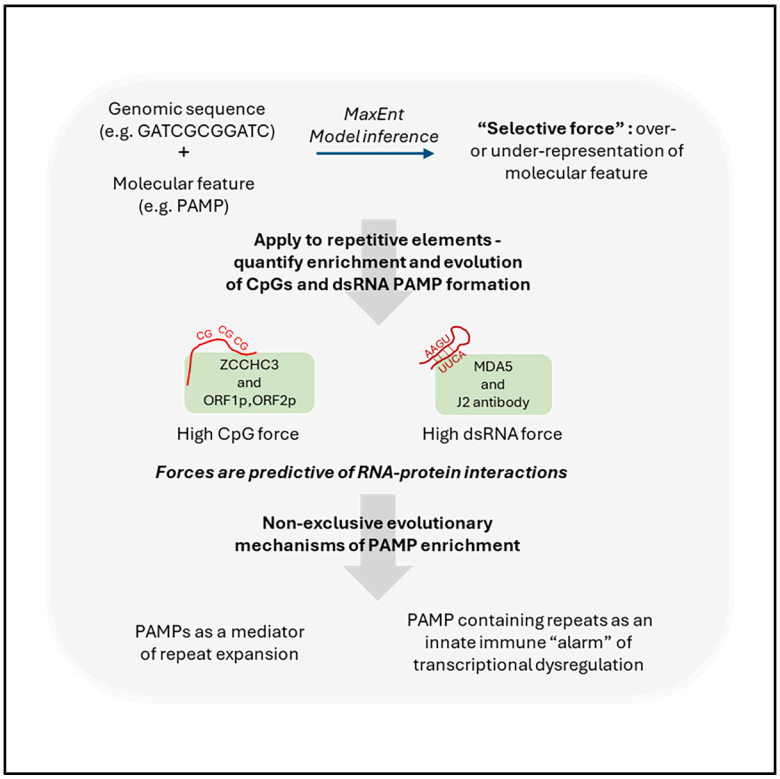
Cell Genomics

Article

Repeats mimic pathogen-associated patterns across a vast evolutionary landscape

Graphical abstract



Highlights

- Statistical physics framework for viral mimicry in genomes using "selective forces"
- Repeats that bind different pattern recognition receptors can be predicted
- Mimicry across families suggests shared mechanisms of emergence and retention
- We suggest two non-exclusive evolutionary hypotheses related to PAMPs and viral mimicry

Authors

Petr Šulc, Andrea Di Gioacchino, Alexander Solovyov, ..., Rémi Monasson, Simona Cocco, Benjamin D. Greenbaum

Correspondence

psulc@asu.edu (P.Š.), andrea.dgioacchino@gmail.com (A.D.G.), jlacava@rockefeller.edu (J.L.), daniel.decarvalho@uhn.ca (D.D.D.C.), simona.cocco@phys.ens.fr (S.C.), greenbab@mskcc.org (B.D.G.)

In brief

An emerging hallmark of disease is transcription of pathogen-associated molecular patterns from within the genome–known as viral mimicry. We propose a statistical physics framework to measure "selective forces" that enrich mimicry. We validate our predictions and show mimicry across eukaryotes. We put forth shared mechanisms driving emergence and evolution.

Šulc et al., 2025, Cell Genomics 5, 101011
December 10, 2025
https://doi.org/10.1016/j.xgen.2025.101011

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Repeats mimic pathogen-associated patterns across a vast evolutionary landscape

Petr Šulc,[1,2,19,*] Andrea Di Gioacchino,[3,19,*] Alexander Solovyov,[4,19] Siyu Sun,[4,20] Stephen Martis,[4,20] Sajid A. Marhon,[5,20] Håvard T. Lindholm,[5,18,20] Raymond Chen,[5] Amir Hosseini,[5,6] Hua Jiang,[7] Bao-Han Ly,[8] Martin S. Taylor,[9,10,11] Parinaz Mehdipour,[5,6] Omar Abdel-Wahab,[12,13] Nicole Rusk,[4] Nicolas Vabret,[14] John LaCava,[7,21,22,*] Daniel D. De Carvalho,[5,15,21,22,*] Rémi Monasson,[3,21,22] Simona Cocco,[3,21,22,*] and Benjamin D. Greenbaum[4,16,17,21,22,23,*]

[1]School of Molecular Sciences and Center for Molecular Design and Biomimetics, The Biodesign Institute, Arizona State University, Tempe, AZ, USA
[2]School of Natural Sciences, Department of Bioscience, TU Munich, Garching, Germany
[3]Laboratoire de Physique de l'Ecole Normale Supérieure, PSL & CNRS, Sorbonne Université, Université de Paris, Paris, France
[4]The Halvorsen Center for Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[5]Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1L7, Canada
[6]Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, OX3 7DQ Oxford, UK
[7]Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, NY, USA
[8]European Research Institute for the Biology of Ageing, University Medical Center Groningen, Groningen, the Netherlands
[9]Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA
[10]Brown Center on the Biology of Aging, Brown University, Providence, RI, USA
[11]Legorreta Cancer Center, Brown University, Providence, RI, USA
[12]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[13]Leukemia Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[14]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[15]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
[16]Physiology, Biophysics & Systems Biology, Weill Cornell Medicine, Weill Cornell Medical College, New York, NY, USA
[17]The Olayan Center for Cancer Vaccines, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[18]Present address: Department of Pathology, Oslo University Hospital - Rikshospitalet, 0372 Oslo, Norway
[19]These authors contributed equally
[20]These authors contributed equally
[21]These authors contributed equally
[22]Senior author
[23]Lead contact
*Correspondence: psulc@asu.edu (P.Š.), andrea.dgioacchino@gmail.com (A.D.G.), jlacava@rockefeller.edu (J.L.), daniel.decarvalho@uhn.ca (D.D.D.C.), simona.cocco@phys.ens.fr (S.C.), greenbab@mskcc.org (B.D.G.)
https://doi.org/10.1016/j.xgen.2025.101011

SUMMARY

An emerging hallmark of many human diseases is transcription of typically silenced repetitive DNA containing pathogen-associated molecular patterns (PAMPs). These PAMPs engage the innate immune system via pattern recognition receptors (PRRs)—a phenomenon known as viral mimicry. We propose a statistical physics framework to quantify viral mimicry by measuring "selective forces" that enrich PAMPs compared to a genome-wide reference distribution. We validate our predictions by identifying repeats that bind different PRRs and show potential viral mimics in different repeat families across eukaryotic genomes, suggesting shared mechanisms drive emergence and retention. We propose two non-exclusive evolutionary hypotheses. The first "repeat-centric" hypothesis posits PAMPs are integral to the repeat life cycle and are therefore enriched as they mediate repeat expansion. The second "organism-centric" hypothesis proposes viral mimicry functions as a cell-intrinsic feedback mechanism for sensing and reacting to transcriptional dysregulation, which provides a selective pressure to maintain PAMPs in genomes.

INTRODUCTION

Repetitive DNA sequences compose over half of the human genome and are often derived from integrated viruses and genetic parasites,[1] including transposable elements (TEs). In healthy cells, these sequences are normally inactive or transcriptionally silenced by epigenetic mechanisms but, in certain diseases, they can become de-repressed and transcribed. The aberrant accumulation of immunostimulatory repeat RNAs is frequently observed in cancers[2] as well as in inflammatory

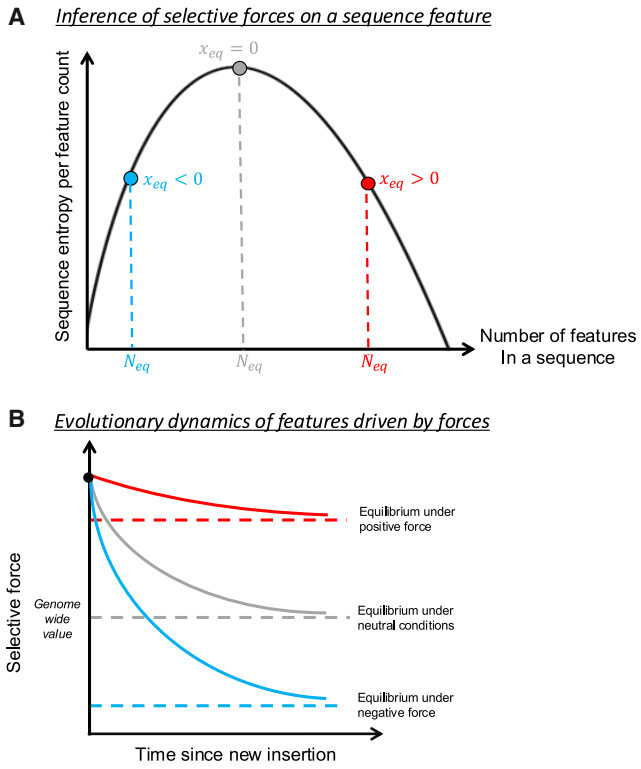


Figure 1. Theoretical framework for the evolution of features in a genomic sequence

(A) Representation of selective versus entropic forces on a genomic feature. At equilibrium ($x_{eq} = 0$), N_{eq} is the typical number of features in a sequence of fixed length. For negative ($x_{eq} < 0$) or positive ($x_{eq} > 0$) selective force, the typical number of features is lower or higher, respectively, than would be expected under the $x_{eq} = 0$ null model, where sequences relax to a genome-wide value dictated by the entropic forces on that feature.

(B) Over the course of evolution, sequences equilibrate to a force value in accordance with the type of selection that acts upon them.

(C) Under the assumptions of a Wright-Fisher model, the fitness, $f(\boldsymbol{s})$, as a function of a genomic sequence, $\boldsymbol{s}$, is given by the number of features contained in that sequence, $N_m(\boldsymbol{s})$, times the conjugate force on that feature, x_m, rescaled by the effective population size, $2N$.

diseases, such as autoimmunity[3–5]; in neurodegenerative disorders[6–8]; in viral infection[9–11]; and in aging.[12–14] These repeat RNAs may exhibit viral mimicry: "self" nucleic acids displaying pathogen-associated molecular patterns (PAMPs) that are recognized as "non-self" by pattern recognition receptors (PRRs) of the innate immune system, which initiate inflammation.[15–19] Collectively, these diverse data imply viral mimicry is a fundamental feature associated with cellular dysfunction. Viral mimicry can also potentially be leveraged therapeutically: for instance, the de-repression of immunostimulatory repeats by epigenetic drugs can trigger innate immune sensors and induce an interferon response, eliminating tumor cells.[15–20]

Recent work has suggested viral mimicry evolved convergently in different species as a response to viral infection. For example, convergent evolution of short interspersed elements (SINEs) in primate and rodent placentas seem to trigger viral mimicry in the form of double-stranded RNAs (dsRNAs) that induce interferon production,[21] providing fetal protection from viral infection at vulnerable moments during development. Such hypotheses are consistent with the importance of cell-intrinsic innate immunity for antiviral defense in undifferentiated stem cells,[22,23] which have not yet developed a multicellular immune system. Notably, a subset of cancer cells are often phenotypically compared to such stem cells,[24,25] implying that such cancer cells could encounter viral mimicry and be forced to adapt.[26,27]

Given the breadth of biological contexts associated with viral mimicry, an appropriate mathematical framework to quantify its extent would be broadly useful. To achieve this end, we model sequence statistics in the face of strongly biased germline mutational processes that repress immunostimulatory features.[28] We present a theoretical framework based on statistical physics to quantify the "selective forces" that lead to enrichment of non-self PAMP sequence features when compared to a genome-wide null model, thereby identifying candidate viral mimic loci. We experimentally verify that the endogenous sequences predicted by our model likely bind PRRs, and quantify viral mimicry across eukaryotic species, suggesting PAMP-enriched repeats have indeed convergently emerged. In the discussion, we qualitatively describe candidate mechanisms that could have driven this apparent convergent evolution.

RESULTS

Quantifying the enrichment of PAMPs in repeats with a statistical physics model

To quantify the landscape of PAMPs in repetitive regions of the human genome, we created the method of selective forces,[29] based on a maximum-entropy model of genomic sequences. We outline the theoretical approach here and, in the sections that follow, apply our approach to the specific cases of two PAMPs: CpG dinucleotides (a sequence motif) and dsRNAs (a structure). Genomic sequences are generally subject to entropic forces: random, but often biased, mutational processes that push sequences containing unusual features (such as PAMPs) to converge to a "typical" sequence composition. An example of such a bias is the fast-acting mutational process of cytosine deamination.[30] Selective forces oppose such sequence drift, leading to an over- or underrepresentation of atypical molecular features in a sequence depending on whether the selective force is positive or negative, respectively (Figure 1A). Based on these observations, we can construct a maximum-entropy distribution for each genomic sequence of length, L, constrained to have that sequence's particular feature count while being consistent with genome-wide nucleotide usage. The parameters of this distribution, which we call selective forces, precisely quantify the degree of over- or underrepresentation of a particular feature within that sequence.

To compute the maximum-entropy distribution and selective forces, we use exact-transfer matrix methods,[31] which are computationally efficient and scale with sequence length, enabling analysis of longer sequences, more complex features, and larger datasets compared to earlier approaches.[32]

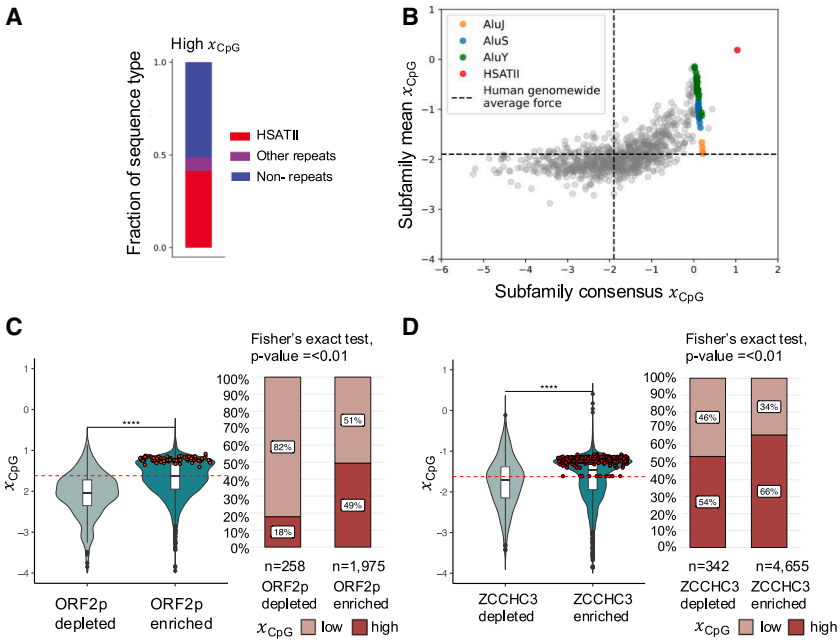


Figure 2. Landscape of selective forces on CpG dinucleotides for repeats

(A) Annotation of high-x_{CpG} ($x_{CpG} > 0$) sequences in the human genome according to their overlap with repeats annotated in the Dfam database. Non-repeat bases do not overlap with any repeat in the database. Only sequences on the main contigs were used.

(B) Average x_{CpG} computed on all inserts annotated in hg38 for each repeat family versus x_{CpG} of the consensus repeat reported in the Dfam database. Alus and HSATII are highlighted with higher mean x_{CpG}. Dashed line indicates genome-wide average force.

(C) Distribution of x_{CpG} of L1 ORF2p-binding L1 transcripts in embryonal carcinoma cell line (N2102Ep). Functional intact LINEs in red (****$p < 0.001$ for t test). ORF2p-enriched and -depleted transcripts are selected by differential expression analysis between ORF2p-IP versus both mock IP and total RNA with |log2FC| greater than 3 and adjusted $p < 0.05$. Fisher's exact test on proportion of x_{CpG} high versus x_{CpG} low of ORF2p-enriched and -depleted transcripts, $p = 9 \times 10^{-23}$, odds ratio 4.4, 95% confidence interval 3.2–6.3.

