viruses

MDPI

Article

Viral Long-Term Evolutionary Strategies Favor Stability over Proliferation

Stéphane Aris-Brosou [1,2,*], Louis Parent [1] and Neke Ibeh [1,†]

[1] Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
* Correspondence: sarisbro@uottawa.ca
† Current address: Princess Margaret Cancer Centre, the University Health Network, Toronto, ON M5G 2M9, Canada.

Received: 30 May 2019; Accepted: 20 July 2019; Published: 24 July 2019

Abstract: Viruses are known to have some of the highest and most diverse mutation rates found in any biological replicator, with single-stranded (ss) RNA viruses evolving the fastest, and double-stranded (ds) DNA viruses having rates approaching those of bacteria. As mutation rates are tightly and negatively correlated with genome size, selection is a clear driver of viral evolution. However, the role of intragenomic interactions as drivers of viral evolution is still unclear. To understand how these two processes affect the long-term evolution of viruses infecting humans, we comprehensively analyzed ssRNA, ssDNA, dsRNA, and dsDNA viruses, to find which virus types and which functions show evidence for episodic diversifying selection and correlated evolution. We show that selection mostly affects single stranded viruses, that correlated evolution is more prevalent in DNA viruses, and that both processes, taken independently, mostly affect viral replication. However, the genes that are jointly affected by both processes are involved in key aspects of their life cycle, favoring viral stability over proliferation. We further show that both evolutionary processes are intimately linked at the amino acid level, which suggests that it is the joint action of selection and correlated evolution, and not just selection, that shapes the evolutionary trajectories of viruses—and possibly of their epidemiological potential.

Keywords: correlated evolution; positive selection; evolutionary strategy; functional analysis

1. Introduction

Humanity is regularly reminded of the epidemiological toll of viruses, in part due to recent and ongoing viral outbreaks of influenza [1], Ebola [2], and Zika [3]. Thanks to recent technological and analytical developments, it is now possible to elucidate, almost in real time [4], their epidemiological dynamics, and unravel their evolutionary dynamics [5]. However, while the evolutionary dynamics of RNA viruses are well documented [6], those of other viruses are not so well known, in particular in a unifying context including major viral types such as double-stranded (ds) and single-stranded (ss) DNA and RNA viruses.

To date, one of the most salient evolutionary features shared by all viruses is the existence of a negative correlation between mutation rate and genome size [6]. This is a critical result as it suggests that selection is driving the evolution of mutation rates, establishing a trade-off between mutational load and availability of adaptive mutations [7]. However, the processes driving the evolution of different types of viruses are multiple. As already argued, both ssDNA and ssRNA viruses share small genome sizes, high mutation rates, but also large effective population sizes, little to no gene duplication or recombination, and overlapping reading frames [6]. This last point suggests that a less frequently explored evolutionary process in viral studies, correlated evolution, could be as critical as positive selection. Correlated evolution happens when mutations at two locations in a genome occur one after the other, in a quick succession [8], repeatedly [9]. Typical examples include drug resistance mutations that have a fitness cost, and that require a second mutation to compensate for the first one [10], or tRNAs that require a specific base-pairing to maintain their secondary and tertiary structures, so that a mutation in the stem region necessitates a second mutation to restore the correct, functional, structure [11]. This process is of particular interest as correlated evolution can be underlain by epistasis, which occurs when the fitness effects of these two mutations are non-additive [12], as in the two examples above. As such, correlated evolution can help us understand the relationship between genotype and fitness, which is a key determinant of evolutionary trajectories [13]. To date, however, correlated evolution has only sporadically been investigated in viral evolution, and these rare instances only focused on ssRNA viruses. Indeed, recent work uncovered pervasive evidence for correlated evolution in influenza viruses [14,15], and both the Zika [16] and the Ebola viruses [17]. Intriguingly, in this latter case (Ebola), evidence was found that sites evolving in a correlated manner could also be under positive selection—bearing the question as to how frequently these two processes, correlated evolution and positive selection, occur, possibly jointly, and if this co-occurrence is limited to ssRNA viruses, or can be generalized to all viruses.

To better understand the role of correlated evolution and positive selection in the evolutionary dynamics of viruses infecting humans, we constructed a nearly exhaustive viral data set spanning all dsDNA, dsRNA, ssRNA, and ssDNA viruses deposited in GenBank (as of August 2017), and conducted an extensive survey of correlated evolution and diversifying selection in these viruses, asking more specifically about the prevalence of these two processes in each viral type, independently or jointly, with the specific hypothesis that the genes affected by both processes encode functions that are most critical to each viral life cycle.

