Published in final edited form as:

Methods. 2012 May ; 57(1): 3–10. doi:10.1016/j.ymeth.2012.05.002.

Computational Detection and Analysis of Sequences with Duplex-Derived Interstrand G-quadruplex Forming Potential

Kajia Cao¹, Paul Ryvkin², and F. Brad Johnson³

¹Department of Cell and Molecular Biology, University of Pennsylvania School of Medicine. Philadelphia, PA. USA

²Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine. Philadelphia, PA. USA

³Department of Pathology and Laboratory Medicine. University of Pennsylvania School of Medicine. Philadelphia, PA. USA

Abstract

Bioinformatic approaches to the identification of genomic sequences having G-quadruplex forming potential (QFP) has enabled important tests of the structure of these sequences *in vitro* and of their behavior under conditions where the formation or function of G-quadruplexes is modulated *in vivo*. Several similar approaches to identifying intramolecular QFP (i.e. forming among G-runs on one strand of DNA) have been developed previously, but none appears to perfectly predict G-quadruplex formation. Here we describe a new approach, which complements and differs from prior approaches in that it identifies motifs containing G-runs on both strands of duplex DNA that could contribute to G-quadruplex structures. We call these motifs duplex-derived interstrand QFP (ddiQFP), and illustrate their potential applications by describing their genomic distribution and an example of their correspondence to loci targeted by a G-quadruplex-unwinding DNA helicase in yeast.

Keywords

G-quadruplex; G4 DNA; interstrand G-quadruplex; ddiQFP; Bioinformatics

Introduction

Interest in the biological functions of G-quadruplex forms of nucleic acids, G4 DNA and G4 RNA, has blossomed over the past five years as strong support for their impact on transcription, translation, replication, recombination, chromatin structure and telomere maintenance has emerged (1-3). Bioinformatic approaches have played an important role by identifying within genomes candidate sequences with G-quadruplex forming potential (QFP). Once identified, the genomic distribution of QFP loci, including their relationships to other genomic features, can be assessed, and the structures adopted by examples of these sequences can be determined *in vitro*. Moreover, whether they might actually form G4

© 2012 Elsevier Inc. All rights reserved.

Corresponding author: F. Brad Johnson, 405A Stellar Chance Labs, 422 Curie Boulevard, Philadelphia PA, 19104-6100, johnsonb@mail.med.upenn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

structures *in vivo* can be addressed by testing for statistically significant associations between sets of QFP loci and those bound physically or impacted functionally by agents or genetic manipulations that target G4 structures. For example, if the set of all QFP loci are immunoprecipitated by a G4-specific antibody from chromatin more efficiently than non-QFP sequences, or if QFP loci preferentially cause replication forks to stall when a G4 DNA-unwinding helicase is inactivated, then it may be inferred that the loci form G4 DNA to some degree (4, 5). Although indirect, this sort of evidence for *in vivo* G4 structures has the strength that the test is applied in a fashion repeated across the genome, yielding statistically robust overall associations between sets of loci.

Initial bioinformatic approaches to the identification of QFP sequences have focused on those that might form intramolecular (i.e. intrastrand) G4 structures (6-11). In simplified terms, intramolecular QFP sequences have been defined as those possessing on one strand four runs each containing three or more consecutive guanines (Gs), with the G-runs separated by short loops consisting of one-to-seven nucleotide(s) of any base composition. This definition is based reasonably on kinetic and thermodynamic considerations, i.e. that G-runs tethered to one another within a strand should more rapidly assemble G4 structures, and with lower entropic cost, than G-runs on separate strands. Essentially similar variants of this definition have used runs of only two consecutive Gs, or have allowed more flexibility in loop lengths by requiring only that the overall length of the QFP motif fall within a certain length (e.g. 35 nucleotides).

