Cell Research (2015) 25:753-756.

www.nature.com/cr

LETTER TO THE EDITOR

Network of co-mutations in Ebola virus genome predicts the disease lethality

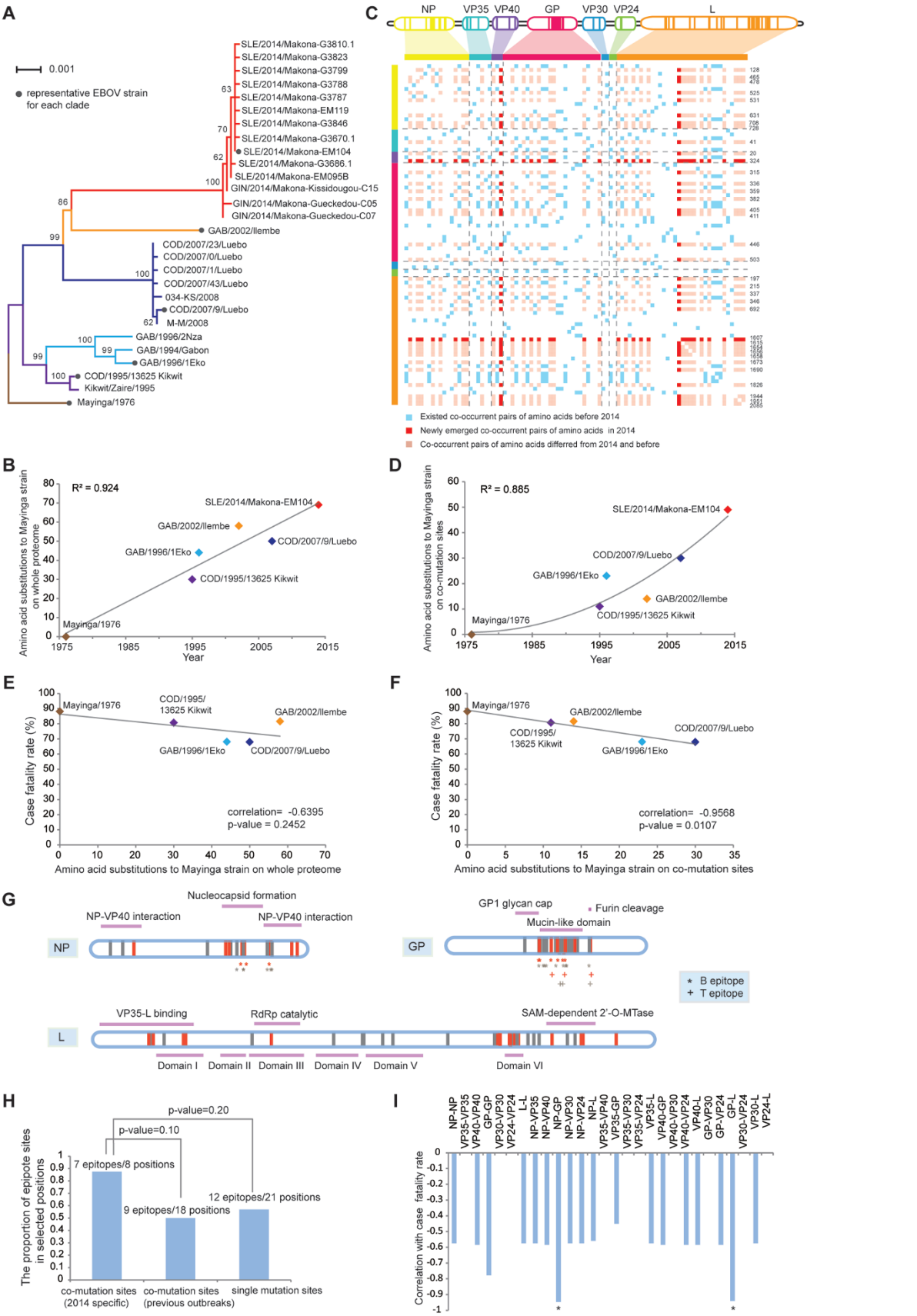
Cell Research (2015) **25**:753-756. doi:10.1038/cr.2015.54; published online 15 May 2015

