Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability

Share Your Story

Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability

Yotam Drier,[1,2] Michael S. Lawrence,[2] Scott L. Carter,[2] Chip Stewart,[2] Stacey B. Gabriel,[2] Eric S. Lander,[2] Matthew Meyerson,[2,3] Rameen Beroukhim,[2,3] and Gad Getz[2,4]

[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel; [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA

Whole-genome sequencing using massively parallel sequencing technologies enables accurate detection of somatic rearrangements in cancer. Pinpointing large numbers of rearrangement breakpoints to base-pair resolution allows analysis of rearrangement microhomology and genomic location for every sample. Here we analyze 95 tumor genome sequences from breast, head and neck, colorectal, and prostate carcinomas, and from melanoma, multiple myeloma, and chronic lymphocytic leukemia. We discover three genomic factors that are significantly correlated with the distribution of rearrangements: replication time, transcription rate, and GC content. The correlation is complex, and different patterns are observed between tumor types, within tumor types, and even between different types of rearrangements. Mutations in the *APC* gene correlate with and, hence, potentially contribute to DNA breakage in late-replicating, low %GC, untranscribed regions of the genome. We show that somatic rearrangements display less microhomology than germline rearrangements, and that breakpoint loci are correlated with local hypermutability with a particular enrichment for C↔G transversions.

[Supplemental material is available for this article.]

Alterations in DNA drive much of cancer development. Many of these alterations are "structural," leading to fusions between distant regions of the genome. Many alterations are deletions and amplifications, which introduce copy-number changes. Others, such as inversions and balanced translocations, maintain copy number. Multiple mechanisms can cause these alterations, including deterioration of DNA repair and replication mechanisms (Hoeijmakers 2001; DePinho and Polyak 2004).

Recently, whole-genome sequencing became affordable enough to allow mapping of rearrangements for large cancer cohorts. This provides the opportunity to answer several key questions on DNA breakage in cancer. We and others have started to approach this by analyzing tumors from individual tumor types (Campbell et al. 2008, 2010; Stephens et al. 2009; Bass et al. 2011; Chapman et al. 2011; Stransky et al. 2011; Totoki et al. 2011; L Wang et al. 2011; Banerji et al. 2012; Berger et al. 2012), and we have specifically applied the initial version of the present method in a recent analysis of prostate cancer (Berger et al. 2011). Here, we study breakpoint patterns across cancer (95 samples of seven types of cancer) and extend our previous analysis.

We find three genomic factors that significantly affect the distribution of DNA breakpoints along the genome: replication time; proximity to transcribed genes; and GC content. These correlations allow us to hypothesize about the causes and cell-cycle timing (mitosis/interphase) of the breakage events, and serve as a basis for future modeling of passenger rearrangements in cancer.

We also identified a significant correlation between breakpoints and somatic point mutations. Although we cannot formally distinguish between cause and effect, we ruled out the possibility that the correlation is merely due to genomic variation in the susceptibility to acquire both types of genome alteration. Furthermore, pinpointing the precise breakpoints of rearrangements allows characterization of microhomology, which may suggest potential mechanisms of rearrangement.

Results

Detecting somatic rearrangements

The growing number of whole-genome sequencing efforts in cancer is raising the need to accurately pinpoint rearrangement breakpoints without additional experimental measurements, particularly due to the high number of breakpoints found. Several studies to date (Campbell et al. 2008, 2010; Stephens et al. 2009; Totoki et al. 2011) either published approximate breakpoint locations or performed additional experiments to pinpoint the breakpoints (e.g., by amplification of the region and resequencing). We recently published several other studies (Bass et al. 2011; Berger et al. 2011, 2012; Chapman et al. 2011; Stransky et al. 2011; L Wang et al. 2011; Banerji et al. 2012) in which we pinpoint the breakpoints to base-pair resolution using BreakPointer, described here in detail for the first time (Supplemental Methods; Supplemental Fig. 1).

In this study, we perform a pan-cancer analysis of rearrangement breakpoints based on WGS data from 95 matched tumor/normal samples: 24 breast samples sequenced at the Sanger Institute (Stephens et al. 2009) and 71 sequenced at the Broad Institute from various tumor types: 23 multiple myeloma (Chapman

et al. 2011), 22 breast carcinomas (Banerji et al. 2012), nine colorectal carcinomas (Bass et al. 2011), seven prostate (Berger et al. 2011), five melanoma (Berger et al. 2012), three chronic lymphocytic leukemia (L Wang et al. 2011), and two head and neck (Stransky et al. 2011). A total of 4996 candidate and approximate somatic rearrangements were detected using dRanger (Supplemental Methods) in the 71 Broad Institute samples. Out of these, 4368 (87%) were successfully pinpointed to single base-pair resolution using BreakPointer (Supplemental Table 1). We successfully validated the existence of 1580 out of 1880 (84%) rearrangements randomly selected for validation by PCR and targeted pyrosequencing (Methods), and confirmed the exact pinpointing of 1503 (95%) by aligning the pyrosequencing results to the fused sequence predicted by BreakPointer. In the analyses presented below, we used different data sets—either the 4368 successfully pinpointed breakpoints or, when relevant, the 4996 candidate rearrangements (though the additional 628 rearrangements did not significantly change the results). The additional 24 samples by Stephens et al. (2009) are used only for the analysis of factors determining the distribution of breakpoints.

Microhomology of rearrangements

Rearranged DNA segments occasionally share a short stretch of identical sequence, known as an overlapping microhomology (Zhu et al. 2002). The base pairing between the two segments being fused is thought to guide the exact location of the fusion. Knowing the exact breakpoint allowed us to measure the microhomology for every rearrangement.

