PLOS BIOLOGY

RESEARCH ARTICLE

The antiviral state has shaped the CpG composition of the vertebrate interferome to avoid self-targeting

Andrew E. Shaw[1,2☯‡], **Suzannah J. Rihn**[1☯‡], **Nardus Mollentze**[1,3☯‡], **Arthur Wickenhagen**[1], **Douglas G. Stewart**[1], **Richard J. Orton**[1], **Srikeerthana Kuchi**[1], **Siddharth Bakshi**[1], **Mila Rodriguez Collados**[1], **Matthew L. Turnbull**[1], **Joseph Busby**[1], **Quan Gu**[1], **Katherine Smollett**[1], **Connor G. G. Bamford**[1], **Elena Sugrue**[1], **Paul C. D. Johnson**[1,3], **Ana Filipe Da Silva**[1], **Alfredo Castello**[1], **Daniel G. Streicker**[1,3], **David L. Robertson**[1], **Massimo Palmarini**[1], **Sam J. Wilson**[1]*

1 MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow, United Kingdom, **2** The Pirbright Institute, Woking, United Kingdom, **3** Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

☯ These authors contributed equally to this work.
‡ These authors share first authorship on this work.
* Sam.Wilson@glasgow.ac.uk

Citation: Shaw AE, Rihn SJ, Mollentze N, Wickenhagen A, Stewart DG, Orton RJ, et al. (2021) The antiviral state has shaped the CpG composition of the vertebrate interferome to avoid self-targeting. PLoS Biol 19(9): e3001352. https://doi.org/10.1371/journal.pbio.3001352

Academic Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: October 7, 2020

Accepted: July 7, 2021

Published: September 7, 2021

Data Availability Statement: The underlying data are available from the University of Glasgow Enlighten database (http://dx.doi.org/10.5525/gla.researchdata.1159). The underlying code is openly available (https://doi.org/10.5281/zenodo.5035607 and https://doi.org/10.5281/zenodo.5036224). The raw fastq files generated during this project have been submitted to the European Bioinformatics Institute (EBI) under project accession numbers PRJEB29677 and PRJEB39825.

Abstract

Antiviral defenses can sense viral RNAs and mediate their destruction. This presents a challenge for host cells since they must destroy viral RNAs while sparing the host mRNAs that encode antiviral effectors. Here, we show that highly upregulated interferon-stimulated genes (ISGs), which encode antiviral proteins, have distinctive nucleotide compositions. We propose that self-targeting by antiviral effectors has selected for ISG transcripts that occupy a less self-targeted sequence space. Following interferon (IFN) stimulation, the CpG-targeting antiviral effector zinc-finger antiviral protein (ZAP) reduces the mRNA abundance of multiple host transcripts, providing a mechanistic explanation for the repression of many (but not all) interferon-repressed genes (IRGs). Notably, IRGs tend to be relatively CpG rich. In contrast, highly upregulated ISGs tend to be strongly CpG suppressed. Thus, ZAP is an example of an effector that has not only selected compositional biases in viral genomes but also appears to have notably shaped the composition of host transcripts in the vertebrate interferome.

Introduction

Vertebrates have evolved a multitude of strategies to sense invading pathogens and deploy the appropriate immune defenses. A common outcome of pathogen sensing is the secretion of type I interferons (IFNs), which upregulate hundreds of interferon-stimulated genes (ISGs) [1–4]. Many ISGs interfere with viral replication, creating a hostile antiviral state within the IFN-stimulated cell [3].

The replication of every virus involves at least 1 viral RNA, and these molecules are frequent targets of both antiviral sensors and antiviral effectors [5]. To sense or target viral RNAs, host

Funding: This study was funded by Medical Research Council (MRC) grants MR/K024752/1 (to SJW), MC_UU_12014/10 (to SJW and MP), MC_UU_12014/12 (to DLR), MR/R021562/1 (to ACP) and MR/P022642/1 (to SJW and SJR) as well as Wellcome Trust support 201366/Z/16/Z (to SJR) and 217221/Z/19/Z (to DG Streicker). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: CAI, Codon Adaptation Index; CDS, coding sequence; DE, differentially expressed; DMEM, Dulbecco's Modified Eagle Medium; FCS, fetal calf serum; FDR, false discovery rate; GC, guanine–cytosine; GO, gene ontology; HSV-1, human alphaherpesvirus 1; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; KO, knockout; PAMP, pathogen-associated molecular pattern; RNA-seq, RNA sequencing; RVFV, Rift Valley fever phlebovirus; sgRNA, single-guide RNA; VSV, vesicular stomatitis virus; ZAP, zinc-finger antiviral protein.

factors must possess the ability to discriminate self-RNAs from nonself-RNAs [5], and, to achieve this, the host exploits molecular features that are rare or absent in host transcripts. One of these features is the CpG dinucleotide (a cytosine base followed by a guanine), which is remarkably underrepresented in vertebrate genomes [6]. Synthetic viruses with enriched CpG content are severely attenuated in human cells as they succumb to CpG-targeting host defenses [7–10]. The zinc-finger antiviral protein (ZAP), an ISG, has evolved to contain a binding pocket that can accommodate a CpG dinucleotide, but no other dinucleotide [11]. ZAP binds to specific CpGs in viral RNAs and can target their degradation [10,12,13]. ZAP, and its binding partner TRIM25 [14,15], possess no intrinsic nuclease activity and mediate RNA degradation by recruiting cofactors such as the putative nuclease KHNYN [16]. Crucially, not all CpGs are equally targeted by ZAP, and the context of the CpGs, as opposed to the number of CpGs, appears to define whether a particular CpG or transcript is targeted by ZAP [10,17,18]. Nonetheless, the number of CpGs is a useful feature, as increased CpG frequency increases the likelihood that a CpG is presented in a context recognised by ZAP. Although the contextual features underlying ZAP targeting are not fully understood, it has recently been proposed that the CpG motif must be presented in a region of single-stranded RNA [19].

It is widely believed that the majority of RNA viruses infecting vertebrates have evolved to possess a relatively low CpG content [20] to escape CpG-targeting defenses [8–10,21], and DNA viruses might also suppress their transcribed CpG content to escape ZAP [22]. However, the extent to which antiviral defenses have influenced the composition of the host genome is unknown. Both ZAP and TRIM25 are evolutionarily conserved "core" ISGs [4]. We therefore hypothesised that ISGs, which are highly upregulated during the IFN response, have been selected to contain fewer CpGs, as they have functioned in the presence of abundant ZAP for over 300 million years [4]. Similarly, although most studies of the transcriptional response to IFN (the "interferome") focus on ISGs, the abundance of many host transcripts actually decreases following exposure to IFNs [1,2,4], and these genes are classified as interferon-repressed genes (IRGs). We thus additionally hypothesised that cellular mRNAs with abundant CpGs are targeted by ZAP during the IFN response. Thus, ZAP targeting of self-RNAs could be the mechanism through which some IRGs are downregulated.

Here, we show that across multiple species, ISG mRNAs are compositionally distinct, with the most important difference being that the CpG content of ISG transcripts is typically strongly suppressed, whereas IRGs tend to have a relatively high CpG content. We further reveal that IFN-stimulated ZAP expression appears to mediate the IFN-induced repression of multiple host transcripts. We propose that the antiviral state targets specific CpGs in mRNAs, and over millions of years, this has driven highly upregulated ISGs to possess very low CpG contents. Together, our data indicate that the impact of ZAP on compositional bias extends far beyond viruses to include significant impacts on the composition of the vertebrate interferome.

Results

The host transcriptional response to interferons has a CpG bias

We initially examined the CpG content of human ISGs and IRGs in relation to their induction and repression in response to IFN treatment. Using open-access RNA sequencing (RNA-seq) data we previously generated using IFN-treated primary human fibroblasts [4], we observed that the most highly upregulated ISGs tended to possess a highly suppressed CpG content (S1 Fig). We therefore selected the 50 most upregulated and 50 most downregulated genes for subsequent analysis as this compositional bias appeared most extreme within these groups (S1 Fig). Because CpG abundance is just one feature, we also examined the compositional

differences among the transcripts produced from these genes more generally (across 10 different species, alongside 50 randomly sampled, robustly expressed, but not differentially expressed [DE] genes from each species; S1 Table). For our gene lists, we again used a publicly available RNA-seq dataset that we previously generated using cells from these 10 species, which were stimulated with type I IFN under standardised conditions [4]. We trained 3 supervised machine learning classifiers to distinguish between (1) ISGs and IRGs; (2) ISGs and random genes; and (3) IRGs and random genes, based on 185 compositional features (encompassing dinucleotide, codon, and amino acid biases). Remarkably, across 50 replicates of 5-fold cross-validation, highly upregulated ISGs were reliably distinguished from the most repressed IRGs (>77% accuracy), and from random genes (with >74% accuracy), based upon their composition alone (Fig 1A and 1B; $N = 1{,}000$ in all cases, i.e., 50 genes from each of the 2 classes being distinguished, across 10 species). This is considerably better than the 50% accuracy expected by chance. Thus, even in the absence of important contextual features (such as the promoter region), responsivity to IFN was predictable based on the composition of the coding sequences (CDSs) alone. The majority of this signal seemed to originate from the ISGs, as IRGs were less reliably distinguished from random genes (accuracy >60%) (Fig 1C).

Strikingly, when the relative importance of each compositional feature was examined, the CpG content was the most definitive feature that distinguished ISGs from IRGs or from random genes across all 10 species used for the training (Fig 1D and 1E). Multiple other compositional differences were also utilised by the classifiers, suggesting that the selective pressures that sculpt the broader composition of ISGs may be distinct. Importantly, the distinctive composition of ISGs was not merely a consequence of the most upregulated ISGs using codons that are conducive to efficient protein expression. When the codon adaptation indices [25] of the most DE human ISGs and IRGs were compared, the ISGs tended to use similar or less optimal codons than the IRGs (Fig 1F). Thus, the compositional features of ISGs that enable efficient expression in the antiviral state appear distinct from those that simply promote efficient expression in unstimulated cells.

Because the CpG composition appeared to be the most defining feature of ISG transcripts, we analysed the dinucleotide composition of the most DE human ISGs and IRGs (S2 Fig). In accordance with the relative importance estimates (Fig 1D and 1E), CpG was the most variable dinucleotide, with IRGs containing approximately 50% more CpGs than ISGs (Fig 1G). As observed above, most of this signal appeared to be associated with ISGs. While CpG was among the most informative features in the classifier trained to distinguish IRGs from random genes, it was not the most important feature (S3 Fig). To examine the utility of the machine learning approach (Fig 1A–1E), using an interferome from a species not used to develop the model, we examined the ability of the model to correctly classify the 50 most DE mouse ISGs and IRGs [2,26]. Remarkably, the composition-based discriminatory power extended to a species not present during training, with 49 out of 50 mouse ISGs and 41 out of 50 mouse IRGs accurately classified using composition alone [2,26] (Fig 1H).

The relative importance of the CpG dinucleotide across all classifiers (Fig 1D and 1E, S3 Fig) led us to examine the CpG composition of the most DE ISGs and IRGs more closely. There are multiple methods for calculating CpG composition, and we have used a simple CpG frequency measurement (normalised to the length of the transcript) throughout. In addition, most experiments are also presented using CpG frequency normalised to the guanine–cytosine (GC) content of the transcript (equivalent to observed/expected), which is a standard method for calculating CpG composition.

