ANALYSIS

nature structural & molecular biology

DNA secondary structures and epigenetic determinants of cancer genome evolution

Subhajyoti De[1,2] & Franziska Michor[1,2]

An unstable genome is a hallmark of many cancers. It is unclear, however, whether some mutagenic features driving somatic alterations in cancer are encoded in the genome sequence and whether they can operate in a tissue-specific manner. We performed a genome-wide analysis of 663,446 DNA breakpoints associated with somatic copy-number alterations (SCNAs) from 2,792 cancer samples classified into 26 cancer types. Many SCNA breakpoints are spatially clustered in cancer genomes. We observed a significant enrichment for G-quadruplex sequences (G4s) in the vicinity of SCNA breakpoints and established that SCNAs show a strand bias consistent with G4-mediated structural alterations. Notably, abnormal hypomethylation near G4s-rich regions is a common signature for many SCNA breakpoint hotspots. We propose a mechanistic hypothesis that abnormal hypomethylation in genomic regions enriched for G4s acts as a mutagenic factor driving tissue-specific mutational landscapes in cancer.

Loss of genomic integrity is a common hallmark of cancer genomes[1]. Recent technological advances have led to several large-scale cancer genome profiling studies[2–5] that have identified genome-wide patterns of alterations in many cancer samples. Notably, DNA breakpoints in cancer genomes, and also in the genomes of apparently healthy subjects, are distributed nonrandomly[2,5–7], suggesting that some regions within the human genome—so-called breakpoint hotspots—are exquisitely prone to rearrangement of genetic material. Some of these regions are common across many cancer types, whereas others are specific to particular types, indicating that genomic instability may manifest itself differentially in neoplasms of diverse origin.

Many exogenous factors (such as nicotine exposure in lung cancer) and endogenous factors (such as repeat elements) as well as molecular mechanisms can cause double strand breaks and erroneous DNA repair, leading to genomic alterations[1,8–10]. Under certain circumstances, DNA can adopt non-B conformations, and recently two such secondary structures (H-DNA and Z-DNA) were shown to contribute to DNA damage[11–13]. Guanine-rich sequences ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$), which are frequent in the human genome, can adopt four-stranded structures called G-quadruplexes (G4) both *in vivo* and *in vitro*[14–16]. G4 structures obstruct the movement of DNA polymerase[17], thereby increasing the risk of DNA breakage or nonallelic homologous recombination. Indeed, G4 structures have been implicated in germline deletion[18,19] and recombination[20] events. However, the role of G4 structures in genomic instability in cancer has so far not been systematically investigated.

In addition to genetic factors, various epigenetic factors are also associated with genomic instability both during the somatic evolution of cancer[21,22] and in germline evolution during speciation[23]. Moreover, epigenetic patterns differ between cell types and thus possess the potential to generate tissue-specific patterns of alterations. Selective epigenetic states such as CpG methylation interact with G4 (refs. 24–26) and other non–B-DNA structures[27,28], potentially interfering with their formation and stability. The D4Z4 region, for instance, which is hypomethylated in some cancer types and hypermethylated in others, contains a subregion that is resistant to hypermethylation and harbors G4s motifs[29]. Furthermore, the CpG dinucleotide frequently resides within G4s, whose CpG methylation is usually low—especially at gene promoters, exons and untranslated regions[30]. These findings raise the possibility that the mutagenic potential of DNA secondary structures may be modulated by epigenetic states.

Here we set out to systematically investigate the role of DNA secondary structures in genomic instability in cancer. We integrated published data on genomic alterations from over 2,700 cancer samples, as well as potentially G-quadruplex–forming sequences (PG4s) and DNA methylation. We propose that hypomethylation and G4 structures together could have a causal role in genomic instability in cancer, thus representing one of the mechanistic bases for tissue-specific mutational landscapes of cancers.

RESULTS

DNA breakpoints in cancer are often clustered in hotspots

We obtained data for 663,446 SCNA breakpoints from Beroukhim *et al.*[2]. Although the breakpoints of some SCNAs occur adjacent to known oncogenes and tumor suppressor genes, some SCNAs span tens of kilobases containing multiple genes or gene desert regions. As examples, **Figure 1** shows the frequency distributions of SCNA breakpoints around *EGFR* (**Fig. 1a**), a gene commonly mutated in many cancer types[2], and *PAX5* (**Fig. 1b**), which is altered primarily in acute lymphoblastic leukemia[2,31]. In some cases, the SCNAs may include additional previously undescribed target genes or functional elements, which are important for tumorigenesis or the development

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to F.M. (michor@jimmy.harvard.edu).