Bioinformatic methods for applying the intramolecular QFP definitions described above have been reviewed elsewhere (12, 13), which we will not repeat here. These methods have been quite useful in that they have successfully predicted sequences that have proved to form G4 structures *in vitro* and that overall have shown strong statistical associations genome-wide with biological manipulations predicted to impact G4 formation or function. However, by the same criteria, these QFP definitions are neither fully specific nor sensitive. Examples of imperfect specificity include the inability of some QFP sequences to form stable G4 DNA *in vitro*, and the fact that most QFP loci are not associated *in vivo* with replication defects or altered gene expression when G4-targeting helicases are inactivated or G4 ligands are introduced (4, 14, 15). Similarly, examples of imperfect sensitivity include the c-kit1 G-quadruplex, which falls outside the definition but which nonetheless forms stable G4 DNA (16), and that most of the loci affected in cells lacking G4 DNA-unwinding helicases or exposed to G4 ligands do not have intramolecular QFP. Whereas likely explanations for lack of full apparent specificity and sensitivity *in vivo* include that the G4-targeting helicases and ligands 1) also affect other types of nucleic structures and 2) target only a subset of actual G4 structures, it is also clear that current definitions for intramolecular QFP do not fully capture the capacity of nucleic acids to form G4 structures, either *in vitro* or *in vivo*. For example *in vitro* studies have demonstrated recently that stable G-quadruplexes may form from a sequence with an intramolecular QFP motif and one very long loop (up to 30 nt in length, and perhaps longer) provided one or both of the other loops are very short (17). Further, conditions that mimic the crowded molecular and dehydrated conditions within nuclei promote G4 formation (18), and these – along with numerous other features of the *in vivo* environment including chromatin proteins, the duplex nature of most genomic DNA sequences, and superhelical stresses – very likely change the rules for G4 structure formation in ways not yet well appreciated. Addressing all of these issues is far beyond the scope of this methods paper. Here we focus on novel methods that address the potential for G4 DNA to assemble from G-runs contained within the two strands of duplex DNA, which we name “duplex-derived interstrand QFP (ddiQFP)”. We suggest that considering sequences containing ddiQFP motifs may improve the ability of bioinformatic approaches to predict sequences that may be impacted by G-quadruplex formation and function.

Materials and Methods

A sliding window approach was used to identify ddiQFP sequences, based on our previous approach (8). In the method, we first defined two parameters, the length of the guanine runs (G-runs) (or cytosine runs (C-runs) on the other strand) $k = 3$ nt and the window size $w = (30, 50, 100, \text{ or } 200 \text{ nt})$. For each window, the sequence takes the form of one of 14 patterns (see Table 1, below) where $a, b, c = 0$ and the length of the sequence is less than or equal to w . Note that in the case of fully compact ddiQFP motifs, a, b , and c are equal to zero. The following Perl code shows an example of how ddiQFP motifs can be detected using a 50 nt sliding window.

```
# searches for 4 sets of 3-6 consecutive G's or C's with loops up to 50nt long
# such that the resulting motif is <50nt long
if ($nongreedy == 1) {
    $regex = "(G{3,6}|C{3,6})({1,50}?(G{3,6}|C{3,6})({1,50}?(G{3,6}|C{3,6})({1,50}?(G{3,6}|C{3,6})))";
} else {
    $regex = "(G{3,6}?|C{3,6}?){1,50}(G{3,6}?|C{3,6}?){1,50}(G{3,6}?|C{3,6}?){1,50}(G{3,6}?|C{3,6}?){1,50}";
}
while ($seq =~ /$regex/g) {
    $start = $-[0]+1;
    $motif = substr($seq, $-[0], $+[0] - $-[0]);
    $len = length($motif);
    # enforce window size
    if ($len <= 50) {
        print "$start\t$motif"
    }
    # restart search after the 1st run
    $nextmotifsearchstart = $+[2];
    pos($seq) = $nextmotifsearchstart;
}
```

Within each sliding window, we used non-greedy and greedy string matching methods to identify ddiQFP sequences. The greedy matching identifies the ddiQFP motifs that have more than six continuous Gs or Cs as one G-run or C-run in the motif, while the non-greedy matching method partitions $G_{>6}$ and/or $C_{>6}$ into two or more G-runs or C-runs as one G/C is treated as N, e.g. $GGGGGGGG = G_3GG_4$. We note that together the two methods identified intramolecular as well as ddiQFP motifs and also sometimes counted the same ddiQFP motif twice. We eliminated intramolecular QFP and duplicated ddiQFP motifs using the following three steps:

1. If the number of G_{3+} runs is four and the number of C_{3+} runs is zero, or vice versa, remove the motif from the list.
2. For any duplicated motifs (same chromosome, start site and length) in the output, remove the duplicate.
3. For each motif in the list, if another motif starts at same site but has a longer length; remove the shorter motif from the list.

To evaluate the significance of association between ddiQFP motifs and other features, the observed overlap was compared to the overlap found when sequence blocks of the same number and length of the real ddiQFP sites were randomly placed within *S. cerevisiae*

genomic and mitochondrial sequences 1000 times. We used BEDTools to randomly sample the genomic regions (19). ddiQFP motifs were identified using Perl and downstream analyses were conducted using R (<http://www.r-project.org/>). The program required only a few seconds to collect the ddiQFP motifs for a defined window size using the Oct. 2003 *S.cerevisiae* genome assembly (SacCer1, <http://genome.ucsc.edu>). The code and instructions are available upon request.