When we examined the CpG composition of the human interferome, we observed that the top 50 DE ISGs possessed significantly fewer CpGs than the corresponding group of 50 IRGs (Fig 2A). Importantly, ISGs tended to have even fewer CpGs than the median of all human

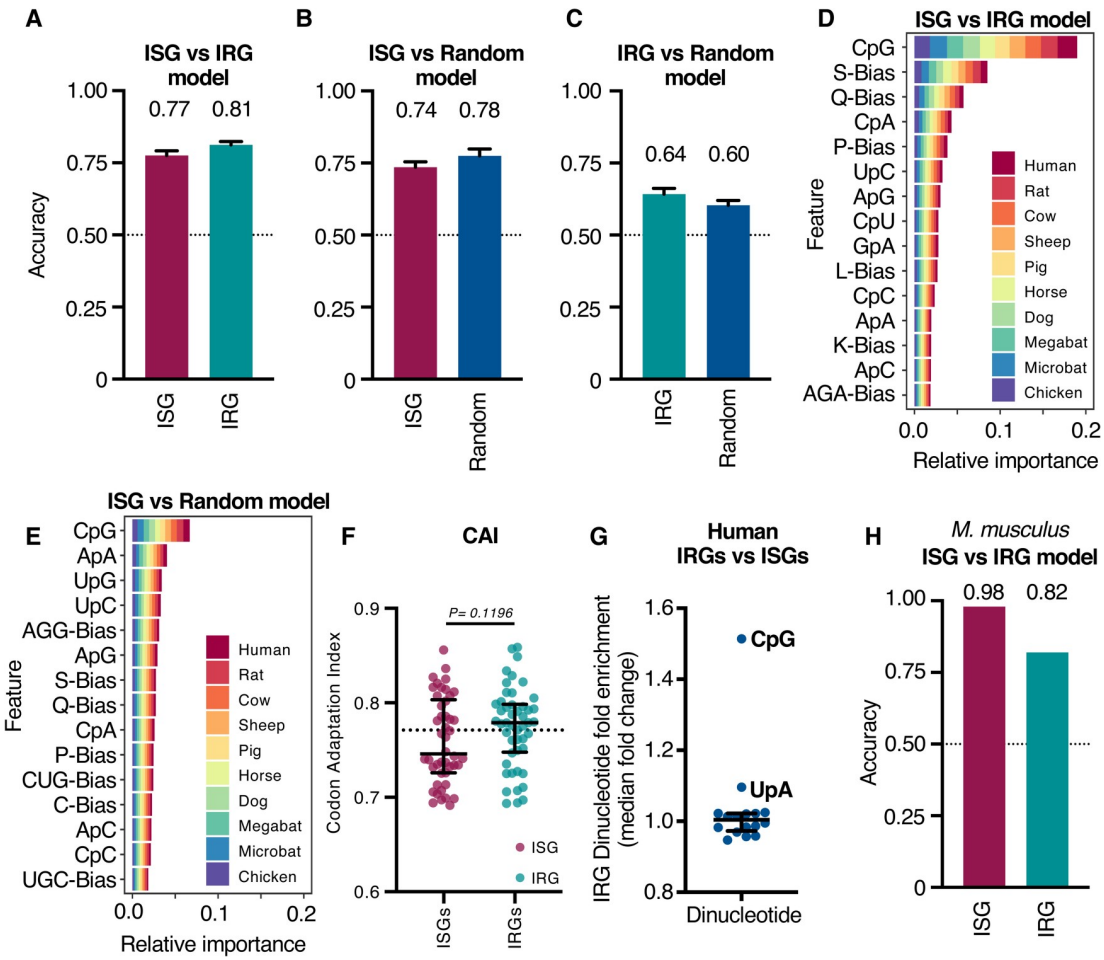

Fig 1. Compositional features predict expression class following IFN treatment. (A–C) Accuracy of classifiers trained to distinguish the top 50 most DE ISGs from the top 50 most DE IRGs (A), ISGs from 50 random genes (from each species) (B), or IRGs from 50 random genes (from each species) (C). Bars show the average proportion of genes in each class that were accurately identified across 50 replicates of 5-fold cross-validation, while error bars show the region containing 95% of observed accuracy values. Dashed lines indicate the expected performance of a null (i.e., uninformative) model. **(D, E)** The 15 most important features used by classifiers to distinguish ISGs from IRGs (D) or ISGs from random genes (E). The equivalent panel with the features used by classifiers to distinguish IRGs from random genes is shown in S3 Fig. Feature importance was quantified for individual genes using the SHAP approach [23,24], before summing their magnitude across all genes from a given species. All classifiers were trained and evaluated on 500 genes from each class, representing the top 50 ISGs or IRGs, or 50 random genes, from each species. Three-letter codes of the form "CpG" indicate measures of dinucleotide composition, single letters followed by the word "bias" (e.g., S-Bias) indicate amino acid composition biases, and 3 letters followed by the word "bias" (e.g., AGA-Bias) indicate codon usage biases. **(F)** The CAI of the 50 most DE ISGs and IRGs calculated as described previously [25]. The CAI is a measure of optimal codon usage; a higher CAI indicates a more optimal usage of codons. The horizontal dotted line represents the median for all transcripts in the human genome. Statistical significance was assessed using the Wilcoxon rank sum test with continuity correction. **(G)** The fold change in median dinucleotide composition between the top 50 most DE ISGs and IRGs in humans (summarising S2 Fig). The underlying RNA-seq data analysed in (A–G) were our previously published open-access data [4]. Briefly, primary fibroblasts derived from human, rat, cow, sheep, pig, horse, dog, little brown bat, and chicken, as well as immortalised large flying fox cells, were treated with type I IFNs (1,000 U/ml universal IFN, 200 ng/ml canine IFNα, 1,000 U/ml porcine IFNα, or 200 ng/ml chicken IFNα) for 4 hours before being analysed using RNA-seq [4]. **(H)** Accuracy of classifiers trained to distinguish the top 50 most DE ISGs from the top 50 most DE IRGs in a microarray dataset from a species not used to develop the model (murine NIH 3T3 cells +/− 100 units IFN [26], extracted from the interferome database [2]). The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). CAI, Codon Adaptation Index; DE, differentially expressed; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; RNA-seq, RNA sequencing; SHAP, SHapley Additive exPlanations.

https://doi.org/10.1371/journal.pbio.3001352.g001

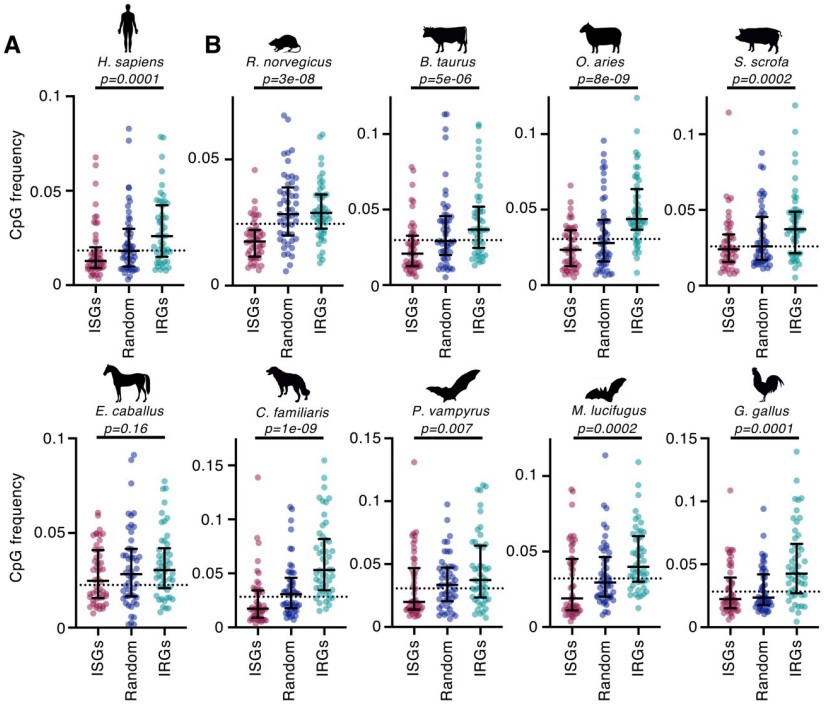

Fig 2. The vertebrate interferome has a CpG bias. (A) The length-normalised (cDNA) CpG frequency (see Materials and methods) of the top 50 most DE human ISGs and IRGs (ranked by mean Log_2FC) is shown. The dashed line represents the median CpG frequency of all transcripts in the relevant genome, a random sample of non-DE genes is included for reference, and whiskers represent the median and interquartile range for the analysed group. **(B)** The interferomes of the remaining 9 vertebrate species are plotted as in (A). The underlying RNA-seq data used were previously published open-access data [4] and were also described in the Fig 1 legend. Significance was determined using the Wilcoxon rank sum test with continuity correction. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; RNA-seq, RNA sequencing.

https://doi.org/10.1371/journal.pbio.3001352.g002

transcripts, whereas IRGs tended to be relatively enriched in CpGs (Fig 2A). The observed difference in the CpG content of ISGs and IRGs remained strongly evident regardless of whether we considered the frequency of CpGs (Fig 2A) or the frequency of CpGs normalised to overall GC content (S4 Fig).

We next examined the CpG content of the most DE ISGs and IRGs from 9 additional species (encompassing over 300 million years of evolution). Interestingly, this evolutionary divergence has resulted in considerable variation in the overall levels of CpG suppression in each species. The most extreme example was the large flying fox (*Pteropus vampyrus*), whose median CpG frequency was approximately 65% higher than in the human genome (Fig 2A and 2B). Despite this variation, in every species, the median CpG content of the most DE IRGs was noticeably higher than the median of all transcripts in that species (Fig 2B). Moreover, the CpG content of IRGs was also significantly higher than that of ISGs in 8 of the 9 additional species and was highly significant in multiple species, such as dogs and sheep (Fig 2B). Notably, the median CpG content of the most DE canine IRGs was 3-fold higher than that of the corresponding ISGs. The exception to the overall trend was the horse, in which the CpG composition of IRGs remained marginally higher than that of ISGs ($p = 0.16$; Fig 2B). This weaker effect may represent a biological difference in the equine interferome, but it might also reflect the methods used to annotate the horse genome (as misannotated sequences could result in substantial errors in the calculated CpG composition). Notably, the overall trends were very similar when the normalised measure of CpG composition was used (S4 Fig). Overall, the

observed dinucleotide bias in interferomes from divergent species suggests that this bias is an ancient property of vertebrate interferomes.

Type I interferons exhibit extreme CpG suppression

The observation that potently up-regulated ISGs tended to be strongly CpG suppressed led us to consider whether other genes playing pivotal roles in immunity might also share this property. We first calculated the dinucleotide frequency of every transcript in the human genome and confirmed the overall tendency for CpGs to be suppressed (S5 Fig). Then, in order to investigate the most CpG-suppressed genes, and determine if such genes shared any biological characteristics, we conducted gene ontology (GO) enrichment analysis [27] of the 1,000 human transcripts with the lowest CpG content (that were longer than 100 nucleotides in length). Remarkably, of all the overrepresented GO processes and functions, all but one function were involved in immune responses (Fig 3A). Importantly, nearly all the overrepresented terms were due to multiple type I IFN genes being among the 1,000 most CpG-suppressed genes, as observed previously [28]. GOrilla enrichment analysis of the same subset, in the absence of IFNα2 and IFNα14 (whose GO annotation and presence in the 1,000 most CpG suppressed likely accounted for these overrepresented terms), resulted in no significant overrepresentation of any immune processes (unpublished observations). Similar results were obtained when the normalised CpG frequency was considered (S5 Fig).

To explore whether extreme CpG suppression was a general property of immune genes, we considered the CpG content of 1,678 known immune genes (collated by ImmPort [29]), together with all significantly DE ISGs and IRGs we described previously [4]. The CpG content of the "immune genes," ISGs and IRGs, were very similar to when all transcripts in the genome were considered together (Fig 3B). Importantly, many genes are classified as significantly DE ISGs or IRGs, even though their transcript abundance varies only subtly following IFN stimulation (S1 Fig). These subtly modulated genes may not play a substantial role in IFN responses. In contrast, the CpG content of the 50 most DE ISGs and IRGs (who likely play a more important role in IFN responses) was noticeably different from the median of all human transcripts (Fig 3B). Moreover, in both cases, the CpG content was significantly different from that of all human transcripts (Fig 3B, S6 Fig). Importantly, the dinucleotide bias observed in the 50 most DE ISGs and IRGs is not observed in the whole interferome and is not present in most immune genes. Thus, the compositional bias has likely been selected for in genes whose expression is highly responsive to IFN stimulation.

In accordance with the GO analysis, the CpG content of type I IFN transcripts was remarkably low, and these genes were among the most CpG-suppressed transcripts in the entire human genome (Fig 3B). We therefore examined the CpG suppression in type I IFNs from a selection of the 10 species from Figs 1 and 2, whose type I IFN loci were well-defined [30]. Strong CpG suppression was also observed in mammalian type I IFN genes from cows, pigs, and large flying foxes (S7 Fig), despite the differential expansion of diverse IFN subtypes in these species. Again, a notable exception was the horse, where, similar to highly expressed ISGs, CpG suppression was not unusually strong (S7 Fig). In contrast, chicken type I IFNs were relatively CpG rich. These differences do not appear to be due to the different IFN subtypes, as IFNα is the predominant subtype in both humans and chickens, and IFNω is the most common subtype in cows and horses. Thus, CpG suppression in IFNs appears to be more evolutionarily heterogeneous than in highly expressed ISGs, where relatively high CpG suppression is uniformly observed.

