nature biotechnology

High-throughput sequencing of DNA G-quadruplex structures in the human genome

Vicki S Chambers[1,5], Giovanni Marsico[2,5], Jonathan M Boutell[3], Marco Di Antonio[1,2], Geoffrey P Smith[3] & Shankar Balasubramanian[1,2,4]

G-quadruplexes (G4s) are nucleic acid secondary structures that form within guanine-rich DNA or RNA sequences. G4 formation can affect chromatin architecture and gene regulation and has been associated with genomic instability, genetic diseases and cancer progression[1–4]. Here we present a high-resolution sequencing–based method to detect G4s in the human genome. We identified 716,310 distinct G4 structures, 451,646 of which were not predicted by computational methods[5–7]. These included previously uncharacterized noncanonical long loop and bulged structures[8,9]. We observed a high G4 density in functional regions, such as 5′ untranslated regions and splicing sites, as well as in genes previously not predicted to contain these structures (such as *BRCA2*). G4 formation was significantly associated with oncogenes, tumor suppressors and somatic copy number alterations related to cancer development[10]. The G4s identified in this study may therefore represent promising targets for cancer intervention.

The formation of DNA secondary structures can influence biological processes such as replication, translation and splicing[11,12]. G4 secondary structures arise in guanine-rich sequences where four guanine bases interact to form planar G-tetrads, which can self-stack[13]. G4 formation is kinetically fast, and these structures are thermodynamically stable under physiological conditions, particularly in the presence of K^+ (ref. 13). Recent studies using immunofluorescence to visualize G4s revealed their active formation in human cells and tissues and highlighted potential functional implications of these structures in diseases such as cancer[14–16]. The prevalence and distribution of G4s in the human genome is a key question that is currently addressed primarily on the basis of computational prediction[5–7]. G4 formation in a DNA template can be assessed using polymerase stop assays, which measure polymerase stalling at G4 sites[17]. However, no high-throughput, genome-wide method for G4 detection is currently available.

Here we describe such a method, called G4-seq, which we developed by combining features of the polymerase stop assay with Illumina next-generation sequencing[18]. We sequenced DNA from primary human B lymphocytes under conditions that either promote or disfavor G4 formation. Because polymerase stalling at G4 sites was found to affect base calling, sequencing readouts from both conditions were compared to elucidate the exact position of G4 structures (**Fig. 1**). We used two independent approaches to promote DNA G4 stabilization: adding K^+ or the G4-stabilizing ligand pyridostatin (PDS, 1 μM) to the sequencing buffers (ref. 19). For each condition, we compared sequencing quality and base calling before and after G4 stabilization in a human genomic DNA library spiked with four known control sequences (Online Methods and **Fig. 1**): two containing stable G4 structures (c-myc and c-kit), one mutated to prevent G4 formation (c-myc-mut) and the complementary C-rich strand of c-myc (c-myc-opp) that cannot fold into a G4.

We supplemented standard Illumina sequencing buffers with either 50 mM LiCl or NaCl, which do not cause strong G4 stabilization, or KCl, which stabilizes G4 structures[20]. The ionic strength of all buffers was kept constant. The overall sequencing quality, as quantified by Phred quality scores[21] (Q), was not globally affected by any of the added cations (**Supplementary Fig. 1**). However, quality was reduced in the presence of K^+ for a subset of sequences, including the G4-positive controls c-myc and c-kit and sequences computationally predicted to form G4s[5]. Conversely, the G4-negative controls c-myc-opp and c-myc-mut showed no change in quality under any condition (**Supplementary Fig. 2a**). Sequencing of the controls under Li^+ and Na^+ conditions revealed no alterations compared to the known input sequences (i.e., base mismatches <2%), whereas under K^+ conditions the G4-positive controls c-kit and c-myc had 34% and 46% mismatches, respectively (**Supplementary Fig. 2b**). Therefore, we sequenced each genomic DNA template twice—with an initial sequencing run (read 1) in Na^+ to ensure accurate sequencing and correct identification by alignment to the human reference genome (hg19), and a second sequencing run (read 2) under G4-stabilizing conditions (K^+) to detect structure formation by mismatch quantification on the basis of the sequence obtained in read 1.

We next explored whether specific stabilization of G4s by the ligand PDS, previously shown to induce polymerase stalling at G4 sites in cells[22], could also induce targeted sequencing errors. We sequenced read 1 in Na^+ and read 2 under the same cation conditions but with addition of PDS (1 μM, Online Methods). We measured mismatches of 45% for c-kit and 66% for c-myc but observed little effect (<5% mismatches) for c-myc-opp and c-myc-mut (**Supplementary Fig. 3**).

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. [2]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK. [3]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, UK. [4]School of Clinical Medicine, University of Cambridge, Cambridge, UK. [5]These authors contributed equally to this work. Correspondence should be addressed to S.B. (sb10031@cam.ac.uk).

