementary Table S1):

i) The nonstop decay (NSD) mRNA are transcripts without a stop codon at their 3′ end that will be eliminated by the NSD surveillance pathway;
ii) The nonsense-mediated decay (NMD) mRNA are mutant transcripts (nonsense mutation) that will be eliminated by the NMD surveillance pathway, thus preventing the formation of deleterious proteins;
iii) Protein coding (PC) mRNA are all other mRNA transcripts that result in protein.

These subclasses are independent. NMD and NSD are transcripts that should have been translated into protein. However, due to the presence of some mutations, these transcripts are degraded prior to translation. Other subclasses of mRNA exist like T-cell receptor genes and immunoglobulin genes, but they have highly variable sequences between individuals. For this reason, they are not studied here. However, it should be noted that some studies have reported G4 structures in immunoglobulin DNA and RNA (36,37).

The distribution of pG4r in the remaining subclasses, PC, NMD and NSD, were first compared independently of their location types in transcripts (Figure 4). A comparable distribution for PC and NMD was observed (∼0.16 per kb), with pG4r contained in 85.3% and 77.5% of the transcripts, respectively. However, the NSD subclass has a lower pG4r density (0.1 per kb) with pG4r in 65.5% of the NSD transcripts.

Next, the pG4r distributions in different location types in the transcripts were evaluated in order to investigate the differences between PC, NMD and NSD, and to verify if the pG4r density was uniform along the transcripts. As de-

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

Table 1. Match between pG4r and rG4 datasets

Condition	Dataset	G4 Number	Number of hits with the other dataset	% of correspondance
K^+	rG4	3 114	1644	52.8
	pG4r	3 181	712	22.4
K^+PDS	rG4	10 481	3727	35.6
	pG4r	5 558	2490	44.8

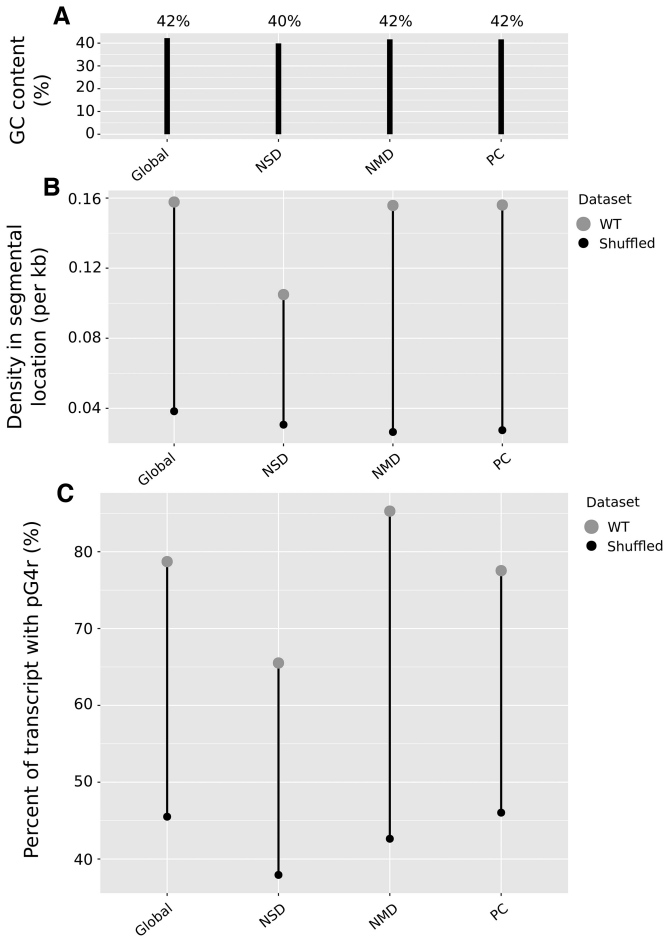

Figure 4. PG4r in some mRNA subclasses. (**A**) GC content, (**B**) pG4r density and (**C**) ratio of transcripts with pG4r, for the whole transcriptome and each subclass. Subclasses are NSD, NMD and PC. For (B) and (C), the values are shown for the WT dataset (gray point) and the shuffled dataset (black point).

scribed earlier, the pG4r were subdivided into several collections depending on their location type: exons, introns, 5′UTR, 3′UTR and CDS for segmental location, and start and stop codons, donor and acceptor splice sites and splicing junctions for point locations. The densities of pG4r for each location type are presented in Figure 5. The densities are comparable between the intron and exon locations for the whole mRNA class level (i.e. 0.16 and 0.15 per kb), and inside each subclass (i.e. 0.16–0.15 per kb for PC; 0.17–0.12 per kb for NMD; and 0.10–0.10 per kb for NSD). Except for the latter case, the results support that the distribution of pG4r is independent of the mRNA subclass. These results do not provide an explanation for the low G4 density observed in the NSD subclass. Note that the NSD RNA do not have a 3′UTR region because they have no stop codon.

Subsequently, the pG4r density within the different exon locations (5′UTR, 3′UTR and CDS) was evaluated (see Figure 5). For the whole mRNA class, the pG4r density in the 5′UTR is greater than that in the 3′ UTR which in turn is greater than the density in the CDS region (respectively, 0.53, 0.19 and 0.12 per kb). The same result is observed for the subclasses PC and NMD. It should be noted that the NSD subclass has a null density in pG4r in the 3′UTR location due to the absence of the 3′UTR region.

The abundance of G4 sequences in the 5′UTR has also been demonstrated at the genomic level in several previous studies (11,12). We also described previously an overrepresentation of G4 in the 3′UTR of genes in the case where the 5′UTR of a second gene is found close to the 3′end of the first gene (38). Our interpretation was that there was a possible role for G4 in the termination of gene transcription. In the same direction, the enrichment in pG4r observed in the two untranslated regions (5′UTR and 3′UTR) of transcripts could support the hypothesis that G4 are involved in the translational control at both the start and the end of transcripts. Regarding the NSD subclass, a similar depletion in all types of segmental locations was observed as compared to all other subclasses.

