G-quadruplex structures mark human regulatory chromatin

Robert Hänsel-Hertsch[1], Dario Beraldi[1], Stefanie V. Lensing[1], Giovanni Marsico[1], Katherine Zyner[1], Aled Parry[1], Marco Di Antonio[2], Jeremy Pike[1], Hiroshi Kimura[4], Masashi Narita[1], David Tannahill[1] & Shankar Balasubramanian[1,2,3]

1. Cancer Research UK, Cambridge Institute, Li Ka Shing Centre, Cambridge, United Kingdom.
2. Department of Chemistry, University of Cambridge, Cambridge, United Kingdom.
3. School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom.
4. Cell Biology Unit, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan

Correspondence and requests for materials should be addressed to S.B. (sb10031@cam.ac.uk).

G-quadruplex (G4) structural motifs have been linked to transcription[1,2], replication[3] and genome instability[4,5] and are implicated in cancer and other diseases[6-8]. However, it is crucial to demonstrate the *bona fide* formation of G4 structures within an endogenous chromatin context[9,10]. Herein, we address this via the development of G4 ChIP-seq, an antibody-based G4 chromatin immunoprecipitation and high-throughput sequencing approach. We find that ~10,000 G4s are predominantly present in human regulatory, nucleosome-depleted chromatin. G4s were enriched in the promoters and 5'UTR regions of highly transcribed genes, particularly in genes related to cancer and in somatic copy number amplifications, such as *MYC*. Strikingly, *de novo* and enhanced G4 formation is associated with increased transcriptional activity as revealed by small molecule-induced chromatin relaxation and in immortalized versus normal cellular states. Our findings show for the first time that regulatory, nucleosome-depleted chromatin and elevated transcription shape the endogenous human G4 DNA landscape.

Regulatory regions in chromatin are characterized by nucleosome depletion to allow access of proteins directing gene transcription, replication and epigenetic plasticity[11]. These accessible regions are quantitative indicators of cellular fate, origin and identity,[12] yet the underlying DNA structural features remain largely uncharacterized. G4 DNA structures have recently been visualized in human cells[13,14] and directly mapped in purified genomic DNA[15]. Employing G4-promoting conditions, ~700,000 G4-induced polymerase-stalling sequences have been observed in the human genome *in vitro* using purified single-stranded DNA as a template for G4-seq, a high-throughput sequencing method for the discovery of structural features in DNA[16]. G4s have been computationally predicted to form in replication origins[17,18] and nucleosome-depleted regions[19]. Endogenous proteins, such as human ATRX[20] and XPB/XPD[1] and yeast Pif-1[21] and Rif-1[22] have been mapped by ChIP-seq to G-rich genomic sequences predicted to adopt G4s *in*

vivo, however, direct capture of G4 structure formation at high-resolution within a chromatin context is missing.

We set out to elucidate the relationship between G4 structure, chromatin and transcription in human cells. To achieve this, we developed a G4 ChIP-seq protocol employing an engineered G4 structure-specific antibody (BG4)[13] to map the genome-wide location of G4s in the chromatin of the spontaneously immortalized, non-oncogenic, human epidermal keratinocyte HaCaT cell line[23] (Fig. 1, Supplementary Fig. 1). We mapped nucleosome-depleted regions, using Formaldehyde Assisted Isolation of Regulatory Elements and sequencing (FAIRE-seq)[24] and Assay for Transposase Accessible Chromatin and sequencing (ATAC-seq)[25] (Supplementary Fig. 1) and determined transcriptional output by RNA-seq. Overall in HaCaT chromatin 10,560 high-confidence G4 ChIP-seq peaks were identified with 87 % conforming to an *in vitro* observed G4 structural motif (Fig. 1, Supplementary Fig. 2, see method section, '*G4 motif and enrichment analysis*')[16]. Analysis of G4 motifs revealed that 21 % of all G4 ChIP-seq peaks contain G4 motifs, with loop sizes of 1–7 (canonical G4s) (Supplementary Fig. 2)[26]. Canonical G4s were enriched more in G4 ChIP-seq peaks than other G4 motifs (Supplementary Fig. 2), e.g. with longer loops[27] or bulges in G-tracts[28]. We also validated the G4-structure specific enrichment observed in the HaCaT G4 ChIP-seq peaks in control experiments by G4 ChIP-qPCR, such as pre-incubation of BG4 with a G4-forming or single-stranded DNA sequence prior to G4 ChIP (Supplementary Fig. 3), and by bioinformatic analysis of recurrent MEME-derived motifs and their density in G4 ChIP-seq regions (Supplementary Fig. 2). Strikingly, most (98 %) G4 ChIP-seq peaks overlapped with regions as defined by the union of FAIRE- and ATAC-seq regions (Fig. 1b). We considered the possibility that BG4 might simply target the most accessible open chromatin regions, rather than true G4 sites. To rule out that BG4 enriches the most abundant accessible chromatin sites, regardless of G4 presence, we overlapped the high-confidence G4 ChIP-seq peak set (10,560) with the 10,560 most accessible FAIRE sites. FAIRE accessibility was ranked according to q-value for peak enrichment assigned by the MACS2 peak caller. We found that ~44 % of the G4 ChIP-seq peaks did not overlap with these highly accessible FAIRE sites but are actually found in relatively less accessible FAIRE sites than the 10,560 most accessible FAIRE sites. This suggests that chromatin accessibility alone is not sufficient for BG4 binding. We further validated the presence of G4 structures in nucleosome-depleted chromatin using immunofluorescence microscopy colocalization for BG4 and another G4 antibody (1H6)[14], with eu- and heterochromatin markers in HaCaT cells. We found that both antibodies significantly colocalized with transcriptionally active euchromatin (H3K4me3 and/or RNA Pol2),

while they showed no significant colocalization with heterochromatin (H3K9me3) (Supplementary Fig. 4). Collectively, these results demonstrate that G4 formation in human cells is predominantly restricted to regulatory nucleosome-depleted regions in euchromatin.

Overall, HaCaT G4 ChIP-seq peaks represent about 1 % of all sequences detected by G4-seq[16] (hereon referred to as **O**bserved **Q**uadruplex **s**equences) (Supplementary Fig. 5). However, of all OQs that map to nucleosome-depleted regions in HaCaT cells, only a subset (26 %) overlapped with G4 ChIP-seq peaks (Supplementary Fig. 5). Importantly, the remaining 74 % OQs not detected by BG4 in nucleosome-depleted regions showed a comparable representation of G4 motifs (sequence and structure) to the positive G4 ChIP subset (26 %) (Supplementary Fig. 5). This suggests that additional genomic features/events, besides nucleosome-depletion, are likely to be important for G4s to form stably in chromatin. While the G4 ChIP-seq peaks are considerably enriched in promoter and 5'UTR regions, i.e. found more often than expected by random chance, they are mostly located outside of these regions (Fig. 1c). This raises the possibility that the transcriptional state of a gene might affect G4 formation in nucleosome-depleted regions. Indeed, we observed genes (4,522) that display a G4 ChIP-seq peak in their promoter have on average significantly ($p = < 0.0001$) higher transcriptional levels than genes (4,345) lacking a promoter G4 ChIP-seq peak, yet are still found nucleosome-depleted and contain an OQ (Fig. 1d). Since regulatory proteins shape open chromatin, we anticipated that published ChIP-seq for XPB/XPD[1] or SP1[29] that unwind or bind G4s *in vitro* or ΔNP63 a master regulator of keratinocyte transcription[30], would be enriched in the G4 ChIP-seq data. Indeed, a notable enrichment of all, XPB/XPD (60-fold), SP1 (330-fold) and ΔNP63 (72-fold) high-confidence peaks were observed in the HaCaT G4 ChIP-seq data, suggesting that these might directly interact with endogenous G4s (Fig. 1b). In contrast and as anticipated, the H3K9me3 and H3K27me3 ChIP-seq peaks showed no correlation with the G4 ChIP-seq sites (Fig. 1b). Taken together, these results suggest that the chromatin context predominantly restricts G4 formation to regulatory nucleosome-depleted regions associated with genes showing elevated transcription. This is consistent with transcriptional up-regulation of predicted G4-forming genes observed in G4 helicase-deficient (i.e. WRN, BLM) human cells[31] and binding sites of G4 helicases XPB, XPD and the yeast PIF-1 homolog, Pfh1, in transcriptionally active chromatin[1,32]. Fig. 1a shows example profiles for G4 ChIP-, ATAC- and FAIRE-seq aligned with peak profiles that mark OQs for the *SRC* and *MYC* oncogenes, that have previously been suggested to be regulated by G4s[2,10]. Here, we directly confirm the presence of G4 structures in the nuclease hypersensitivity element of the oncogenes *MYC*, and in the upstream element and gene body of *SRC* (Fig. 1a). Recently, we reported an