Received 6 December 2010; accepted 4 May 2011; published online 3 July 2011; doi:10.1038/nsmb.2089

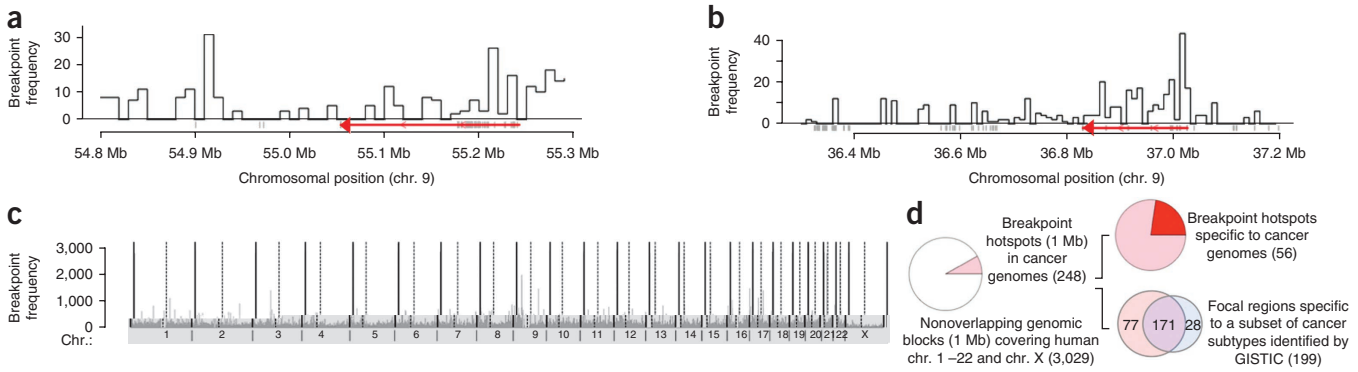

Figure 1 Spatial distribution of breakpoint hotspots in cancer genomes and genomes of healthy human subjects. (**a**) SCNA breakpoints can occur at high frequencies tens of kilobases away from *EGFR* (a known cancer gene shared across multiple cancer subtypes), shown in red. The direction of the red arrow shows the direction of transcription of *EGFR*. (**b**) SCNA breakpoints can occur at high frequency tens of kilobases away from *PAX5* (a known cancer gene specific to acute lymphoblastic leukemia), shown in red. The red arrow shows the direction of transcription of *PAX5*. (**c**) SCNA breakpoint densities calculated over 1-Mb nonoverlapping genomic blocks across the human genome. Dotted vertical lines mark centromeres. (**d**) Summary statistics for SCNA breakpoint hotspots. Frequencies are shown in parentheses.

of a precancerous state. However, another and not mutually exclusive scenario is that some regions in the genome are particularly prone to rearrangement of genetic material, leading to the presence of inherent genomic instability in one or more tissue types. In order to investigate the genome-wide distribution of breakpoints, we first divided the cancer genomes into 1-Mb nonoverlapping blocks and determined the number of SCNA breakpoints within each block. We found that 248 of the 3,029 genomic blocks, covering almost 8% of the human genome, were significantly enriched (FDR-corrected $P < 0.05$) for SCNA breakpoints in cancer (**Fig. 1c**; see Online Methods). We dubbed these regions breakpoint hotspots. Using cancer type–specific analyses, Beroukhim *et al.*[2] identified 199 frequently altered regions across 3,131 cancer samples, 177 of which shared their boundaries with the breakpoint hotspots we identified (**Fig. 1d**). Using data from three additional cancer genomes and three personal genomes, we found that many breakpoint hotspots were shared across samples, suggesting that they are perhaps inherently unstable during both somatic and germline evolution (**Supplementary Methods** and **Supplementary Table 1**). The observation that the SCNA breakpoints are organized in hotspots led us to investigate whether some genomic properties of those regions drive their instability.

Breakpoint hotspots are associated with PG4s

To assess whether G4 structures are associated with DNA breakpoints in cancer, we overlaid information about PG4s[32–34] with the SCNA breakpoint data and analyzed the joint distribution of PG4s and SCNA breakpoint frequencies within 1-Mb nonoverlapping windows tiling the genome. We found that the breakpoint hotspots were significantly enriched for PG4s (FDR-corrected $P = 9.19 \times 10^{-6}$, Mann-Whitney test; **Table 1**, **Supplementary Methods** and **Supplementary Table 2**). To evaluate whether this association is independent of other covariates, we overlaid information about other factors, such as repeat sequences[35], recombination frequency[35] and fragile sites[36]. We found that the breakpoint hotspots were also enriched for simple repeats, Alu repeats and CR1 repeats[35] (FDR-corrected $P < 0.005$, Mann-Whitney test; **Table 1**). We also observed that these sites were moderately enriched for sites of frequent recombination (recombination hotspots)[35] and common fragile sites[36] (FDR-corrected $P < 0.05$, Mann-Whitney test; **Table 1**). For two of the three personal genomes we analyzed, comprehensive structural variation data was available. When focusing on cancer-only breakpoint hotspots—those for which no structural variation was present in those two personal genomes—we found a significant enrichment for PG4s and a moderate depletion of repeats (**Table 1** and **Supplementary Methods**).