In general, rearrangements display an increased level of microhomology, with an average of 1.7 bp instead of the 0.7 bp expected by chance (a 2.4-fold increase; Wilcoxon P-value $< 10^{-250}$; see Methods). To study whether this excess of homology occurs in all types of rearrangements, we classified them into five categories: (1) short deletions (<5 kb); (2) inversions; (3) tandem duplications; (4) all other intrachromosomal rearrangements (mostly deletions); and (5) interchromosomal translocations. All types showed more microhomology than expected by chance (2.2- to 2.8-fold increase; Wilcoxon P-value $< 10^{-25}$). This is true also for every type of cancer separately—except for intrachromosomal rearrangements in CLL, all types with 10 or more rearrangements showed significant increase, FDR < 10%. The short microhomologies imply the involvement of nonhomologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ) in almost all somatic rearrangements (only 0.2% of detected rearrangements displayed >20 bp homology). MMEJ is rare, while NHEJ is quite frequent (only 2.5% of rearrangements had >5-bp microhomology, 44.2% at least 2 bp, but at most 5 bp). Even when comparing only with nonhomologous germline rearrangements in 185 human genomes (Mills et al. 2011), we found that the microhomologies of somatic rearrangements detected in our cohort were shorter (average of 1.7 bp vs. 2.2 bp, Mann–Whitney P-value $< 5.4 \times 10^{-14}$), and MMEJ less frequent (6.6% of nonhomologous germline rearrangements had >5-bp microhomology, 46.6% at least 2 bp, but at most 5 bp). Recently, complex rearrangements in the germline were characterized in several individuals (Chiang et al. 2012), which showed less microhomology than Mills et al. (2011). These complex germline events are closer to the somatic events described here in terms of the overall microhomology distribution (average 1.43 bp, Mann–Whitney P-value < 0.012), probably due to less NHEJ and more MMEJ (5.7% had >5-bp microhomology, 28.6% at least 2 bp, but at most 5 bp).

The distribution of microhomology lengths varied by the type of rearrangement (Scholz-Stephens' P-value $< 10^{-6}$; see Methods). Tandem duplications had the most distinctive distribution, with 2 bp (typical for nonhomologous end joining) being the most common overlap across all tumor types (as we previously reported in colorectal cancer) (Bass et al. 2011). Short deletions and inversions displayed a similar pattern (Fig. 1A). Difference in microhomologies, and specifically more frequent microhomologies in tandem duplications, was previously reported for breast cancer (Stephens et al. 2009).

Each sample had a different composition of rearrangement types (Supplemental Fig. 2), and therefore differences between the microhomology distributions of different samples are to be expected. However, even when controlling for the sample-specific composition and using the overall microhomology pattern for each type, six of the 71 samples (8%) still had a significantly different distribution (FDR < 4%) (Fig. 1B; Methods). Three prostate samples displayed less microhomology than expected by their composition, while three breast samples displayed more, suggesting mechanistic differences not only between the different types of rearrangements, but also between prostate, breast, and other cancers. Indeed, when pooling all breast samples together, they show more microhomology than expected by their composition ($P < 10^{-6}$), and all prostate samples pooled show less microhomology than expected ($P < 10^{-6}$).

Factors determining the distribution of breakpoints

Next, we examined genomic features to identify ones that may affect, or at least are correlated with the density of rearrangement breakpoints along the genome. First, we examined whether the distribution of breakpoints was correlated with local transcription levels typical for that tumor type (Methods). As for microhomologies, we observed strong sample-specific effects with different samples showing opposite behaviors—some with significant enrichment of breakpoints near transcribed genes (most pronouncedly within 10-kb windows) and others with significant depletion (Fig. 2).

Subsequently, we examined the correlation with two additional factors that may affect the location of rearrangements—DNA replication time and GC content. We first considered the effect of each factor separately and partitioned the genome into three or four distinct parts according to the level of each factor. We then calculated, for each sample, the relative rate of breakpoints in every part of the genome (represented as log fold-change to the genome-wide average) and a significance level (Fig. 3A; Methods). Interestingly, the majority of samples showed enrichment of breakpoints either at early replicating, high %GC, transcribed regions of the genome (EHT), or at late-replicating, low %GC, untranscribed (LLU) regions. The fact that the effects of these three variables are correlated is not surprising since they are mostly correlated along the genome. Studying the enrichment patterns across cancer revealed tumor-type-specific patterns; CLL, and breast cancer samples tend to have breakpoints at EHT regions, while colorectal cancer, melanoma, and head and neck cancer samples tend to have breakpoints in the LLU regions (Fig. 3B).

