



Germline de novo mutation rates on exons versus introns in humans

Miguel Rodriguez-Galindo ¹, Sònia Casillas ^{2,3}, Donate Weghorn ^{1,4}✉ & Antonio Barbadilla ^{2,3}✉

A main assumption of molecular population genetics is that genomic mutation rate does not depend on sequence function. Challenging this assumption, a recent study has found a reduction in the mutation rate in exons compared to introns in somatic cells, ascribed to an enhanced exonic mismatch repair system activity. If this reduction happens also in the germline, it can compromise studies of population genomics, including the detection of selection when using introns as proxies for neutrality. Here we compile and analyze published germline de novo mutation data to test if the exonic mutation rate is also reduced in germ cells. After controlling for sampling bias in datasets with diseased probands and extended nucleotide context dependency, we find no reduction in the mutation rate in exons compared to introns in the germline. Therefore, there is no evidence that enhanced exonic mismatch repair activity determines the mutation rate in germline cells.

¹Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain. ²Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), 08193 Barcelona, Spain. ³Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), 08193 Barcelona, Spain. ⁴Universitat Pompeu Fabra (UPF), Barcelona, Spain. ✉email: dweghorn@crg.eu; antonio.barbadilla@uab.es

One of the most general and widely accepted predictions of the neutral theory of molecular evolution is that “the more sequence conservation, the more functional (selective) constraint on the sequence”¹. This principle explains why different functional regions in the genome have different levels of polymorphism and divergence, such as the lower variation at nonsynonymous vs synonymous sites in protein-coding genes or in exonic vs intronic sequences². This relationship between constraint and variation constitutes one of the most powerful approaches in the current search for functional regions in the genome and the detection of natural selection at the molecular level. An integral part of estimating constraint, or purifying selection, on functional genomic regions is the comparison of the observed number of mutations to the expectation under neutral evolution. In genes, this neutral expectation is usually estimated from putatively nonfunctional regions or sites, including intronic sequence^{3,4}. A main requirement for the validation of this assumption is that mutation rate on exons and introns does not correlate with that sequence function.

Mutation rate can vary strongly across the human genome, with regional differences up to threefold in the germline⁵ and at least up to fivefold in tumor cells⁶. It is influenced by several factors, including replication time, chromatin state, and expression level^{6–8}. A priori, none of these factors are expected to correlate directly with genetic sequence function (exonic vs intronic). Another important determinant of mutation rate is DNA sequence composition. Recent studies have addressed mutational processes and their associated sequence-dependent signatures, both in the soma and the germline^{9–14}. Germline and many cancer tumor signatures exhibit a higher relative rate of C > T transitions for single nucleotide variants (SNV)^{6,9,15}. Consequently, due to their higher G/C content, exonic regions show a context-driven relative increase in mutation rate compared to intronic regions¹⁶, which can be corrected for with the proper mutational model.

A differential mutation rate between intronic and exonic DNA beyond the context dependence would require a mutational process that recognizes the difference between the two functional sequence categories. Surprisingly, Frigola et al.¹⁷ found in tumoral DNA, primarily from skin melanomas and DNA-polymerase- ϵ (POLE)-mutant colorectal cancers, that mutation rates are lower in exons than in introns after accounting for the trinucleotide-context-dependent mutational signature. This reduced mutation rate in exons is similar both in synonymous and nonsynonymous sites, which rules out purifying selection as an explanation. The study suggests that the lower mutation rate in exons results from an enhanced mismatch repair (MMR) activity in exons compared to introns. In turn, the increased repair activity is attributed to different amounts of H3K36me3 epigenetic marks on exons and introns¹⁷.

In the germline, whether originating from replication errors or mediated by DNA damage, the dominant mutational processes are expected to produce mismatches^{18–22}. Hence, any MMR-related mechanism is expected to play an important role in germline DNA damage repair. If the enhanced somatic exonic MMR activity found by Frigola et al.¹⁷ could be extrapolated to the germline, as the study suggests, then population and functional genomics studies would be compromised, and they should include differential exonic and intronic mutation rates as an integral part of their explanatory models.

Here, we investigate the relative mutation rates of exons and introns in the human germline using de novo mutation (DNM) data. We show that DNM densities do not differ between exons and introns after accounting for trinucleotide sequence composition and an excess of nonsynonymous exonic variation arising from sampling bias. We further explore factors that can impact

DNM densities on exons and adjacent introns, namely extended sequence context dependency and several chromatin features, including H3K36me3 epigenetic marks. Finally, we provide estimates of exonic and intronic DNM rates.

Results

No evidence of reduced exonic DNM rate compared to introns.

To study the distribution of germline mutations across exons and introns, we collected a total of 679,547 SNV DNMs from seven family-based WGS datasets, consolidating a high-density, high-quality DNM map across the human genome (see “Methods”). The compiled datasets show highly similar mutation spectra and an enrichment with CpG > TpG transitions (Supplementary Fig. 1). We first analyzed whether exonic and intronic mutation densities differ among human DNMs after accounting for sequence composition. For that purpose, we computed the observed total mutation burden at exonic and intronic sites by summing over 95,633 internal exon-centered sequences of size 2001 base pairs (bp), carrying a subset of 50,780 genetic mutations. Since per-nucleotide mutation probability is influenced by the neighboring sequence context, we derived the expected mutation burden at each position of each 2001-bp internal exon-centered window from a context-dependent model (see “Methods”). We initially used a trinucleotide-context-dependent germline whole-genome mutation signature model, in line with the analysis presented in Frigola et al.¹⁷ for somatic mutations.

Frigola et al.¹⁷ found that the mutation burden of POLE-mutant tumors in positions dominated by exonic DNA is lower than expected (Fig. 1a). In contrast, in our study the observed germline exonic mutation burden was significantly increased by 7.2% (s.d. 1.4%; $P = 0.001$, permutation-based test) compared to the expectation across introns and exons (Fig. 1b). This result is robust to biases in mutation calling due to region mappability differences (Supplementary Table 1, Supplementary Fig. 2) and effects of transcription-coupled repair on the mutational pattern (Supplementary Fig. 3). This suggests that the hypothesized mechanism of enhanced repair on exons compared to introns in POLE-aberrant tumors is not determining mutation rate in the germline.

