Viral outbreaks involve destabilised viruses: evidence from Ebola, Influenza and Zika

Stéphane Aris-Brosou[1,2,⋆], **Neke Ibeh**[1] **and Jessica Noël**[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

⋆Author for correspondence: `sarisbro@uottawa.ca`

Running Head: Destabilised viruses

Statistics

Abstract length: 99 words

Total length: 2,123 words

Abstract

Although viral outbreaks and pandemics have plagued humans and other organisms for billions of years, such events are not only still impossible to predict, but the ultimate reasons why outbreaks happen are not understood. Based on recent viral outbreaks and pandemics (Ebola, Zika and Influenza), we searched for a common denominator to these events, positing that the genome of outbreak viruses is far from an evolutionary equilibrium, which is ultimately maintained by a dense network of correlated substitutions. We show here that genes of outbreak viruses are characterised by destabilised correlation networks, a result that might improve outbreak surveillance.

Keywords: Ebola virus, Influenza virus, Zika virus, outbreak, pandemic, correlated evolution

1 Introduction

Viruses are engaged in a form of arms race with their host, in which each endeavours to outpace the other [1]. Regular epidemics can therefore be seen as an equilibrium situation, where neither the virus nor the host populations are at risk of extinction. Such a stable evolutionary strategy can however be broken when the virus becomes extremely virulent, which can lead to a severe outbreak or even a pandemic. Recent history is rich in such instances with an Ebola virus outbreak in 2014 [2], a Zika outbreak since 2015 [3], and an Influenza pandemic in 2009 [4]. Despite all of these instances, we still do not know what causes outbreaks and pandemics. The question we address here is whether we can find commonalities to these three outbreaks, while still setting them apart from non-pandemic or "regular" viruses.

As theory tells us that regular epidemics are the result of a dynamic equilibrium [5], we posit that outbreaks are the results of a disequilibrium, not just in their population dynamics but also at the genomic level. More specifically, we suggest that outbreaks involve destabilised viral genomes, where evolutionary stability is maintained by compensatory mutations, that can be epistatic or not, but that result in signals of correlated evolution. We predict that such signals are weakened during an outbreak. As these signals often lead to complex networks of interactions [6, 7], we test how the structure of these correlation networks is affected during an outbreak. We show that during an outbreak, viral genes are destabilised.

2 Material and methods

(a) Sequence retrieval

Nucleotide sequences were retrieved for three viruses: Ebola, Zika, and Influenza A, for select protein-coding genes, chosen because they represent the most sequenced genes for each of these viruses. All sequences were downloaded in May 2016 (table S1).

For Ebola, the virion spike glycoprotein precursor, GP, was retrieved as follows. A GP sequence (KX121421) was drawn at random from the 2014 strain used previously [7] and was employed as a query for a BLASTn search [8] at the National Center for Biotechnology Information. A conservative E-value threshold of 0 ($E < 10^{-500}$) was used, which led to 1,181 accession numbers. As most of these accession numbers correspond to full genomes, while only GP is of interest, we (i) retrieved all corresponding GenBank files, (ii) extracted coding sequences with ReadSeq [9] of all genes, (iii) concatenated the corresponding FASTA files into a single file, (iv) which was then used to format a sequence database for local BLASTn searches, and (v) used GP from KX121421 in a second round of BLASTn searches ($E < 10^{-250}$, coverage $> 75\%$).

In the case of Zika, sequences of 252 complete genomes were retrieved from the Virus Pathogen Resource (`www.viprbrc.org`). The RNA-dependent RNA polymerase NS5 was specifically extracted by performing local BLASTn searches as described above.

Full-length Influenza A sequences were retrieved directly from the Influenza Virus Resource [10]. Only H1N1 sequences circulating in humans for the hemagglutinin (HA) and neuraminidase (NA) genes were downloaded. Two types of data sets were constructed: one containing pandemic and non-pandemic sequences circulating in 2009, the pandemic year, and one containing pandemic sequences circulating from August 1 to July 31 of each season in the Northern temperate region between 2009/2010 and 2015/2016 (seven

seasons in total). Only unique sequences were retrieved.