Because the above genomic features are functionally interdependent, establishing a primary association is challenging. For instance, recombination hotspots can form G4 structures[37] and often overlap with cancer breakpoints. Further, recombination hotspots, G4 structures and fragile sites are enriched for specific repeat sequences[36,38]; such repetitive elements show relatively low levels of evolutionary conservation[35]. We found that the variation in PG4s explains most of the variation in the density of cancer-specific breakpoint hotspots, and that the association between PG4s and breakpoint density exists even after controlling for other genomic features, such as meiotic recombination rate (**Supplementary Methods** and **Supplementary Table 3**).

Table 1 Association between genomic features and breakpoint hotspots

	FDR-corrected *P* value	
Genomic features	SCNA breakpoint hotspots in cancer[a]	SCNA breakpoint hotspots occurring only in cancer genomes[a]
Sequence features		
Common repeats	3.66×10^{-3}	1.50×10^{-2}
Alu elements	6.88×10^{-2}	
CR1 elements	8.96×10^{-3}	
L1 elements	3.19×10^{-3}	9.08×10^{-3}
L2 elements	4.54×10^{-3}	6.95×10^{-2}
Secondary structure features		
G4 secondary structures	9.19×10^{-6}	4.35×10^{-3}
DNA breakage and recombination features		
Fragile sites	3.52×10^{-2}	
Meiotic recombination rate	1.87×10^{-2}	
Evolutionary features		
28 way most evolutionarily conserved elements	2.43×10^{-2}	

[a]The second column represents the statistical significance corresponding to all breakpoint hotspots found in cancer genomes, while the third column represents the statistical significance for those breakpoint hotspots that occur in cancer genomes but not in the three personal genomes analyzed.

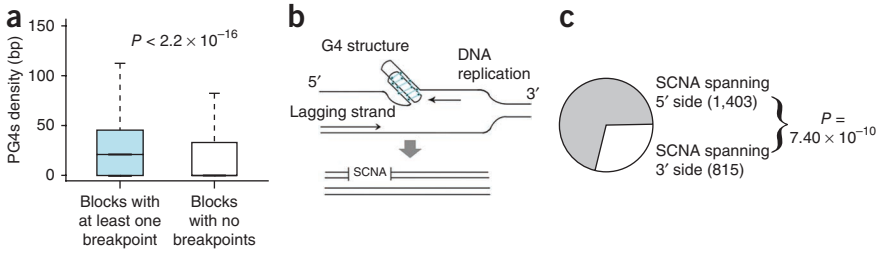

Figure 2 Association between G-quadruplex–forming sequences and breakpoint hotspots. (**a**) The distribution of the density (bp) of PG4s in 10-kb genomic blocks that have at least one SCNA breakpoint in cancer is markedly higher than the distribution of PG4s in those genomic blocks that harbor no breakpoints. The whiskers of the box plots represent the range of the PG4s density for the respective groups. (**b**) A schematic representation of DNA replication near a G4 structure and generation of an SCNA. Arrows indicate the direction of motion of the DNA polymerase. Only the leading strand obstructs the motion of the DNA polymerase and therefore SCNAs are more likely to occur at the 5′ side of G4 structures. (**c**) Cancer SCNAs with at least two PG4s within 10 kb are significantly likely to occur at the 5′ side of the G4 structures, an observation that is consistent with the hypothesis that these structures inhibit the action of DNA polymerase. Frequencies are shown within parentheses. The pattern is independent of the choice of parameters (see **Supplementary Table 5**).

We then performed our analysis at a higher resolution: we identified 10-kb windows centering on DNA breakpoints for each SCNA, and for each such window, we counted the total number of base pairs belonging to PG4s. When the windows from multiple breakpoints partially overlapped, we fused them and determined whether these regions were enriched for PG4s relative to the genome-wide distribution of such sequences. Indeed, the vicinity of SCNA breakpoints was significantly enriched for PG4s compared to the numbers expected by chance (**Fig. 2a**; $P < 2.2 \times 10^{-16}$, Mann-Whitney test). We found similar results at the resolution of 20 kb, 50 kb and 100 kb (**Supplementary Methods** and **Supplementary Table 4**). The association existed even after we controlled for SNP density on the Affymetrix chip and when we excluded centromeric and telomeric regions (**Supplementary Methods**). These findings suggest that the association between PG4s and SCNA breakpoints is probably genuine.

