

1 **Viral outbreaks involve destabilized viruses:**  
2 **evidence from Ebola, Influenza and Zika**

3 Stéphane Aris-Brosou,<sup>1,2,\*</sup> Neke Ibeh,<sup>1</sup> and Jessica Noël<sup>1</sup>

4 <sup>1</sup>Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada

5 <sup>2</sup>Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N  
6 6N5, Canada

7 \*Correspondence and requests for materials should be addressed to S.A.B.  
8 (sarisbro@uottawa.ca).

## 9 Abstract

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

25 *Keywords:* Ebola virus, Influenza virus, Zika virus, outbreak, pandemic, correlated evo-  
26 lution

## 27 Introduction

28 Viruses are engaged in a form of arms race with their hosts, in which each party endeavors  
29 to outpace the other [1]. Regular epidemics can therefore be seen as an equilibrium  
30 situation, where both the virus and the hosts coexist. Such a stable evolutionary strategy  
31 can however break down when the virus becomes extremely virulent, which can lead to a  
32 severe outbreak or even a pandemic. Recent history is rich in such examples with an Ebola  
33 virus outbreak in 2014 [2], a Zika outbreak in 2015 [3], and an Influenza pandemic in 2009  
34 [4]. Despite all of these recent examples, in which the phylodynamics of these events were  
35 meticulously reconstructed, we still do not know what characterizes the evolutionary  
36 dynamics of outbreaks and pandemics. Here we address this outstanding question by  
37 contrasting the evolutionary dynamics of pandemic and non-pandemic viruses.

38 As theory tells us that regular epidemics are the result of a dynamic equilibrium  
39 [5], we posit that outbreaks are associated with a disequilibrium at the genomic level.  
40 More specifically, we suggest that outbreaks involve destabilized viral genomes, where  
41 evolutionary stability is maintained by compensatory mutations, that can be epistatic or  
42 not, but that result in signals of correlated evolution. We predict that such signals are  
43 severely weakened during an outbreak. As these signals often lead to complex networks  
44 of interactions [6, 7], we test how the structure of these correlation networks is affected  
45 during an outbreak. We show that during an outbreak, viral genes are destabilized.

## 46 Results

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

50 before, and during 2014/2016 outbreak. For this, we identified the pairs of nucleotides  
51 that show evidence for correlated evolution in each data set, before and during the out-  
52 break. As in previous work [6, 7], we found that these pairs of sites form a network. A  
53 first inspection of these networks of correlated sites revealed a striking difference between  
54 pre-2014 and outbreak sequences: in particular at weak correlations, the pre-2014 inter-  
55 action networks are very dense and involve most sites of GP, while only a small number  
56 of sites are interacting in outbreak viruses (Figure 1). Furthermore, at increasing corre-  
57 lation strengths, outbreak networks become completely disconnected faster: at posterior  
58 probability  $Pr = 0.80$  some sites still interact in pre-2014 proteins, while all interactions  
59 have disappeared from  $Pr = 0.60$  in outbreak proteins (Figure 1). Similar patterns for  
60 the Influenza (both HA and NA) and Zika viruses (Figures S3-S5) suggest that during  
61 a severe outbreak, a destabilization of viral genes occurs, especially among sites that  
62 entertain weak interactions.

63 **Destabilization affects weakly correlated sites.** To further investigate this desta-  
64 bilization hypothesis, we analyzed the structure of these networks with the tools of social  
65 network analysis [8]. Again, we found a consistent pattern when contrasting regular and  
66 outbreak viruses: at weak to moderate interactions ( $Pr \leq 0.50$ ), outbreak viruses have  
67 networks of smaller diameter, shorter path length, and reduced eccentricity (Figure 2a-  
68 c, columns 1-5). All these patterns point to fewer connected sites in outbreak viruses.  
69 Betweenness is smaller for outbreak viruses (except Ebola), and transitivity tends to be  
70 larger (except Zika). These last two measures also suggest that interactions among sites  
71 are weakened in outbreak viruses. Other networks statistics failed to show a clear pattern  
72 (Figure S6): in particular, there were no clear differences in terms of degree, centrality or  
73 homophily – all properties that are not directly related to network stability.

