Is the Bombali virus pathogenic in humans?

Henry J Martell*, Stuart G Masterson*, Jake E. McGreig, Martin Michaelis§, Mark N Wass§

Industrial Biotechnology Centre and School of Biosciences, University of Kent, Canterbury, Kent, CT2 7NJ, UK.

§To whom correspondence should be addressed

m.michaelis@kent.ac.uk +44 1227 827804

m.n.wass@kent.ac.uk +44 1227 827626

*equal contribution

Supplementary Material

Supplementary Methods

Ebolavirus Nomenclature

The virus nomenclature in this report follows the recommendations set by Kuhn et al., *Filoviridae* is the family, in the order *Mononegavirales*. Both of these terms are always italicised when referenced. The genus is known as *Ebolavirus*, and is only italicised when referring to the genus, but not when referring to physical viruses, virus properties, or constituent virus parts such as proteins or genomes. Ebolavirus Disease (EVD) also remains unitalicised. The five individual species are subsequently referred to as *Bundibugyo ebolavirus* (type virus: *Bundibugyo virus*, BDBV), *Reston ebolavirus* (type virus: *Reston virus*, RESTV), *Sudan ebolavirus* (type virus: *Sudan virus*, SUDV), *Tai Forest ebolavirus* (type virus: *Tai Forest virus*, TAFV) and *Zaire ebolavirus* (type virus: *Ebola virus*, EBOV) (Kuhn et al. 2014).

Collection of *Ebolavirus* Genomes

All ebolavirus genome sequences were obtained from the National Center for Biotechnology Information (NCBI) (Brister et al. 2015), the Virus Pathogen Resource (ViPR) (Pickett et al. 2012), as well as taken from a repository obtained from (Urbanowicz et al. 2016), available here: https://github.com/ebov/space-time. Duplicate sequences present in >1 of the databases were filtered out during initial sample collection, with the order of source preference being NCBI > ViPR > Urbanowicz et al. Supplementary Table 1 summarises the sources used for the set of *Ebolavirus* genomes.

Two Bombali virus sequences were obtained from NCBI (GenBank IDs MF319185, MF319186 respectively). MF319185 was used as the reference sequence and the residues from this were used whenever the two sequences had different amino acids at any given SDP.

Genome Processing and Filtering

For each sample genome, open reading frames (ORFs) were identified using the EMBOSS getorf tool (Rice et al. 2000), and the resulting ORFs were matched to the UniProt *Ebola virus* reference protein sequences using BLAST (Camacho et al.

2009; Bateman et al. 2015). The top ORF hit for each *Ebola virus* protein was then used as the protein sequence for that sample, for all proteins except GP. The ebolavirus GP protein is the result of mRNA editing, due to a slippery 7A-motif that is translated as eight A nucleotides, with the regular ORF containing an early stop codon (Volchkov et al. 1995). The GP ORF hits were further processed by editing the identified ORF to swap the 7A-motif for 8 A nucleotides. ORFs were then re-identified for the edited sequence and BLAST was used to search against the *Ebola virus* reference proteins.

After these steps, ebolavirus samples that did not have a BLAST hit with >90% coverage compared to the *Ebola virus* reference protein, for each of the seven proteins, was removed. Samples with poor metadata, such as unknown host or data were also removed (partial dates were allowed, e.g. if only the year of collection was known). This was to ensure that only high-quality samples were analysed, as incomplete data could affect subsequent analyses. Supplementary Table 2 summarises the samples that were removed in this step, and a full list of the samples that were retained can be found in Supplementary File 2.

Genome Sequence Alignment and Identification of Specificity Determining Positions

Clustal Omega was used to generate sequence alignments for each of the ebolavirus proteins (Sievers et al. 2011), and the individual sequence identities were obtained from the Clustal Omega output. Jensen-Shannon divergence scores were then calculated for each protein (Capra & Singh 2007). S3det was used in supervised mode to find specificity determining positions (SDPs), with sequences assigned to two groups prior to running S3Det (Rausell et al. 2010). Group 1 contained all of the human pathogenic sequences (*Ebola virus*, *Sudan virus*, *Bundibugyo virus*, and *Tai Forest virus*) and group 2 contained all of the human non-pathogenic sequences (*Reston virus*). All SDPs are referred to by the amino acid in the *Ebola virus* protein sequence, the position in the *Ebola virus* reference protein sequence, and the corresponding amino acid in the *Reston virus* protein sequence,

e.g. G20A meaning at position 20 *Ebola virus* has a glycine residue and *Reston virus* has an alanine residue.

Structural Analysis of SDPs

All available ebolavirus protein structures were downloaded from the Protein Databank (PDB) (Berman et al. 2000), and SDPs were mapped to the highest quality structure available, based on structure resolution and coverage. Multimeric protein structures were used to analyse the effects of the SDPs on partner interactions. Where structures were unavailable from the PDB, proteins were modelled using Phyre2 with default settings (Kelly et al. 2015). Supplementary Table 3 summarises the structures used for analysis. PyMOL (https://pymol.org/2/) was used to visualise the identified SDPs in the protein structures and generate images.

For the subset of SDPs mapped to structures, multiple computational tools were used to predict the functional effects of each SDP. mCSM was used to predict the effect on protein stability (Pires et al. 2014), where the change in stability ($\Delta\Delta G$) is measured in kcal/mol, with negative values being destabilising and positive values being stabilising. Relative solvent accessibility of SDP residues was also calculated using mCSM. BLOSUM62 scores were assigned to each SDP, with the score calculated for the change between the *Ebola virus* sequence and the *Reston virus* sequence wherever there was variation amongst the pathogenic species.

Phylogenetic Trees

Whole genome alignments were performed using Clustal Omega, whilst the alignments for each protein were performed using TranslatorX (Abascal et al. 2010), which aligns protein-coding nucleotide sequences based on their corresponding amino acid translations. Bayesian trees, for each protein and genome, were then produced using BEAUTI and BEAST 1.10.4 (Suchard et al. 2018) , performed on the CIPRES Science Gateway (Miller et al. 2010). The consensus tree for each set of 10,000 trees was then calculated using TreeAnnotator, and the nodes were labelled with the posterior probabilities. These trees were then analysed and plotted in R using the “ape” package.

The Maximum Phylogenetic trees were produced using RaxML8.2.10 on the CIPRES Science Gateway, with 1000 Bootstrap replicates run to obtain the best scoring ML tree for each set of sequences. These were plotted and annotated in FigTree [http://tree.bio.ed.ac.uk/software/figtree/]. Species labels were added to all trees using Inkscape.

References

Abascal F, Zardoya R, Telford MJ, 2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research,* 38(W7) pp.13.

Bateman, A. et al., 2015. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1), pp.D204–D212.

Berman, H.M. et al., 2000. The protein data bank. *Nucleic acids research*, 28(1), pp.235–242.

Brister, J.R. et al., 2015. NCBI viral Genomes resource. *Nucleic Acids Research*, 43(D1), pp.D571–D577.

Camacho, C. et al., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10, p.421.

Capra, J.A. & Singh, M., 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15), pp.1875–1882.

Kelly, L.A. et al., 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10(6), pp.845–858.

Kuhn, J.H. et al., 2014. Nomenclature- and database-compatible names for the Two Ebola virus variants that emerged in guinea and the Democratic Republic of the Congo in 2014. *Viruses*, 6(11), pp.4760–4799.

Miller, M. A., Pfeiffer, W. & Schwartz, T., 2010 In *Proceedings of the Gateway Computing Environments Workshop (GCE)* 1–8

Pickett, B.E. et al., 2012. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(D1), pp.593–598.

Pires, D.E. V, Ascher, D.B. & Blundell, T.L., 2014. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3), pp.335–342.

Rausell, A. et al., 2010. Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences*, 107(5), pp.1995–2000.

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(1), pp.276–277.

Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1), p.539.

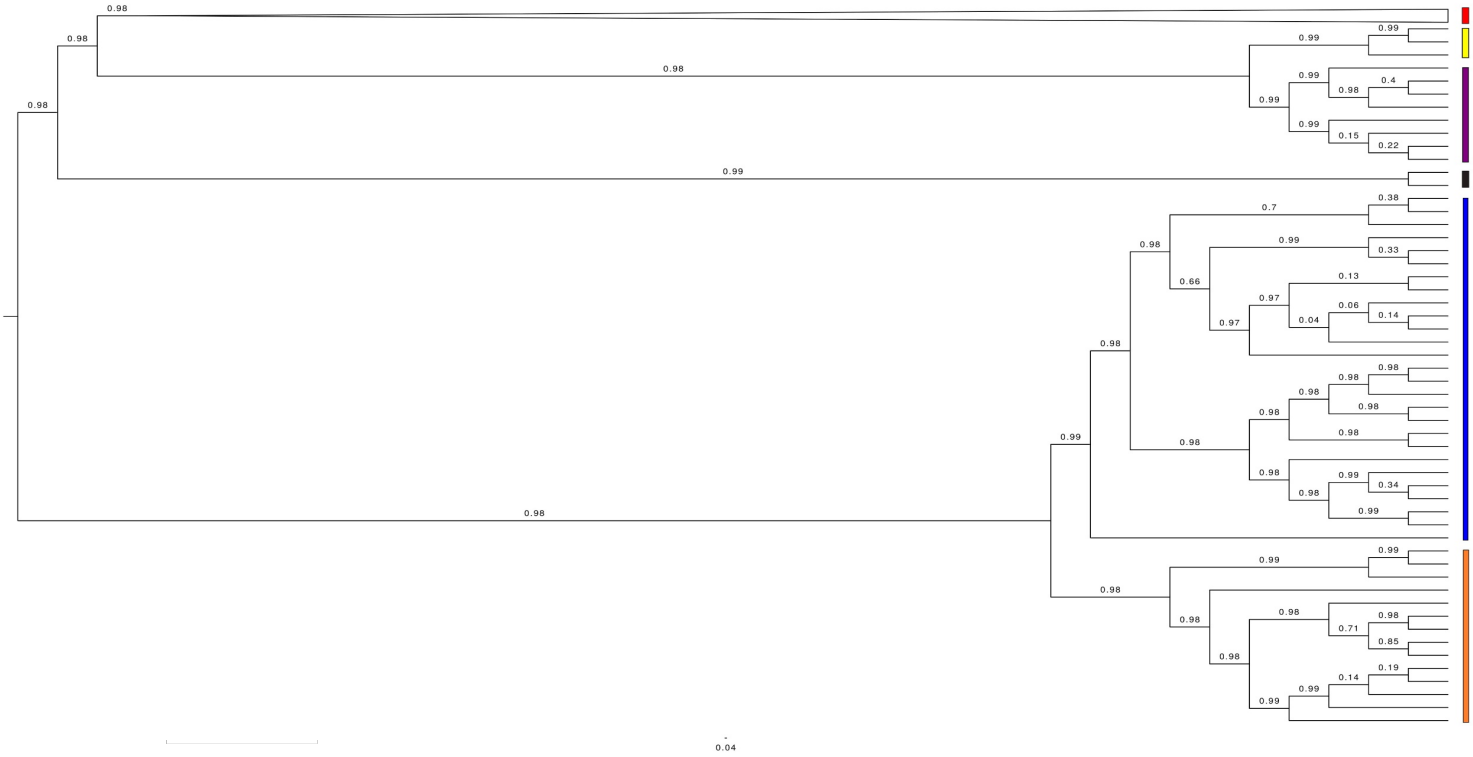
Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ & Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10 *Virus Evolution* **4**, vey016.

