Clustered Mutations in Human Cancer

Steven A Roberts, *National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina, USA*

Dmitry A Gordenin, *National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina, USA*

Advanced article

Article Contents

Online posting date: 15th January 2014

Mutations are the frequent cause of cancer. They are mostly viewed as independent events distributed randomly across chromosomes. However, mutation distribution can be affected by permanent or transient features of genome structure and function. The extreme form of nonrandom distributions is a mutation cluster with multiple mutations concentrated in a tiny fraction of the genome. Multiple lesions in abnormally long regions of transient single-stranded deoxyribonucleic acid (DNA) can cause mutation clusters, which have been found in a majority of human cancer types. Mutation spectra indicated that many clusters in cancer genomes were caused by a subclass of apolipoprotein B mRNA-editing polypeptide-like (APOBEC) cytidine deaminases. These enzymes function to restrict retroviruses and retrotransposons by converting cytidine to uridine in single-stranded complementary DNAs (cDNAs). The simple mutation spectra in clusters aided in highlighting APOBECs among the complex set of mutagenic mechanisms operating throughout many cancer genomes. Thus, clusters are an analytical tool for deciphering cancer mutation mechanisms.

Introduction

Mutations, slowly accumulated over billions of years of biological history, have generated a variety of deoxyribonucleic acid (DNA) sequences in living organisms. Rare mutations increasing fitness are thought to be an important factor of evolution. Mutations can also have significant biological effects over the lifetime of an organism by causing genetic disease or cancer. In the latter case, a cell or a group of cells escape organism control and descends into unrestrained proliferation. The importance of mutations in the incidence and progression of malignant tumours was underscored by the recent sequencing of multiple cancer genomes and exomes, where hundreds of thousands of mutations can be found in a single cancer. These efforts identified many mutations that occurred in genes that have been previously identified as oncogenes or tumour suppressors (Meyerson *et al*., 2010; Stratton, 2011). Moreover, various groups attempted to find more genes potentially important for cancer through creating large databases of cancer mutations (Futreal *et al*., 2004; Forbes *et al*., 2011) and by statistical analysis of mutation distribution in large numbers of cancer samples (Lawrence *et al*., 2013). One of the parameters important for such an analysis is the distribution of mutation probabilities across the genome. Although the frequently used simple assumption is that mutations are random and independent throughout the genome, it is clear now that the chance of a mutation occurring in a given nucleotide may depend on many genomic features, such as replication timing, chromatin organisation, local DNA structure, transcription in the region, transcribed strand orientation, etc. (Thoma, 2005; Stamatoyannopoulos *et al*., 2009; Rochette and Brash, 2010; Koren *et al*., 2012; Lawrence *et al*., 2013). A transient increase in mutability can also occur next to double-strand breaks (DSBs) (Malkova and Haber, 2012; Drier *et al*., 2013). Some regions of the B-cell genomes undergo hypermutation caused by activation-induced cytidine deaminase (AID) targeting immunoglobulin genes as well as secondary chromosomal targets (Conticello *et al*., 2007; Maul and Gearhart, 2010). It has turned to be very important to account for the mutational heterogeneity of genomes for statistical analysis aimed at identifying 'significantly mutated genes' in human cancers (Lawrence *et al*., 2013). The importance of understanding mechanisms underlying the nonrandom distribution of mutations in cancer genomes goes beyond fighting cancer, as it can be applied to noncancer somatic genome dynamics on the organism scale as well as to germ line mutations on population and even on evolution scales. **See also**: Characterising Somatic Mutations in Cancer Genome by Means of Next-generation Sequencing; Mutagenesis Mechanisms; Non-B DNA Structure and Mutations

eLS subject area: Genetics & Disease

How to cite:
Roberts, Steven A; and Gordenin, Dmitry A (January 2014) Clustered Mutations in Human Cancer. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0024941

Causing Human Genetic Disease; Somatic Hypermutation in Antibody Evolution; Somatic Hypermutation of Antigen Receptor Genes: Evolution

Extreme manifestations of nonrandom mutation distributions in genomic space are mutation clusters, where several mutations are found in a tiny fraction of the genome. First findings of unusually high density of mutations in a small fraction of the genome were made through sequencing LacZ mutants in the reporter locus integrated into the mouse genome. Surprisingly, a small fraction (approximately 0.2%) of the LacZ mutant alleles recovered from cells carried multiple mutations. This phenomenon was called a 'mutation shower' by analogy with meteor showers concentrated in time and in a limited sector of the sky (Wang *et al*., 2007). Until recently, there were limited observations of mutation clusters and no molecular mechanism accounting for their formation. Studies in model organism followed by analysis of whole-genome mutation distributions in yeast and then in cancer genomes led to an understanding of the first molecular mechanism underlying cluster formation and to the conclusion that mutation clusters are common in several types of human cancers (Nik-Zainal *et al*., 2012; Roberts *et al*., 2012). This article summarises the research that led to highlighting a mutation cluster mechanism, the role of clusters in understanding mutation processes in human cancers and future developments anticipated in this field.

Clusters Stemming from Long Single-Stranded DNA

Studies in model yeast systems established that clustered multiple mutations can occur in regions of long, up to 20 Kb, artificially created single-stranded DNA (ssDNA) (Yang *et al*., 2010; Burch *et al*., 2011; **Figure 1a** and **1b**). Such regions were generated by a well-understood process of 5′→3′ resection of a DNA strand starting at the ends of DSB or at uncapped telomeres (Mimitou and Symington, 2011). The key element of experimental design enabling clustered mutagenesis was high levels of DNA damage applied to the yeast genome. Lesions in double-stranded DNA (dsDNA) were repaired by one of the excision DNA repair systems utilising the template of undamaged DNA strand to restore wild-type sequence. It was long known that living cells can repair thousands of simultaneous lesions in chromosomal dsDNA generated by acute chemical or radiation damage (Friedberg *et al*., 2006). Surprisingly, yeast cells were also capable of restoring kilobases of ssDNA with dozens of lesions to viable dsDNA. Error-prone translesion synthesis (TLS) (Plosky and Woodgate, 2004) across multiple lesions in ssDNA resulted in multiple changes within the newly synthesised strand, which were then fixed as mutations in dsDNA by subsequent templated repair or replication synthesis. Various classes of lesions, which are substrates for different kinds of excision repair, were capable of causing clustered mutations. Initially, clustered mutagenesis in ssDNA was found for lesions generated by ultraviolet (UV) light and by methyl methanesulfonate (MMS), which serve in dsDNA as substrates for nucleotide and base excision repair, respectively (Yang *et al*., 2008; Yang *et al*., 2010; Burch *et al*., 2011). Later, mutation clusters were also identified with ssDNA-specific lesions caused by chemical cytidine deamination with sodium bisulfite or enzymatically by human APOBEC3G cytidine deaminase expressed in yeast (Chan *et al*., 2012). Based on mutation spectra and a requirement for TLS, it was concluded that uridines converted from cytidines in ssDNA by an apolipoprotein B mRNA-editing polypeptide-like (APOBEC) deaminase were efficiently converted into abasic sites by uracil-DNA glycosylase (Ung1). Copying AP sites in ssDNA resulted in either adenines or cytosines placed by Pol–ζ or Rev1-dependent TLS, respectively, which after fixing mutations left C→T or C→G changes in the mutagenised single strand (Chan *et al*., 2013). **See also**: DNA Damage; DNA Strand Break Repair and Human Genetic Disease; Eukaryotic Recombination: Initiation by Double-strand Breaks; Mutagenesis Mechanisms; Mutation; Recombinational DNA Repair in Eukaryotes

