From Mutational Mechanisms in Single Cells to Mutational Patterns in Cancer Genomes

Cheng-Zhong Zhang[1,2,3,4] and David Pellman[1,3,4,5]
[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215
[3]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115
[4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142
[5]Howard Hughes Medical Institute, Boston, Massachusetts 02115
Correspondence: chengz@broadinstitute.org; david_pellman@dfci.harvard.edu

Analysis of mutations in thousands of cancer genomes has revealed many characteristic patterns of mutagenesis. The search for the molecular mechanisms underlying these mutational patterns has not only generated novel biological insight but also led to the development of new experimental strategies to study cell-to-cell variation and genome evolution. In this essay, we discuss recent progress in the study of mutational mechanisms with a particular emphasis on the analysis of mutagenesis at the single-cell level.

Large-scale cancer genome analyses have greatly expanded the knowledge of somatic mutations. An average cancer genome contains about 10^3–10^4 point mutations, 10–10^2 small insertions or deletions, and 1–10 large-scale chromosome rearrangements (including copy-number alterations) (Alexandrov et al. 2013; Garraway and Lander 2013; Kandoth et al. 2013; Lawrence et al. 2013; Vogelstein et al. 2013; Zack et al. 2013; Martincorena and Campbell 2015). As most cancers are driven by a limited number (three to six) of oncogenic driver mutations (Davoli et al. 2013; Vogelstein et al. 2013), the majority of the observed mutations are passengers (i.e., mutations not under positive selection). These passenger mutations reflect the signatures of different mutational processes that are active during tumor development (Alexandrov et al. 2013; Alexandrov and Stratton 2014; Helleday et al. 2014).

Each mutation is a specific alteration of the nucleotide sequence at a particular position in the genome. Thus, mutational signatures are described both by the type of nucleotide alterations and by their distribution across the genome (Fig. 1A). The type of nucleotide alterations reflects the molecular process of mutagenesis, whereas the distribution of mutations informs on how chromatin affects the activity of a given mutational process. We hereafter refer to the classification of mutations based on the nucleotide changes as the "spectrum" and refer to the distribution of the positions of mutations as the "distribution." We first discuss the genome-wide patterns of mutations.

GLOBAL PATTERNS OF SOMATIC MUTATIONS

If mutagenesis is completely random, mutations will be uniformly distributed, and the number of events in each category of nucleotide change (e.g., C>T) should be proportional to the nucleotide content in the genome (i.e., the number of potential substrates for a given type of nucleotide change). We now know that neither holds true. The deviation of the observed mutational patterns from the distribution expected from completely random mutagenesis is either due to selection or is generated by different mechanisms of DNA damage and/or DNA repair (Watson et al. 2013; Helleday et al. 2014).

Global Patterns of Point Mutations

For point mutations (including small insertion or deletion events), selection makes a minimal contribution to the overall pattern of these events because the majority of mutations do not occur in protein-coding genes (~1% of the genome) and are thus under minimal selection. Furthermore, even mutations that do alter coding sequences show a similar frequency of synonymous and nonsynonymous substitutions, suggesting that they are by and large not under strong selection (Rubin and Green 2009). Mutations in cancer genes are more frequently nonsynonymous, but such events only make a small fraction of the total set of mutations because the number of cancer genes is limited. (Estimates for the number of cancer genes range from ~200 [Vogelstein et al. 2013; Lawrence et al. 2014] to ~500 [Davoli et al. 2013; Forbes et al. 2015].) Therefore, most point mutations in cancer genomes are passengers and their patterns reflect the underlying mutational processes.

The distribution of point mutations shows several patterns of regional variation. First, the frequency of mutations is negatively correlated with the expression level of genes (Pleasance et al. 2010; Lawrence et al. 2013),

Copyright © 2015 Cold Spring Harbor Laboratory Press; all rights reserved; doi: 10.1101/sqb.2015.80.027623

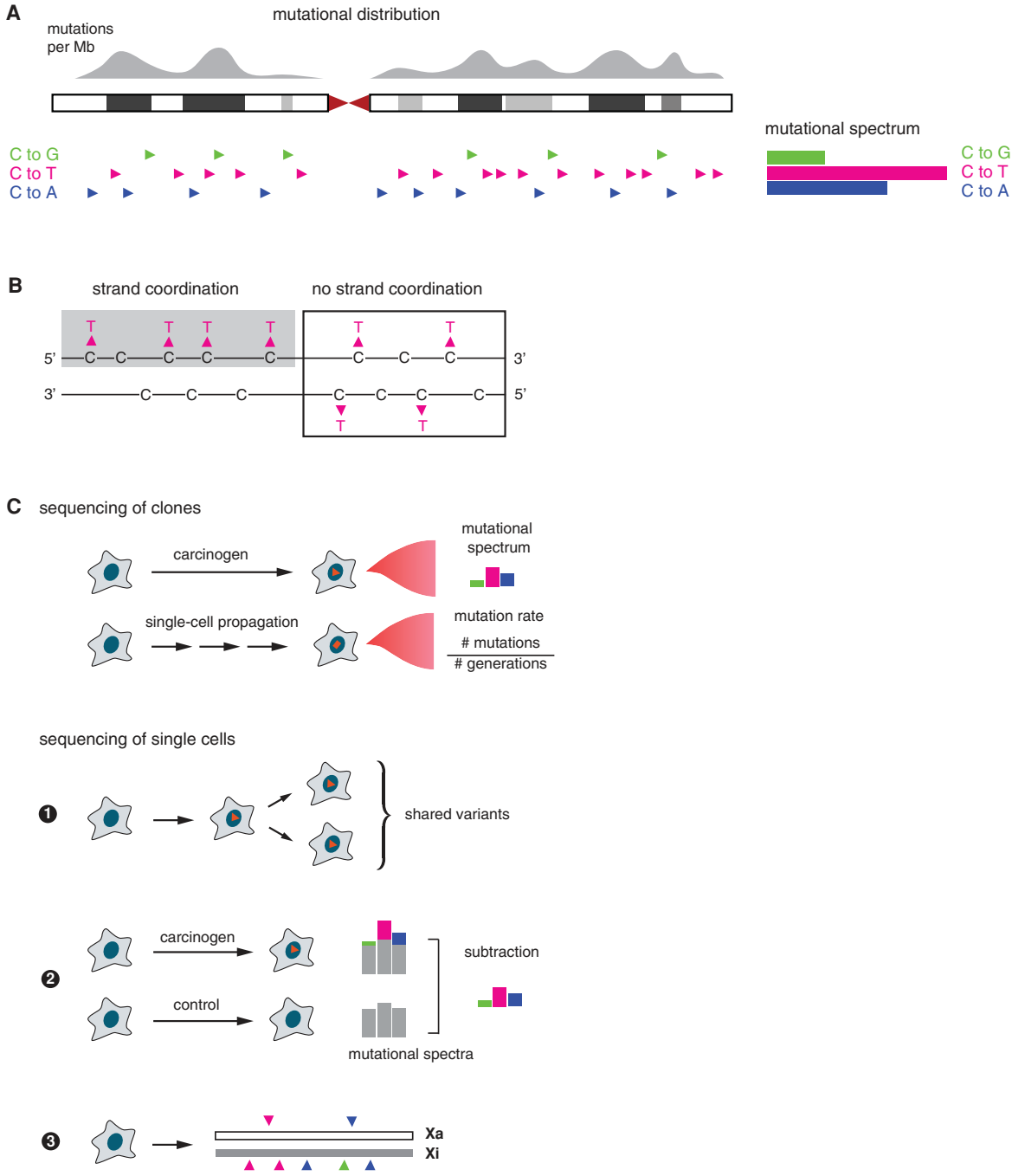

Figure 1. Genome-wide patterns of point mutations. (*A*) Mutational distribution and mutational spectrum. For the analysis of mutational spectra, point mutations are classified based on the 5′ and 3′ flanking bases (16 different contexts); nucleotide changes such as C>T in ApCpA and G>A in TpGpT are counted only in one representative category. (*B*) Strand-coordinated mutations generate opposite spectra on opposite strands (even though the nucleotide content is similar on both strands). In the example shown here, one strand is dominated by C>T transitions and the other strand is dominated by G>A transitions. Because the nucleotide content is similar on both strands, a given mutational process causing C>T transitions would affect both DNA strands, generating roughly equal numbers of C>T transitions (G>A on the opposite strand). The strand asymmetry either reflects asymmetric resection on each side of a double-strand break, asymmetry between the leading and lagging strands during DNA replication, or asymmetry between the transcribed and nontranscribed strands (Haradhvala et al. 2016). (*C*) Different strategies to study mutagenesis. Mutations accumulated in a single cell can either be detected by sequencing of the clonal progeny ("sequencing of clones") or by single-cell sequencing ("sequencing of single cells"). Amplification of single-cell genomic DNA generates artifacts. Random amplification artifacts can be controlled by sequencing a pair of daughters and only including shared mutations. It is also possible to compare the spectrum of de novo mutations in a cell that is subject to a particular mutational process (e.g., exposure to a carcinogen) with the spectrum of a control cell. Mutagenesis differentially affecting the different homologs (e.g., between the active X chromosome [Xa] and the inactive X chromosome [Xi]) can also be analyzed if de novo variants can be directly associated with a homologous chromosome ("haplotype phasing").

which is attributed to transcription-coupled DNA repair that suppresses mutations in highly expressed genes (Fousteri and Mullenders 2008; Hanawalt and Spivak 2008). Second, at the megabase scale, the frequency of mutations is positively correlated with the relative timing of DNA replication (Woo and Li 2012; Lawrence et al. 2013; Liu et al. 2013) and the level of the heterochromatin-associated histone modification H3K9me3 (Schuster-Böckler and Lehner 2012). Consistent with this observation, the very late-replicating inactivated X chromosome frequently undergoes hypermutation (Jäger et al. 2013) with a mutational frequency that is at least twice the frequency on the active X chromosome. An enrichment of mutations in late-replicating, heterochromatic regions is also observed in the germline (Stamatoyannopoulos et al. 2009; Chen et al. 2010; Hodgkinson et al. 2012). One postulated mechanistic explanation for these observations is that these regions experience replication stress in late S phase and accumulate more errors. Consistent with this hypothesis, late-replicating regions, including fragile sites (Glover et al. 2005), are also enriched for chromosomal translocations and copy-number losses (Debatisse et al. 2012; Donley and Thayer 2013; Mankouri et al. 2013). Another nonexclusive possibility is that certain DNA repair mechanisms may be more effective in euchromatic regions with better access to the repair factors. This proposal is supported by the finding that the frequency of mutations in mismatch-repair deficient tumors is less correlated with late replication timing (Supek and Lehner 2015). Finally, epigenomic features, including chromatin accessibility and replication timing, vary between cells of different tissue origins; accordingly, the distribution of mutations also shows tissue-type specificity both in cancers (Polak et al. 2015) and in somatic cells (Behjati et al. 2014; Martincorena et al. 2015). This tissue-type specificity of the mutational distribution can be used to directly identify the cell type from which a tumor is derived (Polak et al. 2015). This finding also implies that epigenetic remodeling in cancer genomes mostly occurs after the accumulation of mutations.

Computational analysis of the spectra of point mutations has also revealed different mutational signatures (Nik-Zainal et al. 2012; Alexandrov et al. 2013). Some of the signatures are associated with well-defined endogenous or exogenous mechanisms. For example, C>T transitions at CpG sites (underlined nucleotides are mutated) are correlated with patient age and are attributed to spontaneous deamination of 5-methylcytosine (Alexandrov et al. 2013, 2015); exposure to ultraviolet radiation leads to C>T transitions at dipyrimidines (CpC or TpC) and exposure to tobacco smoke causes frequent C>A transversions (Lawrence et al. 2013; Alexandrov and Stratton 2014; Helleday et al. 2014; Martincorena et al. 2015; Nik-Zainal et al. 2015). For some signatures, underlying mechanisms are postulated but not conclusively established. For example, two signatures involving C>T or C>G substitutions in a TpC context match the motif of deamination by the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) enzymes (Harris et al. 2002), suggesting an association with APOBEC activity (Nik-Zainal et al. 2012; Roberts et al. 2013). This association has gained support from the observation that at least one member of the APOBEC family, APOBEC3B, is up-regulated in many types of cancer (Burns et al. 2013b; Swanton et al. 2015), and the experimental finding that the expression level of APOBEC3B is correlated with the number of C>T transitions (Burns et al. 2013a). Nevertheless, a direct demonstration that APOBEC3B activity is responsible for this mutational signature in cancer is still lacking. Many mutational signatures have no clear biological explanation (Alexandrov et al. 2013). These signatures are either due to unknown carcinogens or reflect unidentified endogenous mutational processes.

DNA sequencing can directly determine whether a mutational process generates a specific mutational signature. As the same mutational process generates different mutations in different cells, most mutations are only detectable at the single-cell level, either by sequencing a single-cell genome directly or by sequencing a single-cell derived clone.

