Published online 16 August 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D135–D140***
https://doi.org/10.1093/nar/gkac657

GAIA: G-quadruplexes in alive creature database

Anaïs Vannutelli[1,2], Lauriane Lucienne Noele Schell[2], Jean-Pierre Perreault[1,*] and Aïda Ouangraoua[2,*]

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

Received June 01, 2022; Revised July 08, 2022; Editorial Decision July 12, 2022; Accepted August 08, 2022

ABSTRACT

G-quadruplexes (G4) are 3D structures that are found in both DNA and RNA. Interest in this structure has grown over the past few years due to both its implication in diverse biological mechanisms and its potential use as a therapeutic target, to name two examples. G4s in humans have been widely studied; however, the level of their study in other species remains relatively minimal. That said, progress in this field has resulted in the prediction of G4s structures in various species, ranging from bacteria to eukaryotes. These predictions were analysed in a previous study which revealed that G4s are present in all living kingdoms. To date, eleven different databases have grouped the various G4s depending on either their structures, on the proteins that might bind them, or on their location in the various genomes. However, none of these databases contains information on their location in the transcriptome of many of the implicated species. The GAIA database was designed so as to make this data available online in a user-friendly manner. Through its web interface, users can query GAIA to filter G4s, which, we hope, will help the research in this field. GAIA is available at: https://gaia.cobius.usherbrooke.ca

INTRODUCTION

G-quadruplexes (G4s) are non-canonical structures that can be formed by both DNA and RNA. They are composed of Hoogsten base pairings, instead of classical Watson & Crick base pairings, between four guanines. This results in the formation of a planar structure that is called a G-quartet. The G-quartets then stack on top of each other, forming a G4 (1,2). As this structure is primarily composed of guanine residues, and the presence of four guanines are required in order to form a G-track, it creates a motif in the primary sequence, known as a canonical sequence, that follows the pattern: $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$.

Further research revealed the existence of G4s that did not match the canonical pattern. More precisely, four types of non-canonical G4s have been discovered: G4s with two Gtracks instead of the minimum three; G4s with loops longer than seven nucleotides; G4s with a bulge in a G-track; and, G4s possessing a nucleotide other than guanine that can contribute to the Hoogsten pairing (3–6). All of these non-canonical G4s should be able to fold, but they might be less stable than the canonical one. As a consequence, all of these different patterns were thus used to predict G4s (7,8). Even so, a lot of the predicted G4s (pG4s) were found not to fold *in vitro*. This highlighted the fact that G4s need to be located in a low cytosine environment (9). Otherwise, a competition between the Watson & Crick and Hoogsten pairings occurs. These G4s might be less stable or might encounter difficulties to fold. Thus, new prediction tools were developed in order to overcome this problem (10–12). That said, there is more and more evidence showing that a balance between G4s and other secondary structures is possible (13–15).

Both genome-wide or transcriptome-wide predictions showed that G4s are omnipresent. Indeed, G4s are present inside telomeres, centromeres, microsatellites and promoters, and have been predicted to be found in both at least 60% of all transcripts and in almost all transcript classes (5–7,15). This wide presence in sequences shows just how important G4s are. Due to their presence in both DNA and RNA, G4s can regulate transcription and translation, and thus they can impact a significant number of biological processes. For example, the folding of G4s structures located in the 5′UTRs of mRNAs have been shown to repress mRNA translation by preventing ribosome scanning (9). Moreover, G4s located in pre-miRNA can prevent their maturation and thus have an important impact on mRNA translation. A good example of this phenomenon is the case of the human pre-miRNA-1229. Since the resulting miRNA can inhibit the translation of the SORL1 mRNA, which inhibits β-secretase, a decrease in the miRNA-1229 level results in

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

a decrease in the level of β-amyloid. Therefore, the folding of this particular G4 has potential as a target to reduce the production of β-amyloid in Alzheimer's disease (13).