(b) Phylogenetic analyses

Sequences were all aligned with Muscle [11] with fastest options (-maxiters 1 -diags). Alignments were visually inspected with AliView [12] to remove rogue sequences and sequencing errors. Phylogenetic trees were inferred by maximum likelihood under the General Time-Reversible model with amongst-site rate variation [13] with FastTree [14]. As outbreak sequences (Ebola and Zika viruses) cluster away from non-pandemic sequences, we used the `subtreeplot()` function in APE [15] to retrieve accession numbers of pandemic sequences and hence separate them from non-pandemic sequences with minimal manual input. FastTree was used a second time to estimate phylogenetic trees of the subset alignments, with the same settings as above.

(c) Network analyses of correlated sites

Amino acid positions ("sites") that evolve in a correlated manner were identified with the Bayesian graphical model (BGM) in SpiderMonkey [16] as implemented in HyPhy [17]. Briefly, ancestral mutational paths were first reconstructed under the MG94×HKY85 substitution model [18] along each branch of the tree estimated above at non-synonymous sites. These reconstructions were recoded as a binary matrix in which each row corresponds to a branch and each column to a site of the alignment. A BGM was then employed to identify which pairs of sites exhibit correlated patterns of substitutions. Each node of the BGM represents a site and the presence of an edge indicates the conditional dependence between two sites. Such dependence was estimated locally by a posterior probability. Based on the chain rule for Bayesian networks, such local posterior distribu-

tions were finally used to estimate the full joint posterior distribution [19]. A maximum of two parents per node was assumed to limit the complexity of the BGM. Posterior distributions were 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 1,000 steps for inference. Analyses were run in duplicate to test for convergence (figures S1-S2).

The estimated BGM can be seen as a weighted network of coevolution amongst sites, where each posterior probability measures the strength of coevolution. Each probability threshold gives rise to a network whose topology can be analysed based on a number of measures [20] borrowed from social network analysis. We focused in particular on six: average diameter: length of the longest path between pairs of nodes; average betweenness: measures the importance of each node in their ability to connect to dense subnetworks; assortative degree: measures the extent to which nodes of similar degree are connected to each other (homophyly); eccentricity: is the shortest path linking the most distant nodes in the network; average strength: rather than just count the number of connections of each node (degree), strength sums up the weights of all the adjacent nodes; average path length: measures the shortest distance between each pair of nodes. All measures were computed using the igraph package [21]. Thresholds of posterior probabilities for correlated evolution ranged from 0.01 (weak) to 0.99 (strong). LOESS regressions were then fitted to the results.

3 Results

In search for differences between regular epidemics and severe outbreaks, we started off by contrasting GP sequences of the Ebola virus that circulated before and since 2014. A visual inspection of the networks of correlated sites revealed a striking difference between

pre-2014 and outbreak sequences, in particular at weak correlations: while in pre-2014 networks interactions are very dense and involve most sites of the GP protein, only a small number of sites are interacting in outbreak viruses (figure 1). Furthermore, with increasing strengths of interactions, outbreak networks become completely disconnected faster: at posterior probability $P = 0.80$ some sites still interact in pre-2014 proteins, while all interactions disappear from $P = 0.60$ in outbreak proteins (figure 1). Similar patterns for the Influenza (both HA and NA) and Zika viruses (figures S3-S5) suggest that during a severe outbreak, a destabilisation of viral genes occurs, especially amongst sites that entertain weak interactions.

To investigate this destabilisation hypothesis further, we analysed the structure of these networks with the tools of social network analysis. Again, we found a consistent pattern when contrasting regular and outbreak viruses: at weak to moderate interactions ($P \leq 0.50$), outbreak viruses have networks of smaller diameter, path length, and eccentricity (figure 2a-c, columns 1-5). All these patterns point to fewer connected sites in outbreak viruses. Betweenness is smaller for outbreak viruses (except Ebola), and transitivity tends to be larger (except Zika). These last two measures also suggest that interactions amongst sites are weakened in outbreak viruses. Other networks statistics failed to show a clear pattern (figure S6); in particular, there were no clear differences in terms of degree, centrality or homophyly, properties that are not directly related to network stability.

Should these weak interactions play a critical role in the stabilisation of viruses outside of pandemics, we would expect to observe the strengthening of all the network statistics after the outbreak, as years go by. To test this prediction and estimate how long this re-stabilisation process can take, we analysed in a similar way all influenza seasons in the Northern hemisphere following the 2009 pandemic. Consistent with our prediction,

both HA and NA genes show a gradual transition between a typical pandemic state to a regular state in two-to-three seasons (figure 2, column 6-7, respectively).

