Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells

Antoine Tanne[a], Luciana R. Muniz[a], Anna Puzio-Kuter[b], Katerina I. Leonova[c], Andrei V. Gudkov[c], David T. Ting[d], Rémi Monasson[e], Simona Cocco[f], Arnold J. Levine[b,g,1], Nina Bhardwaj[a,2], and Benjamin D. Greenbaum[a,g,h,1,2]

[a]Tisch Cancer Institute, Department of Medicine, Hematology, and Medical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [b]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903; [c]Roswell Park Cancer Institute, Buffalo, NY 14263; [d]Massachusetts General Hospital, Charlestown, MA 02129; [e]Laboratoire de Physique Théorique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [f]Laboratoire de Physique Statistique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [g]The Simons Center for Systems Biology, School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [h]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Contributed by Arnold J. Levine, September 10, 2015 (sent for review April 27, 2015; reviewed by Chakraborty Arup and Curtis G. Callan, Jr.)

Recent studies have demonstrated abundant transcription of a set of noncoding RNAs (ncRNAs) preferentially within tumors as opposed to normal tissue. Using a novel approach from statistical physics, we quantify global transcriptome-wide motif use for the first time, to our knowledge, in human and murine ncRNAs, determining that most have motif use consistent with the coding genome. However, an outlier subset of tumor-associated ncRNAs, typically of recent evolutionary origin, has motif use that is often indicative of pathogen-associated RNA. For instance, we show that the tumor-associated human repeat HSATII is enriched in motifs containing CpG dinucleotides in AU-rich contexts that most of the human genome and human adapted viruses have evolved to avoid. We demonstrate that a key subset of these ncRNAs functions as immunostimulatory "self-agonists" and directly activates cells of the mononuclear phagocytic system to produce proinflammatory cytokines. These ncRNAs arise from endogenous repetitive elements that are normally silenced, yet are often very highly expressed in cancers. We propose that the innate response in tumors may partially originate from direct interaction of immunogenic ncRNAs expressed in cancer cells with innate pattern recognition receptors, and thereby assign a new danger-associated function to a set of dark matter repetitive elements. These findings potentially reconcile several observations concerning the role of ncRNA expression in cancers and their relationship to the tumor microenvironment.

noncoding RNA | genome evolution | cancer immunology

The recent development of total RNA sequencing has allowed a better appreciation of the complexity and breadth of the entire transcriptome (1–4). Analysis by the Encyclopedia of DNA Elements (ENCODE) consortium unexpectedly showed that far more of the mammalian genome than previously appreciated is transcribed into noncoding RNA (ncRNA). Several short ncRNAs have conserved metabolic and regulatory functions, and some antiviral properties have been assigned to novel ncRNA classes, such as eukaryotic piRNA, piwi-interacting RNA, and prokaryotic CRISPR RNA (5). In eukaryotes, long noncoding RNA (lncRNA), such as long-intergenic ncRNA, has been associated with transcriptional, posttranscriptional, and epigenetic regulation (6, 7).

It is now evident that germ-line and cancer cells can have atypical ncRNA transcription, including repetitive elements from regions usually silenced in steady state (8, 9). In eukaryotes, transcription of endogenous retroviruses and mobile elements is mostly repressed epigenetically through processes such as histone modification and DNA methylation, preventing disruptive or deregulatory effects due to integration into coding regions. In mammals, DNA methylation targets the Cyd in CpG motifs to form 5-methyl Cyt, contributing to down-regulation of transcription for methylated sequences (10). Epigenetic regulation is strongly associated with the developmental process, whereas its deregulation, such as by disruption of DNA methylation, can be associated with dedifferentiation and carcinogenic processes (11, 12).

In cancers, such as those cancers driven by p53 mutations and epigenetic alterations, ncRNA associated with repetitive elements can be induced (8, 9). In a study of mouse and human epithelial malignancies by Ting et al. (9), several repetitive elements emanating from genomic dark matter and often repressed in steady-state conditions, particularly in pericentromeric repeats, such as GSAT (major satellite) in mouse and HSATII in humans, were only transcribed in cancer cells. Leonova et al. (8) demonstrated a strong induction of repetitive elements from the mouse genome (particularly GSAT, B1, and B2), along with several other ncRNAs, in cells bearing p53 oncogenic mutations and exposed to epigenome-altering demethylating agents. Anomalous expression of the murine repetitive element GSAT was shown to trigger transcription of the repeat-dependent activated IFN response, which can regulate apoptosis-related cell death. Similarly, when expressed, endogenous retroviral RNA can activate the innate immune response via several pathways (13). Altogether, these studies suggest that certain ncRNAs may also have attributes of immunostimulatory nucleic acid sequences.

We use a set of novel mathematical tools originally developed to analyze potentially immunostimulatory motif use in viral and host genome coding sequences. These methods were recently recast in the language of statistical physics and are extended here to analyze ncRNA motif use (14, 15). We analyze for the first time, to our knowledge, large-scale patterns of motif use in human and

Significance

Using an approach derived from theoretical statistical physics, we quantify transcriptome-wide motif usage in human and murine noncoding RNAs (ncRNAs), determining that most have motif usage consistent with the coding genome. However, an outlier subset of tumor-associated ncRNAs comprises repetitive elements whose motif usage patterns are more typically associated with the genomes of inflammatory pathogens. We demonstrate that a key subset of these elements directly activates the cellular innate immune response. We propose that the innate response in tumors partially originates from direct interaction of immunogenic ncRNAs preferentially expressed in cancer cells with innate pattern recognition receptors.

Author contributions: A.T., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. designed research; A.T., L.R.M., A.P.-K., R.M., S.C., and B.D.G. performed research; D.T.T., R.M., S.C., and B.D.G. contributed new reagents/analytic tools; A.T., L.R.M., A.P.-K., K.I.L., A.V.G., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. analyzed data; and A.T., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. wrote the paper.

Reviewers: C.A., MIT; and C.G.C., Princeton University

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: alevine@ias.edu or benjamin.greenbaum@mssm.edu.