Sequencing of single-cell derived clones can detect mutations present in the ancestor cell without the confounding problem of potential artifacts stemming from amplifying the genome of a single cell. In a proof-of-principle study, Meier et al. (2014) analyzed *Caenorhabditis elegans* clones that were exposed to carcinogens causing different types of DNA lesions (including aflatoxin, mechlorethamine, and cisplatin) or were propagated over multiple generations in genetic backgrounds with different DNA repair deficiencies (Fig. 1C). The authors found that mutations generated by different carcinogens had characteristic mutational spectra that can be readily explained by the chemistry of these reagents. With impaired nucleotide excision repair or under higher concentrations of the carcinogen, the mutation burden increased but the mutational spectra were preserved. This latter result showed that the spectra of mutations generated by the carcinogens are quite specific. Similar findings were also obtained by selecting immortalized mouse fibroblasts with Trp53 inactivation after exposure to various mutagens (Nik-Zainal et al. 2015).

To accurately infer the genome-wide distribution of mutations, a sufficient number of mutations are needed. For example, to accurately infer the frequency of mutations at the megabase scale it is necessary to have on average $\geq$10 mutations per megabase. (The signal-to-noise ratio is approximately $N/\sqrt{N}$ for a Poisson process with a mean of N events. For an average of 10 mutations per megabase, the signal-to-noise ratio is $\sim$3; whereas one mutation per megabase gives a signal-to-noise ratio of $\sim$1.) Generating such a high mutation load (10^4–10^5 mutations in a human genome of 3.2 Gb) in a single cell takes many generations and may impair cell viability and clonal expansion. A more suitable strategy is to generate mutations independently in multiple ancestor cells and infer the mutational frequency from the clonal progeny. The ultimate solution is to directly profile mutations present in single cells, bypassing clonal expansion and the confounding effects of selection.

The main challenge in single-cell sequencing is the presence of amplification errors that can outnumber true de novo mutations. Estimates for the frequency of single-cell amplification errors range from 10^{-4} to 10^{-6} (Paez et al. 2004; Zong et al. 2012; Voet et al. 2013; de Bourcy et al. 2014; Wang et al. 2014; Lodato et al. 2015). This will result in 10^3–10^5 false variants in a diploid human cell consisting of 6.4 billion base pairs. (For reviews of single-cell DNA sequencing methods and applications, see Kalisky et al. 2011; Lasken 2012; Blainey 2013; Van Loo and Voet 2014; Wang and Navin 2015.) Besides improving the fidelity of single-cell amplification, two strategies can be used to control for amplification errors (Fig. 1C). Progeny cells descended from a common ancestor can serve as biological replicates to identify mutations present in the ancestor cell (Zong et al. 2012). With two or three replicates, the frequency of random errors could be brought down to 10^{-8}–10^{-12} (assuming the error frequency in a single amplification to be 10^{-4}). It may also be possible to directly infer the spectrum of true de novo mutations after subtracting the spectrum of amplification errors. Although a systematic analysis of single-cell amplification errors is not available, a recent study showed that for postmitotic neurons, the amplification errors have a very different spectrum than true variants (Lodato et al. 2015). In particular, in vitro amplification errors should be similar between the homologous chromosomes and between the forward and the reverse DNA strands. This feature can be used to characterize mutagenesis with strand coordination (e.g., when mutations primarily accumulate on the nontranscribed DNA strand or on the lagging strand during DNA replication [Fig. 1B; Haradhvala et al. 2016]) or mutagenesis showing homolog asymmetry, such as hypermutation in the inactivated X chromosome (Fig. 1C).

In summary, somatic point mutations in cancer genomes show heterogeneity both in their distribution across the genome and in the type of nucleotide change. Mutations are enriched in regions that are late-replicating and heterochromatic. The mutational spectra reflect the outcomes of different mutagens and/or effects of endogenous DNA metabolism. Both the distribution and the spectrum of mutations show a certain level of tissue specificity. It is therefore possible to gain information on the cell type of origin by directly analyzing the pattern of somatic mutations. This provides a strategy to infer the cancer type from the spectrum of somatic mutations in blood biopsy samples, such as circulating tumor cells or cell-free DNA (Haber and Velculescu 2014). Finally, genome-wide mutagenesis assays are starting to validate the patterns and reveal new features of many mutagens and mutational processes.

Because base substitutions are more abundant than insertions/deletions (indels), the above results are primarily derived from the patterns of base substitutions. The relative abundance of indel variants and substitution variants is uncertain. Comparative genomic analyses suggested that indels may contribute up to 20% of genetic variants (Dawson et al. 2001; Mills et al. 2006; Cartwright 2009). But human population studies by short-read sequencing identified only 3% of variants as indels (1000 Genomes Project Consortium 2012). This discrepancy reflects a high rate of false indel detection due to inaccurate short-read alignment that also affects indel detection in cancer genomes. Because of the uncertainty in indel detection, we restrict our discussion to two classes of indels that are relatively frequent in cancer. One major class of indels is expansion or contraction of short nucleotide repeats ("microsatellites"). These regions are highly polymorphic in the germline (Willems et al. 2014) and frequently mutated in cancers with defective DNA mismatch repair (Boland and Goel 2010). Several patterns have emerged from the analysis of microsatellite mutations in mismatch-repair deficient tumors (Kim et al. 2013; Zhao et al. 2014). Indels are more frequent than substitutions (~4:1 ratio) and the majority are deletions (~81%) (Zhao et al. 2014). Interestingly, microsatellite alterations are more frequent in euchromatic domains than heterochromatic domains and are depleted in regions with nucleosome occupancy (Kim et al. 2013; Zhao et al. 2014). One hypothesis is that faster replication fork progression in euchromatic and nucleosome-depleted regions leads to more slippage errors during DNA replication (Kim et al. 2013). Another class of indels is deletions (<50 bp) with microhomology at the breakpoint junctions. These are frequently observed in cancers with inactivating mutations in *BRCA1* and *BRCA2*. This mutation signature is attributed to deficiency in homologous recombination and more frequent DNA repair by error-prone end-joining mechanisms (Alexandrov et al. 2013). In general, the frequency of indels rapidly decreases with the length of indel (Cartwright 2009; Zhao et al. 2014). Whereas short indels are mostly generated by DNA replication slippage, large indels can also arise from erroneous repair of DNA breaks. A systematic characterization of indel events in the 10–1000-bp range could provide insight into the relative contributions of these two mechanisms.

Global Patterns of Chromosomal Rearrangements and Copy-Number Alterations

In contrast to point mutations or short insertion/deletion events, structural variants are more difficult to classify both because of the complexity of these alterations and because their fitness consequences are more difficult to predict. Common structural variants in the germline are classified as deletions, tandem duplications, transpositions, inversions, and insertions. This classification is based on the fact that most of these events affect only a small region of the genome and are nonoverlapping. In contrast, cancer genomes often harbor large-scale chromosomal alterations including extensive aneuploidy and complex karyotypes (Fig. 2A). Multiple structural alterations may occur to the same region of a chromosome and the resulting overlap makes it difficult to classify individual events (Fig. 2B). On the phenotypic level, large-scale copy-number alterations and chromosomal rearrangements can cause varying changes in fitness that are tumor

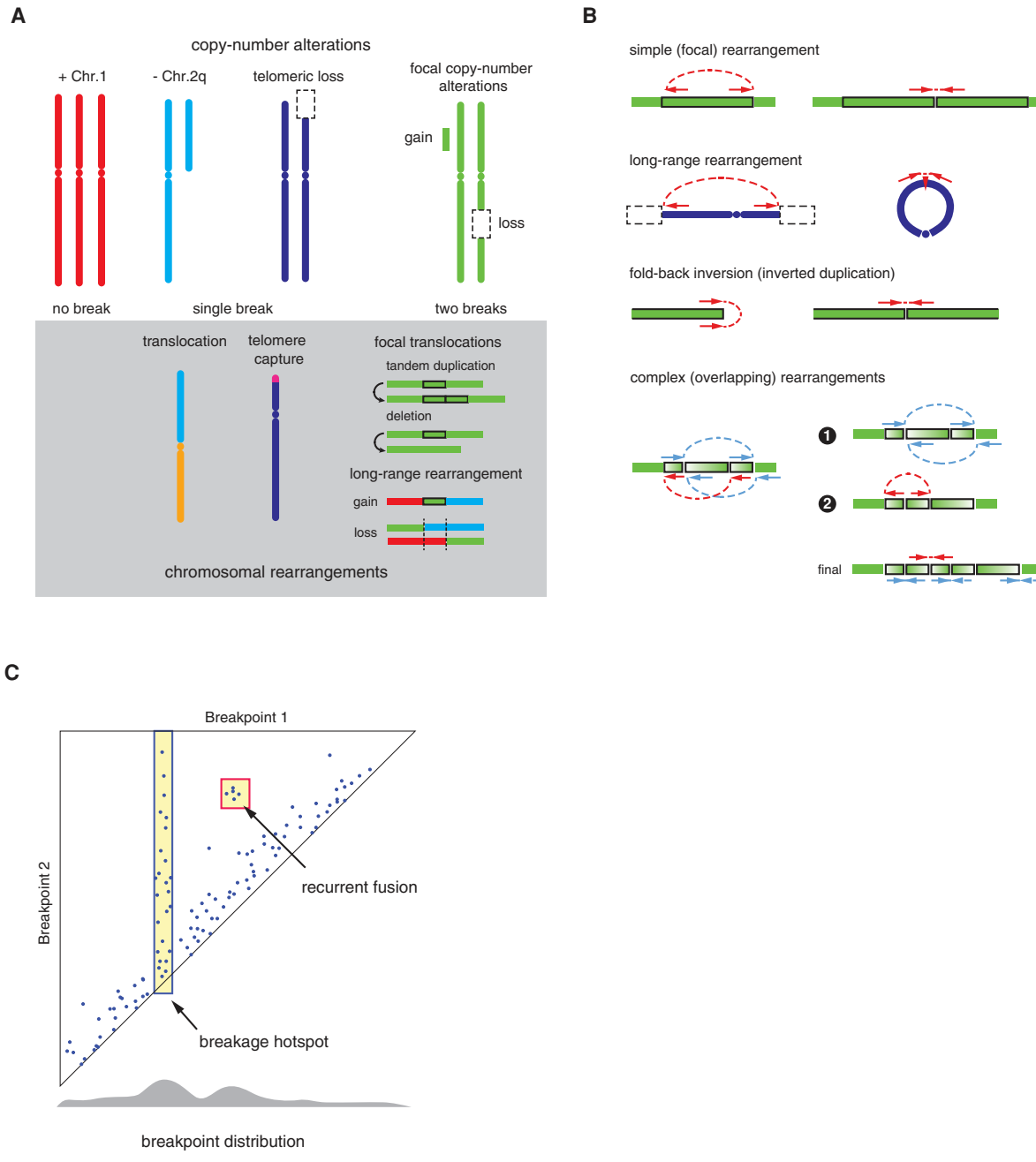

Figure 2. Genome-wide patterns of chromosomal rearrangements. (*A*) Somatic copy-number alterations and the corresponding chromosomal rearrangements. Whole-chromosome aneuploidy does not lead to any chromosomal rearrangement. Arm-level copy-number alterations or telomere-bound copy-number alterations are generated by a single chromosomal break and are associated with a single chromosomal rearrangement. Focal copy-number alterations have two chromosomal breaks at the boundaries; these events can arise from either focal rearrangements, such as tandem duplications or deletions, or long-range rearrangements. (*B*) Simple (focal) and complex rearrangements. A rearrangement is focal when the two ends of the rearranged segment are fused together and the segment does not contain a centromere. For example, a focal "head–tail" connection (red arrows) indicates a tandem duplication (*top*); if the segment contains a centromere, the same "head–tail" connection indicates a ring chromosome (*top middle*). (*Middle*) A fold-back inversion is generated by a fusion between sister chromatids, resulting in an inverted duplication. (*Bottom*) When multiple events overlap, individual events can only be classified after the order of these events is determined. In the example shown here, three inverted-type rearrangements overlap. One possible deconvolution is shown on the *right*: a simple inversion (two discordant pairs in blue) followed by a tandem duplication (discordant pair in red). In this scenario, the red pair represents a tandem duplication even though it has a "head–head" orientation when mapped to the reference genome. (*C*) Rearrangements in cancer genomes can be represented as a two-dimensional scatter plot based on the breakpoint positions. The density (or probability) of rearrangements between a pair of loci is determined by the frequency of breakage at each locus and the contact probability between these two loci. As the contact probability decays rapidly as the two loci are further apart, the majority of rearrangements are between adjacent loci (close to the diagonal). Recurrent oncogenic fusions are identified as a tight cluster (red square); breakage hotspots that can fuse with many other loci manifest as a vertical stripe (blue rectangle).

type–dependent (Varetti et al. 2014; Santaguida and Amon 2015).

Because of the complexity of somatic chromosomal alterations, a comprehensive analysis of their mutational patterns will require a large number of samples. Here, we review the current knowledge of the patterns of somatic structural variants in cancer genomes from the analysis of relatively small cohorts. These conclusions will be refined by ongoing large-scale studies including the analysis of whole-genome sequencing data in the International Cancer Genome Consortium (2010) (https://icgc.org/) and the pan-cancer analysis of The Cancer Genome Atlas (http://cancergenome.nih.gov/).