4 Discussion

To understand how evolutionary dynamics are affected during a viral outbreak, we compared non-outbreak and outbreak viruses. Based on the hypothesis that non-outbreak viruses converge towards a stable evolutionary strategy with their host, and that such a stability is mediated by correlated evolution amongst pairs of sites in viral genes, we reconstructed the coevolution patterns in genes of non-outbreak and outbreak viruses. In line with our prediction, results show that outbreak viruses exhibit fewer coevolving sites than their non-outbreak counterparts, and that these interactions are gradually restored after the outbreak, at least in the case of the Influenza (2009 H1N1) virus for both HA and NA.

Two lines of evidence further support the destabilisation hypothesis. First, all three viruses showed temporary increases in their rate of molecular evolution during each outbreak [2–4]; such increases can be expected to tear down the coevolution pattern, and hence, destabilise viral genomes. Second, a probable cause can be identified in all cases studied here. For Influenza, the 2009 pandemic was caused by a chain of reassortment events that affected the two genes studied here, HA (triple-reassortant swine) and NA (Eurasian avian-like swine) [4]. Such exchanges of segments can very well destabilise the evolutionary dynamics, at least of the implicated segments. A similar argument could be put forward for both Ebola and Zika viruses, as a change of host was implicated in the Ebola outbreak [2], and a change of continent in the case of Zika [3]. These corresponding changes of environment (*sensu lato*) might have triggered the destabilisations observed

here.

One outstanding question is about the importance of weak patterns of coevolution within a gene: how can it be explained that it is essentially weak correlations (around $P = 0.25$) that distinguish non-outbreak from outbreak viruses? In recent work with mice, four phenotypes were quantitatively analysed following large intercrosses, and linear regressions on pairs of quantitative trait loci were used to detect non-additive effects, *i.e.*, epistasis; it was then showed that most epistatic interactions were weak and, critically, tended to stabilise phenotypes towards the mean of the population [22]. Viruses are not mice, and all correlations that we detect are probably not signalling epistasis, but this work in mice and the evidence presented here go in the same direction: weak interactions have a stabilising effect on viral genes and their phenotype (epidemics). It is further possible that the intricate nature of these weak correlation networks has higher-order effects [22], that in turn increase canalisation and hence may help viruses weather environmental and genotypic fluctuations [23]. The elimination of these many weak interactions has a destabilising effect that may lead to outbreaks. While the evidence shown here does not support the causal nature of this relationship, monitoring correlation networks could help forecast imminent outbreaks.

Data accessibility

Scripts and sequence alignments used are available from `github.com/sarisbro`.

Authors' contributions

S.A.B. designed the study, and wrote the paper. S.A.B., N.I. and J.N. performed research and analyses, and edited the paper. All authors approved the final version of the manuscript, and agree to be held accountable for the content therein.

Competing interests

We have no competing interests.

Funding

This work was supported by the Natural Sciences Research Council of Canada and by the Canada Foundation for Innovation (S.A.B.) and by the University of Ottawa (N.I., J.N.).

References

1. Van Valen, L., 1973 A new evolutionary law. *Evolutionary theory* **1**, 1–30.

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.*, 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–72. (doi:10.1126/science.1259657).

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.*, 2016 Zika virus in the americas: Early epidemiological and genetic findings. *Science* **352**, 345–9. (doi:10.1126/science.aaf5036).

4. 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.*, 2009 Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–5. (doi:10.1038/nature08182).

5. Nowak, M. & May, R. M., 2000 *Virus dynamics: mathematical principles of immunology and virology*. Oxford University Press, UK.

6. Poon, A. F. Y., Lewis, F. I., Pond, S. L. K. & Frost, S. D. W., 2007 An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol* **3**, e231. (doi:10.1371/journal.pcbi.0030231).

7. Ibeh, N., Nshogozabahizi, J. C. & Aris-Brosou, S., 2016 Both epistasis and diversifying selection drive the structural evolution of the Ebola virus glycoprotein mucin-like domain. *J Virol* **90**, 5475–84. (doi:10.1128/JVI.00322-16).