Urbanowicz, R.A. et al., 2016. Human Adaptation of Ebola Virus during the West African Outbreak. *Cell*, 167(4), pp.1079–1087.

Volchkov, V.E. et al., 1995. GP mRNA of Ebola Virus Is Edited by the Ebola Virus Polymerase and by T7 and Vaccinia Virus Polymerases. *Virology*, 214(2), pp.421–430.

Supplementary Figures

Supplementary Figure 1. Phylogenetic tree of the Ebolavirus genomes and individual proteins. Bayesian and Maximum Likelihood phylogenetic trees are shown for the Ebolavirus genomes and each of the Ebolavirus proteins. A) genome Bayesian tree. B) genome maximum likelihood tree, C) Bayesian tree for protein L, D) Maximum likelihood tree for protein L, E) Bayesian tree for protein GP, F)Maximum likelihood tree for protein GP, G) Bayesian tree for protein NP, H), Maximum likelihood tree for protein NP, I) Bayesian tree for protein VP24, J) Maximum likelihood tree for protein VP24, K) Bayesian tree for protein VP30, L) Maximum likelihood tree for protein VP30, M) Bayesian tree for protein VP35, N) Maximum likelihood tree for protein VP35, O) Bayesian tree for protein VP40. P) Maximum likelihood tree for protein VP40.



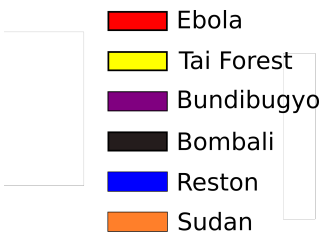


Fig S1A. Genome Bayesian Tree

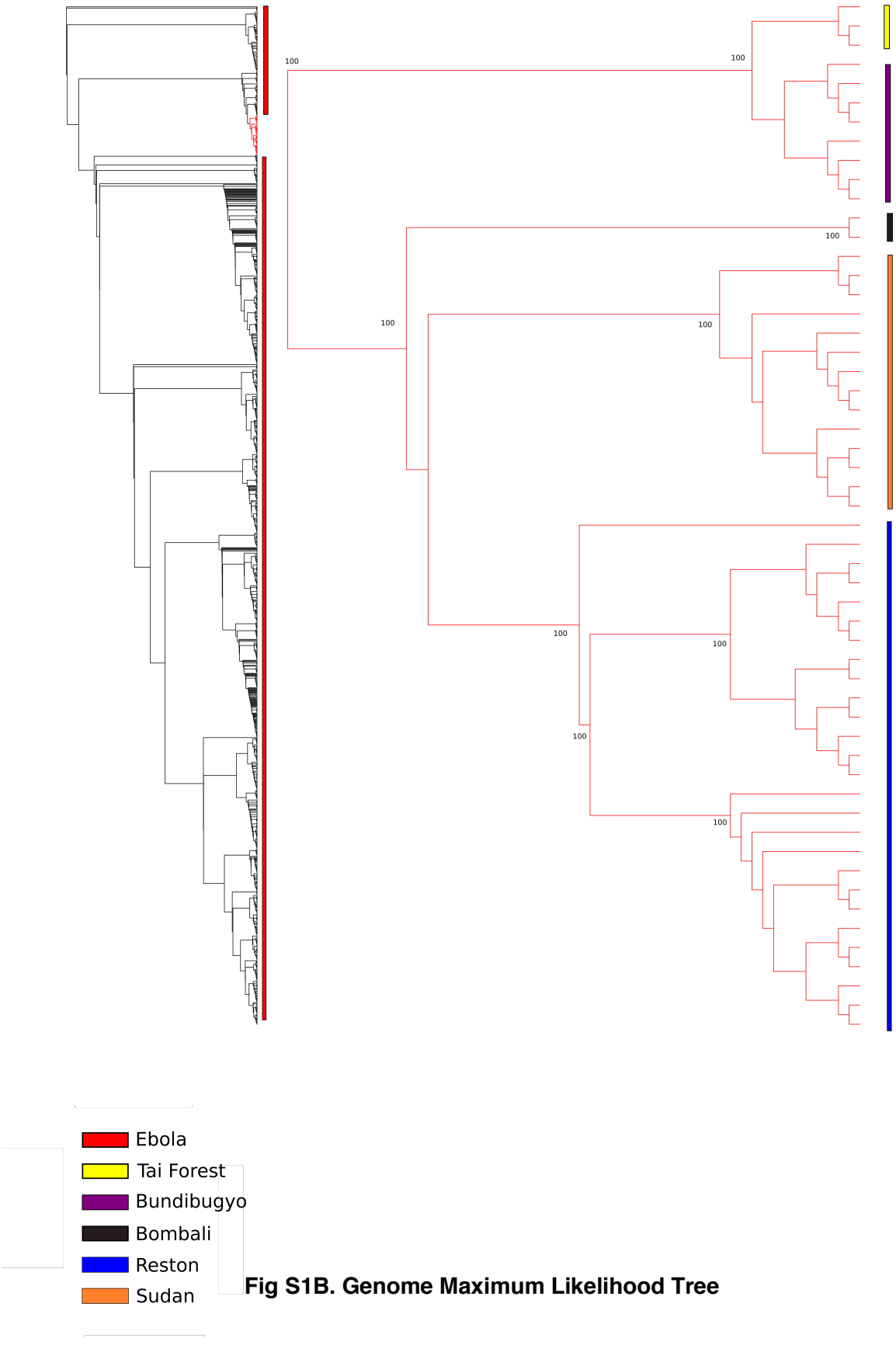
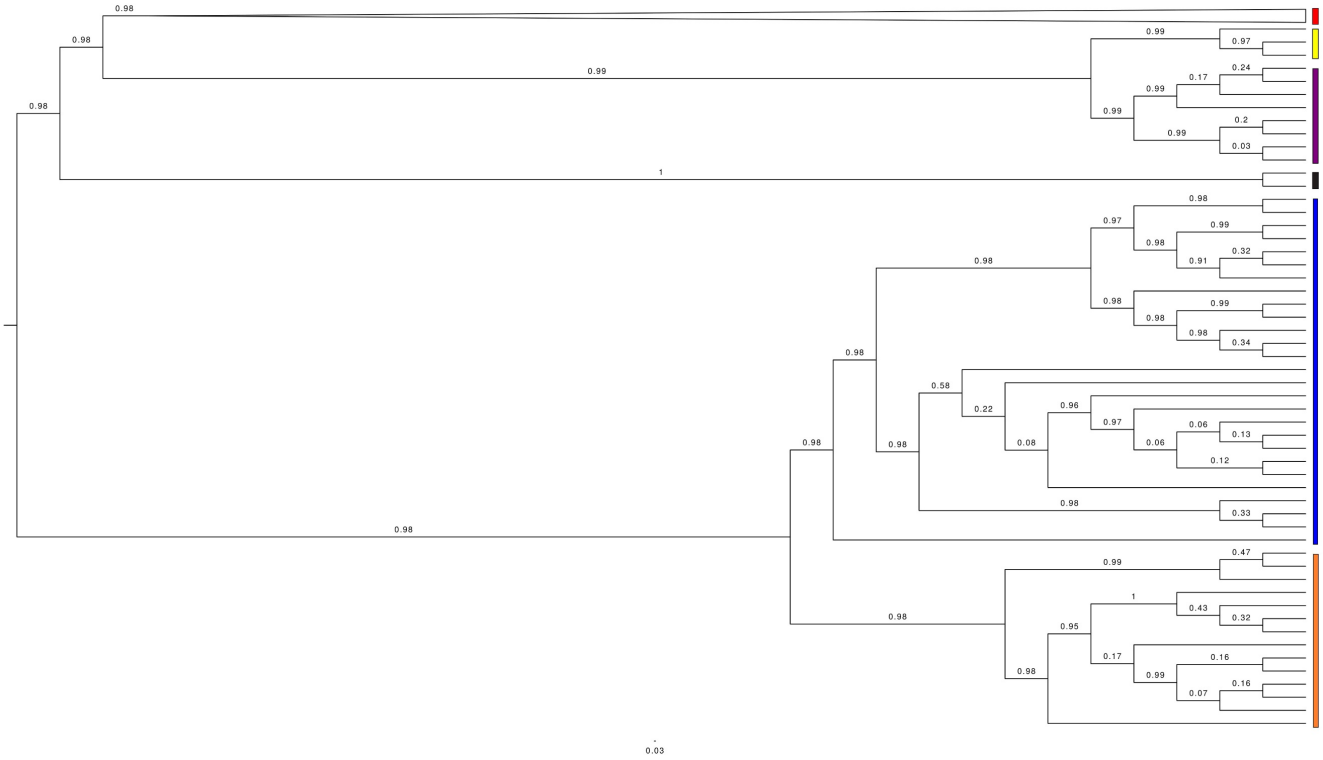


