Adaptive Evolution in the Interferon Response

Elena Judd
Laboratory of Dr. Sara Sawyer
Department of Molecular, Cellular, Developmental Biology
University of Colorado Boulder

Defense Date: April 6, 2018

Primary Thesis Advisor: Sara Sawyer, PhD. Department of Molecular, Cellular, and Developmental Biology and the BioFrontiers Institute
Honors Council Representative: Brian DeDecker, PhD. Department of Molecular, Cellular, and Developmental Biology
Thesis Committee Member: Alison Vigers, PhD; Department of Molecular, Cellular, and Developmental Biology and Department of Neuroscience

Abstract

Interferons and the genes they upregulate are involved in many diverse mechanisms that suppress viral replication. Upon cellular recognition of a virus, interferons are expressed, secreted, and taken up by neighboring cells. Interferons then stimulate the expression of hundreds of genes that target every step of a viral life cycle. In turn, viruses have evolved strategies to inhibit the interferon response by evading recognition, impairing signaling, or hindering restrictive functions of proteins that are expressed in response to interferon signaling. Genes that induce the response and interferon-stimulated genes must evolve in turn to maintain an effective defense against viruses in spite of high viral mutation rates. This evolutionary battle for the upper hand creates a situation of coevolution. In this study, I have analyzed genes that induce the interferon response and interferon-stimulated genes for signatures of enhanced natural selection. I have detected loci which have significantly elevated rates of non-synonymous (amino acid altering) substitutions. Interestingly, the highest signal for elevated rates of non-synonymous mutations was found in interferon-stimulated genes, and not induction genes when compared to a set of random genes. Evolution that favors increased rates of non-synonymous mutations may indicate that viruses are intensifying selective pressure for mutations, especially in interferon-stimulated genes. I experimentally test this hypothesis with several primate alleles of a rapidly evolving interferon-stimulated gene and the viral protein it targets. Alleles of this gene from Humans, Chimpanzees, and Gibbons are able to degrade the viral protein to some extent. Further testing on this degradation phenotype is required.

Introduction

Viruses are abundant and radically diverse pathogens that continuously threaten public health (Bowie and Unterholzner 2008). Viruses exert evolutionary pressure on their hosts. Hosts must adapt to reduce death by infection. The mammalian innate immune system is a mechanism for identifying and fighting such pathogens. Interferons belong to a class of innate immune system genes called cytokines and are necessary for the survival of vertebrates because they act hours to days before other immune responses (Stark et al. 1998). These cytokines are produced and secreted following the recognition of a pathogen. Interferons bind to cell-surface receptors, leading to a signaling cascade that results in the increased expression of a variety of diverse antiviral effectors (Stark et al. 1998, Schoggins and Rice 2011). The amplified expression of hundreds of genes, called interferon-stimulated genes (ISGs), can mitigate viral infection by enabling clearance of infected cells and lessening viral replication.

Viruses have evolved strategies to sabotage the interferon response by dodging recognition, impairing signaling, or obstructing methods of host restriction. Conversely, genes that induce the interferon response and interferon-stimulated genes must evolve in turn to prevent or neutralize infection. It has been shown that many rapidly evolving human genes are involved in immunity (Sawyer and Meyerson 2011). This is likely because targeting or avoiding a variable substance (e.g. a quickly mutating viral protein-coding gene) requires a certain amount of flexibility. I predict that some sites of rapid evolution in the interferon response exist at the site of protein-protein contact with viruses. I hope to determine 1. how the primate interferon response may have evolved due to viral pressures by determining what aspects of the response harbor genes that evolve more rapidly than genes in other cellular pathways, and 2. whether signals of high amino acid variability in a protein correlate to interactions with a virus.

Natural selection allows randomly occurring beneficial genetic mutations to endure and expand throughout a population. Mutations at a genetic level can occasionally lead to functional

differences in the encoded protein. Amino acid variation often leads to structural and functional changes whereas silent mutations, that result in the same amino acid sequence, do not. Enhanced proteomic variability leads to novel phenotypes and therefore may exist at host-virus molecular interfaces. In this case, random mutations in host genes are selected for when these genes produce proteins which are able to interact with a pathogenic protein in a more advantageous manner. This pattern can be characterized by an increased retention of amino acid altering (non-synonymous) mutations over mutations that do not alter the encoded amino acid (synonymous) called positive selection. This term will be used interchangeably with "rapid evolution". Contrary to positive selection is negative, or purifying, selection in which evolution favors conservation of a protein coding sequence. Some of the predicted host-virus interfaces have been experimentally validated by other groups (Elde et al. 2009, Patel et al. 2012) and I will be experimentally validating a predicted interface that has not been previously reported.

I have used a bioinformatics pipeline to predict sites of positive selection in primate interferon response-related genes. Primates are a prime dataset for such analysis because they are closely related to humans and many of the viruses that antagonize our species have evolved from viruses that infected our primate ancestors. Also, primates diverged long enough ego that mutations have become fixed in the populations (now called substitutions), so true interspecies genetic diversity, not only single nucleotide polymorphisms, can be identified. Computational analysis references primate gene alignments and a phylogenetic tree to measure rates of mutations. Alignments of primate protein coding sequences allow the program to identify substitutions, while phylogenies allow the program to determine when substitutions occurred in evolutionary time. The pipeline used here detects positive selection as a function of increased ratios of the rates of non-synonymous (dN) to synonymous (dS) mutations (dN/dS). This metric can be measured as a whole gene average or on a codon by codon basis.

As proof of concept, I will analyze a particular gene that was found to be under positive selection and determine whether this, presumably adaptive, evolution might have been driven by interaction with a viral protein. I chose one gene that is highly variable across primates and is also known to interact with a viral protein. The primate gene, IFI27 (ISG12a), produces a protein that targets a hepatitis C viral (HCV) protein for degradation. The viral protein, Nonstructural Protein 5a (NS5a), is important for replication of hepatitis C virus and modulation of the host-cell interferon response (He et al. 2006). While hepatitis C virus naturally infects humans, chimpanzees can also be experimentally infected (Couto and Kolykhalov 2006). GB viruses are hepatitis C-related viruses that do infect primates, although may not be pathogenic (Patel et al.2012). I predict that some primate alleles will show different functional phenotypes in their ability to degrade the hepatitis C virus NS5a protein based on the overall hypothesis that adaptation in this gene has occurred over the course of primate divergence due to ancient hepatitis C-like viruses. These phenotypes will be tested in a co-transfection assay. If there is a protein-protein interaction, the viral protein will be degraded by the primate protein. Degradation is measured by disappearance of the protein on a Western blot following co-transfection. I expect to see one of three results. Degradation of the NS5a protein could be present only in primates who are susceptible to infection, suggesting that this interaction evolved as a way to constrain infection. The second possible result is that degradation will be much more effective for proteins encoded by primates that are not susceptible to infection, suggesting that this interaction can act as a barrier to cross-species transmission of current day hepatitis C virus. Furthermore, it is possible that all primate proteins degrade to the same extent, indicating that hepatitis C virus or HCV-like viruses did not put selective pressure on the IFI27 gene to change over time.

