Bioinformatics, 31(15), 2015, 2421–2425
doi: 10.1093/bioinformatics/btv189
Advance Access Publication Date: 2 April 2015
Discovery Note

OXFORD

Genome analysis

Three minimal sequences found in Ebola virus genomes and absent from human DNA

Raquel M. Silva[1,*,†], Diogo Pratas[1,2,†], Luísa Castro[1], Armando J. Pinho[1,2] and Paulo J. S. G. Ferreira[1,2]

[1]IEETA and [2]DETI, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on December 11, 2014; revised on March 26, 2015; accepted on March 27, 2015

Abstract

Motivation: Ebola virus causes high mortality hemorrhagic fevers, with more than 25 000 cases and 10 000 deaths in the current outbreak. Only experimental therapies are available, thus, novel diagnosis tools and druggable targets are needed.
Results: Analysis of Ebola virus genomes from the current outbreak reveals the presence of short DNA sequences that appear nowhere in the human genome. We identify the shortest such sequences with lengths between 12 and 14. Only three absent sequences of length 12 exist and they consistently appear at the same location on two of the Ebola virus proteins, in all Ebola virus genomes, but nowhere in the human genome. The alignment-free method used is able to identify pathogen-specific signatures for quick and precise action against infectious agents, of which the current Ebola virus outbreak provides a compelling example.
Availability and Implementation: EAGLE is freely available for non-commercial purposes at http://bioinformatics.ua.pt/software/eagle.
Contact: raquelsilva@ua.pt; pratas@ua.pt
Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ebola virus (EBOV) is a negative strand-RNA virus from *the Filoviridae* family that causes high mortality hemorrhagic fevers, for which no vaccine or treatment currently exist (Sarwar *et al.*, 2014). There are five *Ebolavirus* species, namely, *Zaire ebolavirus*, *Sudan ebolavirus*, *Bundibugyo ebolavirus*, *Tai Forest ebolavirus* and *Reston ebolavirus*, with the first (1976) and major (2014) outbreaks caused by the type species *Zaire ebolavirus* (Baize *et al.*, 2014).

The numbers of the largest ever EBOV outbreak are worrying and continue escalating, with over 25 000 cases and 10 000 deaths from the virus mainly in Guinea, Liberia and Sierra Leone, according to the World Health Organization. The current outbreak is also the first where transmission has occurred outside Africa, with reported cases in Europe (Spain) and America (USA; Butler and Morello, 2014). Promising vaccine candidate tests are being rushed to face the epidemics and could be available within a few months (Gulland, 2014). These yet experimental therapies include, for example, recombinant viral vectors (Jones *et al.*, 2005) or antibodies that target the viral glycoprotein (GP; Friedrich *et al.*, 2012; Sarwar *et al.*, 2014), but innovative approaches are still needed for the development of diagnosis tools and identification of druggable targets.

Minimal absent words are the shortest sequence fragments that are not present in the genomic data of a given organism. They have been studied before to describe properties of prokaryotic and eukaryotic genomes and to develop methods for phylogeny construction or PCR primer design (Chairungsee and Crochemore, 2012; Falda *et al.*, 2014; Garcia *et al.*, 2011; Herold *et al.*, 2008; Pinho *et al.*, 2009; Wu *et al.*, 2010). Here, we introduce minimal relative absent words (RAW), a concept which has not been used so far in the context of personalized medicine, but which is deemed useful for differential identification of sequences that are derived from a pathogen genome but absent from its host.

We use the current EBOV outbreak sequences, which were recently published (Gire *et al.*, 2014), to discover and characterize the minimal RAWs that are present in EBOV genomes but absent from the human genome. Moreover, we show that these words are also absent from the other *Ebolavirus* species and even from the genomes obtained from previous outbreaks. Thus, the sequences that we identify are species-specific and important for future development of diagnosis or therapeutic strategies for EBOV. The method that we introduce can be applied to other emerging pathogens or to show evidence of evolutionary patterns and signatures across species.

