Viral outbreaks involve destabilized 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

⋆Correspondence and requests for materials should be addressed to S.A.B. (sarisbro@uottawa.ca).

Abstract

Recent history has provided us with two severe viral outbreaks (Ebola and Zika) and one pandemic (Influenza A/H1N1). In all three cases, post-hoc analyses have given us deep insights into what triggered these outbreaks, their timing, evolutionary dynamics, and their phylogeography, but the genomic characteristics of outbreak viruses are still unclear. To address this outstanding question, we searched for a common denominator of these recent outbreaks, positing that genomes of outbreak viruses are in an unstable evolutionary state, while those of non-outbreak viruses are stabilized by a network of correlated substitutions that have been found to be prevalent. Here, we show that during regular epidemics, viral genomes are indeed stabilized by a dense network of weakly correlated sites, and that these networks disappear during pandemics and outbreaks when rates of evolution increase transiently. Post-pandemic, these evolutionary networks are progressively re-established. We finally show that destabilization is not caused by mutations targeting epitopes, but more likely by changes in the environment *sensu lato*. Our results prompt for a new interpretation of pandemics as being caused by, from an evolutionary standpoint, destabilized, unhealthy viruses.

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

Introduction

Viruses are engaged in a form of arms race with their hosts, in which each party endeavors to outpace the other [1]. Regular epidemics can therefore be seen as an equilibrium situation, where both the virus and the hosts coexist. Such a stable evolutionary strategy can however break down when the virus becomes extremely virulent, which can lead to a severe outbreak or even a pandemic. Recent history is rich in such examples with an Ebola virus outbreak in 2014 [2], a Zika outbreak in 2015 [3], and an Influenza pandemic in 2009 [4]. Despite all of these recent examples, in which the phylodymanics of these events were meticulously reconstructed, we still do not know what characterizes the evolutionary dynamics of outbreaks and pandemics. Here we address this outstanding question by contrasting the evolutionary dynamics of pandemic and non-pandemic viruses.

As theory tells us that regular epidemics are the result of a dynamic equilibrium [5], we posit that outbreaks are associated with a disequilibrium at the genomic level. More specifically, we suggest that outbreaks involve destabilized 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 severely 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 destabilized.

Results

Networks of correlated sites are destabilized during outbreaks. In search for evolutionary differences between regular epidemics and severe outbreaks, we first contrasted the glycoprotein precursor (GP) sequences of the Ebola virus that circulated

before, and during 2014/2016 outbreak. For this, we identified the pairs of nucleotides that show evidence for correlated evolution in each data set, before and during the outbreak. As in previous work [6, 7], we found that these pairs of sites form a network. A first inspection of these networks of correlated sites revealed a striking difference between pre-2014 and outbreak sequences: in particular at weak correlations, the pre-2014 interaction networks are very dense and involve most sites of GP, while only a small number of sites are interacting in outbreak viruses (Figure 1). Furthermore, at increasing correlation strengths, outbreak networks become completely disconnected faster: at posterior probability $Pr = 0.80$ some sites still interact in pre-2014 proteins, while all interactions have disappeared from $Pr = 0.60$ 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 destabilization of viral genes occurs, especially among sites that entertain weak interactions.

Destabilization affects weakly correlated sites. To further investigate this destabilization hypothesis, we analyzed the structure of these networks with the tools of social network analysis [8]. Again, we found a consistent pattern when contrasting regular and outbreak viruses: at weak to moderate interactions ($Pr \leq 0.50$), outbreak viruses have networks of smaller diameter, shorter path length, and reduced 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 among 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 – all properties that are not directly related to network stability.

Post-outbreak re-stabilization. Should these weak interactions play a critical role in the stabilization 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-stabilization process can take, we analyzed in a similar way all influenza seasons in the Northern hemisphere following the 2009 pandemic (until 2015-16). 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 5-6, respectively).