2. Materials and Methods

2.1. Data Retrieval

Lists of dsDNA, dsRNA, ssRNA, and ssDNA viruses infecting humans were retrieved from the viruSITE database [18] in August 2017 (Tables S1–S4); subtypes/genotypes/clades were treated as independent data sets. Although some of these viruses could be segmented or not, with circular or linear genome, with positive or negative strands, and with or without overlapping reading frames, accounting for these structural features would have led to smaller and smaller data sets, precluding any statistical analysis, so that the data were not split beyond viral type. Each list contained the virus names, the length of their genome, their number of protein-coding genes (CDS's), and was associated with a reference coding sequence (see `query_sequences.zip` at [19]). In order to obtain corresponding sequence alignments of orthologous genes, BLASTn searches were performed on a custom database limited to viral genes present in the National Center for Biotechnology Information nucleotide database with blast-2.6.0+. For this, all gbvrl*.seq.gz files were downloaded from reference [20], while querying viruSITE, and were concatenated into a single GenBank file, then converted into a FASTA file with `readseq` to specifically extract CDS's [21]. This was done to avoid retrieving 5′ and 3′ untranslated regions that would cause problems to the downstream codon analyses. BLASTn searches were performed for each viruSITE viral sequence with a stringent E-value threshold of 10^{-100}, keeping a maximum of 100 sequences with at least 80% coverage with each query; this ensured that subtype/genotype/clade boundaries were not crossed. As the viruses retrieved from the viruSITE also included viruses that require a vector (e.g., arboviruses such as the Dengue and Yellow fever viruses), or viruses that circulate in non-human hosts but that can lead to zoonoses (e.g., the Camel alphacoronavirus leading to MERS [22,23]), these stringent thresholds also ensured that sequences contained in each alignment mostly came from a single host. Only data sets with at least 20 hits were kept for downstream analyses. Sequences corresponding to each accession number were retrieved from the FASTA file obtained with `readseq` [21].

Because this file contained partial sequences, each set of retrieved sequences was first aligned with `Muscle` 3.8.31 [24]. Each alignment was then quality checked ensuring that (i) its length is a multiple of three, (ii) it starts with an ATG and stops with a stop codon. Alignments failing at least one condition were discarded. Within each alignment, mean numbers of indels $\hat{n}_{indels}$ were computed for each sequence, and those containing number of indels $\geq \hat{n}_{indels} + 1$SEM (standard error of the mean) were eliminated. The remaining nucleotide sequences were then re-aligned with TranslatorX [25], at the amino acid level, using the `Muscle` aligner and their heuristics to determine the correct reading frame. Alignment files were then cleaned-up with `Gblocks` 0.91b at the codon level using the stringent default settings [26]. Both trees and alignments are available at [19].

To obtain gene annotations, Gene Ontology (GO) terms were retrieved from gene sequences with HMMER2GO ([27]), relying on the Hidden Markov Models in [28] (ver. 23 Febuary 2017). This is equivalent to performing a Blast2GO search [29], but without the limitations of proprietary software. Individual mapping files coming from our four viral types (ssRNA, dsRNA, ssDNA, and dsDNA viruses) were then merged to create a custom annotation file, used for GO term enrichment testing with topGO [30], based on Fisher's exact test. The reference gene list was always the entire set of genes within each viral type.

2.2. Phylogenetic Analyses

Phylogenetic trees were reconstructed with `FastTree` 2.1.7 [31] under the GTR+Γ model of evolution [32]; note that `FastTree` was recompiled locally to use double-precision arithmetics, as recommended by its authors to estimate very short branch lengths accurately. Those with a nonzero tree length were midpoint rerooted using the phytools package ver. 0.4–60 [33] in R ver. 3.2.3 [34].

Patterns of correlated evolution among sites were identified with the Bayesian graphical model (BGM) implemented in SpiderMonkey [35] (SM), which is part of HyPhy ver. 2.3.3 [36]. Default HyPhy scripts were slightly modified to read in codon data, which were used to reconstruct mutational paths under the MG94 × HKY85 substitution model [37] at nonsynonymous sites along each branch of the estimated trees. Ambiguous reconstructions were resolved by considering all possible resolutions and averaging them. These reconstructed mutational paths were then recoded as a binary matrix, with rows corresponding to branches and columns to each site of the alignment. The BGM was then used to identify the pairs of sites that exhibit correlated patterns of nonsynonymous substitutions according to their posterior probability, estimated with a Markov chain Monte Carlo sampler that was run for 10^5 steps, with a burn-in period of 10,000 steps sampling every 1000 steps for inference [16].