To determine the position-specific density of ddiQFP motifs in ORFs and their flanking regions, the location of the motifs (defined using a 35 nt window) was determined in regions 1 kb upstream and downstream from the start and end, respectively, of every ORF. A sliding window of 150 nt was moved in single nt increments from the upstream and downstream boundaries of the regions towards the start of ORFs, and the number of QFP motifs in each window was counted.

For Gene ontology enrichment analysis terms were computed from a gene list using SGD Gene Ontology Slim Mapper (<http://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl>). The enrichment of GO terms was calculated based on the proportion of genes in each GO category divided by the proportions for all yeast ORF loci. We applied Fisher's exact test to obtain one-tailed *p*-value for each enriched GO term.

Theory/Calculations

Intramolecular G4 DNA is thought to form most readily within DNA strands that lack or have first become separated from a complementary strand (e.g. at the 3' telomere overhang or within the parental duplex strands that have become separated during DNA replication). Nonetheless there is evidence that it may also coexist in direct equilibrium with duplex DNA, which attests to the high stability of certain G4 DNA structures (20). In the latter case, the energetic tradeoff between duplex and G4 DNA forms can be roughly approximated by the noting that 14 base-paired "segments" (each corresponding to a G-run or loop) in the duplex are traded for four such segments in the G-quadruplex (Figure 1A), a net loss of ten segments. For simplicity, this calculation ignores many potentially important factors that could impact the equilibrium for any given case, including contributions that might be made by DNA structures formed by the C-rich strand (i.e. i-motifs), differences between the free energy per bp of the G-quadruplex vs. duplex, and by proteins bound to the different structures.

Within this framework, what would be the energetic tradeoff of forming G4 DNA comprising G-runs on each of the strands in the duplex (ddiQFP)? First, the entropic and kinetic barriers to G-quadruplex formation from ddiQFP sequences should be similar to those for intramolecular QFP sequences, because the two strands are bound to one another and nearby one another in both the duplex and G4 forms. Second, for a QFP motif in which the G-runs are allowed to exist on either strand (Figures 1B and 2) and loop sequences are interposed amongst the G-runs, the energetic "cost" is similar to intrastrand QFP: four segments are paired and 14 paired segments are lost, yielding the same net loss of ten segments. Third, and moreover, for interstrand G4 DNA, nucleotides between adjacent G and C-runs are unnecessary (because the C-runs act as loops), and thus for such "compact" motifs as few as four paired segments are lost (Figure 1C), raising the possibility that sequences with ddiQFP motifs might form particularly stable G-quadruplexes. Of course it is possible that the G and C-runs comprising ddiQFP motifs could form intrastrand hairpins, but these should be energetically similar to the duplex form, and thus should not preclude the possibility of interstrand G4 DNA any more than does the duplex form in the first place.

The ddiQFP motifs fall into 14 patterns when examining a single strand of genomic sequence (Table 1). However, as they are derived from a DNA duplex, some of these motifs

are equivalent because they are the reverse-complements of one another, i.e. a ddiQFP motif identified on one genomic strand is identical to its reverse complement identified on the other strand. For example, Class I comprises the motifs $C_3+N_aG_3+N_bG_3+N_cG_3+$ and $C_3+N_aC_3+N_bC_3+N_cG_3+$. The 14 motifs can thus be further grouped into nine classes, based on the 5'-to-3' order of transitions between the occurrence of the G-runs on each strand. These are described, along with the sequences composing their loops, in Table 1.

Results and Discussion

Using the approaches described above, we examined the occurrence of ddiQFP sequences within *S. cerevisiae* DNA sequences, including genomic and mitochondrial DNA.

Identification of ddiQFP motifs in *S. cerevisiae*

We used a sliding window approach to assess the frequency of ddiQFP motifs in *S. cerevisiae*. We have mentioned that the method depends on two parameters, k (the length of the guanine runs; default is 3nt), w (window size; we tried four different sizes: 30nt, 50nt, 100nt and 200nt). Instead of running the sliding window in a defined number of nucleotides, we applied Hershman's algorithm (8) which examines every sliding window that starts from a G-run. The number of intramolecular QFP and ddiQFP motifs within the yeast genome (including mitochondrial DNA) is shown in Table 2. The number of ddiQFP motifs is ~2-3 fold larger than intramolecular QFP at any given window size, consistent with the greater flexibility and potentially more compact nature of the ddiQFP motifs.

