Repeats Mimic Immunostimulatory Viral Features Across a Vast Evolutionary Landscape

Petr Šulc, Alexander Solovyov, Sajid A Marhon, Andrea Di Gioacchino, Siyu Sun, John Lacava, Omar Abdel-Wahab, Nicolas Vabret, Daniel D de Carvalho, Rémi Monasson, et al.

Repeats Mimic Immunostimulatory Viral Features Across a Vast Evolutionary Landscape

Petr Šulc*,[1], Alexander Solovyov[*,2], Sajid A. Marhon[*,3], Andrea Di Gioacchino*,[4], Siyu Sun[2], John LaCava[5,6], Omar Abdel-Wahab[7,8], Nicolas Vabret[9], Daniel D. De Carvalho[#,3,10], Rémi Monasson[#,4], Simona Cocco[#,4], Benjamin D. Greenbaum[#,2,11]

[1] School of Molecular Sciences and Center for Molecular Design and Biomimetics, The Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA
[2] Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[3] Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1L7, Canada
[4] Laboratoire de Physique de l'Ecole Normale Supérieure, PSL & CNRS UMR8063, Sorbonne Université, Université de Paris, Paris, France
[5] Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, NY 10065, USA
[6] European Research Institute for the Biology of Ageing, University Medical Center Groningen, Groningen, The Netherlands
[7]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA
[8]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA
[9] Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10065 USA
[10] Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
[11] Physiology, Biophysics & Systems Biology, Weill Cornell Medicine, Weill Cornell Medical College, New York, NY 10065, USA
[*] Denotes Equal Contributions
[#] Denotes Senior Author
Correspondence: psulc@asu.edu, simona.cocco@lps.ens.fr, greenbab@mskcc.org

ABSTRACT

An emerging hallmark across many human diseases - such as cancer, autoimmune and neurodegenerative disorders – is the aberrant transcription of typically silenced repetitive elements. Once transcribed they can mimic pathogen-associated molecular patterns and bind pattern recognition receptors, thereby engaging the innate immune system and triggering inflammation in a process known as "viral mimicry". Yet how to quantify pathogen mimicry, and the degree to which it is shaped by natural selection, remains a gap in our understanding of both genome evolution and the immunological basis of disease. Here we propose a theoretical framework that combines recent biological observations with statistical physics and population genetics to quantify the selective forces on virus-like features generated by repeats and integrate these forces into predictive evolutionary models. We establish that many repeat families have evolutionarily maintained specific classes of viral mimicry. We show that for HSATII and intact LINE-1 selective forces maintain CpG motifs, while for a set of SINE and LINE elements the formation of long double-stranded RNA is more prevalent than expected from a neutral evolutionary model. We validate our models by showing predicted immunostimulatory inverted SINE elements bind the MDA5 receptor under conditions of epigenetic dysregulation and that they are disproportionately present during intron retention when RNA splicing is pharmacologically inhibited. We conclude viral mimicry is a general evolutionary mechanism whereby genomes co-opt features generated by repetitive sequences to trigger the immune system, acting as a quality control system to flag genome dysregulation. We demonstrate these evolutionary principles can be learned and applied to predictive models. Our work therefore serves as a resource to identify repeats with candidate immunostimulatory features and leverage them therapeutically.

MAIN TEXT

The ability to predict the presence of patterns sensed by the innate immune system is of considerable theoretical and practical interest[1]. For instance, mathematical models of the evolution of human H1N1 influenza since the 1918 pandemic showed an attenuation of CpG motifs, leading to the prediction such motifs are targeted by pattern recognition receptors (PRRs)[2,3], and trigger pro-inflammatory responses. It was subsequently discovered that the protein ZAP (*ZC3HAV1*) is a PRR targeting CpG motifs, indicating inferences drawn from genome evolution can predict new receptor specificities relevant to emerging and adapting viruses[4,5], including SARS-CoV-2[6]. It has been more difficult to predict PRR specificities from structure prediction. There are multiple receptors known to recognize long and short double stranded RNAs (dsRNAs). For example, MDA-5 (*IFIH1I*) recognizes long dsRNA segments present during RNA virus replication and TLR-3 recognizes shorter segments, on the order of tens of base pairs[7]. Surprisingly, it recently became clear that repetitive elements, which represent most of the human genome and may derive from integrated viruses, can display “non-self" pathogen-associated molecular patterns (PAMPs). Under aberrant conditions such as in cancer[8], repeats are frequently overexpressed, where they may display PAMPs, such as anomalous CpG content and dsRNA[9-14]. Consistently, a growing body of literature has demonstrated the aberrant expression of immunostimulatory repeats across an array of human diseases, such as in aging[15] and autoimmunity[16], implying “viral mimicry” may be a fundamental feature of inflammatory diseases. Moreover, viral mimicry can be leveraged therapeutically: the expression of immunostimulatory repeats is inducible by epigenetic drugs, leading to the triggering of innate sensors and induction of an interferon response[10-14].

Several fundamental questions remain, such as which human sensors can be activated by which repeats, if viral mimicry serves a functional role in the genome as an evolved checkpoint for loss of epigenetic regulation or genome fidelity, and whether tumors and pathogens have learned to manipulate mimicry to their own selective advantage[17,18]. In one evolutionary scenario, repeats which form features in somatically silenced, low-complexity regions can create PAMPs that offer a fitness advantage to cells due to their ability to trigger PRRs under epigenetic stress, eliminating dysregulated cells and maintaining tissue homeostasis[17,18]. Such features would then be maintained by natural selection. Alternatively, in a neutral scenario, it may be that high RNA concentration resulting from dysregulation can engage PAMPs non-specifically, and their sensing is a convenient byproduct of dysregulation rather than selection acting on specific sequence features. Discriminating between these scenarios is key to understanding how non-self mimicry by the self-genome has evolved, and how it can be leveraged for emerging therapies and honed for existing ones. There is therefore a pressing need for new approaches to quantify the presence of viral mimics, infer parameters defining their immunological features, and quantify their evolutionary dynamics in this reduced feature space. We propose a theoretical approach to quantifying immunostimulatory nucleic-acid motifs and double-stranded structures under selection, and present two models for describing the evolutionary dynamics of an immunological feature generated by repeats. In doing so we define specific categories of repeat families that most likely were retained by natural selection to trigger specific receptors of the innate immune system under aberrant conditions.

Inference and evolutionary dynamics of immunostimulatory features

We generalize the framework of selective and entropic forces to infer anomalous sequence features[3]. In our approach, genome segments, subject to constraints such as local nucleic acid content, are randomized by entropic forces to resemble, on average, self-genomic material and are ordered by selective forces acting on sequence features to oppose such randomization. Rather than using p-values to compare the strength of avoidance or enhancement of a certain candidate immunostimulatory feature, the "selective force" is an intensive parameter that can be readily compared between sequences and is easily interpretable as the information theoretic cost of avoiding or enhancing specific features in a sequence. To calculate selective forces, one uses exact transfer matrix methods from statistical physics which, unlike previous approaches[2], are computationally efficient (scaling with the length of the sequence) and facilitate the analysis of longer sequences and large databases. We calculate the degree to which any sequence displays a feature bias (as defined in Methods). To apply this formalism to the evolutionary dynamics of immunostimulatory features, we use this parameter for two approaches to study the population genetics of immunostimulatory features in an ensemble of genome sequences.

The first approach uses relaxation dynamics for the evolution of repeats in the genome. In this formalism, a new repeat with a force on an immunostimulatory feature will evolve until its force value reaches an equilibrium determined by the specificity of PRRs in its host. For an analogy, the 1918 H1N1 influenza virus had one set of features in its original avian host, and then evolved towards a new equilibrium in humans, where PRRs target CpG with greater affinity and therefore exert a greater selective force[3]. The second approach uses a Wright-Fisher (WF) model that considers the evolution of the probability of a sequence with given immunostimulatory feature content[19]. While relaxation dynamics was applied to the evolution of dinucleotides motifs under selective pressure in viral genomes[3], here we connect selective forces to intrinsic molecular mutational processes in human genomes by use of population genetics (Methods). We implement the WF model numerically and evolve, assuming haploid reproduction of sequences, a set of sequences according to a neutral mutation model without a selection term to provide a null model of repeat evolution in the human genome. For each simulation step, we pick a random base for each sequence in the ensemble 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 being mutated into, as well as on the local nucleotide context. Additionally, in vertebrates and plants, mutations in CpG context are known to be more common due to methylation induced hypermutability[20]. Hence, we use different ratios of mutation rates corresponding to nucleotide transitions and transversions in a CpG context and to transitions and transversion in non-CpG context[20]. We calculated the dinucleotide distribution stationary value, obtained as the stationary vector of the stochastic matrix with entries corresponding to probabilities of mutating from one dinucleotide to another dinucleotide (see Methods and Table 1).

Landscape of repeats with selective forces on CpG dinucleotides

A repetitive element is primarily defined by the presence of multiple copies (inserts) of its sequence. We compare the evolution of dinucleotide motifs (quantified by calculating the selective force, x_s, on a dinucleotide motif, s, as defined in Methods) between the original consensus sequence, representing the sequence most likely to be close to the founding ancestral insertion, and its subsequent copies in the genome (Fig. 1). We analyzed all repeat families annotated in the DFAM database and calculated the dinucleotide forces for their consensus sequences as well the mean force on all inserts from a given family[21], finding outliers such as a set of Alu repeats and HSATII, the later consistent with previous results[9] (Fig. 1). The greatest differences between the forces on dinucleotides for a consensus sequence and its subsequent inserts were observed for CpG (Fig. 1A). For all other dinucleotides, the force change with respect to the consensus is approximately 0, as illustrated in Supplementary Figure 1. Typically, CpG content in the human genome is highly

underrepresented (Extended Data Fig. 1) and CpG sites mutate at a much faster rate than the rest of the genome due to their aforementioned hypermutability[20,22-23]. As a result, understanding whether the CpG content of a repeat has “relaxed" to a typical level or is held fixed by selection can indicate whether a repeat transcript can be recognized by a PRR that senses CpG motifs. We evaluated the mean force for all other annotated repeats longer than 150 bases. We plot the mean difference in CpG force per repeat family versus the CpG force of the consensus ancestral insert (Fig. 1B). Consistently, we see that families where the selective force on CpG dinucleotides for the progenitor insert was greater than −1.9 have decreased their force to this value, while those less than −1.9 have increased their value. We therefore establish a genome-wide equilibrium in line with equilibria observed for human adapted viruses such as influenza and SARS-CoV-2[2,2,6]. If a repeat is not subject to selection, one would expect its insertion to evolve according to a WF model with respective mutation rates for transitions and transversions. This approach has been used in several sequence evolution models to explain lower CpG content in vertebrate genomes[24-26]. However, CpG motifs are also functional. Methylation of CpGs in DNA is an important regulator of gene expression[27,28], and CpG-rich RNA can have immunostimulatory properties[9]. Therefore, one could expect selection to act against depletion of functional CpG motifs, as observed in CpG islands located in gene promoters of vertebrates[29]. Indeed, most repeat families show relaxation to the mean genome force expected from the neutral model, further implying HSATII and Alu repeats may be specifically under selection to trigger PRRs (Fig. 1C).

As LINE-1 elements have the most copies in the genome, they are most amenable to our approach. They are estimated to constitute about 20% of human genome[30]. Here we only consider full-length inserts, as annotated in L1Base2, and contrast those designated as fully intact (denoted FLI), from those full-length sequences designated as non-intact (FLnI)[31]. Fully functional LINE-1 DNA sequences are regulated by promoter hyper-methylation, which occurs at CpGs, to inhibit their transcription[32]. Indeed, we find FLI LINE-1 have higher CpG content than FLnI (Fig. 2A). We calculated the mean Kimura distance[33] to all FLI sequences for each of the FLnI sequences as proxy for time since insertion, finding that as a LINE-1 genome insertion ceases to contain an intact copy, its CpG content decays to the genome mean in a predictable way (Fig. 2B), reaching a plateau of -2.0, within the margin of error for the equilibrium of -1.9. We would expect the most recent inserts into the human genome to not have equilibrated. It is important to identify all such cases because the families that have not saturated are candidates for viral mimicry such as, for example, when overexpressed in tumors[8,35-37]. The clearest instance is HSATII. The evolutionary dynamics of the force relaxation fit for HSATII (Fig. 2B) corresponds to saturation at force approximately equal to −0.4, well above the equilibrium distribution given by the WF model simulations (Fig. 2B, green line), implying its ability to stimulate PRRs is maintained by selection. Other outliers comprise repeat families that are still close in age to the original CpG-rich insert or families whose CpG force is decreased at lower rate than observed for other repeat families, implying its features are maintained by selection. For most of families the data points are scarce and noisy, making a relaxation fit such as the one shown for HSATII and LINE-1 difficult. The full genome atlas of CpG-rich repeat families is listed in the Supplementary Table 1 and the distribution of anomalous CpG hotspots is show in Fig. 2C, showing an enhancement in introns and depletion in intergenic regions (Fig. 2D). Most hotspot loci have a Kimura distance from the consensus of less than 0.1 and belong to Alu subfamilies, these species likely maintain their anomalous sequence features due to being evolutionary young compared to the founding member of their repeat family. Other families besides HSATII with higher-than-average Kimura distance from the consensus larger are MER21, TAR-1, and LTR6B families, which may have CpG dinucleotides maintained by selection to trigger PRRs in a dysregulated state.