The first major class of chromosomal alterations is whole-chromosome aneuploidies, including copy-number changes of single-chromosome arms (Fig. 2A; Beroukhim et al. 2010; Zack et al. 2013). These events account for the majority of copy-number alterations. Whole-chromosome aneuploidy occurs much more frequently than intrachromosomal copy-number alterations of a similar length. This reflects distinct mechanisms generating these events. Whole-chromosome aneuploidy is generated by chromosome missegregation events, which occurs more frequently than the fusion of two distal chromosomal breaks resulting in intrachromosomal translocations (Ganem and Pellman 2012). The prevalence of aneuploidies of single-chromosome arms further suggests that centromeres may be more prone to breakage. Little is known about centromeric breaks, as current genomic analysis cannot reveal the exact breakpoints within the centromeric repeats. Finally, because whole-chromosome aneuploidies span hundreds or thousands of genes, it is difficult to assess their net fitness effects (Torres et al. 2010; Davoli et al. 2013). Some aneuploidies are recurrent in certain tumor types and almost certainly under positive selection (Gordon et al. 2012). Why these events promote tumorigenesis is a major unsolved puzzle of cancer evolution. It is now possible to study these aneuploidies in vitro by genome engineering (Jiang et al. 2013; Kotini et al. 2015).

The second class of chromosomal alterations is so-called telomere-bound copy-number alterations (Zack et al. 2013): gains or losses of telomere proximal segments including the telomere (Fig. 2A). These events originate from single chromosomal breaks. A survey of these events in cancer genomes revealed that the breaks occur more frequently in subtelomeric regions or near the centromere but are distributed almost uniformly across the rest of the chromosome arm. This pattern cannot be due to selection as breaks further away from the telomere will affect a larger number of genes. Instead, this pattern suggests different mutational mechanisms. Telomere attrition and erosion may extend into the subtelomeric regions and cause frequent breaks in these regions. Telomere loss can further lead to sister-chromatid fusion and dicentric chromosomal bridges (see below for an in-depth discussion on chromosomal bridges). Resolution of these bridges can cause breaks at any locus on the chromosome arm but may be enriched at centromeres because of their fragility.

The third class of chromosomal alterations is chromosomal rearrangements that fuse two internal loci in the genome (Fig. 2B). In the simplest scenario, a rearrangement of a single intrachromosomal segment can be classified as deletion, inversion, or tandem duplication based on the alteration to the DNA sequence. A special subgroup of intrachromosomal rearrangements is fold-back inversions. These events can be generated by a fusion between sister chromatids broken at the same locus (McClintock 1941a,b), or by a "U-turn" of stalled replication forks (Narayanan et al. 2006; Mizuno et al. 2009, 2013). Rearrangements between loci on different chromosomes or different chromosome arms should be classified as long-range as these events often disrupt the chromosomal structure and are accompanied by additional rearrangements. In more complex scenarios when multiple rearrangements overlap, it is necessary to first infer the order of these events before each individual event can be classified.

Because each rearrangement is generated by the apposition of two DNA breaks, the pattern of rearrangements with regard to their locations on the genome is best represented as a two-dimensional distribution based on the location of each break (Fig. 2C). The probability of generating a rearrangement between any given pair of loci is given by the product of the probabilities of breakage at both loci and the probability of their spatial apposition (Zhang et al. 2010, 2012; Hakim et al. 2012). There are several parameters that affect the breakage frequency across the genome. From the analysis of the distribution of rearrangement breakpoints in 95 cancer genomes, Drier et al. (2013) found that breakpoints are enriched either at early-replicating, high-GC content and transcribed regions or at late-replicating, low-GC content and untranscribed regions, depending on the tumor type. This bimodal pattern is interesting as it potentially indicates two mechanisms. Late-replicating regions often contain fragile sites that are difficult to replicate; these regions are frequently underreplicated and deleted under replication stress (Arlt et al. 2012). The enrichment of breaks at transcribed regions was previously uncovered by high-throughput genome-wide translocation capture experiments (Chiarle et al. 2011; Klein et al. 2011) and likely reflects endogenous mechanisms of DNA breakage generation related to gene transcription. Consistent with these models, Drier et al. (2013) found that in samples where breakpoints are enriched in both categories, deletion events are enriched in late-replicating regions but other events are enriched in transcribed regions. A recent study found that the frequency of chromosomal breakage is also modulated by histone modifications. By looking at the histone marks of 74 frequently translocated genes, Burman et al. (2015) identified H3K4me1, H3K4me3, H3K27ac, and DNase I hypersensitivity as being enriched at the translocated genes. H3K4me1 and H3K4me3 are marks for open chromatin and actively transcribed regions, but the frequently translocated genes do not show elevated expression. This led the authors to hypothesize that H3K4 methylation promotes DNA breakage by inducing chromosome decondensation, which was supported by in vitro experiments (Burman et al. 2015).

The other factor contributing to the distribution of rearrangements is the contact probability between two breaks. The contact probability between two loci is determined by the chromatin structure (Meaburn et al. 2007). Interphase chromatin is partitioned into different chromosome territories (Cremer and Cremer 2010), and within each territory, the contact probability between two loci can be measured by genome conformation capture (or "Hi-C") experiments (Dekker et al. 2002; Lieberman-Aiden et al. 2009; Rao et al. 2014). The initial Hi-C experiments suggested that the contact probability between two loci on a single chromosome is inversely proportional to their separation on the chromosome (Lieberman-Aiden et al. 2009). This scaling relationship of contact probability is similar to the probability of de novo translocations between two loci estimated from in vitro translocation capture experiments (Klein et al. 2011; Zhang et al. 2012) and to the distribution of somatic copy-number alterations of different lengths (Beroukhim et al. 2010; Zack et al. 2013). Recent studies of chromatin structure further revealed that chromosomes are compartmentalized into topologically associating domains (TADs) spanning hundreds of kilobases (Dixon et al. 2012; Dekker et al. 2013) with distinct histone modifications (Rao et al. 2014). Loci in the same TAD interact frequently, but loci from different TADs do not; the contact probability therefore follows different scaling relationships (Sanborn et al. 2015). The higher-resolution map of the chromatin structure should now enable a more systematic analysis of the mechanistic factors contributing to somatic chromosomal rearrangements.

In summary, somatic chromosomal alterations can be classified as whole-chromosome aneuploidies, telomere-bound copy-number alterations, and chromosomal rearrangements. Compared with the analysis of point mutations, the analysis of chromosomal alterations is preliminary because of the small number of samples with available whole-genome sequencing data. The current analysis suggests that chromosomal breakage occurs frequently at centromeres and is enriched either in early-replicating, gene-rich regions or late-replicating, gene-poor regions, possibly reflecting different underlying mechanisms. A systematic analysis of chromosomal alterations in larger data sets is going to provide a much more detailed landscape of somatic chromosomal rearrangements. It will then be possible to perform correlation analysis of the rearrangement breakpoints with different DNA sequence motifs, epigenetic features, and chromatin structure.

PATTERNS OF LOCALIZED MUTAGENESIS AND THEIR MECHANISMS

Besides regional variation in the mutational frequency, there are several unique mutational phenomena where mutations form dense clusters in small regions of the genome. These include clusters of point mutations (also known as "kataegis") and several types of complex chromosome rearrangements, including chromothripsis, chromoplexy, chromoanasynthesis, and breakage–fusion–bridge (BFB) cycles. In each case, tens or hundreds of mutations are concentrated in local regions spanning 10 kb (for kataegis) to a whole chromosome (chromothripsis or chromoanasynthesis). The concentration of mutations to a single chromosome or a pair of sister chromatids further indicates spatial localization of mutagenesis. This spatial localization is different from mutational hotspots (e.g., fragile sites) that are associated with DNA sequence or epigenomic features. In addition to spatial localization, the clustered mutations also have a unique spectrum. For example, most mutations in kataegis are C>T or C>G substitutions in a TpC context. Rearrangements in chromothripsis, chromoanasynthesis, and chromoplexy all have random orientations, but they are accompanied by predominantly segmental losses in chromothripsis, segmental gains in chromoanasynthesis, or minimal copy-number changes in chromoplexy. In contrast, BFB patterns are characterized by a series of inverted duplications. Both the spatial localization of mutagenesis and the unique mutational spectra suggest that these mutational phenomena are not generated by random mutagenesis but reflect novel mutational mechanisms.

Kataegis

Kataegis was first reported in breast cancers (Nik-Zainal et al. 2012). Multiple mutational clusters, each spanning 0.1–1 kb but consisting of 10–100 mutations (a dramatic mutational frequency of 10^{-1}–10^{-2} per base pair), are interspersed in regions with low mutational frequency (10^{-5}–10^{-6}). Such concentration far exceeds regional variation (<10-fold) in mutational frequency. The individual mutational clusters are dominated by C>T or C>G substitutions (G>A or G>C on the opposite strand), and show strand coordination (Fig. 1B). (For example, mutations in each cluster are either mostly C>T mutations or mostly G>A mutations, but not a mixture.) This has led to the hypothesis that each mutational cluster resulted from concurrent DNA damage on single-strand DNA (ssDNA). This hypothesis is further supported by the observation that a fraction of kataegis is found near rearrangement breakpoints, where ssDNA can result from resection during the repair of DNA double-strand breaks (Fig. 3A). The unique mutational spectrum of mutations matches the motif of cytosine deamination by the APOBEC family enzymes. Together, the genomic observation suggests that kataegis is generated DNA damage on single-strand DNA, most likely through the activity of APOBEC enzymes.

Experimental studies in yeast have recapitulated several features of kataegis. Roberts et al. (2012) showed that it is possible to generate localized mutation clusters showing strand coordination by chronic, low-dosage exposure to methyl methanesulfonate (MMS). Various reporter constructs containing multiple adjacent selection markers were introduced to select clones that need closely spaced mutations in these markers for drug resistance. In the surviving clones, the mutations required for drug resistance were accompanied by many other adjacent muta-

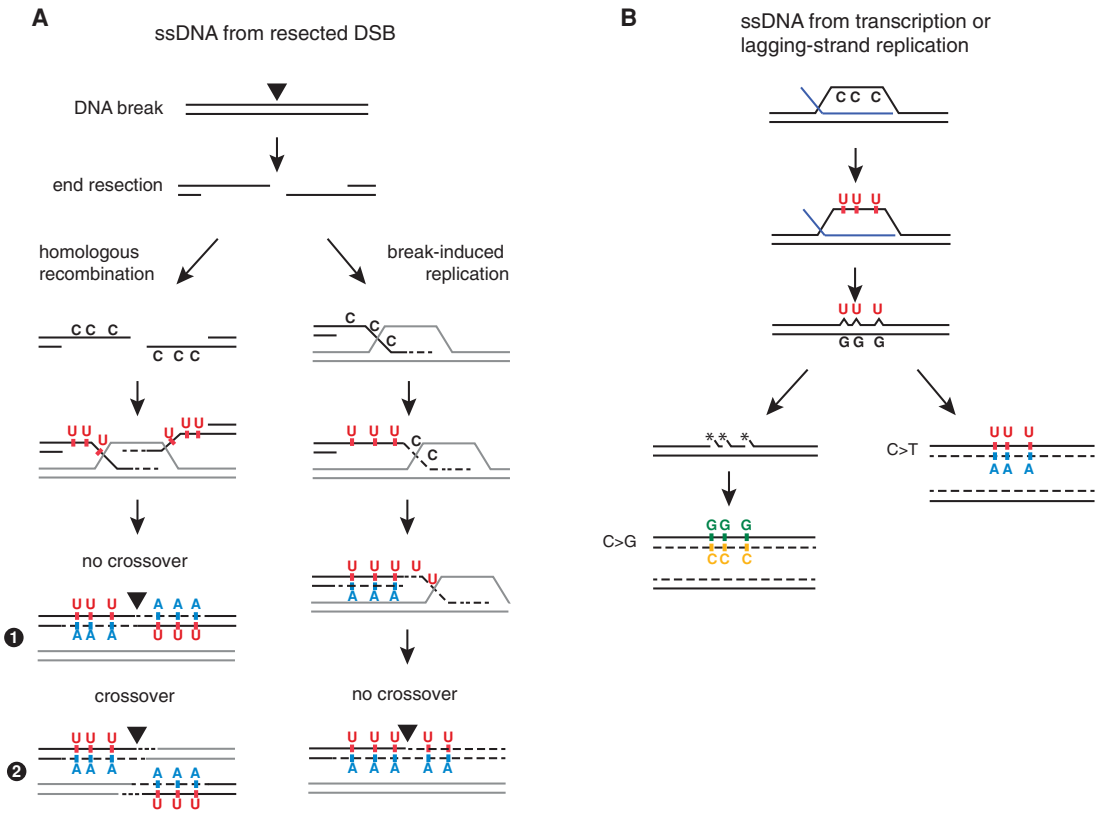

Figure 3. Potential mechanisms generating kataegis, strand-coordinated mutational clusters. (*A*) Kataegis on resected double-strand breaks (DSBs). Deamination of cytosines (e.g., by APOBEC enzymes) on single-strand DNA can be converted to U:A pairs (causing C>T mutations) after homologous recombination or break-induced replication. If the two breaks are repaired by homologous recombination, the C>T transitions occur on opposite strands across the break. The opposite strand coordination is either observed in a single daughter or separately in two daughters, depending on whether there is crossover between the sister chromatids. If a single break undergoes break-induced replication, the cluster of C>T transitions can extend beyond the breakpoint with no switching of strand coordination. (*B*) Kataegis on single-strand DNA during gene transcription or DNA replication. Deamination of cytosines generates uracils and U:G mispairs. If left unrepaired, the U:G mispairs can be converted to U:A pairs (or C>T transitions) during DNA replication. The uracil bases can also be removed by base-excision repair. The abasic sites after uracial excision may be filled with guanines on either strand during DNA replication (Taylor et al. 2013), generating a C>G transversion.

tions. The density of these clustered mutations is more than 500-fold higher than the average frequency of non-clustered mutations. In a subsequent study, Taylor et al. (2013) recapitulated the same features without selection by introducing a hyperactive activation-induced cytidine deaminase (AID). In addition to C>T transitions resulting from replication over uracils generated by cytidine deamination, the authors also observed clustered C>G transversions that were likely generated by uracil excision during base-excision repair (Fig. 3B). Both studies indicated that DNA breaks followed by 5′-to-3′ end-resection during homologous DNA repair could generate single-strand DNA templates that accumulate the clustered mutations. Error-prone break-induced replication, including potential template-switching events (Smith et al. 2007; Hicks et al. 2010), could also contribute to localized mutagenesis around DNA breaks. Sakofsky et al. (2014) showed that break-induced replication in the presence of alkylating DNA damage can generate mutation clusters that are accompanied by further DNA breakage and rearrangements.