(D) x_{CpG} on *ZCCHC3*-binding L1 transcripts in N2102Ep. Functional intact L1s in red (****$p < 0.001$ for t test). *ZCCHC3*-enriched and -depleted transcripts selected by differential expression analysis between *ZCCHC3*-IP versus both mock IP and total RNA with a |log2FC| greater than 3 and adjusted $p < 0.05$. Fisher's exact test, $p = 1.3 \times 10^{-5}$, odds ratio 1.6, 95% confidence interval [1.3:2.1].

Conjugate forces have the advantage that they can be readily compared between groups of sequences within and between genomes, independent of sequence length. They therefore enable the quantitative study of complex phenotypes, like viral mimicry, at scale and across genomes. Traditional measures of selection like the integrated haplotype score (iHS)[33] are primarily restricted to features sampled at the population scale, like single-nucleotide polymorphisms, while our framework allows for inference of statistical properties of sequences with a shared ancestor (e.g., repetitive DNA families such as long interspersed element-1 or LINE-1) from a single reference genome. Intrinsic to an equilibrium maximum-entropy model is the notion of "relaxational dynamics," or the decay dynamics of an (out-of-equilibrium) sequence to a sequence concordant with equilibrium feature content, for a given selective force. In a hypothetical example, we show how an insertion with an initially high positive force value (i.e., over-representation of a particular feature) evolves to an equilibrium when under pressure to maintain that feature (positive force value), to deplete that feature (negative force value), or under neutral evolution to a genome-wide value purely driven by entropic forces (Figure 1B).

Given that repeat insertions can be dated by measuring their distance from their family consensus, we can infer whether selection acts on genome-encoded PAMPs according to the following framework (see STAR Methods for detailed description). First, we infer the selective forces on specific PAMPs for all loci of a particular genomic repeat family. Next, we predict the evolution of that repeat family's lineage in the genome since the time of its insertion by the aforementioned force-relaxation dynamics[29] and compare this to a Kimura-based model of neutral evolution that is consistent with genome-wide nucleotide-usage statistics and with mutational biases. If the equilibrium value for the selective force on a PAMP is greater than (or less than) the equilibrium value obtained from the neutral model, we predict that this PAMP is being maintained (or purged) through selection. In the limit of strong selection and weak mutation where sites along the genome are typically monoallelic, as they are in humans,[34] selective forces map onto appropriately scaled selection coefficients in an underlying population genetic model[35,36] (Figures 1C and STAR Methods). However, it should be clearly stated that this framework cannot identify the particular mechanisms that generate these effective selection coefficients. We address hypotheses regarding putative selective processes that can drive the maintenance of immunostimulatory repeats in the discussion.

Landscape of forces on CpG dinucleotides in repeats

CpG dinucleotides are known to act as PAMPs, recognized by toll-like receptor 9 (TLR9) in DNA,[37] and by the zinc-finger antiviral protein (ZAP) in RNA,[38] and they are highly underrepresented in the human genome, as CpG sites mutate at a much faster rate than other dinucleotides.[39–41] We calculated the forces on CpG dinucleotides for regions across the human genome (STAR Methods; Table S1) and found that high-x_{CpG} genomic regions were primarily in intergenic regions. Strikingly, many such regions with a propensity to over-represent CpG dinucleotides were not annotated as "CpG islands"[42] or specific regulatory regions (Figure 2A). The pericentromeric satellite repeat HSATII is the most representative repeat among those with high-x_{CpG} genomic regions in the human genome (Figure 2A), consistent with previous results.[19] We further quantified the selective forces on CpGs in repeats by comparing the evolution of individual dinucleotide motifs between the original consensus sequence, likely to resemble the founding ancestral

Table 1. Ratios of dinucleotide mutation rates and a corresponding value of the equilibrium force on the CpG dinucleotide

$\mu_{TiCpG}:\mu_{TvCpG}:\mu_{Ti}:\mu_{Tv}$	x_{CpG}^{eq}
40:10:4:1	−1.9
40:4:4:1	−1.8
40:1:4:1	−1.7
4:4:4:1	−0.3
20:4:4:1	−1.2
27:2:4:1	−1.4

Rates are shown for transition and transversion with and outside of CpG context.

insertion, and subsequent genomic copies (Table S2). We found most repeat families to have a force on CpG dinucleotides expected from the neutral model (Figure 2B), with notable exceptions, particularly HSATII and young Alu repeats. The abundant family of LINE-1 (L1) elements, which make up ∼20% of the human genome,[43] is of particular interest. They fall into two categories: fully intact and non-intact copies. Approximately 146 copies of L1 DNA sequences are fully intact and protein coding (i.e., contain potentially functional promoters and open reading frames [ORFs], consisting of approximately 6 kb[44]). These sequences encode proteins that enable autonomous retrotransposition. They represent 1% of full-length L1 DNA sequences (the remaining 99% contain features such as inactivating mutations in the promoters and/or ORFs) and 0.1% of all L1 nucleotides in the genome.[44] In general, fully intact L1 DNA sequences are highly enriched for CpG content (Figures S1 and S2). In part, this can be explained by enrichment of CpG sites in their internal promoter, which are typically hypermethylated to inhibit their transcription.[45,46] Yet CpGs outside of the promoter region also contribute to high-CpG forces of intact L1s.

While our analysis predicts over-representation of PAMP-encoding DNA sequences in host genomes, in most cases, the effector molecules are the transcribed nucleic acid sequences, such as the PAMPs within RNAs that directly engage PRRs. We had previously shown HSATII transcription and immunostimulatory capacity is CpG dependent.[8] Here, we sought to investigate active L1 retrotransposons, motivated by their enrichment for CpG dinucleotides. Fully intact L1 RNAs encode two proteins: ORF1p, a protein with nucleic acid chaperone activity,[47] and ORF2p, an endonuclease and reverse transcriptase.[48] These proteins, along with their encoding L1 RNA, assemble into ribonucleoproteins in *cis*.[49,50] Therefore, we performed RNA co-immunoprecipitation sequencing (RIP-seq) to identify and measure the CpG forces of the RNAs associated with ORF1p and ORF2p (STAR Methods). RIP-seq was conducted in N2102Ep human embryonic carcinoma cell lines,[51] chosen because they homeostatically express L1 at analytically tractable levels,[52] and enriched transcripts were determined by differential expression analysis, comparing RIP-seq to total RNA and mock immunoprecipitation (IP) controls (STAR Methods). As expected, intact L1 transcripts are exclusively recovered in the ORF1p (Figure S3) and ORF2p-enriched fraction (Figure 2C). ORF1p- and ORF2p-enriched L1 transcripts have CpG forces consistent with high x_{CpG} observed in fully intact transcripts, significantly higher than controls. We then probed whether L1 RNA with high x_{CpG} is bound by PRRs that activate innate immune responses. We selected *ZCCHC3*, a protein recently described as a co-sensor of cGAS,[53] since it has been found to interact with and inhibit L1 in an RNA-dependent fashion.[54,55] With RIP-seq, we found a substantial enrichment in high x_{CpG} L1 RNA associating with *ZCCHC3*, compared to controls and non-intact L1 (Figure 2D). Taken together, our findings show that high x_{CpG} L1 RNA is more likely to associate with L1 proteins and with a putative innate immune sensor of L1 RNA, validating our predictions and providing mechanistic insight into previous observations.[56–59] We therefore suggest that high x_{CpG} is associated with both replication-competent L1 and innate immune sensing of L1-associated transcripts.

Evolutionary dynamics of forces on CpG dinucleotides

To gain an understanding of how selective forces could affect evolution of repeat families, we first simulated the strong-selection, weak-mutation (or sequential fixation) dynamics of a single 3,000-bp stretch of DNA under two scenarios (Figure 3A). In the first scenario, all sites are free to mutate, while, in the second scenario, 200 CpG sites are maintained by strong selection (i.e., mutating these corresponds to an effectively infinitely deleterious fitness effect). In the limit of long times (as measured by Kimura distance), trajectories drawn from the two scenarios separate from each other along the force axis. We simulated each scenario 20 times to highlight the relative trajectory-level consistency of this stochastic process.

We next calculated the relaxation dynamics for all single HSATII and L1 inserts (Figure 3B), with a neutral background model where insertions evolve according to a Kimura model with respective mutation rates for transitions and transversions and increased mutation probability in CpG loci (STAR Methods; Table 1).[60] Such known and experimentally measured mutational biases have been invoked to explain reduced CpG content in some vertebrate genomes.[61–63] For HSATII, we observed higher x_{CpG} than expected from the Kimura-based neutral model, which implies selection acting on these repeats to maintain their high CpG content (regression analysis described in STAR Methods). For non-intact full lengh L1s, we observed that their CpG content decays to the genomic mean in a predictable way, which is consistent with an underlying neutral model of sequence evolution (Figure 3B). Taken together, this suggests that intact functional L1 inserts are selected to maintain high x_{CpG} even outside of their promoter regions (as most preserved CpGs do not occur there; Figure S1). We further probed the evolutionary dynamics of x_{CpG} for all repeat families annotated in the Dfam database[64] as a function of the Kimura distance to the consensus insert. Alu repeats, when considered as one family, show a pattern of relaxation to equilibrium, indicative of neutral evolution. However, the younger AluY and, to a lesser extent AluS, are far from their predicted equilibria and still possess PAMP-like high CpG content (Figures 3C and 3D).

To demonstrate the general consistency of our approach with viral mimicry, we quantified motif-usage mimicry between the human genome and human-infecting viruses.[29,32] We quantify the selective forces on all single, di-, and tri-nucleotide motifs[65]

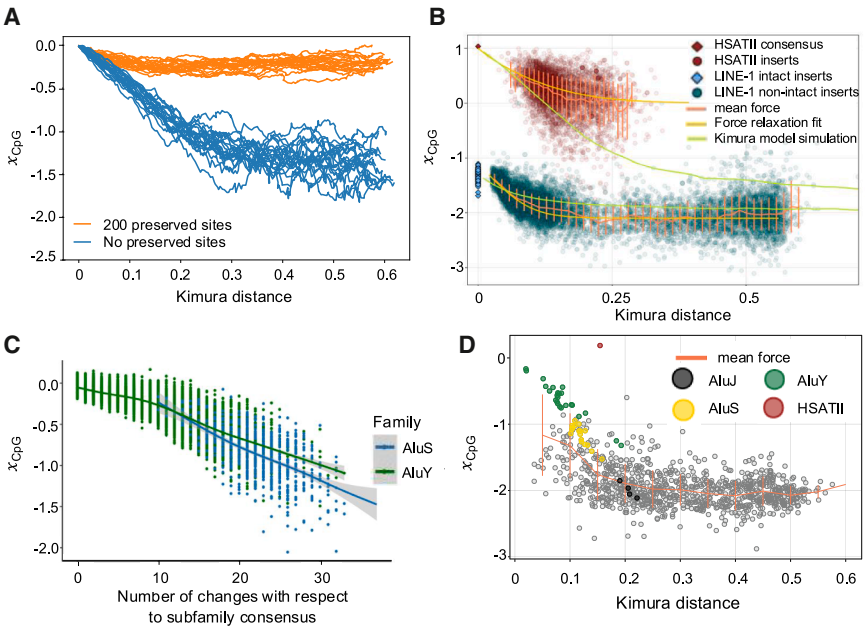


Figure 3. Evolutionary dynamics of CpG dinucleotides in repeats

(A) Temporal dynamics of CpG force in strong-selection, weak-mutation simulations of a single 3,000-bp stretch of DNA under two scenarios, one with and one without 200 CpG sites preserved by selection (20 trajectories).

(B) x_{CpG} for the full length non-intact inserts of LINE-1 and HSATII in human genome as a function of average distance from the full length intact sequences (for LINE-1) or the distance from the consensus sequence (for HSAT-II). The fit from the force-relaxation model is shown for both sequence families together with the force dynamics from the Kimura model for neutral evolution.

(C) Decay of CpG force with increasing distance from the subfamily consensus for the full-length Alu elements that are likely retrotransposition competent (sequence data from Bennett et al.). $p < 2 \times 10^{-16}$ (linear regression with CpG force as dependent variable and Alu family and the number of changes from the subfamily consensus as independent variables, $t = 166$, 12,528 degrees of freedom, two tailed t test for the slope; 95% confidence interval for the slope [−0.0306:−0.0299]).

(D) The mean x_{CpG} of all inserts in a repeat family as a function of the Kimura distance from the consensus sequence for each family (gray points). The binned mean is given in orange, and repeat families of note are highlighted with different colors. Repeat annotations are taken from Dfam.

in viruses that infect humans (Figure S4A) and compare them to different regions in our genome. Consistently, we find broad mimicry of motifs used by infectious viruses, with some variation between viral families.[29] These viruses, which devote most of their genomes to encoding proteins, show stronger overall similarity to motifs in the coding regions of the human genome, yet large variation exists in non-coding regions. Of note, specific repeat classes, identified above by our methods, such as Alu repeats and HSATII, exhibit an enrichment of non-self motifs that exceeds what is typical in the actual human-infecting viruses under consideration (Figure S4B).

Landscape and evolution of selective forces on dsRNA

In addition to motifs, PRRs also detect nucleic acid structures, such as dsRNA, a PAMP associated with virus infection.[66] We therefore extended our framework to quantify biases in RNA secondary-structure formation, defining the double-stranded force, x_{ds}, to quantify the presence of longer double-stranded segments that exceeded their expected frequency from a null model (STAR Methods). We scanned the entire genome in windows of 3,000 bp, comparable to the typical lengths of long non-coding RNAs[67] (Figure 4A). 88% of regions with high double-strand force ($x_{ds} > 0.5$) show a mean length of 40 bp, overlap with known repeats, and are depleted in coding regions (Figure 4B). More than 43% of these segments correspond to AluS inverted repeats (IRs). 75% of all IR Alu in the human genome are AluS IRs, and, if we filter for only IR Alu with high x_{ds}, the proportion increases to 86%. Under a model in which Alus are inserted into the genome in positions chosen uniformly at random,[68] this high fraction of IR-AluS with high x_{ds} can be statistically explained by the high fraction of AluS with respect to the other Alu subfamilies (STAR Methods). Another source of complementary fragments is *ORF2*, fragments of which are generated by 5′ truncated insertions of L1 and are statistically likely to be co-located. We quantified the sequence complexity of complementary segments based on an estimate of their Kolmogorov complexity, the shortest possible description based on a fixed language (STAR Methods; Figure S5A), further verifying that regions able to form long dsRNAs are not exclusive to simple repeat regions. We also detected several previously unannotated IRs prone to forming long dsRNA with high double-strand force (Table S3).

