Genome-wide identifying of G-quadruplex structures directly by whole-genome resequencing

Jing Tu (✉ jtu@seu.edu.cn)

State Key Laboratory of Bioelectronics, Southeast University <https://orcid.org/0000-0002-1310-5915>

Mengqin Duan

State Key Laboratory of Bioelectronics, Southeast University

Wenli Liu

State Key Laboratory of Bioelectronics, Southeast University

Na Lu

Southeast University

Xiao Sun

Southeast University

Zuhong Lu

Southeast University

Article

Keywords: DNA G-quadruplex (G4), profiling, sequencing

Posted Date: January 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-149290/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on October 14th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-26312-w>.

Abstract

We present a convenient genome-wide DNA G-quadruplex (G4) profiling method that identifies G4 structures from ordinary whole-genome resequencing data by seizing the slight fluctuation of sequencing quality. We identified 736,689 G4 structures within human genome, in which 44.9% of all predicted canonical G4-forming sequences were contained. We observed that some of the single nucleotide variations (SNVs) influenced the formation of G4 structures, including homozygous SNVs and heterozygous SNVs.

Due to SNVs contain individual differences, the given approach is available to identify and characterize G4s genome-wide for specific individuals.

Introduction

G-quadruplex (G4) structure is an alternative conformation of the double helical structure of B-form DNA arisen in guanine-rich sequences¹. G4s are also discovered to arise in single-stranded RNA sequences^{2,3}. One four-stranded G4 is the stacks of two or more planar G-tetrads, and one G-tetrad forms through Hoogsteen hydrogen-bonding of four guanines⁴. G4 structures are considered to be knot-like structures and have been showed stable formed under near-physiological conditions *in vitro*^{5,6}. G4 structures are stabilized by cations with stabilizing preference of monovalent cations according to the following order: $K^+ > Na^+ > Li^+$ ⁷. The formation of G4 structures in genome can influence chromatin architecture and many fundamental biology processes, such as DNA replication and gene regulation⁸⁻¹⁰. Identifying G4s genome-wide or transcriptome-wide will accelerate the study on the formation, stability and function of these structures.

In the last decade, several methods have been developed to identify G4 in whole genome or whole transcriptome, including computational sequence analyses and experiments. By searching simple consensus sequence^{11,12} or accommodate structural variants¹³⁻¹⁵, computational tools have helped to identify potential G4s in given genome^{16,17}. Computational predictions employ algorithms derived from experimental data on a small number of sequences, and require a thorough experimental validation. Using chromatin immunoprecipitation followed by high-throughput sequencing (G4 ChIP-seq), DNA G4s can be detected and mapped to genome^{18,19}. Antibodies against known G4-binding proteins are used to infer G4s in these approaches. To be noticed, G4 ChIP-seq depends on specific antibodies and potential biases may be introduced^{20,21}. The presence of folded G4 in DNA can stall a DNA polymerase^{22,23}. Coupling with high-throughput sequencing (G4-seq), more than 700,000 DNA G4 sites were identified based on genome-wide DNA polymerase-stop assay²⁴. In G4-seq, each DNA template is sequenced twice. The initial sequencing run uses the monovalent cation Na^+ that do not stabilize G4s to ensure accurate sequencing, and the second sequencing run is under G4-stabilizing conditions (in K^+ or with addition of pyridostatin). By mismatch quantification based on the results of the first sequencing run, the folded G4s are detected genome-wide *in vitro*. Similarly, RNA G4 structures were detected in high-throughput using

reverse transcriptase stalling (rG4-seq) ²⁵. G4-seq is a fast and powerful approach to identify G4 sites in genome, which included various non-canonical G4 structures that are difficult to predict. Optimized version of G4-seq has been used to generate whole genome G4 maps for 12 species ²⁶. However, G4-seq depends on the accommodation of sequencing buffer, which is unreachable to most researchers. Two sequencing runs also mean double cost. If G4 structures can be directly profiled from standard high-throughput sequencing data, the procedure will be more friendly and the cost will be greatly decreased.