Received 12 January; accepted 19 June; published online 20 July 2015; doi:10.1038/nbt.3295

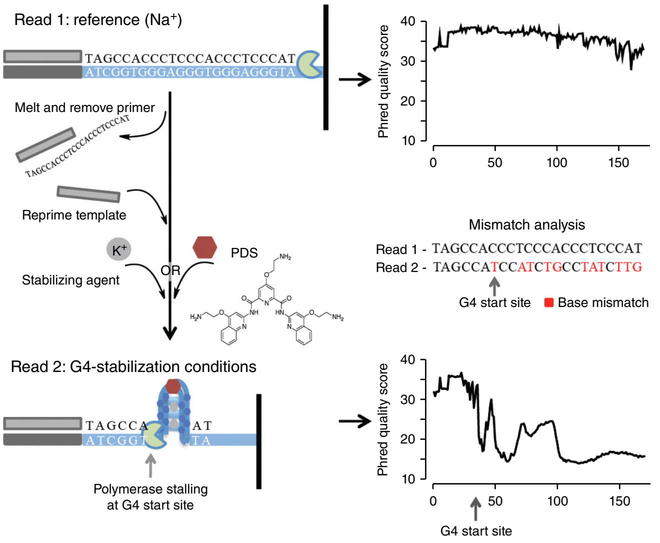

Figure 1 A schematic of the G4-seq method. In a typical G4-seq experiment, sequencing is done twice. Read 1 enables accurate sequencing and alignment of DNA fragments. Subsequently, the DNA synthesized during sequencing is removed and the original template resequenced (read 2) under conditions that promote G4 stabilization, either by the addition of PDS or by supplementing sequencing buffers with K^+. G4-induced polymerase stalling alters the sequencing readout from the beginning of the G4 structure resulting in a drop in sequencing quality from that point in read 2. Differences in sequencing quality and mismatches between read 1 and read 2 are analyzed to provide a map of G4 structures in the human genome.

The inspection of mismatches along the c-kit control, which contains two independent G4 motifs (c-kit*1* and c-kit*2*)[23,24], revealed that sequencing errors accumulated only after the G4 start sites, suggesting that under both K^+ and PDS conditions, the formation of DNA G4s causes polymerase stalling and mismatches in sequencing readout (**Fig. 2a**). When the polymerase encounters a stable G4 in the DNA template, a pause is induced, which can effectively truncate the reading of the template sequence. When this happens, the sequencer will continue to generate what appears to be a scrambled sequence beyond this point (**Supplementary Figs. 4** and **5**). Ordinarily such reads are removed during the data analysis, but we retained them in our experiment to detect G4 sites. Our approach therefore enables both the identification of G4-containing sequences and the exact location of the structure. Notably, only PDS addition induced substantial polymerase stalling at c-kit*1*, in agreement with the relative stability of the two G4s[23].

The analysis of 32 million reads comprising a subset of ~110,000 predicted quadruplexes[5] (PQs) showed higher mismatch levels (median of 20% in K^+ and 35% in PDS) in sequences containing PQs than in those that do not (non-PQs; < 2%) (**Fig. 2b**). Mismatch levels were generally high (>38%) for the sequence immediately following the start of the PQ motif and negligible (<1%) for the sequence in front (**Fig. 2c**), confirming a G4-dependent effect, as observed for c-kit. Although mismatch levels for non-PQs were low on average (<2%), a small fraction (~0.01) was found to have relatively high mismatch levels (>20%; ~149,000 sequences in K^+ and ~216,000 in PDS). The number of these non-PQs displaying >20% mismatches was much greater than the number of predicted PQs (~110,000; **Fig. 2b**), suggesting that the number and nature of human genomic G4s is substantially broader than previously predicted[5].

We applied G4-seq to generate a high-resolution map of G4 structures in the human genome (isolated from primary B lymphocytes (NA18507), Online Methods), using the Illumina HiSeq platform under Na^+ conditions in read 1 and either K^+ or PDS in read 2. Each experiment was performed in duplicate and yielded at least 285 million reads with an average coverage of 14× for the human genome (**Supplementary Table 1**). We set thresholds of 25% and 18% mismatches for PDS and K^+, respectively, to ensure a similar false positive rate of ~2%. Thus, any read with mismatches above these thresholds is considered a reliable indication of G4 formation and is termed observed G4 sequence (OQ). By applying these criteria, we identified 716,310 OQs in PDS and 525,890 OQs in K^+ within the human genome. Furthermore, 73% (in PDS) and 60% (in K^+) of all 361,424 predicted canonical G4-forming sequences (PQs) were present in the experimentally detected OQs (**Supplementary Table 2**). Ninety percent of PQs found in K^+ were also detected in PDS, and 383,984 of the total number of OQs were common to both conditions ($P < 10^{-16}$). The high overlap between distinct G4-stabilizing conditions provides independent validation of the assignment of OQs. Our data indicate that the OQs detected exclusively with PDS show much higher mismatch levels in K^+ than do random genomic intervals, and vice versa for OQs detected exclusively in K^+ (**Supplementary Fig. 6**). This suggests that it is the extent of stabilization under a given set of conditions that affects the likelihood of a G4 being detected by G4-seq. The OQs detected in the presence of PDS could also reflect the binding properties and specificity of the small molecule for G4 stabilization[25]. The use of a different G4-stabilizing ligand, PhenDC3 (ref. 26), showed a strong overlap (85%) with OQs

