Comparative analysis of the L, M and S RNA segments of Crimean-Congo haemorrhagic fever virus isolates from southern Africa

Dominique Goedhals[1], Phillip A. Bester[1], Janusz T. Paweska[2,3], Robert Swanepoel[4], Felicity J. Burt[1*]

[1]Department of Medical Microbiology and Virology, National Health Laboratory Service/University of the Free State, Bloemfontein, South Africa

[2]Center for Emerging and Zoonotic Pathogens, National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa

[3]School of Pathology, Faculty of Health Sciences, University of the Witswatersrand, South Africa

[4]Zoonoses Research Unit, Department of Medical Virology, University of Pretoria, South Africa

Institution at which work was performed:

Department of Medical Microbiology and Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, 9301, South Africa

*Corresponding author: Felicity Burt

Department of Medical Microbiology and Virology, Faculty of Health Sciences, National Health Laboratory Service/University of the Free State, Bloemfontein, 9301, South Africa

Email: burtfj@ufs.ac.za

Telephone: + 27 51 405 3162

Fax: +27 51 4442433

Running head: Comparative analysis of the CCHFV genome

Abstract

Crimean-Congo haemorrhagic fever virus (CCHFV) is a member of the *Bunyaviridae* family with a tripartite, negative sense RNA genome. This study used predictive software to analyse the L (large), M (medium) and S (small) segments of 14 southern African CCHFV isolates. The OTU-like cysteine protease domain and the RdRp domain of the L segment are highly conserved among southern African CCHFV isolates. The M segment encodes the structural glycoproteins, G_N and G_C, and the non-structural glycoproteins which are post-translationally cleaved at highly conserved furin and subtilase SKI-1 cleavage sites. All of the sites previously identified were shown to be conserved among southern African CCHFV isolates. The heavily O-glycosylated N-terminal variable mucin-like domain of the M segment shows the highest sequence variability of the CCHFV proteins. Five transmembrane domains are predicted in the M segment polyprotein resulting in three regions internal to and three regions external to the membrane across the G_N, NS_M and G_C glycoproteins. The corroboration of conserved genome domains and sequence identity among geographically diverse isolates may assist in the identification of protein function and pathogenic mechanisms, as well as the identification of potential targets for antiviral therapy and vaccine design. As detailed functional studies are lacking for many of the CCHFV proteins, identification of functional domains by prediction of protein structure and identification of amino acid level similarity to functionally characterised proteins of related viruses or viruses with similar pathogenic mechanisms are a necessary step for selection of areas for further study.

Key words: Protein domain, RNA-dependant RNA polymerase, glycoprotein

Introduction

Crimean-Congo haemorrhagic fever virus (CCHFV) is a tick-borne virus belonging to the genus *Nairovirus* in the family *Bunyaviridae*. Other genera in this family are *Orthobunyavirus, Hantavirus, Phlebovirus* and *Tospovirus* [Nichol et al., 2006]. The single-stranded, negative sense RNA genomes of the *Bunyaviridae* consist of three segments namely the small (S), medium (M) and large (L) segments [Clerx et al., 1981]. The CCHFV S segment encodes the nucleocapsid protein which encapsidates both viral RNA (vRNA) and complementary RNA (cRNA) to form ribonucleoprotein complexes. The vRNA serves as a template for both mRNA and cRNA, while the cRNA then serves as template for further vRNA [Bergeron et al., 2010]. The nucleocapsid protein has a racket-shaped structure consisting of a head or globular domain and a stalk domain which are both composed predominantly of α-helices [Carter et al., 2012, Guo et al., 2012]. In addition to binding RNA for the formation of the ribonucleoprotein complexes, the nucleocapsid protein also shows endonuclease activity attributed to the head domain [Guo et al., 2012]. The polyprotein encoded by the M segment is co- and post-translationally cleaved into the two structural glycoproteins, G_N and G_C, and three non-structural proteins, NS_M, a mucin-like domain and GP38 [Sanchez et al., 2006, Altamura et al., 2007]. Two further non-structural glycoproteins, GP85 and GP160, contain GP38 and the mucin-like variable domain. These four non-structural glycoproteins are likely released from infected cells as secretory products [Sanchez et al., 2006]. The L segment encodes the RNA-dependent RNA polymerase (RdRp) which is responsible for the synthesis of both mRNA for translation into proteins and cRNA for genome replication [Honig et al., 2004].