Four samples showed contradictory patterns of LLU and EHT, deviating from the above pattern, suggesting that, at least in these cases, more than one factor is required to explain the density of breakpoints. The colorectal sample CRC-3 was enriched for breakpoints in late-replicating, untranscribed regions, but depleted in regions of low %GC. The multiple myeloma sample MMRC0421
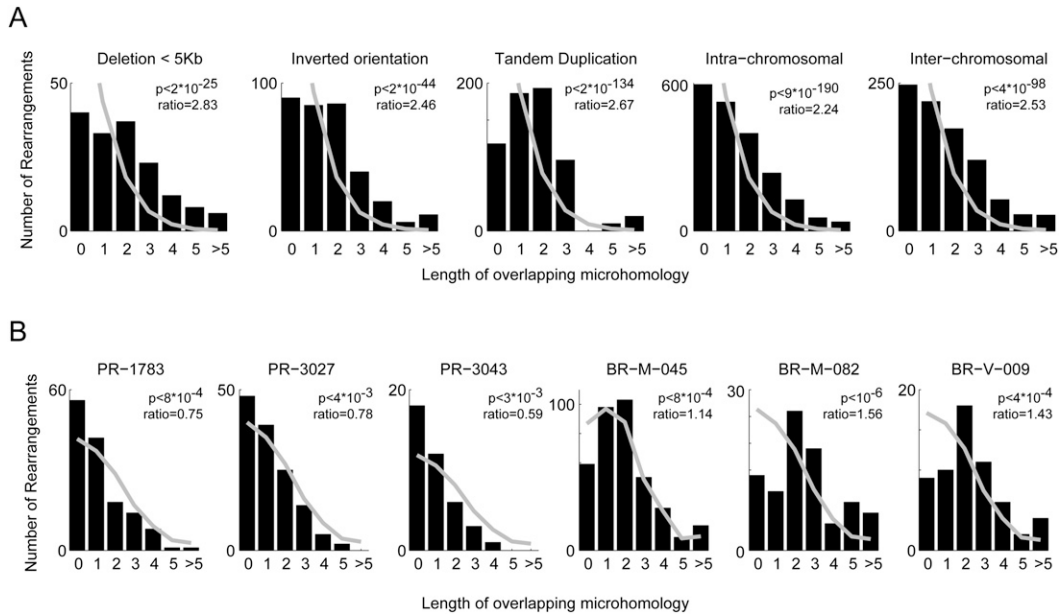
Figure 1. Overlapping microhomology. (*A*) By rearrangement type. (Gray line) The expected distribution, by permuting rearrangement pairs. All rearrangement types show higher microhomology than expected by chance. Tandem duplications display the highest microhomology rate with microhomology of length 2 being the most common case. Short deletions (up to 5 kb) and inversions show more microhomology than other rearrangements. Scholz-Stephens *P*-value for significant difference between histograms is <10^{-6}. (*B*) Rearrangement count by sample for six extreme samples. (Gray line) The expected distribution, controlled for the composition of the different rearrangement types. The three prostate samples show less microhomology than expected (notice the high fraction of breakpoints with no microhomology), and the three breast samples show more (low fraction of breakpoints with no microhomology). Expected distribution was constructed to control for the different rearrangement types and the homologies they display in our cohort. These are the only samples passing FDR < 10% (and in fact satisfy FDR < 4%).

and melanoma sample ME0032 harbored breakpoints in untranscribed regions, but also in regions with high GC content, and the breast sample PD3668a was depleted in both low %GC and high %GC.

These inconsistent patterns can be somewhat explained by examining the contribution of each type of rearrangement separately. Surprisingly, in these samples, different types of rearrangements follow different patterns of enrichment. For the melanoma sample ME0032 (Supplemental Fig. 3), interchromosomal translocations and intrachromosomal inversions and tandem duplications were enriched in regions of high %GC, while other intrachromosomal events were skewed toward low %GC and untranscribed regions. Similarly, for multiple myeloma sample MMRC0421 intrachromosomal rearrangements contributed to enrichment in untranscribed, low %GC regions, while inversions were enriched in high %GC.

In order to quantify the joint contribution of all three parameters and attain a compact representation, we used logistic

regression (Methods; Supplemental Fig. 4). This type of analysis requires a large number of rearrangements in order to uncover significant results, and so only the most highly rearranged samples are amenable. To cope with this challenge, we pooled together several samples of the same cancer type. In contrast to the outliers described above, it seems that the general rule is for rearrangements of different types (deletion, inversion, etc.) to be distributed similarly to each other; however, we cannot rule out some cancellation of opposite effects due to pooling of samples.

Next, we searched for genes with mutations that are correlated with the different patterns of rearrangements (LLU or EHT). Interestingly, we identified *APC* as the only gene whose mutations (in the coding region or promoter) are significantly associated with the LLU-enriched samples (q < 0.05; Methods). The adenomatous polyposis coli (*APC*) gene is mutated in eight of the 71 samples, seven of which are included in the 19 LLU samples (Fisher's exact test; $P = 10^{-4}$, q = 0.017). APC binds to and stabilizes microtubules, and is necessary to keep chromosomal integrity
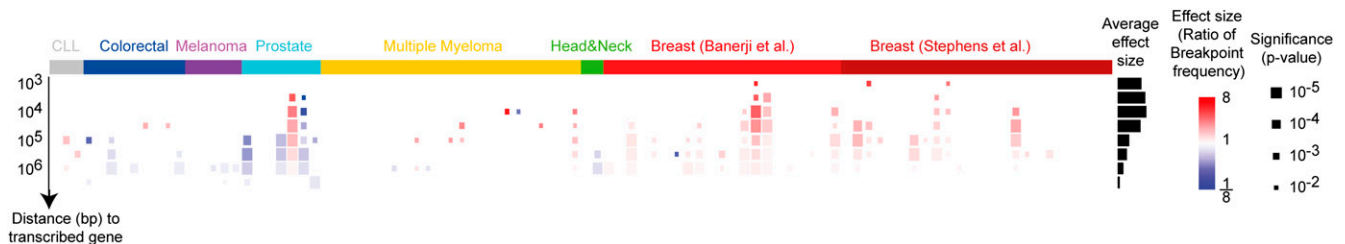
Figure 2. Breakpoints in transcribed and untranscribed regions. Each square represents enrichment (red) or depletion (blue) of breakpoints in transcribed regions defined by maximal distance to transcribed gene. Size represents *P*-value, and color represents ratio. Only tests that passed 10% FDR are shown. Notice that regions of ~10^4 bp were often significantly enriched or depleted. (*Right*) The average ratio (across samples). The colored bar *above* specifies the type of cancer for each sample.
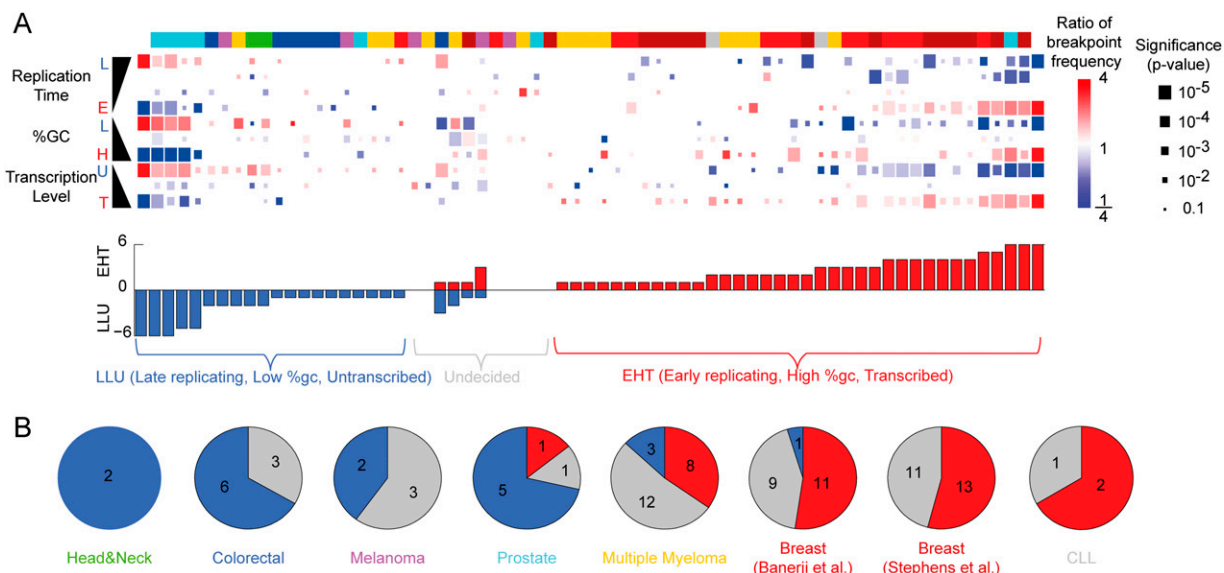
Figure 3. Breakpoint distribution as a function of transcription, replication, and GC content across samples. (*A*) Each row represents a different bin of replication time, GC content, or distance from transcribed gene. Each square represents significant (FDR < 10%) enrichment or depletion, size represents *P*-value, and color represents ratio. Only samples with at least one significant bin are shown. The colored bar *above* specifies the type of cancer for each sample. Most samples are either enriched for breakpoints in early replicating, high %GC transcribed regions of the genome (EHT), or in late replicating, low %GC untranscribed regions (LLU), as can be seen in the bar chart. The samples are sorted by the agreement with that pattern. (*B*) The breaking of each cancer to EHT (red), LLU (blue) and gray samples (without any significant extreme bin, or with contradicting enrichments).