Fig S1B. Genome Maximum Likelihood Tree



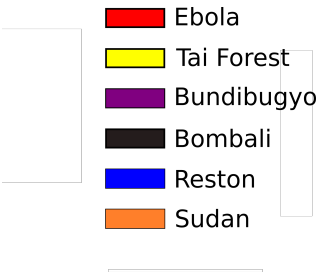


Fig S1C. L Bayesian Tree

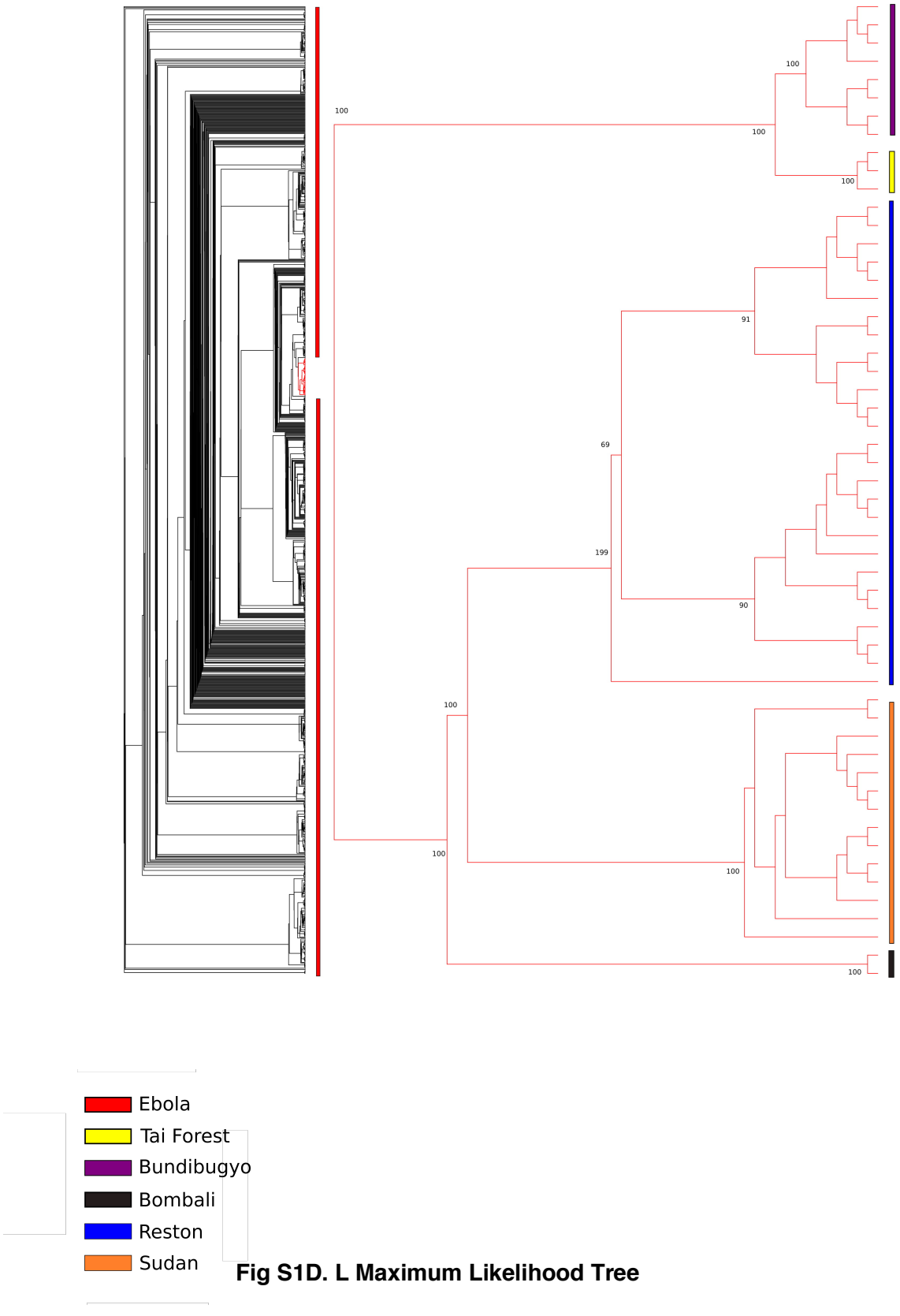
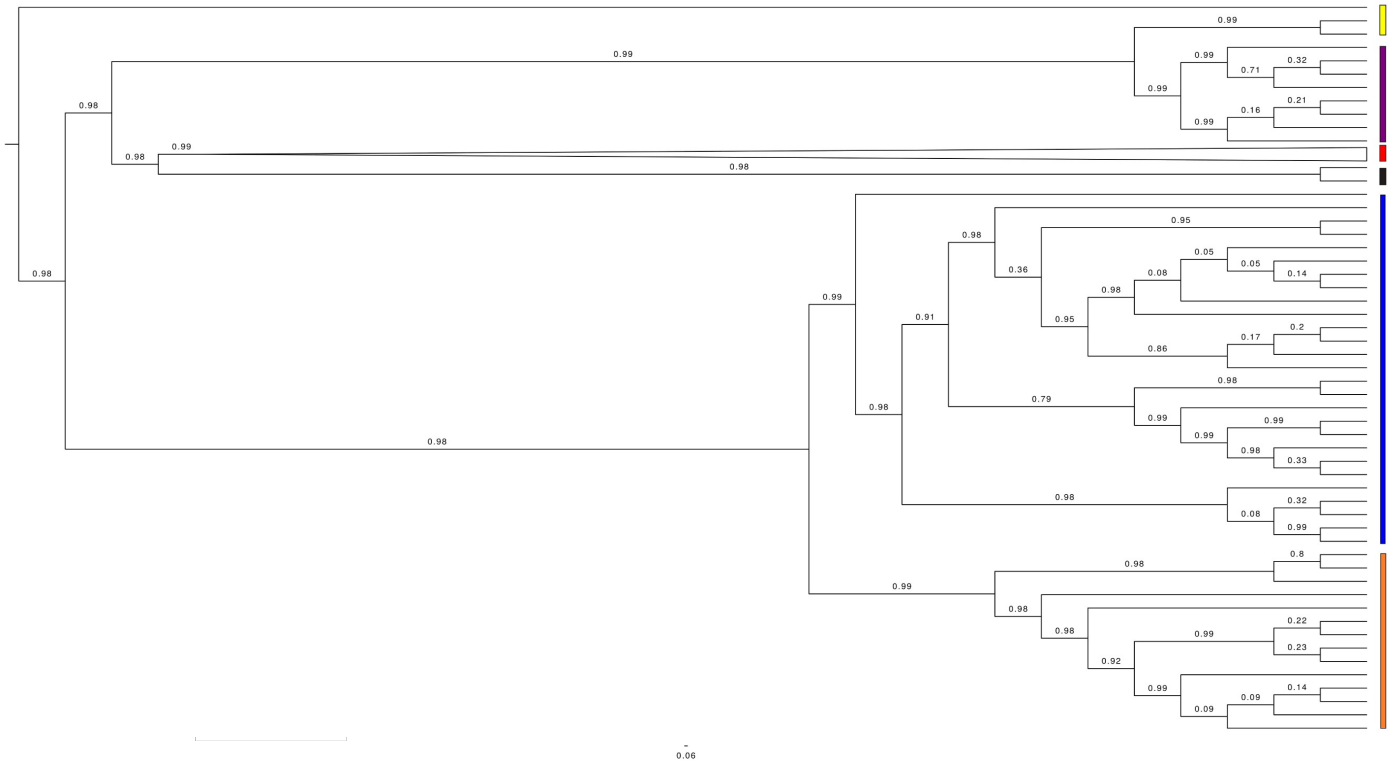