Sampling bias explains exonic mutation density excess. Even in the absence of the supposed enhanced MMR effect on exonic DNA, we would expect a very slight deficit of exonic relative to intronic DNMs, due to strong purifying selection on lethal de novo variants in the early stages of embryonic development²³. Therefore, we next investigated potential factors that could explain the observed increase in exonic relative to intronic mutation burden in DNM datasets, such as technical differences in sequencing and calling (Supplementary Table 2) and enrichment with diseased probands. As detailed in Table 1, the analyzed DNM datasets are heterogeneous regarding their study conditions, including disease cohorts. Diseased probands, e.g. those with autism spectrum disorder (ASD) or preterm birth, are more likely than average to carry mutations with functional impact^{24,25}. This ascertainment bias in the data is expected to entail an enrichment with exonic nonsynonymous variants^{26,27}. To test this, we classified the 4669 mutations in the internal exons as synonymous and nonsynonymous changes, resulting in 3488 nonsynonymous and 1170 synonymous DNMs (corresponding to a ratio of 2.98:1). We then repeated the internal exon-centered analysis for each of the two exonic mutation categories.

Figure 2a shows that the observed synonymous profile matches the expected profile almost perfectly, with a slight nonsignificant deficit (−1.1%, $P = 0.353$). However, exonic

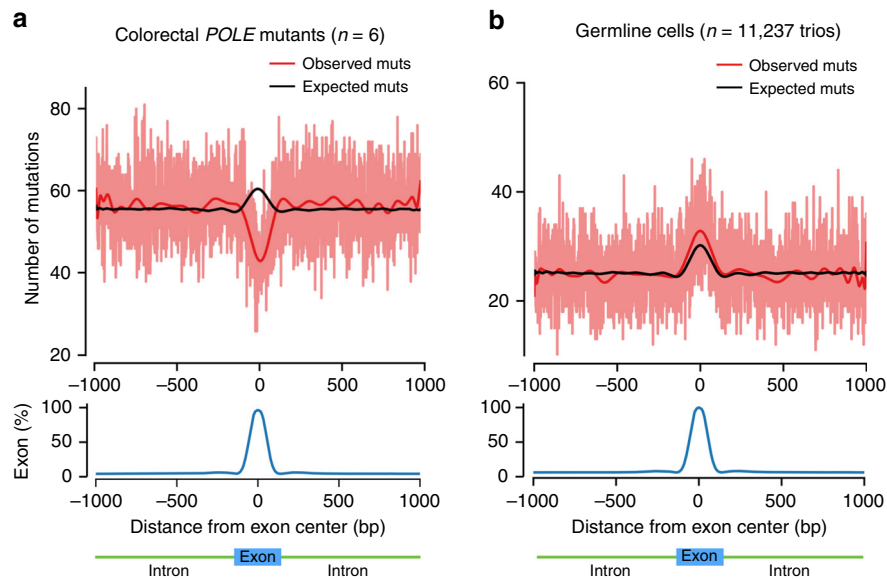


Fig. 1 Internal exon-centered analyses on somatic and germline de novo mutations. Exon-centered 2001-nt-wide observed and expected mutational profiles (top) and exon density (bottom) in **a** somatic and **b** germline cells. The light red line represents the observed number of mutations at each position, whereas the dark red and black lines represent smoothed numbers of observed and expected mutations, respectively, obtained from a polynomial fit. **a** Profile of mutations in six *POLE*-mutant colorectal tumors, reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Genetics, Reduced mutation rate in exons due to differential mismatch repair, Frigola et al.¹⁷. **b** Profile of mutations in the germline of 11,237 trios.

Table 1 Properties of analyzed DNM datasets including excess in exonic burden.

Dataset	Phenotypic condition	Control/proband DNMs	Exonic excess [%]	Emp. <i>p</i> value
Halldorsson ³⁶	Mixed (multiple diseases)	180,151	11.7 ± 2.8	0.001
Yuen ³⁸	Mostly ASD	127/117,612	8.0 ± 4.1	0.015
Goldmann ³⁵	Mixed (preterm birth)	35,793	13.0 ± 6.6	0.016
Sasani ³⁹	Random (mostly healthy)	27,454/0	12.3 ± 7.9	0.052
An ³⁷	ASD + healthy sibling	115,697/117,942	3.5 ± 2.2	0.070
GoNL ¹⁶	Random (mostly healthy)	11,016/0	12.1 ± 14.5	0.181
Goldmann ²⁹	Healthy	73,755/0	2.5 ± 4.4	0.291
Autism probands	ASD	0/235,554	8.4 ± 2.5	0.001
Representative sample	Non-ASD + ASD	98,300/1,700	−2.5 ± 3.5	0.233
Healthy probands	Healthy	189,579/0	−0.5 ± 2.5	0.417

The seven used datasets and their references are shown above the line, while below the results for the composed datasets (see “Methods”) are given. In each group, datasets are ordered from most to least significant exonic mutation excess. Errors of the exonic excess denote 1 s.d. from 1000 permutations (see “Methods”).

nonsynonymous mutations show a large and statistically significant excess compared to the expectation under the trinucleotide-context model (10.4%, $P = 0.001$, Fig. 2b). Note that this stratification by functional mutation category entails a reduction of the number of both synonymous and nonsynonymous mutations relative to flanking introns across the window of stacked sequences. This is why the number of mutations, when moving outwards from the center, converges to that of Fig. 1b. Overall, Fig. 2 suggests that disease ascertainment during data acquisition may be responsible for the overall excess of 7.2% of exonic variants. Therefore, we next repeated the analysis for all seven DNM datasets individually, as well as for assembled samples with only healthy, only ASD, or a representative mixture of probands (see “Methods”). We found a significant exonic mutation excess only in cohorts with a high fraction of diseased probands or those assembled purely from diseased samples, while all cohorts with mostly or exclusively healthy probands show no signal (Table 1). Moreover, when we stratify exonic mutations by functional impact, we find a statistically significant excess only in nonsynonymous mutations among ASD individuals (Supplementary Table 3).

