Reduced mutation rate in exons due to differential mismatch repair

Joan Frigola1,2,4, Radhakrishnan Sabar Nathan1,2,4, Loris Mularoni1,2, Ferran Muiños1,2, Abel Gonzalez-Perez1,2 & Núria López-Bigas1–3,*

While recent studies have identified higher than anticipated heterogeneity of mutation rate across genomic regions, mutations in exons and introns are assumed to be generated at the same rate. Here we find fewer somatic mutations in exons than expected from their sequence content and demonstrate that this is not due to purifying selection. Instead, we show that it is caused by higher mismatch-repair activity in exonic than in intronic regions. Our findings have important implications for understanding of mutational and DNA repair processes and knowledge of the evolution of eukaryotic genes, and they have practical ramifications for the study of evolution of both tumors and species.

Genetic variation in exonic regions is lower than in intronic ones both across species and within populations. This differential exon–intron variation rate is attributed to the action of stronger purifying selection on exonic nucleotide changes, whereas the rate of generation of variants—which precedes the effect of selection—is generally assumed to be overall homogeneous between these two genic regions. This assumption lies at the heart of evolutionary biology and cancer genomics approaches that compare the rates of intronic and exonic variation to estimate the strength of selection acting on coding genes1–5.

Recent studies have shown that the rate of mutations across genomic regions is highly heterogeneous. Replication timing6,7, the level of gene expression8 and the degree of chromatin compaction9,10 have been described as features that affect mutation rate at the mega-base scale. Our group and others recently demonstrated that the local efficiency of DNA repair is influenced by factors that affect accessibility of the repair machinery11–14.

The assumption that introns and exons have similar basal rates of mutation before the action of purifying selection is a reasonable one because exonic and intronic regions are replicated at the same time and transcribed equally. DNA repair mechanisms associated with the advance of the replication fork, as well as transcription-coupled repair, are therefore expected to have equivalent access to both regions. Nevertheless, several features of the chromatin structure—including some that have been related to the recruitment of DNA repair machineries15–17—vary widely between exons and introns18,19. This motivated us to question the long-standing assumption that introns and exons have similar rates of mutation before selection.

Somatic mutations detected in tumors20 are an ideal ground to explore whether exonic and intronic variants appear at the same rate. Tumor cells, upon clonal expansion, accumulate somatic mutations at accelerated rates as compared to the germ line. We demonstrate here that, even in the absence of purifying selection, exons acquire fewer mutations than expected given their nucleotide composition. We show that this decreased exonic mutation burden is detectable across seven tumor types. We also demonstrate that the cause of this reduction is that the mismatch repair (MMR) system acts more efficiently in exons than in introns, and we propose that this differential repair is caused by the differential positioning of histone marks in these two genic regions.

These findings imply that the differential genetic variation in exonic and intronic regions across species and within populations is caused by a combination of different sequence context, rate of DNA repair and purifying selection. This has ramifications of a technical nature for evolutionary methods that rely on the calculation of intronic variation to estimate the strength of selection on genes or to detect cancer driver genes1–3,5,21,22. More generally, these findings have profound implications for knowledge of gene evolution and DNA repair mechanisms.

RESULTS

Differential distribution of chromatin features in exons and introns

We first sought to identify chromatin features with the most different distributions between exons and introns, using data generated by the Roadmap Epigenomics23 and the Encyclopedia of DNA Elements (ENCODE)24. We analyzed 32 chromatin features—comprising 30 histone modifications, the presence of a histone variant (H2A.Z) and DNase I–hypersensitive sites (DHSs)—in 127 cell lines and primary cells from different tissue types and nucleosome density obtained in a lymphoblastoid cell line (Supplementary Table 1). We computed the coverage (fraction of bases overlapping peaks) of each feature on exons and introns located at different positions along the structure of genes (the results of this calculation for three chromatin features are shown in Fig. 1a; Online Methods). Then, we defined the difference...
in the exonic and intronic coverage of each mark in each cell type as the P value of the two-tailed Mann–Whitney test of their comparison (box plots in Fig. 1a). Several chromatin marks exhibited a significant overall difference in exonic and intronic coverage (Fig. 1a,b). In particular, nucleosome density and trimethylation of histone H3 at lysine 36 (H3K36me3) were significantly higher in exons than in introns across the gene, and H3K36me3 was the histone mark with higher coverage across all exons in the gene. This behavior of H3K36me3 was consistent across the majority of the 127 cell types in Roadmap Epigenomics (Fig. 1b and Supplementary Tables 1 and 2). Moreover, H3K36me3 coverage decreased steeply in flanking introns (Fig. 1c). Interestingly, the hMutS protein of MMR machinery, involved in the recognition of mismatches, has recently been described as being recruited to the chromatin through interaction of its hMSH6 subunit with H3K36me3 (refs. 15,17).

We therefore hypothesized that the exonic enrichment of certain chromatin features, in particular H3K36me3, might result in increased recruitment of the MMR machinery to exons. This, in turn, would lead to a reduction in the quantity of exonic mutations with respect to the number of mismatches expected from the exonic sequence content alone.

**Internal exons exhibit decreased exonic mutation burden in POLE-mutant tumors**

POLE-mutant tumors, owing to the decreased proofreading capabilities of DNA polymerase ε, sustain a substantial number of mismatches during DNA replication, which make up a sizable part of their somatic mutations. Therefore, to determine whether the rate of somatic mutations caused by mismatches differs in exonic and intronic regions, we first explored the mutations detected across the whole genomes of six POLE-mutant colorectal tumors, sequenced by The Cancer Genome Atlas (TCGA). We stacked exon-centered 2,001-nucleotide (nt) sequences and computed the mutation burden at each position of this window as the number of mutations overlapping the position. This analysis showed that the mutation burden in positions dominated by exonic sequences was lower than that observed along flanking intronic regions (Fig. 2a, red line).