CCHFV infection in humans can result in haemorrhagic fever with a fatality rate of up to 30%. In recent years, the distribution of disease has expanded to numerous countries in Africa, Asia, the Middle East and Europe [Bente et al., 2013]. With more than 7000 cases identified in Turkey since the emergence of CCHFV in this country (Maltezou et al., 2010), this likely represents true expansion of the virus to new endemic regions as the high number of cases cannot be attributed to increased surveillance and improved diagnostic capacity. The expanding endemicity in the absence of an effective vaccine or antiviral therapy has prioritized CCHFV research to identify possible conserved targets for such interventions. Antiviral drugs targeting enzymes involved in viral replication are now available for viral infections such as human immunodeficiency virus (HIV), the herpes viruses and hepatitis C. Molecules inhibiting viral adsorption and entry into host cells, such as the CCR-5 inhibitors used for the treatment of HIV, and immunotherapy for pre- and post-exposure prophylaxis are other approaches which have been shown to be effective for managing diverse viral infections. If similar approaches are to be followed for the management of CCHFV infection, then detailed information regarding viral proteins and their functions in geographically distinct isolates will be required. A thorough understanding of the pathogenesis of CCHFV disease is also paramount and this may, in part, be elucidated by the comparison of gene products and protein functions of other viral haemorrhagic fevers such as Ebola and Lassa viruses. In this study, nucleotide and deduced amino acid sequences were analysed using a range of predictive software to identify conserved domains, O- and N-linked glycosylation sites, transmembrane helices, cleavage sites and genetic distances among southern African CCHFV isolates.

Materials and Methods

Sequence data set

Complete genome sequences for 14 southern African CCHFV isolates were included in the analysis. The sequence data for ten of the CCHFV isolates were determined in a previous study , namely SPU431/85, SPU383/87, SPU556/87, SPU18/88, SPU45/88, SPU497/88, SPU130/89, SPU48/90, SPU187/90 and SPU44/08, in which genetic diversity was determined and segment reassortment confirmed using phylogenetic analysis [Goedhals et al., 2014]. A further four complete sequences were retrieved from GenBank, namely SPU415/85, SPU97/85, SPU103/87 and SPU4/81. All of the CCHFV isolates were from humans, with one isolate originating in Namibia while the remaining isolates were from South Africa including the Free State, Northern Cape and North West Provinces. Sequences were aligned using ClustalX version 2.0 [Larkin et al., 2007] and manually edited using BioEdit version 7.2.3 (available at http://www.mbio.ncsu.edu/bioedit/bioedit.html).

Sequence analysis

Genome cyclization due to interactions between the complementary 5' and 3' non-coding regions (NCR) was investigated by joining the 5' NCR with an approximately equivalent number of bases from the relevant 3' NCR using a 50 base poly-A spacer [Khromykh et al., 2001]. The poly-A linked construct for each segment was analyzed using the Mfold Web Server (available at http://mfold.rna.albany.edu/?q=mfold) with default parameters and folding predictions at 37^{o}C [Zuker, 2003]. Sequence diversity within specific motif regions was calculated with Molecular Evolutionary Genetics Analysis v5 (MEGA5) using the *P* distance option [Tamura et al., 2011].

Protein analysis

The deduced amino acid sequences of the L segment were submitted to InterProScan 4 [Zdobnov and Apweiler, 2001] and PSI-BLAST [Altschul et al., 1997] to identify functional sites and conserved protein domains. Prediction of transmembrane helices was performed on the M segment amino acid sequences using TMHMM 2.0 [Krogh et al., 2001] and signal sequence cleavage sites were predicted using SignalP version 4.1 [Petersen et al., 2011]. Potential sites of N-linked and O-linked glycosylation were determined using N-Glycosite (available at www.hiv.lanl.gov/sequence/GLYCOSITE/glycosite.html) [Zhang et al., 2004] and NetOGlyc 4.0 [Steentoft et al., 2013] respectively. Visual inspection of sequences was performed using BioEdit.

Results

Complete genome sequences

The characteristics of the southern African CCHFV complete genome sequences are summarized in Table I. The complete L segment, including the 5' and 3' NCR, ranged from 12157 – 12170 nucleotides in length for the southern African CCHFV isolates. The open reading frame (ORF) was 11838 nucleotides in length therefore encoding a protein of 3945 amino acid residues (approximate nucleotide position 77 – 11911). The variability in length of the L segment was chiefly due to variations in size of the 3' NCR (243 – 256 nucleotides) while the 5' NCR (76 nucleotides) and ORF were conserved in length. The 3945 amino acid ORF of the L segment had a GC content of 41%.

Table I : Southern African CCHFV genome characteristics.

Characteristic	**S segment**	**M segment**	**L segment**
Length (nucleotides)	1671 - 1673	5344 – 5364	12155 – 12170
% GC content	45.9 – 47.7	43.4 – 45.0	41.2 – 41.4
5' NCR length – nucleotides	54 – 55	77 – 93	76
ORF length – nucleotides (amino acids)	1449 (482)	5055 - 5070 (1684 – 1689)	11832 - 11838 (3943 - 3945)
3' NCR length - nucleotides	168 - 169	198 – 218	243 - 256

The complete M segment, including the 5' and 3' NCR, ranged from 5344 – 5365 nucleotides in length. The M segment 5' and 3' NCR were more variable than the L segment, at 77 – 91 nucleotides and 198 – 218 nucleotides in length respectively. The M segment coding region was between 5055 – 5070 nucleotides in length, encoding a polyprotein of 1684 – 1689 amino acid residues. The GC content of the M segments ranged from 43.4 – 45.0%.