De novo variants show extended-context dependency. The stratification into synonymous and nonsynonymous changes entailed a polarization of the exonic excess (Fig. 2), intensifying the signal for nonsynonymous variants (10.4%) with a concomitant decrease for synonymous variants (−1.1%). This type of polarization could be due to the incompleteness of our mutational model. Our mutational model for the expected number of mutations was constructed using trinucleotide-context-dependent mutation probabilities. However, it has been shown that SNPs segregating in the human population are affected by the extended flanking sequence, with a heptameric context explaining a majority of the observed mutation rate variability^{13,14}. Figure 3 and Supplementary Figs. 4, 5 show that this is confirmed by DNMs based on the relative frequencies of all four nucleotides around mutations in our DNM dataset, although the effect of the extended flanking sequence is small compared to the one observed in *POLE*-mutated tumor genomes²⁸.

We assessed the impact of context dependency by expanding our mutational signature model to incorporate the pentameric and heptameric mutational sequence context (based on exact computation and a likelihood decomposition approach, respectively; see

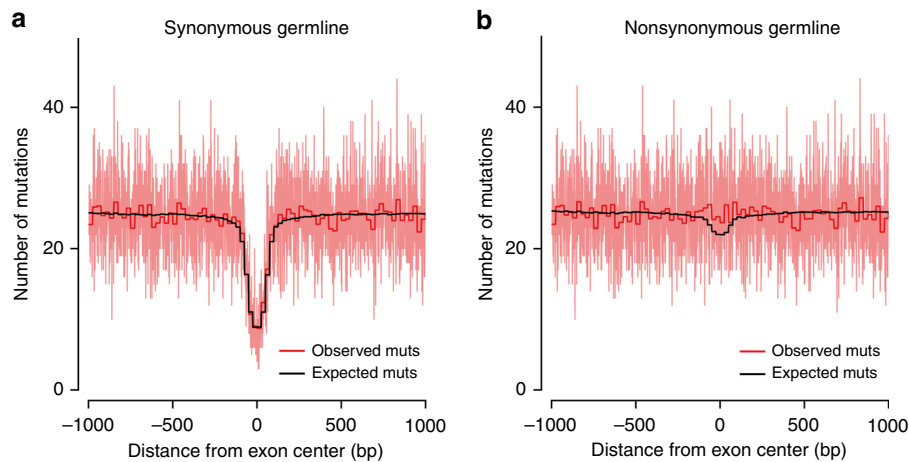


Fig. 2 Internal exon-centered analysis for synonymous and nonsynonymous DNMs. The light red line represents the observed number of mutations at each nucleotide position, while the dark red and black lines represent averages in bins of size 25 positions for observed and expected mutations, respectively. **a** Synonymous DNM profile with observed exonic mutation difference of -1.1% ($P = 0.353$). **b** Nonsynonymous DNM profile, showing an exonic excess of 10.4% ($P = 0.001$). Due to the removal of nonsynonymous and synonymous mutations in panels **(a)** and **(b)**, respectively, the total number of exonic mutations relative to flanking introns is reduced. The number of mutations, when moving away from the center, converges to that of Fig. 1b.

Table 2 Extended sequence context dependency for Goldmann et al.²⁹.

Model	Exonic excess [%]	Emp. p value	Log-likelihood	# param	AIC
1-mer	16.2 ± 5.4	0.001	-68,111	12	136,247
CpG	1.7 ± 4.3	0.348	-66,594	18	133,224
3-mer	2.7 ± 4.3	0.281	-66,315	192	133,014
5-mer	2.9 ± 4.6	0.258	-66,054	1344	134,797
7-mer	2.8 ± 4.3	0.266	-65,981	2496	136,955

Errors of the exonic excess denote 1 s.d. from 1000 permutations (see “Methods”).

“Methods”, Supplementary Fig. 6). We applied these extended-context models to the largest DNM dataset that had no diseased probands, Goldmann et al.²⁹, as well as to this dataset and the pooled dataset stratified by synonymous and nonsynonymous variants. We found that while the overall likelihood increases for increasing context size, penalization due to the additional parameters of the larger context models entails that the trinucleotide-context-dependent model is found to be the best model for the current datasets based on the Akaike information criterion (AIC) (Table 2, Supplementary Tables 4–6).

H3K36me3 does not correlate with exonic mutation density.

The enhanced exonic MMR activity compared to introns in *POLE*-aberrant tumors was proposed to be mediated by the H3K36me3 mark¹⁷. Using our dataset of mutations from healthy probands, we therefore investigated the relative exonic mutation density as a function of H3K36me3 and nucleosome density. These two features show differential coverage between (mainly internal) exons and introns (Supplementary Fig. 7), and had been previously described to contribute to the recognition of splice marks at internal exon–intron boundaries^{30,31}. We observed no significant correlation ($r = -0.03$, $P = 0.84$) between the exonic mutation enrichment and the H3K36me3 mark (Fig. 4b). This contrasts with the recruitment mechanism described in somatic cells³², which is invoked as the mechanistic hypothesis behind the findings in Frigola et al.¹⁷ (Fig. 4a). Conversely, we find a negative, nearly significant correlation ($r = -0.28$, $P = 5.38 \times 10^{-2}$) with nucleosome coverage (Supplementary Fig. 8). This result

complements the previously reported influence of nucleosome organization on human germline DNMs^{5,33}.