G4 structures are strand specific—the G-rich DNA strand forms a G4 structure that can obstruct the movement of the DNA polymerase[17,20] and cause mutagenic events[18–20] (**Fig. 2b**). Therefore, G4-mediated deletion and duplication events occur predominantly toward the 5′ direction of the G4 sequence[18–20]. Although the C-rich strand can potentially form an i-motif structure[39], a mutagenic potential of this structure has not been demonstrated. We therefore tested whether the SCNA events in cancer associated with G4 structures also show a strand bias. We identified SCNA breakpoints that had at least two PG4s within a 10-kb window and found that in more than two-thirds of these cases, the breakpoints resided on the same strand. For these cases, we observed a significant enrichment (**Fig. 2c**; $P = 7.40 \times 10^{-10}$, binomial test) for structural alterations to extend to the 5′ direction relative to that expected by chance. This finding was independent of how the enrichment of G4 structures on one strand was determined (**Supplementary Methods** and **Supplementary Table 5**). We obtained similar results at the resolutions of 20 kb and 50 kb, but the statistical significance decayed quickly for larger window sizes, suggesting that the effect is local (**Supplementary Methods** and **Supplementary Table 5**). Our observations point toward a causal role of PG4s in the generation of structural alterations in cancer.

G4 structure formation is facilitated by negative DNA supercoiling[25,40], which occurs not only during DNA replication[41,42] but also during repair and transcription[43]. Although the effects of replication are genome wide, transcription-associated events are likely to be localized to the neighborhood of transcribed regions. Overlaying transcription start sites (TSS), PG4s and SCNA breakpoints onto the human genome, we found that the 1-Mb genomic blocks with above-median PG4s density and above-median TSS density had significantly higher SCNA breakpoint densities than the remaining blocks ($P = 1.727 \times 10^{-15}$, Mann-Whitney test). We obtained similar results using different block sizes and cutoffs for PG4s and TSS densities (**Supplementary Methods** and **Supplementary Table 6**). Thus, gene promoters with high PG4s density are at an increased risk of DNA breakage in cancer.

G4-dense breakpoint hotspots are hypomethylated in cancer

Because the genomic sequence is the same in all cells within an individual, one might expect that PG4s-driven genomic aberrations would be recurrent in both normal and cancer cells and would be similar across all tissue types. In contrast, the mutational patterns observed in cancer genomes differ between cancer types—as do the patterns of epigenetic states such as DNA methylation[44]. In normal tissues, the genome is usually hypermethylated and does not show genomic instability, whereas genome-wide hypomethylation is a hallmark of many cancer types[21,22]. Notably, almost 50% of PG4s motifs contain CpG dinucleotides, and for a majority of those cases, the guanine participates in G-quadruplex formation (**Supplementary Methods**). PG4s motifs show depletion of CpG methylation and nucleosome occupancy[26,30], and DNA methylation patterns have a role in the stability of other noncanonical DNA structures such as Z-DNA and H-DNA[27,28]. Furthermore, chemical modifications such as O^6-methylguanine inhibit G4 structure formation[45]. Therefore, we investigated the patterns of DNA methylation in several normal and cancer tissues in the context of PG4s and DNA breakpoints.

We obtained methylation data for several healthy colon, brain, liver and spleen samples and DNA methylation data for 13 pairs of colorectal cancer samples and their matched normal colonic mucosa[44]. We first analyzed the patterns of hypomethylation in the normal brain, liver and spleen samples. We found that in those tissues, regions of hypomethylation were in general depleted for PG4s relative to their genome-wide distribution ($P = 4.16 \times 10^{-5}$, Kolmogorov-Smirnov test). Moreover, genomic blocks that harbor an above-median PG4s density were significantly more hypermethylated ($P < 2.2 \times 10^{-16}$, Mann-Whitney test) than those blocks that have below-median PG4s density in all three normal tissue types. Taken together, our data indicate that extensive hypomethylation and high PG4s density rarely co-occur in normal tissues.

We then overlaid differential methylation patterns, PG4s and cancer breakpoint densities for 13 colorectal cancer samples, and we found that sites of acute hypomethylation and high PG4s density often overlap with breakpoint hotspots (**Fig. 3a**). Furthermore, sites with both above-average PG4s density and differential hypomethylation harbored significantly more breakpoints than would be expected from the genome-wide distribution (**Fig. 3b**; $P = 1.55 \times 10^{-8}$, Mann-Whitney test). Our observation was independent of the threshold for hypomethylation and PG4s density (**Table 2**) and remained

