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

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

1. Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK.
2. Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK.
3. Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, Essex, CB10 1XL, UK.
4. School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK.
* These authors contributed equally to this work.

During active transcription and replication chromatin architecture is altered, allowing formation of DNA secondary structures[1]. G-quadruplexes (G4s) have emerged as important regulatory DNA structures and have been associated with genomic instability, genetic diseases and cancer progression[2-4]. Experimental evidence for G4 prevalence in the entire human genome is still lacking. We present a high-resolution sequencing-based method that detected 716,310 distinct G4s in the human genome, more than predicted by computational methods[5-7], including structural variants previously uncharacterised in a genomic context[8,9]. We observed high G4-density in functional regions, such as 5' UTRs and splicing sites, and in genes not predicted to have such structures (*BRCA1* and *BRCA2*). We found a significant association of G4 formation with oncogenes and tumor suppressors, and with Somatic Copy-Number Alterations (SCNAs) that act as cancer drivers[10]. Our results support that G4s are promising targets for cancer intervention and suggest novel candidates for further biological and mechanistic studies.

The formation of DNA and RNA secondary structures is of vital importance to fundamental biological processes, such as replication, translation and splicing[11,12]. While RNA structure-mapping on a genomic-scale is established[13,14], extending these methodologies to interrogate DNA secondary-structure formation remains a challenge. G4s are a particular class of DNA secondary structures that is emerging as a regulatory element for key biological processes and an important therapeutic target[2-4]. G4 structures can form in guanine-rich sequences from the interaction of four guanine bases to generate a planar G-tetrad, which can subsequently self-stack[15]. G4 formation is kinetically fast and they are thermodynamically very stable under physiological conditions, particularly in the presence of K^+ [15]. Recently, G4 formation has been visualised in human cells and tissues by means of immuno-fluorescence[16-18]. These and other studies highlight the importance of G4 formation in specific genes, underpinning the value of studying these structures at a larger scale. The formation of G4s can be assessed *in vitro* by measuring the stalling of a polymerase along its template at G4 sites (polymerase stop assay)[19]. Here, we adapt the polymerase stop assay together with Illumina® next-generation sequencing[20] to establish G4-Seq, the first method to detect and map DNA secondary structures on a genome-wide scale. We altered sequencing conditions to either disfavour or promote G4 formation on the sequencing array, comparing the respective sequencing readouts to elucidate the exact position of the DNA structure (Fig. 1). We used two independent approaches to promote DNA G4 stabilisation: i) adding K^+; ii) adding the G4 stabilising ligand pyridostatin (PDS, 1 μM)[21]. For each condition, we compared sequencing quality and base calling before and after G4 stabilisation in a human genomic DNA library spiked with four known control sequences (Methods, 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.

In our experiments, we supplemented standard Illumina sequencing buffers with either 50 mM LiCl or NaCl, which do not cause strong G4 stabilization, or KCl that does stabilizes G4 structure[22] (Methods), keeping the ionic strength of all buffers constant. The overall sequencing quality, as quantified by Phred Quality scores[23] (Q, Methods), was not globally affected by any of the added cations (Extended Data Fig. 1). However, quality was reduced *only* 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 a G4 [5]. Conversely, the G4-negative controls *c-myc-opp* and *c-myc-mut* showed no change in quality under any condition (Extended Data 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* displayed 34% and 46% mismatches respectively (Extended Data 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 stabilising conditions (K^+), to detect structure formation by mismatch quantification based on the sequence obtained in Read-1.

We next explored whether specific stabilisation of G4s by the ligand PDS, previously shown to induce polymerase stalling at G4 sites in cells[24], could also induce targeted sequencing errors. We performed Read-1 in Na^+ and Read-2 under the same cation conditions but with addition of PDS (1 μM, *Methods*). Herein, we measured mismatches of 45% for *c-kit* and 66% for *c-myc* but little effect (< 5% mismatches) for *c-myc-opp* and *c-myc-mut* that are unable to form G4s (Extended Data Fig. 3). The inspection of mismatches along the *c-kit* control, which contains two independent G4 motifs *c-kit1* and *c-kit2*,[25,26] revealed that sequencing errors accumulated only after the G4 start sites, suggesting that under both K^+ and PDS conditions the formation of DNA G4s cause polymerase stalling and mismatches in sequencing readout (Fig. 2a). In fact, when the polymerase encounters a stable G4 in the DNA template a pausing 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, as illustrated by Supplementary Figures 1 and 2. Ordinarily such reads are removed during the data analysis, whereas we have 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. Interestingly, only PDS addition induced significant polymerase stalling at *c-kit1* in agreement with the relative stability of the two G4s[25].