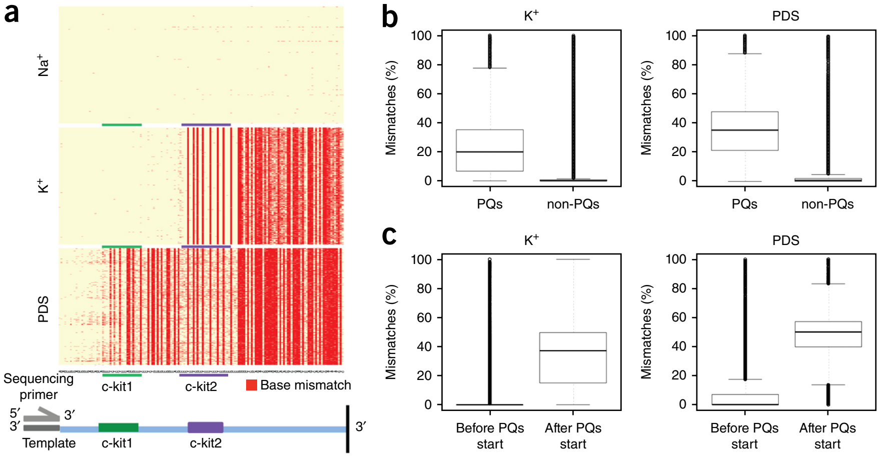

Figure 2 Analysis of G4-seq for known G4 sequences. (**a**) Identification of base mismatches for the c-kit control sequence depicted in a heatmap plot of sequencing in Na^+ (top), K^+ (middle) and PDS (bottom). Each row is an independent sequenced template, and each column corresponds to the sequenced bases. Mismatches are shown in red. (**b**) Box plots showing the mismatch percentage between read 1 and read 2 for reads with PQs (N = ~110,000) and non-PQs (N = ~32 million) for K^+ (left) and PDS (right). (**c**) Box plots representing the percentage mismatches for the reads containing a PQ, before or after the motif start site, for K^+ (left) and PDS (right). Center lines represent median values; box limits represent the interquartile range; whiskers extend each 1.5 times the interquartile range (**b**,**c**).

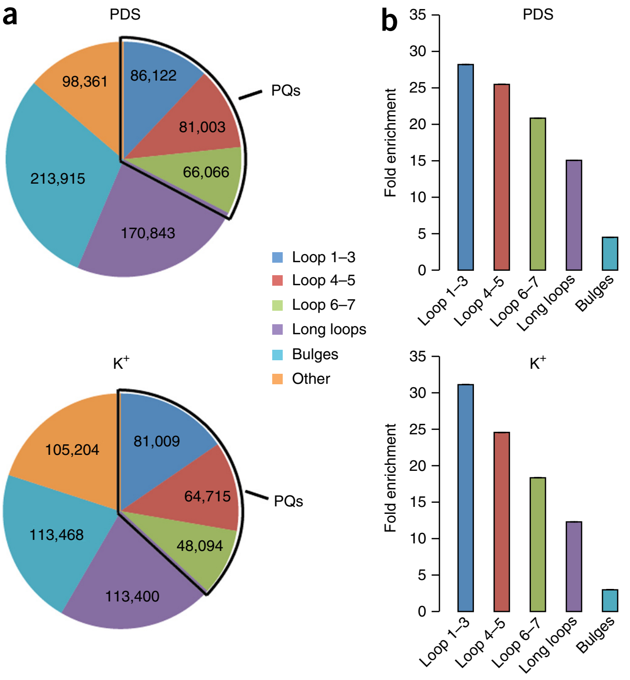

Figure 3 Structural analysis of OQs. (**a**) Number of OQs found in different G4 structural families, for Na+ with PDS or K+ sequencing conditions (Online Methods). Loop 1–3, Loop 4–5 and Loop 6–7, OQs with at least one loop of the indicated length; long loops, OQs with any loop of length >7; bulges, OQs with a bulge of 1–7 bases in one G-run or multiple 1-base bulges; other, sequences that do not fall into the categories above. (**b**) Fold enrichment (ratio) of each structural family represented in OQs over random genomic sequences measured for Na+ with PDS (top) and K+ (bottom) conditions. Error bars, s.e.m. from three independent randomizations. Fold enrichment values follow the relative thermodynamic stability of the different G4 families, with highest enrichment for G4 structures with short loops.

detected in PDS (**Supplementary Fig. 7**), suggesting that no major differences in binding specificity were observed with these two ligands.

Notably, ~70% of the OQs were not predicted from a classical description of a G4 structure[5]. Recent structural and biophysical studies have identified a small number of cases of stable noncanonical G4 structures in which either the loops are exceptionally long (>7 bases)[9,27] or a discontinuity in the G-tracts leads to bulges[8] (**Supplementary Fig. 8**). To elucidate distinct structural features, we categorized the OQs as follows (**Fig. 3a**): (i) canonical PQs, broken down in three subcategories according to loop length; (ii) long loops, sequences with any loop >7 bases; (iii) bulges, sequences with singe-nucleotide interruptions in one or more of the G-runs or a longer interruption in one G-run (for example, $GGH_{1–7}G$); (iv) other, sequences not belonging to the previous categories. Structural families were defined by a hierarchical assignment on the basis of sequence only. There is potential for multiple folding scenarios or polymorphism that was not accounted for in our assignment but could be assessed by dedicated structural studies on a case-by-case basis. Long loops and bulges accounted for 21.5% and 21.6% of total OQs in K+ and 24% and 30%, respectively, in PDS. The remaining OQs (in the 'other' category) may have the potential to form G4s, such as structures containing multinucleotide bulges, two-tetrad G4s or topologies

