JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 0, 2019

Pp. 1–7
DOI: 10.1089/cmb.2019.0010

Research Article

Plant-GQ: An Integrative Database of G-Quadruplex in Plant

FANGFANG GE,[1,*] YI WANG,[2,*] HUAYANG LI,[1] RUI ZHANG,[1] XIAOTONG WANG,[1] QINGYUN LI,[1] ZHENCHANG LIANG,[2] and LONG YANG[1]

ABSTRACT

G-quadruplex (G-Q) is advanced DNA or RNA secondary structures frequently found in plant and involved in important biological processes such as transcription, translation, and telomere maintenance. Although some databases and tools were developed for predicting and studying G-Q, none of them was for plant. With the development of next-generation sequencing technology, a large number of plant genomes have been assembled and annotated to provide opportunities for mining G-Q. Plant G-quadruplex database (Plant-GQ) was constructed for predicting G-Q in 195 plants. It has a total of 626,341,645 predicted G-Qs. The database contains four major parts: Search, Tools, JBrowse, and Download. Not only G-Q information but also online forecasting tool can be retrieved and obtained from Plant-GQ. It can also browse and analyze G-Q information by JBrowse in a graph visualization interface. Considering the key role of G-Q in plant, this database will play an important status in the study of the structure, function, and biological relevance of G-Q in plant.

Keywords: database, G-quadruplex, plant.

HIGHLIGHTS

1. Plant-GQ is the first database for G-quadruplexes in plant. 2. It contains 195 plant species, a total of 626,341,645 predicted G-quadruplexes. 3. It also provides a user-friendly platform for querying, browsing, and downloading G-quadruplex information.

1. INTRODUCTION

PLANT DNA, IN ADDITION TO DOUBLE HELICAL B-DNA, has various extra-helices (Bochman et al., 2012). G-quadruplex (G-Q) is one of the secondary structures that is ubiquitous and plays an important role in the physiological functions of all living organisms. Structurally, G-Qs are formed by π–π stacking of G-quartet composed of four guanines connected by Hoogsteen hydrogen bonds (Guedin et al., 2018). The layout and bonding to make up G-Q are not random and have very unusual functional purposes (Bochman

[1]Agricultural Big-Data Research Center and College of Plant Protection, Shandong Agricultural University, Taian, China.
[2]Beijing Key Laboratory of Grape Science and Enology and Key Laboratory of Plant Resource, Institute of Botany, Chinese Academy of Sciences, Beijing, China.
*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

et al., 2012; Cheng et al., 2018). Depending on the direction of the strands that makes up the quartets, structures can be described as parallel or antiparallel (Moye et al., 2015). These quadruplex structures are stabilized by cations, especially potassium ion (K+). Cations are located in the central channel between each pair of quartets (Bhattacharyya et al., 2016). The complex structure of G-Q determines the complexity of its function.

Since Sen and Gilbert discovered G-Q (Sen and Gilbert, 1988), the study of G-Q has become increasingly hot and developed into a field of preface. The reason for this phenomenon is that G-Qs are widely present in eukaryotic and prokaryotic and concentrate on many functional areas (Beaume et al., 2013; Jackowiak et al., 2017). For example, a high proportion of potential G-Q structures are in G-rich regions of telomeres (Moye et al., 2015; Noer et al., 2016), promoters (Beaume et al., 2013; Voter et al., 2018), mitotic and meiotic double-strand break sites (Lemmens et al., 2015), ribosomal DNA (Wallgren et al., 2016), transcriptional start sites (Morgan et al., 2016), and untranslated regions of messenger RNA (Stefanovic et al., 2015). In human and mammal, the structure of G-Q is related to the instability of the genome (Yadav et al., 2016) and its widespread existence and occurrence in proto-oncogenes (Thandapani et al., 2015). These proto-oncogenes include c-myc (Borgognone et al., 2010), c-myb (Miyazaki et al., 2012), c-kit (Wei et al., 2015), bcl-2 (Feng et al., 2016), KRAS (Cogoi and Xodo, 2006), RET (Garg et al., 2016), VEGF (Sun et al., 2005), and HIF-1 (Du et al., 2009). Distribution of these G-Qs suggests their key role in cancer progression (Wolfe et al., 2014). In bacteria, G-Q is likely to concentrate on the promoter region and regulate the transcription, signaling, and other special functions (Rawal et al., 2006; Beaume et al., 2013). In addition, G-Q plays an important role in *Neisseria gonorrhoeae* antigenic variation (Cahoon and Seifert, 2009).