The mutation probability at individual DNA positions is influenced by sequence context. Therefore, differences in nucleotide composition between exons and introns could provide a plausible explanation for the observed difference between exonic and intronic mutation counts. To compute the expected mutation burden at each position of the 2,001-nt exon-centered window, we distributed the mutations observed in each sequence in the stack, taking into account the conditional probability that each of the 2,001 positions was mutated given the adjoining 5′ and 3′ bases. This sequence-wise distribution of expected mutations (details in the Online Methods) avoids potential biases resulting from aggregating genomic regions with different mutation rates and exon/intron proportions (Supplementary Fig. 1). The distribution of these synthetically generated ‘expected’ mutations in POLE-mutant tumors across exons and their flanking introns showed that more mutations are expected in exons than in introns, as represented by the black line in Figure 2a.

We then set out to compare the number of observed exonic mutations to the expected quantity in POLE-mutant tumors and to assess the statistical significance of the deviation between the two (Fig. 2b and Online Methods). We carried out this comparison at the level of individual genes to guarantee that its results were free from the aforementioned caveat. (Known cancer driver genes \(^{25,26}\) were excluded from this and subsequent analyses to eliminate any deviation due to positive selection.) First, we randomly distributed a number of mutations equal to that observed in each gene across its exons and introns, according to the probability of each nucleotide being mutated. A second method to obtain expected mutation burden based on permutations of observed mutations yielded similar results (Online Methods, Supplementary Fig. 2 and Supplementary Table 3). We then computed the difference between the observed and expected mutation burdens for each gene (Fig. 2c). Most genes (77%) possessed fewer exonic mutations than expected from their sequence content, resulting in a negative difference. After aggregating the numbers of observed and expected mutations across all genes (Fig. 2d), we discovered that, whereas internal exons bore only 5,616 mutations in the six POLE-mutant tumors, 8,996 exonic mutations were expected, given (i) the total number of genic mutations, (ii) the nucleotide composition of exons and introns, and (iii) the mutational processes operating in these tumors. This represents a decrease of 37.6% for the observed exonic mutation burden with respect to the expected burden. Employing a likelihood-based statistical approach (Online Methods), we found this decrease to be statistically significant (P < 0.0001). We have named this phenomenon ‘decreased exonic mutation burden’, and we quantify it globally as the percentage decrease with respect to the expected mutation burden.

We next tested whether the decreased exonic mutation burden was due to increased selective pressure on exons resulting in purifying selection of mutations in these regions during tumor evolution. To determine the impact of purifying selection on the exonic mutation burden, we separated exonic mutations on the basis of their consequence types. We found that the 5,616 exonic mutations in the six POLE-mutant tumors corresponded to 950 synonymous and 4,666 nonsynonymous mutations. If the decreased exonic mutation burden were caused by purifying selection, we would expect it to consist mostly of a decrease in nonsynonymous mutations. Nevertheless, when redistributing genic mutations across intronic, synonymous and nonsynonymous sites according to their mutational probabilities, we found a 35.7% decrease in nonsynonymous mutations, along with a 45.4% decrease in synonymous mutations (P < 0.0001; Fig. 2d). On the other hand, when redistributing solely exonic mutations on the basis of their mutational probability, we found that the expected number of nonsynonymous mutations was very close to the actual number observed: 4,562 (with the remaining 1,054 expected to yield synonymous variants). The results of these two tests support the conclusion that the decrease in the exonic mutation burden is not due to negative selection (Fig. 2d). This result was maintained across bins of genes with different mutation rates and was observable for all individual POLE-mutant tumors (Supplementary Tables 4 and 5).

We then checked that the decreased exonic mutation rate was not driven by a subset of genes at either extreme of the mutation rate range. To do this, we binned the genes into ten groups of increasing mutation rate (Fig. 2e, top). We then aggregated the mutations of the genes in each bin and confirmed that the decreased exonic mutation burden remained around 40% across all bins. Finally, we found that very similar values of decreased exonic mutation burden were observed across groups of genes with increasing replication times, expression levels and H3K36me3 coverage, and also across exons at different positions along genes (Fig. 2e, second to bottom panel). Furthermore, the decrease in exonic mutation burden was not driven by one or few POLE-mutant tumors, as it was observable and significant for each of them (Fig. 3a; this analysis also included a POLE-mutant tumor of uterine adenocarcinoma origin).
In summary, we found a significant decrease in the exonic mutation rate in POLE-mutant tumors. This decrease is not due to sequence content and cannot be explained by negative selection acting on exonic mutations, and it is maintained across genes with all levels of mutation rate and across exons at different positions of the gene.