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 as opposed to those without (non PQs; < 2%) (Fig. 2b). Mismatch levels were generally high (> 38%) immediately after the PQ motif and negligible (< 1%) beforehand (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), far greater than the number of predicted PQs (~110,000; Fig 2b). Thus, suggesting that the number and nature of human genomic G4s is substantially broader than previously predicted[5].

This method, which we call G4-Seq, was applied to generate a high-resolution map of G4 structures in the human genome (*NA18507*, 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 14x 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% (Methods). 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 (Extended Data Table 1). 90% of PQs found in K^+ were also detected in PDS and 383,984 of the overall OQs were common to both conditions ($p<10^{-16}$). The high overlap between distinct G4 stabilising conditions provides independent validation of the assignment of OQs. Our data indicates that the OQs detected exclusively with PDS do in fact also display significantly high mismatch levels in K^+ (compared to random genomic intervals) and accordingly for OQs detected exclusively in K^+ (Supplementary Figure 3), suggesting that it is the extent of stabilisation 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 stabilisation[27]. The use of a different G4-stabilising ligand, PhenDC3[28], showed a strong overlap (85%) with OQs detected in PDS (Supplementary Figure 4), suggesting that no major differences in binding specificity were observed with these two ligands.

Notably, the majority (~70%) of the OQs were actually *not* predicted from a classical description of G4 structure[5]. Recent structural and biophysical studies have identified a small number of cases of stable non-canonical G4 structures in which either the loops are exceptionally long (>7 bases)[9,29], or a discontinuity in the G-tracts leads to bulges[8] (Extended Data Fig. 4). To elucidate distinct structural features, the OQs were grouped as follows (Methods): 1) Canonical PQs: in three categories according to loop length; 2) Long loops: sequences with any loop > 7 bases; 3) Bulges: sequences with singe-nucleotide interruptions in one or more of the G-runs or a longer interruption in one G-run (e.g. $GGH_{1-7}G$); 4) Other: sequences not belonging to the previous categories (Fig. 3a). Structural families are defined by a hierarchical assignment based on sequence only (Methods). There is potential for multiple folding scenarios or polymorphism, that is not accounted for in our assignment, but which could be assessed by dedicated structural

studies on a case-by-case basis. Long loops and Bulges accounted respectively for 21.5% and 21.6% of total OQs in K^+ and 24% and 30% in PDS. The remaining OQs (category Other) may have the potential to form G4s, such as structures containing multi-nucleotide bulges, two-tetrads G4s, or topologies comprising both long loops and bulges (Extended Data Table 2). Collectively, these findings have unraveled a dataset 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 to be detected by G4-Seq (Methods). Sequences with short loops have high enrichment (>25 fold) under both PDS and K^+ conditions, whereas sequences with longer loops or bulges displayed lower enrichment (<15 fold; Fig. 3b) consistent with the relative thermodynamic stability of the different G4 structures[8,9,30]. Also, less stable G4s were more easily detected by PDS (Extended Data Fig. 5).

To understand the potential functions of G4s we evaluated the existence of OQs in genomic regions associated with promoters, 3' and 5'-UTRs, exons, introns and splicing junctions (Extended Data Table 3). Notably, a large proportion of these regions (up to 49% in PDS and 46% in K^+) comprise *exclusively* non-canonical G4s (i.e. Long loops or Bulges). The highest density of G4s was found in 5' UTRs and splicing sites, consistent with a role in post-transcriptional regulation, as supported by the recent finding in the 5' UTR of *eIF4A*[2].