G-Q is also abundant in plant. The whole-genome discovery and analysis of G-Q results from 15 monocotyledons and dicotyledons indicated that G-Q was involved in important physiological processes such as plant development, cell growth and size, and gene expression regulation (Garg et al., 2016).

In recent years, a number of databases concerning G-Qs are currently available. Greglist lists genes that contain promoter G-Qs from genomes of different species, including human, mouse, rat, and chicken (Zhang et al., 2008). GRSDB is a database of G-Q near RNA processing sites in human and mouse (Kostadinov et al., 2006). G4RNA was created to help meet the need for known RNA G-Q data (Garant et al., 2015). G4IPDB is a database for G-Q forming nucleic acid interacting proteins (Mishra et al., 2016). In addition, prediction algorithms such as QuadParser (Scaria et al., 2006), G4 calculator (Eddy and Maizels, 2006), QGRS-Conserve (Frees et al., 2014), QGRS mapper (Kikin et al., 2006), and QuadBase2 (Dhapola and Chowdhury, 2016) can also be easily accessible and used widely. However, these databases are all for human and mammal. Therefore, it is urgent to establish a comprehensive and unified database of G-Q in plant.

With the reduction in the cost of sequencing, more and more plant genomes have been assembled and annotated. For better study of G-Qs in plant, mining and annotation of G-Qs in these genomes are essential. In this study, a structure-based approach was used for G-Q determination and annotation in the genomes of sequenced plant. Plant G-quadruplex database (Plant-GQ) contains a series of user-friendly webpage for query and presentation of G-Q. Plant-GQ will greatly accelerate the study of various regulatory effects of G-Qs. The database is obtained from http://biodb.sdau.edu.cn/plantgq/index.php

2. METHODS

2.1. Data sets

Genome and GFF3 files were downloaded from public data platforms (http://biodb.sdau.edu.cn/plantgq/Public/table/Supplementary%20Table1.xlsx). For each plant species, the G-Q information was mined and annotated in genome and GFF3.

2.2. Identification of G-Q

Based on the following G-Q motif, the G-Qs were identified by Perl scripts (Huppert and Balasubramanian, 2007):

$$\text{Gx-Ny-Gx-Nz-Gx-Nr-Gx}\ (x = 2-4, y = 1-10, z = 1-10, r = 1-10)$$

x is the number of guanine tetrads in each short G-tract. In G-Q motif, the four groups of guanines are of equal length. Ny, Nz, and Nr can be any combination of four residues A, T, G, and C, forming the loops. y, z, and r are the lengths of the loops.

2.3. System implementation

The server of Plant-GQ was constructed using Centos 6.5, Apache 2.4.27, MySQL 14.4, PHP 7.1.9, ThinkPHP 3.2.3, and xampp 3.2.2. For efficient management, query, and presentation, all G-Q information was stored in MySQL tables. The web interface was developed by HTML5, CSS, and JavaScript languages. Common Gateway Interface programs were built by Perl and PHP programming languages. A complete flowchart portraying the data collection and content management steps in building the Plant-GQ is provided with Figure 1.

3. RESULTS

3.1. Summary of G-Q

A total of 626,341,645 putative G-Qs were identified with 195 plant species. The total G-Qs of two G-tracts, three G-tracts, and four G-tracts were 610,897,949, 14,326,347, and 1,117,349, respectively. Apparently, most of them are two G-tracts, accounting for 97.43% of the total. The G-Q formed by four G-tracts is the most stable, but the proportion of the smallest, which accounted for 0.19%.

3.2. Database features and data retrieval and display tools

A complete flowchart depicts the main contents of the database (Fig. 2). To facilitate the query and browsing of G-Q information, Plant-GQ provides query function, online tool, and graphical visualization page.