**Decreased exonic mutation rate is caused by differential mismatch repair**

We reasoned that the decreased exonic mutation burden observed in POLE-mutant tumors could be caused by elevated activity of MMR in exons with respect to their neighboring introns. MMR is
Figure 2 Decreased exonic mutation burden in POLE-mutant colorectal tumors. (a) Exon-centered 2,001-nt-wide observed and expected profiles of mutations in six POLE-mutant colorectal tumors. The light red line represents the actual number of mutations at each position, while the red and black lines represent smoothed mutation numbers using a polynomial fit. Bottom, distribution of the percentage of exonic bases at each position across the 2,001-nt window. (b) Schematic of the method used to compute observed and expected numbers of mutations in exons and introns at the gene level. Rescf, represents the rescaled expected frequency of mutation of each nucleotide in the gene. The conditional probability of mutation of each site (P in the figure) is proportional to this quantity. EExonic and EIntronic represent the expected numbers of exonic and intronic mutations, respectively (see Online Methods for details). (c) Density plot representation of the distribution of gene-level differences in the numbers of observed and expected exonic mutations versus total mutations in POLE-mutant tumors. The distribution of exonic mutation difference (right-hand 1D density plot) is biased toward negative values, indicating that a majority of genes possess lower than expected numbers of mutations in exons. An analogous plot, restricted to genes with at least one expected exonic mutation (Supplementary Fig. 2b) shows similar results. (d) An overall highly significant 37.6% decreased in exonic mutation burden (3,380 fewer observed exonic mutations than the 8,996 expected) is observed in these tumors (top). Both synonymous and nonsynonymous mutations account for this decrease (bottom), with their observed values significantly below expected (left and middle plots), and no fewer nonsynonymous mutations than expected when the latter are computed solely from observed exonic mutations (right plot). (e) The decreased exonic mutation burden is maintained around the overall computed value (37.6%) across groups of genes with different mutation rate (first row), level of expression (second row), replication time (third row), the number of genomic bases covered by H3K36me3 peaks (fourth row) and across exons at different positions in the gene (fifth row). The values in the abscissa in each graph represent the ordinal number of the bins of genes, sorted in the direction indicated in each panel.
the main mechanism responsible for the repair of errors generated by
the polymerase during DNA replication. Colorectal tumors, as well
as other cancer types, acquire a microsatellite instability (MSI)
phenotype when mismatches introduced by the DNA polymer-
ase are not corrected, owing to deficiencies in the MMR system 27.
MSI tumors are normally classified on the basis of the level of five
biomarkers into MSI-H (high, with over 40% of the biomarkers of
MSI) and MSI-L (low, with less than 40%), although the latter have
recently been shown to not significantly differ from microsatellite
stable (MSS) tumors in numbers of gained microsatellite alleles28.
Thus, if our hypothesis were true, we would expect tumors with an
impaired MMR function (MSI-H) to show lower decreased exonic
mutation burden than MMR-competent tumors, such as
POLE
-mutant or MSS tumors.

We proceeded to compute the decreased exonic mutation burden
of six colorectal and eight uterine MSI-H tumors in the TCGA
cohort. We found, as predicted by our hypothesis, that MSI-H
tumors exhibited a decreased exonic mutation burden of around 20%
(Fig. 3a,b), close to half of the decrease observed in the MMR-
proficient POLE
-mutant tumors. Several reasons may explain why
the decreased exonic mutation burden did not disappear completely
in MSI-H tumors. On the one hand, the impairment of the MMR system
may not be complete and has probably not existed throughout the
entire history of the tumor. On the other hand, alternative mutational
processes may also contribute to the mutation load.

Then, we computed the decreased exonic mutation burden of
two POLE
-mutant and two POLD
-mutant glioblastomas from chil-
dren with inherited biallelic mismatch-repair deficiency (bMMRD)
sequenced by the International BMMRD Consortium29. These tumors
have been MMR-deficient throughout their entire history, and their
POLE
 or POLD
 mutations guarantee a preponderance of mismatch-
caused mutations. Their decreased exonic mutation burden was
indeed close to zero (Fig. 3a,b), with independence of the mutation
rate of genes (Supplementary Fig. 3 and Supplementary Table 4).
Mismatches in these tumors were generated at a rate comparable to
that in previously analyzed POLE
-mutant tumors. However, most of
these mismatches remained uncorrected and turned into mutations.
In other words, the mutations observed in these tumors follow the
pattern of mismatch generation, corroborating our hypothesis that
they appear with higher probability in exons than introns and that it
is MMR, with its increased efficiency in the former, that causes the
decreased exonic mutation burden.

In summary, the decreased exonic mutation burden differs between
different scenarios of MMR activity, with higher decrease in
MMR-proficient tumors to none in MMR-deficient tumors. These
results indicate that the increased activity of MMR in exons is the

![Figure 3](https://example.com/figure3.png)

**Figure 3** Decreased exonic mutation burden across scenarios of MMR activity. (a) Tumor-level decreased exonic mutation burden in POLE
-mutant (left) and, MSI-H (center) colorectal and uterine tumors and bMMRD glioblastomas (right). Dots represent individual tumors. Broken red lines represent the
median decreased exonic mutation burden of each group of tumors. (b) Exon-centered 2,001-nt-wide observed and expected profiles of mutations in
six colorectal and ten uterine MSI-H tumors and two POLE
-mutant and two POLD
-mutant bMMRD glioblastomas. (c) Schematic showing the increased
efficiency of MMR at exons and the decreased exonic mutation burden.
be explained, at least in part, by the exonic enrichment of this histone mark in cells of the colon epithelium. If true, we should be able to observe the biggest decrease in exonic mutation burden in genes with the strongest exonic enrichment for H3K36me3 in MMR-proficient tumors. To test this, we first computed the exon-to-intron ratio of H3K36me3 read count for primary cells from the colonic mucosa (E075; Fig. 4a)23. Then, we grouped the genes into bins of increasing H3K36me3 exon-to-intron ratio, and we computed the aggregated decrease in exonic mutation burden of the genes in each bin for POLE-mutant colorectal tumors (Fig. 4a and Supplementary Figs. 4 and 5). As predicted by our hypothesis, we found a significant negative correlation between the H3K36me3 exon-to-intron ratio and the decrease in exonic mutation burden (correlation coefficient = −0.68, \( P = 6.7 \times 10^{-8} \)). A much weaker, non-significant correlation (Supplementary Fig. 5, bottom) was observed between the exon-to-intron ratio of nucleosomes and the decrease in exonic mutation burden. This suggests that the H3K36me3 histone mark and not just the presence of nucleosomes underlies the increased level of MMR in exons that results in the decreased exonic mutation burden. The correlation with other histone marks was also lower (Supplementary Table 6). On the other hand, the negative correlation between the H3K36me3 exon-to-intron ratio and the decreased exonic mutation burden was absent in MSI-H colorectal tumors (correlation coefficient = 0.12, \( P = 0.46 \)) and bMMRD tumors (exon-to-intron H3K36me3 read count ratio computed from cells of the brain angular gyrus, E067; correlation coefficient = 0.07, \( P = 0.64 \)) (Fig. 4b,c).