The S segment ORF encoded a single nucleoprotein of 482 amino acid residues (1449 nucleotides) in length for all available southern African CCHFV isolates. The complete S segment was 1671 nucleotides long, including a 5'NCR of 55 nucleotides and a 3'NCR of 169 nucleotides and with a GC content of 45.9 – 47.7%.

5' and 3' NCR cyclization analysis

Mfold analysis confirmed that the complementary 5' and 3' UTR of each segment allow for cyclization to form a panhandle RNA structure. Figure 1 shows representative examples of the three most stable structures according to the estimated change in Gibbs free energy predicted by Mfold and ranged from -37.17 to -51.91 using SPU97/85 as an example.

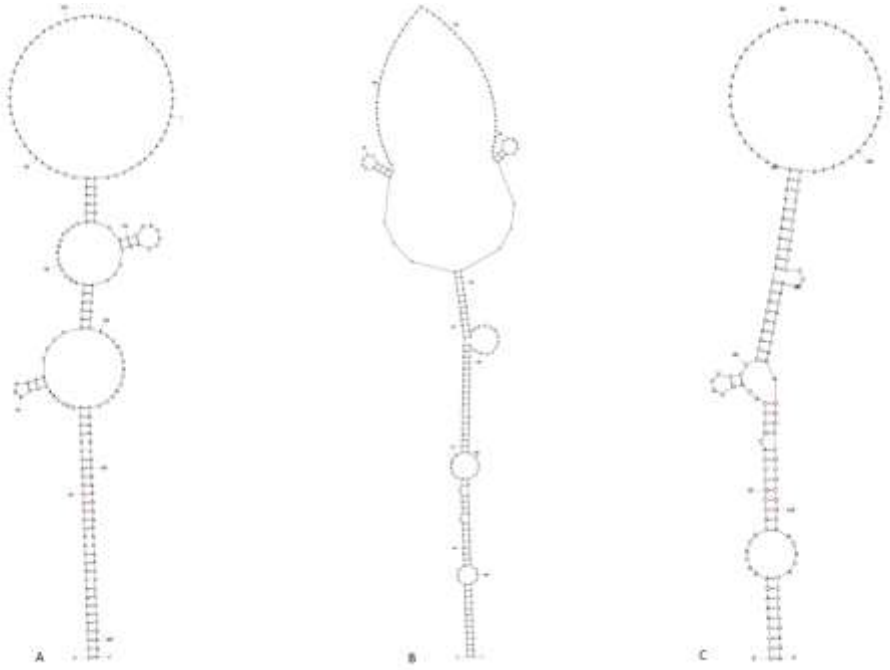

Figure 1 : 5' and 3' NCR complementary regions forming panhandle structures as modelled using Mfold Web Server. A) L segment, dG = -40.48. B) M segment, dG = -51.91. C) S segment, dG = -37.17.

Molecular characterization of the L segment ORF

Conserved domains identified by InterProScan and PSI-BLAST analysis included an N-terminal cysteine-protease motif of the ovarian tumour (OTU) protein superfamily and the RNA-dependent RNA polymerase (RdRp) motif. The OTU domain was located at approximate amino acid residues 29 – 158 of the L segment ORF and showed little sequence diversity, with a maximum of two amino acid differences between isolates over the 118 amino acid region and calculated p-distances of 0 – 1.7%. The catalytic triad of the CCHFV OTU-like cysteine protease is predicted to contain cysteine, histidine and aspartic acid residues, initially identified potentially as G**D**GN**C**FYHSIAX$_{100}$**H**FD with the position of the catalytic triad indicated in bold [Kinsella et al., 2004, Duh et al., 2008]. A recent study, however, has identified the catalytic triad as GDGN**C**FYHSIAX$_{100}$**H**F**D** by means of kinetic studies using OTU mutated at the candidate residues [Capodagli et al., 2011]. The amino

acid sequence of the OTU-like cysteine protease catalytic triad was conserved among the southern African CCHFV isolates. The approximate position of the catalytic domain of the RdRp was predicted to occur at amino acid residue positions 2043 – 2776 using PFAM and 2342 – 2551 using PROFILE relative to SPU97/85. The exact position of the RdRp catalytic domain has yet to be experimentally determined. Sequence diversity within the catalytic site ranged from 0 – 1.6% between isolates.