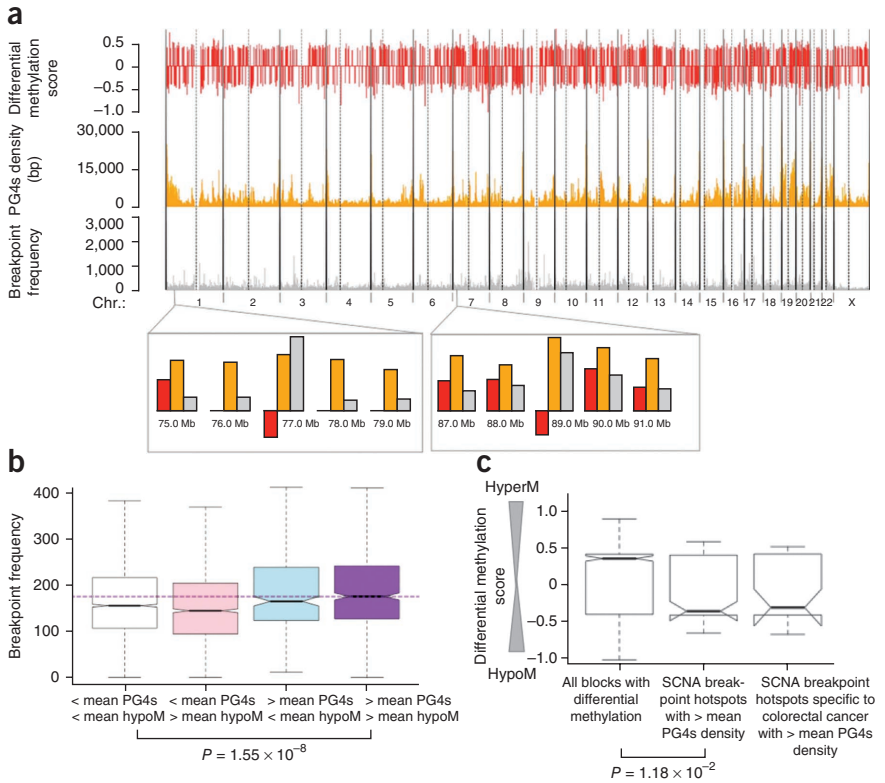

Figure 3 Role of G-quadruplex structures in the generation of breakpoint hotspots. (**a**) Extent of differential methylation in colon cancer relative to normal colon (red), density of G4 sequences (orange) and density of DNA breakpoints in cancer (gray) are shown across the human chromosomes. Vertical dotted lines mark centromeres. A negative value of differential methylation indicates differential hypomethylation. (**b**) The density of DNA breakpoints in cancer is higher in genomic blocks that have both above-average hypomethylation and above-average PG4s density than that in genomic blocks that do not have above-average representation of either of the factors. The purple horizontal dashed line shows the median breakpoint density corresponding to the rightmost group. The whiskers of the box plots represent the range of the breakpoint frequencies for the respective groups. (**c**) SCNA breakpoint hotspots with above-average PG4s density are significantly differentially hypomethylated (low differential methylation score) relative to the genome-wide background. SCNA breakpoint hotspots specific to colorectal cancers with above-average PG4s density show a similar trend (*P* value > 0.05 because there are fewer data points).

significant even after we excluded genomic blocks that were within 1 Mb of telomeres or centromeres ($P = 2.36 \times 10^{-7}$). Moreover, SCNA breakpoint hotspots with above-median PG4s density showed a significant enrichment for differential hypomethylation compared to the genome-wide background (**Fig. 3c**; $P = 1.18 \times 10^{-2}$, Mann-Whitney test). As these SCNA breakpoints were derived from various cancer types, we then focused our analysis on the colorectal cancer–specific breakpoint hotspots[2], and we found a similar trend (**Fig. 3c**). The association of PG4s with SCNA breakpoint density is significant even after controlling for DNA methylation, and the association between methylation and SCNA breakpoint density is marginally significant after controlling for PG4s (**Supplementary Methods** and **Supplementary Table 7**). We then repeated our analysis using breast cancer[46] and osteosarcoma[47] data (**Supplementary Methods**, **Supplementary Table 8** and **Supplementary Figs. 1** and **2**) and obtained similar results. Finally, we analyzed copy number, DNA methylation and gene expression data for glioblastoma samples[5] and found that loss of methylation in the CpG dinucleotides within PG4s was associated with genomic alterations (**Supplementary Methods**, **Supplementary Fig. 3** and **Supplementary Table 9**). Taken together, our data indicate that hypomethylation near regions of high PG4s density, which is rare in normal tissues but common in cancer genomes, is a signature of many DNA breakpoints across many cancer types.

DISCUSSION

Here we have established that SCNA breakpoints in cancer are often clustered into hotspots, which are markedly enriched for PG4s. The strand bias of the SCNAs relative to PG4s indicates that G4 structures are likely to have a causal role in cancer genome instability. Furthermore, we found that genomic regions rich in PG4s are on average hypermethylated in normal tissue, but hypomethylation in those regions is substantially associated with DNA breakpoint hotspots across a wide range of cancer types. Our results show that G4 structures and aberrant hypomethylation have a key role in generating genomic alterations in cancer.

