Published in final edited form as:
Nat Genet. 2020 September 01; 52(9): 958–968. doi:10.1038/s41588-020-0674-6.

DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers

David Mas-Ponte[1], **Fran Supek**[1,2,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

Abstract

Certain mutagens, including the APOBEC3 (A3) cytosine deaminase enzymes, can create multiple genetic changes in a single event. Activity of A3s results in striking ‘mutation showers’ occurring near DNA breakpoints, however less is known about mechanisms underlying the majority of A3 mutations. We classified the diverse patterns of clustered mutagenesis in tumor genomes, which identified a novel A3 pattern: nonrecurrent, diffuse hypermutation (*omikli*). This mechanism occurs independently of the known focal hypermutation (*kataegis*), and is associated with activity of the DNA mismatch repair (MMR) pathway, which can provide the single-stranded DNA substrate needed by A3 and contributes to a significant portion of A3 mutations genome-wide. Because MMR is directed towards early-replicating, gene-rich domains, A3 mutagenesis has a high propensity to generate impactful mutations, which exceeds other common carcinogens such as tobacco smoke and UV exposure. Cells direct their DNA repair capacity towards more important genomic regions, thus carcinogens that subvert DNA repair can be remarkably potent.

Introduction

Many types of mutation patterns in somatic cells are linked either with exposure to DNA damaging agents, or with genome instability resulting from failures of DNA repair. Both are causal factors for carcinogenesis due to increases in mutation rates. In addition, dysregulated activity of certain enzymes may be mutagenic. For example, many tumors as well as the human germline bear signatures of error-prone DNA polymerases [1–4]. However, the most striking example of endogenous mutagens is the APOBEC family of cytosine deaminases.

[*]Correspondence: fran.supek@irbbarcelona.org.

Author contributions
F.S. and D.M.P. conceptualized the study and devised the methodology. D.M.P. carried out the formal analysis, the investigation, operated the software and performed data visualization. D.M.P. and F.S. wrote and edited the draft manuscripts. F.S. acquired the funding and supervised the study.

Competing interests
The authors declare no competing interests.

Reporting Summary
Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

They defend against viruses and retrotransposons by damaging their genetic material; additionally, APOBEC1 is an mRNA editing enzyme (reviewed in ref. [5]).

The protein products of *APOBEC3* (A3) paralogs were implicated as mutagens in many human cancer types [6–10]. This is consistent with their ability to deaminate DNA [11,12] when it is single-stranded (ss) [13,14]. Tumors have a highly variable burden of the A3 mutational spectrum, which is associated with differential A3 activity: an activating germline polymorphism in *APOBEC3A* and *APOBEC3B* genes results in a higher mutation burden [15], and there is some correlation thereof with tumoral mRNA expression level of *APOBEC3A* and *APOBEC3B* [4,7,16,17]. In addition to the A3 activity, the availability of its ssDNA substrate is a requirement for mutagenesis. One known source of such ssDNA are intermediates of DNA repair of double-stranded breaks [10,18,19], where A3 results in ‘mutation showers’ or *kataegis* (greek for thunderstorm), local hypermutation events that may consist of tens of mutations [8,10]. While *kataegis* is striking, it is not common: very few of the A3-signature mutations are accounted by the mutation showers [10,20]. Additionally, DNA secondary structures can generate A3 mutational hotspots [21], however, the processes that generate global, abundant ssDNA substrate for A3 mutagenesis need to be further explored.

Clues are provided by the peculiarities of the A3 mutation patterns. Most mutation types are enriched in late-replicating domains, because DNA mismatch repair (MMR) and possibly nucleotide excision repair are more active in early-replicating domains [22,23]. A3 signature mutations run counter to this trend [20]. Additionally the A3 mutations have a curiously strong DNA replication strand bias [24–26]. These biases, considered together with experimental evidence [27–29], suggest that A3 mutagenic activity is coupled to DNA replication. Expressing A3 enzymes in *E. coli* and yeast produced a mutational bias at replication origins [30,31], suggesting that ssDNA exposed during discontinuous DNA synthesis may be vulnerable to A3. In addition, another source of A3 substrate ssDNA was suggested by experiments in which the repair of a lesion-bearing DNA by base excision repair (BER) and MMR promoted A3 signature mutagenesis in flanking segments [32]. Identifying the mechanisms that allow access of A3s to nuclear DNA is important because A3 enzymes generate cancer driver mutations[21,33–35] and promote tumor heterogeneity [36–38].

Kataegis illustrates how mutation clustering patterns can be used to detect ssDNA generating mechanisms [10,18]. We introduce a sensitive statistical method to detect non-random mutation distribution that results from localized mutagenic events. Applying this to human cancer genomes uncovered a ubiquitous pattern of diffuse A3 mutation clusters, which we named *omikli* (greek: ομίχλη, meaning “fog”). This ‘mutation fog’, *omikli*, is more common than *kataegis*, however it occurs via a distinct mechanism. We present evidence that the activity of DNA mismatch repair (MMR) promotes A3 mutagenic activity, evident in the *omikli* pattern, and that the same process is responsible for the majority of unclustered A3 mutations. They are surprisingly likely to impact cancer genes – more so than the changes resulting from common external mutagens – because DNA repair directs A3 mutagenesis towards early-replicating, gene rich domains.

Results

Detection of two distinct types of local hypermutation

Our aim was to systematically characterize the different types of mutation clustering in human cancer cells. To this end, we developed a statistical approach (HyperClust) that has two distinguishing features (Fig. 1a; Extended Data Fig. 1a, b). Firstly, it accounts for the heterogeneity of mutation rates and of trinucleotide composition across chromosomal domains, which is an extension of our recent approach [4] with additional support for local false discovery rate (*lfdr*) statistics. Secondly, it draws on the signal present in allelic frequencies of mutations –serving as a proxy for mutation timing – to enforce that mutations constituting one clustered event must occur simultaneously (Methods). We tested these improvements in HyperClust using simulated data with spiked-in mutation clusters, generating precision-recall curves(Extended Data Fig. 1c-e), comparing HyperClust to two previous approaches for detecting clustered mutations [8,10,29]. Our simulation studies suggest that HyperClust compares favorably in calling shorter clusters consisting of two mutations (at various intermutational distance (IMD) distributions, Extended Data Fig. 1e). Therefore our method supports systematic studies of diverse types of clustered mutagenesis.

We used HyperClust to identify clustered somatic single-nucleotide variants in whole-genome sequences of 22 tumor types, detecting a total of 108,401 clustered mutations in 699 tumors (at a *lfdr*≤20%). Henceforth, we defined the A3 spectrum as C>T and C>G changes in a T<u>C</u>W context (W is A or T). Overall 45% of all clustered mutations are in A3 contexts, consistent with A3 enzymes being an important cause of local hypermutation, however 55% of mutation clusters are not in the canonical A3 context, supporting that additional processive agents including error-prone DNA polymerases commonly mutagenize human cells[1–4,39] (we note that A3 may also rarely generate C>A changes [40]). In contrast to prior heuristic rules [29,41,42] that required e.g. at least 5 mutations with an IMD ≤1kb, importantly, the majority of A3 clusters do not meet this definition and instead consist of pairs and triplets (Fig. 1b, c). The distribution of A3 mutation cluster lengths (number of consecutive mutations) was significantly better described by a mixture of two distributions than by a single distribution (Fig. 1d; Extended Data Fig. 1f, g). This suggests that there are at least two types of mutagenesis generating tracts of A3-context changes, which we estimate to have a mean length of 2.2 mutations and 7.1 mutations.

While the latter distribution neatly fits current notions of *kataegis*, the former one does not. We named this type of diffuse mutation clustering *omikli* (fog), by analogy to the focused *kataegis* (thunderstorm) events. Henceforth, we classify mutation clusters with 2, 3 or 4 variants as *omikli* (the short-tract Poisson mixture component predominates; Fig. 1d), and clusters with 5 or more single-nucleotide variants as *kataegis* (with ≥95% contribution of the component with long tracts; Fig. 1d). *Omikli* is ubiquitous, occuring in more tumors (76% tumors contain at least three A3 *omikli* mutations; by random expectation approx. 14% would do so; Fig. 1e) than A3 *kataegis* (48% samples with at least three A3 *kataegis* mutations). In tumors in which they occur, A3 *omikli* are similarly abundant per genome (Q_1-Q_3: 4-36 mutations) as A3 *kataegis* (6-36 mutations; Fig. 1f, Extended Data Fig. 1h).

Distinct mechanisms for *kataegis* and *omikli* A3 mutagenesis

Multiple lines of genomic evidence suggest that A3 *omikli* clusters are generated by a mechanism distinct from *kataegis*. First, *kataegis* is, expectedly [8,10], enriched near rearrangement breakpoints, a proxy for locations of chromosome breaks [43], but not so for *omikli* (Fig. 1g). Second, the burden of A3 *omikli* clusters appears uncoupled from *kataegis* across individual tumors and is weakly correlated (R^2=0.11) with long *kataegis* events (≥8 mutations; Fig. 1h), suggesting that short clusters derive from a different mechanism than the intermediate and long ones, which share a common mechanism (R^2=0.52; Fig. 1h). Third, correlation of A3 mutation burden with APOBEC3A and APOBEC3B mRNA levels is stronger for *omikli* (Spearman rho=0.31 and 0.45, respectively) than for *kataegis*(rho=0.04 and 0.14). This suggests that for *omikli* the A3 expression is commonly limiting, while for *kataegis* another factor becomes limiting, plausibly the source of ssDNA that is available only rarely, e.g. during repair of ds breaks [10,18,44]. Fourth, the 5' mutational context of A3 *omikli* mutationshad a significant enrichment of the A3A-like context over the A3B-like context [45] in five cancer types, compared to *kataegis*(Extended Data Fig. 2a–c; the converse was not the case in any cancer type), thus A3A and A3B may have preferential roles in causing *omikli* and *kataegis*, respectively. We also note overall tissue-specific differences A3A-like *versus* A3B-like contexts, as reported [4,45](Extended Data Fig. 2c). Fifth, the unclustered A3 mutation burden is highly correlated with *omikli* (rho=0.66) but less with *kataegis* (rho=0.27). The numerous unclustered A3 mutations can be seen as a mixture of three components: singletons created by the *omikli* process (henceforth, A3-O), singletons created by the *kataegis* process (A3-K), and the remainder (A3-X) would encompass mutations caused by A3s independently of *kataegis* and *omikli* mechanisms plus the TCW>K mutations not caused by A3s. Consistently, the distribution of the numbers of mutations per cluster in *omikli* (Fig. 1d; >98% are pairs or triplets) suggests that A3-O generates many A3 singletons while A3-K generates few.

Regional distribution of A3 clusters suggests a link to MMR

To gain insight into the process generating *omikli*, we studied its distribution across the genome. A3-context *omikli* mutations were strongly enriched in early-replicating regions (2.0-fold and 2.5-fold for C>T and C>G respectively, Fig. 2a, b), in contrast to unclustered TCW (0.54 and 0.72-fold) and to the control, non-A3 context (VCN, where V is not T; 0.56 and 0.47-fold). These latter enrichments are similar to various other unclustered mutation types (Extended Data Fig. 3a), which are known to be depleted from early-replicating domains [46–48]. Protection of early-replicating domains from mutations stems from the differential activity of DNA mismatch repair (MMR) [4,22,49]. The enrichment of diffuse clustered A3 mutations (*omikli*), uniquely, matches the genomic gradient of increasing MMR activity, rather than that of decreasing MMR activity, as for most other mutation types (this is not explained by the genomic distribution of the TCW trinucleotide; Extended Data Fig. 3b).

MMR is directed towards the regions bearing the H3K36me3 histone mark [50], which is enriched at gene bodies of expressed genes [51,52], lowering their mutation rates [4,53]. Consistently with higher MMR activity, we find a significant enrichment of A3 *omikli* clusters at H3K36me3 regions, after conditioning on replication time and gene expression levels (Fig. 2c; Methods). However, the mRNA level, after conditioning on H3K36me3 and replication time, was not associated with higher A3 *omikli* burden (Fig. 2c). This agrees with prior data [20,31] suggesting that transcription is not a common source of ssDNA substrate for A3 enzymes, even though ssDNA generated during transcription can be prone to mutagenic spontaneous deamination [54]. Regarding A3 *kataegis*, the enrichment in H3K36me3 regions (Extended Data Fig. 3c,d) might stem from recruitment of the homologous recombination machinery (that can generate ssDNA tracts) by this histone mark [55].

We further examined a set of regions proximal to CpG dinucleotides, proposed to be linked with differential MMR activity [56]. There were more A3 *omikli* clusters in the top genomic tertile by CpG density (Extended Data Fig. 3e). Consistently with MMR activity causing the mutations, this difference was more pronounced within early-replicating regions. The mutation rate of the control V$\underline{C}$H context in CpG-dense regions was, in contrast, lowered (Extended Data Fig. 3e) [56].

Next, we examined the replication strand bias [24,25] of A3 clusters. The ratio of A3 *omikli* in the leading *versus* the lagging DNA strand closely matched that observed in MMR-deficient (microsatellite instable, MSI) tumors (1.006-fold difference, Fig. 2d), but was less compatible with strand bias associated with mutated proofreading domain of the leading strand-specific DNA polymerase epsilon (POLE, 0.81-fold difference). This suggests that the strand asymmetry of postreplicative MMR activity [57] rather than the asymmetry of DNA replication itself [58] underlies *omikli*; see Supplementary Note.

APOBEC mutagenesis hotspots can occur in DNA sequences that form hairpin secondary structures [21]. Our data do not reflect this: *omikli* after excluding hairpin loci maintained the early replication time enrichment at 2.16-fold.

Coupling of A3 mutagenic mechanisms with DNA replication.

We hypothesized a mechanism by which MMR promotes A3 mutagenesis. MMR generates a single-stranded (ss) DNA intermediate during excision of a mutated DNA segment [59,60]. This provides an opportunity for A3 enzymes to cause DNA damage that converts into clustered mutations, wherein such mutation tracts are short (*omikli*) because the ssDNA segments are short. The widespread occurrence of A3 *omikli* clusters is consistent with most tumors being largely MMR-proficient [61–63]. This is in contrast to *kataegis*, which is known to also stem from DNA repair intermediates, however, these longer segments result from processing of double-strand breaks [10,18,19,40]. The MMR mechanism would explain the enrichment of A3 diffuse clustered mutations in early-replicating domains, and also enrichment in the lagging DNA strand, both associated with higher MMR activity [22,57]. Because MMR is largely replication-coupled [64,65], the MMR-associated A3 mutagenesis is consistent with the greater vulnerability to A3 damage in dividing cells [27].

An additional hypothesis was proposed to explain the associations of A3 mutations with DNA replication-related genomic features [20,47]: ssDNA exposed during discontinuous synthesis of the lagging strand would be mutagenized by A3. This was proposed based on strand-biased mutations that result from expressing human A3s in *Escherichia coli* [30] and in yeast [31]. Because length of eukaryotic Okazaki fragments is known, and length of MMR intermediates has been characterized in eukaryotic systems reconstituted *in vitro* [66,67], we next examined the length distribution of inter-mutational distances (IMD) in the A3 clustered mutations.

The IMD distribution for A3 *omikli* has a global peak at 355 nt, closely matching the peak (378 nt) of a simulated IMD distribution resulting from 800 nt long ssDNA segments (Fig. 2e, Methods). The length of MMR excision tracts was estimated at 800 nt using *in vitro* studies of human and yeast MMR [66,68]. Additionally, we approximated the length of MMR tracts by an analysis of somatic hypermutation events in lymphomid genomes (Methods); this suggested an approx. 400-1000 nt length range (Extended Data Fig. 4a, b). In contrast, the global peak in *omikli* IMD was not compatible with the approx. 200 nt long Okazaki fragments [67], which would generate a peak at 96 nt (Fig. 2e).(Of note, in *kataegis* events, IMD are devoid of the peak corresponding to ~800 nt length tracts (Fig. 2e), thus *kataegis* would result independently of MMR). These data suggest that discontinuous lagging strand synthesis is not the main mechanism supplying ssDNA that yields A3 clustered mutations because the observed IMDs are too long. However the IMDs are compatible with MMR-supplied ssDNA. Moreover, the proposed mechanism agrees with the early replication time enrichment of A3 *omikli*, which is consistent with higher MMR activity.