Patterns of episodic selection were identified based on the Mixed Effects Model of Evolution [38] (MEME), also as implemented in HyPhy. The default script from ver. 2.2.6 was used, still with the 2.3.3 HyPhy engine, to infer nonsynonymous to synonymous rate ratios ω assuming that these rates can vary across lineages and among sites. In this implementation, two categories of sites were assumed, those for which $\omega_{neg} \leq 1$, in proportion p, and those for which $\omega_{pos} > 1$ that are under positive selection, in proportion $1 - p$. Evidence for selection was derived by means of a likelihood ratio test between this model, and a null model where ω_{pos} was constrained to take its value between 0 and 1. Linear models were fitted through robust regressions [39]. All R scripts and HyPhy source files are available from reference [19] (file "API_scripts.zip," in "data").

3. Results and Discussion

First, in order to better understand the genomic characteristics of the four types of viruses known to infect humans, dsDNA ($n = 94$), dsRNA ($n = 15$), ssRNA ($n = 354$), and ssDNA ($n = 84$), and understand how these characteristics can impact the evolutionary dynamics of these viruses (Methods; Figure S1), we examined the distribution of four of their genomic features. While dsDNA viruses had the largest and the most variable genome lengths, ssDNA were the smallest genomes by at least an order of magnitude, with both dsRNA and ssRNA exhibiting intermediate sizes (Figure 1a). Because of the compact structure of these genomes, these differences were also reflected in the number of genes, protein-coding genes, and RNAs encoded by the genomes of these viruses (Figure S2). In turn, these characteristics suggest that, with their large genomes, dsDNA viruses potentially have a higher degree of functional redundancy than ssDNA or even ssRNA viruses, possibly due to duplication events [40], and thereby be under less stringent selective pressures than viruses with smaller genomes. On the other hand, the extent of correlated evolution and its interaction with selection is more difficult to predict.

To address these questions, we first tested for the presence of episodic diversifying selection in each viral type. Altogether, we found extensive differences among all four viral types in terms of the number of genes under selection (Figure 1b; $X^2 = 99.61$, $df = 3$, $p < 2.2 \times 10^{-16}$), and that single stranded viruses were more subject to selection than double stranded viruses ($X^2 = 95.14$, $df = 1$, $p < 2.2 \times 10^{-16}$). Indeed, and contra our original hypothesis, there were no differences between dsDNA and dsRNA viruses ($X^2 = 0.26$, $df = 1$, $p = 0.6110$), or between ssDNA and ssRNA viruses in terms of prevalence of diversifying selection ($X^2 = 2.07$, $df = 1$, $p = 0.1503$). Note that these differences cannot be attributed to genetic diversity, as dsDNA and dsRNA, which have similar levels of selection, have however different levels of diversity (Figure S3). However, it is unlikely that "strandedness" (single vs. double stranded genetic material) alone drives selection, even if greater instability can be postulated in single-stranded nucleic acids [41]. Indeed, strandedness is negatively correlated with genome size ($t = -4.98$, $df = 108.35$, $p = 2.43 \times 10^{-6}$), which is itself correlated with mutation rates [42,43]. As a result, episodic diversifying selection is mostly driven by structural aspects of viral genomes, which condition mutation rates across all virus types [44].

To understand if similar aspects drive intragenic correlated evolution, we counted in the same way the number of genes for which we could find evidence for interactions. Here, again, we found extensive differences among all four viral types (Figure 1c; $X^2 = 29.79$, $df = 3$, $p < 1.5 \times 10^{-16}$), with no difference between dsDNA and ssDNA viruses ($X^2 = 0.0005$, $df = 1$, $p = 0.9827$), or between dsRNA and ssRNA viruses ($X^2 = 0.0001$, $df = 3$, $p = 0.9903$). Again, diversity is not driving these differences, as both dsDNA/ssDNA and dsRNA/ssRNA have significantly different levels of diversity (Figure S3). As a result, the prevalence of correlated evolution seems to be mostly driven by the nature of viral genetic material. At least two processes can underpin correlated evolution: linkage and epistasis [12]. As recombination is pervasive in some dsDNA viruses [45], linkage alone may not explain the high prevalence of correlated evolution in DNA viruses (close to 40%: Figure 1c). Rather, this pattern suggests that intragenic constraints are higher in DNA viruses for unknown structural reasons, maybe due to protein structure [46], or in the same way that recombination in RNA viruses may be driven by mechanistic constraints associated with genome structures and viral life cycles [47].

