DNA secondary structures and epigenetic determinants of cancer genome evolution

An unstable genome is a hallmark of many cancers. It is unclear, however, whether some mutagenic features in the genome sequence are encoded in a tissue-specific manner. We performed a genome-wide analysis of 663,446 DNA breakpoints from 2,792 cancer samples associated with somatic copy-number alterations (SCNAS) and established that SCNAS show a strong enrichment in cancer genomes. Many SCNA breakpoints are spatially clustered in cancer genomes. Notably, abnormal hypomethylation near SCNA breakpoints establishes a strand bias consistent with G4-mediated structural alterations. We observed a significant enrichment for G-quadruplex sequences (G4s) in the vicinity of many SCNA breakpoints and established a common mechanism for the enrichment of G4s in genomic regions. This mechanism is a strand-biased structural alteration that acts as a mutagenic factor in cancer.

1. Loss of genomic integrity is a common hallmark of cancer genomes. Recent technological advances have led to several large-scale cancer genome profiling studies 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, 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.

2. 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. 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. Guanine-rich sequences (G\textsuperscript{4}N G\textsubscript{4}N1-7G\textsubscript{4}N1-7G\textsubscript{4}1-714-10), which are frequent in the human genome, can adopt four-stranded structures called G-quadruplexes (G\textsubscript{4}Q) both in vitro and in vivo. Guanine-rich sequences obstruct the movement of DNA polymerase\textsubscript{a,b} thereby increasing the risk of DNA breakage or nonallelic homologous recombination. Indeed, G\textsubscript{4}Q structures have been implicated in germline recombination events. However, the role of G\textsubscript{4}Q structures in genomic instability in cancer has so far not been systematically investigated.

2.1 In addition to genetic factors, various epigenetic factors are also associated with genomic instability both during the somatic evolution of cancer and in germline evolution during speciation. Moreover, epigenetic patterns differ between cell types and thus possess the potential to generate tissue-specific patterns of alterations. Selective epigenetic changes may include additional previously undescribed target genes or functional elements, which are important for tumorigenesis or the development of cancer.

Subhajyoti 1,2 & Franziska Deitrich

1,2 Michor
Figure 1 Spatial distribution of breakpoint hotspots. (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 frequencies 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. identified 199 frequently altered regions across 3,131 cancer samples,
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 structures6,38 and often overlap with cancer breakpoints. Further, recombination hotspots, G4 structures and fragile sites are enriched for specific repeat sequences8,38; such repetitive elements show relatively low levels of evolutionary conservation. 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

<table>
<thead>
<tr>
<th>Genomic features</th>
<th>SCNA breakpoint hotspots in cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common repeats</td>
<td>3.66 × 10⁻³</td>
</tr>
<tr>
<td>Alu elements</td>
<td>6.88 × 10⁻²</td>
</tr>
<tr>
<td>CR1 elements</td>
<td>8.96 × 10⁻³</td>
</tr>
<tr>
<td>L1 elements</td>
<td>3.19 × 10⁻³</td>
</tr>
<tr>
<td>L2 elements</td>
<td>4.54 × 10⁻³</td>
</tr>
<tr>
<td>G4 secondary structures</td>
<td>9.19 × 10⁻⁶</td>
</tr>
<tr>
<td>DNA breakage and recombination features</td>
<td>2.43 × 10⁻²</td>
</tr>
<tr>
<td>Fragile sites</td>
<td>3.52 × 10⁻²</td>
</tr>
</tbody>
</table>

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.

To assess whether G4 structures are associated with DNA breakpoints in cancer, we overlaid information about PG4s 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 \( P = 9.19 \times 10^{-5} \), 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, recombination frequency8,36 and fragile sites35. We found that the breakpoint hotspots were also enriched for simple repeats, Alu repeats and CR1 repeats35 (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) and common fragile sites36 (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
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 × 10⁻¹⁰, 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 and cause mutagenic events (18-20) (Fig. 2b). Therefore, G4-mediated deletion and duplication events occur predominantly toward the 5’ direction of the G4 sequences. Although the G-rich strand can potentially form an l-motif structure, 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 × 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.
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 methylation21,22. In normal tissues, the genome is usually hypermethylated and does not show genomic instability, whereas genome-wide hypomethylation is a hallmark of many cancer types26,29. 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 occupancy27,28, and DNA methylation patterns have a role in the stability of other noncanonical DNA structures such as Z-DNA and H-DNA26. Furthermore, chemical modifications such as O-methylguanine inhibit G4 structure formations29. 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 tissues. 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 ($P = 4.16 \times 10^{-16}$, 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, $P = 1.55 \times 10^{-3}$, Mann-Whitney test). Our observation was independent of the threshold for hypomethylation and PG4s density (Table 2) and remained
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.
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).