The study of G4s is thus clearly important in order to understand how they are used as a regulatory feature in both the genome and in gene expression. In addition, this knowledge might help in their development as new therapeutic targets. Clearly, to improve our knowledge on G4s, all data on them should be easily accessible. To our knowledge, to date, there are eleven databases available on G4s (see Supplementary Table S1). Among them, three are no longer accessible (Plant-GQ (17), TTS mapping (18) and GRSDB/GRSutrDB (19)). Eight of the 11 G4s databases involve human sequences and are DNA oriented. Thus, only three databases contain RNA data, and even then only a few pG4s are annotated in them (specifically between 48 and 261 G4 were annotated).

It is for this reason that the GAIA (G-quAdruplexes In Alive creature) database was developed. GAIA is a user-friendly database with a web interface that can be used to query a data set containing pG4s in diverse species from all three living kingdoms.

IMPLEMENTATION

GAIA stores pG4s in a relational database using PostgreSQL 14.0 (20). GAIA datasets were obtained through G4RNA screener, a tool developed to predict the presence of G4s in RNA. The pG4s are available with both their sequence and chromosomal coordinates (i.e. start, end, strand and chromosome name). G4s were predicted on each transcript available in release 46 of the Ensembl Compara database (21). From this release, 61 species were selected. For more details on the prediction method, see (16).

All annotated transcripts in the Ensembl Compara database were kept for the prediction run. All annotated transcripts in Ensembl possess a Transcript Support Level (TSL) which provides information as to their quality. The TSL can be selected by the user, as can the transcript class and the transcript location. A pG4 can be present in two locations depending on the transcript's maturation. On a premature RNA, the pG4 will be located inside either an exon or an intron, while in a mature RNA, the pG4 will be located either in the coding sequences (CDS) or in an untranslated region (UTR). If several transcripts overlap because of constitutive splicing, the pG4 will be annotated many times.

G4s are predicted using G4RNA screener (11), a tool that produces three different scores: G4NN, cGcG and G4 Hunter. G4NN is an artificial neural network that learns from the G4RNA database (22) whether or not a given sequence looks like it will fold into a G4 or not. Both cGcC and G4 Hunter (10,12) were primarily developed to reflect G4 stability by measuring G and C skewness. Altogether, the higher the scores, the higher the probability that the sequence in question will fold and be able to generate a stable G4.

It is expected that the release of a user-friendly website platform is going to help the scientific community build projects focused around G4s.

WEBSITE

The Gaia database can be accessed using the following URL: https://gaia.cobius.usherbrooke.ca. When the database is accessed, a default query is automatically made and outputted. This default query is made in order to retrieve all G4s predicted in *Escherichia coli* (Ecol) (16). The default query can always be made again by accessing the 'Home' page. These pG4s can be filtered through the query field (see Supplementary Figure S1), which contains two parts: a 'search by' part; and, a 'display field' part. For all options, a help button is available which provides additional information on possible criteria. In the search by area, pG4s can be selected by genes, transcripts, sequences, transcript classes and transcript locations. The results can also be filtered depending on the species. The user can select either multiple species or only one. If the user is interested in a specific region, its coordinates may be entered so as to restrict the queries to this particular area. In this case, only one species must be selected. Finally, it is possible to restrict the research to a higher score than those used during the prediction. All of the previous options help to filter the database. In the display field, the user is also given the option to choose which information to display. This is accomplished by selecting the column name in the display field section.

Once the user makes a query, the output is displayed in an HTML table in which 100 rows are displayed at a time through multiple tabs. This table can be downloaded in three different formats. The first is a csv file that corresponds to the table displayed on the web page. The second is a fasta file in which all of the selected columns are concatenated into an identifier, followed by the sequence of the pG4. Finally, a bed file format is created with the chromosome name and the start/end coordinates for the first three columns. If more columns are selected, they are also included in the output file, after the mandatory columns.

The website was developed using Django 3.2.8 as both a back-end and front-end tool, and Apache 2.4 as a web server on a Ubuntu 18.04.4 LTS system.