The above studies suggest that single-strand DNA is a major substrate for generating clustered mutagenesis (Chan and Gordenin 2015). Several questions remain to be addressed. First, the studies in yeast were all performed in conditions where elevated mutagenesis was artificially induced, either by an alkylating agent (Roberts et al. 2012; Sakofsky et al. 2014) or a hyperactive deaminase (Taylor et al. 2013). It remains to be determined if the patterns of clustered mutations observed in cancer genomes can be generated in a more physiologic context. In particular, it is uncertain if the clustered mutations were generated episodically through a transient burst of mutagenesis or gradually over multiple generations. Second, the molecular mechanisms that convert damage on single-strand DNA into mutations on both DNA strands are unclear. Normal DNA replication is semiconservative but break-induced replication is conservative (Donnianni and Symington 2013; Saini et al. 2013). The most straightforward strategy to distinguish between the contributions of these two replicative modes to kataegis is to compare the patterns of mutations in daughter cells after a single-cell division

(Fig. 3A). Strand-coordinated mutations followed by semiconservative replication should result in mutation clusters in both daughters but on opposite strands (Fig. 3A, left); in contrast, conservative break-induced replication will result in identical mutations in both daughters if it occurs in G_1 or in one daughter only if it occurs in G_2 (Fig. 3A, right). Finally, the current models associating kataegis with DNA breaks are mostly based on replicative mechanisms including error-prone translesion synthesis and break-induced replication. However, it is unclear what fraction of somatic rearrangements in cancer are generated by these mechanisms. Based on the junction sequences, it has been suggested that the majority of somatic translocations may be generated by end-joining mechanisms (Yang et al. 2013), including nonhomologous end joining and microhomology-mediated alternative end joining. Resolving this discrepancy requires an in-depth analysis of kataegis and somatic rearrangements in cancer genomes as well as a better understanding of their mechanisms.

Chromothripsis

Chromothripsis is characterized by extensive rearrangement of a single chromosome, a chromosomal segment, or occasionally a few chromosomes, accompanied by interspersed DNA deletions (Stephens et al. 2011; Korbel and Campbell 2013). Similar patterns were also identified in congenital disorders in which clustered rearrangements were accompanied with little DNA loss (Kloosterman et al. 2011; Chiang et al. 2012). The rearrangement pattern led to a hypothetical model that the affected chromosome had been fragmented into many pieces, followed by random ligation (Stephens et al. 2011). A compelling biological mechanism for chromothripsis was provided by the partitioning of lagging chromosomes into abnormal nuclear structures called micronuclei (Crasta et al. 2012; Leibowitz et al. 2015). Micronuclei have multiple defects that cause delayed DNA replication, impaired nuclear import, and accumulation of DNA damage (Crasta et al. 2012; Hatch et al. 2013). After mitosis, damaged chromatids from micronuclei can be reincorporated into the daughter cells' main nuclei, where DNA damage on the micronucleated chromatid can potentially be converted into extensive rearrangements via different DNA repair pathways (Fig. 4A).

To test this model, we recently performed single-cell DNA sequencing to show that DNA damage accumulated in micronuclei can cause a range of rearrangements on the lagging chromosome, including chromothripsis (Zhang et al. 2015b). Micronuclei were identified by live-cell imaging but a critical challenge in the genomic analysis was to infer the sequence identity of the missegregated chromosome in the micronucleus. An important hint came from the cytological observation that the chromosome in the micronucleus was underreplicated; this implies that the micronucleated chromosome would undergo asymmetric segregation to the daughter cells, resulting in uneven DNA copy number for this chromosome in the daughter cells. The copy-number asymmetry was validated by the sequencing results, which further showed that the chromosomes in micronuclei were severely underreplicated. Furthermore, by comparing the sequence coverage at heterozygous sites, we were able to determine the DNA copy number of both parental homologs and uniquely identify the homolog that was partitioned into the micronucleus (Fig. 4B–D). With the identity of the micronucleated chromatid known, it then became straightforward to test whether there is a significant concentration of chromosomal rearrangements occurring on this chromosome. We found that chromosomal rearrangements were significantly enriched on the micronucleated chromatid and only on this chromatid. Uneven segregation and extensive rearrangements both occur on a single chromatid, leaving the other homolog intact. This asymmetry between a pair of homologous chromosomes was not observed in any daughter pairs of mothers without micronuclei and cannot be due to single-cell amplification artifacts as homologous chromosomes have almost identical DNA sequences and will be subject to similar amplification bias. Moreover, chromosome fragmentation, manifested by reciprocal distribution of chromosome segments between both daughter cells, was observed in three cases; in one case, fragmentation resulted in more than 40 segments being reciprocally distributed, exactly matching the copy-number pattern in chromothripsis. Together, these results showed that DNA damage from micronuclei can generate chromothripsis.

The combination of live-cell imaging and DNA sequencing of daughter pairs after a single-cell division is a powerful approach to relate phenotype and genotype. By comparing the sequence coverage at heterozygous sites, we were able to determine the integer copy number for each parental homolog and identify the missegregated chromosome from copy-number asymmetry. The missegregated chromatid monitored by live-cell imaging can therefore be directly assigned a specific parental haplotype (the genotypes at heterozygous sites). Mutations phased to this haplotype are directly related to DNA damage and mutagenesis on the missegregated chromosome. Importantly, the intact homolog and the normally segregated chromosomes in the same cell serve as a control for sequence-dependent artifacts generated during whole-genome amplification, sequencing, or the bioinformatic analysis.

Whole-genome sequencing not only represents a high-throughput assay for mutagenesis and DNA damage but also generates base-level resolution of the rearrangement junctions that can provide insight into the mechanisms of DNA repair. The majority of rearrangement junctions show little or no homology (<6 bp), consistent with the possibility of ligation of the DNA breaks by an end-joining mechanism (Stephens et al. 2011). Strikingly, several rearrangements on the micronucleated chromosomes contained insertions of short templated sequences, which is a feature frequently associated with template-switching events during break-induced replication (Hastings et al. 2009). Thus, at least two mechanisms of DNA repair may have been involved in the generation of translocations.

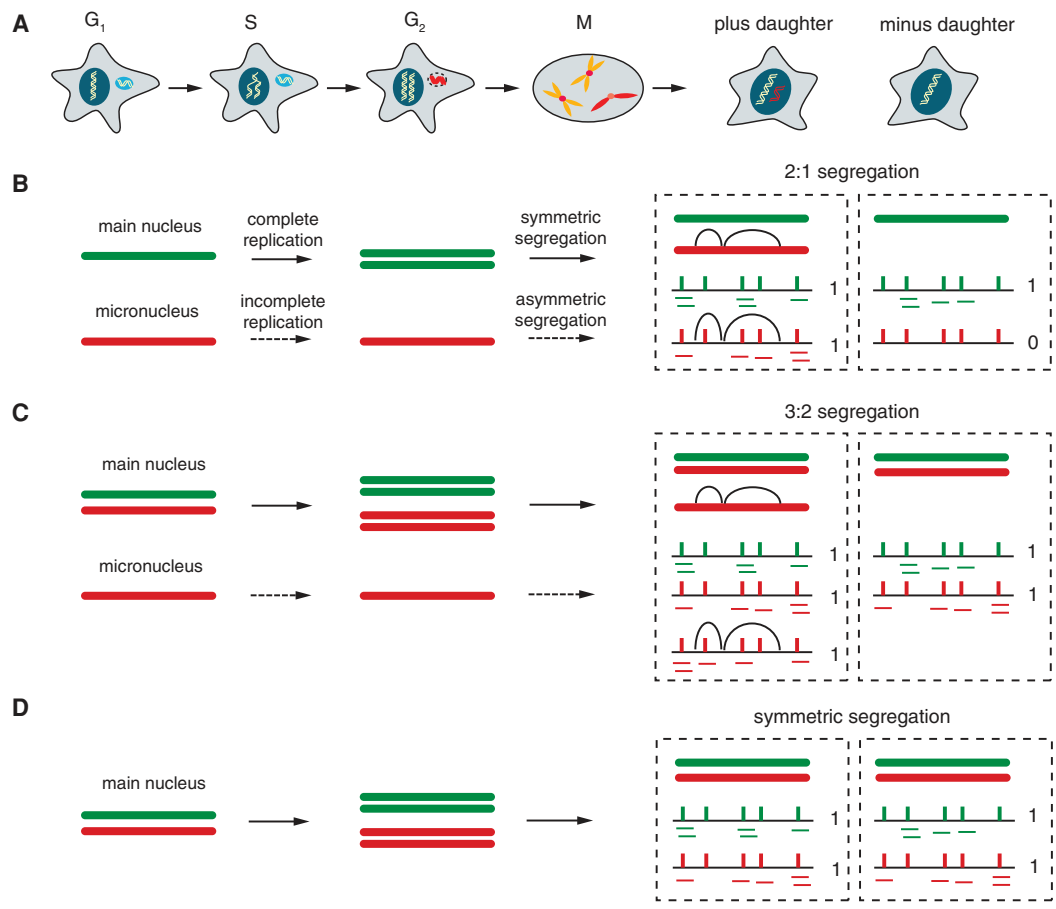

Figure 4. Characterization of chromosome missegregation and DNA damage by haplotype-based single-cell genomic analysis. (*A*) Chromothripsis from DNA damage in micronuclei. DNA damage can accumulate in micronuclei because of either defective DNA replication or irreversible rupture of the nuclear envelope; after mitosis, the damaged chromatid can be reincorporated into the daughter cells' primary nuclei, generating extensive rearrangements. (*B*,*C*) Single-cell sequencing enables identification of the missegregated chromatid by haplotype DNA copy number and characterization of the DNA damage by chromosomal rearrangement detection. The missegregated chromatid in the micronucleus undergoes incomplete replication and segregates asymmetrically into the daughters. This asymmetry is observed only for the missegregated chromatid (red) but not for the intact homolog (green). Chromosomal rearrangements (links) are further associated with the red homolog, indicating that DNA damage is accumulated on the missegregated chromatid. (*D*) Control chromosomes in the main nucleus are evenly segregated and do not contain rearrangements.

It is now possible to measure the contribution of these two mechanisms by analyzing the rearrangement patterns after one or both of them are inhibited.

Chromothripsis may also be generated by mechanisms other than the formation of micronuclei. Li et al. (2014) reported that chromothripsis could also result from dicentric chromosomes that form bridges at mitosis (Fig. 5). This association has gained further support from the analysis of chromosomal bridges induced by telomere attrition. Mardin et al. (2015) induced telomere attrition by siRNA (small interfering RNA) knockdown of the shelterin complex protein TRF2 in hTERT (human telomerase reverse transcriptase)-immortalized RPE-1 cells and found two transformed clones containing chromothripsis. Maciejowski et al. (2015) generated chromosomal bridges by telomere crisis induced by a dominant-negative *TRF2* mutation. Postcrisis clones frequently showed chromothripsis. (Of 10 clones selected for telomeric fusion and karyotypic abnormalities, five harbored chromothripsis.) Interestingly, kataegis was also frequently observed in the postcrisis clones, often near rearrangement breakpoints. The generation of kataegis was consistent with the detection of single-strand DNA at the center of chromosomal bridges by replication protein A (RPA) accumulation. Single-strand DNA formation was dependent on TREX1, a cytoplasmic exonuclease that also contributed to the bridge resolution. Taken together, resolution of dicentric chromosomal bridges by TREX1 nuclease activity is proposed to generate single-strand DNA and chromosome fragmentation, which can lead to kataegis and chromothripsis (Fig. 5, left scheme).