[2]N.B. and B.D.G. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1517584112/-/DCSupplemental.

murine transcriptomes, which we use to find anomalies in ncRNA expressed in cancer transcriptomes (5, 16). As a result, we are able to characterize features of ncRNA overexpressed in cancerous cells relative to normal cells (8, 9, 17). Our analysis includes several large datasets of functionally characterized ncRNA, in addition to pseudogenes and repetitive elements, such as satellite DNA, endogenous retroviruses, and long and short interspersed elements. We demonstrate many ncRNAs preferentially expressed in cancerous cells display anomalous motif use patterns compared with the vast majority of ncRNAs whose patterns of motif use we show to be consistent with those patterns of motif use in coding regions. Based on their unusual pattern of motif use and differential expression in cancerous vs. normal cells, we predicted that HSATII and GSAT incorporate immunostimulatory motifs in humans and mice, respectively. Remarkably, we validate our prediction demonstrating that both directly stimulate antigen-presenting cells and accordingly label them immunostimulatory ncRNAs (i-ncRNAs).

Results

General Motif Use Patterns in lncRNAs. Using the GENCODE database of lncRNA transcripts from humans and mice (versions 19 and 2 for humans and mice, respectively) we calculated the strength of statistical bias (referred to as a force) on sequence motif use for all contained lncRNAs as described in *Materials and Methods*. GENCODE lncRNA established a baseline of sequence motif use expressed in a broad array of cells and tissues so that we could compare these patterns of motif use with those patterns of motif use of ncRNAs expressed in certain cancers. For each sequence, we calculate the force on all two- and three-nucleotide motifs and use Eq. **5** in *Materials and Methods* to calculate the probability of observing a sequence with that number of motifs. The number of sequences in GENCODE for which a given dinucleotide is aberrantly expressed is illustrated in Fig. 1*A*. CpG dinucleotides are vastly underrepresented, as indicated by their negative forces in *SI Appendix*, Table S1. UpA dinucleotides are often underrepresented, although to a lesser extent. As in our previous work, these patterns cannot be explained by nucleotide frequencies, such as ~~GC~~ content, which are accounted and normalized for in our method.

These dinucleotide motif use patterns are similar in human and mouse genomes across the wide array of cells and cell lines contained in GENCODE (2, 3). Strikingly, avoidance of the CpG and UpA dinucleotide motifs in this dataset is stronger than in coding regions (*SI Appendix*, Fig. S1). One can conclude that the patterns previously observed in virus and host coding genes are not due to effects from coding regions, such as codon use patterns (18–20). Rather, such constraints in coding regions likely weaken the strength of a statistical bias that comes from the same underlying mechanisms. This ■■■ suggests selective restrictions on dinucleotide frequencies observed in ncRNAs preserving a function or avoiding a detrimental consequence, such as a chronic autoinflammatory response that could result from presenting danger-associated molecular patterns (DAMPs). Adaptation of dinucleotide motif use in these elements over time is analogous to the viral mimicry of host patterns of sequence motif use (14, 21). When an avian influenza virus enters the human population, one can observe adaptation to analogous patterns emerging over time (14, 15, 22, 23). In that case, mutation rates in influenza are very high, so one can follow these evolutionary adaptations over far shorter time periods.

Trinucleotide motifs with significant forces are listed in the *SI Appendix*, Table S1, along with dinucleotide motifs. Trinucleotide motifs with significant forces acting on them are conserved between humans and mice, as was the case for dinucleotides, with the exception of UAC and UAG (which are significant in humans but less so in mice). Except for UAG (chain termination codons used in coding RNAs), whenever a trinucleotide motif is significantly enhanced or avoided in humans, its reverse complement is also significantly enhanced or avoided, suggesting avoidance of complementary motifs. The strongest forces suppress CpG and CpG-containing trinucleotides particularly when an A or U is next to the core CpG motif. ~~This ■■■ is~~ consistent with the avoidance of CpGs in AU contexts observed in influenza viruses replicating in humans (15, 22, 23). Given the apparent bias against CpG and UpA, we sought to determine if these motifs were linked. Pearson correlation between these forces across all GENCODE ncRNA in humans and mice showed no correlation between CpG and UpA biases ($r = 0.0006$; *SI Appendix*, Fig. S2). Therefore, the forces on CpG and UpA are likely independent. Moreover, every significant trimer across the GENCODE is correlated to CpG, UpA, or both. As a result, all significant trimers can be explained by their CpG or UpA motif use.

Cancer-Enriched Noncoding Repeat RNA May Have Anomalous Motif Use. Prior work revealed aberrant expression of ncRNA across a spectrum of mouse and human cancers (8, 9). These sequences were found in the Repbase database of human and murine repetitive elements and the Functional Annotation of Mouse (FANTOM) database of murine noncoding elements (currently NONCODE) (24, 25). We also found high induction of GSAT in a murine testicular teratoma and liposarcoma tumor model (8, 9) (*SI Appendix*, Fig. S3). Focusing on these cancer-expressed repeats, we surprisingly found a significant enrichment of anomalous motif use patterns compared with other ncRNAs. In the Repbase database, we tested whether the bias on dinucleotide and trinucleotide motifs observed in repetitive element sequences fell outside the distribution obtained from GENCODE lncRNA. Remarkably, we found hundreds of sequences falling outside of this distribution. Many have high use of CpG dinucleotides, including a set of endogenous viruses (*SI Appendix*, Table S2) recently implicated in the innate immune response in tumors (13). We conclude that although the portions of the noncoding regions typically expressed as lncRNAs have similar motif use patterns as RNA from coding regions, there are many genomic regions with atypical motif use that are not transcribed in normal cells or tissues.