Here, we present G4-miner, a genome-wide approach to directly identify G4 structures from standard sequencing data. G4-miner inspects sequencing quality in whole genome, seizes unexpected fluctuations in local, and ascribes some of the quality fluctuations to the unstable formation of G4. We sequenced DNA from primary human B lymphocytes under standard Illumina sequencing condition. Although the standard Illumina sequencing buffers do not cause strong G4 stabilization ^{24,27}, a slight and inconstant drop of sequencing quality was recorded from the beginning of the G4 structure (**Fig. 1**). We mined G4 sites in human genome, appraised the results with the results of computational prediction and G4-seq, and evaluated the influence of single nucleotide variations (SNVs) in G4 formation using the seized G4 sites. The given method provides a friendly and convenient way to identify G4 structure genome-wide.

Results

Principle and process of G4-miner

In G4-miner, genomic DNAs are sequenced only one time using standard Illumina sequencing protocol and standard buffers. The standard sequencing buffers has been optimized and do not cause strong G4 stabilization ^{24,27}. However, destabilized G4 structures still influence the sequencing reaction. Instead of altering sequencing readout, a slight and inconstant drop of sequencing quality was recorded from the beginning of the G4 structure (**Fig. 1**). In a standard sequencing experiment, DNA fragments in clusters are sequenced and the phred quality scores are calculated for each read (cluster). The unstably formed G4 causes unexpected fluctuations of sequencing quality, but do not cause mismatch at most time.

G4-miner inspects sequencing quality in whole genome, seizes unexpected fluctuations in local, and ascribes some of the quality fluctuations to the unstable formation of G4. Median quality scores of bases mapping to each locus in the genome are calculated after reads mapping. For each site, we seize the quality scores of the next 75 nt, recorded the number of loci whose median quality score significantly dropped in comparing with that of the specific site. G4 sites are mined after setting the thresholds of the number of loci and the quality score dropped.

Figure 1 A schematic of the G4-miner method. In a standard sequencing experiment, DNA fragments in clusters are sequenced and the phred quality scores are calculated for each read (cluster). The unstable formation of G4 causes unexpected fluctuations of sequencing quality, but do not cause mismatch. Reads mapping is performed to calculate the median quality score of bases mapping to each locus in the genome. For each site in the genome, the number of loci in the next 75 nt whose median quality score

significantly dropped in comparing with that of the specific site is calculated. G4 sequences in the whole genome can be mined after setting the threshold of the number of loci and the quality score dropped.

Validation by known sequences

We validated our approach by inspecting the sequencing quality scores of four known control sequences: two containing stable G4 structures (src and myc), one conforming requirement of canonical G4 but has been demonstrated to prevent G4 formation (a-repeat)²⁸ and a guanine-rich strand that cannot form a G4 (g-rich)²⁹ (**Supplementary Table 1**). Phred quality scores³⁰ were used to quantify the sequencing quality. We calculated the median quality scores for each locus in the template/genome to seize and represent the inconstant sequencing quality fluctuation. Although the readouts do not alter in all control sequences, the median quality score was slightly fluctuated down from or before the beginning of the G4 structure in positive controls (**Fig. 2a, b**). For src and myc, median quality score dropped no less than 4 at 55 loci and 25 loci among the 75 bases (half of the read length) from beginning of G4 structure. In contrast, sequencing of the negative controls observed less than 6 loci (**Fig. 2c, d**).

Figure 2 Sequencing quality of four known controls. **(a)** src; **(b)** myc; **(c)** a-repeat; **(d)** g-rich. The yellow shadows present the regions of the sequences.