The search interface allows users to filter G-Qs through four categories, including species, chromosome, G-Q type, and start and end. For species searches, a drop-down menu lists all species in Latin alphabetical order (Fig. 3A). The predicted G-Qs that comprise the search results will be listed in search results interface (Fig. 3B). The search results interface is a list of 30 rows at a time. Each G-Q entry contains details of the species, chromosome, type, start, end, motif, and location. If more than 30 results are generated in a search, users can navigate between pages using the paging bar at the top of search results interface. In addition, the

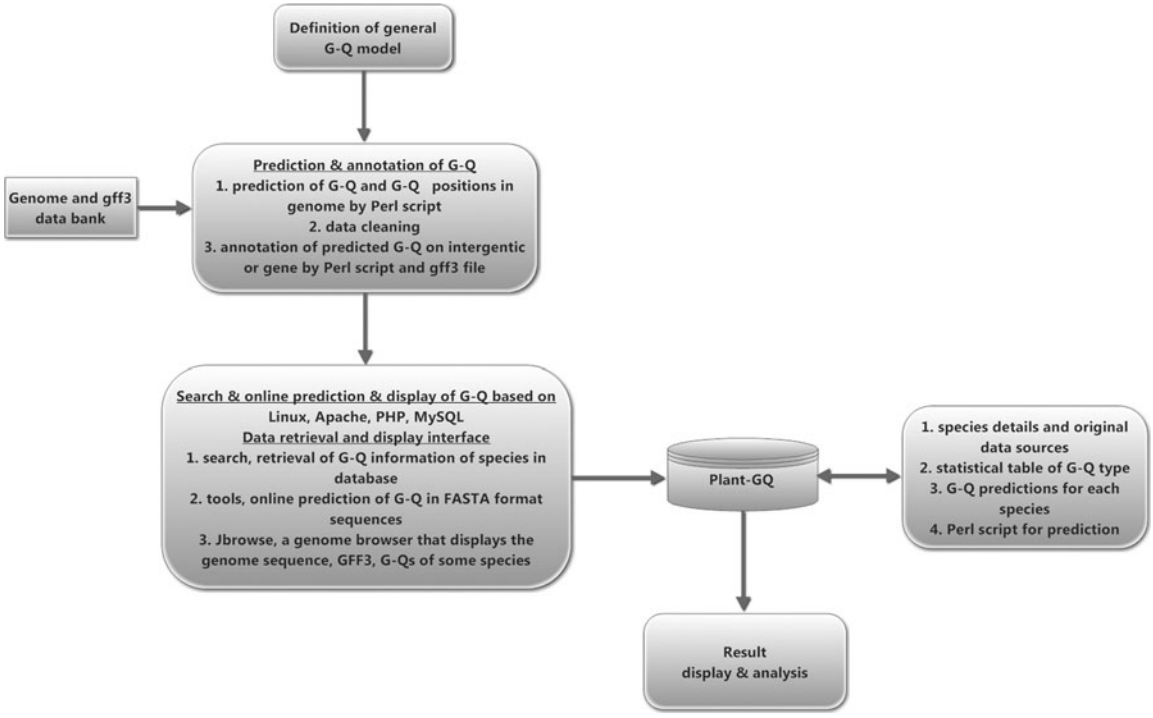

FIG. 1. Schematic representation of Plant-GQ. G-Q, G-quadruplex; Plant-GQ, Plant G-quadruplex database.