74 **Post-outbreak re-stabilization.** Should these weak interactions play a critical role in  
75 the stabilization of viruses outside of pandemics, we would expect to observe the strength-  
76 ening of all the network statistics after the outbreak, as years go by. To test this prediction  
77 and estimate how long this re-stabilization process can take, we analyzed in a similar way  
78 all influenza seasons in the Northern hemisphere following the 2009 pandemic (until 2015-  
79 16). Consistent with our prediction, both HA and NA genes show a gradual transition  
80 between a typical pandemic state to a regular state in two-to-three seasons (Figure 2,  
81 column 5-6, respectively).

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

## 94 Discussion

95 To understand how evolutionary dynamics are affected during a viral outbreak, we com-  
96 pared non-outbreak and outbreak viruses. Based on the hypothesis that non-outbreak

97 viruses are in a stable evolutionary equilibrium, and that such a stability is mediated by  
98 correlated evolution among pairs of sites in viral genes, we reconstructed the coevolution  
99 patterns in genes of non-outbreak and outbreak viruses. In line with our prediction, we  
100 found that outbreak viruses exhibit fewer coevolving sites than their non-outbreak coun-  
101 terparts, and that these interactions are gradually restored after the outbreak, at least in  
102 the case of the Influenza (2009 H1N1) virus for both HA and NA.

103 Two independent lines of evidence are consistent with our destabilization hypothesis.  
104 First, all three viruses showed temporary increases in their rate of molecular evolution  
105 during each outbreak [2, 3, 4]; such increases can be expected to tear down the coevolu-  
106 tionary structure, and hence, destabilize viral genomes. We showed that epitopes were  
107 not particular targets of this mutational process, which is hence most likely affecting sites  
108 randomly. Second, a probable cause of the epidemics can be identified in all cases studied  
109 here. For Influenza, the 2009 pandemic was caused by a chain of reassortment events  
110 that affected the two genes studied here, HA (triple-reassortant swine) and NA (Eurasian  
111 avian-like swine) [4]. Such exchanges of segments can very well destabilize the evolution-  
112 ary dynamics, at least of the implicated segments. A similar argument can be made for both  
113 Ebola and Zika viruses, as a change of host was implicated in the Ebola outbreak [2],  
114 and a change of continent in the case of Zika [3, 9, 10]. These corresponding changes  
115 of environment (*sensu lato*) might have triggered the destabilizations observed here. In  
116 addition to such environmental changes, it is very likely that destabilization reflects a  
117 complex interaction between the genetics of viruses, their demographic fluctuations and  
118 environmental changes.

119 One outstanding question is about the importance of weak patterns of coevolution  
120 within a gene: how can it be explained that it is essentially weak correlations (around  
121  $Pr = 0.25$ ) that distinguish non-outbreak from outbreak viruses? In recent study on

122 mice, four phenotypes were quantitatively analyzed following large intercrosses, and linear  
123 regressions on pairs of quantitative trait loci were used to detect non-additive effects, *i.e.*,  
124 epistasis; it was then showed that most epistatic interactions were weak and, critically,  
125 tended to stabilize phenotypes towards the mean of the population [11]. Viruses are not  
126 mice, and all correlations that we detect are probably not involved in epistatic interactions,  
127 but both this work in mice and the evidence presented here go in the same direction:  
128 weak interactions have a stabilizing effect on viral genes and their phenotype (epidemics).  
129 It is further possible that the intricate nature of these weak correlation networks has  
130 higher-order effects [11], that in turn increase canalization and hence may help viruses  
131 weather environmental and genotypic fluctuations [12]. The elimination of these many  
132 weak interactions has a destabilizing effect that may be caused or lead to outbreaks. This  
133 calls for a new interpretation of pandemics that, from an evolutionary point of view,  
134 appeared to be caused by unhealthy or diseased viruses. While the evidence shown here  
135 does not support the causal nature of this relationship, monitoring correlation networks  
136 could help forecast imminent outbreaks.