The inspection of quality scores and readouts along positive control revealed that sequencing accuracy slightly reduced after the G4 start sites but the readouts did not alter (**Fig. 1**). At the same time, the reduction of quality scores was not observed in all reads which covered G4 structure and the sequence downstream. Some of the reads had constant high quality scores. The observations suggested that using standard Illumina sequencing buffers, the formation of G4s is not stable and the polymerase stalling is not occurred in most template molecules. Sequencing quality score of a read was generated by comprehensively evaluating the accuracy of synthesis reactions in tens of thousands of molecules in a cluster. Under G4-destabilizing condition (standard Illumina sequencing buffers), the quadruplex structures were unstably formed in only a small portion of template molecules in a cluster. The pause of polymerase stalling occurred in these molecules cause phase shift signals, which interfered the real signals and induced the drop of quality score.

Whole genome seizing the quality fluctuation

Next, we explored whether quality fluctuation could be specifically seized genome-wide. A set of 75nt (half of the read length and longer than 99% PG4s, **Supplementary Fig. 1**) segments which contain about 178,606 predicted G-quadruplexes (PG4s) and downstream sequences at 3' end were cut from the plus strand of the human genome. The rest regions of plus strand which do not contain PG4s were cut into 75nt segments. The analysis based on about one million reads (about 45× of the plus strand) generated by standard pipeline showed more significant quality fluctuations in the segments containing PG4s than the rest ones. Over 40% PG4 positive segments contained over 15 loci whose quality scores decrease no less than 4 in comparing with the loci before PG4s, and only 4.05% non-PG4 segments passed the criteria (**Supplementary Table 2**). Considering that the guanine content is greater than 28% in over 99% PG4

positive segments (**Supplementary Fig. 2**), we filtered the non-PG4 segments with low guanine content ($G\% < 28\%$), and less than 1% of the non-PG4 segments remained under the same criteria (**Supplementary Table 3**). Although the sequence quality for most non-PG4s were stable with no observable decrease, the remained 1% non-PG4 segments were found to have significant quality fluctuation (~370,000 segments). The number of these non-PG4 segments showing the considerable decreasing amplitude in considerable number of loci, was much greater than the number of predicted PG4s (178,606), which is consistent with the previous approach in G4-stablizing conditions²⁴, indicating that the human genome contains more G4s than predicted¹¹.

We applied G4-miner to profile DNA G4s structures in the human genome by resequencing primary B lymphocytes (GM12878, Online Methods), using the Illumina Hiseq platform under standard condition. To identify reliable G4 sequences, we inspected the quality scores of 75 bases at 3' end of a specific site and screened median quality scores of each site in both strands of the human genome. We set individual threshold for the two strands of the parallel runs to ensure the false positive rate similar and smaller than 1% (**Supplementary Table 4**). The influence evaluation of sequencing depth revealed that PG4 detection ratio got plateau when sequencing depth was greater than 20 folds for one stand (**Supplementary Fig. 3**). Therefore, two duplicates of the experiment were performed, at least 1.88 billion reads were generated with an average coverage of 45× for each strand of human genome (**Supplementary Table 5**). We set the threshold of 15 loci quality score decreased 4 (15, 4) for plus stand of experiment 1, (15, 3) for the minus stand of experiment 1, (12, 4) for the minus stand of experiment 2 and (11,4) for the minus stand of experiment 2, respectively. Any segment of the genome passed these thresholds is termed mined G4 sequence (MG4). Within both strands of the human genome, we identified 1,054,941 MG4s and 936,545 MG4s in the two parallel experiments (**Supplementary Table 6**). About 52% and 49% of all 356,298 predicted canonical quadruplexes were also detected by G4-miner and present in MG4s (**Supplementary Table 6**). Moreover, 736,689 of the total number of MG4s (70% for experiment 1 and 78% for experiment 2) were detected in both experiments (**Fig. 3a, b**). The high overlap between parallel experiments indicated that G4-miner is stable in quadruplex assigning.