enrichment of OQs at cancer-related genes and somatic copy number alterations (SCNAs)[16], we now extend these findings to show endogenous G4 (i.e. G4 ChIP) enrichment in cancer-related genes, such as *MYC, TP53, JUN, HOXA9, FOXA1, RAC1* (Supplementary Fig. 6, Supplementary Table 1) and SCNAs (Supplementary Fig. 7, Supplementary Table 2). Among all cancer-related SCNA amplifications and oncogenes, *MYC* shows the highest G4 ChIP density, supporting an already suggested role[33] for G4 structures in cancer progression (Supplementary Fig. 6, 7).

We reasoned that if G4 structures are coupled to nucleosome-depletion and their transcriptional status, then changing the chromatin landscape would cause a concomitant shift in the G4 profile. We induced chromatin relaxation of HaCaT cells using the histone deacetylase (HDAC) inhibitor Entinostat[34] to stabilize transcriptionally active chromatin through promoter-specific histone H3K27 acetylation[35]. HaCaT cells treated with 2 μM Entinostat for 48h[35] were analyzed by G4 ChIP-, ATAC-, and RNA-seq (Fig. 2, Supplementary Fig. 8). Genome-wide changes in G4 ChIP-seq peaks, open chromatin and mRNA levels were quantified bioinformatically by Differential Binding Analysis (DBA)[36] (Fig. 2a, b, Supplementary Fig. 8). HDAC inhibition resulted in the appearance of 4,117 new, or more intense G4 ChIP-seq sites and 7,970 open chromatin regions (Fig. 2a, b). The emergent 4,117 G4 ChIP-seq sites are located in new or pre-existing nucleosome-depleted regions, e.g. *SIGIRR*, *GRIN1* and a non-coding region (Fig. 2a, c), however, importantly none of these sites were found in nucleosome-depleted regions in closed chromatin after Entinostat treatment (Fig. 2a, c). Next, we explored the relationship between G4s identified in promoters of nucleosome-depleted regions and their transcriptional status of associated genes by comparing RNA-seq data to G4 ChIP-seq peaks between untreated and Entinostat treated cells. Consistent with the outcome in untreated cells (Fig. 1d), we observed that genes with OQs in promoters that overlap with both a G4 ChIP-seq peak and an ATAC-seq peak have on average a significantly higher transcriptional output ($P<0.0001$) than promoters without a G4 ChIP-seq peak in Entinostat treated HaCaTs (Supplementary Fig. 8). Importantly, for promoters that showed a new or larger G4 ChIP-seq peak, but no significant change in their open chromatin state after Entinostat treatment, transcriptional output also increased in comparison to their untreated state ($P<0.0001$) (Fig. 2d). This indicates that there is a positive and dynamic relationship between G4 structure and transcriptional activity independent of the degree of chromatin accessibility (Fig. 2d), and further suggests that G4s epigenetically mark the genome, whereby four-stranded structure formation rather than the underlying sequence *per se* is directly linked to elevated transcription.

We next investigated how an altered cellular state results in chromatin, G4 and transcriptional changes by comparing normal human epidermal keratinocytes (NHEKs) and their spontaneously immortalized counterpart, the HaCaT cell line. Using G4 ChIP-seq, ATAC/FAIRE-seq (Fig. 3, Supplementary Fig. 9), we found that while NHEKs exhibit 85,668 more nucleosome-depleted regions than HaCaT cells (Supplementary Fig. 9), there are less ChIP-seq peaks in NHEKs as compared to HaCaTs (1,496 vs. 10,560) (Fig. 3b). The G4 ChIP-seq peaks identified in HaCaT cells (8,478, 80%), but absent from NHEKs, were found located entirely within nucleosome-depleted regions common to both cell lines (Fig. 3a), suggesting that additional mechanism(s), beyond having a nucleosome-depleted environment, control G4 formation in NHEKs relative to HaCaTs. Indeed, genes comprising promoter G4 ChIP-seq peaks with OQs that were present exclusively in HaCaTs showed an overall increase in transcription as compared to genes containing promoter G4 ChIP-seq peaks common to both cellular states (Fig. 3c, Supplementary Fig. 9).

It is noteworthy that we identified G4s in many cancer genes, for example *MYC*, *PTEN* and *KRAS*, in immortalized HaCaT cells but not in NHEKs (Fig. 3a), suggesting a link between increased proliferative capacity/immortalization[23] and G4s. We determined whether the increased presence of G4s in HaCaT cells versus NHEKs was also reflected in the G4 foci observed in nuclei visualized by BG4 immunofluorescence microscopy[13]. Consistent with our G4 ChIP-seq data, we found that HaCaT cells showed ~4-fold (20 ± 8 vs. 5 ± 2) more G4 foci than NHEKs (Fig. 3d, Supplementary Fig. 8). We have previously shown that the G4-selective small molecule pyridostatin (PDS) binds G4 structures, inducing a transcription- and replication-dependent DNA damage response, thus inhibiting cell growth[10,37]. We therefore investigated PDS sensitivity of HaCaTs vs. NHEKs. We found that HaCaTs are ~7-times more sensitive to growth inhibition by PDS as determined by GI_{50} (0.9 ± 0.3 vs. 6.5 ± 0.4 μM) values than NHEKs and is consistent with an increased number of G4s in HaCaTs (Fig. 3e). This increased abundance of G4s mirrors our previous observation that G4 structures can be more prevalent in cancer as compared to matched normal tissue[38] and suggests a potential rationale for selective cancer intervention by G4 targeting.

Here, we provide the first high-resolution genome-wide map of G4 structures in human chromatin. The number of G4 ChIP-seq sites (~10,000) is substantially lower than predicted by computation[26] or observed by G4-seq[16], and likely reflects the generally suppressive role of heterochromatin for G4 formation in human cells (e.g. nucleosome density), which may be different in other species[39]. More G4 ChIP-seq peaks are observed than BG4 IF foci[13] and is

explained by the higher sensitivity and resolution of the ChIP-seq method. We conclude that G4s show hallmarks of dynamic epigenetic features in chromatin primarily found in regulatory, nucleosome-depleted regions and correlated with genes showing elevated transcription (Fig. 4). While small-molecule G4 stabilization can promote a DNA damage response[10,40] and can cause transcriptional repression[2], our study suggests that endogenous G4 structures in promoters are ordinarily linked to elevated transcriptional activity. We have also discovered that the endogenous G4 landscape is dynamically altered depending on chromatin relaxation or cell status; and that G4s are particularly enriched in cancer-related genes and regions predisposed to amplification in cancer. Our study further illustrates the potential of the G4 structural motif as a *bona fide* target for disease, diagnosis and intervention.