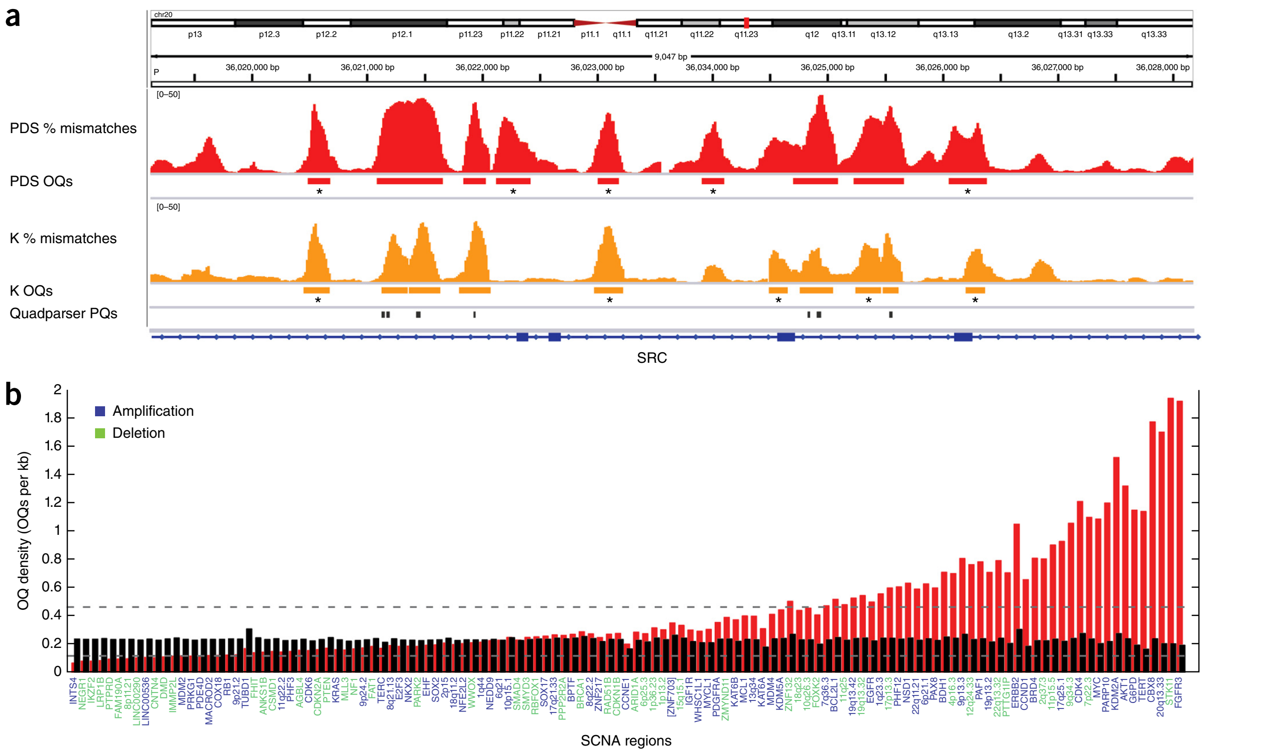

Figure 4 Genomic distribution of experimentally determined OQs. (**a**) Genome browser view of PQs and OQs across the *SRC* oncogene. Tracks for mismatches in reads aligning to the reverse strand (–) for PDS and K+ are shown in red and orange, respectively. Red and orange bars indicate OQ regions above threshold; black bars indicate PQs. OQs not predicted are indicated by asterisk (sequences in **Supplementary Table 5**). (**b**) OQ density (red) in different SCNAs[10] compared to random intervals (black), measured as number of OQs per kilobase. Blue labels indicate SCNAs representing amplifications; green labels indicate SCNAs representing deletions. Dotted lines show values corresponding to 0.5 and 2 times the average random density (0.22). Bars are sorted according to the fold enrichment of OQs density over random (**Supplementary Table 8**).

comprising both long loops and bulges (**Supplementary Table 3**). Collectively, these findings unravel a data set of stable G4 sequences that could not have been easily identified a priori in genomic DNA by computational approaches.

We measured the fold enrichment of OQs compared to random genomic intervals to assess the likelihood of each class being detected by G4-seq. Sequences with short loops had high enrichment (>25-fold) under both PDS and K^+ conditions, whereas sequences with longer loops or bulges displayed lower enrichment (<15-fold; **Fig. 3b**) which is consistent with the relative thermodynamic stability of the different G4 structures[8,9,28]. Also, less stable G4s were more easily detected by PDS (**Supplementary Fig. 9**).

To understand the potential functions of G4s, we quantified the prevalence of OQs in genomic regions associated with promoters, 3′ and 5′ untranslated regions (UTRs), exons, introns and splicing junctions (**Supplementary Table 4**). Notably, a large proportion of these regions (up to 49% in PDS and 46% in K^+) comprised exclusively noncanonical G4s (i.e., long loops or bulges). The highest density of G4s was found in 5′ UTRs and splicing sites, which is consistent with a role in post-transcriptional regulation, as supported by the recent finding in the 5′ UTR of *EIF4A1* (ref. 2).