We do not exclude however that the discontinuous synthesis of the lagging strand contributes to A3 mutagenesis because the *omikli* IMD distribution has a secondary peak corresponding to 200 nt segment lengths (Fig. 2e). Modelling the IMD as a mixture of gamma distributions (Fig. 2f) suggests that up to one-quarter of A3 clusters might be generated by a process corresponding to ~200 nt long segments (Extended Data Fig. 4c, d). Notably, the mixture modelling also suggests a minor component in *omikli* IMD at very short peak lengths (~25 nt, Fig. 2f). It is tempting to speculate that this reflects the binding of the ssDNA protective protein RPA, which has a 24-30 nt footprint [69,70]. A secondary IMD peak of this length is observed also in *kataegis* (Fig. 2e; see Methods for limitations of use of IMD measure for *kataegis* analyses).

MMR deficiencies are associated with lower A3 mutagenesis

We next examined the tumors exhibiting microsatellite instability (MSI), which are MMR deficient; we took care to adjust for different statistical power to detect clusters in these high mutation burden tumors (Extended Data Fig. 4e, f) making the following analyses conservative.

We compared the fraction of A3 *omikli* mutations in MSI and microsatellite stable (MSS, MMR-proficient) tumors of the matched cancer types (Fig. 3a). Supporting our hypothesis, the fraction of A3 *omikli* clusters in the MSI samples was significantly lower than in the MSS tumors (p<0.001 by Mann-Whitney test; 5.52-fold difference between the median of samples), but there was no significant difference in the non-A3-context (V$\underline{\text{C}}$N>K) clusters (p=0.34, 1.2-fold difference; Fig. 3a). Of note, comparing absolute, i.e. not normalized to overall number of mutations, *omikli* A3 burdens were also lower in MSI (p<0.01, Extended Data Fig. 4g). Therefore, the depletion of A3 clusters is in contrast with the overall increase of mutation load in MSI tumors: MMR normally protects against many types of mutations but provides an opportunity for A3. The MSI-MSS difference is consistently observed across three cancer types (4.0, 3.7 and 12.1-fold enrichment of A3 *omikli* in MMR proficient MSS tumors, Fig. 3a) and the overall difference is significant after stratifying by cancer type (Fig. 3b, pooled p<0.001, Fisher's method for combining p-values).

The early replication enrichment of *omikli* is not observed in MSI (Fig. 3c), but instead a profile more similar to unclustered mutations is seen, further supporting that MMR directs the A3 mutagenesis. Consistently, A3 *omikli* burden associates with expression levels and copy number status of MMR genes *MSH6, MSH2* and *EXO1* (Fig. 3d, e; Extended Data Fig. 3f, g; discussed in Supplementary Note).

We have further validated findings on an independent set of 2,304 tumor whole genome sequences (WGS, Methods). This supported the dichotomy between A3 *kataegis* and *omikli* clustering in tract lengths (Extended Data Fig. 5a-c). The key evidence that links A3 mutagenesis to MMR activity validates: there is a strongly increased A3 *omikli* fraction in MSS *versus* MSI cancers, in a data set stratified by cancer type, here also including additional tissues such as prostate and breast; this difference is however modest in the control, non-A3 context (Extended Data Fig.5d, e). Moreover, additional supporting evidence of MMR involvement validates in these data: significantly increased A3 *omikli* burdens in tumors with copy number gains in *MSH6* and *MSH2* and *EXO1* genes (Extended Data Fig. 5f), and the altered regional distribution of A3 *omikli* between MSS (enriched in early-replicating) and MSI cancers (less enriched) (Extended Data Fig. 5g). The IMD distributions of A3 *omikli* similarly have a peak corresponding to approx. 800 nt long vulnerable DNA segments (Fig. 2e; Extended Data Fig. 5h). Finally, an analysis of >3,000 whole-exome sequences showed a 3.02-fold excess of nearby TCW mutation pairs (within 1 kb), compared to more distant TCW pairs, in MSS over MSI samples; we also note the overall differences in TCW mutation burden in MSS *versus* MSI (Extended Data Fig. 5i, j). This further supports the association between A3 local hypermutation and MMR activity, which – as suggested by our IMD analysis – may stem from the ssDNA excision tracts generated during MMR. However other molecular mechanisms may similarly be able to explain the MMR-associated A3 mutagenesis, such as changes in replication fork dynamics.

Contribution towards the global A3 mutation burden

While *kataegis* and *omikli* clusters are informative markers of certain mutational processes, their numbers are low. We quantified the contribution of the two clustered A3 processes to the (much more abundant) unclustered mutational burden using a regression analysis, similar to ref. [4]; see Methods. Informally, a correlation between clustered burden of tumor samples and unclustered burden in the same mutational context suggests that the same process underlies the clustered and unclustered component (Fig. 4a shows A3 *omikli* and *kataegis* fits for lung adenocarcinoma; the former is a good fit, while the latter a poor one).

In the pan-cancer data, we estimated that the *omikli* process contributes approximately two-thirds of all A3 context mutations (A3-O, 66.4%, Fig. 4b), while the *kataegis* contribution is negligible (A3-K, ~0%) and an unknown process (or a mix thereof) contributes the remaining nearly one-third of A3 context mutations (A3-X, 32.4%; Fig. 4b). The lack of *kataegis* contribution is not unexpected, given that this process generates long tracts but almost never pairs or triplets (Fig. 1d) and thus by extension singletons would not be generated. The presence of mutations originating from the A3-X process, which is not associated with *omikli* and thus likely independent of MMR, suggests that the MMR hypothesis is one of the possible explanations for the mechanisms that generate the global pool of ssDNA vulnerable to A3.

We also considered cancer types individually (Extended Data Fig. 6), showing that the relative contribution of A3-O was strongly correlated with the absolute A3 mutation burden across cancer types (Fig. 4c). This further supported that a MMR-dependant, likely A3A-driven process which can be diagnosed via *omikli* is the major source of APOBEC mutagenesis in human cancer. This creates very high A3 mutation burdens in lung, breast, bladder and head-and-neck cancers (Fig. 4c), while other cancer types such as prostate – even though *kataegis* is known to occur therein – exhibit less *omikli* and lower overall A3 mutation burdens.

A3 mutagenesis has a high functional impact per mutation

Certain mutational processes – including A3 activity, MMR failures and use of translesion DNA polymerases – were reported to, atypically, produce many mutations in early-replicating, gene-rich chromosomal domains [4,26]. Such 'mutation redistribution' [71] means that at an equal global mutation burden, different mutagens may have different potential for affecting genes, thus having varied functional consequences. To quantify this, we introduce a concept of 'functional impact density' (FID) of a mutational process: the fraction of putatively impactful mutations among all mutations observed.

In case of cancer, a simple estimate of the oncogenic FID is the fraction of changes affecting coding regions of known cancer genes ('oncogenic mutations per thousand', henceforth OMPK; Methods). This is based on the reasonable assumption that many mutations occurring in a typical cancer gene are oncogenic and also that the set of 299 frequently mutated cancer genes [72] contains many of the driver mutations found in a tumor.

We examined the oncogenic FID of A3-O and A3-K mutations, as estimated from total A3 burden in tumors that harbor predominantly *omikli* or predominantly *kataegis* clusters(Methods). This was compared to common mutagenic processes [6] associated with tobacco smoking (C>A in lung), UV exposure (C>T in skin), exposure to gastric acid (A>C in stomach) and finally with aging (C>T changes at CpG dinucleotides). A3 mutations derived either from *omikli* or from *kataegis* processes have very high oncogenic FID: 0.47 and 0.46 OMPK, respectively (Fig. 5a, Methods), approximately twice that of common external mutagens: tobacco smoking and stomach acid-associated mutations, both at 0.24 OMPK, and of UV at 0.19 OMPK.

In addition to A3, another endogenous mutagenic process – the aging-associated C>T changes at CpG dinucleotides – also had high oncogenic FID per mutation (Fig. 5a). This is in line with a high frequency of CpG dinucleotides in coding regions in the human genome (Extended Data Fig. 7a); consistently, aging-related mutagenesis was suggested to have a higher risk of generating coding mutations than cancer chemotherapeutics did [73]. Of note, the A3 TCW context is not markedly enriched in coding regions so the high FID of A3 mutations is irrespective of trinucleotide composition therein.

We asked if the high FID of A3 mutagenesis stems from increased positive selection on oncogenic changes introduced by A3. Using intronic mutation rates as a baseline [74] (Methods), we find that selection on A3 mutations is not stronger than on external mutagen-induced changes (Extended Data Fig. 7b), which agrees with recent reports [33].

Instead, we hypothesized the higher FID of A3 results from the increased susceptibility of the affected genes to DNA repair as they are more often located in early-replicating euchromatic domains [22,23,25,75] than intergenic regions are. The high intronic/intergenic ratio shows that A3 mutagenesis is strongly redistributed towards genic DNA, compared to the various external mutagens (Extended Data Fig. 7b). The difference of FID of A3 processes *versus* external mutagens is exaggerated in cancer genes that reside in early-replicating regions (Extended Data Fig. 7c). This suggests that the *omikli*-driven A3 mutations are impactful due to an enrichment in gene-dense, early replicating domains, which are protected from many other mutation types. In addition to cancer genes, because somatic mutations might play a role in aging and neurodegeneration [76,77], we also examined a set of known essential genes, and a set of genes linked with neurodegeneration (Methods). Overall, we observed very similar results, with FID increases of A3 over the external mutagens ranging from 2 to 11-fold (Extended Data Fig. 7d, e).

A3 mutagenesis affects genes encoding chromatin modifiers

FID is a measure of the relative impact of a mutational process (expressed per mutation), however the absolute mutational burden of a process also needs to be considered. While tobacco smoking and UV mutations are less impactful, they are abundant. Aging-associated mutations are impactful per mutation but lowly abundant. The two A3 processes are however both impactful and abundant (Fig. 5a; error bars show variation across those tumors that were affected by a mutagenic process).

The absolute mutation burden strongly differentiates the *omikli* from the *kataegis* mutagenesis (A3-O and A3-K, respectively) even though their FID is similar. We estimate that the MMR-associated *omikli* process can generate, in tumors where it is highly active, approximately twice as many mutations with oncogenic potential (2.72 per tumor) than the DNA break repair-mediated *kataegis* process (1.32 per tumor) on average. Moreover, *omikli* generates twice as many oncogenic mutations as the aging-associated CpG mutagenesis. Notably, the A3 *omikli* process generates a comparable number of putatively oncogenic mutations per sample as the tobacco smoking (2.14 per tumor, in smokers' lung adenocarcinoma) and UV light (3.54 per tumor, in melanoma). This suggests that A3– considering jointly the (major) *omikli* and the (minor) *kataegis* components – may be an important carcinogen because, in exposed cells, it is able to create larger numbers of mutations in cancer genes than common external mutagens.

We observed a significant association between *omikli* burden and mutation occurrence (Methods) in 22 cancer genes at FDR<5%, and in 30 at FDR<10% (of 61 testable genes with ≥3 T$\underline{C}$W>K coding mutations in our data; Fig. 5b; Supplementary Table 1). However, no genes were significantly associated with *kataegis* burden (Extended Data Fig. 8a), supporting that *omikli* is more oncogenic than *kataegis*. The genes linked with *omikli* are enriched in tumor suppressors (n=14, *versus* 5 oncogenes; Fig. 5c) and are commonly chromatin modifiers (e.g. *KMT2A/C/D, NCOR1, SETD2, MECOM*) or chromatin remodelers (e.g. *PBRM1, ARID2*) (Fig. 5c) which have a higher count of TCW motifs in the coding sequence (Extended Data Fig. 8b). These associations do not however show the direction of the effect. We thus examined the control V$\underline{C}$N mutations, which were significantly associated in only 3 genes (Fig. 5b; Extended Data Fig. 8c). This suggests that the MMR-mediated A3 mutagenic pathway is an important source of cancer driver events. Consistently, cancer gene mutations in early-replicating regions are more strongly associated with overall *omikli* burden than those in late replicating regions (Extended Data Fig. 8d).

Discussion

Clustered mutations, even though rare, can occur in different types of clustering patterns, which serve as markers of different mutagenic processes. *Kataegis* originates from repair of double-stranded DNA breaks by the homologous recombination or break-induced replication pathways, which expose long tracts of ssDNA [18,40,78]. Here we propose that another DNA repair pathway –MMR –promotes A3 mutagenesis, generating *omikli* clusters and the bulk of A3 unclustered context mutations in human tumors. A different link of A3 with DNA repair was proposed recently, resulting from DNA lesions processed by the base excision repair (BER) pathway (abasic sites, uracils, or T:G mismatches), which generated A3-context mutations flanking the repaired site [32]. MMR was suggested to be able to‘hijack’ the BER intermediates to provide additional ssDNA substrate for A3 [32]. Our data suggest that MMR may generate A3 substrate ssDNA more generally, which could occur by processing mismatches occurring during DNA replication. We do not exclude that BER-processed lesions result in A3 mutagenesis in cancer; indeed this may help explain the approximately one-third of the unclustered A3 mutations (A3-X) that we do not account for via *omikli*. Another likely contributor to this MMR-independent A3 mutation fraction is A3 activity at ssDNA occurring discontinuous synthesis of the lagging strand in DNA replication**[24,25,30,31]**, which finds some support in our IMD distribution analyses.

MMR activity preferentially protects early-replicating, euchromatic regions from mutations [22,79,80] and additionally transcribed gene bodies therein, because it is recruited by the H3K36me3 histone mark [4,53]. Therefore, mutagenic processes that subvert MMR would be particularly dangerous because they are directed to active genes. One example of this is non-canonical MMR that recruits the error-prone DNA polymerase η (POLH protein) [81,82], who semutational signatures are seen across human tumors [2,4]. Here we provide another example of MMR activity leading to mutagenesis, in this case by promoting APOBEC activity. Based on the enrichment of MMR-associated A3-context mutations in early-replicating gene-rich chromosome domains, we propose that the MMR-A3A coupling has particularly high potential for generating impactful mutations, exceeding common exogenous mutagens. In addition to oncogenes and tumor suppressor genes, A3-context mutations were directed towards essential genes and neurological disease-associated genes, suggesting possible roles for APOBEC mutagenesis not only in cancer, but also more generally in aging-related pathologies.

Online methods

Data sources

Mutation calls for TCGA-WGS were obtained as in ref. [22]. In brief, BAM files were downloaded from the cgHub repository (now superseded by the NCI Genomic Data Commons) for normal and tumor samples, and somatic single-nucleotide variants were called with Strelka 1.0.6 [83]. Also as previously [4,22] we excluded mutations in blacklisted regions by UCSC (Duke and DAC) and in difficult-to-align genomic regions by the 'CRG Alignability 36' criterion, meaning we required genomic 36-mers to be unique in the hg19 genome assembly (even after allowing up to two mismatches).

SNP6 Affymerix microarray data were downloaded from the GDC legacy portal (portal.gdc.cancer.gov/legacy-archive) for matched donors, with both normal and tumor data available. The final dataset contained 699 TCGA samples with WGS mutations and SNP6 array data available. One of the donors (TCGA-CZ-5454) was excluded from those analyses that required external metadata as two different aliquots were available and metadata could not be unambiguously matched. This change makes the number of total samples equal to 697 in some analyses.

MSI status and other metadata for hypermutated tumors (i.e. POLE status) was obtained as described in ref. [22]. In total, our TCGA-WGS dataset contained 24 MSI samples (Supplementary Table 2).

An additional dataset, comprising WGS single nucleotide variants, purity estimates, and copy number alterations was obtained from the Hartwig Medical Foundation [84], was used for validation analyses in Extended Data Fig. 5a-h. This dataset has been processed similarly to our TCGA WGS (Strelka version 1.0.14 was used to call single-nucleotide variants) and additionally the Purple tool was used to infer purity and obtain CNA estimates [84] (Supplementary Table 3).

Inferred MSI/MSS labels [85] were obtained from the supplementary data of the corresponding publication [84]. We additionally discarded samples (n = 53) that were treated with temozolomide (TMZ), which is known to positively select for MMR deficient cells in brain tumors [86].