Landscape and evolution of repeats with selective forces on double-stranded RNA formation

We extend our approach to the evolution of repeats that can trigger PRRs via double-stranded RNA (dsRNA) formation. Known dsRNA receptors include TLR-3, RIG-I, and MDA-5[7]. While the detailed mechanism of dsRNA motif recognition and receptor activation are still a subject of active research, it is generally accepted that TLR-3 is activated by short (approx. 30 bp) endosomal dsRNA and RIG-I (*DDX58*) by short (tens of bases) cytoplasmic dsRNA accompanied by a triphoshphate[37], while MDA-5 recognizes longer cytoplasmic dsRNA[38]. We study the distribution of double-stranded segments in annotated regions in human genome, quantified by the double stranded force, x_{ds} (Methods). It is analogous to forces on dinucleotide motifs, where $x_{ds} = 0$ if, for a given sequence, the length of its longest complementary segments corresponds to what one would expect from a neutral model of a random sequence with the same nucleotide distribution and length. We quantified x_{ds} for repetitive families as well as ncRNA and mRNA sequences. The histogram of observed double-stranded forces is shown in Fig. 3A, along with a histogram of randomly generated sequences of different lengths. While the mean value and standard deviation of functional mRNA and ncRNA sequences is essentially random, the consensus sequences of repeats contain multiple families with long complementary segments, contributing to an increased average x_{ds} value (Fig. 3A). Such repeats therefore entered the genome with the potential ability to form dsRNA segments and, as with CpG motifs, typically lost that ability over time due to mutations. While the general trend is to relax the double-stranded force towards zero (Fig. 3B), there are several repeat families with large x_{ds} values, indicating a possible reservoir of double-stranded segments being maintained by selection (Fig. 3C, Extended Data Fig. 2A). Many of these families were not detected by the selective force on CpG dinucleotides, implying the selective forces on dinucleotides and RNA structures are largely independent and detected by distinct PRRs. Several outliers have a high positive x_{ds} values, including the species Tigger4a and HSMAR (Extended Data Fig. 2B). While they are DNA transposons, we found also their RNA transcripts in The Cancer Genome Atlas (TCGA - https://www.cancer.gov/tcga), and hence their RNA may still be immunostimulatory when transcribed.

To locate possible sources of double-stranded segments originating from the same transcript, we scan the entire genome (HG38 assembly), using a window of transcripts of length 3000bp, comparable to typical lengths of long ncRNAs[39]. We scan these windows for two fully complementary segments (through Watson-Crick or wobble base pairs). We quantified the sequence complexity of such complementary segments (based on Kolmogorov complexity, as described in Methods), as shown in Fig. 3D. The segments close to the low complexity limit typically contain a repeating motif of only a few nucleic acids (such as poly(AT)) while the longest segments have higher complexity, i,e. the long dsRNA are not exclusively being formed by simple repeats. The longest inserts with high complexity correspond to segments that do not overlap with any known insert, annotated gene or ncRNA. An atlas of all families of repeats analyzed are summarized in Supplementary Table 2&3.

We specifically explored which specific genome loci, as opposed to consensus repeats, can stimulate MDA-5 receptors by forming long dsRNA segments, as their transcription has been implicated as a response to genome-wide DNA demethylation[10]. Using a sliding window of the entire human genome, with transcript length of 3000bp, we observed the two x_{ds} peaks, a major one close to 0 and a smaller around 0.5 (Fig 4A), consistent with the results for consensus repeats found in Fig. 3. We found that for the majority (74%) of regions with $x_{ds} > 0.5$ the complementary segments in the 3000 bases long regions overlap with known repeats. Greater than 90% of identified complementary segments correspond to AluS and AluY, two inserts from Alu families, where a copy has inserted in a positive orientation close to one in a negative orientation (inverted-repeat Alus IR-Alus) (Fig 4B). These results, based solely on evolutionary analysis using our framework, are strikingly predictive of the experimental observations that IR-Alus are the major source of self-RNA that form MDA-5 agonists[10]. To test this hypothesis, we plotted a histogram of the transcripts found experimentally in Ref. 10 to

bind MDA-5, both at baseline and after treatment with a DNA demethylating agent (Fig. 4A). Those experimentally validated MDA-5 agonist dsRNAs indeed have a clear x_{ds} peak at 0.5 (Fig. 4A), providing strong experimental support to the predictive power of our evolutionary model and, in turn, the hypothesis that evolution selected this feature as an epigenetic checkpoint[17,18]. The mean length of the longest complementary segments found in the dataset with x_{ds} > 0.5 is 40 base pairs. We further investigated a subset, consisting of regions that can form 100 base pairs or longer double-stranded segments. In this subset, only 20% of the complementary segments overlap with known annotated repeat segments. Besides the Alu subfamilies (which constitute about 40% of long complementary segments that overlap with known inserts), we also identified complementary fragments inserted in their positive and negative orientation from the ORF2 open reading frame of LINE-1, which is lowly expressed compared to ORF1, the other LINE-1 open reading frame, in human cancers.[40]

In addition, we observed most regions (56.9%) with $x_{ds} > 0.5$ were over-represented at intronic regions (Fig 4C). These results are consistent with a recent hypothesis that intronic repeats can form dsRNA and induce viral mimicry as a checkpoint against intron retention[41,42]. We therefore hypothesized that predicted repeats with high dsRNA force would be disproportionately present when introns are retained as a checkpoint against splicing abnormalities[18]. To test this hypothesis, we analyzed the effects of a class of inhibitors of RNA splicing which induce intron retention and exhibit synthetic lethal interactions in cancers with mutations in RNA splicing factors such as SF3B1[43]. We examined RNA sequencing data from SF3B inhibitors (including the drugs E7107 and H3B-8800) which cause the retention of introns in SF3B1 K700E mutant cells. Consistent with our model, we found splicing agents which lead to intron retention over express the high double-stranded force intronic repeats we predicted (Fig. 4D-E), simultaneously supporting the evolutionary role of inverted SINE elements in guarding against intron retention and the potential ability to manipulate this feature using a cancer therapeutic targeting RNA splicing. Consistently, for inhibitors less associated with intron retention the effect was either weakened or not present (Extended Data Figs. 3&4).

Finally, we annotated long dsRNA segments formed by bidirectional transcription, which have been implicated as potentially forming dsRNA due to their perfect complementarity[12]. To find plausible sources of regions that can be transcribed in both directions, we analyzed available transcription datasets from TCGA (Methods), finding multiple regions with long (over hundred base pairs) regions that are transcribed bidirectionally, indicating a possible source of antagonists (Extended Data Fig. 5). We found different inserts of MIR, Alus and LINE-1, i.e., some of the most abundant repeat families, to be the most represented among such transcripts. The respective loci for the top 1% highest bidirectional transcript counts, along with the number of reads transcribed from either the negative or positive strand, are listed in Supplementary Table 4.

DISCUSSION

We quantify the evolution of non-self, pathogen-associated patterns, based on competition between selective and entropic forces, within repeat families in the human genome. In doing so we find the high-copy satellite RNA HSATII is likely under selection to maintain its pathogen-associated CpG dinucleotide content and functional LINE-1 inserts maintain higher CpG content than expected. LINE-1 promoters are controlled at the DNA-level by CpG methylation, and it has an internal, bi-directional promoter transcribed with the 5'-UTR of the RNA[44,45], ensuring the promoter co-mobilizes with the protein coding regions. HSATII may have a DNA regulatory function as well, as its DNA sequences can sequester chromatin regulatory proteins and trigger epigenetic change[46]. However, at the RNA-level, CpGs can function evolutionarily as a danger signal to maintain fitness of tissues under epigenetic stress for both LINE-1 and HSATII, whose immunostimulatory properties have been documented.

Furthermore, we incorporate RNA secondary structure into evolutionary models and identify a reservoir of anomalous repeats with likely immunostimulatory dsRNAs. We attempt to exhaustively annotate regions where repeats evolutionarily maintain the ability to form long dsRNAs or present anomalous CpG motifs, providing an atlas for mapping transcriptomes of cells which exhibit stimulation of PRRs so one can identify the potential source of causal immunostimulatory self-transcripts. As strong validation of our approach, repeats predicted through evolutionary analysis to be dsRNA-forming were found to be MDA-5 agonists in a recently published MDA-5 protection assay that profiled ligands induced upon response to epigenetic cancer therapy by DNA demethylating agents[10]. The repeats that are induced by epigenetic therapy come from regions of the genome which may selectively maintain the ability to form dsRNA, implying the therapeutic condition mimics the evolutionary role of these RNA species to safeguard tissue homeostasis by killing dysregulated cells. Moreover, we find such repeats disproportionately arise within introns and can be disproportionately induced by intron retaining splice-inhibitors[43], where they may be localized as a checkpoint against intron retention[18,41,42]. Furthermore, CpG sequences may make intronic repeats better targets for RNA-binding proteins – without such insulation, repetitive elements within the introns of protein-coding genes could lead to deleterious RNA processing, which is ultimately relieved as the elements age by (presumably neutral) mutational decay[47,48].

Our work therefore has several implications for how we understand self versus non-self discrimination. When one quantifies pathogen-associated features, specific repeats in the genome not only display PAMPs capable of stimulating PRRs but, in some instances, seemingly maintain such features under selection. For multicellular organisms with a high degree of epigenetic regulation and chromosomal organization, this offers an opportunity to maintain stimulatory features to release a danger signal when epigenetic control is lost, such as during the release of repeats after p53 mutations, where immunostimulatory repeats may offer a back-up for p53 functions such as senesence[12,49]. Our work supports the hypothesis that repeats are selected to maintain "non-self" PAMPs to act as sensors for loss of heterochromatin as an epigenetic checkpoint of quality control system and avoid genome instability generally[17,18]. With our framework one may learn how to identify which pathogen-associated features the genome maintains, which receptors they ligate, and, thereby, learn what pathways the genome has evolved to agonize and when.

Specific genome repeats, such as HSATII and inverted SINE elements have been disproportionately implicated in the ability to stimulate non-self detection pathways and we predict that they are maintained under natural selection to do so. Each repeat likely engages a different receptor family. For CpG motifs the ZAP receptor and TLR7/8 have been implicated, and inverted SINE elements are likely detected by long dsRNA sensors such as MDA-5. Decoding viral mimicry by repeats using a combination of physically interpretable machine learning and predictive evolutionary models may

therefore shed light on the function of genomic “dark matter” across disease indications, in a manner which may be further exploited therapeutically. For instance, it had been observed that early-stage melanoma may manipulate epigenetic regulators to suppress immunostimulatory repeat expression, and recent work has shown the possibility of targeting those proteins to reinvigorate the immune response[50,51]. Furthermore, viruses and late-stage tumors may have learned to manipulate viral mimicry to their own advantage: Y-RNAs have been implicated in RIG-I sensing during RNA virus infection[52] and herpesviruses derive a fitness advantage from induction of HSATII, which is also often overexpressed in tumors[53]. The implication is that we can learn a “repeat code” of self-agonists within our genome held by selection to stimulate receptors under specific circumstances. We provide both an annotated atlas of predicted repeats under selection (Supplementary Tables) and software for building predictive models for this purpose. The lack of unbiased sequencing of repeats, which can easily be missed in RNA sequencing that focuses only on mRNA or in whole exome or short read whole genome DNA sequencing, is therefore a critical bottleneck. Once decoded we can better understand the evolution of these surprisingly non-self features encoded within families of repeats in our genome.