2 Methods

2.1 Relative absent words

Consider a target sequence (e.g. a virus sequence), x, and a reference sequence (e.g. the human genome), y, both drawn from the finite alphabet $\sum = \{A, C, G, T\}$. We say that α is a factor of x if x can be expressed as $x = u\alpha v$, with uv denoting the concatenation between sequences u and v.

We denote by $W_k(x)$ the set of all k-size words (or factors) of x. Also, we represent the set of all k-size words *not in* x as $\overline{W_k(x)}$. For each word size k, we define the set of all words that exist in x but do not exist in y by

$$R_k(x, \bar{y}) = W_k(x) \cap \overline{W_k(y)} \quad (1)$$

and the subset of words that are minimal, in the sense presented in Pinho *et al.* (2009), as

$$M_k(x, \bar{y}) = \{\alpha \in R_k(x, \bar{y}) : W_{k-1}(\alpha) \cap M_{k-1}(x, \bar{y}) = \phi\} \quad (2)$$

i.e. a minimal absent word of size k cannot contain any minimal absent word of size less than k. In particular, $l\alpha r$ is a minimal absent word of sequence x, where l and r are single letters from $\sum$, if $l\alpha r$ is not a word of x but both $l\alpha$ and αr are (Pinho *et al.*, 2009). In this work, we were particularly interested in the non-empty set $M_k(x, \bar{y})$ corresponding to the smallest k. These are referred as RAWs.

2.2 Protein structural models

Protein 3D structural models were built by homology modeling as previously described (Duarte-Pereira *et al.*, 2014). Appropriate templates were selected from PDB (www.rcsb.org; Berman *et al.*, 2000), where several nucleoprotein (NP) structures from viruses within *Mononegavirales* (negative-sense genome single-stranded RNA viruses) are available, whereas for the region of interest in L-protein only structures from more distant viruses exist.

Structures from the Nipah virus NP (PDB ID:4CO6; Yabukarski *et al.*, 2014) and the BVDV (bovine viral diarrhea virus) RNA polymerase (PDB ID:1S48; Choi *et al.*, 2004) were used as templates in MODELLER (Eswar *et al.*, 2006; Sali and Blundell, 1993), to predict the structure of the N-terminal regions of Ebola virus NP (residues 1–380) and RNA-polymerase (residues 177–805), respectively (Supplementary Figs. S3 and S4). Accuracy of the predicted models (Supplementary Fig. S5) was estimated using ProSA-web (https://prosa.services.came.sbg.ac.at/prosa.php; Sippl, 1993; Wiederstein and Sippl, 2007) and structures were visualized with PyMOL (Schrodinger, 2010).

3 Results

To identify RAWs, we have developed the EAGLE tool that implements the method described above (Supplementary Data). We have used the full GRC-38 human reference genome (Church *et al.*, 2011) downloaded from the NCBI, including the mitochondrial, unplaced and unlocalized sequences. The sequences of 99 EBOV genomes from the current outbreak in Sierra Leone (Gire *et al.*, 2014) and additional 66 *Ebolavirus* genomes have been also downloaded from NCBI (Supplementary Table S1). The code used in this analysis is available (Pratas, 2015).

Figure 1 shows the computation for word sizes 12, 13 and 14 (for computer characteristics see Supplementary Section Software and Hardware). As expected, the number of absent words decreases as the k-mer size decreases. Specifically, for $k = 11$ (not represented), there are no EBOV RAW. On the other hand, for $k = 12$, three groups of points emerge (RAW1, RAW2 and RAW3) representing the position of a RAW in each of the 99 unaligned viral genomes (Fig. 1a).

Alignments of 124 *Ebolavirus* sequences (Gire *et al.*, 2014), including additional EBOV genomes from the current outbreak in Guinea (Baize *et al.*, 2014) and from previous outbreaks, show that the identified minimal RAWs fall into conserved protein regions (Fig. 1b). However, several mutations can be found in the genomes that discriminate between the different species of *Ebolavirus* and even between EBOV sequences from the current and previous outbreaks (Fig. 1c). The identification of these viral genome signatures is important for quick diagnosis in outbreak scenarios. Additional analysis with all 165 *Ebolavirus* genomes confirmed these results (Supplementary Fig. S1). In particular, RAW1 is conserved within EBOV and can distinguish EBOV from other *Ebolavirus* species. RAW2 is conserved in all sequences from the West African 2014 outbreak in Guinea, Sierra Leone and Liberia, and only one nucleotide difference exists between these sequences and unrelated outbreak genomes. RAW3 is also conserved at the species level, excluding the four EBOV 1976/77 genomes, and can distinguish between all *Ebolavirus* species (Supplementary Fig. S2).