For the functional impact of UV mutations we additionally obtained WGS variant calls of 70 melanomas tumors from the MELA-AU study [87] within PCAWG. For the somatic hypermutation analyses, we additionally obtained WGS variant calls of blood tumors CLLE-ES and MALY-DE from the PCAWG dataset [88] available as controlled files in the ICGC data portal (https://dcc.icgc.org/pcawg). We selected the SANGER pipeline calls (Supplementary Table 4).

We obtained exonic mutations from the TCGA mc3 dataset, available at (https://gdc.cancer.gov/about-data/publications/mc3-2017) [89]. This dataset contains unified somatic mutation calls for approximately 10,000 whole-exome sequences (WES). We selected cancer types that had at least one sample classified as MSI (see below), therefore the subset used in this analysis comprised 5,831 tumors from 16 cancer types. Only 6% of the WES samples overlap with the WGS cohort. We obtained the MSI status from ref. [61], which contains experimentally determined MSI labels (for ESCA, UCEC, COAD, READ and STAD) and additionally inferred MSI status labels at 80% confidence level that covered additionally 11 cancer types (Supplementary Table 5).

The acronyms used for cancer types in this analysis are as listed in the ICGC Project portal page (https://docs.icgc.org/submission/projects/).

HyperClust, a randomization-based FDR estimation for local hypermutation detection.

The process of detecting local hypermutation (or mutation clusters) aims to distinguish those pairs of mutations that occurred in the same event from those that occurred independently. The classification is based primarily on intermutational distances (IMD) on the genomic sequence but other sources of information can be used such as the allelic fraction of the mutations.

We developed HyperClust building upon our recent approach [4] which employs a trinucleotide context-preserving randomization of mutations within megabase-sized chromosomal domains, obtaining a baseline frequency of mutation cluster occurrence at a certain IMD (Extended Data Fig. 1a). While the original approach applied a single IMD threshold at which every genome was evaluated, in HyperClust we compute significance estimates at the level of each mutation, meaning that many more samples could be analyzed while retaining acceptable false discovery rates.

HyperClust provides a rigorous estimate of the local FDR (*lfdr*) for each clustered mutation event, given its IMD and the baseline distribution of IMDs in that genome. It is also possible to stratify mutations pairs in each tumor sample into smaller sets according to different features. Because A3 mutagenesis occurs primarily in coordinated cytosines within ssDNA fragments [8,10], we stratified of mutation pairs according to base types (C:G and A:T) and to strand-coordinated bases. We additionally stratified by mutation clonal fraction, as it should be shared by the mutations occuring contemporaneously in a cluster (Supplementary Note).

We evaluated the different stratification features of HyperClust together with other local hypermutation detection approaches from the literature using 48 randomized tumor samples with simulated spiked-in mutation clusters. The stratification with both the strand-coordinated base types and clonal fraction of the mutations outperforms the other tested set ups and was therefore used to obtain mutations for the rest of the analysis (Supplementary Note).

Our method is designed to test pairs of mutations, instead of on larger groups, which leads to balanced power of detection for shorter clusters and longer clusters (*kataegis*-like), while previous methods tend to be better adapted to calling the latter.

Poisson mixture modelling of number of mutations per tract.

The aim of this analysis is to examine whether there exist multiple mechanisms generating clustered mutations, resulting in tracts of different lengths. The number of mutations per cluster can be modeled with a Poisson distribution. We considered only clustered events consisting of two or more mutations at T<u>C</u>W>K, which are likely to be a highly pure set of the A3 mutations. Then, we modeled the probability that *x* mutations occur in a fragment of ssDNA when two mutations are already present P(x| x = 2) = Pois (λ), meaning that 0 represents a cluster pair, 1 represents a triplet etc. If more than one biological mechanism generates clustered mutations at different tract lengths (number of mutations), the observed distribution would be better modeled as a mixture of two or more Poisson distributions, than by a single Poisson distribution.

We used the R package *flexmix*[90] to fit a mixture model, testing the range of components from 1 to 5. We transformed the Akaike Information Criterion (AIC) values extracted from the models to relative likelihoods by calculating the exponential of the difference between each AIC value and the minimum AIC (Extended Data Fig. 1f).

We performed a bootstrap likelihood test (*LR_test* function in *flexmix*) with 500 iterations. This test yields a p-value for the difference of the log-likelihood distributions between the selected model and one more or one less component.

The λ of each Poisson component is the exponential of the fitted intercept in the regression. The confidence intervals of the λ values were obtained by transforming the standard error of that value at C.I.= 95%. We used the λ values to compute density distributions of each component.

We then used the posterior probabilities to obtain the proportion of events with a given track length that can be attributed to each Poisson component(relevant for Fig. 1d, bars). We also obtained a random Poisson distribution for each component based on the λ (relevant for Fig. 1d, lines).

Samples from skin cancer (SKCM) and B-cell lymphoma (DLBC) were excluded from this analysis as they contain particular mutation properties that may confound our analysis. Skin cancer has a high percentage UV signature mutations which overlap with the APOBEC TCW>T context. Somatic hypermutation (SHM) is common in lymphomas and some mutations therein may present a similar profile to the APOBEC mutagenesis.

Association of increased A3 clustered burden with various genomic regions.

Genomic segments and bins extracted from chromatin marks were computed as in ref. [4]. In brief, data for epigenetic marks (H3K36me3) were downloaded from the Roadmap Epigenomics repository, stratified according to the fold-enrichment (FE) of that mark over the input, into three equal-sized bins where the FE>1, and additionally the bin 0, which correspond to regions with FE<1. Expression values were obtained from Roadmap Epigenomics for genic and intergenic regions and processed in a similar manner to the ChipSeq data. Replication time bins were computed from wavelet-smoothed RepliSeq signal tracks from the ENCODE dataset. Again, we binned the genome into equal-frequency bins where bin 1 is the latest-replicating quartile, and bin 4 is earliest-replicating quartile. These data were averaged over the 8 cell lines, as in ref. [4].

To detect significant associations of mutations in specific regions of the genome we used a negative binomial regression[4] (*glm.nb* from the *MASS* R package). In brief, combinatorial intersections between the genomic region sets were computed, 4 bins for each feature. In each set, the number of TCW>K mutations were stratified by the four A3 mutation types (TCA>T, TCA>G, TCT>T and TCT>G). These values (mutation counts stratified by mutation type) are used as the dependent variable in the regression and has a total length of 256, corresponding to 64 x 4 mutation types. The number of susceptible genomic sites in 64 bins was also computed and multiplied by the number of samples, thus representing the exposure variable. The three independent variables were the genomic bins of each feature, encoded as factors. This same approach was used for the control contexts (VCN>T). The 95% confidence intervals of the regression coefficient were computed with the *confint* function in R.

For this analysis, we excluded the DLBC (lymphoma) dataset and we discarded mutations in the somatic hypermutation (SHM) off-targets extracted from ref. [91] which might derive from tumor-infiltrated lymphocytes.

Determining IMD distributions of mutation tracts by simulation.

The IMD distribution of a clustered mutational process will be dependent on the length of the vulnerable DNA segment (for A3, the length of the ssDNA). To determine the expected IMD distribution we randomly sampled with replacement 1,000 times from a set of possible positions and computed the distance between random pairs. We used three sets representing three lengths of ssDNA fragments: short (25 bp), mid-length (200 bp) meant to represent the approximate length of ssDNA between Okazaki fragments in eukaryotes [67] and a long ssDNA (800 bp) meant to represent the ssDNA segments generated during the MMR process [66]. We note that, in order to draw conclusions about ssDNA tract lengths underlying *kataegis*, the cluster span (distance from the first to the last mutation) would be a more appropriate measure. However in case of *omikli*, which consists predominantly of two-mutation clusters, the IMD measure can for practical purposes be considered equivalent to the cluster span measure. For this analysis we considered samples in the APOBEC-prone cancer types in our TCGA dataset: bladder, breast, lung (LUAD and LUSC), cervical, head-and-neck and mismatch repair proficient uterus cancers.

Gamma mixture modelling of IMD distributions.

It is expected the distance between 2 mutations occuring in a single hypermutation event will follow a gamma distribution. Thus, to quantify different mechanisms generating clustered mutations we modelled the observed IMD distributions as a gamma mixture.

We selected only the TCW>K mutations with IMD lower than 1kb. We also required TCW coordination, meaning that at least 70% of the mutations in that clustered event must have occurred at TCW sites.

We used the R package *mixtools* (*gammamixEM*) that implements an Expectation Maximization (EM) based algorithm for the detection of different components. We obtained estimates for mixtures that ranged from 1 up to 8 components. As initial parameters, we used alpha = 0.2, 100 maximum iterations and an epsilon (convergence difference) of 0.01. We re-simulated the original IMD distributions (see above) for 10,000 iterations and re-computed the parameters. Based on the log-likelihood and the matching shape parameters of the distributions we extracted a total of three components, because the log-likelihood value suggests a strong increase from 1 to 2, and from 2 to 3 components, while the increase from 3 to 4 is more modest; we cannot however rule out a four-component model based on these data. Next, we computed the density of the components using the extracted parameters and the proportions of each component.

Same as the IMD distribution analysis we used samples in the APOBEC prone cancer types, bladder, breast, lung (LUAD and LUSC), cervical, head and neck and mismatch repair proficient uterus cancers.

Contribution of A3 clustered mutagenic process to the unclustered mutation burden.

In order to estimate how much the clustered processes contributed to the unclustered burden, which is the main contributor to the overall tumor mutation burden (TMB), we adapted a method that we recently introduced [4]. In brief, we used a robust linear regression (*rlm* function in the R MASS package) to predict the overall unclustered burden in the TCW>K context (dependent variable) from the counts of each clustered process (TCW>K *kataegis* and *omikli* burden, as separate independent variables (predictors), and additionally an interaction term.

From the fitted model, the intercept is the number of unclustered mutation that cannot be explained by the presence of either *omikli* or *kataegis* clusters, thus, these mutations likely occur independently from the mechanisms that generate either *omikli* or *kataegis*. We named this mutational process A3-X. Similarly, we obtained estimates of the average unclustered mutation burden when one of the two types of clusters (either *omikli* or *kataegis*) is not present but the other type is. These estimates represent the contribution of the *omikli* (A3-O) and *kataegis* (A3-K) processes to the unclustered A3 mutation burden. By adjusting for the total predicted unclustered mutations we can obtain estimates of the contribution of *kataegis* and *omikli* to unclustered burden. Note that because the A3 trinucleotide context (here defined as T$\underline{C}$W>K) overlaps with signatures of certain other mutagens, presence of these non-A3-derived unclustered mutations may inflate the estimate of the intercept in the fits (Fig. 4a), causing a downward bias in the estimated *omikli* contribution to global A3 burden (A3-O). For further details, see Supplementary Note.

Parsimony suggests that unclustered (singleton) mutations are generated by the clustered processes of the same mutational context (T$\underline{C}$W>K). However, we cannot rule out the possibility that the two processes (*omikli* and unclustered) are mechanistically distinct but tightly co-regulated thus co-occuring in the same tumor samples.

We extracted the 95% prediction intervals of the unclustered values (representing the number of mutations at the average value of each variable) by the R function *predict*. We then used the upper and lower ends of the interval to compute upper and lower bounds of the contribution in percentage. Error bars (Fig. 4 a-c) represent the SEM extracted from this interval.

Functional impact density of mutational processes.

We define the functional impact density (FID) as the putative functionally relevant mutations that occur in a certain set of genes which are associated with a selected mutational process. For a set of genes *G* and a mutational process S, the FID is computed as the number of mutations falling in the coding sequences (CDS) of *G* divided by the total number of mutations from *S*. For sake of clarity, this value can be represented as the number of mutations that fall in a gene coding sequence per thousand mutations.

This measure reports the joint effect of the mutational spectrum, the trinucleotide composition of the gene coding sequence (CDS) and, importantly for the A3 example, the regional preferences of the mutational process. For instance, if the trinucleotide composition of *G* matches with the trinucleotide propensity of *S* it will increase the FID. Also, if *S* is enriched in certain parts of the genome where *G* is also enriched, it will also yield a higher FID.

We selected three disease associated gene sets from the literature, (i) a set of 299 cancer genes, including tumor suppressor genes and oncogenes, which were recurrently mutated in TCGA cancer genomes [72],(ii) a set of genes associated with neurodegenerative disease (n = 39) [92], and finally (iii) a set of cell essential genes extracted from CRISPR/Cas9 genetic screens (n = 683) [93].

In order to obtain mutations that are putatively generated by a given mutational process, we selected those mutations matching the susceptible trinucleotides in a set of tumor samples where the mutational process was reported to occur. In total, we defined four mutational processes: (i) the aging associated process, (ii) “smoking”, (iii) “UV” and (iv) Signature 17. For the ageing process the trinucleotide set was NCG>T and the sample set was comprised by all samples (*n*= 697). For the “smoking” process the trinucleotide subset was NCN>A and the sample set was comprised by lung (LUAD and LUSC) tumor patients with at least three years of tobacco smoking [94] (self-reported data; *sub*21). For the “UV” process the trinucleotide subset was TCC>T (thus minimizing overlap with other mutational processes) and the sample sets were the skin cancer patients from the TCGA (n = 13) and a set of melanomas PCAWG dataset (MELA-AU, *n* =70) that were included to increase the number of mutations. For the Signature 17 process the trinucleotide subset was defined as AAN>C and the sample set was the stomach cancers available in our TCGA-WGS data (*n* =20).

Note that estimates from this analysis are likely conservative because we use a stringent A3 trinucleotide context of TCW>K, and moreover because we examined only unclustered A3 mutations but did not explicitly consider the A3 clustered *omikli* and *kataegis* events in this analysis, on the basis of their lower abundance (Fig. 1f) relative to the unclustered A3 mutations.

Logistic regression approach to determine susceptibility in cancer genes.

We used a logistic regression to determine if the occurrence of a mutation in a cancer gene was associated with a higher burden of either *omikli* or *kataegis*. We examined the set of 299 cancer genes [72] and selected mutations in their coding sequence (CDS) matching the A3 context TCW>K (W is A or T; K is T or G). If a gene contained at least one of these mutations in the CDS it was classified as mutated by an A3 process. We tested only the61 cancer genes (Supplementary Table 1) that bore A3 context mutations in at least 3 samples from the TCGA-WGS dataset. As negative control we also counted mutations in the cancer genes at the non-A3 context VCN>K (V is not T).

Next, we performed a multiple logistic regression using the square-rooted burdens of *omikli* and *kataegis* as independent variables to predict the mutation status of the gene (dependent variable). The independent variables were always restricted to the A3 (TCW>K) context to represent the A3 activity of either *omikli* or *kataegis*. The mutation status was tested both with genes harboring A3 mutations and the control context (VCN>K). The p-values for each gene were FDR adjusted using the Benjamini-Hochberg correction.