REFERENCES

1. Vabret, N., Bhardwaj, N. & Greenbaum, B.D. Sequence-specific sensing of nucleic acids. *Trends in Immunology* **38**, 53-65 (2017).
2. Greenbaum, B.D., Levine, A.J., Bhanot, G. & Rabadan, R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathogens* **4**, e1000079 (2008).
3. Greenbaum, B.D., Cocco, S., Levine, A.J., & Monasson, R. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proceedings of the National Academy of Sciences* **111**, 5054-5059 (2014).
4. Takata, M.A. et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124-127 (2017).
5. Stern, A. et al. The evolutionary pathway to virulence of an RNA virus. *Cell* **169**, 35-46 (2017).
6. Di Gioacchino, A. et al. The heterogeneous landscape and early evolution of pathogen-associated CpG dinucleotides in SARS-CoV-2. *Molecular Biology & Evolution* **38**, 2428-2445 (2021).
7. Jensen, S. & Thomsen, A.R. Sensing of RNA viruses: a review of innate immune receptors involved in recognizing RNA virus invasion. *Journal of Virology* **86**, 2900-2910 (2012).
8. Ting, D.T. et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593-596 (2011).
9. Tanne, A. et al. Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *Proceedings of the National Academy of Sciences* **112**, 15154-15159 (2015).
10. Mehdipour, P. et al. Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. *Nature* **588**, 169-173 (2020).
11. Roulois, D. et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* **162**, 961-973 (2015).
12. Leonova, K.I. et al. p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. *Proceedings of the National Academy of Sciences* **110**, E89-E98 (2013).
13. Chiappinelli, K.B. et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162**, 974-986 (2015).
14. Sheng, W. et al. LSD1 Ablation stimulates anti-tumor immunity and enables checkpoint blockade. *Cell* **174**, 549-563 (2018).
14. De Cecco, M. et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566**, 73-78 (2019).
15. Rice, G.I. et al. Reverse-Transcriptase Inhibitors in the Aicardi–Goutières Syndrome. *New England Journal of Medicine* **379**, 2275-2277 (2018).
17. Ishak, C.A. & De Carvalho, D.D. Reactivation of endogenous retroelements in cancer development and therapy. *Annual Review of Cancer Biology* **4**, 159-176 (2020).
18. Chen, R., Ishak, C.A. & De Carvalho, D.D. Endogenous retroelements and the viral mimicry response in cancer therapy and cellular homeostasis. *Cancer Discovery*, in press (2021).
19. Tran, T.D., Hofrichter, J. & Jost, J. An introduction to the mathematical structure of the Wright–Fisher model of population genetics. *Theory in Biosciences* **132**, 73-82 (2013).
20. Sved, J. & Bird, A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of* Sciences **87**, 4692-4696 (1990).
21. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**, D81-D89 (2016).
22. Shen, J.C., Rideout III, W.M. & Jones, P.A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Research* **22**, 972-976 (1994).
23. Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8**, 1499-1504 (1980).

24. Baele, G., Van de Peer, Y. & Vansteelandt, S. Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. *BMC Evolutionary Biology* **10**, 244 (2010).
25. Arndt, P.F. Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny. *Gene* **390**, 75-83 (2007).
26. Bérard, J. & Guéguen, L. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Systematic Biology* **61**, 510-521 (2012).
27. Robertson, K.D. & Wolffe, A.P. DNA methylation in health and disease. *Nature Reviews Genetics* **1**, 11-19 (2000).
28. Ziller, M.J. et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-481 (2013).
29. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes and Development* **25**, 1010-22 (2011).
30. Lander, E.S et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
31. Penzkofer, T. et al. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Research* **45**, D68-D73 (2017).
32. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227–234 (1997).
33. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111-120 (1980).
34. Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-971 (2012).
35. Tubio, J.M. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
36. Solovyov, A. et al. Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and T cell suppressive classes. *Cell Reports* **23**, 512-521 (2018).
37. Pichlmair, A. et al. RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. *Science* **314**, 997-1001 (2006).
38. Wu, B. et al. Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. *Cell* **152**, 276-289 (2013).
39. Novikova, I.V., Hennelly, S.P. & Sanbonmatsu, K.Y. Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture* **2**, 189-199 (2012).
40. Ardeljan, D. et al. LINE-1 ORF2p expression is nearly imperceptible in human cancers. *Mobile DNA* **11**, 1 (2020).
41. Bowling, E.A. et al. Spliceosome-targeted therapies trigger an antiviral immune response in triple-negative breast cancer. *Cell* **184**, 384-403 (2021).
42. Arrowsmith, C. et al. Altered RNA splicing initiates the viral mimicry response from inverted SINEs following type I PRMT inhibition in Triple-Negative Breast Cancer. *Research Square* **PPR368631** (2021).
43. Seiler, M. et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nature Medicine* **24**, 497-504 (2018).
44. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227-34 (1997).
45. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology* **21**, 1973-1985 (2001).
46. Hall, L.L. et al. Demethylated HSATII DNA and HSATII RNA foci sequester PRC1 and MeCP2 into cancer-specific nuclear bodies. *Cell Reports* **18**, 2943-2956 (2017).

47. Attig, J. et al. Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. *Cell* **174**, 1067-1081 (2018).
48. LaCava, J. RNA Binding proteins as regulators of retrotransposon-Induced exonization. *BioEssays* **42**, 1800263 (2019).
49. Levine, A.J. & Greenbaum, B. The maintenance of epigenetic states by p53: the guardian of the epigenome. *Oncotarget* **3**, 1503-1504 (2012).
50. Badal, B. et al. Transcriptional dissection of melanoma identifies a high-risk subtype underlying TP53 family genes and epigenome deregulation. *JCI Insight* **2**, e92102 (2017).
51. Zhang, S.M. et al. KDM5B promotes immune evasion by recruiting SETDB1 to silence retroelements. *Nature* in press (2021).
52. Vabret, N. et al. Y-RNAs lead an endogenous program of RIG-I agonism mobilized upon RNA virus infection and targeted by HIV. *bioRxiv* **773820** (2019).
53. Nogalski, M.T. et al. A tumor-specific endogenous repetitive element is induced by herpesviruses. *Nature Communications* **10**, 90 (2019).

Data Availability Statement: Original data will be made available upon reasonable request.

Acknowledgements:
This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748 (A.S., S.S., B.G.); NIH grants R01AI081848 (N.V., B.G.), R01CA240924 (A.S., B,G); Fondation de la Recherche Médicale: ANR-Flash Covid, Project SARS-Cov-2immunRNAs (S.C., R.M.); P50 254838-01 (O.A.-W.) and U01CA228963 (A.S., S.S., B.G.); the V Foundation for Cancer Research (A.S.); and the Pershing Square Sohn Prize-Mark Foundation Fellowship (A.S., O.A.-W., N.V., B.G.), and the Edward P. Evans Foundation (O.A-W.). Canadian Institute of Health Research (CIHR), New Investigator salary award (201512MSH360794-228629 to D.D.C.), Canada Research Chair (to D.D.C.), CIHR Foundation Grant (FDN 148430 to D.D.C.), CIHR Project Grant (PJT 165986 to D.D.C.), NSERC (489073 to D.D.C) The authors would like to acknowledge productive conversations with David Ting, Arnold Levine and the Greenbaum laboratories and thank Nicole Rusk for reading and editing the manuscript.

Author Contributions:
Conceptualization: P.S., R.M., S.C., B.G.; Research Plan: P.S., R.M., S.C., B.G; Mathematical Modeling: P.S., R.M., S.C., B.G.; Double-stranded Force Calculation: P.S., R.M., S.C.; Model Implementation: P.S.; Data Analysis: P.S., A.S., S.M., S.S.; Interpretation: P.S., N.V., J.L., O.A.-W., D.D.C., R.M., S.C., B.G.; Writing: P.S., R.M., S.C., B.G.; Reviewing & Editing: P.S., J.L., O.A.-W., D.D.C., R.M., S.C., B.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.G. has received honoraria for speaking engagements from Merck, Bristol Meyers Squibb, and Chugai Pharmaceuticals; has received research funding from Bristol Meyers Squibb and Merck; and has been a compensated consultant for Darwin Health, Merck, PMV Pharma and Rome Therapeutics of which he is a co-founder. A.S. has been a compensated consultant for Rome Therapeutics. D.D.C received research funding from Pfizer and Nektar therapeutics; is a shareholder, co-founder and CSO of Adela (former DNAMx).

Correspondence: psulc@asu.edu, greenbab@mskcc.org

FIGURE LEGENDS

Figure 1|Landscape of forces on CpG dinucleotides in the genome. A, Histogram of changes in the force on CpG motifs across all repetitive elements in the human genome. **B,** Change in the CpG force as a function of the force on the original (consensus) repeat insert over its evolutionary history. Each point represents a family of repetitive elements, along with a linear fit. All repeats whose consensus is above the mean force on CpG dinucleotides (−1.9) have decreased their CpG content. Alu repeats (green) and HSATII (red) are highlighted as exceptions to the general trend. **C,** The mean CpG force of all inserts in a repeat family as a function of the Kimura distance from the consensus sequence for each family.

Figure 2|The evolution of CpG dinucleotides for LINE-1 and HSATII repeats. A, Scatter plot of forces on CpG and UpA dinucleotides for LINE-1 functional (red) and non-functional (blue) elements in the human genome. The white ellipse corresponds to one standard deviation distance (in the principal axes directions along CpG and UpA forces) from the mean for the CpG and UpA forces on FLI and FLnI LINE-1 inserts respectively. **B,** Force on CpG motif for FLnI inserts of LINE-1 and HSAT-II in human genome as a function of average distance from the intact FLI sequences (for LINE-1) or the distance from the consensus sequence (for HSAT-II, marked with a red diamond). The force relaxation evolutionary model fit is shown for both sequence families. **C,** Calculated forces acting on CpG motifs for all inserts from the DFAM database in the human genome. The segments with no inserts present are colored in grey, and the colorbar shows colors assigned to selected force values. The length of the colored segment in the plot is proportional to the length of the sequence of the insert. The mean CpG forces acting on all mRNAs in human genome is −1.1, corresponding to white in color code of the heatmap. **D,** Distribution of genomic regions containing repeats with high CpG force greater than one standard deviation from the mean, illustrating an over-representation of repeats from intronic regions, and a depletion from intergenic regions.

Figure 3| Double-stranded forces in the human genome. A, Histogram of dsRNA force calculated for the following human genome transcripts: mRNA coding sequences (blue), non-coding RNAs (green), inserts (red), consensus sequences of repeats (cyan), and sequences obtained by randomly reshuffling mRNA coding sequences (violet). **B,** Mean of double-stranded force calculated for each family of repeats as a function of the mean Kimura distance of all inserts for a repeat family from the consensus sequence. The red curve corresponds to mean value (and standard deviation from it) for all families binned into the same distance from consensus. **C,** Calculated double stranded forces all inserts from the DFAM database in the human genome. The segments with no inserts present are colored in grey, and the colorbar shows the colors assigned to selected force values. **D,** Complexity of sequences in complementary regions found in the human genome (grey dots) as a function of the segment length. The complementary regions that overlap with known repeat element or ncRNA or mRNA are highlighted as red dots, and the ones where both regions contain insert of a repetitive element from the same family are highlighted in red. The dashed lines correspond to the complexity of a completely random sequence (top line) and trivial region consisting of a single nucleotide (bottom). Complexity of both complementary segments are similar, so we only include the complexity of one of the complementary transcripts.

Figure 4|Classes and genomics origins of repeats with large double-stranded forces. A, The double-stranded force histograms in human genome (sliding window with transcript of length of 3000) and compared to MDA-5 binding RNA transcripts. **B,** Distribution of classes of repeats from the peak of large double-stranded forces in **(B)** illustrating an over-representation of repeats emanating from SINE elements. **C,** Distribution of genomic regions containing repeats with high double-stranded forces, illustrating an over-representation of repeats from intronic regions. **D,** Volcano plot of SINE element expression of elements with double stranded force greater than 0.5 in SF3B inhibitor (H3B-

8800) versus control (DMSO) treated SF3B1 K700E mutant K562 cell lines. **E,** Distribution of genomic regions containing expressed repeats with high double-stranded forces and fold-change of greater than 0.5 in H3B-8800 versus DMSO treated SF3B1 K700E mutant K562 cell line. Double-stranded SINE elements from intronic regions are over-represented.

EXTENDED DATA FIGURE LEGENDS

Extended Data Figure 1|Dinucleotide distribution in human genome. Counts of dinucleotides across the human genome (HG38 genome assembly).

Extended Data Figure 2| A, The mean of maximum lengths in a secondary structure in a single-stranded RNA sequence (green line) and the mean maximum length of complementary segments (blue line), along with respective fits of Eq. 12 from Methods. **B,** Double-stranded force on repeat family Tigger4a. The force relaxation evolutionary model fit shows the relaxation of the inserts compared to the relaxation simulated by neutral Wright-Fisher model.

Extended Data Figure 3| The genomic distribution of regions which under-represent dsRNA sequences in the human genome. Only exons were shown to significantly under-represent dsRNA formation.