Concerning the distribution of G4 in the point locations, an enrichment of G4 is observed at the start codon and at the donor splice site location for the whole mRNA class, and for both the PC and the NMD subclasses. The NSD subclass again presents an exception with a low pG4r density at the start codon location. Beside this exception, the G4 density in point locations agrees with the high density reported in the 5′UTR located adjacent to the start codon.

The comparison between the results for the WT and the shuffled datasets shows that, despite the large densities of pG4r in 5′UTR and start codon locations, there are less pG4r predicted in these locations than would be expected randomly.

Competition between G4 and canonical secondary structure in short ncRNA

The non-coding transcripts category is subdivided into two groups: short (<200 nt) and long (>200 nt) ncRNA. The short ncRNA group is composed of transcripts from eight subclasses. Among these eight classes, only two were kept to avoid classes with less than 50 annotated transcripts. Here is a description of the two classes kept (see also Supplementary Table S1):

i) microRNA and their precursors (pre-miRNA). pre-miRNA is the precursor of miRNA, which modulates the translation of some transcripts by the cleavage, destabilization or less efficient translation of the reverse complement transcript.

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

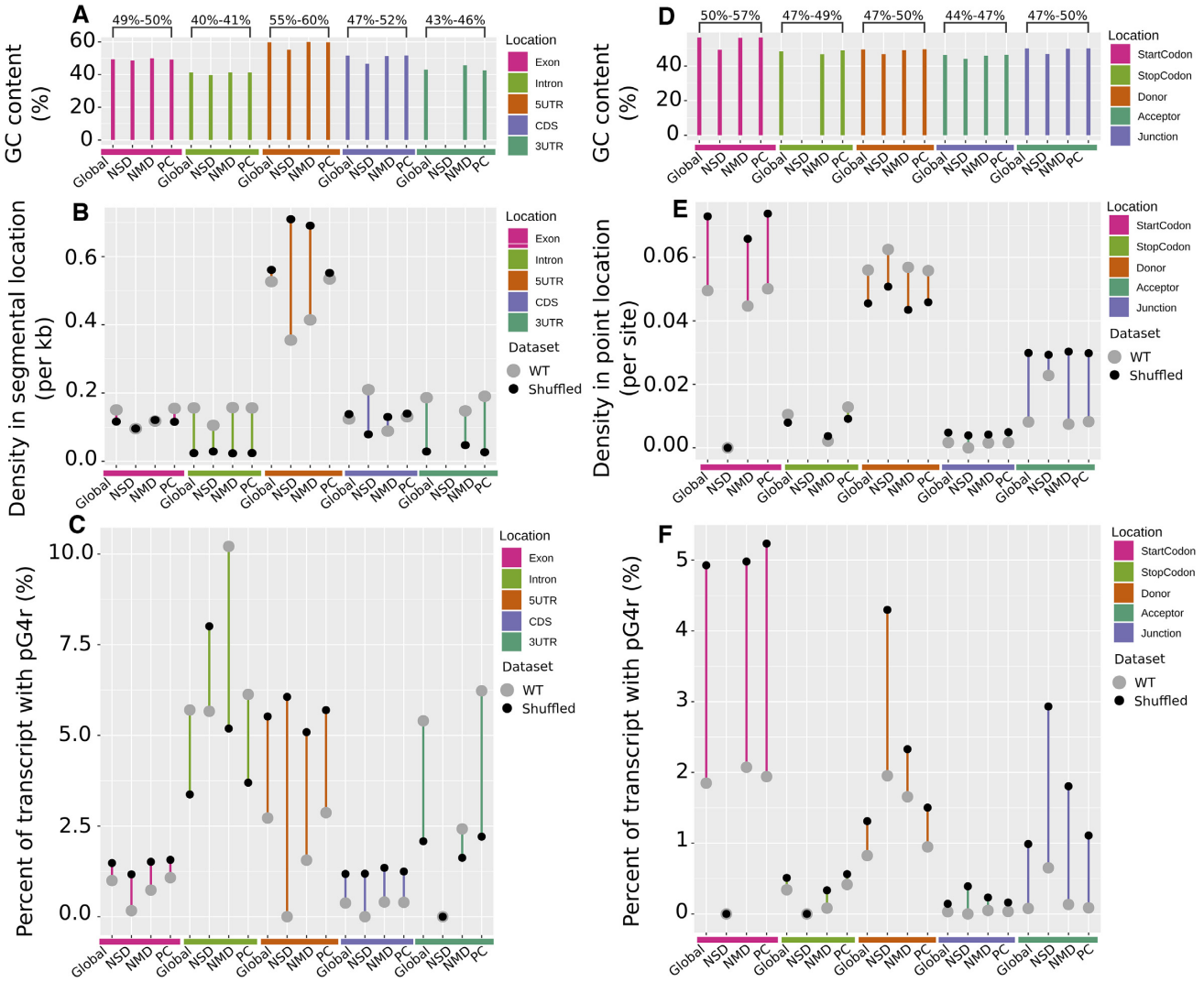

Figure 5. PG4r in some mRNA subclasses by location. (**A**) GC content, (**B**) pG4r densities and (**C**) percent of transcript with pG4r for each segmental location (exon, intron, 5′UTR, CDS and 3′UTR) and for each subclass. Subclasses are NSD, NMD and PC. Same graphics are made for point location (donor, acceptor and junction), respectively (**D**), (**E**) and (**F**). For (B), (C), (E) and (F) values are shown for WT dataset (gray points) and for shuffled data set (black points).

ii) miscellaneous RNA (miscRNA) are short ncRNA that seem to be transcribed but cannot be classified at the moment.

The long ncRNA group is composed of longer transcripts from various subclasses that are described later.