These results indicate that the exonic enrichment for H3K36me3, possibly in combination with other chromatin features, could act as a driver of the enhanced MMR activity in exons that ultimately results in the decreased exonic mutation rate of POLE-mutant tumors. When cells become MMR deficient, either during tumor evolution (MSI-H colorectal samples) or before its emergence (bMMRD glioblastomas), the link between the H3K36me3 exonic enrichment and the decreased exonic mutation burden is thus severed. This results in uncorrected mismatches accumulating and, ultimately, mutations appearing more frequently in exons.

**Tumors of other cancer types also exhibit decreased exonic mutation rate**

Our observations in previous sections have been limited to colorectal and uterine carcinomas, the mutational spectra of which are dominated by the interplay between the generation of mismatches in the course of DNA replication and their correction by the MMR machinery. The mutational processes of other somatic tissues are dominated by different types of damage dealt with by other DNA repair systems. Nevertheless, somatic cells in a human body, as well as the gametes, are the result of millions of cell divisions involved in organism development and tissue renewal. Therefore, MMR must have a role—although with different relative contribution—in shaping the mutational processes of all human tissues. We then asked whether tumors originated from other tissues exhibit a decreased exonic mutation rate. To do this, we first clustered the samples of eight tumor types on the basis of their mutational signatures (Supplementary Fig. 6).

For the tumors in each cluster, we next computed the decreased exonic mutation burden (Fig. 5a, top). All clusters except the one grouping POLE-mutant bMMRD glioblastomas exhibited significantly decreased exonic mutation burden. This global trend was corroborated for individual samples (Fig. 5b). Interestingly, we found that the decreased exonic mutation rate was apparent also in the somatic mutations detected in a normal skin sample (Fig. 5b, white-crossed

---

**Figure 4** Decreased exonic mutation burden and H3K36me3 exon-to-intron ratio. (a–c) Decreased exonic mutation burden computed in POLE-mutant colorectal tumors (a), MSI-H colorectal tumors (b) and bMMRD glioblastoma tumors (c) for 50 groups of genes with increasing exon-to-intron ratios of H3K36me3 coverage (in the corresponding cell of origin; Online Methods). The trendline and its confidence interval in graphs was added using the seaborn package of Python, while the correlation coefficient and its significance were computed using an iteratively reweighted least-squares approach.

**A role for H3K36me3 in the differential activity of MMR in exons and introns**

The results of the previous two sections demonstrate that the enhanced efficiency of the MMR system in exons is the cause of the observed decreased exonic mutation burden of colorectal POLE-mutant tumors. On the basis of formerly established mechanistic links between H3K36me3 and the recognition of mismatches, we then hypothesized that the decreased exonic mutation burden could
Figure 5 Decreased exonic mutation burden across cancer types. (a) Top, decreased exonic mutation burden of groups of tumors clustered according to their underlying mutational processes. Dots represent clusters of tumors denoted in the bottom panel. Middle, decreased nonsynonymous mutation burden of the same groups of tumors, computed as in Figure 2d. Bottom, ratio of exonic and intronic mutation rates of the same groups of tumors. Clusters of tumors in the three panels are sorted following their decreased exonic mutation burden. (b) Decreased exonic mutation burden of individual tumors versus their mutational burden. Dots representing individual tumors are colored according to their cancer type; dots of tumors with significantly decreased exonic mutation burden are encircled by a black ring. The table to the top right presents the total number of samples, the subset of them with decreased exonic mutation burden and the subset of these with a significant decrease.