Fig S1D. L Maximum Likelihood Tree



Ebola
Tai Forest
Bundibugyo
Bombali
Reston
Sudan

Fig S1E. GP Bayesian Tree

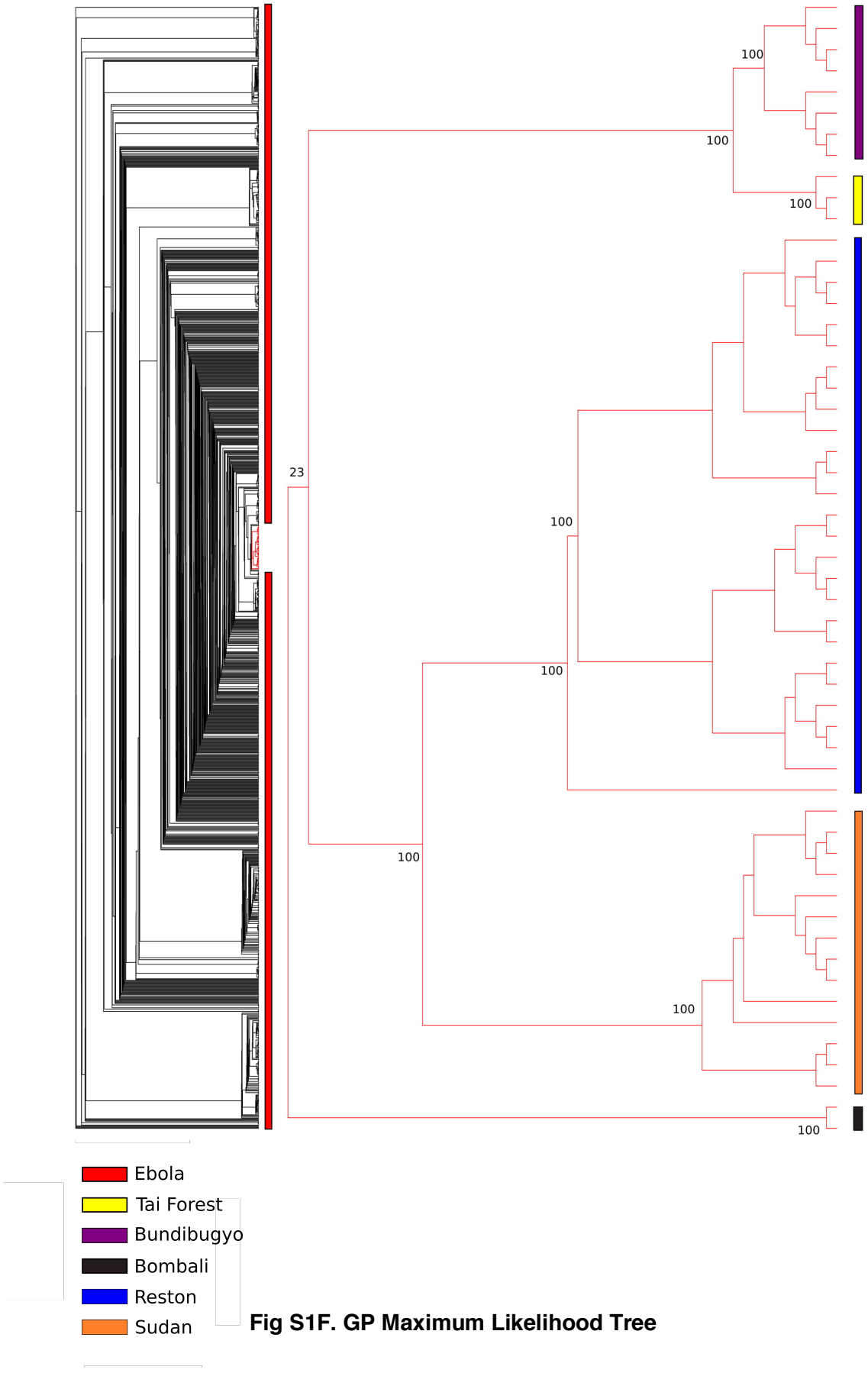


Fig S1F. GP Maximum Likelihood Tree