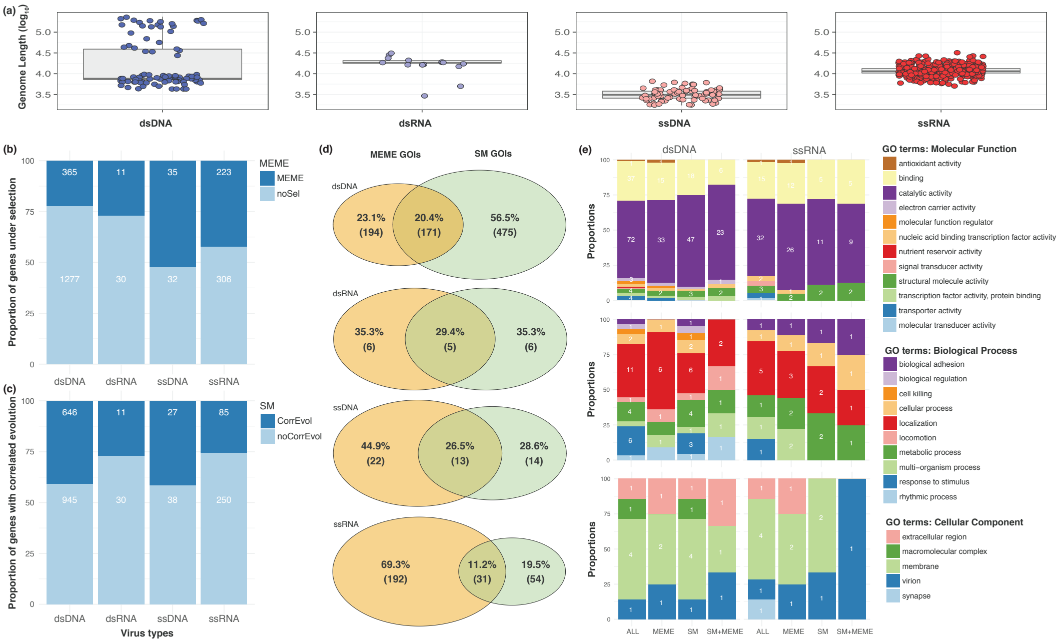

Figure 1. Modes of evolution of viruses. Modes of evolution of viruses. (**a**) distribution of genome size, on a $\log_{10}$ scale ssRNA (blue), ssDNA (purple), dsRNA (orange), and dsDNA (red); (**b**) proportion of genes detected to be under diversifying selection (top, dark hues); actual numbers of genes are shown within each column; (**c**) proportion of genes detected to be evolving in a correlated manner (top, dark hues); actual numbers of genes are shown within each column; (**d**) Venn diagrams showing the Genes of Interest (GOIs) either under selection (MEME GOIs: detected in the Mixed Effects Model of Evolution analyses), or evolving in a correlated manner (SM GOIs: detected in the SpiderMonkey analyses), or both (intersect) for each virus type; (**e**) Gene Ontology (GO) enrichment tests for the genes that are both under selection and evolving in a correlated manner for their Molecular Function (top), Biological Processes (middle), and Cellular Component (bottom), all at level 2 of the ontology.

While previous work showed that both diversifying selection and correlated evolution can affect the same gene and even the same site in a viral genome such as Ebola's [17], an ssRNA virus, the generality of this association is still unknown. At the gene level, we found strong heterogeneity among all four viral types for gene numbers showing evidence for selection, correlated evolution or both ($X^2 = 206.54$, $df = 6$, $p < 2.2 \times 10^{-16}$). Surprisingly, ssRNA viruses are those that show the least overlap between selection and correlated evolution, with only 11% of the genes evolving under both mechanisms (Figure 1d). Patterns are, however, much more difficult to extract here, mostly because the number of genes involved becomes quite small, in particular for the virus type with most overlap, the dsRNA viruses (Figure 1d: 29.4%, i.e., five genes).

To assess the extent to which some of these differences at the gene level are functionally driven, we extracted the Gene Ontology (GO) annotations, or GO terms, associated with the genes analyzed, as well as those under selection, correlated evolution, or both—focusing exclusively on the virus types for which we had the largest samples sizes, dsDNA and ssRNA viruses (Figure 1e). Each of the three parts of the ontology (Molecular Function [MF], Biological Process [BP], and Cellular Component [CC]) was first limited to the second level of GO in order to derive a high-level interpretation (low-level descriptions are shown in Tables S1–S3). Figure 1e shows that GO terms related to catalytic activity (MF), involved in the establishment or maintenance of a certain location (BP) at the membrane level (CC) are predominantly present in both dsDNA and ssRNA genomes. However, only a subset of these GO terms is mostly under episodic diversifying selection: helicase activity/binding (MF) and replication (BP) in the host cell nucleus (CC) in dsDNA, while it is mostly peptidase activity and methylation on the viral envelope in ssRNA viruses (Table S1; $p < 0.01$). In spite of these differences, we note that episodic diversifying selection mostly affects genes involved in viral replication (Table S1).