A depletion in pG4r is observed in the short ncRNA, which exhibits a density of 0.10 per kb (Figure 3). This low density in pG4r in short ncRNA highlights that this group of transcripts is not a favorable host for G4 structure. It has been shown that short ncRNA harbor-specific secondary structures that are related to their functions (39). Thus, the low density of pG4r in this group can be explained by the competition between the G4 structures and the short ncRNA secondary structures.

In order to investigate if the depletion in pG4r was uniform in all subclasses of the group, the densities of pG4r in each subclass of the short ncRNA group was evaluated (Figure 6). Among the shortNC class, the pre-miRNA subclasses have a density of 0.11 per kb and the miscRNA have a density of 0.05 per kb. The results obtained are in agreement with previous results (8)) and support the idea of a competition between the formation of the G4 structure and the formation of the secondary structure of short non-coding transcripts.

In order to confirm this observation, and to complete the investigation, the same method of G4 detection was applied to the short non-coding transcripts that are not present in the Ensembl database. The Ensembl database does not contain information about transfer RNA (tRNA), Piwi-interacting RNA (piRNA) or circular RNA (circRNA).

tRNA constitutes the most well-known class of non-PC RNA genes (39). Their complexity is related to the presence of isoforms, and chemical modifications provide an interesting case for the study of G4 in this subclass of transcripts (40). They are characterized by a cloverleaf-like secondary

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

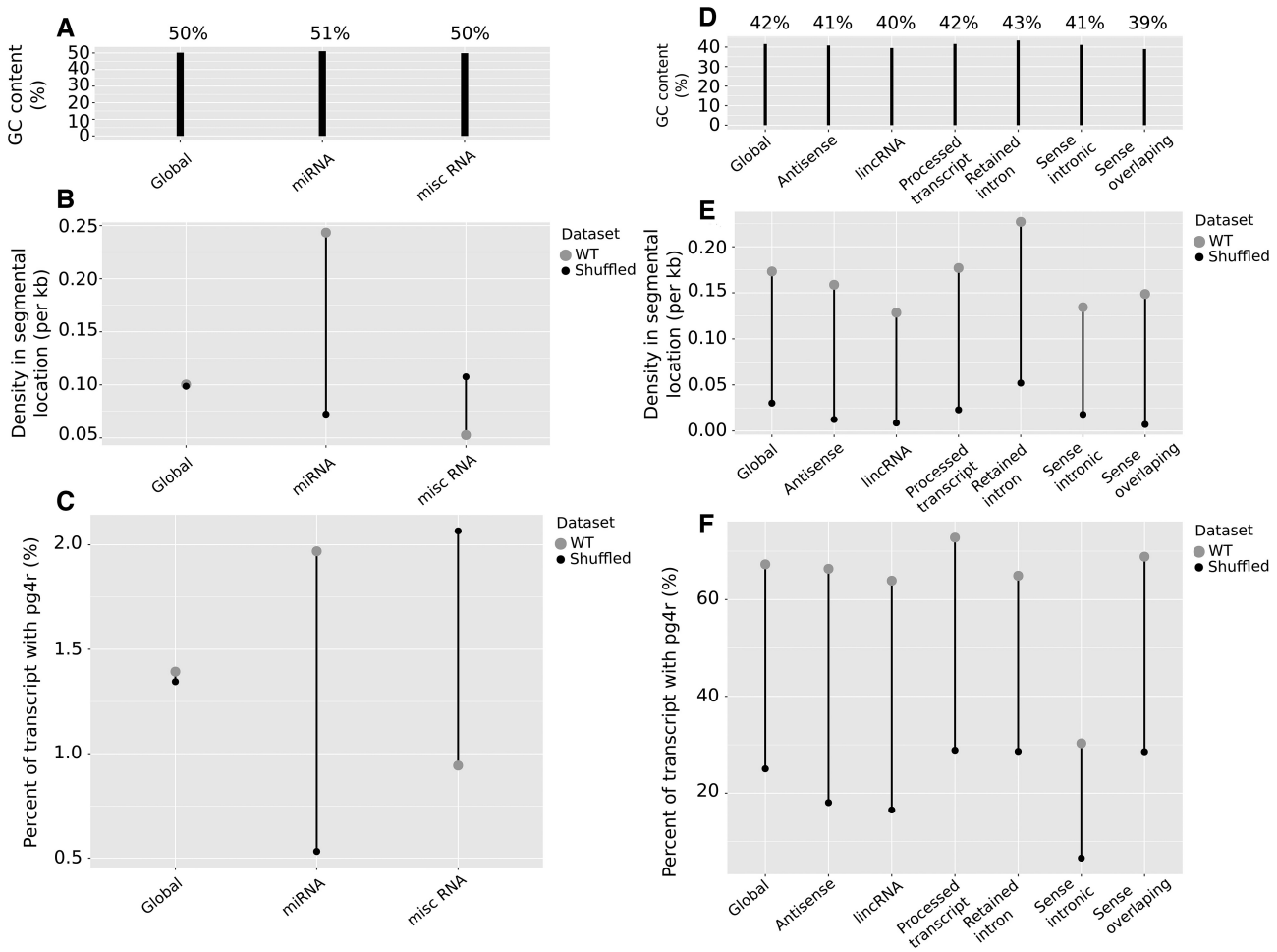

Figure 6. PG4r in ncRNA subclasses. (**A**) GC content, (**B**) pG4r density and (**C**) ratio of transcripts with pG4r, for the whole transcriptome and subclass (miRNA and miscRNA). (**D**, **E** and **F**): same as (A), (B) and (C) for long ncRNA. For (B), (C), (E) and (F), the values are shown for the WT dataset (gray point) and the shuffled dataset (black point).

structure which gives them their property of aminoacylation reactions by aminoacyl tRNA synthetases during protein synthesis. The Genomic tRNA Database (GtRNAdb) is a repository of all sequences identified by tRNAscan-SE in several organisms (41). Of the 420 tRNAs present in the database, no pG4r were detected in either the precursor or the mature form of the tRNA (data not shown). With our method, we could not predict intermolecular rG4, which means we could miss them. Indeed, some rG4 can fold between two or more tRNA (42).