The C2H2-type zinc finger motif (amino acid residue positions 609 – 632) identified previously [Duh et al., 2008, Ozdarendeli et al., 2010, Yadav et al., 2013] was not predicted using either InterProScan or PSI-BLAST although the amino acid sequence in this region was highly conserved. In comparison to Turkey-Kelkit06, Kosova Hoti, NIVA118594, NIVA118595 and NIV112143 CCHFV isolates, a single, semi-conserved amino acid residue substitution of arginine to lysine at position 611 was present in 2/14 southern African isolates and the C2H2 motif was identified visually in all isolates. Similarly, the leucine zipper motif (amino acid residue positions 1386 – 1407) identified in strain IbAr10200 was not predicted in the southern African isolates although a maximum of two amino acid substitutions were present in the region [Honig et al., 2004, Kinsella et al., 2004]. Isoleucine was substituted for leucine at position 1386 in 4/14 isolates and serine for glycine at position 1389 in 13/14 isolates in comparison with IbAr10200. The conserved protein domains in the L ORF are illustrated in Figure 2A.

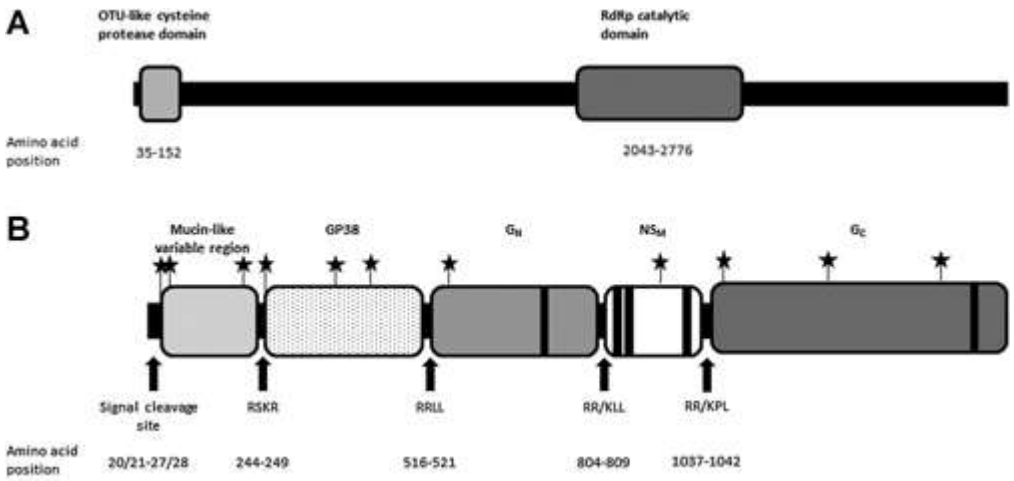

Figure 2 : Schematic representation of the protein analysis of the complete genome of southern African CCHFV isolates (not drawn to scale). A) L segment. B) M segment. Cleavage sites are indicated by arrows with the amino acid residues and positions indicated below. Transmembrane helices are indicated as black lines in the relevant ORFs. N-glycosylation sites are indicated by stars.

Molecular characterization of the M segment

Analysis of the southern African M segments using SignalP identified a signal peptide cleavage site at the N terminus for all isolates with the exception of SPU415/85. The exact position of the cleavage site differed between isolates, occurring variably between amino acid residues 20/21 (SPU187/90 and SPU48/90), 22/23 (SPU4/81 and SPU103/87), 24/25 (SPU383/87, SPU130/89, SPU497/88, SPU431/85, SPU97/85, SPU556/87, SPU45/88 and SPU18/88) or 27/28 (SPU44/08). Signal peptides are commonly found in proteins which are targeted to organelles such as the endoplasmic reticulum or Golgi, and in many membrane-bound proteins. The mucin-like variable region stretches from the signal peptide cleavage site to a furin cleavage site motif, RSKR, which was conserved among all southern African isolates at either amino acid residue position 244 - 247 or 249 - 252. The calculated p-distance for the mucin-like domain of southern African isolates showed nucleotide sequence diversity of 0 – 40.7% (amino acid diversity 0 – 60.1%) for this region. The furin cleavage site

is followed by GP38 which was 804 nucleotides (268 amino acid residues) in length for all southern African CCHFV isolates studied. The nucleotide and amino acid sequences in this region were more conserved than the mucin-like region with p-distances calculated at 0 – 12.7% at the nucleotide level and 0 – 16.8% at the amino acid level. The N terminus of G_N is cleaved from GP38 at a protease SKI-1 cleavage site motif, RRLL, which was conserved among southern African isolates at amino acid residue position 516 – 519 or 521 – 524. The smaller of the two envelope glycoproteins, G_N, was 852 nucleotides (268 amino acid residues) in length and showed nucleotide variation of 0 – 8.1% (amino acid 0 – 8.5%). The C terminus of the G_N protein is cleaved at a further SKI-1 cleavage site motif, RRLL or RKLL, at amino acid residue position 804 – 807 or 809 – 812. Calculated p-distances for the 687 nucleotide (229 amino acid) long NS_M protein were 0 – 10.5% at the nucleotide level and 0 – 11.5% at the amino acid level. A final cleavage site, RRPL or RKPL, was situated at amino acid residue position 1037 - 1040 or 1042 - 1045, at the G_C N terminus. The 1935 nucleotide long G_C protein showed the highest sequence conservation of the M segment polyprotein with amino acid divergence of 0 – 4.8% (nucleotide divergence 0 – 6.2%). The genome organization of the M segment is illustrated in Figure 2B.