We next considered the level of CpG conservation in the different gene classes where one-to-one orthologs existed in 9 primate species [31]. In addition to possessing very few CpGs, the CpGs in primate type I IFNs were also short-lived, with fewer than 1 CpG/kb present in the common ancestor being conserved in all 9 extant species (approximately 43 million years

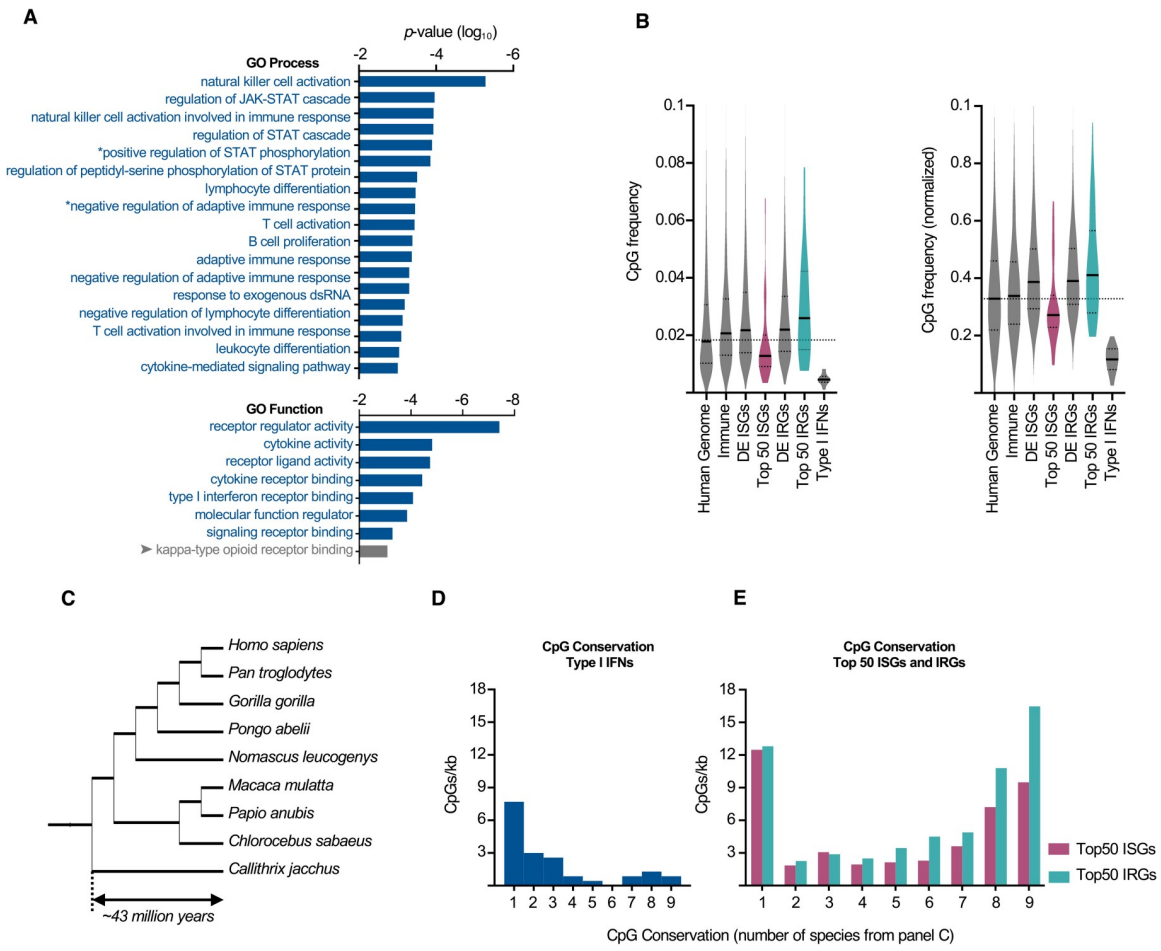

Fig 3. Type I IFNs exhibit extreme CpG suppression. (A) Significantly enriched GO processes and functions identified through GOrilla enrichment analysis of the 1,000 most CpG-suppressed cDNAs >100 bp (compared to all cDNAs >100 bp). **(B)** The CpG frequency in immune genes, ISGs, IRGs, and type I IFNs (for ISGs and IRGs, all significantly DE genes and the top 50 most DE genes are plotted). Matrices highlighting the significance (Kruskal–Wallis rank sum test) of potential comparisons in Fig 3B are displayed in S6 Fig. **(C)** A phylogenetic tree of the 9 primate species used to quantify CpG conservation in Figs 3D, 3E, and 5D. **(D, E)** Conservation is plotted as the number of CpGs per kb, binned by the number of species that possess that specific CpG, where 1:1 orthologs exist. Bin "1" represents CpGs present in only 1 of the species, whereas bin "9" represents CpGs conserved in all of the 9 species considered. CpG conservation is plotted for (D) type I IFNs (4 1:1 orthologs) and (E) top 50 ISGs (30 1:1 orthologs) and IRGs (34 1:1 orthologs). * Full GO process names: "positive regulation of peptidyl-serine phosphorylation of STAT protein" and "negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains." Arrow indicates a function enriched by genes other than the IFN genes IFNA2 and IFNA14. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; dsRNA, double-stranded RNA; GO, gene ontology; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene.

https://doi.org/10.1371/journal.pbio.3001352.g003

of divergence [32]) (Fig 3C and 3D). Similarly, CpGs appeared to be rapidly purged from ISGs, as IRGs contained approximately 70% more 43 million–year-old CpGs/kb than ISGs (Fig 3E). Thus, the most upregulated ISGs not only possess fewer CpGs (Fig 2), but these CpGs are also more rapidly lost (Fig 3E), consistent with stronger selection against CpGs in ISGs.

The ZAP network targets specific host mRNAs and mediates interferon-induced repression of gene expression

The reduced abundance and rapid evolutionary turnover of CpGs in strongly induced ISGs led us to consider whether ZAP (a CpG-targeting effector) might influence the transcript

abundance of IFN-regulated genes. To investigate this possibility, we depleted ZAP from human cells using CRISPR/Cas-9. We opted to use A549 cells for these analyses as previous work has shown that the replication of an influenza A virus engineered to possess an elevated CpG content is severely restricted in these cells [9]. We observed that ZAP protein expression was efficiently depleted when "bulk" A549 populations, transduced with vectors encoding ZAP-targeting guide RNAs, were compared to Cas-9 expressing transduced controls without a ZAP guide (Fig 4A). To examine the effect ZAP depletion had on the A549 transcriptome, we carried out RNA-seq analysis of these "bulk" populations. We opted to limit our analysis to bulk populations in this instance as clonal variants (from knockout [KO] clones) often have distinct transcriptomic signatures that could obscure the signal from ZAP depletion. Comparison of the transcriptomes of ZAP KO and "No guide" controls revealed a number of significantly DE genes. Notably, the CpG content of transcripts whose expression increased following ZAP depletion tended to be higher than the median CpG content of all transcripts in the genome (Fig 4B). This observation is consistent with the notion that ZAP-mediated RNA surveillance might increase the turnover of CpG-rich mRNAs in resting cells. Interestingly, a large number of genes were downregulated following ZAP depletion (as has been reported previously [33]). We speculate that multiple genes involved in transcriptional regulation (possibly upregulated following ZAP depletion) might indirectly downmodulate these transcripts.

To investigate whether ZAP targeting of host mRNAs might be involved in the dinucleotide bias observed in ISGs and IRGs, we generated 5 clonal A549 complete ZAP KO cell lines (Fig 4C) and determined their transcriptional response to IFN alongside 5 clones of transduced "No guide" clonal control cell lines generated in parallel (Fig 4C). In addition, we also defined the interferome of our bulk KO cells. We then analysed the CpG content of the significantly DE genes. ISGs tended to possess fewer CpGs than IRGs, regardless of the presence or absence of ZAP (Fig 4D and 4E). Similarly, the composition of IRGs in IFN-stimulated cells with and without ZAP was remarkably similar in terms of overall CpG content (Fig 4D and 4E). Moreover, the ISG versus IRG classifier did not appear to identify ZAP-independent IRGs any less reliably than the top 50 IRGs (Fig 4F). Thus, our classifier does not specifically detect the CpGs targeted by ZAP. Strikingly, however, fewer genes were repressed following IFN treatment when ZAP was depleted. The total number of genes classified as IRGs was approximately 40% lower in bulk ZAP KO cells and was approximately 75% lower in ZAP KO clones, suggesting that ZAP is involved in reducing the abundance of a sizable subset of IRG transcripts (Fig 4G). Conversely, ZAP KO cells had a similar number of ISGs to control cells (Fig 4H). Importantly, the continued presence of IRGs in the absence of ZAP expression reveals that not all IRGs are dependent upon ZAP.

Interestingly, the identity of the IRGs in bulk and ZAP KO clones was far more variable than that of ISGs in the equivalent cells (S8 Fig). This likely reflects the fact that ISGs are very strongly upregulated (sometimes >1,000-fold), whereas IRGs are seldom repressed more than 2-fold, both here (S1 Fig) and in previous studies [1,2,4]. Moreover, downmodulation of a transcript is dependent on sufficient expression of that transcript prior to IFN stimulation, making the IRG class more dependent on cell state, lineage, and culture conditions than ISGs. Thus, to identify putative ZAP targets, we focused on the IRGs that were most consistently downregulated across our datasets. First, we filtered our A549 transcriptomic data to include only DE genes also observed in primary human cells [4] (S8 Fig). Among these genes, we selected those that were also IRGs, in both A549 "bulk" and clonal control cells (present in the overlapping populations on the left-hand side of panel B, S8 Fig). This approach identified 15 genes that were significantly downregulated in all 3 RNA-seq experiments (Fig 5A). Remarkably, when ZAP was knocked out, two-thirds of these consistent IRGs were no longer significantly downregulated by IFN in any ZAP KO condition (either "bulks" or individual "clones")