On the basis of our analyses and supporting lines of evidence, we propose a mechanistic model of the potential contribution of hypomethylation and PG4s to the generation of genomic instability, thus bridging the roles of genetic and epigenetic factors driving tumorigenesis (**Fig. 4**). In normal tissues, the genome is generally hypermethylated, which is a marker for closed chromatin—a state generally unfavorable for G4 formation. In contrast, hypomethylation and open chromatin create a favorable condition for G4 structure formation in the presence of stabilizing proteins and negative supercoiling—for example, during transcription[25,40,43] or replication[41,42]. In addition, the CpG dinucleotide often occurs within PG4s, and methylation of those CpGs may also play a direct role in the stability of G4 structures through chemical and steric effects, as is the case for several other non–B-DNA structures[27,28]. Furthermore, CpG dinucleotide methylation regulates local nucleosome occupancy and chromatin structure[48], which may in turn influence DNA accessibility, G4 formation and DNA breakage. Aberrant genome-wide DNA hypomethylation, which can arise during aging[49] and tumorigenesis[46], exposes

Table 2 Enrichment for cancer breakpoint hotspots in genomic blocks with over-representation of G4 sequences and hypomethylation

Total no. genomic blocks	SCNA breakpoint hotspots (A)	Genomic blocks with: PG4s density	Differential methylation	No. blocks (B)	Overlap between groups A and B	Corresponding *P* value
3,029	248	> mean	< mean	579	72	2.47×10^{-5}
3,029	248	> mean + s.d.	< mean	205	30	5.09×10^{-4}
3,029	248	> mean	< mean – s.d.	272	39	1.26×10^{-4}
3,029	248	> mean + s.d.	< mean – s.d.	122	22	1.69×10^{-4}

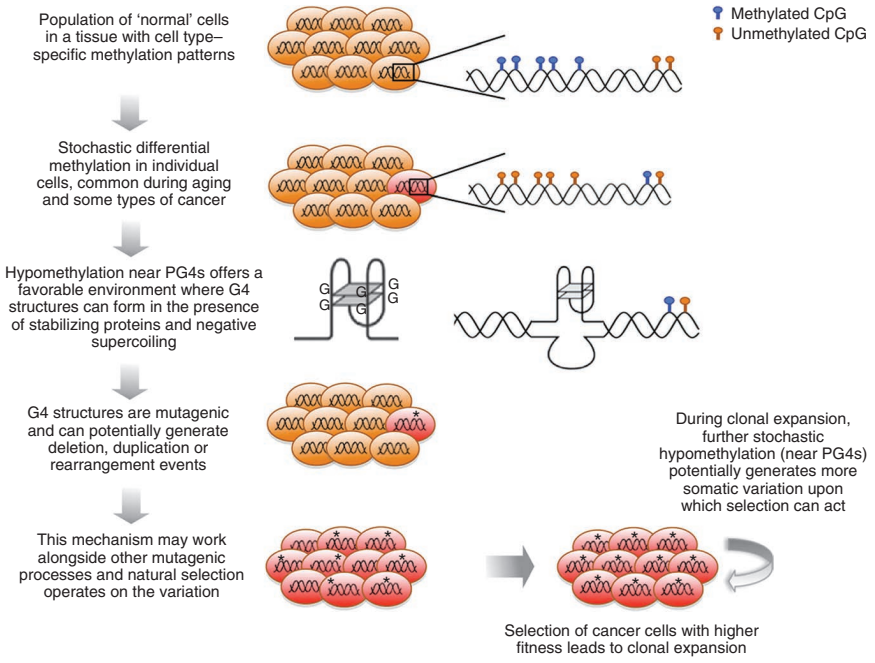

Figure 4 A mechanistic hypothesis of epigenetic involvement in the generation of breakpoints in cancer genomes. Genomes in normal tissue are generally hypermethylated and stable. Genome-wide hypomethylation, which occurs stochastically during aging and tumorigenesis, offers a favorable environment in which PG4s can fold into G4 structures in the presence of stabilizing proteins and negative supercoiling. G4 structures are mutagenic and have the potential to generate deletion, insertion or rearrangement events of genetic material on which selection can act to drive cancer evolution. See Discussion for further details.

large genomic regions where G4 structures can form frequently and perhaps nonspecifically; aberrant hypomethylation may work in concert with other epigenetic modifications, which cannot yet be systematically investigated because of insufficient data. Although such events are likely to be crucial during replication, some G4 structures formed during transcription may be recognized and mis-repaired by transcription-coupled repair or persist until subsequent replication. During replication, G4 structures obstruct DNA polymerase[17], increasing the risk of fork stalling and template switching (FosTes), erroneous microhomology-mediated replication-dependent recombination (MMRDR) and nonhomologous end joining (NHEJ)[9]; these processes then increase the risk of genomic alterations.