during mitosis (Kaplan et al. 2001; Guerrero et al. 2010). Defects in APC might, therefore, lead to chromosome breakage during aberrant mitosis, or disrupt mechanisms that protect or repair heterochromatin regions. *APC* is known to be highly mutated in colorectal cancers (~70%–80%) (Fearon 2011; The Cancer Genome Atlas Network 2012) and, indeed, all six colorectal samples with an *APC* mutation were LLU, and the remaining three were not. This explains the high prevalence of LLUs in colorectal cancer. This might suggest that the correlation to *APC* mutations is merely due to colorectal cancer being a confounding variable, and require further study on larger cohorts.

Hypermutability near breakpoints

Analysis of the relationship between the sites of somatic mutations and rearrangements showed that the rate of somatic single-nucleotide variations is significantly elevated near breakpoints (Fig. 4A; Methods). The effect can be detected in very close proximity to the breakpoint, but it becomes even stronger when calculated across 100 bp–1 kb surroundings. Notice that the windows are nonoverlapping, i.e., each window has a "hole" in the middle associated with the previous smaller window, and therefore the hypermutability is detectable also in regions far from the breakpoint. The increase in mutation frequency in a 1-kb window around breakpoints often reaches a staggering 100×–3000× fold for several samples (Fig. 4B). The relationship between hypermutability and rearrangements was noted previously in various contexts (De and Babu 2010), and we also previously showed it specifically for prostate cancer (Berger et al. 2011). Here we demonstrate that this is true across many cancer types.

The hypermutation cannot simply be explained by rearrangement and mutations occurring in the same regions of the genome that are hyper-susceptible to all forms of genomic aberrations in all cases. We examined regions defined by the rearrangements of any given sample, and looked for mutations in those regions in all other samples of the same cancer type. While sometimes we indeed noted elevated mutation rates (coinciding with the hypothesis of fragile and hypermutable genomic regions), there were almost always significantly more mutations (~16× increase in density) in samples identified by comparing with the genome-wide average (Fig. 4C).

The spectrum of the mutations surrounding breakpoints is significantly different from the spectrum over the entire genome, as can be seen in Figure 4D, with C↔G transversions being most highly enriched. C↔G transversions were suggested to be caused by oxidative DNA damage (Kino and Sugiyama 2001, 2005) and by base excision repair via uracil-DNA glycosylase and REV1 translesion synthesis (Jansen et al. 2006; Ross and Sale 2006).

C↔G transversions are known to be enriched in breast cancer (Stephens et al. 2005), where they tend to occur in a TpC (or GpA for G → C) dinucleotide context. A similar context-specific pattern also holds in lung cancer, ovarian cancer, and melanoma (Greenman et al. 2007; Rubin and Green 2009). Mutations in that context are consistent with a DNA deamination by apolipoprotein B mRNA-editing enzymes (APOBEC1 and several APOBEC3 proteins) (Beale et al. 2004; Bishop et al. 2004). We confirmed the enrichment of C↔G transversions in the TpC context, but also observed that this effect is significantly higher near breakpoints. Out of 25 samples that have more than five C↔G transversions near breakpoints (1 kb or less), nine (five breast cancer, two melanoma, and two multiple myeloma) displayed significant enrichment of TpC context compared with transversions far from breakpoints (FDR < 5%, Fisher's exact test).