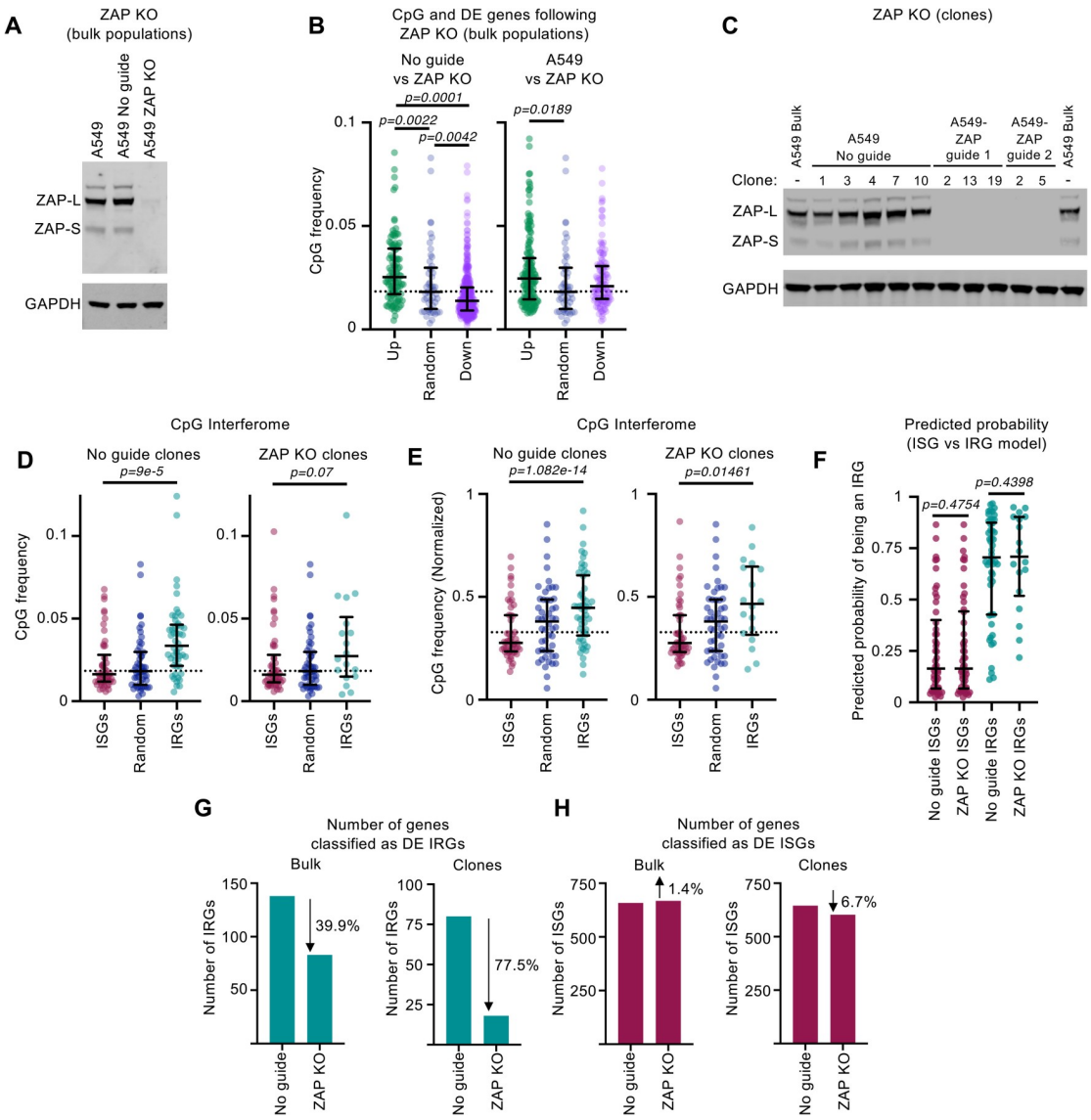

Fig 4. ZAP mediates the IFN-induced repression of a subset of IRGs. (A) WB analysis of ZAP and GAPDH expression in human "bulk populations" of A549 cells transduced with ZAP-targeting CRISPR sgRNAs or transduced Cas-9 expressing "No guide" controls. ZAP-L (PARP13.1) and ZAP-S (PARP13.2) bands are indicated [34]. **(B)** The CpG content of significantly DE genes (identified using RNA-seq in the absence of IFN using edgeR and an FDR <0.05), comparing triplicate "bulk" KO cells and transduced controls or unmodified cells. A random selection of 50 genes is included for comparison. Bars represent the median values, and whiskers represent the interquartile ranges. Horizontal dotted lines represent the median of all transcripts in the human genome. Significance was determined using a Kruskal–Wallis rank sum test (only significant comparisons are shown). **(C)** WB analysis (as in A) of clonal lines modified with ZAP-targeting CRISPR sgRNAs or parallel clonal lines derived from Cas-9 expressing transduced "No guide" controls. **(D, E)** The CpG content (D) or the normalised CpG content (E) of the 50 most significantly DE ISGs and IRGs (ranked by mean Log2FC), determined using RNA-seq of the 10 clones in (C) stimulated with 1,000 units/ml of IFNβ (4 hours) are shown, alongside 50 random genes. Where fewer than 50 significantly DE genes were detected, all significantly DE genes are plotted. Significance was assessed using the Wilcoxon rank sum test with continuity correction. **(F)** The predicted probability that the IRGs classified in the presence of ZAP or the absence of ZAP, from (D and E), are IRGs based on their nucleotide composition. Significance was assessed using the Kruskal–Wallis rank sum test (only insignificant comparisons are shown). **(G, H)** Numbers of significantly DE IRGs (G) or ISGs (H) from the RNA-seq of cells in either (A) ("bulk") or (C) ("clones") stimulated with 1,000 units/ml of IFNβ (4 hours). The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; FDR, false discovery rate; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; KO, knockout; RNA-seq, RNA sequencing; sgRNA, single-guide RNA; WB, western blot; ZAP, zinc-finger antiviral protein.

https://doi.org/10.1371/journal.pbio.3001352.g004

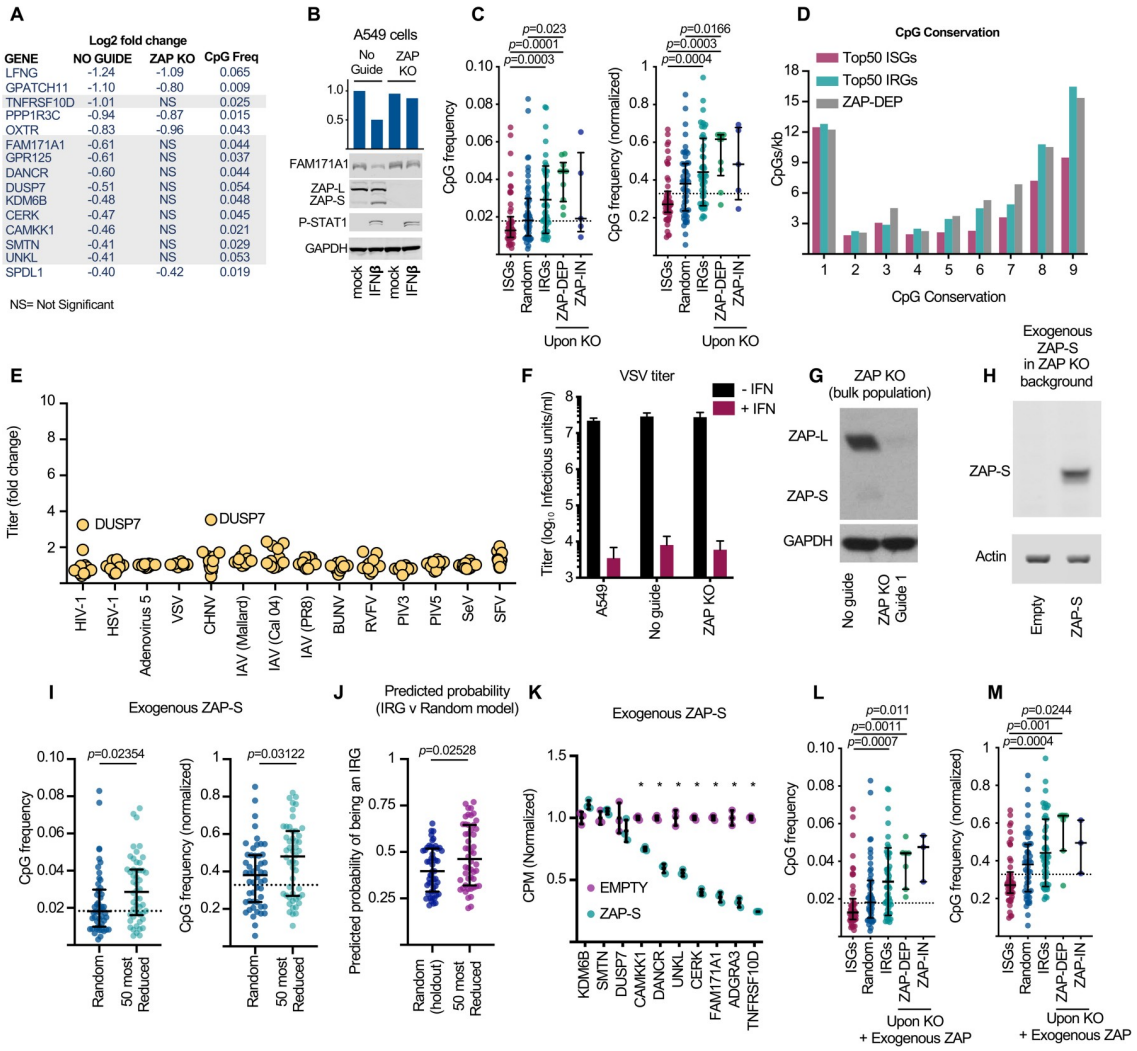

Fig 5. Identification of host transcripts consistently targeted by IFN-stimulated ZAP. (A) The identity and differential expression of the 15 IRGs identified in Shaw et al. [4], as well as in bulk and clone NO GUIDE controls (derived from the overlapping populations, left-hand side of panel B in S8 Fig). Shaded genes represent IRGs whose repression is not observed following ZAP KO. **(B)** WB of FAM171A1, ZAP, P-STAT1, and GAPDH in A549 NO GUIDE or ZAP KO clones stimulated with 1,000 units/ml of IFNβ (24 hours). **(C)** The CpG contents of 10 ZAP-dependent (ZAP-DEP) IRGs and 5 ZAP-independent (ZAP-IN) IRGs (identified in A) are shown alongside the most DE ISGs and IRGs from Fig 2A. Significance was assessed using the Kruskal–Wallis rank sum test; significant differences are shown (all comparisons listed in S9 Fig). **(D)** CpG conservation among 9 primate species (as in Fig 3C–3E) is plotted as the number of CpGs per kb, binned by the number of species that possess that specific CpG, where 1:1 orthologs exist (ZAP-DEP $n = 5$). "1" represents CpGs present in only 1 of the species, whereas "9" represents CpGs conserved in all of the 9 species considered (top 50 ISGs includes 30 1:1 orthologs and top 50 IRGs includes 34 1:1 orthologs). **(E)** The 10 putative ZAP targets (identified in A) encoded by lentiviral vectors were used to transduce human cells prior to serially diluted challenge with a GFP-expressing virus. The resulting titres were calculated using flow cytometry and normalised to the empty vector control. These data were plotted as the fold change in titre (y-axis) relative to the empty vector control cells (yellow circles represent each cDNA). **(F)** The impact of ZAP KO (see Materials and methods) on the replication of *Indiana vesiculovirus* in the presence and absence of IFNβ. **(G)** WB for ZAP in bulk A549 cells KO'd for ZAP (ZAP KO guide 1) compared to mock treated cells (No guide). **(H)** WB of A549 ZAP KO cells (guide 1, clone 2; Fig 4C) transduced with an empty lentiviral vector or a lentiviral vector encoding CRISPR-resistant ZAP-S. **(I)** The transcriptomes of the transduced cells in (H) were defined using RNA-seq and the CpG compositions of the 50 most downregulated transcripts are shown. Significance was determined using the Wilcoxon rank sum test with continuity correction. **(J)** The predicted probability that the most reduced transcripts in (I) are IRGs. A sample of 50 random genes (not used for training) are included as a comparator. Significance was determined as in (I). **(K)** The transcript abundance of the 10 ZAP targets identified in (A) in the cells from (H) and (I) is shown. **(L)** The CpG frequency and **(M)** normalised CpG frequency of the 7ZAP-DEP IRGs and 5 ZAP-independent (ZAP-IN) IRGs (identified in A and J) are shown alongside the most DE ISGs and IRGs from Fig 2A. Significance was assessed as in (C). The dashed line in (C), (I), (L) and (M) represents the median CpG frequency of all transcripts in the genome, a random sample of non-DE genes is included for reference, and whiskers represent the median and interquartile range for the analysed group. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). BUNV,

Bunyamwera orthobunyavirus; CHNV, *Chandipura vesiculovirus*; DE, differentially expressed; HIV-1, *Human immunodeficiency virus 1*; HSV-1, *Human alphaherpesvirus 1* (formerly known as herpes simplex virus 1); IAV, Influenza A virus; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; KO, knockout; P-STAT1, phosphorylated STAT1; PIV-3, *Human respirovirus 3* (formerly parainfluenza virus 3); PIV-5, *Mammalian orthorubulavirus 5* (formerly parainfluenza virus 5 or simian virus 5); RNA-seq, RNA sequencing; RVFV, *Rift Valley fever phlebovirus*; SeV, *Murine respirovirus* (formerly Sendai virus); SFV, *Semliki Forest virus*; VSV, *Indiana vesiculovirus* (formerly vesicular stomatitis virus); WB, western blot; ZAP, zinc-finger antiviral protein; ZAP-DEP, ZAP-dependent.

https://doi.org/10.1371/journal.pbio.3001352.g005

(Fig 5A). The functions of these IRGs are summarised in S2 Table. This observed change in transcript abundance might also be reflected at the protein level as one readily detectable antigen appeared similarly repressed by IFN in a ZAP-dependent fashion (Fig 5B). These observations suggest that ZAP plays a role in the downregulation of a subset of IRGs. Notably, the only previously well-defined cellular target of ZAP, TNFRSF10D [33], is a known IRG [2,4] (S9 Fig) and was among the genes that we identified using this approach (Fig 5A).

When the CpG content of the 10 putative ZAP-dependent targets ("ZAP-DEP," identified in Fig 5A) was examined, all were relatively CpG rich, with all 10 transcripts exhibiting a CpG frequency greater than the median of all transcripts (Fig 5C, S9 Fig). Moreover, the putative ZAP targets also possessed a substantial number of evolutionarily conserved CpGs (Fig 5D). We hypothesised that downregulation of the putative ZAP targets contributes to the antiviral state. We therefore examined their ability to promote virus replication. Each IRG was exogenously expressed, and their ability to promote the replication of a panel of 14 viruses was considered (Fig 5E). Although some relatively weak enhancement of virus replication was observed in specific instances (Fig 5E), IFN still potently inhibited *Indiana vesiculovirus* (formerly known as vesicular stomatitis virus [VSV]) infection when ZAP was depleted (Fig 5F and 5G). Thus, the ZAP-mediated downregulation of host transcripts does not appear critical for a functional "anti-VSV state." However, the importance of an IRG to virus replication may not be revealed through exogenous expression, and many more viruses would need to be considered in the absence of IRG expression (i.e., IRG KO) before a key role for IRGs in the "antiviral state" can be excluded.