From the three EBOV sequence motifs absent in the human genome, the first (RAW1) is included in the virus NP, while the other two (RAW2 and RAW3) fall within the sequence of the viral RNA-polymerase (L-protein; Fig. 1c). Previous studies show that the N-terminal region of EBOV NP participates in both the formation of nucleocapsid-like structures through NP–NP interactions and in the replication of the viral genome (Watanabe *et al.*, 2006), and RAW1 sequence (TTTCGCCCGACT) is part of this N-terminal region. The L-protein (LP) produces the viral transcripts to be translated by host ribosomes and is involved in the replication of the viral genome as well. The LP contains the two additional minimal RAWs, RAW2 (TACGCCCTATCG) and RAW3 (CCTACGCGCAAA).

Both NP and LP are critical for the virus life cycle and constitute good targets for therapeutic intervention. Screening for new antiviral compounds could benefit from knowledge of their protein structures. For EBOV, most protein structures are unknown except for the C-terminal domain of NP, GP, VP24 and VP35 (Shurtleff *et al.*, 2012), thus, we have predicted the structure of the N-terminal regions of the EBOV NP and LP by homology modeling (Supplementary Figs. S3–S5). These structural models show that the amino acids corresponding to the RAW1 motif are enclosed within the structure, while RAW2 and RAW3 are exposed at the protein surface, which can justify its higher degree of conservation.

4 Discussion

The personalized medicine field is now closer to clinical practice with the advances of next-generation sequencing technologies. Personalized therapeutics are a possibility and their development is essential with the emergence of resistance to current available drugs. Additionally, quick diagnosis is required for emerging pathogens and in epidemics