Five transmembrane helices were predicted for all southern African CCHFV isolates analyzed, resulting in three regions internal to and three regions external to the membrane. Using isolate SPU97/85 as an example, the transmembrane helices were located at approximate amino acid residue positions 700 – 722, 825 – 847, 862 – 884, 974 – 996, and 1600 – 1622. These regions were located within the G_N, NS_M, and G_C proteins. Predicted mucin-type GalNAc O-glycosylation sites were identified in 5.4 – 6.5% of the M segment polyprotein, predominantly in the mucin-like variable region where 29.4 – 34.0% of amino

acid residues were predicted to be O-glycosylated. In contrast, only 1.3 – 1.9% of the rest of the M segment was predicted to have O-linked glycosylation. O-linked glycosylation involves the addition of N-acetyl-galactosamine to serine or threonine residues. The serine and threonine content in the M segment ORF averaged 9.1% (range 8.9 – 9.3%) and 8.6% (8.3 – 9.1%), respectively. In comparison, the mucin-like variable region showed increased serine and threonine content at 14.6% (range 12.1 – 16.9%) and 17.4% (16.1 – 19.0%), correlating with the distribution of predicted O-linked glycosylation sites. The positions of predicted N-glycosylation sites for all the aligned M segment sequences are summarized in Table II. Eleven of these predicted sites were conserved in 13 - 14 of the southern African isolates as indicated in Figure 2B, excluding position 760 which was within a predicted transmembrane helix region. The M segment of the southern African CCHFV isolates contained 79 cysteine residues which were highly conserved. This may suggest the presence of a large number of disulphide linkages, however the G_N cytoplasmic tail has been shown to contain two ββα-type zinc fingers with a **C**X_2**C**$X_{11\text{-}12}$**H**X_3**C** motif of cysteine and

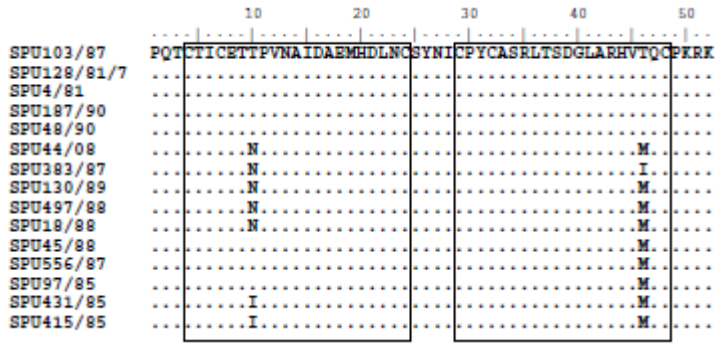

Figure 3 : The amino acid sequences of the dual CCHC-type zinc finger motifs in the G_N tail. Conserved residues are indicated by a dot, while substitutions are indicated by the relevant amino acid. The conserved **C**-X_2-**C**-$X_{11\text{-}12}$-**H**-X3-**C** motifs of each zinc finger are outlined.

Table II : N-linked glycosylation sites on the M segment ORF as predicted by N-Glycosite. Amino acid residue positions are based on the aligned ORF sequences to allow uniformity of numbering. Position 760 forms part of a predicted transmembrane helix and is therefore unlikely to be N-glycosylated.

	30	**35**	**37**	**46**	**82**	**108**	**116**	**186**	**201**	**205**	**229**	**248**	**381**	**431**	**562**	**760**	**981**	**1059**	**1350**	**1568**
SPU4/81	N	N		N					N	N		N	N	N	N	N	N	N	N	N
SPU187/90		N		N					N	N		N	N	N	N	N	N	N	N	N
SPU48/90	N	N		N				N		N		N	N	N	N	N	N	N	N	N
SPU103/87	N	N								N		N	N	N	N	N	N	N	N	N
SPU44/08	N	N								N		N	N	N	N	N	N	N	N	N
SPU383/87	N	N								N		N	N	N	N	N	N	N	N	N
SPU130/89	N		N							N		N	N	N	N	N	N	N	N	N
SPU497/88	N	N								N		N	N	N	N	N	N	N	N	N
SPU18/88	N	N								N		N	N	N	N	N	N	N	N	N
SPU45/88	N	N				N				N		N	N	N	N	N	N	N	N	N
SPU556/87	N	N			N		N			N		N	N	N	N	N	N	N	N	N
SPU97/85	N	N			N		N			N		N	N	N	N	N	N	N	N	N
SPU431/85	N	N			N					N	N	N	N	N	N	N	N	N	N	N
SPU415/85	N	N			N					N	N	N	N	N	N	N	N	N	N	N

histidine residues [Estrada and De Guzman, 2011]. These motifs were conserved in the southern African isolates (Figure 3) beginning at amino acid residue positions 736/741 and 761/766 respectively and separated by a 4 amino acid linker, specifically SYNI.