We use the forces that quantify the strength of the statistical bias on the often underrepresented CpG and UpA dinucleotides to differentiate between ncRNAs found preferentially in cancerous cells and the total lncRNA referenced in GENCODE for humans and mice, because these two dinucleotides essentially account for all significant trinucleotide motifs in this set. We use the distribution of forces on CpG and UpA to define a null hypothesis, which we approximate by a Gaussian (Fig. 2). Many ncRNAs from cancerous cells are clearly outside the distribution, often to a large extent. In particular, HSATII, the main ncRNA up-regulated in human pancreatic cancers, is far outside the human distribution, and GSAT, the main murine ncRNA implicated in murine tumoral

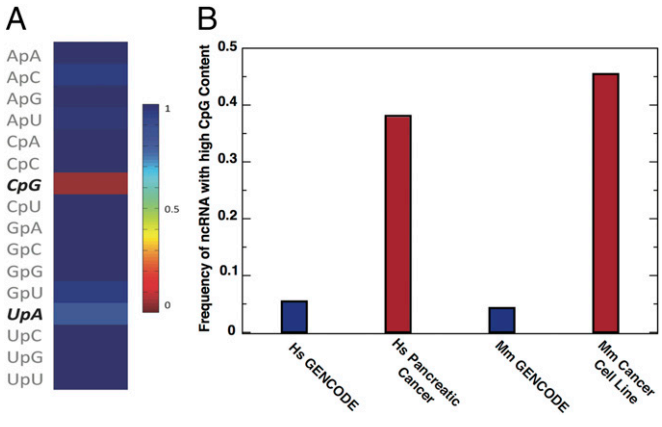

Fig. 1. ncRNAs expressed in cancer differ from general lncRNA motif use patterns. (*A*) Fraction of GENCODE human lncRNA sequences where a motif occurs the expected number of times as defined by corresponding to a probability greater than 0.05 (Eq. **5**). (*B*) Fraction of GENCODE lncRNA sequences in humans and mice where CpG motifs occur the expected number of times compared with the CpG motifs expressed in human cancerous cells and mouse cancer cell lines. ~~Hs, ■■■; Mm, ■■■.~~

cell lines, is well outside the mouse distribution. Within our null hypothesis, the P values for all ncRNAs considered here are less than 10^{-61} for human pancreatic cancer data and less than 10^{-2} for murine cell line data.

Many of the ncRNAs from the studies of Leonova et al. (8) and Ting, et al. (9) are outliers of at least three SDs with respect to at least one of the significant motifs implicated in the previous section, accounting for 70.46% of the modulated Repbase RNA expression induced in pancreatic cancer, along with even higher percentages (74.86% and 85.30%, respectively) in the smaller sets of prostate and lung cancers. HSATII is the most differentially expressed (by a considerable margin) in the pancreatic cancer data, and HSATII and BSR are the highest in prostate and lung cancer data. In p53 KO murine cell lines treated with demethylation agents, around 68 ncRNAs are significantly modulated (8). Among those ncRNAs, 78.96% of the total expression comes from outliers as defined above, with the vast majority coming from GSAT and B2. Overall, we observed that repetitive sequences containing unusual motif use had varying degrees of conservation. However, the subset preferentially expressed in cancerous cells and tissues is encoded by sequences of more recent evolutionary origin. HSATII and GSAT are only conserved back to primates and mice, respectively, and 21 of the 22 ncRNAs from the study of Ting et al. (9) are conserved in humans and primates but extend no further back in evolution. Any function is likely to be species-specific.

ncRNAs with Unusual Motif Use Highly Expressed in Cancers Are Immunostimulatory. Our analysis highlights that many ncRNAs up-regulated in cancer display abnormal nucleotide motif use that we had previously related to immunogenic properties in viruses. The innate immune system contains several effector cells that react to immunogenic nucleic acids, such as exogenous viral and bacterial nucleic acids, as well as endogenous nucleic acids that can be released upon cell death (6). Among those effectors, the mononuclear phagocytic system [macrophages, monocytes, and dendritic cells (DCs)] contains key regulators of innate immune activation and adaptive immunity (26–28). DCs efficiently sense and sample their environment to integrate information and mount a proper response, which may be tolerogenic or immunogenic. To test whether ncRNA with highly unusual motif use could be recognized as a DAMP by some nucleic acid-sensing pattern recognition receptors (PRRs), we studied the effect of human HSATII and murine GSAT following transfection in human monocyte-derived DCs (moDCs) and murine bone marrow-derived macrophages. Liposomal transfection was required for stimulation, whereas naked RNA had no effect, implying recognition is consistent with activation via an endosomal or intracellular sensor (*SI Appendix*, Fig. S4). The general sets of recognition pathways tested are indicated in the *SI Appendix*, Fig. S5.

We generated different ncRNAs by in vitro transcription using minigenes coding for the two main candidate outliers computationally predicted to have immunogenic motif use (HSATII and GSAT). As controls, we derived RNA from minigenes encoding scrambled (sc) versions with the same nucleotide content but having normal motif use (labeled HSATII-sc and GSAT-sc) and repetitive elements of comparable length but having normal motif use patterns (RMER33 and UCON18), as described in *SI Appendix*. In human moDCs, liposomal transfection of HSATII induced significant production of IL-6, IL-12, and TNF-alpha relative to both endogenous controls and their scrambled versions (Fig. 3*A*). A similar profile of cytokines was elicited by moDCs in response to selected Toll-like receptor (TLR) agonists (*SI Appendix*, Fig. S6*A*). The candidate murine immunogenic ncRNA, GSAT, had less pronounced immunogenic properties but still induced IL-12 (Fig. 3*A*). Upon liposomal transfection of the same ncRNA into immortalized murine bone marrow-derived macrophages (imBMs), the immunogenic properties of HSATII were strongly attenuated, whereas the murine GSAT induced high levels of TNF-alpha (Fig. 3*B*) and monocyte chemotactic protein 1 (MCP-1), but not IFN-gamma, IL-6, or IL-12. The imBM almost exclusively regulates TNF-alpha in response to PRR agonists (*SI Appendix*, Fig. S6*B*).

HSATII and GSAT ncRNA induced IL-12 in human moDCs similar to the TLR3 ligand poly-IC (a synthetic dsRNA mimic; *SI Appendix*, Fig. S5). The absence of an effect by ncRNA with normal motif use [i.e., the scrambled forms (Fig. 3 *A* and *B*)] suggests specific sequence patterns within the RNA, such as CpG and UpA motifs, regulate immunostimulatory activity. Such motif use could