Estimation of exonic and intronic de novo mutation rate. We estimated germline DNM rates for exons and introns separately. Using the largest dataset with only healthy probands²⁹, we estimated 1.38×10^{-8} and 1.11×10^{-8} mutations per site per generation for exons and introns, respectively. This difference reflects the higher mutability of exons with respect to introns due to sequence differences, namely higher CpG and overall GC content¹⁶. These estimates are consistent with a previously reported whole-genome based rate of 1.2×10^{-8} mutations per site per generation²⁰. Conversely, estimates obtained from the pooled dataset reflect the disease ascertainment bias, with a similar intronic mutation rate (1.15×10^{-8}), but a much larger exonic rate (1.52×10^{-8}), in line with previous findings in diseased cohorts³⁴.

Discussion

We compiled human DNM data to show that the rate of generation of new genetic variants, the mutation rate, does not significantly vary between exons and adjacent introns when accounting for sequence context. Moreover, we went beyond previous analyses that used extreme rare variants as a proxy for DNMs^{13,14} and described directly germline mutation patterns based on a large aggregated DNM dataset. We corroborated earlier findings, in particular an extended-context dependence of germline variants. At the same time, the internal exon-centered analysis, with its relatively low number of mutations compared to the entire dataset, is still adequately described by a trinucleotide-context-dependent model. Beyond context dependence, the sampling bias introduced by enrichment with diseased probands is one of the most important confounding factors of DNM analyses. We showed that its effects can lead to significant deviations from the null model and, depending on the application, should be addressed through an informed choice of samples.

Our analysis shows that the results found in the soma cannot be directly extrapolated to the germline, and the MMR-dependent process that was proposed as an explanation for the decreased exonic mutation burden in somatic cells does not seem to determine germinal cell mutation rates. Last, this study provides a

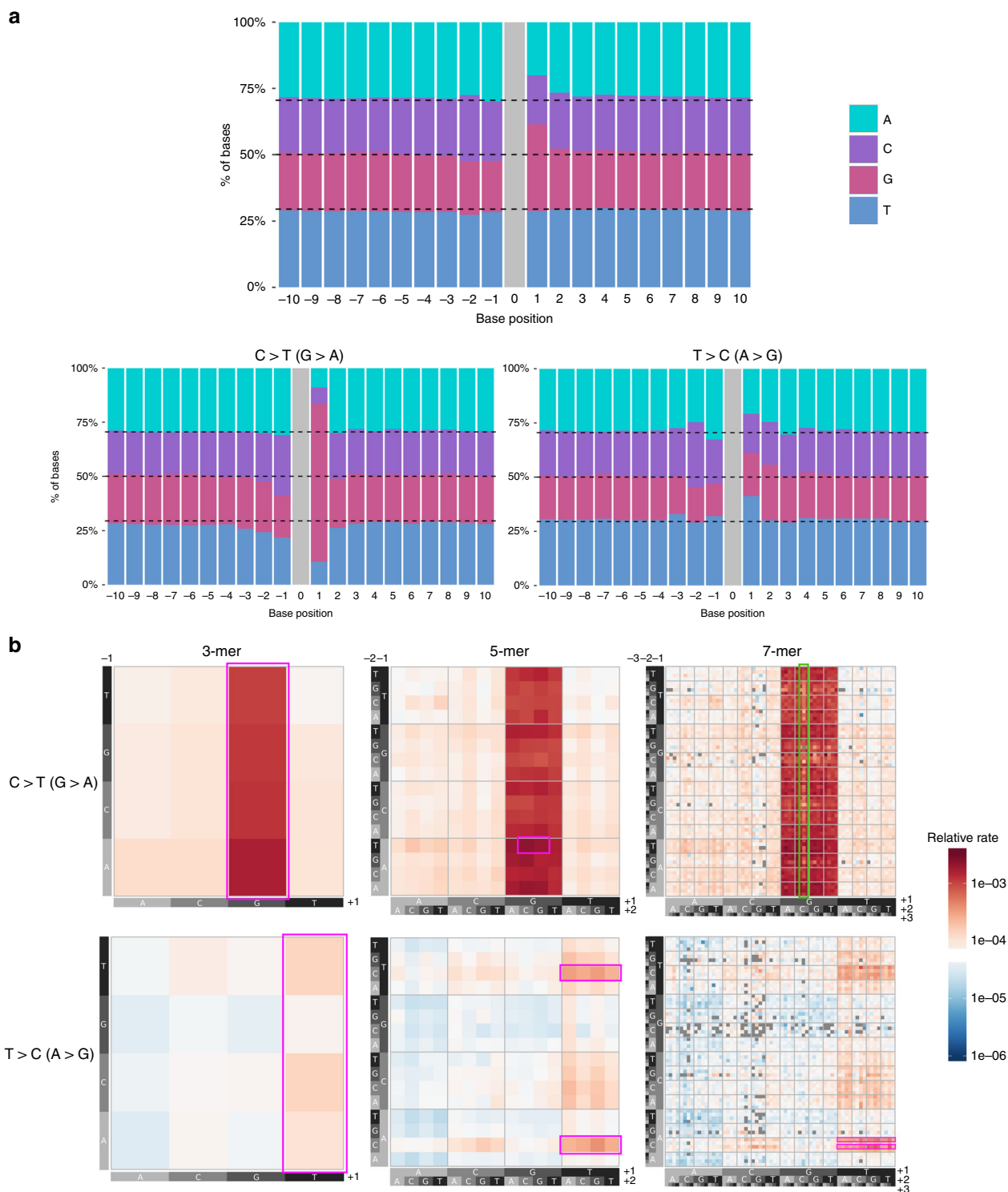


Fig. 3 Mutational classes with highest sequence context dependency around de novo mutations. **a** Frequencies of nucleotides neighboring our entire set of 679,547 DNMs and subsets belonging to C > T (G > A) and T > C (A > G) 1-mer classes in a window of size 21 bp. Black dashed lines represent the whole-genome background frequencies for the four nucleotides. The extended sequence context dependency varies across 1-mer mutation classes. **b** Heatmap of estimated relative mutation rates, corrected by abundance of reference k-mer, for C > T (G > A) (top) and T > C (A > G) (bottom) 1-mer classes, up to a 7-mer resolution. For each 1-mer class, each of the three grids delineates mutation contexts of different length, defined by the upstream sequence (y-axis) and downstream sequence (x-axis) from the central (mutated) nucleotide. Boxed regions indicate motifs previously identified as hypermutable (pink) or hypomutable (green). Supplementary Figs. 4 and 5 show the corresponding plots for the other 1-mer classes.