## 137 **Methods**

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

142 For Ebola, the virion spike glycoprotein precursor, GP, was retrieved as follows. A  
143 GP sequence (KX121421) was drawn at random from the 2014 strain used previously  
144 [7] and was employed as a query for a BLASTn search [13] at the National Center for

145 Biotechnology Information. A conservative  $E$ -value threshold of 0 ( $E < 10^{-500}$ ) was used,  
146 which led to 1,181 accession numbers. As most of these accession numbers correspond  
147 to full genomes, while only GP is of interest, we (i) retrieved all corresponding GenBank  
148 files, (ii) extracted coding sequences with ReadSeq [14] of all genes, (iii) concatenated the  
149 corresponding FASTA files into a single file, (iv) which was then used to format a sequence  
150 database for local BLASTn searches, and (v) used GP from KX121421 in a second round  
151 of BLASTn searches ( $E < 10^{-250}$ , coverage  $> 75\%$ ).

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

155 Full-length Influenza A sequences were retrieved directly from the Influenza Virus  
156 Resource [15]. Only H1N1 sequences circulating in humans for the hemagglutinin (HA)  
157 and neuraminidase (NA) genes were downloaded. Two types of data sets were constructed:  
158 one containing pandemic and non-pandemic sequences circulating in 2009, the pandemic  
159 year, and one containing pandemic sequences circulating from August 1 to July 31 of  
160 each season in the Northern temperate region between 2009/2010 and 2015/2016 (seven  
161 seasons in total). Only unique sequences were retrieved.

162 **Phylogenetic analyses.** Sequences were all aligned with Muscle [16] with fastest op-  
163 tions (-maxiters 1 -diags). Alignments were visually inspected with AliView [17] to remove  
164 rogue sequences and sequencing errors. Phylogenetic trees were inferred by maximum  
165 likelihood under the General Time-Reversible model with among-site rate variation [18]  
166 with FastTree [19]. As outbreak sequences (Ebola and Zika viruses) cluster away from  
167 non-pandemic sequences, we used the `subtreepplot()` function in APE [20] to retrieve  
168 accession numbers of pandemic sequences and hence separate them from non-pandemic



169 sequences with minimal manual input. FastTree was used a second time to estimate  
170 phylogenetic trees of the subset alignments, with the same settings as above.

171 **Network analyses of correlated sites.** Amino acid positions (“sites”) that evolve  
172 in a correlated manner were identified with the Bayesian graphical model (BGM) in  
173 SpiderMonkey [21] as implemented in HyPhy [22]. Briefly, ancestral mutational paths  
174 were first reconstructed under the MG94×HKY85 substitution model [23] along each  
175 branch of the tree estimated above at non-synonymous sites. These reconstructions were  
176 recoded as a binary matrix in which each row corresponds to a branch and each column  
177 to a site of the alignment. A BGM was then employed to identify which pairs of sites  
178 exhibit correlated patterns of substitutions. Each node of the BGM represents a site and  
179 the presence of an edge indicates the conditional dependence between two sites. Such  
180 dependence was estimated locally by a posterior probability. Based on the chain rule for  
181 Bayesian networks, such local posterior distributions were finally used to estimate the full  
182 joint posterior distribution [24]. A maximum of two parents per node was assumed to  
183 limit the complexity of the BGM. Posterior distributions were estimated with a Markov  
184 chain Monte Carlo sampler that was run for  $10^5$  steps, with a burn-in period of 10,000  
185 steps sampling every 1,000 steps for inference. Analyses were run in duplicate to test for  
186 convergence (Figures S1-S2).