The density of damage-induced mutations within clusters occurring in transient ssDNA was up to 1000-fold greater than that in the rest of the genome (Burch *et al*., 2011). Because mutagens usually have preference or even complete specificity for certain bases (or even to a single base), multiple mutations of certain bases (or a base) were more frequently found in the same DNA strand of a mutation cluster generated in ssDNA, that is, mutations were strand coordinated. The example of ultimate strand coordination can be illustrated by the spectrum of mutations induced in yeast by APOBEC3G cytidine deaminase creating uracils from cytosines in ssDNA triggered by controlled telomere uncapping followed by 5′→3′ DNA end resection (Chan *et al*., 2012; **Figure 2a**). In the absence of Ung1 glycosylase, all uracils stay in ssDNA and after replication result in clusters of exclusively C→T mutations, always in the single strand exposed to the mutagen, whereas the complementary strand would carry clusters of strand-coordinated G→A mutations. It is worth mentioning that strand-coordinated mutations can be caused by action of a processive as well as a distributive enzyme or chemical factor, where each change is the result of an event independent of others. Coordination (similarity) of mutated bases or even motifs would be the consequence of simultaneous incidence of lesions in a common ssDNA stretch.

Mutagenesis pathways enabling the formation of vast strand-coordinated mutation clusters were found to operate in cells proliferating in the presence of DNA base alkylation by MMS (Roberts *et al*., 2012; **Figure 2b**). The size and mutation density in such 'naturally' occurring clusters were dramatic, with more MMS-induced mutations in a small (approximately 1%) fraction of yeast genome than in the remaining 99%. Cytosine base specificity and strand coordination of mutations within clusters

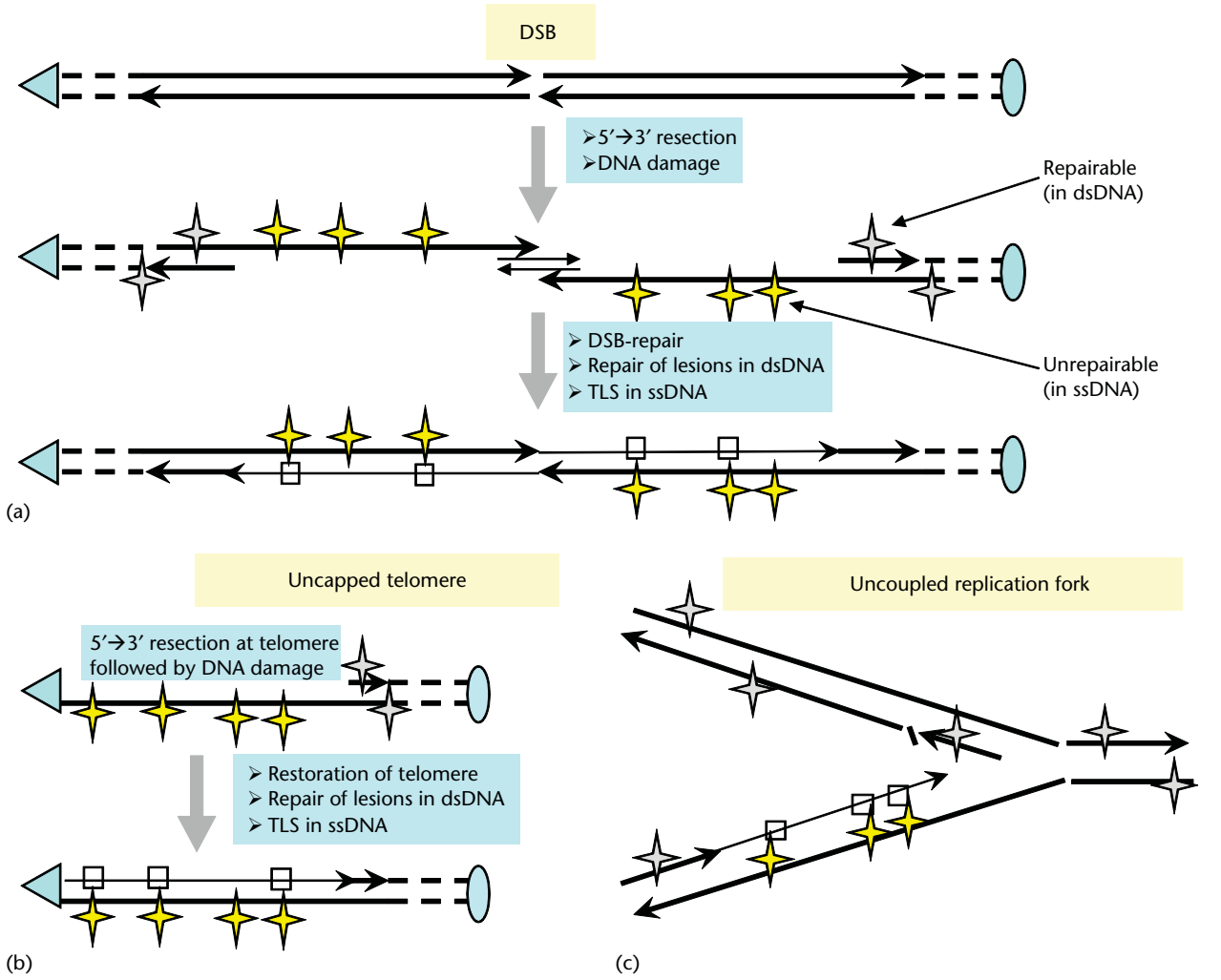
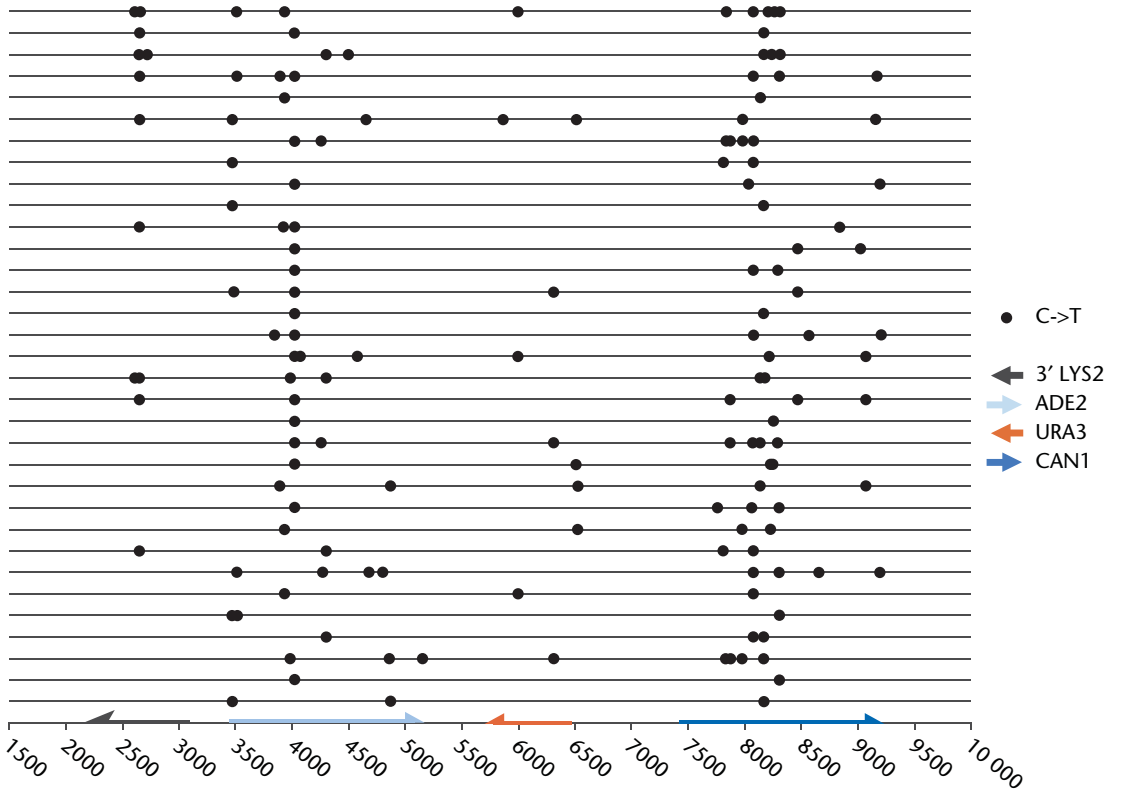