URLs Reprints and permissions information are available at www.nature.com/reprints. A sample sheet describing the detailed experimental design is available at https://github.com/sblab-bioinformatics/dna-secondary-struct-chrom-lands. Details of data analysis have been deposited at https://github.com/sblab-bioinformatics/dna-secondary-struct-chrom-lands. Lists of oncogenes and tumor-suppressor genes were obtained respectively from the COSMIC database (http://cancer.sanger.ac.uk/census).[41] and tumor suppressor gene database (http://bioinfo.mc.vanderbilt.edu/TSGene/)[42]. Peak correlations were performed using the bedtool package (bedtools.readthedocs.org).

Accession codes The data reported in this paper are available at the NCBI's GEO repository, accession number GSE76688.

Acknowledgements The authors would like to thank the staff in the Genomic and Light Microscopy and Biorepository core facilities at Cancer Research UK Cambridge Institute. We are grateful to the European Molecular Biology Organization for funding R.H.H. with the EMBO Long-Term Fellowship. We acknowledge support from University of Cambridge and Cancer Research UK program. The Balasubramanian and Narita laboratories are supported by core funding from Cancer Research UK (C14303/A17197).

Author Contributions R.H.H. developed the G4 ChIP-seq method with help from S.L.. R.H.H. carried out all experiments except for the immunofluorescence microscopy and growth inhibition experimental work and analysis. K.Z., A.P. and M.D.A. carried out immunofluorescence microscopy experiments. M.D.A. performed growth inhibition experiments. R.H.H., D.B. and G.M. designed, implemented and performed the bioinformatic analysis. R.H.H., D.B., S.L., D.T. and S.B. designed epigenome experiments. R.H.H., K.Z., A.P., M.D.A. and M.N. designed

immunofluorescence experiments. J.P. analyzed and quantified colocalization immunofluorescence microscopy experiments. H.K. provided the antibodies α-H3K9me3 (Clone CMA304), α-H3K9me3 (clone CMA318) and α-RNA polymerase II carboxy terminal-domain (clone CMA601). All authors interpreted the results. R.H.H. wrote the manuscript with support and contributions from all authors.

References

1 Gray, L. T., Vallur, A. C., Eddy, J. & Maizels, N. G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat Chem Biol* **10**, 313-318 (2014).

2 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).

3 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).

4 Ribeyre, C. *et al.* The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* **5**, e1000475 (2009).

5 Cheung, I., Schertzer, M., Rose, A. & Lansdorp, P. M. Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nat Genet* **31**, 405-409 (2002).

6 Haeusler, A. R. *et al.* C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**, 195-200 (2014).

7 Maizels, N. G4-associated human diseases. *EMBO Rep* **16**, 910-922 (2015).

8 Wolfe, A. L. *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**, 65-70 (2014).

9 Rodriguez, R. & Miller, K. M. Unravelling the genomic targets of small molecules using high-throughput sequencing. *Nat Rev Genet* **15**, 783-796 (2014).

10 Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat Chem Biol* **8**, 301-310 (2012).

11 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

12 Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888-903 (2013).

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

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

15 Lam, E. Y., Beraldi, D., Tannahill, D. & Balasubramanian, S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat Commun* **4**, 1796 (2013).

16 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877-881 (2015).

17 Cayrou, C. *et al.* The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* **25**, 1873-1885 (2015).

18 Foulk, M. S., Urban, J. M., Casella, C. & Gerbi, S. A. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing

between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res* **25**, 725-735 (2015).

19 Hershman, S. G. *et al.* Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res* **36**, 144-156 (2008).

20 Law, M. J. *et al.* ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell* **143**, 367-378 (2010).

21 Paeschke, K., Capra, J. A. & Zakian, V. A. DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678-691 (2011).

22 Kanoh, Y. *et al.* Rif1 binds to G quadruplexes and suppresses replication over long distances. *Nat Struct Mol Biol* **22**, 889-897 (2015).

23 Boukamp, P. *et al.* Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. *J Cell Biol* **106**, 761-771 (1988).

24 Hogan, G. J., Lee, C. K. & Lieb, J. D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet* **2**, e158 (2006).

25 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

26 Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**, 2908-2916 (2005).

27 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).

28 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).

29 Raiber, E. A., Kranaster, R., Lam, E., Nikan, M. & Balasubramanian, S. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res* **40**, 1499-1508 (2012).

30 McDade, S. S. *et al.* Genome-wide analysis of p63 binding sites identifies AP-2 factors as co-regulators of epidermal differentiation. *Nucleic Acids Res* **40**, 7190-7206 (2012).

31 Johnson, J. E., Cao, K., Ryvkin, P., Wang, L. S. & Johnson, F. B. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. *Nucleic Acids Res* **38**, 1114-1122 (2010).

32 Sabouri, N., Capra, J. A. & Zakian, V. A. The essential Schizosaccharomyces pombe Pfh1 DNA helicase promotes fork movement past G-quadruplex motifs to prevent DNA damage. *BMC Biol* **12**, 101 (2014).

33 Balasubramanian, S., Hurley, L. H. & Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat Rev Drug Discov* **10**, 261-275 (2011).

34 Saito, A. *et al.* A synthetic inhibitor of histone deacetylase, MS-27-275, with marked in vivo antitumor activity against human tumors. *Proc Natl Acad Sci U S A* **96**, 4592-4597 (1999).

35 Tomazou, E. M. *et al.* Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. *Cell Rep* **10**, 1082-1095 (2015).

36 Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393 (2012).

37 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).

38 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).
39 Hoffmann, R. F. *et al.* Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res* **44**, 152-163 (2016).
40 Gomez, D. *et al.* Telomestatin-induced telomere uncapping is modulated by POT1 through G-overhang extension in HT1080 human tumor cells. *J Biol Chem* **281**, 38721-38729 (2006).