Secondly, the piRNA transcripts protect the genome from parasite invasion by forming the piRNA-induced silencing complex (43). The database piRBase collects piRNA transcripts from various organisms and contains 32 826 transcripts from humans (44). In all these piRNA sequences, only 296 pG4r were detected, and these were found in only 296 species (i.e. 1 pG4 in each of the 296 sequences). The pG4r density in piRNA is 0.31 per kb because of the short length of sequences (data not shown).

Thirdly, circRNA are only expressed at low levels depending of the tissue, but they constitute a good model with which to confirm the competition between G4 structure and secondary structure formation in short ncRNA (45). Indeed, their circular form confers more stability than that of linear RNA in cells. These RNAs are created from precursor mRNA by the back-splicing of exons and are classified in the non-coding RNA class even if some studies support a possible capacity of circRNA to generate protein (46). Circular RNA from eukaryotes are listed in circBase (47). Of the 140 790 circRNA in the database, 36 050 pG4r were detected, leading to a density of pG4r of 0.03 per kb, lower than the density of the whole ncRNA class (data not shown).

Together, these results show that the G4 depletion in the tRNA, piRNA and circRNA subclasses are in accordance with the results obtained for the entire short ncRNA group. Indeed, low pG4r density is expected for the short ncRNA category because RNAs from this category are known to adopt compact, canonical secondary structures that are not compatible with the G4 secondary structure.

In short ncRNA, GC content is really high (i.e. between 47 and 51%). This lead, into all short ncRNA subclasses,

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

to a shuffled density higher than WT densities. Thus, this is consistent with the previous expectation to get low pG4r densities in these subclasses, due to their function.

Competition between the G4 and the stem-loop structure at Dicer cleavage site

At this time, miRNA is the second most studied type of non-coding RNA after tRNA, and several studies report that the presence of G4 inside its precursors has a deleterious effect on mature miRNA formation (48,49). The mature miRNA is about 20 nt in length and results from cascades of molecular processing involving several actors. First, due to the action of RNA polymerase, the miRNA gene is transcribed into the pri-miRNA, a long hairpin structure (for a review, see (50)). Next, a cleavage due to a complex containing Drosha allows for pre-miRNA formation, which then releases the final miRNA. After migration outside of the nucleus, and under the action of a complex containing Dicer, the hairpin of the precursor is cleaved leading to the creation of two strands, one of which will be selected and matured, forming the final, active miRNA. This final structure acts as a regulator of post-transcriptional gene expression by targeting an mRNA, resulting in either its translational inhibition or its degradation. Several mechanisms of the regulation of the Dicer-mediated maturation have been demonstrated, but all without including a possible effect of G4. The presence of G4 overlapping the future active miRNA inside the precursor has been demonstrated to be an inhibitor of Dicer complex binding on the stem-loop structure. On the one hand, Mirihana *et al.* worked on the human pre-miRNA 92b which is involved in lung cancer (48). They showed that a pG4 conserved in the pre-miRNA can fold and inhibit the formation of the canonical stem-loop structure, leading to a modulation at the protein level. On the other hand, Pandey *et al.* worked on the equilibrium between the G4 and stem-loop structures in several pre-miRNA (49). They demonstrated that the formation of a G4 which overlaps the corresponding mature miRNA can have a negative influence on the maturation mechanism.

In order to investigate this hypothesis, efforts were focused on the pre-miRNA subclass and mature miRNA (45), especially on the location of pG4r in those transcripts. Out of the 1879 pre-miRNA sequences reported in the human transcriptome, 37 pG4r were detected. This means that only 2% of pre-miRNA contain pG4r. In average, a pre-miRNA is 95 nt long and its pG4r is 61 nt long. This show that the pG4r always overlap either a Dicer site or a Drosha site. These results led to the observation of pG4r overlapping the Dicer cleavage site in 9% of the miRNAs. In 2016, a G4 analysis was performed on the pre-miRNAs contained in the primary miRNA sequence database (51). PG4r were detected based in the canonical definition, with a requirement for at least 2 G-tetrads (48). This analysis resulted in the prediction of G4 in 16% of human pre-miRNA stem loop regions. The higher number of pG4 as compared to our result could be due to the fact that their prediction did not account for GC content of sequences. The results of the present analyses confirm a competition between G4, and the secondary structure specifically recognized by Dicer. The results also support the idea of a new mechanism of regulation of miRNA at the precursor level. Finally, it has been shown that a similar modulation of the mature miRNA levels can also be influenced by the presence of G4 overlapping the Drosha cleavage site located in the miRNA primary form (8). G4 were also predicted in mature miRNA, using miRbase annotations (45). PG4r were detected in 9.3% of mature miRNA. The increase of pG4r quantity from 2% in pre-miRNA to 9.3% in mature miRNA can be explained by the fact that mature miRNA are shorter than pre-miRNA, and then the nucleotide environment has a less important contribution in the prediction of pG4r in mature miRNA. In a biological context, we could hypothesize that this increase could point to a regulation mechanism on mature miRNA. Pre-miRNA would not be regulated by pG4 because some of them contain two mature miRNA. The regulation of those pre-miRNA could stop both mature miRNA while if the regulation is only on mature miRNA the regulation could be more specific. Clearly, this is a hypothesis that requires be tested experimentally.

Variable enrichment of pG4 depending on the transcript type in long non-coding RNA

Following the high density of pG4r found in the long ncRNA group (0.17 pG4r per kb), the densities of pG4r in each transcript subclass of the group were further evaluated in order to investigate any possible variation of the densities between the subclasses. The long ncRNA group is composed of long transcripts from six subclasses (see also Supplementary Table S1):

i) antisense RNAs are located on the opposite strand of a PC gene;
ii) lincRNA corresponds to long intergenic RNA.
iii) processed transcripts are transcripts without any ORF;
iv) retained introns are sequences that may be either spliced or kept. They are not an exon because they are not flanked by introns. However, if they are translated, the protein will be non-functional;
v) sense intronic RNA are located in a PC gene's intron without overlapping any exons. They are on the same strand as the PC gene;
vi) sense overlapping RNA contain a coding gene in the intron region on the same strand.