Figure 1 Clusters can result from hypermutation in long regions of ssDNA. (a) DSB followed by 5′→3′ degradation (resection) eliminates one DNA strand, thus leaving regions of ssDNA around a DSB. DNA lesions could occur in both ssDNA (yellow stars) and dsDNA (grey stars). DSB with long resected ends can be repaired by homologous recombination with a sister chromatid or with an ectopic homologous region elsewhere in the genome. Shown are short oligonucleotides used to repair artificially created DSB in yeast (Yang *et al.*, 2008, 2010). Restoration to dsDNA involves error-prone TLS. Mutations generated by error-prone TLS (blue squares) can be copied by excision repair of lesions (not shown), creating mutant sequence in the second DNA strand. Mutagenic lesions from an agent with base specificity could result in a single type of base mutated exclusively (or predominantly) on one side of the break (strand-coordinated mutations). Note that this same base specificity would switch to a complementary base on the other side of the break. Indeed, such 'switching' mutation clusters were observed in Roberts *et al.* (2012). (b) ssDNA can be generated by resection at a telomere that has lost its protein cap. Modelling this process, a controlled transient uncapping allowing 5′→3′ resection has been achieved by shifting a temperature-sensitive *cdc13–1* yeast mutant to nonpermissive temperature (Yang *et al.*, 2008; Burch *et al.*, 2011; Chan *et al.*, 2012, 2013). Returning yeast to permissive temperature after applying DNA damage allowed restoration of normal dsDNA at the telomere in a process that involved error-prone TLS and resulted in mutation clusters. (c) Transient ssDNA can be formed at dysfunctional replication forks, if synthesis of one out of two nascent strands is uncoupled from the proceeding of the replicative helicase. In support of this mechanism, switching of mutation strand specificity in mutation clusters was observed in the yeast strains lacking *TOF1* (homologue of *hTIMELESS*) or *CSM3* (homologue of *hTIPIN*) responsible for the replication fork integrity (Roberts *et al.*, 2012).

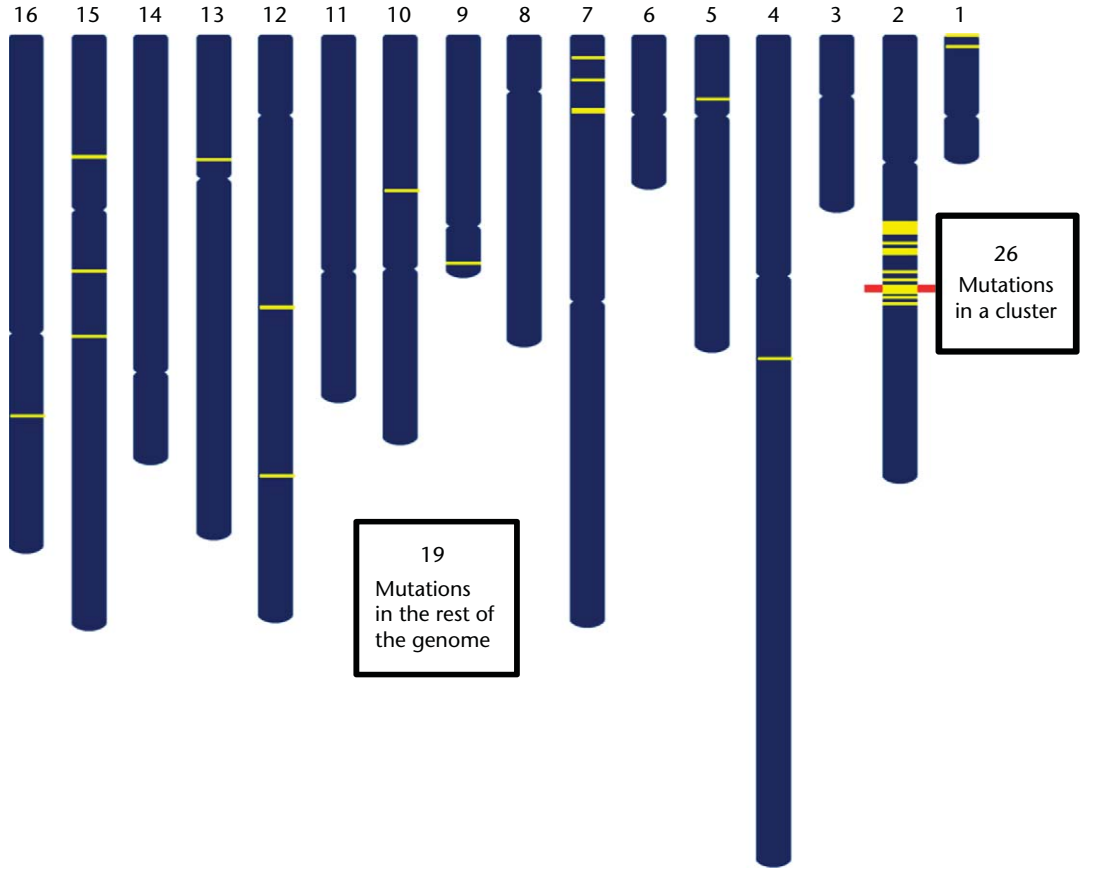
agreed with the previously established MMS signature in artificially created ssDNA (Yang *et al.*, 2010). Mutation distributions and the effects of genetic controls indicated that alkylation damage in ssDNA formed at DSBs and dysfunctional replication forks is a source of mutation clusters. Altogether, it appears that as long as the two factors, ssDNA and a DNA-damaging agent, are present, it leads to formation of mutation clusters. In support of this suggestion, mutation clusters were also found in proliferating yeast expressing heterologous APOBEC cytidine deaminases from humans (Taylor *et al.*, 2013) and lamprey (Lada *et al.*, 2013).