187 The estimated BGM can be seen as a weighted network of coevolution among sites,  
188 where each posterior probability measures the strength of coevolution. Each probability  
189 threshold gives rise to a network whose topology can be analyzed based on a number  
190 of measures [8] borrowed from social network analysis. We focused in particular on six:  
191 average diameter: length of the longest path between pairs of nodes; average betweenness:  
192 measures the importance of each node in their ability to connect to dense subnetworks;

193 assortative degree: measures the extent to which nodes of similar degree are connected to  
194 each other (homophily); eccentricity: is the shortest path linking the most distant nodes  
195 in the network; average strength: rather than just count the number of connections of  
196 each node (degree), strength sums up the weights of all the adjacent nodes; average path  
197 length: measures the shortest distance between each pair of nodes. All measures were  
198 computed using the igraph package ver. 1.0.1 [25]. Thresholds of posterior probabilities  
199 for correlated evolution ranged from 0.01 (weak) to 0.99 (strong). LOESS regressions  
200 were then fitted to the results.

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

## 210 References

- 211 1. Van Valen, L. A new evolutionary law. *Evolutionary theory* **1**, 1–30 (1973).
- 212 2. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmis-  
213 sion during the 2014 outbreak. *Science* **345**, 1369–72 (2014).
- 214 3. Faria, N. R. *et al.* Zika virus in the americas: Early epidemiological and genetic  
215 findings. *Science* **352**, 345–9 (2016).
- 216 4. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin  
217 H1N1 influenza A epidemic. *Nature* **459**, 1122–5 (2009).

- 218 5. Nowak, M. & May, R. M. *Virus dynamics: mathematical principles of immunology*  
219 *and virology* (Oxford University Press, UK, 2000).
- 220 6. Poon, A. F. Y., Lewis, F. I., Pond, S. L. K. & Frost, S. D. W. An evolutionary-  
221 network model reveals stratified interactions in the V3 loop of the HIV-1 envelope.  
222 *PLoS Comput Biol* **3**, e231 (2007).
- 223 7. Ibeh, N., Nshogozabahizi, J. C. & Aris-Brosou, S. Both epistasis and diversifying  
224 selection drive the structural evolution of the Ebola virus glycoprotein mucin-like  
225 domain. *J Virol* **90**, 5475–84 (2016).
- 226 8. Newman, M. *Networks: an introduction* (OUP Oxford, 2010).
- 227 9. Zhang, Q. *et al.* Spread of Zika virus in the Americas. *Proc Natl Acad Sci U S A*  
228 (2017).
- 229 10. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil  
230 and the Americas. *Nature* (2017).
- 231 11. Tyler, A. L., Donahue, L. R., Churchill, G. A. & Carter, G. W. Weak epistasis  
232 generally stabilizes phenotypes in a mouse intercross. *PLoS Genet* **12**, e1005805  
233 (2016).
- 234 12. Waddington, C. H. Canalization of development and the inheritance of acquired  
235 characters. *Nature* **150**, 563–565 (1942).
- 236 13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local  
237 alignment search tool. *J Mol Biol* **215**, 403–10 (1990).
- 238 14. Gilbert, D. Sequence file format conversion with command-line readseq. *Curr*  
239 *Protoc Bioinformatics* **Appendix 1**, Appendix 1E (2003).
- 240 15. Bao, Y. *et al.* The influenza virus resource at the National Center for Biotechnology  
241 Information. *J Virol* **82**, 596–601 (2008).
- 242 16. Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high  
243 throughput. *Nucleic Acids Res* **32**, 1792–7 (2004).
- 244 17. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large  
245 datasets. *Bioinformatics* **30**, 3276–8 (2014).
- 246 18. Aris-Brosou, S. & Rodrigue, N. The essentials of computational molecular evolu-  
247 tion. *Methods Mol Biol* **855**, 111–52 (2012).
- 248 19. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2—approximately maximum-  
249 likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

