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 the mechanisms underlying the majority of A3 mutations. We classified the diverse patterns of clustered mutagenesis in tumor genomes, which identified a new 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 pathway, which can provide the single-stranded DNA substrate needed by A3, and contributes to a substantial proportion of A3 mutations genome wide. Because mismatch repair is directed towards early-replicating, gene-rich chromosomal domains, A3 mutagenesis has a high propensity to generate impactful mutations, which exceeds that of other common carcinogens such as tobacco smoke and ultraviolet exposure. Cells direct their DNA repair capacity towards more important genomic regions; thus, carcinogens that subvert DNA repair can be remarkably potent.

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. However, the most striking example of endogenous mutagens is the APOBEC family of cytosine deaminases. They defend against viruses and retrotransposons by damaging their genetic material; additionally, APOBEC1 is a messenger RNA (mRNA)-editing enzyme (reviewed in ref.5). The protein products of APOBEC3 (A3) paralogs were implicated as mutagens in many human cancer types. This is consistent with their ability to deaminate DNA when it is single stranded. 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, and there is some correlation thereof with the tumoral mRNA expression levels of APOBEC3A and APOBEC3B. In addition to the A3 activity, the availability of its single-stranded DNA (ssDNA) substrate is a requirement for mutagenesis. A known source of such ssDNAs is intermediates of DNA repair of double-strand breaks, where A3 results in ‘mutation showers’ or kataegis (Greek for thunderstorm)—local hypermutation events that may consist of tens of mutations. While kataegis is striking, it is not common: very few of the A3 signature mutations are accounted for by the mutation showers. Additionally, DNA secondary structures can generate A3 mutational hotspots; however, the processes that generate global, abundant ssDNA substrates for A3 mutagenesis need to be explored further.

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. A3 signature mutations run counter to this trend. Additionally, the A3 mutations have a curiously strong DNA replication strand bias. These biases, considered together with experimental evidence, suggest that A3 mutagenic activity is coupled with DNA replication. Expressing A3 enzymes in Escherichia coli and yeast produced a mutational bias at replication origins, 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 and MMR promoted A3 signature mutagenesis in flanking segments. Identifying the mechanisms that allow access of A3s to nuclear DNA is important because A3 enzymes generate cancer driver mutations and promote tumor heterogeneity.

Kataegis illustrates how mutation clustering patterns can be used to detect ssDNA-generating mechanisms. 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 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 MMR promotes A3 mutagenic activity, as is 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 and Extended Data Fig. 1a,b). First, it accounts for the heterogeneity of mutation rates and of trinucleotide composition across chromosomal domains, which is an extension of our recent approach\(^1\) with additional support for local false discovery rate (FDR) statistics. Second, 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 have occurred 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 with two previous approaches for detecting clustered mutations\(^2,3,5\). 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 local FDR of \( \leq 20\% \)). Henceforth, we defined the A3 spectrum as C\(\rightarrow T\) and C\(\rightarrow G\) changes in a TCW context (where 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\(^6\). We note that A3 may also rarely generate C\(\rightarrow A\) changes\(^7\). In contrast with previous heuristic rules\(^8,9\), that required, for example, at least five mutations with an IMD of \( \leq 1\) kilobase (kb), 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 and 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, respectively.

While the latter distribution neatly fits current notions of kataegis, the former does not. We named this type of diffuse mutagenesis clustering omikli (fog), by analogy to the focused kataegis (thunderstorm) events. Henceforth, we classify mutation clusters with two, three or four variants as omikli (the short-tract Poisson mixture component predominates; Fig. 1d), and clusters with five or more single-nucleotide variants as kataegis (\( \geq 29\% \) contribution of the component with long tracts; Fig. 1d). A3 omikli is ubiquitous, occurring in more tumors (76% tumors contain at least three A3 omikli mutations; by random expectation, \( \sim 14\% \) 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 have similar abundance per genome (\(Q_1\rightarrow Q_2\): 4–36 mutations) as A3 kataegis (6–36 mutations; Fig. 1f and 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\(^6,10\), enriched near rearrangement breakpoints (a proxy for locations of chromosome breaks\(^11\)), but this is not so for omikli (Fig. 1g). Second, the burden of A3 omikli clusters appears uncoupled from kataegis across individual tumors: it is weakly correlated (coefficient of determination, \( R^2 = 0.11 \)) with long kataegis events (\( \geq 8\) mutations; Fig. 1h), suggesting that short clusters derive from a different mechanism than the intermedi-ate and long ones, which share a common mechanism (\( R^2 = 0.53 \); Fig. 1b). Third, correlation of A3 mutation burden with A3A and A3B mRNA levels is stronger for omikli (Spearman’s \( \rho = 0.31 \) 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; for example, during repair of double-strand breaks\(^12,13,14\). Fourth, the 5′ mutational context of A3 omikli mutations had a significant enrichment of the A3A-like context over the A3B-like context\(^15\) in five cancer types, compared with 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 for A3A-like versus A3B-like contexts, as reported\(^15\) (Extended Data Fig. 2c). Fifth, the unclustered A3 mutation burden is highly correlated with omikli (\( \rho = 0.66 \)) but less so 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), encompassing mutations caused by A3s independent of kataegis and omikli mechanisms plus the TCW\(\rightarrow\)K (where K is T or G) mutations not caused by A3s. Consistently, the distribution of cluster lengths 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- and 2.5-fold for C\(\rightarrow T\) and C\(\rightarrow G\), respectively; Fig. 2a,b), in contrast with unclustered TCW (0.54- and 0.72-fold) and the control, non-A3 context (VGN, 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\(^16\). Protection of early-replicating domains from mutations stems from the differential activity of MMR\(^4,18,44\). Fourth, the 5′ end of A3-context omikli mutations was enriched in regions proximal to CpG dinucleotides (Fig. 2c and 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 previous data\(^13,14\) 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\(^14\). 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\(^15\).