Figure 3 Overlap between MG4s and PG4s in two parallel runs and different conditions. **(a)** Overlap between PG4s which were detected in MG4s in two parallel runs. **(b)** Overlap between MG4s in two parallel runs. **(c)** Overlap between PG4s which were detected in MG4s in standard sequencing buffers and K^+ or Na^+ + PDS conditions. **(d)** Overlap between MG4s in standard sequencing buffers and K^+ or Na^+ + PDS conditions.

Comparing to the result of the G4-seq, which combined polymerase stop assays with next-generation sequencing²⁴, over 75% and 82% of PG4s detected by G4-miner are also detected by adding K^+ and PDS to sequencing buffers, separately (**Fig. 3c, d**). More PG4s are detected under G4-stabilizing conditions (K^+ or PDS) than under standard sequencing condition, which confirmed to the existing knowledge that G4 structure are thermodynamically stable in the presence of K^+ and PDS.

Structural analysis and genome browser of mined G4 sequences

We noticed that only 17% of the MG4s were predicted as classical described G4s. In recent, noncanonical G4s structures such as long loops, bulges and two quartets have been identified by structural biological and biophysical studies³¹⁻³³. Therefore, we hierarchically assigned the MG4 to five categories to elucidate the structural features (**Fig. 4a**). Canonical G4s, long loops, bulges and two quartets accounted for 17%, 6%, 23% and 34% of total MG4s based on the threshold of sequencing quality. The distribution of MG4 categories is highly consistent with the result of the rG4-seq²⁵ which exploited reverse transcriptase stalling reactions containing K⁺ with G4-stabilizing ligand pyridostatin (PDS) (**Supplementary Fig. 4**). Canonical G4 sequences with short loops were more easily identified than sequences with longer loops (**Fig. 4b**), which is keeping with their relative thermodynamic stability^{32,33}. The MG4s in 'other' category do not contain known G4 structures, but also cause slight and inconstant drop of sequencing quality. The sequences analysis revealed that most of them contained other DNA secondary structures, such as hairpin structure, poly-adenine/thymine and i-motif (**Supplementary Table 7**). To ensure fidelity, this category was removed for the following analyses. Our results revealed that the G4-miner is able to mine and identify G4 sequences in whole genome from ordinary sequencing data.

Figure 4 Structural analysis and genome browser of MG4s. **(a)** Prevalence of MG4s by different G4 structural families. **(b)** Detection ratio of canonical G4 sequences categorized by loop length. **(c)** Genome browser view of MG4 and PG4 across two oncogenes, MEGF6 and SRC. The median sequencing quality score dropped in each loci is tracked along the whole gene sequence. The regions above the threshold are shown as red bars (MG4s). The PG4s are shown as black bars.

We quantified the OQs in exons, introns, promoters, untranslated regions (UTRs) and splicing junctions (**Table 1**). MG4s was found to be densest in 5' UTRs and splicing junctions, where G4s functioned in post-transcriptional regulation^{34,35}. The density of G4s in genomic regions is consistent with the result by adding K⁺ or PDS to sequencing buffers²⁴. Visual inspection of genes rich in PG4s revealed that G4-miner is effective in mining G4 regions (**Fig. 4c**).

Table 1 Distribution of MG4s in different genomic regions.

Region	# regions	Total region size ¹	# MG4s	MG4s density ²	# Mined PG4s	Mined PG4s density ²	# canonical MG4s	canonical MG4s density ²
3'UTR	46,297	38.3	6,171	0.16	1,565	0.04	2,308	0.06
5'UTR	66,404	15.6	10,361	0.66	1,776	0.11	5,143	0.33
Exon	243,071	81.6	25,349	0.31	3,664	0.04	11,513	0.14
Intron	211,307	1494.1	169,429	0.11	44,966	0.03	37,005	0.02
CDS	111,622	34.6	13,592	0.39	606	0.02	6,397	0.18
Promoter 1000 up	35,634	71.3	31,976	0.45	17,800	0.25	13,005	0.18
TSS 1000 up down	35,634	35.6	15,994	0.45	7,670	0.22	6,544	0.18
Splice 50	385,429	38.5	19,748	0.51	4,607	0.12	10,181	0.26

¹ The total length of the specific kind of region (million base pairs).