- 250 20. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolu-  
251 tion in r language. *Bioinformatics* **20**, 289–290 (2004).
- 252 21. Poon, A. F. Y., Lewis, F. I., Frost, S. D. W. & Kosakovsky Pond, S. L. Spi-  
253 dermonkey: rapid detection of co-evolving sites using bayesian graphical models.  
254 *Bioinformatics* **24**, 1949–50 (2008).
- 255 22. Pond, S. L. K., Frost, S. D. W. & Muse, S. V. Hyphy: hypothesis testing using  
256 phylogenies. *Bioinformatics* **21**, 676–9 (2005).
- 257 23. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of  
258 methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**, 1208–22  
259 (2005).
- 260 24. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible infer-*  
261 *ence* (Morgan Kaufmann, 1988).
- 262 25. Csardi, G. & Nepusz, T. The igraph software package for complex network research.  
263 *InterJournal, Complex Systems* **1695**, 1–9 (2006).
- 264 26. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte  
265 epitope prediction. *BMC bioinformatics* **8**, 1 (2007).
- 266 27. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances  
267 circular visualization in r. *Bioinformatics* btu393 (2014).

## 268 **Author contributions**

269 S.A.B. designed the study, and wrote the paper. S.A.B., N.I. and J.N. performed re-  
270 search and analyses, and edited the paper. All authors approved the final version of the  
271 manuscript.

## 272 **Acknowledgements**

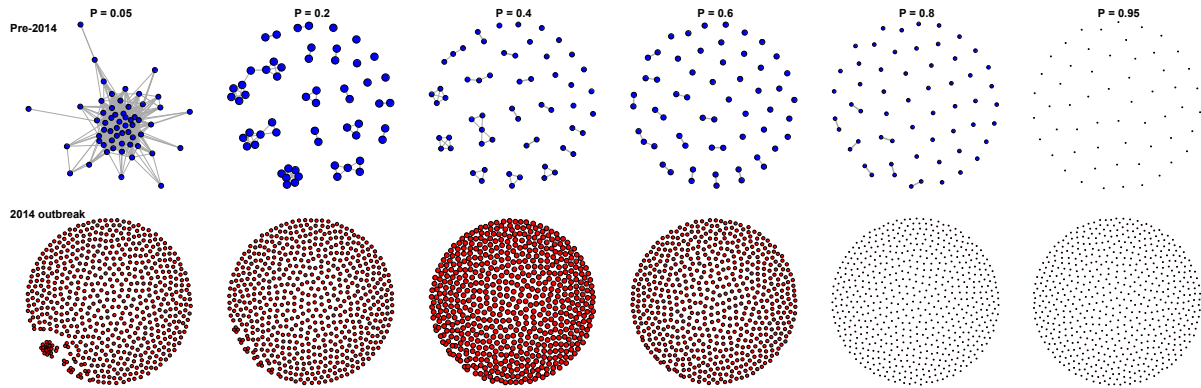
273 We thank Jonathan Dench, and Berthin Bitja for discussions. This work was supported  
274 by the Natural Sciences Research Council of Canada and by the Canada Foundation for  
275 Innovation (S.A.B.) and by the University of Ottawa (N.I., J.N.).

