Data in Brief 50 (2023) 109550

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of bulged G-quadruplex forming sequences in the human genome

Csaba Papp[a], Piroon Jenjaroenpun[b,c], Vineeth T. Mukundan[d], Anh Tuân Phan[d,e], Vladimir A. Kuznetsov[a,c,*]

[a] Department of Urology, Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University, Syracuse, NY 13210, USA
[b] Division of Bioinformatics and Data Management for Research, Research Group and Research Network Division, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
[c] Bioinformatics Institute, A*STAR Biomedical Institutes, Singapore, Singapore
[d] School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
[e] NTU Institute of Structural Biology, Nanyang Technological University, Singapore 636921, Singapore

ARTICLE INFO

Article history:
Received 2 May 2023
Revised 24 August 2023
Accepted 1 September 2023
Available online 6 September 2023

Dataset link: Potential bulged G-quadruplex forming sequences (pG4-BS) in the human genome (Original data)

Keywords:
G-quadruplex dataset
G4-bulge structures
DNA
Coordinates
Computational modelling and search algorithm

ABSTRACT

When several continuous guanine runs are present closely in a nucleic acid sequence, a secondary structure called G-quadruplex can form (G4s). Such structures in the genome could serve as structural and functional regulators in gene expression, DNA-protein binding, epigenetic modification, and genotoxic stress. Several types of G4-forming DNA sequences exist, including bulged G4-forming sequences (G4-BS). Such bulges occur due to the presence of non-guanine bases in specific locations (G-runs) in the G4-forming sequences. At present, search algorithms do not identify stable G4-BS conformations, making genome-wide studies of G4-like structures difficult. Data provided in this study are related to a published article "Stable bulged G-quadruplexes in the human genome: Identification, experimental validation and functionalization" published by Nucleic Acids Research [DIO.org/10.193/nar/gkad252]. Based on our studies in vitro and G4-seq and G4 CUT&Tag data analysis, we have specified and validated three pG4-BS models. In this article, a large collection of 'raw' (unfiltered) dataset is presented, which includes three subfamilies of pG4-BS. For each of pG4-BS, we provide strand-specific genomic boundaries. Data on

* Corresponding author.
E-mail address: kuznetsv@upsate.edu (V.A. Kuznetsov).

https://doi.org/10.1016/j.dib.2023.109550

pG4-BS might be useful in elucidating their structural, functional, and evolutionary roles. Furthermore, they may provide insight into the pathobiology of G4-like structures and their potential therapeutic applications.

Specifications Table

Subject	G-quadruplex, bulged G-quadruples, G4 forming DNA sequence
Specific subject area	Bioinformatics/Computational Biology/Genomics
Type of data	Table
How the data were acquired	Initially, a set of artificial G4-rich sequences with and without non-guanine nucleotide insertions in one or two short G-repeats were chemically sequenced on ABI 394 DNA synthesizer and the samples were prepared and then studied using NMR, UV melting, and circular dichroism (CD) experiments described in[1,2]. Three sequence models (motifs) were established that correspond to tested sequences that displayed bulge-containing G-quadruplex forming ability in our thermodynamic stability assays (nuclear magnetic resonance (NMR), Circular dichroism (CD), DNA UV melting). Characteristic NMR and CD spectra, and UV melting curve profiles indicate whether the tested sequence folds into a stable G4 structure or not. We used these experimental methods to identify DNA sequences that form G4 conformations in vitro and optimize the nucleotide composition of G4-BS, providing a selection of stable pG4-BS models. The in-house motifs sequence search algorithm is written in Python 2.7 and is available at https://doi.org/10.6084/m9.figshare.22110965.
Data format	Raw (unfiltered)
Description of data collection	The in-house motifs sequence search algorithm was written in Python 2.7. We included any pG4-BS that matched one of the three sequence models, and passed the following filters: - The sequence must contain at least four guanine clusters with at least three guanines present in each - Non-guanine insertions in guanine clusters must be between 0 to 3 non-guanine nucleotides - The lengths of the loops connecting guanine clusters must be between 1 to 3 nucleotides - The number of bulges in a sequence must be equal to or less than two - pG4-BS with contiguous cytosines are not allowed - pG4-BS with more than one loop consisting of a single guanine are not allowed
Data source location	SUNY Upstate Medical University Syracuse, NY, USA
Data accessibility	Dataset: Papp, Csaba; Jenjaroenpun, Piroon; Mukundan, Vineeth T.; Phan, Anh T.; Kuznetsov, Vladimir (2023), "Potential bulged G-quadruplex forming sequences (pG4-BS) in the human genome", Mendeley Data, V3, doi:10.17632/w37rx9hpb7.3 https://data.mendeley.com/datasets/w37rx9hpb7 Software: https://doi.org/10.6084/m9.figshare.22110965
Related research article	C. Papp, V.T. Mukundan, P. Jenjaroenpun, F.R. Winnerdy, G.S. Ow, A.T. Phan and V.A. Kuznetsov. Stable bulged G-quadruplexes in the human genome: identification, experimental validation and functionalization, Nucleic Acids Research, doi:10.193/nar/gkad252