Materials and methods

Sequence acquisition and alignments

Genes involved in the interferon response are split into two categories for this study. The first category includes genes whose proteins induce the response, i.e. those that identify pathogens, produce interferons, and signal for the increased expression of hundreds of genes. The second category is made up of proteins that are upregulated by interferon signaling— including repressors of translation, pro-apoptotic proteins, and factors that sequester viral proteins. We did not place the same gene in more than one category. If a gene is upregulated by interferons and implicated in canonical induction pathways, it was placed in the induction category. The list of induction genes was curated from reviews of interferon signaling pathways (Hornung 2014, Kawai and Akira 2009, Shuai and Liu 2003, Barber 2011). The interferon-stimulated gene list was hand-curated from published literature (Kane et al. 2016, Schneider et al. 2014, Schoggins et al. 2011, Schoggins and Rice 2011). The list was then confirmed by the *Interferome* database (Rusinova et al. 2013). Each interferon-stimulated gene is upregulated at least twofold by type 1 interferons. A list of random human genes was formed using a random gene set generator (Čermák 2016).

We analyzed one-hundred interferon-stimulated genes, ninety-three interferon induction genes, and one hundred random genes for signatures of positive selection. The longest human isoform of each gene was collected from the NCBI Gene database. There were few exceptions where a shorter human isoform was retained because it aligned better to available primate sequences (e.g. SCOC). We gathered predicted sequences of other primate species, belonging to the three main clades of primates: Hominoids, New World Monkeys, or Old World Monkeys (Figure 1a). In each case, we retained the sequence that best aligned to the human isoform. The cDNA sequences were then translated to amino acids and aligned using the Unipro UGENE software (Okonechnikov et al. 2012). Pal2Nal (Mikita et al. 2006) referenced the amino acid alignment in order to align cDNA by codon. This corrects for insertions or deletions in the cDNA sequences better than aligning cDNA directly. Pal2Nal outputs were used for the analysis of each gene.

Evolutionary analysis

Positive selection is detected using the Phylogenetic Analysis by Maximum Likelihood (PAML) program. The algorithm accommodates for the differences in rates of transition/ transversion (same type or different type) base pair changes, unequal codon frequencies, and the probabilities of mutation across the codon (Yang et al. 2007). The third base pair is more likely to change because several amino acids have codons that are redundant across the first two bases and not the third. PAML requires the codon alignment be accompanied by a phylogenetic tree to accurately identify rates of mutations. A master phylogenetic tree with the twenty possible primate species was made using Perelman et al. 2011 as a reference and modified as necessary for each gene (Figure 1a). In order to reduce the possibility of false positives, no alignments with less than 10 sequences were retained (McBee et al. 2015) (Figure 2a).

PAML fits the codon alignments of primate cDNA to multiple models— M0, M2, M8, M8a— of nucleotide substitution (Yang 2007) (Figure 2b). M0 is a measure of the average rate of non-synonymous mutations (dN) to the rate of synonymous mutations (dS) (dN/dS) of every codon in a gene. M2, a simple model that allows for positive selection, puts all codons into three bins at a dN/dS of less than one (conserved), one (neutral), and greater than one (positive selection). M8a is a null model that puts all of a gene's codons into a distribution of dN/dS

values that does not allow any codons to fall beyond a dN/dS ratio greater than one. M8, a more complex positive selection model, also allows codons to fall into a distribution of dN/dS values less than 1, however, some codons can move beyond a dN/dS value of one (Yang 2007).

Likelihood ratio tests determine which model, M8 or M8a, best fits the data. PAML provides a log likelihood (lnl) value for each alignment in both the null and positive selection models. The difference of these values is then doubled, referred to here as "2Δlnl", and used to perform Chi-Square tests with a single degree of freedom. Genes that pass the maximum likelihood ratio test have a p-value of less than 0.05, allowing us to reject the null hypothesis that there is no difference in how well models M8 and M8a fit the data. Genes that pass have codons which fall outside the M8a distribution ending at a dN/dS of 1, thus, these genes are better represented by the model of positive selection, M8.

When the model of best fit is that of positive selection (M8), specific codons are identified which have elevated rates of non-synonymous mutations. This is determined by the Bayes empirical Bayes (BEB) method which accounts for sampling errors in the parameters of the model (Yang et al. 2007). Generally, a dN/dS ratio greater than one indicates positive selection.

All genes predicted to be under positive selection via PAML were confirmed to contain orthologous sequences through reciprocal best hit. This entailed using the UCSC BLAT tool (Kent 2002) to confirm each primate sequence was most similar to the correct human gene. If a sequence was not most homologous to the proper human gene, it was removed from the alignment and PAML was rerun (Figure 2a).

Amplification of IFI27 and preparation of plasmid constructs

Primers were made for IFI27 that anneal to conserved regions of the gene's untranslated regions. These primers amplified IFI27 from prepared cDNA of Human, Chimp, Gibbon, Rhesus Macaque, Crab Eating Macaque, Black Mangabey, Baboon, and Colobus. The constructs were re-amplified with primers containing linkers and then gateway cloned into V5 tagged PLPCX vectors. I was unable to sequence IFI27 from any New World Monkeys. HCV NS5a was prepared in a PCDNA6.2 plasmid with a V5 tag (a gift from Sonya Best's Lab). Plasmids were grown up in DH5-α competent *E. coli* and isolated using Qiagen Midi Preparation Kits. The plasmids were ethanol precipitated for purification purposes before transfection.