## 276 **Additional information**

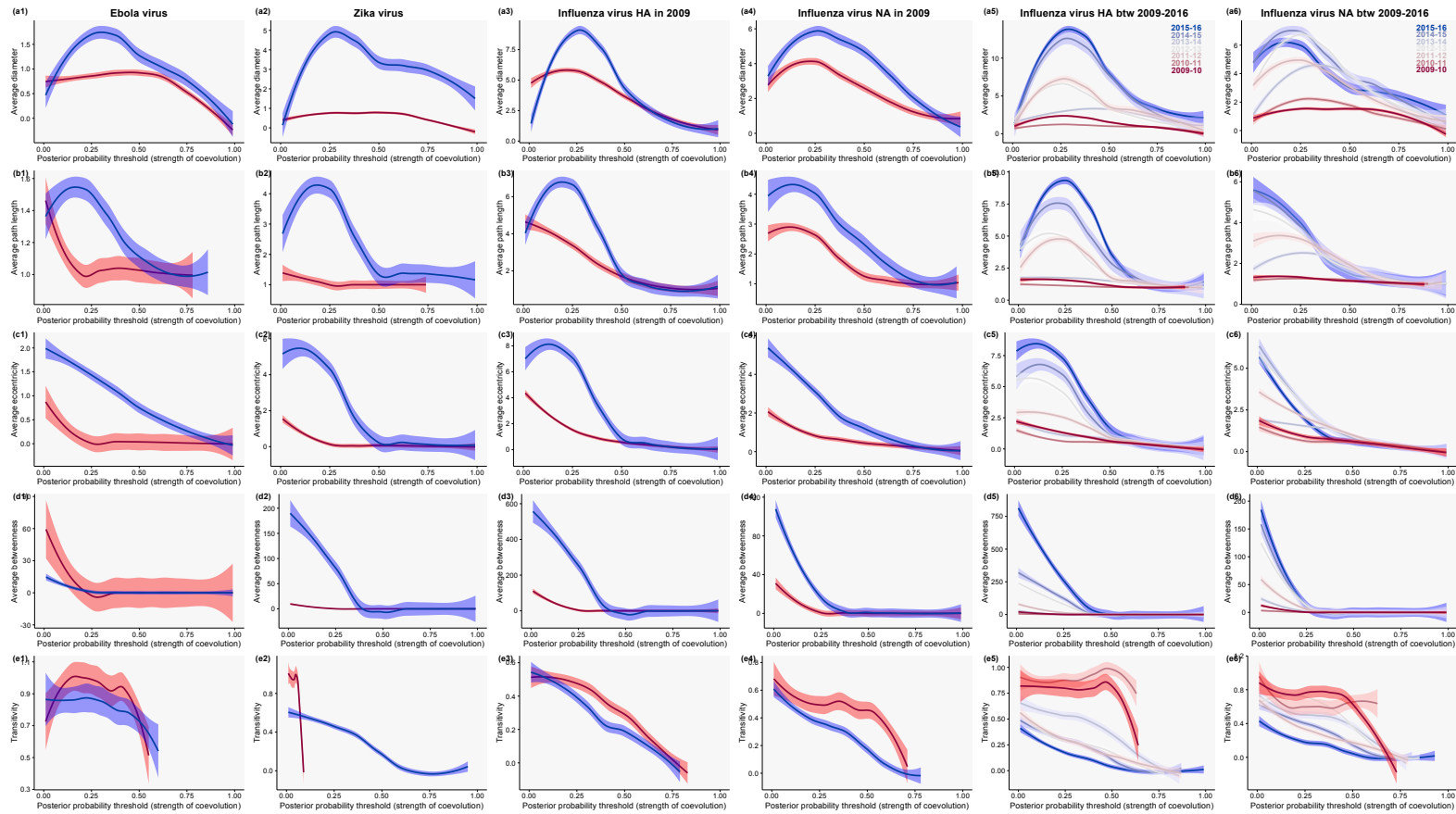
277 Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunication>

278 **Competing interests:** The authors declare no competing financial interests.