STATISTICS

GAIA regroups a total of 3 854 805 pG4s spread throughout all living kingdoms, among which there are 12 archaeal species with 6 155 pG4s, 24 bacterial species with 11 002 pG4s and 25 eukaryotic species with 3 837 648 pG4s. The number of pG4s in all species and their location/transcript classes are presented in Table 1. The number of pG4s is higher in eukaryotes because G4s are more predicted than expected by chance (for more details on the method, see (16)), but also because there are more annotated eukaryotic transcripts. Indeed, over all eukaryotes, there is an average of 2 transcripts per gene (for humans, the average go up to 4), while only 1 transcript per gene is annotated for both archaea and bacteria.

All species possess pG4s, except for *Francisella tularensis* and *Staphylococcus aureus* (Ftu and Sau, respectively), in which no pG4 were retrieved. In Table 1, pG4 absence can be due to either the lack of annotation ('-') or to the real lack of pG4s (0). This shows that pG4s are present in all loca-

Table 1. Number of pG4 contained in GAIA. Numbers of pG4s per species and locations/transcript class are shown in this table

Specie	Exon	Intron	5′UTR	CDS	3′UTR	Start codon	Stop codon	Donor	Acceptor	Junction	Coding	Long non coding	Short non coding
Aaeo	80	-	-	78	-	3	8	-	-	-	165	4	0
Acar	2338	32758	961	629	490	154	22	981	40	136	33828	4669	3
Aful	426	-	-	423	-	20	24	-	-	-	890	3	0
Amel	1475	8990	92	1165	207	75	12	70	48	70	12203	0	1
Anid	1583	57	5	1540	27	8	171	86	9	59	3539	0	0
Apha	42	-	-	39	-	5	7	-	-	-	91	1	1
Atha	1769	93	57	1600	89	58	25	54	169	147	4022	38	1
Babo	53	-	-	52	-	5	6	-	-	-	116	0	0
Bbur	8	-	-	8	-	0	1	-	-	-	17	0	0
Bsub	19	0	-	20	-	0	6	0	0	0	45	0	0
Caur	229	-	-	219	-	8	47	-	-	-	495	8	0
Cele	610	1858	5	395	13	13	14	184	17	118	2963	128	94
Cjej	8	0	-	4	-	0	1	0	0	0	9	3	0
Ckor	699	-	-	696	-	28	44	-	-	-	1462	5	0
Crei	48466	66405	1103	30516	16689	214	949	41303	532	1573	207750	0	0
Csym	386	-	-	389	-	27	21	-	-	-	823	0	0
Ctra	8	-	-	8	-	0	2	-	-	-	18	0	0