DISCUSSION

In this work, we provide, to the best of our knowledge, the first demonstration that the generation of somatic mutations—in the absence of negative selection—is lower in exons than expected given their nucleotide composition. In other words, somatic cells exhibit a decreased exonic mutation burden. We have also shown that the reason is that mismatches in exonic DNA are repaired more efficiently than their intronic counterparts. These results represent an important contribution to the body of research that in recent years has revealed higher than anticipated heterogeneity in the mutation rate across different regions of the genome. Several recent seminal studies exploiting whole-genome germline and somatic mutations and the availability of nucleotide-resolution maps of DNA repair33,34 have provided glimpses at a complex relationship between chromatin conformation, basic cellular processes like gene expression, DNA replication, the binding of transcription factors, and DNA repair 5,7,10–12,14,32,35–39. It is the complicated interplay between these processes that determines that mutations accumulate heterogeneously across the genome. The results we present here show that the interaction of the most basic structural feature of eukaryotic genes, namely their segmentation in exons and introns, and its correlative chromatin structural differences, results in these two regions being repaired at very different rates.
As a possible explanation of the mechanisms through which the segmented structure of genes influences the activity of the DNA repair machinery, we have shown a pervasive enrichment of the trimethylation of H3K36 in exons of normal tissues, which correlates with the decrease in exonic mutation burden in the corresponding tumors. As we show here, H3K36me3, possibly in combination with other chromatin features, may participate in shaping the observed depletion of exonic mutations. The enrichment of H3K36me3 for exonic regions, which appears in both germline and somatic tissues, has been proposed to be ultimately responsible for the correct recognition of exon-intron boundaries by the splicing machinery. Nevertheless, H3K36me3 is bound by the MutSε protein via the PWWP domain of its MSH6 subunit. A concomitant factor may thus have acted in the evolution of H3K36me3-enriched exons. Indeed, our results suggest that the increased recruitment of the MMR machinery to exonic regions as a result of higher levels of this histone mark would result in a reduction of the exonic mutation burden after DNA replication and, ultimately, in an increase of fitness.

Our results demonstrate that the decreased exonic mutation burden is not due to negative selection in the generation of cancer somatic mutations across all tumor types analyzed. This finding suggests that the mutational landscape of cancer-related genes is not strongly influenced by negative selection, in agreement with a recent report. Nevertheless, we expect that, in the germ line, purifying selection has influenced by negative selection, in agreement with a recent report.41. This, in turn, would bring into question the use of rates of intronic substitution (Kι) as a proxy for neutral evolution,44–46 with important implications for understanding of the evolution of genes. Further implications may be extracted for methods aimed at detecting cancer driver genes that model the background mutation rate of exonic elements from their surrounding areas to identify signals of positive selection in the coding region of genes. Some of these methods,21,22, which use intronic mutations as estimators of the exonic background mutation rate, may be strongly affected by the differential generation of mutations in these regions.

In summary, we demonstrate that the differential MMR in exons and introns in somatic cells causes the former to harbor fewer mutations than expected from their nucleotide composition. This finding advances knowledge of the interplay between mutational processes and the DNA repair machinery. Moreover, our results have important implications regarding the way we study the forces that shape the development of tumors and the evolution of the genome.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge funding from the Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE), La Fundació la Marató de TV3, EU H2020 Programme 2014-2020 under grant agreements 634143 (MedBiobinformatics) and by the European Research Council (Consolidator Grant 682398). IRB Barcelona is the recipient of a Severo Ochoa Centre of Excellence (MINECO, Government of Spain) and is supported by CERCA (Generalitat de Catalunya). R.S. is supported by an EMBO Long-Term Fellowship (ALTF 568-2014) cofunded by the European Commission (EMBOCOFUND2012, GA-2012-600394) with support from Marie Curie Actions. A.G.-P. is supported by a Ramón y Cajal contract from the Spanish Ministry of Economy and Competitiveness (RYC-2013-14554). We acknowledge the contribution of I. Reyes-Salazar to refactoring and cleaning all code produced in the study for publication. We are grateful to R. Campbell and U. Taborí for help in obtaining the mutation calls for bMMRD samples sequenced by the International BMMRD Consortium. The results published here are in part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

J.F. and R.S. participated in the design and execution of analyses, produced the figures, participated in the interpretation of results and edited the manuscript. L.M. developed computational code employed in the analyses. E.M. developed the statistical framework to compute the significance of the decreased exonic mutation burden and its correlation with chromatin features. A.G.-P. participated in the design of analyses, the interpretation of results, the oversight of analyses, and drafted and edited the manuscript. N.L.-B. conceived the study, participated in the design of analyses, oversaw the study and the interpretation of results, and drafted and edited the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.
21. Hodis, E. et al. A landscape of driver mutations in melanoma. Cell 150, 251–263 (2012).
22. Lanzós, A. et al. Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features. Sci. Rep. 7, 41544 (2017).
23. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
25. Futreal, P.A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).
26. Rubio-Perez, C. et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell 27, 382–396 (2015).
27. Li, G.-M. Mechanisms and functions of DNA mismatch repair. Cell Res. 18, 85–98 (2008).
28. 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).
29. Shlien, A. et al. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. Nat. Genet. 47, 257–262 (2015).
30. Martincorena, I. et al. High burden and pervasive somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. Nat. Genet. 47, 257–262 (2015).
31. Kolasinska-Zwierz, P. et al. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. 41, 376–381 (2009).
32. Chamary, J.V., Parmley, J.L. & Hurst, L.D. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7, 98–108 (2006).
33. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87 (2005).
34. Hoffman, M.M. & Birney, E. Estimating the neutral rate of nucleotide substitution using introns. Mol. Biol. Evol. 24, 522–531 (2007).
ONLINE METHODS

Whole-genome expression and mutation data. Whole-genome somatic mutations and expression data for 38 skin cutaneous melanomas (SKCM), 46 lung adenocarcinomas (LUAD), 45 lung squamous cell carcinomas (LUSC), 42 colorectal adenocarcinomas (CRC), 96 breast carcinomas (BRCA), 21 bladder carcinomas (BLCA), 47 uterine corpus squamous cell carcinomas (UCEC), 27 glioblastomas (GBM), 18 low-grade gliomas (LGG), 20 prostate adenocarcinomas (PRAD), 34 thyroid carcinomas (THCA) and 27 head and neck squamous cell carcinomas (HNSC) probed by TCGA were obtained from Fredriksson et al.\textsuperscript{29}. Cohorts of tumors with fewer than 5,000 genic mutations or fewer than 1,000 exonic mutations (HNSC, GBM, KIRC, THCA, LGG and PRAD) were discarded from the analysis. The somatic mutations detected in four BMMRD pediatric glioblastomas sequenced by the International BMMRD Consortium\textsuperscript{28} were obtained through personal communication from the authors. Finally, we obtained the somatic mutations detected across the whole genome of a normal skin sample from Martincorena et al.\textsuperscript{30}.