Visual inspection of genes with biologically important G4s, such as *SRC* and *MYC*[22,29], or genes rich in PQs, such as *MYL5* and *MYL9* (**Fig. 4a**, **Supplementary Fig. 10**), confirmed that G4-seq is able to identify both predicted and uncharacterized G4s and is highly specific for the G-rich strand (**Supplementary Fig. 11** and **Supplementary Table 5**). Biophysical characterization using circular dichroism (CD) spectroscopy on selected noncanonical OQs confirmed the formation of G4 structures in these sequences (**Supplementary Fig. 12**). We found noncanonical G4s within many genes that have few or no PQs (**Supplementary Table 6**), including important cancer-related genes such as *BRCA1*, *BRCA2* and *MAP3K8*. Genes with a high number of G4s may be particularly sensitive to treatment with G4-stabilizing ligands, as shown for the oncogene *SRC*[22]. Our experimental map also identified oncogenes and tumor suppressors with a notably high G4 density, such as *CUL7*, *FOXA1*, *TUSC2* and *HOXB13* (**Supplementary Table 7**). This map further revealed significant enrichment of G4s ($P = 4.5 \times 10^{-8}$) in somatic copy number (SCN) alterations (SCNAs), which are signatures of cancer[10] (**Fig. 4b**). In particular, we observed high G4 density in regions containing oncogenes such as *MYC*, *TERT*, *AKT1*, *FGFR3* and *BCL2L1* (**Supplementary Table 8**) that specifically relate to SCN amplifications ($P = 2 \times 10^{-7}$) rather than deletions ($P = 0.01$). This is consistent with a mechanistic link between G4s and the sites of genomic instability, a hallmark of cancer[3,30].

We have established a high-throughput, genome-wide method for profiling G4 DNA secondary structures with high resolution. Our study provides insights into the nature of G4 formation, including noncanonical structural features. Our experimental data set shows enrichment of G4s in regulatory regions, in addition to oncogenes and SCNAs, and provides a resource of genomic targets for further biological and mechanistic studies and potential future therapeutic intervention. This approach is applicable to the study of G4 prevalence in any given genome. G4-seq may also be extended to the detection of other DNA secondary structures and to DNA–small molecule interactions.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene Expression Omnibus: the data reported in this paper are available under accession number GSE63874.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Lowe and D. Tannahill for critical reading of the manuscript and D. Beraldi for technical support. We thank P. McCauley (Illumina) who prepared the custom sequencing buffers. We are grateful to the Biotechnology and Biological Sciences Research Council (BBSRC) and Illumina for the studentship supporting V.S.C. (BB/I015477/1). The S.B. research group is supported by program funding from Cancer Research UK and the European Research Council and project funding from BBSRC.

AUTHOR CONTRIBUTIONS

V.S.C. and J.M.B. carried out the experiments. G.M. designed, implemented and performed the analysis. V.S.C., J.M.B., G.M., M.D.A., S.B. and G.P.S. designed the experiments. V.S.C., G.M., M.D.A. and S.B. interpreted the results and co-wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Rodriguez, R. & Miller, K.M. Unravelling the genomic targets of small-molecules using high-throughput sequencing. *Nat. Rev. Genet.* **15**, 783–796 (2014).
2. Wolfe, A.L. *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**, 65–70 (2014).
3. Maizels, N. Genomic stability: FANCJ-dependent G4 DNA repair. *Curr. Biol.* **18**, R613–R614 (2008).
4. Haeusler, A.R. *et al.* C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**, 195–200 (2014).
5. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).
6. Eddy, J. & Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* **34**, 3887–3896 (2006).
7. Kikin, O., D'Antonio, L. & Bagga, P.S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* **34**, W676–W682 (2006).
8. 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).
9. Guédin, 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).
10. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
11. 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).
12. Cruz, J.A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604–609 (2009).
13. Davis, J.T. G-quartets 40 years later: from 5′-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed. Engl.* **43**, 668–698 (2004).
14. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–186 (2013).
15. Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.* **42**, 860–869 (2014).
16. Biffi, G., Tannahill, D., Miller, J., Howat, W.J. & Balasubramanian, S. Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PLoS ONE* **9**, e102711 (2014).
17. Weitzmann, M.N., Woodford, K.J. & Usdin, K. The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand tetraplex formation. *J. Biol. Chem.* **271**, 20958–20964 (1996).
18. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
19. Rodriguez, R. *et al.* A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J. Am. Chem. Soc.* **130**, 15758–15759 (2008).
20. Hud, N.V., Smith, F.W., Anet, F.A.L. & Feigon, J. The selectivity for K^+ versus Na^+ in DNA quadruplexes is dominated by relative free energies of hydration: A thermodynamic analysis by H-1 NMR. *Biochemistry* **35**, 15383–15390 (1996).
21. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using Phred. 1. Accuracy Assessment. *Genome Res.* **8**, 175–185 (1998).

22. Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.* **8**, 301–310 (2012).
23. Fernando, H. *et al.* A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry* **45**, 7854–7860 (2006).
24. Rankin, S. *et al.* Putative DNA quadruplex formation within the human c-kit oncogene. *J. Am. Chem. Soc.* **127**, 10584–10589 (2005).
25. Marchand, A. *et al.* Ligand-induced conformational changes with cation ejection upon binding to human telomeric DNA G-quadruplexes. *J. Am. Chem. Soc.* **137**, 750–756 (2015).
26. De Cian, A., DeLemos, E., Mergny, J.-L., Teulade-Fichou, M.-P. & Monchaud, D. Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J. Am. Chem. Soc.* **129**, 1856–1857 (2007).
27. 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).
28. Bugaut, A. & Balasubramanian, S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry* **47**, 689–697 (2008).
29. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA* **99**, 11593–11598 (2002).
30. Paeschke, K. *et al.* Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **497**, 458–462 (2013).

ONLINE METHODS

Design of control sequences. Full-length control sequences (sequence of interest underlined) are as follows:

Control 1 (Positive): c-kit
5′-Adapter 1-AGAGCCGCGAGCGGCGAGCAGCAGCCCTCTCCTCCC
AGCGCCCTCCCTCTGCGCGCCGG
CCACGCCCCTCCTCGCCTCCCTCCCTCCGCCCGCCCGGGGCTCG
CG-Adapter 2-3′.