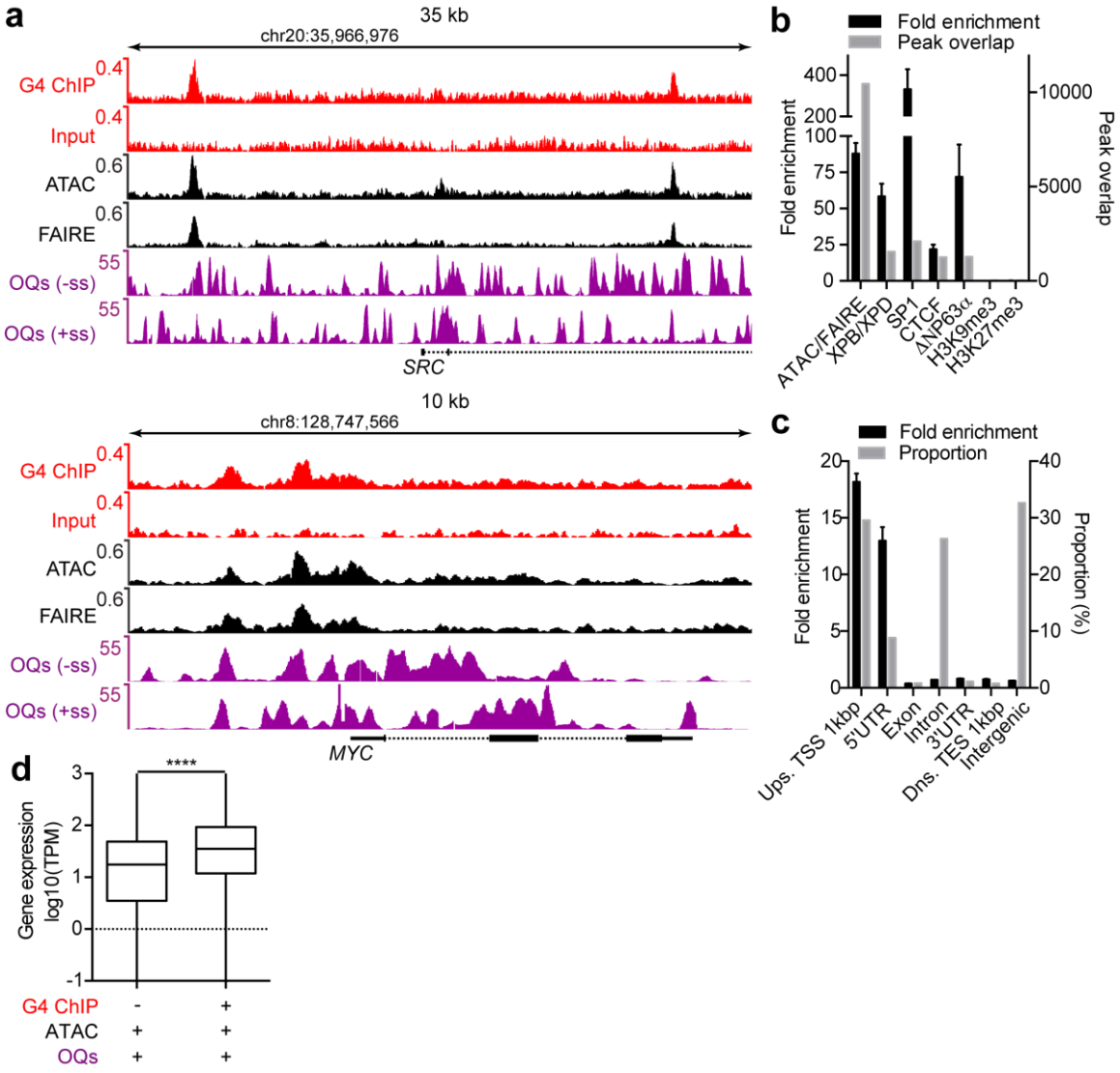
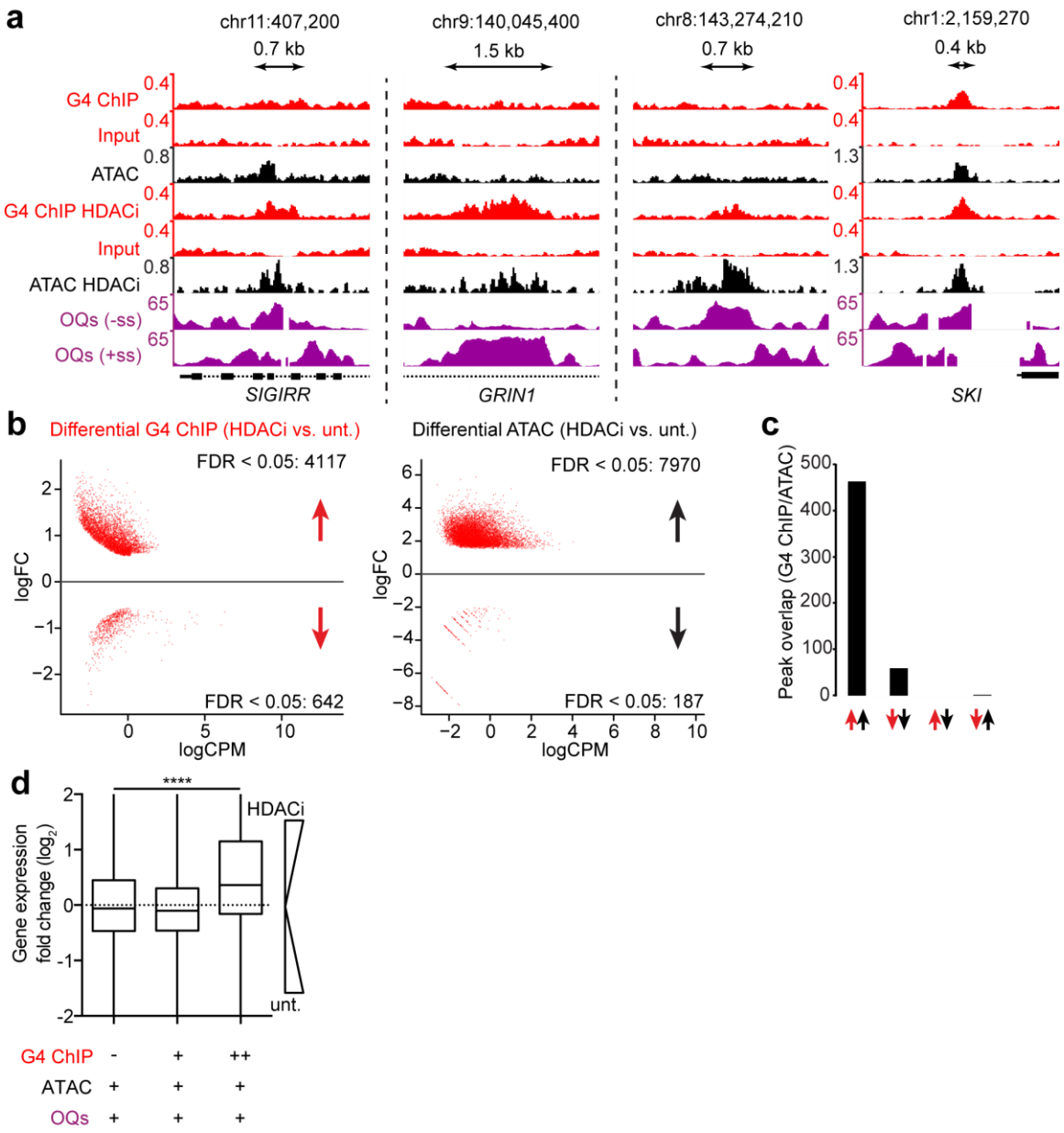


Figure 1 G4s are prevalent in regulatory chromatin regions. (**a**) Example genome browser screenshots for *SRC* and *MYC*. Tracks are shown for HaCaT G4 ChIP-seq (top, red) and control input below (red); regulatory chromatin sites by ATAC and FAIRE-seq (tracks 3 and 4, black) and G4-seq peaks (Pyridostatin derived OQs) (purple, % mismatches in reads aligned) on the

reverse (-ss) and forward strand (+ss)[16], respectively (tracks 5 and 6, purple). (**b**) Enrichment and peak overlap of HaCaT G4 ChIP-seq peaks with different ENCODE protein ChIP-seq data sets[43]; s.d. (N = 3) (**c**) Enrichment of HaCaT G4 ChIP-seq peaks relative to what is expected by randomly shuffling G4 ChIP-seq peaks in OQs and the absolute proportion of G4 ChIP-seq peaks across different genomic features (N = 3, error bars indicate standard deviation). (**d**) Distribution of mRNA levels (displayed in transcripts per million, log10 scale) are shown for promoter associated genes (4,345) that feature an ATAC-seq peak and an OQs signature in comparison to genes (4,522) that exhibit a promoter G4 ChIP peak, ATAC-seq and OQs feature in HaCaTs. **** indicates significance (P<0.0001; unpaired two-tailed T-test).