Extended Data Figure 4| A, Volcano plot of SINE element expression between DMSO treated SF3B1 K700E mutant and wild-type in K562 cell lines. **B,** Volcano plot of SINE element expression between H3B-8800 and DMSO treated SF3B1 wild-type K562 cell lines. **C,** Volcano plot of SINE element expression between H3B-8800 treated SF3B1 mutant & wild-type in K562 cell lines. **D,** Volcano plot of SINE element expression between H3B-8800 and DMSO treated SF3B1 K700E in Nalm6 cell lines. **E,** Volcano plot of SINE element expression between E7107 and DMSO treated SF3B1-K700 in Nalm6 cell lines.

Extended Data Figure 5| A, The distribution of lengths of bidirectional transcripts identified in TCGA. **B,** Scatter plot of maximum length of complementary segments in a single transcript vs the number of occurrences of such transcript in respective TCGA datasets. We only considered transcripts that come from genome regions of length 3000 that have double-stranded RNA force larger than 0.5. Only transcripts with one or more occurrences are shown.

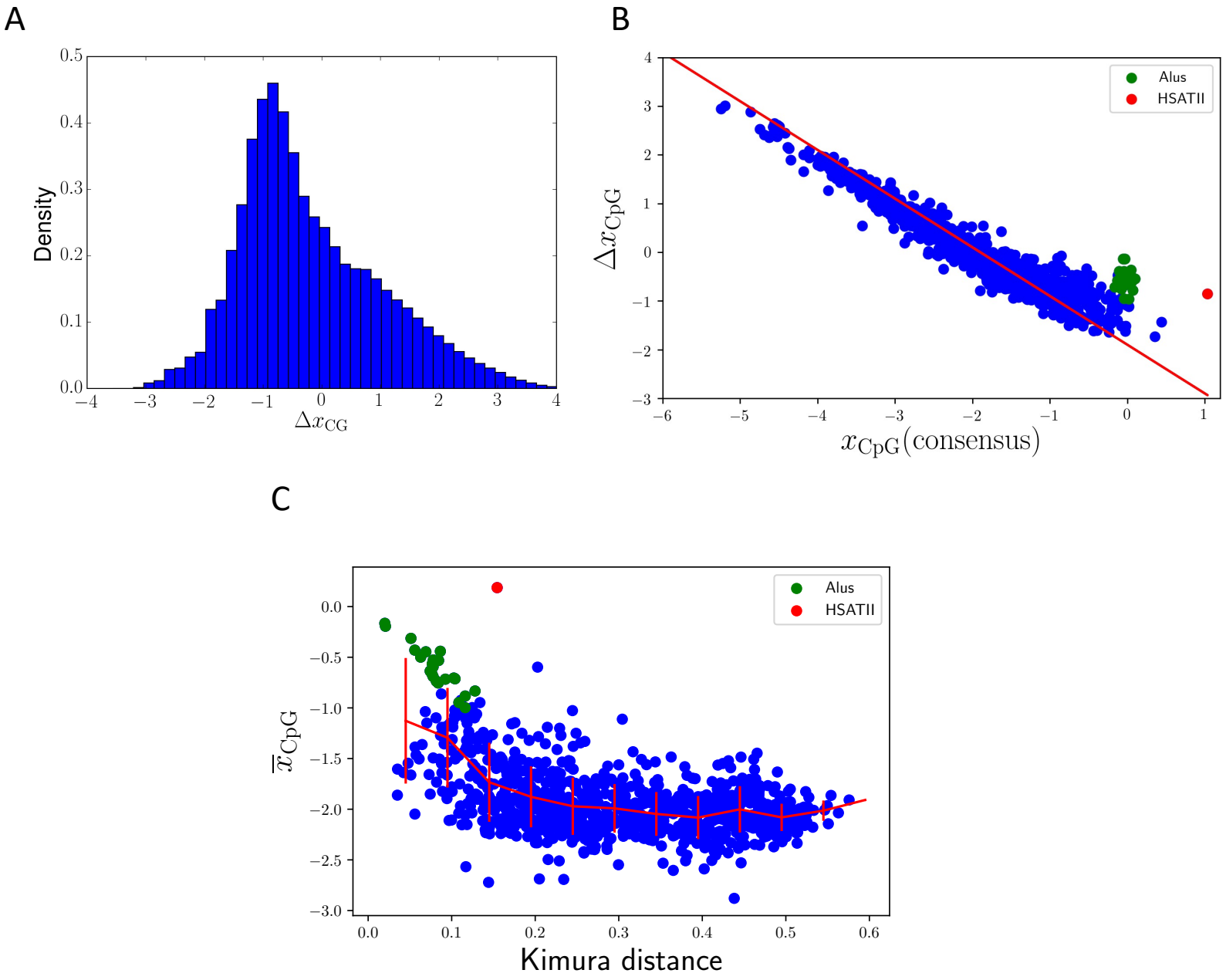


Figure 1|Landscape of forces on CpG dinucleotides in the genome. A, Histogram of changes in the force on CpG motifs across all repetitive elements in the human genome. **B,** Change in the CpG force as a function of the force on the original (consensus) repeat insert over its evolutionary history. Each point represents a family of repetitive elements, along with a linear fit. All repeats whose consensus is above the mean force on CpG dinucleotides (−1.9) have decreased their CpG content. Alu repeats (green) and HSATII (red) are highlighted as exceptions to the general trend. **C,** The mean CpG force of all inserts in a repeat family as a function of the Kimura distance from the consensus sequence for each family.

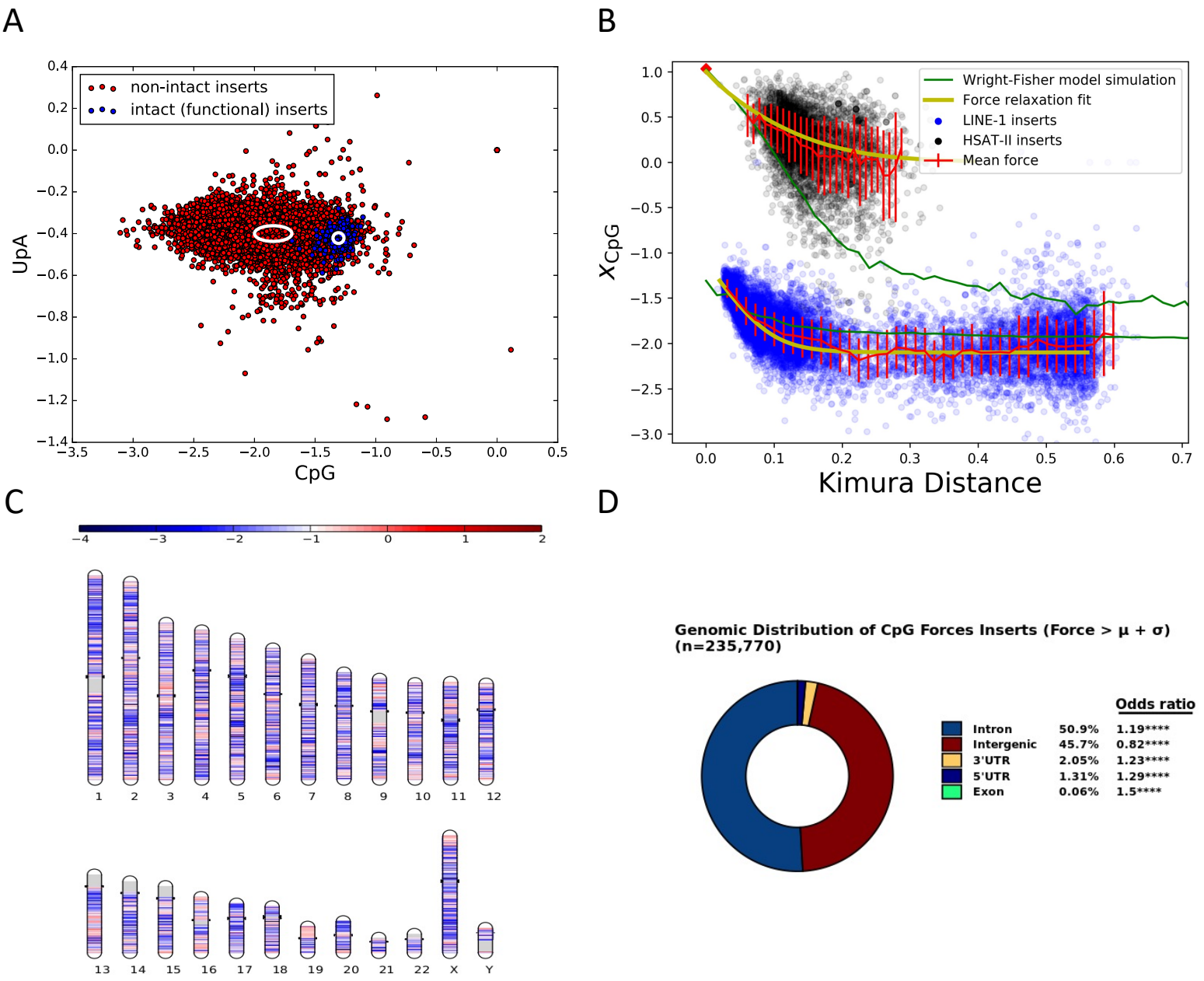


Figure 2|The evolution of CpG dinucleotides for LINE-1 and HSATII repeats. A, Scatter plot of forces on CpG and UpA dinucleotides for LINE-1 functional (red) and non-functional (blue) elements in the human genome. The white ellipse corresponds to one standard deviation distance (in the principal axes directions along CpG and UpA forces) from the mean for the CpG and UpA forces on FLI and FLnI LINE-1 inserts respectively. **B,** Force on CpG motif for FLnI inserts of LINE-1 and HSAT-II in human genome as a function of average distance from the intact FLI sequences (for LINE-1) or the distance from the consensus sequence (for HSAT-II, marked with a red diamond). The force relaxation evolutionary model fit is shown for both sequence families. **C,** Calculated forces acting on CpG motifs for all inserts from the DFAM database in the human genome. The segments with no inserts present are colored in grey, and the colorbar shows colors assigned to selected force values. The length of the colored segment in the plot is proportional to the length of the sequence of the insert. The mean CpG forces acting on all mRNAs in human genome is −1.1, corresponding to white in color code of the heatmap. **D,** Distribution of genomic regions containing repeats with high CpG force greater than one standard deviation from the mean, illustrating an over-representation of repeats from intronic regions, and a depletion from intergenic regions.