Dear Editor,

Zaire ebolavirus (EBOV) is one of the most lethal human viruses with observed high case fatality rate (CFR) of up to 90%. The elucidation of how the genetic changes of EBOV link to its lethality will not only help us to rapidly assess the severity of the current and future EBOV outbreaks but also facilitate a deep understanding of the virus evolution. The EBOV together with the other four strains of ebolaviruses belongs to the *Filoviridae* RNA virus family [1] and was first identified in 1976 in the Ebola river region of Zaire (now Democratic Republic of Congo) in the Central Africa [2]. The current outbreak of Ebola virus disease (EVD) that started in December of 2013 in Guinea, West Africa [3] has caused 10 587 deaths among 25 550 reported cases of infections as of April 5, 2015. It is the largest known epidemic of EVD, which appears to expand exponentially, demonstrating a new trend of EBOV epidemics. Thus, there is an urgent need to understand how genetic evolution of EBOV contributes to the severity of the current outbreak.

The EBOV genome is a single-stranded RNA approximately 19 000 nucleotides in length, coding for 4 virion structural proteins (nucleoprotein NP, minor nucleoprotein VP30, the RNA-dependent RNA polymerase L and polymerase cofactor VP35) and 3 membrane-associated proteins (matrix proteins VP40 and VP24, and glycoprotein GP) [4]. To investigate how genetic mutations in the EBOV genome underlie its evolution process, we collected complete genomes available in the public databases [5, 6] for 126 EBOV isolates obtained from 1976, 1995, 1996, 2002, 2007-2008 and 2013-2014 outbreaks (Supplementary information, Table S1). By grouping isolates with the genome encoding identical proteins together, 28 distinct strains were formed. Whole-genome phylogenetic tree analysis further showed that these 28 EBOV strains fall into 6 distinct clades (Figure 1A and Supplementary information, Figure S1). By plotting the number of individual mutations against the year of the outbreak, we observed a near-linear increase in the numbers of amino acid changes over the time (Figure 1B), suggesting that there is an intrinsic time clock controlling the overall evolution of the viral genome. Similar patterns have also been observed with nucleotide substitutions (Supplementary information, Figure S2). However, when a co-mutation network [7] was constructed by analyzing pair-wise co-occurrence of amino acids within a single EBOV protein or between two different EBOV proteins (Figure 1C and Supplementary information, Figure S3), we found to our surprise the change in the numbers of amino acids involved in the co-mutation networks (denoted as nodes) occurs in a non-linear shape over the time (Figure 1D). Intriguingly, when considering the change in the numbers of connections in the co-mutation networks (denoted as edges), the non-linear shape is more evident with a trend of accelerated changes in the more recent outbreaks, in particular the 2014 outbreak (Supplementary information, Figure S4).

The past EBOV epidemics were reported to have CFRs ranging from 44% to 88% (Supplementary information, Figure S5 and Table S1). Due to typically clustered infection, short latency time and rapid development of severe disease symptom, the CFR of EVD can be reasonably estimated based on direct observation. In African countries, particularly because of the lack of access to medical care services, the observed CFR reported by WHO well-reflects the natural course of disease development mostly determined by the property and behavior of the virus itself. Thus, CFR could be regarded as an indicator of the virulence of the virus. To find out whether there exist significant correlations between CFR and genetic changes, we plotted CFRs (Supplementary information, Table S2) against the numbers of genetic changes in terms of either individual mutations or the network of co-mutations. The numbers of individual mutations do not have any significant correlation with CFRs (Figure 1E). However, plotting CFRs against the numbers of the nodes (Figure 1F) or edges (Supplementary information, Figure S6) of the co-mutation network of the whole viral genome