This proposed mutagenic process may work alongside other endogenous and exogenous mutagens to create genomic instability and generate mutations on which selection can operate during tumor evolution. If G4-associated structural alterations involve cancer genes or other functional elements, they may alter cellular fitness and thus change the course of cancer progression by leading to clonal expansion. Recent experimental findings that aberrant methylation promotes tumorigenesis[21] and is associated with PG4s[24,26,30], and that the formation of G4 structures is mutagenic[18–20], are consistent with our model. Because PG4s are widespread in the genome[38], methylation patterns differ between tissue types[44] and between cells within a tissue[49], and G4-mediated structural alteration is a stochastic event[45], this mechanism has the potential to generate tissue-specific mutational landscapes in cancer as well as heterogeneity among single cells within a tumor. Our model has attractive preventative, diagnostic and therapeutic implications, as agents that counteract hypomethylation and/or dissolve G4 structures may stabilize the rates of genomic aberrations and thus contribute to preventing cancer progression and the evolution of resistance. Our findings also contribute to the ongoing debate about epigenetic origins of cancer[8].

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We would like to thank G. Parmigiani, J. Widom, N. Maizels, G.-Ch. Yuan, R. Beroukhim and D. Patel for discussions and comments. S.D. is a recipient of Human Frontier Science Program long-term fellowship and is a Research Fellow at King's College, Cambridge. This work was funded by the US National Cancer Institute's initiative to found Physical Science–Oncology Centers (U54CA143798).

AUTHOR CONTRIBUTIONS

S.D. and F.M. designed the research and wrote the manuscript. S.D. performed the research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
2. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
3. Leary, R.J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. USA* **105**, 16224–16229 (2008).
4. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
6. Kim, J.I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
7. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
8. Feinberg, A.P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**, 21–33 (2006).
9. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
10. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
11. Wang, G., Christensen, L.A. & Vasquez, K.M. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. USA* **103**, 2677–2682 (2006).
12. Wang, G. & Vasquez, K.M. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl. Acad. Sci. USA* **101**, 13448–13453 (2004).
13. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K.M. Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.* **67**, 43–62 (2010).
14. Huppert, J.L. Structure, location and interactions of G-quadruplexes. *FEBS J.* **277**, 3452–3458 (2010).
15. Lipps, H.J. & Rhodes, D. G-quadruplex structures: in vivo evidence and function. *Trends Cell Biol.* **19**, 414–422 (2009).

16. Maizels, N. Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.* **13**, 1055–1059 (2006).
17. Sun, D. & Hurley, L.H. Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods Mol. Biol.* **608**, 65–79 (2010).
18. Kruisselbrink, E. *et al.* Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective *C. elegans*. *Curr. Biol.* **18**, 900–905 (2008).
19. Pontier, D.B., Kruisselbrink, E., Guryev, V. & Tijsterman, M. Isolation of deletion alleles by G4 DNA-induced mutagenesis. *Nat. Methods* **6**, 655–657 (2009).
20. Boán, F. & Gomez-Marquez, J. In vitro recombination mediated by G-quadruplexes. *ChemBioChem* **11**, 331–334 (2010).
21. Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455 (2003).
22. Kanai, Y. Genome-wide DNA methylation profiles in precancerous conditions and cancers. *Cancer Sci.* **101**, 36–45 (2010).
23. Carbone, L. *et al.* Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet.* **5**, e1000538 (2009).
24. Halder, K., Halder, R. & Chowdhury, S. Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. *Mol. Biosyst.* **5**, 1703–1712 (2009).
25. Huppert, J.L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406–413 (2007).
26. Wong, H.M. & Huppert, J.L. Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. Biosyst.* **5**, 1713–1719 (2009).
27. Behe, M. & Felsenfeld, G. Effects of methylation on a synthetic polynucleotide: the B–Z transition in poly(dG-m5dC)·poly(dG-m5dC). *Proc. Natl. Acad. Sci. USA* **78**, 1619–1623 (1981).
28. Vargason, J.M. & Ho, P.S. The effect of cytosine methylation on the structure and geometry of the Holliday junction: the structure of d(CCGGTACm5CGG) at 1.5 A resolution. *J. Biol. Chem.* **277**, 21041–21049 (2002).
29. Tsumagari, K. *et al.* Epigenetics of a tandem DNA repeat: chromatin DNaseI sensitivity and opposite methylation changes in cancers. *Nucleic Acids Res.* **36**, 2196–2207 (2008).
30. Halder, R. *et al.* Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.* **6**, 2439–2447 (2010).
31. O'Neil, J. & Look, A.T. Mechanisms of transcription factor deregulation in lymphoid cell transformation. *Oncogene* **26**, 6838–6849 (2007).
32. Sen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334**, 364–366 (1988).
33. Sundquist, W.I. & Klug, A. Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature* **342**, 825–829 (1989).
34. Williamson, J.R., Raghuraman, M.K. & Cech, T.R. Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell* **59**, 871–880 (1989).
35. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613–D619 (2010).
36. Durkin, S.G. & Glover, T.W. Chromosome fragile sites. *Annu. Rev. Genet.* **41**, 169–192 (2007).
37. Mani, P., Yadav, V.K., Das, S.K. & Chowdhury, S. Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS ONE* **4**, e4399 (2009).
38. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).
39. Gehring, K., Leroy, J.L. & Gueron, M. A tetrameric DNA structure with protonated cytosine•cytosine base pairs. *Nature* **363**, 561–565 (1993).
40. Sun, D. & Hurley, L.H. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J. Med. Chem.* **52**, 2863–2874 (2009).
41. Crabbe, L., Verdun, R.E., Haggblom, C.I. & Karlseder, J. Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. *Science* **306**, 1951–1953 (2004).
42. Sarkies, P., Reams, C., Simpson, L.J. & Sale, J.E. Epigenetic instability due to defective replication of structured DNA. *Mol. Cell* **40**, 703–713 (2010).
43. Basundra, R. *et al.* A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *FEBS J.* **277**, 4254–4264 (2010).
44. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
45. Mekmaysy, C.S. *et al.* Effect of O6-methylguanine on the stability of G-quadruplex DNA. *J. Am. Chem. Soc.* **130**, 6710–6711 (2008).
46. Shann, Y.J. *et al.* Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome Res.* **18**, 791–801 (2008).
47. Sadikovic, B. *et al.* In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. *PLoS ONE* **3**, e2834 (2008).
48. Chodavarapu, R.K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388–392 (2010).
49. Maegawa, S. *et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res.* **20**, 332–340 (2010).