Co-transfection and degradation assay

293T immortalized cell lines were used for co-transfections. Cells were plated in 12 well dishes approximately 24hrs prior to transfection at a density of 500,000 cells per well in 10% fetal bovine serum (FBS), antibiotic free Dulbecco's Modified Eagle Media (DMEM), supplemented with 1% L-Glutamine. 150μLs of OptiMem, 4μL Mirius Bio TransIT-293, and 800ng of plasmid DNA were added to each well. Empty V5 tagged PLCPX vectors were used to ensure that all wells were transfected with 8 ng of DNA when the target plasmids did not add up to the full amount. The cells were collected ~48hrs after transfection with Dulbecco's Phosphate Buffered Saline (DPBS). The cells were then pelleted (5mins at 1000rcf) and lysed with radio-immunoprecipitation assay (RIPA) buffer containing a protease inhibitor (cOmplete). Whole cell lysates were sonicated at 40% amplitude for 7 seconds, spun down at 4ºC for 15 mins and the supernatant was removed. I made a 1:10 dilution of the supernatant to perform a Bradford assay which normalizes the amounts of protein added to each well. 10-15μg of protein was loaded to each well in a 12% BioRad stain free gel. Transfer was performed using the TransTurbo BioRad

machine on the low molecular weight setting (IFI27) and as a wet transfer at 100V for 1 hour (NS5a). I detected the target proteins using a 1:1000 dilution of anti-V5 tag antibody [SV5-Pk1] (Abcam ab27671) in 0.1% TBST and a 1:10000 dilution of goat anti-mouse secondary antibody in 0.1% TBST. The enhanced chemiluminescent (ECL) substrate was placed on the nitrocellulose (NS5a) and PVDF (IFI27) membranes and incubated at room temperature for 5 mins before imaging. I used different membranes to accommodate for the different protein weights in an attempt to optimize the western blotting protocol. The intensity of protein bands detected by the anti-V5 antibody were normalized to total protein loaded in each lane.

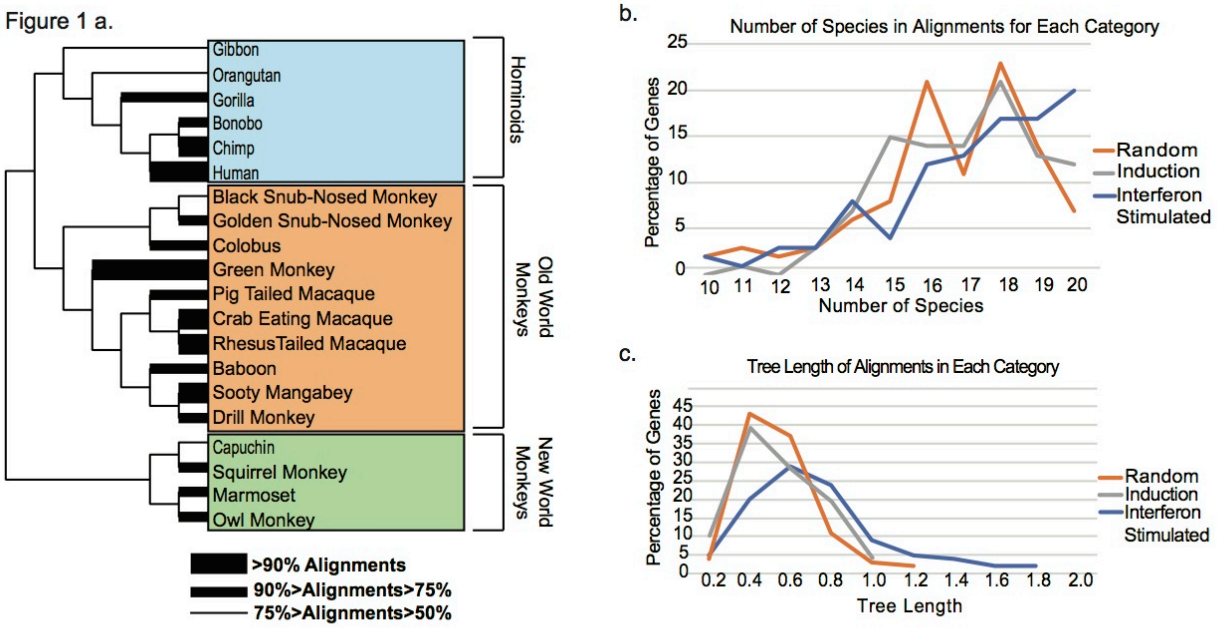


Figure 1. Collection of primate data sets for interferon-related genes and random genes. a. Twenty primate species have available genome sequences on NCBI. Many are used commonly, while others had few usable sequences. b. The number of species used in the evolutionary analysis for the genes in each category. c. Distribution of the tree lengths of the gene alignments in each category (determined by M8).

Results

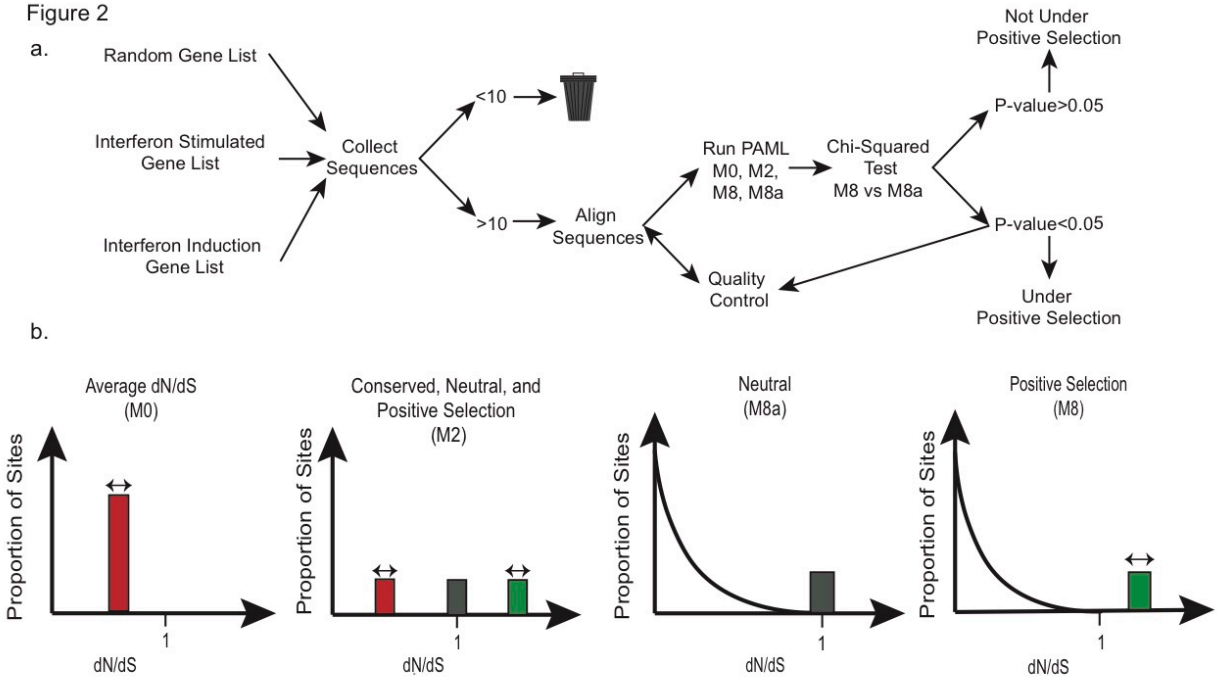
Interferon-Stimulated Genes are evolving more rapidly than Induction or Random Genes

The number of sequences and species represented in an alignment can alter the probability of a gene passing the maximum likelihood ratio test (McBee et al. 2015). All three clades are represented in 75% of alignments and each primate species is represented in at least 50% of alignments (Figure 1a). However, interferon-stimulated genes have the greatest proportion of alignments with the maximum number of species, 20 (Figure 1b). This may allow for more statistical power in identifying sites of positive selection and could be a confounding factor. We will look into this in the future. The tree lengths of genes in the three categories were also compared. Tree length is a measure of the number of amino acid substitutions per codon site across the alignment (Yang et al. 2007). The interferon-stimulated gene category does appear to have alignments with tree lengths of greater value than either the induction or random gene sets (Figure 1c), indicating that the alignments in this category have increased numbers of substitutions.