8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J., 1990 Basic local alignment search tool. *J Mol Biol* **215**, 403–10. (doi:10.1016/S0022-2836(05)80360-2).

9. Gilbert, D., 2003 Sequence file format conversion with command-line readseq. *Curr Protoc Bioinformatics* **Appendix 1**, Appendix 1E. (doi:10.1002/0471250953.bia01es00).

10. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. & Lipman, D., 2008 The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**, 596–601. (doi:10.1128/JVI.02005-07).

11. Edgar, R. C., 2004 Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–7. (doi:10.1093/nar/gkh340).

12. Larsson, A., 2014 AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–8. (doi:10.1093/bioinformatics/btu531).

13. Aris-Brosou, S. & Rodrigue, N., 2012 The essentials of computational molecular evolution. *Methods Mol Biol* **855**, 111–52. (doi:10.1007/978-1-61779-582-4_4).

14. Price, M. N., Dehal, P. S. & Arkin, A. P., 2010 Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490. (doi:10.1371/journal.pone.0009490).

15. Paradis, E., Claude, J. & Strimmer, K., 2004 APE: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290.

16. Poon, A. F. Y., Lewis, F. I., Frost, S. D. W. & Kosakovsky Pond, S. L., 2008 Spidermonkey: rapid detection of co-evolving sites using bayesian graphical models. *Bioinformatics* **24**, 1949–50. (doi:10.1093/bioinformatics/btn313).

17. Pond, S. L. K., Frost, S. D. W. & Muse, S. V., 2005 Hyphy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–9. (doi:10.1093/bioinformatics/bti079).

18. Kosakovsky Pond, S. L. & Frost, S. D. W., 2005 Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**, 1208–22. (doi:10.1093/molbev/msi105).

19. Pearl, J., 1988 *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

20. Newman, M., 2010 *Networks: an introduction*. OUP Oxford.

21. Csardi, G. & Nepusz, T., 2006 The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9.

22. Tyler, A. L., Donahue, L. R., Churchill, G. A. & Carter, G. W., 2016 Weak epistasis generally stabilizes phenotypes in a mouse intercross. *PLoS Genet* **12**, e1005805. (doi:10.1371/journal.pgen.1005805).

23. Waddington, C. H., 1942 Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565.

Figures

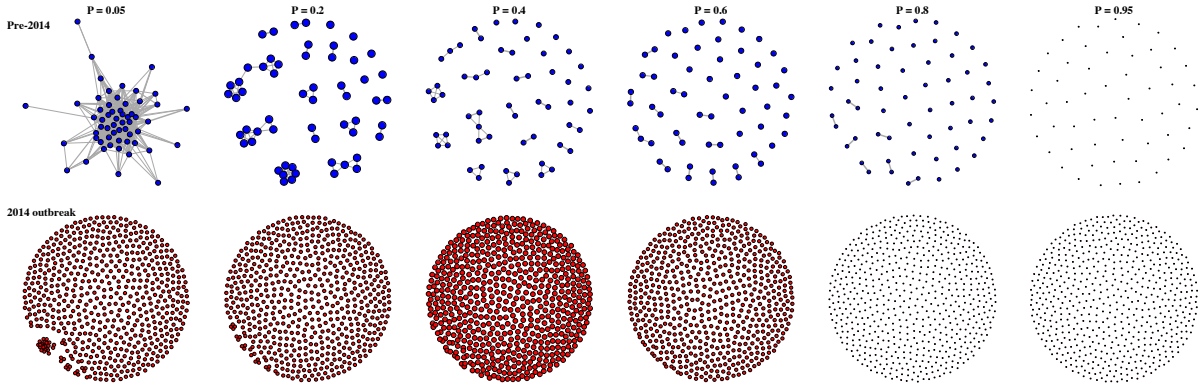

Figure 1. Correlation network of pre-outbreak and outbreak Ebola viruses. Networks of correlated sites in the GP protein are shown in each panel. The top row shows networks for the viruses circulating before the 2014 outbreak (blue); the bottom row shows networks for outbreak viruses (red). Each column shows networks for different strengths of correlation, from weak ($P = 0.05$) to strong ($P = 0.95$). Nodes represent animo acid sites, and edges correlations. Node sizes are proportional to diameter.