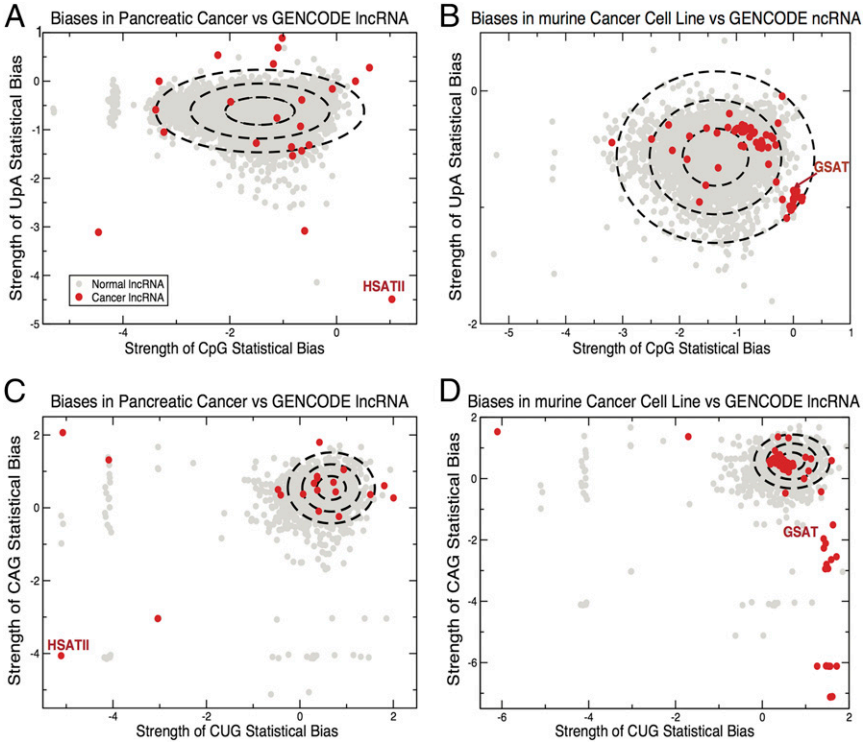

Fig. 2. ncRNA from cancer cells contains outliers from normal motif use. Distribution of UpA and CpG bias in lncRNA taken from human tumors (*A*) and murine cell lines (*B*) (indicated in red) plotted against lncRNA from GENCODE (indicated in gray). Each ellipse indicates 1 SD from the mean value in the GENCODE dataset. The forces on CAG and CUG are also shown for human tumors (*C*) and murine cell lines (*D*).

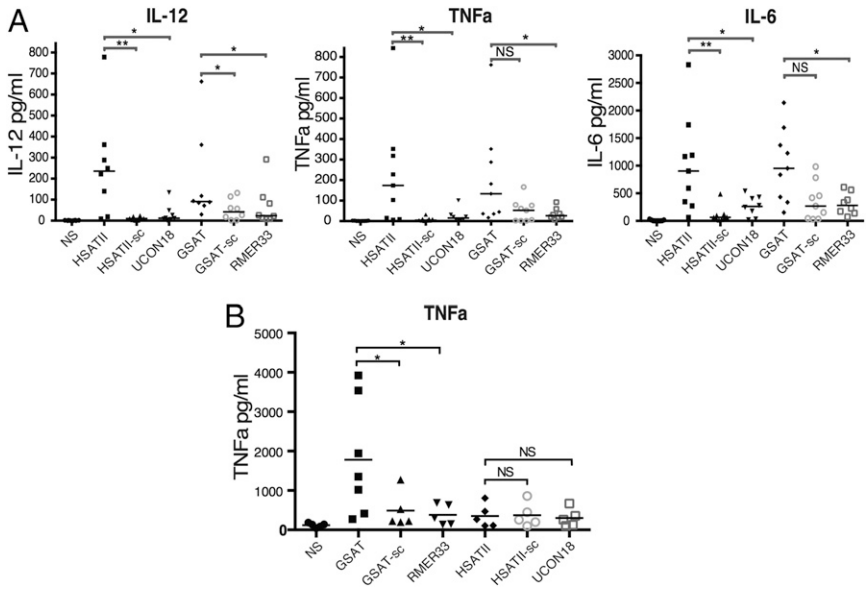

Fig. 3. i-ncRNA stimulates human moDC cytokine production. Quantification of inflammatory cytokine production in human moDCs (*A*) and murine imBMs (*B*) upon liposomal transfection of human i-ncRNA (HSATII) and murine i-ncRNA (GSAT) vs. their scrambled and endogenous controls. Each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median. The significance of i-ncRNA stimulation is analyzed by the nonparametric Mann–Whitney test to compare their effect vs. their scrambled and endogenous controls. NS, ~~RMER, UCON18,~~ *P ; **P .

also influence secondary conformation that may contribute to immunogenic properties, although we checked that the scrambled sequences did not lower the RNA minimum folding energy. Based upon these observations, we refer to HSATII and GSAT as immunogenic ncRNA or i-ncRNA. Interestingly, our study corroborates previous findings by Leonova et al. (8) that ncRNA, such as GSAT, can induce an innate response, although in those studies, the type I IFN pathway was also activated. Our initial investigations into this pathway were inconclusive (*SI Appendix*, Fig. S6*C*).

Dissection of the Immunostimulatory Properties of i-ncRNA. Pathogen-associated molecular patterns and DAMPs activate innate immune cells through PRRs. To characterize better the mechanisms involved in sensing i-ncRNA, we studied the immunomodulatory properties of HSATII and GSAT on a panel of imBMs that lack specific PRRs or effector molecules in their downstream signaling pathways (*SI Appendix*, Fig. S5). Whereas GSAT induced a TNF-alpha response, HSATII did not induce differential cytokine expression in these immortalized cells, indicating that there is either a species-specific effect, because the cells are murine, or a cell type-specific effect, because these cells are macrophages. This is perhaps unsurprising, because different species and cell types express different PRRs, and HSATII and GSAT have different sequence compositions. Significantly, the absence of two key adaptor and regulatory proteins, MYD88 and UNC93B1:UNC93B3d (UNC93b), respectively, eliminated the differential response to GSAT in imBMs (Fig. 4).

MYD88 is a key cytosolic adaptor protein that is used by all TLRs except TLR3 to activate the transcription factor NF-κB. Similarly, the mutated form of UNC93b essentially eliminated inflammatory responses in imBMs. Although less well characterized than MYD88, this protein is known to interact with several endosomal TLRs (TLR3, TLR7, and TLR9) and has been implicated in TLR trafficking between the endoplasmic reticulum and endosomes, and their resultant maturation (29–31). We tested the requirement for TLR3, TLR7, and TLR9, which are known to recognize dsRNA, ssRNA, and CpG DNA, respectively (32–34) (*SI Appendix*, Fig. S7*A* and S8). None of these receptors were required for GSAT to activate TNF-alpha production from imBMs. Additional pathways investigated, including the stimulator of IFN genes (STING) and inflammasome pathways, are discussed in *SI Appendix* and did not contribute to i-ncRNA stimulatory activity. Altogether, our data are consistent with a requirement for i-ncRNA activation through signaling pathways that rely upon MYD88 and UNC93b. The precise receptor involved in initial recognition remains to be determined.