All genes that were analyzed had values recorded for the number of species included in the alignment, whole gene average dN/dS (measured by M0), tree length, p-value of the

maximum likelihood ratio test, two times the difference of the log likelihood value (2Δlnl), as well as the dN/dS value and percentage of codon sites in both the M2 bin less than 1 and the M8 bin greater than one. Example data for a gene under positive selection and not under positive selection can be seen in Figure 2c. The trend seen here is fairly robust. Genes under positive selection often have higher dN/dS values in the M0, M2, and M8 models. These genes also generally have higher tree lengths and percentage of sites in the M8 bin greater than one. The percentage of sites in the M2 bin less than one is consistently less than genes with similar

Figure 2



c.

Gene ID	Number of Species	M0 dN/dS	Tree Length	P-value M8 vs M8a	2Δ (lnl)	Pass M8 vs M8a	dN/dS M2 Bin < <1(%)	dN/dS M8 Bin >1(%)	Sites Under Selection >90% BEB
MAVS	18	0.74	0.85	1.41E - 06	23	Yes	0.25 (67.8%)	2.9 (20.1%)	14-R, 60-N, 82-V, 100-S, 163-T, 285-S, 310-L, 328-T, 398-T, 423-V, 498-D, 504-R
IFIH1	18	0.31	0.44	>0.05	0.68	No	0.076 (79.4%)	1.35 (20.1%)	N/A

Figure 2. Selection analysis. a. Methods of obtaining sequence alignments and positive selection data. Quality control includes determining if sequences align well and are predicted to be the correct gene. b. PAML codon models of evolution. c. Example of data for gene under positive selection (MAVS) and gene not under positive selection (IFIH1).

sequence data that are not evolving under positive selection.

Of the 100 interferon-stimulated genes analyzed, 43 were under positive selection. Only 22 of 100 random genes and 24 of 93 induction genes (25.8%) analyzed appear to be evolving rapidly. Pair-wise contingency tests confirm that the proportion of genes under positive selection to those that are not is significantly greater for interferon-stimulated genes than random genes

(Chi-square=9.12, df=1, p-value=0.0025) and induction genes (Chi-square=5.55, df=1, p-value=0.019) (Figure 3a, b). There is no significant difference in the proportion of genes under selection in the induction and random categories (Chi-Square=0.2, df=1, p-value=0.65).

Figure 3

a.

Category	Under Positive Selection	Not Under Positive Selection	Total	Ratio
Induction	24	69	93	0.35
Interferon Stimulated	43	57	100	0.79
Random	22	78	100	0.28

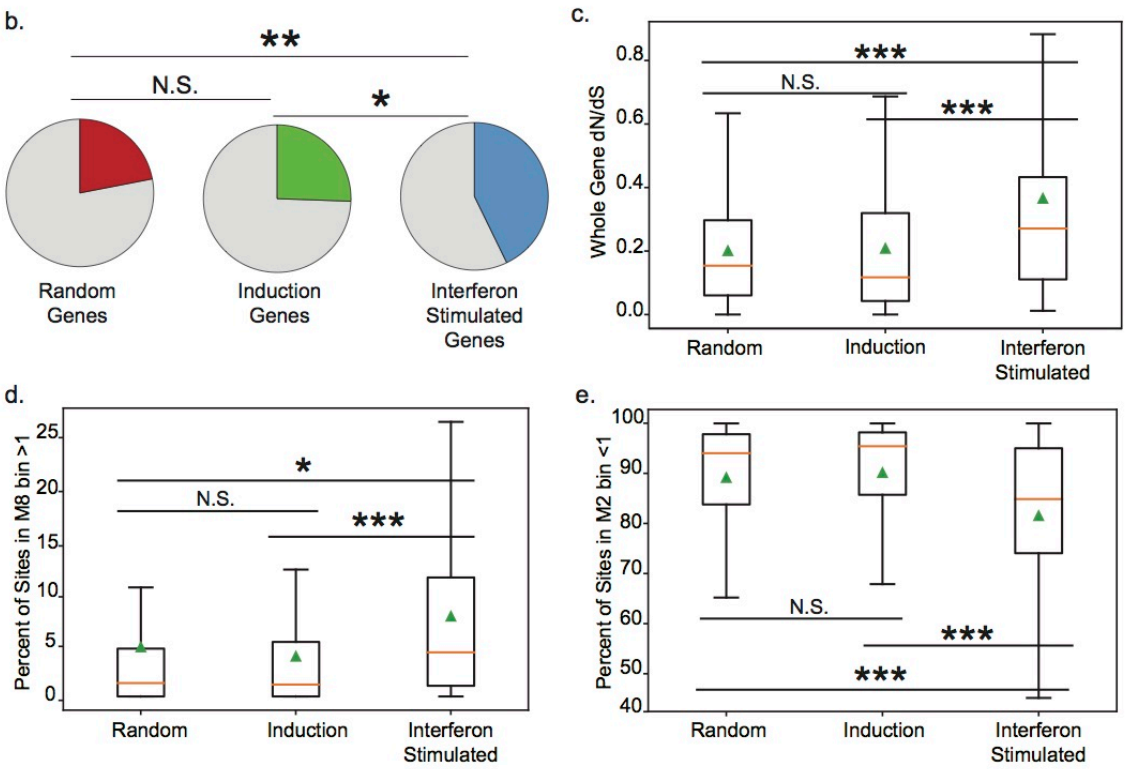


Figure 3: Evolutionary analysis across random, induction, and interferon stimulated genes. a. Contingency table. b. Proportion of genes in each category that are under positive selection (colored). c. Box plot of the whole gene average dN/dS values determined by M0 in each category. d. Box plot of proportion of codon sites per gene in the M8 bin greater than one. e. Box plot of the proportion of codon sites per gene in the M2 bin less than one. *, p-value<0.05, **, p-value<0.01, ***, P-value<0.001, ▲, average.