Ddic	112	1	0	106	-	4	1	2	1	1	228	0	0
Dmel	3317	29758	245	2181	556	25	67	327	5	69	35435	1109	1
Drer	4455	57477	308	2985	911	121	54	392	323	576	62275	5249	1
Ecol	78	0	-	73	-	0	15	0	0	0	159	1	0
Efae	5	-	-	5	-	2	1	-	-	-	13	0	0
Ftul	0	0	-	0	-	0	0	0	0	0	0	0	0
Gacu	4345	26921	228	3834	247	104	82	2483	856	942	40034	0	1
Ggal	15010	162695	3193	6909	3097	1314	500	16567	691	1232	191402	17754	34
Gsul	594	0	-	585	-	15	66	0	0	0	1255	4	1
Hbut	135	0	-	133	-	9	21	0	0	0	296	2	0
Hinf	18	-	-	8	-	0	1	-	-	-	17	10	0
Hsal	209	-	-	206	-	14	15	-	-	-	441	1	0
Hsap	54679	1253928	9202	8747	13754	1594	811	58017	774	2920	939945	456877	64
Lmaj	1201	4	-	1142	-	36	41	1	1	1	2362	45	0
Lpne	20	-	-	19	-	0	4	-	-	-	42	0	0
Mace	126	-	-	126	-	5	12	-	-	-	269	0	0
Mdom	10991	186214	3330	2594	3729	381	118	7052	402	527	193604	21685	7
Mmus	29261	672515	5166	4425	8461	796	359	28190	408	1255	559933	186666	36
Mpne	26	-	-	27	-	0	2	-	-	-	55	0	0
Msmi	1	-	-	1	-	0	0	-	-	-	2	0	0
Mtub	564	-	-	550	-	15	22	-	-	-	1138	4	1
Mxan	1638	-	-	1630	-	104	45	-	-	-	3416	0	1
Ncra	6299	238	145	4592	1511	19	307	72	26	90	13265	0	0
Nequ	4	-	-	3	-	0	0	-	-	-	6	1	0
Nmen	43	0	-	42	-	0	9	0	1	0	95	0	0
Oana	1088	34991	203	346	434	29	18	853	131	119	35490	2717	1
Osat	14847	5271	2346	10572	1308	1903	260	263	1190	448	38402	5	1
Pabe	8461	148582	2100	2626	3490	408	154	8756	255	485	175173	2	22
Paer	631	-	-	621	-	30	26	-	-	-	1299	9	0
Phor	158	0	-	153	-	16	17	0	0	0	339	5	0
Ppat	31600	11014	15057	13654	2296	2326	512	1769	2625	1437	82289	1	0
Ptro	17863	407584	5894	6170	5090	1078	395	22497	724	1328	458626	9889	55
Sau	0	0	-	0	-	0	0	-	-	-	0	0	0
Scer	26	0	0	20	-	3	1	0	0	0	44	6	0
Slyc	1710	3526	115	1345	223	123	71	116	130	151	7494	16	0
Spne	4	-	-	5	-	0	0	-	-	-	9	0	0
Spom	64	2	6	10	11	0	1	0	0	0	58	36	0
Ssol	124	-	-	116	-	5	8	-	-	-	248	5	0
Taci	21	-	-	21	-	0	5	-	-	-	47	0	0
Tthe	1594	-	-	1588	-	55	95	-	-	-	3326	4	2
Vcho	25	-	-	24	-	1	4	-	-	-	53	1	0
Vvin	3205	8638	233	2758	184	290	82	84	212	206	15880	10	2
Wend	6	-	-	6	-	0	1	-	-	-	13	0	0
Ypes	193	-	-	186	-	6	8	-	-	-	386	7	0