All the subclasses in the analysis contain pG4r: sense intronic with a density of 0.13 per kb, antisense RNA with a density of 0.16 per kb, lincRNA with a density of 0.13 per kb, processed transcripts with a density of 0.18 per kb, retained intron with a density of 0.23 per kb and sense overlapping RNA with a density of 0.15 per kb (Figure 6E).

Next, an analysis of the G4 distribution across different locations was performed, in the same manner as for the mRNA class. First, the analysis was performed for the exon and intron locations. In comparison to the mRNA transcripts where the enrichment was similar between the exon and intron locations, the pG4r density seems to be higher in the exons of the long non-coding transcripts (Figure 7). This observation holds for all of the long ncRNA subclasses. In the long non-coding RNA, the G4 structure seems to be more involved in the splicing mechanism. PG4r are over-

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

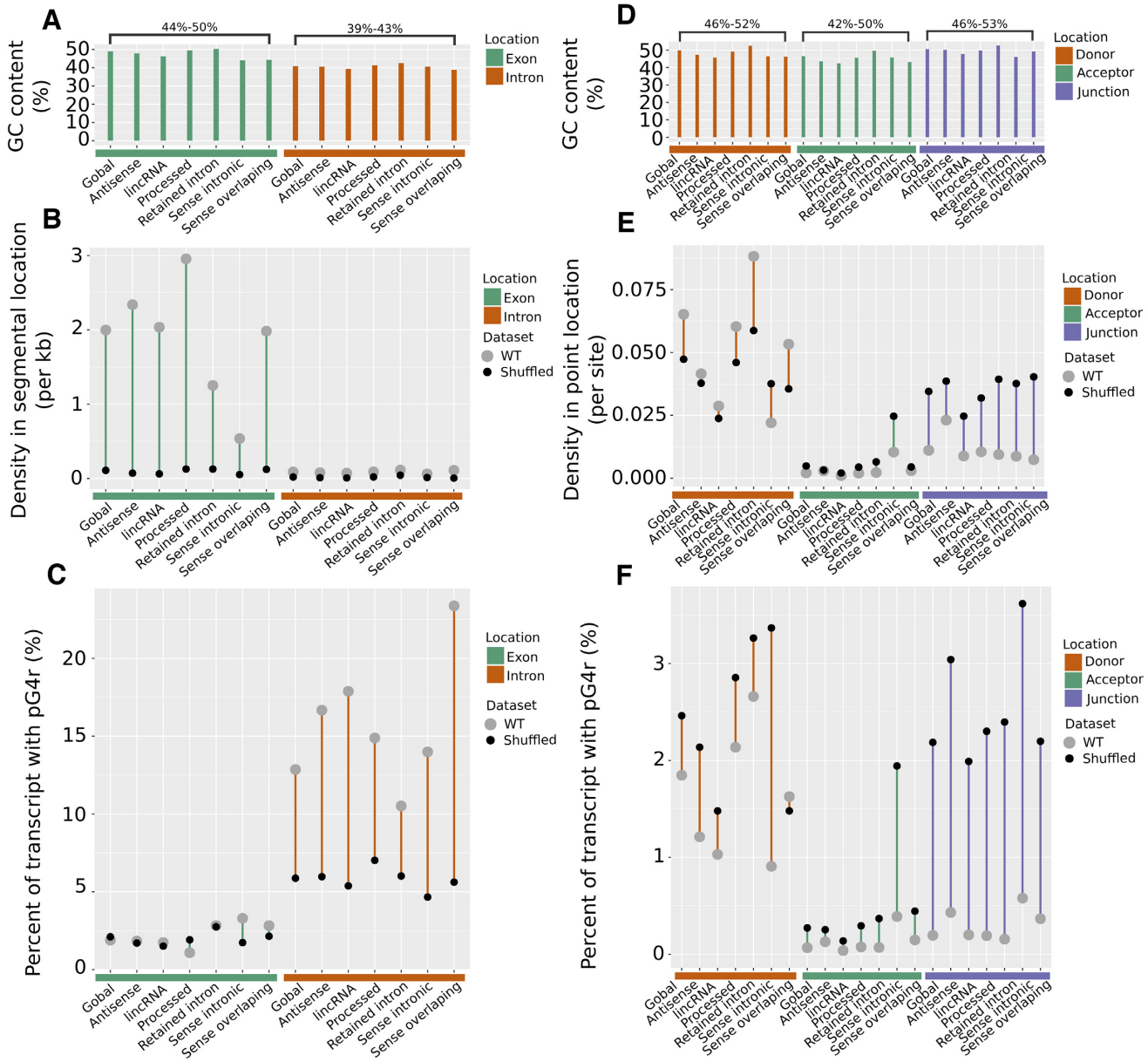

Figure 7. PG4r in the long ncRNA subclass by location. (**A**) GC content, (**B**) pG4r densities and (**C**) percent of transcript with pG4r for each segmental location (exon, intron, 5′UTR, CDS and 3′UTR) and for each subclass (antisense, lincRNA, processed, retained intron, sense intronic and sense overlaping). (D–F): same as (A–C) for point locations (donor, acceptor and junction). For (B), (C), (E) and (F) values are shown for WT dataset (gray points) and for shuffled data set (black points).

represented in the donor splice site, with a density of 0.06 per site. In fact, the study reveals that the results are the same for all of the subclasses of the group. The retained intron subclass has similar results to the whole ncRNA class, with a G4 density of 0.09 per site at the donor splice site, 0.002 per site at the acceptor splice site and 0.009 per site at the splicing junction. However, it is hard to explain all these observations by biological events or mechanisms since there is still very little knowledge about the biology of long ncRNA (52).