Strand-Coordinated Mutation Clusters in Cancers

Over the past several years, large datasets of mutations, rearrangements and copy number variations from



(a)



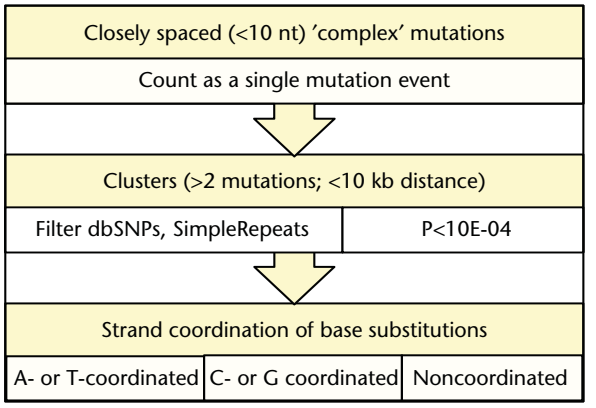
(b)

thousands of human malignant tumours have been created (Zhang *et al*., 2011; Chang *et al*., 2013). These data further support the consensus that genome changes are a strong factor enabling the hallmarks of cancer (Hanahan and Weinberg, 2011). However, the complex spectra of genome changes in tumours make the task of reading this 'archaeological record' far from straightforward (Stratton, 2011). One way to decipher the complexity of the mutational record in human cancer is to look for the presence of a signature left by known mutagenic mechanisms. This approach is based on prior mechanistic knowledge and led to prompt identification of mutation signatures associated with 5-methyl cytosine deamination in CpG, UV light and tobacco in the very first whole-genome-sequenced cancers (Pleasance *et al*., 2010a, 2010b). **See also**: Characterising Somatic Mutations in Cancer Genome by Means of Next-generation Sequencing; Genomic Rearrangements: Mutational Mechanisms

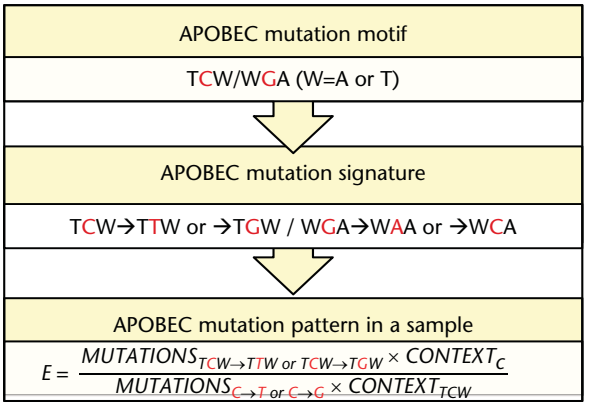
Based on the ssDNA-associated mutation cluster mechanisms uncovered in yeast, Roberts *et al*. (2012) suggested that if ssDNA-associated hypermutation also exists in human cancers, it will include formation of completely strand-coordinated clusters, that is, clusters in which all base substitution mutations have replaced the same kind of nitrogenous bases of the same DNA strand. The general principle of using hypothesis-based algorithms to pinpoint specific mutagenic mechanism within a mixed mutation spectrum can be likened to a succession of affinity columns used in the biochemical purification of specific macromolecules from a complex mix extracted from cells. Each step is designed to remove contaminants from the desired molecule even at the cost of reducing the yield. In the first step, potential clusters in cancer datasets were identified using parameters (intermutation distance and *p*-value under an assumption of random mutation positioning) with the same or even increased stringency as those characteristic of damage-induced clusters in yeast (**Figure 3a**). In addition, groups of mutations separated by just a few nucleotides were considered as a single complex event, because such groups could be often generated by multiple polymerase errors at the site of a single lesion. In four types of cancers, where whole-genome datasets were analysed (**Figure 4a** and **4b**; Roberts *et al*., 2012, 2013), clustered mutations were rare, approximately 1% of all mutations in the dataset and were scattered all over cancer genomes. In independent studies, Stratton and co-authors also found mutation clusters in 152 out of 507 samples from 8 out of 10 whole-genome-sequenced types of cancer (Nik-Zainal *et al*., 2012; Alexandrov *et al*., 2013). Statistical definition of clusters by the Stratton team was based on calculations of probability in assumption of random distribution of mutations along chromosomes. They suggested a very convenient visualisation of mutation distribution as a 'rainfall plot', where clusters appear as areas of high 'rain' density (**Figure 4c**). This as well as the earlier hypothesis of 'mutation showers' (Wang *et al*., 2007) suggested the term 'kataegis' (which means thunderstorm in Greek). It is worth noting that the close spacing of mutations on the chromosome could reflect special features of mutagenic mechanisms as well as artifacts caused by selection or even by amplification of the region containing mutations (Alexandrov *et al*., 2013). Thus, downstream filtering is essential to find additional nonrandom features that could altogether indicate a mechanism.

The third filtering step enriching for events potentially associated with ssDNA would be identification of clusters with complete strand coordination (**Figure 3a**). Such clusters represented a significant (up to 50%) fraction of all clusters in whole-genome datasets of four cancer types, as analysed in Roberts *et al*. (2012, 2013) (**Figure 4a**). The majority of strand-coordinated clusters contained only mutations in cytosines (C coordinated) or in complementary guanines (G coordinated). A- or T-coordinated clusters were less frequent and found mostly in multiple myelomas. Tight clusters of strand-coordinated changes are unlikely to occur by coincidence of mutations generated through several independent pathways. C- and G-coordinated clusters displayed additional features indicating a likely mutagenic mechanism leading to their formation. Mutated cytosines (C) were very often preceded by a thymine (T) and followed by either an adenine or a thymine (designated as W in IUPAC code) (**Figure 4b**). This TCW (or complimentary WGA) motif has been reported as the preferred substrate for a subclass of APOBEC cytidine deaminases, namely APOBEC1/3A/3B/3C/3DE/3F/3H (throughout this article this subclass shall be referred to as APOBEC without gene-specifying suffix). These enzymes function normally to convert cytosine bases to uracils during ribonucleic acid (RNA) editing as well as in the single-stranded cDNA of retroviruses and retrotransposons (Smith *et al*., 2012 and references therein). Hypermutation or degradation of cDNAs resulting from multiple deaminations prevents integration of retroelements into chromosomes (Refsland and Harris, 2013). Importantly, APOBECs have a very strong preference to ssDNA over dsDNA. It is worth noting that as expected

Figure 2 Mutation clusters in yeast. (a) An example of complete strand coordination of mutations in clusters obtained by controlled telomere uncapping combined with transient expression of APOBEC3G cytidine deaminase in *cdc13–1* yeast strains lacking uracil-DNA glycolsylase (Ung1) (reproduced from Figure 3B of Chan *et al.*, 2012, with modifications). Each thin horizontal line represents a cluster. Small black circles represent individual base substitutions identified in a strand that was transiently lacking the complement. All mutations are C→T in agreement with cytidines being converted to uridines by APOBEC3G and after round of replication being cemented as thymidines in the progeny DNA. Also shown are the positions of open reading frames composing the multigene mutation reporter and distances from the left tip of chromosome V. (b) An example of a mutation cluster found by whole-genome sequencing of a yeast cell isolated after growth for approximately 25 generations in the presence of alkylating agent methyl methanesulfonate (Roberts *et al.*, 2012). Thick vertical blue lines represent yeast chromosomes sized proportional to DNA length. Thin yellow lines across chromosomes show positions of mutations. Several lines depicting mutations in a cluster as well as in some other genomic positions are merging because of close positioning. © PLoS.