The distributions of the model M0 whole gene average dN/dS value in each category were compared using a two-sample t-test (TT) and a Kruskal-Wallis rank sum statistical test (KW) (Figure 3c). The average whole gene dN/dS value of interferon-stimulated genes is significantly higher than the random set ($p\text{-value}_{TT}$=3.2E-5, $p\text{-value}_{KW}$=3.4E-4) in both tests. Induction and interferon-stimulated genes are also significantly different ($p\text{-value}_{TT}$=2.8E-4, $p\text{-value}_{KW}$=6.23E-5), but the induction genes are not significantly different from the random genes ($p\text{-value}_{TT}$=0.80, $p\text{-value}_{KW}$=0.29). The average number of sites per gene in the M8 bin greater than one and M2 bin less than one were also compared using two-sample t-tests. Interferon-stimulated genes have a significantly greater average number of codons per gene in the M8 bin greater than one than either the induction or random gene set (p-value=4.9E-4, 0.015) (Figure

3d). This indicates that interferon-stimulated genes have, on average, a greater proportion of sites that fall outside the M8a distribution ending at dN/dS=1. Random genes have a greater proportion of sites in the M2 bin less than one than interferon-stimulated genes (p-value=3.9E-4), but not induction genes (p-value=0.57) (Figure 3e). Consistent with other findings, this suggests that, on average, random and induction genes have more conserved codon sites per gene than interferon-stimulated genes. Despite interferon-stimulated genes having a stronger signal of positive selection across the entire category, some induction genes are also under positive selection (Figure 4).

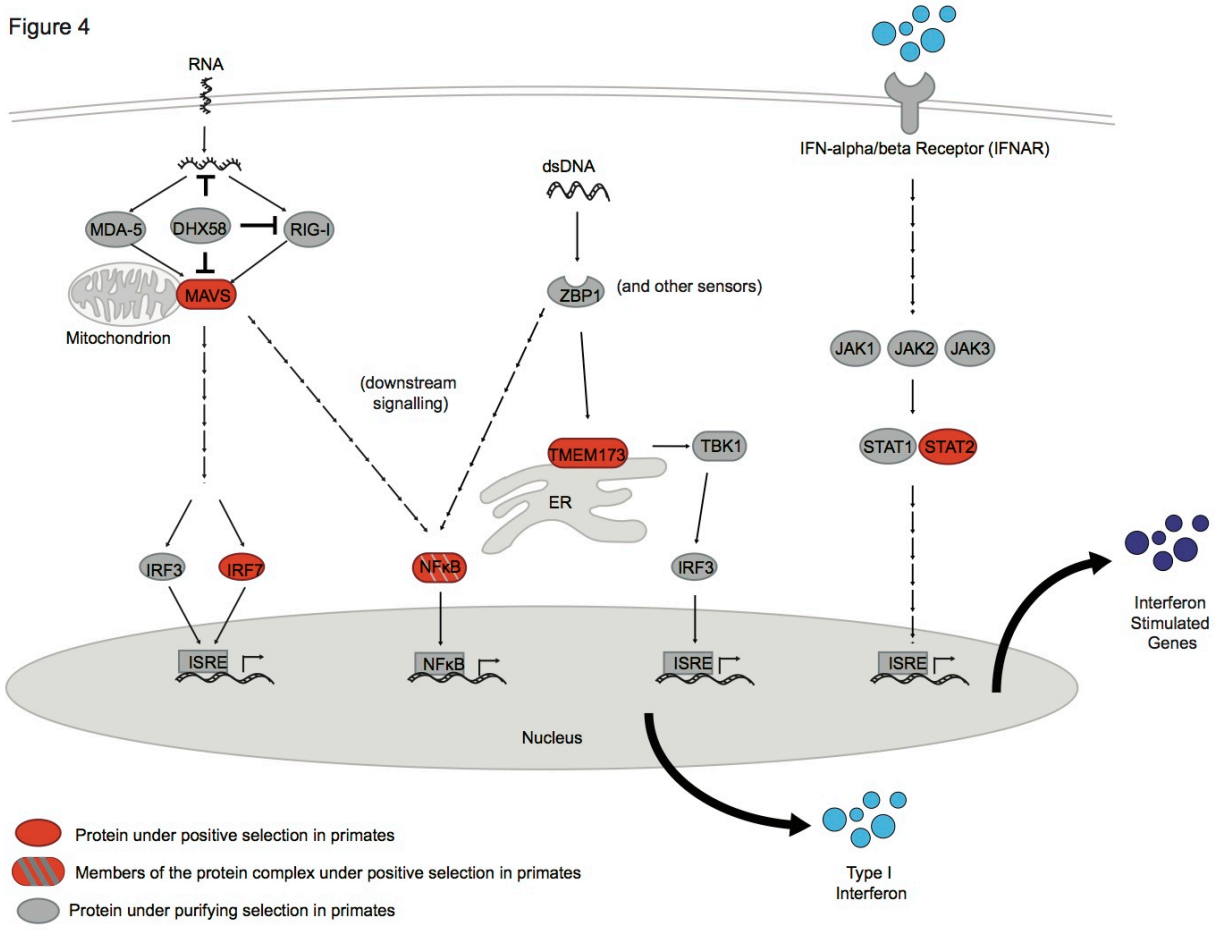


Figure 4. Interferon induction pathway. Several genes in the induction pathway are under positive selection.