To validate the putative ZAP targets, we examined the ability of exogenous ZAP expression to decrease the abundance of these specific transcripts. We used the ZAP-S isoform, as this isoform is preferentially upregulated by type I IFNs (Fig 5B) [35]. The shorter ZAP-S mRNA utilises the intron between exons 9 and 10 as a 3′ UTR and does not encode the PARP domain [34,36]. We used a lentiviral vector to express a CRISPR-resistant ZAP-S in A549 ZAP KO guide 1 clone 2 cells (Figs 4C and 5H). We then used RNA-seq to compare the transcriptome of cells expressing exogenous ZAP to control cells transduced with an empty vector. When ZAP-S was expressed (Fig 5H), similar to IFN treatment, the mRNAs that were reduced in abundance were relatively CpG rich (Fig 5I). Moreover, the majority of these downmodulated transcripts have been identified as IRGs in multiple transcriptomic studies (S9 Fig). Accordingly, these putative ZAP targets were predicted to have an increased probability of being IRGs (Fig 5J) using the relevant model from Fig 1. Crucially, as well as reducing the expression of a large number of relatively CpG-rich transcripts, the majority (7/10) of the putative ZAP targets identified in Fig 5A were also significantly downregulated by exogenous ZAP-S (Fig 5K). Although modest in some instances, the observed magnitude of downregulation was similar to the level of repression following IFN treatment (Fig 5A and 5B, S9 Fig). The CpG composition of the 3 IRG transcripts whose abundance did not decrease in the presence of exogenous ZAP-S was similar to the 7 genes whose transcript abundance was reduced by exogenous ZAP (Fig 5L and 5M), suggesting that the overall CpG composition is not the sole determinant of sensitivity to ZAP-S.

When the loss of IRG repression (following ZAP KO) and the reduced IRG transcript abundance (following exogenous ZAP-S expression) are considered together, these data strongly suggest that ZAP is mechanistically responsible for the repression and/or regulation of a subset of IRGs.

Discussion

The hypothesis that the recognition of nonself influences the composition of innate immune genes was proposed in 2009 [28]. Following the identification of ZAP as a CpG-targeting antiviral effector by Takata and colleagues [10], Stephen Goff commented that "the most far-reaching suggestion arising from this study [Takata et al.] is that ZAP constitutes a factor in the evolution of low CG content in the DNA of host cells" [21]. Given that CpG-targeting antiviral surveillance/effector mechanisms are likely to be at their most active during the IFN-induced antiviral state, we reasoned that the most extreme evolutionary pressures (that could select for biased nucleotide composition) would be exerted on IFN-regulated transcripts. Consistent with these hypotheses, a supervised machine learning approach identified that strongly upregulated ISGs tended to be notably more CpG suppressed than the average gene. Moreover, we demonstrate that IRGs tend to be relatively rich in CpGs and that IFN-stimulated CpG-targeting antiviral defenses downmodulate many CpG-rich host transcripts, leading to their classification as IRGs. Specifically, ZAP KO led to a considerable reduction in the total number of IRGs, suggesting a direct involvement in the suppression of a sizable subset of IRGs. Considering the significant CpG bias present in the interferome, together with the evidence that CpG-targeting effectors modulate gene expression, it seems plausible that IFN-stimulated antiviral effectors have selected for bias in the composition of the interferome. Thus, it is likely that antiviral effectors have influenced the genome composition of both virus and host. Interestingly, CpG composition was not the only discriminatory feature identified using the machine learning approach, and more work is required to understand the selective pressures that may have selected for other compositional biases in the interferome. Notably, the UpA dinucleotide (which is also underrepresented in vertebrate genomes and in viruses with vertebrate hosts) was surprisingly unimportant for distinguishing ISGs from the other gene classes, suggesting that the underrepresentation of UpAs in viral genomes might not be due to the IFN-induced "antiviral state." Alternatively, the resistance of ISGs to degradation via antiviral pathways, such as the OAS-RNaseL system [37,38], may be achieved by a mechanism that does not require the depletion of UpA and UpU dinucleotides.

Strong selection for high IFN stimulation (or maintenance of efficient IFN expression in the face of autocrine stimulation) could plausibly select against the presence of targeted CpGs. Over evolutionary timescales, the stepwise elimination of targeted CpGs would leave these genes in a highly CpG-suppressed space. Accordingly, we observed lower levels of CpG conservation in IFN genes and ISGs than in IRGs. Moreover, the biased CpG content is an ancient property of the interferome that we observed in multiple vertebrates, some of whom have been separated by over 300 million years of evolution. Although limiting the analysis to the 50 most DE genes was arbitrary, the compositional bias was only easily observed at the extremes of the interferome. We believe that the visualisation of the same trend in multiple species supports the relevance of this approach. Importantly, this trend was not due to repeated sampling of the same genes, as the bias remained constant even though more than 600 different orthologous clusters are represented in Figs 1 and 2.

As a word of caution, it is difficult to use contemporary observations to ascribe causation to the selection of genomic signatures that have accumulated over deep timescales. We are unable to rule out that other processes, such as reduced CpG methylation within the genomic loci of

highly expressed ISGs (which could be more conducive to rapid induction of gene expression), have selected for low CpG composition in ISGs. Thus, although self-targeting represents a plausible and parsimonious explanation that links our observed transcriptomic changes to the observed compositional biases, this link is currently only correlative.

Based on our RNA-seq data, ZAP targeting of host mRNAs appears to be mechanistically responsible for the IFN-induced repression of multiple IRGs. Whether ZAP-mediated mRNA degradation represents "collateral damage" (i.e., a tolerated consequence of the antiviral state), or whether specific transcripts have been selected to be regulated by ZAP (i.e., ZAP-targeted repression of specific transcripts enhances immune responses), is not yet fully resolved. It is worth noting some key differences between these gene classes. Many ISGs are expressed at low or undetectable levels prior to IFN induction. This means that following IFN stimulation, an increase in transcript abundance can correspond to remarkable fold changes (sometimes >1,000-fold). This "zero-to-hero" lifestyle provides added consistency/robustness to RNA-seq studies of ISGs as this class is relatively indifferent to the cell type or cellular state prior to IFN stimulation. In contrast, IRGs, by definition, must be robustly expressed prior to IFN stimulation, and IRGs are seldom repressed more than 2-fold. This technical difference contributes to making IRGs a more variable class of genes. Moreover, many individual ISGs are conserved over large phylogenetic distances and are consistently upregulated in divergent cell lineages [2,4,39], while this is usually more variable for IRGs [2,4]. Crucially, although the identity of the individual transcripts that are repressed is variable, the compositional characteristics of IRGs appear relatively constant. Thus, we believe collateral damage (as opposed to regulation) underlies the downregulation of most IRGs. A collateral damage hypothesis is also consistent with the reduced ability of our machine learning classifier to robustly distinguish IRGs from random genes. In contrast, the consistent downregulation of a small number of IRGs implies a regulatory role. We thus speculate that ZAP-mediated regulation of specific IRGs occurs within the broader context of "off-target" downregulation of many relatively CpG rich "bystander" mRNAs (not necessarily related to the antiviral state).

While the biological consequences of the regulation imposed by ZAP targeting are largely uncharacterised, it is logical to predict that they will be aligned with the previously defined functions of the interferome. For example, a proapoptotic programme is an ancient function of the interferome [4,40]. Thus, ZAP-mediated downregulation of TNFRSF10D and CAMKK1, both of which have been ascribed antiapoptotic functions [33,41,42], would likely potentiate the proapoptotic effect of IFNs. Moreover, the ability to directly interfere with viral replication is a key function of the interferome. As both CAMKK1 and KDM6B are necessary for efficient herpesvirus replication or reactivation [43,44], the reduced levels of these gene products could potentially inhibit viral replication. In support of this notion, ablation of CAMKK1 expression has recently been shown to inhibit cytomegalovirus replication in vitro [45]. In contrast, the observed ability of IFN to inhibit *Indiana vesiculovirus* (VSV) infection in the absence of ZAP indicates that ZAP-targeted IRGs are not essential for a functional antiviral state. Importantly, redundancy in the antiviral state means that the removal of a single antiviral pathway does not always result in a rescue of infection or replication. Notably, VSV (like many viruses) is sensitive to multiple antiviral ISGs [46–48]. Thus, much more work will be required to determine the contribution (if any) that IRGs might make to the inhibition of specific viruses.

Although the most DE IRGs tended to be relatively CpG rich, we observed no absolute relationship between the CpG content of a transcript and whether ZAP was required for its repression. Clearly, not all CpGs are targeted equally by ZAP, and it is likely that RNA structure/context is crucial for CpG-targeted degradation [17,19,21,49]. Thus, the tendency for ZAP to

target relatively CpG-rich IRGs likely represents the increased probability that these IRGs have the "right kind" of CpGs.

Interestingly, ZAP-independent IRGs also tended to be enriched in CpGs. Moreover, using a supervised machine learning approach, ZAP-dependent and ZAP-independent IRGs were compositionally indistinguishable. Thus, we are currently unable to predict which CpGs in host transcripts might be recognised as pathogen-associated molecular patterns (PAMPs). Given the similarity between ZAP-dependent and ZAP-independent IRGs, it is tempting to speculate that other IFN-stimulated CpG-targeting antiviral effectors might also exist, or that the ZAP-mediated regulation of IRGs is redundant. Importantly, other CpG-targeting effectors could also select for compositional biases in host transcripts, and the relative contribution of ZAP may only become apparent once the context of ZAP-mediated CpG recognition is fully mechanistically understood. Therefore, examining whether all synthetically deoptimised, CpG-rich viruses are rescued by ZAP depletion is likely to be an important avenue of future research.

Whether ZAP influences the expression of IFN genes directly has not been tested here. Intriguingly, it has been proposed that ZAP-S might regulate the expression of human IFN genes [36]. This study focused on the relatively CpG rich type III IFN genes, although the authors also proposed that type I IFNs could be regulated by ZAP [36]. Thus, it is possible that CpG-targeting effectors might also have influenced the CpG composition of IFN genes, although this does not appear to be the case in all the species examined. Although extreme CpG suppression is observed in the type I IFN genes of humans and some other mammals, this trend appears to be more evolutionarily variable than the bias in the composition of the interferome. One potential explanation for this could be that IFNs transiently trigger the antiviral state, and, therefore, might not need to be expressed for long periods in the face of autocrine signalling (and may even have been selected as targets in some species). Conversely, ISG transcripts must persist in the IFN-stimulated cell to potentiate the antiviral state. More studies will be needed to detail how/whether all IFN transcripts escape inhibition by the antiviral state and how these processes might vary between species. Similarly, although the CpG bias in the interferome was observed in divergent animal species, the horse was a notable exception to this trend. Both equine ISGs and type I IFNs were not strongly CpG suppressed and exhibited levels of CpG suppression that were very similar to the equine transcriptome as a whole. Notably, we could not detect any obvious features at the horse ZC3HAV1 locus that suggested equine ZAP might be inactive in horses. However, we have equally been unable to definitively support or refute the existence of ZAP-S in horses (although this was the case for multiple species under consideration in this study). Thus, further studies are required to investigate whether CpG-targeting effectors are active in equine cells. Conversely, we currently have no explanation for the more pronounced CpG bias observed in the ovine and canine interferomes, and more work is required to investigate whether CpG-based pattern recognition is particularly effective in these species.

Viral RNAs are simultaneously under surveillance from a multitude of host defenses and sensors. Because these highly upregulated ISGs must be efficiently expressed in the face of all these defenses, it is likely that ISGs and viruses will share multiple compositional characteristics (to facilitate efficient gene expression during an active antiviral state). The influence of self versus nonself discrimination on the composition of host genomes may therefore be an underappreciated evolutionary process. Once a nucleic acid–based pattern recognition system emerges, the removal of said patterns will be selected for in both host and pathogen. Thus, the observed underrepresentation of features exploited for virus recognition (in contemporary host genomes) may have become exaggerated over evolutionary timescales and become particularly exaggerated in ISG transcripts. Identifying the shared features of viral RNAs and highly

upregulated ISGs could, in the future, reveal novel mechanisms used by the host to sense and destroy invading pathogens, help identify viral reservoirs, and improve vaccine strategies.

Materials and methods

Determining dinucleotide frequencies

A database of CpG dinucleotide frequencies was prepared for 10 species for which ISG/IRG data were available [4] (http://isg.data.cvr.ac.uk). Briefly, for all species, all cDNAs and CDSs were downloaded from Ensembl via their FTP website (release 92: ftp://ftp.ensembl.org/pub/release-92/fasta), except the 100 mouse genes, which were analysed at a later date using Ensembl release 99. For each sequence, both the relative frequency of CpGs and the observed/expected ratio (normalised CpG frequency) were calculated using the following formulas. The CpG dinucleotide frequency (CpG frequency) was counted for each transcript, giving a raw count that was then normalised by cDNA length. Thus, CpG frequency = #CG dinucleotides/(length-1). A second value normalising for GC richness was also calculated, where CpG frequency (normalised) = CpG frequency/ (proportion C * proportion G), equivalent to observed/expected or odds ratio. Where observed CpG and expected CpG are plotted, the observed = #CG dinucleotides, while the expected CpG = Prop(C)*Prop(G)*(Length of transcript-1), where Prop(C) and Prop(G) are the proportions of C and G nucleotides in the transcript of that specific gene. These databases were then matched against ISGs/IRGs according to Ensembl gene ID. The analysis of the CpG composition of human cDNAs was also conducted using cDNA sequences from Ensembl (ftp://ftp.ensembl.org/pub/release-92/fasta/homo_sapiens/). The location of each transcript was mapped, and redundant transcripts for each gene were removed, retaining the longest transcript variant of each gene. In multiple panels, we compared the composition of the 50 most DE ISGs and IRGs. We opted to compare the top 50 most DE ISGs and IRGs based upon S1 Fig. Due to the non-normal distribution of CpG frequencies, the nonparametric Wilcoxon rank sum test was used for comparing independent groups of genes. Similarly, where multiple groups were compared, we used a Kruskal–Wallis rank sum test followed by post hoc analysis using the Dunn test including Benjamini–Hochberg correction for multiple comparisons. The frequency of each dinucleotide was calculated for the top 50 human ISGs and IRGs, and the normalised dinucleotide frequency was used (equivalent to the observed/expected or odds ratio). Each set of 50 random genes was selected to be representative of genes unaffected by IFN treatment [4]. First, genes DE at false discovery rate (FDR) <0.05 were subtracted from the remainder of the genome. Non-DE genes were further subsetted to leave only those for which there was evidence of transcription in every sample [4]. Finally, 50 random genes were selected using the sample function in R.