1. Value of the Data

- This dataset of stable bulged G-quadruplex forming sequences (pG4-BS) was generated via a computational prediction algorithm utilizing three experimentally validated stable G4-BS

models. This provides an unbiased identification of potential sites of bulged G-quadruplex structures genome-wide and in artificial sequences.
- This dataset is – to the best of our knowledge - the first computationally derived genome-wide map of pG4-BS sites in the human genome (Assembly GRCh38).
- This genome-wide map of pG4-BS provides a novel sequence resource for biologists and technologists with an interest in the study and potential uses of stable G4-like structures in genome biology, disease diagnosis, drug development, and treatment.

2. Objective

Bulged G4s are a subset of G4-like structures that do not adhere to the canonical G4 motif. The identification of canonical G4 sequences was the first instance of using a search algorithm to provide genome-wide maps of any type of G4 [3]. Following this, several alternative computational tools have been developed for the identification of G4-like sequences (reviewed in [4]). In addition, an increasing number of studies have shown that sequences exist in the human genome that could theoretically form bulged G4s [1,2,5-7]. Despite these advances, the constraints of bulged G4 formation have not been elucidated, preventing the development of genome-wide search algorithms for these sequences. The objective of this work is to provide access to the original collection of data representing the stable pG4-BS identified in the human genome that related to a published article "Stable bulged G-quadruplexes in the human genome: Identification, experimental validation and functionalization" published by Nucleic Acids Research [https://doi.org/10.1093/nar/gkad252]. Our search identified 1,935,686 individual pG4-BS in the human genome. For each of pG4-BS, we provide strand-specific genomic boundaries. The publication of this data provides full access to this novel dataset in its entirety, for which only processed versions are available with the original research article.

3. Data Description

We utilized a sliding-window approach allowing us to separately detect pG4-BS that had any degree of overlap between their sequences. The columns included in the dataset are the chromosome the pG4-BS is located on (**chr**), the start site (**start**) and end site (**end**) of the pG4-BS, the pG4-BS sequence (**Sequence**), a unique identifier (**ID**), the strand where the pG4-BS is located (**strand**), the number of intact guanine clusters in the sequence (**Intact_G_clusters**), the number of bulges in the sequence (**Bulges**). In addition, the dataset contains several columns that contain the pG4-BS sequence formatted in different ways to highlight the sequence composition of guanine clusters and the nucleotides connecting them (**Sequence_breakdown**), and the identity (**Non_G_cluster_nucleotides**) and length (**Length_of_non_G_cluster_nucleotides**) of non-guanine cluster nucleotides in the sequence. Finally, the **G4BS_model** column indicates which of the three G4-BS models the given sequence belongs to.

4. Experimental Design, Materials, and Methods

Our search algorithm utilized a regular expression to capture potential bulged G4 forming sequences (pG4-BS). This regular expression matches any sequence where there are at least four guanine clusters consisting of at least three guanines, the guanine clusters contain between 0 to 3 non-guanine nucleotides and the lengths of the loops connecting said guanine clusters are between 1 to 3 nucleotides. This initial search does not discriminate between pG4-BS based on the number and length of bulge sites. For example, both of the following sequences will be captured: (a) GGG-T-GAGG-T-GGG-T-GGG, (b) GATGG-GAT-GACCGCCG-T-GAGTG-T-GGCCG. Based on available literature and our experiments, we set up filtering steps to specifically capture pG4-BS with