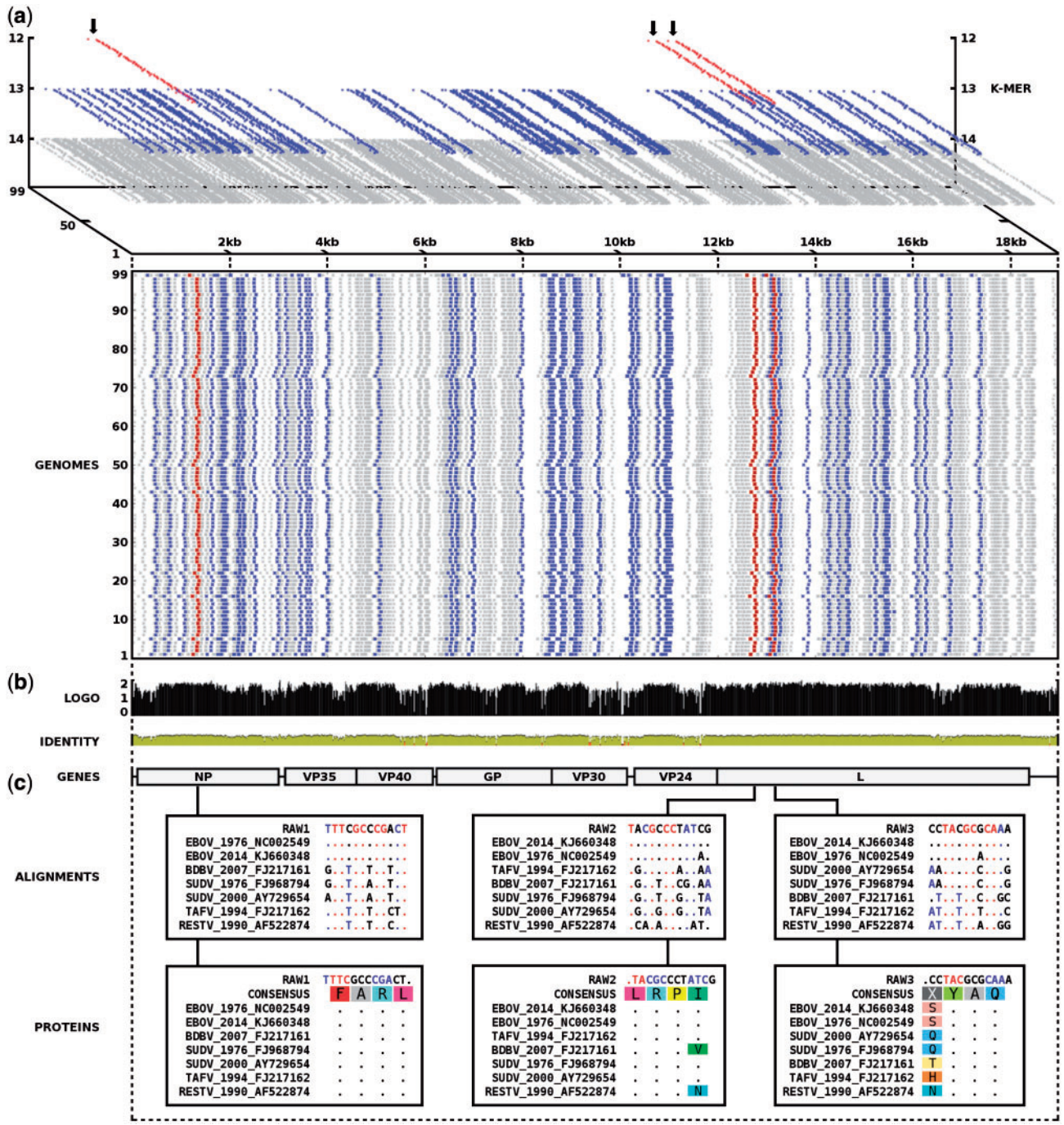

Fig. 1. Ebola virus minimal absent words relatively to the human complete genome. (**a**) RAWs were identified in 99 unaligned genomes from the current outbreak in Sierra Leone (2014) and are highlighted in red ($k=12$, arrows), blue ($k=13$) and grey ($k=14$). (**b**) Whole genome alignments from 124 published *Ebolavirus* genomes were obtained from Gire *et al.* (2014) and visualized in Geneious (created by Biomatters, available from http://www.geneious.com). Sequence logos and identity define conserved regions. (**c**) Regions corresponding to the identified RAWs are shown in genome location and both as nucleic acid and protein alignments. The *Ebolavirus* reference genomes are displayed, as well as selected representative sequences where nucleotide differences are observed

such as the current Ebola outbreak. Here, we have detected minimal RAWs in the human genome that are present in EBOV genomes, and identified nucleotide differences in some of these sequences that can distinguish between *Ebolavirus* species and outbreaks. Also, we show that the corresponding amino acid sequences are conserved within EBOV. These results can now be further explored for diagnosis and therapeutics, sometimes mentioned as theranostics (Picard and Bergeron, 2002). Namely, RAW nucleotide sequences can be used in diagnosis to design primers that identify *Ebolavirus* infections or distinguish between *Ebolavirus* species. For PCR-based methods, longer sequences and multiplex reactions can be developed to avoid primer binding bias. Additional nucleotide or protein-based strategies for therapeutics can be envisaged, as discussed below.