Discussion

Characterization of the deduced amino acid sequences for the L segment of 14 southern African CCHFV isolates confirmed two conserved functional domains as previously identified, namely the OTU cysteine protease domain and the RdRp catalytic domain [Honig et al., 2004, Kinsella et al., 2004, Duh et al., 2008, Ozdarendeli et al., 2010, Yadav et al., 2013]. Analysis of protein domains showed that the domains were conserved among both southern African and geographically distinct isolates. It has been shown that the OTU cysteine protease domain situated in the N terminus of the CCHFV L protein is not required for viral replication or RdRp function [Bergeron et al., 2010]. Rather, these proteases hydrolyze ubiquitin and interferon-stimulated gene product 15 (ISG15) thereby allowing the virus to evade the type 1 interferon and tumour necrosis factor alpha cytokine pathways [Frias-Staheli et al., 2007]. This domain is highly conserved, showing 98.3 - 100% sequence identity among southern African CCHFV isolates. Similarly, the predicted catalytic domain of the RdRp is highly conserved among isolates as would be expected of the functional site of an enzyme which is essential for viral replication. Despite the sequence conservation of the previously identified C2H2-type zinc finger motif and leucine zipper among geographically diverse isolates [Honig et al., 2004, Kinsella et al., 2004, Duh et al., 2008, Ozdarendeli et al., 2010, Yadav et al., 2013], these motifs were not identified by the predictive software used in this analysis. To confirm that sequence differences were not the reason for this

discrepancy, the published isolates were analysed and the zinc finger and leucine zipper were not predicted. This may be due to the dynamic nature of the hidden Markov model based databases such as Pfam [Finn et al., 2010] and the extensive variation in the primary protein structure between nucleotide polymerases and RdRp. The L segment translational product is poorly defined with regards to posttranslational modifications including proteolytic cleavage and maturation and experimental studies are required to confirm such motifs.

The genome organization of CCHFV M segment follows the sequence: 5'NCR-mucin like region-GP38-G_N-NS_M-G_C-3'NCR. Cleavage of the M polyprotein is accomplished at four cleavage sites including one furin cleavage site (RSKR) and two subtilase SKI-1 cleavage sites (RRLL, RR/KLL) [Vincent et al., 2003, Sanchez et al., 2006]. The final cleavage site (RR/KPL), which separates G_C from NS_M, has been shown to be poorly cleaved by SKI-1 although the protease involved in this cleavage event has not been identified [Vincent et al., 2003]. The four cleavage sites are conserved among all CCHFV isolates including those from southern Africa. The structural glycoproteins, G_N and G_C, were the most conserved of the M segment proteins with up to 91.9% nucleotide (91.5% amino acid) and 95.2% nucleotide (93.8% amino acid) sequence identity between southern African isolates respectively. The role of the non-structural or secretory glycoproteins (mucin-like variable domain, GP38, NS_M, GP85 and GP160) in CCHFV infection is currently unclear, but a similar mucin-like domain glycoprotein of Ebola virus plays a role in viral pathogenesis by causing endothelial cell disruption resulting in increased vascular permeability [Yang et al., 2000]. Sequence diversity was highest in the mucin-like variable region with divergence of up to 40.7% at the nucleotide level and 60.1% at the amino acid level. CCHFV utilizes various vertebrate and

tick hosts in distinct geographical regions. This ability may relate to the variation within the mucin-like region [Ozdarendeli et al., 2010]. This is supported by grouping of isolates based on phylogenetic analysis of the mucin-like region which correlates with the complete M segment (data not shown) showing geographic linkage of isolates within groups. The mechanism underlying the variablility of mucin-like regions of negative stranded RNA viruses including CCHFV has been explained chiefly by a relaxation of purifying selection in this region. This implies that the amino acid composition of the region is not important for its function as long as the O-glycosylation is maintained [Wertheim and Worobey, 2009].

Although the high cysteine content of the M protein which is conserved in all available CCHFV isolates may point to extensive disulphide bonds and a complex secondary structure, the presence of two CCHC-type zinc finger motifs in the G_N cytoplasmic tail also contribute to this cysteine content. The G_N tail has been shown to bind RNA and likely plays a role in RNA packaging and viral assembly by associating with the viral ribonucleoprotein complexes, as do the glycoprotein zinc fingers of hantaviruses, also belonging to the *Bunyaviridae* family [Estrada and De Guzman, 2011].