Yet, the comparison of WT dataset and shuffled dataset highlight the fact that more pG4r are found in WT dataset than by chance, for exons and donor junction. For the remaining locations (introns, acceptors and junctions), densities between those two datasets are mainly the same, or in some cases, the shuffled densities are above WT densities.

Over-representation of pG4r in the pseudogene class

Pseudogenes are transcribed genes that should result in a protein. That said, the ORF of these transcripts is disrupted, which is why they are classified differently. The pseudogene class can be divided into seven subclasses of transcripts:

i) The processed pseudogenes are the pseudogenes which are reintegrated into DNA from spliced mRNA;

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

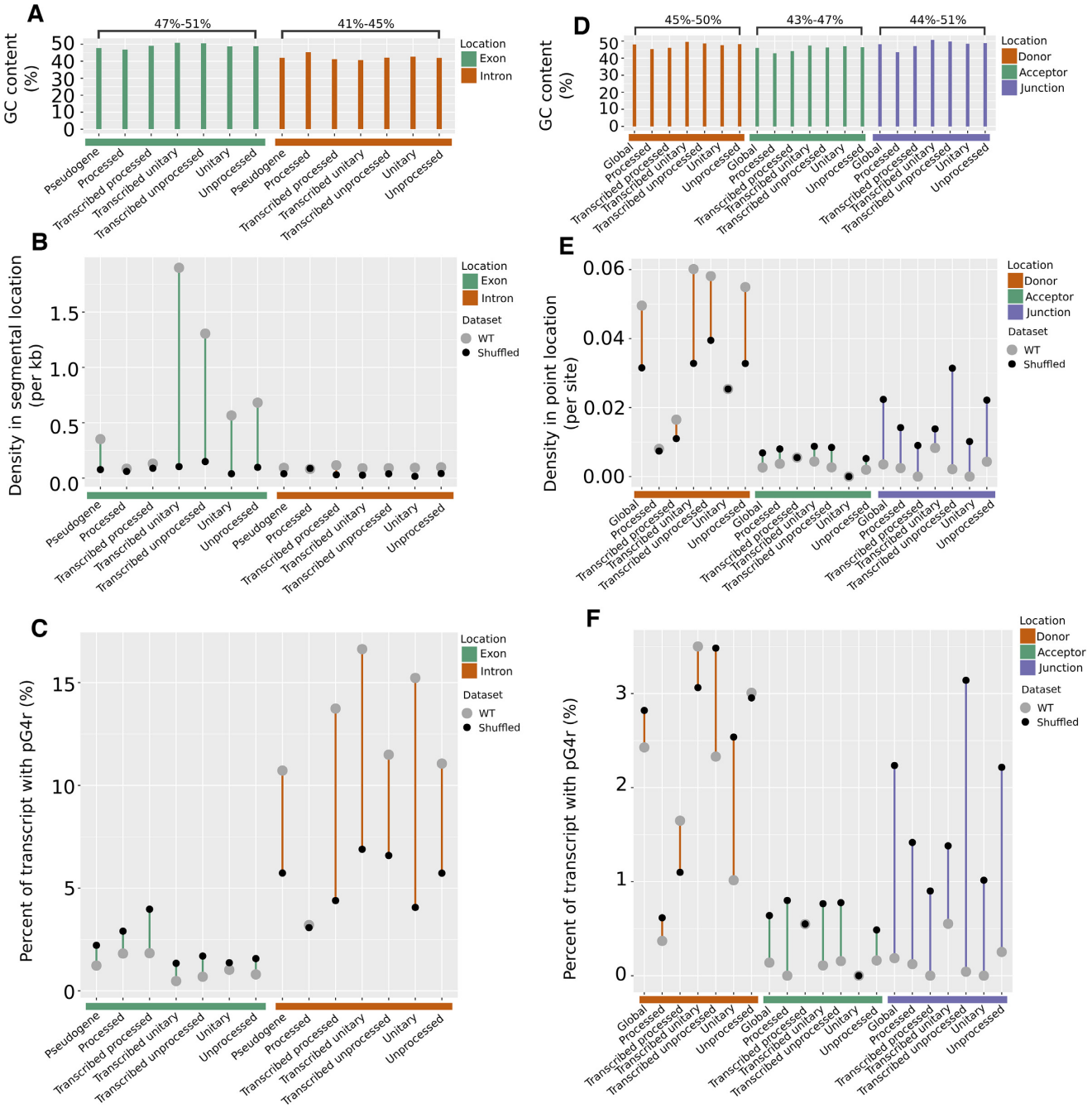

Figure 8. PG4r in the pseudogene subclass by location. (**A**) GC content, (**B**) pG4r densities and (**C**) percent of transcript with pG4r for each segmental location (exon, intron, 5′UTR, CDS and 3′UTR) and for each subclass (processed, transcribed processed, transcribed unitary, unitary, transcribed unprocessed and unprocessed). (D–F): same as (A–C) for point locations (donor, acceptor and junction). For (B), (C), (E) and (F) values are shown for WT dataset (gray points) and for shuffled data set (black points).

ii) The unprocessed pseudogenes are those produced by gene duplication and containing introns;
iii) The transcribed pseudogenes are those that protein homology or genomic structure indicates a pseudogene (processed or unprocessed), but that the presence of locus-specific transcripts suggests that there is expression;
iv) The unitary pseudogenes are inactive in humans but have an active ortholog in other species.

For the other types of pseudogene transcripts, the densities of pG4r for each type of location are shown in Figure 8. These results suggest an over-representation of pG4r in the pseudogene class. The study of the pG4r distribution in the different locations leads to several additional observations for this class. Like most other transcript classes, the pG4r density seems higher in exons than in introns. Furthermore, pG4r are over-represented in the donor splice site for almost all of the pseudogene subclasses. At the global level,

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

the pseudogene transcripts have a pG4r density of 0.05 per site for donor splice sites, compared to 0.007 for acceptor splice sites and 0.02 for splicing junctions. As in the case of the long non-coding transcripts, these results observed for pseudogene transcripts cannot be explained by either biological events or mechanisms because there exists little knowledge about the biology of pseudogenes. Indeed, this class was considered for a long time as a 'junk RNA' and was not investigated (52), but recent studies have brought support to their having possible functions.