We further examined a set of regions proximal to CpG dinucleotides, proposed to be linked with differential MMR activity\(^8\). There were more A3 omikli clusters in the top genomic tertile by CpG density (Extended Data Fig. 3e). In line with MMR activity causing the mutations, this difference was more pronounced within early-replicating regions. In contrast, the mutation rate of the control VCH (where H is not G) context in CpG-dense regions was lowered (Extended Data Fig. 3e)\(^8\).
Fig. 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 modeling of the number of A3-context mutations per cluster. A solution with two distributions is shown (teal: kataegis; orange: omikli). The stacked bars show component proportions and the curves are density estimates. The gray curve is the baseline solution with one component. The P values are from a two-sided bootstrap test. LL, log likelihood. e, Cumulative percentage of tumor samples that contained 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 mutations 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 CIs. Numbers of mutations are listed in parentheses. h, Pearson’s correlations between the burden of two-mutation omikli and of long kataegis events (left) and between the burden of kataegis of different lengths (right). Statistical significance was determined by two-tailed t-test on the Fisher-transformed correlation coefficients.
Fig. 2 | Association of A3 clustered mutation density with genomic features. 

a. Mutation rates in replication time (RT) quartiles, relative to the latest RT quartile, for A3 mutation trinucleotide contexts (top) and control contexts (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’. b. Mutation enrichment in the earliest versus latest RT quartile for A3-context clusters (top) and non-A3-context clusters (bottom). c. Relative density of A3 and non-A3 mutation types across genomic regions. All enrichments are relative to the lowest bin (the latest RT quartile for replication time), which is not shown. Points are coefficients from negative binomial regression. Error bars are 95% CIs. d. Replication strand bias (ratio of the mutation count on the leading versus the lagging DNA strand) of clustered TCW mutations. Error bars are binomial 95% CIs. As a control, the reciprocal of the strand bias for mSI-H (orange; 24 samples) and PODE mutant (purple; nine samples) tumors is shown as a dashed line. Values in parentheses are mutation counts used to estimate the ratios. mSI-H, microsatellite instability-high. 

e. Left: distributions of IMD in A3-context kataegis and omikli clusters. Right: expected IMD distributions from simulations using three different segment lengths. f. Gamma mixture modeling of the omikli IMD distribution using three components. The bar shows the proportions of the components. The curves show their densities at various IMDs. BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma; HNSC, head and neck squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; SARC, sarcoma; UCEC, uterine corpus endometrial carcinoma.
Next, we examined the replication strand bias$^{42,25}$ of A3 clusters. The ratio of A3 omikli in the leading versus lagging DNA strand closely matched that observed in MMR-deficient (microsatellite-instable) tumors (1.006-fold difference; Fig. 2d), but was less compatible with strand bias associated with a 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$^{43}$, rather than the asymmetry of DNA replication itself$^{46}$, underlies omikli (see Supplementary Information for a discussion).

APOBEC mutagenesis hotspots can occur in DNA sequences that form hairpin secondary structures$^{41}$. 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 an ssDNA intermediate during excision of a mutated DNA segment$^{46,60}$, which 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$^{42,44,46}$. This is in contrast with kataegis, which is known to also stem from DNA repair intermediates; however, these longer segments result from processing of double-strand breaks$^{43,44,46}$. 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 of which are associated with higher MMR activity$^{42,44}$. Because MMR is largely replication coupled$^{44,46}$, the MMR-associated A3 mutagenesis is consistent with the greater vulnerability to A3 damage in dividing cells$^{47}$.

An additional hypothesis was proposed to explain the association of A3 mutations with DNA replication-related genomic features$^{45,46}$: 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 *E. coli*$^{48}$ and yeast$^{45}$. Because the length of eukaryotic Okazaki fragments is known, and the length of MMR intermediates has been characterized in eukaryotic systems reconstituted in vitro$^{49,60}$, we next examined the length distribution of IMDS in the A3 clustered mutations.