One of the features of the translesion synthesis that we suggested above is that it acts upon one of the strands, and therefore only this strand will be mutated by the deamination. Indeed, two multiple myeloma samples (MMRC0344 and MMRC0392) and four breast samples (BR-V-004, BR-V-006, BR-V-008, and BR-V-010) had a least one breakpoint with significant strand specificity (FDR < 10%, see Methods).
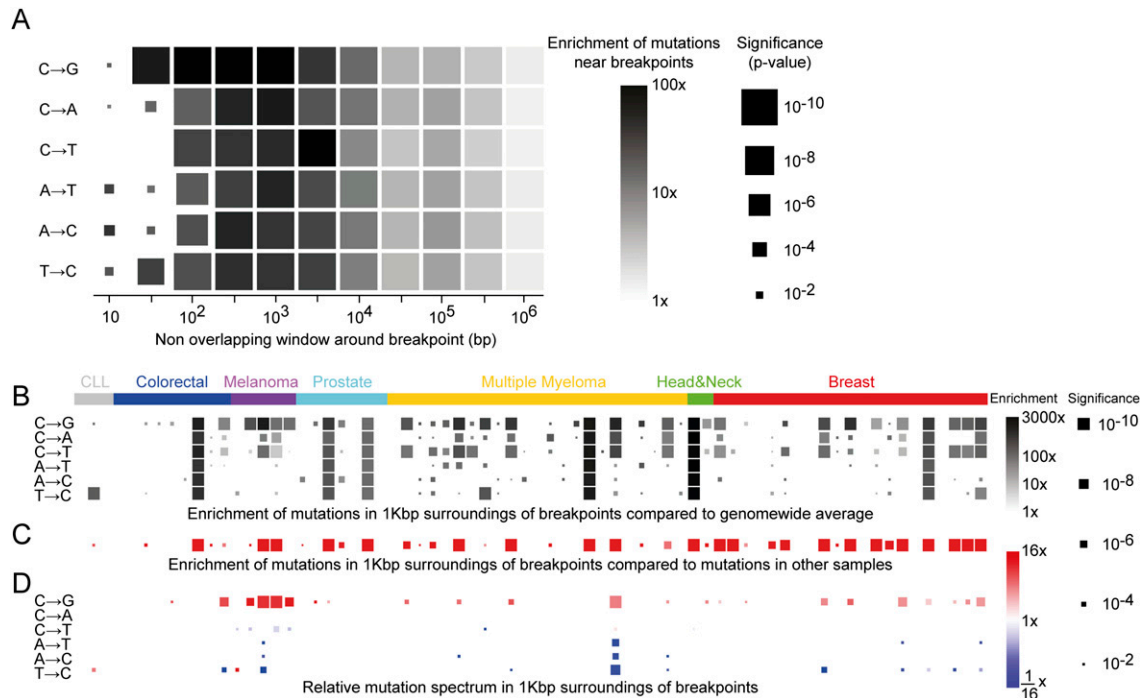
Figure 4. Hypermutability near breakpoints. (*A*) Enrichment of mutations across all samples by mutation type. Square represents mutation rate in concentric nonoverlapping exponential windows around each breakpoint, compared with overall mutation rates in the 71 samples cohort, aggregating them together. Size represents *P*-value, and color represents ratio. Only significant (FDR < 10%) results are shown. Hypermutation can be seen in a close proximity of the breakpoint, but it is even stronger in 100 bp to 1 kb surroundings. (*B*) Similar analysis per sample in 1-kb windows reveals that for some samples the mutation rate can reach 1000×–3000× fold. (*C*) Hypermutation is not only due to rearrangement and mutations occurring in the same "bad" regions of the genome. For each sample we defined the 1-kb regions according to their rearrangements and measured the mutations in those regions in all other samples of the same cancer type, aggregating them together. Squares represent *P*-value (by size) and ratio (by color) comparing the mutation rate in each selected sample to the mutation rate at the other samples of the same cancer type. Any sample with significant hypermutation displays significant elevation in mutation rate near breakpoints of that sample. (*D*) Mutation spectrum near breakpoints compared with spectrum across the genome of that sample. Hypermutated samples are often skewed toward C↔G transversions near breakpoints. Melanoma samples show depletion of C → T transitions near breakpoints due to high C → T transitions across the genome.

Recently, clusters of mutations were discovered in breast cancer, a phenomenon termed kataegis (Nik-Zainal et al. 2012) as well as in yeast and other types of cancer (Roberts et al. 2012). Some of those clusters were colocalized with rearrangements. Nik-Zainal et al. (2012) identified five mutational signatures by statistical inference, two of which (B and E) were found to be enriched in the kategis events. Signature E is mainly C↔G transversions in TpC context, and signature B is a combination of C↔G and C → T in TpC context. Our results are consistent with their findings, namely, hypermutation near breakpoints, enrichment of C↔G mutations in TpC context, and strand specificity.