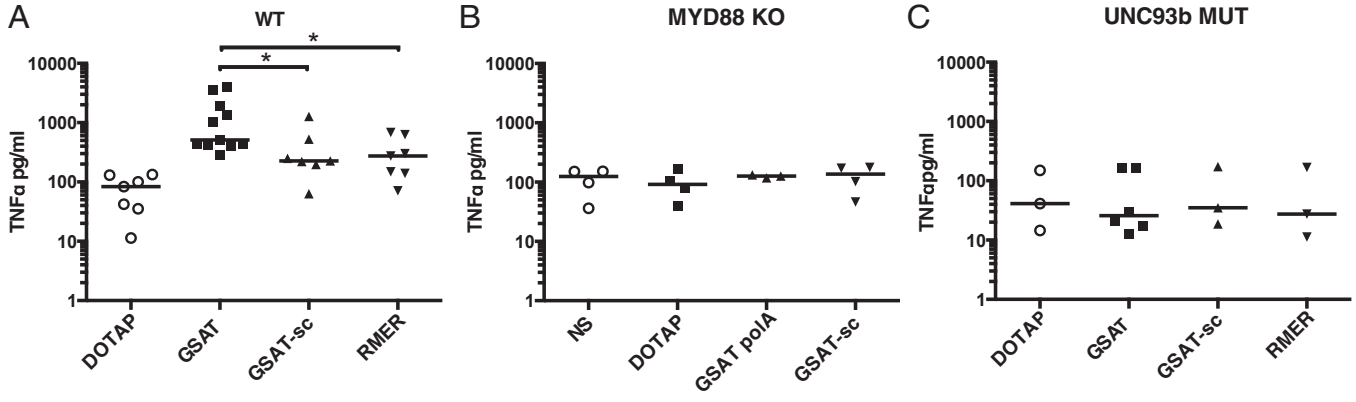

Fig. 4. MYD88 and UNC93b control GSAT i-ncRNA stimulation. Genetic screen of the innate immune pathway related to i-ncRNA function in murine imBMs. The imBM cells of different genotypes (WT, MYD88 KO, and UNC93b$^{3d/3d}$ MUT) have been stimulated by liposomal transfection of the murine i-ncRNA (GSAT). TNF-alpha (TNF-a) production in the supernatant has been quantified, and each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median. ~~DOTAP,~~ *P .

Discussion

There is a surprising similarity to be drawn between foreign viral nucleotide sequences and select ncRNAs silent in normal cells, yet transcribed in cancer cells, activating innate immunity (23, 29, 35–37). We determined that ncRNAs expressed predominantly in normal cells from humans and mice reflect patterns of nucleotide sequence motif avoidance, such as underrepresentation of CpG-containing sequences and reduced UpA, similar to protein-coding RNA. ~~This~~ [redacted] often includes a many-fold underrepresentation of CpG-containing sequences and reduced UpA motif use compared with expected levels. However, the genome also harbors repetitive elements, which often have abnormal use of CpG and UpA motifs compared with the use of CpG and UpA motifs observed in RNA expressed in normal cells and tissues. Sets of these ncRNAs, typically newer genome entries over evolutionary time scales, can be expressed at very high levels in cancerous cells and tumors. ~~This [redacted] is why~~ human and mouse elements expressed in cancer cells can have different sequences but can share high CpG content and are not generally observed in the human or mouse transcriptome in normal cells.

We previously proposed immunostimulatory and proinflammatory properties of highly inflammatory influenza and other RNA viruses derive, in part, from RNA containing CpGs in AU-rich contexts, which are avoided in RNA viruses circulating in humans. Experimental evidence has supported this hypothesis (23, 38, 39). We recently recast our analysis in the language of statistical physics in a way that is theoretically insightful and computationally efficient (15). In this language, the evolution and optimization of nucleotide sequence motifs are driven by the interplay between selective and entropic forces. The latter randomize motif frequencies in a genome under constraints, whereas the former are largely Darwinian, optimizing for functions enhancing viral replication and spreading. However, ncRNAs mostly transcribed in cancerous cells would not be exposed to the same selective and entropic forces as coding RNAs and ncRNAs transcribed in normal cells. Based on motif use patterns, we predicted many ncRNAs may have immunogenic properties, presenting DAMPs.

We focused experimentally on HSATII and murine GSAT, because they are preferentially and highly expressed in carcinogenic processes and exhibit abnormal patterns of motif use. In particular, human HSATII is enriched in CpG motifs in AU-rich contexts avoided in genomes of humans and human-adapted viruses. We demonstrate that their computationally predicted immunogenic properties lead to the induction of inflammatory cytokines in human and murine innate cells (Fig. 3 *A* and *B*). Our observations, together with previous work by Leonova et al. (8), strongly suggest that these endogenous i-ncRNAs are recognized as DAMPs by cellular nucleic acid PRRs.

We identified a key role for MYD88 and UNC93b as regulators of GSAT immunogenicity, but without evidence for the common endosomal nucleic acid sensors typically regulated by UNC93b or associated with the MYD88 adaptor (TLR2, TLR4, TLR7, and TLR9). Our results indicate that in the murine imBM background, there is potent induction of TNF-alpha. Further studies will be required to elucidate whether TLR13, which has been identified in murine cells and recognizes ribosomal bacterial and viral RNA, is involved, or whether there exist intracellular sensors of i-ncRNA associated with MYD88 (40–42), as there are for dsDNA (DHX-9 or DHS-36) (43). Interestingly, we find alignment of GSAT contains a subsequence conserved in immunogenic RNA isolated from bacterial ribosomal RNA, which specifically activates murine TLR13 (41).