317

318 **Figure 2 Chromatin relaxation increases G4 prevalence in regulatory chromatin regions.**

319 (**a**) Genomic view of *SIGIRR*, *GRIN1* and a non-coding region (chr8:143,274,210) showing
320 overlap between new G4 sites and regulatory chromatin in Entinostat-treated cells. The *SKI*
321 promoter (right) exemplifies a gene not significantly changed in G4 ChIP-seq or nucleosome-
322 depletion. Top two tracks: untreated HaCaT cells with G4 ChIP-seq/input (red) and ATAC-seq
323 (black). Middle three tracks: Entinostat-treated HaCaT cells, G4 ChIP-seq/input HDACi (red),
324 ATAC-seq HDACi (black). Bottom two tracks- OQs (PDS-derived, purple), reverse (-ss) and
325 forward strand (+ss)[16], respectively. (**b**) Differential Binding Analysis (DBA) showing significant

(FDR <0.05) differences in G4 ChIP-seq and ATAC-seq peaks between Entinostat-treated versus untreated HaCaT cells. Red dots represent peaks where G4 ChIP peaks or nucleosome-depletion (ATAC) is significantly changed in Entinostat-treated compared to untreated HaCaT cells. Red and black arrows indicate increase or decrease of G4 formation and changes in nucleosome-depletion (ATAC-seq), respectively. (**c**) Peak overlap between increased and decreased G4 ChIP-seq peaks and open chromatin regions, derived from (**b**). (**d**) Differential gene expression for promoter-associated genes that i) **do not** contain high-confidence G4 ChIP-seq peaks, but have at least one OQs and ATAC-seq peak unaltered in size between conditions (1,734 promoters), ii) as in i) but contains at least one high-confidence G4 ChIP-seq peak shared between conditions and **do not** contain G4 ChIP-seq peaks significantly increased in size (3,627 promoters), or iii) as in i) but contains G4 ChIP-seq peaks significantly increased in size across conditions (373 promoters). (**** P<0.0001; ordinary one-way ANOVA).

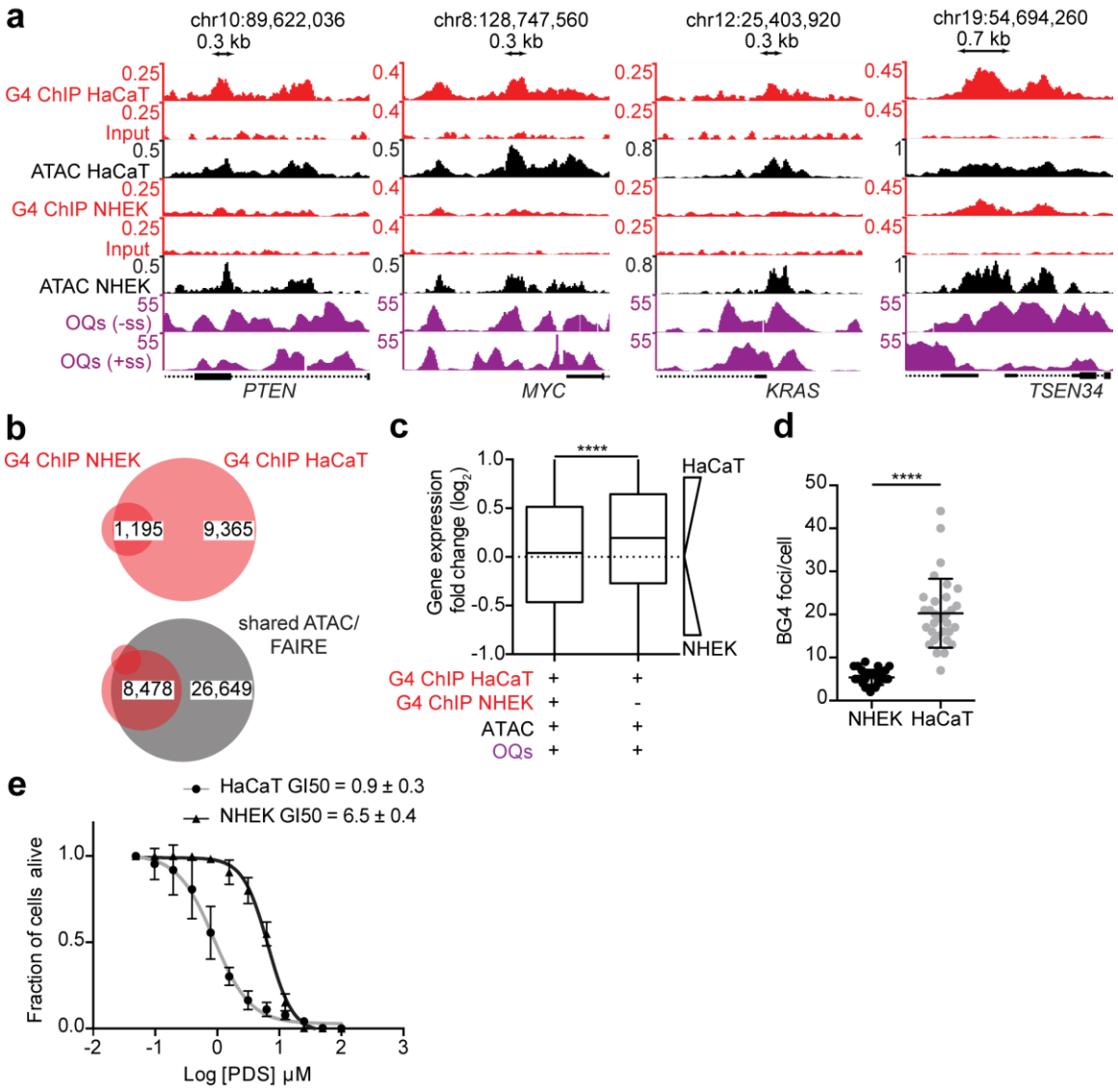


Figure 3 G4 prevalence is significantly increased in immortalized compared to normal human epidermal keratinocytes.

(**a**) Genomic view of *PTEN*, *MYC*, *KRAS*, and *TSEN34* in HaCaT and NHEK cells. Tracks are G4 ChIP-seq/input HaCaT (top two, red), ATAC-seq HaCaT (third, black), G4 ChIP-seq/input NHEK (fourth and fifth, red), ATAC-seq NHEK (sixth, black), OQs (PDS-derived, bottom two tracks in purple) on the reverse (-ss) and forward strand (+ss)[16], respectively. (**b**) Overlap of G4 ChIP-seq peaks between HaCaT and NHEK (top), and overlap between G4 ChIP-seq peaks and shared open chromatin regions (ATAC/FAIRE) (bottom); while both NHEKs and HaCaTs share G4 and active chromatin regions, NHEKs have ~7-fold less G4s. (**c**) mRNA levels for promoter-associated genes featuring i) a G4 ChIP and ATAC-seq peak and an OQ signature (503 genes)

common to NHEK and HaCaT cells and ii) a unique G4 ChIP/ATAC/OQs peak signature in HaCaT cells vs. NHEK (3,617 genes) (**** P<0.0001, unpaired two-tailed T-test). (**d**) Quantification of immunolocalization (see Supplementary Fig. 9) showing the average number of BG4 foci per cell increases in HaCaT vs. NHEK cells. 109 and 333 HaCaT nuclei and 186 and 326 NHEK nuclei were analyzed (N = 2). Error bars indicate standard deviation and (**** P<0.0001, unpaired two-tailed T-test). (**e**) Growth inhibition response curves for treatment with PDS (n = 6, error bars indicate standard deviation). The concentration to give 50 % growth inhibition (GI_{50}) is indicated for HaCaTs (light grey curve) and NHEKs (dark grey curve).