A
0.001
representative EBOV strain for each clade
SLE/2014/Makona-G3810.1
SLE/2014/Makona-G3823
SLE/2014/Makona-G3799
SLE/2014/Makona-G3788
SLE/2014/Makona-G3787
SLE/2014/Makona-EM119
SLE/2014/Makona-G3846
SLE/2014/Makona-G3670.1
SLE/2014/Makona-EM104
SLE/2014/Makona-G3686.1
SLE/2014/Makona-EM095B
GIN/2014/Makona-Kissidougou-C15
GIN/2014/Makona-Gueckedou-C05
GIN/2014/Makona-Gueckedou-C07
GAB/2002/Ilembe
COD/2007/23/Luebo
COD/2007/0/Luebo
COD/2007/1/Luebo
COD/2007/43/Luebo
034-KS/2008
COD/2007/9/Luebo
M-M/2008
GAB/1996/2Nza
GAB/1994/Gabon
GAB/1996/1Eko
COD/1995/13625 Kikwit
Kikwit/Zaire/1995
Mayinga/1976
C
NP
VP35
VP40
GP
VP30
VP24
L
Existed co-occurrent pairs of amino acids before 2014
Newly emerged co-occurrent pairs of amino acids in 2014
Co-occurrent pairs of amino acids differed from 2014 and before
B
R² = 0.924
Amino acid substitutions to Mayinga strain on whole proteome
Year
SLE/2014/Makona-EM104
GAB/2002/Ilembe
GAB/1996/1Eko
COD/2007/9/Luebo
COD/1995/13625 Kikwit
Mayinga/1976
D
R² = 0.885
Amino acid substitutions to Mayinga strain on co-mutation sites
Year
SLE/2014/Makona-EM104
COD/2007/9/Luebo
GAB/1996/1Eko
GAB/2002/Ilembe
COD/1995/13625 Kikwit
Mayinga/1976
E
Case fatality rate (%)
Amino acid substitutions to Mayinga strain on whole proteome
Mayinga/1976
COD/1995/ 13625 Kikwit
GAB/2002/Ilembe
GAB/1996/1Eko
COD/2007/9/Luebo
correlation= -0.6395
p-value = 0.2452
F
Case fatality rate (%)
Amino acid substitutions to Mayinga strain on co-mutation sites
Mayinga/1976
GAB/2002/Ilembe
COD/1995/ 13625 Kikwit
COD/2007/9/Luebo
GAB/1996/1Eko
correlation= -0.9568
p-value = 0.0107
G
Nucleocapsid formation
NP-VP40 interaction
NP-VP40 interaction
NP
GP1 glycan cap
Furin cleavage
Mucin-like domain
GP
* B epitope
+ T epitope
VP35-L binding
RdRp catalytic
SAM-dependent 2'-O-MTase
L
Domain I
Domain II
Domain III
Domain IV
Domain V
Domain VI
H
p-value=0.20
p-value=0.10
The proportion of epipote sites in selected positions
7 epitopes/8 positions
9 epitopes/18 positions
12 epitopes/21 positions
co-mutation sites (2014 specific)
co-mutation sites (previous outbreaks)
single mutation sites
I
Correlation with case fatality rate
NP-NP
VP35-VP35
VP40-VP40
GP-GP
VP30-VP30
VP24-VP24
L-L
NP-VP35
NP-VP40
NP-GP
NP-VP30
NP-VP24
NP-L
VP35-VP40
VP35-GP
VP35-VP30
VP35-VP24
VP35-L
VP40-GP
VP40-VP30
VP40-VP24
VP40-L
GP-VP30
GP-VP24
GP-L
VP30-VP24
VP30-L
VP24-L

gave rise to a striking linear regression curve with correlation factor of approximately –0.95 and *P* value of around 0.01. Such a striking correlation between co-mutation network and CFR was also observed when we used the partial correlation analysis to control the effect of time (Supplementary information, Data S1), applied a more stringent cutoff to remove highly similar sequences (Supplementary information, Tables S3-S7 and Data S2) or used an alternative method, the CAPS method, to detect correlated amino acid changes (Supplementary information, Figure S7 and Table S8). As the overall sample size is small, we performed a stringent leave-one-out cross validation test. The results showed that the predicted CFRs agreed well with the observed ones (Supplementary information, Figure S8), indicating that the linear relationship between the co-mutation network and the CFR is robust and reliable. Given the novel linear correlation revealed between a co-mutation network of EBOV genome and CFR, we predicted a CFR of ~52% for the 2014 EBOV outbreak. As the 2014 EBOV outbreak remains ongoing, it is difficult to accurately estimate its CFR based on the numbers of the infected cases and deaths collected during the epidemic progression (Supplementary information, Data S1). Our sequence-based approach provides a novel and rapid way of assessing the CFR of EBOV even at the very beginning of the outbreak once its genome sequence is available.