² The density of the specific sequences in that region (# per kilo base pairs).

Evaluation of the influence of SNVs in G4 formation

Single nucleotide variations (SNVs) are the differences at specific nucleotide locations in the genome, and are one of the most important genetic variations among individuals. The alteration of single nucleotides might have different consequences at the phenotypic level ³⁶. Some of them might have different consequences in G4s, including controlling the formation, altering the structure and influencing the stability. Therefore, we inspected SNVs in the whole human genome of GM12878, both homozygous SNVs and heterozygous SNVs. For homozygous SNVs, we predicted G-quadruplexes again using the modified genome which altered nucleotides in specific locations. Due to the alternation of homozygous SNVs, 1,502 new PG4s were predicted in this time, where 48% of them (728 PG4s) were identified as MG4s, and 1,077 existing PG4s were not predicted, where 60% of them (643 PG4s) were not recognized as MG4s (experiment 1, **Fig. 5a, b**).

Figure 5 Influence of SNVs in G4 formation. **(a)** The alteration of a homozygous SNV (rs11264880) causes the formation of G4 structure and was mined by G4-miner. **(b)** The alteration of a homozygous SNV (rs10915961) prevents the formation of G4 structure and was not detected. **(c)** Two alleles of the heterozygous SNV (rs369859441, G/A) have different influence on the formation of G4 structure and was demonstrated by G4-miner (formed: allele G; not formed: allele A).

For heterozygous SNVs, we classified the sequencing reads into two groups based allele type to identify MG4s separately. Of all heterozygous SNVs, 7.17% (186,904) of them revealed to having different influence on the formation of G-quadruplexes in plus stand, and 6.28% (163,666) for the minus stand (**Fig. 5c, Table 2**). All those homozygous and heterozygous SNVs associated G4s are individual specific G4s which might contain individual differences in chromatin architecture and many fundamental biology processes, such as DNA replication and gene regulation. Our results show that G4-miner can characterize G4s for specific individuals.

Table 2 The influence of heterozygous SNVs in MG4s detection.

Strand	# Total heterozygous SNVs	# SNV (++) ¹	# SNV (+-) and SNV (-+) ²	# SNV (--) ³	# MG4 (+-) and MG4 (-+) ⁴
Plus	2,608,087	123,529	186,904	2,297,654	153,533
Minus		109,410	163,666	2,335,191	134,532

¹ The number of heterozygous SNVs of which the MG4s were identified in both two groups of alleles.

² The number of heterozygous SNVs of which the MG4s were identified only in the group of allele 1 or group of allele 2.

³ The number of heterozygous SNVs of which the MG4s were not identified in both two group of alleles.

⁴ The number of heterozygous MG4s which were identified only in the group of allele 1 or group of allele 2.

Discussion

Many techniques have been developed to detect and map G4s including chemical biology and genomic technologies. However, to date most of these techniques' protocols are complicated by relying on G4-stabilizing conditions or G4 antibodies. To address this conundrum, we have established a convenient, high-throughput approach for identifying DNA G4 structures from ordinary next-generation sequencing data directly. Our results confirm that G4 structures are unstably formed in the standard sequencing buffer, which cause sequencing quality fluctuations. We have conveniently identified canonical G4s and numerous non-canonical structures in whole genome by inspecting the quality fluctuations.