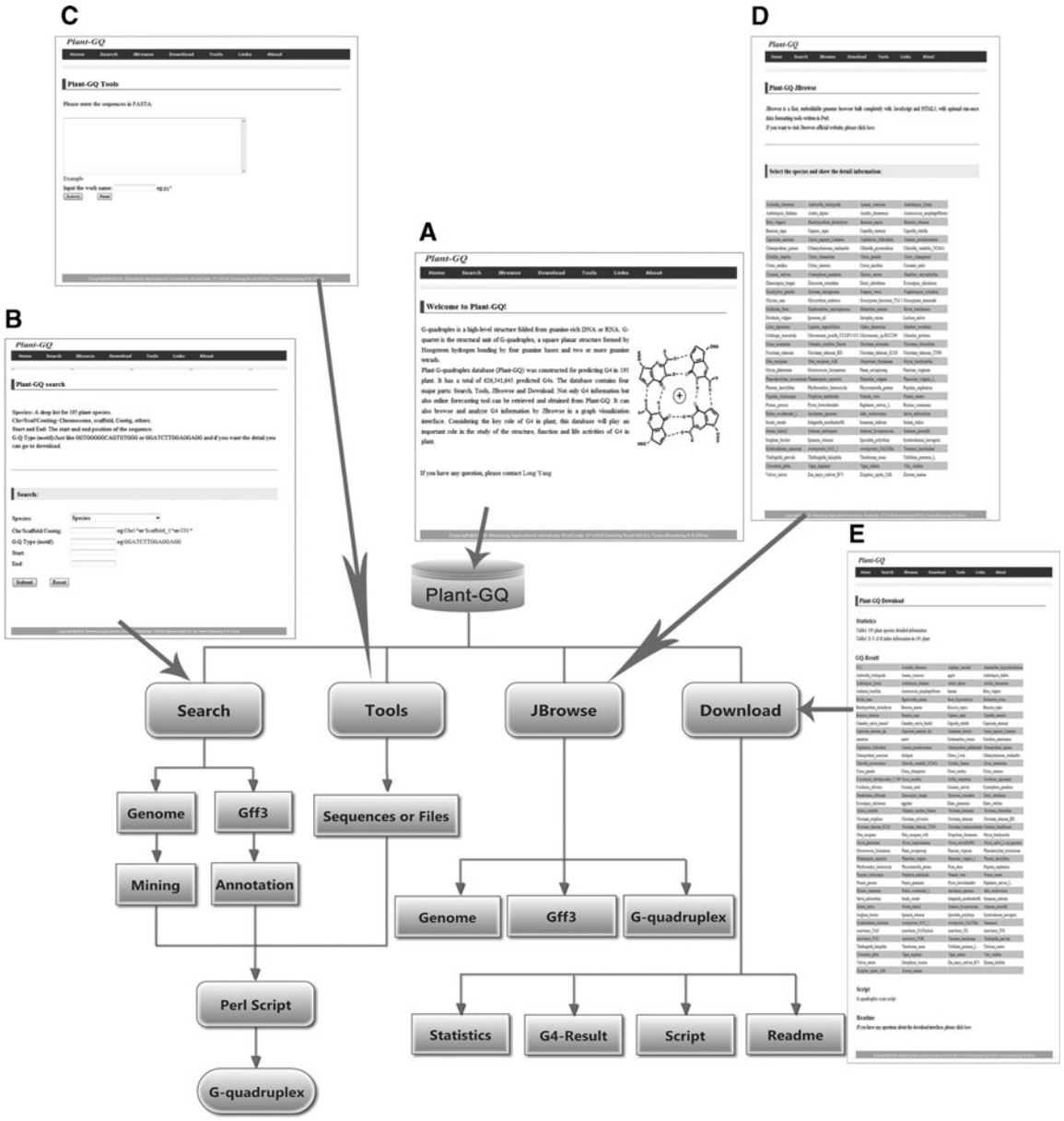

FIG. 2. Overview of Plant-GQ. This database contains four major parts: Search, Tools, JBrowse, and Download. **(A)** Home interface. The interface contains G-Q and Plant-GQ summary. In addition, the schematic diagrams of G-Q and its structural unit G-quartet are also shown. **(B)** Search interface. The interface is mainly divided into two parts, introduction of search conditions and search box. Users can search G-Q information by species name, chromosome, G-Q type, and start and end position of the sequence. After clicking submit, the query results will be displayed in a tabular form of a new interface. **(C)** Tools interface. An online mining tool is provided to query G-Q. Sequences in FASTA format can be pasted directly in textarea. The query results will open with a new interface. **(D)** JBrowse interface. JBrowse provides a convenient genome browser that displays the genome sequence, GFF3, G-Qs of some species. Users can select a species in the table for large-scale browsing with graphical visual interface. **(E)** Download interface. G-Q information, including statistics and tables, G-Q motif for all species, and script for mining G-Q structures, is provided in download interface.

total number of filtered results is shown at the top of the paging bar. All results of search results interface can be downloaded in batches of ''GQ-result'' in download interface.