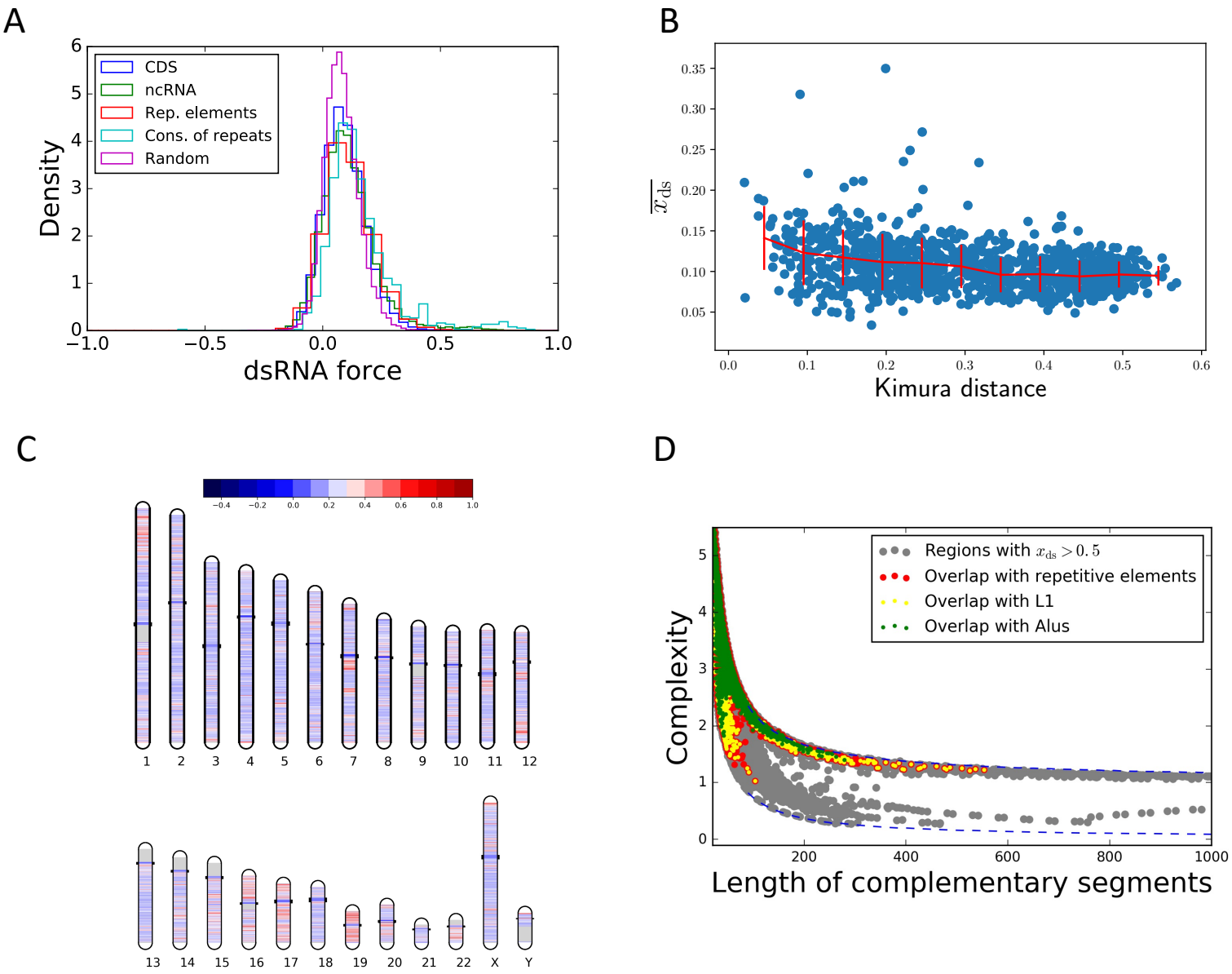


Figure 3| Double-stranded forces in the human genome. A, Histogram of dsRNA force calculated for the following human genome transcripts: mRNA coding sequences (blue), non-coding RNAs (green), inserts (red), consensus sequences of repeats (cyan), and sequences obtained by randomly reshuffling mRNA coding sequences (violet). **B,** Mean of double-stranded force calculated for each family of repeats as a function of the mean Kimura distance of all inserts for a repeat family from the consensus sequence. The red curve corresponds to mean value (and standard deviation from it) for all families binned into the same distance from consensus. **C,** Calculated double stranded forces all inserts from the DFAM database in the human genome. The segments with no inserts present are colored in grey, and the colorbar shows the colors assigned to selected force values. **D,** Complexity of sequences in complementary regions found in the human genome (grey dots) as a function of the segment length. The complementary regions that overlap with known repeat element or ncRNA or mRNA are highlighted as red dots, and the ones where both regions contain insert of a repetitive element from the same family are highlighted in red. The dashed lines correspond to the complexity of a completely random sequence (top line) and trivial region consisting of a single nucleotide (bottom). Complexity of both complementary segments are similar, so we only include the complexity of one of the complementary transcripts.

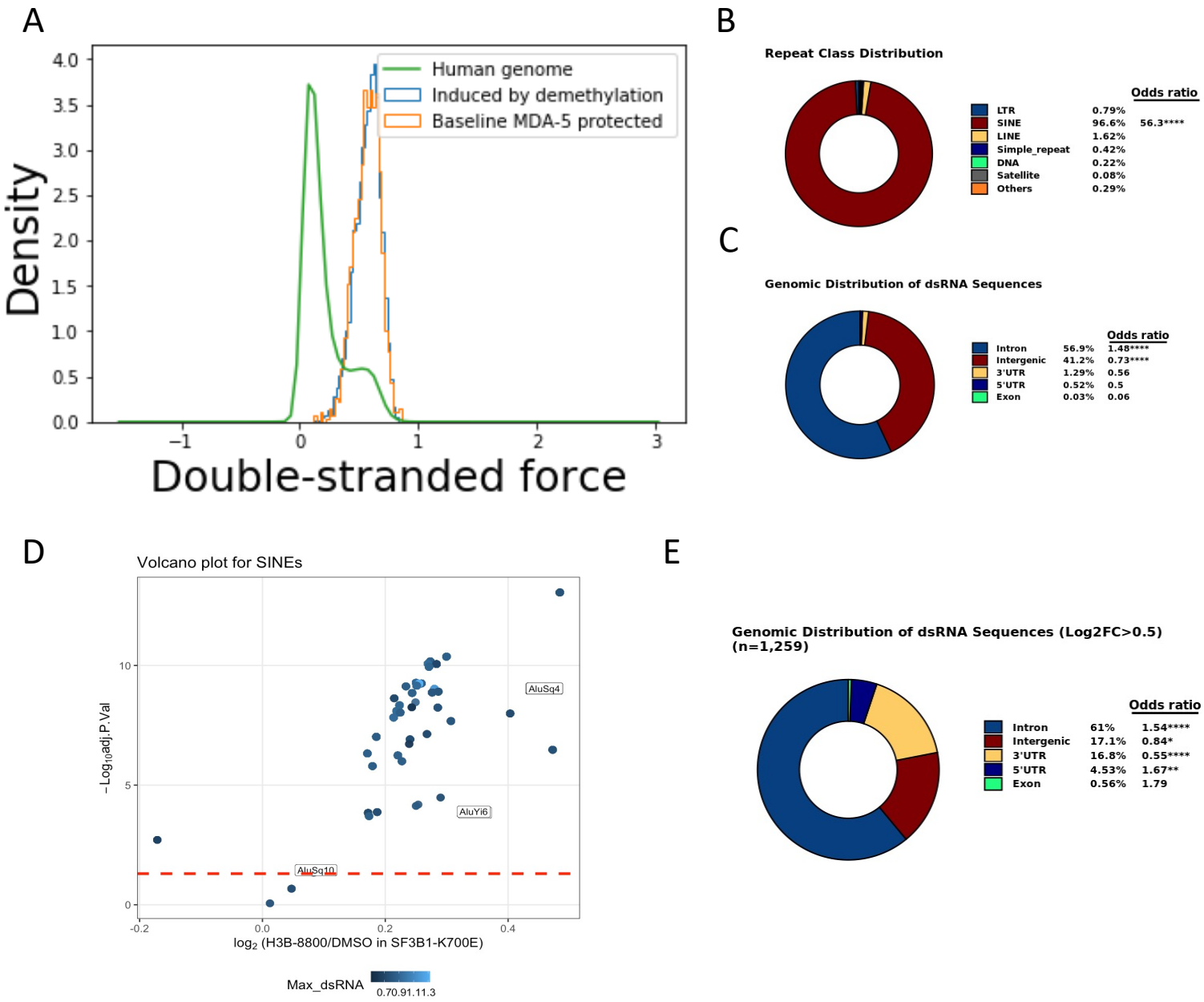
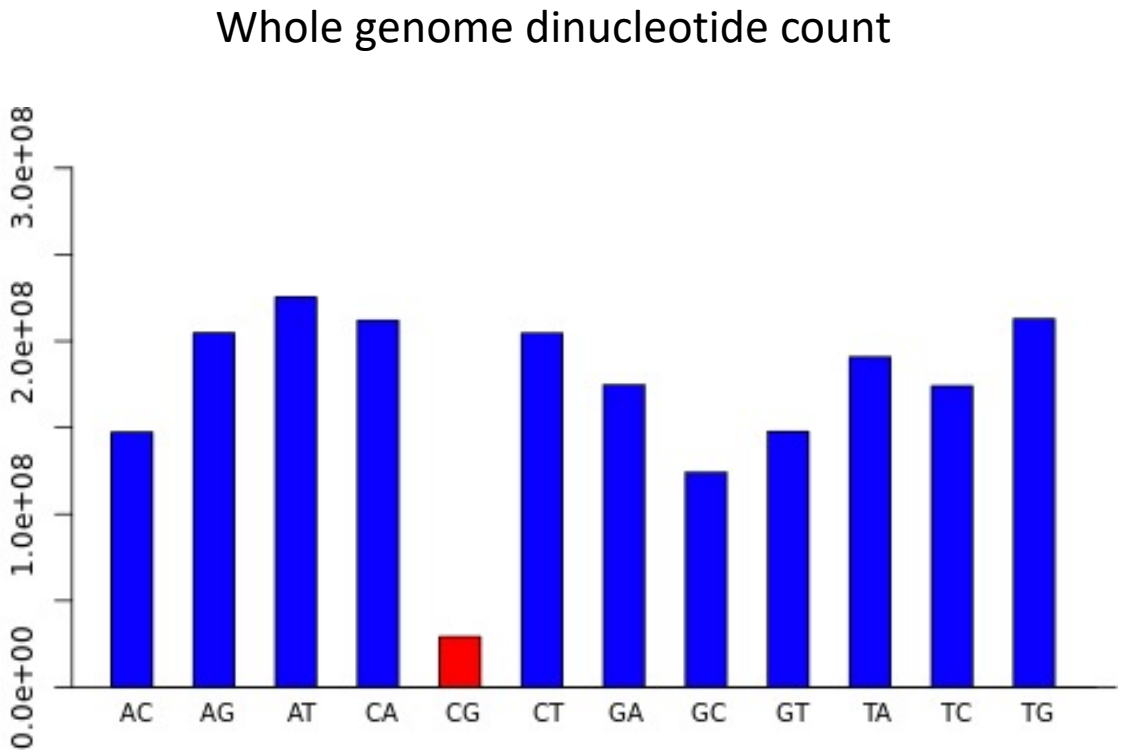
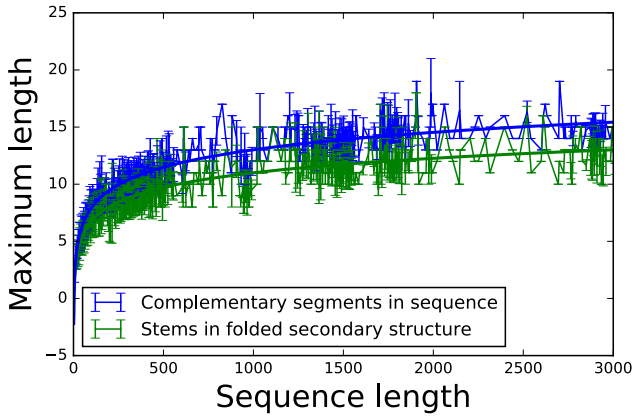


Figure 4|Classes and genomics origins of repeats with large double-stranded forces. A, The double-stranded force histograms in human genome (sliding window with transcript of length of 3000) and compared to MDA-5 binding RNA transcripts. **B,** Distribution of classes of repeats from the peak of large double-stranded forces in **(B)** illustrating an over-representation of repeats emanating from SINE elements. **C,** Distribution of genomic regions containing repeats with high double-stranded forces, illustrating an over-representation of repeats from intronic regions. **D,** Volcano plot of SINE element expression of elements with double stranded force greater than 0.5 in H3B-8800 versus DMSO treated SF3B1-K700 mutant K562 cell lines. **E,** Distribution of genomic regions containing expressed repeats with high double-stranded forces and fold-change of greater than 0.5 in H3B-8800 versus DMSO treated SF3B1-K700 mutant K562 cell line. Double-stranded SINE elements from intronic regions are over-represented.

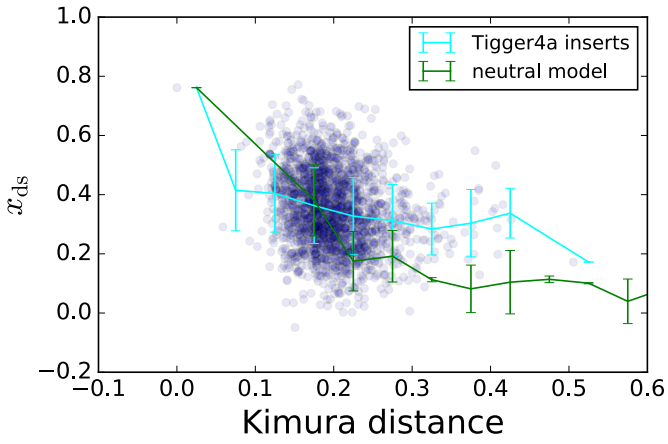


Extended Data Figure 1|Dinucleotide distribution in human genome. Counts of dinucleotides across the human genome (HG38 genome assembly).