For some locations or classes, no pG4 are available in GAIA, for two reasons: the annotation was not available (represented by a '-') or no pG4 were predicted (represented by a 0).

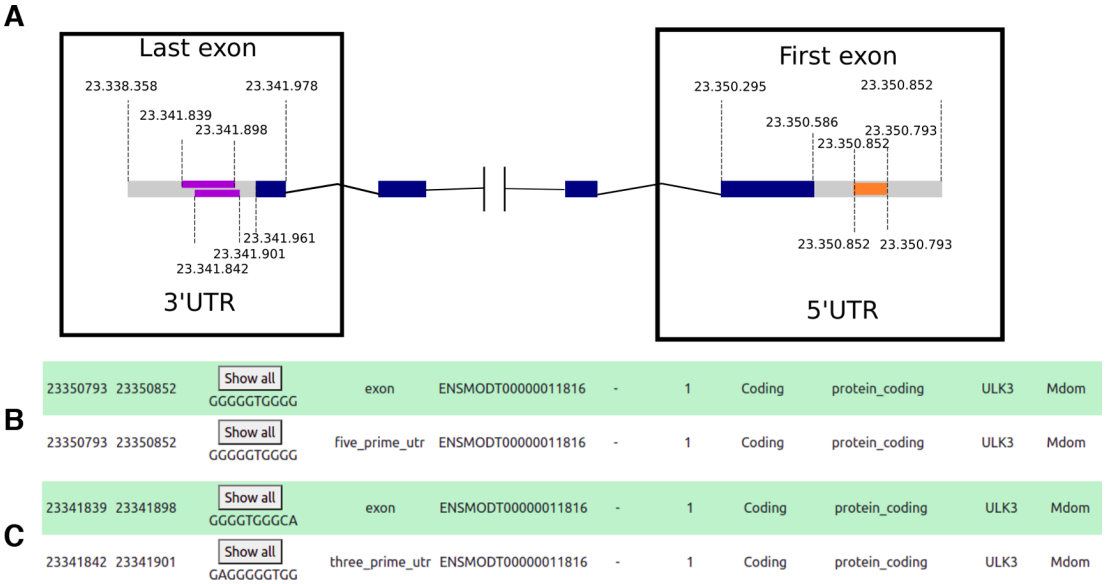

Figure 1. Example of overlapping pG4. (**A**) shows a schematic representation of ULK3 gene structure in Mdom, where pG4 are represented in purple for the 3′UTR or orange for the 5′UTR. CDS are blue and UTRs are grey; (**B**) example of perfect overlapping between a 5′UTR and an exon; (**C**) example of the same pG4 with different coordinates because of exon and 3′UTR coordinates differences.

tions, and in almost all species, when the annotation allows its investigation. The locations most altered by the annotation are the introns in bacteria and archaea because they are less common in both of these kingdoms, and the UTRs because only CDS are annotated. For a more detailed analysis of the pG4s distribution, see (16).

USE CASE

By default, a query is made on the server in order to retrieve pG4s predicted in *Ecol*, without any restrictions. The user can customise the query to make it fit his needs. If the user desires to retrieve all pG4s from GAIA, either all species or the 3 domains of life need to be selected.

If only human pG4s are required, they can all be downloaded by selecting 'Human' in the species selection. Other filters can be applied at the same time, like choosing only pG4s located in exons, or by choosing only pG4s located in a particular gene. When looking at a specific gene, selecting the desired species, or using the gene's ID, might be useful since a gene name can be found in many species. For example, the gene 'ULK3′ is annotated in many species and would return pG4s in Gacu, Ggal, Hsap, Mmus, Mdom, Pabe and Ptro. In that case, if the user had unchecked the species name from the column, the results could be misinterpreted.

Column selection might also impact the number of rows returned. Indeed, the same pG4 can be annotated in an exon and in its CDS/UTR, and thus would appear twice in the results (see Figure 1). If the location name column was not selected, the pG4 would have been unique. Similarly, in Ensembl annotation, many transcripts are annotated for a gene, and pG4s will appear for each transcript if the transcript's ID is shown. Taken altogether, this means that to get a unique pG4, and to avoid the redundancy of transcript/location, the user needs to select a minimal number of columns: coordinates; strand; chromosomes; and, sequences. In some cases, overlapping pG4s can have different coordinates if there is a shift between the exon and UTR/CDS coordinates (see Figure 1C).

In order to help users, keep track of their queries, a query file can be downloaded. This file contains options and the criteria that were selected for the last query run.

DISCUSSION

Currently, no accessible databases provide a user-friendly interface which permits to query pG4s data sets in the transcripts from a wide variety of species. GAIA was developed to complete the ensemble of existing G4 databases (see Supplementary Table S1). In addition, it is important to note that more than half of the currently available databases do not grant access to experimental data but to predicted data. GAIA could be improved in that way by providing predicted G4 as well as experimental data, but also by providing a form for the submission of user data. The development, improvement and usage of high-throughput methods for the detection of G4s in either the whole transcriptome or the genome could help for the development of more databases that include experimental data. This would help not only the global knowledge on G4s, but it would also improve G4s prediction, a subject that is becoming more complex. Indeed, as our knowledge on G4s has progressed they have gone from a simple canonical pattern usage, to more relaxed patterns and then finally to a score that takes into account the nucleotide landscape. Few years ago, a study tried to show that G4 folding *in vivo* is controlled by multiple DNA and RNA binding proteins as they are globally unfolded in cells (23). A global view of G4-binding protein landscape would be primordial in order to understand G4 folding and cell usage. A recent database, G4IPDB (24) con-

tains 130 DNA binding proteins and 75 RNA binding proteins and is a good start in this regard.

Finally, research on diverse species showed quantified G4 conservation in genomes (25–31). That said, no information is available on G4 evolution, since their sequences are not well conserved. Similarly, little research has been performed on G4 sequence evolution. If families of G4 are found, GAIA could be updated with it.