Figure 5

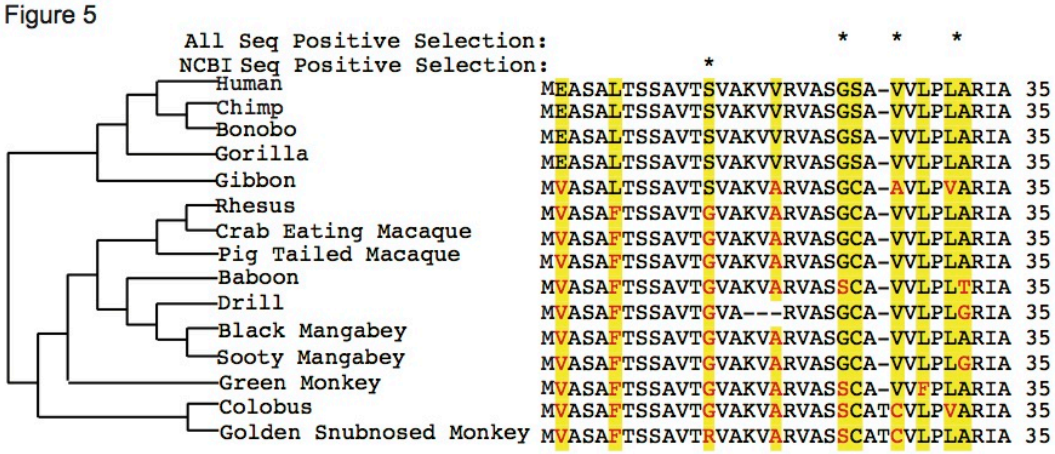
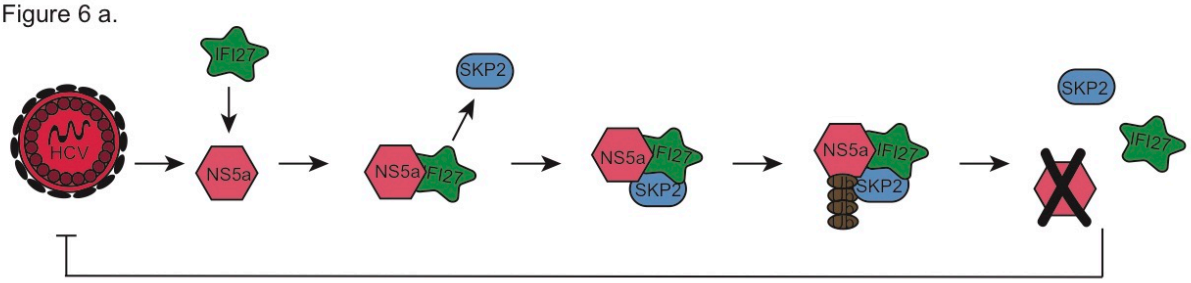


Figure 5. IFI27 (ISG12a) is an interferon stimulated gene that is under positive selection. Asterisks indicate sites that are >90% likely to be evolving rapidly. Sites containing non-synonymous mutations are highlighted. Amino acids that are different from the human sequence at that site are colored red.

Positive selection and functional analysis of IFI27

Positive selection analysis on IFI27, also called ISG12a, was performed twice. Once using sequences obtained from NCBI in the original scan of interferon response related genes and again after I amplified and sequenced several primate alleles from cDNA in the lab. With the 13 predicted sequences acquired from NCBI, the codon alignment of IFI27 passes the maximum likelihood test for belonging in the M8 model of codon evolution (p-value=3.4E-3). The whole gene average dN/dS value was calculated to be 0.942. 73% of codons fell into the M2 bin less than one, while 27% of sites fell into the M8 bin greater than one. PAML identified four sites to be evolving rapidly (S13, A45, A62, I110) with a Bayesian Empirical Bayes probability of >90%. Among replacement/addition of the amplified and sequenced alleles from Human, Chimpanzee, Gibbon, Rhesus Macaque, Crab Eating Macaque, Black Mangabey, Baboon, and Colobus these results changed slightly. With the addition of two new sequences (Gibbon and Black Mangabey), the gene continues to pass the maximum likelihood test (p-value= 3.4E-4). The tree length increased to 0.76 from 0.54 and the whole gene dN/dS value decreased slightly to 0.928. 84% of codons fell into the M2 bin less than 1, while 16% of sites fell into the M8 bin greater than 1. Predicted codons that are under positive selection are subject to change when analysis is performed with a different set of sequences. The new rapidly evolving sites (Bayesian Empirical Bayes >90%) numbered 5 (23G, 26V, 31A 45A, 110T). The protein alignment of IFI27 shows many sites that are different across the alignment (highlighted) with amino acids that have changed from the human sequence indicated in red. A characteristic feature of positive selection is resampling, the mutation of one amino acid to another multiple times across a phylogeny and can be seen in the alignment at sites 24G, 30L, and 31A in the human protein (Figure 5).

Figure 6 a.



b.

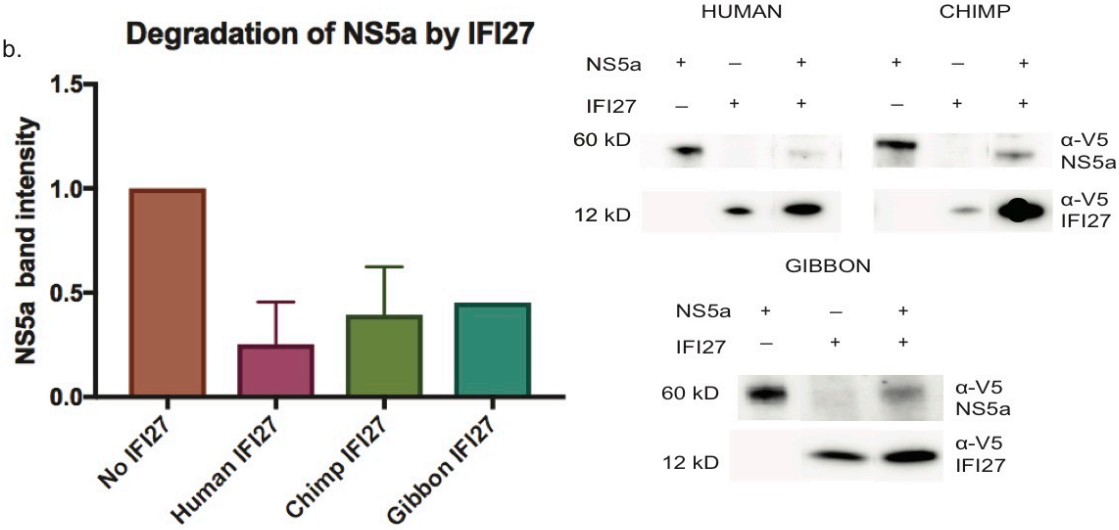


Figure 6. a. IFI27 targets HCV-NS5a for degradation via an E-3 ubiquitin ligase, SKP2. b. Primate alleles of IFI27 target NS5a for degradation to varying extents. Proportion of NS5a degraded is measured as intensity of the NS5a band on a Western blot. Intensity of bands is normalized to total protein loaded. Error bars are standard error of the mean (SEM). There are two biological replicates plotted for Human and Chimp. Only one experiment has been completed for Gibbon. NS5a and IFI27 alone were transfected at 200ng. Degradation was assessed with 200ng NS5a and 600ng IFI27 transfected. These are preliminary results. Experiments are ongoing.

Human IFI27, also known as ISG12a, is upregulated by interferons in response to many viruses: hepatitis C virus, Newcastle Disease virus, West Nile virus, hepatitis E virus, influenza A virus, Japanese encephalitis virus, and Sindbis virus (Xue et al. 2016). I chose to analyze IFI27 functionally because it was identified as being under positive selection (previously unreported) and has been shown to target hepatitis C virus nonstructural protein 5a (NS5a) for degradation by SKP2, a ubiquitin ligase (Xue et al. 2016) (Figure 6a). IFI27 is experiencing positive selection in primate species while SKP2 is not (p-value= 0.945), supporting the conclusion that IFI27 acts as a molecular link between the ubiquitin ligase and target viral protein and may therefore be participating in an evolutionary arms race between hepatitis C virus and primates.