Visual inspection of genes with biologically important G4s (*SRC*, *MYC*)[24,31] or genes rich in PQs (*MYL5*, *MYL9*; Fig. 4a, Extended Data Fig. 6) confirmed that G4-Seq is a powerful tool to identify both predicted and uncharacterised G4s, and is highly specific for the G-rich strand (Extended Data Fig. 7, Supplementary Table 2). We found non-canonical G4s within many genes that have few or no PQs (Supplementary Table 3), 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-stabilising ligands, as shown for the oncogene *SRC*[24]. Our experimental map also identified oncogenes and tumor suppressors with a notably high G4 density, such as *CUL7*, *FOXA1*, *TUSC2* and *HOXB13* (Supplementary Table 4). This map further revealed significant enrichment of G4s ($p = 4.5e^{-8}$) in somatic copy number alterations (SCNAs), which are signatures of cancer[10] (Fig. 4b). In particular, high G4 density is observed in regions containing oncogenes such as *MYC*, *TERT*, *AKT1*, *FGFR3* and *BCL2L1* (Supplementary Table 5) that specifically relate to SCN amplifications ($p = 2e^{-7}$) 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[3,32].

We have established a high-throughput, genome-wide method that profiles G4 DNA secondary structure with high resolution. Our study reveals new insights into the nature of G4s that form in the human genome, including non-canonical structural features. Our experimental dataset shows enrichment of G4s in regulatory regions, in addition to oncogenes and SCNAs and provides a resource of novel genomic targets for further biological and mechanistic studies and potential future therapeutic intervention. We anticipate that our approach can be extended to study the prevalence of G4s, and potentially other DNA secondary structures, in any genome. Furthermore, G4-Seq can be exploited to detect DNA-small molecules interaction in a genomic context.

1 Rodriguez, R., Miller, K. M. Unravelling the genomic targets of small-molecules using high-throughput sequencing *Nat. Rev. Genet.* **15**, 783-96, (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**, R613-614, (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**, W676-682, (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 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).

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 Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696-700, (2014).

14 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716, (2009).

15 Davis, J. T. G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed.* **43**, 668-698, (2004).

16 Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182-186, (2013).

17 Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.* **42**, 860-869, (2014).

18 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).
19 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).
20 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, (2008).
21 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).
22 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).
23 Ewing, B., Hillier, L., Wendl, M. C., Green, P. Base-Calling of Automated Sequencer Traces Using Phred. 1. Accuracy Assessment *Genome Research* **8**, 175-185, (1998).
24 Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.* **8**, 301-310, (2012).
25 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).
26 Rankin, S. *et al.* Putative DNA quadruplex formation within the human c-kit oncogene. *J. Am. Chem. Soc.* **127**, 10584-10589, (2005).
27 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).
28 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).
29 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).
30 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).
31 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. U S A* **99**, 11593-11598, (2002).
32 Paeschke, K. *et al.* Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **497**, 458-462, (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Dr. Chris Lowe and Dr. David Tannahill for critical reading of the manuscript and Dr. Dario Beraldi for technical support. We thank Patrick 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.C (BB/I015477/1). The S.B. research group is supported by programme funding from Cancer Research UK and from the European Research Council and project funding from BBSRC.

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

Author information The data reported in this paper is available at the NCBI's GEO repository, accession number GSE63874 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63874). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.B. (sb10031@cam.ac.uk).