(a)



(b)

Figure 3 Bioinformatics analysis used to identify mutation clusters and cluster-associated mutation signatures (Roberts *et al.*, 2012, 2013). (a) Detection and classification of mutation clusters. The first step involves filtering out complex mutations, most of which are composed of several very closely spaced changes caused by a single lesion. The second step highlights groups of closely spaced mutations defined as clusters by intermutation distance and the probability to occur by random genomic positioning of the mutations in a dataset. Mutations identical to small nucleotide polymorphisms (SNPs) previously found in the human population (dbSNPs) as well as mutations falling into a SimpleRepeat track (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=204690969&c=chr8&g=simpleRepeat) are excluded from whole-genome cluster detection, because they have a greater chance to represent germ line mutations or false positive mutation calls, respectively. Altogether these two categories rarely exceed 15% of all mutation calls. For exome mutation data, where mutation calls are based on very high coverage, only dbSNPs were excluded. (b) Enrichment with the APOBEC mutation signature. Analysis can be applied to the whole-genome or exome mutation catalogues as well as to a defined part of a catalogue (e.g. to C- or G-coordinated mutation clusters). The number of mutated nucleotides (shown in red) as well as the number of corresponding nucleotides and nucleotide motifs in the immediate (+/− 20 nt) vicinity of mutated bases is counted and used to calculate enrichment (E). $MUTATIONS_{TCW\rightarrow TTW\ or\ TCW\rightarrow TGW}$: the number of mutations of cytosines to thymines or guanines in TCW context (including complements); $MUTATIONS_{C\rightarrow T\ or\ C\rightarrow G}$: the number of mutations of cytosines to thymines or guanines; $CONTEXT_C$: the number of cytosines (including complements) in the immediate vicinity of the mutated cytosines; $CONTEXT_{TCW}$: the number of TCW motifs (including complements) in the immediate vicinity of mutated cytosines.

from prior studies, the mutation motifs of activation-induced cytidine deaminase operating in somatic hypermutation of immunoglobulin genes was also found in noncoordinated clusters in multiple myeloma samples, but unlike strand-coordinated clusters, AID-enriched clusters occurred only in a small number of known primary and secondary targets of AID (Roberts *et al.*, 2012 and references therein).

The next piece of corroborating evidence linking C- or G-coordinated clusters with APOBEC enzymes was the high preference for T or G bases to replace cytosines in TCW with very little TCW→TAW mutations (Nik-Zainal *et al.*, 2012; Roberts *et al.*, 2012, 2013; Alexandrov *et al.*, 2013). This was in complete agreement with a previously established pathway leading from cytosine deamination in ssDNA to base substitutions in the descendant dsDNA molecule (Chan *et al.*, 2012 and references therein). First, uracil is created from cytosine by deamination, and then uracil-DNA glysosylase removes the uracil base, resulting in an abasic site. Error-prone TLS inserts either adenine or cytosine across from the abasic site, which ultimately results in either a C→T or a C→G mutation. Another unusual feature of C- or G-coordinated clusters was revealed when cluster localisation was compared with the positions of hundreds of chromosomal rearrangement breakpoints identified from the same Illumina sequencing that produced mutation calls (**Figure 4a**). Nearly half of C- or G-coordinated clusters colocalised with breakpoints in contrast with very few noncoordinated clusters and no A- or T-coordinated clusters registering colocalisation. Such colocalisation is in good agreement with ssDNA intermediates, which can be formed either at DNA breaks leading to rearrangements (e.g. **Figure 1a**) or/and by the enhanced chance of break formation in or next to ssDNA stretches. Altogether, C- or G-coordinated clusters are abundant in many cancer types and carry several features indicating APOBEC mutagenesis in long ssDNA. The main indicative feature is enrichment of APOBEC signature mutations (TCW→TTW or TCW→TGW) calculated over that which is expected for random mutation of cytosines residing in the immediate (+/− 20 nt) vicinity of the mutated nucleotides (**Figure 3b** and Roberts *et al.*, 2012, 2013). This sampling method reduced statistical power but allowed concentrating on the reliably sequenced part of the genome. It also accounted for short size of APOBEC scanning tracks (Chelico *et al.*, 2009). It turned out that C- or G-coordinated clusters carried distinct APOBEC mutation signature even if picked from exome sequencing, which covers only approximately 1% of the genome (Roberts *et al.*, 2013). Although segments containing strand-coordinated clusters identified in exomes have a greater chance than the whole-genome clusters to contain additional mutations, because they could have been missed in the nonsequenced part of the genome, a statistically significant enrichment with the APOBEC signature was observed even for exome clusters containing just 2 C- (or G-) mutations. This enrichment increased with an increase in cluster size to more than threefold excess over that expected for random mutation of cytosines. Thus, at least two out of three C- (or G-) mutations in these exome clusters were likely to be caused by APOBECs.