Control 2 (Negative): c-myc-opp
5′-Adapter 1-ATTAGCGAGAGAGGATCTTTTTTCTTTTCCCCCACGC
CCTCTGCTTTGGGAACCCGGGA
GGGGCGCTTATGGGGAGGGTGGGGAGGGTGGGGAAGGGGGAGG
AGAG-Adapter 2-3′.

Control 3 (Positive): c-myc
5′-Adapter 1-TCTCCTCCCCACCTTCCCCACCCTCCCCACCCTCCCC
ATAAGCGCCCCTCCCGGGTTCCC
AAAGCAGAGGGCGTGGGGGAAAAGAAAAAAGATCCTCTTCGCT
AATAG-Adapter 2-3′.

Control 4 (Negative): c-myc-mut
5′-Adapter 1-CTCCTCTTCACCTTCTTCACTCTCTTCACTCTCTTCAT
AAGCGCCCCTCCCGGGTTCCCAA
AGCAGAGGGCGTGGGGGAAAAAAAAAAAGATCCTCTCTCGCTA
ATAG-Adapter 2-3′.
Adapter 1-5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCC
TACACGACGCTCTTCCGATCT-3′
Adapter *2*-5′-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAC
TGATATATCTCGTATGCCGTCTT
CTGCTTG-3′

The c-myc and c-kit positive controls were designed based on the human genomic sequence of two regions in the promoter of the oncogenes *MYC* and *KIT*, respectively, which are well-studied examples of G-quadruplex (G4)-forming motifs[23,24,29]. Crucially, controls were designed complementary to the G4 motif, i.e., the C-rich sequence to ensure that during Illumina cluster generation the G-rich sequence becomes immobilized to the flow cell surface and acts as the template for sequencing. This protocol is necessary to allow the study of G4 structures on polymerase procession. Two negative control sequences were also designed based on the c-myc sequence: (i) c-myc-opp: the complementary G-rich strand of the c-myc G4, which becomes the C-rich template sequence upon cluster generation; (ii) c-myc-mut: a mutant of c-myc that can no longer form a G4.

Control sequence library preparation. Synthetic oligonucleotides of the control sequences, and their complement sequences, with a 5′ phosphate group and an A overhang (Biomers) were prepared using nuclease-free water at the final concentration of 1 μg/ml. The two complementary oligonucleotide sequences of each control (100 ng/μl) were annealed in 10 mM Tris, 50 mM NaCl buffer by heating to 95 °C for 10 min and then cooled to 20 °C at 1 °C/min. The annealed DNA was prepared for Illumina sequencing by ligation of Illumina adapters using a T4 DNA ligase at 30 °C for 10 min. Following AMPure bead clean-up, the adapted sequences were PCR amplified using standard Illumina PCR primers and gel purified (Qiagen MinElute Gel Extraction kit). Purified fragments were ligated into Life Technologies PCR-Blunt Vectors and transformed according to standard methods. Plasmid DNA was purified from selected clones (Thermo Scientific GeneJET plasmid Miniprep Kit), followed by Sanger sequencing (GATC) to confirm the sequence identity and directionality. DNA inserts of the chosen clones (C-rich variant of the insert in the case of c-myc, c-kit and c-myc-mut and G-rich for c-myc-opp) were isolated by EcoRI-HF digestion and gel purification to generate sequences ready for use in sequencing. Sequences were quantified using a Qubit Fluorimeter (Life Technologies) and denatured according to standard Illumina protocols. Control sequences were spiked into a human genomic library at a final concentration of 0.01 pM for all sequencing experiments.

Genomic library preparation. Purified human genomic DNA isolated from primary human B lymphocytes (NA18507) was purchased from Coriell Institute for Medical Research and prepared for sequencing using TruSeq DNA sample prep kit (Illumina) according to the manufacturer's protocol. Human template DNA was denatured as in standard Illumina protocols and used at 8 pM for sequencing on MiSeq instruments (Illumina) and 12 pM for all sequencing on an Illumina HiSeq 2500 in Rapid Run mode (with the addition of 0.01 pM of each control sequence).

Modified sequencing buffer preparation. In collaboration with Illumina, the standard sequencing buffers (incorporation, wash and cleavage buffers) were supplemented with K^+, Na^+ or Li^+ at a final concentration of 50 mM for the incorporation and wash buffers and 1 M for the cleavage buffer. In addition, for small-molecule experiments with PDS, all buffers were prepared using Na^+ at 50 mM final concentration, and PDS[4] (1 μM) was added to the incorporation buffer on the instrument. All other reagents used were from standard proprietary Illumina sequencing kits.