To validate that predicted dsRNA-forming regions can engage PRRs, we sequenced ligands to the J2 monoclonal antibody in a set of patient-derived colorectal cancer cell lines (STAR Methods). J2 binds dsRNA greater than 40 bp, nearly identical in length to the average length of complementary segments when $x_{ds} > 0.5$ (Figure 4C). Consistent with predictions, we show an enrichment of high-x_{ds} regions in J2 antibody-binding transcripts and with a similar profile to the previously published MDA5 agonists. We found AluS repeats constitute 84% of the enriched IR-Alus, again in agreement with the value predicted for high-x_{ds} sequences with our framework. As an additional source of validation, we used two published genome-wide datasets of dsRNA agonists[16,69] for the MDA5 receptor, a PRR that senses long dsRNA.[70] MDA5 ligands show a bivalent distribution with a peak at $x_{ds} = 0$ and a predominant peak at $x_{ds} > 0.5$. The ligands derived from IRs, and notably from AluS IRs, are only present in the high-x_{ds} peak. In agreement with our theoretical prediction that 86% of all IR-Alus in the human genome with high x_{ds} are AluS, we found that 89% of enriched IR-Alus in the MDA5 agonist experiment are AluS IRs, consistent with a similar profile to those found by the J2 antibody (Figure 4D). When studying the enrichment under DNA hypomethylating 5-azacytidine, we find a bivalent peak; when only looking at IR-Alus, we find they share

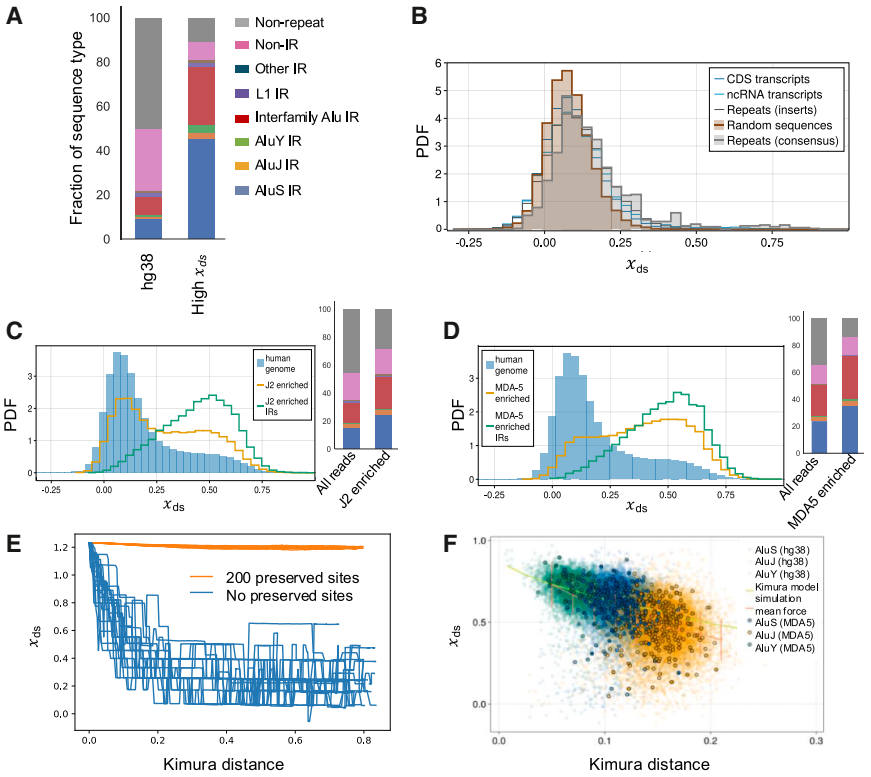


Figure 4. Landscape of selective forces on dsRNA for repeats

(A) Type of repeat (as annotated in RepeatMasker) with the longest overlapping sequence in complementary sequences for high-x_{ds} ($x_{ds} > 0.5$) windows in hg38.

(B) Histogram of x_{ds} calculated for mRNA coding sequences, non-coding RNAs, inserts, consensus sequences of repeats, and sequences obtained by randomly reshuffling mRNA coding sequences (red).

(C) x_{ds} histograms in human genome (sliding window with transcript of length of 3 kb) compared to J2 binding RNA transcripts. Enriched transcripts have a positive log enrichment with respect to the control experiment. Inverted repeat (IR) transcripts are annotated repeats with another repeat of the same family in opposite genomic sense within 3 kb.

(D) x_{ds} histograms in human genome (sliding window with transcript of length of 3 kb) compared to MDA5 binding RNA transcripts. Enriched transcripts have a positive log enrichment with respect to the control experiment. There is a strong association between high x_{ds} and ability to bind MDA5 (Fisher's exact test, odds ratio 4.5, 95% confidence interval [4.43:4.64], $p < 2.2 \times 10^{-16}$).

(E) Temporal dynamics of the double-stranded force in strong-selection, weak-mutation simulations of a single 3,000-bp stretch of DNA under two scenarios, one with and one without 200 CpG sites preserved by selection (20 trajectories).

(F) x_{ds} for IR-Alu inserts in human genome as a function of average distance from the consensus sequence of each subfamily forming the IR. The Kimura model neutral dynamics is averaged over Alu subfamilies. Points highlighted are the Alu inserts that were experimentally found to bind MDA5.

the same x_{ds} peak, indicating IR-Alus is the predominant foldable RNA species induced (Figure S5B).

When double-stranded segments are preserved, we expect x_{ds} to be much higher than expected under a neutral Kimura model. As done for CpG dinucleotides, we first performed a numerical simulation to understand how selective forces can be used to detect selection on double-strandedness (Figure 4E). Simulations were performed as described above for CpG evolution, with the only difference being that a 200-bp reverse-complementary stretch of sequence is held fixed. We further found that multiple subfamilies, particularly AluY and AluS, maintain high positive x_{ds} values, predicting that their transcripts would form dsRNA (Figures 4F; Tables S2 and S3). These positive x_{ds} values are driven by the proximity between two Alus, where a copy in the positive orientation is inserted adjacent to one in the negative orientation, occurring in the same transcript and, by their complementary nature, forming dsRNA (STAR Methods).

Forces on PAMP-enriched repeats across genomes

To demonstrate the power of our approach and explore how viral mimicry is shaped over evolutionary timescales, we examined the presence and conservation of repeats with high forces on CpG dinucleotides and dsRNA across genomes in 20 species[71,72] (Figures S6 and S7). Consistent with our assumptions, we found that forces on CpG dinucleotides and dsRNA are largely independent (Figure S8). For humans and mice, we show that forces on CpG dinucleotides are robust even when we exclude enhancers, regulatory regions from the FANTOM[73] and ENCODE databases,[74] or CpG islands. We found that high x_{ds} outlier regions occur in many species, even those lacking the Alu family, suggesting that foldable regions are not specifically a function of Alu repeats but could be a generic byproduct of transposon machinery across genomes (Figures S9A and S9B). As an example, we plot the x_{ds} distribution for the zebrafish genome, which has a clear set of high-x_{ds} regions and does not have Alu repeats (Figure 5A). Moreover, we find many of the genomic regions prone to dsRNA formation are not simply low-complexity regions. To the best of our knowledge, this is the first quantification of the presence of potential PAMP-forming repeat regions outside of primates and mouse models. The full list of regions with $x_{ds} > 0.5$ in the zebrafish genome is reported in Table S4.

Next, we compared the selective forces on repeat families in Hominoidea, whose most recent common ancestor dates to about 16 million years ago (mya).[75] We scanned the genomes of five small and great apes and compared sequences with high x_{CpG} and x_{ds} values. The number of high-x_{CpG} sequences conserved between humans and other apes decreases exponentially with their evolutionary distance (Figure 5B), an expected result given the high mutation rate in CpG motifs. Almost half of the high-x_{CpG} subsequences in the human genome overlap with HSATII, found in primates after the branching of the *Pongo* genus from the other great apes. This allowed us to pinpoint the HSATII insertion in the primate

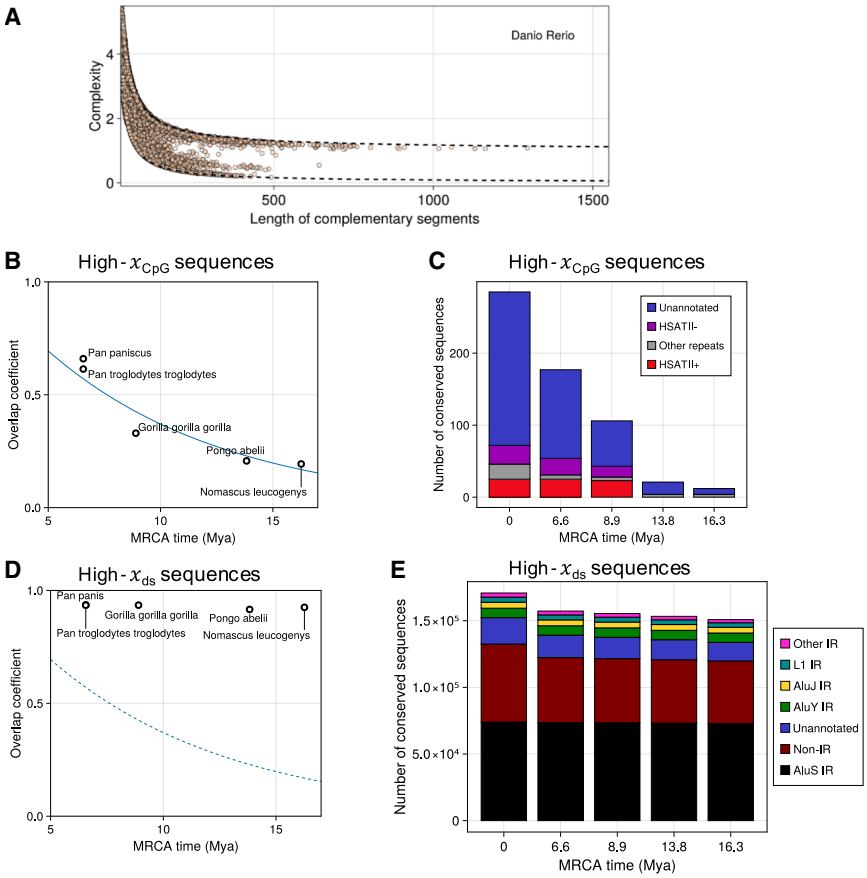


Figure 5. Evolution and conservation of forces on PAMPs

(A) Complexity of sequences in complementary regions found in the *Danio rerio* genome as a function of segment length. Dashed lines correspond to the complexity of a completely random sequence (top line) and trivial region consisting of a single nucleotide (bottom).

(B) Scatterplot of the overlap coefficient between the high-x_{CpG} ($x_{CpG} > 0$) sequences in the human genome and those of other primates versus the most recent common ancestor (MRCA) time.[56] Two high-x_{CpG} sequences are considered overlapping if they result as a hit from BLAST (STAR Methods). The blue curve denotes an exponential fit.

(C) Barplot presenting overlap with repeats of conserved high-sequences. The *x* axis indicates the MRCA time (0 mya are human sequences). Sequences are counted as repeats if they overlap with annotations in the Dfam database. The + or − sign after the repeat name indicates the sense in which the repeat is annotated in the database. Non-repeat sequences do not overlap with any repeat in the database.

(D) Same analysis as (B) but with high-x_{ds} sequences (greater than 0.5).

(E) Barplot presenting overlap with repeats of conserved high-x_{ds} sequences. The *x* axis indicates the MRCA time (0 mya are human sequences). Sequences are counted as repeats if they overlap with annotations in the RepeatMasker database. Sequences are counted as IRs if the two complementary regions overlap with two annotations in the RepeatMasker database with the same name but inverted sense (+/− or −/+). "Non-IR" indicates cases where the two repeats in the two inverted senses and in the same window have a different name. "Non-repeat" indicates cases where one or both the two complementary regions do not overlap with any known repeat.

genomes between 13.8 and 8.9 mya. Since its insertion into the genome, HSATII sequences have been conserved to a much greater extent than other sequences in the high-CpG pool, while other CpG-containing regions have typically rapidly decayed (Figure 5C). We therefore conclude that PAMPs are maintained in HSATII. Unlike high-x_{CpG} regions, which, apart from HSATII and Alus, rapidly equilibrate and are generally not conserved, we found high-x_{ds} windows more generally conserved across Hominoidea for all sequence families. Our findings imply that their demonstrated ability to engage dsRNA innate receptors is a conserved phenotypic feature along the lineage (Figure 5D). When focusing on high-x_{ds} regions that overlapped with repeats, we confirmed that inverted Alu repeats are highly conserved since their appearance in primate genomes more than 16.3 mya (Figure 5E).

DISCUSSION

Quantifying the landscape of viral mimicry and how it is shaped by evolution is an open theoretical question. Our framework enables large-scale quantification of enrichment of immunogenic sequence features, which we have leveraged to assign function to non-coding regions of the genome.[28] In our previous work, we used this framework to explore viral evolution, in particular the evolutionary history of the 1918 pandemic influenza strain. Our models showed that, after H1N1 entered its human host from an avian reservoir, its genome was depleted of CpG motifs to better resemble its new host genome. This led to the prediction, subsequently validated,[38] that CpGs in RNA are PAMPs in humans, and that viruses adapt to mimic host-sequence features and eliminate CpGs so they can avoid PRR detection.[32] Here, we build on this theory in viral evolution to develop a framework to quantify selective forces on repeats in metazoan genomes, detect PAMP-enriched regions within those genomes, and show how selection likely has shaped their evolution. We quantify viral mimicry across 20 species, finding strong enrichment of PAMPs across species-specific repeat classes, strongly suggesting a shared evolutionary constraint acting across all species. It is natural, then, to ask what this shared constraint might be. We propose two non-exclusive hypotheses for the phenomena driving PAMP enrichment, one "repeat-centric" and one "organism-centric."

The repeat-centric hypothesis states that PAMPs are beneficial to the repeat family themselves, with PAMP enrichment favoring repeat-family survival and expansion. The foundation of this hypothesis is that PAMP enrichment can be correlated with retrotransposition competence. Indeed, this has been shown in Alu subfamilies where retrotransposition rapidly

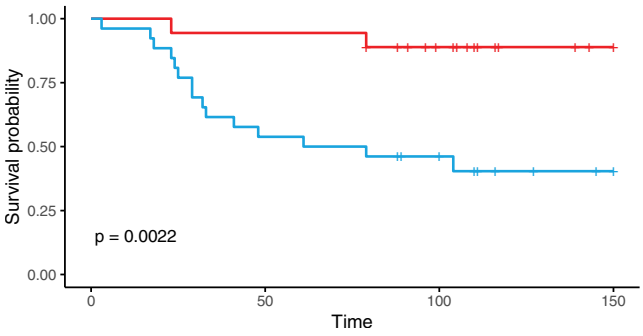


Figure 6. Using the double-stranded force framework to interpret clinical data

Kaplan-Meier curves stratified by mean expression of SINEs originating from loci with predicted high dsRNA force ($x_{ds} > 0.5$) in red ($N = 18$) and low predicted form in blue ($N = 26$), based on survival data in the Badal et al. melanoma cohort ($p = 0.0022$, log-rank test; STAR Methods).

decays with distance from the consensus sequences, which are typically enriched for CpG dinucleotides. We showed that younger Alu subfamilies, which are more likely to retain the ability to retrotranspose,[76] tend to be enriched for CpGs, suggesting that CpG enrichment is correlated with retrotransposition competence of Alu subfamilies. Likewise, functional copies of L1 maintain atypically high CpG content compared to non-functional copies. Our coIP assays with L1 ORF1p and ORF2p proteins as bait indeed capture high x_{CpG} L1 RNA, as does the innate immune sensor *ZCCHC3*, which was recently linked to the CpG-dependent ZAP sensor,[77] further validating our approach. These data confirm that our predicted functional L1 transcripts are indeed binding the L1 ORF2p reverse transcriptase and are innate agonists. Moreover, CpG-rich 5′ UTRs in L1 have been found to be prone to Z-DNA formation and interaction with ZBP1, indicating a second mode by which regions that over-represent CpGs can signal innate immunity.[5] These results suggest another aspect of the repeat-centric model whereby active L1 species derive a selective benefit from innate immune restriction, as undetected L1 species would have a deleterious effect on cellular and potentially organismal fitness (e.g., through ORF2p-derived endonuclease-induced DNA damage and a highly processive insertion mechanism).[78] Therefore, several lines of evidence presented here suggest that there are aspects of the repeat life cycle that favor PAMP enrichment. However, we would expect these selective pressures to be repeat family specific, given differences in different family lifecycles.