Genes that are affected by correlated evolution include: helicase activity and binding at the interface of multiple compartments in dsDNA viruses, or transferase activity and protein modifications on the envelope in ssRNA viruses (Table S2). Again, most of these functions and processes are involved in viral replication (Table S2). At the intersection of these evolutionary processes however, the genes that are jointly affected by selection and correlated evolution are involved in structural integrity and assembly within or outside a cell at the capsid level for dsDNA viruses, or in interacting with host cell surface via antigen activity on the viral envelope for ssRNA viruses—functions and processes that are mostly involved in "viral stability" (cell entry, integrity, assembly, immune escape; Table S3). This suggests that, despite key differences in life history strategies adopted by dsDNA and ssRNA, there is a certain unity across viral types, where genes involved in replication are mostly under either selection or correlated evolution, while those involved in viral stability are mostly affected by both evolutionary processes—as has been shown for the influenza [10] and the Ebola [17] viruses (both cases involved a glycoprotein required for cell entry). Because these two evolutionary processes are required by epistasis, it is possible that the genes involved in viral stability are the most likely to show evidence for non-additive fitness effects—and hence shape viral fitness landscapes. This tension between replication and stability is evocative of the existence of trade-offs between capsid stability and proliferation within a host [48], or fecundity and lifespan [49] in dsDNA viruses, which supports the idea that it is the joint action of selection and of correlated evolution that shapes viral life histories.

The previous analyses were at the gene level. One outstanding question is whether this relationship between selection and correlated evolution also holds at the amino acid level that is, if the amino acids under selection are also involved in correlated evolution. This question was addressed in two different ways, all virus types confounded (in order to have larger sample sizes), by focusing on one process at a time, and finding at what point evidence supporting the second process becomes significant. First, we searched the list of pairs of sites evolving in a correlated manner (the SM sites) to see if at least one pair member was potentially evolving under episodic diversifying selection (the MEME sites), irrespective of its probability of being a MEME site. For this, we identified pairs of SM sites, that is those with a posterior probability $\geq$0.95, and plotted their posterior probability as a function of the $-\log_{10}$ probability of each pair member to be under selection (Figure 2a). Both the least-square ($\hat{\rho}_{\text{slope}} = 0.005$, $t = 3.90$, $p = 0.0001$) and the robust regressions ($\hat{\rho}_{\text{slope}} = 0.005$, $t = 2.62$, $p = 0.0089$) have positive and significant slopes, hereby demonstrating the existence of a relationship between correlated evolution and episodic diversifying selection at the amino acid level—even if most of the SM sites show weak evidence of selection (at $p \leq 0.01$, gray broken line in Figure 2a). Thus, to validate the existence of this relationship, we then took the list of MEME sites, and searched the list of SM sites to see if at least one pair member was potentially a MEME site, irrespective of its posterior probability of being in an SM site pair. For this, we identified the MEME sites, that is those with a p-value ≤ 0.01 ($\geq$2 on a $-\log_{10}$ scale), and plotted this (on the x-scale) vs. their posterior probability of being in an SM pair (on the y-scale; Figure 2b). Both the least-square ($\hat{\rho}_{\text{slope}} = 0.026$, $t = 3.64$, $p = 0.0003$) and the robust regressions ($\hat{\rho}_{\text{slope}} = 0.031$, $t = 3.37$, $p = 0.0008$) have positive and significant slopes, further confirming the existence of a relationship between correlated evolution and episodic diversifying selection at the amino acid site level. Note, however, that this latter analysis is more informative than the former, as the density also shows that most of the sites under selection are involved in weak interactions. This is intriguingly reminiscent of the involvement of weakly interacting pairs of sites in severe outbreaks or pandemics [16].