Genomic coordinates of exons and introns. GENCODE\textsuperscript{47} v19 coordinates for 20,345 protein-coding genes were retrieved. Genes without introns, overlapping genes and cancer driver genes, according to the Cancer Gene Census and other sources\textsuperscript{25,26}, were discarded, leaving a filtered set of 12,104 genes. All transcripts per gene were merged, generating meta-exon and meta-intron coordinates. Finally, 5’ and 3’ exons were removed, as well as all UTRs (except for the analysis shown in Fig. 1), thus leaving only internal exons and their flanking introns. We then identified all genic regions where mutation calling would be technically challenging because of low sequence complexity, ambiguous mappability of sequencing reads or low sequencing coverage. Regions of low complexity or low mappability were obtained from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMtapability). The former included repetitive regions defined by RepeatMasker, while the latter comprised regions with low unique mappability for 36-mer sequences (CRG Alignability 36’ track, score <1). Finally, regions covered by fewer than eight reads in any of five randomly selected tumor samples of each tumor type (the requirement to make somatic calling in Fredriksson et al.\textsuperscript{29}) were considered of low coverage. Regions of any of these three types were removed from introns and exons.

Clusters of tumors with different somatic mutational processes. To group the tumors of each cancer type in the cohort according to their underlying mutational processes, we first built a matrix of the frequencies of the 96 trinucleotide changes across tumors, as in a previous work\textsuperscript{12}. We carried out hierarchical clustering (using a Euclidean distance and average method to compute the similarity between clusters) of this matrix. We then manually separated the clusters of tumors and identified their underlying mutational processes through visual comparison with previously obtained\textsuperscript{25,26} mutational signatures across cancer types. Clusters of tumors with fewer than 5,000 genic mutations or fewer than 1,000 exonic mutations were discarded for downstream analyses.

Chromatin features. We downloaded peak (narrow) coordinates and genome-wide read coverage for the 32 chromatin features presented in Supplementary Table 1 across 127 cell lines and primary cell types from Roadmap Epigenomics\textsuperscript{23} and the nucleosome density from ENCODE\textsuperscript{24}. Peaks and reads (see below) obtained from http://egg2.wustl.edu/roadmap/data/byFileType/narrowPeak and http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated, respectively, for each feature were mapped to intronic and exonic regions of genes. The primary cell types closest to colorectal tumors and glioblastomas in Roadmap Epigenomics were selected to represent the exon–intron distribution of chromatin features. Genome-wide nucleosome positioning signals (density graphs) for the ENCODE cell line GM12878 (lymphoblastoid cell line) were obtained via the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/). Further, by using the bwttool find program (with parameters localextrema -maxima -min-sep = 150), nucleosome peak regions were identified across the genome, and the 146-bp regions flanking each peak (73 bp per side) were considered as regions covered by a nucleosome.

We numbered the exons and introns in each gene according to their positions with respect to the transcriptional start site (TSS). Exons and introns that occupied different positions in different transcripts and those in the lower quartile of length were discarded. We then stacked all exons and all introns separately and computed the aggregated coverage (fraction of bases covered by peaks for each mark) at the center of the stack corresponding to the number of bases of the shortest exon or intron remaining after filtering. Finally, the difference between exonic and intronic coverage was computed via the two-tailed Mann–Whitney $P$ value for the comparison of the two distributions.

Classification of colorectal tumors according to MMR level. Colorectal samples were separated into four subtypes on the basis of MMR levels. MSI-H ($n = 6$), MSI-L ($n = 4$) and MSS ($n = 26$) groups were defined on the basis of clinical information from TCGA (https://portal.gdc.cancer.gov/query). The POLE-mutant group ($n = 6$) was defined by identifying samples with missense mutations of the POLE (DNA polymerase $\epsilon$) gene.

Exon-centered mutational analyses. We stacked 2,001-nt sequences centered on the middle position of internal exons. In this analysis, we did not exclude regions that overlapped any of the three types of technically challenging regions mentioned above. Thus, we obtained a stack of 95,164 sequences centered on exons. We then counted the observed and expected (distributed across each sequence of the 2,001-nt window following the mutational probability for each nucleotide, as explained below for individual genes) mutations associated with each nucleotide of these sequences. With these counts across the selected windows, we produced exon-centered plots as shown in Figures 2a and 3b.