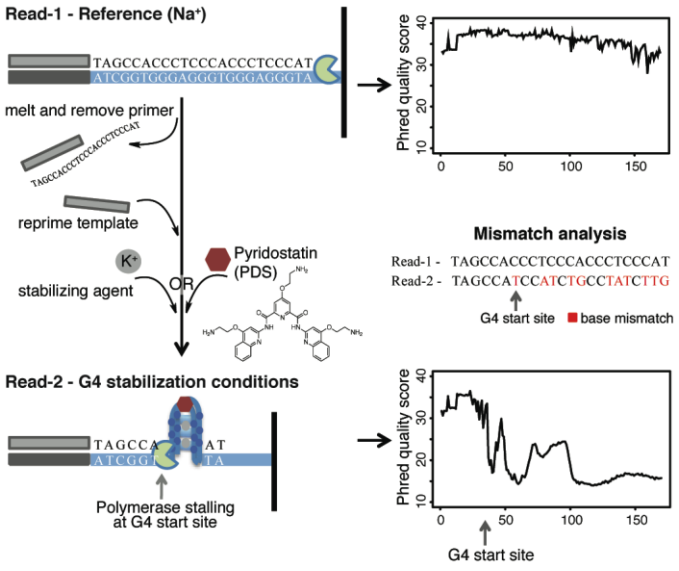

Figure 1: A schematic of the G4-Seq method. In a typical G4-Seq experiment sequencing is performed twice. A first sequencing run under Na^+ conditions (Read-1) enables accurate sequencing and alignment of DNA fragments. Subsequently, the DNA synthesised during sequencing is removed and the original template re-sequenced (Read-2) under conditions that promote G-quadruplex (G4) stabilization: either by the addition of the G4-ligand 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 only. Differences in sequencing quality and mismatches between Read-1 and Read-2 are analysed to provide a map of G4 structures in the human genome.

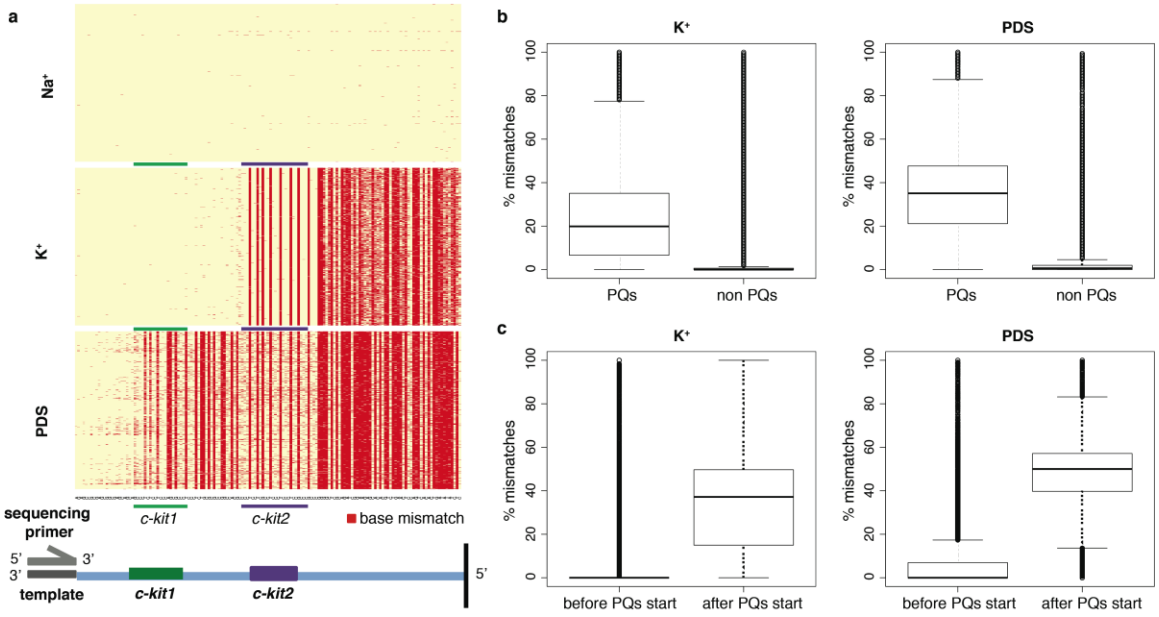

Figure 2: Analysis of G4-seq for known G-quadruplex sequences. a) Identification of base mismatches for the *c-kit* control sequence depicted in a heat-map plot under different sequencing conditions. Each row is an independent sequenced template, while each column corresponds to each the sequenced bases. Yellow background indicates no difference to the known input sequence and the sequence experimentally obtained, while red indicates mismatches. The two G-quadruplex motifs are indicated in green (*c-kit1*) and purple (*c-kit2*). Top: sequencing in Na^+, where negligible mismatches were observed. Middle: sequencing in K^+ showed mismatches accumulation starting at *c-kit2*, thus suggesting polymerase stalling. Bottom: sequencing in presence of PDS revealed stalling already at the first G4 motif (*c-kit1*) and significant mismatch accumulation. b) Boxplots showing the mismatch percentage between Read-1 and Read-2 for reads with *Quadparser*-predicted PQs (PQs; *N*~110,000) and without (non-PQs; *N*~32 million) for K^+ (left) and PDS (right). c) Boxplots representing the percentage mismatches for the reads containing a PQ, before or after the motif start site, for K^+ (left) and PDS (right).