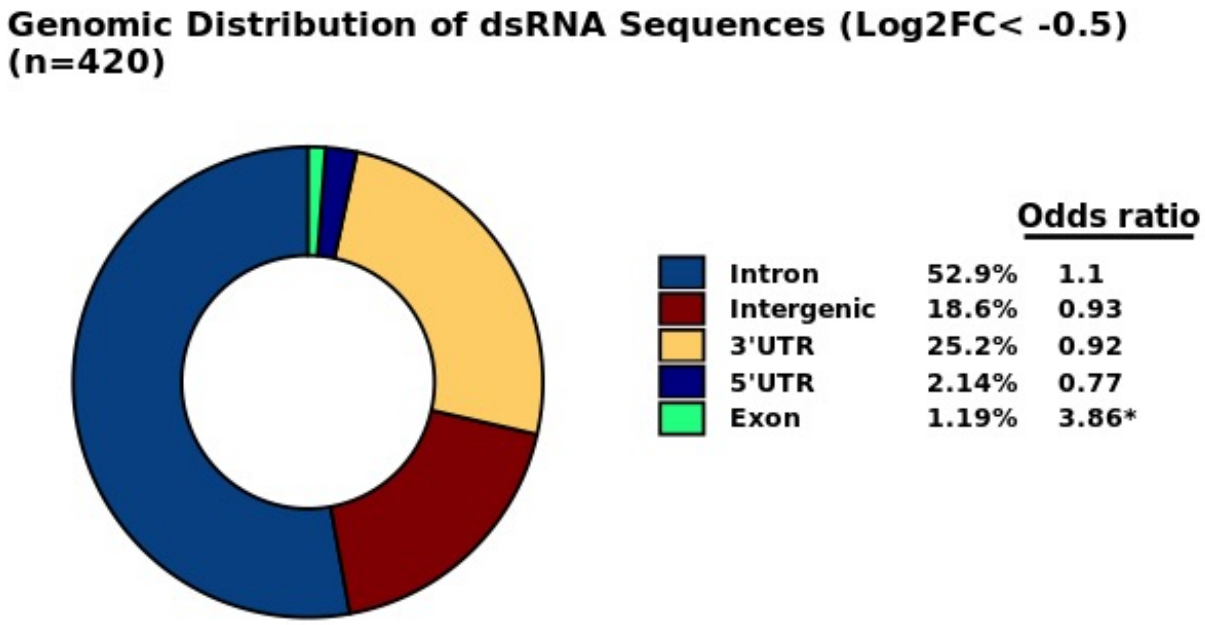
A



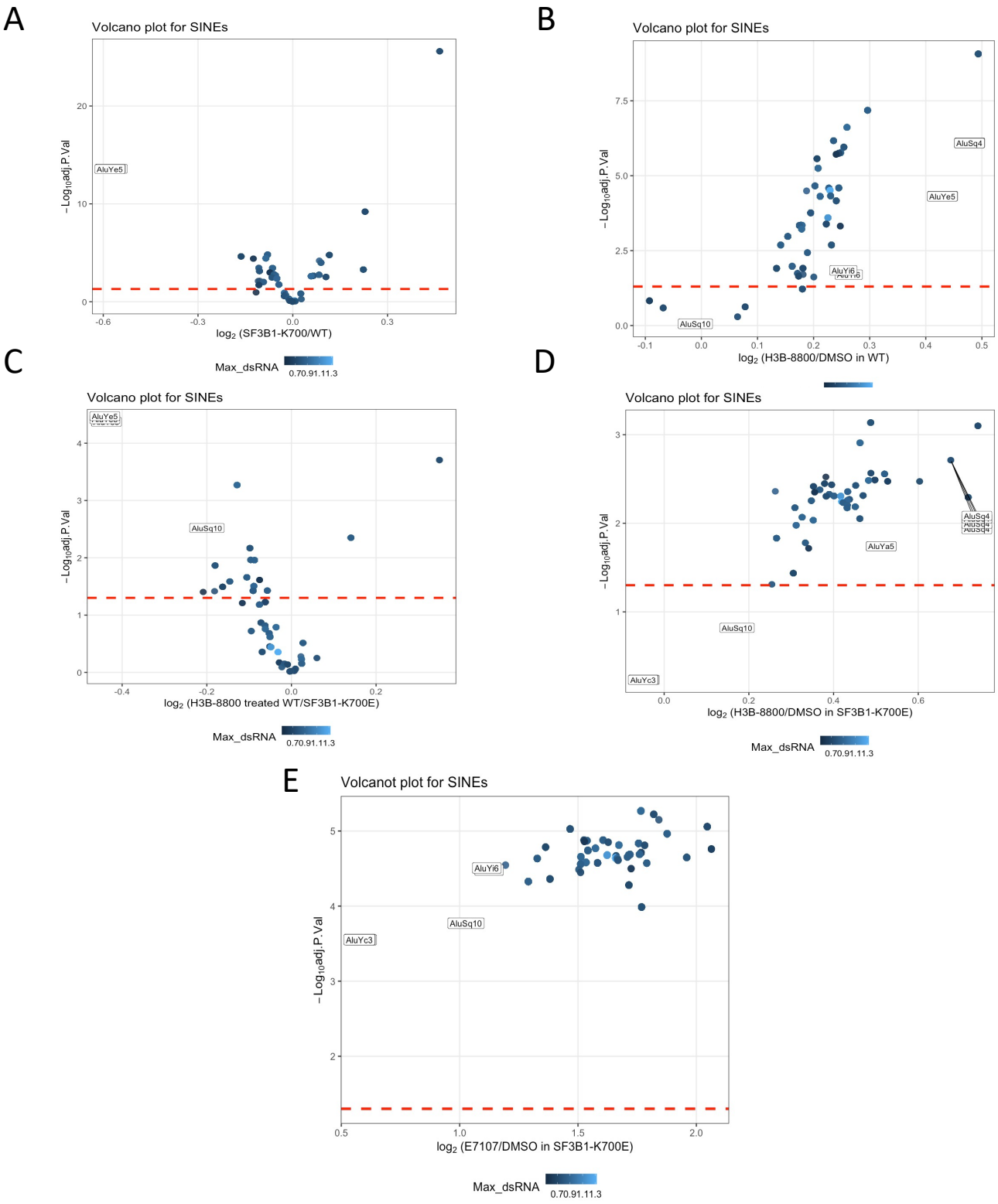
B



Extended Data Figure 2| A, The mean of maximum lengths in a secondary structure in a single-stranded RNA sequence (green line) and the mean maximum length of complementary segments (blue line), along with respective fits of Eq. 12 from Methods. **B,** Double-stranded force on repeat family Tigger4a. The force relaxation evolutionary model fit shows the relaxation of the inserts compared to the relaxation simulated by neutral Wright-Fisher model.

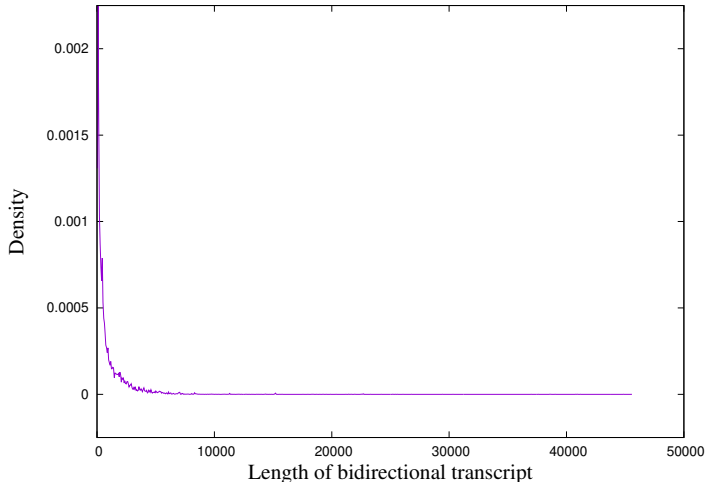


Extended Data Figure 3| The genomic distribution of regions which under-represent dsRNA sequences in the human genome. Only exons were shown to significantly under-represent dsRNA formation.

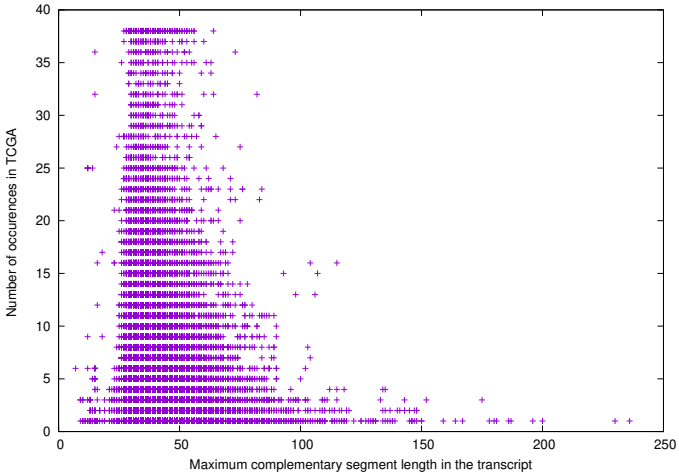


Extended Data Figure 4| A, Volcano plot of SINE element expression between DMSO treated SF3B1-K700 mutant and wild-type in K562 cell lines. **B,** Volcano plot of SINE element expression between H3B-8800 and DMSO treated wild-type K562 cell lines. **C,** Volcano plot of SINE element expression between H3B-8800 treated SF3B1 mutant & wild-type in K562 cell lines. **D,** Volcano plot of SINE element expression between H3B-8800 and DMSO treated SF3B1-K700 in Nalm6 cell lines. **E,** Volcano plot of SINE element expression between E7107 and DMSO treated SF3B1-K700 in Nalm6 cell lines.

A



B



Extended Data Figure 5| A, The distribution of lengths of bidirectional transcripts identified in TCGA. **B,** Scatter plot of maximum length of complementary segments in a single transcript vs the number of occurrences of such transcript in respective TCGA datasets. We only considered transcripts that come from genome regions of length 3000 that have double-stranded RNA force larger than 0.5. Only transcripts with one or more occurrences are shown.

Methods

Quantification of sequence constraints

We quantify the sequence constraints as a parameter (force) x_s acting on a particular sequence motif m, which can be for instance CpG in a DNA or RNA sequence, but it could also correspond to longer motifs (such as codons for example). Using the maximum entropy principle, the probability of occurrence of sequence σ with N_σ^m motifs is

$$P\left(\sigma|x_s\right) = \frac{1}{Z(x_s)} \prod_{i=1}^{L} f\left(c_i^\sigma\right) \exp\left(x_s N_\sigma^m\right) \tag{1}$$

where and $f\left(c_i^\sigma\right)$ is the probability of finding nucleotide c on i-th position in the sequence. In the case of sequences considered here, we use frequency of a given nucleotide (A, C, G, U) (i.e. the number of occurrences divided by sequence length) for a given sequence σ to obtain the estimates of f. $Z(x_s)$ is the normalization factor

$$Z(x_s) = \sum_{\tau} \prod_{i=1}^{L} f\left(c_i^\sigma\right) \exp\left(x_s N_\tau^m\right) \tag{2}$$

where the sum is carried out over all possible sequences τ of the same length as sequence σ.

For instance, in the case of the respective sequences of LINE-1 and HSATII elements considered in this work, the respective frequencies f correspond to the number of occurrence of a given nucleotide (A,C,G or U) in the sequence divided by the sequence length. Force x_s associated with a given motif m (such as CpG) is then obtained so that the number of occurrences of the motif, N_m, satisfies

$$N_\sigma^m = \sum_{\tau} P\left(\sigma|x_s\right) N_\tau^m = \frac{\partial \log Z(x_s)}{\partial x_s}. \tag{3}$$

A Newton-method based algorithm to efficiently ($\sim O(L^2)$) calculate x_s such that (3) is satisfied was derived in [1]. Note the positive value of force x_s signifies that associated motif is enriched compared to what would be expected from a random sequence with given frequencies of nucleotides, and *vice versa* for a negative value. In this work, we use x_{CpG} to denote the value of force x_s for CpG dinucleotides.

Evolutionary dynamics of a sequence motif with force relaxation formalism

It is possible to harness the formalism developed in (1) - (3) to study the evolutionary dynamics of number of motifs N^m, as it approaches the steady state (equilibrium) value N_{avg}^m. We assume that the sequence is evolving under action of two competing effects, a selective force x_{eq}^m and an 'entropic force' x_s.

$$x_s(N^m) = \frac{\partial S}{\partial N^m} = \frac{\partial S(x_s)}{\partial x_s} \frac{\partial x_s}{\partial N^m} \tag{4}$$

where S is the entropy (using P from (1)):

$$S(x_s) = -\sum_{\tau} P\left(\tau|x_s\right) \log P\left(\tau|x_s\right). \tag{5}$$

The value of x_s that maximizes entropy S corresponds to the most likely sequence and corresponding number of motifs N^m that one would obtain given the respective frequencies f for a

sequence of length L. However, there is also a selection pressure acting on a motif m, specified by a given value x_{eq}^m: for a positive value the motif is under a positive selection, and for a negative value the motif is under negative selection.