A tools interface is mainly used for online prediction of G-Qs. Common sequences in FASTA format, such as gene or CDS sequences, can be pasted directly into textarea (Fig. 3C). The query results include chromosomes, positive and negative chains, start and end, motif, G-tract, and loops type which will be

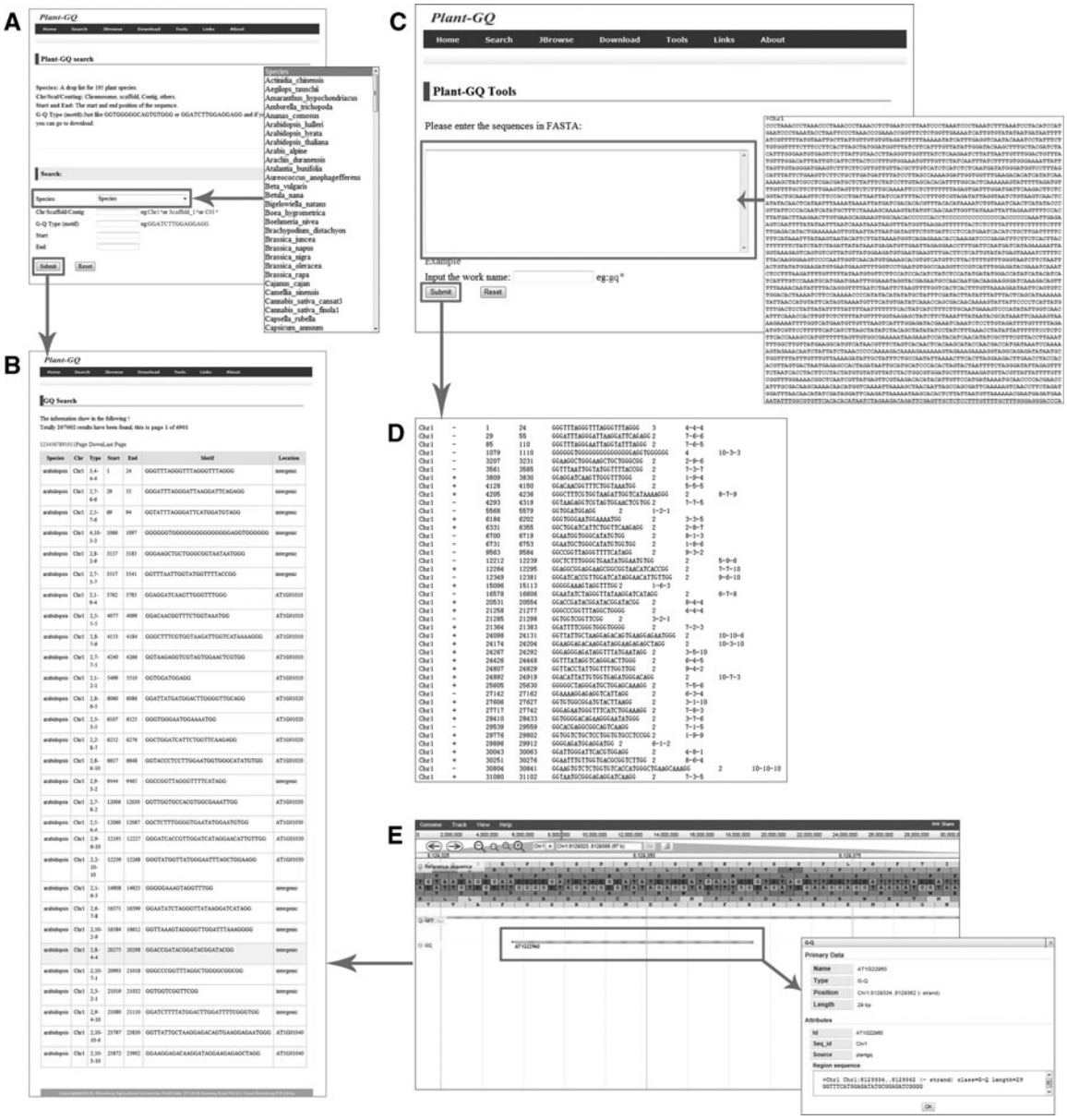

FIG. 3. Screenshots of data retrieval and display interface. **(A)** Screenshots of search interface. **(B)** Screenshots of a sample search results interface. Filtering options are provided by users to filter results for a list of G-Qs. Users can quickly view each G-Q information on search results interface. Search results include species name, chromosome, G-Q type, start, end, motif, and location. **(C)** Screenshots of tools interface. **(D)** Screenshots of a sample tools results interface. FASTA sequences are provided by users to predict G-Qs, the predicted results include chromosome, positive and negative chains, start, end, motif, G-tract, and loops type. **(E)** Screenshots of a sample JBrowse results interface. Through the genome browser JBrowse, users can conveniently view the elaborate information of G-Qs by simply clicking the species name form JBrowse interface. In addition, the elaborate information of genome and GFF3 can also be provided for analysis.

opened with a new interface (Fig. 3D). Full information about the predicted G-Qs can also be downloaded in tools results interface.