Discussion

We identified three genomic factors that significantly affect, in a sample-specific manner, the distribution of breakpoints: GC content, transcription, and replication time. The scales on which transcription affects the distribution of breakpoints suggest that the main effect is through the 3D DNA structure of the genome, i.e., the different open/closed chromatin compartments (present mostly during interphase). DNA replication time suggests colocalization, mostly during replication (Meister et al. 2006; Ryba et al. 2010), and was shown to affect rearrangements in bacteria (Eisen et al. 2000; Tillier and Collins 2000) and has been recently suggested for cancer as well (De and Michor 2011). GC content might affect breakpoint distribution by sequence-dependent mecha-

nisms (such as homology), or may simply be correlated to other biologically relevant factors. We show that the three factors, although highly correlated, are not redundant, and each may contribute differently in different contexts, e.g., in different samples and in different rearrangement types. We previously showed some correlation between breakpoints and transcription for prostate samples (Berger et al. 2011); here, we extend our analysis and offer a possible explanation. This genomic scale is consistent with recent discoveries that during interphase, transcription occurs in distinct compartments in the nucleus, and that untranscribed regions occupy other compartments (Lanctot et al. 2007; Guelen et al. 2008; Lieberman-Aiden et al. 2009; Yaffe and Tanay 2011). It is known that breakpoint-pairs of individual rearrangements often occur in nearby segments of the DNA (Meaburn et al. 2007; Mani and Chinnaiyan 2010; De and Michor 2011; Fudenberg et al. 2011; Klein et al. 2011). However, we find that many breakpoints, belonging to different rearrangements, also tend to occur in some samples in transcribed/early replicating compartments, and in others in untranscribed/late replicating compartments. This is not an artifact due to the vicinity of breakpoint-pairs of individual rearrangements, as a similar pattern is observed when randomly selecting only one breakpoint of each rearrangement and repeating the analysis (Supplemental Fig. 5). This observation is consistent with a model in which one or more events has occurred, each causing several breakpoints within the same compartment, perhaps due to a strong DNA damaging event (as suggested to cause

chromothripsis [Stephens et al. 2011] when occurring during metaphase). Incorrect fusion of the resulting nearby fragments then yields the observed rearrangements. Moreover, we suggest *APC* deficiency as a mechanism that may contribute to DNA breakage in late replicating, low %GC, untranscribed regions of the genome, during or after mitosis. However, the accuracy of our findings regarding transcription and replication may be imperfect, as we did not measure transcription and replication in the specific tumor samples analyzed here. Reassuringly, time of replication is mostly constant in different tissues (Farkash-Amar and Simon 2010). To make sure that the small difference in the patterns of transcription does not have a big impact, we have deduced the expression profile for each type of cancer separately. Moreover, repeating the analysis with different expression profiles yielded similar results (data not shown).

This data-driven model of the breakpoint distribution is not predictive at this point and requires the full analysis of breakpoints in each sample. Due to the complexity of the effects, we believe that such an approach is necessary to assess the significance of driver rearrangements across cancer. Since the cohort size is still a limiting factor, statistical inference of the causes of the different behavior of different samples is not yet possible. However, with the large number of cancer whole-genome sequences becoming available, this is expected to change in the near future, allowing similar methodology to provide an understanding of different biological processes that contribute to the variability across samples and types of alterations.

Integrative analysis of mutations and exact breakpoints revealed a global hypermutability near breakpoints, common to almost all samples. We suggest that the hypermutability might be related to base excision repair caused by *APOBEC* deamination, which can cause both DNA breakage and the mutations that we observe near breakpoints. Moreover, the strand-specific pattern in some of the samples may suggest that it is caused by translesion synthesis, which is known to occur in base excision repair. This emphasizes the complexity of understanding the deterioration of genome stability, the effect of different DNA repair mechanisms, and the need to integrate the different data types in order to better understand them. It also has a practical impact on modeling of background mutation rates in cancer. The different mutation spectrums near rearrangements suggests that different mechanisms generate or repair these mutations, and may help point to these mechanisms. Further study is required to understand the relationship between breakpoints and the processes that govern mutation spectra.

Methods

Data and preprocessing

The data used for the analysis was whole-genome shotgun sequencing performed as described in the references (Berger et al. 2011; Chapman et al. 2011). Candidate chromosomal rearrangements were identified from the observation of multiple discordant read pairs using dRanger (Supplemental Methods). BreakPointer was originally designed to use MAQ (Li et al. 2008) alignments, but was also adapted to BWA (Li and Durbin 2009). BWA later introduced advanced clipping features, making the identification of split reads easier by allowing the use of alternative rearrangement detection algorithms such as CREST (J Wang et al. 2011). Breast and head and neck samples were aligned using BWA, all other samples using MAQ.

Validation

Rearrangements predicted by dRanger (with at least three supporting discordant reads) were validated by PCR, followed by pooled 454 Life Sciences (Roche) sequencing. PCR primers were designed using Primer3 (Rozen and Skaletsky 2000), such that they spanned the predicted chimeric junction and would produce an amplicon ~300–350 bp long. PCRs were performed on whole-genome amplified product for both tumor and normal DNA (For somatic breakpoints, only the tumor DNA would be expected to yield a product). Each PCR product was quantified using a NanoDrop Spectrophotometer (Thermo Scientific). PCR products were pooled such that: (1) Equal amounts of tumor products were combined, (2) the same volumes were taken from the corresponding normal products, and (3) matching tumor and normal products were placed in separate pools. Libraries for 454 sequencing were prepared from each pool and sequenced in separate regions of a 454 Genome Sequencer FLX System (454 Life Sciences). Primer sequences served as unique barcodes for identifying the source PCR product for each 454 read. A rearrangement was judged to be somatic if the predicted chimeric product was detectable in tumor DNA and not normal DNA.

To validate BreakPointer results, the fused sequence generated by BreakPointer was aligned by Smith and Waterman (1981) to all of the sequences of the appropriate amplicons (or their reverse complement). For each amplicon, the alignment was declared to be successful if it contained no gaps in a 20-bp window around the breakpoint (to ensure exact pinpointing) and at least 95 matches in a 100-bp window. Notice that since BreakPointer fuses the reference genome, some mismatches with cancer genomes are expected (due to germline and somatic point variations).