CONCLUSION

To date, no online database presents RNA pG4s of many species (32–39). GAIA is a new database that contains the predicted RNA G4s in all living kingdoms. The data set in GAIA has been analysed in depth, and its annotation is now fully available. The GAIA data set is accessible via a user-friendly web interface which allows the user to easily query the database. It is hoped that GAIA will help stimulate the scientific community to work on G4s and expand our knowledge of them outside of the human perspective.

DATA AVAILABILITY

Database freely available at https://gaia.cobius.usherbrooke.ca.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Canada Research Chair in Computational and Biological Complexity [CRC Tier2 Grant 950-230577 to A.O.]; Chaire de recherche de l'Universite de Sherbrooke en Structure et Genomique 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.); Universite de Sherbrooke. Funding for open access charge: Canada Research Chair in Computational and Biological Complexity [CRC Tier2 Grant 950-230577 to A.O.]; Chaire de recherche de l'Universite de Sherbrooke en Structure etGenomique 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 [NSERCRGPIN-155219-17 to J.P.P., RGPIN-05552-17 to A.O.]; Centre de Recherche du CHUS (to J.P.P.); Universite de Sherbrooke.

Conflict of interest statement. None declared.

REFERENCES

1. Kim,J., Cheong,C. and Moore,P.B. (1991) Tetramerization of an RNA oligonucleotide containing a GGGG sequence. *Nature*, **351**, 331–332.
2. Cheong,C. and Moore,P.B. (1992) Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. *Biochemistry*, **31**, 8406–8414.
3. 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.
4. 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.
5. Lim,K.W., Amrane,S., Bouaziz,S., Xu,W., Mu,Y., Patel,D.J., Luu,K.N. and Phan,A.T. (2009) Structure of the human telomere in K+ solution: a stable basket-type G-Quadruplex with only two G-Tetrad layers. *J. Am. Chem. Soc.*, **131**, 4301–4309.
6. Lim,K.W., Alberti,P., Guédin,A., Lacroix,L., Riou,J.-F., Royle,N.J., Mergny,J.-L. and Phan,A.T. (2009) Sequence variant (CTAGGG)n in the human telomere favors a G-quadruplex structure containing a G.C.G.C tetrad. *Nucleic Acids Res.*, **37**, 6239–6248.
7. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
8. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.
9. Beaudoin,J.-D. and Perreault,J.-P. (2010) 5′-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.*, **38**, 7022–7036.
10. 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.
11. 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.
12. Bedrat,A., Lacroix,L. and Mergny,J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.
13. Imperatore,J.A., Then,M.L., McDougal,K.B. and Mihailescu,M.R. (2020) 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.*, **21**, 767.
14. De Nicola,B., Lech,C.J., Heddi,B., Regmi,S., Frasson,I., Perrone,R., Richter,S.N. and Phan,A.T. (2016) Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome. *Nucleic Acids Res.*, **44**, 6442–6451.
15. Vannutelli,A., Belhamiti,S., Garant,J.-M., Ouangraoua,A. and Perreault,J.-P. (2020) Where are G-quadruplexes located in the human transcriptome? *NAR Genomics Bioinformatics*, **2**, lqaa035.
16. Vannutelli,A., Ouangraoua,A. and Perreault,J.-P. (2022) G-Quadruplex occurrence and conservation: more than just a question of guanine-cytosine content. *NAR Genomics Bioinformatics*, **4**, lqac010.
17. Ge,F., Wang,Y., Li,H., Zhang,R., Wang,X., Li,Q., Liang,Z. and Yang,L. (2019) Plant-GQ: an integrative database of G-Quadruplex in plant. *J. Comput. Biol.*, **26**, 1013–1019.
18. Jenjaroenpun,P. and Kuznetsov,V.A. (2009) TTS mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics*, **10**, S9.
19. Kikin,O., Zappala,Z., D'Antonio,L. and Bagga,P.S. (2008) GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. *Nucleic Acids Res.*, **36**, D141–D148.
20. Stonebraker,M., Rowe,L.A. and Hirohama,M. (1990) The implementation of POSTGRES. *IEEE Trans. Knowledge Data Eng.*, **2**, 125–142.
21. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
22. Garant,J.-M., Luce,M.J., Scott,M.S. and Perreault,J.-P. (2015) G4RNA: an RNA G-quadruplex database. *Database (Oxford)*, **2015**, bav059.
23. Guo,J.U. and Bartel,D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**, aaf5371.
24. Mishra,S.K., Tawani,A., Mishra,A. and Kumar,A. (2016) G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Scientific Rep.*, **6**, 38144.
25. Bartas,M., Čutová,M., Brázda,V., Kaura,P., Šťastný,J., Kolomazník,J., Coufal,J., Goswami,P., Červeň,J. and Pečinka,P.