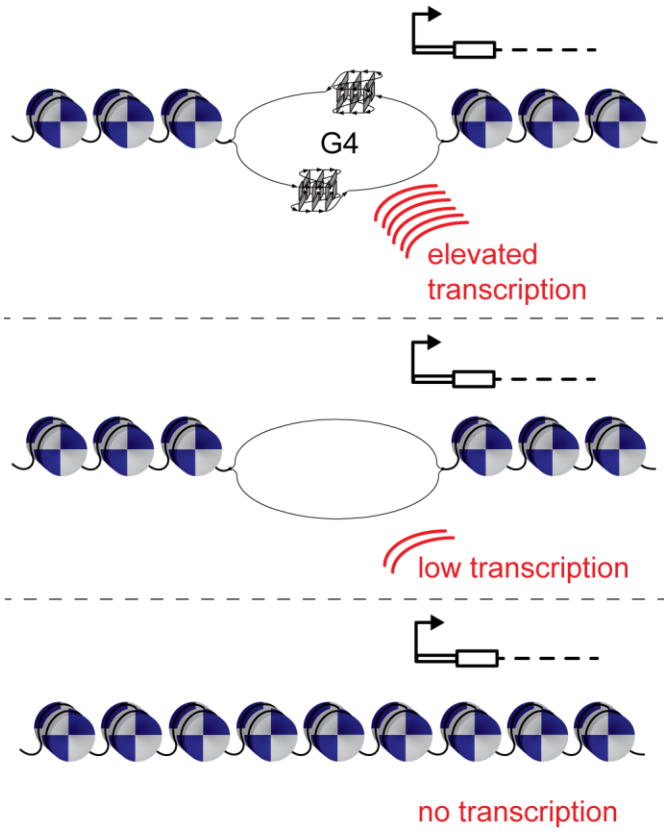


Figure 4 G4 DNA formation in chromatin. G4 DNA formation is highly dependent on chromatin structure and is frequently found in regulatory nucleosome-depleted regions in proximity to transcription start sites of genes that undergo elevated transcription.

Online Methods

Cell culture

HaCaT cells were kindly provided by Prof. Fiona Watt and cultured in Dulbecco's Modified Eagle Medium **(**Thermofisher, DMEM, cat. no. 41965-039) supplemented with 10% fetal bovine serum (FBS) (Thermofisher). U2OS cells were cultured in DMEM (Thermofisher, cat. no. 41966-029) supplemented with 10 % FBS. Normal human epidermal keratinocytes, pooled from multiple donors, were purchased from Thermofisher and cultured in EpiLife medium supplemented with human keratinocyte growth supplement (HKGS) (Thermofisher). Cell line genotypes were certified by the supplier and STR profiling. Cells lines were confirmed mycoplasma-free by RNA-capture ELISA.

G4 ChIP-seq protocol and library preparation

20 million cells were fixed in DMEM (cat no. 41965-039) containing 1 (v/v) % formaldehyde and 10 (v/v) % FBS for 15 min at room temperature (RT) followed by quenching with 0.13 M glycine at RT for 10 min. Cells were washed in ice-cold PBS and chromatin isolated and prepared as described by the manufacturer (Chromatrap). Chromatin was sonicated using a Bioruptor Plus (Diagenode). 250 µL of lysed nuclei suspension was sonicated for 25 cycles (30s on/60s off) in 1.5 mL TPX sonication tubes (Diagenode) to an average DNA size of 100–500 bp. For chromatin immunoprecipitation (ChIP), 2.5 µL chromatin was blocked at 16 °C for 20 min in 44.5 µL intracellular salt solution (25 mM HEPES pH 7.5, 10.5 mM NaCl, 110 mM KCl, 130 nM $CaCl_2$, 1 mM $MgCl_2$), containing 1 (w/v) % bovine serum albumin (BSA) (Sigma cat. no. B4287), referred as blocking buffer. Recombinant BG4 and control phage display antibody were expressed as described[14]. 200 ng of recombinant BG4 or control phage display antibody were then added and incubated for 1h at 16 °C. Meanwhile, 90 µL beads (SIGMA-ALDRICH Anti-FLAG® M2 Magnetic Beads, cat. no. M8823) were washed 3x with 900 µL blocking buffer and stored on ice in 900 µL blocking buffer (blocked bead solution). Next, 50 µL of blocked bead solution was added to the ChIP reaction (final volume 100 µL) and incubated for 1h at 16 °C with rotation. Beads were magnetically captured and the supernatant discarded and the beads washed three times in 200 µL wash (4 °C) buffer (10mM Tris pH 7.4, 100mM KCl, 0.1 (v/v) % Tween 20) with manual agitation. Beads were resuspended in 200 µL wash buffer and incubated on a rotating platform at 37 °C for 10 min. The warm wash procedure was repeated, and beads subjected to a final wash (4 °C) before removing the supernatant and re-suspension in 75 µL of elution buffer, containing 1x TE buffer with 20 µg proteinase K (Thermofisher). Elution was performed at 37 °C for 1h and at 65 °C for additional 2h. Beads were magnetically captured and

eluted DNA purified from supernatant using a MinElute kit (Qiagen). For each technical replicate, eluted DNA from four ChIP reactions were combined and the pool subjected to Nextera library preparation as described by the manufacturer (Illumina, cat. no. FC-121-1030). Three independent technical replicates were pulled-down and sequenced for each of the two biological replicates.

G4 ChIP-qPCR

Purified and sonicated DNA (as above) were used to quantify G4 enrichment via qPCR, using Fast SYBR PCR mix (Applied Biosystems, UK), with a BioRad CFX96 quantitative PCR machine. Cycling conditions were 95 °C for 20 s followed by 40 cycles of 3 s at 95°C and 30 s at 60°C. We employed primer pairs that target G4 ChIP positive and negative regions (Supplementary Table 3). Relative enrichments were derived with respect to their inputs and normalized to a G4-free enhancer region associated with the *ESR1* gene (Supplementary Table 3). RNase A has been shown to digest G4 RNA[13,44,45] and G4 DNA/RNA hybrids[45]. For DNase or RNase A treatments, 2 U of TURBOTM DNase (2 U/µL; Thermo Fisher Scientific) or 1µg of RNase A (1 mg/mL, Ambion® Thermo Fisher Scientific) were added to the ChIP blocking buffer chromatin mixture and chromatin digestion, instead of chromatin blocking, was performed at 37 ℃ for 20 min.

Epigenome mapping

FAIRE-seq was performed essentially as described, with minor modifications[46]. Cell lysis was performed using a PRECELLYS® 24 homogeniser and phenol-chloroform extraction was performed using Heavy Phase-Lock Gels (5-PRIME™). Transposase-accessible chromatin using sequencing (ATAC-seq) was performed essentially as described, with minor modifications[25]. Tagmented DNA samples were amplified (11 PCR cycles) using the Nextera index kit (Illumina, cat. no. FC-121-1011) and open chromatin fractions were size selected (190–300 bp) via 2 (w/v) % agarose E-Gel® SizeSelect™ gels (Themofisher). Total RNA for RNA-seq experiments was extracted using the RNeasy kit (Qiagen, cat. no. 74104), following the manufacturer's instructions. RNA-seq libraries were generated using the Illumina Truseq RNA HT (stranded mRNA) kit (cat. no. RS-122-2103). Overall, 18 G4 ChIP-seq, 18 ATAC-seq, 4 FAIRE-seq and 12 RNA-seq libraries were made (Supplementary Fig. 1). All epigenome libraries were sequenced in single-end or paired-end using 75 bp reads and the NextSeq500 platform.

Data analysis

Mapping, peak calling and peak processing. Raw fastq reads were trimmed to remove adapter contamination and aligned to the human reference genome version hg19 using cutadapt[47] and bwa[48], respectively. Reads with a mapping quality below 10 and marked as positional duplicates were excluded from further analysis. FAIRE-seq, ATAC-seq and G4-ChIP peaks were mapped using MACS2.0[49]. RNA-seq reads were aligned to the human reference genome with tophat2[50]. The Galaxy cistrome platform[51] was employed to process MACS2.0 called peak intervals and to generate high-confidence peak overlaps between biological replicates and area-proportional Venn diagrams. Find Individual Motif Occurrences (FIMO)[52] analysis was used to quantify the density of the Multiple Em for Motif Elicitation (MEME)[53] motifs in the high-confidence G4 ChIP-seq data (HaCaT).