By comparing WT dataset and shuffled dataset, it can be shown that pG4r are found less than by chance in intron and donor junction location, while in exon and acceptor junction it appears that pG4r densities are similar. Yet, for exon–exon junction, pG4r are found more than by chance.

CONCLUDING REMARKS

To our knowledge, this is the first extensive prediction of pG4r across the complete human transcriptome. The study reveals a prevalence of pG4r in the human transcriptome, with pG4r being present in all types of transcripts: mRNA, ncRNA and pseudogenes. The analysis of pG4r present in mRNA confirmed their enrichment in the 5′UTR, the 3′UTR and in the splicing junctions, as reported previously (for examples see refs (9,15,16)). In ncRNA, the analysis reveals a depletion of pG4r in short ncRNA as compared to both mRNA and long ncRNA due to a competition between G4 structures and RNA secondary structures. More specifically, the analysis shows that there is a competition between the G4 and stem-loop structures at the Dicer or Drosha cleavage site in pre-miRNA sequences. This is shown to have an effect on the Dicer complex's binding on the stem-loop structure, which supports the hypothesis of a new mechanism of regulation of miRNA at the precursor level. The study also reports a high density of pG4r in long ncRNA, with a higher density in the exon as compared to the intron, and an over-representation of pG4r in the donor splice sites. This suggests an involvement of G4 in splicing mechanisms. In pseudogene transcripts, the results show an over-representation of pG4r as compared to the other classes of transcripts, with a particular enrichment in introns and at the donor splice sites. This further suggests that G4 are involved in splicing mechanisms. A search for conserved G4 motifs located in splicing regions was inconclusive (data not shown). Further investigations in that direction are required and could lead to interesting discoveries.

Overall, this study provides exhaustive results on the distribution of pG4r across the entire human transcriptome, divided according to the various types of transcripts and their locations in the transcripts. It confirms the results of previous studies on G4 in mRNA and ncRNA and brings to light new hypotheses on the role of G4 in the transcriptome. It is anticipated that this study will lead to several new experiments, and to the discovery of new biological mechanisms involving G4 in the transcriptome. In particular, numerous discoveries involving the G4 located in the long non-coding RNA and in the pseudogene transcripts that are currently undergoing extensive investigation are foreseen.

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/G4HumanTranscriptome').

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

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.

Conflict of interest statement. None declare.

REFERENCES

1. Neidle,S. and Balasubramanian,S. (2007) *Quadruplex Nucleic Acids*. Royal Society of Chemistry (RSC), Cambridge.
2. Rouleau,S., Jodoin,R., Garant,J.-M. and Perreault,J.-P. (2017) RNA G-Quadruplexes as key motifs of the transcriptome. *Adv. Biochem. Eng. Biotechnol.*, **170**, 1–20.
3. Paeschke,K., Simonsson,T., Postberg,J., Rhodes,D. and Lipps,H.J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat. Struct. Mol. Biol.*, **12**, 847–854.
4. Wang,Q., Liu,J., Chen,Z., Zheng,K., Chen,C., Hao,Y. and Tan,Z. (2011) G-quadruplex formation at the 3′ end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase. *Nucleic Acids Res.*, **39**, 6229–6237.
5. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
6. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
7. Millevoi,S., Moine,H. and Vagner,S. (2012) G-quadruplexes in RNA biology. *Wiley Interdiscip. Rev. RNA*, **3**, 495–507.
8. Rouleau,S.G., Garant,J.-M., Bolduc,F., Bisaillon,M. and Perreault,J.-P. (2018) G-Quadruplexes influence pri-microRNA processing. *RNA Biol.*, **15**, 198–206.
9. Kwok,C.K., Marsico,G., Sahakyan,A.B., Chambers,V.S. and Balasubramanian,S. (2016) rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods*, **13**, 841–844.
10. Guo,J.U. and Bartel,D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**, aaf5371.
11. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
12. Chambers,V.S., Marsico,G., Boutell,J.M., Antonio,M.D., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
13. Marsico,G., Chambers,V.S., Sahakyan,A.B., McCauley,P., Boutell,J.M., Di Antonio,M. and Balasubramanian,S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.*, **47**, 3862–3874.