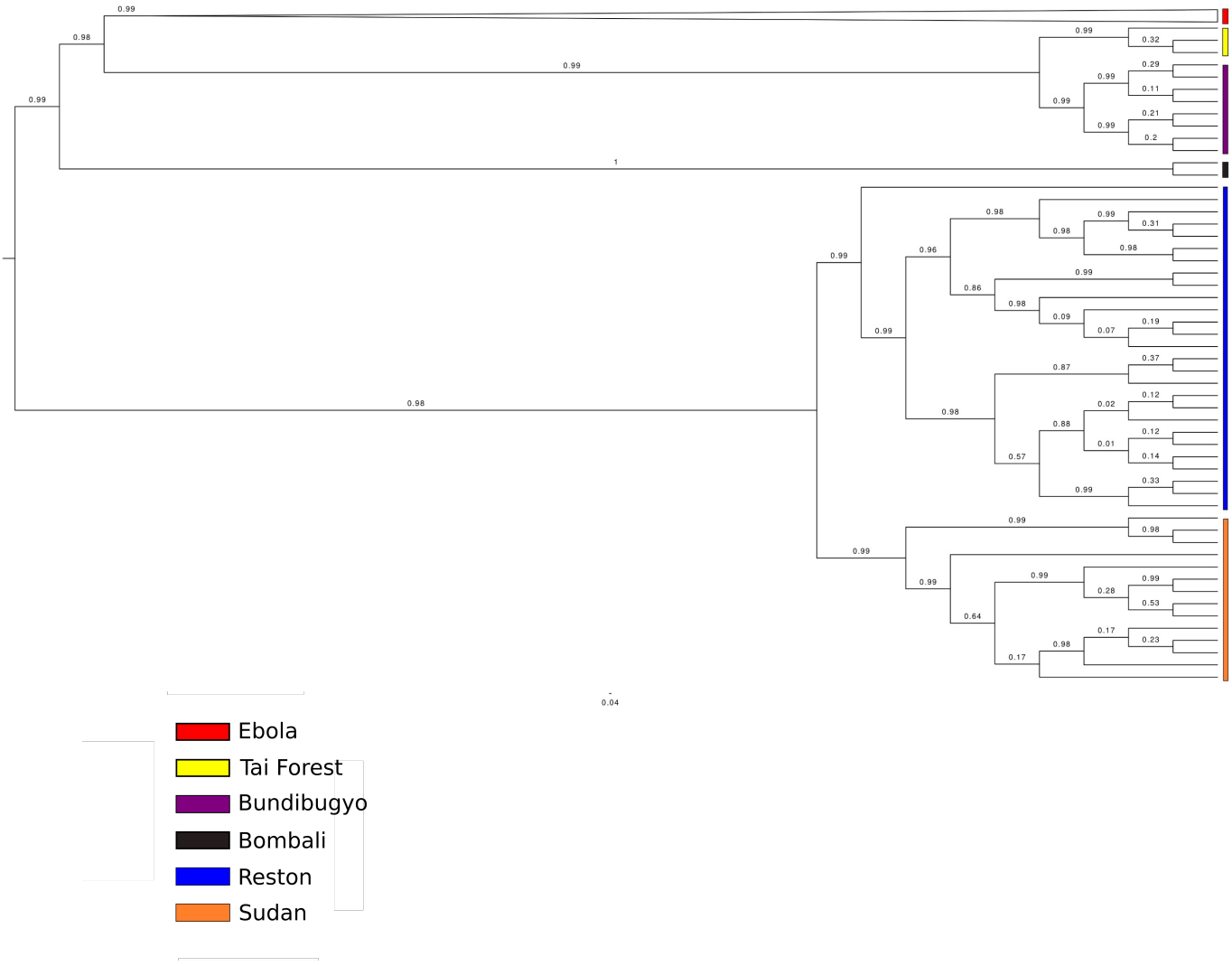


Fig S1G. NP Bayesian Tree

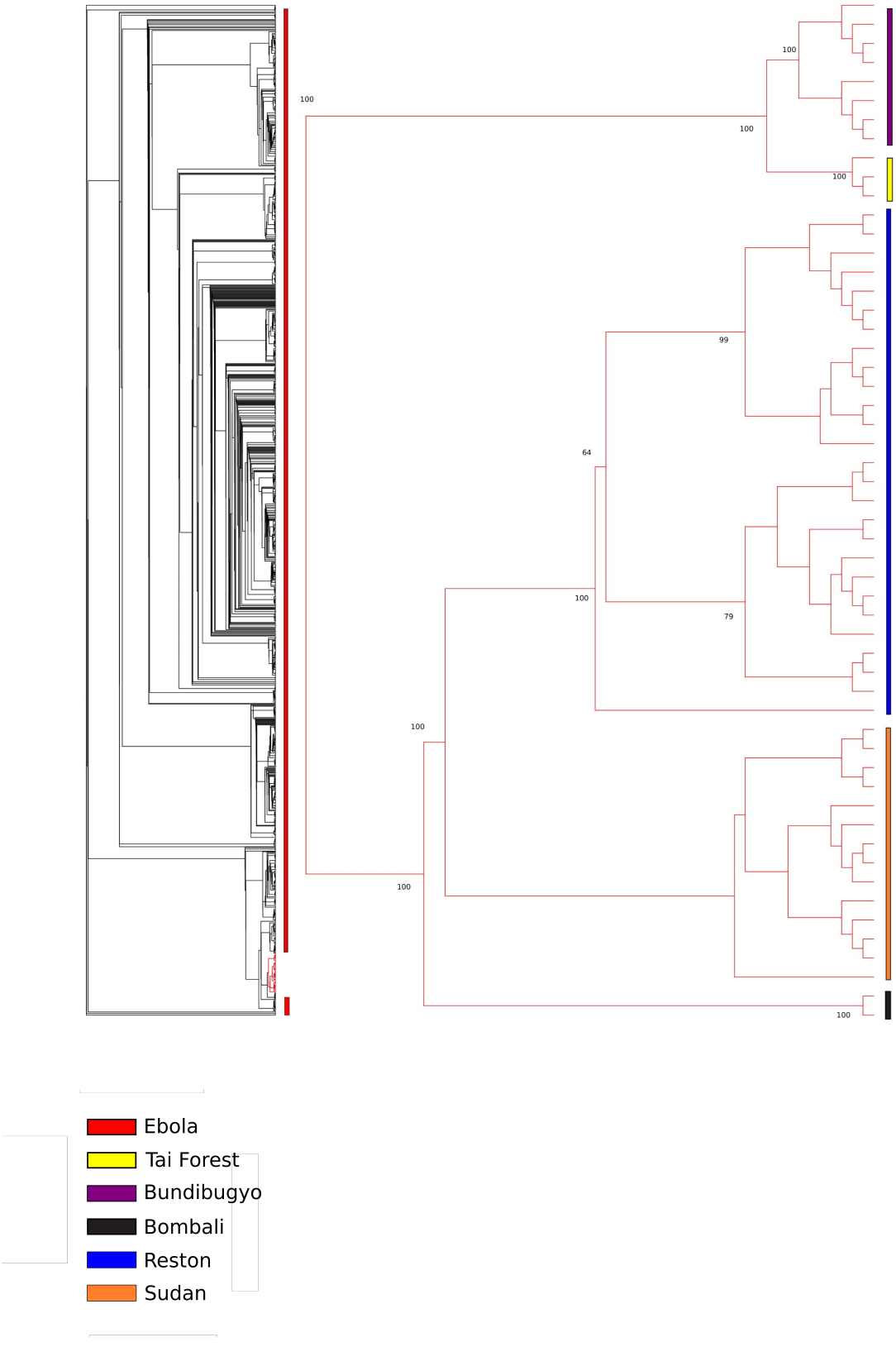


Fig S1H. NP Maximum Likelihood Tree

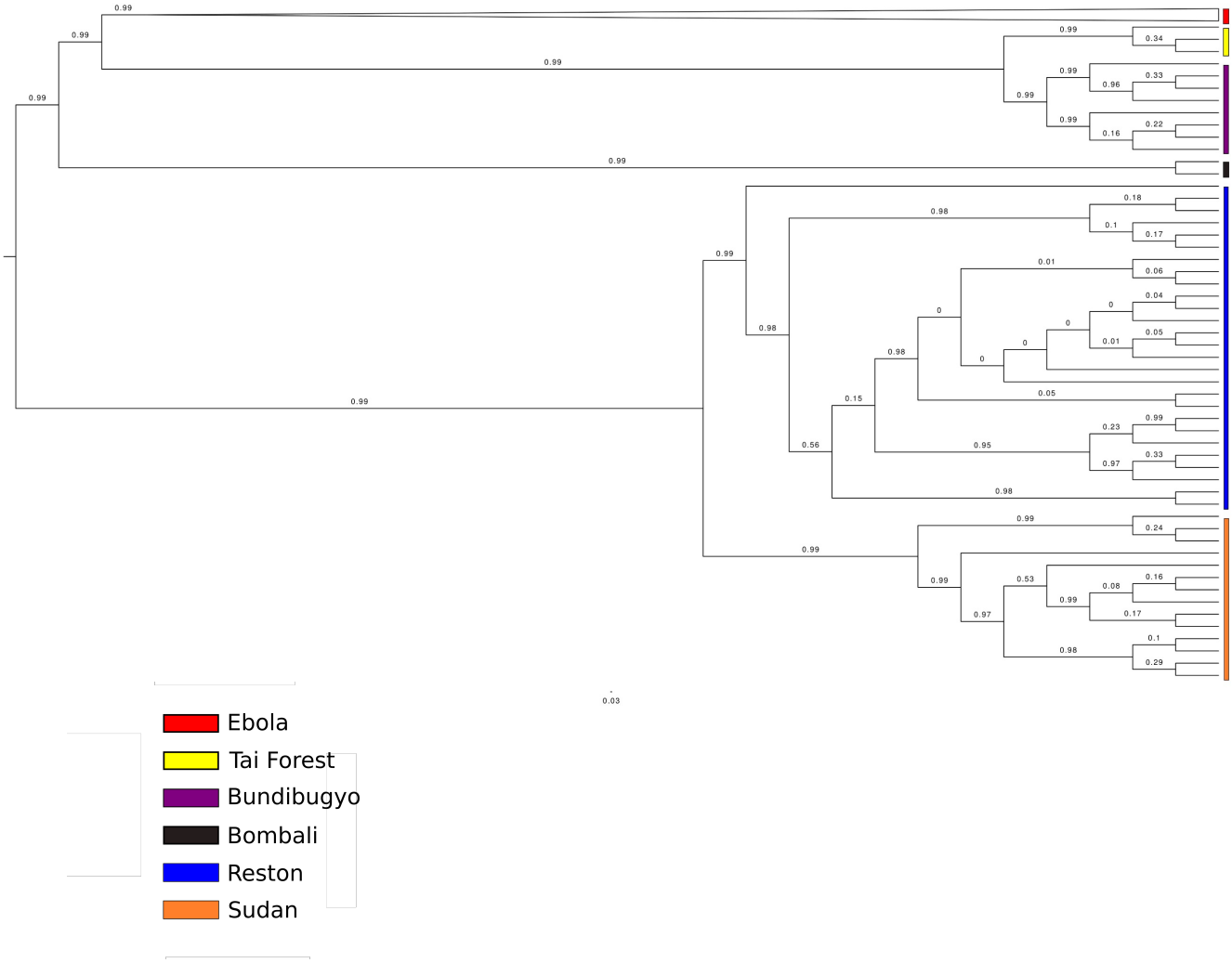
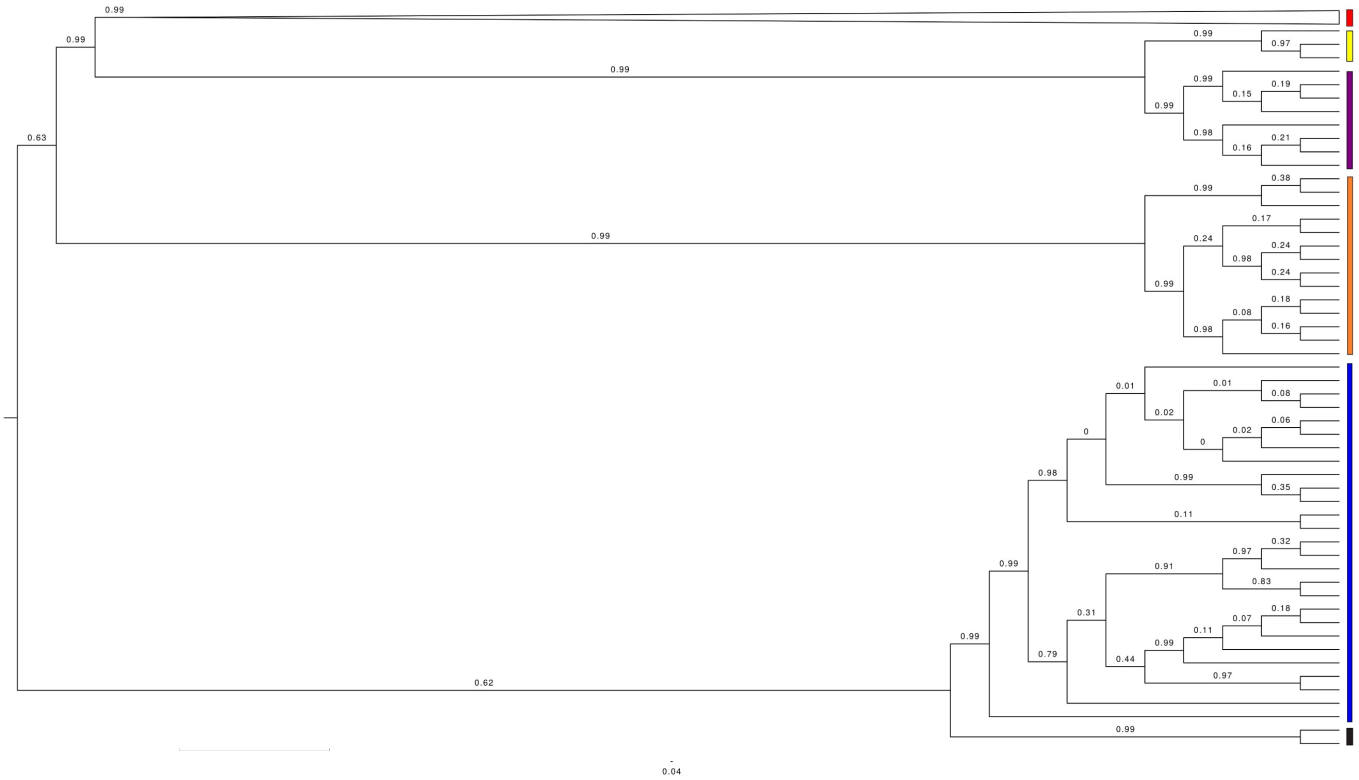