JBrowse provides a convenient genome browser that displays the genome sequence, GFF3, G-Qs of some species. Select a species in the table for large-scale browsing in the graphical visual interface. In graphical visualization interface, the left column lists the reference sequence of the plant species genome, GFF3 choice, and G-Q locations. The GFF3 information and G-Q locations will be displayed in the right column as a histogram when the selection item in the left column is clicked. The top column allows to

select the species' chromosomes and zoom in or out of the results. The detail information about G-Q can also be conveniently browsed by clicking on the G-Q name of the visual interface (Fig. 3E).

4. CONCLUSIONS AND FUTURE DIRECTIONS

Plant-GQ provides a user-friendly platform for querying, browsing, and downloading G-Q information. The database contains 626,341,645 G-Qs. In addition, Plant-GQ accelerates scientific research by mining G-Q for analysis of its various regulatory effects in plant. In the future, more G-Q structures will be identified as the information on genomes continues to improve. Furthermore, Plant-GQ will be upgraded and will optimize the database interface at regular intervals.

ACKNOWLEDGMENTS

This work was supported by the Foundation of Shandong Province Modern Agricultural Technology System Innovation Team (SDAIT-25-02 to L.Y.) and the Open Project Program of Beijing Key Laboratory of Grape Science and Enology.

AUTHOR DISCLOSURE STATEMENT

The authors declare that there are no competing financial interests.

REFERENCES

Beaume, N., Pathak, R., Yadav, V.K., et al. 2013. Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: Radioresistance of *D. radiodurans* involves G4 DNA-mediated regulation. *Nucleic Acids Res.* 41, 76–89.

Bhattacharyya, D., Mirihana Arachchilage, G., and Basu, S. 2016. Metal cations in G-quadruplex folding and stability. *Front Chem.* 4, 38.

Bochman, M.L., Paeschke, K., and Zakian, V.A. 2012. DNA secondary structures: Stability and function of G-quadruplex structures. *Nat Rev Genet.* 13, 770–780.

Borgognone, M., Armas, P., and Calcaterra, N.B. 2010. Cellular nucleic-acid-binding protein, a transcriptional enhancer of c-Myc, promotes the formation of parallel G-quadruplexes. *Biochem J.* 428, 491–498.

Cahoon, L.A., and Seifert, H.S. 2009. An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. *Science.* 325, 764–767.

Cheng, M., Cheng, Y., Hao, J., et al. 2018. Loop permutation affects the topology and stability of G-quadruplexes. *Nucleic Acids Res*. 46, 9264–9275.

Cogoi, S., and Xodo, L.E. 2006. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.* 34, 2536–2549.

Dhapola, P., and Chowdhury, S. 2016. QuadBase2: Web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.* 44, W277–W283.

Du, Z., Zhao, Y., and Li, N. 2009. Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Res.* 37, 6784–6798.

Eddy, J., and Maizels, N. 2006. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* 34, 3887–3896.

Feng, Y., Yang, D., Chen, H., et al. 2016. Stabilization of G-quadruplex DNA and inhibition of Bcl-2 expression by a pyridostatin analog. *Bioorg Med Chem Lett.* 26, 1660–1663.

Frees, S., Menendez, C., Crum, M., et al. 2014. QGRS-Conserve: A computational method for discovering evolutionarily conserved G-quadruplex motifs. *Hum. Genomics.* 8, 8.

Garant, J.M., Luce, M.J., and Scott, M.S. 2015. G4RNA: An RNA G-quadruplex database. *Database*. 2015, bav059.

Garg, R., Aggarwal, J., and Thakkar, B. 2016. Genome-wide discovery of G-quadruplex forming sequences and their functional relevance in plants. *Sci Rep.* 6, 28211.