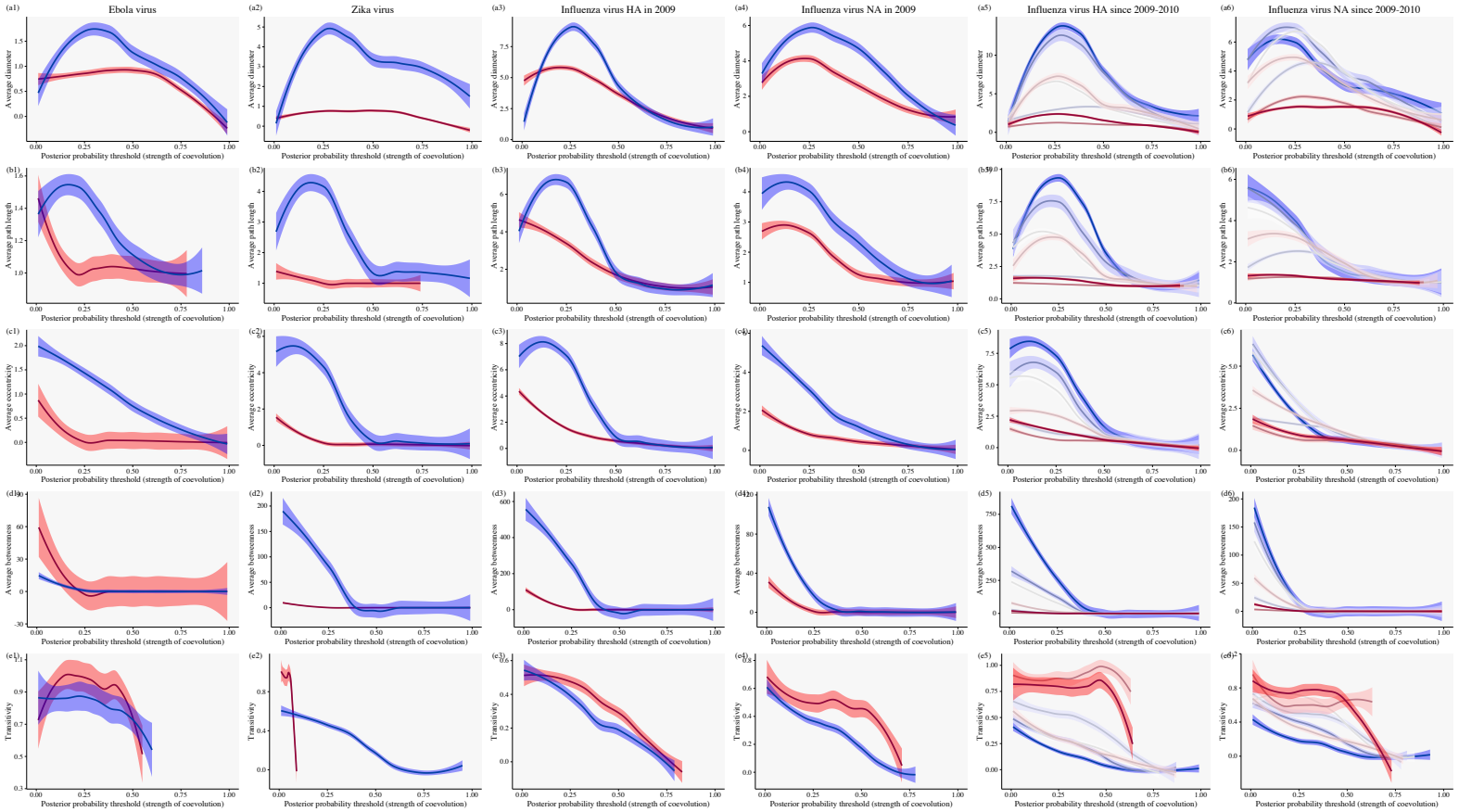

Figure 2. Network properties between pandemic and non-pandemic viruses. Results are shown for Ebola (column 1), Zika (2) and Influenza viruses (for HA and NA circulating in 2009 in (3) and (4), respectively, and for pandemic viruses circulating since then, season by season (5-6)). Pandemic viruses are show in red, while non-pandemic ones are in blue. Shading: 95% confidence envelopes of the LOESS regressions. Five network measures are shown: (a) diameter, (b) average path length, (c) eccentricity, (d) betweenness, and (e) transitivity.