Genomic distribution of ddiQFP sequences

The number of mitochondrial motifs is striking given that the mitochondrial genome is approximately 140-fold smaller than the nuclear genome and has a higher AT base content (83% vs. 62%). The overall density of ddiQFP (length = 35 nt) is 0.23% and 0.0041% for the mitochondrial and nuclear genomes, respectively. This 56-fold difference is even greater than the ten-fold difference in the density of intramolecular QFP in the yeast mitochondrial vs. nuclear genomes noted previously (21), which using our definitions is five-to-six fold for 35 – 50 nt windows.

In both the nuclear and mitochondrial genomes, the density of ddiQFP is approximately two-fold higher in upstream promoter regions (defined as 50 to 850 bp upstream from the start of the ORF) than in ORFs. We determined the overall distribution of sequences with ddiQFP with respect to the start of all yeast ORFs by calculating the number of loci having ddiQFP (= 35nt) within a 150nt sliding window at each nucleotide position upstream and downstream of the start of all nuclear genome ORFs (Figure 3A). Generally, ddiQFP sequence appears much more frequently in the upstream of ORF. A particularly high peak of ddiQFP runs from ~300 to 500nt upstream of ORFs, similar to what was observed previously for intramolecular QFP (8). A similar distribution of ddiQFP is observed with respect to mitochondrial ORFs, when examined separately (Figure 3B).

Functions of genes with ddiQFP

We evaluated the enrichment within different Gene Ontology (GO) categories of *S. cerevisiae* genes having ORF or promoter ddiQFP (window = 50 nt). Results for ORFs and promoters are shown in Tables 3A and 3B, respectively. Genes with ORF and/or promoter ddiQFP tend to be involved in mitochondrial function, whereas those with promoter ddiQFP are involved with ribosome or transcription factor function. The high ddiQFP density of the mitochondrial DNA does not alone explain the associations, because most of the relevant genes lie within the nuclear genome, perhaps indicating that ddiQFP motifs might be involved in co-regulation of nuclear and mitochondrial genome functions.

The frequencies of the classes of ddiQFP motifs

We next calculated the proportion of each ddiQFP class among all the ddiQFP motifs in *S. cerevisiae* (Figure 4A). At the largest window lengths, the proportions are those predicted by chance (~14% for each of the classes having two motifs that are the reverse complements of one another, and ~7% for those where the motif is its own reverse complement). However, at smaller window lengths, there is skewing in the frequencies, and in particular classes 3 and 6 are over-represented and class 4 is under-represented. Strikingly, when the analysis is applied to the nuclear and mitochondrial genomes separately (Figure 4B), a clear enrichment for class 6, and depletion of classes 1, 2 and 4, is observed in the mitochondrial genome. Over a third of mitochondrial QFP motifs are class 6, which appear related to the so-called mitochondrial GC-boxes or GC-clusters, which are of uncertain function but might be involved in transcriptional regulation (22).

The distribution of G-runs and C-runs in ddiQFP motifs

We next assessed the distributions of G-runs and C-runs among all yeast ddiQFP motifs. The number of G-runs and C-runs per strand can be either of two combinations as described in Table 1: two of each, or three of one and one of the other (the “three and one” types and the “one and three” types are equivalent based on reverse-complementary relationships). Assuming the chances of having each combination are equal, 8/14 (57.14%) of ddiQFP motifs are expected to have either three G-runs and one C-run or one G-run and three C-runs, and 6/14 (42.86%) of motifs to have two G-runs and two C-runs. We estimated the number of G-runs and C-runs in ddiQFP motifs in different window sizes. At the larger window sizes (where the motifs would be expected to decay to near-random behavior), the expected frequencies were observed, but at the 30 nt window, the two G- and C-run motifs were more frequent than expected (Figure 5). One possible reason is that the 458 motifs in the 30 nt window are not enough to avoid random skewing in the distribution, but it is also possible that the skewing reflects important biological behaviors of the sequences (since shorter ddiQFP motifs would be expected to form *bona fide* G-quadruplexes more readily than longer ones).