Guedin, A., Lin, L.Y., Armane, S., et al. 2018. Quadruplexes in ''Dicty'': Crystal structure of a four-quartet G-quadruplex formed by G-rich motif found in the *Dictyostelium discoideum* genome. *Nucleic Acids Res.* 46, 5297–5307.

Huppert, J.L., and Balasubramanian, S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* 35, 406–413.

Jackowiak, P., Hojka-Osinska, A., Gasiorek, K., et al. 2017. Effects of G-quadruplex topology on translational inhibition by tRNA fragments in mammalian and plant systems in vitro. *Int J Biochem Cell Biol.* 92, 148–154.

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.

Kostadinov, R., Malhotra, N., Viotti, M., et al. 2006. GRSDB: A database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic Acids Res.* 34, D119–D124.

Lemmens, B., van Schendel, R., and Tijsterman, M. 2015. Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. *Nat Commun.* 6, 8909.

Mishra, S.K., Tawani, A., Mishra, A., et al. 2016. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep.* 6, 38144.

Miyazaki, T., Pan, Y., Joshi, K., et al. 2012. Telomestatin impairs glioma stem cell survival and growth through the disruption of telomeric G-quadruplex and inhibition of the proto-oncogene, c-Myb. *Clin Cancer Res.* 18, 1268–1280.

Morgan, R.K., Batra, H., Gaerig, V.C., et al. 2016. Identification and characterization of a new G-quadruplex forming region within the kRAS promoter as a transcriptional regulator. *Biochim Biophys Acta.* 1859, 235–245.

Moye, A.L., Porter, K.C., Cohen, S.B., et al. 2015. Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat Commun.* 6, 7643.

Noer, S.L., Preus, S., Gudnason, D., et al. 2016. Folding dynamics and conformational heterogeneity of human telomeric G-quadruplex structures in Na+ solutions by single molecule FRET microscopy. *Nucleic Acids Res.* 44, 464–471.

Rawal, P., Kummarasetti, V.B., Ravindran, J., et al. 2006. Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation. *Genome Res.* 16, 644–655.

Scaria, V., Hariharan, M., Arora, A., et al. 2006. Quadfinder: Server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res.* 34, W683–W685.

Sen, D., and Gilbert, W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature.* 334, 364.

Stefanovic, S., DeMarco, B.A., Underwood, A., Williams, K.R., Bassell, G.J., and Mihailescu, M.R. 2015. Fragile X mental retardation protein interactions with a G quadruplex structure in the 3′-untranslated region of NR2B mRNA. *Mol Biosyst.* 11, 3222–3230.

Sun, D., Guo, K., Rusche, J.J., et al. 2005. Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res.* 33, 6070–6080.

Thandapani, P., Song, J., Gandin, V., et al. 2015. Aven recognition of RNA G-quadruplexes regulates translation of the mixed lineage leukemia protooncogenes. *Elife.* 4, 4.

Voter, A.F., Qiu, Y., Tippana, R., et al. 2018. A guanine-flipping and sequestration mechanism for G-quadruplex unwinding by RecQ helicases. *Nat Commun.* 91, 4201.

Wallgren, M., Mohammad, J.B., Yan, K.P., et al. 2016. G-rich telomeric and ribosomal DNA sequences from the fission yeast genome form stable G-quadruplex DNA structures in vitro and are unwound by the Pfh1 DNA helicase. *Nucleic Acids Res.* 44, 6213–6231.

Wei, D., Husby, J., and Neidle, S. 2015. Flexibility and structural conservation in a c-KIT G-quadruplex. *Nucleic Acids Res.* 43, 629–644.

Wolfe, A.L., Singh, K., Zhong, Y., et al. 2014. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature.* 513, 65–70.

Yadav, P., Owiti, N., and Kim, N. 2016. The role of topoisomerase I in suppressing genome instability associated with a highly transcribed guanine-rich sequence is not restricted to preventing RNA:DNA hybrid accumulation. *Nucleic Acids Res.* 44, 718–729.

Zhang, R., Lin, Y., and Zhang, C.T. 2008. Greglist: A database listing potential G-quadruplex regulated genes. *Nucleic Acids Res.* 36, D372–D376.

Address correspondence to:
Dr. Long Yang
Agricultural Big-Data Research Center and College of Plant Protection
Shandong Agricultural University
Taian 271018, China

E-mail: lyang@sdau.edu.cn