The IMD distribution for A3 omikli has a global peak at 355 nucleotides, closely matching the peak (378 nucleotides) of a simulated IMD distribution resulting from 800-nucleotide-long ssDNA segments (Fig. 2e and Methods). The length of MMR excision tracts was estimated at 800 nucleotides using in vitro studies of human and yeast MMR$^{42,48}$. Additionally, we approximated the length of MMR tracts by an analysis of somatic hypermutation events in lymphoid cancer genomes (Methods); this suggested a length range of ~400–1,000 nucleotides (Extended Data Fig. 4a,b). In contrast, the global peak in omikli IMD was not compatible with the Okazaki fragments of ~200 nucleotides in length$^{46}$, which would generate a peak at 96 nucleotides (Fig. 2e). (Of note, in kataegis events, IMDS are devoid of the peak corresponding to tracts of ~800 nucleotides (Fig. 2e); thus, kataegis would result independent 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.

However, we do not exclude that the discontinuous synthesis of the lagging strand contributes to A3 mutagenesis because the omikli IMD distribution has a secondary peak corresponding to 200-nucleotide segment lengths (Fig. 2e). Modeling 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 segments of ~200 nucleotides (Extended Data Fig. 4c,d). Notably, the mixture modeling also suggests a minor component in omikli IMD at very short peak lengths (~25 nucleotides; Fig. 2f). It is tempting to speculate that this reflects the binding of the ssDNA protective protein RPA (replication protein A), which has a 24- to 30-nucleotide footprint$^{50,59}$. A secondary IMD peak of this length is observed also in kataegis (Fig. 2e; see Methods for limits of the use of the IMD measure for kataegis analyses).

**MMR deficiencies are associated with lower A3 mutagenesis.** Next, we 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 ($P<0.001$; Mann–Whitney U-test; 5.52-fold difference between the median of samples), but there was no significant difference in the non-A3-context (VCN$>K$) clusters ($P=0.34$; 1.2-fold difference; Fig. 3a). Of note, comparing absolute omikli A3 burdens (that is, not normalized to the overall number of mutations) 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); instead, a profile more similar to unclustered mutations is seen, further supporting that MMR directs A3 mutagenesis. Consistently, A3 omikli burden associates with the expression levels and copy number status of the A3 genes *MSH6*, *MSH2* and *EXO1* (Fig. 3d,e and Extended Data Fig. 3f,g; discussed in detail in the Supplementary Information).

We have further validated findings on an independent set of tumor whole-genome sequences (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. In particular, there is a strongly increased A3 omikli fraction in MSS versus MSI cancers, in a dataset stratified by cancer type, here also including additional tissues such as the prostate and breast; however, this difference is modest in the control, non-A3 context (Extended Data Fig. 5d,e). Moreover, additional supporting evidence of MMR involvement validates in these data, which show that there are significantly increased A3 omikli burdens in tumors with copy number gains in *MSH6*, *MSH2* and *EXO1* genes (Extended Data Fig. 5f). Consistently, there is altered regional distribution of A3 omikli between MSS (enriched in early replication) and MSI cancers (less enriched) (Extended Data Fig. 5g). The IMD distributions of A3 omikli similarly have a peak corresponding to vulnerable DNA segments of ~800 nucleotides in length (Fig. 2e and 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 with 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-status-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. 1 (see Methods). Informally, a correlation between the clustered burden of tumor samples and the 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%); 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); 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 is thus probably 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

---

**Fig. 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. The results for LUAD tumor samples (points) are shown (other cancers are shown in Extended Data Fig. 6). For clarity, combinations of two variables are shown (center: omikli versus unclustered; right: kataegis versus unclustered), even though the regression was performed on the three variables simultaneously (schematic in leftmost panel; Methods). The 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. The dotted line is the average omikli (center) and kataegis (right) mutation burden across tumors. Dashed lines are the estimated contributions for each process (also shown as bars on the right part of the plot). The bottom panels show the same data as the 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, kataegis and a remainder (intercept of regression fit) not explained by either process. Error bars show s.e. of regression coefficients (n = 646 tumors). **c**, The 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 in A3 burden between tissues. Error bars show s.e. of regression coefficients. The shaded band is the 95% CI of the linear fit. GBm, glioblastoma multiforme; KICH, kidney chromophobe cancer; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; pRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; THCA, thyroid carcinoma.
correlated with the absolute A3 mutation burden across cancer types (Fig. 4c). This further supported that an MMR-dependent, probably 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 cancer exhibit less omikli and thus lower overall A3 mutation burdens—even though kataegis is known to occur therein.