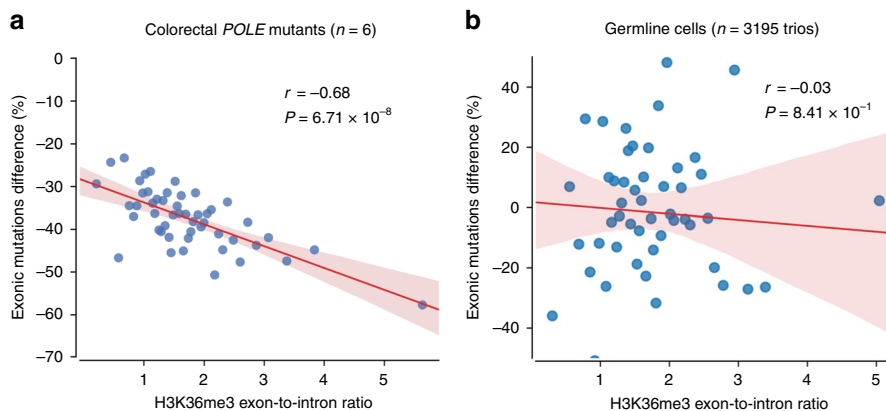


Fig. 4 Deviation in the exonic mutation burden as a function of the H3K36me3 exon-to-intron ratio. Blue dots denote 50 groups of genes binned by their exon-to-intron ratio of H3K36me3 coverage (*x*-axis). The relative difference between the total observed and expected number of exonic mutations (computed using a 3-mer model) per group is shown on the *y*-axis. The trend line and its confidence interval were added using the seaborn package of Python, while the correlation coefficient and its significance were computed using the same iteratively re-weighted least-squares approach as used by Frigola et al.¹⁷ to ensure comparability. **a** *POLE*-mutant colorectal tumors. The H3K36me3 histone mark is derived from colonic mucosa (E075), reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Genetics, Reduced mutation rate in exons due to differential mismatch repair, Frigola et al.¹⁷. **b** DNMs from healthy probands (a total of 3195 trios). The H3K36me3 histone mark is derived from H1 stem cells (E003).

clear-cut answer to the challenge posed by Frigola et al.¹⁷: It validates the main assumption of molecular population genetics—that genomic mutation rate does not depend on sequence function—and demonstrates different mutational dynamics in somatic vs germinal cells.

Methods

De novo mutation data. We aggregated DNMs from seven family-based WGS datasets coming from multiple centers and projects: the Genomes of the Netherlands (GoNL) project¹⁶, the Inova Translational Medicine Institute Preterm Birth Study³⁵, Inova Translational Medicine Institute’s Longitudinal Childhood Genome Study²⁹, deCODE genetics³⁶, Simons Simplex Collection (SCC) and Korean ASD cohort³⁷, the Autism Genetic Research Exchange (AGRE) repository³⁸ and Centre d’Etude du Polymorphisme Humain (CEPH)³⁹. Mutation datasets were downloaded from the supplementary tables of the respective papers^{35–39} or by direct request to the authors²⁹. Data from the GoNL were downloaded from <http://www.nlgenome.nl>. Most of datasets were originally mapped to hg19, with exception of the data from An et al.³⁷ and Halldorson et al.³⁶, which were mapped to hg38. Subsequently, coordinates in these datasets were lifted over to hg19, the most common reference genome in our data. To avoid possible biases arising from mutation calling on sexual chromosomes, only autosomal SNVs were used, leaving a total of 679,547 germline SNV DNMs coming from 11,237 trios.

Effect prediction of de novo mutations. The predicted consequence class of all DNMs was obtained using the Ensembl Variant Effect Predictor (VEP)⁴⁰ for the GRCh37/hg19 assembly. Since some DNMs were reported to have more than one consequence, e.g. different transcripts or overlapping genes, only one predicted consequence for each DNM was retrieved (according to VEP criteria). Predictions are classified from major to mild according to the Ensembl Variation hierarchy.

Whole-genome de novo mutation spectrum. All mutations were divided into nine classes, considering the fact that CpG sites are highly mutagenic. The number of mutations were corrected by the relative abundance of the context in the whole genome, e.g. the total number of C > T (G > A) transitions occurring at CpG sites divided by the relative abundance of CpG sites in the genome. We performed the mutational analysis across all used studies (Supplementary Fig. 1). For all subsequent analyses, extended nucleotide-context-dependent mutational models were used.

Genomic coordinates of internal exons and flanking introns. Coordinates for a total of 20,345 protein-coding genes were obtained from GENCODE v19⁴¹. Genes without introns and overlapping genes were discarded, leaving a filtered set of 13,474 genes. Genes located on chromosomes X, Y and on the mitochondrial genome were removed from the analysis, leaving a total of 12,754 autosomal genes. Finally, all transcripts per gene were merged into meta-exon and meta-intron coordinates, both 5’ and 3’ flanking exons were removed as well as UTRs. Only internal exons (unfiltered by mappability issues, see below) were used for the main internal exon-centered mutational analyses.

Moreover, positions where mutation calling would be technically challenging because of mappability issues were removed, leaving a total of 10,237 genes for the gene by gene analyses. We also filtered out the internal exon-centered 2001-nt windows that overlapped at least one nucleotide with regions with mappability issues (Supplementary Table 1) for the supplementary internal exon-centered analysis restricted to highly mappable regions (Supplementary Fig. 2). Coordinates of unreliable regions⁴² were obtained from the UCSC Genome Browser, available at <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability>.