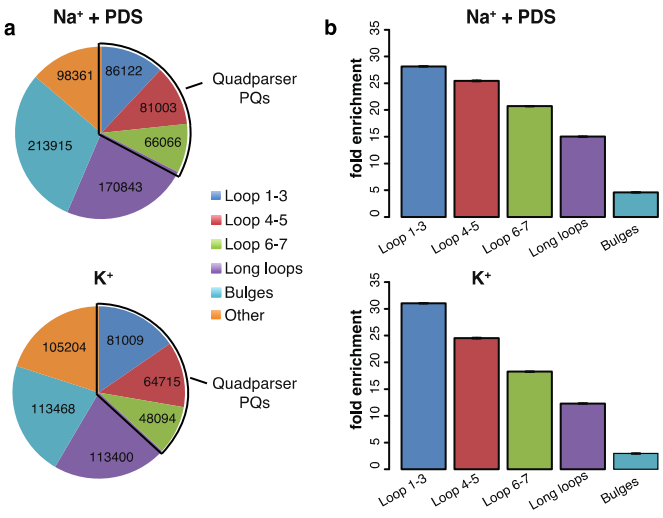

Figure 3: Structural analysis of Observed G-quadruplex sequences (OQs). a) Number of OQs found in different G-quadruplex structural families, for Na^+ + PDS or K^+ sequencing conditions (Methods). The different families are defined as follows. Loop 1-3; Loop 4-5; Loop 6-7: OQs with at least one loop 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 which 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^+ + PDS (top) and K^+ (bottom) conditions. Error bars are SEM of 3 independent randomizations. Fold enrichment values follow the relative thermodynamic stability of the different G4 families, with highest enrichment for G4 structures with short loops compared to longer loop counterparts. Treatment with PDS enables the detection of G4 structural variants with lower intrinsic stability.