To gain further mechanistic insight, we subsequently mapped the co-mutation sites to each of the coding regions of EBOV (Figure 1G and Supplementary information, Figure S9 and Table S9). Clearly, the co-mutation sites form distinct clusters in all 7 EBOV-encoded proteins, most abundantly in NP, GP and L, which harbor multiple densely clustered regions. Furthermore, according to the annotation of functional domains extracted from literature, the co-mutation clusters are enriched in protein-protein interaction domains, including domains involved in self-assembly, oligomerization, inter-protein interaction/binding and membrane binding. Interestingly, the polymerase L contains co-mutation clusters overlapping with the RNA-dependent RNA polymerase (RdRp) catalytic domain. As for GP, which is located on the envelope of virus particles, all known B cell/antibody epitopes are localized in the densely clustered region of co-mutation area (Figure 1G), suggesting that EBOV is subjected to immune selection by antibodies during the course of their infection and transmission. Although in infected humans, the latency is short (only 3-21 days) and the infection is often catastrophic with virus destroying immune cells early in the infection, antibodies appeared to be induced in survivors and likely play certain protective roles in limiting EBOV infection and spreading. It is worthy of note that, for the 2014 Ebola outbreak, 7 of the 8 co-mutation sites on GP are located in the known epitope regions, while for all other 5 clades of ebolaviruses only 9 of 18 co-mutation sites on GP are located in the known epitope regions (Figure 1H). This may suggest that the 2014 EBOV strain has undergone stronger immune selection than previous EBOV strains.

Given the striking correlation between edges of the co-mutation network and CFRs, we further looked into the association of intragenic and intergenic sub-networks with CFR. Correlation analysis revealed that NP-GP and GP-L sub-networks showed highly significant correlation with CFR (Figure 1I; $P = 0.015$ for NP-GP and $P = 0.017$ for GP-L sub-networks). Supplementary information,

Figure 1 Co-mutation networks correlate with CFRs of EBOV. **(A)** The maximum likelihood tree of complete proteomes for 28 EBOV strains, with branches colored by clades. Values above branches represent percent support based on 1 000 maximum likelihood bootstrap trees. EBOV strains Mayinga/1976 (accession number: AF086833), COD/1995/13625 Kikwit (KC242796), GAB/1996/1E-ko (KC242793), COD/2007/9/Luebo (KC242784), GAB/2002/Ilembe (KC242800), and SLE/2014/Makona-EM104 (KM233035) are selected as representative strains for each of the six clades. **(B)** The number of amino acid substitutions of the whole proteome for each of the six clades relative to that of the Mayinga strain. The gray line is fitted with a linear model ($R^2 = 0.9240$). **(C)** The matrix representation of co-mutation network for 2014 EBOV outbreak. Co-occurrent pairs of amino acids existed before 2014 are coded in blue. Newly emerged pairs in 2014 are coded in red, and amino acid pairs differed between 2014 and previous outbreaks are coded in pink. **(D)** The number of amino acid substitutions on co-mutation sites for each of six clades relative to that of the Mayinga strain. The gray line is fitted with a non-parametric regression model ($R^2 = 0.8849$). **(E)** Correlation between CFR and the number of amino acid substitutions on the whole proteome (Pearson's correlation coefficient = −0.6395, $P = 0.2452$). **(F)** Correlation between CFR and the number of amino acid substitutions on co-mutation sites (Pearson's correlation coefficient = −0.9568, $P = 0.0107$). **(G)** Co-mutation sites on NP, GP and L proteins. Functional regions of these three proteins are drawn as purple horizontal bars. Co-mutation sites are indicated as gray vertical lines, and co-mutation sites specific to the SLE/2014/Makona-EM104 strain are highlighted in red. If a co-mutation site locates in an epitope region, it is marked with '*' to indicate B epitope or '+' to indicate T epitope. Epitope sites specific to the SLE/2014/Makona-EM104 strain are also highlighted in red. **(H)** The proportion of co-mutation sites on GP that are located in epitope regions for 2014 and previous EBOV outbreaks. Data for single mutations on GP over all outbreaks were also shown. *P* values were calculated by Fisher's exact test. **(I)** Correlations of CFR with intragenic and intergenic co-mutation numbers. $^*P < 0.05$.