Our alternative organism-centric hypothesis is that repeat-associated PAMPs directly enhance organismal fitness by acting as a signal of dysregulated cells, triggering inflammation and improving the chances of eliminating dysregulated cells by programmed cell death or by immunosurveillance. Thereby, repeats are consistently selected to be enriched in PAMPs across many independent introductions across species. This scenario represents a novel form of repeat "domestication," a reversal of the typical scenario where a genetic parasite selfishly co-opts host functions: here, the host co-opts features of the parasite, leveraging them with potentially profound consequences for the organism. Repeats may be domesticated for other purposes, particularly in differentiated cell states. For instance, Alus can become hotspots for RNA editing or altered gene expression,[79–82] or HSATII may have a DNA-regulatory function, as its DNA sequences can sequester chromatin-regulatory proteins and trigger epigenetic change.[83] We argue that the innate immune system provides an independent explanation for how the endogenous repeat landscape is broadly shaped, as it has a non-specific function, separated from the regulation of a particular gene. We hypothesize that repeats are prone to forming PAMPs through their sequence similarity and rate of generation, such as the high rate of inversion from L1-mediated reverse transcription.[48,84]

A consequence of this hypothesis is that viral mimicry is capable of exerting a selective pressure on cells, as has been proposed during development.[21,22] In this sense, our framework assigns a quantitative phenotype to classes of non-coding RNA that can be used to help interpret RNA expression data. Such a genotype-phenotype map is an important open problem in RNA biology.[85] Our recent work has shown viral mimics impact cancer evolution in pancreatic cancer[25]; however, a lack of early-stage cancer datasets often prevents an analysis of how the mimicry response acts earlier in cancer evolution. We analyzed a published early-stage dataset in melanoma,[86]which had shown early expression of repeats to be associated with inflammatory signaling and better outcomes. Similarly, we found that premalignant lesions of the fallopian tube show early expression of dsRNA-forming repeats and the consequent chronic viral mimicry response shapes the antitumor immune response post-transformation.[27] Likewise, chronic interferon signaling has been linked to resistance to immunotherapy.[87] An initial analysis shows that stratifying patients using the dsRNA force annotations and RNA expression is associated with survival (Figure 6), consistent with our hypothesis that viral mimicry is a fitness challenge for tumors with translational implications. Additionally, in cell lines where drugs inhibited RNA splicing and induced intron retention,[88] we found the overexpression of repeats for which we had predicted high x_{ds} (Figure S10). Previous work hypothesized that dsRNA formed from introns is a checkpoint against intron retention.[89–91] For drugs less associated with intron retention, the effect was consistently weakened or absent. Collectively taken, this evidence is supportive of the organism-centric hypothesis.

The analysis presented here supports our previous hypothesis that regions with the capacity to form dsRNA and enriched for CpGs are spread across the genome where they can act as triggers for a broader viral mimicry warning system upon disruption of several cellular processes, including DNA methylation, chromatin organization, and RNA metabolism.[92–95] One such warning signal is the expression of transcriptionally repressed repeats after mutations in p53, a key sensor of DNA damage. Therefore, repeat expression can serve as a backup for p53 loss of function.[20,27,96] Indeed, the ancient age of some innate receptors, such as dsRNA receptors; the clear co-evolution of zinc-finger proteins with transposons[94]; and evidence of potential dsRNA-forming SINEs other than Alus (Figures S7 and S8) together suggest that viral mimicry is an ancient and broad metazoan characteristic. The DNA-damage response and innate immunity may well have co-evolved over long evolutionary timescales. By

quantifying loci prone to forming PAMPs, future work will be able to define the mechanisms of PAMP formation in the genome with greater specificity, such as the coupling of specific active LINE elements to dsRNA formation by SINE elements in *trans*, their evolutionary age, and specific SINE subfamilies.

We hypothesize that this repeat-mediated immune response is particularly important during development, a time when an acute infection, transcriptional dysregulation, or genome-destabilizing endogenous retroelements are potentially catastrophic, particularly since the cooperative response of a fully differentiated immune system is still lacking. The stem-like cells in early cancer may reactivate this response, as we show in melanoma, and as tumors progress they likely adapt to evade it. While we focused here on two PAMPs, our framework is generalizable to other, more complex patterns,[11] which comprise the full extent of viral mimics in the genome. Our ultimate goal is to learn the "PAMP code" maintained within genomes.

Limitations of the study

Our study has demonstrated a theoretical framework for capturing viral mimicry. However, future work will be needed to fully elucidate the breadth and mechanisms underlying this phenomenon. This can be challenging, as viral mimicry often emanates from repetitive regions of the genome, where traditional genetic tools can lack in efficacy. An additional limitation is the lack of complete differentiation between innate immune function and other regulatory functions that repeats may perform as they are domesticated into the genome, particularly given one is not entirely exclusive of the other. Our study also suffers from less thorough annotation of repeats and study of their function outside of humans and model organisms such as mice, zebrafish, and yeast.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Benjamin D. Greenbaum (greenbab@mskcc.org).

Materials availability

This study did not generate new materials.

Data and code availability

- J2 RIP-seq data of the POP92 cell line are deposited under Gene Expression Omnibus (GEO) series GEO: GSE305619, and matched RNA-seq control data are deposited under series GEO: GSE145639. ZCCHC3 and ORF1p RIP-seq data for N2102Ep cell line with 250-bp reads are deposited under GEO series GEO: GSE305618. ORF1p and ORF2p RIP-seq data with 50-bp reads are deposited under series GEO: GSE280626.
- Code for the computation of dinucleotide forces is available in the following public repository: https://github.com/crankycrank/DimerForces (DOI: https://doi.org/10.5281/zenodo.16576248).
- Code for the computation of dsRNA forces is available in the following public repository: https://github.com/adigioacchino/DSForces.jl (DOI: https://doi.org/10.5281/zenodo.16616688).

ACKNOWLEDGMENTS

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748 (A.S., S.S., and B.D.G.); NIH grants R01AI081848 (N.V. and B.D.G.), R01CA240924 (A.S. and B.D.G.), R01GM126170 (J.L.), R01AG078925 (J.L.), P50 254838-01 (O.A.-W.), and U01CA228963 (A.S., S.S., and B.D.G.); Fondation de la Recherche Médicale: ANR-Flash Covid, Project SARS-Cov-2immunRNAs (S.C. and R.M.); the V Foundation for Cancer Research (A.S.); the Mark Foundation ASPIRE award (B.D.G.); the Pershing Square Sohn Prize-Mark Foundation Fellowship (A.S., O.A.-W., N.V., B.D.G.); the Edward P. Evans Foundation (O.A-W.); the Marie-Josée Kravis Fellowship in Quantitative Biology (S.M.), the Canadian Institutes of Health Research, New Investigator salary award 201512MSH360794-228629 (D.D.D.C.); Canada Research Chair (D.D.D.C.); CIHR Foundation grant FDN 148430 (D.D.D.C.); CIHR project grant PJT 165986 (D.D.D.C.); NSERC 489073 (D.D.D.C); NSF grant no. CHE 2155095 (P.S.); and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 101026293 (A.D.G.). The authors would like to acknowledge productive conversations with Jef Boeke, Kathleen Burns, Katherine Chiappinelli, Arnold Levine, John Moran, Charles M. Rice, William Schneider, Phil Sharp, David T. Ting, and the De Carvalho, Cocco, Greenbaum, LaCava, and Monasson laboratories. We would also like to acknowledge support from the Halverson Center for Computational Oncology, the National Center for Dynamic Interactome Research (NIH P41GM10982); the Genome Technology Center at NYULH, a shared resource partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center; and the UMCG Research Sequencing Facility and Utrecht Sequencing Facility (USEQ; USEQ is subsidized by the University Medical Center Utrecht and The Netherlands X-omics Initiative [NWO project 184.034.019]). This work utilized resources from the High Performance Computing Group at Memorial Sloan Kettering Cancer Center.

AUTHOR CONTRIBUTIONS

Conceptualization, P.S., R.M., S.C., and B.D.G.; research plan, P.S., R.M., S.C., and B.D.G.; mathematical modeling, P.S., A.D.G., S.M., R.M., S.C., and B.D.G.; double-stranded force calculation, P.S., A.D.G., R.M., and S.C.; model implementation, P.S., A.D.G., S.M., and S.S.; population genetics interpretation and simulations, S.M.; RIP-seq and RNA-seq data analysis and interpretation, S.S.; comparative genomic analysis, A.S.; data analysis, P.S., A.D.G., A.S., H.L., S.M., and S.S.; L1 experimental design and execution, J.L., H.J., and B.-H.L.; dsRNA experimental design, A.H., R.C., P.M., and D.D.D.C.; interpretation, P.S., M.S.T., O.A.-W., N.V., J.L., D.D.D.C., R.M., S.C., and B.D.G.; writing, A.D.G., A.S., S.M., R.M., S.C., and B.D.G.; writing – review & editing, P.S., A.D.G., A.S., S.S., S.M., J.L., O.A.-W., D.D.D.C., P.M., N.R., R.M., S.C., and B.D.G.

DECLARATION OF INTERESTS

O.A.-W. has performed consulting for Incyte, Prelude Therapeutics, AstraZeneca, Merck, Janssen, Pfizer Boulder, and LoxoOncology/Eli Lilly and is on the Scientific Advisory Board of AIChemy and Harmonic Discovery Inc. B.D.G. has received honoraria for speaking engagements from Merck, Bristol Meyers Squibb, and Chugai Pharmaceuticals; he has received research funding from Bristol Meyers Squibb, Merck, and ROME Therapeutics; and has been a compensated consultant for Darwin Health, Merck, PMV Pharma, Shennon Biotechnologies, Synteny AI, and ROME Therapeutics of which he is a co-founder. A.S. has been a compensated consultant for ROME Therapeutics. A.S. and M.S.T. hold ROME stock options. D.D.D.C. received research funding from Pfizer and Nektar therapeutics and is a shareholder, co-founder, and CSO of Adela (former DNAMx). J.L. received research funding from ROME Therapeutics, Ribon Therapeutics, and Refeyn; he received compensation from Transposon Therapeutics, ROME Therapeutics, and Oncolinea.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Experimental methods

- cDNA library preparation and RNA-Immunoprecipitation sequencing (RIP-seq)
- Immunoprecipitation of double-stranded RNA by J2 antibody
- Analysis of J2 immunoprecipitation RNAseq
- Analysis of MDA5 protection assays

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of forces on sequence features
  - Relationship between Maximum Entropy models and population genetic models
  - Evolutionary dynamics of a sequence motif with force relaxation formalism
  - Kimura-based model of population genetics for the evolution of sequence motifs
  - Analysis of forces across species
  - Analysis of evolutionary conserved sequences with high x_{CpG} or x_{ds}
  - Sequence ensembles
  - Search of long transcripts with complementary regions
  - Sequence complexity quantification
  - Estimate of genome regions with high double stranded force
  - Analysis of repeats for splice inhibitors
  - Analysis of repeats for melanoma cohort
  - Counts filtering, normalization and statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.101011.

Received: November 3, 2023
Revised: September 25, 2024
Accepted: August 28, 2025
Published: September 24, 2025

REFERENCES

1. Hoyt, S.J., Storer, J.M., Hartley, G.A., Grady, P.G.S., Gershman, A., de Lima, L.G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., et al. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science *376*, eabk3112. https://doi.org/10.1126/science.abk3112.
2. Ting, D.T., Lipson, D., Paul, S., Brannigan, B.W., Akhavanfard, S., Coffman, E.J., Contino, G., Deshpande, V., Iafrate, A.J., Letovsky, S., et al. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science *331*, 593–596. https://doi.org/10.1126/science.1200801.
3. Jiao, H., Wachsmuth, L., Wolf, S., Lohmann, J., Nagata, M., Kaya, G.G., Oikonomou, N., Kondylis, V., Rogg, M., Diebold, M., et al. (2022). ADAR1 averts fatal type I interferon induction by ZBP1. Nature *607*, 776–783. https://doi.org/10.1038/s41586-022-04878-9.
4. de Reuver, R., Verdonck, S., Dierick, E., Nemegeer, J., Hessmann, E., Ahmad, S., Jans, M., Blancke, G., Van Nieuwerburgh, F., Botzki, A., et al. (2022). ADAR1 prevents autoinflammation by suppressing spontaneous ZBP1 activation. Nature *607*, 784–789. https://doi.org/10.1038/s41586-022-04974-w.
5. Zhang, T., Yin, C., Fedorov, A., Qiao, L., Bao, H., Beknazarov, N., Wang, S., Gautam, A., Williams, R.M., Crawford, J.C., et al. (2022). ADAR1 masks the cancer immunotherapeutic promise of ZBP1-driven necroptosis. Nature *606*, 594–602. https://doi.org/10.1038/s41586-022-04753-7.
6. Saleh, A., Macia, A., and Muotri, A.R. (2019). Transposable Elements, Inflammation, and Neurological Disease. Front. Neurol. *10*, 894. https://doi.org/10.3389/fneur.2019.00894.
7. Jönsson, M.E., Garza, R., Johansson, P.A., and Jakobsson, J. (2020). Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. Trends Genet. *36*, 610–623. https://doi.org/10.1016/j.tig.2020.05.004.
8. Takahashi, T., Stoiljkovic, M., Song, E., Gao, X.-B., Yasumoto, Y., Kudo, E., Carvalho, F., Kong, Y., Park, A., Shanabrough, M., et al. (2022). LINE-1 activation in the cerebellum drives ataxia. Neuron *110*, 3278–3287. https://doi.org/10.1016/j.neuron.2022.08.011.
9. Chiang, J.J., Sparrer, K.M.J., van Gent, M., Lässig, C., Huang, T., Osterrieder, N., Hopfner, K.-P., and Gack, M.U. (2018). Viral unmasking of cellular 5S rRNA pseudogene transcripts induces RIG-I-mediated immunity. Nat. Immunol. *19*, 53–62. https://doi.org/10.1038/s41590-017-0005-y.
10. Nogalski, M.T., Solovyov, A., Kulkarni, A.S., Desai, N., Oberstein, A., Levine, A.J., Ting, D.T., Shenk, T., and Greenbaum, B.D. (2019). A tumor-specific endogenous repetitive element is induced by herpesviruses. Nat. Commun. *10*, 90. https://doi.org/10.1038/s41467-018-07944-x.
11. Vabret, N., Najburg, V., Solovyov, A., Gopal, R., McClain, C., Šulc, P., Balan, S., Rahou, Y., Beauclair, G., Chazal, M., et al. (2022). Y RNAs are conserved endogenous RIG-I ligands across RNA virus infection and are targeted by HIV-1. iScience *25*, 104599. https://doi.org/10.1016/j.isci.2022.104599.
12. De Cecco, M., Ito, T., Petrashen, A.P., Elias, A.E., Skvir, N.J., Criscione, S.W., Caligiana, A., Brocculi, G., Adney, E.M., Boeke, J.D., et al. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature *566*, 73–78. https://doi.org/10.1038/s41586-018-0784-9.
13. Simon, M., Van Meter, M., Ablaeva, J., Ke, Z., Gonzalez, R.S., Taguchi, T., De Cecco, M., Leonova, K.I., Kogan, V., Helfand, S.L., et al. (2019). LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation. Cell Metab. *29*, 871–885. https://doi.org/10.1016/j.cmet.2019.02.014.
14. Fukuda, S., Varshney, A., Fowler, B.J., Wang, S.-B., Narendran, S., Ambati, K., Yasuma, T., Magagnoli, J., Leung, H., Hirahara, S., et al. (2021). Cytoplasmic synthesis of endogenous Alu complementary DNA via reverse transcription and implications in age-related macular degeneration. Proc. Natl. Acad. Sci. USA *118*, e2022751118. https://doi.org/10.1073/pnas.2022751118.
15. Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A., et al. (2015). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell *162*, 974–986. https://doi.org/10.1016/j.cell.2015.07.011.
16. Mehdipour, P., Marhon, S.A., Ettayebi, I., Chakravarthy, A., Hosseini, A., Wang, Y., de Castro, F.A., Loo Yau, H., Ishak, C., Abelson, S., et al. (2020). Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. Nature *588*, 169–173. https://doi.org/10.1038/s41586-020-2844-1.
17. Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J., et al. (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell *162*, 961–973. https://doi.org/10.1016/j.cell.2015.07.056.
18. Sheng, W., LaFleur, M.W., Nguyen, T.H., Chen, S., Chakravarthy, A., Conway, J.R., Li, Y., Chen, H., Yang, H., Hsu, P.-H., et al. (2018). LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade. Cell *174*, 549–563. https://doi.org/10.1016/j.cell.2018.05.052.
19. Tanne, A., Muniz, L.R., Puzio-Kuter, A., Leonova, K.I., Gudkov, A.V., Ting, D.T., Monasson, R., Cocco, S., Levine, A.J., Bhardwaj, N., and Greenbaum, B.D. (2015). Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. Proc. Natl. Acad. Sci. USA *112*, 15154–15159. https://doi.org/10.1073/pnas.1517584112.
20. Leonova, K.I., Brodsky, L., Lipchick, B., Pal, M., Novototskaya, L., Chenchik, A.A., Sen, G.C., Komarova, E.A., and Gudkov, A.V. (2013). p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. Proc. Natl. Acad. Sci. USA *110*, E89–E98. https://doi.org/10.1073/pnas.1216922110.