Statistical analysis of microhomologies

Wilcoxon rank-sum test was used to compare the observed microhomology distribution with the expected background for each type of rearrangement separately. The background for each test is based on hypothetical rearrangements constructed by taking all possible breakpoint pairs among the breakpoints belonging to a particular rearrangement type, and then computing the distribution of microhomologies in this set of hypothetical rearrangements. To evaluate the difference between the histograms of the different rearrangement types, we used Scholz-Stephens' k-sample Anderson-Darling statistic (Scholz and Stephens 1987) to measure the similarity between the histograms. We then tested the significance of this value based on 10^6 sets of "permuted" histograms generated under the null hypothesis in which the histograms are, in fact, not different. To generate each set of histograms, we randomly permuted the observed microhomology among the five rearrangement types. We then computed the Anderson-Darling statistic for each set, and the *P*-value is simply the fraction of sets with greater or equal Anderson-Darling statistics than the original five histograms. To evaluate the contribution of the short deletions and the tandem duplications to the significance, we repeated the analysis omitting one or both. Excluding the short deletions and keeping the tandem duplications yields histograms that are still significantly different ($P < 10^{-6}$). However, when removing the tandem duplications and keeping the short deletions, the results are less significant ($P = 0.03$), and when omitting both the histograms, are no longer significantly different ($P = 0.15$).

To detect a significant deviation of the average microhomology from that expected, in a given sample we calculated empirical *P*-values by comparing the observed average microhomology to a background distribution that controlled the composition of rearrangement types in the sample. For each sample, the background distribution was constructed by sampling 10^6 times the appropriate number of rearrangements of each type. We capped the microhomology at 6 bp to eliminate the unwanted effect of inflating the average due to a few rearrangements with large homology. Similarly, for the cancer-type-specific analysis, all samples of the same cancer-type were pooled together and deviations from the appropriate background of the pool were calculated.

Usually microhomology is defined by perfect homology. However, biological mechanisms mediating microhomology might induce imperfect homology (i.e., sequence similarity of <100%). To assess our sensitivity we also attempted to define rearrangement with microhomology by requiring at least five matches and up to two mismatches (i.e., sequence similarity >71%). Only 12% of those rearrangements did not have 2-bp perfect microhomology, yielding no significant change in any aspect by adding this new definition. No correlation ($\rho = -0.009$, $P = 0.53$) was detected between microhomology and coverage, excluding the possibility of a detection bias due to coverage.

Breakpoint distribution statistical analysis

Enrichment and depletion of breakpoints in different regions of the genome, defined by replication time, GC content, and distance to transcribed gene, were computed by random generated distributions controlled for chromosome and coverage. First, nearby breakpoints (up to 2500 bp away) were consolidated into a single "event." This was needed since nearby breakpoints were probably a result of one DNA breakage event. Controlling for chromosome was required to avoid artifacts resulting just from the chromosome identity. These steps are specifically important for short deletions and in the presence of complex events (such as balanced translocations [Berger et al. 2011] or variants of chromothripsis [Stephens et al. 2011]) that occur in several of the samples, as we previously reported (Bass et al. 2011; Berger et al. 2011, 2012). Controlling for coverage (using the average coverage of all samples aligned to the same genome build) was needed because the ability to detect rearrangements depended on coverage (Supplemental Fig. 6).

For each event, 100,000 locations (one per iteration) were generated uniformly from all locations on the same chromosome having the same coverage (quantizing the average coverage across all samples mapped to the same genome build in steps of five). The genome was binned to the following bins: low GC (0%–36%), medium GC (36%–45%), and high GC (45%–100%). Transcribed regions (transcribed gene in 100 kb), medium (100–500 kb), and untranscribed regions (no transcribed gene in 500 kb) (see definition below of transcribed gene). Replication time was binned according to late/early ratio (Ryba et al. 2010) at $(-\infty, -0.8]$, $(-0.8, 0]$, $(0, 0.8]$, $(0.8, \infty)$. Changing the thresholds did not affect the essence of the results other than losing sensitivity for too small or too big bins (data not shown). For every bin we counted the number of breakpoints for both the observed breakpoints and the random breakpoints. All of these counts were used to compute nonparametric P-values (observed rates). Enrichment or depletion was determined by picking the lower of the one-sided P-values, and P-values were then corrected for multiple hypotheses by the Benjamini-Hochberg FDR procedure (Benjamini and Hochberg 1995). Logistic regression was used to study the effect of each parameter on the probability of having a breakpoint. The number of breakpoints that fall in each bin is modeled as a binomial distribution with probability p. Logistic regression models $log(p/(1-p))$ as a linear combination of the binned covariates (%GC, transcription and replication time), where each covariate is assigned a value (evenly spaced between -1 and 1) to represent its bin. To train the model we used the observed breakpoints as "successes" and the permuted breakpoints (from the enrichment test, see above) to represent the "failures."

Transcribed genes were identified by picking the 10,000 most-expressed genes on average from a matching data set, as described in the Supplemental Methods. DNA replication time data for H7 hESC cells was obtained from reference (Ryba et al. 2010), remapping to hg19 build was done via UCSC Genome Browser's tool liftOver. The GC content was called in 100-kb windows. The only noticeable effect of using 10-kb or 1-Mbp windows was equivalent to slightly shifting the bin thresholds (as the smaller the windows size is, the more disperse the GC distribution).

To look for genes mutated in LLU or EHT samples, we examined all mutations within genes, other than silent mutations or mutations in introns (but including mutations in promoters and UTR). We chose only genes that have the potential to be differentially mutated, i.e., those mutated in at least three samples, which are not mutated in at least three samples within our LLU and EHT samples. Fisher's exact test was used to calculate the probability of a gene to be mutated in as many LLU or EHT samples.