G4-seq protocol. Illumina sequencing was performed using either MiSeq or HiSeq 2500 Rapid Run instrumentation, using the same basic protocol. A human genomic library containing synthetic control sequences (prepared as above) was used as template. Cluster generation and amplification were carried out according to standard procedures. The template DNA was then sequenced using buffer conditions containing Na^+ (read 1) for 250 cycles (MiSeq) or 150 cycles (HiSeq 2500). The newly synthesized DNA strand was removed by denaturation to leave the original template DNA strand. The read 1 sequencing primer (HP10) was then added to the flow-cell and hybridized as per standard sequencing protocols. Annealing buffer (10 mM Tris and 100 mM KCl, pH 7.4) was added to the flow cell and the temperature increased to 65 °C for 5 min, followed by cooling to 20 °C at 1 °C/min, in order to promote G4 formation in immobilized template DNA. For sequencing experiments with PDS or PhenDC3, the small molecule was added to the flow cell (1 μM in annealing buffer) and equilibrated for 30 min at room temperature. Sequencing was then performed on the template DNA (read 2) in G4-stabilization conditions (i.e., either K^+ sequencing buffers or with PDS addition in Na^+ buffer). The sequencing read length was 250 and 150 bp for the MiSeq and HiSeq 2500, respectively. Base-calling log (bcl) files from the sequencing run were processed to generate FASTQ files for further analysis. Sequencing scripts for the MiSeq experiments (**Supplementary Script 1**) and HiSeq experiments performed for G4-seq (**Supplementary Scripts 2** and **3**) are provided.

FASTQ files. The FASTQ format[31] consists of: (i) a read identifier to allow identification of sequences from the same cluster when performing different sequencing reads, hence read 1 and read 2; (ii) a measure of base-calling quality—the Phred quality score, Q, which is inversely related to the probability that the corresponding base-call is incorrect (i.e., a high Q score indicates a low probability of erroneously calling the given base, whereas a lower Q score indicates greater probability that the given base is incorrectly called); (iii) the actual base call, where the nucleotide with highest confidence is assigned to each sequencing position. Read quality was calculated as the average Phred quality of all bases; the quality difference was calculated as read 1 quality minus read 2 quality; the percentage of mismatches was calculated comparing base calling at read 1 and read 2 and counting the fraction of different calls across the whole read.

Different cation analysis. Sequencing was performed in Li^+, Na^+ and K^+ as described above. Experiments were done in duplicate for K^+ and Li^+ conditions and in triplicate for Na^+ conditions. FASTQ files were obtained from MiSeq 250-bp single-end reads. Files were aligned to the human genome (hg19) using the bwa mem aligner with default parameters (http://bio-bwa.sourceforge.net/).

K^+ and PDS genomic analysis. Sequencing was performed as described above. Two technical replicates were performed for each G4-stabilization condition on HiSeq instrumentation. FASTQ files were obtained from HiSeq 2500 150-bp single-end reads. FASTQ files from read 1 were aligned to the human genome (hg19) using the bwa mem aligner with default parameters (http://bio-bwa.sourceforge.net/). Bam alignment files were processed using bedtools (https://code.google.com/p/bedtools/): (i) bam files were converted to bed files

doi:10.1038/nbt.3295

(command bamToBed); (ii) bed files were expanded 30 bases downstream (command slopBed -s -r 30); (iii) expanded bed files were grouped to keep only the best alignments for each read (command groupBy -g 4 -c 5 -o max); (iv) FASTA sequence files were extracted from the bed intervals (command bedtools getfasta -s); (v) FASTA sequence files and the FASTQ files from both read 1 and read 2 were loaded in R (http://www.r-project.org/) for analysis. Sequence tails beyond poly(A) tails (≥9 bases) were trimmed, as they represent the end of the DNA fragment attached to the flow cell. The difference in the quality score and percentage of mismatches (% mismatches) between read 1 and read 2 for each individual base was calculated and stored for each read, as well as a coverage count incremented by one. All single-base values calculated from the processed reads were then pooled to generate genomic tracks of mismatch percentage (average of values) and total coverage (sum of values). To ease data handling, genomic tracks were binned in intervals of length 15 bases and smoothed with a moving average of order 15 (i.e., window size around the point value to be smoothed).

Control sequences analysis. FASTQ files were generated from the MiSeq (cations experiments) or the HiSeq 2500 (K$^+$ and PDS experiments) sequencing platforms. FASTQ were aligned to a FASTA file containing only the control sequences by using the bwa mem aligner with default parameters (http://bio-bwa.sourceforge.net/). The Phred quality score (Q) and the base-calling extracted from reads were successfully aligned to each control sequence then were analyzed.

PQ identification and positional analysis. For each sequencing read, the aligned sequence information was extracted as above and PQs were identified according to the Quadparser algorithm by searching for the regular expression '(G{3,}[ATGC]{1,7}){3,}G{3,}'. For positional analysis, 'before PQs start' is defined as the sequence up to 12 bases upstream of the PQ start site (12 bases is the approximate footprint of DNA polymerase). 'After PQs start' is defined as the remaining sequence, from 12 bases upstream the PQ start site until the end of the sequence (excluding any sequencing beyond the poly(A) tail).

OQ detection. Quadparser-predicted PQs were considered as a positive set (PQs) and reads without PQs as a negative set (non-PQs). For all reads, % mismatches were calculated (range 0%–100%). For each threshold t_i, the following numbers were calculated: TP$_i$, true positives (i.e., reads with PQs above the threshold t_i); FP$_i$, false positives (i.e., reads without PQs above the threshold t_i); FN$_i$, false negatives (i.e., reads with PQs below the threshold t_i); and TN$_i$, true negatives (i.e., reads without PQs below the threshold t_i). The false positive rate, FPR$_i$ = (FP$_i$ / (FP$_i$ + TN$_i$) was calculated for each threshold t_i and the thresholds for OQ detection were set in order to have FPR ~0.02 (high specificity), i.e., 2% of the non-PQs would be detected as OQs. This yielded thresholds of 18% and 25 for K$^+$ and PDS sequencing respectively. A sequence with a % mismatch value above these thresholds was defined as an Observed G-quadruplex Sequence (OQs). For the genomic analysis, continuous regions with a maximal peak summit above the threshold (18% for K$^+$ and 25% for PDS) were considered as OQ regions. OQ regions displaying multiple peak were split into separated OQs using PeakSplitter (http://www.ebi.ac.uk/research/bertone/software). Regions from two replicates were analyzed independently, keeping strand information separated. We only considered high-confidence OQ regions in genomic intervals common to both replicates for further analysis (command intersectBed -s of the bedtools).