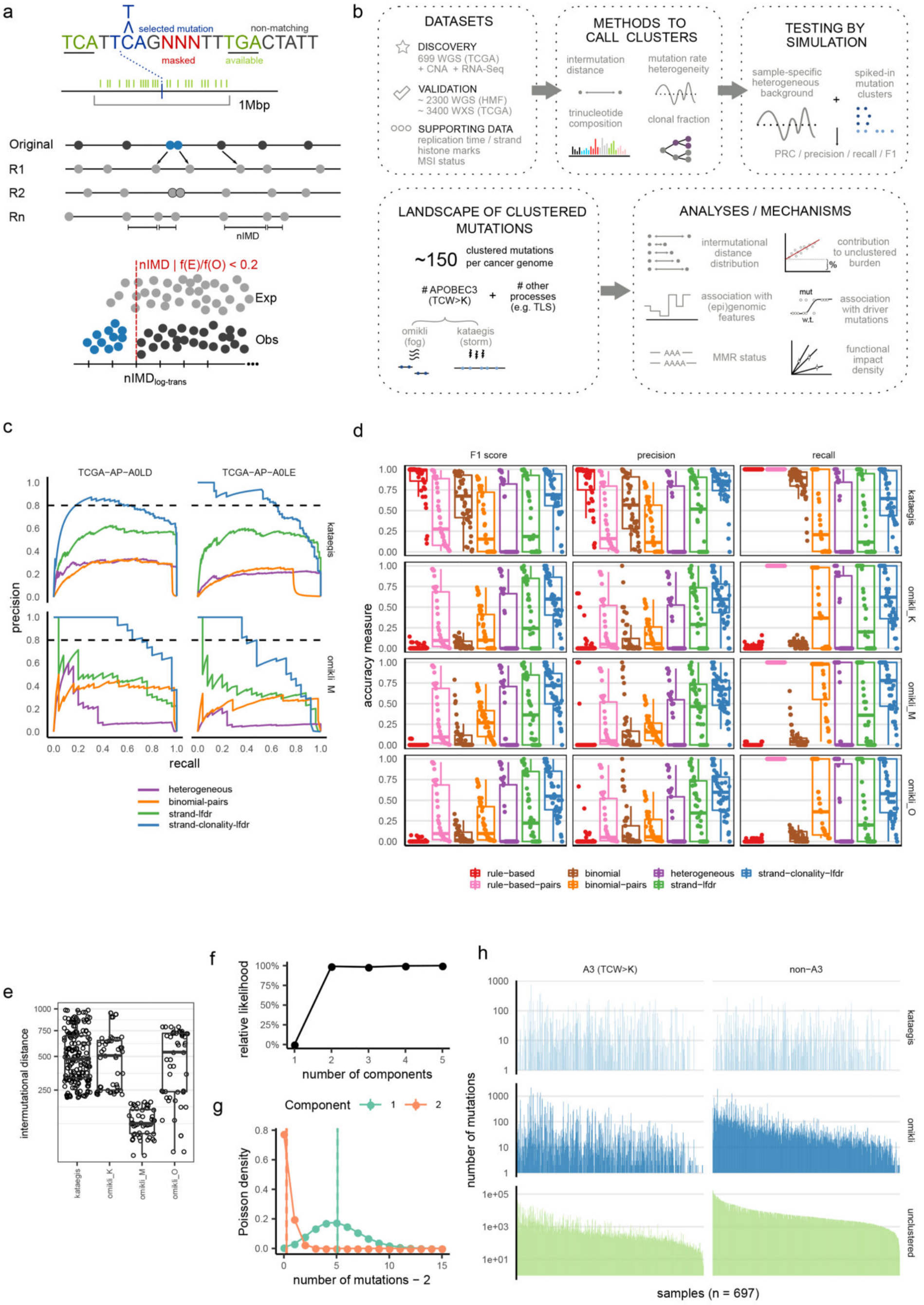
We also divided the CDS fragments from the cancer genes according to their replication time and then used logistic regression to predict if any of the CDS located in that specific replication time bin was mutated. We used the number of *omikli* mutations (square-rooted) as predictor.

Statistics

If not stated otherwise, the comparison of two distributions of continuous values was tested with a two-tailed Mann-Whitney U test. Pooling p-values obtained from stratified data groups was performed with the Fisher's method for combining P-values. P values are shown as exact values or otherwise referenced as symbol according to this scale: *** < 0.001, ** < 0.01, * < 0.05, "." < 0.1.

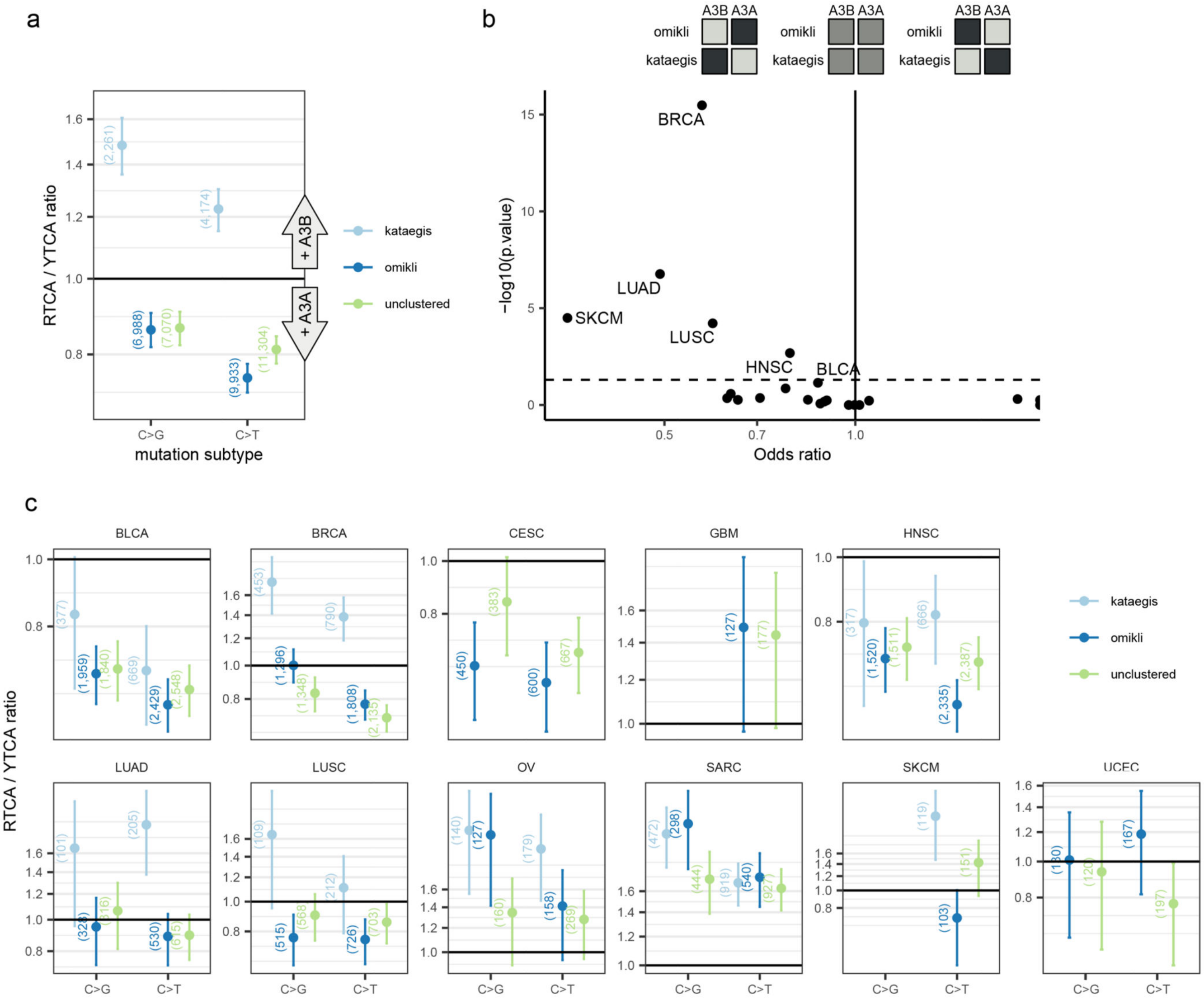
All boxplots used in the current analysis are represented according to the standard boxplot notation in the R statistical environment (*ggplot2* package): the central box represents the inter quartile range (IQR), the central line is the median value of the distribution, the outlier points are instances higher or lower than 1.5 times the IQR from the median value and the whiskers are the lowest and highest points of the distribution after removing the outliers. If the boxplot has notches, the notch width is 1.58 times the IQR divided by the square root of the sample size, which is an estimate of the 95% C.I. of the median.

Extended Data



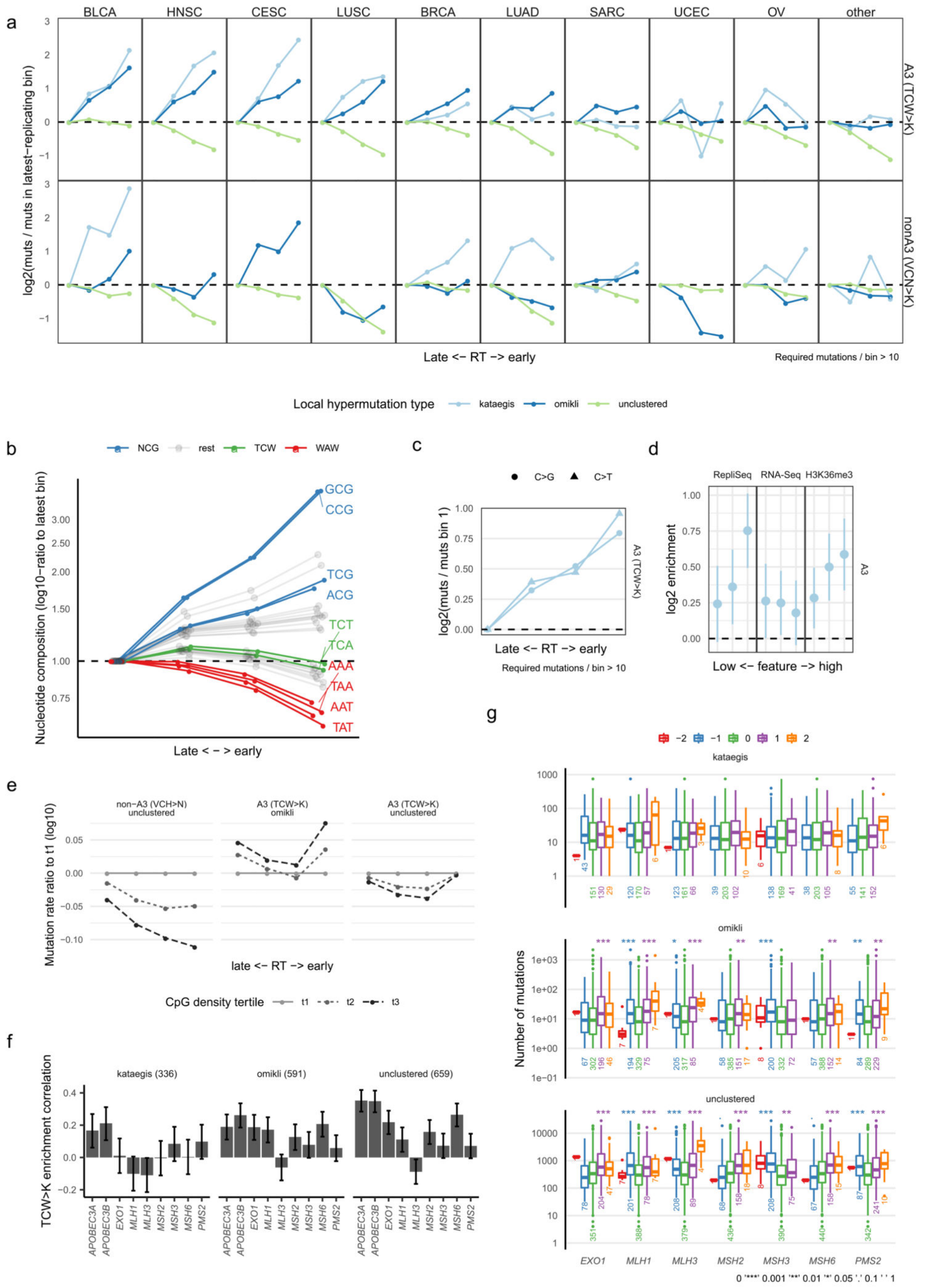
Extended Data Fig. 1. Detecting clustered mutations and simulating processes that generate clustered mutations.

a, Method to determine significant mutation clustering in HyperClust. A baseline distribution is generated by shuffling mutations within 1 Mbp windows multiple times (R1, R2, …, Rn) to matching trinucleotide context. For every mutation, the observed intermutational distance to its nearest neighbour (nIMD) is compared with distributions of expected IMDs (from randomized data) to determine a local FDR (lfdr). Thresholding by lfdr yields clustered mutation calls (blue). **b**, Overview of study. **c**, Precision-recall curves for models in Fig. 1a, derived from simulated data with spiked-in mutation clusters: *kataegis* (top; with five mutations per cluster at an average 600 bp pairwise distance) or *omikli*_M (bottom; two mutations at 101 bp). Two examples of high mutation burden tumors (TCGA-AP-A0LD, TCGA-AP-A0LE) were used here to generate the background mutation distributions. **d-e**, Testing accuracy of mutation cluster calling methods using simulated data. Points represent randomized tumor samples into which spiked-in mutation clusters were introduced. Samples are ordered according to total mutation burden (d). Columns show different performance metrics: F1 score, precision, and recall, all at lfdr=20%. Rows represent different types of spiked-in mutation clusters (IMD distributions plotted in panel e, where *kataegis* have five mutations and *omikli*_K/M/O two mutations. Boxplots compare cluster calling methods, including implementations of some previous methodologies (details in Methods). The strand-clonality-lfdr (blue) is the HyperClust method used throughout our work. **f-g**, Poisson mixture modelling (related with Fig. 1d) of the number of mutations per cluster, showing relative likelihood (panel **f**) of models with increasing number of components and the density functions (panel **g**) of a model with two Poisson components. solid line represents mean and dashed lines the 95% C.I.. **h**, Number of mutation events per tumor sample (X-axis, n) per local hypermutation type (rows), either the A3 context T<u>C</u>W>K, or the remaining mutations (columns).



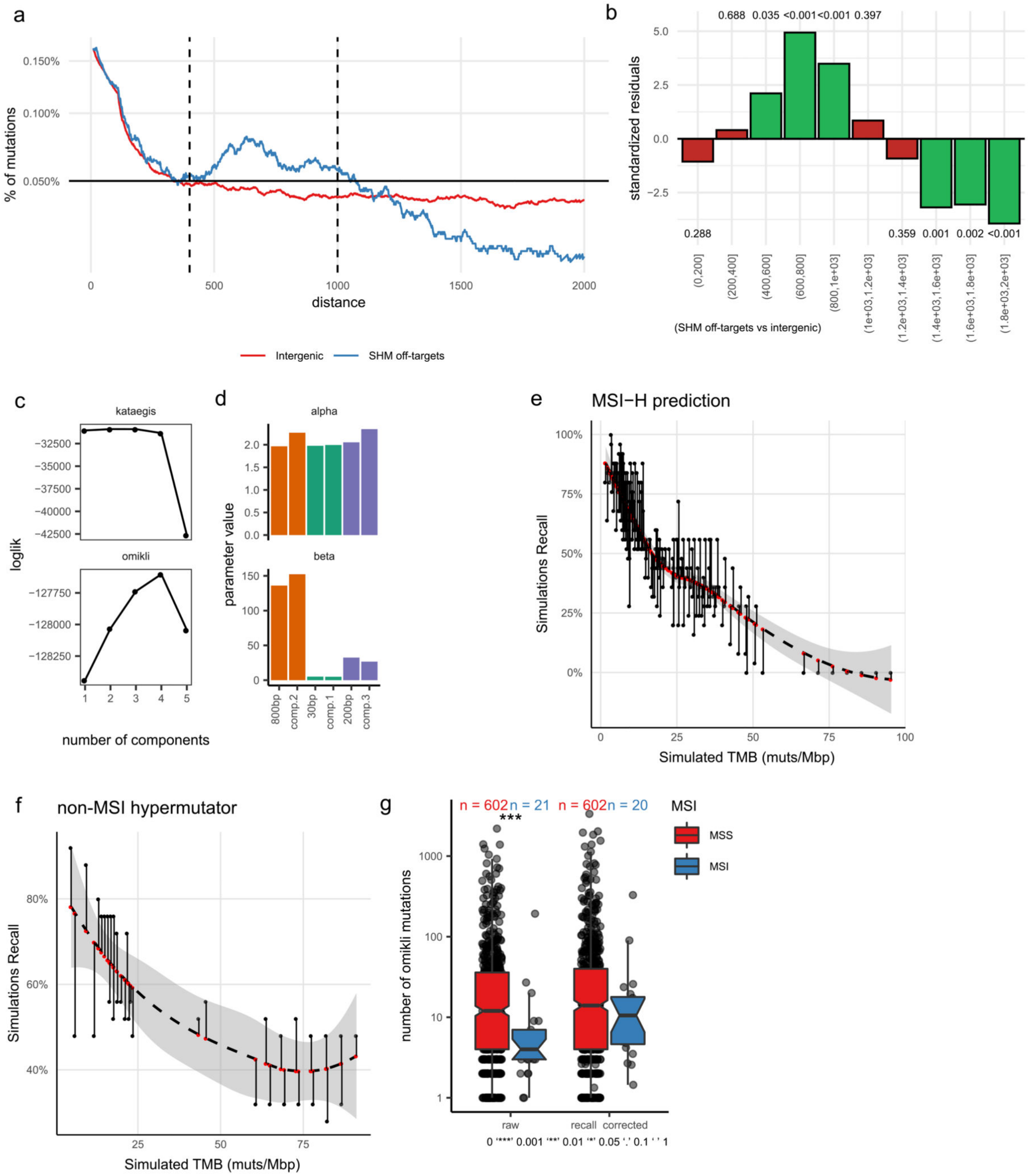
Extended Data Fig. 2. Tetranucleotide context suggests a role for the A3A enzyme in generating *omikli* and for A3B in *kataegis* mutations.

a, c, Ratios of the YTCA (A3A-like) and RTCA (A3B-like) mutation frequencies suggest differential mutagenic activity of A3A versus A3B enzymes in cancer samples. The C>T and the C>G changes in the two A3 contexts are shown in a pan-cancer analysis (panel a) and broken down by cancer type (panel c). At least 100 TCW mutations of a certain type across all tumor samples were required to perform analyses on that tissue (number of mutations in brackets). Error bars are the bootstrap 95% C.I. of the ratio. KICH and THCA cancer types are not shown due to low overall number of A3-context mutations. b, Across multiple cancer types, *omikli* shows a tendency towards A3A-like, lower RTCA/YTCA-ratios than does *kataegis*. Difference tested by Fisher's exact test (per tumor type), two-tailed; p-values were adjusted for multiple testing. Dashed line is FDR = 20%. Lower odds ratios (<1) denote relative enrichment of YTCA (A3A-like) mutations in *omikli* compared to *kataegis*; see illustration above plot.