Computing decrease in exonic mutation burden. We first computed the relative frequencies of the 192 possible trinucleotide changes, $f(A_i X C_j \rightarrow A_i X C_k)$, across each cluster of tumors as

$$f(A_i X C_j \rightarrow A_i X C_k) = \frac{N(A_i X C_j \rightarrow A_i X C_k)}{T}$$

where $N(A_i X C_j \rightarrow A_i X C_k)$ was the number of such changes among all mutations observed in the tumors and $T$ was the total number of substitutions observed across tumors. Then, we made $f$ relative to the abundance of each trinucleotide in the genome, $G(A_i X C_j)$,

$$\tilde{f}(A_i X C_j \rightarrow A_i X C_k) = \frac{f(A_i X C_j \rightarrow A_i X C_k)}{G(A_i X C_j)}$$

Next, for each genic site, we summed the relative frequency of its three possible changes given its 5’ and 3’ flanking bases.

$$\overline{\sum}_{mut}(A_i X C_j \rightarrow A_i X C_k) = \sum_{l=0}^{4} \tilde{f}(A_i X C_j \rightarrow A_i X C_k)$$

We rescaled the relative frequency of change for each site to 1 by multiplying each frequency by factor $k$.

$$k = \frac{1}{\overline{\sum}_{site}}$$

The rescaled frequency (Rescf) of each nucleotide in the gene is proportional to the conditional probability that the reference nucleotide changes to the alternative given its 5’ and 3’ nucleotides. Finally, for each independent gene, we redistributed all observed mutations ($N_{muts}$) across exonic and intronic sites following these summed rescaled frequencies of each site to be mutated.

$$EE_{xonic} = N_{muts} \sum_{i \in \text{xonic sites}} \text{Rescf}_i \forall \text{xonic sites}$$

$$E_{intronic} = N_{muts} \sum_{i \in \text{intronic sites}} \text{Rescf}_i \forall \text{intronic sites}$$

Note that this redistribution process could be performed equivalently for the mutations observed in one tumor (for single-tumor analysis; Fig. 5b) or across a group of tumors (for group or cluster analyses; Figs. 2, 3, 4 and 5a). The process yielded the number of expected exonic (EE_{xonic}) and intronic (E_{intronic}) mutations in the gene. We employed a second method to compute the expected number of exonic mutations based on the average of 1,000 random permutations of the observed mutations in each gene following the probability of each site to acquire a mutation (Supplementary Table 3).
Summing the observed and expected exonic mutations over all genes, we computed the difference between the observed and expected numbers of exonic mutations, which we refer to as the decrease in exonic mutation burden (as in most tumors there was a negative difference). Throughout the paper, we express this decrease as the percentage of the total number of observed exonic mutations.

To compute the significance of this decrease, we employed two tests: (i) a $G$ test of independence comparing the numbers of observed and expected mutations in exons and introns, under the null hypothesis that the observed and theoretical distributions of the variables are equal, and (ii) for the expected number of exonic mutations computed using the permutations approach, we computed an empirical $P$ value as the fraction of the iterations with fewer expected than observed exonic mutations.

Test for negative selection on exonic mutations. The consequence type of all observed exonic mutations was obtained using the Ensembl Variant Effect Predictor (VEP; v.70). We subsequently separated exonic mutations into two groups: those with synonymous consequence and those with a consequence ranking higher than synonymous in the Ensembl Variation hierarchy (http://www.ensembl.org/info/genome/variation/predicted_data.html), which were collectively deemed nonsynonymous. All possible nucleotide changes in a gene were then divided into three categories: (i) synonymous; (ii) nonsynonymous (with the consequences defined above); and (iii) intronic. We redistributed the mutations observed in each gene across these three types of sites following the probability of occurrence of each change computed as explained above.

Through the difference between observed and expected synonymous and nonsynonymous mutations, we were able to compute the decrease in the burden of both types of mutations (expressed as the percentage of the expected number, as explained above for all exonic mutations). Finally, a $G$ test of independence was used on the null hypothesis that fewer nonsynonymous mutations should be observed than expected.

We also redistributed only the exonic mutations across synonymous and nonsynonymous sites according to the probability of change for each type of site. In this case, we used the $G$ test of independence on the null hypothesis that the number of expected nonsynonymous mutations was not smaller than the observed number.

Stratification of genes by mutation rate and several covariates. The mutation rate of each gene was computed as the quotient between the number of observed mutations and the number of bases in the gene. Genes were subsequently grouped into ten bins according to their mutation rate.

We computed the 75th percentile of the expression of each gene across the tumors in each cohort. Genes with a 75th percentile of expression equal to 0 were considered to be non-expressed and were grouped together. All other genes were sorted on the basis of their previously computed expression percentile and divided into nine bins of equal size. Non-expressed genes were subsequently added as a tenth bin.

Replication time data across the human genome measured in lymphoblastoid cell lines were obtained from Koren et al. Using these data, the mean replication time per gene was computed. Next, genes were sorted on the basis of this value and divided into ten groups of equal size.

Finally, we also grouped the genes into ten bins according to H3K36me3 peak coverage.

Relationship between decreased exonic mutation rate and exonic enrichment of nucleosomes and histone marks. For each gene, we computed the read-count-based exonic enrichment for any chromatin feature as the ratio between the exonic and intronic read counts (total number of bases covered by reads of the chromatin feature). (This read-count-based exonic enrichment was used to compute the correlations shown in Fig. 4 and Supplementary Fig. 4.) We computed the peak-based exonic enrichment of any chromatin feature as the ratio of exonic and intronic bases covered by peaks of the feature (to compute the correlation shown in Supplementary Fig. 5). The exonic and intronic numbers of bases covered by reads or peaks of the chromatin feature for colorectal and bMMRD glioblastoma tumors were computed from colonic mucosa (E075) and brain angular gyrus (E067) cells, respectively, both obtained from Roadmap Epigenomics. (In the case of nucleosomes, peaks were obtained from occupancy values as explained above.) Genes were grouped into 10, 25 or 50 bins according to their exonic H3K36me3 enrichment, and the aggregated decrease in the exonic mutation rate of the genes in each bin was computed as explained above for colorectal POLE-mutant, MSI-H and bMMRD tumors. We then computed the correlation between the median exonic chromatin feature enrichment and the decreased exonic mutation rate across the bins. The trendline and its confidence intervals were added to each plot using the bootstrapping functions of the Python seaborn package, which confers equivalent weights in the regression to all points. To guarantee that the trend was not the result of a few outliers, the correlation coefficient and its significance were computed using an iteratively reweighted least-squares approach, letting the variance in exonic H3K36me3 enrichment of the bins influence the weight of each point.