We also tested for the frequencies of compact ddiQFP motifs, and found only one fully compact motif (CCCGGGCCCGGG), which is located at Chr4: 1059054-1059065. However, we found more than 23% of the ddiQFP motifs in the 30nt window contain connected C-runs and G-runs that are in same length (for example, CCCGGG and GGGCCC). Not surprisingly, the proportion of these partially compact ddiQFP motifs decreases at larger windows, with fewer than 5% in the 200nt window. We also found only five and eight ddiQFP motifs containing “CCCGGGCCCG” in the 100nt and 200nt windows respectively, but none in 30nt and 50nt windows. We found two ddiQFP motifs containing “GGGCCCCGGG” sequences in 30nt and 50nt windows, three motifs in 100nt window, and four in 200nt window. Unexpectedly, among all partially compact forms at all window sizes, “GGGCCCC” has 4.35 times the frequency of “CCCGGG” in 30nt window and 2.4~2.9 times in 50nt ~ 200nt windows. Not surprisingly, forms “C₃₊G₃₊” and “C₃₊G₃₊C₃₊” are not found at telomeres (because they comprise TG₁₋₃ repeats), but they are also not present in mitochondrial sequences, indicating a significant difference between the nuclear and mitochondrial genomes.

Pif1-K264A preferentially targets ddiQFP sites

The Pif1 helicase unwinds G-quadruplexes efficiently (23); although yeast Pif1 has apparently been tested on only intramolecular G-quadruplex substrates, the human PIF1 helicase actively unwinds intermolecular G-quadruplexes (24), raising the possibility that the yeast protein also has this activity. Recently, Paeschke et al. used a helicase-inactive mutant version of Pif1, Pif1-K264A to assess Pif1 binding to chromatin, including its

targeting to QFP sites (4). They found significant associations between intramolecular QFP and Pif1-K264A target sites, using a QFP window of approximately 100 nt, and defining associations between QFP sites and Pif1-K264A target peaks as those that are within 500 bp of one another. We reexamined their data (using the same $q = 0.05$ cutoff for peak definition) and calculated the association between Pif1-K264A binding sites and intramolecular QFP and ddiQFPs. However, instead of defining associations as those within 500 bp, we required that QFP and Pif1 peaks actually intersect by at least one bp.

We tested for intersection of the 1148 peaks of Pif1-K264A targets with sets of intramolecular QFP and ddiQFP sites identified using sliding windows of 30nt, 50nt, 100nt and 200nt. The statistical significance of the intersections was assessed by comparing the observed frequencies to 1000 examples of those obtained when the real set of QFP sites was randomly redistributed within the *S. cerevisiae* genome. We found significant overlap for intramolecular QFP at the 100nt and 200nt windows (Table 4), consistent with the published findings. Interestingly, overlap with ddiQFP is highly significant at the 30, and 50 nt but not larger windows. The higher significance for ddiQFP than intramolecular QFP at the smaller windows can be in part attributed to the greater number of ddiQFP motifs at each window length, but might also reflect a stronger association between Pif1 targets and ddiQFP, because the relative percentages of Pif1 targets intersecting ddiQFP vs. intramolecular QFP are somewhat larger than the overall relative numbers of ddiQFP and intramolecular QFP motifs (compare Tables 2 and 4). Overall, these findings indicate that Pif1 targets ddiQFP motifs significantly.

Conclusions

Here we have described methods for defining and identifying ddiQFP motifs. We find that they are found in association with sites targeted by the Pif1 G-quadruplex-unwinding helicase in yeast, and to a degree of significance at least as great as intramolecular QFP motifs. We described some ways in which different classes of ddiQFP motifs are distributed in yeast genomic and mitochondrial sequences, and also demonstrated associations with upstream promoter regions and genes involved in mitochondrial functions. However, we did not explore most of the potential ways in which the classes may differ with respect to propensity to form G4 DNA or association with particular biological processes. More work is needed to understand the structural and biological significance of ddiQFP motifs, and we hope that the concepts we have presented will stimulate these efforts.

Acknowledgments

We thank Li-San Wang for advice and discussions. This work was supported by NIH grants R01 AG021521 and P01 AG031862.