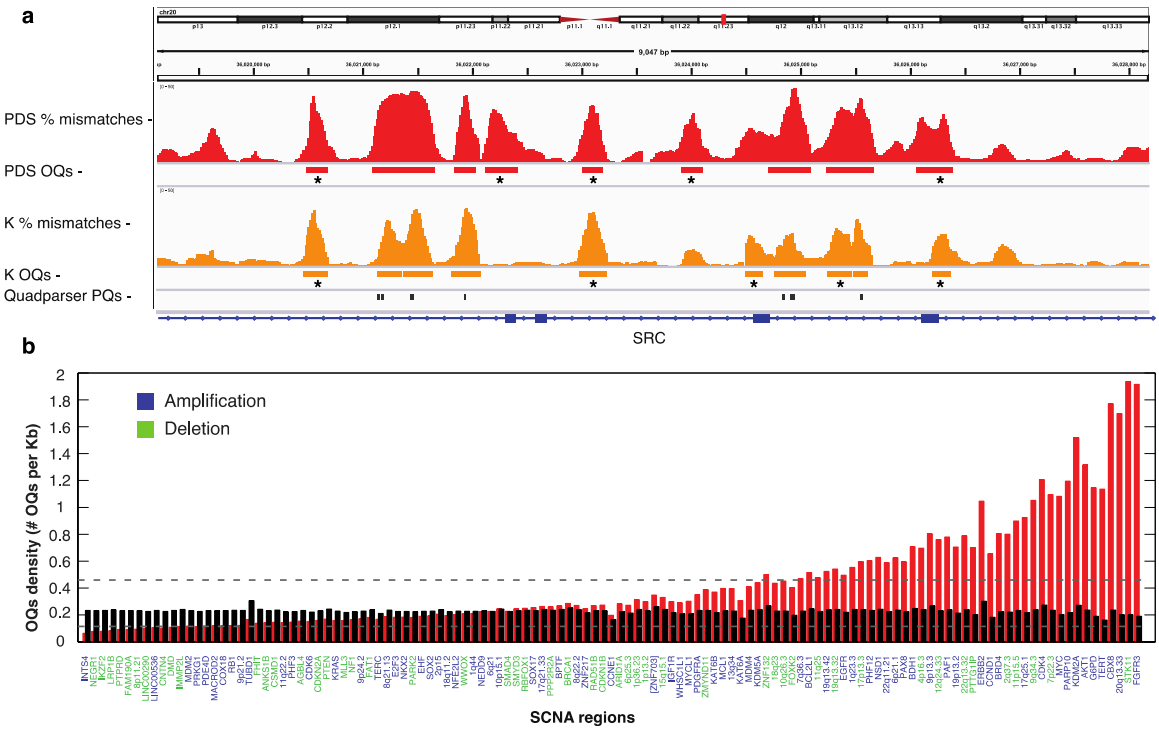

Figure 4: Genomic distribution of experimentally-determined OQs. a) Genome browser view of PQs and OQs across the *SRC* oncogene. Red and orange show tracks for mismatches in reads aligning to the reverse strand (-) for PDS and K^+, respectively. Bars of the same colors indicate OQ regions above threshold, while black bars indicate Quadparser PQs. OQs not predicted by Quadparser are indicated by * (sequences in Supplementary Table 2). G4-Seq enables detection of G4 sequences that were not previously predicted by computational algorithms in this oncogene. b) OQs density (red) in different SCNAs[10] compared to random intervals (black), measured as number of OQs per kilobase. Blue gene labels: SCNAs representing amplifications. Green gene labels: SCNAs representing deletions. Dotted lines: 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 5). A correlation was observed between the OQs density per kilobase and with SCNAs associated with amplifications, suggesting a potential role of G4 structure in carcinogenesis.

Methods

Design of control sequences

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

Control 1 (Positive): c-kit

5'-Adapter 1-AGAGCCGCGAGCGGCGAGCAGCAGCCCTCTCCTCCCAGCGCCCTCCCTCTGCGCGCCGG
CCACGCCCCTCCTCGCCTCCCTCCCTCCGCCCGCCCGGGGCTCGCG-Adapter 2-3'.

Control 2 (Negative): c-myc-opp

5'-Adapter 1- ATTAGCGAGAGAGGATCTTTTTTCTTTTCCCCCACGCCCTCTGCTTTGGGAACCCGGGA
GGGGCGCTTATGGGGAGGGTGGGGAGGGTGGGGAAGGGGGAGGAGAG-Adapter 2-3'.

Control 3 (Positive): c-myc

5'-Adapter 1- TCTCCTCCCCACCTTCCCCACCCTCCCCACCCTCCCCATAAGCGCCCCTCCCGGGTTCCC
AAAGCAGAGGGCGTGGGGGAAAAGAAAAAAGATCCTCTTCGCTAATAG-Adapter 2-3'.

Control 4 (Negative): c-myc-mut

5'-Adapter 1- CTCCTCTTCACCTTCTTCACTCTCTTCACTCTCTTCATAAGCGCCCCTCCCGGGTTCCCAA
AGCAGAGGGCGTGGGGGAAAAAAAAAAAGATCCTCTCTCGCTAATAG-Adapter 2-3'.

where:

Adapter 1- 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
Adapter *2*- 5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTATGCCGTCTT
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[25,26,28]. 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 immobilised 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: 1) *c-myc-opp*: the complementary G-rich strand of the *c-myc* G4, which becomes the C-rich template sequence upon cluster generation; 2) *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 manufacturers 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 synthesised 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 hybridised as per standard sequencing protocols. Annealing buffer (10mM Tris and 100mM 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 immobilised 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-stabilisation conditions, i.e. either K^+ sequencing buffers or with PDS addition in Na^+ buffer. The sequencing read length was 250 and 150 base pairs (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.