Meta-exon and meta-intron coordinates of genes with at least five meta-exons (not only internal) were extracted from GENCODE v19 for the analyses with chromatin features across genic regions (Supplementary Fig. 7).

Sequence context model. We directly computed the probability of a mutation into an alternative nucleotide, H_a where $a \in \{1, 2, 3\}$, given the reference nucleotide H_r and its flanking sequence $\mathbf{X} = (\mathbf{X}_5', \mathbf{X}_3')$, where \mathbf{X}_5' and \mathbf{X}_3' are the 5’ and 3’ flanking sequences, respectively. Here, $H_r \in \mathbf{M} = \{A, C, G, T\}$, where the latter denote nucleotides adenine, cytosine, guanine, and thymine, and $H_a \in \mathbf{M}' = \{m \in \mathbf{M}, m \neq H_r\}$. For example, for the 5-mer GCACG > GCTCG mutation $H_r = A, H_a = T, \mathbf{X}_5' = (G, C), \mathbf{X}_3' = (C, G)$ and $\mathbf{X} = (G, C, C, G)$. Therefore, the probability of each of the possible *k*-mer changes, normalized by the abundance of each reference *k*-mer in the genome, was computed as follows:

$$P(H_a|H_r, \mathbf{X}) = \frac{N(H_a, H_r, \mathbf{X})}{G(H_r, \mathbf{X})}, \tag{1}$$

where $N(H_a, H_r, \mathbf{X})$ is the genome-wide number of observed mutations into alternate allele H_a with given reference allele H_r and flanking sequence \mathbf{X} . $G(H_r, \mathbf{X})$ is the abundance of the reference *k*-mer with reference nucleotide H_r and flanking sequence \mathbf{X} in the genome. We computed the relative abundance of each reference *k*-mer in the autosomal genome using the pyFasta package. We also computed strand-wise signatures restricted to mutations falling in genic regions (exons at the canonical CDS and the respective introns) for the supplementary analysis in Supplementary Fig. 3. To compute strand-wise signatures, we polarized mutations according to the transcription strand on which the canonical CDS is annotated.

For some *k*-mer models, given limitations imposed by the amount of DNMs, we used a decomposition approach to compute the probability. For a *k*-mer model of sequence length *k*, let H_r be a reference core *h*-mer of length *h* and H_a an alternate core *h*-mer, where $1 \leq h < k$. Let the tuple $\mathbf{X} = (x_1, \dots, x_g)$ with $g = (k - h)$ elements represent again the flanking sequence of the core *h*-mer, where $x_i \in \mathbf{M} \forall i \in \{1, \dots, g\}$. In other words, $\mathbf{X} \in \mathbf{M}^g$ where \mathbf{M}^g is the *g*-fold Cartesian product. For example, with $k = 7, h = 3$ and the mutation ACTGACT > ACTCACT, then $H_r = TGA, H_a = TCA$ and $\mathbf{X} = (x_1 = A, x_2 = C, x_3 = C, x_4 = T)$. We then approximate the probability $P(H_a|H_r, \mathbf{X})$ by:

$$P(H_a|H_r, \mathbf{X}) \approx P(H_a|H_r) \cdot \prod_{i=1}^g \frac{P_i(H_a|H_r, x_i)}{P(H_a|H_r)}, \tag{2}$$

with

$$P(H_a|H_r) = \frac{\sum_{\mathbf{Z} \in \mathbf{M}^k} N(H_a, H_r, \mathbf{Z})}{\sum_{\mathbf{Z} \in \mathbf{M}^k} G(H_r, \mathbf{Z})}, \tag{3}$$

and

$$P_i(H_a|H_r, x_i) = \frac{\sum_{Z \in Y_{x_i}} N(H_a, H_r, Z)}{\sum_{Z \in Y_{x_i}} G(H_r, Z)}, \quad (4)$$

where

$$Y_{x_i} = \left\{ (y_1, \dots, y_j, \dots, y_g) \quad \forall j \in \{1, \dots, g\} \begin{cases} \text{if } j = i, y_j = x_i \\ \text{otherwise, } y_j \in \mathbf{M} \end{cases} \right\}. \quad (5)$$

We implemented this framework using custom Python code. The composite likelihood model was applied to the 7-mer analysis of the data pooled across all cohorts using $k = 7$ and $h = 5$. Also in the analysis of the largest single dataset purely composed of healthy probands²⁹ and the largest single dataset³⁶, in each for 5-mers with $k = 5$ and $h = 3$ and for 7-mers with $k = 7$ and $h = 3$. The rest of probabilities were computed using the direct approach. Supplementary Fig. 6 shows the relationship between the exact computations of mutational probabilities and the composite likelihood model for the pooled dataset.

The number of parameters for the direct approach increases exponentially as k increases following $f(x) = 4^x \cdot 3$, where $x = k$. For the decomposition approach they increase linearly as $k - h$ increases from a fixed h following $g(y, x) = f(x) \cdot (1 + 3(y - x))$, where $y = k$ and $x = h$.

Comparison of sequence context models. We selected the sequence context dependency model that best explains mutations across our set of exonic and intronic sequences by means of the AIC. We interrogated each of the 191,361,633 (~6.4% of the whole-genome length) exonic and intronic sites on the 95,633 2001-nt windows for the state in the observed data: mutated (and type) or not mutated. For recurrent sites, we chose one observed mutation at random. For a given model, we computed the log-likelihood as the sum across sites of the logarithm of the estimated probability of the observed state at the site. Probabilities were estimated with mutations from the entire dataset, through the direct or the decomposition approach as stated above.