Feature-based machine learning

To gauge the relative importance of various compositional features, 3 classifiers were trained to distinguish (1) the top 50 ISGs from the top 50 IRGs of each species; (2) the top 50 ISGs from 50 randomly selected genes from each species; and (3) the top 50 IRGs from the same set of 50 randomly selected genes (ISGs and IRGs were ranked by mean Log_2FC (fold change), whereas random genes (see above) were randomly selected from the remaining transcribed genes that were not DE in response to IFN). For one spliced/mature transcript representing each gene, 35 features were calculated across the entire sequence describing GC content, AT content, dinucleotide biases ($N = 16$; calculated as described previously [50]), and the proportion of bases dedicated to each dinucleotide (dinucleotide proportion, $N = 16$). An additional 133 features were calculated to specifically describe the open reading frames of all protein

CDSs, including open reading frame–specific versions of GC content, AT content, dinucleotide biases ($N = 16$), and dinucleotide proportion ($N = 16$), along with codon usage biases ($N = 64$), amino acid biases ($N = 21$), and position-specific dinucleotide biases at either codon bridge ($N = 16$) or non-bridge positions ($N = 16$) [50].

These features were used to train 3 independent boosted regression tree classifiers using the DART algorithm of the XGBoost library, combining genes from all species into a single dataset [51,52]. A search for the optimal combination of tuning parameters was performed using the caret library in R, searching across 100 randomly selected parameter combinations [53,54]. Models were evaluated by randomly splitting the data into 5 equally sized bins and using each bin, in turn, as a test dataset to score the prediction accuracy of a model trained on the remaining 4 bins (i.e., 5-fold cross-validation). This procedure was repeated 50 times to assess the variation in predictive accuracy before training the final model using all available data.

The relative importance of the different genome features in the classification of individual genes was assessed in this final model by calculating approximate Shapley values using the TreeSHAP algorithm [23,24]. To obtain a single value for each broad feature class, often consisting of highly correlated features, Shapley values were summed across all features making up the broader feature class. For example, the total importance of ApU dinucleotides was calculated as the sum of mean absolute Shapley values for ApU bias across the entire gene, ApU proportion across the entire gene, ApU bias in open reading frames, ApU proportion in open reading frames, ApU bias at codon bridge positions, and ApU bias at non-bridge positions. Finally, the importance of each feature class was calculated as the mean of absolute Shapley values across all genes in the data. To allow comparison across independent classifiers, values were rescaled to lie between 0 and 1.

GOrilla analysis and immune gene selection

Enrichment of GO terms was assessed using GOrilla [27]. Lists of Ensembl gene IDs, ranked according to Log_2FC or CpG content, as appropriate, were used as inputs. Transcripts shorter than 100 nucleotides in length were excluded from these analyses. Default parameters were then used to generate lists of enriched terms under "Process" and "function." Genes were classified as either immune genes, ISGs/IRGs, or IFNs using the ImmPort database [29], previously published transcriptomics analyses [4] or text mining gene names containing "IFN," respectively.

Cells and gene editing

Human A549 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 9% fetal calf serum (FCS) supplemented with gentamicin. Gene editing was achieved using the lentiGuide-Puro system in accordance with the Zhang Lab protocols [55]. Oligos specifying the following ZAP target sequences were used: 5′- ATGTGGAGTCTTGAACACGG-3′ (1), 5′-GCAACTATTCGCAGTCCGAG-3′ (2), and 5′-ACTCTCTGGACTGAACAAAG-3′ (5) to generate single-guide RNA (sgRNA) encoding vectors. A549 cells were either transduced with vectors encoding Cas-9 and the relevant ZAP-targeting sgRNAs or transduced in parallel with vectors encoding Cas-9 and no sgRNA (no guide). To maximise the frequency of gene disruption, the bulk KO cells examined in Fig 3 were transduced with a mixture of both sgRNA 1 and sgRNA 5 vectors. Single cell clones were generated using limiting dilution.

RNA-seq

For the majority of experiments, we selected a sample size of 3 independent replicates (3 separate cultures per condition) as an optimal balance between experimental power and the

resources available. The exception to this was the transcriptomic analysis of ZAP KO clones, where 5 control single cell clones were compared to 5 ZAP KO single cell clones, with each clone treated as a replicate. This experiment was initially designed to compare 6 clones v 6 clones. However, following principle component analysis, one ZAP KO clone was an extreme outlier (guide 2 clone 3). Thus, the guide 2 clone 3 and a randomly selected control clone (No guide clone 11) were discarded from all subsequent analyses.

Cells were treated or untreated with 1,000 units/ml IFNβ (PBL Assay Science, New Jersey, USA) for 4 hours prior to RNA extraction (RNeasy, Qiagen, Hilden Germany). Poly(A)-enriched Illumina RNA-seq libraries were prepared and run. Briefly, RNA concentration and integrity were determined using a Qubit Fluorimeter (Life Technologies, California, USA) and 4200 TapeStation (Agilent, California, USA). All samples had an RNA integrity number of 8 or above. Moreover, 500 ng of total RNA was used to prepare libraries for sequencing, using an Illumina TruSeq Stranded mRNA HT kit (Illumina, California, USA), according to the manufacturer's instructions. Libraries were pooled in equimolar concentrations and sequenced using an Illumina NextSeq 500 sequencer (high output cartridge), generating single end reads with a length of 75 bp. The raw FASTQ read output was first assessed using FASTQC. Sequencing adapter sequences were removed from reads using TrimGalore prior to mapping against the human genome (GRCh38.p12) using HISAT2. Mapped reads aligning to genes annotated in a .gtf file of the human genome were counted using HTseq-Count. Reads mapping at a frequency of <1 in greater than one half of the samples were removed prior to differential expression analysis using edgeR. To correct for multiple testing in these analyses, we applied the Benjamini–Hochberg correction to generate an FDR value for each gene. Genes DE with an FDR <0.05 were considered significant. The raw FASTQ files generated during this project have been submitted to the European Bioinformatics Institute (EBI) under project accession numbers PRJEB29677 and PRJEB39825.

Exogenous ZAP expression

Three independent cultures of A549 ZAP KO cells (ZAP guide 1 clone 2) were transduced with lentiviral SCRPSY vectors expressing either ZAP-S or an empty backbone as control. Cells were transduced in the presence of polybrene (final concentration: 8 μg/ml) and spinoculated at 400 × g for 1 hour at room temperature. After 48 hours, cells were harvested in TRIzol (Ambion, Thermo Fisher Scientific, Texas, USA) prior to RNA extraction (RNeasy, Qiagen) and RNA-seq analysis, as described above. Additional samples were harvested for FACS analysis to confirm the percentage of transduction (EMPTY 92.2% and ZAP-S 76%) alongside samples for western blot analysis.

Viruses and titrations

Human A549 cells and derivatives were stimulated with 1,000 units of IFNβ (PBL Assay Science) for 24 hours before being infected with a serially diluted challenge of a single cycle *Indiana vesiculovirus* system, rVSVΔG-GFP [56]. At 16 hours postinfection, the levels of infection were quantified using flow cytometry.

Human MT4 cells were transduced with IRG-encoding SCRPSY lentiviral vectors [57] and 48 hours later were infected with 7-point serially diluted doses of a panel of GFP-encoding viruses. Infected cells were enumerated using flow cytometry and the titre of each virus (in cells transduced with each IRG) was extrapolated from 3 points in the linear range. Each titre is plotted as a fold change in titre (relative to cells transduced with the empty vector). The virus panel consisted of HIV-1 (NHG, JQ585717.1), human alphaherpesvirus 1 (HSV-1) [58], adenovirus 5 [59], *Indiana vesiculovirus* [56], and *Chandipura vesiculovirus* [60]; *Bunyamwera*

orthobunyavirus and *Rift Valley fever phlebovirus* (RVFV 35/74) [61]; and parainfluenza virus 3, parainfluenza virus 5, Sendai virus (ViraTree), and *Semliki Forest Virus* [62]. In addition, the design and rescue of the NS1-eGFP expressing A/Puerto Rico/8/1934 (H1N1) virus has been described previously [47]. The A/California/04-061-MA/2009(H1N1) NS1-eGFP virus was rescued by replacing the parental NS segment (GenBank accession KX134783.1) with the PR8 NS1-eGFP segment. An additional NS1-eGFP expressing NS segment was designed using the NS1 and NEP sequences of A/mallard/Netherlands/10-Cam/1999(H1N1) (GenBank accession KC209519.1) and used to rescue the Mallard-GFP virus in the same fashion as the PR8 NS1-eGFP virus described previously [47]. The IRG-encoding vectors were synthesised (GENEWIZ, New Jersey, USA), and the following sequences were used: NM_003840 (TNFRSF10D), NM_001947 (DUSP7), NM_032294 (CAMKK1), NM_022766 (CERK), NM_134269 (SMTN), NM_001037125 (UNKL), NM_145290 (ADGRA3), NM_001010924 (FAM171A1), NM_001080424 (KDM6B), and NR_024031 (DANCR, note that DANCR is not protein coding).

Western blotting

Western blotting was carried out as described previously [63] using a Li-COR Odyssey scanner and the following antibodies: ZAP (Proteintech (Illinois, USA) 16820-1-AP), Phospho-STAT1, (Cell Signaling Technology 9167), GAPDH, (Cell Signaling Technology (Massachusetts, USA) #2118), FAM171A1 (Abcam (Cambridge, UK) ab229247), or actin (JLA20 hybridoma, courtesy of the Developmental Studies Hybridoma Bank, University of Iowa). In Fig 5G, proteins were visualised using an ECL system (GE Healthcare (Illinois, USA) RPN2106) and the above antibodies.

CpG conservation

Alignments available for the 50 most DE IRGs ($n = 34$), ISGs ($n = 30$), and type I IFNs ($n = 4$) were obtained from existing primate alignments [31] and used to estimate the degree of CpG conservation. Each alignment includes 9 primate species: *Callithrix jacchus*, *Chlorocebus sabaeus*, *Gorilla gorilla*, *Homo sapiens*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis*, and *Pongo abelii*. For each site in the alignment that had at least 1 CpG, the percentage of CpG conservation was computed by normalising the number of CpGs by the total number of sequences in the alignment. CpG/kb was then determined from the ratio of the number of CpGs at each degree of conservation to the total gene length of all the sequences.

Codon Adaptation Index

The Codon Adaptation Index (CAI) for each CDS in the genome was calculated using an established method [25]. Briefly, a relative adaptiveness weight [w] for each codon [c] is calculated using the frequency of the codon [f_c] and the frequency of the most frequent synonymous codon for the corresponding amino acid [$\max(f_c)$] across all CDSs in the genome:

$$w_c = \frac{f_c}{\max(f_c)}$$

The CAI of each individual CDS is then calculated as the geometric mean of the weight of each codon in the sequence over its codon length [L]:

$$\mathrm{CAI} = \frac{\prod_{c=1}^{L} w_c}{L}$$

Data retrieval from the "Interferome" database

To visualise the expression levels of genes upon IFN treatment in other datasets, the Interferome database was used. Data from the Interferome v2.01 [2] were downloaded from the web application http://www.interferome.org/. The database was searched for a list of Ensembl IDs, and the following additional search criteria were used: Interferon Type I, Species *Homo Sapiens*, Fold Change Up/Down 1.0. The retrieved experimental data of those genes was downloaded as a text file and used for downstream analysis.