Differential gene expression. Gene expression levels were quantified as transcripts per million (TPM). Differentially expressed genes between i) HaCaT treated with Entinostat and untreated HaCaT and ii) NHEK vs HaCaT were identified using with the Bioconductor package edgeR (FDR < 0.05; fold change > 1.5)[54].

G4 transcriptional analysis. Promoter transcription start site (TSS) coordinates, 1kb (+/-) from TSS, were generated for 22,483 genes using hg19. To reveal absolute gene expression values (TPM, transcripts per million) for promoters with and without a G4 ChIP-seq signature, we divided promoters into two categories: 1) promoters that overlap with at least one high-confidence ATAC-seq peak, at least one OQs (potassium- and PDS-induced) and that overlap with at least one high-confidence G4 ChIP-seq peak, and 2) promoter regions as in 1) that **do not** overlap with the high-confidence G4 ChIP-seq data set of untreated HaCaTs (command intersect, subtract bedtools package).

Differential gene expression levels for promoter-associated genes were prepared into 3 subsets as follows: Promoters that 1) **do not** overlap with G4 ChIP-seq peaks, but have at least one OQs and ATAC-seq peak unaltered in size (log2 fold change = -0.6-to-0.6, FDR<0.05 differential binding analysis[36]) between untreated and Entinostat treated HaCaTs, 2) as in 1) but **do not** overlap with G4 ChIP-seq peaks significantly increased in size for Entinostat treated vs. untreated HaCaTs and overlap with at least one high-confidence G4 ChIP-seq peak shared between untreated and Entinostat treated HaCaTs, or 3) as in 1) but overlap with G4 ChIP-seq peaks significantly increased in size (log2 fold change = >0.6, FDR<0.05 differential binding analysis[36]) for Entinostat treated vs. untreated HaCaTs (command intersect, subtract bedtools package). **** indicates significance (P<0.0001; ordinary one-way ANOVA).

Epigenome enrichment analysis. ChIP-seq data from the ENCODE project was retrieved from the NCBI's GEO repository as follows: XPB/XPD (GSE44849); SP1 (ENCSR991ELG; ENCSR000BJX; ENCSR000BHK); H3K9me3 (ENCSR000EYF; ENCSR000ARN; ENCSR000APE); H3K27me3 (ENCSR000EWB; ENCSR000DWU; ENCSR000DUE); CTCF (ENCSR000DWX; ENCSR000EGM; ENCSR000DUG); DeltaNp63 (GSE32061). Overlap between the high confidence HaCaT G4 ChIP-seq peak file and the ENCODE ChIP-seq data sets was tested using the Galaxy cistrome platform[51]. Common high confidence ENCODE ChIP-seq peak files across the three different cell lines were calculated, if applicable (see GEO accession numbers). ENCODE ChIP-seq peak files were randomly shuffled (N = 6) across the genome and potential overlaps tested with the G4 ChIP-seq peak file (command shuffleBed of the bedtools package). Enrichments between G4 ChIP-seq and ENCODE ChIP-seq data sets were calculated from the ratio of the direct overlaps with the randomly shuffled overlaps.

G4 ChIP-seq peak annotation and enrichment analysis. PAVIS[55] was used to annotate the HaCaT G4 ChIP-seq peaks. The G4 ChIP-seq file was randomly shuffled across a file containing all OQs (potassium and PDS induced) (command shuffleBed of the bedtools package) three independent times and their genomic annotations retrieved via PAVIS, respectively. To calculate enrichments of annotated features, ratios of the annotated G4 ChIP-seq proportions were performed with the proportions of the randomly shuffled features (N = 3).

G4 motif and enrichment analysis. Sequences within peak regions were analyzed by regular expression matching and assigned to one of the following classes: Loop size 1–3, 4–5 and 6–7: sequences with at least one loop of the respective length; long loop: sequences with a G4 with any loop of length >7 (up to 12 for any loop and 21 for the middle loop); simple bulge: sequences with a G4 with a bulge of 1–7 bases in one G-run or multiple 1-base bulges; 2-tetrads / Complex bulge: sequences with a G4s with two G-bases per G-run or several bulges of 1–5 bases; and other: other G4-types that do not fall into the former categories. Sequences in each peak were assigned to these classes in a hierarchical fashion: for instance, if a sequence matched both the Loop 1–3 and Long loop categories, it was assigned to the former most category, in this case Loop 1–3. The fold enrichment analysis was calculated by comparing actual counts for each class within peak regions to counts of the same peak regions after random reshuffling throughout the genome using the bedtools shuffle command. Results are shown as the average of 10 randomizations, and the fold enrichment bar plot displays the ratio of real counts divided by average random counts. For fold enrichment analysis of G4 motifs compared to similar motifs with lower G4 forming potential, we compared respectively: 1) G4 with 4 G-runs of at least 3 Gs

each and loop size 1–7 ($G_{3+}L_{1–7}$) compared to a similar motif with 3 G-runs of at least 3 Gs each and loop size 1–7; 2) as 1), but with loop length 1–12 ($G_{3+}L_{1–12}$); 3) as 1), but with G-runs of 2 Gs each, i.e. 2-tetrads ($G_2L_{1–7}$); 4) canonical G-quadruplex motifs as in 1) ($G_{3+}L_{1–7}$) but with a single bulge of 1–7 bases ($GGH_{1–7}G$, with H=A,T or C) in only one G-run compared to a similar motif with a single G ($H_{1–2}GH_{1–2}$) instead of the bulge motif; 5) same G-quadruplex as in 3) ($G_2L_{1–7}$) compared to a similar motif with one G-run having a single G ($H_{1–2}GH_{1–2}$).

Cancer-related gene analysis. After remapping to hg19, 555 oncogenes and 1211 tumor suppressors (Supplementary Table 1) were considered. The occurrence of ChIP-seq peaks in each gene was determined and compared to the occurrence within not annotated as cancer genes nor tumor-suppressors. To account for different region size, the peak density was calculated as number of peaks per Megabase (Mb). Peaks occurring within the gene body (i.e., exons + introns) and in proximal promoter regions (i.e., 3000 bases upstream of the transcription start site) were counted separately. P-values for statistical testing were calculated with the non-parametric Wilcoxon rank sum test (implemented as wilcox.test() function in the R software).

SCNAs analysis. Somatic copy number alteration (SCNA) regions associated with cancer (*n=108*), either amplifications (*n=54*) or deletions (*n=54*), (i.e. copy number gains or losses of any length and amplitude), were obtained from the Cancer Genome Atlas Pan-Cancer dataset[56]. ChIP-seq peak abundance in each region was compared to the occurrence of randomly reshuffled peak regions (5 different randomization) in a similar manner to the G4 motif enrichment analysis. To account for different region sizes, peak density was calculated as the number of peaks per megabase (Mb). For visual comparison, we also determined the average genome-wide peak density per Mb (*avg_density = 3.91*), calculated as total number of peaks divided by the effective genome size ($2.7x10^9$ for *hg19*), multiplied by 1,000,000 bp (Supplementary Table 2).