Fig S1I. VP24 Bayesian Tree

Fig S1J. VP24 Maximum Likelihood Tree



Ebola
Tai Forest
Bundibugyo
Bombali
Reston
Sudan

Fig S1K. VP30 Bayesian Tree

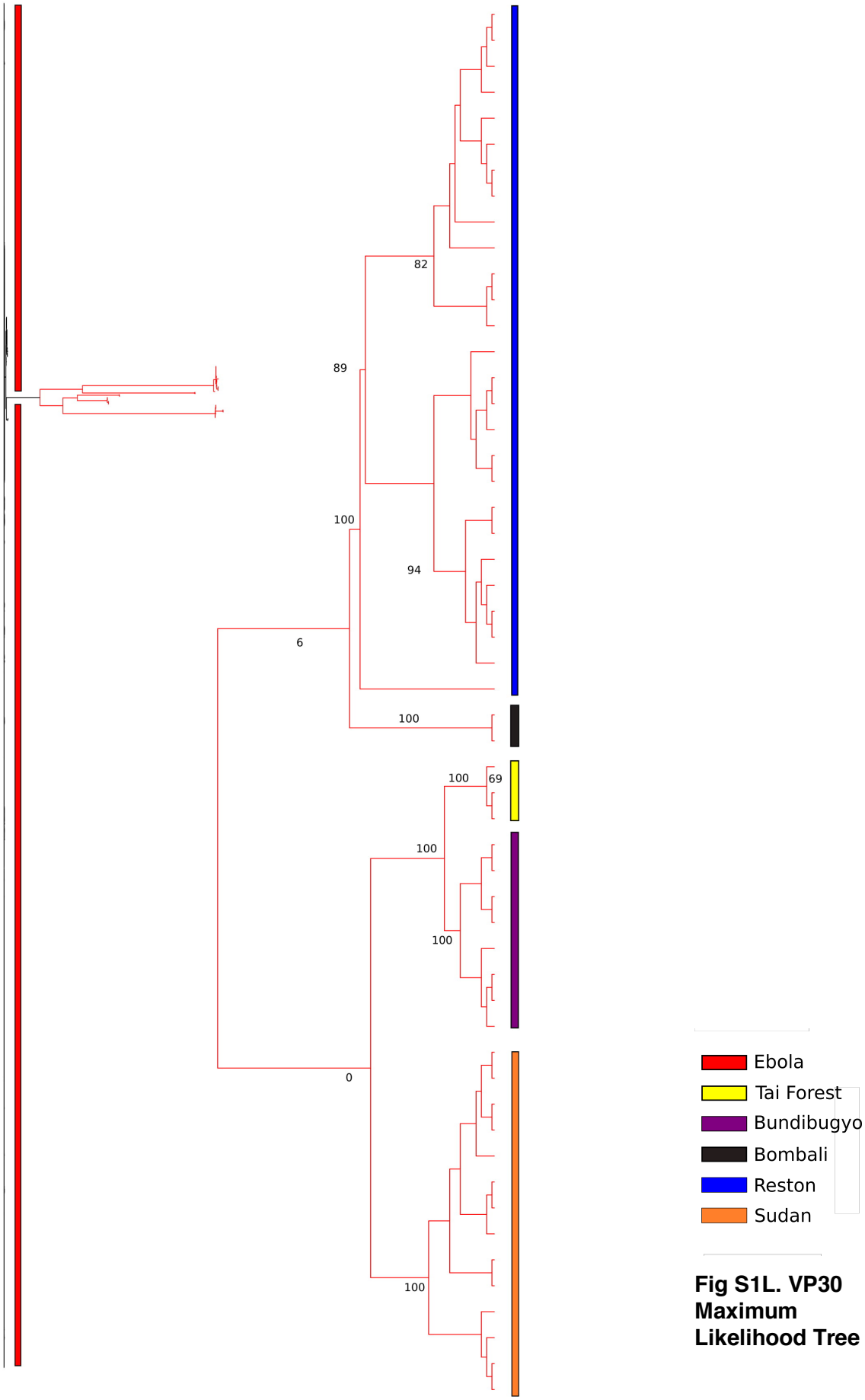


Fig S1L. VP30 Maximum Likelihood Tree

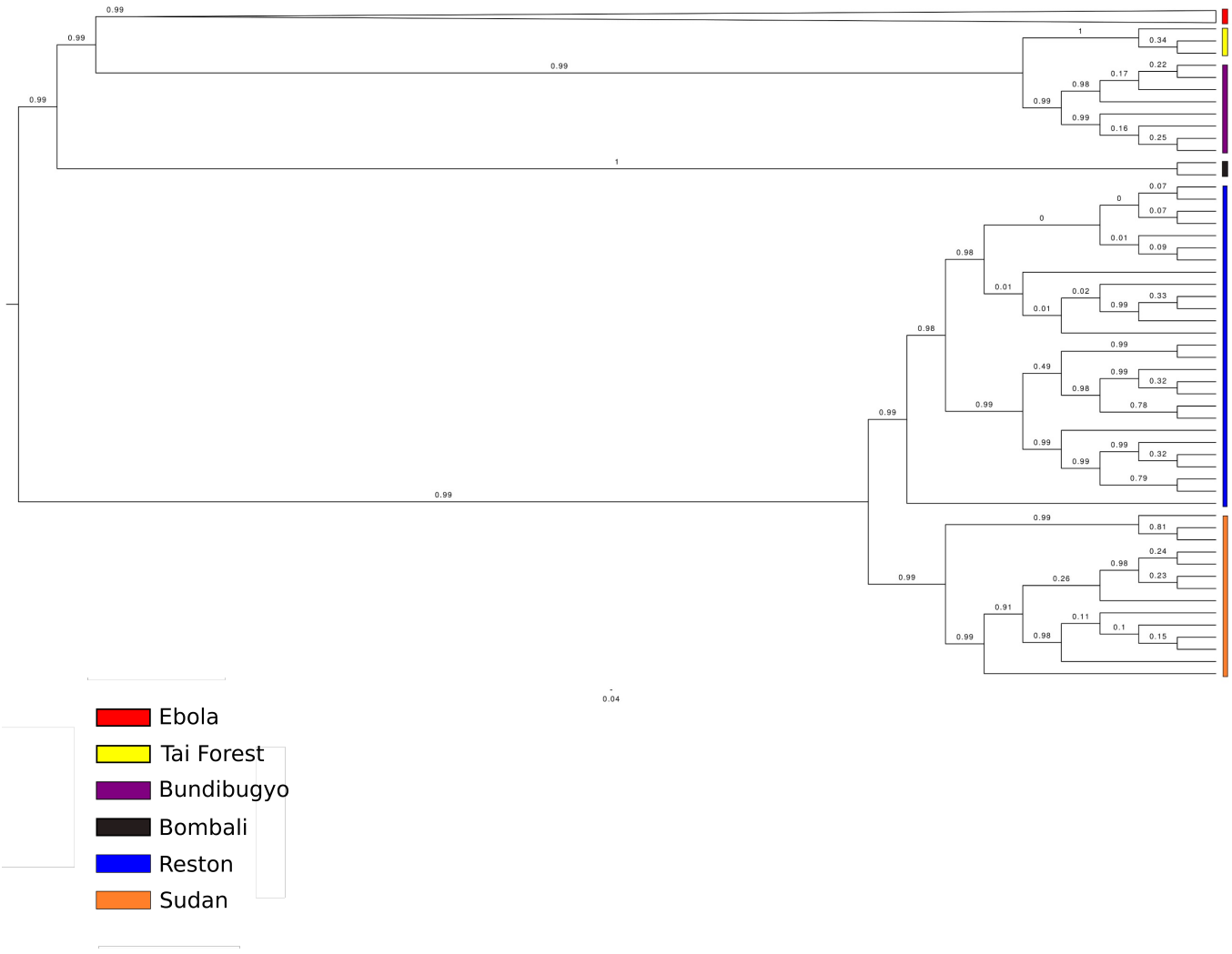


Fig S1M. VP35 Bayesian Tree