Supporting information

S1 Fig. Sliding window analysis of the CpG composition of ISGs and IRGs. (A) The CpG frequency (upper) and normalised (to GC content) frequency (lower) in transcripts from primary human fibroblasts DE in response to type I IFN. CpG values were extracted for each human ISG and IRG defined in [4]. **(B)** The median CpG frequency (left) and normalised frequency (right) of windows of 50 genes. ISGs and IRGs were ranked according to fold change (FC) and the average CpG value calculated for a window of 50 genes. Window medians were calculated at 10-gene intervals. **(C)** CpG frequency (upper panels) and normalised frequency (lower panels) distributions for the 50-gene windows defined in (B). Each distribution is compared to the same 50 human genes selected at random. Statistical significance between ISG and IRG distributions of CpG frequencies was assessed by Wilcoxon rank sum test. The horizontal dotted line in each plot represents the median value for all transcripts in the human genome. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; GC, guanine–cytosine; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene.
(TIF)

S2 Fig. Normalised dinucleotide composition of the most DE ISGs and IRGs. The 50 most DE human ISGs and IRGs (ranked by Log_2FC) were selected and the frequency of each dinucleotide calculated. Each row represents the first nucleotide and each column the second nucleotide. The median frequency of each dinucleotide in the human transcriptome is indicated by the horizontal dotted line. Statistical significance between ISG and IRG distributions of dinucleotide frequencies was assessed by Wilcoxon rank sum test. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; IRG, interferon-repressed gene; ISG, interferon-stimulated gene.
(TIF)

S3 Fig. The relative importance of features used to distinguish IRGs from random genes. The relative discriminating power of the most significant features identified using a supervised machine learning approach. Features are ranked according to their ability to discriminate the 50 most DE IRGs from 50 non-DE transcripts selected randomly from the remainder of the transcribed human genome. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; IRG, interferon-repressed gene.
(TIF)

S4 Fig. The vertebrate interferome has a CpG bias. (A) The CpG values normalised for both length and GC content of the top 50 most DE human ISGs and IRGs (ranked by Log_2FC). The dotted line represents the median frequency of all transcripts in the relevant genome. A random sample of non-DE genes is included for reference. Bar and whiskers represent the median and interquartile range of each distribution, respectively. **(B)** The remaining 9 vertebrate

species plotted as in (A). The underlying RNA-seq data used were previously published open-access data [4] and were also described in the Fig 1 legend. Significance was determined using the Wilcoxon rank sum test with continuity correction. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; GC, guanine–cytosine; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; RNA-seq, RNA sequencing.
(TIF)

S5 Fig. CpG suppression in human transcripts and GO analysis of the most suppressed human transcripts. (A) Observed over expected CpG values (equivalent to CpG frequencies normalised for GC content) for every transcript in the human genome. The diagonal dotted line represents equal observed and expected CpG values (3 cDNAs were excluded from the plot due to excessive length but did not display different trends). **(B)** Significant GO terms associated with 1,000 human transcripts over 100 nucleotides in length with the lowest CpG content (normalised for GC composition). Blue bars indicate GO processes and functions associated with the immune response. Removal of IFNα2 and IFNα14 ablated the significance of these terms. * Full GO process names: "positive regulation of peptidyl-serine phosphorylation of STAT protein" and "negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains". The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). GC, guanine–cytosine; GO, gene ontology; IFN, interferon.
(TIF)

S6 Fig. Statistical analysis of Fig 3B. Matrices highlighting the significance of potential comparisons in Fig 3B. Significance was assessed using the Kruskal–Wallis rank sum test followed by post hoc analysis using the Dunn test including Benjamini–Hochberg correction for multiple comparisons. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159).
(TIF)

S7 Fig. CpG composition of type I IFN genes. (A) The CpG frequency and (B) the CpG frequency normalised for GC composition of type I IFN genes relative to 50 genes selected at random from the remainder of the genome. The dotted line represents the median CpG frequency of all transcripts in the relevant genome. Statistical significance between IFN gene and random gene distributions of CpG frequencies was assessed by Wilcoxon rank sum test. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). GC, guanine–cytosine; IFN, interferon.
(TIF)

S8 Fig. Selecting IRGs for further investigation. Venn diagrams representing the numbers of **(A)** significantly DE ISGs or **(B)** significantly DE IRGs from the RNA-seq of A549 cells from either ("bulk") or ("clones") stimulated with 1,000 units/ml of IFNβ (4 hours). To increase confidence in the classification of these DE genes, in all cases, these data were filtered to only consider genes that were significantly DE in either bulk populations or clonal lines and were independently classified as DE in primary fibroblasts [4]. Arrows highlight the percentages of genes that differ (no longer DE) between the ZAP KO and "no guide" controls. The 15 IRGs present in bulk and clonal populations in the "no guide" control (left-hand side of panel B) are expanded in Fig 5A. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). DE, differentially expressed; IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; RNA-seq, RNA sequencing; ZAP, zinc-

finger antiviral protein.
(TIF)

S9 Fig. The putative ZAP targets are documented IRGs. (A) A heat map describing the response of the putative ZAP targets to IFN treatment is shown based on previous studies extracted from the Interferome database (www.interferome.org). Colour and intensity reflects the direction and extent of differential expression observed in the individual studies: ISGs are orange, and IRGs are blue. **(B)** As in (A), a heat map is shown summarising the previous measurements in the Interferome database of 46 genes (46 transcripts of the top 50 were retrievable from the database), which are reduced in abundance following exogenous ZAP expression. Putative ZAP targets are highlighted using red typeface. **(C)** As in (A) and (B), a heat map is shown summarising the previous measurements in the Interferome database of 46 genes which are increased in abundance following exogenous ZAP expression. **(D–G)** Matrices highlighting the significance of potential comparisons in Fig 5C (D and E), 5L (F) and 5M (G). Significance was assessed using the Kruskal–Wallis rank sum test followed by post hoc analysis using the Dunn test including Benjamini–Hochberg correction for multiple comparisons. The underlying data from this figure are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). IFN, interferon; IRG, interferon-repressed gene; ISG, interferon-stimulated gene; ZAP, zinc-finger antiviral protein.
(TIF)

S1 Table. The ISGs and IRGs from Fig 2. IRG, interferon-repressed gene; ISG, interferon-stimulated gene.
(XLSX)

S2 Table. Details of putative ZAP targets. ZAP, zinc-finger antiviral protein.
(XLSX)

S1 Raw Images. Raw images of the WBs presented in this study. The underlying image files are openly available (http://dx.doi.org/10.5525/gla.researchdata.1159). WB, western blot.
(TIF)

Acknowledgments

We thank Ron Fouchier, Paul Digard, Laurence Tiley, Ed Hutchinson, Daniel Perez, Andrew Easton, Chris Boutell, and Jeroen Kortekaas for viruses and Joseph Hughes for helpful discussions.

Author Contributions

Conceptualization: Andrew E. Shaw, Suzannah J. Rihn, David L. Robertson, Massimo Palmarini, Sam J. Wilson.

Data curation: Andrew E. Shaw, Richard J. Orton.

Formal analysis: Andrew E. Shaw, Suzannah J. Rihn, Nardus Mollentze, Richard J. Orton, Paul C. D. Johnson.

Funding acquisition: Suzannah J. Rihn, Alfredo Castello, Daniel G. Streicker, David L. Robertson, Massimo Palmarini, Sam J. Wilson.

Investigation: Andrew E. Shaw, Suzannah J. Rihn, Nardus Mollentze, Arthur Wickenhagen, Douglas G. Stewart, Richard J. Orton, Siddharth Bakshi, Mila Rodriguez Collados, Matthew

L. Turnbull, Joseph Busby, Quan Gu, Katherine Smollett, Connor G. G. Bamford, Elena Sugrue, Sam J. Wilson.

Methodology: Andrew E. Shaw, Suzannah J. Rihn, Nardus Mollentze, Richard J. Orton, Srikeerthana Kuchi.

Project administration: Suzannah J. Rihn, Sam J. Wilson.

Software: Nardus Mollentze, Richard J. Orton, Srikeerthana Kuchi.

Supervision: Suzannah J. Rihn, Ana Filipe Da Silva, Daniel G. Streicker, David L. Robertson, Massimo Palmarini, Sam J. Wilson.

Visualization: Andrew E. Shaw, Suzannah J. Rihn, Nardus Mollentze, Arthur Wickenhagen.

Writing – original draft: Andrew E. Shaw, Suzannah J. Rihn, Sam J. Wilson.

Writing – review & editing: Andrew E. Shaw, Suzannah J. Rihn, Nardus Mollentze, Arthur Wickenhagen, Douglas G. Stewart, Richard J. Orton, Srikeerthana Kuchi, Siddharth Bakshi, Mila Rodriguez Collados, Matthew L. Turnbull, Joseph Busby, Quan Gu, Katherine Smollett, Connor G. G. Bamford, Elena Sugrue, Paul C. D. Johnson, Ana Filipe Da Silva, Alfredo Castello, Daniel G. Streicker, David L. Robertson, Massimo Palmarini, Sam J. Wilson.

References

1. Der SD, Zhou AM, Williams BRG, Silverman RH. Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. Proc Natl Acad Sci U S A. 1998; 95 (26):15623–8. https://doi.org/10.1073/pnas.95.26.15623 WOS:000077697200087. PMID: 9861020
2. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. Nucleic Acids Res. 2013; 41(D1):D1040–D6. https://doi.org/10.1093/nar/gks1215 WOS:000312893300148. PMID: 23203888
3. Schneider WM, Chevillotte MD, Rice CM. Interferon-Stimulated Genes: A Complex Web of Host Defenses. Annu Rev Immunol. 2014; 32:513–45. https://doi.org/10.1146/annurev-immunol-032713-120231 WOS:000336427400017. PMID: 24555472
4. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. PLoS Biol. 2017; 15(12). ARTN e200408610.1371/journal.pbio.2004086. WOS:000418943900021. https://doi.org/10.1371/journal.pbio.2004086 PMID: 29253856
5. Gebhardt A, Laudenbach BT, Pichlmair A. Discrimination of Self and Non-Self Ribonucleic Acids. J Interferon Cytokine Res. 2017; 37(5):184–97. https://doi.org/10.1089/jir.2016.0092 WOS:000400694200002. PMID: 28475460
6. Karlin S, Mrazek J. Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci U S A. 1997; 94(19):10227–32. https://doi.org/10.1073/pnas.94.19.10227 WOS: A1997XX39900042. PMID: 9294192
7. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus attenuation by genome-scale changes in codon pair bias. Science. 2008; 320(5884):1784–7. https://doi.org/10.1126/science.1155761 WOS:000257121200041. PMID: 18583614
8. Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P. 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. 2014; 42(7):4527–45. https://doi.org/10.1093/nar/gku075 WOS:000334761100042. PMID: 24470146
9. Gaunt E, Wise HM, Zhang HY, Lee LN, Atkinson NJ, Nicol MQ, et al. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. Elife. 2016; 5. ARTN e12735 https://doi.org/10.7554/eLife.12735 WOS:000371872300001. PMID: 26878752
10. Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017; 550(7674):124–7. Epub 2017/09/28. https://doi.org/10.1038/nature24039 PMID: 28953888.
11. Meagher JL, Takata M, Goncalves-Carneiro D, Keane SC, Rebendenne A, Ong H, et al. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-

rich viral sequences. Proc Natl Acad Sci U S A. 2019; 116(48):24303–9. Epub 2019/11/14. https://doi.org/10.1073/pnas.1913232116 PMID: 31719195; PubMed Central PMCID: PMC6883784.

12. Gao GX, Guo XM, Goff SP. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. Science. 2002; 297(5587):1703–6. https://doi.org/10.1126/science.1074276 WOS:000177819100050. PMID: 12215647

13. Guo XM, Ma J, Sun J, Gao GX. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. Proc Natl Acad Sci U S A. 2007; 104(1):151–6. https://doi.org/10.1073/pnas.0607063104 WOS:000243456300029. PMID: 17185417

14. Li MMH, Lau Z, Cheung P, Aguilar EG, Schneider WM, Bozzacco L, et al. TRIM25 Enhances the Antiviral Action of Zinc-Finger Antiviral Protein (ZAP). PLoS Pathog. 2017; 13(1). ARTN e1006145 https://doi.org/10.1371/journal.ppat.1006145 WOS:000395743500046. PMID: 28060952

15. Zheng XJ, Wang XL, Tu F, Wang Q, Fan ZS, Gao GX. TRIM25 Is Required for the Antiviral Activity of Zinc Finger Antiviral Protein. J Virol. 2017; 91(9). UNSP e00088 https://doi.org/10.1128/JVI.00088-17 WOS:000399474400007. PMID: 28202764

16. Ficarelli M, Wilson H, Pedro Galao R, Mazzon M, Antzin-Anduetza I, Marsh M, et al. KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. Elife. 2019; 8. Epub 2019/07/10. https://doi.org/10.7554/eLife.46767 PubMed Central PMCID: PMC6615859. PMID: 31284899

17. Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, et al. CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and -independent mechanisms. J Virol. 2019. Epub 2019/11/22. https://doi.org/10.1128/JVI.01337-19 PMID: 31748389.