References

1. Lipps HJ, Rhodes D. G-quadruplex structures: in vivo evidence and function. *Trends Cell Biol.* 2009; 19:414–22. [PubMed: 19589679]
2. Wu Y, Brosh RM Jr. G-quadruplex nucleic acids and human disease. *FEBS J.* 2010; 277:3470–88. [PubMed: 20670277]
3. Bugaut A, Balasubramanian S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Res.* 2012
4. Paeschke K, Capra JA, Zakian VA. DNA replication through G-quadruplex motifs is promoted by the *Saccharomyces cerevisiae* Pif1 DNA helicase. *Cell.* 2011; 145:678–91. [PubMed: 21620135]
5. Fernando H, Sewitz S, Darot J, Tavaré S, Huppert JL, Balasubramanian S. Genome-wide analysis of a G-quadruplex-specific single-chain antibody that regulates gene expression. *Nucleic Acids Res.* 2009; 37:6716–22. [PubMed: 19745055]

6. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* 2005; 33:2908–16. [PubMed: 15914667]
7. Yadav VK, Abraham JK, Mani P, Kulshrestha R, Chowdhury S. QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.* 2008; 36:D381–5. [PubMed: 17962308]
8. Hershman SG, Chen Q, Lee JY, Kozak ML, Yue P, Wang LS, Johnson FB. Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in *Saccharomyces cerevisiae*. *Nucl Acids Res.* 2008; 36:144–56. [PubMed: 17999996]
9. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research.* 2007; 35:406–13. [PubMed: 17169996]
10. Todd AK, Johnston M, Neidle S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* 2005; 33:2901–7. [PubMed: 15914666]
11. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* 2006; 34:3887–96. [PubMed: 16914419]
12. Ryvkin P, Hershman SG, Wang LS, Johnson FB. Computational approaches to the detection and analysis of sequences with intramolecular G-quadruplex forming potential. *Methods Mol Biol.* 2010; 608:39–50. [PubMed: 20012414]
13. Todd AK. Bioinformatics approaches to quadruplex sequence location. *Methods.* 2007; 43:246–51. [PubMed: 17967695]
14. Verma A, Halder K, Halder R, Yadav VK, Rawal P, Thakur RK, Mohd F, Sharma A, Chowdhury S. Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J Med Chem.* 2008; 51:5641–9. [PubMed: 18767830]
15. Johnson JE, Cao K, Ryvkin P, Wang LS, Johnson FB. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. *Nucleic Acids Res.* 2010; 38:1114–22. [PubMed: 19966276]
16. Phan AT, Kuryavyy V, Burge S, Neidle S, Patel DJ. Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J Am Chem Soc.* 2007; 129:4386–92. [PubMed: 17362008]
17. Guedin A, Gros J, Alberti P, Mergny JL. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.* 2010; 38:7858–68. [PubMed: 20660477]
18. Miyoshi D, Nakao A, Sugimoto N. Molecular crowding regulates the structural switch of the DNA G-quadruplex. *Biochemistry.* 2002; 41:15017–24. [PubMed: 12475251]
19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 26:841–2. [PubMed: 20110278]
20. Shirude PS, Okumus B, Ying L, Ha T, Balasubramanian S. Single-molecule conformational analysis of G-quadruplex formation in the promoter DNA duplex of the proto-oncogene c-kit. *J Am Chem Soc.* 2007; 129:7484–5. [PubMed: 17523641]
21. Capra JA, Paeschke K, Singh M, Zakian VA. G-quadruplex DNA sequences are evolutionary conserved and associated with distinct genomic features in *Saccharomyces cerevisiae*. *PLoS Comput Biol.* 6:e1000861. [PubMed: 20676380]
22. Santangelo GM, Tornow J. A *Saccharomyces cerevisiae* mitochondrial DNA fragment activates Reg1p-dependent glucose-repressible transcription in the nucleus. *Curr Genet.* 1997; 32:389–98. [PubMed: 9388294]
23. Ribeyre C, Lopes J, Boule JB, Piazza A, Guedin A, Zakian VA, Mergny JL, Nicolas A. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet.* 2009; 5:e1000475. [PubMed: 19424434]
24. Sanders CM. Human Pif1 helicase is a G-quadruplex DNA-binding protein with G-quadruplex DNA-unwinding activity. *Biochem J.* 2010; 430:119–28. [PubMed: 20524933]

Highlights

- A new method is described for detection of sequences with G-quadruplex forming potential (QFP) that derive from G-runs present on both strands of duplex DNA, which we call duplex-derived interstrand QFP (ddiQFP).
- Application of this method to detect ddiQFP in *S. cerevisiae* reveals that particular classes of ddiQFP motifs occur in an apparently non-random fashion, including differences between nuclear and mitochondrial genomes.
- Reexamination of a published dataset of sites targeted by the G-quadruplex-unwinding helicase Pif1, demonstrates that it not only preferentially targets intramolecular QFP but also ddiQFP sites.