**A3 mutagenesis has a high functional impact per mutation.** Certain mutational processes, including A3 activity, MMR failures and the use of translesion DNA polymerases, were reported to, atypically, produce many mutations in early-replicating, gene-rich chromosomal domains\(^1\). Such mutation redistribution\(^1\) 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 the 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\(^2\) 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 the total A3 burden in tumors that harbor predominantly omikli or predominantly kataegis clusters (Methods). This was compared with common mutagenic processes\(^6\) associated with tobacco smoking (C>T in skin), exposure to gastric acid (A>C in stomach) and aging (C>T changes at CpG dinucleotides). A3 mutations derived either from omikli or from kataegis processes have a very high oncogenic FID (0.47 and 0.46 OMPK, respectively; Fig. 5a and Methods)—approximately twice that of common external mutagens, particularly tobacco smoking and stomach acid–associated mutations (both 0.24 OMPK) and ultraviolet exposure (0.19 OMPK).

In addition to A3, another endogenous mutagenic process—the aging–associated C>T changes at CpG dinucleotides—also had high

---

**Fig. 5 | APOBEC mutagenesis generates many impactful mutations.** a. Top: the functional impact density (FID) 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: number of mutations estimated to result from each process across tumor samples. Hollow circles in box plots (bottom panel) and on lines (top panel) are the average mutation burden of that process in the affected tumor samples (definition in Methods). APOBEC-O4, A3 mutagenesis in omikli-rich tumors; APOBEC-K2, A3 mutagenesis in kataegis-rich tumors; S17 stale, Signature 17 mutagenesis in stomach adenocarcinomas; SKIN, UV mutagenesis in melanoma; age, aging-associated mutagenesis (details in Methods). Error bars are s.e.m. b. The 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. This is based on the reasonable assumption that many mutations occurring in a typical cancer gene exhibit less omikli and thus lower overall A3 mutation burdens—even though kataegis is known to occur therein.
AnAl ysis NATurE GENETiCS

A3. Using intronic mutation rates as a baseline (Methods), we increased positive selection on oncogenic changes introduced by trinucleotide composition therein. Data Fig. 7a), so the high FID of A3 mutations is irrespective of examined a set of known essential genes, and a set of genes linked found that selection on A3 mutations was not stronger than on Fig. 7b). The difference of FID of A3 processes versus external muta- than intergenic regions are. The high intronic-to-intergenic ratio n with omikli are enriched in tumor suppressors (Extended Data Fig. 8a), support-

Discussion

Clustered mutations, although 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. Here, we propose that another DNA repair pathway—MMR—promotes A3 mutagenesis, generating omikli clusters and the bulk of A3 unclu-

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary infor-

Received: 22 August 2019; Accepted: 30 June 2020;
Published online: 3 August 2020
References

1. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multilocusinstability in humans. *Genome Res.* **24**, 1445–1454 (2014).

2. Rogozin, I. B. et al. DNA polymerase \( \eta \) mutational signatures are found in a variety of different types of cancer. *Cell Cycle* **17**, 348–355 (2018).

3. Seplyarskiy, V. B. et al. Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nat. Genet.* **51**, 36–41 (2019).

4. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Cell Rep.* **1228**, 1282–1294.e20 (2019).

5. Cortez, L. M. et al. APOBEC3A is a prominent cytidine deaminase in breast cancer. *PLoS Genet.* **15**, e1008545 (2019).

6. Glaser, A. P. et al. APOBEC-mediated mutagenesis in urothelial carcinoma is driven by passenger hotspot mutations in cancer driven by passenger hotspot mutations in cancer. *Nat. Genet.* **46**, 478–491 (2014).

7. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).

8. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

9. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).

10. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).

11. Landry, S., Narvaez, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO Rep.* **12**, 444–450 (2011).

12. Súspé, R. et al. Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc. Natl. Acad. Sci. USA* **108**, 4858–4863 (2011).

13. Byeon, I.-J. L. et al. NMR structure of human restriction factor APOBEC3A causes bursts of genome-wide clustered hypermutation. *Science* **315**, 131–142 (2009).

14. Yang, Y. et al. DNA repair targeting mutations to human papillomavirus-driven tumor development. *Nature* **518**, 977–983 (2013).

15. Chen, J., Miller, B. F. & Furano, A. V. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* **3**, e02001 (2014).

16. Cannataro, V. L. et al. APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene* **38**, 3475–3487 (2019).

17. Henderson, S., Chakravarty, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.* **7**, 1833–1841 (2014).

18. Li, Z. et al. APOBEC signature mutation generates an oncogenic enhancer that drives LMO1 expression in T-ALL. *Leukemia* **31**, 2057–2064 (2017).

19. de Bruin, E. C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).

20. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).

21. Ullah, I. et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. *J. Clin. Invest.* **128**, 1355–1370 (2018).

22. Byeon, I.-J. L. et al. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **4**, 1890 (2013).

23. Hofman, M. M. et al. Driver hotspot mutations in cancer are strongly enriched on the lagging DNA strand during replication. *Cell Cycle* **17**, 534–547.e23 (2017).

24. Morin, A., Murray, S. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Front. Microbiol.* **5**, 334 (2014).

25. Morganella, S. et al. The topography of mutational processes in breast cancer reveals mechanisms of DNA damage and repair. *Cell Rep.* **14**, 1273–1282 (2016).

26. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).