21. Wickramage, I., VanWye, J., Max, K., Lockhart, J.H., Hortu, I., Mong, E. F., Canfield, J., Lamabadu Warnakulasuriya Patabendige, H.M., Guzeloglu-Kayisli, O., Inoue, K., et al. (2023). SINE RNA of the imprinted miRNA clusters mediates constitutive type III interferon expression and antiviral protection in hemochorial placentas. Cell Host Microbe *31*, 1185–1199. https://doi.org/10.1016/j.chom.2023.05.018.

22. Wu, X., Dao Thi, V.L., Huang, Y., Billerbeck, E., Saha, D., Hoffmann, H.-H., Wang, Y., Silva, L.A.V., Sarbanes, S., Sun, T., et al. (2018). Intrinsic Immunity Shapes Viral Resistance of Stem Cells. Cell *172*, 423–438. https://doi.org/10.1016/j.cell.2017.11.018.

23. Lefkopoulos, S., Polyzou, A., Derecka, M., Bergo, V., Clapes, T., Cauchy, P., Jerez-Longres, C., Onishi-Seebacher, M., Yin, N., Martagon-Calderón, N.-A., et al. (2020). Repetitive elements trigger RIG-I-like receptor signaling that regulates the emergence of hematopoietic stem and progenitor cells. Immunity *53*, 934–951. https://doi.org/10.1016/j.immuni.2020.10.007.

24. Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111. https://doi.org/10.1038/35102167.

25. Batlle, E., and Clevers, H. (2017). Cancer stem cells revisited. Nat. Med. *23*, 1124–1134. https://doi.org/10.1038/nm.4409.

26. Sun, S., You, E., Hong, J., Hoyos, D., Del Priore, I., Tsanov, K.M., Mattagajasingh, O., Di Gioacchino, A., Marhon, S.A., Chacon-Barahona, J., et al. (2024). Cancer cells restrict immunogenicity of retrotransposon expression via distinct mechanisms. Immunity *57*, 2879–2894. https://doi.org/10.1016/j.immuni.2024.10.015.

27. Ishak, C.A., Marhon, S.A., Tchrakian, N., Hodgson, A., Loo Yau, H., Gonzaga, I.M., Peralta, M., Lungu, I.M., Gomez, S., Liang, S.-B., et al. (2025). Chronic viral mimicry induction following p53 loss promotes immune evasion. Cancer Discov. *15*, 793–817. https://doi.org/10.1158/2159-8290.CD-24-0094.

28. Vabret, N., Bhardwaj, N., and Greenbaum, B.D. (2017). Sequence-Specific Sensing of Nucleic Acids. Trends Immunol. *38*, 53–65. https://doi.org/10.1016/j.it.2016.10.006.

29. Greenbaum, B.D., Cocco, S., Levine, A.J., and Monasson, R. (2014). Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. Proc. Natl. Acad. Sci. USA *111*, 5054–5059. https://doi.org/10.1073/pnas.1402285111.

30. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Turki, S.A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al. (2016). Timing, rates and spectra of human germline mutation. Nat. Genet. *48*, 126–133. https://doi.org/10.1038/ng.3469.

31. Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. Phys. Rev. *106*, 620–630. https://doi.org/10.1103/PhysRev.106.620.

32. Greenbaum, B.D., Levine, A.J., Bhanot, G., and Rabadan, R. (2008). Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. *4*, e1000079. https://doi.org/10.1371/journal.ppat.1000079.

33. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72. https://doi.org/10.1371/journal.pbio.0040072.

34. Collins, F.S., and Mansoura, M.K. (2001). The Human Genome Project. Revealing the shared inheritance of all humankind. Cancer *91*, 221–225. https://doi.org/10.1002/1097-0142(20010101)91:1+<221::aid-cncr8>3.3.co;2-0.

35. Mustonen, V., and Lässig, M. (2005). Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc. Natl. Acad. Sci. USA *102*, 15936–15941. https://doi.org/10.1073/pnas.0505537102.

36. Sella, G., and Hirsh, A.E. (2005). The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA *102*, 9541–9546. https://doi.org/10.1073/pnas.0501865102.

37. Bauer, S., Kirschning, C.J., Häcker, H., Redecke, V., Hausmann, S., Akira, S., Wagner, H., and Lipford, G.B. (2001). Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. Proc. Natl. Acad. Sci. USA *98*, 9237–9242. https://doi.org/10.1073/pnas.161293498.

38. Takata, M.A., Gonçalves-Carneiro, D., Zang, T.M., Soll, S.J., York, A., Blanco-Melo, D., and Bieniasz, P.D. (2017). CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature *550*, 124–127. https://doi.org/10.1038/nature24039.

39. Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504. https://doi.org/10.1093/nar/8.7.1499.

40. Shen, J.C., Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22*, 972–976. https://doi.org/10.1093/nar/22.6.972.

41. Sved, J., and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. USA *87*, 4692–4696. https://doi.org/10.1073/pnas.87.12.4692.

42. Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022. https://doi.org/10.1101/gad.2037511.

43. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

44. Penzkofer, T., Jäger, M., Figlerowicz, M., Badge, R., Mundlos, S., Robinson, P.N., and Zemojtel, T. (2017). L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. *45*, D68–D73. https://doi.org/10.1093/nar/gkw925.

45. Hata, K., and Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene *189*, 227–234. https://doi.org/10.1016/s0378-1119(96)00856-6.

46. Sanchez-Luque, F.J., Kempen, M.-J.H.C., Gerdes, P., Vargas-Landin, D. B., Richardson, S.R., Troskie, R.-L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., Salvador-Palomeque, C., et al. (2019). LINE-1 Evasion of Epigenetic Repression in Humans. Mol. Cell *75*, 590–604. https://doi.org/10.1016/j.molcel.2019.05.024.

47. Martin, S.L. (2006). The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. J. Biomed. Biotechnol. *2006*, 45621. https://doi.org/10.1155/JBB/2006/45621.

48. Baldwin, E.T., van Eeuwen, T., Hoyos, D., Zalevsky, A., Tchesnokov, E. P., Sánchez, R., Miller, B.D., Di Stefano, L.H., Ruiz, F.X., Hancock, M., et al. (2024). Structures, functions, and adaptations of the human LINE-1 ORF2 protein. Nature *626*, 194–206. https://doi.org/10.1038/s41586-023-06947-z.

49. Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H. H., Boeke, J.D., and Moran, J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. Mol. Cell Biol. *21*, 1429–1439. https://doi.org/10.1128/MCB.21.4.1429-1439.2001.

50. Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat. Struct. Mol. Biol. *13*, 655–660. https://doi.org/10.1038/nsmb1107.

51. Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P., and Moran, J.V. (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769–773. https://doi.org/10.1038/nature09209.

52. Di Stefano, L.H., Saba, L.J., Oghbaie, M., Jiang, H., McKerrow, W., Benitez-Guijarro, M., Taylor, M.S., and LaCava, J. (2023). Affinity-based interactome analysis of endogenous LINE-1 macromolecules. Methods Mol. Biol. *2607*, 215–256. https://doi.org/10.1007/978-1-0716-2883-6_12.

53. Lian, H., Wei, J., Zang, R., Ye, W., Yang, Q., Zhang, X.-N., Chen, Y.-D., Fu, Y.-Z., Hu, M.-M., Lei, C.-Q., et al. (2021). Author Correction: ZCCHC3 is a co-sensor of cGAS for dsDNA recognition in innate immune response. Nat. Commun. *12*, 5526. https://doi.org/10.1038/s41467-021-25397-7.

54. Taylor, M.S., Altukhov, I., Molloy, K.R., Mita, P., Jiang, H., Adney, E.M., Wudzinska, A., Badri, S., Ischenko, D., Eng, G., et al. (2018). Dissection of affinity captured LINE-1 macromolecular complexes. eLife *7*, e30094. https://doi.org/10.7554/eLife.30094.

55. Taylor, M.S., LaCava, J., Mita, P., Molloy, K.R., Huang, C.R.L., Li, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T., et al. (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell *155*, 1034–1048. https://doi.org/10.1016/j.cell.2013.10.021.

56. Moldovan, J.B., and Moran, J.V. (2015). The Zinc-Finger Antiviral Protein ZAP Inhibits LINE and Alu Retrotransposition. PLoS Genet. *11*, e1005121. https://doi.org/10.1371/journal.pgen.1005121.

57. Mavragani, C.P., Sagalovskiy, I., Guo, Q., Nezos, A., Kapsogeorgou, E. K., Lu, P., Liang Zhou, J., Kirou, K.A., Seshan, S.V., Moutsopoulos, H. M., and Crow, M.K. (2016). Expression of Long Interspersed Nuclear Element 1 Retroelements and Induction of Type I Interferon in Patients With Systemic Autoimmune Disease. Arthritis Rheumatol. *68*, 2686–2696. https://doi.org/10.1002/art.39795.

58. Zhang, X., Zhang, R., and Yu, J. (2020). New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. Front. Cell Dev. Biol. *8*, 657. https://doi.org/10.3389/fcell.2020.00657.

59. Lagisquet, J., Zuber, K., and Gramberg, T. (2021). Recognize Yourself-Innate Sensing of Non-LTR Retrotransposons. Viruses *13*, 94. https://doi.org/10.3390/v13010094.

60. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120. https://doi.org/10.1007/BF01731581.

61. Arndt, P.F. (2007). Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny. Gene *390*, 75–83. https://doi.org/10.1016/j.gene.2006.11.022.

62. Baele, G., Van de Peer, Y., and Vansteelandt, S. (2010). Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. BMC Evol. Biol. *10*, 244. https://doi.org/10.1186/1471-2148-10-244.

63. Bérard, J., and Guéguen, L. (2012). Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. Syst. Biol. *61*, 510–521. https://doi.org/10.1093/sysbio/sys024.

64. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*, D81–D89. https://doi.org/10.1093/nar/gkv1272.

65. Di Gioacchino, A., Lecce, I., Greenbaum, B.D., Monasson, R., and Cocco, S. (2025). Deciphering the code of viral-host adaptation through maximum entropy models Molecular Biology & Evolution *42*, 1–22. https://doi.org/10.1093/molbev/msaf127.

66. Jensen, S., and Thomsen, A.R. (2012). Sensing of RNA viruses: a review of innate immune receptors involved in recognizing RNA virus invasion. J. Virol. *86*, 2900–2910. https://doi.org/10.1128/JVI.05738-11.

67. Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. (2012). Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? BioArchitecture *2*, 189–199. https://doi.org/10.4161/bioa.22592.

68. Riba, A., Fumagalli, M.R., Caselle, M., and Osella, M. (2020). A Model-Driven Quantitative Analysis of Retrotransposon Distributions in the Human Genome. Genome Biol. Evol. *12*, 2045–2059. https://doi.org/10.1093/gbe/evaa201.

69. Ahmad, S., Mu, X., Yang, F., Greenwald, E., Park, J.W., Jacob, E., Zhang, C.-Z., and Hur, S. (2018). Breaching Self-Tolerance to Alu Duplex RNA Underlies MDA5-Mediated Inflammation. Cell *172*, 797–810. https://doi.org/10.1016/j.cell.2017.12.016.

70. Wu, B., Peisley, A., Richards, C., Yao, H., Zeng, X., Lin, C., Chu, F., Walz, T., and Hur, S. (2013). Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. Cell *152*, 276–289. https://doi.org/10.1016/j.cell.2012.11.048.

71. Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C. C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., et al. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Res. *47*, D752–D758. https://doi.org/10.1093/nar/gky1099.

72. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22. https://doi.org/10.1186/s13059-014-0560-6.

73. Abugessaisa, I., Ramilowski, J.A., Lizio, M., Severin, J., Hasegawa, A., Harshbarger, J., Kondo, A., Noguchi, S., Yip, C.W., Ooi, J.L.C., et al. (2021). FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Res. *49*, D892–D898. https://doi.org/10.1093/nar/gkaa1054.

74. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

75. Thinh, V.N., Mootnick, A.R., Geissmann, T., Li, M., Ziegler, T., Agil, M., Moisson, P., Nadler, T., Walter, L., and Roos, C. (2010). Mitochondrial evidence for multiple radiations in the evolutionary history of small apes. BMC Evol. Biol. *10*, 74. https://doi.org/10.1186/1471-2148-10-74.

76. Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active Alu retrotransposons in the human genome. Genome Res. *18*, 1875–1883. https://doi.org/10.1101/gr.081737.108.

77. Goodier, J.L., Wan, H., Soares, A.O., Sanchez, L., Selser, J.M., Pereira, G.C., Karma, S., García-Pérez, J.L., Kazazian, H.H., Jr., and García Cañadas, M.M. (2023). ZCCHC3 is a stress granule zinc knuckle protein that strongly suppresses LINE-1 retrotransposition. PLoS Genet. *19*, e1010795. https://doi.org/10.1371/journal.pgen.1010795.

78. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025. https://doi.org/10.1038/sdata.2016.25.

79. Capshew, C.R., Dusenbury, K.L., and Hundley, H.A. (2012). Inverted Alu dsRNA structures do not affect localization but can alter translation efficiency of human mRNAs independent of RNA editing. Nucleic Acids Res. *40*, 8637–8645. https://doi.org/10.1093/nar/gks590.

80. Daniel, C., Silberberg, G., Behm, M., and Öhman, M. (2014). Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol. *15*, R28. https://doi.org/10.1186/gb-2014-15-2-r28.

81. Kim, D.D.Y., Kim, T.T.Y., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S., and Gabriel, A. (2004). Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res. *14*, 1719–1725. https://doi.org/10.1101/gr.2855504.

82. Yakovchuk, P., Goodrich, J.A., and Kugel, J.F. (2009). B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. Proc. Natl. Acad. Sci. USA *106*, 5569–5574. https://doi.org/10.1073/pnas.0810738106.