279 **Figures**

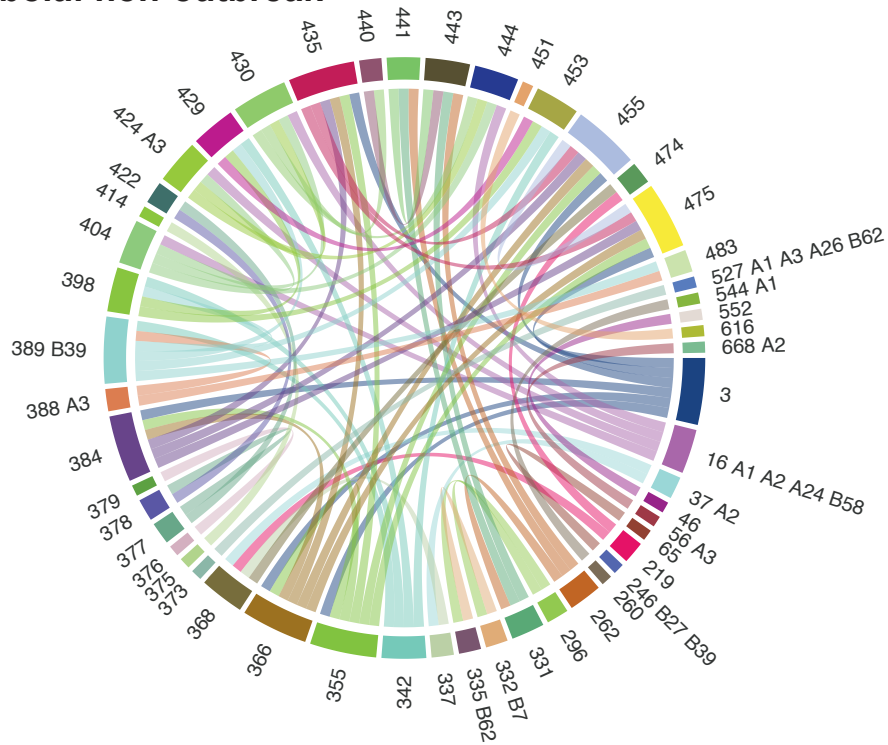


**Figure 1. Correlation network of pre-outbreak and outbreak Ebola viruses.** Networks of correlated sites in the GP protein are shown in each panel. The top row shows networks for the viruses circulating before the 2014 outbreak (blue); the bottom row shows networks for outbreak viruses (red). Each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

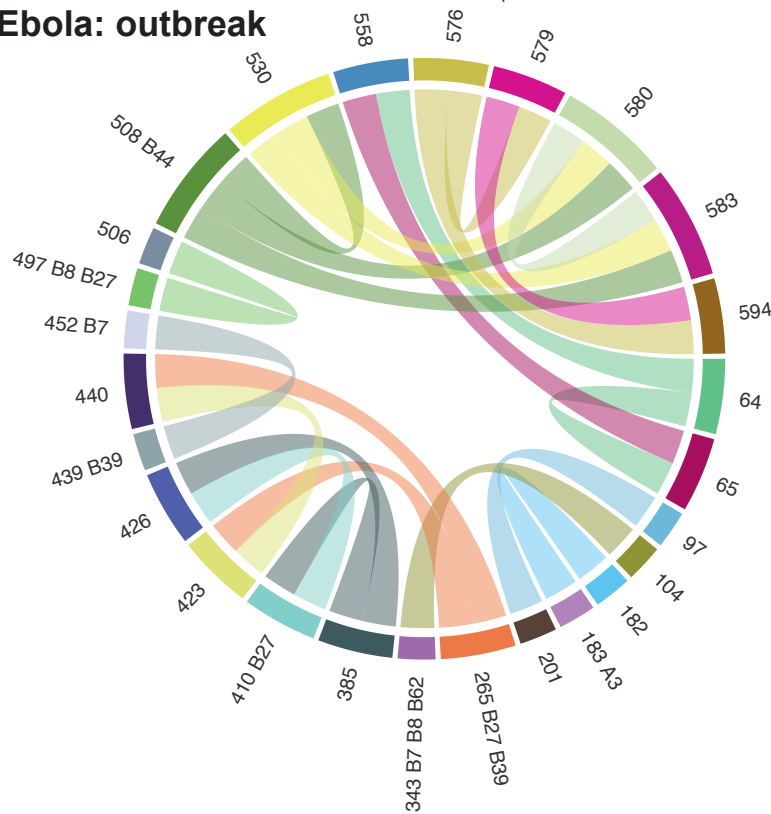


**Figure 2. Network properties between pandemic and non-pandemic viruses.** Results are shown for Ebola (column 1), Zika (2) and Influenza viruses: for HA and NA circulating in 2009 in (3) and (4), respectively, and for pandemic viruses circulating between the 2009-10 (deep red) and the 2015-16 (deep blue) season in (5) and (6). Pandemic viruses are shown 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.

### A. Ebola: non-outbreak



### B. Ebola: outbreak



**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.