27. Zhang, Y. et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. *Cell Rep.* **24**, 351–517 (2018).

28. Chen, J. & Furano, A. V. Breaking bad: the mutagenic effect of DNA repair. *DNA Repair* **31**, 2176–2181 (2016).

29. Chen, J., Gu, L. & Li, G.-M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *J. Biol. Chem.* **293**, 7811–7823 (2018).

30. Mugal, C. F., von Grünberg, H.-H. & Peifer, M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol. Biol. Evol.* **26**, 131–142 (2009).

31. Páfrk, K. et al. SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability. *Cell Rep.* **7**, 2006–2018 (2014).

32. Chen, J. & Furano, A. V. Breaking bad: the mutagenic effect of DNA repair. *DNA Repair* **32**, 43–51 (2015).

33. Andranov, M. A., Bazykin, G. A., Nikolaev, S. I. & Seplyarskiy, V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome Res.* **27**, 1336–1343 (2017).

34. Shinholt, E. et al. Eukaryotic DNA polymerase epsilon reveals reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).

35. Jiricny, J. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.* **17**, 335–346 (2006).

36. Tran, P. T., Erdenzin, N., Symington, L. S. & Liskay, R. M. EXO1-A multi-tasking eukaryotic nuclease. *DNA Repair* **3**, 1549–1559 (2004).
61. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. Nat. Commun. 8, 15180 (2017).
62. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. Nat. Med. 22, 1342–1350 (2016).
63. Maruvka, Y. E. et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat. Biotechnol. 35, 951–959 (2017).
64. Hombauer, H., Srivatsan, A., Putnam, C. D. & Kolodner, R. D. Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. Science 334, 1713–1716 (2011).
65. Hombauer, H., Campbell, C. S., Smith, C. E., Desai, A. & Kolodner, R. D. Visualization of eukaryotic DNA mismatch repair reveals distinct recognition and repair intermediates. Cell 147, 1040–1053 (2011).
66. Jeon, Y. et al. Dynamic control of strand excision during human DNA mismatch repair. Proc. Natl Acad. Sci. USA 113, 3281–3286 (2016).
67. Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature 483, 434–438 (2012).
68. Bowen, N. et al. Reconstitution of long and short patch mismatch repair reactions using Saccharomyces cerevisiae proteins. Proc. Natl Acad. Sci. USA 110, 18472–18477 (2013).
69. Brosey, C. A. et al. A new structural framework for integrating replication protein A into DNA processing machinery. Nucleic Acids Res. 41, 2313–2327 (2013).
70. Fan, J. & Pavletich, N. P. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. Genes Dev. 26, 2337–2347 (2012).
71. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair 81, 102647 (2019).
72. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. Cell 173, 371–385.e18 (2018).
73. Pich, O. et al. The mutational footprints of cancer therapies. Nat. Genet. 51, 1732–1740 (2019).
74. Hodis, E. et al. A landscape of driver mutations in melanoma. Cell 150, 251–263 (2012).
75. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science 358, 234–238 (2017).
76. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559 (2018).
77. Verheijen, B. M., Vermulst, M. & van Leeuwen, F. W. Somatic mutations in neurons during aging and neurodegeneration. Acta Neuropathol. 135, 811–826 (2018).
78. Lei, L. et al. APOBEC3 induces mutations during repair of CRISPR–Cas9-generated DNA breaks. Nat. Struct. Mol. Biol. 25, 45–52 (2018).
79. Relfield, E. J. et al. DNA mismatch repair preferentially protects genes from mutation. Genome Res. 28, 66–74 (2018).
80. Lujan, S. A. et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. Genome Res. 24, 1751–1764 (2014).
81. Peña-Díaz, J. et al. Noncanonical mismatch repair as a source of genomic instability in human cells. Mol. Cell 47, 669–680 (2012).
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 43, 649–662 (2011).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
Methods

Data sources. Mutation calls for TCGA whole-genome sequences were obtained as in ref. 2. In brief, BAM files were downloaded from the Cancer Genomics Hub repository (now superseded by the NCBI Genomic Data Commons) for normal and tumor samples, and somatic single-nucleotide variants were called with Strelka 1.0.6 (ref. 1). Also as previously described, we excluded UCSC blacklisted regions in hg19 (Duke and DARC) 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).

SNP 6 Affymetrix array 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 whole-genome-sequenced mutations and SNP 6 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 made the number of total samples equal to 697 in some analyses.

MSI status and other metadata for hypermutated tumors (that is, POLE status) were obtained as described in ref. 2. In total, our TCGA whole-genome-sequenced dataset contained 24 MSI samples (Supplementary Table 2).