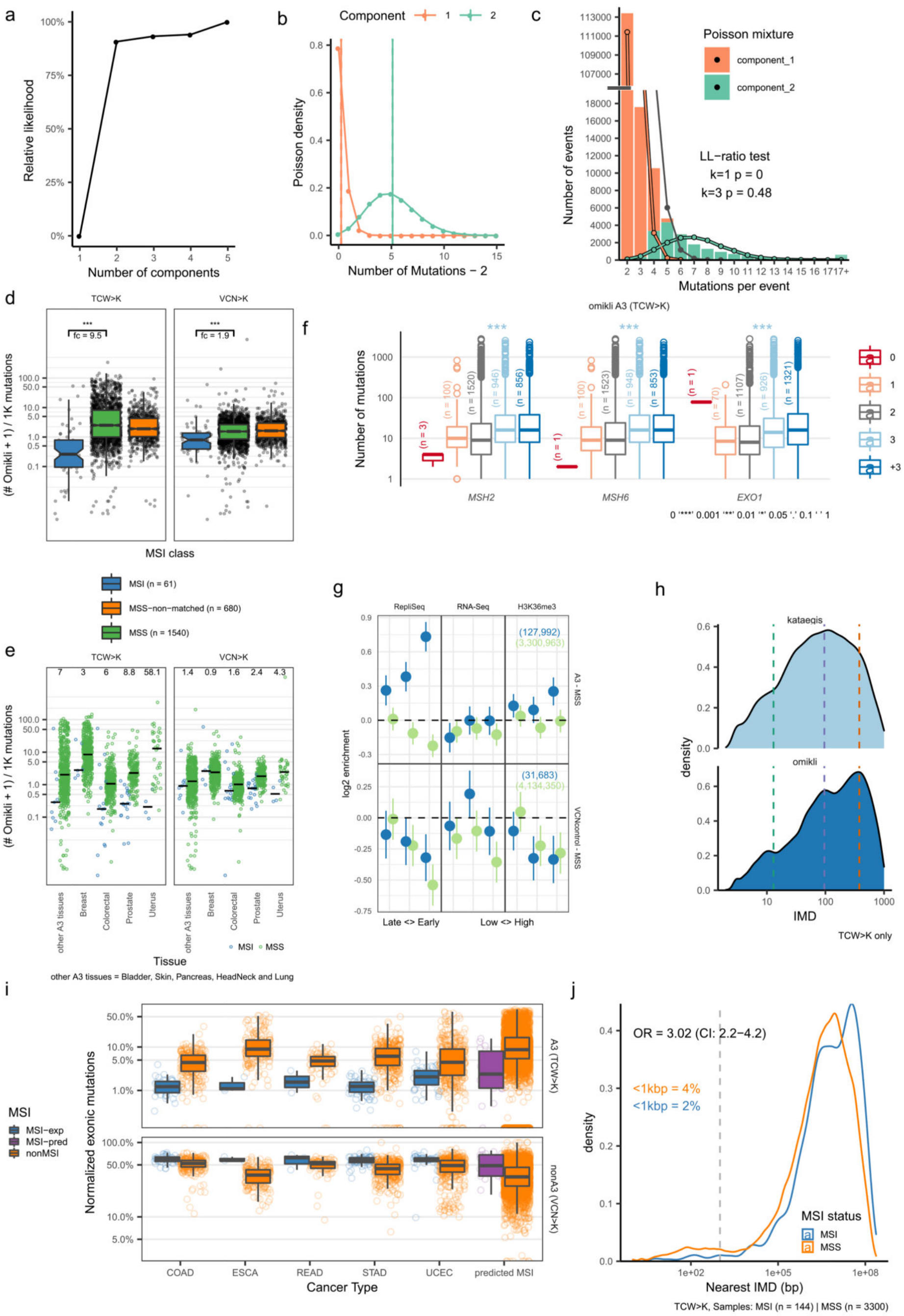
Extended Data Fig. 3. Association of clustered mutation rates with replication time (RT).

a, RT association per cancer type. Number of mutations per replication time bin in each context: A3 (top row) and the non-A3 control context at C:G nucleotide pairs (bottom row). RT bins are ordered from the latest-replicating quartile to the earliest-replicating quartile; mutation rates are shown relative to the latest bin. Enrichments not shown when the mutation count was lower than 10. **b**, Trinucleotide composition of the human reference genome in four replication time bins, normalized to the latest quartile (leftmost point). The A3 trinucleotide contexts ($T\underline{C}W$, green) are similarly abundant in the late and in the early-replicating regions of the genome. **c-d**, Enrichment of A3-context *kataegis* clusters, considering only RT (**b**), or jointly considering RT, mRNA levels and the H3K36me3 histone mark levels; points are coefficients from negative binomial regression, and error bars are 95% C.I. **e**, Mutation rates in genomic bins with different CpG density (determined per 10 kb segment), stratified by RT quartiles. Y-axis shows mutation densities relative to the first bin (“t1”, lowest tertile by CpG content). **f**, Spearman correlation between mRNA expression of A3A, A3B and MMR genes, and the TCW context enrichment of clustered mutations in a tumor. Error bars are 95% C.I. from the Fisher transformation of the correlation coefficient. **g**, Association of A3 mutation burden (clustered and unclustered) with copy number alterations of MMR genes. Significance by a two-tailed Mann-Whitney test, comparing tumor samples with neutral (0) versus gain/amplification (+1 and +2) states (blue stars show p-values according to legend), and independently, comparing samples with neutral (0) versus loss (−1 and −2) states (purple stars). P-values were not adjusted.



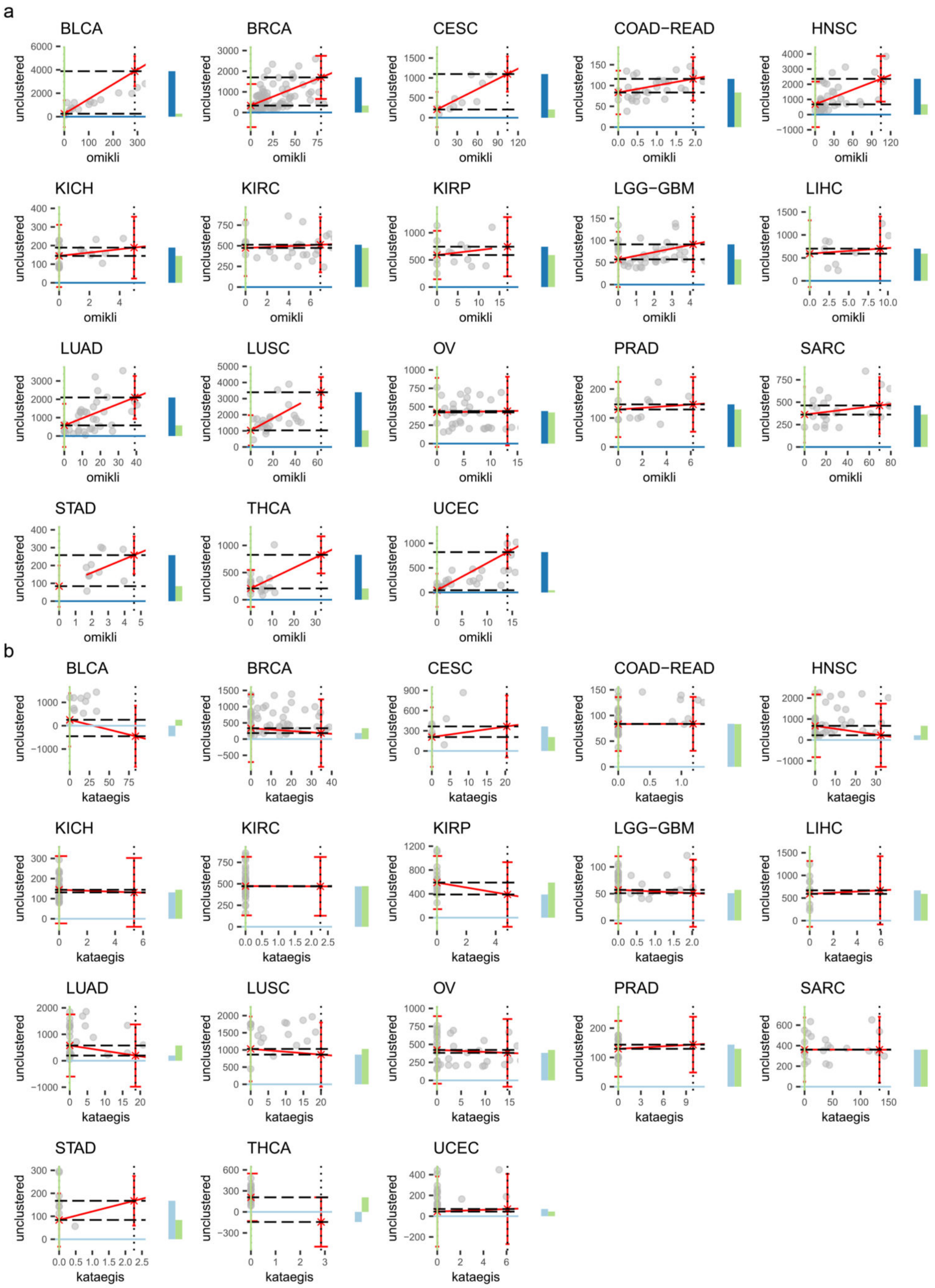
Extended Data Fig. 4. Simulations estimate power to detect mutation clusters and deconvolute their IMD distributions.

a-b, An analysis of somatic hypermutation (SHM) events in lymphoid cancers suggests length of MMR excision tracts in human cells. The distance from the initiating AID mutation (here, WNCYN>N context) to the flanking mutation introduced by error-prone MMR (here, any mutation at a A:T pair) is plotted, in known SHM off-target regions (blue) and, as a control, in intergenic regions (red) (panel **a**). A statistically significant enrichment is seen in the bins of the distance to central AID mutation (X-axis) between 400-1000 nt (panel **b**). Numbers above/below bars are p-values by Chi-square test on the standardized residuals. **c**, Gamma mixture modelling of the IMD distributions. Log-likelihood values for different number of components when modelling IMD of the A3 *kataegis* and *omikli* mutations. **d**, The alpha and beta parameters of the three fitted Gamma distributions (“comp.1”, “comp.2” and “comp. 3”) approximately match the alpha and beta parameters expected from simulated distributions with IMD at 30 bp, 800 bp and 200 bp, respectively. **e**-**f**, Simulations using spiked-in clustered mutations into genomes obtained by randomizing and subsampling mutations from MSI-H hypermutated tumors (panel e) and other hypermutators (panel **f**), with the goal of determining the recall (sensitivity; Y-axis) of recovering mutation clusters at various global mutation burdens (X-axis). Dashed line is a loess fit and shaded area is its 95% C.I. Vertical lines are residuals of the fit. **g**, Difference between MSI and MSS tumor samples in the absolute burden of clustered A3 *omikli* mutations; significance by Mann-Whitney (two-tailed).



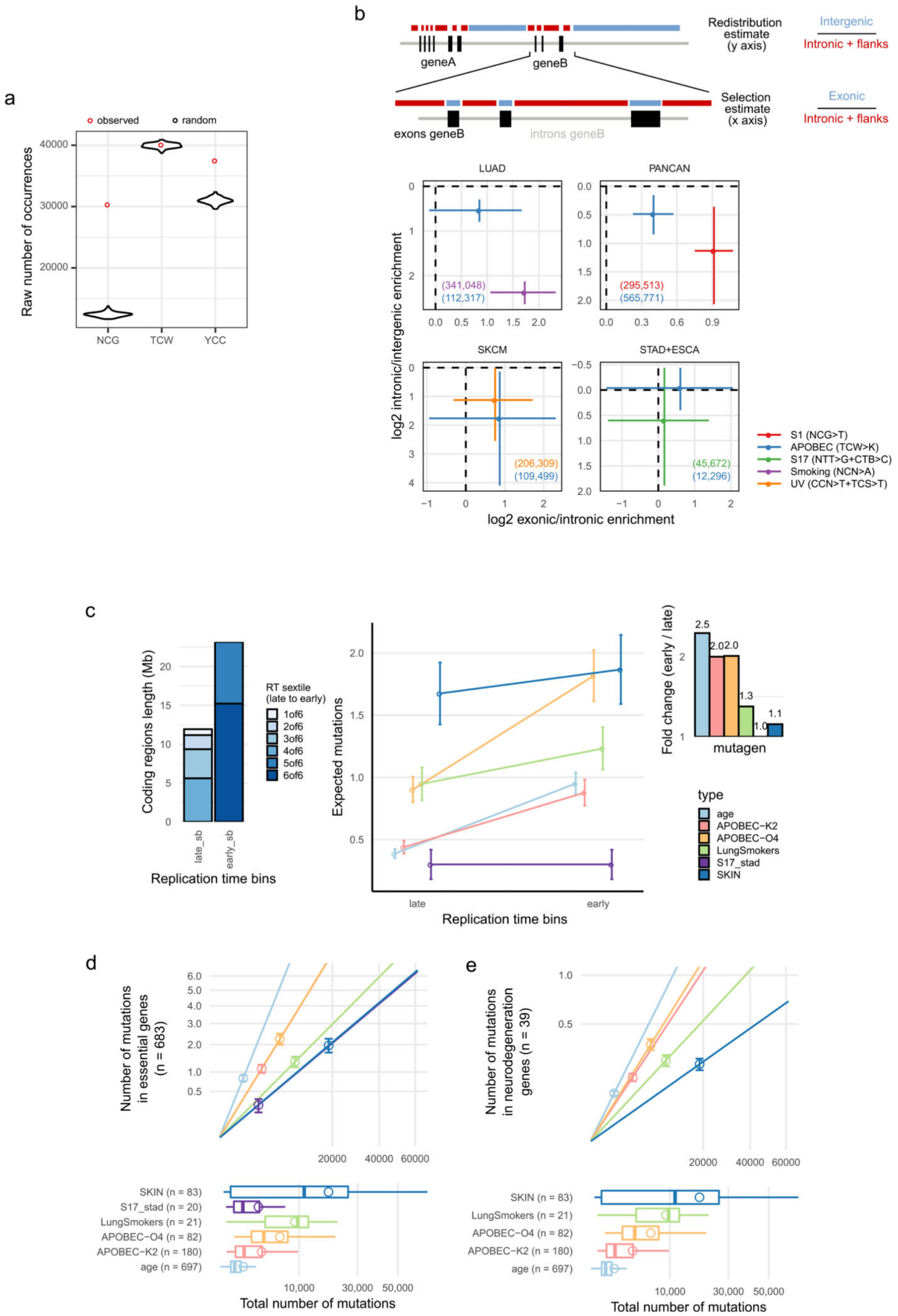
Extended Data Fig. 5. Validation analyses using independent genomic data sets.

a-c, Fitting a Poisson distribution mixture to the number of mutations per cluster in the Hartwig Medical Foundation (HMF) dataset. The near-maximum log likelihood (LL) is obtained with two components (panel **c**) and the increase to three components is not statistically supported; p-values are from a two-sided bootstrap test. **d-e**, The relative density of A3 context (left) clustered mutations is higher in MSS (MMR-proficient) than in MSI (MMR-deficient) samples of the same tumor type (left column) in the HMF data. The difference is smaller for the non-A3, control context (right). Significance by Mann-Whitney (two-tailed), n is the number of samples, *** is $p < 0.001$. Numbers show fold-difference between MSS and MSI samples. The “other A3 tissues” are lung, head-and-neck, skin, pancreas and bladder cancer. **f**, In HMF data, the A3-context *omikli* clustered mutations are enriched in tumors with amplified MMR genes; significance by Mann-Whitney test (two-tailed) comparing the neutral (0) versus the gain states (+1 and +2, considered jointly); n is the number of samples. **g**, In HMF data, A3-context *omikli* are enriched in early replicating, H3K36me3-marked genomic regions; error bars are 95% C.I. **h**, Intermutational distance distributions for *kataegis* (top) and *omikli* (bottom) A3 context mutations in the HMF data. Dashed lines show peaks of the simulated distributions (Fig. 2) with segment lengths of 25bp (green), 200bp (purple) and 800bp (orange). **i-j**, Whole-exome sequences in the TCGA data show an excess of A3 context (T<u>C</u>W) mutation fraction in MSS compared to MSI cancers (panel **i**), and an excess of T<u>C</u>W mutations at distances <1000 bp, normalized to longer distances, in MSS over MSI samples (panel **j**). “MSI-exp” (152) denotes the experimentally established MSI-H statuswhile “MSI-pred” (18) is the MSI status predicted using machine learning(ref. [61]), “nonMSI” (5,661) is neither of these cases.