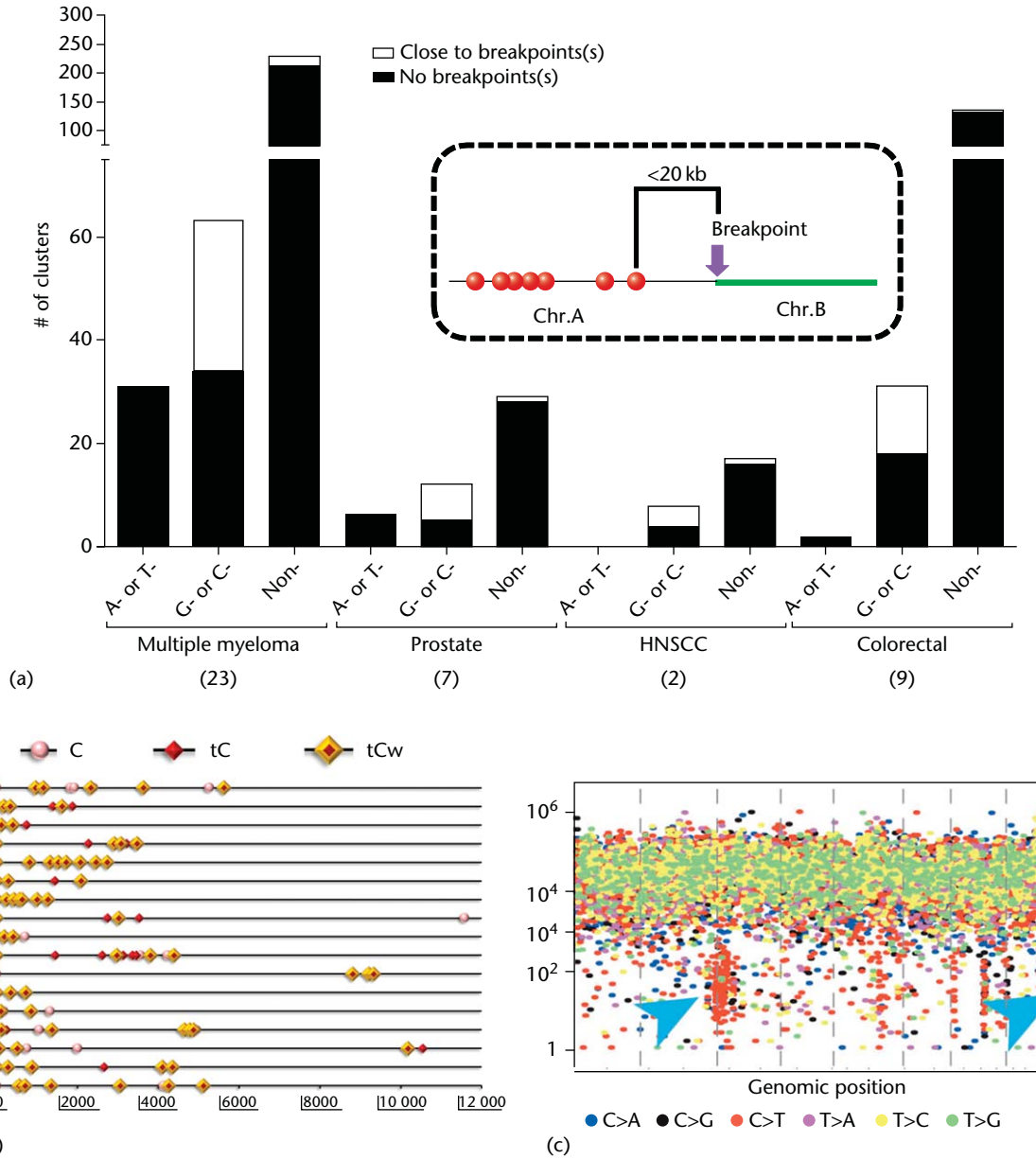


Figure 4 Mutation clusters in cancers. (a) Mutation clusters identified by Roberts *et al.* (2012, 2013) in whole-genome-sequenced samples of multiple myelomas (Chapman *et al.*, 2011), prostate adenocarcinomas (Berger *et al.*, 2011), head and neck squamous cell carcinomas (HNSCC) (Stransky *et al.*, 2011) and colorectal adenocarcinomas (Bass *et al.*, 2011). Clusters are separated by type of strand coordination ('A- or T-', 'G- or C-' and 'non' coordinated). White bars indicate the number of clusters colocalised with breakpoint(s). The number of sequenced samples for each cancer type is shown in parentheses. Colocalisation (schematically shown in a box insert) was registered when the region covered by the cluster plus left and right flanks of 20 000 nucleotides contained at least one breakpoint. Black bars depict the number of clusters not associated with a specific breakpoint. (b) The distribution of mutations within 17 C-coordinated clusters with greater than 3 mutations from multiple myeloma. Mutated cytosines are categorised by their presence in a TC motif (red diamonds), TCW motif (yellow highlighted red diamonds) or no identified motif (pink circles). Reproduced from Figure 6C of Roberts *et al.* (2012). © Cell Press. (c) An example of kataegis (mutation clusters) graphically represented as 'rainfall' plots. Each dot represents a single somatic mutation in a lung cancer sample. Dots are ordered on the horizontal axis according to the rank of the mutation's position in the human genome. The vertical axis denotes the genomic distance of each mutation from the previous mutation. Arrowheads indicate clusters of mutations in a kataegis event. Reproduced from Figure 6 of Alexandrov *et al.* (2013) with modifications. © Nature Publishing Group.

Mutation Clusters – A Tool for Deciphering Mutagenesis Pathways in Human Cancers

Strand-coordinated clusters can offer a permanent record of a single mutation mechanism that operated at some point in the cancer sample. Thus, in addition to their potential biological impact during cancer progression, mutation signatures in clusters provide a simple and powerful tool for statistical exploration of whole-genome or exome mutation catalogues. Enrichments with mutation signatures, calculated directly from the nucleotide and motif counts in sample's mutation catalogue and in genomic context obtained from the human genome reference sequence (**Figure 3b**), can be evaluated by simple statistical methods, for example, Fisher's exact test. Sample-specific *p*-values can be then corrected for multiple hypotheses testing using standard methods, such as Benjamini–Hochberg false discovery rate (Benjamini and Hochberg, 1995). Roberts *et al*. (2013) applied such an analysis to evaluate the presence of APOBEC mutagenesis pattern in mutation catalogues from 2680 exomes, mainly from The Cancer Genome Atlas. They found that 6 out of 14 analysed cancer types contained significant fractions of samples displaying an APOBEC mutagenesis pattern (**Figure 5**). Enrichment with APOBEC signature mutations in some samples approached a theoretical maximum (assuming a random distribution of nucleotides in the genome) of 5.5-fold over random mutagenesis. Some samples contained up to 1000 APOBEC mutations representing approxiamately 70% of the sample's mutation load. Another group found that the same cancer types showed an increased number of mutations in APOBEC motifs when mutation catalogues were pooled together within each cancer type (Burns *et al*., 2013). Both studies used sample-specific evaluations of APOBEC mutagenesis to correlate with the presence of messenger RNA (mRNA) of each member of the APOBEC family and