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023

14. Pandey,S., Agarwala,P. and Maiti,S. (2013) Effect of loops and G-Quartets on the stability of RNA G-Quadruplexes. *J. Phys. Chem. B*, **117**, 6896–6905.
15. Beaudoin,J.-D., Jodoin,R. and Perreault,J.-P. (2014) New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res.*, **42**, 1209–1223.
16. Beaudoin,J.-D. and Perreault,J.-P. (2010) 5′-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.*, **38**, 7022–7036.
17. Bedrat,A., Lacroix,L. and Mergny,J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.
18. Jodoin,R., Bauer,L., Garant,J.-M., Laaref,Mahdi, Phaneuf,A. and Perreault,J.-P. (2014) The folding of 5′-UTR human G-quadruplexes possessing a long central loop. *RNA*, **20**, 1129–1141.
19. Bolduc,F., Garant,J.-M., Allard,F. and Perreault,J.-P. (2016) Irregular G-quadruplexes found in the untranslated regions of human mRNAs influence translation. *J. Biol. Chem.*, **291**, 21751–21760.
20. Mukundan,V.T. and Phan,A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017–5028.
21. Meier,M., Moya-Torres,A., Krahn,N.J., McDougall,M.D., Orriss,G.L., McRae,E.K.S., Booy,E.P., McEleney,K., Patel,T.R., McKenna,S.A. *et al.* (2018) Structure and hydrodynamics of a DNA G-quadruplex with a cytosine bulge. *Nucleic Acids Res.*, **46**, 5319–5331.
22. Lightfoot,H.L., Hagen,T., Cléry,A., Allain,F.H.-T. and Hall,J. (2018) Control of the polyamine biosynthesis pathway by G2-quadruplexes. *Elife*, **7**, e36362.
23. Fleming,A.M., Ding,Y., Alenko,A. and Burrows,C.J. (2016) Zika virus genomic RNA possesses conserved G-quadruplexes characteristic of the flaviviridae family. *ACS Infect. Dis.*, **2**, 674–681.
24. Zhang,Y., Liu,S., Jiang,H., Deng,H., Dong,C., Shen,W., Chen,H., Gao,C., Xiao,S., Liu,Z.-F. *et al.* (2020) G 2 -quadruplex in the 3′UTR of IE180 regulates Pseudorabies virus replication by enhancing gene expression. *RNA Biol.*, **17**, 816–827.
25. Faudale,M., Cogoi,S. and Xodo,L.E. (2012) Photoactivated cationic alkyl-substituted porphyrin binding to g4-RNA in the 5′-UTR of KRAS oncogene represses translation. *Chem. Commun.*, **48**, 874–876.
26. Garant,J.-M., Perreault,J.-P. and Scott,M.S. (2017) Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics*, **33**, 3532–3537.
27. Garant,J.-M., Luce,M.J., Scott,M.S. and Perreault,J.-P. (2015) G4RNA: an RNA G-quadruplex database. *Database (Oxford)*, **2015**, bav059.
28. Garant,J.-M., Perreault,J.-P. and Scott,M.S. (2018) G4RNA screener web server: user focused interface for RNA G-quadruplex prediction. *Biochimie*, **151**, 115–118.
29. Belmonte-Reche,E. and Morales,J.C. (2020) G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. *NAR Genomics Bioinform.*, **2**, 1–12.
30. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of ensembl. *Genome Res.*, **14**, 925–928.
31. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
32. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
33. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
34. Rodriguez,R., Müller,S., Yeoman,J.A., Trentesaux,C., Riou,J.-F. and Balasubramanian,S. (2008) A novel small molecule that alters shelterin integrity and triggers a DNA-Damage response at telomeres. *J. Am. Chem. Soc.*, **130**, 15758–15759.
35. Kwok,C.K., Marsico,G. and Balasubramanian,S. (2018) Detecting RNA G-quadruplexes (rG4s) in the transcriptome. *Cold Spring Harb. Perspect. Biol.*, **10**, a032284.
36. Mizuta,R., Iwai,K., Shigeno,M., Mizuta,M., Uemura,T., Ushiki,T. and Kitamura,D. (2003) Molecular visualization of immunoglobulin switch region RNA/DNA complex by atomic force microscope. *J. Biol. Chem.*, **278**, 4431–4434.
37. Dunnick,W., Hertz,G.Z., Scappino,L. and Gritzmacher,C. (1993) DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Res.*, **21**, 365–372.
38. Beaudoin,J.-D. and Perreault,J.-P. (2013) Exploring mRNA 3′-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res.*, **41**, 5898–5911.
39. Bhartiya,D. and Scaria,V. (2016) Genomic variations in non-coding RNAs: structure, function and regulation. *Genomics*, **107**, 59–68.
40. Schimmel,P. (2018) The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nat. Rev. Mol. Cell Biol.*, **19**, 45–58.
41. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
42. Lyons,S.M., Gudanis,D., Coyne,S.M., Gdaniec,Z. and Ivanov,P. (2017) Identification of functional tetramolecular RNA G-quadruplexes derived from transfer RNAs. *Nat. Commun.*, **8**, 1127.
43. Iwasaki,Y.W., Siomi,M.C. and Siomi,H. (2015) PIWI-Interacting RNA: Its biogenesis and functions. *Annu. Rev. Biochem.*, **84**, 405–433.
44. Zhang,P., Si,X., Skogerbø,G., Wang,J., Cui,D., Li,Y., Sun,X., Liu,L., Sun,B., Chen,R. *et al.* (2014) piRBase: a web resource assisting piRNA functional study. *Database (Oxford)*, **2014**, 1–7.
45. Li,X., Yang,L. and Chen,L.-L. (2018) The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell*, **71**, 428–442.
46. Pamudurti,N.R., Bartok,O., Jens,M., Ashwal-Fluss,R., Stottmeister,C., Ruhe,L., Hanan,M., Wyler,E., Perez-Hernandez,D., Ramberger,E. *et al.* (2017) Translation of CircRNAs. *Mol. Cell*, **66**, 9–21.
47. Glažar,P., Papavasileiou,P. and Rajewsky,N. (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.
48. Mirihana Arachchilage,G., Dassanayake,A.C. and Basu,S. (2015) A potassium ion-dependent RNA structural switch regulates human Pre-miRNA 92b maturation. *Chem. Biol.*, **22**, 262–272.
49. Pandey,S., Agarwala,P., Jayaraj,G.G., Gargallo,R. and Maiti,S. (2015) The RNA stem–loop to G-Quadruplex equilibrium controls mature MicroRNA production inside the cell. *Biochemistry*, **54**, 7067–7078.
50. Zhang,J., Li,S., Li,L., Li,M., Guo,C., Yao,J. and Mi,S. (2015) Exosome and exosomal MicroRNA: trafficking, sorting, and function. *Genomics Proteomics Bioinformatics*, **13**, 17–24.
51. Weldon,C., Eperon,I.C. and Dominguez,C. (2016) Do we know whether potential G-quadruplexes actually form in long functional RNA molecules? *Biochem. Soc. Trans.*, **44**, 1761–1768.
52. Tutar,Y. (2012) Pseudogenes. *Comp. Funct. Genomics*, **2012**, 424526.

Downloaded from https://academic.oup.com/nargab/article/2/2/lqaa035/5843801 by Université de Sherbrooke user on 04 October 2023