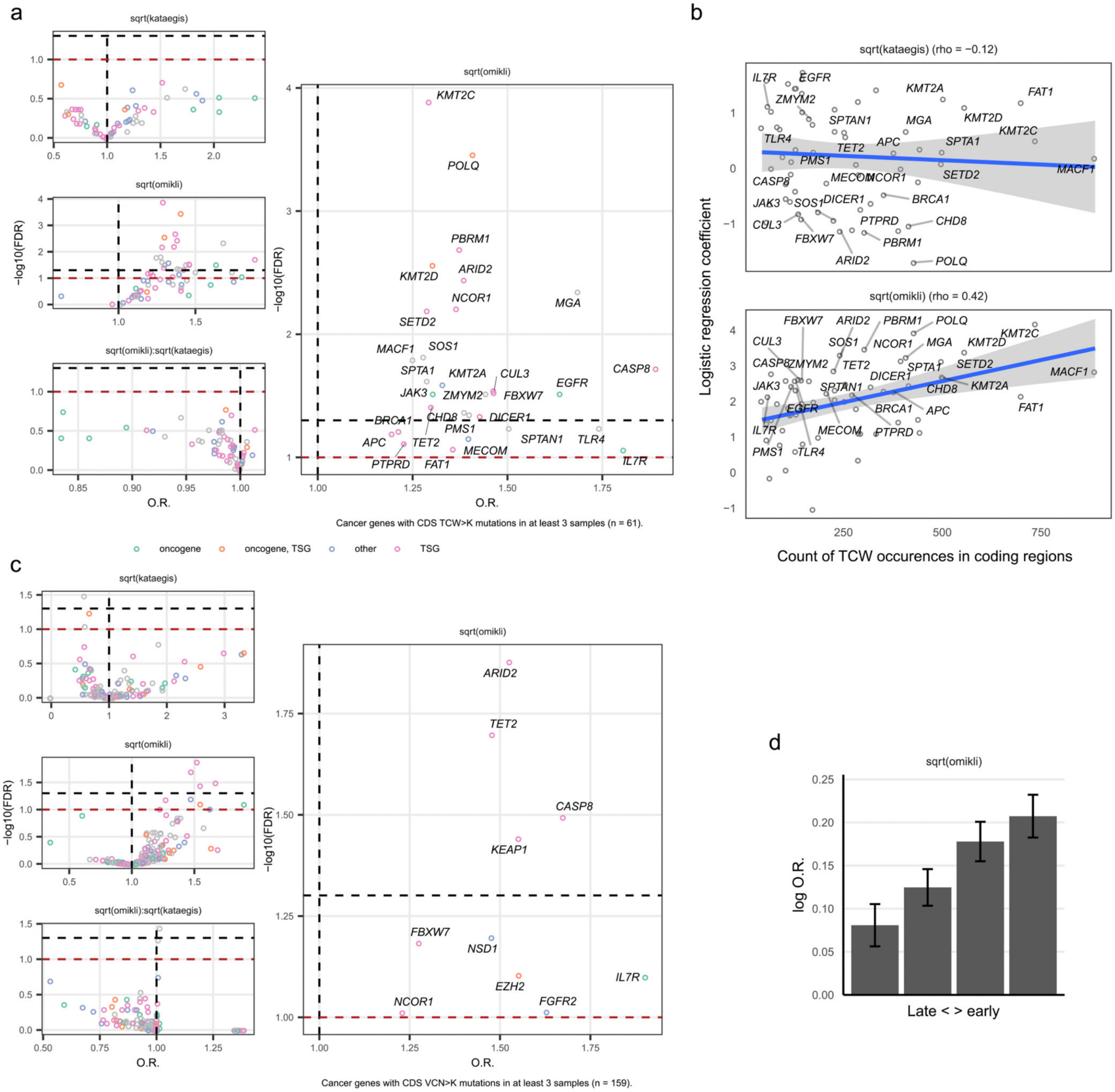
Extended Data Fig. 6. Contribution of the *omikli* and the *kataegis* mechanism to the unclustered A3 mutation burden in various tissues.

a, The *omikli* mechanism generates many unclustered mutations (“A3-O”) in various cancer types. **b**, The *kataegis* mechanism generates comparatively few unclustered mutations (“A3-K”). Panels show the fit (red line) of the unclustered A3 burden (Y-axis) to the clustered A3 burden (X-axis), (see Methods). Error bars are 95% prediction intervals at x = 0, and at x = mean burden of A3 clustered mutations for that cancer type. Horizontal dashed lines are the predicted numbers of unclustered A3 mutations at those two points (for clarity also shown in blue/green bars next to each plot). Fits use robust regression (rlm function in R). For visual clarity, only the part of the plot up to the mean of unclustered mutation burden plus a margin is shown, however the fit uses all data points (i.e. tumor samples) including ones not visualized.



Extended Data Fig. 7. Mechanisms underlying A3 clustered mutations generate many impactful changes, affecting disease genes.

a, Coding regions in the human genome are enriched for CpG dinucleotide (NCG), but not with the A3-context TCW trinucleotides, compared to random expectation. **b**, Enrichment of mutations in exons *versus* introns (estimate of selection strength, X-axis) and the enrichment in intergenic regions versus introns (estimate of redistribution of mutations towards genic DNA, Y-axis; flipped). The comparison of mutagenic agents against APOBEC was performed for selected tissues, matching the relevant tissue for the particular mutagen (tumor samples listed in Supplementary Table 7). Error bars are 95% C.I. from negative binomial regression; numbers in parenthesis are the tally of mutations. **c**, The differential functional impact of the tested mutagens across replication time (RT) bins. Left: total length of CDS in the late and early RT bins, shaded by the RT sextiles that were merged to create the two bins (where 1 is the latest and 6 is the earliest). Middle: expected number of cancer gene CDS-affecting mutations in an average tumor sample (same sets of samples, genes and mutations as in Fig. 5a; Y-axis) for the late versus early RT bin (X-axis), for various mutagens (colors); error bars are S.E.M. Right: fold-difference between the functional impact at the late versus early bin, for various mutagen types. **d-e**, The functional impact density (FID) of various mutational processes in a set of cell-essential genes (panel **d**) and neurodegenerative disease-associated genes (panel **e**). Slope shows the fraction of impactful genetic changes i.e. those affecting the coding region of at least one gene in the set. Points show the expected number of impactful changes resulting from a mutational process, on average, in a tumor genome affected by the mutational process. Error bars are S.E.M. “APOBEC-O4” is A3 mutagenesis in *omikli*-rich tumors. “APOBEC-K2” is A3 mutagenesis in *kataegis*-rich tumors.



Extended Data Fig. 8. Associations between genic mutations and global burden of clustered mutations.

a, Associations between A3-context TCW>K mutations in coding regions of each cancer gene, and the global burden of A3 *kataegis* (top left) or *omikli* (middle left) and their interaction term (bottom left). Right panel is same as middle-left panel, but showing only the significant genes with labels. Volcano plots show logistic regression coefficients (transformed to odds ratio) on the X-axis and the log FDR on the Y-axis. Genes bearing coding mutations in at least three tumor samples were tested. **b**, Number of TCW sites in a gene coding sequence (CDS; X-axis) predicts the association of cancer gene mutations (Y-axis) with A3 *omikli* burden (bottom) but not with A3 *kataegis* burden (top). Error bands are 95% C.I. of the linear fit. **c**, Same association analysis as panel a but for the control, non-A3 context VCN>K mutations in the gene CDS. **d**, Early RT cancer genes are more affected by A3 mutagenesis. Cancer genes were stratified into RT quartiles (X-axis) and logistic regression coefficient (log odds ratio, Y-axis) linking A3 *omikli* burden with the presence of a mutation in the CDS of any cancer gene in that RT bin was determined. Error bars are 95% C.I. from logistic regression (on n = 593 tumor samples).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the members of the Genome Data Science group and Ben Lehner for comments and discussions. This work was funded by the ERC Starting Grant HYPER-INSIGHT (757700) and by the Spanish Ministry of Economy and Competitiveness (grant no. BFU2017-89833-P). F.S. was funded by the ICREA Research Professor programme. D.M.P. was funded by a Severo Ochoa FPI fellowship (MCIU/Fondo Social Europeo ref. BES-2017-079820). The authors acknowledge support of the Severo Ochoa Centres of Excellence programme to the IRB Barcelona.

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analysed and made available data for this research.

Data availability statement

For the current study we used publicly available data described in the Methods. In brief, we used a set of whole genome sequences from TCGA available through cgHub repository (superseded by the NCI Genomic Data Commons, https://gdc.cancer.gov/). SNP arrays for the same data set were downloaded from the GDC legacy portal (portal.gdc.cancer.gov/legacy-archive). We used two validation sets: (i) the whole genome tumor cohort from the Hartwig Medical Foundation available at hartwigmedicalfoundation.nl (DR-069) upon request and (ii) the whole exome TCGA cohort through the MC3 dataset available at https://gdc.cancer.gov/about-data/publications/mc3-2017. Data generated by the analyses in this study are available in the Supplementary Tables.

Code availability

Code to generate clustered mutation calls was implemented in Python (version 3.6) and R environments (version 3.6). Relevant packages are biopython (version 1.73) and numpy (version 1.15.4) for Python, and Biostrings (2.52.0), VariantAnnotation (1.30.1) and GenomicRanges (1.36.0) for R. Code is available at https://github.com/davidmasp/hyperclust.

Statistical analysis of the data was performed using custom scripts in R (version 3.6); relevant packages are mclust (version 5.4.4), mixtools (version 1.1.0), MASS (version 7.3-51.4) and flexmix (version 2.3-15).

References

1. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014; 24:1445–1454. [PubMed: 25079859]
2. Rogozin IB, et al. DNA polymerase η mutational signatures are found in a variety of different types of cancer. Cell Cycle. 2018; 17:1–31. [PubMed: 29108451]
3. Seplyarskiy VB, et al. Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. Nat Genet. 2019; 51:36. [PubMed: 30510240]
4. Supek F, Lehner B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell. 2017; 170:534–547.e23. [PubMed: 28753428]
5. Moris A, Murray S, Cardinaud S. AID and APOBECs span the gap between innate and adaptive immunity. Front Microbiol. 2014; 5:534. [PubMed: 25352838]
6. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–421. [PubMed: 23945592]
7. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013; 45:977–983. [PubMed: 23852168]
8. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149:979–993. [PubMed: 22608084]
9. Roberts SA, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45:970–976. [PubMed: 23852170]
10. Roberts SA, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell. 2012; 46:424–435. [PubMed: 22607975]
11. Landry S, Narvaiza I, Linfesty DC, Weitzman MD. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. EMBO Rep. 2011; 12:444–450. [PubMed: 21460793]
12. Suspène R, et al. Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proc Natl Acad Sci. 2011; 108:4858–4863. [PubMed: 21368204]
13. Byeon I-JL, et al. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. Nat Commun. 2013; 4
14. Holtz CM, Sadler HA, Mansky LM. APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. Nucleic Acids Res. 2013; 41:6139–6148. [PubMed: 23620282]
15. Nik-Zainal S, et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat Genet. 2014; 46:487–491. [PubMed: 24728294]
16. Glaser AP, et al. APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. Oncotarget. 2017; 9:4537–4548. [PubMed: 29435122]
17. Cortez LM, et al. APOBEC3A is a prominent cytidine deaminase in breast cancer. PLOS Genet. 2019; 15
18. Sakofsky CJ, et al. Break-induced replication is a source of mutation clusters underlying kataegis. Cell Rep. 2014; 7:1640–1648. [PubMed: 24882007]
19. Sakofsky CJ, et al. Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation. PLOS Biol. 2019; 17
20. Kazanov MD, et al. APOBEC-induced cancer mutations are uniquely enriched in early-replicating, gene-dense, and active chromatin regions. Cell Rep. 2015; 13:1103–1109. [PubMed: 26527001]
21. Buisson R, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science. 2019; 364
22. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015; 521:81–84. [PubMed: 25707793]
23. Zheng CL, et al. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. Cell Rep. 2014; 9:1228–1234. [PubMed: 25456125]

24. Haradhvala NJ, et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell. 2016; 164:538–549. [PubMed: 26806129]

25. Morganella S, et al. The topography of mutational processes in breast cancer genomes. Nat Commun. 2016; 7

26. Seplyarskiy VB, et al. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. Genome Res. 2016; 26:174–182. [PubMed: 26755635]

27. Green AM, et al. APOBEC3A damages the cellular genome during DNA replication. Cell Cycle. 2016; 15:998–1008. [PubMed: 26918916]

28. Kanu N, et al. DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. Genome Biol. 2016; 17:185. [PubMed: 27634334]

29. Nikkilä J, et al. Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. Br J Cancer. 2017; 117:113–123. [PubMed: 28535155]

30. Bhagwat AS, et al. Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. Proc Natl Acad Sci. 2016; 113:2176–2181. [PubMed: 26839411]

31. Hoopes JI, et al. APOBEC3A and APOBEC3B preferentially deaminate the lagging strand template during DNA replication. Cell Rep. 2016; 14:1273–1282. [PubMed: 26832400]

32. Chen J, Miller BF, Furano AV. Repair of naturally occurring mismatches can induce mutations in flanking DNA. eLife. 2014; 3

33. Cannataro VL, et al. APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. Oncogene. 2019; 38

34. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. Cell Rep. 2014; 7:1833–1841. [PubMed: 24910434]

35. Li Z, et al. APOBEC signature mutation generates an oncogenic enhancer that drives LMO1 expression in T-ALL. Leukemia. 2017; 31:2057–2064. [PubMed: 28260788]

36. de Bruin EC, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014; 346:251–256. [PubMed: 25301630]

37. McGranahan N, et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci Transl Med. 2015; 7:283ra54–283ra54.

38. Ullah I, et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. J Clin Invest. 2018; 128:1355–1370. [PubMed: 29480816]

39. Reijns MAM, et al. Lagging strand replication shapes the mutational landscape of the genome. Nature. 2015; 518:502–506. [PubMed: 25624100]