Of the 11 conserved N-linked glycosylation sites predicted for the M segments of southern African CCHFV isolates, three have thus far been confirmed experimentally namely one site located on G_N (amino acid residue position 557) and two sites located on G_C (amino acid residue positions 1054 and 1563). The N-linked glycosylation site on G_N was shown to be important for glycoprotein localization and for transport of both G_N and G_C proteins from the endoplasmic reticulum to the Golgi apparatus [Erickson et al., 2007].

The conserved nucleotide complementarity of the 5' and 3' NCR of the L, M and S segments of CCHFV were modelled using Mfold and showed cyclization to form panhandle structures. This can be compared to the complementary genomic ends of another member of the *Bunyaviridae* family, Bunyamwera virus. In the case of Bunyamwera virus, the cooperation of the complementary 5' and 3' NCR were required for RNA synthesis to generate both mRNA and cRNA [Barr and Wertz, 2004]. A similar role in CCHFV replication seems likely. The first nine nucleotides of the L, M and S segments are identical in CCHF, as in other nairoviruses such as Dugbe virus and Hazara virus, and may function as an RdRp recognition site for the initiation of viral mRNA transcription and replication [Marriott et al., 1992, Lasecka and Baron, 2013].

As shown in a related study, the genetic distances for the complete segments at amino acid level are highest for the M segment at 0-15.6% between southern African isolates, with the S segment being more conserved at 0-1.7% and the L segment at 0-2.1% [Goedhals et al., 2014]. Further analysis of the genetic diversity within specific protein domains indicates that the amino acid variation is not uniformly distributed and may give an indication of the potential significance of such sites as targets for future interventions. The OTU-like cysteine protease and RdRp domains of the L segment represent likely targets for further study, showing both a high level of conservation and significant functional roles, contrasting with the mucin-like region of the M segment which shows divergence of up to 60.1% at amino acid level.

Genetic characterization of CCHFV isolates can help to identify protein functions and to suggest pathogenic mechanisms which require further examination. The development of

diagnostic assays, vaccine design and identification of potential targets for antiviral interventions may also benefit, particularly when information from geographically distinct areas is available. It is clear that further experimental studies are required to investigate the significance and function of many of the CCHFV genome properties identified in this way. However, this examination of southern African CCHFV isolates provides corroboration of conserved genome domains and sequence identity which may direct further studies.

Funding

This project was funded by the National Health Laboratory Service Research Trust, the Polio Research Foundation, South Africa and University of the Free State Cluster funding.

Competing interests

The authors have no competing interests to declare.

References

Altamura LA, Bertolotti-Ciarlet A, Teigler J, Paragas J, Schmaljohn CS, Doms RW. 2007. Identification of a novel C-terminal cleavage of Crimean-Congo hemorrhagic fever virus $PreG_N$ that leads to generation of an NS_M protein. J Virol 81:6632-6642.

Altschul SF, Madden TL, Schäffer AA, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Barr JN, Wertz GW. 2004. Bunyamwera Bunyavirus RNA synthesis requires cooperation of =- and 5-terminal sequences. J Virol 78:1129-1138.

Bente DA, Forrester NL, Watts DM, McAuley AJ, Whitehouse CA, Bray M. 2013. Crimean-Congo hemorrhagic fever: History, epidemiology, pathogenesis, clinical syndrome and genetic diversity. Antiviral Res 100:159-189.

Bergeron E, Albariño CG, Khristova ML, Nichol ST. 2010. Crimean-Congo hemorrhagic fever virus-encoded ovarian tumor protease activity is dispensable for virus RNA polymerase function. J Virol 84:216-226.

Capodagli GC, McKercher MA, Baker EA, Masters EM, Brunzelle JS, Pegan SD. 2011. Structural analysis of a viral ovarian tumor domain protease from the Crimean-Congo hemorrhagic fever virus in complex with covalently bonded ubiquitin. J Virol 85:3621-3630.

Carter SD, Surtees R, Walter CT, Ariza A, Bergeron E, Nichol ST, Hiscox JA, Edwards TA, Barr JN. 2012. Structure, function and evoluation of the Crimean-Congo hemorrhagic fever virus nucleocapsid protein. J Virol 86:10914-10923.

Clerx JP, Casals J, Bishop DH. 1981. Structural characteristics of nairoviruses (genus *Nairovirus, Bunyaviridae*). J Gen Virol 55:165-178.

Duh D, Nichol ST, Khristova ML, Saksida A, Hafner-Bratkovic I, Petrovec M, Dedushaj I, Ahmeti S, Avsic-Zupanc T. 2008. The complete genome sequence of a Crimean-Congo hemorrhagic fever virus isolated from an endemic region in Kosovo. Virol J 5:7.