ONLINE METHODS

Data sets. We obtained data for DNA breakpoints associated with somatic copy-number alterations from several different cancer types from a published report[2]. The authors originally studied a total of ~130,000 cases of gains and losses in 3,131 samples classified into 26 histological types, and each cancer type in this dataset was represented by at least 20 samples[2]. A subset of the data (~10% of samples) were not publicly available, and therefore we restricted our analysis to the set of publicly available 663,446 DNA breakpoints from 2,792 samples (89% of the complete dataset). SCNAs were obtained by comparing the signal intensities from the Affymetrix 250k array data of each cancer sample to the matched normal tissue[2]; the boundaries of alterations, which we denote as SCNA breakpoints, were determined with a precision of 8–10 kb. We also obtained a list of structural variations in 24 breast cancer samples from Stephens *et al.*[7], who used a paired-end sequencing strategy to identify somatic rearrangements. We obtained methylation data for colon cancer and also from healthy brain, liver and spleen samples from Irizarry *et al.*[44]; these authors performed a high-throughput array-based relative methylation analysis (CHARM) and also pyrosequencing-based revalidation analysis on an additional set of colon cancer samples.

Identification of breakpoint hotspots. To identify breakpoint hotspots, first we divided the cancer genomes into 1-Mb nonoverlapping blocks and counted the number of SCNA breakpoints in each block. Next, we randomized the position of the breakpoints 100,000 times for each chromosome and generated a distribution of breakpoint densities for the 1-Mb blocks. The genomic blocks that had a higher breakpoint frequency than that expected from the top 5% from the simulation across the whole genome were identified as breakpoint hotspots.

Genomic features. We obtained the genomic locations of PG4s from (http://www.quadruplex.org/)[38], where the PG4s were predicted using the Quadparser algorithm, which is based on the Folding rule postulating that a sequence of the form $d(G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3)$ will fold into a quadruplex under near-physiological conditions, where G is guanine and N is any nucleotide (A, T, G or C). We obtained the list of fragile sites from Durkin and Glover[36]. Common fragile sites are loci that preferentially show chromosomal aberrations visible as gaps and breaks on metaphase chromosomes after partial inhibition of DNA synthesis, and are present in normal individuals. Different families of repeat elements, recombination rate and 28-way evolutionary conservation information were obtained from the UCSC Genome Browser[35]. Data on recombination rate and fragile sites had about megabase resolution. The list of the genes causally implicated in cancer was obtained from The Cancer Gene Census database[50].

Analysis. All statistical analyses were performed using R. The **Supplementary Methods**, **Supplementary Figures 1–3** and **Supplementary Tables 1–9** contain details of all analyses.

50. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

doi:10.1038/nsmb.2089