40. Taylor BJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife. 2013; 2

41. D’Antonio M, Tamayo P, Mesirov JP, Frazer KA. Kataegis expression signature in breast cancer is associated with late onset, better prognosis, and higher HER2 levels. Cell Rep. 2016; 16:672–683. [PubMed: 27373164]

42. Petljak M, et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell. 2019; 176:1282–1294.e20. [PubMed: 30849372]

43. Zhang Y, et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. Cell Rep. 2018; 24:515–527. [PubMed: 29996110]

44. Yang Y, Sterling J, Storici F, Resnick MA, Gordenin DA. Hypermutability of damaged single-strand dna formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. PLOS Genet. 2008; 4

45. Chan K, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015; 47

46. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol. 2011; 29:1103–1108. [PubMed: 22101487]

47. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018; 19:129. [PubMed: 30201020]

48. Woo YH, Li W-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun. 2012; 3

49. Zou X, et al. Validating the concept of mutational signatures with isogenic cell models. Nat Commun. 2018; 9:1–16. [PubMed: 29317637]

50. Li F, et al. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. Cell. 2013; 153:590–600. [PubMed: 23622243]

51. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]

52. Vavouri T, Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol. 2012; 13:1–12.

53. Huang Y, Gu L, Li G-M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. J Biol Chem. 2018; 293:7811–7823. [PubMed: 29610279]

54. Mugal CF, von Grünberg H-H, Peifer M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol Biol Evol. 2009; 26:131–142. [PubMed: 18974087]

55. Pfister SX, et al. SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability. Cell Rep. 2014; 7:2006–2018. [PubMed: 24931610]

56. Chen J, Furano AV. Breaking bad: The mutagenic effect of DNA repair. DNA Repair. 2015; 32:43–51. [PubMed: 26073774]

57. Andrianova MA, Bazykin GA, Nikolaev SI, Seplyarskiy VB. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. Genome Res. 2017; 27:1336–1343. [PubMed: 28512192]

58. Shinbrot E, et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 2014; 24:1740–1750. [PubMed: 25228659]

59. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol. 2006; 7:335–346. [PubMed: 16612326]

60. Tran PT, Erdeniz N, Symington LS, Liskay RM. EXO1-A multi-tasking eukaryotic nuclease. DNA Repair. 2004; 3:1549–1559. [PubMed: 15474417]

61. Cortes-Ciriano I, Lee S, Park W-Y, Kim T-M, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017; 8

62. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med. 2016; 22:1342–1350. [PubMed: 27694933]

63. Maruvka YE, et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat Biotechnol. 2017; 35:951–959. [PubMed: 28892075]

64. Hombauer H, Srivatsan A, Putnam CD, Kolodner RD. Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. Science. 2011; 334:1713–1716. [PubMed: 22194578]

65. Hombauer H, Campbell CS, Smith CE, Desai A, Kolodner RD. Visualization of eukaryotic DNA mismatch repair reveals distinct recognition and repair intermediates. Cell. 2011; 147:1040–1053. [PubMed: 22118461]

66. Jeon Y, et al. Dynamic control of strand excision during human DNA mismatch repair. Proc Natl Acad Sci. 2016; 113:3281–3286. [PubMed: 26951673]

67. Smith DJ, Whitehouse I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature. 2012; 483:434–438. [PubMed: 22419157]

68. Bowen N, et al. Reconstitution of long and short patch mismatch repair reactions using Saccharomyces cerevisiae proteins. Proc Natl Acad Sci U S A. 2013; 110:18472–18477. [PubMed: 24187148]

69. Brosey CA, et al. A new structural framework for integrating replication protein A into DNA processing machinery. Nucleic Acids Res. 2013; 41:2313–2327. [PubMed: 23303776]

70. Fan J, Pavletich NP. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. Genes Dev. 2012; 26:2337–2347. [PubMed: 23070815]

71. Supek F, Lehner B. Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair. 2019; 81

72. Bailey MH, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018; 173:371–385e.18. [PubMed: 29625053]

73. Pich O, et al. The mutational footprints of cancer therapies. Nat Genet. 2019; 51:1732–1740. [PubMed: 31740835]

74. Hodis E, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]

75. Drost J, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017; 358:234–238. [PubMed: 28912133]

76. Lodato MA, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science. 2018; 359:555–559. [PubMed: 29217584]

77. Verheijen BM, Vermulst M, van Leeuwen FW. Somatic mutations in neurons during aging and neurodegeneration. Acta Neuropathol (Berl). 2018; 135:811–826. [PubMed: 29705908]

78. Lei L, et al. APOBEC3 induces mutations during repair of CRISPR–Cas9-generated DNA breaks. Nat Struct Mol Biol. 2018; 25:45. [PubMed: 29323274]

79. Belfield EJ, et al. DNA mismatch repair preferentially protects genes from mutation. Genome Res. 2017; doi: 10.1101/gr.219303.116

80. Lujan SA, et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. Genome Res. 2014; 24:1751–1764. [PubMed: 25217194]

81. Peña-Diaz J, et al. Noncanonical mismatch repair as a source of genomic instability in human cells. Mol Cell. 2012; 47:669–680. [PubMed: 22864113]

82. Zlatanou A, et al. The hMSH2-hMSH6 complex acts in concert with monoubiquitinated PCNA and pol η in response to oxidative DNA damage in human cells. Mol Cell. 2011; 43:649–662. [PubMed: 21855803]

83. Saunders CT, et al. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics. 2012; 28:1811–1817. [PubMed: 22581179]

84. Priestley P, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019; 575:210–216. [PubMed: 31645765]

85. Huang MN, et al. Msiseq: software for assessing microsatellite instability from catalogs of somatic mutations. Sci Rep. 2015; 5:1–10.

86. Wang J, et al. Clonal evolution of glioblastoma under therapy. Nat Genet. 2016; 48:768–776. [PubMed: 27270107]

87. Hayward NK, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017; 545:175–180. [PubMed: 28467829]

88. Campbell PJ, et al. Pan-cancer analysis of whole genomes. Nature. 2020; 578:82–93. [PubMed: 32025007]

89. Ellrott K, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 2018; 6:271–281.e7. [PubMed: 29596782]

90. Grün B, Leisch F. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw. 2008; 28:1–35. [PubMed: 27774042]

91. Khodabakhshi AH, et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget. 2012; 3:1308–1319. [PubMed: 23131835]

92. Krüger S, et al. Rare variants in neurodegeneration associated genes revealed by targeted panel sequencing in a german ALS cohort. Front Mol Neurosci. 2016; 9

93. Hart T, et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3 Genes Genomes Genet. 2017; 7:2719–2727.

94. Liu J, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018; 173:400–416.e11. [PubMed: 29625055]

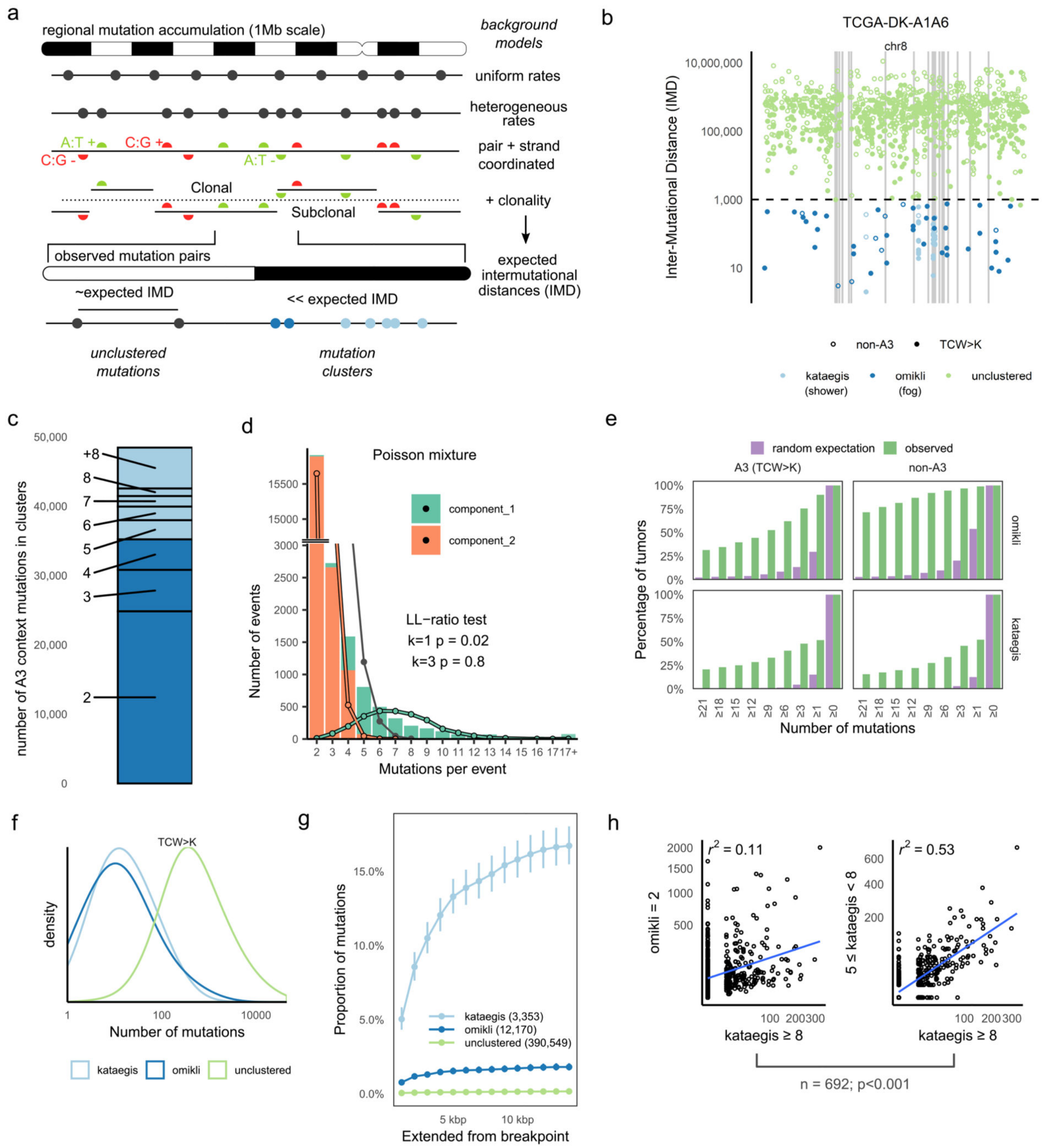


Figure 1. Two types of local hypermutation in human tumors.

a, The HyperClust framework detects mutation clustering by accounting for heterogeneous mutation rates at the megabase scale, further stratifying mutations by type, and additionally by their approximate timing (clonal fraction). **b**, *Kataegis*(thunderstorm) and *omikli* (fog) mutation clusters in an example tumor genome segment (chromosome 8 of TCGA-DK-A1A6). Vertical lines are rearrangement loci. **c**, Distribution of the number of A3-context TCW>K mutations in *omikli* (bottom) and *kataegis* (top) of different sizes (number of mutations per cluster; callouts). **d**, Poisson mixture modelling of number of A3 context mutations per cluster. Solution with two distributions is shown (*kataegis*, teal and *omikli*, orange). Stacked bars show component proportions and curves are density estimates. Grey curve is the baseline solution with one component; p-values are from a two-sided bootstrap test; LL, log likelihood. **e**, Cumulative percentage of tumor samples that contain at least the given number of clustered mutations, either observed, or expected at random. **f**, Distribution of the burden of A3 context somatic mutations per tumor, across tumors; samples with no *omikli* or no *kataegis* mutations were not considered. **g**, Cumulative fraction of A3 mutations within the neighborhood (width on X-axis) of a rearrangement breakpoint. Error bars are 95% binomial C.I.; number of mutations listed in parenthesis. **h**, Pearson correlation between the burden of two-mutation *omikli* and of long *kataegis* events (left) and the correlation between burden of *kataegis* of different lengths (right). Significant difference by a two-tailed t-test on the Fisher-transformed correlation coefficients.

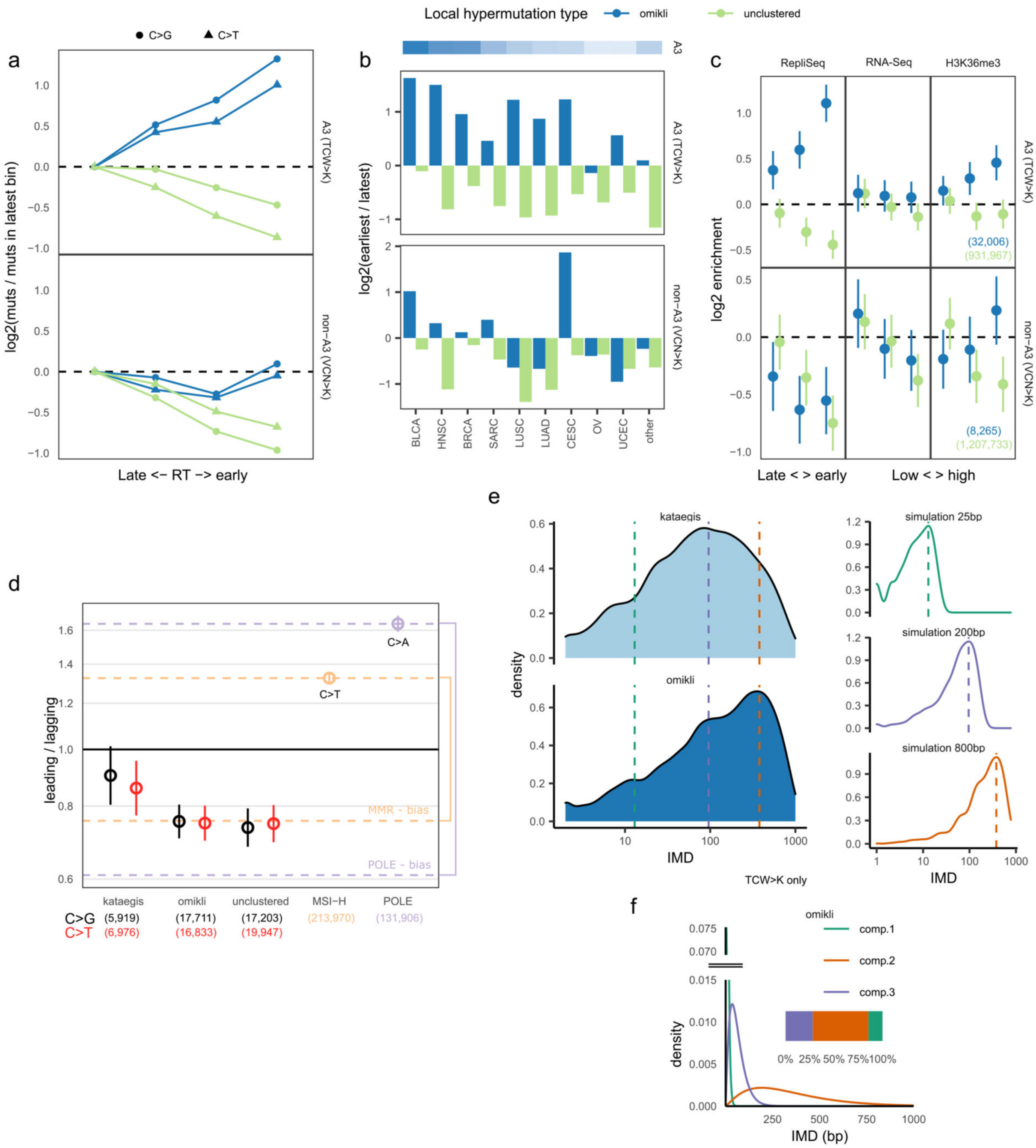


Figure 2. Association of A3 clustered mutation density with genomic features.

a, Mutation rates in replication time (RT) quartiles, relative to the latest RT, for A3 mutation contexts (top) and control contexts (bottom). **b**, Mutation enrichment in the earliest *versus* latest RT quartile for A3 context clusters (top) and non-A3 context clusters (bottom). Cancer types are ordered by total A3 burden across all tumors (shading in top bar). Moderate/low-A3 burden cancer types are pooled into the group “other”. **c**, Relative density of A3 and non-A3 mutation types across genomic regions. All enrichments are relative to the lowest bin (the latest-replicating quartile for RT), which is not shown on figure. Points are coefficients from negative binomial regression, and error bars are 95% C.I. **d**, Replication strand bias (ratio of mutation count on the leading *versus* lagging DNA strand) of clustered TCW mutations. Error bars are binomial 95% C.I. As a control, the reciprocal of the strand bias for MSI-H (orange; 24 samples) and POLE-mutant (purple; 9 samples) tumors is shown as a dashed line. Values in parentheses are mutation counts used to estimate the ratios. **e**, Distributions of intermutation distances (IMD) in A3 context *kataegis* and *omikli* clusters (left). Expected IMD distributions from simulations using three different segment lengths (right). **f**, Gamma mixture modeling of the *omikli* IMD distribution using three components. Bar shows proportions of the three components, while curves show their densities at various IMDs.