Mutation rate statistical analysis

To test for enrichment of mutations near breakpoints, the same generated background distribution described above was used to count how many breakpoints had at least one mutation in any given window around the breakpoint. As breakpoints with a nearby mutation are rare events, Poisson distribution was assumed to infer P-values. When comparing to several samples together (Fig. 4A,C), mutations were aggregated into one virtual sample with all the mutations. To test for the enrichment/depletion of transitions and transversions near breakpoints, we performed a Fisher's exact test for each sample on the number of mutations of each type near breakpoints versus their distribution over all of the genome. A similar Fisher's exact test was used to compare TpC → TpG out of all C↔G transversions, near breakpoints and over all of the genome. Fisher's exact test was also used to compare the mutation enrichment near breakpoints with the enrichment of mutations in other samples of the same tumor type in the same regions. To compute the frequency of mutations over all of the genome and near rearrangements, mutations of each type were counted and divided by the total number of base pairs of the appropriate type that were covered enough to call for mutations.

To estimate the strand specificity of mutations near breakpoints, we examined all of the 10-kb windows around breakpoints that had at least 15 mutations. Mutation rate in the window was calculated on both strands (e.g., C → T and G → A together), and then binomial distribution was used to estimate the probability of having as many mutations on a single strand in that window (e.g., either C → T or G → A).

Software availability

BreakPointer is available at http://www.broadinstitute.org/cancer/cga/BreakPointer.

References

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, et al. 2012. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486:** 405–409.

Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, et al. 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43:** 964–968.

Beale RC, Petersen-Mahrt SK, Watt IN, Harris RS, Rada C, Neuberger MS. 2004. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: Correlation with mutation spectra in vivo. *J Mol Biol* **337:** 585–596.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. 2011. The genomic complexity of primary human prostate cancer. *Nature* **470:** 214–220.

Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, et al. 2012. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485:** 502–506.

Bishop KN, Holmes RK, Sheehy AM, Davidson NO, Cho SJ, Malim MH. 2004. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Curr Biol* **14:** 1392–1396.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40:** 722–729.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML, et al. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467:** 1109–1113.

The Cancer Genome Atlas Network. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487:** 330–337.

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, et al. 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature* **471:** 467–472.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR, et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44:** 390–397.

De S, Babu MM. 2010. A time-invariant principle of genome evolution. *Proc Natl Acad Sci* **107:** 13004–13009.

De S, Michor F. 2011. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol* **29:** 1103–1108.

DePinho RA, Polyak K. 2004. Cancer chromosomes in crisis. *Nat Genet* **36:** 932–934.

Eisen JA, Heidelberg JF, White O, Salzberg SL. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1:** RESEARCH0011.

Farkash-Amar S, Simon I. 2010. Genome-wide analysis of the replication program in mammals. *Chromosome Res* **18:** 115–125.

Fearon ER. 2011. Molecular genetics of colorectal cancer. *Annu Rev Pathol* **6:** 479–507.

Fudenberg G, Getz G, Meyerson M, Mirny LA. 2011. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol* **29:** 1109–1113.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453:** 948–951.

Guerrero AA, Gamero MC, Trachana V, Futterer A, Pacios-Bras C, Diaz-Concha NP, Cigudosa JC, Martinez AC, van Wely KH. 2010. Centromere-localized breaks indicate the generation of DNA damage by the mitotic spindle. *Proc Natl Acad Sci* **107:** 4159–4164.

Hoeijmakers JH. 2001. Genome maintenance mechanisms for preventing cancer. *Nature* **411:** 366–374.

Jansen JG, Langerak P, Tsaalbi-Shtylik A, van den Berk P, Jacobs H, de Wind N. 2006. Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. *J Exp Med* **203:** 319–323.

Kaplan KB, Burds AA, Swedlow JR, Bekir SS, Sorger PK, Nathke IS. 2001. A role for the Adenomatous Polyposis Coli protein in chromosome segregation. *Nat Cell Biol* **3:** 429–432.

Kino K, Sugiyama H. 2001. Possible cause of G-C→C-G transversion mutation by guanine oxidation product, imidazolone. *Chem Biol* **8:** 369–378.

Kino K, Sugiyama H. 2005. UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutat Res* **571:** 33–42.

Klein IA, Resch W, Jankovic M, Oliveira T, Yamane A, Nakahashi H, Di Virgilio M, Bothmer A, Nussenzweig A, Robbiani DF, et al. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147:** 95–106.

Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T. 2007. Dynamic genome architecture in the nuclear space: Regulation of gene expression in three dimensions. *Nat Rev Genet* **8:** 104–115.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.

Mani RS, Chinnaiyan AM. 2010. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat Rev Genet* **11:** 819–829.

Meaburn KJ, Misteli T, Soutoglou E. 2007. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* **17:** 80–90.

Meister P, Taddei A, Gasser SM. 2006. In and out of the replication factory. *Cell* **125:** 1233–1235.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149:** 979–993.

Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, et al. 2012. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46:** 424–435.

Ross AL, Sale JE. 2006. The catalytic activity of REV1 is employed during immunoglobulin gene diversification in DT40. *Mol Immunol* **43:** 1587–1594.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365–386.

Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. *Proc Natl Acad Sci* **106:** 21766–21770.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20:** 761–770.

Scholz FW, Stephens MA. 1987. K-sample Anderson-Darling tests. *J Am Stat Assoc* **82:** 918–924.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195–197.

Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, et al. 2005. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37:** 590–592.

Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, et al. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462:** 1005–1010.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144:** 27–40.

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence M, Sougnez C, McKenna A, et al. 2011. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333:** 1157–1160.

Tillier ER, Collins RA. 2000. Genome rearrangement by replication-directed translocation. *Nat Genet* **26:** 195–197.

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T, et al. 2011. High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* **43:** 464–469.

Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, et al. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8:** 652–654.

Wang J, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, et al. 2011. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365:** 2497–2506.

Yaffe E, Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43:** 1059–1065.

Zhu C, Mills KD, Ferguson DO, Lee C, Manis J, Fleming J, Gao Y, Morton CC, Alt FW. 2002. Unrepaired DNA breaks in p53-deficient cells lead to oncogenic gene amplification subsequent to translocations. *Cell* **109:** 811–821.