FASTQ files

The FASTQ format[33] consists of: 1) a read identifier to allow identification of sequences from the same cluster when performing different sequencing reads, hence Read-1 and Read-2; 2) 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, while a lower Q score indicates greater probability that the given base is incorrectly called); 3) 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. Two replicates were performed for K^+ and Li^+ conditions and three replicates for Na^+ conditions. FASTQ files were obtained from MiSeq 250 bp single-end reads. Files were aligned to the human genome (*hg19*) by 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-stabilisation 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/): 1) bam files were converted to bed files (command bamToBed); 2) bed files were expanded 30 bases downstream (command slopBed -s -r 30); 3) expanded bed files were grouped to keep only the best alignments for each read (command groupBy -g 4 -c 5 -o max); 4) FASTA sequence files were extracted from the bed intervals (command bedtools getfasta -s); 5) 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, together with coverage count of +1. 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 finally 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 analysed.

PQs 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 analysed independently, keeping strand information separated. We only considered high confidence OQ regions in genomic intervals common to both replicates for further analyses (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 $GH_{1-7}GG$ 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 sub-categories containing OQs having either multiple bulges with more than one nucleotide (e.g., GH{2,5}GGN{1,7}GGH{2,5}G) or two-tetrads motifs (GGN{1,7}GGN{1,7}GGN{1,7}GG) (Extended Data Table 2). Finally, the ratio of the numbers of each category in PDS and K^+ was calculated (Extended Data Fig. 5).

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 standard error of the mean (SEM) of three random replicates, and each SEM 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, TSSs 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 non-canonical PQs (i.e., Long loops and Bulges) were calculated (Extended Data Table 3). 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 1000 (i.e. density is the number of structures per kilobase). For oncogene analyses, we considered 498 oncogenes and 766 tumour suppressors[24]. Genes with a PQs density less than half of *SRC* PQs density but with a OQs density higher than *SRC* OQs 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 SCNA less than 10 Mb in size were analysed, leaving a total of 123 regions (50 deletions and 73 amplifications). For each region t

he 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 OQs and random-OQs counts were divided by each region size and multiplied by 1000, to give a density per kilobase. The OQs and random-OQs 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 5; 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 only, respectively).

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).

Extended Data Figure 1: Overall sequencing quality in sequencing experiments with the different cations Li^+, Na^+ and K^+. Each plot visually shows base calling quality (Phred quality score, Q; y-axes) for the 250 sequenced bases (x-axes), in two independent experiments, with sequencing buffers containing Li^+ (top), Na^+ (middle) and K^+ (bottom), as generated by the program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Yellow bars and black whiskers are box plots for the respective base positions; red lines are median values; blue lines are mean values.

Extended Data Figure 2: Sequencing quality and sequencing errors (% mismatches) for control sequences. Bar plots showing the average Phred quality score (Q) (a) and % mismatches (b) for the 4 control sequences when sequencing with different cations Li^+ (left), Na^+ (middle) and K^+ (right). *c-kit*, *c-myc*: positive controls; *c-myc-opp*, *c-myc-mut*: negative controls (see Methods). Data is taken from a number of independent sequencing experiments: 3 for Na^+, 2 for Li^+ and K^+. The numbers of different control sequences (i.e. independent sequencing clusters on the flow cell) in the combined experiments are (order; *c-kit*, *c-myc-opp*, *c-myc*, *c-myc-mut*): 2741, 1139, 1040, 10945 for Li^+; 8235, 3076, 2787, 26974 for Na^+; 2935, 1315, 1, 12809 for K^+. Bars are standard deviations. No error bar present for *c-myc* in K^+ (n=1).

Extended Data Figure 3: Sequencing errors for controls in PDS conditions. % mismatches for the control sequences in the same sequencing experiment with Na^+ sequencing buffers during the first read (Read-1; left) followed by the addition of the small-molecule PDS in Na^+ throughout the second read (Read-2; right). Error bars are SEMs (respectively: 0.16, 0.02, 0.18 and 0.07 for left plot; 0.12, 0.08, 0.15 and 0.09 for right plot). N = 948, 367, 367 and 3990 for *c-kit*, *c-myc-opp*, *c-myc*, *c-myc-mut*.