18. Kmiec D, Nchioua R, Sherrill-Mix S, Stürzel CM, Heusinger E, Braun E, et al. CpG Frequency in the 5′ Third of the *env* Gene Determines Sensitivity of Primary HIV-1 Strains to the Zinc-Finger Antiviral Protein. MBio. 2020; 11. https://doi.org/10.1128/mBio.02903-19 PMID: 31937644

19. Luo X, Wang X, Gao Y, Zhu J, Liu S, Gao G, et al. Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. Cell Rep. 2020; 30(1):46–52 e4. Epub 2020/01/09. https://doi.org/10.1016/j.celrep.2019.11.116 PMID: 31914396.

20. Karlin S, Doerfler W, Cardon LR. Why Is Cpg Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses but Not in Those of Large Eukaryotic Viruses. J Virol. 1994; 68(5):2889–97. WOS: A1994NF45200013. https://doi.org/10.1128/JVI.68.5.2889-2897.1994 PMID: 8151759

21. Goff SP. Zapping viral RNAs. Nature. 2017; 550(7674):46–7. WOS:000412214100038. https://doi.org/10.1038/nature24140 PMID: 28953872

22. Murphy EA, Lin Y-T, Chiweshe S, McCormick D, Raper A, Wickenhagen A, et al. Human cytomegalovirus evades ZAP detection by suppressing CpG dinucleotides in the major immediate early 1 gene. PLoS Pathog. 2020; 16(9):e1008844. https://doi.org/10.1371/journal.ppat.1008844 PMID: 32886716

23. Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:180203888. 2018.

24. Lundberg SM, Lee S-I, editors. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Proces Syst. 2017 20172017.

25. Sharp PM, Li WH. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987; 15(3):1281–95. Epub 1987/02/11. https://doi.org/10.1093/nar/15.3.1281 PMID: 3547335; PubMed Central PMCID: PMC340524.

26. Dolken L, Ruzsics Z, Radle B, Friedel CC, Zimmer R, Mages J, et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA. 2008; 14(9):1959–72. Epub 2008/07/29. https://doi.org/10.1261/rna.1136108 PMID: 18658122; PubMed Central PMCID: PMC2525961.

27. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10. Artn 48 https://doi.org/10.1186/1471-2105-10-10 WOS:000264007400001. PMID: 19133123

28. Greenbaum BD, Rabadan R, Levine AJ. Patterns of Oligonucleotide Sequences in Viral and Host Cell RNA Identify Mediators of the Host Innate Immune System. PLoS ONE. 2009; 4(6). ARTN e5969 https://doi.org/10.1371/journal.pone.0005969 WOS:000267079500016. PMID: 19536338

29. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen JM, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data. 2018;5. ARTN 180015 https://doi.org/10.1038/sdata.2018.15 WOS:000426155900001. PMID: 29485622

30. Zhou P, Tachedjian M, Wynne JW, Boyd V, Cui J, Smith I, et al. Contraction of the type I IFN locus and unusual constitutive expression of IFN-alpha in bats. Proc Natl Acad Sci U S A. 2016; 113(10):2696–701. Epub 2016/02/24. https://doi.org/10.1073/pnas.1518240113 PMID: 26903655; PubMed Central PMCID: PMC4790985.

31. van der Lee R, Wiel L, van Dam TJP, Huynen MA. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. Nucleic Acids Res. 2017; 45(18):10634–48. Epub 2017/10/05. https://doi.org/10.1093/nar/gkx704 PMID: 28977405; PubMed Central PMCID: PMC5737536.

32. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017; 34(7):1812–9. Epub 2017/04/08. https://doi.org/10.1093/molbev/msx116 PMID: 28387841.

33. Todorova T, Bock FJ, Chang P. PARP13 regulates cellular mRNA post-transcriptionally and functions as a pro-apoptotic factor by destabilizing TRAILR4 transcript. Nat Commun. 2014;5. ARTN 5362 https://doi.org/10.1038/ncomms6362 WOS:000345655100002. PMID: 25382312

34. Kerns JA, Emerman M, Malik HS. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. PLoS Genet. 2008; 4(1). ARTN e21 https://doi.org/10.1371/journal.pgen.0040021 WOS:000255378700005. PMID: 18225958

35. Hayakawa S, Shiratori S, Yamato H, Kameyama T, Kitatsuji C, Kashigi F, et al. ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses. Nat Immunol. 2011; 12(1):37–U56. https://doi.org/10.1038/ni.1963 WOS:000285465100011. PMID: 21102435

36. Schwerk J, Soveg FW, Ryan AP, Thomas KR, Hatfield LD, Ozarkar S, et al. RNA-binding protein isoforms ZAP-S and ZAP-L have distinct antiviral and immune resolution functions. Nat Immunol. 2019; 20 (12):1610–20. Epub 2019/11/20. https://doi.org/10.1038/s41590-019-0527-6 PMID: 31740798.

37. Burke JM, Moon SL, Matheny T, Parker R. RNase L Reprograms Translation by Widespread mRNA Turnover Escaped by Antiviral mRNAs. Mol Cell. 2019; 75(6):1203–17.e5. https://doi.org/10.1016/j.molcel.2019.07.029 PMID: 31494035

38. Chitrakar A, Rath S, Donovan J, Demarest K, Li Y, Sridhar RR, et al. Real-time 2-5A kinetics suggest that interferons β and λ evade global arrest of translation by RNase L. Proc Natl Acad Sci U S A. 2019; 116(6):2103–11. https://doi.org/10.1073/pnas.1818363116 PMID: 30655338

39. Aso H, Ito J, Koyanagi Y, Sato K. Comparative Description of the Expression Profile of Interferon-Stimulated Genes in Multiple Cell Lineages Targeted by HIV-1 Infection. Front Microbiol. 2019;10. https://doi.org/10.3389/fmicb.2019.00010 PMID: 30728810

40. Kayagaki N, Yamaguchi N, Nakayama M, Eto H, Okumura K, Yagita H. Type I interferons (IFNs) regulate tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) expression on human T cells: A novel mechanism for the antitumor effects of type IIFNs. J Exp Med. 1999; 189(9):1451–60. https://doi.org/10.1084/jem.189.9.1451 WOS:000080200900010. PMID: 10224285

41. Marsters SA, Sheridan JP, Pitti RM, Huang A, Skubatch M, Baldwin D, et al. A novel receptor for Apo2L/TRAIL contains a truncated death domain. Curr Biol. 1997; 7(12):1003–6. https://doi.org/10.1016/s0960-9822(06)00422-2 WOS:A1997YL44000035. PMID: 9382840

42. Yano S, Tokumitsu H, Sodeling TR. Calcium promotes cell survival through CaM-K kinase activation of the protein-kinase-B pathway. Nature. 1998; 396(6711):584–7. https://doi.org/10.1038/25147 WOS:000077466800060. PMID: 9859994

43. McArdle J, Schafer XL, Munger J. Inhibition of Calmodulin-Dependent Kinase Kinase Blocks Human Cytomegalovirus-Induced Glycolytic Activation and Severely Attenuates Production of Viral Progeny. J Virol. 2011; 85(2):705–14. https://doi.org/10.1128/JVI.01557-10 WOS:000285554300006. PMID: 21084482

44. Rossetto CC, Pari G. KSHV PAN RNA Associates with Demethylases UTX and JMJD3 to Activate Lytic Replication through a Physical Interaction with the Virus Genome. PLoS Pathog. 2012; 8(5). ARTN e1002680 https://doi.org/10.1371/journal.ppat.1002680 WOS:000305322900014. PMID: 22589717

45. Dunn DM, Rodriguez-Sanchez I, Schafer X, Munger J, Goodrum F. Human Cytomegalovirus Induces the Expression of the AMPKa2 Subunit To Drive Glycolytic Activation and Support Productive Viral Infection. J Virol. 2021; 95(5). https://doi.org/10.1128/jvi.01321-20 PMID: 33268515

46. Liu SY, Sanchez DJ, Aliyari R, Lu S, Cheng G. Systematic identification of type I and type II interferon-induced antiviral factors. Proc Natl Acad Sci U S A. 2012; 109(11):4239–44. https://doi.org/10.1073/pnas.1114981109 PMID: 22371602

47. Rihn SJ, Aziz MA, Stewart DG, Hughes J, Turnbull ML, Varela M, et al. TRIM69 Inhibits Vesicular Stomatitis Indiana Virus. J Virol. 2019; 93(20). https://doi.org/10.1128/JVI.00951-19 PubMed Central PMCID: PMC6798119. PMID: 31375575

48. Kueck T, Bloyet L-M, Cassella E, Zang T, Schmidt F, Brusic V, et al. Vesicular Stomatitis Virus Transcription Is Inhibited by TRIM69 in the Interferon-Induced Antiviral State. J Virol. 2019; 93(24). https://doi.org/10.1128/JVI.01372-19 PMID: 31578292

49. Chen SD, Xu YH, Zhang K, Wang XL, Sun J, Gao GX, et al. Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA. Nat Struct Mol Biol. 2012; 19(4):430–5. https://doi.org/10.1038/nsmb.2243 WOS:000302514400011. PMID: 22407013

50. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science. 2018; 362(6414):577–80. https://doi.org/10.1126/science.aap9072 PMID: 30385576; PubMed Central PMCID: PMC6536379.

51. Chen TQ, Guestrin C. XGBoost: A Scalable Tree Boosting System. Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016:785–94. https://doi.org/10.1145/2939672.2939785 WOS:000485529800092.

52. Rashmi KV, Gilad-Bachrach R, editors. DART: Dropouts meet Multiple Additive Regression Trees. AISTATS; 2015 20152015.

53. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008; 28(5):1–26. WOS:000260799600001.

54. R-Core-Team. R: A Language and Environment for Statistical Computing. Vienna2018 2018.

55. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science. 2014; 343(6166):84–7. https://doi.org/10.1126/science.1247005 WOS:000329162000053. PMID: 24336571

56. Whitt MA. Generation of VSV pseudotypes using recombinant DeltaG-VSV for studies on virus entry, identification of entry inhibitors, and immune responses to vaccines. J Virol Methods. 2010; 169 (2):365–74. https://doi.org/10.1016/j.jviromet.2010.08.006 PMID: 20709108; PubMed Central PMCID: PMC2956192.

57. Kane M, Zang TM, Rihn SJ, Zhang F, Kueck T, Alim M, et al. Identification of Interferon-Stimulated Genes with Antiretroviral Activity. Cell Host Microbe. 2016; 20(3):392–ss. https://doi.org/10.1016/j.chom.2016.08.005 PMID: 27631702; PubMed Central PMCID: PMC5026698.

58. Pasdeloup D, Beilstein F, Roberts AP, McElwee M, McNab D, Rixon FJ. Inner tegument protein pUL37 of herpes simplex virus type 1 is involved in directing capsids to the trans-Golgi network for envelopment. J Gen Virol. 2010; 91(Pt 9):2145–51. https://doi.org/10.1099/vir.0.022053-0 PMID: 20505007; PubMed Central PMCID: PMC3066548.

59. de Martin R, Raidl M, Hofer E, Binder BR. Adenovirus-mediated expression of green fluorescent protein. Gene Ther. 1997; 4(5):493–5. https://doi.org/10.1038/sj.gt.3300408 PMID: 9274728.

60. Marriott AC, Hornsey CA. Reverse genetics system for Chandipura virus: tagging the viral matrix protein with green fluorescent protein. Virus Res. 2011; 160(1–2):166–72. https://doi.org/10.1016/j.virusres.2011.06.007 PMID: 21704089.

61. Feng JJ, Wickenhagen A, Turnbull ML, Rezelj VV, Kreher F, Tilston-Lunel NL, et al. Interferon-Stimulated Gene (ISG)-Expression Screening Reveals the Specific Antibunyaviral Activity of ISG20. J Virol. 2018; 92(13). ARTN e02140-17 https://doi.org/10.1128/JVI.02140-17 WOS:000435100400034. PMID: 29695422

62. Tamberg N, Lulla V, Fragkoudis R, Lulla A, Fazakerley JK, Merits A. Insertion of EGFP into the replicase gene of Semliki Forest virus results in a novel, genetically stable marker virus. J Gen Virol. 2007; 88(Pt 4):1225–30. https://doi.org/10.1099/vir.0.82436-0 PMID: 17374766; PubMed Central PMCID: PMC2274952.

63. Rihn SJ, Foster TL, Busnadiego I, Aziz MA, Hughes J, Neil SJD, et al. The Envelope Gene of Transmitted HIV-1 Resists a Late Interferon Gamma-Induced Block. J Virol. 2017; 91(7). UNSP e02254 https://doi.org/10.1128/JVI.02254-16 WOS:000398833200018. PMID: 28100611