Structural analysis of OQ categories. OQ sequences were stratified into different OQ categories by searching for different regular expressions (**Fig. 3**). To assign univocally an OQ region to a specified category and avoid considering the same region multiple times, we followed priority rules based on the predicted stability from high to low (loop 1–3 > loop 4–5 > loop 6–7 > long loops > bulges > other). The different categories were defined as follows: loop 1–3, (G{3,}N{1,3}){3,}G{3,}, with N = [ATCG]; loop 4–5: (G{3,}N{1,5}){3,}G{3,} and not in previous category; loop 6–7, (G{3,}N{1,7}){3,}G{3,} and not in a previous category; long loops, (G{3,}N{1,12}){3,}G{3,} or G{3,}N{1,7}G{3,}N{13,21}G{3,}N{1,7}G{3,} and not in a previous category; bulges, OQ sequences with any G-run being GH$_{1\text{-}7}$GG or GHGGN{1,7}GGHG, with H = [ATC] and not in a previous category; other, not in any other category. The 'other' category was further stratified into subcategories containing OQs having either multiple bulges with more than one nucleotide (for example, GH{2,5}GGN{1,7}GGH{2,5}G) or two-tetrads motifs (GGN{1,7}GGN{1,7}GGN{1,7}GG) (**Supplementary Table 3**). Finally, the ratio of the numbers of each category in PDS and K$^+$ was calculated (**Supplementary Fig. 9**).

Fold-enrichment analysis of OQ structural categories. The 525,890 K$^+$ OQ intervals were randomly shuffled three times across the genome (command shuffleBed in bedtools) to generate random sequences of the same size distribution as the OQs. This was also done for the 716,310 PDS OQ intervals. The different OQ categories were identified and counted in both the experimental OQs and the three randomized intervals. For each category, the ratio of real OQ over the average of three random cases was calculated and plotted as fold-enrichment for PDS and K$^+$ (**Fig. 4b**). Error bars were calculated for each category as the s.e.m. of three random replicates, and each s.e.m. was then divided by the average of random counts in the category to adapt it to the fold-enrichment plot.

Genomic regions analysis. Gene annotation files were downloaded from the UCSC genome browser website (https://genome.ucsc.edu/), genome version hg19, and different genomic regions (5′ UTRs, 3′ UTRs, exons, introns, promoters, translational start sites (TSS) and splice regions) were extracted and stored as genomic intervals (bed file format). For each region, the total number of regions, the total region size and the number of PDS or K$^+$ OQs overlapping to the region intervals (command intersectBed of the bedtools) were calculated. The number of regions overlapping exclusively with Quadparser PQs and with noncanonical PQs (i.e., long loops and bulges) were calculated (**Supplementary Table 4**). Any intervals overlapping sequences from both categories were excluded from analysis to avoid ambiguity.

Genes and oncogenes analysis. For each gene annotated in the version hg19 of the human genome, the number of Quadparser-predicted PQs of OQs in PDS and OQs in K$^+$ were counted. The density of PQs or OQs was calculated by dividing the respective counts by the gene body length and multiplying by 1,000 (i.e., density is the number of structures per kilobase). For oncogene analysis, we considered 498 oncogenes and 766 tumor suppressors[22]. Genes with a PQ density less than half of *SRC* PQ density but with a OQ density higher than *SRC* OQ density were extracted (**Supplementary Table 4**).

Somatic copy number alteration (SCNA) analysis. 140 SCNAs previously identified as being associated with cancer were considered[10], of which 70 were amplifications and 70 were deletions. Only SCNAs smaller than 10 Mb were analyzed, leaving a total of 123 regions (50 deletions and 73 amplifications). For each region the number of OQs was counted. OQ genomic intervals were then randomly reshuffled three times (random-OQs) and the number of random-OQs in each SCNA was calculated and averaged. The OQ and random-OQ counts were divided by each region size and multiplied by 1,000, to give a density per kilobase. The OQ and random-OQ densities were then compared and their ratio calculated such that SCNA regions with ratio >1 are enriched in OQs compared to random, whereas SCNAs with ratio <1 are depleted (**Supplementary Table 8** and **Fig. 4b**). The difference between OQs and random densities was statistically assessed for the 123 regions using the two-tailed *t*-test; SNCA amplifications (n = 73) and deletions (n = 50) were also tested in the same way against their counterpart (random-OQs for amplification and deletion regions, respectively).

Statistical analysis. For experiments shown in **Figure 2b**,**c** we performed the nonparametric two-sample Wilcoxon rank sum test (function wilcox.test in R). For each comparison, given the large sample size and the strong difference, we obtained P values $<2.2 \times 10^{-16}$. This is the minimal value possible returned by the test implementation.

Supplementary code. A collection of scripts for performing raw sequencing analysis, OQ detection, control sequence analysis and structural analysis of OQ categories is provided as **Supplementary Code**.

31. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).