83. Hall, L.L., Byron, M., Carone, D.M., Whitfield, T.W., Pouliot, G.P., Fischer, A., Jones, P., and Lawrence, J.B. (2017). Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. Cell Rep. *18*, 2943–2956. https://doi.org/10.1016/j.celrep.2017.02.072.

84. Solovyov, A., Behr, J.M., Hoyos, D., Banks, E., Drong, A.W., Thornlow, B., Zhong, J.Z., Garcia-Rivera, E., McKerrow, W., Chu, C., et al. (2025). Pan-cancer multi-omic model of LINE-1 activity reveals locus heterogeneity of retrotransposition efficiency. Nat. Commun. *16*, 2049. https://doi.org/10.1038/s41467-025-57271-1.

85. Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell *157*, 77–94. https://doi.org/10.1016/j.cell.2014.03.008.

86. Badal, B., Solovyov, A., Di Cecilia, S., Minhow Chan, J., Chang, L.-W., Iqbal, R., Aydin, I.T., Rajan, G.S., Chen, C., Abbate, F., et al. (2017). Transcriptional dissection of melanoma identifies a high-risk subtype underlying TP53 family genes and epigenome dysregulation. JCI Insight *2*, e92102. https://doi.org/10.1172/jci.insight.92102.

87. Benci, J.L., Xu, B., Qiu, Y., Dada, H., Twyman-Saint Victor, C., Cucolo, L., Lee, D.S.M., Pauken, K.E., Huang, A.C., Gangadhar, T.C., et al. (2016). Tumor interferon signaling regulates a multigenic resistance program to immune checkpoint blockade. Cell *167*, 1540–1554. https://doi.org/10.1016/j.cell.2016.11.022.

88. Seiler, M., Yoshimi, A., Darman, R., Chan, B., Keaney, G., Thomas, M., Agrawal, A.A., Caleb, B., Csibi, A., Sean, E., et al. (2018). H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat. Med. *24*, 497–504. https://doi.org/10.1038/nm.4493.

89. Wu, Q., Nie, D.Y., Ba-Alawi, W., Ji, Y., Zhang, Z., Cruickshank, J., Haight, J., Ciamponi, F.E., Chen, J., Duan, S., et al. (2022). PRMT inhibition induces a viral mimicry response in triple-negative breast cancer. Nat. Chem. Biol. *18*, 821–830. https://doi.org/10.1038/s41589-022-01024-4.

90. Bowling, E.A., Wang, J.H., Gong, F., Wu, W., Neill, N.J., Kim, I.S., Tyagi, S., Orellana, M., Kurley, S.J., Dominguez-Vidaña, R., et al. (2021). Spliceosome-targeted therapies trigger an antiviral immune response in triple-negative breast cancer. Cell *184*, 384–403. https://doi.org/10.1016/j.cell.2020.12.031.

91. Ishak, C.A., Loo Yau, H., and De Carvalho, D.D. (2021). Spliceosome-Targeted Therapies Induce dsRNA Responses. Immunity *54*, 11–13. https://doi.org/10.1016/j.immuni.2020.12.012.

92. Chen, R., Ishak, C.A., and De Carvalho, D.D. (2021). Endogenous Retroelements and the Viral Mimicry Response in Cancer Therapy and Cellular Homeostasis. Cancer Discov. *11*, 2707–2725. https://doi.org/10.1158/2159-8290.CD-21-0506.

93. Ishak, C.A., and De Carvalho, D.D. (2020). Reactivation of Endogenous Retroelements in Cancer Development and Therapy. Annu. Rev. Cancer Biol. *4*, 159–176. https://doi.org/10.1146/annurev-cancerbio-030419-033525.

94. Lindholm, H.T., Chen, R., and De Carvalho, D.D. (2023). Endogenous retroelements as alarms for disruptions to cellular homeostasis. Trends Cancer *9*, 55–68. https://doi.org/10.1016/j.trecan.2022.09.001.

95. Jacobs, F.M.J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242–245. https://doi.org/10.1038/nature13760.

96. Levine, A.J., and Greenbaum, B. (2012). The maintenance of epigenetic states by p53: the guardian of the epigenome. Oncotarget *3*, 1503–1504. https://doi.org/10.18632/oncotarget.780.

97. Ardeljan, D., Wang, X., Oghbaie, M., Taylor, M.S., Husband, D., Deshpande, V., Steranka, J.P., Gorbounov, M., Yang, W.R., Sie, B., et al. (2020). LINE-1 ORF2p expression is nearly imperceptible in human cancers. Mob. DNA *11*, 1. https://doi.org/10.1186/s13100-019-0191-2.

98. LaCava, J., Jiang, H., and Rout, M.P. (2016). Protein complex affinity capture from cryomilled mammalian cells. J. Vis. Exp. *118*, 54518. https://doi.org/10.3791/54518.

99. Taylor, M.S., LaCava, J., Dai, L., Mita, P., Burns, K.H., Rout, M.P., and Boeke, J.D. (2016). Characterization of L1-ribonucleoprotein particles. Methods Mol. Biol. *1400*, 311–338. https://doi.org/10.1007/978-1-4939-3372-3_20.

100. Nielsen, M.I., Wolters, J.C., Bringas, O.G.R., Jiang, H., Di Stefano, L.H., Oghbaie, M., Hozeifi, S., Nitert, M.J., van Pijkeren, A., Smit, M., et al. (2025). Targeted detection of endogenous LINE-1 proteins and ORF2p interactions. Mob. DNA *16*, 3. https://doi.org/10.1186/s13100-024-00339-4.

101. Cristea, I.M., and Chait, B.T. (2011). Conjugation of magnetic beads for immunopurification of protein complexes. Cold Spring Harb. Protoc. *2011*, pdb-prot5610. https://doi.org/10.1101/pdb.prot5610.

102. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

103. Yang, W.R., Ardeljan, D., Pacyna, C.N., Payer, L.M., and Burns, K.H. (2019). SQuIRE reveals locus-specific regulation of interspersed repeat expression. Nucleic Acids Res. *47*, e27. https://doi.org/10.1093/nar/gky1301.

104. O'Brien, C.A., Kreso, A., Ryan, P., Hermans, K.G., Gibson, L., Wang, Y., Tsatsanis, A., Gallinger, S., and Dick, J.E. (2012). ID1 and ID3 regulate the self-renewal capacity of human colon cancer-initiating cells through p21. Cancer Cell *21*, 777–792. https://doi.org/10.1016/j.ccr.2012.04.036.

105. Gao, S., Soares, F., Wang, S., Wong, C.C., Chen, H., Yang, Z., Liu, W., Go, M.Y.Y., Ahmed, M., Zeng, Y., et al. (2021). CRISPR screens identify cholesterol biosynthesis as a therapeutic target on stemness and drug resistance of colon cancer. Oncogene *40*, 6601–6613. https://doi.org/10.1038/s41388-021-01882-7.

106. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

107. Jaynes, E.T. (1957). Information theory and statistical mechanics. II. Phys. Rev. *108*, 171–190. https://doi.org/10.1103/physrev.108.171.

108. Kimura, M., and Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. Genetics *61*, 763–771. https://doi.org/10.1093/genetics/61.3.763.

109. Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics *61*, 893–903. https://doi.org/10.1093/genetics/61.4.893.

110. Ohta, T., and Tachida, H. (1990). Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics *126*, 219–229. https://doi.org/10.1093/genetics/126.1.219.

111. Berg, J., Willmann, S., and Lässig, M. (2004). Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. *4*, 42. https://doi.org/10.1186/1471-2148-4-42.

112. Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. *10*, 195–205. https://doi.org/10.1038/nrg2526.

113. Suzuki, Y., Gojobori, T., and Kumar, S. (2009). Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. Mol. Biol. Evol. *26*, 2275–2284. https://doi.org/10.1093/molbev/msp133.

114. Subramanian, S., and Kumar, S. (2006). Higher intensity of purifying selection on >90% of the human genes revealed by the intrinsic replacement mutation rates. Mol. Biol. Evol. *23*, 2283–2287. https://doi.org/10.1093/molbev/msl123.

115. Di Gioacchino, A., Šulc, P., Komarova, A.V., Greenbaum, B.D., Monasson, R., and Cocco, S. (2021). The heterogeneous landscape and early evolution of pathogen-associated CpG dinucleotides in SARS-CoV-2. Mol. Biol. Evol. *38*, 2428–2445. https://doi.org/10.1093/molbev/msab036.

116. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).

Scikit-learn: Machine learning in Python. The Journal of machine Learning research *12*, 2825–2830.

117. Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., et al. (2023). The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. *51*, D1188–D1195. https://doi.org/10.1093/nar/gkac1072.

118. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889. https://doi.org/10.1093/nar/gkz1062.

119. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

120. Smit, AFA, Hubley, R & Green, P. (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

121. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA *12*, 2. https://doi.org/10.1186/s13100-020-00230-y.

122. Sippl, M.J. (1999). Biological sequence analysis. In Probabilistic Models of Proteins and Nucleic Acids, R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson, eds. (Cambridge University Press), p. 356. https://doi.org/10.1110/ps.8.3.695.

123. Kimura, M., and Weiss, G.H. (1964). The Stepping Stone model of population structure and the decrease of genetic correlation with distance. Genetics *49*, 561–576. https://doi.org/10.1093/genetics/49.4.561.

124. Li, M., and Vityani, P. (2019). An introduction to Kolmogorov complexity and its applications. Texts in Computer Science (Springer). https://doi.org/10.1007/978-3-030-11298-1.

125. Dingle, K., Camargo, C.Q., and Louis, A.A. (2018). Input-output maps are strongly biased towards simple outputs. Nat. Commun. *9*, 761. https://doi.org/10.1038/s41467-018-03101-6.

126. Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Cancer Genome Atlas Research Network; Buonamici, S., and Yu, L. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep. *23*, 282–296. https://doi.org/10.1016/j.celrep.2018.01.088.

127. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11. https://doi.org/10.1186/s13100-015-0041-9.

128. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467. https://doi.org/10.1159/000084979.

129. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

130. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

131. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCES	SOURCE	IDENTIFIER
Cell lines		
N2102Ep Clone 2/A6 cells	Merck	06011803
POP92	Princess Margaret Cancer Center	N/A
Antibodies		
α-ORF1p 4H1 monoclonal ab	Millipore, Sigma	MABC1152; RRID: AB_2941775
α-ORF2p	Abcam	in house[97]
α-ZCCHC3	Proteintech	29399-1-AP; RRID: AB_2918294
Naive polyclonal rabbit IgG (control for α-ZCCHC3)	Millipore, Sigma	I5006; RRID: AB_1163659
Naive polyclonal mouse IgG (control for α-ORF1p)	Millipore, Sigma	I5381; RRID: AB_1163670
Naive polyclonal rabbit IgG (control for α-ORF2p)	Innovative Research	RBIGGAP10MG; (no RRID)
J2	SCICIONS	10010500; RRID: AB_2651015

METHOD DETAILS

Experimental methods

RNA extractions and co-immunoprecipitations (RIP) from embryonal carcinoma cells

Embryonal carcinoma cells were cultured at 37°C in humidified incubators maintained with 7% CO2 atmosphere. N2102Ep Clone 2/A6 cells (Merk, #06011803) were cultured in DMEM (high glucose, no sodium pyruvate; Thermo Fisher, #11965092), supplemented with 10% (v/v) fetal bovine serum, 1x penicillin/streptavidin and 2 mM Glutamine (Thermo Fisher, #25030024). Large-scale growth, harvesting, cryo-milling, and co-IP was achieved as previously described,[52,98–100] and summarized as follows. α-ORF1p-, α-ORF2p-, and α-ZCCHC3-targeted co-IPs used in-house made[101] magnetic affinity media: for α-ORF1p [15 μg antibody/mg magnetic beads], we used the 4H1 monoclonal antibody (Millipore Sigma, #MABC1152); for α-ORF2p [10 μg/mg magnetic beads] we used the clone 9 monoclonal antibody[97]; for α-ZCCHC3 [10 μg/mg magnetic beads] we used the rabbit polyclonal antibody (Proteintech, #29399-1-AP). Co-IPs were conducted using 100 mg cell powder, extracted at 25% (w/v) in 20 mM HEPES pH 7.4, 500 mM or 300 mM NaCl, 1% (v/v) Triton X-100, 1x protease inhibitors (Roche, #1187358001), and 0.4% (v/v) RNasin (Promega, #2515). Centrifugally clarified cell extracts were incubated with affinity medium (20 μL of slurry for α-ORF1p and α-ORF2p, later used in 50bp RIP-seeq as described[26] and 15 μL of slurry for α-ORF1p and α-ZCCHC3 later used in 250bp RIP-seq) for 30 min at 4°C. The solutions were made with nuclease-free H2O and experiments were conducted using nuclease-free tubes and pipette tips. Macromolecule extractions performed on this cell line as described typically yielded between 450 and 500 μL of soluble extract at 6–8 mg/mL of protein as assessed by Bradford assay (Thermo Fisher, #23200). After target capture, washing the media was performed with the same solution without protease inhibitors and with RNasin at 0.1% (v/v). RNAs were eluted from the affinity media after RIP with 250 μL of TRIzol Reagent (Thermo Fisher, #15596026). After adding chloroform to the TRIzol eluate, the separated aqueous phase (containing RNAs) was obtained using Phasemaker tubes (per manufacturer's instructions; Thermo Fisher, #A33248), and was then combined with an equal volume of ethanol and further purified using a spin-column according to the manufacturer's instructions (Zymo Research, #R2060). For α-ZCCHC3 co-IP, two 100 mg-scale preparations were pooled prior to spin column purification. Eluates from α-ORF1p and α-ORF2p co-Ips, later used in 50-bp RIP-seq, were not treated with DNase I on-column, this was done during the sequencing library preparation (described, below); eluates from α-ORF1p and α-ZCCHC3 later used in 250-bp RIP-seq were DNase I treated on-column. Purified nucleic acids from all co-IPs were eluted in 6-10μL of nuclease-free water; 0.5-1μL was used for quality analysis and the remainder conserved for RNA-seq. Mock RNA co-IP controls were prepared in an identical manner using either naive polyclonal mouse IgG (control for α-ORF1p; Millipore Sigma, #I5381) or naive polyclonal rabbit IgG (control for α-ORF2p: Innovative Research, #IRBIGGAP10MG; control for α-ZCCHC3: Millipore Sigma, #I5006). Total RNA controls were prepared by combining up to 35 μL of the clarified cell extracts with up to 500 μL of Trizol, vortex mixing for 1 min, then snap freezing in liquid N2 - and then later proceeding as above.

cDNA library preparation and RNA-Immunoprecipitation sequencing (RIP-seq)

50bp α-ORF1p and α-ORF2p RIP-seq

RNA extractions were quantified and quality controlled using RNA Pico Chips (Cat. #5067-1513) on an Agilent 2100 BioAnalyzer. RNA-Seq cDNA libraries were prepared using the Trio RNA-Seq Library Prep kit (Tecan, #0357-A01) with AnyDeplete Probe Mix-Human rRNA (Tecan, #S02305). DNase treatment preceded cDNA synthesis. cDNA synthesis: 3 - 5ng of input RNA from α-ORF1p RIPs and mock IPs, 1 ng from α-ORF2p RIPs and mock IPs, and 50 ng of total RNA were used with 8 (2 + 6) cycles of pre-depletion PCR library amplification and 8 (2 + 6) cycles of post-depletion amplification; the libraries were purified using Agencourt AMPure XP beads (Beckmann Coulter), quantified by qPCR, and the size distribution was checked using the Agilent TapeStation 2200 system. Final libraries were sequenced, paired-end, at 50 bp read-length on an Illumina NovaSeq 6000 v1.5 with 2% PhiX spike-in.