Non-genetic sources of destabilization. To understand what the potential sources of this destabilization are, we assessed the involvement of viral antigenic determinants / epitopes. Should mutations accumulating in such epitopes be responsible for destabilization, we would expect (i) that weak interactions in non-pandemic viruses involve mostly epitopes, and (ii) that pandemics be associated with the disappearance of these interactions at epitopes first. Figure 3 shows no evidence supporting this hypothesis ($X^2 = 0.0663$, $df = 1$, $P = 0.7967$): non-pandemic viruses show a small number of predicted epitopes in their interaction network, that do not act as central hubs of these networks, while pandemic viruses may actually show an enrichment in interacting epitopes. This suggest that non-genetic factors are likely responsible for the initial destabilization of the genome of pandemic viruses. Changes in their ecology / environment (vector) cannot be ruled out.

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 are in a stable evolutionary equilibrium, and that such a stability is mediated by correlated evolution among pairs of sites in viral genes, we reconstructed the coevolution patterns in genes of non-outbreak and outbreak viruses. In line with our prediction, we found 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 independent lines of evidence are consistent with our destabilization hypothesis. First, all three viruses showed temporary increases in their rate of molecular evolution during each outbreak [2, 3, 4]; such increases can be expected to tear down the coevolutionary structure, and hence, destabilize viral genomes. We showed that epitopes were not particular targets of this mutational process, which is hence most likely affecting sites randomly. Second, a probable cause of the epidemics 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 destabilize the evolutionary dynamics, at least of the implicated segments. A similar argument can made 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, 9, 10]. These corresponding changes of environment (*sensu lato*) might have triggered the destabilizations observed here. In addition to such environmental changes, it is very likely that destabilization reflects a complex interaction between the genetics of viruses, their demographic fluctuations and environmental changes.

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 $Pr = 0.25$) that distinguish non-outbreak from outbreak viruses? In recent study on

mice, four phenotypes were quantitatively analyzed 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 stabilize phenotypes towards the mean of the population [11]. Viruses are not mice, and all correlations that we detect are probably not involved in epistatic interactions, but both this work in mice and the evidence presented here go in the same direction: weak interactions have a stabilizing 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 [11], that in turn increase canalization and hence may help viruses weather environmental and genotypic fluctuations [12]. The elimination of these many weak interactions has a destabilizing effect that may be caused or lead to outbreaks. This calls for a new interpretation of pandemics that, from an evolutionary point of view, appeared to be caused by unhealthy or diseased viruses. While the evidence shown here does not support the causal nature of this relationship, monitoring correlation networks could help forecast imminent outbreaks.

Methods

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 [13] 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 [14] 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 [15]. 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.

Phylogenetic analyses. Sequences were all aligned with Muscle [16] with fastest options (-maxiters 1 -diags). Alignments were visually inspected with AliView [17] to remove rogue sequences and sequencing errors. Phylogenetic trees were inferred by maximum likelihood under the General Time-Reversible model with among-site rate variation [18] with FastTree [19]. As outbreak sequences (Ebola and Zika viruses) cluster away from non-pandemic sequences, we used the `subtreeplot()` function in APE [20] 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.

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 [21] as implemented in HyPhy [22]. Briefly, ancestral mutational paths were first reconstructed under the MG94×HKY85 substitution model [23] 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 distributions were finally used to estimate the full joint posterior distribution [24]. 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 among sites, where each posterior probability measures the strength of coevolution. Each probability threshold gives rise to a network whose topology can be analyzed based on a number of measures [8] 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 ver. 1.0.1 [25]. Thresholds of posterior probabilities for correlated evolution ranged from 0.01 (weak) to 0.99 (strong). LOESS regressions were then fitted to the results.

Epitope analyses. Epitopes were predicted using the NetCTL 1.2 Server [26]. Briefly, Cytotoxic T lymphocyte (CTL) epitopes are predicted based on a neural network algorithm trained on a database of human MHC class I ligands. Epitopes can be predicted for 12 MHC supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, B62), that are broad families of very similar peptides for which independent neural network models have been generated. As such, we ran the epitope prediction for each supertype independently, on non-outbreak and outbreak viruses. Circos plots were generated with the circlize package ver. 0.3.10 in R [27]. Scripts and sequence alignments used are available from `github.com/sarisbro`.