Activation of innate immune signaling can contribute to either carcinogenesis or antitumoral immunity. TLR signaling and MYD88 have been associated with tumor development (44). Given that HSATII and GSAT expression has been found to be pervasive in many tumor types and induces responses that differ by species or cell type, the role of i-ncRNA in tumorigenesis is likely dependent on the particular RNA expressed and other properties of the tumor microenvironment. For instance, HSATII activates macrophages and monocytes in our study, suggesting it may be a mechanism for attraction and retention of tumor-associated macrophages. These macrophages have consistently been shown to be a poor prognostic in cancer, leading to increased tumorigenesis, metastasis, and immunoevasion (45). Under this hypothesis, HSATII is used by the tumor to keep macrophages in the tumor microenvironment while driving out T cells. Interestingly, the viral-like behavior of HSATII transcripts is found not only in the immune response to these elements but also in their ability to reverse-transcribe in cancer cells akin to retroviruses. Q:19

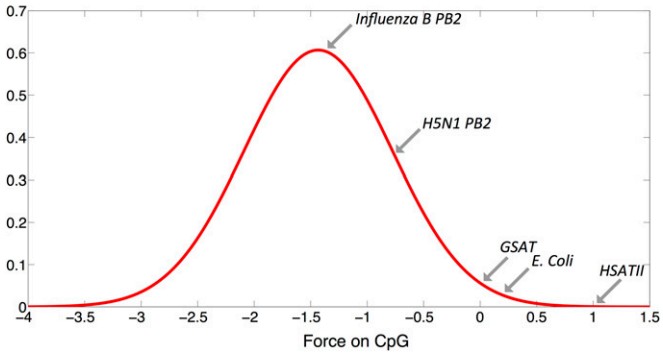

Fig. 5. Motif use in HSATII and GSAT clusters with foreign RNA. A comparison of the forces on CpG dinucleotides is plotted against the distribution of forces on all GENCODE lncRNA relative to a sequences nucleotide bias. The force on CpG dinucleotides for HSATII and GSAT is shown on the distribution, along with the average values for the longest gene (PB2) in human influenza B and avian H5N1 and all *Escherichia coli* coding regions.

The i-ncRNA, not subject to the same forces as ncRNA transcribed in steady state, may retain or evolve to mimic features of foreign RNA, as seen by comparing HSATII and GSAT with typical human ncRNA and foreign genomic material in Fig. 5 (15, 46). Indeed, HSATII and GSAT cluster more closely, in terms of motif use patterns, with bacterial rather than human RNA. Such RNA may have been selected to identify and eliminate cells when their epigenetic state is disrupted. Essentially self-"junk" RNA may have been maintained or may have evolved to mimic non–self-pathogen–associated patterns to create a danger signal. We propose that such a mechanism would be a ~~new~~ aspect of "genetic mimicry," where the host is, for all practical purposes, mimicking pathogen-associated nucleic acid patterns. HSATII and GSAT emanate from the pericentromeres, which harbor new repetitive elements with no known function (47). This region, unlike centromeres or regions critical for structure or regulation, may dynamically produce unusual repetitive elements that can adapt to a particular organism's PRRs. Our studies indicate that under the "extraordinary" circumstances where these repetitive elements are expressed, they could play a critical role in the regulation of immune responses against cancer. Q:20

Materials and Methods

~~**Entropy of Nucleotide Sequences for a Given Motif.**~~ We consider an RNA sequence of length L, hereafter called S_0, and a motif m [a series of contiguous nucleotides (e.g., CpG)]. Our objective is to define a probabilistic model over the set of the 4^L sequences, $S = (s_1\ s_2 \ldots\ s_i \ldots\ s_L)$, such that the average value of the number, $N_m(S)$, of occurrences of the motif m in S coincides with the number, $N_m(S_0)$, of occurrences of that motif in S_0. To do so, we consider a random-nucleotide model, where nucleotides are independently distributed according to the frequencies $f^0(s)$, with $s = A,C,G,U$, found in S_0. We then introduce the weakest bias that allows us to reproduce $N_m(S_0)$ on average. Q:21

The probability of a sequence S in this least-constrained, maximum entropy model is

$$P(S|x,m) = \frac{1}{Z_m(x)} \prod_{i=1}^{L} f^0(s_i)\ exp(x\ N_m(S)), \qquad [1]$$

where

$$Z_m(x) = \sum_{sequences\ S} \prod_{i=1}^{L} f^0(s_i)\ exp(x\ N_m(S)) \quad [2]$$

ensures the probability is correctly normalized. Parameter x, referred to as a selective force (or just force) on the motif m, introduces a statistical bias over P (15). The force quantifies the strength of statistical bias, which may be due to selection on a motif. In the absence of bias ($x = 0$), the probability of S simplifies to the product of its nucleotide frequencies, and the number of motifs is what one would expect in a typical sequence with nucleotide frequencies given by $f^0(s)$. Positive values for x push the distribution toward sequences with $N_m(S)$ larger than what one would expect, whereas negative values for x favor sequences with a smaller $N_m(S)$ than expected.

The value of the force, $x(S_0)$, is computed by maximizing the probability $P(S_0|x,\ m)$ of the sequence S_0 over x. This is equivalent to finding the value of x such that the average number of motifs,

$$N_m^{av}(x) = \sum_{sequences\ S} P(S|x,\ m)\ N_m(S) = \frac{\partial \log Z_m}{\partial x}(x), \quad [3]$$

equals $N_m(S_0)$. By scanning the sequences S_0 in the GENCODE database, we obtain the forces $x(S_0)$ shown in Fig. 2.

The logarithm of the number of sequences having $N_m(S)$ repetitions of m is bounded from above by the entropy of the random-nucleotide model; the equality is reached in the absence of bias only ($x = 0$). The difference between those entropies is the entropy cost corresponding to the constraint on the average number of occurrences of m, and is denoted by σ_m. It is the Legendre transform of $\log Z_m(x)$ (Eqs. **2** and **3**):

$$\sigma_m = x(S_0)\ N_m(S_0) - \log Z_m(x(S_0)). \quad [4]$$

Efficient computational techniques allow us to calculate the sum over the 4^L sequences in Eq. **2** in a time growing only linearly with L.