Internal exon-centered mutational analysis. A total of 95,633 stacked 2001-nt sequences centered on the middle position of internal meta-exons were used to compare the observed and expected mutational profiles across exons and introns. We computed the frequency of mutation at a site l with reference core sequence H_r^l and flanking sequence \mathbf{X}^l as

$$f_l = \sum_{a=1}^3 P(H_a^l|H_r^l, \mathbf{X}^l), \quad l \in \{1, \dots, L\}, \quad (6)$$

where $L = 2001$ denotes the total number of considered sites. Then, each frequency was normalized by the total frequency on the sequence:

$$f_l^{\text{resc}} = \frac{f_l}{\sum_{l=1}^L f_l}. \quad (7)$$

Finally, the total number of observed mutations n_s on each of the 2001-nt sequence s , of a total of $S = 95,633$ stacked sequences, was redistributed across both middle exonic and flanking intronic sites according to the normalized frequencies:

$$\hat{n}_s^l = f_l^{\text{resc}} \cdot n_s, \quad s \in \{1, \dots, S\}, \quad (8)$$

thus yielding the expected number of mutations at site l of a given sequence s . By adding up the values of all the stacked sequences, we obtain the cumulative number of expected mutations at site l ,

$$\hat{n}^l = \sum_{s=1}^S \hat{n}_s^l. \quad (9)$$

For the internal exon-centered analysis on synonymous or nonsynonymous mutations, we separated all possible exonic mutations in middle exon sequences into two groups: those with synonymous consequence and those with a consequence ranking higher than synonymous in the Ensembl Variation hierarchy. Then we computed the expected numbers by only adding frequencies for either synonymous or nonsynonymous mutations.

Computation of effect size and statistical significance. We performed 1000 random permutations of the observed mutations in each stacked sequence based on the probability of each site to acquire a mutation. The effect size, defined as the relative increase or decrease in observed exonic mutations with respect to the expected number, was computed based on the simulation mean expected value. The error of this estimate is given as one standard deviation derived from the 1000 permutations. Moreover, we computed an empirical one-sided p value as the fraction of the simulations with more (or fewer) exonic mutations than the observed number of exonic mutations.

Composed datasets. Mutations were only resampled from datasets with known conditions of the probands, either from healthy probands or those with ASD. Given an ASD prevalence in humans of 1.7%, we created a random sample of 100,000 whole-genome DNMs, taking 98,300 mutations classified as strictly from

healthy probands and 1700 classified as ASD and repeated the internal exon-centered analysis. To generate the purely healthy and purely ASD cohorts, we used solely mutations from probands with the respective condition.

Nucleosome and H3K36me3 histone mark data. We downloaded narrow peak coordinates and genome-wide read-coverage of H3K36me3 from human embryonic stem cell H1-hESC (E003), as proxy for germline cells, from the Epigenome Roadmap consortium⁴³ data portal (<http://www.roadmapepigenomics.org/data>). The genome-wide nucleosome positioning density graph of ENCODE⁴⁴ cell line GM12878 (lymphoblastoid cell line) was obtained via the UCSC genome browser (<https://hgdownload.soe.ucsc.edu/downloads.html>). Nucleosome peak regions were identified across the genome by using the bwtool program (with parameters local-extrema -maxima -min-sep = 150). The window of 146 bp flanking the peak coordinate (73 bp per side) was considered the region covered by a nucleosome.

Coverage of chromatin features across exons and introns. Exons and introns in each gene were classified according to their position with respect to the transcription start site, where the ones that occupy different positions in different transcripts were discarded. We also discarded exons and introns at the lower quartile of length to compute the coverage for a set of exons or introns of heterogeneous lengths in a given position: the fraction of bases covered by H3K36me3 and nucleosomes at the center of the stack corresponding to the window defined by the shortest exon or intron remaining after the filtering. Finally, the difference between the exonic and intronic coverage was computed via the two-tailed Mann-Whitney p value of the comparison of both distributions.

We also computed the positions in the genome covered by H3K36me3 or nucleosomes across 95,633 internal exon-centered 4001-nt windows. By stacking sequences, we obtained middle exon-centered profiles of coverage across exons and introns (Supplementary Fig. 7).

Nucleosome and H3K36me3 binned gene analysis. For each gene, we computed the readcount-based exonic enrichment of H3K36me3 or nucleosomes as the ratio between the exonic and intronic total number of bases covered by reads of the chromatin feature. Genes with no exonic and intronic bases covered by reads were removed from the analysis, as well as genes without any observed exonic or intronic mutation. Thus, a total of 7215 and 6529 genes remained for the H3K36me3 and nucleosome analysis, respectively.

For a given gene, we computed the exonic expected number of mutations as follows:

$$\hat{n}_e = P_e \cdot n, \quad (10)$$

where n is the total number of mutations (both exonic and intronic) observed on the gene and P_e is the (binomial) probability of a mutation to fall on the exonic region of the gene, which in turn is computed as:

$$P_e = \frac{\mathcal{L}_e}{\mathcal{L}_e + \mathcal{L}_i}, \quad (11)$$

with

$$\mathcal{L}_e = \sum_{l_e=1}^{L_e} \sum_{a=1}^3 P(H_a^l|H_r^l, \mathbf{X}^l), \quad (12)$$

and

$$\mathcal{L}_i = \sum_{l_i=1}^{L_i} \sum_{a=1}^3 P(H_a^l|H_r^l, \mathbf{X}^l). \quad (13)$$

Here, $l_e \in \{1, \dots, L_e\}$ and $l_i \in \{1, \dots, L_i\}$ denotes the set of all exonic and intronic positions of a given gene, respectively. \mathcal{L}_e and \mathcal{L}_i represent the exonic and intronic target size, respectively, expressed as the sum of the probability $P(H_a^l|H_r^l, \mathbf{X}^l)$ of all possible three mutations that can happen across all exonic or intronic sites of the gene. The probability was computed under a 3-mer model for each of the genes as explained above only with mutations from the composed dataset of healthy probands.