References

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

2. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–72 (2014).

3. Faria, N. R. *et al.* Zika virus in the americas: Early epidemiological and genetic findings. *Science* **352**, 345–9 (2016).

4. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–5 (2009).

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

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

7. 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* **90**, 5475–84 (2016).

8. Newman, M. *Networks: an introduction* (OUP Oxford, 2010).

9. Zhang, Q. *et al.* Spread of Zika virus in the Americas. *Proc Natl Acad Sci U S A* (2017).

10. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* (2017).

11. Tyler, A. L., Donahue, L. R., Churchill, G. A. & Carter, G. W. Weak epistasis generally stabilizes phenotypes in a mouse intercross. *PLoS Genet* **12**, e1005805 (2016).

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

13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–10 (1990).

14. Gilbert, D. Sequence file format conversion with command-line readseq. *Curr Protoc Bioinformatics* **Appendix 1**, Appendix 1E (2003).

15. Bao, Y. *et al.* The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**, 596–601 (2008).

16. Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–7 (2004).

17. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–8 (2014).

18. Aris-Brosou, S. & Rodrigue, N. The essentials of computational molecular evolution. *Methods Mol Biol* **855**, 111–52 (2012).

19. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

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

21. 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* **24**, 1949–50 (2008).

22. Pond, S. L. K., Frost, S. D. W. & Muse, S. V. Hyphy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–9 (2005).

23. 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* **22**, 1208–22 (2005).

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

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

26. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC bioinformatics* **8**, 1 (2007).

27. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in r. *Bioinformatics* btu393 (2014).

Author 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.

Acknowledgements

We thank Jonathan Dench, and Berthin Bitja for discussions. 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.).

Additional information

Supplementary Information accompanies this paper at `http://www.nature.com/naturecommunication`

Competing interests: The authors declare no competing financial interests.

Figures

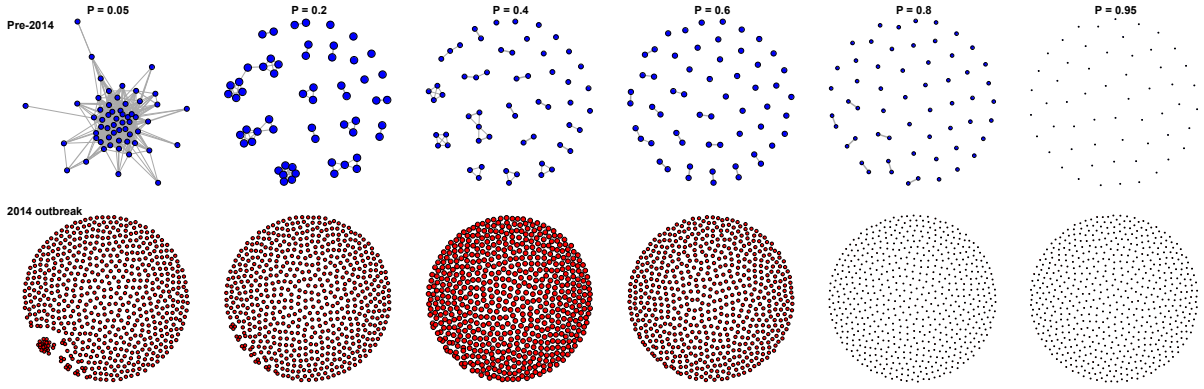

Figure 1. Correlation network of pre-outbreak and outbreak Ebola viruses. Networks of correlated sites in the GP protein are shown in each panel. The top row shows networks for the viruses circulating before the 2014 outbreak (blue); the bottom row shows networks for outbreak viruses (red). Each column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