(2019) The presence and localization of G-Quadruplex forming sequences in the domain of bacteria. *Molecules*, **24**, E1711.
26. Brázda,V., Luo,Y., Bartas,M., Kaura,P., Porubiaková,O., Šťastný,J., Pečinka,P., Verga,D., Da Cunha,V., Takahashi,T.S. *et al.* (2020) G-Quadruplexes in the Archaea domain. *Biomolecules*, **10**, 1349.
27. Marsico,G., Chambers,V.S., Sahakyan,A.B., McCauley,P., Boutell,J.M., Di Antonio,M. and Balasubramanian,S.Whole genome experimental maps of DNA G-quadruplexes in multiple species(2019). *Nucleic Acids Res.*, **47**, 3862–3874.
28. Wu,F., Niu,K., Cui,Y., Li,C., Lyu,M., Ren,Y., Chen,Y., Deng,H., Huang,L., Zheng,S. *et al.* (2021) Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Commun. Biol.*, **4**, 98.
29. Puig Lombardi,E., Holmes,A., Verga,D., Teulade-Fichou,M.-P., Nicolas,A. and Londoño-Vallejo,A. (2019) Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species. *Nucleic Acids Res.*, **47**, 6098–6113.
30. Ding,Y., Fleming,A.M. and Burrows,C.J. (2018) Case studies on potential G-quadruplex-forming sequences from the bacterial orders deinococcales and thermales derived from a survey of published genomes. *Sci. Rep.*, **8**, 15679.
31. Dey,U., Sarkar,S., Teronpi,V., Yella,V.R. and Kumar,A. (2021) G-quadruplex motifs are functionally conserved in cis-regulatory regions of pathogenic bacteria: an in-silico evaluation. *Biochimie*, **184**, 40–51.
32. Lavezzo,E., Berselli,M., Frasson,I., Perrone,R., Palù,G., Brazzale,A.R., Richter,S.N. and Toppo,S. (2018) G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide. *PLoS Comput. Biol.*, **14**, e1006675.
33. Ghosh,A., Largy,E. and Gabelica,V. (2021) DNA G-quadruplexes for native mass spectrometry in potassium: a database of validated structures in electrospray-compatible conditions. *Nucleic Acids Res.*, **49**, 2333–2345.
34. Yadav,V.K., Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: genome-wide database of G4 DNA–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*, **36**, D381–D385.
35. Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.
36. Zhang,R., Lin,Y. and Zhang,C.-T. (2008) Greglist: a database listing potential G-quadruplex regulated genes. *Nucleic Acids Res.*, **36**, D372–D376.
37. Hong,X., Zheng,J., Xie,J., Tong,X., Liu,X., Song,Q., Liu,S. and Liu,S. (2021) RR3DD: an RNA global structure-based RNA three-dimensional structural classification database(2021). *RNA Biol.*, **18**, 738–746.
38. Li,Q., Xiang,J.-F., Yang,Q.-F., Sun,H.-X., Guan,A.-J. and Tang,Y.-L. (2013) G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res.*, **41**, D1115–D1123.
39. Wang,Y.-H., Yang,Q.-F., Lin,X., Chen,D., Wang,Z.-Y., Chen,B., Han,H.-Y., Chen,H.-D., Cai,K.-C., Li,Q. *et al.* (2021) G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands(2022). *Nucleic Acids Res.*, **50**, D150–D160.