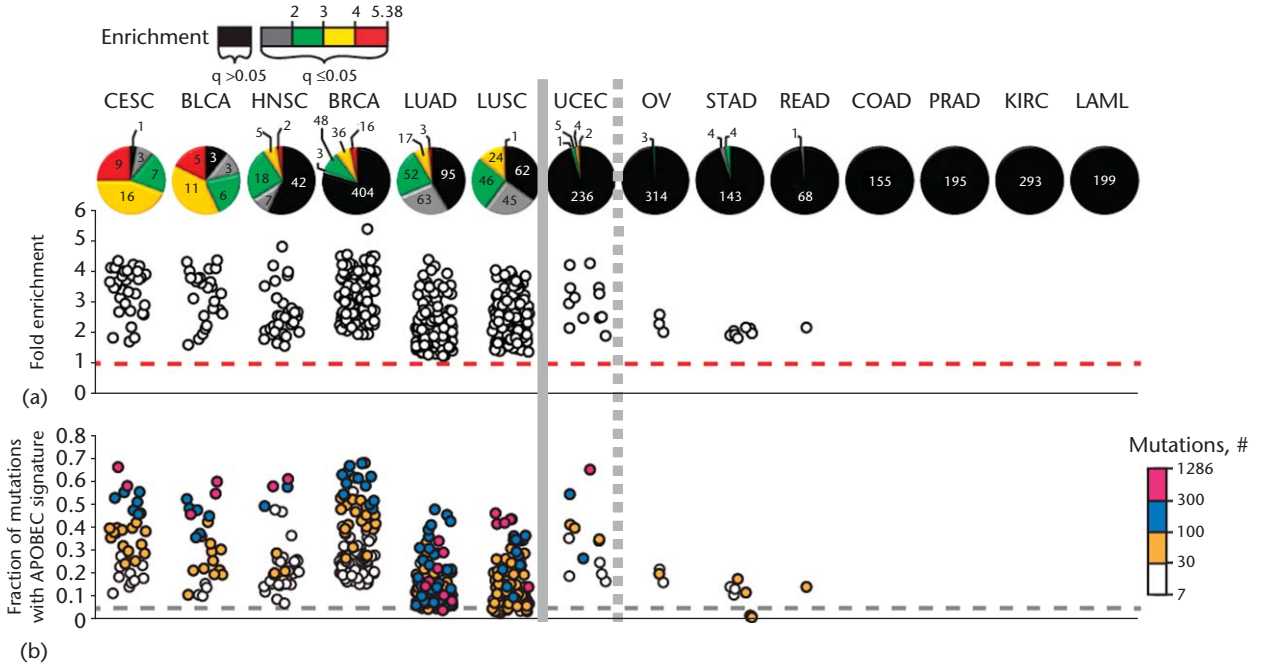


Figure 5 The APOBEC mutagenesis pattern is widespread across several types of human cancers. (a) The fold enrichment of the APOBEC mutagenesis signature as determined within each of 2680 whole-exome-sequenced tumours, representing 14 cancer types. Samples were first categorised by the statistical significance of the APOBEC mutation pattern (see **Figure 4** and text) and then by fold enrichment for samples with and FDR-corrected *q*-value <0.05. Pie charts show the distribution of fold-enrichment categories within a given type. The colour code for fold-enrichment categories is shown above the charts. Samples displaying Benjamini–Hochberg corrected *q*-values >0.05 are represented in black. These samples are excluded from the scatter graphs. (b) Fractional load of APOBEC mutation signature. The colour scale to the right of the graph indicates the number of APOBEC signature mutations for samples with $q<0.05$. Horizontal dashed lines in (a) and (b) indicate effects expected for random mutagenesis. Cancer types are abbreviated as in TCGA. *Abbreviations*: CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; BLCA, bladder urothelial carcinoma; HNSC, head and neck squamous cell carcinoma; BRCA, breast invasive carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; UCEC, uterine corpus endometrioid carcinoma; OC, ovarian serous cystadenocarcinoma; STAD, stomach adenocarcinoma; READ, rectum adenocarcinoma; COAD, colon adenocarcinoma; PRAD, prostate adenocarcinoma; KIRC, kidney renal clear cell carcinoma; and LAML, acute myeloid leukaemia (LAML). Cancer types to the left of vertical solid line show statistically significant presence of samples with APOBEC mutagenesis pattern in Roberts *et al.* (2013) as well as in Alexandrov *et al.* (2013). Cancer types to the right of vertical dashed line did not show statistically significant presence of samples with APOBEC mutagenesis pattern in Roberts *et al.* (2013) as well as in Alexandrov *et al.* (2013). APOBEC mutagenesis pattern in UCEC was on the marginal level in Roberts *et al.* (2013) analysis, whereas it was highlighted as carrying APOBEC mutation signature in Alexandrov *et al.* (2013). Reproduced from Figure 2 from Roberts *et al.* (2013) with modification. © Nature Publishing Group.

came to the conclusion that APOBEC3B is most likely to be the cause of these mutations.

Sample-specific enrichment values calculated by the method of Roberts *et al*. (2013) showed nearly perfect correlation (see their Supplementary **Figure 1**) with the presence of the APOBEC mutagenesis signature determined by mathematical signature decomposition of mutations from 21 breast cancers in the study of Nik-Zainal *et al*. (2012). When signature decomposition method was applied to 7042 cancer samples of 30 cancer types, it highlighted at least 20 distinct mutational signatures (Alexandrov *et al*., 2013). Two of those signatures clearly showed the presence of an APOBEC mutagenesis component. APOBEC-like signatures were detected in 16 out of 30 cancer types analysed. In this regard, there was nearly complete agreement with the hypothesis-based analysis in Roberts *et al*. (2013) (**Figure 5**). Currently, the two methods are complementary. Signature decomposition is capable of finding new mutagenesis patterns in large pools of samples. The hypothesis-based approach utilises a single mutagenesis signature, which can be derived from any combination before mechanistic knowledge or even from an a priori signature decomposition to generate sample-specific *p*-values that allow distinguishing signature-containing samples within cancer types and subtypes. The latter allowed (Roberts *et al*., 2013) to support the statement that APOBEC mutagenesis extends into cancer genes as well as to highlight the HER2-enriched subtype of breast cancer as containing the highest fraction of samples with APOBEC mutagenesis pattern.

Conclusions and Future Questions

In summary, it is clear that the phenomenon of mutation clusters does exist and that it is widespread in human cancers. Only a single mechanism has been identified so far with a sum of indications pointing to a subclass of APOBEC cytidine deaminases as a source of C- or G-coordinated clusters scattered across cancer genome. The signature of a specific mutagenic mechanism rectified through analysis of mutation clusters increases the statistical power of mining mutation catalogues of individual cancer samples and enables the statistical evaluation of correlations with biological features, such as cancer type and subtype or gene expression. Progress in understanding and utilising the mutation cluster phenomenon could continue to emerge while answering questions that grew from cluster findings.

1. Other sources of ssDNA vulnerable to damage-induced mutagenesis and clustered mutations. Recent studies revealed that long ssDNA associated with DSBs can result not only from strand degradation (5′→3′ resection) around a DSB but also in the unusual break-induced replication (BIR) fork. BIR is initiated by homologous pairing and strand invasion of only one end of a DSB into intact homologous dsDNA, thereby initiating replication from the end of the break using the intact molecule as a template (Malkova and Ira, 2013). During this process, BIR generates kilobases of ssDNA detectable by microscopy and 2-D gel electrophoresis (Saini *et al*., 2013). It remains to establish whether this ssDNA is sufficiently persistent to tolerate multiple lesions to result in clusters of multiple mutations. Potential sources of ssDNA could be abnormal nucleotide excision repair (Ma *et al*., 2013) and/or R-loops formed by the annealing of transcripts to one DNA strand leaving the other strand unpaired (Aguilera and Garcia-Muse, 2012). However, it remains to establish whether these processes could account for formation of kilobases of ssDNA needed for cluster formation. **See also**: Immunoglobulin Gene Rearrangements
2. Sources of ssDNA-assoicated clusters other than APOBECs. Widespread mutagenesis by ssDNA-specific APOBEC cytidine deaminases indicates that this DNA form can be targeted by other mutagens. Mutations originated in ssDNA of cancer cells may locate in clusters or scatter as single events across the genome. These could be mutations induced by endogenous mutagenic sources or environmental factors, some of which may be as specific to ssDNA as APOBECs. The latter can be identified in model systems relying on transient ssDNA created *in vivo* through the use of special conditions and/or genetic defects (Chan *et al*., 2012).
3. Mechanisms of cluster formation other than damaged ssDNA. Mutation clusters formed by mechanisms other than damage to ssDNA are elusive at the moment but can be expected if there are areas of the genome in which dsDNA lesion repair is completely inhibited, for example, by special features of chromatin (Thoma, 2005) or genome locus (Rochette and Brash, 2010).
4. Somatic mutations accumulated in human tissues over a lifetime. It is unclear at the moment how many somatic mutations can be accumulated through a lifetime's proliferative history or through dozens of quiescent states for cells in healthy tissues (e.g. see discussion in Fox *et al*., 2010; Shibata and Lieber, 2010). If these numbers are comparable to the mutation load in cancer, the question of clustered mutagenesis can and should be addressed.
5. Mutation clusters in the germ line on population and evolution scales. The effect of simultaneously induced mutation clusters on a population scale would require accumulation of a significant number of germ line mutation catalogues for human triads of both parents and a child. Because the number of *de novo* mutations per germ line generation is low, there may be a significant wait before having sufficient statistical power to explore clustering. Selection for specific mutations that happen to be closely spaced could make addressing this question even more difficult. Nevertheless, there has been a recent report indicating an apparent clustering component in *de novo* germ line mutations in autistic patients (Michaelson *et al*., 2012). Spatial

concentration of mutations during primate evolution – human accelerated regions (Pollard *et al*., 2006) could also involve mutagenic mechanisms with the potential to cause mutation clusters, in addition to more frequently cited explanations, such as selection and biased gene conversion.

In general, understanding the abundance of and mechanisms forming mutation clusters appears of interest beyond the field of mutagenesis mechanisms. It extends into more general issues of genome structure and function, organism development, cancer and evolution. Applying a combination of mechanistic research with bioinformatics mining and analysis of genomics data to address question concerning this phenomenon turned to be productive and promises additional interesting results in the future.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. Steven A. Roberts was also supported by NIH Pathway to Independence Award K99ES022633–01. The authors declare no conflict of interest.

References

Aguilera A and Garcia-Muse T (2012) R loops: from transcription byproducts to threats to genome stability. *Molecular Cell* **46**(2): 115–124.

Alexandrov LB, Nik-Zainal S, Wedge DC *et al*. (2013) Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415–421.

Bass AJ, Lawrence MS, Brace LE *et al*. (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature Genetics* **43**(10): 964–968.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate – A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B:Methodological* **57**(1): 289–300.