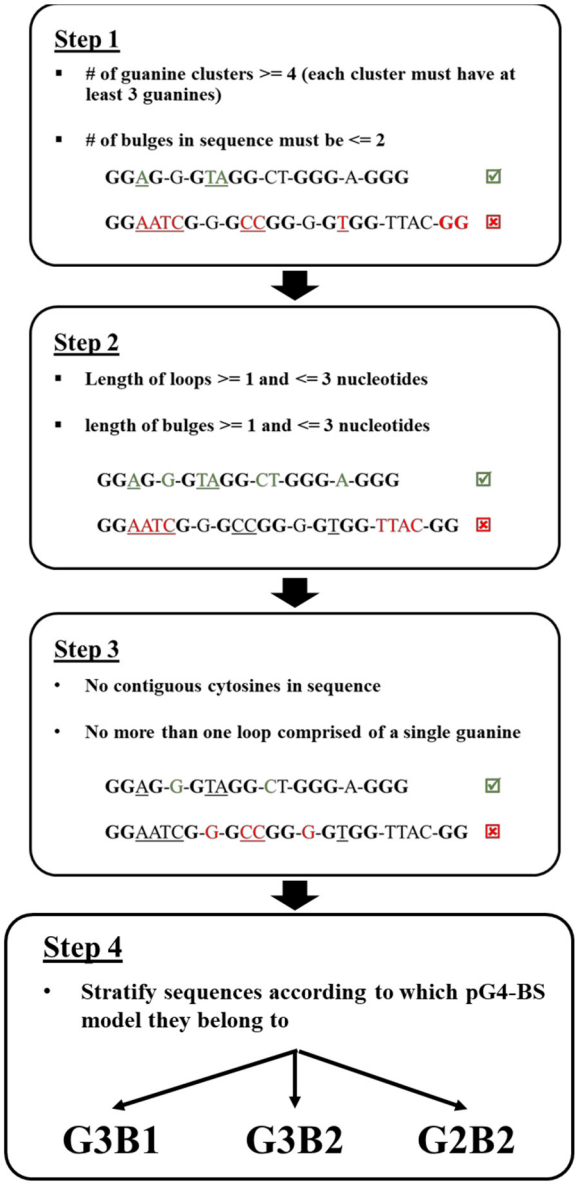

Fig. 1. Flowchart illustrating the filtering criteria utilized by our algorithm to identify pG4-BS in the human genome. As shown in the example sequences, nucleotides that obey or violate the rules of a given step are colored green or red respectively. Nucleotides located in the guanine clusters and loop regions are separated from each other by dashes. The nucleotides comprising the bulges are indicated by underline in the examples. The classification of three pG4-BS models takes place in Step 4.

a higher probability of being able to form bulged G4s [1,2,8]. These are: (i) the number of bulge sites in a sequence must be equal to or less than two (e.g., *GGG-T-G**a**GG-T-GG**t**G-T-GGG*: allowed; *G**t**GT**t**G-T-GGG-T-G**a**GG-T-GGG*: not allowed), (ii) pG4-BS with contiguous cytosines are not allowed (e.g., *G**cc**GG-T…*), (iii) pG4-BS with more than one loop consisting of single guanine are not allowed (e.g., *GGaG-T-GGG-**G**-GGaG-**G**-GGG*) (Fig. 1). Our search algorithm captures pG4-BS from both DNA strands. The code used is available at https://doi.org/10.6084/m9.figshare.22110965.

CRediT Author Statement

Csaba Papp: Software, Formal analysis, Writing – Original Draft, Data Curation; **Piroon Jenjaroenpun**: Software, Algorithm Developing, Formal analysis; **Vineeth Thachappilly Mukundan**: Investigation, Methodology; **Anh Tuân Phan**: Conceptualization, Supervision, Writing – Review and Editing; **Vladimir A. Kuznetsov**: Conceptualization, Supervision, Algorithm Developing, Writing – Review- Final manuscript version, and Editing

Ethics Statement

This work does not involve studies with animals or humans.

Data Availability

Potential bulged G-quadruplex forming sequences (pG4-BS) in the human genome (Original data) (Mendeley Data)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially funded institutionally by the Bioinformatics Institute, Biomedical Institutes/A-STAR, Singapore. Also, V.A.K. was supported by a SUNY EMPIRE innovation program scholar grant, the Upstate Medical University Cancer Center grant, and the Upstate Foundation Turn4ACure Fund. Research in A.T.P. lab was supported by Nanyang Technological University Singapore. P.J. was supported by the Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation (OPS MHESI), Thailand Science Research and Innovation (TSRI) (Grant No. RGNS 64-161).

References

[1] V.T. Mukundan, N.Q. Do, A.T. Phan, HIV-1 integrase inhibitor T30177 forms a stacked dimeric G-quadruplex structure containing bulges, Nucleic. Acids. Res. 39 (2011) 8984–8991.
[2] V.T. Mukundan, A.T. Phan, Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences, J. Am. Chem. Soc. 135 (2013) 5017–5028.
[3] J.L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome, Nucleic. Acids. Res. 33 (2005) 2908–2916.
[4] E. Puig Lombardi, A Londono-Vallejo, A guide to computational methods for G-quadruplex prediction, Nucleic. Acids. Res. 48 (2020) 1–15.
[5] T.Q. Ngoc Nguyen, K.W. Lim, A.T Phan, Duplex formation in a G-quadruplex bulge, Nucleic. Acids. Res. 48 (2020) 10567–10575.
[6] P. Das, K.H. Ngo, F.R. Winnerdy, A. Maity, B. Bakalar, Y. Mechulam, E. Schmitt, A.T. Phan, Bulges in left-handed G-quadruplexes, Nucleic. Acids. Res. 49 (2021) 1724–1736.
[7] A. Varizhuk, D. Ischenko, V. Tsvetkov, R. Novikov, N. Kulemin, D. Kaluzhny, M. Vlasenok, V. Naumov, I. Smirnov, G. Pozmogova, The expanding repertoire of G4 DNA structures, Biochimie 135 (2017) 54–62.
[8] A. Bedrat, L. Lacroix, J.L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, Nucleic. Acids. Res. 44 (2016) 1746–1759.