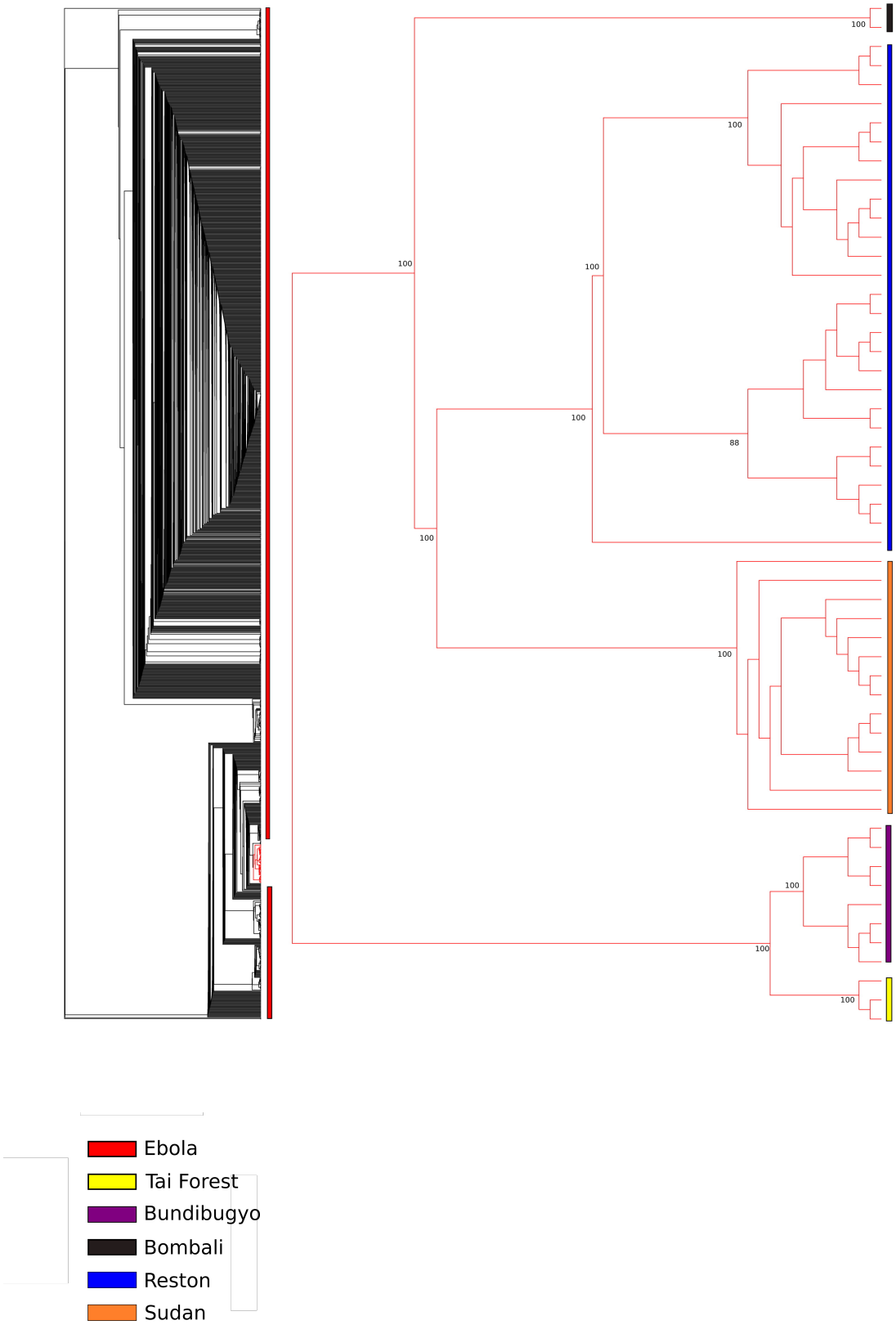


Fig S1N. VP35 Maximum Likelihood Tree

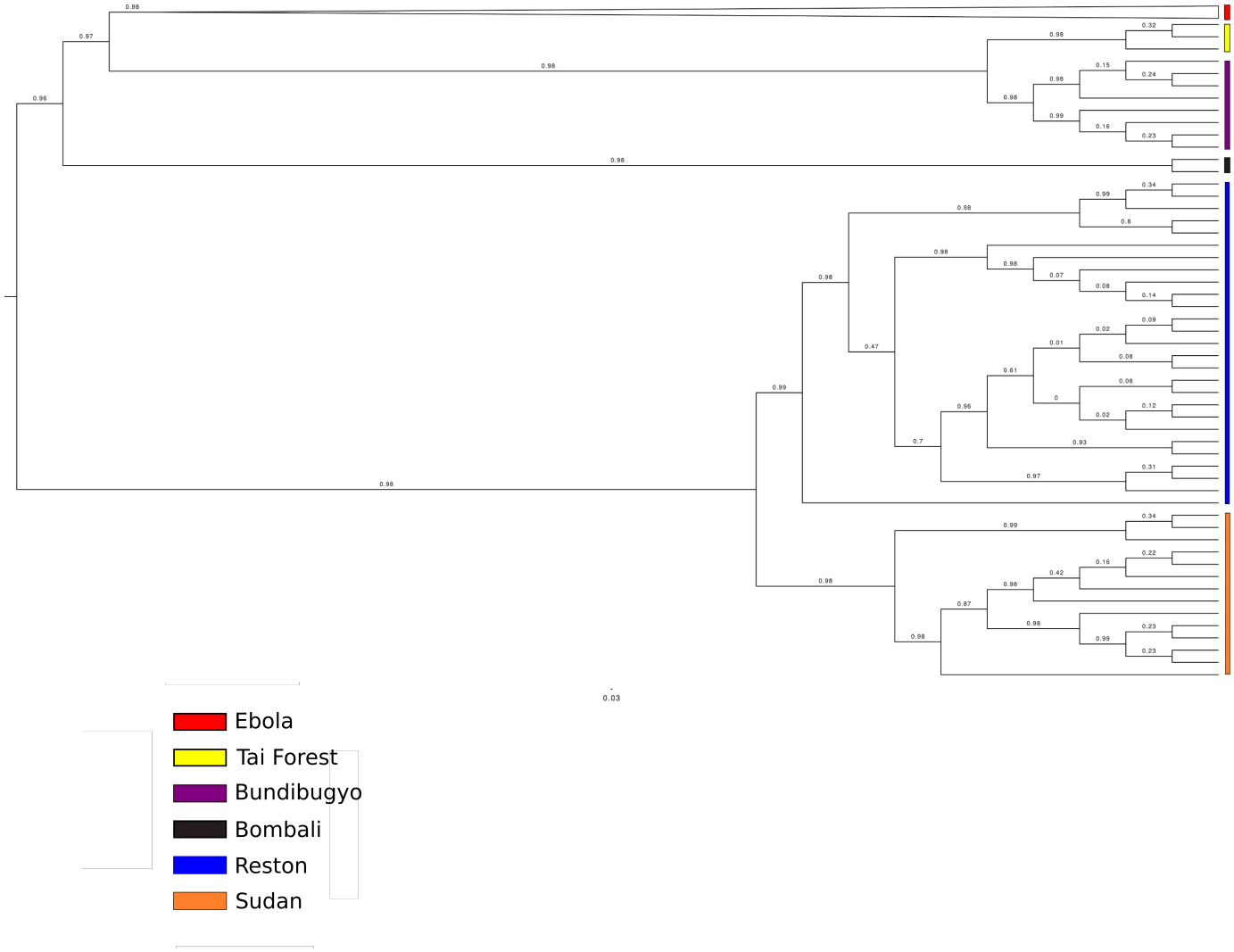


Fig S1O. Vp40 Bayesian Tree

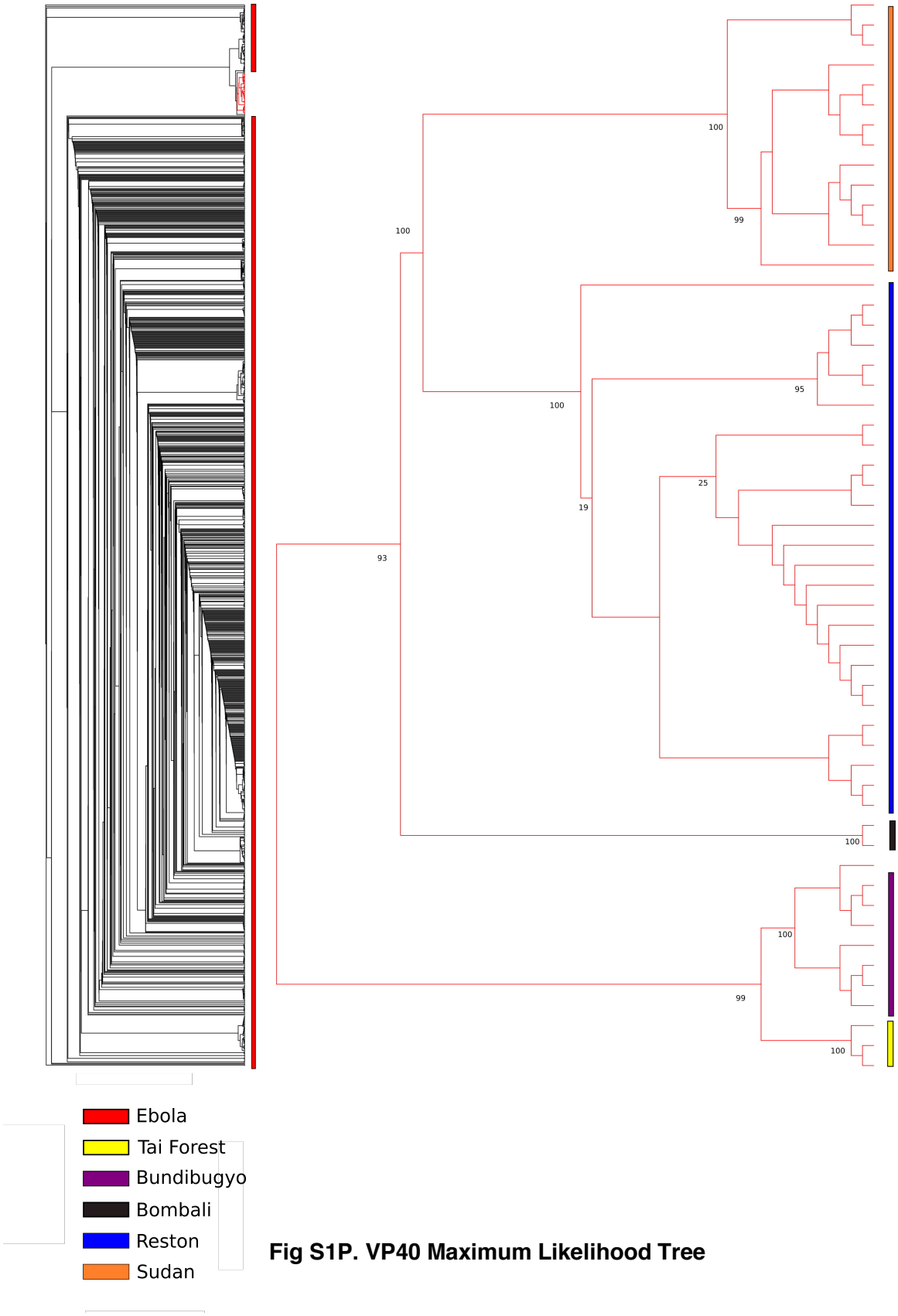
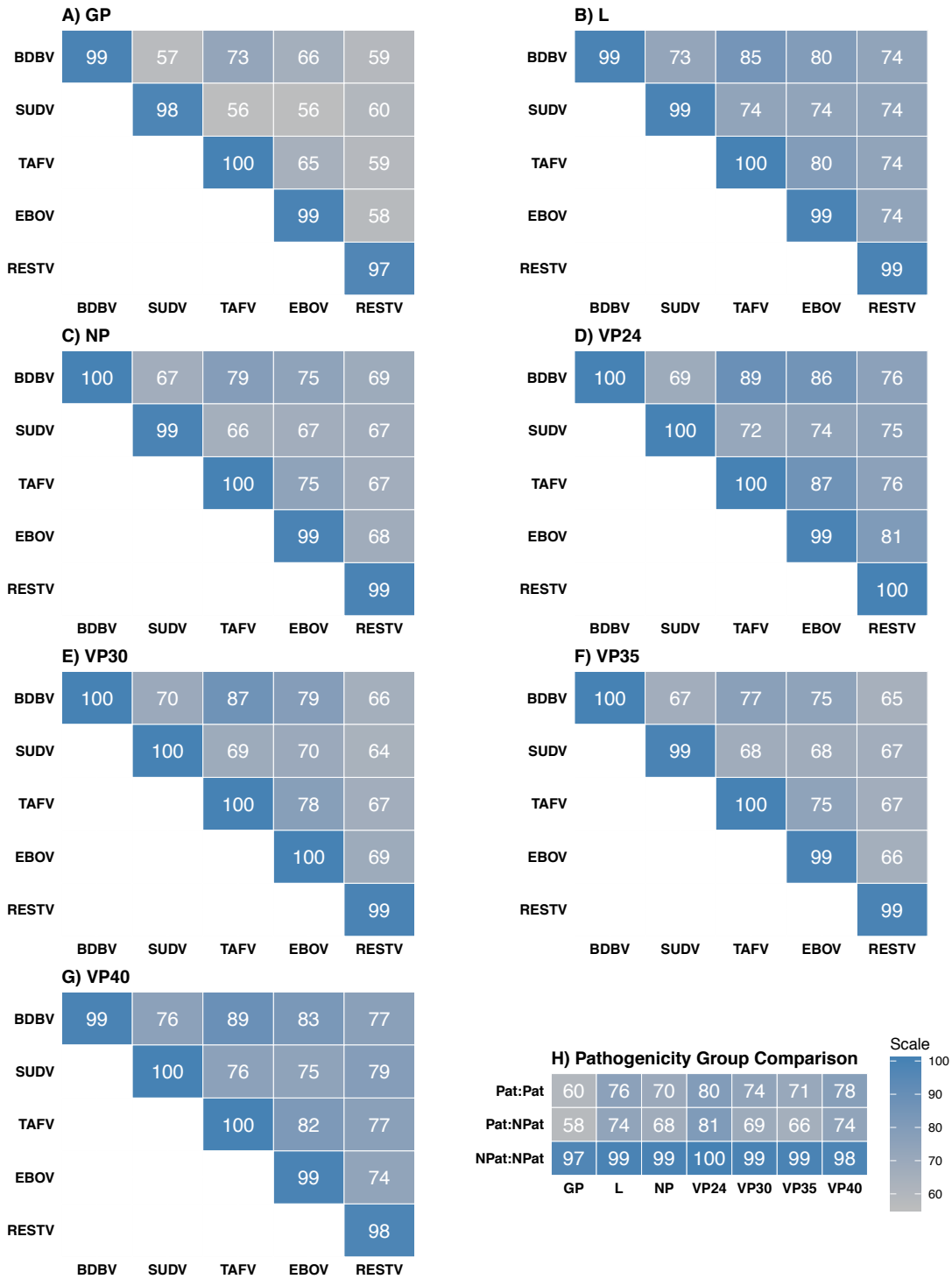
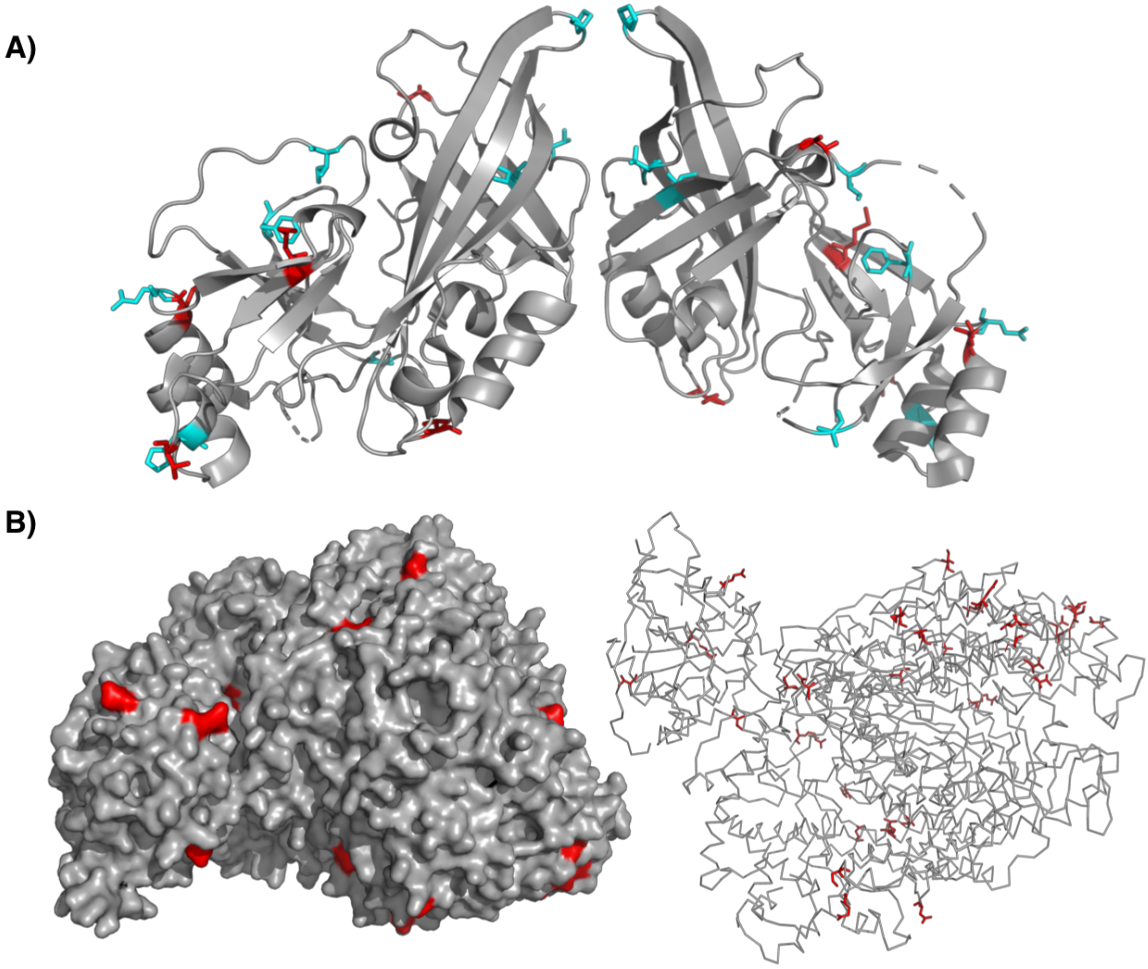