Berger MF, Lawrence MS, Demichelis F *et al*. (2011) The genomic complexity of primary human prostate cancer. *Nature* **470**(7333): 214–220.

Burch LH, Yang Y, Sterling JF *et al*. (2011) Damage-induced localized hypermutability. *Cell Cycle* **10**(7): 1073–1085.

Burns MB, Temiz NA and Harris RS (2013) Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature Genetics* **45**(9): 977–983.

Chan K, Resnick MA and Gordenin DA (2013) The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair*. doi: 10.1016/j.dnarep.2013.07.008.

Chan K, Sterling JF, Roberts SA *et al*. (2012) Base damage within single-strand DNA underlies *in vivo* hypermutability induced by a ubiquitous environmental agent. *PLoS Genetics* **8**(12): e1003149.

Chang K, Creighton CJ, Davis C *et al*. (2013) The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45**(10): 1113–1120.

Chapman MA, Lawrence MS, Keats JJ *et al*. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**(7339): 467–472.

Chelico L, Pham P and Goodman MF (2009) Mechanisms of APOBEC3G-catalyzed processive deamination of deoxycytidine on single-stranded DNA. *Nature Structural and Molecular Biology* **16**(5): 454–455; author reply 455–456.

Conticello SG, Langlois MA, Yang Z and Neuberger MS (2007) DNA deamination in immunity. AID in the context of its APOBEC relatives. *Advances in Immunology* **94**: 37–73.

Drier Y, Lawrence MS, Carter SL *et al*. (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Research* **23**(2): 228–235.

Forbes SA, Bindal N, Bamford S *et al*. (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research* **39**(Database issue): D945–D950.

Fox EJ, Beckman RA and Loeb LA (2010) Reply: is there any genetic instability in human cancer? *DNA Repair* **9**(8): 859–860.

Friedberg EC, Walker GC, Siede W *et al*. (2006) *DNA Repair and Mutagenesis*. Washington: ASM Press.

Futreal PA, Coin L, Marshall M *et al*. (2004) A census of human cancer genes. *Nature Reviews Cancer* **4**(3): 177–183.

Hanahan D and Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**(5): 646–674.

Koren A, Polak P, Nemesh J *et al*. (2012) Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics* **91**(6): 1033–1040.

Lada AG, Stepchenkova EI, Waisertreiger IS *et al*. (2013) Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an APOBEC deaminase. *PLoS Genetics* **9**(9): e1003736.

Lawrence MS, Stojanov P, Polak P *et al*. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**(7457): 214–218.

Ma W, Westmoreland JW and Resnick MA (2013) Homologous recombination rescues ssDNA gaps generated by nucleotide excision repair and reduced translesion DNA synthesis in yeast G2 cells. *Proceedings of the National Academy of Sciences of the USA* **110**(31): E2895–E2904.

Malkova A and Haber JE (2012) Mutations arising during repair of chromosome breaks. *Annual Review of Genetics* **46**: 455–473.

Malkova A and Ira G (2013) Break-induced replication: functions and molecular mechanism. *Current Opinion in Genetics and Development* **23**(3): 271–279.

Maul RW and Gearhart PJ (2010) AID and somatic hypermutation. *Advances in Immunology* **105**: 159–191.

Meyerson M, Gabriel S and Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing.*Nature Reviews Genetics* **11**(10): 685–696.

Michaelson JJ, Shi Y, Gujral M *et al*. (2012) Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* **151**(7): 1431–1442.

Mimitou EP and Symington LS (2011) DNA end resection–unraveling the tail. *DNA Repair* **10**(3): 344–348.

Nik-Zainal S, Alexandrov LB, Wedge DC *et al*. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**(5): 979–993.

Pleasance ED, Cheetham RK, Stephens PJ *et al*. (2010a) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**(7278): 191–196.

Pleasance ED, Stephens PJ, O'Meara S *et al*. (2010b) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**(7278): 184–190.

Plosky BS and Woodgate R (2004) Switching from high-fidelity replicases to low-fidelity lesion-bypass polymerases. *Current Opinion in Genetics and Development* **14**(2): 113–119.

Pollard KS, Salama SR, King B *et al*. (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**(10): e168.

Refsland EW and Harris RS (2013) The APOBEC3 family of retroelement restriction factors. *Current Topics in Microbiology and Immunology* **371**: 1–27.

Roberts SA, Lawrence MS, Klimczak LJ *et al*. (2013) An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics* **45**(9): 970–976.

Roberts SA, Sterling J, Thompson C *et al*. (2012) Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular Cell* **46**(4): 424–435.

Rochette PJ and Brash DE (2010) Human telomeres are hypersensitive to UV-induced DNA Damage and refractory to repair. *PLoS Genetics* **6**(4): e1000926.

Saini N, Ramakrishnan S, Elango R *et al*. (2013) Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature*. doi: 10.1038/nature12584.

Shibata D and Lieber MR (2010) Is there any genetic instability in human cancer? *DNA Repair* **9**(8): 858; discussion 859–860.

Smith HC, Bennett RP, Kizilyer A *et al*. (2012) Functions and regulation of the APOBEC family of proteins. *Seminars in Cell and Developmental Biology* **23**(3): 258–268.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE *et al*. (2009) Human mutation rate associated with DNA replication timing. *Nature Genetics* **41**(4): 393–395.

Stransky N, Egloff AM, Tward AD *et al*. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**(6046): 1157–1160.

Stratton MR (2011) Exploring the genomes of cancer cells: progress and promise. *Science* **331**(6024): 1553–1558.

Taylor BJ, Nik-Zainal S, Wu YL *et al*. (2013) DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2**: e00534.

Thoma F (2005) Repair of UV lesions in nucleosomes–intrinsic properties and remodeling. *DNA Repair* **4**(8): 855–869.

Wang J, Gonzalez KD, Scaringe WA *et al*. (2007) Evidence for mutation showers. *Proceedings of the National Academy of Sciences of the USA* **104**(20): 8403–8408.

Yang Y, Gordenin DA and Resnick MA (2010) A single-strand specific lesion drives MMS-induced hyper-mutability at a double-strand break in yeast. *DNA Repair* **9**(8): 914–921.

Yang Y, Sterling J, Storici F *et al*. (2008) Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. *PLoS Genetics* **4**(11): e1000264.

Zhang J, Baran J, Cros A *et al*. (2011) International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database: the journal of biological databases and curation* 2011: bar026.

Further Reading

Donley N and Thayer MJ (2013) DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Seminars in Cancer Biology* **23**(2): 80–89.

Eifert C and Powers RS (2012) From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. *Nature Reviews Cancer* **12**(8): 572–578.

Ghosal G and Chen J (2013) DNA damage tolerance: a double-edged sword guarding the genome. *Translational Cancer Research* **2**(3): 107–129.

Kuong KJ and Loeb LA (2013) APOBEC3B mutagenesis in cancer. *Nature Genetics* **45**(9): 964–965.

Lambert S and Carr AM (2013) Replication stress and genome rearrangements: lessons from yeast models. *Current Opinion in Genetics and Development* **23**(2): 132–139.

Sale JE (2013) Translesion DNA synthesis and mutagenesis in eukaryotes. *Cold Spring Harbor Perspectives in Biology* **5**(3): a012708.

Setlur SR and Lee C (2012) Tumor archaeology reveals that mutations love company. *Cell* **149**(5): 959–961.

Sprouffske K, Merlo LM, Gerrish PJ *et al*. (2012) Cancer in light of experimental evolution. *Current Biology* **22**(17): R762–R771.