Erickson BR, Deyde V, Sanchez AJ, Vincent MJ, Nichol ST. 2007. N-linked glycosylation of Gn (but not Gc) is important for Crimean Congo hemorrhagic fever virus glycoprotein localization and transport. Virology 361:348-355.

Estrada DF, De Guzman RN. 2011. Structural characterization of the Crimean-Congo hemorrhagic fever virus Gn tail provides insight into virus assembly. J Biol Chem 286:21678-21686.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res 38:D211-D222.

Frias-Staheli N, Giannakopoulos NV, Kikkert M, Taylor SL, Bridgen A, Paragas J, Richt JA, Rowland RR, Schmaljohn CS, Lenschow DJ, Snijder EJ, Garcia-Sastre A, Virgin HW 4th. 2007. Ovarian tumor domain-containing viral proteases evade ubiquitin- and ISG 15-dependent innate immune responses. Cell Host Microbe 2:404-416.

Goedhals D, Bester PA, Paweska JT, Swanepoel R, Burt FJ. 2014. Next-generation sequencing of southern African Crimean-Congo haemorrhagic fever virus isolates reveals a high frequency of M segment reassortment. Epidemiol Infect 142:1952-1962.

Guo Y, Wang W, Ji W, Deng M, Sun Y, Zhou H, Yang C, Deng F, Wang H, Hu Z, Lou Z, Rao Z. 2012. Crimean-Congo hemorrhagic fever virus nucleoprotein reveals endonuclease activity in bunyaviruses. Proc Natl Acad Sci U S A 109:5046-5051.

Honig JE, Osborne JC, Nichol ST. 2004. Crimean-Congo hemorrhagic fever virus genome L RNA segment and encoded protein. Virology 321:29-35.

Khromykh AA, Meka H, Guyatt KJ, Westaway EG. 2001. Essential role of cyclization sequences in flavivirus RNA replication. J Virol 75:6719-6728.

Kinsella E, Martin SG, Grolla A, Czub M, Feldmann H, Flick R. 2004. Sequence determination of the Crimean-Congo hemorrhagic fever virus L segment. Virology 321:23-28.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-580.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez A, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Lasecka L, Baron MD. 2013. The molecular biology of nairoviruses, an emerging group of tick-borne arboviruses. Arch Virol 159:1249-1265.

Maltezou HC, Andonova L, Andraghetti R, Bouloy M, Ergonul O, Jongejan F, Kalvatchev N, Nichol S, Niedrig M, Platonov A, Thomson G, Leitmeyer K, Zeller H. 2010. Euro Surveill 15:19504.

Marriott AC, el-Ghorr AA, Nuttall PA. 1992. Dugbe Nairovirus M RNA: nucleotide sequence and coding strategy. Virology 190:606-615.

Nichol ST, Beaty BJ, Elliott RM, Goldbach R, Plyusnin A, Schmaljohn CS, Tesh RB. 2006. Index of Viruses – *Bunyaviridae*. In: Büchen-Osmond C, editor. ICTVdB – The Universal Virus Database, version 4. New York: Columbia University.
http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fs_index.htm

Ozdarendeli A, Canakoğlu N, Berber E, Aydin K, Tonbak S, Ertek M, Buzgan T, Bolat Y, Aktaş M, Kalkan A. 2010. The complete genome analysis of Crimean-Congo hemorrhagic fever virus isolated in Turkey. Virus Res 147:288-293.

Petersen TN, Brunak S, van Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785-786.

Sanchez AJ, Vincent MJ, Erickson BR, Nichol ST. 2006. Crimean-Congo hemorrhagic fever virus glycoprotein precursor is cleaved by furin-like and SKI-1 proteases to generate a novel 38-kilodalton glycoprotein. J Virol 80:514-525.

Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U,

Brunak S, Wandall HH, Levery SB, Clausen H. 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32:1478-1488.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Vincent MJ, Sanchez AJ, Erickson BR, Basak A, Chretien M, Seidah NG, Nichol ST. 2003. Crimean-Congo hemorrhagic fever virus glycoprotein proteolytic processing by subtilase SKI-1. J Virol 77:8640-8649.

Wertheim JO, Worobey M. 2009. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. J Virol 83:4690-4694.

Yadav PD, Cherian SS, Zawar D, Kokate P, Gunjikar R, Jadhav S, Mishra AC, Mourya DT. 2013. Genetic characterization and molecular clock analyses of the Crimean-Congo hemorrhagic fever virus from human and ticks in India, 2010-2011. Infect Genet Evol 14:223-231.

Yang ZY, Duckers HJ, Sullivan NJ, Sanchez A, Nabel EG, Nabel GJ. 2000. Identification of the Ebola virus glycoprotein as the main viral determinant of vascular cell cytotoxicity and injury. Nat Med 6:886-889.

Zdobnov EM, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847-848.

Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B. 2004. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology 14:1229-1246.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406-3415.