The sequence undergoes mutations, which cause changes in the number of motifs (and hence associated value of x_s). To model the evolutionary dynamics of the sequences, we assume the number of motifs (N^m) evolves according to the relaxation dynamics given by

$$\tau \frac{dN^m(t)}{dt} = -x_s\left(N^m(t)\right) + x_{\mathsf{eq}}^m, \tag{6}$$

where τ sets the timescale. The number of motifs reaches its stationary (equilibrium) value when $x_s = x_{\mathsf{eq}}^m$. It is convenient to express (6) as

$$\tau \frac{\mathsf{d}x_s}{\mathsf{d}t} = -\left(-x_s(t) + x_{\mathsf{eq}}^m\right) \mathsf{var}\left(x_s|N^m\right), \tag{7}$$

where $\mathsf{var}\left(x_s|N^m\right)$ is the variance of x_s for a given N^m.

If we can express $\mathsf{var}\left(x_s|N^m\right)$ as a function of x_s, it is possible to obtain a solution of (7) 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 to $x_s(t)$ (and hence the number of the corresponding sequence motif m) to reach its equilibrium value. For the case of HSATII and LINE-1 we fit the $\mathsf{var}\left(x_s|N^m\right)$ as a quadratic function of x_s.

Quantification of double-stranded RNA content

Following-up on the quantification of dinucleotide content, we develop an analogous framework for quantification of the length of duplex strands. We assume we are given a sequence s of length L. We define frequency in the sequence $f(c)$ for each nucleotide type c (A,C,G or U) in the sequence. If we divide sequence into N segments of maximum length K ($N = L/K$), then the probability that two given of length K segments are fully complementary (i.e. that they can form K base pairs long duplex region) is

$$p_1 = \left(e^{x_{\mathsf{ds}}}\alpha\right)^K, \tag{8}$$

where α is the probability that randomly chosen pair of two nucleotides can form a base pair

$$\alpha = \sum_{\langle cc^* \rangle} f\left(c\right) f\left(c^*\right), \tag{9}$$

where the sum is over all permutations complementary of Watson-Crick or wobble base pairs (A-U, C-G, G-U) with their respective frequencies $f(c)$ ($\alpha = 0.375$ for uniformly distributed nucleotides). The parameter x_{ds} is analogous to the dinucleotide force in Eq. (2), and corresponds to bias that increases (for positive x_{ds}) or decreases ($x_{\mathsf{ds}} < 0$) the typical length of double-stranded region in the sequence with respect to a randomly drawn sequence from nucleotide distribution $f(c)$ (in which case $x_{\mathsf{ds}} = 0$). The probability of having at least one pair of fully complementary segments of length K is

$$p_{\mathsf{ds}}(K) = 1 - \left(1 - \left(e^{x_{\mathsf{ds}}}\alpha\right)^K\right)^{\frac{N(N-1)}{2}}. \tag{10}$$

We are interested in the typical length of the longest segment that is complementary in the ensemble of sequences of length L with given distribution of nucleotides $f(c)$. We look for K such

that $p_{\mathsf{ds}} \approx 1/2$. Assuming that $K \ll L$ and $\alpha^K \ll 1$, we obtain from Eq. (10)

$$K \approx \frac{\log L}{\log \frac{1}{\sqrt{e^{x_{\mathsf{ds}}}\alpha}}} \tag{11}$$

for larger $L \gg K$. Hence we get the maximum length of the longest complementary segments of the ensemble as a function of L, α and x_{ds} as

$$\lambda_{\mathsf{max}} \approx \frac{\log L}{\log \frac{1}{\sqrt{e^{x_{\mathsf{ds}}}\alpha}}} + c_0. \tag{12}$$

We fit the Eq. 12 to a set of mean maximum length of segments of randomly generated RNA sequences of lengths ranging up to 3000 bases (Extended Data Fig. 2A) and obtain $c_0 = -2.2$ and $x_{\mathsf{ds}} = 0.06$. We also fit Eq. 12 to the mean of maximum duplex lengths in a secondary structure (as obtained from folding the sequences by ViennaRNA tool [2]) of set of randomly generated sequences, we obtain $c_0 = -1.7$ and $x_{\mathsf{ds}} = -0.11$. We note that the value of x_{ds} is slightly smaller for the longest double-stranded segment in the folded sequence, because the longest complementary segments will not always form a duplex segment (e.g. due to entropic cost of bringing the two segments together). The longest segments in folded sequences are therefore on average slightly shorter than the lengths of maximum complementary segments.

Therefore, for a sequence of length L with frequencies of bases given by $f(c)$ and with maximum length of complementary segments λ_{max}, we obtain x_{ds} from Eq. (12), thus obtaining a single metric to compare distribution of double-stranded segments across various RNA sequence ensembles and families.

Wright-Fisher model of population genetics for the evolution of sequence motifs

In addition to the force relaxation model introduced above, we present here a different approach to study the evolution of nucleotide sequence motifs based on a Wright-Fisher (WF) population genetics model, which assumes haploid reproduction of sequences. The probability distribution of all sequences evolves in time according to

$$\frac{\partial p_\sigma(t)}{\partial t} = \left(s_\sigma - \langle s \rangle^t\right) p_\sigma(t) + \sum_\gamma \left(p_\gamma(t) T_{\gamma\to\sigma} - p_\sigma(t) T_{\sigma\to\gamma}\right), \tag{13}$$

where $p_\sigma(t)$ is the probability of sequence σ at time t, s_σ is a selection coefficient that depends on the number of motifs N^m, with $\langle s \rangle^t$ as its average value over the probability distribution of all sequences at time t.

We implement the WF model numerically, and evolve a set of sequences according to a neutral mutation model without a selection term to provide a null model of neutral sequence evolution. We evolve a population of sequences (which either start all equal to the same one or from a distribution). For each simulation step, we pick a random base for each sequence in the ensemble, and mutate it to 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 well as on the context (identity of the bases in the neighborhood), as transversion (purine mutating to pyrimidine or vice versa) and transition (purine mutating to purine or pyrimidine mutating to pyrimidine) substitutions in sequences can have different likelihood [3].

μ_{TiCpG}:μ_{TvCpG}:μ_{Ti}:μ_{Tv}	$x^{\text{eq}}_{\text{CpG}}$
40:10:4:1	-2.2
40:4:4:1	-2.0
40:1:4:1	-1.7
4:4:4:1	-0.8
20:4:4:1	-1.5
27:2:4:1	-1.7

Table 1: The ratios of dinucleotide mutation rates (transition and transversion with and outside of CpG context) and a corresponding value of equilibrium force on the CpG dinucleotide

Additionally, in vertebrates and plants, mutations in CpG context are known to be more common due to CpG hypermutability [4]. Hence, for the mutation rates in the WF 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 the following ratios introduced in Ref. [3] and which are listed in Table 1. For each ratio, we constructed the stochastic matrix, with entries corresponding to probabilities of mutating from one dinucleotide into another dinucleotide. We calculated the stationary dinucleotide distributon from the stationary vector of this matrix, from which one can calculate the corresponding CpG force x_{eq} using the fact that the force is approximately equal to the logarithm of relative frequency of the dinucleotide motif $x \approx \log(f(\text{CpG})/f(\text{C})f(\text{G}))$ [5]. 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 Wright-Fisher model evolution comparison of LINE-1 and HSATII inserts in Figure 2 in the main text, we used the 40:10:4:1 ratio as it was closer to the saturated value of x_{CpG} of the LINE-1 elements.

Sequence ensembles

The LINE-1 sequences were obtained from L1Base2 database [6]. 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). 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 [7]. We then calculate the average distance for each of the non-intact sequences from the intact-sequences, and furthermore calculate 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 DFAM database [8] (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 insert 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 [7] between the consensus and the inserts.

We note that we use the Kimura distance [9] 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 other tentative transcripts that can be possible source of long (longer than 100 bp) duplex formation. For each window of length 3000 bases (taken in the positive sense of the read), we calculate the double-stranded force by expressing x_{ds} from (12), and using the respective nucleotide frequency to obtain α from Eq. (9). We find the maximum length of complementary segments L in the given sequence of length N by filling an $N \times N$ base compatibility matrix with 1 if two nucleotides can form a base pair (either by Watson-Crick or wobble base pairing). We then find the longest antidiagonal stretch of 1s in the matrix, which corresponds to the maximum length L of complementary segments.

Sequence complexity quantification

We use an approximation of Kolmogorov complexity [10] to quantify how "non-trivial" are the complementary segments. Adopting the approach from Ref. [11], we use the size (in bytes) of the sequence compressed with gzip software as a proxy of the Kolmogorov complexity. Simple sequences, *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).

Transcriptome Analysis

Analysis of repeats from splice inhibitors

Raw RNAseq data (GSE95011) associated with the Seiler, et al., 2018 study [12] were downloaded from NCBI. Briefly, reads were trimmed and quality checked using first and then mapped to the human genome (hg38) and repetitive elements from RepBase. 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 20$^{\text{th}}$ percentile or shorter than 40 bases were discarded. Only paired reads that passed the filtering step were retained. Quality filtered reads were then mapped using STAR aligner and assigned to genes (Gencode annotation) and repeat elements (RepeatMasker annotation) using function of package using the external Ensembl annotation. To check the expression difference for a given repeat in a locus-specific manner, we modified the RepeatMasker reference file and counted the reads that mapped to repeats at different locus separately.

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 as implemented in the function in [13]. These low-count values (CPM < 2) were likely due to sequencing errors and were removed before calculating the size factor for each sample. Then, filtered CPM was log2 transformed and used in heat-map visualization and downstream statistical analysis. On the heatmap, genes (rows) were scaled by z-score scaling. Heat maps were generated by the R statistical

programming package. Differential expression analysis was performed using [14] between splicing modulator H3B-8800 treated VS DMSO treated SF3B1-K700 mutated cell line k562 for a given locus. The adjusted p-value was calculated using the Benjamini & Hochberg correction [15].

Estimate of genome regions with high double stranded force

To estimate the dsRNA force for a given repeat loci, we intersect each repeat loci with the calculated 3kb genomic windows that have high dsRNA forces (> 0.5). Then the Start and End coordinates of the corresponding dsRNA sequence pairs, which overlap with the repeat loci that match the criteria: $|\text{log2FC}(\text{treated}/\text{untreated})| > 0.5$ and FDR < 0.05, were used to annotate different genomic features. We counted the genomic features of the dsRNA 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. Donut plots for the genomic feature proportions of the dsRNA sequences that overlap with upregulated and downregulated repeats were plotted using the R package script.

Whole transcriptome analysis

We analyzed RNA-Seq data for the 38 TCGA patients for whom Total RNA-Seq data exist, as defined in Ref. 34. These patients have one FFPE and one fresh frozen sample each. Unlike the majority of ("canonical") samples in TCGA, these samples were sequenced using stranded protocol. These data are not a part of the "harmonized" samples set, and they are available the from "legacy" section of TCGA archive. Reads were mapped to HG38 genome with Gencode annotation using STAR aligner. Transcripts were assembled taking strandedness of the protocol into account using the stringtie program with the default settings. Reference annotation (Gencode) was used as a guide for assembly. Overlapping transcripts from both strands were identified and the length of the overlapping complementary sequences was computed. We computed intersection of all assembled transcripts with regions having high dsRNA force (shown in Extended Data Figure 5B). We required that the transcript covers at least the region from startA to endB in the regions with high dsRNA force (supp. table 3). For each region with high dsRNA force we computed the number of times it is seen in the analyzed TCGA samples. If the same region overlapped assembled transcripts in two (FFPE and fresh frozen) samples from the same patient, this was counted as one occurrence.

References for Methods Section

[1] Greenbaum, B. D., Cocco, S., Levine, A. J. & Monasson, R. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proceedings of the National Academy of Sciences* **111**, 5054–5059 (2014).

[2] Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for molecular biology* **6**, 26 (2011).

[3] Suzuki, Y., Gojobori, T. & Kumar, S. Methods for incorporating the hypermutability of cpg dinucleotides in detecting natural selection operating at the amino acid sequence level. *Molecular biology and evolution* **26**, 2275–2284 (2009).

[4] Subramanian, S. & Kumar, S. Higher intensity of purifying selection on> 90% of the human genes revealed by the intrinsic replacement mutation rates. *Molecular biology and evolution* **23**, 2283–2287 (2006).

[5] Di Gioacchino, A. *et al.* The heterogeneous landscape and early evolution of pathogen-associated cpg dinucleotides in sars-cov-2. *Molecular Biology and Evolution* **38**, 2428–2445 (2021).

[6] Penzkofer, T. *et al.* L1base 2: more retrotransposition-active line-1s, more mammalian genomes. *Nucleic Acids Research* **45**, D68 (2017).

[7] Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge university press, 1998).

[8] Hubley, R. *et al.* The dfam database of repetitive dna families. *Nucleic acids research* **44**, D81–D89 (2016).