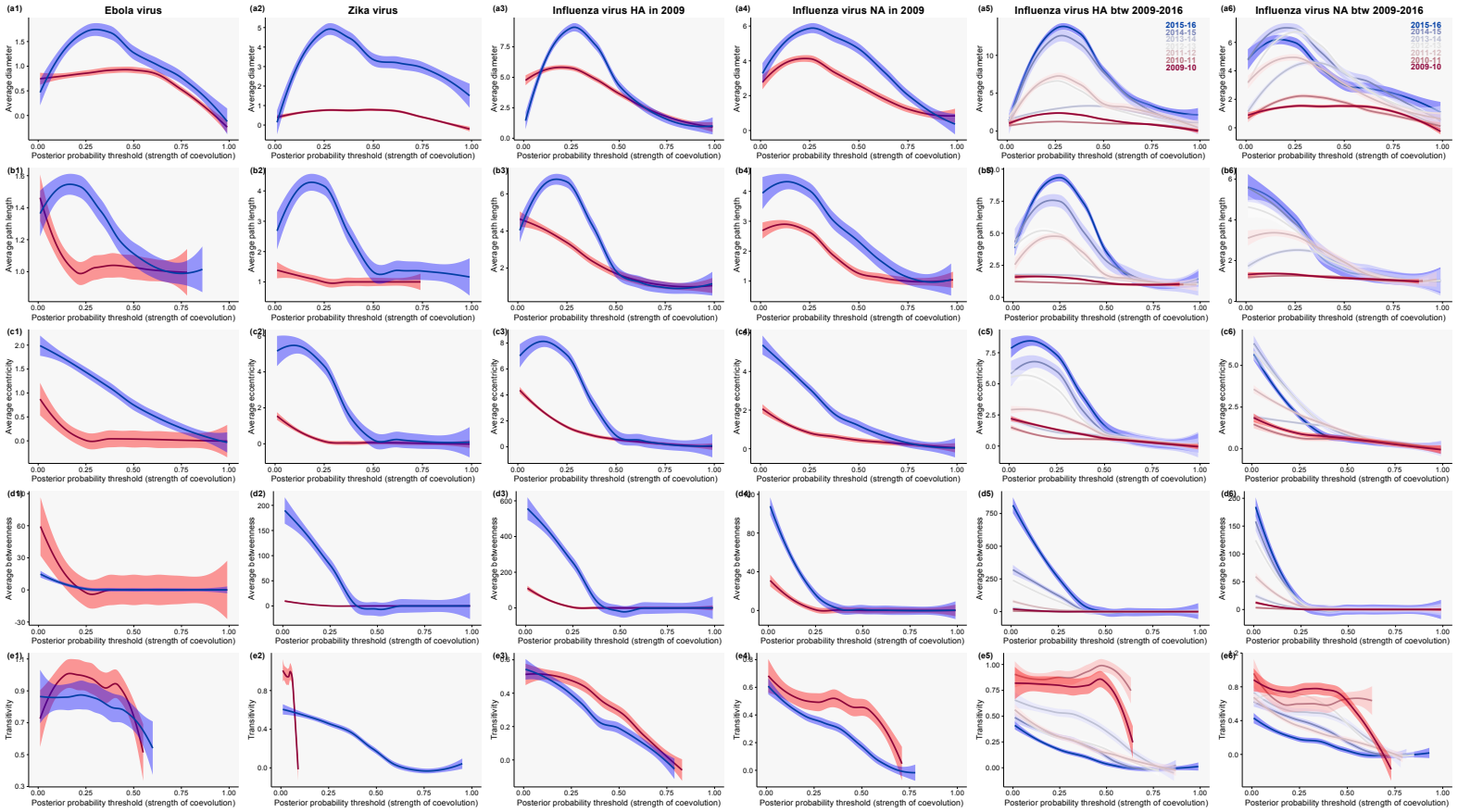

Figure 2. Network properties between pandemic and non-pandemic viruses. Results are shown for Ebola (column 1), Zika (2) and Influenza viruses: for HA and NA circulating in 2009 in (3) and (4), respectively, and for pandemic viruses circulating between the 2009-10 (deep red) and the 2015-16 (deep blue) season in (5) and (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.