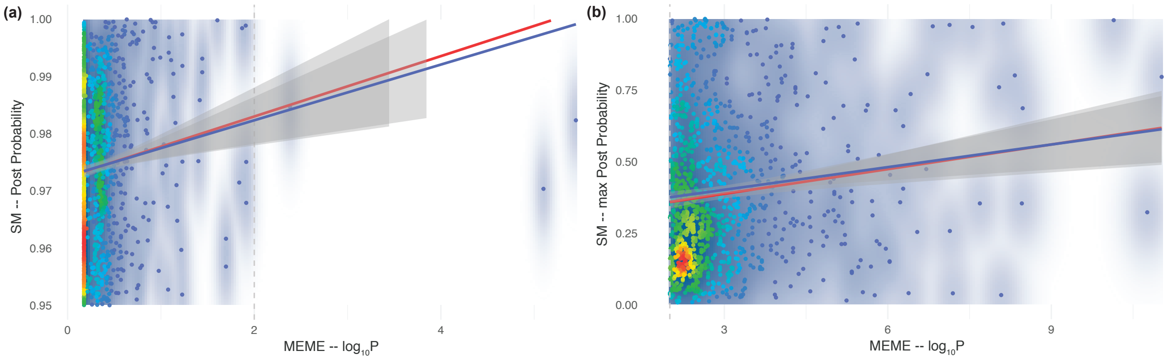

Figure 2. Strength of evidence for sites under both selection (MEME) and correlated evolution (SM). (**a**) SM sites that are also under episodic diversifying selection; (**b**) MEME sites that are also evolving in a correlated manner. In both cases, the posterior probability of a site evolving in a correlated manner is plotted as a function of the *p*-value of a site being under selection (on a $-\log_{10}$ scale). Regular least-square regressions are shown in blue; robust regressions are shown in red; 95% confidence envelopes are shaded in grey. See text for statistical details. Density scale: from low (cool colors) to high (warm colors).

4. Conclusions

Altogether, we showed that episodic diversifying selection is mostly found in single stranded viruses, while correlated evolution is more prevalent in DNA viruses. More critically, we also showed that the genes affected by each process, when acting independently, are involved in viral replication. However, the genes that are jointly affected by both processes are mostly involved in viral stability (cell entry, integrity, assembly, immune escape), and that the same amino acid sites tend to be affected by both processes. In retrospect, this tight relationship between selection and correlated evolution may not be surprising, as correlated evolution can be underlain by epistasis [12,14], which directly involves selection (in a non-additive way). Epistasis being a key determinant of fitness landscapes, and hence of evolutionary trajectories [13], our results suggest that, in the long-term, both processes jointly shape the life history of viruses, favoring stability over proliferation. If so, analyzing viral evolution in the joint light of selection and correlated evolution might help us better predict how viruses that affect humans might evolve [50,51]—as predicting their evolution [52] and epidemiology [53] has a long history fraught with mixed success.

We note however that we neglected some aspects of viral structure: indeed, viruses can be segmented or not, with a circular or linear genome, with positive or negative strands, overlapping reading frames, complications that we could not consider here due to the resulting small sample sizes, even if these factors can impact the mode of evolution of viruses [6]. Future work should however strive to address these limitations. We also neglected the population genetics context in which different viruses evolve, a context that can often be correlated to structural constraints [6,42]. Furthermore, as we solely focused on intragenic interactions, and not intergenic or higher order correlations, it is not impossible that we missed higher-level constraints affecting viral evolution. In particular, it is possible that intergenic correlations reveal the nature of trade-offs shaping the life history strategies in viruses [48]. Future modeling [54] and empirical [55] work should probably focus on elucidating not only these constraints, but also the theoretical basis connecting correlated evolution, if not epistasis, to selection in shaping the evolutionary strategies of biological replicators, as current evidence linking these processes is currently limited to the influenza [10] and the Ebola [17] viruses.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/8/677/s1, Figure S1: Overview of the experimental procedures, Figure S2: Physical characteristics of the four types of viruses, Figure S3: Viral diversity across viral types, Table S1: GO enrichment tests for all the MEME genes of interests, Table S2: GO enrichment tests for all the SM genes of interests, Table S3: GO enrichment tests for all the SM + MEME genes of interests.

Author Contributions: Conceptualization, S.A.-B.; methodology, S.A.-B., L.P. and N.I.; software, S.A.-B., L.P.; validation, S.A.-B. and L.P.; formal analysis, S.A.-B. and L.P.; writing—original draft preparation, S.A.-B. and L.P.; writing—review and editing, S.A.-B., L.P. and N.I.; supervision, S.A.-B. and N.I.; funding acquisition, S.A.-B.

Funding: This research was funded by the Natural Sciences Research Council of Canada (S.A.-B.), and by the University of Ottawa (L.P., N.I.). The Article Processing Charges was funded by the Natural Sciences Research Council of Canada (S.A.-B.).

Acknowledgments: We thank three anonymous reviewers for providing us with constructive comments, as well as Compute Canada and Ontario's Centre for Advanced Computing for giving us access to their servers.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BGM	Bayesian graphical model
GO	Gene ontology
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
MEME	Mixed effects model of evolution (analysis of positive selection)
SEM	Standard error of the mean
SM	SpiderMonkey (analysis of correlated evolution)
ssDNA	Single stranded DNA
ssRNA	Single stranded RNA