Afterwards, genes were grouped into 50 bins according to their exonic enrichment of H3K36me3 or nucleosomes. Then, with the observed n_e and expected \hat{n}_e exonic mutations over all genes in the bin, we computed the relative difference between the observed and expected number of exonic mutations per bin as follows:

$$\text{Exonic mutations difference [\%]} = \frac{n_e - \hat{n}_e}{\hat{n}_e} \cdot 100. \quad (14)$$

Finally, we computed the correlation between the median exonic chromatin feature enrichment and the difference in exonic mutations across the bins. The trend line and its confidence intervals were added using the bootstrapping functions of the python seaborn package, which confers equivalent weights in the regression to all points. In order to guarantee that the trend is not the result of a few outliers, the correlation coefficient and its significance were computed using an

iteratively re-weighted least-squares approach, letting the variance of exonic chromatin feature enrichment of the bins influence the weight of each point.

Estimation of absolute mutation rates. We estimated absolute mutation rate in our set of 95,633 middle exons as a proxy of mean exonic mutation rate and absolute mutation rate on the rest of the 2001-nt window as a proxy of mean intronic mutation rate as follows:

$$\mu = \frac{N_{\text{obs}}}{N_{\text{site}} \cdot N_{\text{gen}}} \quad (15)$$

Here, μ is the mutation rate per site and generation, N_{obs} is the number of observed mutations, N_{site} is the number of sites (we used $N_{\text{site}} = 13,632,264$ exonic sites and $N_{\text{site}} = 177,729,369$ flanking intronic sites) and N_{gen} is the number of generations. A total of $N_{\text{gen}} = 2582$ gametogeneses for the largest dataset with healthy probands²⁹ and $N_{\text{gen}} = 22,474$ gametogeneses in the pooled dataset.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All the analyses in this study were based on published datasets. Mutation data from the Genomes of the Netherlands (GoNL) project¹⁶ was downloaded from (<http://www.nlgenome.nl>). The remaining mutation datasets were either by direct request to the authors²⁹ or downloaded from the supplementary tables of their respective publications^{35–39}. Coordinates of unreliable regions⁴² were obtained from the UCSC Genome Browser, available at <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability>. Narrow peak coordinates and genome-wide read-coverage of H3K36me3 from human embryonic stem cell H1-hESC (E003) were downloaded through the Epigenome Roadmap consortium⁴³ data portal (<http://www.roadmapepigenomics.org/data>). The genome-wide nucleosome positioning density graph of ENCODE⁴⁴ cell line GM12878 (lymphoblastoid cell line) was obtained via the UCSC genome browser (<https://hgdownload.soe.ucsc.edu/downloads.html>).

Code availability

Custom scripts and associated files needed to reproduce all analyses described here are provided together with the code at https://bitbucket.org/weghornlab/germline_intron_exon_mutrate. Segments of the code are based on scripts shared by Frigola et al.¹⁷ to ensure comparability of the main analysis.

Received: 7 January 2020; Accepted: 2 June 2020;

Published online: 03 July 2020

References

- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
- Castle, J. C. SNPs occur in regions with less genomic sequence conservation. *PLoS ONE* **6**, e20660 (2011).
- Mikkelsen, T. S. et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69 (2005).
- 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).
- Smith, T. C., Arndt, P. F. & Eyre-Walker, A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet.* **14**, 1–30 (2018).
- Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
- Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393 (2009).
- Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504 (2012).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
- Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
- Carlson, J. et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **6**, 3753 (2018).
- Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48**, 349–355 (2016).

- Cooper, D. N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**, 151–155 (1988).
- Francioli, L. C. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
- Frigola, J. et al. Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).
- Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
- Ozturk, S. & Demir, N. DNA repair mechanisms in mammalian germ cells. *Histol. Histopathol.* **26**, 505–517 (2011).
- Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 1–8 (2017).
- Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).
- García-Rodríguez, A., Gosálvez, J., Agarwal, A., Roy, R. & Johnston, S. DNA damage and repair in human reproductive cells. *Int. J. Mol. Sci.* **20**, 1–22 (2019).
- Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
- Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237 (2012).
- Li, J., Oehlert, J., Snyder, M., Stevenson, D. K. & Shaw, G. M. Fetal de novo mutations and preterm birth. *PLoS Genet.* **13**, e1006689 (2017).
- Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216 (2014).
- Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504 (2017).
- Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
- Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* **16**, 996–1001 (2009).
- Luco, R. F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
- Li, F. et al. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSa. *Cell* **153**, 590–600 (2013).
- Li, C. & Luscombe, N. M. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat. Commun.* **11**, 1–13 (2020).
- Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242 (2012).
- Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).
- Halldórsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
- An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
- Yuen, R. K. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
- Sasani, T. A. et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife* **8**, 1–24 (2019).
- McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
- Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
- Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

Acknowledgements

We thank especially Núria López-Bigas for her insight and suggestions during the development of this work. We also thank her research group for sharing the software code in all their contributions, especially their Frigola et al.¹⁷ paper. We would also like to acknowledge Jakob Goldmann, Christian Gilissen, and Wendy Wong for sharing high-quality mutation data. We are grateful to David Castellano for helpful discussions and Vladimir Seplyarskiy for a critical reading of the manuscript. The project that gave rise to these results received the support of a fellowship from “la Caixa” Foundation (ID 100010434) with fellowship code LCF/BQ/DR19/11740019 (M.R.-G.); by the Ministerio de Economía y Competitividad (Spain) (CGL2017-89160P to M.R.-G. and A.B.); and AGAUR (Generalitat de Catalunya) (2017SGR-1379 to A.B.). We acknowledge the support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the

Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya (M.R.-G. and D.W.).

Author contributions

M.R.-G. performed the analyses. M.R.-G. and D.W. designed the analyses. M.R.-G., D.W., and A.B. wrote the manuscript. D.W., S.C., and A.B. supervised the study. D.W. and A.B. conceived the study. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-17162-z>.

Correspondence and requests for materials should be addressed to D.W. or A.B.

Peer review information *Nature Communications* thanks Edward Hollox and Wendy Wong for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020