An additional dataset, comprising whole-genome-sequenced somatic single-nucleotide variants, purity estimates and copy number alterations (CNAs), was obtained from the Hartwig Medical Foundation23 and used for the validation analyses shown in Extended Data Fig. 5a–h. This dataset has been processed similarly to our TCGA whole-genome-sequenced dataset (Strelka version 1.0.14 was used for all single-nucleotide variant calls). For hypermutation analyses, we additionally used the Purify tool was used to infer purity and obtain CNA estimates (Supplementary Table 3). Inferred MSI/MSS labels were obtained from the supplementary data of the corresponding publication23. We additionally discarded samples (n = 53) that were treated with temozolomide, which is known to positively select for MMR-deficient tumors in humans.

For the functional impact of ultraviolet mutations, we additionally obtained whole-genome-sequence variant calls of 70 melanoma tumors from the MEA-AU (melanoma; Australia) study24 within the Pancancer Analysis of Whole Genomes (PCAWG) study. For the somatic hypermutation analyses, we additionally obtained whole-genome-sequence variant calls of the blood tumors CLLE-ES (chronic lymphocytic leukemia; Spain) and MALY-DE (malignant lymphoma; Germany) from the PCAWG dataset24, which were available as controlled files in the International Cancer Genome Consortium 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 (https://gdcc.cancer.gov/about-data/publications/mc3-2017/). This dataset contains unified somatic mutation calls for approximately 10,000 whole-exome sequences. 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 whole-exome sequencing samples overlap with the whole-genome-sequenced TCGA cohort. We obtained the MSI status from ref. 25, which contains experimentally determined MSI labels (for esophageal carcinoma, uterine corpus endometrial carcinoma, colon adenocarcinoma, rectum adenocarcinoma and stomach adenocarcinoma) and additionally inferred MSI status labels (at the 80% confidence level) that covered an additional 11 cancer types (Supplementary Table 5).

HyperClust, a randomization-based FDR estimate 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 IMDs on the genomic sequence, but other sources of information can be used, such as the allelic fraction of the mutations.

We developed HyperClust by building on our recent approach, 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 relied on the 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 FDRs.

HyperClust provides a rigorous estimate of the local FDR for each mutation pair in each tumor sample into smaller sets according to different features. Because A3 mutagenesis occurs primarily in coordinated cytosines within ssDNA fragments and in APOBEC mutagenesis, we stratified mutation pairs according to base types (C:A and A:T) and strand-coordinated bases. We additionally stratified by mutation clonal fraction, because it should be shared by the mutations occurring contemporaneously in a cluster (Supplementary Information).

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 setups and was therefore used to obtain mutations for the rest of the analysis (Supplementary Information).

Our method is designed to identify sets of mutations, instead of larger groups, which leads to balanced power of detection for smaller clusters and longer clusters (kataegis-like), where previous methods tend to be better adapted to calling the latter (Extended Data Fig. 1d).

Poison mixture modeling of the number of mutations per tract. Our aim was to examine whether there exist multiple mechanisms generating clustered mutations, resulting in mutation 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 TCW > K, which are likely to be a high-purity set of A3 mutations. Then, we modeled the probability that the number of ssDNA given A3 mutagenesis at each position already present P(x = k) = Poiss(λ), meaning that we represent a cluster 1, represents a triplet and so on. 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 flexmix26 to fit a mixture model, testing the range of components from 1–5. 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 a model with one or more or fewer components.

The AIC of each Poisson component is the exponential of the fitted intercept in the model. The confidence interval for the λ value was obtained by transforming the standard error of that value at CI = 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 tract 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 value (relevant for Fig. 1d, lines).

Samples from skin cancer and B cell lymphoma were excluded from this analysis as they contained particular mutation properties that may confound our analysis. Skin cancer has a high percentage of ultraviolet signature mutations, which overlap with the APOBEC TCW > T context. Somatic hypermutation is common in lymphoma and some mutations therein may present a similar trinucleotide profile to the APOBEC mutagenesis.

Association of increased A3 clustered burden with genomic regions. Genomic segments and bins extracted from chromatin marks were computed as in ref. 1. In brief, ChiP-seq data for the epigenetic mark H3K6m3 were downloaded from the Roadmap Epigenomics repository and stratified (according to the fold enrichment of that mark over the input) into three equal-sized bins where fold enrichment > 1 and additionally bin 0, which corresponded to regions with fold enrichment < 1. Expression values were obtained from Roadmap Epigenomics for genomic and intergenic regions and processed in a similar manner to the ChiP-seq data. Replication time bins were computed from wavelet-smoothed Repli-seq signal tracks from the ENCODE dataset. Again, we binned the genome into equal-frequency bins, where bin 1 was the latest-replicating quartile and bin 4 was the earliest-replicating quartile. These data were averaged over the eight cell lines, as in ref. 1.

To detect significant associations of mutations in specific regions of the genome, we used a negative binomial regression (glm.nb from the MASS R package). In brief, combinatorial intersections between the genomic region sets were computed (four bins for each feature). In each set, the numbers of TCW > K mutations were stratified by the four A3 mutation types (TCW > T, TCW > G, TCW > A and TCT > G). These sets of mutation counts for each mutation type were used as the dependent variable in the regression and had a total length of 256, corresponding to 64 × 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 factor. The three independent variables were the genomic bins of each feature, encoded as factors. This same approach was also used for the control contexts (VCN > T). The 95% CI of the coefficient estimates were computed with the confint function in R.