For one thing, to discover the stability of the G4-miner, we carried out two parallel experiments, and 736,689 MG4s (70% for experiment 1 and 78% for experiment 2) were detected in both experiments. The high overlap between them indicated that G4-miner is stable in quadruplex assigning. For another, to verify the validity of the method, we compared G4-miner with the G4-seq. As a result, over 75% and 82% of PG4s also detected by G4-miner, and it confirmed that G4s are stable in the presence of K^+ and PDS. The different overlaps between G4-stabilizing conditions and G4-miner indicated that G4-miner is valid in detecting G4s. Besides, the quality fluctuations seized by G4-miner are not only G4s, but also many unidentified regions. The principle of G4-miner can be used to identify other structures that would influence sequencing quality, and describes genomic maps of other secondary structures by their sequence characterization through quality fluctuations.

One key role of our method is to reveal the effect of SNVs on G4 formation. After inspecting SNVs in the whole human genome, we observed that the presence of some SNVs determines the formation of G4 structure. The homozygous SNV causes the formation of G4 structure or prevents the formation of G4 structure. For some heterozygous SNVs, the two alleles have different influence on the formation of G4 structure, as the allele G caused the quality fluctuations, the allele A did not. The SNVs shows significant individual differences, the specific relation between SNVs and G4s might contain individual differences in chromatin architecture and fundamental biology processes. The G4-miner can be used to detect individualized G4 structures, which will accelerate the discovery of the relation among SNVs, G4s and biology processes.

In summary, our work provides G4-miner, a friendly and convenient method to identify G4 structure genome-wide. Our finding that some quality fluctuations ascribed to the unstable formation of G4, describes a map of G4s in the whole human genome under standard sequencing protocol. Importantly, through G4-miner, we have also confirmed that some of the SNVs can influence the formation of G4 structures. The given approach is available to identify and characterize G4s genome-wide for specific individuals. This approach could also be extended to profile other DNA secondary structures which exist in standard sequencing buffer.

Methods

PG4 identification and region classification

PG4s were predicted from hg19 using g4predict (<https://github.com/mparker2/g4predict.git>) according to the Quadparser algorithm by searching for the '(G{3,}[ATCG]{1,7}){3}G{3,}' pattern. For parameter determining step, the start site of 'positive region' is defined as the sequence up to 12 bases (approximate footprint of DNA polymerase) up-stream of the PG4 start site, and the end site was 75 bases downstream of the start site. After removing the positive regions and 300 bases upstream and downstream of them, the remaining region was split into 75-base windows, which were defined as 'negative regions'.

Low-quality region scanner

For each sequencing read, the aligned sequence information was extracted from BAM file. Reads were separated into forward strands and reverse strands according to the FLAG value. After the quality score was translated into decimal numbers, the median quality of every locus of different strands was calculated. The low-quality region scanner algorithm checked the median quality value of every locus sequentially. The parameter 'M' is designed to describe the degree of low quality, and parameter 'N' limits the number of low-quality bases to set the lower bound of the low-quality region. If the difference of median quality scores between one base and the just base up-stream of it (the orientation of 'up' or 'down' is consistent with the orientation of the strand) is no less than M, the base with higher quality value is 'the flag site' whose quality score is defined as 'the flag value', and the other is 'low-quality site', also 'the start site of low-quality region'. From the start site on, if the number of low-quality sites is greater than or equal to N in a range of 75 bp, a 75nt-low-quality region was marked as a 'hit'.

Parameter determining and MG4 detection

Positive regions (with PG4s) and negative regions (without PG4s) from 'PG4 identification and region classification' step were used to calculate the hit ratio according to the previous low-quality region scanning algorithm. The parameters M and N were set as 1~12 and 1~20 respectively. Under each M-N set, a positive rate and a false positive rate were calculated. For both strands of two replicates, parameter sets which can obtain the max positive rate while keeping the false positive rate lower than 1% would be used in the following steps. To eliminate the influence of other non-G-quadruplex secondary structures, the ratio of G/C bases in detection regions should be at least 28% (99% of positive regions contain 28% or more G/C bases). For the whole genomic analysis, low-quality regions from two strands of two replicates were scanned independently under their most suitable parameter sets. For the more complete detection of positive regions, the low-quality regions in MG4 detection step extended 35 bases (the longest length of canonical G4s) up-stream of the start site. Low-quality regions with overlapped fragments were merged into a single one. The merged regions considered as MG4s.