Our aim is to find anomalous motif use in a sequence where the number of motif occurrences is different from what is expected by chance in the random-nucleotide model (i.e., associated with a significant nonzero force). We express the likelihood of observing the natural sequence S_0 with a given motif count as

$$P(S^0|m) = \max_x \left[P(S^0|x,\ m)\right] = e^{\sigma_m} \prod_i f^0(s_i^0). \quad [5]$$

This likelihood is therefore directly related to the entropic cost: The larger the cost, the more likely is the motif to be statistically significant.

ACKNOWLEDGMENTS. We thank Dr. K. Fitzgerald (University of Massachusetts Medical School), Dr. R. Vance (University of California, Berkeley), Dr. G. Barton (University of California, Berkeley), and BEI resource [American Type Culture Collection/National Institute of Allergy and Infectious Diseases (NIAID)] for helping us collect murine immortalized macrophages. We also thank Dr. N. Vabret for many helpful discussions and A. Munk for all of his assistance. B.D.G. was supported by NIH [National Cancer Institute (NCI)] Grant 5P01CA087497-13; N.B. was supported by NIH (NIAID) Grants 5R01AI081848-05 and 5R01AI081848-05, NCI Grant 1R01CA180913-01A1, and the Cancer Research Institute; D.T.T. was supported by NIH (NCI) Grant K12CA087723-11A1, Department of Defense (US Army) Grant W81XWH-13-1-0237, and the Burroughs Welcome Fund; and R.M. and S.C. were supported by L'Agence Nationale de la Recherche Grant ANR-13-BS04-0012-01.

1. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
2. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
3. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
4. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nat Rev Genet* 12(10):671–682.
5. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.
6. Atianand MK, Fitzgerald KA (2013) Molecular basis of DNA recognition in the immune system. *J Immunol* 190(5):1911–1918.
7. Zhang K, et al. (2014) The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* 547(1):1–9.
8. Leonova KI, et al. (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(1):E89–E98.
9. Ting DT, et al. (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331(6017):593–596.
10. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. *Science* 293(5532):1068–1070.
11. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nat Rev Cancer* 4(2):143–153.
12. Yi L, Lu C, Hu W, Sun Y, Levine AJ (2012) Multiple roles of p53-related pathways in somatic cell reprogramming and stem cell differentiation. *Cancer Res* 72(21):5635–5645.
13. Zeng M, et al. (2014) MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. *Science* 346(6216):1486–1492.
14. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog* 4(6):e1000079.
15. Greenbaum BD, Cocco S, Levine AJ, Monasson R (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proc Natl Acad Sci USA* 111(13):5054–5059.
16. Ulitsky I, Bartel DP (2013) lincRNAs: Genomics, evolution, and mechanisms. *Cell* 154(1):26–46.
17. Levine AJ, Greenbaum B (2012) The maintenance of epigenetic states by p53: The guardian of the epigenome. *Oncotarget* 3(12):1503–1504.
18. Coleman JR, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
19. Mueller S, et al. (2010) Live attenuated influenza virus vaccines by computer-aided rational design. *Nat Biotechnol* 28(7):723–726.
20. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J Virol* 80(19):9687–9696.
21. Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* 68(5):2889–2897.
22. Greenbaum BD, Rabadan R, Levine AJ (2009) Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. *PLoS One* 4(6):e5969.
23. Jimenez-Baranda S, et al. (2011) Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. *J Virol* 85(8):3893–3904.
24. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462–467.
25. Xie C, et al. (2014) NONCODEv4: Exploring the world of long non-coding RNA genes. *Nucleic Acids Res* 42(Database issue):D98–D103.
26. Guilliams M, et al. (2014) Dendritic cells, monocytes and macrophages: A unified nomenclature based on ontogeny. *Nat Rev Immunol* 14(8):571–578.
27. Kroemer G, Galluzzi L, Kepp O, Zitvogel L (2013) Immunogenic cell death in cancer therapy. *Annu Rev Immunol* 31:51–72.
28. Sabado RL, Bhardwaj N (2013) Dendritic cell immunotherapy. *Ann N Y Acad Sci* 1284:31–45.
29. Casrouge A, et al. (2006) Herpes simplex virus encephalitis in human UNC-93B deficiency. *Science* 314(5797):308–312.
30. Lee BL, et al. (2013) UNC93B1 mediates differential trafficking of endosomal TLRs. *eLife* 2:e00291.
31. Tabeta K, et al. (2006) The Unc93b1 mutation 3d disrupts exogenous antigen presentation and signaling via Toll-like receptors 3, 7 and 9. *Nat Immunol* 7(2):156–164.
32. O'Neill LA, Golenbock D, Bowie AG (2013) The history of Toll-like receptors - redefining innate immunity. *Nat Rev Immunol* 13(6):453–460.
33. Broz P, Monack DM (2013) Newly described pattern recognition receptors team up against intracellular pathogens. *Nat Rev Immunol* 13(8):551–565.
34. Gajewski TF, Schreiber H, Fu YX (2013) Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* 14(10):1014–1022.
35. Bogunovic D, et al. (2009) Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci USA* 106(48):20429–20434.
36. Kayagaki N, et al. (2011) Non-canonical inflammasome activation targets caspase-11. *Nature* 479(7371):117–121.
37. Cosset É, et al. (2014) Comprehensive metagenomic analysis of glioblastoma reveals absence of known virus despite antiviral-like type I interferon gene response. *Int J Cancer* 135(6):1381–1389.
38. Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P (2014) The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Res* 42(7):4527–4545.
39. Vabret N, et al. (2012) The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. *PLoS One* 7(4):e33502.
40. Li XD, Chen ZJ (2012) Sequence specific detection of bacterial 23S ribosomal RNA by TLR13. *eLife* 1:e00102.
41. Oldenburg M, et al. (2012) TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. *Science* 337(6098):1111–1115.
42. Shi Z, et al. (2011) A novel Toll-like receptor that recognizes vesicular stomatitis virus. *J Biol Chem* 286(6):4517–4524.
43. Kim T, et al. (2010) Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase A helicases sense microbial DNA in human plasmacytoid dendritic cells. *Proc Natl Acad Sci USA* 107(34):15181–15186.
44. Wang JQ, Jeelall YS, Ferguson LL, Horikawa K (2014) Toll-like receptors and cancer: MYD88 mutation and inflammation. *Front Immunol* 5:367.
45. Noy R, Pollard JW (2014) Tumor-associated macrophages: From mechanisms to therapy. *Immunity* 41(1):49–61.
46. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006.
47. Maumus F, Quesneville H (2014) Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun* 5:4104.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 1

Q: 1_Please contact PNAS_Specialist.djs@sheridan.com if you have questions about the editorial changes, this list of queries, or the figures in your article. Please include your manuscript number in the subject line of all email correspondence; your manuscript number is 201517584.