Figure 1. Comparisons of interconversion between duplex DNA and A) intramolecular, B) ddiQFP, and C) compact ddiQFP forms of G4 DNA.

Figure 2.

Graphical representation of the nine classes of ddiQFP. Classes 1, 2 and 7-9 are equivalent to their reverse complementary forms when considered within the context of an unoriented DNA duplex, and Classes 3-6 are their own reverse-complements. Note that compact forms do not have a loop between contiguous G and C-runs.

Figure 3.

Locations of ddiQFP motifs with respect to the start of ORFs. A) Nuclear genome ORFs. B) Mitochondrial genome ORFs. A 35 nt window was used to define ddiQFP, and then the number of ORFs containing a ddiQFP motif within a 150 bp window surrounding each base from 1000 bp upstream to 1000 bp downstream from the ORF start was plotted.

Figure 4. Frequencies of ddiQFP motifs among the nine different classes as a function of window size. A) The pie charts depict the same data listed in the table, and window sizes are indicated. B) The proportion of ddiQFP motifs of each class within the mitochondrial genome, the nuclear genome, and overall are shown.

Figure 5. The percentages of combinations of G-runs and C-runs. The actual number of motifs is provided in the table, and reflects frequencies for one strand (the “Watson” strand) of genomic and mitochondrial DNA.

Table 1

Fourteen possible patterns of ddiQFP motifs.

Class	Regular expression	Loop1	Loop2	# of G-runs	# of C-runs
1	$C_{3+}N_aG_{3+}N_bG_{3+}N_cG_{3+}$	N_b	N_c	3	1
	$C_{3+}N_aC_{3+}N_bC_{3+}N_cG_{3+}$	$RC(N_b)$	$RC(N_b)$	1	3
2	$G_{3+}N_aG_{3+}N_bG_{3+}N_cC_{3+}$	N_a	N_b	3	1
	$G_{3+}N_aC_{3+}N_bC_{3+}N_cC_{3+}$	$RC(N_c)$	$RC(N_b)$	1	3
3	$C_{3+}N_aC_{3+}N_bG_{3+}N_cG_{3+}$	$RC(N_a)$	N_c	2	2
4	$G_{3+}N_aG_{3+}N_bC_{3+}N_cC_{3+}$	N_a	$RC(N_c)$	2	2
5	$C_{3+}N_aG_{3+}N_bC_{3+}N_cG_{3+}$	$RC(N_aG_{3+}N_b)$	$N_bC_{3+}N_c$	2	2
6	$G_{3+}N_aC_{3+}N_bG_{3+}N_cC_{3+}$	$N_aC_{3+}N_b$	$RC(N_bG_{3+}N_c)$	2	2
7	$C_{3+}N_aG_{3+}N_bG_{3+}N_cC_{3+}$	$RC(N_aG_{3+}N_bG_{3+}N_c)$	N_b	2	2
	$G_{3+}N_aC_{3+}N_bC_{3+}N_cG_{3+}$	$N_aC_{3+}N_bC_{3+}$	$RC(N_b)$	2	2
8	$C_{3+}N_aC_{3+}N_bG_{3+}N_cC_{3+}$	$RC(N_bG_{3+}N_c)$	$RC(N_b)$	1	3
	$G_{3+}N_aC_{3+}N_bG_{3+}N_cG_{3+}$	$N_aC_{3+}N_b$	N_c	3	1
9	$C_{3+}N_aG_{3+}N_bC_{3+}N_cC_{3+}$	$RC(N_c)$	$RC(N_aG_{3+}N_b)$	1	3
	$G_{3+}N_aG_{3+}N_bC_{3+}N_cG_{3+}$	N_a	$N_bC_{3+}N_c$	3	1

Motifs that are the reverse-complements (RC) of one another are grouped together, yielding a total of nine classes. Note that for adjacent G and C-runs, intervening nucleotides (N) are not required (i.e. a, b, and/or c may equal 0).

Table 2

Numbers of intramolecular QFPs and ddiQFPs found using the indicated sliding window lengths.

	30nt	35nt	50nt	100nt	200nt
Intramolecular QFP	146	262	904	4386	18964
Mitochondrial-specific intramolecular QFP	3	9	37	98	200
ddiQFP	458	693	2031	12298	44947
Mitochondrial-specific ddiQFP	177	199	262	326	509

Intramolecular QFP motifs were identified using similar sliding window algorithm (see “Materials and Methods”) except only G-runs within one strand were considered Total motifs in nuclear and mitochondrial genomic DNA are enumerated in the first two rows, and mitochondrial genome-specific motifs are enumerated in the third row.