References

1. Smith, G.J.D.; Vijaykrishna, D.; Bahl, J.; Lycett, S.J.; Worobey, M.; Pybus, O.G.; Ma, S.K.; Cheung, C.L.; Raghwani, J.; Bhatt, S.; et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **2009**, *459*, 1122–1125. [CrossRef] [PubMed]
2. Gire, S.K.; Goba, A.; Andersen, K.G.; Sealfon, R.S.G.; Park, D.J.; Kanneh, L.; Jalloh, S.; Momoh, M.; Fullah, M.; Dudas, G.; et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **2014**, *345*, 1369–1372. [CrossRef] [PubMed]
3. Faria, N.R.; Azevedo, R.D.S.D.S.; Kraemer, M.U.G.; Souza, R.; Cunha, M.S.; Hill, S.C.; Thézé, J.; Bonsall, M.B.; Bowden, T.A.; Rissanen, I.; et al. Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **2016**, *352*, 345–349. [CrossRef] [PubMed]
4. Faria, N.R.; Sabino, E.C.; Nunes, M.R.T.; Alcantara, L.C.J.; Loman, N.J.; Pybus, O.G. Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **2016**, *8*. [CrossRef] [PubMed]
5. Grenfell, B.T.; Pybus, O.G.; Gog, J.R.; Wood, J.L.N.; Daly, J.M.; Mumford, J.A.; Holmes, E.C. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **2004**, *303*, 327–332. [CrossRef] [PubMed]
6. Holmes, E.C. *The Evolution and Emergence of RNA Viruses*; Oxford Series in Ecology and Evolution; Oxford University Press: Oxford, UK, 2009.
7. Holmes, E.C. What does virus evolution tell us about virus origins? *J. Virol.* **2011**, *85*, 5247–5251. [CrossRef] [PubMed]
8. Kryazhimskiy, S.; Dushoff, J.; Bazykin, G.A.; Plotkin, J.B. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **2011**, *7*, e1001301. [CrossRef] [PubMed]
9. Maddison, W.P.; FitzJohn, R.G. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* **2015**, *64*, 127–136. [CrossRef] [PubMed]
10. Gong, L.I.; Suchard, M.A.; Bloom, J.D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2013**, *2*, e00631. [CrossRef]
11. Meer, M.V.; Kondrashov, A.S.; Artzy-Randrup, Y.; Kondrashov, F.A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **2010**, *464*, 279–282. [CrossRef]
12. Dench, J.; Hinz, A.; Aris-Brosou, S.; Kassen, R. The idiosyncratic drivers of correlated evolution. *bioRxiv* **2019**, *2019*, 474536.

13. Li, C.; Qian, W.; Maclean, C.J.; Zhang, J. The fitness landscape of a tRNA gene. *Science* **2016**, *352*, 837–840. [CrossRef] [PubMed]
14. Nshogozabahizi, J.C.; Dench, J.; Aris-Brosou, S. Widespread historical contingency in Influenza viruses. *Genetics* **2017**, *205*, 409–420. [CrossRef] [PubMed]
15. Lyons, D.M.; Lauring, A.S. Mutation and epistasis in Influenza virus evolution. *Viruses* **2018**, *10*. [CrossRef] [PubMed]
16. Aris-Brosou, S.; Ibeh, N.; Noël, J. Viral outbreaks involve destabilized evolutionary networks: Evidence from Ebola, Influenza and Zika. *Sci. Rep.* **2017**, *7*, 11881. [CrossRef] [PubMed]
17. Ibeh, N.; Nshogozabahizi, J.C.; Aris-Brosou, S. Both epistasis and diversifying selection drive the structural evolution of the Ebola virus glycoprotein mucin-like domain. *J. Virol.* **2016**, *90*, 5475–5484. [CrossRef] [PubMed]
18. Stano, M.; Beke, G.; Klucar, L. viruSITE-integrated database for viral genomics. *Database* **2016**, *2016*. [CrossRef]
19. Aris-Brosou, S. Available online: https://github.com/sarisbro (accessed on 30 May 2019).
20. NCBI. Available online: https://ftp.ncbi.nih.gov/genbank/ (accessed on 30 May 2019).
21. Gilbert, D. Sequence file format conversion with command-line readseq. *Curr. Protoc. Bioinform.* **2003**. [CrossRef]
22. van Boheemen, S.; de Graaf, M.; Lauber, C.; Bestebroer, T.M.; Raj, V.S.; Zaki, A.M.; Osterhaus, A.D.M.E.; Haagmans, B.L.; Gorbalenya, A.E.; Snijder, E.J.; et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio* **2012**, *3*. [CrossRef]
23. Zaki, A.M.; van Boheemen, S.; Bestebroer, T.M.; Osterhaus, A.D.M.E.; Fouchier, R.A.M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **2012**, *367*, 1814–1820. [CrossRef]
24. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
25. Abascal, F.; Zardoya, R.; Telford, M.J. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **2010**, *38*, W7–W13. [CrossRef] [PubMed]
26. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552. [CrossRef] [PubMed]
27. Staton, E. Available online: https://github.com/sestaton/HMMER2GO (accessed on 30 May 2019).
28. Pfam. Available online: http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz (accessed on 30 May 2019).
29. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]
30. Alexa, A.; Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconduct. Improv.* **2009**, *27*, 1–26.
31. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef] [PubMed]
32. Aris-Brosou, S.; Rodrigue, N. The essentials of computational molecular evolution. *Methods Mol. Biol.* **2012**, *855*, 111–152. [CrossRef]
33. Revell, L.J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **2012**, *3*, 217–223. [CrossRef]
34. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016.
35. Poon, A.F.Y.; Lewis, F.I.; Frost, S.D.W.; Kosakovsky Pond, S.L. Spidermonkey: Rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* **2008**, *24*, 1949–1950. [CrossRef]
36. Pond, S.L.K.; Frost, S.D.W.; Muse, S.V. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **2005**, *21*, 676–679. [CrossRef]
37. Kosakovsky Pond, S.L.; Frost, S.D.W. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **2005**, *22*, 1208–1222. [CrossRef] [PubMed]
38. Murrell, B.; Wertheim, J.O.; Moola, S.; Weighill, T.; Scheffler, K.; Kosakovsky Pond, S.L. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **2012**, *8*, e1002764. [CrossRef] [PubMed]
39. Yohai, V.J.; Stahel, W.A.; Zamar, R.H. A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics*; Springer: Berlin, Germany, 1991; pp. 365–374.