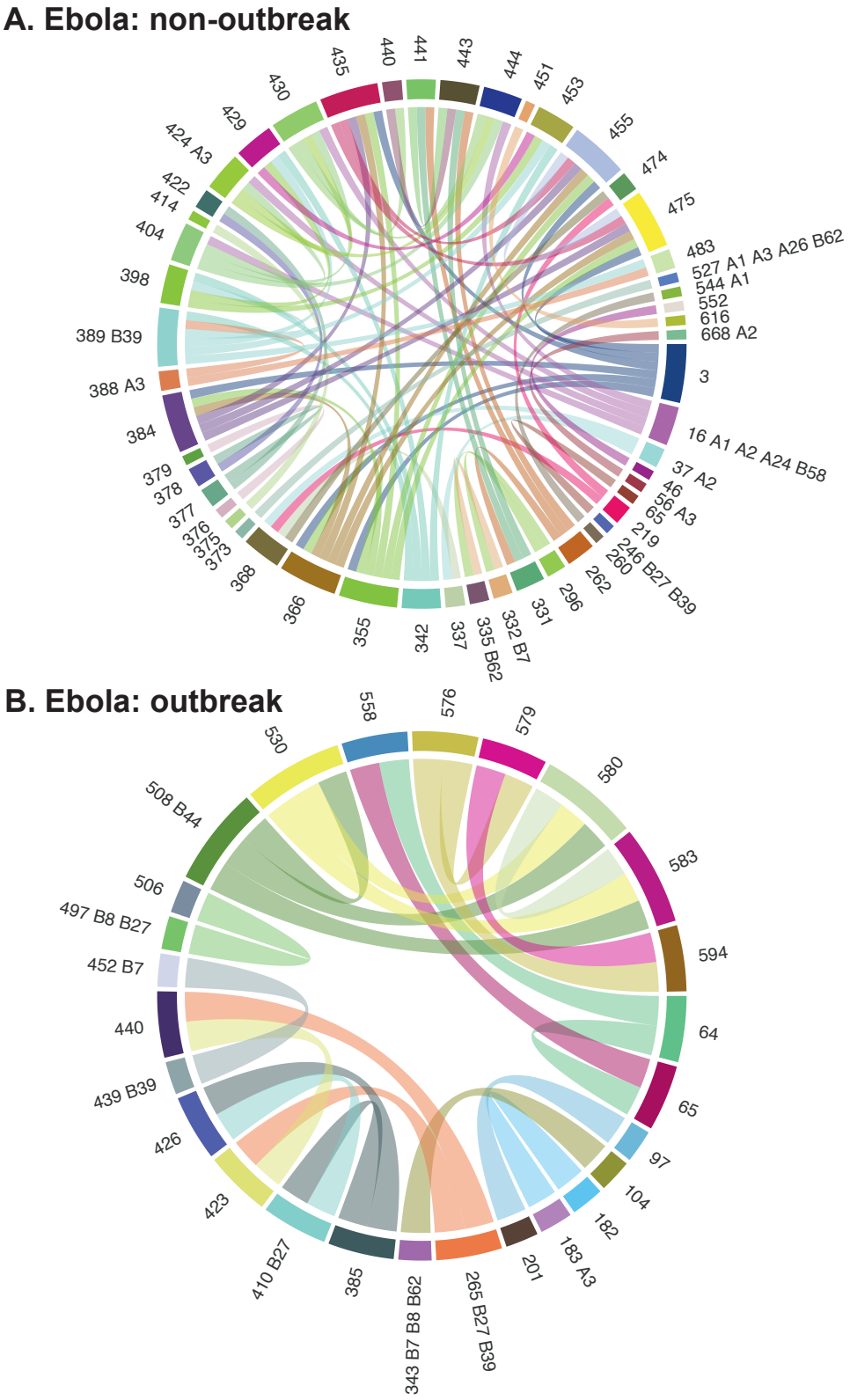

Figure 3. Interacting residues in pandemic and non-pandemic viruses. Results are shown for Ebola at weak correlations ($Pr = 0.20$). Coevolving positions in the alignment are identified with arabic numbers; for those that are predicted to be epitopes, supertypes (A1, A2, *etc.*) are shown.