Chromoanasynthesis

Chromoanasynthesis is characterized by clusters of copy-number gains that are hypothesized to have been generated by microhomology-mediated break-induced

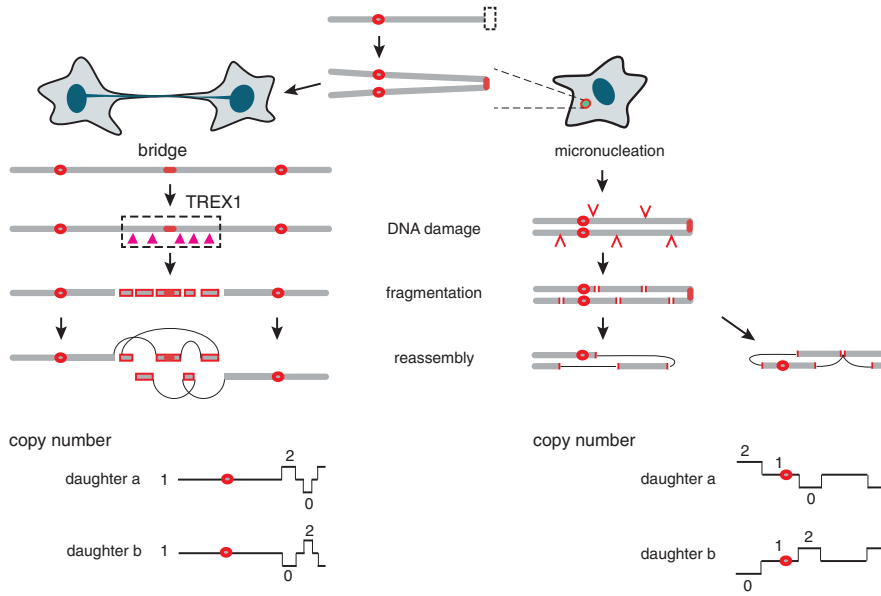

Figure 5. Chromothripsis from dicentric chromosomal bridges. (*Left*) A dicentric chromosomal bridge can persist into interphase and then undergo fragmentation, possibly because of cytoplasmic nucleases such as TREX1 (Maciejowski et al. 2015). This will generate chromothripsis that is localized to the subtelomeric region, but with three copy-number states. (*Right*) A dicentric chromosome can also missegregate into a micronucleus, where it undergoes fragmentation. In this scenario, chromothripsis will affect the whole chromosome and can also generate three copy-number states. In both scenarios, the daughter cells should have mirror-image copy-number profiles.

replication (MMBIR) (Fig. 6; Hastings et al. 2009; Liu et al. 2011). Break-induced replication (BIR) appears to be an important mechanism in restarting stalled or broken replication forks but can also generate complex rearrangements (Smith et al. 2007; Anand et al. 2013; Mayle et al. 2015). BIR requires Pol32 in yeast or POLD3 in mammalian cells (Costantino et al. 2014; Minocherhomji et al. 2015). The mechanistic link between replication fork stalling and chromoanasynthesis gained support from the *C. elegans* sequencing study by Meier et al. (2014). The authors found that exposure to DNA-cross-linking reagents can generate rearrangement patterns that resemble chromoanasynthesis. Notably, the clonal analysis in this study likely only revealed a limited fraction of potential mutational outcomes as progeny survival is severely affected by the cross-linking reagent treatment. Single-cell experiments may provide a more direct and comprehensive analysis of the mutational outcome.

Chromoplexy

Chromoplexy is a chain of translocations where the translocation loci are pairs of adjacent DNA breaks (Baca et al. 2013). In contrast to chromothripsis or chromoanasynthesis where a large number of rearrangements are restricted to one or a few chromosomes, chromoplec-

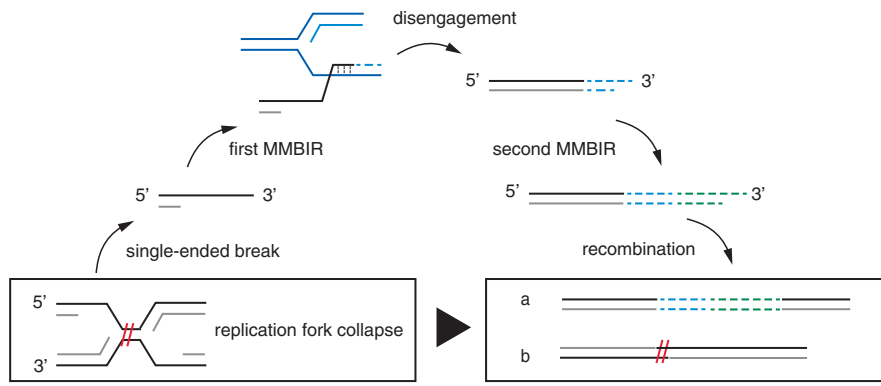

Figure 6. Chromoanasynthesis by microhomology-mediated break-induced replication (MMBIR). Replication fork collapse results in a single-end DNA break, with a 3′ overhang that can invade other sites such as another replication fork. Such invasions are thought to undergo low-processivity DNA replication, with frequent switching of templates. After multiple MMBIR events, the break can re-establish the replication fork or recombine with other breaks. As MMBIR occurs during replication, the rearranged chromosome should be inherited only by one daughter cell.

tic chains often contain fewer translocations that span multiple chromosomes. The sparseness of breakpoints in each chromosome contrasts with the clustering of breakpoints in local regions. The biological mechanism of chromoplexy is unclear. The spatial clustering of chromoplectic breaks from multiple chromosomes led to the proposal that these breaks were generated altogether in interphase nuclei when different chromosomes can interact (Baca et al. 2013). The high frequency of chromoplexy in prostate cancers further led to the hypothesis that chromoplectic breaks were related to the activity of androgen-dependent transcription (Lin et al. 2009; Mani et al. 2009; Haffner et al. 2010). It is now possible to test this model by simultaneously analyzing the genome and the transcriptome of individual prostate cancer cells (Dey et al. 2015; Macaulay et al. 2015) under androgen stimulation. If this model holds, then translocations should occur more frequently at or near genes that are highly expressed. Finally, it will be interesting to determine whether the association of chromoplectic breaks with active transcription is also observed in other cancer types and offers an explanation for the observed enrichment of translocation breaks in transcribed regions (Drier et al. 2013).

Breakage–Fusion–Bridge Cycles and the Resolution of Dicentric Chromosomal Bridges

Although chromothripsis, chromoanasynthesis, and chromoplexy have different rearrangement patterns, the spatial localization of rearrangements is thought to reflect the generation of multiple breaks in a single-cell cycle. In contrast, breakage–fusion–bridge (BFB) cycles were originally proposed to arise from recurrent breakage of dicentric chromosome bridges over multiple cell divisions (McClintock 1939, 1941b). The original BFB model predicted that, after breakage of the dicentric chromosomal bridge and DNA replication, the replicated broken ends in sister chromatids can fuse together and generate a fold-back inversion. Multiple BFB cycles will generate an array of fold-back inversions with amplifications. Notably, the first fold-back inversion should be farthest away from the centromere and always have a "tail–tail" orientation, whereas subsequent fold-back events can have either "tail–tail" or "head–head" orientations (Fig. 7A).

BFB cycles will generate a hierarchy of palindromic structures separated by fold-back inversions (Zakov et al. 2013; Greenman et al. 2015). One such example is the

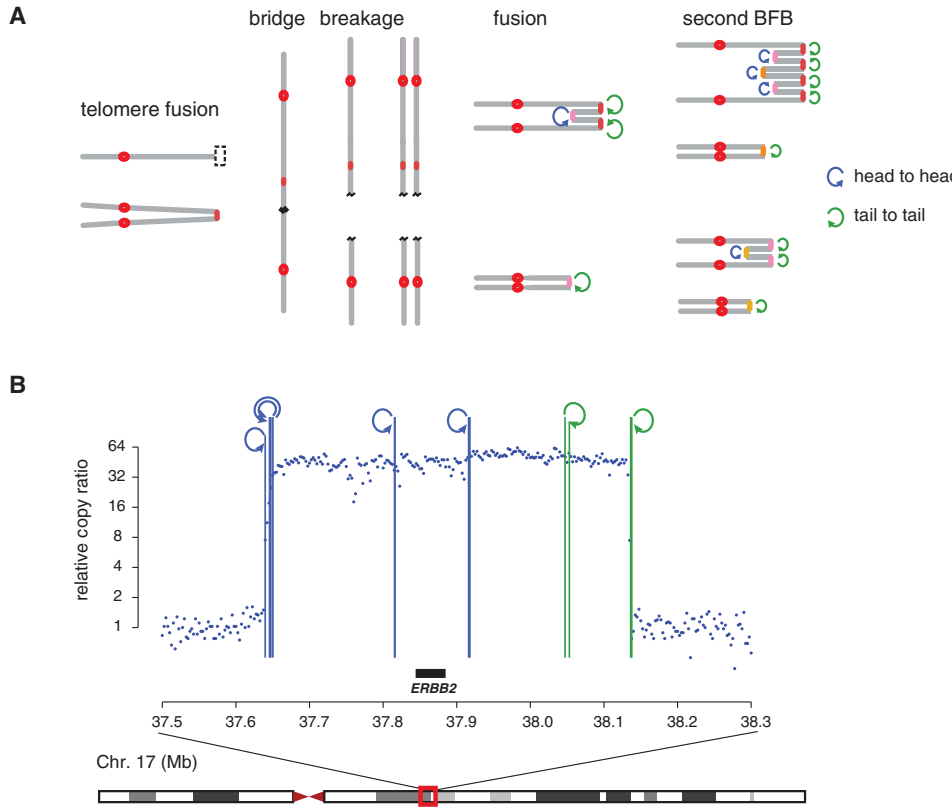

Figure 7. Breakage–fusion–bridge (BFB) cycles in cancer genomes. (*A*) The classical model of BFB cycles assumes that a series of sister-chromatid fusions and single breakage events generate multiple fold-back inversions of "head–head" (blue) or "tail–tail" orientations (green). Although BFB cycles can generate both terminal duplication and terminal loss, only the amplification preserves the fold-back inversion footprint of BFBs. (*B*) The *ERBB2* amplicon in the HCC1954 breast cancer cell line as an example of classical BFB amplification (Bignell et al. 2007; C-Z Zhang, M Imielinski, J Wala, et al., unpubl. data). A series of BFBs results in a ~ 40-fold amplification of this region relative to the flanking region, resulting in >150 copies of the *ERBB2* gene. The rightmost boundary at ~38.14 Mb marks the likely site of the first sister-chromatid fusion event as this fusion is the most amplified.

amplicon at 17q12 in a breast cancer cell line HCC1954 (Fig. 7B; Bignell et al. 2007). This amplicon consists of multiple inverted duplications and contains *ERBB2*, a well-known oncogene in breast cancer. The extremely high DNA copy number (>150) indicates that the BFB palindrome may have undergone subsequent amplifications. BFB cycles are also frequent in pancreatic ductal adenocarcinoma (Campbell et al. 2010; Waddell et al. 2015), and they are inferred to be early events during tumorigenesis. In experimental models of gene amplification induced by selection, BFB amplification is a frequent initiating event (Ma et al. 1993) and can be induced by DNA breakage (Tanaka et al. 2002) or telomere loss (Lo et al. 2002). Together, the cancer sequencing and the in vitro experimental studies suggested that BFB cycles occur early during genome evolution and can cause further genomic instability.

Although dicentric chromosomal bridges are frequently observed in cancer cell lines (Gisselsson et al. 2000), perfect BFB palindromes consisting only of fold-back inversions are surprisingly uncommon in cancer genomes analyzed so far. This could be partially due to difficulties in detecting fold-back inversions by short-read sequencing. But in most cases, fold-back inversions are interspersed with other long-range translocations; these long-range translocations "break" the perfect palindromic structure. One possibility is that the BFB cycles and the long-range translocations occurred at different time points but at the same locus (Fig. 8). For example, a dicentric chromosome resulting from an interchromosomal translocation could initiate BFB cycles that span the translocation junction. One possible example is shown in Figure 8A (C-Z Zhang, M Imielinski, J Wala, et al., unpubl. data). In this example, the translocation is amplified by subsequent BFB cycles that generate fold-back inversions flanking the translocation junction. It is also possible that a long-range translocation occurring after the BFB cycles could disrupt the palindromic structure. Figure 8B shows one example of BFB cycles followed by a chromothripsis that generated interspersed deletions in the BFB amplicon. Another scenario would be that the BFB cycles and the long-range translocations are mechanistically linked and thus concurrent both in space and in time. This latter possibility is supported by recent studies suggesting that dicentric chromosomal bridges can lead to either BFB cycles or chromothripsis or both (Li et al. 2014; Maciejowski et al. 2015; Mardin et al. 2015).