### ANALYSIS

1.0 HyperM

<table>
<thead>
<tr>
<th>SCNA breakpoint density</th>
<th>PG4s density</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; mean PG4s</td>
<td>&gt; mean hypoM</td>
</tr>
<tr>
<td>SCNA breakpoint density</td>
<td>PG4s density</td>
</tr>
<tr>
<td>&gt; mean PG4s</td>
<td>&gt; mean hypoM</td>
</tr>
</tbody>
</table>

Total no. SCNA breakpoint hotspots

<table>
<thead>
<tr>
<th>Total no. SCNA breakpoint hotspots</th>
<th>PG4s density</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>&lt; mean PG4s</td>
</tr>
<tr>
<td>0</td>
<td>&lt; mean hypoM</td>
</tr>
</tbody>
</table>

we ecific the brea density showed a n kpoi significant foci nt enrichment for use hots differential d pots2 hypomethylation our ). anal More genome-wide

lysi over, s SCN P = 1.18 ×
on A 10-2 significant even after we excluded col kpoi genomic blocks that were within 1 Mb of ctal hots telomerases or can pots centromeres (P =
cer with 2.36 × 10^-7 = 1.55 × 10^-10

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. 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 sample and osteosarcoma, 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 cancers, Mann-Whitney test). As these SCNA breakpoints were derived from various cancer types,
0
-0.5
-1.0

<table>
<thead>
<tr>
<th>HypoM</th>
<th>All blocks with differential methylation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25,40,43 or replicates on siRNA</td>
</tr>
</tbody>
</table>

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$3,029 , 248 &gt; \text{mean}$ $\times 10^5$</td>
<td></td>
</tr>
<tr>
<td>$3,029 , 248 &gt; \text{mean} + \text{s.d.}$ $\times 10^4$</td>
<td></td>
</tr>
<tr>
<td>$3,029 , 248 &gt; \text{mean} - \text{s.d.}$ $\times 10^4$</td>
<td></td>
</tr>
</tbody>
</table>

$0.29248 > \text{mean} < \text{mean} 57972.47 \times 10^5$
$3,029 \, 248 > \text{mean} + \text{s.d.} < \text{mean} 20530.509 \times 10^4$
$3,029 \, 248 > \text{mean} - \text{s.d.} 27239.1.26 \times 10^4$
<table>
<thead>
<tr>
<th>Breakpoint frequency</th>
<th>Breakpoint CG density</th>
<th>Differential methylation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Differential methylation
score

Genomic blocks with:
Overlap between Differential methylation blocks (B) groups A and B Corresponding P value
ANALYSIS

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 mispaired by transcription-coupled repair or persist until subsequent replication. During replication, G4 structures obstruct DNA polymerases, increasing the risk of fork stalling and template switching (FosTes), erroneous microhomology-mediated replication-dependent recombination (MMRDR) and nonhomologous end joining (NHEJ); these processes then increase the risk of genomic alterations. G4 structures are mutagenic and can potentially generate deletion, duplication or rearrangement events. This mechanism may work alongside other mutagenic processes and natural selection operates on the variation.

Recent experimental findings that aberrant methylation promotes tumorigenesis and is associated of G4 structures is mutagenic, are consistent with our model. Because PG4s are widespread in between tissue types and between cells within a tissue, and G4-mediated structural alterations have the potential to generate tissue-specific mutational landscapes in cancer as well as heterogeneity and attractive preventative, diagnostic and therapeutic implications, agents that counteract hypomethylation may stabilize the rates of genomic aberrations and thus contribute to preventing cancer progression. This contributes to the ongoing debate about epigenetic origins of cancer.

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

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


ONLINE METHODS

Data sets. We obtained data for DNA breakpoints associated with somatic copynumber alterations from several different cancer types from a published report. 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. 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; the boundaries of alterations, which we denote as SCNAs, 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., 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.; these authors performed a high-throughput array-based relative methylation analysis (CHARM) and also a whole-genome bisulfite-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/), where the PG4s were predicted using the Quadparser algorithm, which is based on the Folding rule postulating that a sequence of the form d(GN3G1–7N4G1–7N3G1–723) 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 Glover55. 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 28S evolutionary conservation information were obtained from the UCSC Genome Browser. 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.

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.