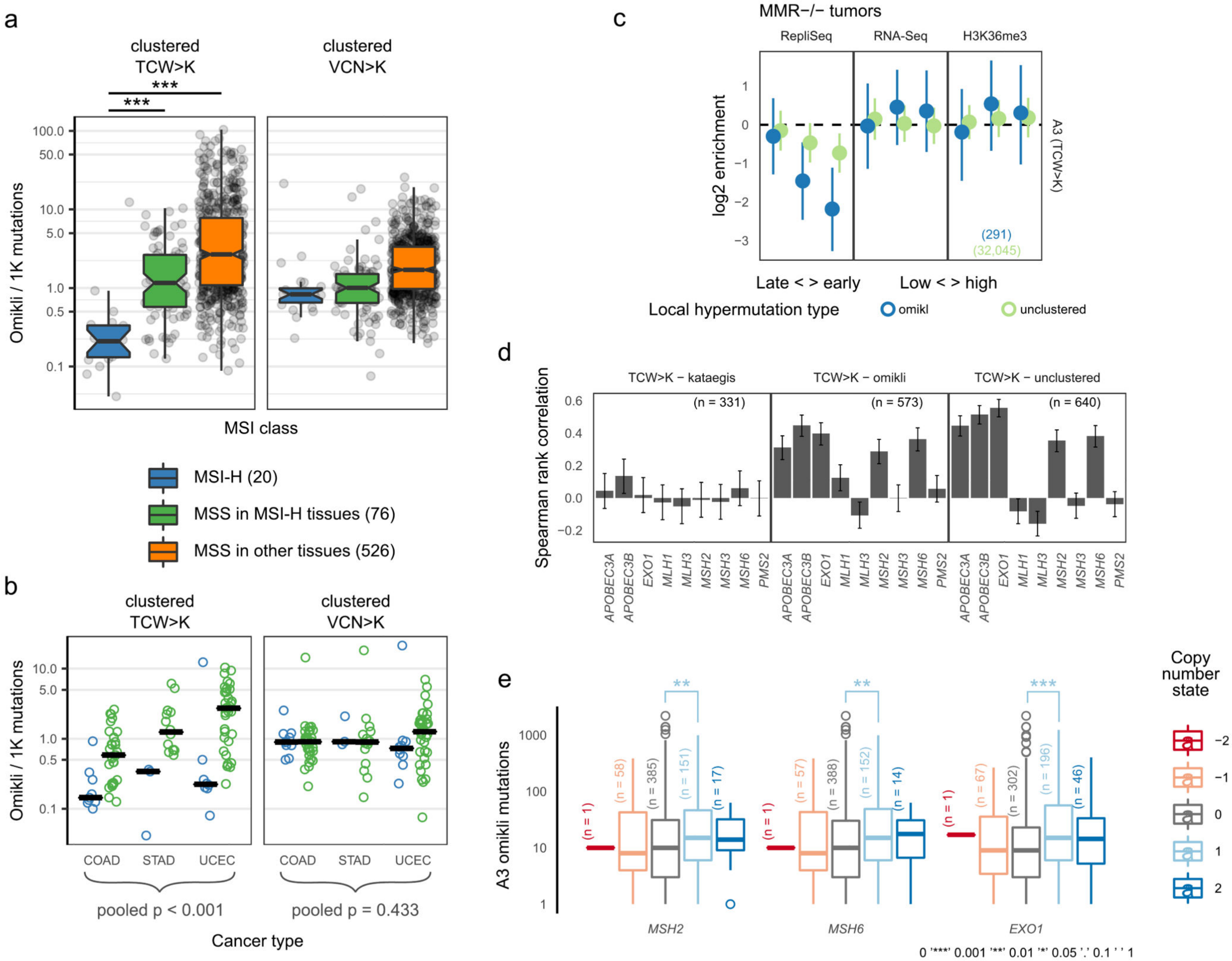


Figure 3. MMR activity in tumors is associated with APOBEC mutagenesis.

a, Proportion of *omikli* clusters in A3 (left) and control non-A3 contexts (right), comparing MMR deficient (MSI-H) samples with MMR-proficient (MSS) samples, in matched tissues (“MSI tissues”, COAD, STAD and UCEC, green) or in non-matched tissues (red). Significance by Mann-Whitney test, two-tailed; p < 0.001 (***); number of tumor samples listed in parenthesis. **b**, Same as (a) but broken down by tissue. UCEC, uterus; STAD, stomach; COAD, colon. Pooled p-value (p < 0.001 for A3; p = 0.433 for control) from two-tailed Mann Whitney tests on stratified data. **c**, Enrichment of A3 *omikli* clusters and unclustered A3 mutations in various genome regions in MMR-deficient samples (MSI-H). Related to Fig. 2c. Coefficients of negative binomial regression are shown (as $\log_2$), indicating enrichments of mutation frequency in a genomic bin *versus* the lowest bin (in case of RT, latest-replicating), where enrichment would equal unity and is thus not shown. Error bars are 95% C.I. **d**, Correlation of the burden of A3-context (T<u>C</u>W>K) *kataegis, omikli*, and unclustered mutations with mRNA levels of MMR genes and of *APOBEC3A* and *APOBEC3B* genes. Error bars are 95% C.I. **e**, Association of copy number alterations (CNA) in selected MMR genes with burden of A3 *omikli*. CNAs are represented as integer copy number differences (Methods); positive values are gains and negative losses. See also Extended Data Fig. 3g. Significance by Mann-Whitney test (two-tailed) comparing the neutral (0) *versus* the gain (+1 and +2) states considered jointly.

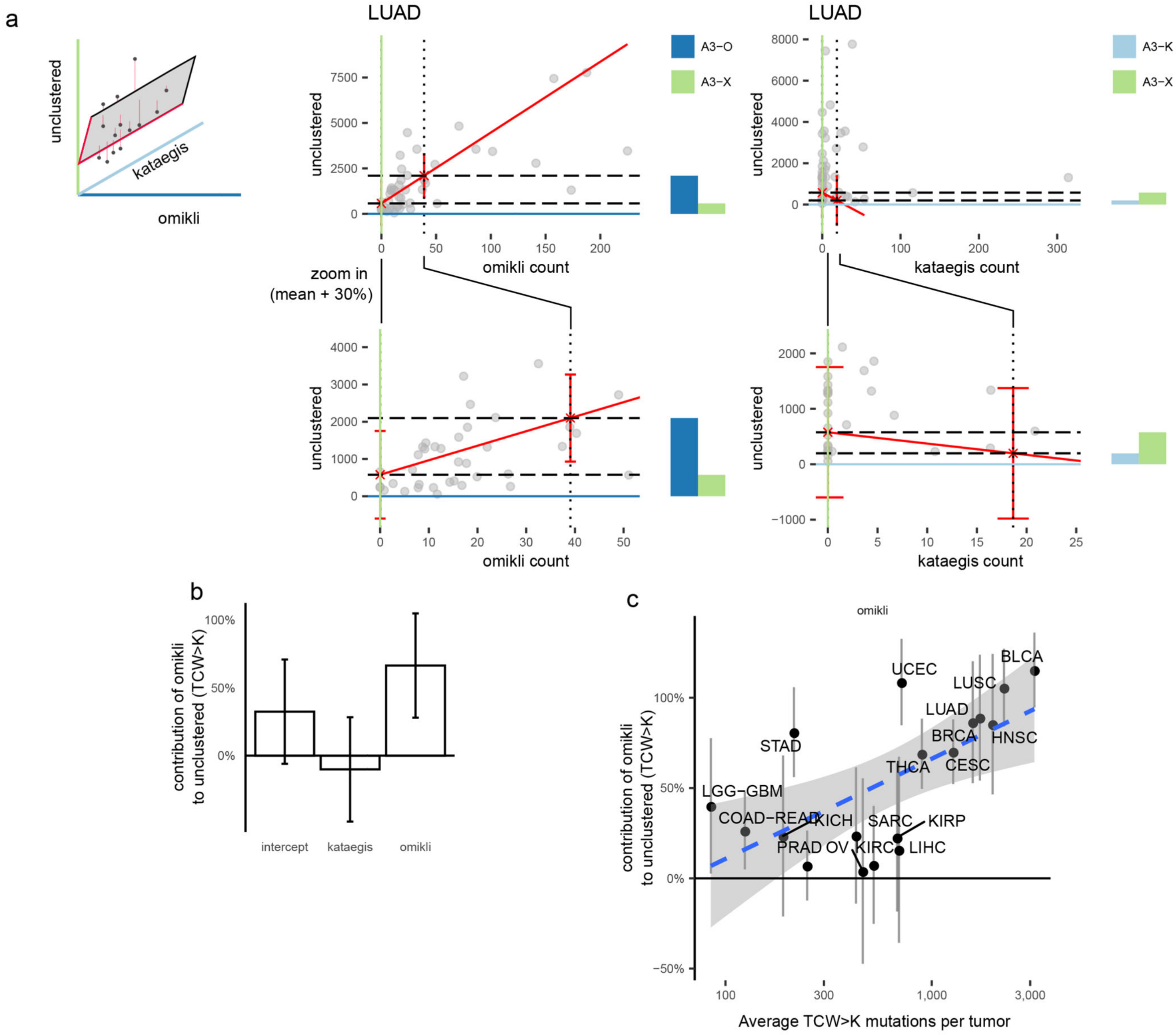


Figure 4. The *omikli* process generates the majority of unclustered A3 mutations across tissues.

a, A regression analysis estimates the contributions of *omikli* and *kataegis* processes towards the unclustered A3 mutation burden, shown for lung adenocarcinoma (LUAD, other cancers in Extended Data Fig. 6) tumor samples (points). For clarity, data panels show combinations of two variables (*omikli versus* unclustered, center; *kataegis versus* unclustered, right), whereas the regression is performed on the three variables simultaneously (schematic in leftmost panel; Methods). Red line is the intersection of the fitted plane with the shown two-dimensional coordinate system. Error bars are 95% prediction intervals of the fit. Dotted line is the average of *omikli* (center) and *kataegis* (right) mutation burden across tumors. Bottom panels have same data as top panels, but zoomed in on the X-axis for clarity. **b**, Pan-cancer regression analysis provides estimates of the fraction of unclustered TCW>K mutations contributed by processes that generate *omikli*(A3-O), that generate *kataegis* (A3-K) and a remainder ("intercept") not explained by either process (A3-X). Error bars are standard errors (S.E.) of regression coefficients; n = 646 tumors. **c**, Relative contribution of the *omikli*-process to the unclustered A3 burden (Y-axis) of cancer types correlates with the overall burden of A3 mutations in that cancer type (X-axis) suggesting that differential activity of the *omikli* mechanism drives differences of A3 burden between tissues. Error bars are S.E. of regression coefficients. Shaded band is 95% C.I. of the linear fit.

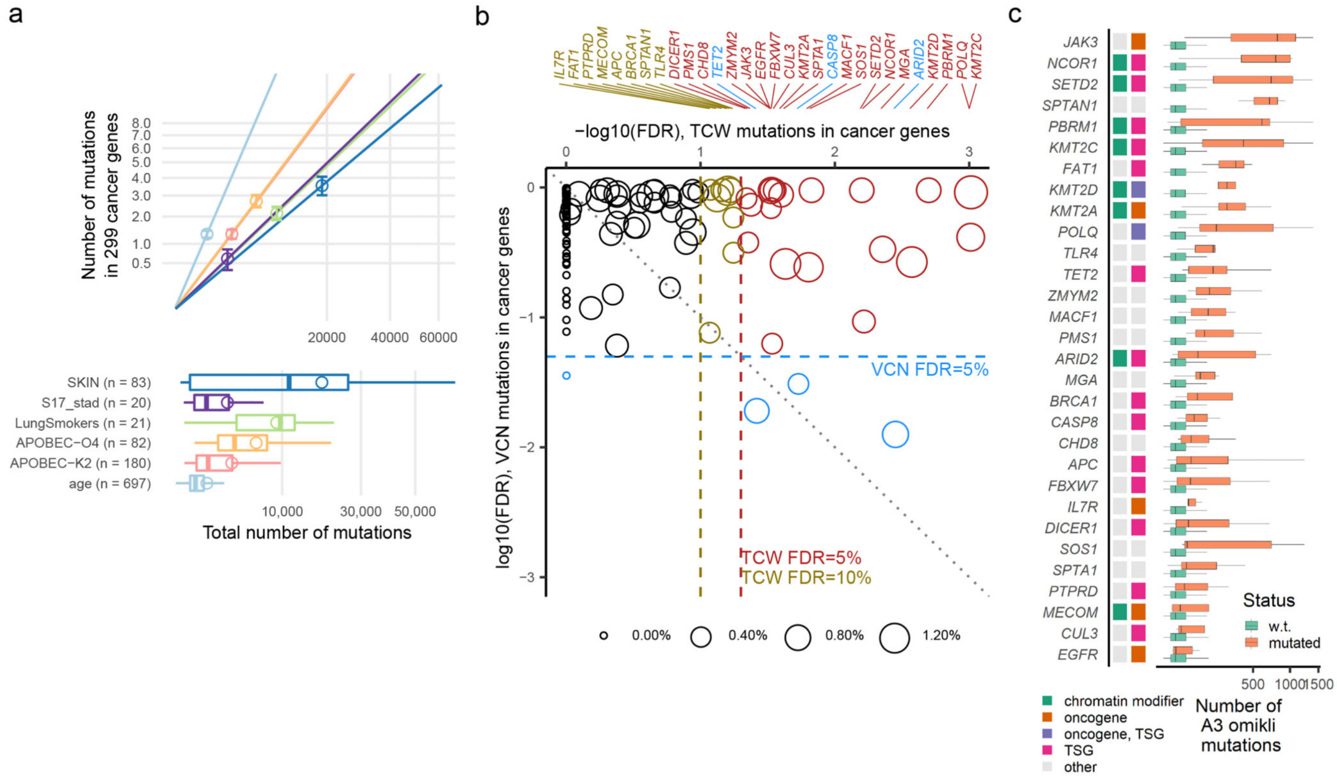


Figure 5. APOBEC mutagenesis generates many impactful mutations.

a, Functional impact density of mutational processes (slope of line), estimated as the number of mutations in coding regions of 299 cancer genes (Y-axis) normalized to the total mutation tally contributed by a process (X-axis). Bottom panel shows the number of mutations estimated to result from each process across tumor samples. Points in boxplots (lower panel) and on lines (upper panel) are the average mutation burden of that process in the affected samples (definition in Methods); error bars are S.E.M. **b**, Occurrence of A3 context mutations in many cancer genes is associated with the genomic burden of A3 *omikli* mutation clusters, suggesting that the *omikli* process generates driver mutations. FDRs are Benjamini-Hochberg adjusted p-values from a logistic regression to predict presence of a TC̲W>K (A3 context, X-axis) or a VC̲N>K (control non-A3 context, Y-axis) mutation in each driver gene. Red and gold, hits at stringent (5%) and permissive (10%) FDR thresholds in the A3 context; blue, hits in the control context (FDR < 5%) suggesting an indirect association with A3 *omikli* burden. Diagonal line denotes equal FDR between the A3 and the control contexts. FDRs were capped at 0.1%. **c**, Burden of A3 *omikli* mutations in tumors which are *wild-type*(teal) or which are mutated (orange) in the driver genes that were significantly associated in the logistic regression in panel **b**.