Structure analysis of MG4 categories

MG4s were stratified into different categories according to the patterns of G-quadruplexes. To avoid calculating the same MG4s multiple times, we followed the order which based on the predicted stability

of different patterns: canonical > long loops > bulges > others. These different categories are defined as follows: canonical, $(G\{3\}, N\{1,7\})\{3\}G\{3\}$, with $N=[ATCG]$; long loops, $G\{3\}, N\{8,12\}(G\{3\}, N\{1,7\})\{2\}G\{3\}$ or $G\{3\}, N\{1,7\}G\{3\}, N\{8,21\}G\{3\}, N\{1,7\}G\{3\}$; bulges, G-quadruplex with any G-tract being $GH\{1,7\}GG$ or multi G-tracts being $GH\{1,2\}GG$, with $H=[ATC]$; others, not in any previous category. MG4s with more than one pattern would only be categorized as the more stable one. The 'others' category was further stratified into subcategories containing non-G-quadruplex structures (for example, i-motifs, hairpins, repeats, poly-A/T). Some of these categories are defined as follows: i-motif, $(C\{3\}, N\{1,7\})\{3\}C\{3\}$; hairpins, stem-loop structures with stems no less than 7 bp and loops no more than 30 bp; poly-A/T, continuous 10 or more A/T bases.

Genomic regions analysis

Gene annotation files were downloaded from the UCSC genome browser website (<https://genome.ucsc.edu/>), genome version hg19. Genomic regions contain 5'UTRs, 3'UTRs, exons, introns, CDSs, promoter regions, TSS (translational start site) regions and splice regions. TSS regions are defined as the 1,000 bp up- and down-stream the TSS. Promoters are 1,000 bases up-stream TSS. Splice regions are 50 bases up- and down-stream splice sites. For these regions, the total number and size of them were calculated. The number of MG4s, including total MG4s and canonical MG4s, and predicted PG4s overlapping to the region intervals were calculated. The number of MG4s were all the intersection of two replications. For each gene annotated in the human genome (hg19), the number of predicted PG4s of MG4s (PG4inMG4s) in two replications were counted. The density of G4s was calculated by dividing the respective counts by the length of genes or regions and multiplying by 1,000.

Declarations

Acknowledgments

This work was supported by the National Key Project of China (2016YFA0501604), Project (61972084, 61971125) of National Natural Science Foundation of China, and Six Talent Peaks Project of Jiangsu Province (2019-SWYY-004).

Author contributions

M.D and W.L implemented and performed the analysis. J.T., X.S. and Z.L. designed the experiments and analysis. J.T., M.D. and N.L. interpreted the results and co-wrote the manuscript.

Additional information

Accession codes: The data reported in this paper are deposited in the Gene Expression Omnibus and are available under accession number GSE159307.

Supplementary Information accompanies this paper at online version.

Competing financial interests: The authors declare no competing financial interests.