For this analysis, we excluded the B-cell lymphoma dataset, and we further discarded mutations in the somatic hypermutation off-target loci from ref. 8, which might derive from tumor-infiltrated lymphocytes.

Simulating IMD distributions of mutation tracts. 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
considered three lengths of ssDNA fragment: short (25 base pairs (bp)); mid-length (200 bp; meant to represent the approximate length of ssDNA between Okazaki fragments in eukaryotes); and long (800 bp; meant to represent the ssDNA segments generated during the MMR process). 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 the case of omikli, which consists predominantly of two-mutation clusters, the IMD measure can, for practical purposes, be considered equivalent to the cluster span. For this analysis, we considered samples in which we had at least 50 mutations.

**Gamma mixture modeling of IMD distributions.** It is expected that the distance between two mutations occurring in a single hypermutation event will follow a gamma distribution. Thus, to quantify different mechanisms generating clustered mutations, we modeled the observed IMD distributions as a gamma mixture. We selected only the TCW>K mutations with an IMD lower than 1 kb. 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), which implements an expectation-maximization-based algorithm for the detection of different Gamma components in a mixed distribution. We obtained estimates for mixtures that ranged from one to eight components. As initial parameters, we used \( n = 0.2 \), 100 maximum iterations and an \( e \) value (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 one to two, and from two to three components, while the increase in log likelihood from three to four 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.

As for the IMD distribution analysis, we used samples in the APOBEC-prone cancer types: bladder, breast, lung (LUAD and LUSC), cervical, head and neck, and MMR-proficient uterus cancers.

**Contribution of the A3 clustered mutagenic process to the unclustered mutation burden.** To estimate how much the clustered processes contributed to the unclustered burden (which is the main contributor to the overall tumor mutation burden), we adapted a method that we recently introduced. In brief, we used a robust linear regression (the 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) and additionally an interaction term thereof.

From the fitted model, the intercept is the number of unclustered mutations that can be explained by the presence of the considered kataegis mutation clusters; thus, these mutations probably occur independent of the mechanisms that generate 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 cluster (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 mutation burden. By adjusting for the total predicted unclustered mutations, we can obtain estimates of the contribution of kataegis and omikli to the unclustered burden. Note that because the A3 trinucleotide context (here defined as TCW>K) overlaps with signatures of certain other mutagens, the presence of these non-A3-derived unclustered mutations may inflate the estimate of the A3 contribution.

We examined the A3-context mutations in three samples from the TCGA whole-genome dataset. As a negative control we also counted mutations in the cancer genes at the non-A3-context VCN=K. Next, we performed 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. The mutation status was tested both with genes harboring A3 mutations and additionally with the control context (VCN=K) mutations. The Pvalues for each gene were FDR adjusted using the Benjamini–Hochberg correction.

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

**Statistics.** If not stated otherwise, the comparison of two distributions of continuous values was tested with a two-tailed Mann–Whitney U-test. Pooled Pvalues obtained from stratified data groups were determined with Fisher’s method for combining Pvalues. Pvalues are shown as exact values or otherwise referenced with a symbol according to the following scale: ***P < 0.001; **P < 0.01; *P < 0.05.

All box plots are presented according to the standard box plot notation in the R statistical environment (ggplot2 package): the central box represents the interquartile range (IQR); the central line is the median; 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 box plot 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% CI of the median.

**Data availability.** Whole-genome sequences from the TCGA project were available through the Cancer Genomics Hub repository (now superseded by the NCI Genomic Data Commons). TCGA mutation and VCF data were downloaded from the GDC legacy portal (https://portal.gdc-cancer.gov/legacy-archive). WGS data from the Hartwig Medical Foundation are available at https://www.hartwigmedicalfoundation.nl/en. The whole-exome sequencing data were downloaded from the GDC legacy portal (https://portal.gdc-cancer.gov/legacy-archive).
of TCGA cohort are available through the MC3 dataset 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
83. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
84. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575, 210–216 (2019).
85. Huang, M. N. et al. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. Sci. Rep. 5, 13321 (2015).
86. Wang, J. et al. Clonal evolution of glioblastoma under therapy. Nat. Genet. 48, 768–776 (2016).
87. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. Nature 545, 175–180 (2017).
88. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020).
89. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 6, 271–281.e7 (2018).
90. Grün, B. & Leisch, F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28, 1–35 (2008).
91. Khodabakhshi, A. H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget 3, 1308–1319 (2012).
92. Krüger, S. et al. Rare variants in neurodegeneration associated genes revealed by targeted panel sequencing in a German ALS cohort. Front. Mol. Neurosci. 9, 92 (2016).
93. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3 (Bethesda) 7, 2719–2727 (2017).
94. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 173, 408–416.e11 (2018).