Extended Data Figure 4: Different families of G-quadruplex structures: Left: canonical PQs predicted by Quadparser (L1-3=N1-7, with N=A|C|T|G). Middle: PQs with longer loops (L1-3=N8-12 or L2=N8-21). Right: PQs with a single bulge B1=H1-7 or multiple bulges B2=H1-5 (H=A|T|C).

Extended Data Figure 5: Detection of OQs representing different G-quadruplex structural families in PDS versus K^+ conditions. Fold enrichment (ratio) between the numbers of OQs in PDS over K^+ for each category (see B). Values > 1 indicate higher numbers in PDS. G-quadruplex structural families: Loop 1-3; Loop 4-5; Loop 6-7: OQs with at least one loop of the

indicated length; Long loops: OQs with any loop of length 8 to 12 for L1-3 or 8 to 21 for L2; Bulges: OQs with one bulge of 1 to 7 bases (A, T, C) or multiple bulges of 1 base.

Extended Data Figure 6: Comparison of genomic regions in PDS and K^+ sequencing conditions. a) Genome browser view of a genomic region within *MYC* oncogene. Red and orange tracks: % mismatches in reads aligning to the reverse strand (-) for PDS and K^+, respectively. OQ intervals are shown as red and orange bars below the corresponding peaks.. b) Genome browser view of a genomic region within the *MYL5-MFSD7* gene. Black and blue tracks: % mismatches in reads aligning to the forward strand (+) for PDS and K^+, respectively. OQ intervals are shown as black and blue bars below the corresponding peaks. c) Genome browser view of a genomic region within the *MYL9* gene. All colours and features as in a). See Supplementary Table 2 for sequence details. For all panels, OQs not predicted by Quadparser are indicated by * and Quadparser PQs are shown as black bars.

Extended Data Figure 7: Comparison of forward versus reverse strands in PDS sequencing conditions. A) Genomic region within the *MYL9* gene. Red and black tracks: % mismatches in reads aligning to the reverse strand (-) and forward strand (+), respectively. OQs intervals are shown as red and black bars below corresponding peaks. Quadparser PQs are shown below in black. OQs not predicted by Quadparser are indicated by asterisks (*). See Supplementary Table 2 for sequence details.

Extended Data Table 1: Quadparser PQs detected by G4-Seq. The number and percentage of Quadparser PQs detected by G4-Seq under PDS or K^+ conditions or common to both. Two replicate Illumina HiSeq sequencing runs were performed for each condition. These data show the high degree of reproducibility and overlap between two different G-quadruplex stabilisation conditions.

Extended Data Table 2: Number of OQs in the category "Other". Multiple bulges = $G[ATC]_{2\text{-}5}GGLGG[ATC]_{2\text{-}5}G$; two-tetrads = GGLGGLGGLGG, with L = $N_{1\text{-}7}$; % G content = percentage of G nucleotides in the OQ sequence.

Extended Data Table 3: Distribution of OQs in different genomic regions. Several measurements reporting the genomic distribution of OQs in PDS (top half) or K^+ (bottom half) are listed in the table; columns are as follows. Region: genomic features- UTR: untranslated region; TSS: transcription start site; promoters 1000 up: 1000 bases upstream the TSS; TSS 1000 up down: 1000 bases up- and down-stream the TSS; splice 50: 50 bases up- and down-stream splice sites (i.e. exon-intron junctions); # regions: number of disjoint genomic regions; total region size: sum of all disjoint genomic regions; OQs density: 1000 * (# OQs) / (total region size); # regions with OQs: number of genomic regions overlapping with at least one OQ. # regions with non-canonical OQs: number of genomic regions overlapping exclusively with OQs having a long loop or a bulge; # regions with PQs: number of genomic regions overlapping exclusively with *Quadparser*-predicted PQs (loop size 1-7). Ratio non-canonical OQs / PQs: ratio of the number of regions with non-canonical and canonical PQs.