Table 3

GO terms that are enriched among genes with ddiQFP (length 50 nt). The enrichment of GO term is calculated based on the proportion of genes with ddiQFP in each GO category divided by the proportions for all yeast ORFs.

A) Promoters with ddiQFP			
GOID	GO term	Enrichment	P-value
<i>Molecular Function</i>			
16491	oxidoreductase activity	1.91	6.71E-06
3735	structural constituent of ribosome	1.63	0.004
1071	nucleic acid binding transcription factor activity	1.56	0.023
3677	DNA binding	1.28i	0.044
5198	structural molecule activity	1.30	0.046
16791	phosphatase activity	1.60	0.050
8134	transcription factor binding	1.70	0.083
51082	unfolded protein binding	1.70	0.083
<i>Biological Process</i>			
6091	generation of precursor metabolites and energy	2.00	0.0004
51186	cofactor metabolic process	1.92	0.001
42221	response to chemical stimulus	1.48	0.004
45333	cellular respiration	2.07	0.005
6811	ion transport	1.76	0.007
6366	transcription from RNA polymerase II promoter	1.37	0.009
6629	lipid metabolic process	1.43	0.018
5975	carbohydrate metabolic process	1.40	0.022
7005	mitochondrion organization	1.36	0.031
9408	response to heat	2.00	0.032
6979	response to oxidative stress	1.77	0.040
2181	cytoplasmic translation	1.48	0.041
55086	nucleobase-containing small molecule metabolic process	1.41	0.044
6873	cellular ion homeostasis	1.58	0.046
16570	histone modification	1.50	0.065
6457	protein folding	1.50	0.084
6417	regulation of translation	1.50	0.089
6353	transcription termination, DNA-dependent	2.00	0.096
<i>Cellular Component</i>			
5740	mitochondrial envelope	1.98	2.54E-07
5739	mitochondrion	1.32	0.0002
16020	membrane	1.25	0.0004
5624	membrane fraction	1.84	0.0006

A) Promoters with ddiQFP			
GOID	GO term	Enrichment	P-value
5737	cytoplasm	1.08	0.005
5840	ribosome	1.38	0.016
5783	endoplasmic reticulum	1.31	0.034
B) ORFs with ddiQFP			
GOID	GO term	Enrichment	P-value
<i>Molecular Function</i>			
22857	transmembrane transporter activity	1.61	0.0004
16746	transferase activity, transferring acyl groups	1.89	0.002
16757	transferase activity, transferring glycosyl groups	1.94	0.004
16491	oxidoreductase activity	1.45	0.009
8134	transcription factor binding	1.90	0.030
16301	kinase activity	1.41	0.038
16874	ligase activity	1.31	0.085
<i>Biological Process</i>			
5975	carbohydrate metabolic process	1.78	3.11E-05
55085	transmembrane transport	1.60	0.005
51186	cofactor metabolic process	1.62	0.009
6091	generation of precursor metabolites and energy	1.64	0.012
6520	cellular amino acid metabolic process	1.45	0.016
6897	endocytosis	1.67	0.033
51049	regulation of transport	1.80	0.033
6865	amino acid transport	2.00	0.035
9311	oligosaccharide metabolic process	1.86	0.061
7005	mitochondrion organization	1.28	0.062
6629	lipid metabolic process	1.29	0.077
55086	nucleobase-containing small molecule metabolic process	1.31	0.089
6310	DNA recombination	1.38	0.094
<i>Cellular Component</i>			
5739	mitochondrion	1.44	2.31E-07
5737	cytoplasm	1.09	0.002
16020	membrane	1.21	0.002
5886	plasma membrane	1.52	0.003
5777	peroxisome	1.90	0.027
5773	vacuole	1.39	0.036
5740	mitochondrial envelope	1.27	0.056
5938	cell cortex	1.41	0.089

Table 4

Comparison of Pif1-K264A targeted intramolecular QFP and ddiQFP.

	Intramolecular QFP identified using the indicated sliding window length			
	30nt	50nt	100nt	200nt
Percent of Pif1-K264A targets with QFP	0.78%	4.36%	22.91%	57.67%
p-value	0.676	0.843	0.065	1
ddiQFP identified using the indicated sliding window length				
	30nt	50nt	100nt	200nt
Percent of Pif1-K264A targets with ddiQFP	3.22%	11.59%	39.98%	76.83%
p-value	<0.001	0.001	1	1