Acknowledgements
We thank the members of the Genome Data Science group and B. Lehner for comments and discussions. This work was funded by the ERC Starting Grant HYPER-INSIGHT (757700) and the Spanish Ministry of Economy and Competitiveness (REGIOMUT, grant number BFI2017-89833-P). The results published here are in whole or part based on data generated by the TCGA Research Network (https://www.cancer.gov/tcga). This publication and the underlying research are partly facilitated by the Hartwig Medical Foundation and Center for Personalized Cancer Treatment (CPCT), which have generated, analyzed and made available data for this research. D.M.P. was funded by a Severo Ochoa FPI fellowship (MCIU/Fondo Social Europeo; BES-2017-079820). F.S. was funded by the ICREA Research Professor program and is a member of the EMBO Young Investigator Program. The authors acknowledge support from the Severo Ochoa Centre of Excellence program to IRB Barcelona.

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

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0674-6.
Correspondence and requests for materials should be addressed to F.S.
Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Detecting clustered mutations and simulating processes that generate clustered mutations. a, Method to determine significant mutation clustering using HyperClust. A baseline distribution is generated by shuffling mutations within 1 Mbp windows multiple times (R1, R2, ..., Rn) to loci with matching trinucleotide contexts. 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 101bp). Two examples of high mutation burden tumors (TCGA-AP-A0LD, TCGA-AP-A0LE) were used 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 (panel 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 TCW>K mutations, or the remaining mutations (columns).
Extended Data Fig. 2 | Tetranucleotide context suggests a role for the A3A enzyme in generating omikli and 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 in a tissue 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 schematic above plot.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Association of clustered mutation rates with replication time (RT). a, RT association per cancer type. Number of mutations per RT bin: A3 context (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 RT bin. Enrichments are not shown when the mutation count was lower than 10. b, Trinucleotide composition of the human reference genome in four RT bins, normalized to the latest RT quartile (leftmost point). The A3 trinucleotide contexts (TGW, 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 (c), or jointly considering RT, mRNA levels and the H3K36me3 histone mark levels (d); 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 TGW 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 (O) versus gain/amplification (+1 and +2) states (blue stars, showing p-values according to legend), and independently, comparing samples with neutral (O) versus loss (−1 and −2) states (purple stars). P-values were not adjusted.
Extended Data Fig. 4 | See next page for caption.
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 (or 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 test (two-tailed).
Extended Data Fig. 5 | See next page for caption.
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 25 bp (green), 200 bp (purple) and 800 bp (orange). i, j, Whole-exome sequences in the TCGA data show an excess of A3 context (TCW) mutation fraction in MSS compared to MSI cancers (panel i), and an excess of TCW mutations at distances <1000 bp, normalized to longer distances, in MSS over MSI samples (panel j). ‘MSI-exp’ (n = 152) denotes the experimentally established MSI-H status while ‘MSI-pred’ (n = 18) is the MSI status predicted using machine learning (ref. 61); ‘nonMSI’ (n = 5,661) is neither of these cases.
Extended Data Fig. 6 | Contribution of the omikli and the kataegis mechanisms 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 (that is tumor samples) including ones not visualized.
Extended Data Fig. 7 | See next page for caption.
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 dinucleotides (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 regions containing genic DNA, y axis; flipped). The comparison of mutagenic agents against APOBEC was performed for selected tissues, matching the relevant tissue with 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 coding sequences (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 RT). 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 CDS 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 that 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 that bore 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).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No special software was used for data collection; we used previously collected data available from either public databases or other publications as described in the Methods.

Data analysis: 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 Biostings (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).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

For the current study we used publicly available data as described in the Methods. In brief, we used a set of whole genome sequences from TCGA available through cGHub repository (now NCI Genomic Data Commons, https://gdc.cancer.gov/). SNP arrays for the same data set were downloaded from the GDC legacy portal.
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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Somatic mutations from 699 whole-genome sequences of tumors were used as a discovery data set. Validation analyses were performed on somatic mutations from an independent set of 2304 tumor whole-genome sequences, and an additional near-independent set of 5831 tumor whole-exome sequences. Sample sizes were determined by data availability.

Data exclusions
Description of excluded samples or cancer types for particular analyses was detailed in the Methods section. A list of cancer sample IDs used in various analyses is listed in the Supplementary Tables.

Replication
Main findings from the discovery set (TCGA-WGS) were validated in two additional data sets (HMF-WGS and TCGA-WES). Please see the Results and Methods sections for details.

Randomization
n/a (there was no randomized allocation of samples to experimental groups -- this is an observational study)

Blinding
n/a (this is an observational study, experimental interventions were not performed)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a  ☒ Involved in the study
☒ Antibodies
☒ Eukaryotic cell lines
☒ Palaeontology and archaeology
☒ Animals and other organisms
☒ Human research participants
☒ Clinical data
☒ Dual use research of concern

Methods

n/a  ☒ Involved in the study
☒ ChIP-seq
☒ Flow cytometry
☒ MRI-based neuroimaging