One problem in developing efficient EBOV treatments is the virus ability to evade the immune system. The viral GP is a major target because it mediates attachment and entry into the host cells. However, in addition to the surface envelope protein, the GP gene also produces fragment, soluble GPs that are secreted and direct the

immune system to produce antibodies for variable and non-essential regions of the virus (Cook and Lee, 2013; Mohan *et al.*, 2012). As current efforts based on the viral GP might prove ineffective, additional targets should be sought. Our results show that the viral NP and polymerase (LP) can be attractive targets. As the amino acid sequences of all three 12-mer RAWs are conserved within EBOV, these regions can be used to screen for small molecule inhibitors. In particular, RAW1 is conserved in all *Ebolavirus* NP proteins, which can indicate a functional or structural role. And, considering that the protein model predicts that RAW2 and RAW3 are relatively close in the 3D structure and in exposed domains, these regions can be used to develop novel antibodies. Also, a recently described mechanism shows that the polymerase (LP) from Ebola and Marburg viruses is capable of editing transcripts, resulting in increased variability in the produced proteins, and that the most edited mRNAs are the Ebola GP and Marburg NP and LP itself (Shabman *et al.*, 2014). Thus, the use of combined therapies towards multiple proteins can be more effective, as suggested by studies to develop vaccines for Lassa virus that target both NP and GP (Fisher-Hoch *et al.*, 2000; Lukashevich, 2012).

RNA-based strategies such as RNA interference (RNAi) or antissense therapies are also promising approaches to silence target-specific gene expression. The RAW sequences that we have identified can be used to develop RNAi or antisense probes that bind viral transcripts and prevent their translation, thus, inhibiting viral replication without blocking the host mRNAs. Translation of these technologies into clinical applications have been slowed by challenges in the delivery of small RNAs into cells, but recent developments in delivery systems are bridging the bench to bedside gap (Hayden, 2014; Yin *et al.*, 2014). Among these, gold or lipid nanoparticles (Conde *et al.*, 2014; Draz *et al.*, 2014) were shown to be effective against cancer and viral infections, including EBOV (Geisbert *et al.*, 2010). Gold-nanobeacons can be applied as a combined diagnosis and therapy tool for effective testing, including in low-cost settings (Costa *et al.*, 2014) and, with this purpose, advances in peptide nucleic acid probes for viral detection are also taking place (Joshi *et al.*, 2013; Zhang *et al.*, 2010).

Whichever the technology, the identification of genome signatures for rapid evolving species such as Ebola viruses will be useful for the development of both diagnosis and therapeutics.

Funding

This work was supported by the European Fund for Regional Development (FEDER) through the Operational Program Competitiveness Factors (COMPETE) and by the Portuguese Foundation for Science and Technology (FCT), in the context of projects PEst-OE/EEI/UI0127/2014 and Incentivo/EEI/UI0127/2014, by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 'RD-Connect: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research', and the project Neuropath (CENTRO-07-ST24-FEDER-002034), co-funded by QREN "Mais Centro" program and the EU.

Conflict of Interest: none declared.

References

Baize,S. *et al.* (2014) Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.*, **371**, 1418–1425.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Butler,D. and Morello,L. (2014) Ebola by the numbers: the size, spread and cost of an outbreak. *Nature*, **514**, 284–285.

Chairungsee,S. and Crochemore,M. (2012) Using minimal absent words to build phylogeny. *Theor. Comput. Sci.*, **450**, 109–116.

Choi,K.H. *et al.* (2004) The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation. *Proc. Natl. Acad. Sci. USA.*, **101**, 4425–4430.

Church,D. *et al.* (2011) Modernizing reference genome assemblies. *Plos Biol.*, **9**, e1001091.

Conde,J. *et al.* (2014) Gold-nanobeacons for gene therapy: evaluation of genotoxicity, cell toxicity and proteome profiling analysis. *Nanotoxicology*, **8**, 521–532.

Cook,J. and Lee,J. (2013) The secret life of viral entry glycoproteins: moonlighting in immune evasion. *Plos Pathogens*, **9**, e1003258.

Costa,M.N. *et al.* (2014) A low cost, safe, disposable, rapid and self-sustainable paper-based platform for diagnostic testing: lab-on-paper. *Nanotechnology*, **25**, 094006.

Draz,M.S. *et al.* (2014) Nanoparticle-mediated systemic delivery of siRNA for treatment of cancers and viral infections. *Theranostics*, **4**, 872–892.

Duarte-Pereira,S. *et al.* (2014) NAMPT and NAPRT1: novel polymorphisms and distribution of variants between normal tissues and tumor samples. *Sci. Rep.*, **4**, 6311.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics.*, **Chapter 5**, Unit 5.6.