When BFB cycles overlap with other long-range translocations, it is possible to infer the chronological order of these events by analyzing the pattern of copy-number changes at translocation sites. Using this strategy, Li et al. (2014) showed that BFB cycles are often the earlier events that trigger other complex rearrangements including chromothripsis. This pattern was also seen in the evolution of ring chromosomes (Garsed et al. 2014), although it is possible that BFB cycles alternate with chromothripsis. The same pattern was observed in experimental studies of the mutational outcomes of dicentric chromosomal bridges and telomere attrition. Meier et al. (2014) analyzed *C. elegans* clones with homozygous mutations in MRT-2, the *C. elegans* RAD1 subunit of the 9-1-1 complex. (The 9-1-1 complex plays critical roles in homologous recombination and DNA damage response and is required for telomere maintenance.) MRT-2 deficiency led to complex rearrangements at the ends of chromosomes that included fold-back inversions indicative of BFB cycles. The BFB cycles were inferred to have been terminated by simultaneous acquisition of multiple rearrangements that resemble chromothripsis. In the study by Mardin et al. (2015), BFB cycles and chromothripsis were induced by either telomere attrition or doxorubicin treatment that induces DNA double-strand breaks. The sequencing data also suggested the dicentric chromosome initially underwent a few BFB cycles, which were presumably concluded by more catastrophic chromosome fragmentation.

The above results suggest that BFB cycles may occur quite frequently during genome evolution but are not commonly observed in the clonal tumors for several reasons. First, BFB amplification may not generate a net gain of fitness and thus will be lost during clonal expansion. Second, BFB cycles generate dicentric chromosomes as intermediates; the unstable dicentric chromosome may undergo chromothripsis and rearrangements that disrupt the perfect palindromic structure. This could either occur by chromothripsis within the bridge or through the partitioning of the whole dicentric chromosome into a micronucleus (Fig. 5). Finally, the induction of telomere crisis could generate massive instability through the loss of telomeres from multiple chromosomes in the same cell. Such instability may cause concurrent BFB cycles, inter- and intrachromosomal translocations, and chromothripsis involving multiple chromosomes.

EXPERIMENTAL STRATEGIES FOR STUDYING MUTAGENESIS

A critical challenge for the field is to define specific mutational signatures for different mutational processes. This has traditionally relied on reporter genes that select for specific mutational outcomes at a defined locus. For example, amplification of the *DHFR* transgene is selected in cells exposed to methotrexate (Alt et al. 1978); hypermutation can be selected by independent mutations occurring in multiple linked reporters (Roberts et al. 2012). It is also possible to study mutagenesis associated with specific break-induced recombination events (Malkova and Haber 2012). For example, using overlapping inactive reporter gene fragments, one can select cells in which the gene fragments are recombined in a predefined manner. The junction DNA sequence can then be used to identify mutations that co-occur with the recombination event (Hicks et al. 2010; Sakofsky et al. 2014). This provides mechanistic insight into the underlying mutational processes by analyzing cells with deficiencies in different DNA repair pathways (Hu et al. 2013; Costantino et al. 2014). Reporter gene strategies thus provide detailed information on the routes to a specific, predetermined outcome. The use of reporters enables rare events

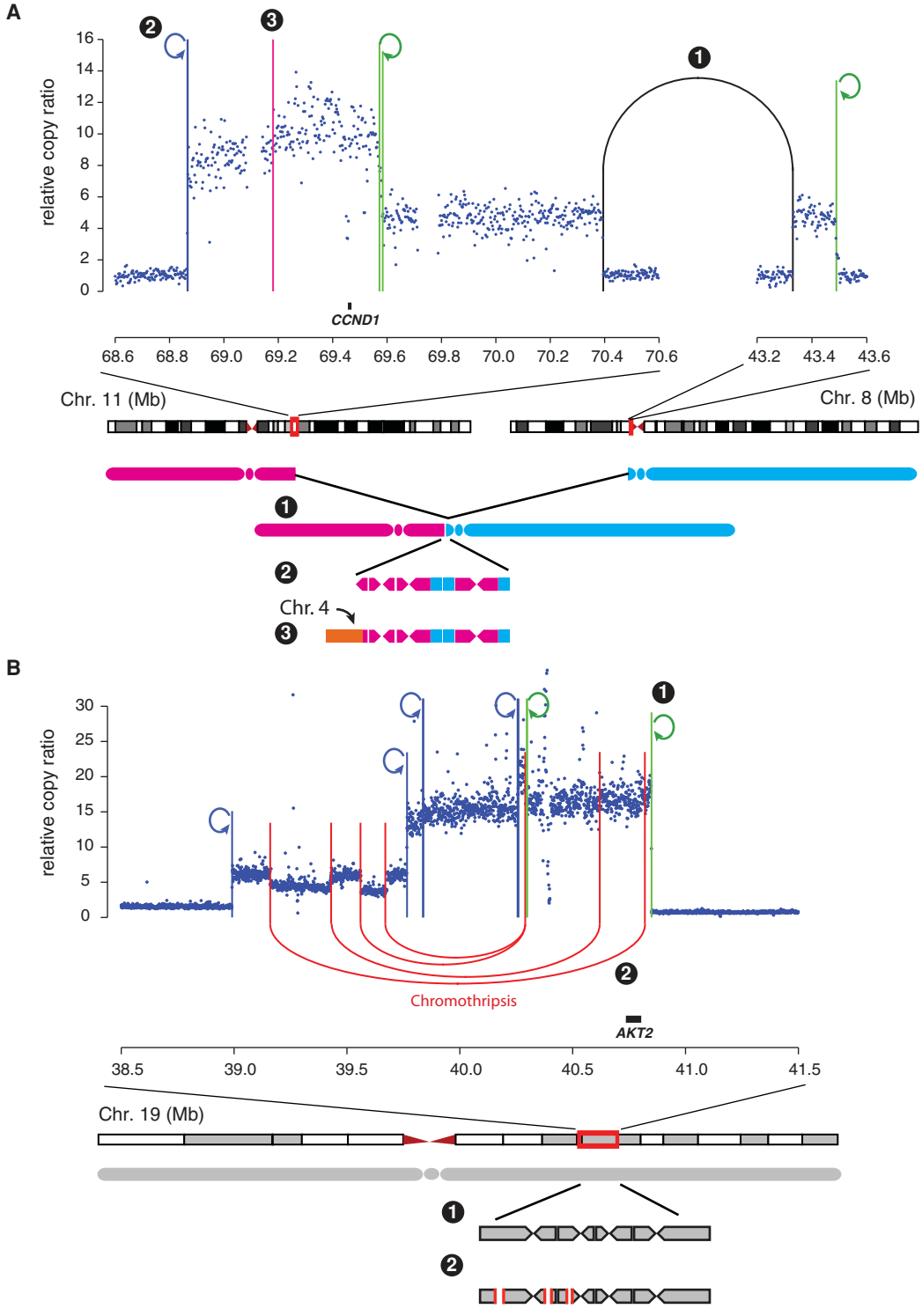

Figure 8. Breakage–fusion–bridge (BFB) cycles overlapping with other rearrangements. (*A*) The *CCND1* amplicon in the HCC1143 breast cancer cell line as an example of BFB amplification occurring after a long-range translocation (C-Z Zhang, M Imielinski, J Wala, et al., unpubl. data). A translocation between Chr. 11 and Chr. 8 is followed by at least two rounds of BFB cycles. One hypothetical model is that the fusion between the 11q arm and the 8p arm (step 1) could have resulted in a dicentric chromosome that initiated the BFB cycles (step 2) and was concluded by a final translocation to Chr. 4 (red line, step 3). (*B*) An example of BFB amplification followed by a chromothripsis event in the PANC-1 pancreatic cancer cell line (C-Z Zhang and M Meyerson, unpubl. data). The chromothripsis most likely occurred after all the BFB cycles as it creates interspersed single-copy DNA losses on different basal copy-number states. It is interesting that chromosomal fragmentation and the BFB amplifications are both restricted to a small region (<2 Mb) in the chromosome. As a matter of fact, McClintock already noted that dicentric bridges in *Zea mays* frequently break at the position of previous fusions (see the Discussion at the end of McClintock 1941b). This might be because the BFB palindrome being at the center of the dicentric bridge is more prone to breakage (see Fig. 9B).

to be selected and analyzed; this is powerful but obviously constrained by knowledge of the selected events.

Alternatively, mutational outcomes can be studied by genome sequencing. Whole-genome sequencing can simultaneously profile many mutations generated by the same process, from which the mutational signatures can be inferred (Fig. 1C). For mutational processes with a constant mutation rate, one can determine both the mutation rate and the mutational signature by analyzing de novo mutations accumulated over a defined number of generations (Taylor et al. 2013; Meier et al. 2014).

In addition to gradual mutagenesis with a consistent rate, it has been recognized that cancer evolution also involves episodic periods of mutagenesis, producing large-scale chromosomal alterations (Wang et al. 2014) or various forms of local hypermutation (Zhang et al. 2013; Chan and Gordenin 2015). Such mutational phenomena are best studied at the single-cell level for the following reasons. First, episodic mutagenesis can disrupt a large number of genes or generate extensive DNA damage, generating negative fitness effects that compromise clonal analysis. Lethal or severely compromised mutational outcomes can only be probed at the single-cell level (Vanneste et al. 2009; Voet et al. 2011). Second, a single event can cause a series of mutagenesis in subsequent cell cycles. To disentangle this kind of complexity, it is necessary to study mutagenesis over a single cell cycle. Finally, live-cell imaging can directly relate the mutational mechanism to the mutational outcomes revealed by DNA sequencing (Zhang et al. 2015b).

Detecting Mutations in Single Cells

The primary challenge in analyzing mutagenesis at the single-cell level is to account for single-cell whole-genome amplification artifacts. This can be overcome by two strategies. The first strategy is to use sibling cells as biological replicates. True mutations will be shared in the progeny but random de novo amplification artifacts will not. This approach has been extensively used to validate mutations shared by a small number of somatic cells (Fig. 1C; Lohr et al. 2014; Wang et al. 2014; Lodato et al. 2015), or by a pair of daughters (Zong et al. 2012; Voet et al. 2013). A second strategy, which we recently developed, controls for amplification artifacts by linking genetic mutations to a specific haplotype (Fig. 9A; Zhang et al. 2015b). Here we focus on the second strategy for its general applicability to the detection of de novo mutations present only in a single cell.

The main idea of the haplotype-phasing approach is to determine whether each homologous chromosome has a mutant or wild-type genotype by phasing sequencing reads covering a given locus to each parental haplotype. A true de novo mutation generated on a single chromosome will be present in all sequencing reads derived from this chromosome. Similarly, sequencing reads from the wild-type chromosome should all show the wild-type genotype. In contrast, amplification errors will accumulate in a fraction of sequencing reads derived from one chromosome (Fig. 9A). The all-or-none assignment of mutant or wild-type genotype to each homolog can therefore control both false-positive and false-negative variant detection.

Haplotype phasing of mutations requires long reads covering both the mutated base and adjacent heterozygous sites. In human genomes, the average density of heterozygous sites is about one per 2 kb (Sachidanandam et al. 2001; Wheeler et al. 2008; 1000 Genomes Project Consortium 2012). The average spacing is well beyond the read length of current high-throughput sequencing platforms (100–300 bp). Application of the haplotype-phasing strategy thus requires long-range Sanger sequencing, which is low-throughput and more suitable as a validation strategy. To generate long sequencing fragments including both the mutated base and adjacent heterozygous variants also requires larger amplicons, such as those generated by multiple-displacement amplification (MDA) (Evrony et al. 2012; Zhang et al. 2015a). For PCR-based amplification methods, the 1–2-kb amplicon size is too short (Navin et al. 2011; Zong et al. 2012). Haplotype phasing of de novo mutations is also becoming feasible with long-read sequencing technologies or diluting strategies (Snyder et al. 2015). The key idea is to dilute long genomic DNA molecules into subhaploid fractions and determine the haplotype phase of DNA fragments in each dilution by sequencing. For single-cell sequencing, it should be possible to apply the same strategy to the long genomic DNA fragments generated by MDA to directly phase the haplotypes of these fragments.

Like genuine point mutations, true copy-number alterations affect a single homolog whereas amplification artifacts will affect both homologs. This means that if the haplotypes are known, by taking the ratio of coverage for each haplotype, regional variation in amplification efficiency will be normalized away (Zhang et al. 2015b). In contrast, true copy-number alterations will be present at an integer copy ratio (Fig. 4B–D). This normalization strategy should in principle eliminate all systematic, sequence-dependent amplification noise, but not random, sequence-independent noise. The random amplification noise can be reduced by performing a window-based average of the read depth signal before the normalization of coverage between homologs.

Calculation of the coverage for each homolog requires knowledge of the whole-chromosome haplotype, which requires complete segregation of sequencing reads from each parental chromosome. This can be accomplished by isolating individual chromosomes and performing single-chromosome sequencing (Fan et al. 2011). It is also possible to induce chromosome missegregation and generate random monosomies in a population of dividing cells and perform single-cell sequencing on monosomic cells to obtain whole-chromosome haplotypes (Zhang et al. 2015b).