Different primate alleles of IFI27 were predicted to show different functional phenotypes in relation to their ability to degrade NS5a in a co-transfection. Human and Chimpanzee IFI27 alleles are capable of degrading NS5a. The degradation of NS5a is visible at lower amounts of transfected IFI27 but is most stark when IFI27 is transfected at three times the amount of NS5a. Gibbon IFI27 appears to degrade NS5a to a lesser extent than Human IFI27, but at a similar level to Chimpanzee IFI27 (Figure 6b). These are preliminary results and need to be repeated in the future for more confidence. Additionally, other cloned primate alleles should be tested for degradation in a co-transfection.

Discussion

Although a ratio of the rate of non-synonymous (dN) to synonymous (dS) mutations greater than one generally indicates positive selection, it is important to note that many genes under positive selection do not have a whole gene average dN/dS value greater than one. Specific codons may be highly variable while others are highly conserved. This is most often the case, especially in closely related species such as primates. However, whole gene dN/dS values much greater than the average for random genes (0.2), or close to one, may indicate that there are enhanced non-synonymous mutation rates within the gene.

We did not test the likelihood of genes belonging to models M0 or M2 because all genes are likely to be better fit by a distribution that allows for codon dN/dS values to fall anywhere between 0 and 1, than at one or three dN/dS values, respectively. It is also important to note that the dN/dS values of the bin less than one in M2 and greater than one in M8 are not significantly different across the three categories (data not shown). This is likely due to the prevalence of both conserved and diverse sites within a gene regardless of category. Therefore, the difference observed in numbers of genes in each category that are under positive selection is attributable to the proportion of sites per gene.

The fact that interferon-stimulated genes have the highest proportion of alignments with the greatest number of sequences may be due to the fact that they have been highly studied. Genes that have been subject to many experiments are likely to have more available primate sequences than less studied random genes. Not all human genes have orthologous sequences in all primate species, thus sequences from all twenty species cannot be obtained for some genes. The quality of the gene sequences can also affect our ability to retain the predicted gene sequence. However, with the available sequence data I was able to perform a relatively inclusive analysis of the interferon response.

I found it surprising that the interferon-stimulated gene set was evolving more rapidly than the induction set. We had hypothesized that viruses have evolved mechanisms to halt the production of antiviral cellular states by antagonizing the initial expression of interferons or their stimulated genes rather than the individual proteins which produce antiviral states. However, it is possible that induction genes are under more evolutionary constraint in order to preserve specific functions in their respective signaling pathways. Interferon-stimulated genes may have more flexibility to obtain and tolerate mutations. It is also possible that there are more viral mechanisms to antagonize interferon-stimulated genes than induction genes, creating more opportunity for interferon-stimulated genes to evolve in response.

Although genes that induce the interferon response do not appear to have significantly elevated signatures of positive selection as a category, some individual genes do. Induction genes have been shown to be antagonized by viruses (Schulz and Mossman 2016) and may be experiencing adaptive evolution. Viruses often degrade proteins in this pathway directly or indirectly through ubiquitination (Schulz and Mossman 2016). Some viruses are capable of inhibiting phosphorylation, dimerization, and translocation of transcription factors (Schulz and Mossman 2016). The proper localization of pattern recognition receptors, which identify pathogen-associated molecular patterns in the extracellular space or the cytoplasm and signal for the expression of interferons, can also be obstructed (Liu et al. 2017). The results of this study could help to identify previously unknown areas of coevolution and host-pathogen interaction in both the induction and interferon-stimulated gene lists.

Several random genes had a high signal of positive selection. Some of these genes have functions involving the immune system (e.g. AZU1, CFLAR) as we did not control for this in the process of obtaining random genes. Random genes under positive selection could also contribute to the phenotypic traits which define our species such as: cognitive abilities (e.g. GRP), vocal organs, bipedalism, opposable thumbs, reproduction, or dietary adaptation (potentially DTD2 and NPC2) (Vallender and Lahn 2004). Genes of these functional categories would also be highly variable in primate species. Many of the random genes have predicted cellular functions, such as cell cycle regulation (GML), which may or may not be related to the larger biological processes listed above. Positive selection is not necessarily indicative of host-pathogen interactions.

Even when positive selection is due to host-pathogen interactions, it can be difficult to determine what host-virus interaction is driving positive selection as many genes are upregulated in response to more than one virus or type of virus. In concordance, signatures of positive selection in IFI27 may also be due to amino acid substitutions that mediate interactions of primate proteins with a virus other than hepatitis C. The fitness of a species is complex and governed by many genes. Even if some proteins experience functional changes, it may not necessarily affect the fitness of a species (Nozawa et al. 2009). However, I do not think that specific mutations would have swept through multiple unique primate populations if they did not improve the fitness of the host. It is important to pair bioinformatics and statistical methods with biological experiments in order to determine if a predicted site of positive selection is biologically significant.

Positive selection in IFI27 as a method of adaptation against viral infection in primates is inconclusive at this point in time. I have observed that Human, Chimpanzee, and Gibbon IFI27 alleles were able to degrade NS5a to some extent. Other primate alleles need to be tested to determine if there are natural variants of IFI27 that show different functional phenotypes in interacting with hepatitis C virus. The final step will include testing human constructs that have naturally occurring amino acids at predicted positive selection loci to determine if specific amino acid changes are capable of augmenting or diminishing degradation. It is possible that some moderately diverse codon sites within the gene mediate this interaction and were not predicted to be evolving rapidly. One nonsynonymous mutation at a specific codon site within a gene can be overshadowed by the number of synonymous mutations across an alignment of multiple species. However, codons which are predicted to be rapidly evolving can be a good indication of advantageous substitutions.

This study has provided a comprehensive analysis of evolution in the interferon response. Interferon-stimulated genes are presumably experiencing more adaptive evolution than genes whose proteins induce the response and ultimately lead to the production of antiviral proteins. Interferon-stimulated genes that are under positive selection are likely to be good indicators of genes that are necessary for effective control of viral infection. Primate alleles of these genes may provide a foundation for identifying amino acid sequences that produce viable proteins and are able to prevent pathogenesis of human viruses. These proteins could improve current antiviral treatments.