References

1. Davis, J.T. G-quartets 40 years later: From 5'-GMP to molecular biology and supramolecular chemistry. *Angew Chem Int Edit* **43**, 668-698 (2004).
2. Wan, Y., Kertesz, M., Spitale, R.C., Segal, E. & Chang, H.Y. Understanding the transcriptome through RNA structure. *Nat Rev Genet* **12**, 641-655 (2011).
3. Kwok, C.K., Tang, Y., Assmann, S.M. & Bevilacqua, P.C. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem Sci* **40**, 221-232 (2015).
4. Gellert, M., Lipsett, M.N. & Davies, D.R. Helix Formation by Guanylic Acid. *P Natl Acad Sci USA* **48**, 2013-& (1962).
5. Sen, D. & Gilbert, W. Formation of Parallel 4-Stranded Complexes by Guanine-Rich Motifs in DNA and Its Implications for Meiosis. *Nature* **334**, 364-366 (1988).
6. Sundquist, W.I. & Klug, A. Telomeric DNA Dimerizes by Formation of Guanine Tetrads Between Hairpin Loops. *Nature* **342**, 825-829 (1989).
7. Sen, D. & Gilbert, W. A Sodium-Potassium Switch in the Formation of 4-Stranded G4-DNA. *Nature* **344**, 410-414 (1990).
8. Wolfe, A.L. et al. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**, 65-70 (2014).
9. Haeusler, A.R. et al. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**, 195-200 (2014).
10. Millevoi, S., Moine, H. & Vagner, S. G-quadruplexes in RNA biology. *Wires Rna* **3**, 495-507 (2012).
11. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**, 2908-2916 (2005).
12. Todd, A.K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* **33**, 2901-2907 (2005).
13. Varizhuk, A. et al. The expanding repertoire of G4 DNA structures. *Biochimie* **135**, 54-62 (2017).
14. Stegle, O., Payet, L., Mergny, J.L., MacKay, D.J.C. & Huppert, J.L. Predicting and understanding the stability of G-quadruplexes. *Bioinformatics* **25**, 1374-1382 (2009).
15. Bedrat, A., Lacroix, L. & Mergny, J.L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* **44**, 1746-1759 (2016).
16. Huppert, J.L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome (vol 35, pg 406, 2006). *Nucleic Acids Res* **35**, 2105-2105 (2007).
17. Eddy, J. & Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res* **34**, 3887-3896 (2006).

18. Hansel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267-1272 (2016).
19. Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* **13**, 551-564 (2018).
20. Meyer, C.A. & Liu, X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**, 709-721 (2014).
21. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* **21**, 459-474 (2020).
22. Woodford, K.J., Howell, R.M. & Usdin, K. A novel K(+)-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes. *J Biol Chem* **269**, 27029-27035 (1994).
23. Han, H., Hurley, L.H. & Salazar, M. A DNA polymerase stop assay for G-quadruplex-interactive compounds. *Nucleic Acids Res* **27**, 537-542 (1999).
24. Chambers, V.S. et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877-881 (2015).
25. Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S. & Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods* **13**, 841-844 (2016).
26. Marsico, G. et al. Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res* **47**, 3862-3874 (2019).
27. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
28. Chen, J.L., Guo, Y.H., Zhou, J. & Ju, H.X. The Effect of Adenine Repeats on G-quadruplex/hemin Peroxidase Mimicking DNAzyme Activity. *Chem-Eur J* **23**, 4210-4215 (2017).
29. Smirnov, I. & Shafer, R.H. Effect of loop sequence and size on DNA aptamer stability. *Biochemistry-Us* **39**, 1462-1468 (2000).
30. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).
31. Palumbo, S.L., Ebbinghaus, S.W. & Hurley, L.H. Formation of a Unique End-to-End Stacked Pair of G-Quadruplexes in the hTERT Core Promoter with Implications for Inhibition of Telomerase by G-Quadruplex-Interactive Ligands. *J Am Chem Soc* **131**, 10878-10891 (2009).
32. Guedin, A., Gros, J., Alberti, P. & Mergny, J.L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res* **38**, 7858-7868 (2010).
33. Mukundan, V.T. & Phan, A.T. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *J Am Chem Soc* **135**, 5017-5028 (2013).
34. Bochman, M.L., Paeschke, K. & Zakian, V.A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780 (2012).

35. Cruz, J.A. & Westhof, E. The Dynamic Landscapes of RNA Architecture. *Cell* **136**, 604-609 (2009).
36. Syvanen, A.C. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**, 930-942 (2001).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures.docx](#)
- [SupplementaryTables.docx](#)