Immunofluorescence microscopy

BG4 immunostaining and fluorescence microscopy for the HaCaT and NHEK cells were performed as previously described[13]. All secondary or tertiary antibodies were obtained from Thermofisher unless otherwise stated. In brief, cells were fixed with 2 (v/v) % formaldehyde in PBS, permeabilized with 0.1 (v/v) % Triton X100 in PBS, and blocked with 2 (w/v) % Marvel milk (Premier Foods plc) in PBS. After blocking, cells were incubated with BG4 followed by incubation with secondary Rabbit α-FLAG (cat. no. 2368, Cell Signaling Technology) and tertiary goat α-rabbit conjugated with Alexa-594 (cat. no. A11037) at 37 °C in a humid chamber for 1h each. Coverslips were mounted on Prolong Gold/DAPI Antifade (Thermofisher). Digital

images were recorded using a DP70 camera (Olympus) on an Axioskop 2 Plus microscope (Zeiss) and analyzed with Fiji[57]. 200–300 nuclei were counted per condition and standard deviations calculated from two biological replicates. Frequency distribution graphs were plotted using GraphPad Prism (GraphPad Software Inc.).

For BG4 or 1H6 colocalization studies with H3K4me3, H3K9me3 and RNA polymerase II, HaCaT cells were grown to 50% confluency on # 1.5 glass coverslips. Cells were fixed and permeabilized as above, treated with 50 µg RNase A for 30 min at 37 °C and incubated in blocking buffer for 1h at 37 °C (PBS/ 0.5% normal goat serum, 0.1% Tween 20). For BG4 studies, coverslips were incubated with BG4 (10 nM) for 1h at 37 °C, washed three times for 5 min with PBST and incubated 37 °C for 1h with rabbit α-FLAG 1:800 (cat. no. 2368, Cell Signalling Technology) and mouse α-H3K9me3 1:200 (Clone CMA304)[58] or mouse α-H3K9me3 1:200 (clone CMA318)[59] or mouse α-RNA polymerase II carboxy terminal-domain (clone CMA601) 1:200[60]. For 1H6 studies, coverslips were incubated with mouse α-quadruplex DNA (cat. no. Ab00389-1.1, Absolute Antibody) and rabbit α-H3K9me3 1:200 (cat. no. 9751, Cell Signaling Technology) or rabbit α-H3K4 me3 1:500 (cat.no. 07-523, Upstate) for 37 °C for 1h. Subsequently, all coverslips were washed as described and incubated for 30 min at 37 °C with the following Alexa Fluor conjugated antibodies: goat α-rabbit Alexa-488 (cat. no. A-11034) and goat α-mouse Alexa-555 (cat. no. A-21424) for BG4 studies and goat α-mouse Alexa-488 (cat. no. A-11001) and goat α-rabbit Alexa-555 (cat. no. A-21429) for 1H6 studies. DAPI counterstain was included in the final antibody incubation. Following washing, all coverslips were mounted onto Superfrost Plus slides (cat. no. 4951PLUS, Thermofisher) with Vectashield antifade mounting media (cat. No. H-1000, Vector Laboratories Ltd.). Three biological replicates were performed.

Confocal z-stack images (15 steps) were acquired using a Leica TCS SP8 microscope with a HC PL APO CS2 1.4NA 100X oil objective (Leica Microsystems), in bidirectional mode, at a scan speed of 400 Hz and sampling rate of 0.06 µm x 0.06 µm x 280 µm. The 405 nm diode laser was used to excite the DAPI channel (at 405 nm) and the white light pulsed laser (SuperK EXTREME, NKT Photonics) was used to excite secondary antibody fluorophores (at 555 nm and 488 nm). Fluorescence detection was performed in sequential acquisition mode with hybrid detectors (Leica HyD Photon Counter) at the collection wavelength ranges of 490–535 nm, 565–630 nm and 410-485 nm for Alexa Fluor 488, 555 and DAPI respectively. The pinhole was set to one Airy unit and laser power and gain settings were consistent between replicates. Five to six Z-

stacks were obtained per replica representing 60–160 cells each. Representative images were processed using Image J[61]/Fiji and Photoshop (Adobe Systems Inc) software and assembled using Adobe Illustrator CS4.

Colocalization analysis was performed using an automated workflow in MATLAB (R2015b, Mathworks Inc.) with a link to Image J through MIJ[62]. First, the nuclear region was isolated by blurring the DAPI channel with a 3D Gaussian (radius 0.2µm) and Otsu thresholding (stack histogram). Noise reduction of the signal of each colocalization channel was performed using the PureDenoise plugin (ImageJ)[63] and isolated by removing background with a rolling ball approach (radius 2µm) and Otsu thresholding. The corrected first Mander's overlap coefficient ($M1_{diff}$) corresponds to the difference between the percentage of the G4 antibody (BG4 or 1H6) signal that colocalizes with euchromatin, heterochromatin or gene promotor signal; the expected value for randomly distributed signal was calculated on the isolated signal. Finally, a two-way t-test was performed for each condition across the mean values of $M1_{diff}$ across the three biological replicates[64], and corrected for multiple hypothesis testing using the Bonforroni correction (n=5). Graphs were created using GraphPad Prism version 6.00.

Small molecule treatment and growth inhibition analysis

HaCaT cells were treated with 2 µM Entinostat (E5477-5 mg, Cambridge Bioscience) for 48h at 30% confluency in 150 mm tissue culture dishes prior to G4 ChIP fixation procedure or ATAC-seq preparation, or in 6-well plates to generate four technical replicates for RNA-seq preparation. The concentration to inhibit by 50% cell growth (GI_{50}) induced by PDS was quantified using the end point Cell Titre Glo assay (Promega). Briefly, 4000 HaCaT and NHEK cells were seeded in a 96-well white plates (Nunc) for 18h prior to small molecule treatment. Cells were then treated with PDS doses ranging from 50 to 0.012 µM for 72h. Cells were then incubated with Cell Titer Glow for 1h at room temperature and luminescence was measured using a PHERAstar FS microplate reader (BMG LabTech). Cell survival curves were plotted and GI_{50} values were calculated using GraphPad Prism (GraphPad Software Inc.). Error bars represent variability within 6 replicates.

Online Methods-only references

41 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811 (2015).

42 Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res* (2015).

43 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

44 Biffi, G., Di Antonio, M., Tannahill, D. & Balasubramanian, S. Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat Chem* **6**, 75-80 (2014).

45 Wanrooij, P. H. *et al.* A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. *Nucleic Acids Res* **40**, 10334-10344 (2012).

46 Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* **7**, 256-267 (2012).

47 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).

48 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

49 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

50 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

51 Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-1455 (2005).

52 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

53 Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697 (2011).

54 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

55 Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D. & Li, L. PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics* **29**, 3097-3099 (2013).

56 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140 (2013).

57 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).

58 Kimura, H., Hayashi-Takanaka, Y., Goto, Y., Takizawa, N. & Nozaki, N. The organization of histone H3 modifications as revealed by a panel of specific monoclonal antibodies. *Cell Struct Funct* **33**, 61-73 (2008).

59 Hayashi-Takanaka, Y. *et al.* Tracking epigenetic histone modifications in single cells using Fab-based live endogenous modification labeling. *Nucleic Acids Res* **39**, 6475-6488 (2011).

60 Stasevich, T. J. *et al.* Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* **516**, 272-275 (2014).

61 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-675 (2012).

62 Sage, D., Prodanov, D., Tinevez, J.-Y., Schindelin, J. . in *ImageJ User \& Developer Conference* 24-26 (2012).

63 Luisier, F., Vonesch, C., Blu, T. & Unser, M. Fast Interscale Wavelet Denoising of Poisson-Corrupted Images. *Signal Processing* **90**, 415-427 (2010).

64 McDonald, J. H. & Dunn, K. W. Statistical tests for measures of colocalization in biological microscopy. *J Microsc* **252**, 295-302 (2013).

639

640 The authors declare no competing financial interests.