Although type I interferons are currently approved for treating HIV-1 (Doyle et al. 2015) and hepatitis infections (Schneider et al. 2014), the extent of interferon treatment is quite limited due to variable and short-lived efficacy (Doyle et al. 2015). Interferons can only confer protection to host cells if they are able to stimulate the expression of proteins which create antiviral environments. A subset of interferon-stimulated genes may be a better treatment for

viral infection because interferon signaling is antagonized by many viruses (Schulz and Mossman 2016) and has not readily adapted in primates. This is complemented by the fact that some combinations of interferon-stimulated genes have been shown to reduce viral replication by more than 90% (Schoggins and Rice 2011). The bioinformatics pipeline outlined here, paired with experimental validation, may identify genes and their subsequent proteins that could result in more successful clinical practice.

Acknowledgements

I would like to express my gratitude and appreciation to Alison Gilchrist, Dr. Nicholas Meyerson, and Vanessa Lamae Bauer DuMont for their help with troubleshooting experiments and critical readings of this manuscript. Alison Gilchrist collected data and ran PAML for many random and induction genes, as well as AIM2 (an interferon-stimulated gene). Alison also made the majority of Figure 4. Thank you to Dr. Sara Sawyer for her assistance and oversight of this project. A huge thank you to the whole Sawyer Laboratory for assisting me with experiments and data interpretation, patiently answering questions, and providing encouragement. Thank you to Dr. Brian DeDecker for his input and assistance. Thank you to Dr. Alison Vigers for her time and support.

Bibliography

1. Bowie A, Unterholzner L. 2008. Viral evasion and subversion of pattern-recognition receptor signaling. Nat Rev Immunol 8:911–922.
2. Čermák, Vladimír. 2016. MolBioTools; Free web app. Random Gene Set Generator.
3. Couto LB, Kolykhalov AA. 2006. Animal Models for HCV Study. In: Tan S-L, editor. Hepatitis C Viruses: Genomes and Molecular Biology. Norfolk (UK): Horizon Bioscience.
4. Kane M, Zang TM, Rihn SJ, Zhang F, Kueck T, Alim M, Schoggins J, Rice CM, Wilson SJ, Bieniasz PD. 2016. Identification of Interferon-Stimulated Genes with Antiretroviral Activity. Cell Host Microbe 20:392–405. doi:10.1016/j.chom.2016.08.005.
5. McBee RM, Rozmiarek SA, Meyerson NR, Rowley PA, Sawyer SL. 2015. The effect of species representation on the detection of positive selection in primate gene data sets. Mol. Biol. Evol. 32:1091– 1096.
6. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A Molecular Phylogeny of Living Primates. PLOS Genetics 7:e1001342.
7. Okonechnikov K, Golosova O, Fursov M, the UGENE team. Uniprro UGENE: a unified bioinformatics toolkit. Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091
8. Rusinova, I.; Forster, S.; Yu, S.; Kannan, A.; Masse, M.; Cumming, H.; Chapman, R.; Hertzog, P.J. INTERFEROME v2. 0: an updated database of annotated interferon-regulated genes. Nucleic Acids Research. 2013 January; 41 (database issue): D1040-D1046.
9. Schneider WM, Chevillotte MD, Rice CM. 2014. Interferon-Stimulated Genes: A Complex Web of Host Defenses. In: Littman DR, Yokoyama WM, editors. Annual Review of Immunology, Vol 32. Vol. 32. Palo Alto: Annual Reviews. p. 513–545.
10. Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, Rice CM. 2011. A diverse array of gene products are effectors of the type I interferon antiviral response. Nature 472:481–485. doi:10.1038/nature09907.
11. Schoggins JW, Rice CM. 2011. Interferon-stimulated genes and their antiviral effector functions. Curr. Opin. Virol. 1:519–525. doi:10.1016/j.coviro.2011.10.008.
12. Stark G, Kerr I, Williams B, Silverman R, Schreiber R. 1998. How cells respond to interferons. Ann Rev Biochem. 67:227–264.
13. Xue B, Yang D, Wang J, Xu Y, Wang X, Qin Y, Tian R, Chen S, Xie Q, Liu N, et al. 2016. ISG12a Restricts Hepatitis C Virus Infection through the Ubiquitination-Dependent Degradation Pathway. J Virol 90:6832–6845. doi:10.1128/JVI.00352-16.
14. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
15. Hornung, V. "Nucleic Acid Immune Sensors" *Immunity* (2014)
16. Kawai, T. and Akira, S. "The roles of TLRs, RLRs and NLRs in pathogen recognition." *International Immunology* (2009)

17. Shuai, K. and Liu, B. "Regulation of JAK–STAT signaling in the immune system." *Nature Reviews Immunology* (2003)
18. Barber, G.N. "Innate Immune DNA Sensing Pathways: STING, AIMII and the Regulation of Interferon Production and Inflammatory Responses." *Current Opinions in Immunology*(2011)
19. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. Trends Microbiol. NIH Pusblic Access; 2011;19: 286–94. doi: 10.1016/j.tim.2011.03.006
20. Elde N.C., Child S.J., Geballe A.P., Malik H.S. Protein kinase R reveals an evolutionary model for defeating viral mimicry. Nature. 2009; 457:485–489
21. Patel M.R., Loo Y.-M., Horner S.M., Gale M., Malik H.S. Convergent evolution of escape from hepaciviral antagonism in primates. PLoS Biol. 2012; 10:e1001282
22. Mikita Suyama, David Torrents, and Peer Bork (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609-W612.
23. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.
24. Eric J. Vallender, Bruce T. Lahn; Positive selection on the human genome, *Human Molecular Genetics*, Volume 13, Issue suppl_2, 1 October 2004, Pages R245–R254.
25. Masafumi Nozawa, Yoshiyuki Suzuki, Masatoshi Nei. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proceedings of the National Academy of Sciences Apr 2009, 106 (16) 6700-6705; DOI: 10.1073/pnas.0901855106
26. Schulz, K. S., & Mossman, K. L. (2016). Viral Evasion Strategies in Type I IFN Signaling – A Summary of Recent Developments. *Frontiers in Immunology*, *7*, 498. http://doi.org/10.3389/fimmu.2016.00498
27. Liu Y, Olagnier D, Lin R. 2017. Host and Viral Modulation of RIG-I-Mediated Antiviral Immunity. Front. Immunol. 7.
28. He Y, Staschke KA, Tan SL. 2006. HCV NS5A: A Multifunctional Regulator of Cellular Pathways and Virus Replication. In: Tan SL, editor. Hepatitis C Viruses: Genomes and Molecular Biology. Norfolk (UK): Horizon Bioscience. Chapter 9.
29. Doyle T, Goujon C, Malim M. 2015. HIV-1 and interferons: who's interfering with whom? Nat Rev Micro 13:403–413.