Several strategies can be applied to analyze de novo rearrangements in single cells and distinguish them from artificial chimeric sequences generated during single-cell genome amplification. Rearrangements that lead to copy-

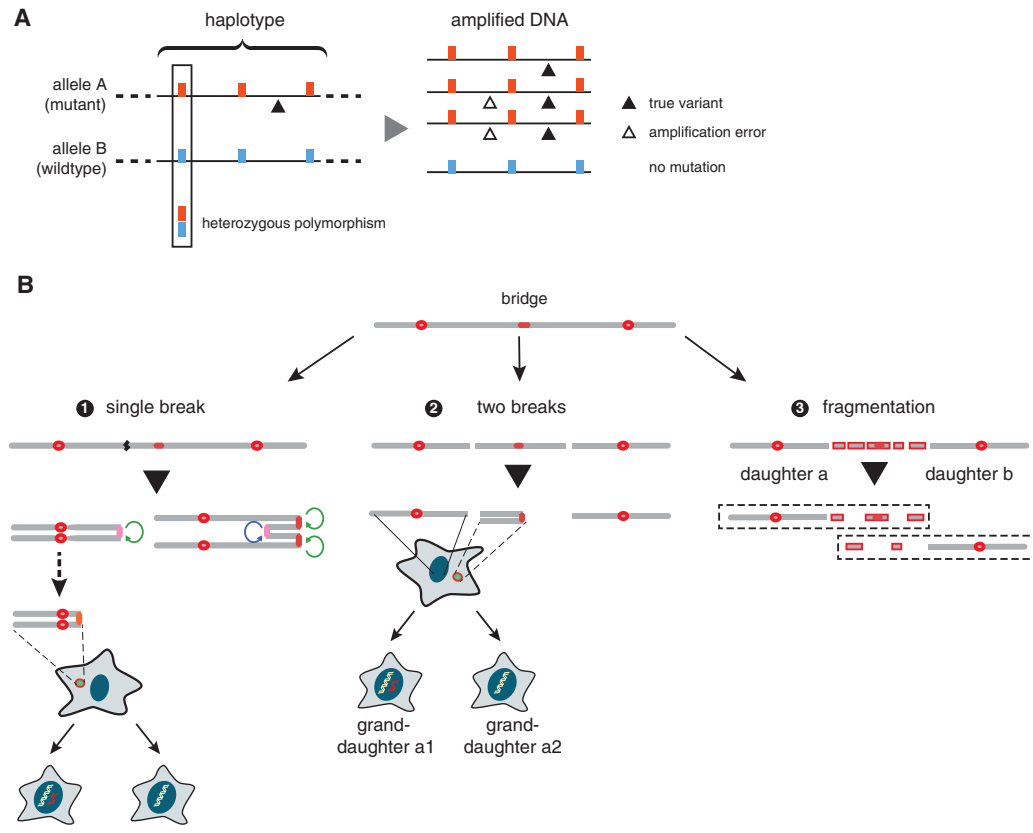

Figure 9. Strategies to study mutagenesis by single-cell DNA sequencing. (*A*) Single-cell whole-genome amplification generates errors (open triangles) and creates uneven representation of the two homologs. The error base is only present in a fraction of amplified DNA; in contrast, the true variant (filled triangle) is present in all amplified DNA. This provides a strategy to control for false-positive variant detection. Furthermore, if the haplotypes of both homologs can be determined, then all the mutations at this locus should be identifiable from the different haplotypes. For example, the blue homolog does not contain any mutation at this locus. This strategy also controls for false-negative variant detection when there is no copy-number amplification of either homolog. (*B*) Different mutational outcomes of a dicentric bridge. A dicentric bridge can be resolved by (1) a single break leading to a BFB cycle, (2) two breaks generating an acentric fragment, or (3) multiple breaks causing chromosome fragmentation. Bridge fragmentation may generate chromothripsis in the subtelomeric region in both daughters. Chromothripsis in the subtelomeric region can also result from the acentric terminal fragment being partitioned into a micronucleus. Finally, BFB cycles can cause successive DNA loss and eventually generate a pseudodicentric chromosome with adjacent centromeres functioning as a single unit. Incorrect merotelic attachment can cause this chromosome to be partitioned into a micronucleus and undergo chromothripsis. By combining live-cell imaging and single-cell sequencing, it should in principle be possible to resolve these different possibilities and the frequencies of these events.

number alterations can be validated by copy-number analysis (Voet et al. 2013; Francis et al. 2014). Because the majority of artificial chimeras in single-cell genome amplification are created between loci within a short distance from each other (Lasken and Stockwell 2007; Evrony et al. 2015; Zhang et al. 2015b), it is possible to control for such chimeras in the study of long-range chromosomal rearrangements. Finally, haplotype phasing can also serve as a validation strategy for chromosomal translocations (Zhang et al. 2015b).

Characterize Mutagenesis at the Single-Cell Level

In the study of mutagenesis, single-cell sequencing enables a direct comparison of genetic mutations in sibling cells. The resulting mirror-image genetic alterations can provide powerful validation of genetic events and resolve the relationship between mutagenesis and DNA replication. Below we propose a few ways that single-cell analysis might bring unique insight into the study of mutational mechanisms.

Our first example would be to use single-cell sequencing to distinguish different mechanisms causing kataegis. One proposed model for kataegis posits that single-strand DNA resulting from resected double-strand breaks accumulates extensive DNA damage that is subsequently converted into strand-coordinated mutations (Roberts et al. 2012; Sakofsky et al. 2014). If the two break ends are recombined using the sister chromatid as a template, then the mutations will have opposite mutational signatures across the break (e.g., C>T on one side of the break, and G>A on the other side [Fig. 3A, left scheme]). One daughter will inherit the mutated chromosome, but the other daughter will only have the intact sister chromosome. If there is a crossover between the sister chromatids at the recombination site, then the two daughters will each inherit half of the mutations. If a single break end undergoes a translocation or break-induced replication,

the mutations occur on a single strand and are inherited by a single daughter only (Fig. 3A, right scheme). An alternative model would be that kataegis occurs on single-stranded DNA without a break, such as during gene transcription (Taylor et al. 2014; Haradhvala et al. 2016). Under this model, after semiconservative DNA replication, one daughter cell will have kataegis with single-strand coordination (but no switching of strand orientation), and the other daughter will not harbor these mutations (Fig. 3B).

Second, single-cell analysis might be able to resolve different mechanisms generating segmental copy-number gains (Figs. 5 and 6). Chromoanasynthesis is hypothesized to be generated when a single-ended break resulting from a collapsed replication fork invades other replication forks and undergoes multiple microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009; Liu et al. 2011). MMBIR is thought to be conservative: The leading strand is synthesized after microhomology-mediated annealing, and the lagging strand is synthesized using the leading strand as the template (Saini et al. 2013). Under this model, the newly synthesized chromosome containing the insertions is different from its sister chromatid. After cell division, one daughter cell will contain the rearranged chromosome including multiple segmental gains due to DNA synthesis, whereas the other daughter will inherit the intact chromatid (Fig. 6). Segmental gains can also result from chromothripsis involving sister chromatids (Li et al. 2014). However, in this case, segmental gains in one daughter should be accompanied by mirror-image segmental losses in the other daughter (Fig. 5). It is thus possible to distinguish these models by single-cell analysis.

Our third proposal would be to understand how chromosomal bridges generate chromothripsis (Fig. 9B). For sister chromatids fused near the telomere, fragmentation of the dicentric bridge can result in chromothripsis in the subtelomeric region (Fig. 5, left scheme; Fig. 9B, right scheme; Maciejowski et al. 2015). It is also possible that the bridge could be resolved by two breaks, generating an acentric fragment (Fig. 9B, middle scheme); this fragment could then be partitioned into a micronucleus and undergo chromothripsis. In this scenario, chromothripsis would also occur in the subtelomeric region. Chromosomal bridges could be resolved by a single break and initiate BFB cycles. After multiple BFB cycles, one progeny cell will experience successive DNA loss from the telomeric end (Fig. 7A). Eventually, as the BFB cycles continue, this could bring the two centromeres close enough that they might fuse into a single functional unit (Fig. 9B, left scheme). Incorrect merotelic attachment of this quasi-dicentric chromosome might cause the entire chromosome to lag and be partitioned into a micronucleus. This whole chromosome might then undergo chromothripsis that is not restricted to the subtelomeric region (as in the other scenarios). By performing live-cell imaging through multiple cell divisions followed by single-cell analysis, it should be possible to specifically detect each of the above scenarios (Fig. 9B) and reconstruct the short-term evolution of the chromosomal bridge.

CONCLUSION

Cancer is a genetic disease fueled by mutations. A cancer genome can be understood as a combination of driver mutations that confer different cancer phenotypes. Different cancers have different driver mutations, but these mutations promote tumorigenesis by recurrently disrupting a small number of pathways. A cancer genome also harbors many passenger mutations. These passenger mutations constitute an archaeological record of the mutational events during tumor evolution. Therefore, a cancer genome can also be interpreted as the outcome of a combination of different mutational processes. Understanding these mutational processes has important implications in cancer prevention and early detection. Because the mutational signatures associated with different carcinogens can be very specific, it is possible to assess the mutational potential of different environmental factors by analyzing the patterns of mutations in cancer patients with exposure to such factors. Both the mutational signature and the mutational distribution show a certain level of tissue specificity. These features may be useful as biomarkers for early cancer detection from circulating tumor cells or cell-free DNA.

The landscape of somatic mutations also presents a wealth of mutational outcomes due to errors in DNA replication, chromosome missegregation, or impaired DNA repair. Hypothetical models inferred from these mutational phenomena can be tested in experimental systems. Such analysis not only provides insight into the mechanism generating these phenomena, but it should also expand our knowledge of the biology of DNA replication and DNA repair. In particular, the combination of live-cell experiments and single-cell genomic analysis represents a highly specific approach to relate visible events on individual chromosomes to alterations in the DNA sequence. We expect this combination to also have broad applications in the study of cell-to-cell variation and the phenotype–genotype relationship.

REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Alexandrov LB, Stratton MR. 2014. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24:** 52–60.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500:** 415–421.

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. 2015. Clock-like mutational processes in human somatic cells. *Nat Genet* **47:** 1402–1407.

Alt FW, Kellems RE, Bertino JR, Schimke RT. 1978. Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. *J Biol Chem* **253:** 1357–1370.

Anand RP, Lovett ST, Haber JE. 2013. Break-induced DNA replication. *Cold Spring Harb Perspect Biol* **5:** a010397.

Arlt MF, Wilson TE, Glover TW. 2012. Replication stress and mechanisms of CNV formation. *Curr Opin Genet Dev* **22:** 204–210.

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. 2013. Punctuated evolution of prostate cancer genomes. *Cell* **153:** 666–677.

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, Alexandrov LB, Gundem G, et al. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513:** 422–425.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463:** 899–905.

Bignell GR, Santarius T, Pole JC, Butler AP, Perry J, Pleasance E, Greenman C, Menzies A, Taylor S, Edkins S, et al. 2007. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res* **17:** 1296–1303.

Blainey PC. 2013. The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37:** 407–427.

Boland CR, Goel A. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology* **138:** 2073–2087.

Burman B, Zhang ZZ, Pegoraro G, Lieb JD, Misteli T. 2015. Histone modifications predispose genome regions to breakage and translocation. *Genes Dev* **29:** 1393–1402.

Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, et al. 2013a. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494:** 366–370.

Burns MB, Temiz NA, Harris RS. 2013b. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* **45:** 977–983.

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.

Cartwright RA. 2009. Problems and solutions for estimating indel rates and length distributions. *Mol Biol Evol* **26:** 473–480.

Chan K, Gordenin DA. 2015. Clusters of multiple mutations: Incidence and molecular mechanisms. *Annu Rev Genet* **49:** 243–267.

Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20:** 447–457.

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 nonhomologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44:** 390–397, S391.

Chiarle R, Zhang Y, Frock RL, Lewis SM, Molinie B, Ho YJ, Myers DR, Choi VW, Compagno M, Malkin DJ, et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147:** 107–119.

Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD. 2014. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343:** 88–91.

Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D. 2012. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482:** 53–58.

Cremer T, Cremer M. 2010. Chromosome territories. *Cold Spring Harb Perspect Biol* **2:** a003889.

Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ. 2013. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155:** 948–962.

Dawson E, Chen Y, Hunt S, Smink LJ, Hunt A, Rice K, Livingston S, Bumpstead S, Bruskiewich R, Sham P, et al. 2001. A SNP resource for human chromosome 22: Extracting dense clusters of SNPs from the genomic sequence. *Genome Res* **11:** 170–178.

Debatisse M, Le Tallec B, Letessier A, Dutrillaux B, Brison O. 2012. Common fragile sites: Mechanisms of instability revisited. *Trends Genet* **28:** 22–32.

de Bourcy CF, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. 2014. A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9:** e105585.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295:** 1306–1311.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat Rev Genet* **14:** 390–403.

Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. 2015. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33:** 285–289.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485:** 376–380.

Donley N, Thayer MJ. 2013. DNA replication timing, genome stability and cancer: Late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin Cancer Biol* **23:** 80–89.

Donnianni RA, Symington LS. 2013. Break-induced replication occurs by conservative DNA synthesis. *Proc Natl Acad Sci* **110:** 13475–13480.

Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, Getz G. 2013. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23:** 228–235.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151:** 483–496.

Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, Yang L, Haseley P, Lehmann HS, Park PJ, et al. 2015. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85:** 49–59.