[9] Kimura, M. & Weiss, G. H. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561 (1964).

[10] Li, M., Vitányi, P. *et al. An introduction to Kolmogorov complexity and its applications*, vol. 3 (Springer, 2008).

[11] Dingle, K., Camargo, C. Q. & Louis, A. A. Input–output maps are strongly biased towards simple outputs. *Nature communications* **9**, 1–7 (2018).

[12] Seiler, M. *et al.* Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell reports* **23**, 282–296 (2018).

[13] Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology* **11**, 1–9 (2010).

[14] Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**, e47–e47 (2015).

[15] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

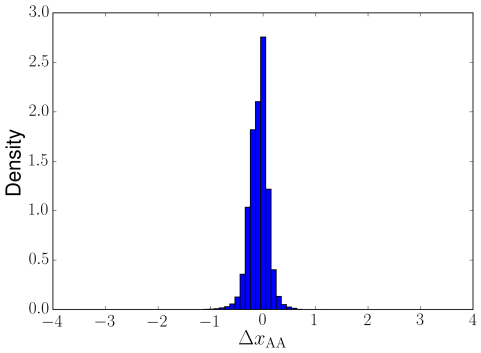
Supplementary Material

Supplementary Tables

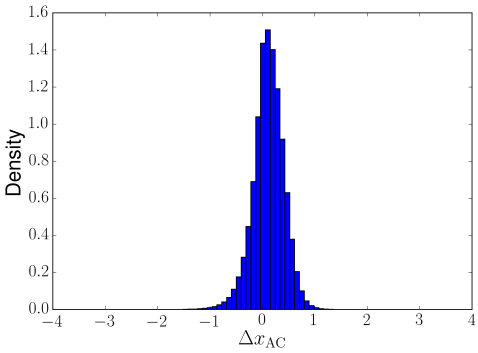
Supplementary Tables are available separately in Microsoft Excel format (xlsx). The description of the respective tables is provided below:

- Supplementary Table 1 contains the dinucleotide force calculated for CpG dinucleotides, x_{CpG}, for each repeat family, as annotated in the DFAM database. For each family, it includes x_{CpG} of the consensus sequence and the mean $\overline{x_{\mathsf{CpG}}}$ calculated for the inserts.
- Supplementary Table 2 contains the double-stranded force x_{ds} calculated for each repeat family. The table includes the x_{ds} of the consensus sequence and mean double-stranded force $\overline{x_{\mathsf{ds}}}$ of the insert sequences.
- Supplementary Table 3 contains a list of all segments identified in 3000 bases long windows in human genome that had associated double-stranded force x_{ds} larger than 0.5.
- Supplementary Table 4 contains all identified repeat transcripts that have been found in The Cancer Genome Atlas (TCGA) total RNA-Seq datasets to be transcribed in both positive and negative sense.

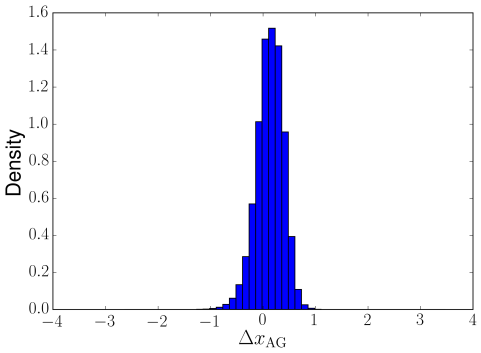
Supplementary Figures



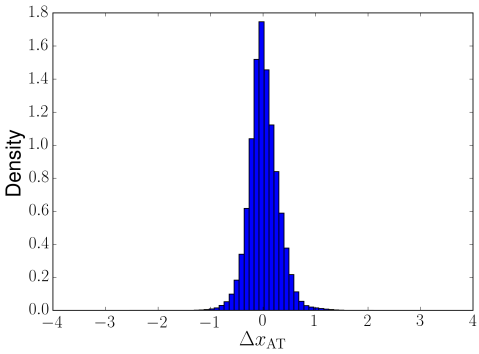
AA



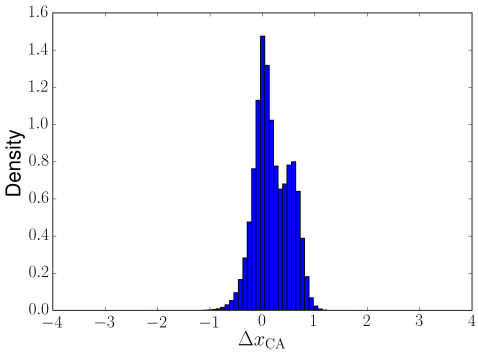
AC



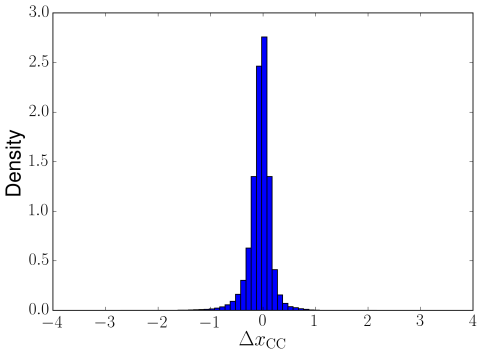
AG



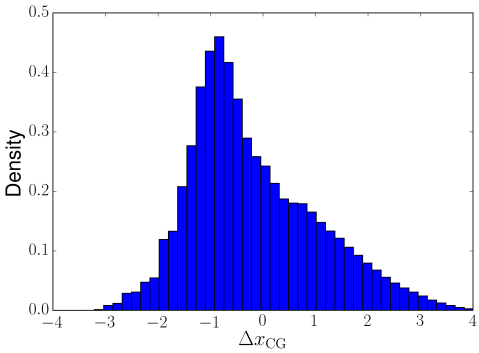
AT



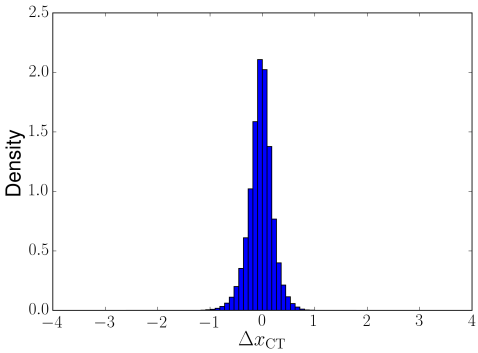
CA



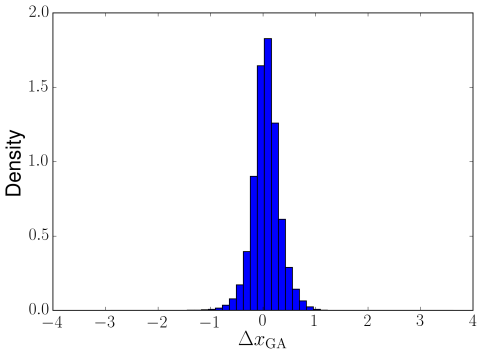
CC



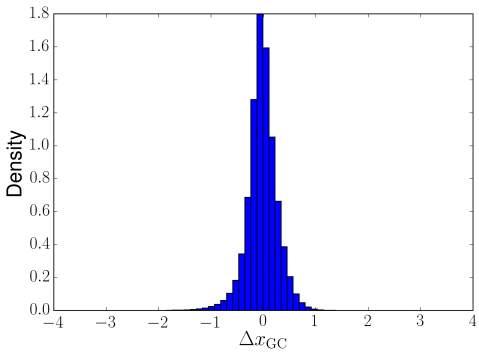
CG



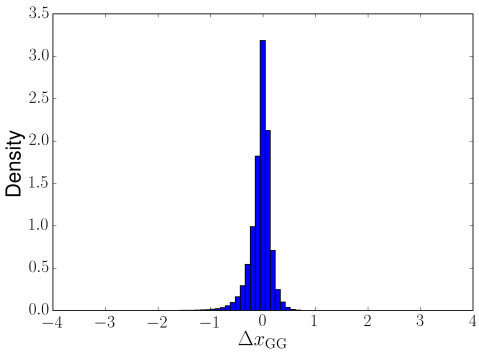
CT



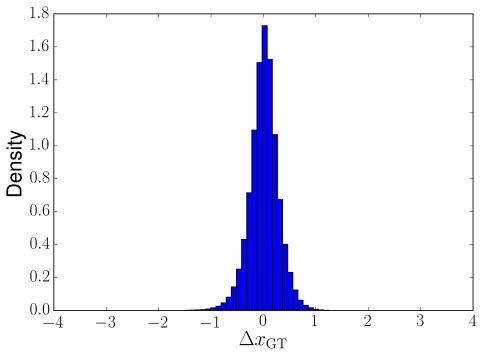
GA



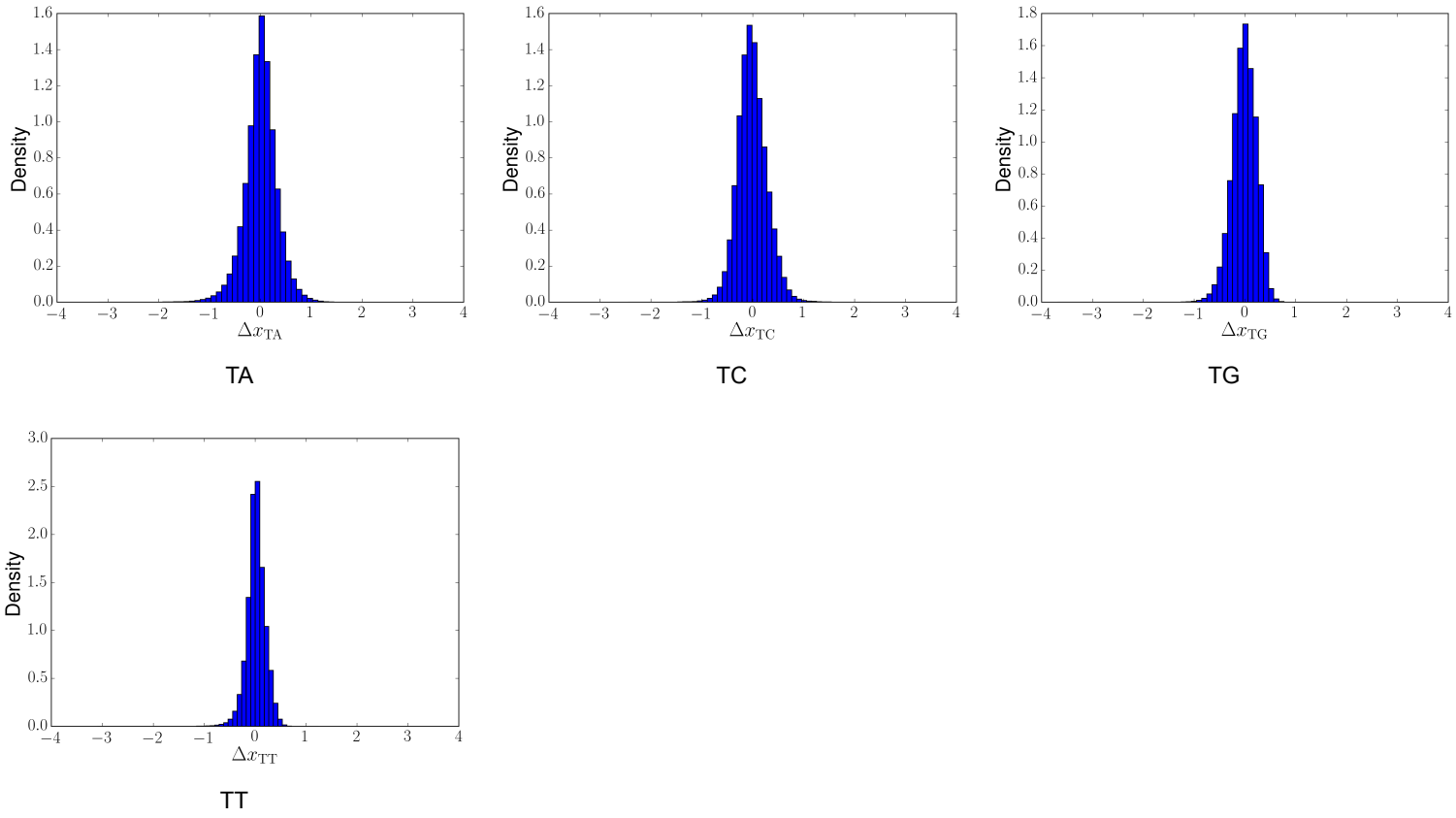
GC



GG



GT



Supplementary Figure 1 | Histograms of dinucleotide force change between the inserts and their consensus sequence for repeat families for respective dinucleotides.