Q: 2_Please (i) review the author affiliation and footnote symbols carefully, (ii) check the order of the author names, and (iii) check the spelling of all author names, initials, and affiliations. Please check with your coauthors about how they want their names and affiliations to appear. To confirm that the author and affiliation lines are correct, add the comment "OK" next to the author line. This is your final opportunity to correct any errors prior to publication. Misspelled names or missing initials will affect an author's searchability. Once a manuscript publishes online, any corrections (if approved) will require publishing an erratum; there is a processing fee for approved erratum.

Q: 3_Please review and confirm your approval of the short title: Properties of noncoding RNAs expressed in cancer. If you wish to make further changes, please adhere to the 50-character limit. (NOTE: The short title is used only for the mobile app and the RSS feed.)

Q: 4_Please review the information in the author contribution footnote carefully. Please make sure that the information is correct and that the correct author initials are listed. Note that the order of author initials matches the order of the author line per journal style. You may add contributions to the list in the footnote; however, funding should not be an author's only contribution to the work.

Q: 5_You have chosen the open access option for your paper and have agreed to pay an additional $1350 (or $1000 if your institution has a site license). Please confirm this is correct and note your approval in the margin.

Q: 6_Please verify that all supporting information (SI) citations are correct. Note, however, that the hyperlinks for SI citations will not work until the article is published online. In addition, SI that is not composed in the main SI PDF (appendices, datasets, movies, and "Other Supporting Information Files") have not been changed from your originally submitted file and so are not included in this set of proofs. The proofs for any composed portion of your SI are included in this proof as subsequent pages following the last page of the main text. If you did not receive the proofs for your SI, please contact **PNAS_Specialist.djs@sheridan.com**.

Q: 7_PNAS allows up to five keywords. You may add two keywords. Also, please check the order of your keywords and approve or reorder them as necessary.

Q: 8_Please confirm whether all units/divisions/departments/laboratories/sections have been included in the affiliations line for each footnote symbol or add if missing. PNAS requires smallest institutional unit(s) to be listed for each author in each affiliation.

Q: 9_PNAS discourages claims of priority; is "novel approach" truly novel in abstract? If not, please either (a) replace the term "novel" with a term such as "previously unidentified" or (b) remove it altogether to avoid the claim of priority.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 2

Q: 10_Please spell out HSATII in abstract. You have the option of following the expanded term with (HSATII) if you wish.

Q: 11_PNAS discourages claims of priority; is "new danger-associated function" truly new in abstract? If not, please either (a) replace the term "new" with a term such as "previously unidentified" or (b) remove it altogether to avoid the claim of priority.

Q: 12_PNAS articles should be accessible to a broad scientific audience. As such, please spell out "piwi" and CRISPR" (Several short ncRNAs have conserved metabolic and regulatory functions, and some antiviral properties have been assigned to novel ncRNA classes, such as eukaryotic siRNA, piwi-interacting RNA, and prokaryotic CRISPR RNA).

Q: 13_Please note that abbreviations for amino acids are used throughout your paper per PNAS style.

Q: 14_PNAS discourages claims of priority; are "novel mathematical tools" truly novel (We use a set of novel mathematical tools originally developed to analyze potentially immunostimulatory motif use in viral and host genome coding sequences)? If not, please either (a) replace the term "novel" with a term such as "previously unidentified" or (b) remove it altogether to avoid the claim of priority.

Q: 15_PNAS articles should be accessible to a broad scientific audience. As such, please spell out GC (these patterns cannot be explained by nucleotide frequencies, such as GC content).

Q: 16_PNAS mandates unambiguous pronoun antecedents. Please provide an appropriate noun after "This" in this sentence (This ■■■ suggests selective restrictions on dinucleotide frequencies observed in ncRNAs preserving a function or avoiding a detrimental consequence) and throughout remaining text whenever an unambiguous pronoun antecedent has been used instead of a noun.

Q: 17_PNAS articles should be accessible to a broad scientific audience. As such, MCP1 has been spelled out as "monocyte chemotactic protein 1" [whereas the murine GSAT induced high levels of TNF-α (Fig. 3*B*) and monocyte chemotactic protein 1 (MCP-1)]. Please revise as necessary.

Q: 18_PNAS articles should be accessible to a broad scientific audience. As such, STING has been spelled out as "stimulator of IFN genes" [Additional pathways investigated, including the stimulator of IFN genes (STING) and inflammasome pathways]. Please revise as necessary.

Q: 19_PNAS no longer allows citations of unsupported data; the citation for unpublished data (Bersani et al, PNAS, in press) has thus been removed from the text. Please (a) provide the data as Supporting Information or (b) provide an "in press" reference if the article has been accepted for publication.

Q: 20_PNAS discourages claims of priority; is "new aspect" truly new (We propose that such a mechanism would be a new aspect of "genetic mimicry,")? If not, please either (a) replace the term "new" with a term such as "previously unidentified" or (b) remove it altogether to avoid the claim of priority.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 3

Q: 21_Per PNAS policy, the use of a single level 2 heading in a level 1 section should be avoided. Please provide a second level 2 heading or delete this single level 2 heading (Wntropy of Nucleotide Sequences for a Given Motif) in this level 1 section (Materials and Methods).

Q: 22_Please spell out BEI in Acknowledgments section.

Q: 23_Please provide an issue number for ref. 5 if possible.

Q: 24_The references by Kayagaki et al. and the reference by Cosset et al. were both numbered reference 36 in the original version of your paper. The paper by Cosset et al. has now been numbered referenced 37 and cited as such in main text. All subsequent references and citations have also been renumbered.

Q: 25_Please provide definitions for Hs and Mm in legend for Fig. 1.

Q: 26_Please provide definitions for NS, RMER, and UCON18 in legend for Fig. 3. Also provide *P* values for * and ** in figure legend.

Q: 27_Please provide definition for DOTAP in legend for Fig. 4. Also provide *P* value for * in figure legend.