Figure S10 shows the dynamic change in edge numbers of intra/intergenic sub-networks during EBOV evolution. Clearly, the co-mutations between the four proteins L, NP, VP40 and GP are highly enriched in 2014 EBOV. L, NP and VP40 are all involved in the viral replication and RNA synthesis [8], and are likely to interact and form a complex during viral replication process [9]. In addition, NP has been reported to be an important molecular determinant for EBOV virulence [10]. Thus, these results demonstrate that adaptive change in viral RNA synthesis and genome replication by coordinated and compensatory mutations in multiple protein components could be one of the major contributing factors to the virulence alteration during the natural course of EBOV disease development. Surprisingly, although the number of edges within the GP-GP intragenic sub-network does not exhibit any significant contribution to the CFR, the number of edges in the GP-L sub-network significantly correlates with the CFR with a correlation factor of −0.941 and a P value of 0.017 (Supplementary information, Figure S11). It is not clear how GP function is associated viral RNA synthesis and genome replication activity. Currently there is no evidence that GP has any direct interaction with L protein, and the mechanism of their potential functional association needs to be elucidated in future studies. Nevertheless the superb correlation of the GP-L co-mutation sub-network with CFR may provide a simplified tool for predicting the trend of EBOV virulence by directly sequencing the GP and L regions of the viral genome. Taken together, our study not only reveals a novel sequence-based approach to assessing the severity of EBOV outbreaks but also provides a new insight into the dynamic network of co-evolving amino acids underlying the virus evolution.

Acknowledgments

We sincerely thank professors Barry Bloom and Marc Lipsitch (Harvard School of Public Health) for discussions, suggestions, and reading of the manuscript. This study was supported by the National Basic Research Program of China (2015CB910501), the Major National Earmark Project for Infectious Diseases (2014ZX10004002-001), and the Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L09-1-2) to TJ, and by the National Major Scientific and Technological Special Project for "Significant New Drugs Development" (2015ZX09102023) to GC.

Lizong Deng[1, *], Mi Liu[2, 3, *], Sha Hua[2, 3, *], Yousong Peng[4], Aiping Wu[2], F Xiao-Feng Qin[1], Genhong Cheng[1, 5], Taijiao Jiang[1, 2]

[1]Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005; Suzhou Institute of Systems Medicine, Suzhou, Jiangsu 215123, China; [2]Key Laboratory of Protein and Peptide Pharmaceuticals, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China; [3]University of the Chinese Academy of Sciences, Beijing 100049, China; [4]College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China; [5]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, CA 90095, USA

*These three authors contributed equally to this work.

Correspondence: Taijiao Jiang[a], Genhong Cheng[b], F Xiao-Feng Qin[c]

[a]E-mail: taijiao@moon.ibp.ac.cn

[b]E-mail:gcheng@mednet.ucla.edu

[c]E-mail: fqin1@foxmail.com

References

1 King AM, eds. Virus taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses. London: Academic Press, 2012.

2 Ebola haemorrhagic fever in Zaire, 1976. *Bull World Health Organ* 1978; **56**:271-293.

3 Rieger T, Koivogui L, Magassouba NF, *et al. N Engl J Med* 2014; **371**:1418-1425.

4 Sanchez A, Kiley MP, Holloway BP, *et al. Virus Res* 1993; **29**:215-240.

5 Maglott D, Ostell J, Pruitt KD, *et al. Nucleic Acids Res* 2005; **33**:D54-D58.

6 Gire SK, Goba A, Andersen KG, *et al. Science* 2014; **345**:1369-1372.

7 Du X, Wang Z, Wu A, *et al. Genome Res* 2008; **18**:178-187.

8 Lamb RA, Parks GD. Paramyxoviridae: The viruses and their replication. In: Knipe DM, Howley PM, eds. Fields Virology. Wolters Kluwer: Lippincott WIlliams & Wilkins, 2007:1449-1496.

9 Noda T, Hagiwara K, Sagara H, *et al. J Gen Virol* 2010; **91**:1478-1483.

10 Ebihara H, Takada A, Kobasa D, *et al. PLoS Pathog* 2006; **2**:e73.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)