Falda,M. *et al.* (2014) keeSeek: searching distant non-existing words in genomes for PCR-based applications. *Bioinformatics*, **30**, 2662–2664.

Fisher-Hoch,S.P. *et al.* (2000) Effective vaccine for lassa fever. *J. Virol.*, **74**, 6777–6783.

Friedrich,B.M. *et al.* (2012) Potential vaccines and post-exposure treatments for filovirus infections. *Viruses*, **4**, 1619–1650.

Garcia,S. *et al.* (2011) Minimal absent words in prokaryotic and eukaryotic genomes. *Plos One*, **6**, e16065.

Geisbert,T.W. *et al.* (2010) Postexposure protection of non-human primates against a lethal Ebola virus challenge with RNA interference: a proof-of-concept study. *Lancet*, **375**, 1896–1905.

Gire,S. *et al.* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, **345**, 1369–1372.

Gulland,A. (2014) Clinical trials of Ebola therapies to begin in December. *BMJ*, **349**, g6827.

Hayden,E.C. (2014) RNA interference rebooted. *Nature*, **508**, 443.

Herold,J. *et al.* (2008) Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, **9**, 167.

Jones,S.M. *et al.* (2005) Live attenuated recombinant vaccine protects nonhuman primates against Ebola and Marburg viruses. *Nat. Med.*, **11**, 786–790.

Joshi,V.G. *et al.* (2013) Rapid label-free visual assay for the detection and quantification of viral RNA using peptide nucleic acid (PNA) and gold nanoparticles (AuNPs). *Anal. Chim. Acta*, **795**, 1–7.

Lukashevich,I.S. (2012) Advanced vaccine candidates for Lassa fever. *Viruses*, **4**, 2514–2557.

Mohan,G. *et al.* (2012) Antigenic subversion: a novel mechanism of host immune evasion by Ebola virus. *Plos Pathogens*, **8**, e1003065.

Picard,F.J. and Bergeron,M.G. (2002). Rapid molecular theranostics in infectious diseases. *Drug Discov. Today,* **7**, 1092–1101.

Pinho,A. *et al.* (2009) On finding minimal absent words. *BMC Bioinformatics*, **10**, 137.

Pratas,D. (2015) eagle: EAGLE v1.1. Zenodo. 10.5281/zenodo.15521.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sarwar,U.N. *et al.* (2014) Safety and immunogenicity of DNA vaccines encoding ebolavirus and marburgvirus wild-type glycoproteins in a phase I clinical trial. *J. Infect. Dis.*, **211**, 549–557.

Schrodinger,L. (2010) The PyMOL molecular graphics system, version 1.3 r1. Py-MOL, The PyMOL Molecular Graphics System, Version 1.

Shabman,R.S. *et al.* (2014) Deep sequencing identifies noncanonical editing of Ebola and Marburg virus RNAs in infected cells. *MBio*, **5**, e02011–e02014.

Shurtleff,A. *et al.* (2012) Therapeutics for filovirus infection: traditional approaches and progress towards in silico drug design. *Expert Opin. Drug Discov.*, **7**, 935–954.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.

Watanabe,S. *et al*. (2006) Functional mapping of the nucleoprotein of Ebola virus. *J. Virol.*, **80**, 3743–3751.

Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407–W410.

Wu,Z. *et al*. (2010) Efficient computation of shortest absent words in a genomic sequence. *Inf. Process. Lett.*, **110**, 596–601.

Yabukarski,F. *et al.* (2014) Structure of Nipah virus unassembled nucleoprotein in complex with its viral chaperone. *Nat. Struct. Mol. Biol.*, **21**, 754–759.

Yin,H. *et al*. (2014) Non-viral vectors for gene-based therapy. *Nat. Rev. Genet.*, **15**, 541–555.

Zhang,N. and Appella,D.H. (2010) Advantages of peptide nucleic acids as diagnostic platforms for detection of nucleic acids in resource-limited settings. *J. Infect. Dis.*, **201**, S42–S45.