250bp α-ORF1p and α-ZCCHC3 RIP-seq

RNA extractions were quantified and quality controlled using an Agilent TapeStation 4200 and High Sensitivity RNA ScreenTape (Agilent, #5067–5579). RNA-seq cDNA libraries were prepared using the SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian (Takara, #634485), including rRNA depletion during library construction. cDNA synthesis: 5 ng of input RNA from α-ORF1p RIPs, α-ZCCHC3 RIPs and total RNA, and 1 ng of input RNA from mock IPs were used with 5 cycles of pre-depletion PCR amplification and 12 cycles (α-ZCCHC3 RIPs and total RNA) or 14 cycles (mock IPs) of post-depletion amplification. Libraries were purified using NucleoMag beads supplied in the library preparation kit and subsequently quantified using the Qubit 4 Fluorometer and the Qubit dsDNA HS assay kit (Invitrogen, #Q32854). The size distribution was checked using the TapeStation 4200; noting that primer dimers ($\sim$ 150bp) were persisted in the mock IP libraries (motivating an additional round of cleaning). To treat all libraries equally, they were pooled in a 4:4:4:1 ratio (α-ORF1p RIPs, α-ZCCHC3 RIPs:total RNA:mock IP) based on molarity of fragments of interest (range $\sim$ 200-1000 bp). An additional round of cleanup with NucleoMag beads was done to remove the primer dimers. A size selection of the final library pool was performed on a 2% E-gel EX (Invitrogen, #G401002) to exclude small fragments (less than about 200bp) and the DNA was eluted from the gel slices using the Zymoclean Gel DNA Recovery Kit (Zymo, #D4001), followed by quantification (Qubit) and quality control (TapeStation). Final libraries were sequenced, paired-end, at 250bp read-length on an Illumina NovaSeq 6000 platform.

RIP-seq read mapping and quantification

Reads were mapped to the hg38 genome for human samples using STAR v2.7.7a taking into account known splice sites from Gencode annotation.[102] STAR mapping parameters were set to as described in Software Quantifying Interspersed Repeat Expression (SQuIRE).[103] Quality filtered reads were mapped to annotated repeat loci in SQuIRE. SQuIRE pipeline quantifies locus-specific repeat expression by redistributing multi-mapping read fractions in proportion to estimated TE expression with an expectation maximization algorithm. Samples extracted at 300mM and 500mM NaCl were used as replicates to increase statistical power after checking for transcript composition similarity (Figure S3).

CpG quantification for repeats in hg38

Sequences of repeats were extracted from hg38 based on the coordinate annotation in RepeatMasker. x_{CpG} was then calculated for those hg38 derived repeat sequences. L1 inserts reported in RepeatMasker are considered intact whenever they have minimum 80% overlap with inserts annotated as full-length intact in the L1Base database.

Selection of RIP-Seq enriched repeats/transcripts

Samples extracted at 300mM and 500mM NaCl were used as replicates to increase statistical power after checking for transcript composition similarity (via hierarchical clustering). Targeted protein enriched transcripts were selected by Log2(Co-IP/mock) > 3, Log2(Co-IP/total RNA) > 3, and Benjamini-Hochberg adjusted p value < 0.05. Similarly, target protein depleted transcripts were selected by Log2(Co-IP/mock) < −3, Log2FC(Co-IP/total RNA) < −3, and Benjamini-Hochberg adjusted p value < 0.05 for samples collected using the same extraction and sequencing protocol and, Log2(Co-IP/mock) < −1, Log2FC(Co-IP/total RNA) < −1 was used for selection when comparing results generated using different protocols.

Immunoprecipitation of double-stranded RNA by J2 antibody

The patient-derived colon cancer stem cell (CSC)-enriched spheroid line POP92 was used for J2 immunoprecipitation.[104,105] Protein G Dynabeads were washed twice and resuspended in antibody conjugation buffer (1x PBS, 2mM EDTA, 0.1% BSA (w/v). 5 μg of anti-dsRNA mAb (J2) (SCICONS, cat# 10010500) were bound to 30 μL of washed beads overnight at 4°C on a rotating wheel. 10^7 POP92 cells per IP were fixed with 0.1% paraformaldehyde at room temperature (RT) for 10 min. Immediately, cells were quenched by adding glycine and washed twice with cold PBS. Crosslinked cells were lysed in lysis buffer (20mM Tris [pH 7.5], 150mM NaCl, 10mM EDTA, 10% Glycerol, 0.1% NP-40, 0.5% Triton X-100, supplemented with protease inhibitor tablet) for 15 min on ice. Following a spin at 12,000*g* at 4°C for 15 min, supernatant was transferred to a new eppendorf. The lysate was then immunoprecipitated using 30 μL antibody-conjugated Dynabeads per IP reaction overnight at 4°C in a rotator. Following magnetic separation, beads were washed three times with high salt wash buffer (20 mM Tris pH 7.5, 500 mM NaCl, 10 mM EDTA, 10% glycerol) and resuspended in 1X TBS. Per IP, 2 μL of Promega RNasinPlus RNase inhibitor (Fisher Scientific, PRN2611) and 0.5 μL of proteinase K (NEB, P8107S) was added. Decrosslinking was performed for all the IP samples at 65°C for 15 min. The Direct-zol RNA MiniPrep kit (Zymo Research, R2051) was used to extract RNA from IP supernatant. Samples were treated with Turbo DNase to remove any DNA contamination in the extracted RNA. Library prep was performed using Illumina Stranded total RNA ligation with Ribo Zero plus according to the manufacturers protocol. Samples were sequenced on a NovaSeq 6000 using paired end reads with 100 cycles.

Analysis of J2 immunoprecipitation RNAseq

RNAseq controls not enriched with J2 antibody from untreated POP92 cells were downloaded from GEO (submission GEO: GSE145639; samples GEO: GSM4322694 and GEO: GSM4322693).[16] 25 bp of J2 RNAseq reads were cut off with cutadapt to match the length of RNAseq control reads. All samples were aligned to the human genome hg38 using STAR with default settings.[102] BAM files were sorted using samtools. The compressed x_{ds} table was used to count fragments in the RNAseq data, see above for details on how this table was generated. The table describes windows for the genome as well as a complementary sequence which can form a double stranded sequence. Every complementary sequence has a seqA and a seqB part which was split into two different files. featureCounts was used to count the number of fragments aligning to seqA and seqB in J2 and RNAseq control BAM files including information about strand and reporting multimapping and muti-overlapping reads as fractional counts.[106] Counts for each seqA and seqB were then merged for each complementary sequence. For each complementary sequence log2FC (mean J2/mean control) was calculated.

Analysis of MDA5 protection assays

Raw sequencing data was downloaded from GEO: GSE103539 and GEO: GSE145639[69] and aligned to hg38 using STAR. The following settings were used to increase the mapping due to the repetitive nature of the data[16]: –outFilterMultimapNmax 1000 –outSAMmultNmax 1 –outFilterMismatchNmax 10 –outMultimapperOrder Random –winAnchorMultimapNmax 1000. After mapping, the data were processed as described above for J2 immunoprecipitation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of forces on sequence features

We define a Maximum Entropy (MaxEnt) framework[107] to determine the least constrained probability distribution over the set of sequences $\boldsymbol{s} = \{s_1, s_2 \ldots s_L\}$ of length L compatible with the observed occurrences (measurement) of a set of M features $\boldsymbol{N} = \{N_1, N_2 \ldots N_M\}$. The distribution is written as

$$P(\boldsymbol{s}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{m=1}^{M} x_m N_m(\boldsymbol{s})\right). \quad \text{(Equation 1)}$$

Here $Z(\boldsymbol{x})$ is a normalization factor:

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{s}'} \exp\left(\sum_{m=1}^{M} x_m N_m(\boldsymbol{s}')\right), \quad \text{(Equation 2)}$$

where the sum runs over all possible sequences $\boldsymbol{s}'$ having length L. The set of parameters $\boldsymbol{x} = \{x_1, x_2, \ldots, x_M\}$, hereafter called *selective forces*, is chosen so that the average value of each feature over the distribution P matches the observed number of this feature $\boldsymbol{N}^{obs}$ in one or more reference sequences:

$$N_m^{obs} = \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{x}) N_m(\boldsymbol{s}) = \frac{\partial \log Z(\boldsymbol{x})}{\partial x_m}. \quad \text{(Equation 3)}$$

The above equalities define a set of M coupled, nonlinear equations, with a unique solution due to the convexity of $\log Z(\boldsymbol{x})$. The forces are analogous to chemical potentials in statistical physics.

We use our formalism to calculate selective forces on sequences due constraints imposed by: (1) the force due to potentially immunogenic CpG motifs, the CpG force, x_{CpG}; (2) the forces on all nucleic acid motifs with length up to three nucleotides; and (3) the force on long complementary sequence stretches the double-stranded force, x_{ds}. For the latter case, we also use an equivalent minimal model to simplify the calculation. Details on these specific calculations are included in the Supplementary Methods.

Relationship between Maximum Entropy models and population genetic models

We have that the MaxEnt (ME) sequence distribution constraining a particular feature, m, over a set of sequences (e.g., insertions) in a single genome is given by:

$$P_{ME}(\boldsymbol{s}) = \frac{1}{Z_{ME}} \prod_{i}^{L} p_0(s_i) \exp[x_m N_m(\boldsymbol{s})]$$

where $p_0(s_i)$ is the null distribution of genomewide nucleotide usage statistics.

An ensemble of independent populations of sequences following Wright-Fisher (WF) dynamics in the weak mutation regime ($\mu\mathcal{N} \ll 1$, where μ is the mutation rate and $\mathcal{N}$ is the effective population size) can be described by the master equation:

$$\partial_t P(\boldsymbol{s}, t) = \sum_{\boldsymbol{r}} [u_{\boldsymbol{r}\to\boldsymbol{s}} P(\boldsymbol{r}, t) - u_{\boldsymbol{s}\to\boldsymbol{r}} P(\boldsymbol{s}, t)]$$

where the assumption is that each population remains essentially monoclonal, i.e., fixation is sufficiently fast. This assumption is supported by the fact that human genomes are not particularly diverse – human pairwise diversity is roughly 0.1%, or roughly 1 single nucleotide variant per 1000 base pairs.

Fundamental to our analysis is that, given the reasonable assumptions that insertions of a single repeat family are approximately uniformly distributed along the genome and are therefore typically well separated from each other (longer than typical recombination length scales and therefore roughly independent from each other), we can make the approximation that a repeat family is an instance of such an ensemble. The transition rates between sequences are given by Kimura and Ohta[108–110]:

$$u_{\boldsymbol{r}\to\boldsymbol{s}} = \mu_{\boldsymbol{r}\to\boldsymbol{s}}\mathcal{N}\frac{1-e^{-2[f(\boldsymbol{s})-f(\boldsymbol{r})]}}{1-e^{-2\mathcal{N}[f(\boldsymbol{s})-f(\boldsymbol{r})]}}.$$

We have that $f(\boldsymbol{s})$ is the absolute fitness of a particular sequence, $\mathcal{N}$ is the population size, and $\mu_{\boldsymbol{r}\to\boldsymbol{s}}$ is the underlying relative rate of the transition $\boldsymbol{r}\to\boldsymbol{s}$. We can derive a stationary distribution:

$$P_{WF}(\boldsymbol{s}) = \frac{1}{Z_{WF}}\prod_i^L p_0(s_i)\exp[2(\mathcal{N}-1)f(\boldsymbol{s})]\approx\frac{1}{Z_{WF}}\prod_i^L p_0(s_i)\exp[2\mathcal{N}f(\boldsymbol{s})]$$

where the last approximation holds for sufficiently large $\mathcal{N}$[111].

The bare nucleotide frequencies p_0 are identical between the ME and WF distributions and describe nucleotide usage statistics (i.e., a neutral null model where $f = 0$). In order for the two distributions to be equivalent, we must have:

$$2\mathcal{N}f(\boldsymbol{s}) = x_m N_m(\boldsymbol{s}).$$

Since x_m and $\mathcal{N}$ do not depend on any particular sequence $\boldsymbol{s}$, for equality to hold for all sequences, we must have:

$$f(\boldsymbol{s}) = \sigma N_m(\boldsymbol{s})$$

where σ can be interpreted as a fitness effect size per unit of feature m. Therefore, the fitness effect size can be given in terms of the conjugate force scaled by the population size $\sigma = x_m/2\mathcal{N}$. Note that the actual fitness landscape describing immunological features in arbitrary regions of the genome is likely to be more complex than this simple linear model. In addition, the human population size $\mathcal{N}$ is certainly not fixed over the timescales we are considering. However, this model can be thought of as a first order approximation beyond neutral sequence drift.

There are good reasons to assume repeats operate in such a 'strong selection-weak mutation' (SSWM) regime to first order. This regime is quantitatively delineated by the conditions:

$$\mu L\mathcal{N} \ll 1 \text{ and } |\delta|\mathcal{N} \gg 1$$

where μ is the per base mutation rate, L is the typical length (in bp) of a repeat locus, $\mathcal{N}$ is the effective population size, and δ is the typical fitness effect size. It has been estimated that $\mu \sim 10^{-8}$[30] and $\mathcal{N} \sim 10^4$[112]. For the repeat classes we consider, $L \lessapprox 10^4$ so that the first relation typically holds. Likewise, the typical fitness effect size for a sequence $\boldsymbol{s}$ is given by the fitness effect per feature m, $\sigma = x_m/2\mathcal{N}$, times the quantification of that feature, $N_m(\boldsymbol{s})$. Plugging these in to the second condition, we obtain:

$$\frac{|x_m|\,N_m(\boldsymbol{s})}{2} \gg 1.$$

As is reported in the main text and below, we infer that typically $|x_m| \sim \mathcal{O}(1)$ for both the CpG and double-stranded forces so that the second condition holds for sequences where $N_m(\boldsymbol{s}) \gg 1$, i.e., sufficient feature enrichment in an insertion is indicative of strong selection. It should be noted the assumptions of our model rely on patterns of *fixed* sequences, while other selection metrics can rely on extant (non-fixed) variation to infer selective coefficients due to selection in progress.[33]

Evolutionary dynamics of a sequence motif with force relaxation formalism

One can harness the formalism developed in Equations 1, 2, and 3 to study the evolutionary dynamics of number of motifs N_m, as it approaches the steady state (equilibrium) value N_m^{avg}[29]. As a sequence evolves it undergoes mutations, which cause changes in the number of motifs (and hence associated value of x_m). To model the evolutionary dynamics of sequences, we assume the number of motifs (N_m) evolves according to a linear response where the rate of change in the number of motifs is proportoinal to the deviation of the selective force from its equilibrium value. Such relaxation dynamics is given by

$$\tau\frac{dN_m(t)}{dt} = -x_m(N_m(t)) + x_m^{\text{eq}}, \qquad \text{(Equation 4)}$$

where τ sets the timescale. The number of motifs reaches its stationary (equilibrium) value when $x_m = x_m^{\text{eq}}$ at which point the selective force is balanced by entropic forces which randomize sequences. For the CpG motif, we set x_m^{eq} according to the value in the first line in Table 1.

It is convenient to express Equation 4 as

$$\tau \frac{dx_m}{dt} = -\left(-x_m(t) + x_m^{eq}\right)\mathrm{var}(x_m|N_m) \quad \text{(Equation 5)}$$

where $\mathrm{var}(x_m|N_m)$ is the variance of x_m for a given N_m.