40. Gao, Y.; Zhao, H.; Jin, Y.; Xu, X.; Han, G.Z. Extent and evolution of gene duplication in DNA viruses. *Virus Res.* **2017**, *240*, 161–165. [CrossRef] [PubMed]
41. Frederico, L.A.; Kunkel, T.A.; Shaw, B.R. A sensitive genetic assay for the detection of cytosine deamination: Determination of rate constants and the activation energy. *Biochemistry* **1990**, *29*, 2532–2537. [CrossRef] [PubMed]
42. Lynch, M. *The Origins of Genome Architecture*; Sinauer Associates: Sunderland, MA, USA, 2007.
43. Duffy, S. Why are RNA virus mutation rates so damn high? *PLoS Biol.* **2018**, *16*, e3000003. [CrossRef] [PubMed]
44. Sanjuán, R. From molecular genetics to phylodynamics: Evolutionary relevance of mutation rates across viruses. *PLoS Pathog.* **2012**, *8*, e1002685. [CrossRef]
45. Robinson, C.M.; Seto, D.; Jones, M.S.; Dyer, D.W.; Chodosh, J. Molecular evolution of human species D adenoviruses. *Infect. Genet. Evol.* **2011**, *11*, 1208–1217. [CrossRef] [PubMed]
46. Woo, J.; Robertson, D.L.; Lovell, S.C. Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants. *Virology* **2014**, *454–455*, 34–39. [CrossRef]
47. Simon-Loriere, E.; Holmes, E.C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **2011**, *9*, 617–626. [CrossRef]
48. De Paepe, M.; Taddei, F. Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol.* **2006**, *4*, e193. [CrossRef]
49. García-Villada, L.; Drake, J.W. Experimental selection reveals a trade-off between fecundity and lifespan in the coliphage Qß. *Open Biol.* **2013**, *3*, 130043. [CrossRef] [PubMed]
50. Weinreich, D.M.; Delaney, N.F.; Depristo, M.A.; Hartl, D.L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **2006**, *312*, 111–114. [CrossRef] [PubMed]
51. Pedruzzi, G.; Barlukova, A.; Rouzine, I.M. Evolutionary footprint of epistasis. *PLoS Comput. Biol.* **2018**, *14*, e1006426. [CrossRef] [PubMed]
52. Sandie, R.; Aris-Brosou, S. Predicting the emergence of H3N2 influenza viruses reveals contrasted modes of evolution of HA and NA antigens. *J. Mol. Evol.* **2014**, *78*, 1–12. [CrossRef] [PubMed]
53. Ben-Nun, M.; Riley, P.; Turtle, J.; Bacon, D.P.; Riley, S. Forecasting national and regional influenza-like illness for the USA. *PLoS Comput. Biol.* **2019**, *15*, e1007013. [CrossRef] [PubMed]
54. Neverov, A.D.; Kryazhimskiy, S.; Plotkin, J.B.; Bazykin, G.A. Coordinated evolution of Influenza A surface proteins. *PLoS Genet.* **2015**, *11*, e1005404. [CrossRef] [PubMed]
55. Ashenberg, O.; Padmakumar, J.; Doud, M.B.; Bloom, J.D. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. *PLoS Pathog.* **2017**, *13*, e1006288. [CrossRef]