Fan HC, Wang J, Potanina A, Quake SR. 2011. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29:** 51–57.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. 2015. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43:** D805–D811.

Fousteri M, Mullenders LH. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: Molecular mechanisms and biological effects. *Cell Res* **18:** 73–84.

Francis JM, Zhang CZ, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, Homer H, Haidar S, Blumenstiel B, Pedamallu CS, et al. 2014. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov* **4:** 956–971.

Ganem NJ, Pellman D. 2012. Linking abnormal mitosis to the acquisition of DNA damage. *J Cell Biol* **199:** 871–881.

Garraway LA, Lander ES. 2013. Lessons from the cancer genome. *Cell* **153:** 17–37.

Garsed DW, Marshall OJ, Corbin VD, Hsu A, Di Stefano L, Schroder J, Li J, Feng ZP, Kim BW, Kowarsky M, et al. 2014. The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26:** 653–667.

Gisselsson D, Pettersson L, Hoglund M, Heidenblad M, Gorunova L, Wiegant J, Mertens F, Dal Cin P, Mitelman F, Man-

dahl N. 2000. Chromosomal breakage–fusion–bridge events cause genetic intratumor heterogeneity. *Proc Natl Acad Sci* **97:** 5357–5362.

Glover TW, Arlt MF, Casper AM, Durkin SG. 2005. Mechanisms of common fragile site instability. *Hum Mol Genet* **14**(suppl 2)**:** R197–R205.

Gordon DJ, Resio B, Pellman D. 2012. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet* **13:** 189–203.

Greenman CD, Cooke SL, Marshall J, Stratton MR, Campbell PJ. 2015. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. *J Math Biol* **72:** 47–86.

Haber DA, Velculescu VE. 2014. Blood-based analyses of cancer: Circulating tumor cells and circulating tumor DNA. *Cancer Discov* **4:** 650–661.

Haffner MC, Aryee MJ, Toubaji A, Esopi DM, Albadine R, Gurel B, Isaacs WB, Bova GS, Liu W, Xu J, et al. 2010. Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat Genet* **42:** 668–675.

Hakim O, Resch W, Yamane A, Klein I, Kieffer-Kwon KR, Jankovic M, Oliveira T, Bothmer A, Voss TC, Ansarah-Sobrinho C, et al. 2012. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature* **484:** 69–74.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: Two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9:** 958–970.

Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, et al. 2016. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164:** 538–549.

Harris RS, Petersen-Mahrt SK, Neuberger MS. 2002. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* **10:** 1247–1253.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10:** 551–564.

Hatch EM, Fischer AH, Deerinck TJ, Hetzer MW. 2013. Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* **154:** 47–60.

Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15:** 585–598.

Hicks WM, Kim M, Haber JE. 2010. Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* **329:** 82–85.

Hodgkinson A, Chen Y, Eyre-Walker A. 2012. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat* **33:** 136–143.

Hu L, Kim TM, Son MY, Kim SA, Holland CL, Tateishi S, Kim DH, Yew PR, Montagna C, Dumitrache LC, et al. 2013. Two replication fork maintenance pathways fuse inverted repeats to rearrange chromosomes. *Nature* **501:** 569–572.

International Cancer Genome Consortium. 2010. International network of cancer genome projects. *Nature* **464:** 993–998.

Jäger N, Schlesner M, Jones DT, Raffel S, Mallm JP, Junge KM, Weichenhan D, Bauer T, Ishaque N, Kool M, et al. 2013. Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell* **155:** 567–581.

Jiang J, Jing Y, Cost GJ, Chiang JC, Kolpa HJ, Cotton AM, Carone DM, Carone BR, Shivak DA, Guschin DY, et al. 2013. Translating dosage compensation to trisomy 21. *Nature* **500:** 296–300.

Kalisky T, Blainey P, Quake SR. 2011. Genomic analysis at the single-cell level. *Annu Rev Genet* **45:** 431–445.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502:** 333–339.

Kim TM, Laird PW, Park PJ. 2013. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155:** 858–868.

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.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20:** 1916–1924.

Korbel JO, Campbell PJ. 2013. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152:** 1226–1236.

Kotini AG, Chang CJ, Boussaad I, Delrow JJ, Dolezal EK, Nagulapally AB, Perna F, Fishbein GA, Klimek VM, Hawkins RD, et al. 2015. Functional analysis of a chromosomal deletion associated with myelodysplastic syndromes using isogenic human induced pluripotent stem cells. *Nat Biotechnol* **33:** 646–655.

Lasken RS. 2012. Genomic sequencing of uncultured microorganisms from single cells. *Nat Rev Microbiol* **10:** 631–640.

Lasken RS, Stockwell TB. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7:** 19.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499:** 214–218.

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505:** 495–501.

Leibowitz ML, Zhang CZ, Pellman D. 2015. Chromothripsis: A new mechanism for rapid karyotype evolution. *Annu Rev Genet* **49:** 183–211.

Li Y, Schwab C, Ryan SL, Papaemmanuil E, Robinson HM, Jacobs P, Moorman AV, Dyer S, Borrow J, Griffiths M, et al. 2014. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508:** 98–102.

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.

Lin C, Yang L, Tanasa B, Hutt K, Ju BG, Ohgi K, Zhang J, Rose DW, Fu XD, Glass CK, et al. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139:** 1069–1083.

Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, et al. 2011. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146:** 889–903.

Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* **4:** 1502.

Lo AW, Sprung CN, Fouladi B, Pedram M, Sabatier L, Ricoul M, Reynolds GE, Murnane JP. 2002. Chromosome instability as a result of double-strand breaks near telomeres in mouse embryonic stem cells. *Mol Cell Biol* **22:** 4836–4850.

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D'Gama AM, Cai X, et al. 2015. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350:** 94–98.

Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, Francis JM, Zhang CZ, Shalek AK, Satija R, et al. 2014. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat Biotechnol* **32:** 479–484.

Ma C, Martin S, Trask B, Hamlin JL. 1993. Sister chromatid fusion initiates amplification of the dihydrofolate reductase gene in Chinese hamster cells. *Genes Dev* **7:** 605–620.

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al. 2015.

G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12:** 519–522.
Maciejowski J, Li Y, Bosco N, Campbell PJ, de Lange T. 2015. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163:** 1641–1654.
Malkova A, Haber JE. 2012. Mutations arising during repair of chromosome breaks. *Annu Rev Genet* **46:** 455–473.
Mani R-S, Tomlins SA, Callahan K, Ghosh A, Nyati MK, Varambally S, Palanisamy N, Chinnaiyan AM. 2009. Induced chromosomal proximity and gene fusions in prostate cancer. *Science* **324:** 1230.
Mankouri HW, Huttner D, Hickson ID. 2013. How unfinished business from S-phase affects mitosis and beyond. *EMBO J* **32:** 2661–2671.
Mardin BR, Drainas AP, Waszak SM, Weischenfeldt J, Isokane M, Stutz AM, Raeder B, Efthymiopoulos T, Buccitelli C, Segura-Wang M, et al. 2015. A cell-based model system links chromothripsis with hyperploidy. *Mol Syst Biol* **11:** 828.
Martincorena I, Campbell PJ. 2015. Somatic mutation in cancer and normal cells. *Science* **349:** 1483–1489.
Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, et al. 2015. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348:** 880–886.
Mayle R, Campbell IM, Beck CR, Yu Y, Wilson M, Shaw CA, Bjergbaek L, Lupski JR, Ira G. 2015. Mus81 and converging forks limit the mutagenicity of replication fork breakage. *Science* **349:** 742–747.
McClintock B. 1939. The behavior in successive nuclear divisions of a chromosome broken at meiosis. *Proc Natl Acad Sci* **25:** 405–416.
McClintock B. 1941a. Spontaneous alterations in chromosome size and form in *Zea mays*. *Cold Spring Harb Symp Quant Biol* **9:** 72–81.
McClintock B. 1941b. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26:** 234–282.
Meaburn KJ, Misteli T, Soutoglou E. 2007. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* **17:** 80–90.
Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR, et al. 2014. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* **24:** 1624–1636.
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16:** 1182–1190.
Minocherhomji S, Ying S, Bjerregaard VA, Bursomanno S, Aleliunaite A, Wu W, Mankouri HW, Shen H, Liu Y, Hickson ID. 2015. Replication stress activates DNA repair synthesis in mitosis. *Nature* **528:** 286–290.
Mizuno K, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev* **23:** 2876–2886.
Mizuno K, Miyabe I, Schalbetter SA, Carr AM, Murray JM. 2013. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature* **493:** 246–249.
Narayanan V, Mieczkowski PA, Kim HM, Petes TD, Lobachev KS. 2006. The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. *Cell* **125:** 1283–1296.
Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90–94.
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.
Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH. 2015. The genome as a record of environmental exposure. *Mutagenesis* **30:** 763–770.
Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, et al. 2004. Genome coverage and sequence fidelity of φ29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* **32:** e71.
Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, et al. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463:** 191–196.
Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahovicek K, Stamatoyannopoulos JA, et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518:** 360–364.
Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680.
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.
Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45:** 970–976.
Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. *Proc Natl Acad Sci* **106:** 21766–21770.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Saini N, Ramakrishnan S, Elango R, Ayyar S, Zhang Y, Deem A, Ira G, Haber JE, Lobachev KS, Malkova A. 2013. Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature* **502:** 389–392.
Sakofsky CJ, Roberts SA, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA, Malkova A. 2014. Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep* **7:** 1640–1648.
Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci* **112:** E6456–E6465.
Santaguida S, Amon A. 2015. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat Rev Mol Cell Biol* **16:** 473–485.
Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488:** 504–507.
Smith CE, Llorente B, Symington LS. 2007. Template switching during break-induced replication. *Nature* **447:** 102–105.
Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: Experimental methods and applications. *Nat Rev Genet* **16:** 344–358.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41:** 393–395.
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.

Supek F, Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521:** 81–84.

Swanton C, McGranahan N, Starrett GJ, Harris RS. 2015. APOBEC enzymes: Mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov* **5:** 704–712.

Tanaka H, Tapscott SJ, Trask BJ, Yao MC. 2002. Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. *Proc Natl Acad Sci* **99:** 8772–8777.

Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, Rada C, Stratton MR, Neuberger MS. 2013. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2:** e00534.

Taylor BJ, Wu YL, Rada C. 2014. Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. *Elife* **3:** e03553.

Torres EM, Williams BR, Tang YC, Amon A. 2010. Thoughts on aneuploidy. *Cold Spring Harb Symp Quant Biol* **75:** 445–451.

Van Loo P, Voet T. 2014. Single cell analysis of cancer genomes. *Curr Opin Genet Dev* **24:** 82–91.

Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, Melotte C, Debrock S, Amyere M, Vikkula M, Schuit F, et al. 2009. Chromosome instability is common in human cleavage-stage embryos. *Nat Med* **15:** 577–583.

Varetti G, Pellman D, Gordon DJ. 2014. *Aurea mediocritas*: The importance of a balanced genome. *Cold Spring Harb Perspect Biol* **6:** a015842.

Voet T, Vanneste E, Van der Aa N, Melotte C, Jackmaert S, Vandendael T, Declercq M, Debrock S, Fryns JP, Moreau Y, et al. 2011. Breakage–fusion–bridge cycles leading to inv dup del occur in human cleavage stage embryos. *Hum Mutat* **32:** 783–793.

Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin ML, Zamani Esteki M, Van der Aa N, Mateiu L, McBride DJ, et al. 2013. Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res* **41:** 6119–6138.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339:** 1546–1558.

Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, et al. 2015. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518:** 495–501.

Wang Y, Navin NE. 2015. Advances and applications of single-cell sequencing technologies. *Mol Cell* **58:** 598–609.

Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512:** 155–160.

Watson IR, Takahashi K, Futreal PA, Chin L. 2013. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet* **14:** 703–718.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452:** 872–876.

Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24:** 1894–1904.

Woo YH, Li WH. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun* **3:** 1004.

Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al. 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153:** 919–929.

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45:** 1134–1140.

Zakov S, Kinsella M, Bafna V. 2013. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc Natl Acad Sci* **110:** 5546–5551.

Zhang Y, Gostissa M, Hildebrand DG, Becker MS, Boboila C, Chiarle R, Lewis S, Alt FW. 2010. The role of mechanistic factors in promoting chromosomal translocations found in lymphoid and other cancers. *Adv Immunol* **106:** 93–133.

Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148:** 908–921.

Zhang CZ, Leibowitz ML, Pellman D. 2013. Chromothripsis and beyond: Rapid genome evolution from complex chromosomal rearrangements. *Genes Dev* **27:** 2513–2530.

Zhang CZ, Adalsteinsson VA, Francis J, Cornils H, Jung J, Maire C, Ligon KL, Meyerson M, Love JC. 2015a. Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat Commun* **6:** 6822.

Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. 2015b. Chromothripsis from DNA damage in micronuclei. *Nature* **522:** 179–184.

Zhao H, Thienpont B, Yesilyurt BT, Moisse M, Reumers J, Coenegrachts L, Sagaert X, Schrauwen S, Smeets D, Matthijs G, et al. 2014. Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *Elife* **3:** e02725.

Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338:** 1622–1626.