If we can express $\mathrm{var}(x_m|N_m)$ as a function of x_m, it is possible to obtain a solution of Equation 5 that can then be fitted to the dataset with timescale τ, thus providing the approximation of relaxation dynamics, along with the estimate of the time it will take (and hence the number of the corresponding sequence motifs m) to reach its equilibrium value. For the case of HSATII and LINE-1 we fit $\mathrm{var}(x_m|N_m)$ as a quadratic function of x_m.

Kimura-based model of population genetics for the evolution of sequence motifs

In addition to the force relaxation model introduced above, we present an approach to study the neutral evolution of nucleotide sequence motifs based on the Kimura model of sequence evolution. We implement the model numerically and evolve a set of sequences to provide a null model of neutral sequence evolution. For each simulation step, we pick a random base and mutate it to a randomly chosen different base with a given probability. We consider different possible mutation probabilities depending on the type of base it is mutating into, as transversion (purine mutating to pyrimidine or vice versa) and transition (purine mutating to purine or pyrimidine mutating to pyrimidine) substitutions in sequences have different likelihood.[113]

Additionally, in vertebrates and plants, mutations in CpG context are known to be more common due to CpG hypermutability.[114] Hence, in the model implementation, we use different ratios of mutation rates μ_{TiCpG}: μ_{TvCpG}: μ_{Ti}: μ_{Tv} (corresponding to nucleotide transitions and transversions in CpG context and to transitions and transversion in non-CpG context). In particular, we consider previously introduced ratios,[113] which are listed in Table 1. The CpG equilibrium force can be computed analytically from the neutral relaxation in the Kimura model. As the transition-transversion bias does not affect the number of CpGs at equilibrium we can write for the evolution of the number of CpG in a sequence

$$N_{\mathrm{CpG}}^{t} = N_{\mathrm{CpG}}^{t-1} - 2\mu N_{\mathrm{CpG}}^{t-1} + \frac{1}{3}\gamma\left(N_{\mathrm{C}}^{t-1} - N_{\mathrm{CpG}}^{t-1}\right) + \frac{1}{3}\gamma\left(N_{\mathrm{G}}^{t-1} - N_{\mathrm{CpG}}^{t-1}\right) \quad \text{(Equation 6)}$$

while the number of C nucleotides N_{C} evolves as

$$N_{\mathrm{C}}^{t} = N_{\mathrm{C}}^{t-1} - \mu N_{\mathrm{CpG}}^{t-1} - \gamma\left(N_{\mathrm{C}}^{t-1} - N_{\mathrm{CpG}}^{t-1}\right) + \frac{1}{3}\gamma\left(L - N_{\mathrm{C}}^{t-1} - N_{\mathrm{CpG}}^{t-1}\right) + \frac{1}{3}\mu N_{\mathrm{CpG}}^{t-1} \quad \text{(Equation 7)}$$

and similarly for the number N_{G} of G nucleotides. In these equations, μ is the probability of a substitution happening in a CpG context, and γ is the probability of the mutation happening outside a CpG context, and L is the length of the sequence. At equilibrium, we find

$$\frac{N_{\mathrm{CpG}}}{L} = \frac{1}{14r+2}$$

$$\frac{N_{\mathrm{C}}}{L} = \frac{N_{\mathrm{G}}}{L} = \frac{3r+1}{14r+2} \quad \text{(Equation 8)}$$

where $r = \mu/\gamma$.

We can now compute the corresponding CpG force x_{CpG}^{eq} using the fact that the force is approximately equal to the logarithm of the relative frequency of the dinucleotide motif $x_{\mathrm{CpG}} \approx \log(f(\mathrm{CpG})/(f(\mathrm{C})f(\mathrm{G})))$[115]. The ratios 40:10:4:1, 40:4:4:1 and 40:1:4:1, provide the closest approximation to relaxation to the force observed in the genome. For the neutral model, we used the 40:10:4:1 ratio as it was closer to the saturated value of x_{CpG} of the LINE-1 elements.

For the schematic figures in the main text (3A and 5A), we simulated the neutral model with the 40:10:4:1 ratio. To illustrate the force dynamics corresponding to the preservation of sequence motifs, we ran a Kimura model simulation with a specified number of motif occurrences held fixed (200 CpG sites and a 200bp inverted segment, respectively), so that the mutation rate at these sites is effectively zero. This corresponds to a large, but inhomogeneous selective pressure on those specific prespecified motifs. For each sequence in the simulation, we compute the Kimura distance from the initial sequence and the motif conjugate force.

Analysis of forces across species

Values of x_{CpG} or x_{ds} were computed for each 3000 kb sliding window of the genomes of the following species: Pan troglodytes troglodytes, Pan paniscus, Gorilla gorilla gorilla, Pongo abelii, Nomascus leucogenys, Canis lupus, Danio rerio, Mus musculus, Rattus norvegicus, Equus caballus, Bos taurus, Gallus gallus, Felis catus, Pteropus vampyrus, Caenorhabditis elegans, Saccharomyces cerevisiae, Meleagris gallopavo, Erinaceus europaeus, and Ornithorhynchus anatinus, in addition to humans.

To compute the distribution of a force across all windows, the values of that force were sorted numerically and every 100th entry was retained (entry number 50, 150, 250, …, etc.). Non-numeric values were excluded. Windows with one or more ambiguous characters were excluded. The distribution density was computed using a Gaussian kernel with bandwidth 0.05 as implemented in scikit-learn package.[116] The density was computed for all points within the target interval ([-5:2] for x_{CpG}, and [-2:3] for x_{CpG}) with a step of 0.005. We compute the FDR as the ratio of the cumulative area of the null model distribution to the right of the cutoff to the cumulative

area of the distribution to the right of the cutoff. The null model distribution is fitted as a Gaussian distribution with the peak at the point with the maximal density. The standard deviation was computed using the 20 points to the right of the peak.

For humans and mice, we also performed an additional analysis which excluded both *cis*-regulatory regions, enhancers and CpG islands. Coordinates of enhancers from the FANTOM database were lifted from hg19 to hg38, and mm9 to mm38 using the LiftOver program and chain files from UCSC.[71,72,117] Coordinates of CpG islands for hg38 and mm38 were downloaded from UCSC (https://genome.ucsc.edu). "Filtered" data for hg38 and mm38 in Extended Data Figure S9 consist only of the windows which have zero overlap with CpG islands and enhancers. Coordinates of *cis*-regulatory regions were downloaded from the ENCODE portal (https://www.encodeproject.org/) for hg38 and mm10.[118]

These regions (FANTOM database, ENCODE database, CpG islands) were merged. Values of dsRNA and CpG forces were then computed for each window. Non-numeric values were excluded. For CpG force windows with at least one ambiguous character were excluded. Sliding windows the forces were computed for were split into overlapping (by any non-zero number of bases) with *cis*-regulatory regions, enhancers or CpG islands and non-overlapping windows.

Analysis of evolutionary conserved sequences with high x_{CpG} or x_{ds}

We considered the genomes of 5 species (Pan troglodytes troglodytes, Pan paniscus, Gorilla gorilla gorilla, Pongo abelii, Nomascus leucogenys) in addition to the human genome to look for conserved regions with high x_{CpG} or high x_{ds}. After computing both forces for each species exactly as we did for the human genome, we considered the set of $x_{CpG} > 0$ and $x_{ds} > 0.5$. We reduced the number of the CpG windows by clustering together all those which overlapped more than 1000 bases, and from each cluster we only considered the window with the highest value of x_{CpG}. For the x_{ds} windows, we first excluded windows with complementary sequences, as being related sequences this would lower the complexity. For each window, we then extracted the subsequence spanning the pair of complementary sequences. Finally, we clustered together all overlapping sequences and from each cluster we only considered the window with the highest value of x_{ds}.

We then ran the Basic Local Alignment Search Tool (BLAST) to compare each of these sequences with high x_{CpG} or high x_{ds} extracted from the human genome and we retained any significant match (whenever one sequence of a given organism matched with more than one human sequences we only kept the match with the highest BLAST score).[119] The result of this procedure consists of two sets of sequences for each organism that are alignable to human sequences, one for the high x_{CpG} and one for the high x_{ds}. We then computed the overlaps between the set A of high x_{CpG} (or high x_{ds}) organism sequences and that of the human, B, defined as

$$O(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}, \quad \text{(Equation 9)}$$

where a sequence of A belongs to the intersection $A \cap B$ if and only if it is alignable with significant score by BLAST to a sequence of B. $|A|$ is the size of the set A.

To plot the overlaps versus the time since the most recent common ancestor, we have taken the latter from,[75] obtaining 6.6 million of years ago for Homo Sapiens, Pan troglodytes troglodytes and Pan paniscus; 8.9 for Homo Sapiens, Pan troglodytes troglodytes, Pan paniscus and Gorilla gorilla gorilla; 13.8 for Homo Sapiens, Pan troglodytes troglodytes, Pan paniscus, Gorilla gorilla gorilla and Pongo abelii; and 16.3 for all the primate species considered here.

Afterward we focused on the annotations in RepeatMasker[120] and Dfam[121] databases to check for potential overlaps of the conserved sequences with annotated repeats in the human genome. For the set of high-CpG sequences we used the Dfam dataset as we found a better annotation for HSATII with respect to the one in RepeatMasker. We associated each conserved high-CpG sequence with a repeat if the corresponding window overlapped with its position as annotated in the database (for the windows overlapping with more than one annotated repeats, the one with the largest overlap was considered).

For the set of x_{ds} sequences we used the RepeatMasker dataset. In this case, each x_{ds} window is characterized by two fully complementary sequences, and we searched for repeats overlapping with each of them (when overlaps with multiple repeats were found, the one with the largest overlap was considered). We observed 3 different cases: one or two of the two sequences do not overlap with any repeat annotated in RepeatMasker; each sequence overlapped with a repeat, both being of the same family and annotated in the two strands of the genome, e.g., one sequence overlapping with AluS+ and one with AluS- (IR repeats); each sequence overlapped with a repeat, but of different families or of the same family but on the same strand of the genome (non IR repeats).

Sequence ensembles

The LINE-1 sequences were obtained from L1Base2 database.[44] We separately downloaded all the sequences annotated as full-length intact and hence are more likely to still be active (146 for human genome and 2811 for mouse genome), and sequences annotated as full-length non-intact (13148 for human genome and 14076 in mouse genome). Full-length intact sequences contain both an intact open reading frame and "the two ORFs, 5′ UTR-located internal promoter and 3′ UTR regions LINE-1s" (https://l1base.charite.de/l1iterature.php#l1s). We separately aligned each of the non-intact sequences with each of the respective intact sequences using pairwise alignment and calculated the Kimura distance between the sequences.[122] We then calculated the average distance for each of the non-intact sequences from the intact-sequences, and furthermore calculated the number of CpG motifs in each sequence.

Sequences of all inserts of HSATII and all other Human Genome repetitive elements considered in this work have been obtained from the Dfam database[64] (version introduced in 2016). Each family of sequences in the Dfam database contains sequences of all its inserts in the human genome and their consensus sequence, as well as with the hidden Markov Chain Model (HMM) that we use to align inserts with respect to the consensus sequence. For comparison of sequences of inserts with respect to their consensus sequence, we only consider inserts of length longer than 150 bases. To quantify the difference between the insert sequence and the consensus sequence, we use the Kimura distance[122] between the consensus and its inserts.

We note that we use the Kimura distance[123] from the consensus sequence (for inserts from Dfam) or from average of all full-length non-intact sequences (for LINE-1s from L1Base2) as a measure of time, assuming that it is proportional to the time since insertion of the particular transposable element into the species genome. All the sequences studied in this work have been obtained from hg38 genome assembly.

Search of long transcripts with complementary regions

We scanned the hg38 genome assembly for transcripts that can be possible source of long duplex formation. To this aim, for each window of length 3000 bases (taken in the positive sense of the read), we calculate the double-stranded force x_{ds} from

$$N_{ds}^{obs}(L, x_{ds}) \approx \frac{\log L}{\log \frac{1}{\sqrt{\exp(x_{ds})\alpha}}} + c,$$

using the window-specific nucleotide frequencies to obtain α as further defined in the Supplementary Methods.

We considered windows resulting in $x_{ds} > 0.5$ as having a high double-stranded force when compared to the rest of the genome.

Sequence complexity quantification

We use an approximation of Kolmogorov complexity[124] to quantify how "non-trivial" complementary segments are. Adopting an approach,[125] we use the size (in bytes) of the sequence compressed with gzip software as a proxy of the Kolmogorov complexity. Simple sequences, e.g.,*e.g.* poly(AT) or poly(C) and poly(G), will have low complexity, as they can be compressed to a smaller size than a completely random sequence of the same length (which would have maximum complexity).

Estimate of genome regions with high double stranded force

To estimate x_{ds} for a given repeat loci, we intersect each repeat loci with the calculated 3kb genomic windows that have high dsRNA forces ($x_{ds} > 0.5$). The Start and End coordinates of the corresponding dsRNA sequence pairs, which overlap with the repeat loci that match the criteria: $|\log2FC(\text{treated/untreated})| > 0.5$ and FDR < 0.05, were used to annotate different genomic features. We counted the genomic features of the predicted double-stranded RNA sequences that overlap with the upregulated repeats (log2FC > 0.5 and FDR < 0.05), and of those that overlap with the downregulated repeats (log2FC < −0.5 and FDR <0.05). These counts have been compared with the genomic feature counts of all dsRNA sequences that overlap with the transcribed repeats to calculate the odds ratio and p value using the Fisher Exact test.

Analysis of repeats for splice inhibitors

Raw RNAseq data (GEO: GSE95011) associated with the Seiler et al. study[126] was downloaded from NCBI. Briefly, reads were trimmed and quality checked using skewer and then mapped to the human genome (hg38) and repetitive elements from RepBase.[127,128] In quality check, Illumina reads were trimmed to remove N's and bases with quality less than 20. After that, the quality scores of the remaining bases were sorted, and the quality at the 20th percentile was computed. Reads quality less than 15 at the 20th percentile or shorter than 40 bases were discarded. Only paired reads that passed the filtering step were retained, and mapped to the reference genome (hg38) using STAR (v2.7) with default parameters.[102] Gene counts were assigned based on Gencode annotation using featureCounts (Subread package) with the external Ensembl annotation. Repeat counts for a given subfamily were primarily quantified against RepeatMasker using featureCounts and then adding the counts of the unassigned reads that mapped to the Repbase consensus sequence. The repeat counts of a given family is therefore the sum of mapped reads to RepeatMasker and unmapped reads against Repbase.

Analysis of repeats for melanoma cohort

Raw RNAseq data (GEO: GSE98394) associated with the Badal et al.[86] study was downloaded from NCBI. Reads were mapped to hg38 using STAR[102] (v2.7) with parameters set as described in Software Quantifying Interspersed Repeat Expression.[103] Kaplan-Meier survival curves were generated using the *survfit* function from the R survival package, with the significance of differences in survival between high and low expression groups evaluated using the log-rank test via the *survdiff* function. Double-stranded forces were calculated for inverted SINE elements near an ancohred SINE element.[26] Mean expression levels of SINE loci with high dsRNA force and fpkm > 30 were calculated to stratify samples.

Counts filtering, normalization and statistical analysis

Gene expression in terms of log2-CPM (counts per million reads) was computed and normalized across samples using the TMM (trimmed-mean of M-values) method using the calcNormFactors() and cpm() functions from edgeR.[129] These low-count values (CPM < 2) were removed before calculating the size factor for each sample. Filtered CPM was log2 transformed and used in heatmap visualization and downstream statistical analysis. On the heatmap, repeats (rows) were scaled by *Z* score scaling. Differential expression analysis was performed using limma[130] between splicing modulator H3B-8800 treated versus DMSO treated SF3B1-K700 mutated cell line k562 for a given locus. The adjusted p-values were calculated using the Benjamini-Hochberg correction.[131]