Exon-to-intron mutation rate ratio. As described above, we stacked all exon-centered and intron-centered sequences. Then, we averaged the total number of mutations observed at each of the 41 central positions of each stack. The selection of 41 central positions guaranteed both a vast majority of exonic sequences contributing mutations and enough mutations for calculation across all clusters at exon-centered stacks. The exon-centered and intron-centered mutation burden averages were then divided by the number of sequences included in each stack to make them comparable. Finally, we computed the exon-to-intron mutation rate ratio as the quotient between the corrected exon-centered and intron-centered mutation burden averages.

Computing the activity of NER from XR–seq data. Genome-wide maps of NER for two UV-induced photoproducts, namely cyclobutane pyrimidine dimers (CPDs) and pyrimidine–pyrimidine (6,4) photoproducts (PP64s), in irradiated skin fibroblast cell lines were obtained from Hu et al. This data set comprises NER maps for the following three cell lines: (i) wild-type NIH3T3 skin fibroblasts, which have active global and transcription-coupled repair mechanisms; (ii) XP-C mutants, which are deficient in the global repair mechanism; and (iii) CS-B mutants, which are deficient in transcription-coupled repair. For each of these cell lines, we extracted the sequencing reads, processed and mapped to the human genome, following the steps mentioned in Hu et al. Further, we selected the mapped reads that were 26 nt in size, which is the typical size of NER-excised oligomers, and classified the reads on the basis of the presence of dipyrimidines (TT, CT, TC and CC) at positions 19–20 or 20–21 of the reads. In addition, we recorded the mapped genomic locations of the nucleotides in positions 19–20 or 20–21 of the reads. This way, we could predict the damage site according to the excised fragments. We mapped this information to the XR–seq exon-centered plot together with the frequencies of the dipyrimidines observed in each column (Supplementary Fig. 7).

Statistics. We used the $G$ test described above to compute $P$ values to test the significance of decreased exonic mutation burden for groups of genes or all genes in a tumor or across groups or clusters of tumors. (All $P$ values computed for all comparisons are provided in Supplementary Table 3, which also includes $P$ values computed using a permutation-based test also described above.) When appropriate, $P$ values computed with this test were corrected using the Benjamini–Hochberg approach. In Figure 1, we used the two-tailed Mann–Whitney test to compare the exonic and intronic distributions of chromatin features (and corrected them when appropriate). Above, we describe the approach employed to compute the correlation coefficient (and its associated $P$ value) for the regression lines shown in Figure 4 and Supplementary Figures 4 and 5.

Code availability. All code needed to reproduce the analyses described in the paper are available at the Bitbucket repository (https://bitbucket.org/bbglab/intron_exon_mutrate). Preprocessed data needed to reproduce all analyses described here are provided together with the code at https://bitbucket.org/bbglab/intron_exon_mutrate. A Life Sciences Reporting Summary is available.
Author Correction: Reduced mutation rate in exons due to differential mismatch repair
Joan Frigola, Radhakrishnan Sabarinathan, Loris Mularoni, Ferran Muiños, Abel Gonzalez-Perez and Núria López-Bigas

Correction to: Nature Genetics https://doi.org/10.1038/ng.3991, published online 6 November 2017.

In the version of this article initially published, the x axis on the fourth plot in Fig. 2e was incorrectly labeled "H3K36me3 exon-to-intron ratio (lower to higher)." The x axis on this plot should read "Genic H3K36me3 coverage bins (higher to lower)."

Published online: 04 July 2018
https://doi.org/10.1038/s41588-018-0123-y
**Experimental design**

1. **Sample size**
   - Describe how sample size was determined.
   - All available data from sources cited in methods was analyzed

2. **Data exclusions**
   - Describe any data exclusions.
   - Cohorts of tumors with fewer than 5000 mutations or fewer than 1000 exonic mutations were excluded

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   - Not applicable

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   - Not applicable

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - Not applicable

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. **Statistical parameters**
   - For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).
   - Confirmed

   - The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - A statement indicating how many times each experiment was replicated
   - The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - Clearly defined error bars

*See the web collection on statistics for biologists for further resources and guidance.*
### Software

**Policy information about availability of computer code**

7. **Software**

Describe the software used to analyze the data in this study.

We used mostly ad hoc python scripts and notebooks (available via bitbucket at https://bitbucket.org/bbglab/intron_exon_mutrate). We also used the bedtools and samtools suites.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

### Materials and reagents

**Policy information about availability of materials**

8. **Materials availability**

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Not applicable.

9. **Antibodies**

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Not applicable.

10. **Eukaryotic cell lines**

   a. State the source of each eukaryotic cell line used.

   b. Describe the method of cell line authentication used.

   c. Report whether the cell lines were tested for mycoplasma contamination.

   d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

   Not applicable

### Animals and human research participants

**Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines**

11. **Description of research animals**

Provide details on animals and/or animal-derived materials used in the study.

Not applicable

**Policy information about studies involving human research participants**

12. **Description of human research participants**

Describe the covariate-relevant population characteristics of the human research participants.

Not applicable