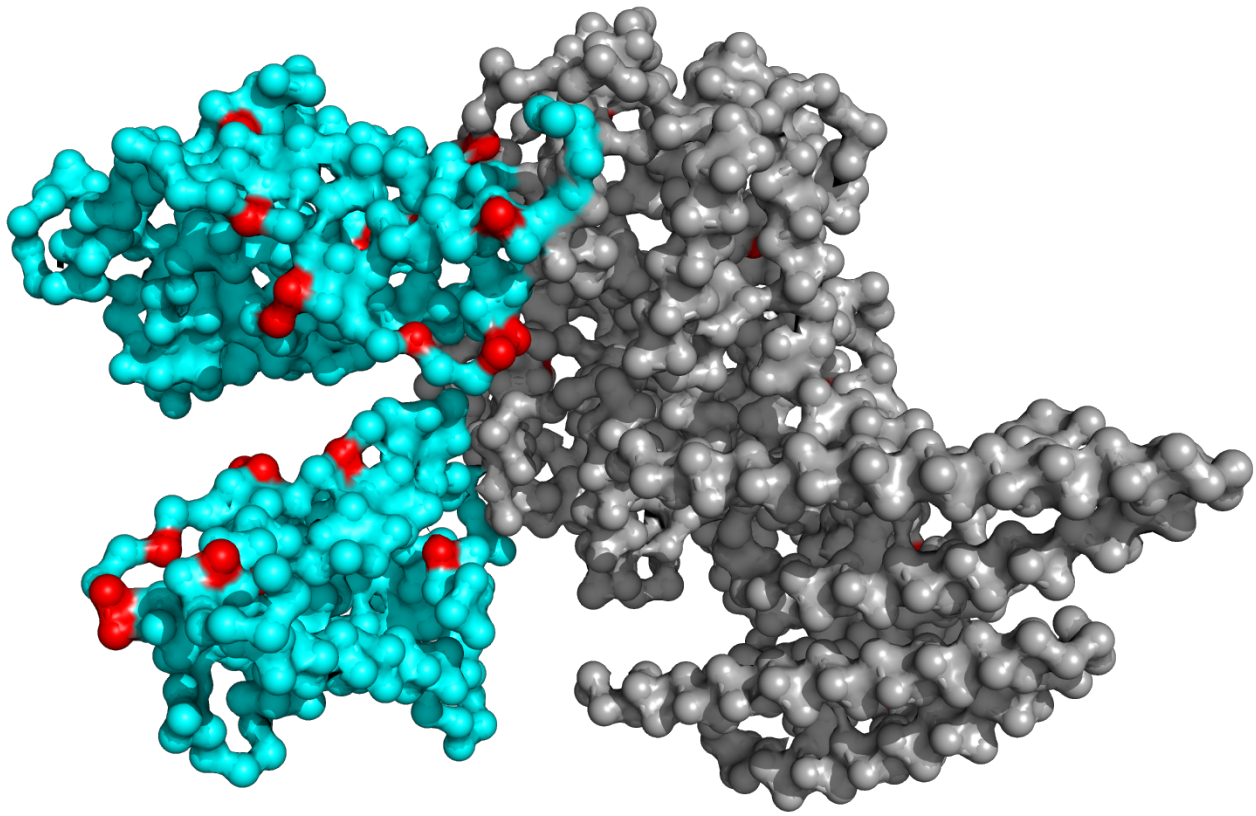
Fig S1P. VP40 Maximum Likelihood Tree



Supplementary Figure 2: Intra- and inter-species protein sequence conservation for each of the 7 ebolavirus proteins, and conservation between pathogenic and non-pathogenic groups. **A)** GP gene. **B)** L gene. **C)** NP gene. **D)** VP24 gene. **E)** VP30 gene. **F)** VP35 gene. **G)** VP40 gene. **H)** Comparison of pathogenicity groups for all 7 proteins.



Supplementary Figure 3: SDPs mapped to VP40 and L. **A)** SDPs identified in VP40 – VP40 is shown in cartoon format and coloured grey, SDPs are shown in stick format with retained SDPs coloured cyan and gained SDPs coloured red. **B)** SDPs mapped to the Phyre2 structure of L, shown as a surface representation and as a ribbon representation – L is shown in grey and SDPs are shown in red.



Supplementary Figure 4: Model of the *Ebola virus* nucleocapsid subunit from recombinant virus-like particles using Cryo-EM (resolution 7.3 angstroms) featuring VP24 (cyan) and Nucleoprotein (grey). SDPs are shown in red.

Supplementary Tables

Supplementary Table 1: SDPs identified for the gene GP. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
2	M1(1356)	M1(14)	M1(8)	M1(3)	G2(27)	N/A	M1	Retained
32	F31(1355);S31(1)	F31(14)	F31(8)	F31(3)	I32(27)	V27	F31	Retained
38	V37(1356)	V37(14)	V37(8)	V37(3)	I38(27)	V33	V37	Retained
46	V45(1356)	V45(14)	V45(8)	V45(3)	A46(27)	V41	V45	Retained
76	V75(1349);A75(7)	V75(14)	V75(8)	V75(3)	I76(27)	I71	V75	Retained

197	S196(1356)	S196(14)	S196(8)	S196(3)	A197(27)	P192	S196	Retained
261	I260(1356)	I260(14)	I260(8)	I260(3)	L261(27)	I256	I260	Retained
270	T269(1356)	T269(14)	T269(8)	T269(3)	S270(27)	T265	T269	Retained
308	X307(1);S307(1355)	S307(14)	S307(8)	S307(3)	H308(27)	S303	S307	Retained
497	X476(4);S476(1352)	S476(14)	S476(8)	L476(3)	P477(27)	I479	S476	Gained
519	R498(1353);X498(3)	R498(14)	R498(8)	R498(3)	K499(27)	R494	R498	Retained
521	X500(2);R500(1354)	R500(14)	R500(8)	R500(3)	K501(27)	K496	R500	Retained
535	N514(1354);X514(2)	N514(14)	N514(8)	N514(3)	D515(27)	N510	N514	Retained
542	X521(1);Q521(1355)	Q521(14)	Q521(8)	L521(3)	V522(27)	H517	Q521	Retained
605	I584(1356)	I584(14)	I584(8)	I584(3)	L585(27)	I580	I584	Retained
628	D607(1354);X607(2)	D607(14)	D607(8)	D607(3)	S608(27)	D603	D607	Retained
643	X622(1);K622(1355)	K622(14)	K622(8)	K622(3)	E623(27)	R618	K622	Retained
659	Q638(1352);L638(1);R638(1);X638(2)	Q638(14)	Q638(8)	Q638(3)	H639(27)	Q634	Q638	Retained
665	X644(2);W644(1354)	W644(14)	W644(8)	W644(3)	L645(27)	W640	W644	Retained
680	A659(1);T659(1354);X659(1)	T659(14)	T659(8)	T659(3)	I660(27)	V655	T659	Retained
3	G2(1356)	G2(11);E2(3)	V2(8)	G2(3)	S3(27)	N/A	G2	Lost
208	X207(57);E207(1299)	E207(14)	T207(8)	T207(3)	D208(27)	N203	E207	Lost
211	S210(1299);X210(57)	S210(14)	S210(8)	S210(3)	T211(27)	S206	S210	Lost
326	R325(1354);X325(2)	R325(14)	V325(8)	V325(3)	G326(27)	E321	R325	Lost
355	H354(1356)	H354(14)	R354(8)	Q354(3)	L355(27)	Q350	H354	Lost

417	X403(10);Q403(1346)	S412(14)	A409(8)	T409(3)	E412(27)	Q397	Q403	Lost
432	S418(1339);X418(15);X417(1);A417(1)	T427(14)	S419(8)	T419(3)	T422(27)	S412	S418	Lost
468	T448(1345);X448(8);A448(3)	-(14)	T451(8)	K451(3)	-(27)	T438	T448	Lost
537	H516(1355);X516(1)	H516(14)	H516(8)	H516(3)	H517(14);Y517(13)	N512	H516	Lost
568	L547(1352);X547(4)	L547(14)	I547(8)	I547(3)	V548(27)	L543	L547	Lost
663	D642(1354);X642(2)	D642(14)	D642(8)	S642(3)	L643(27)	D638	D642	Lost

Supplementary Table 2: SDPs identified for the gene L. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
67	V59(1); V66(1355)	V67(14)	V66(8)	V66(3)	T66(27)	V66	V66	Retained
137	I136(1355);I129(1)	I137(14)	I136(8)	I136(3)	L136(27)	L136	I136	Retained
147	L139(1); L146(1355)	L147(14)	L146(8)	L146(3)	V146(27)	L146	L146	Retained
203	T202(1346);S202(9);T195(1)	T203(14)	T202(8)	T202(3)	I202(27)	E202	T202	Gained
222	A221(1355);A214(1)	A222(14)	A221(8)	A221(3)	S221(27)	A221	A221	Retained
224	Q223(1355);Q216(1)	Q224(14)	Q223(8)	Q223(3)	L223(27)	Q223	Q223	Retained
227	T219(1); T226(1354);A226(1)	T227(14)	T226(8)	T226(3)	S226(27)	K226	T226	Gained
228	H227(1355);H220(1)	H228(14)	H227(8)	H227(3)	Q227(27)	Y227	H227	Retained
237	V236(1355);V229(1)	V237(14)	V236(8)	V236(3)	I236(27)	V236	V236	Gained
284	L283(1355);L276(1)	L284(14)	L283(8)	L283(3)	V283(27)	L283	L283	Retained
331	T323(1); T330(1355)	T331(14)	T330(8)	T330(3)	D330(27)	T330	T330	Retained
351	E343(1); E350(1355)	E351(14)	E350(8)	E350(3)	D350(27)	E350	E350	Retained
362	M361(1);T361(1353);T354(1);X361(1)	T362(14)	T361(8)	T361(3)	S361(27)	T361	T361	Retained
366	L358(1); L365(1354);X365(1)	L366(14)	L365(8)	L365(3)	F365(27)	F365	L365	Retained

380	V379(1353);V372(1);X379(2)	V380(14)	V379(8)	V379(3)	I379(27)	I379	V379	Retained
448	X447(4);Q447(1351);Q440(1)	Q448(14)	Q447(8)	Q447(3)	H447(27)	Q447	Q447	Retained
451	P450(1351);P443(1);X450(4)	P451(14)	P450(8)	P450(3)	S450(27)	P450	P450	Retained
466	D465(1355);D458(1)	D466(14)	D465(8)	D465(3)	N465(27)	D465	D465	Retained
848	X847(1);S847(1354);S840(1)	S848(14)	S847(8)	S847(3)	A847(27)	S847	S847	Retained
869	S861(1);S868(1355)	S869(14)	S868(8)	S868(3)	A868(27)	S868	S868	Retained
1025	T1017(1);T1024(1355)	T1025(14)	T1024(8)	T1024(3)	N1024(27)	N1024	T1024	Retained
1074	R1066(1);R1073(1355)	R1074(14)	R1073(8)	R1073(3)	K1073(27)	R1073	R1073	Retained
1120	A1112(1);A1119(1355)	A1120(14)	A1119(8)	A1119(3)	S1119(27)	A1119	A1119	Retained
1164	P1156(1);P1163(1355)	P1162(14)	P1163(8)	P1163(3)	A1161(27)	P1163	P1163	Retained
1190	D1189(1355);D1182(1)	D1188(14)	D1189(8)	D1189(3)	S1187(27)	D1189	D1189	Retained
1215	A1214(1355);A1207(1)	A1213(14)	A1214(8)	A1214(3)	S1212(27)	A1214	A1214	Retained
1218	R1210(1);R1217(1355)	R1216(14)	R1217(8)	R1217(3)	K1215(27)	R1217	R1217	Retained
1238	D1237(1355);D1230(1)	D1236(14)	D1237(8)	D1237(3)	E1235(27)	D1237	D1237	Retained
1355	R1354(1355);R1347(1)	R1353(14)	R1354(8)	R1354(3)	K1352(27)	R1354	R1354	Retained
1367	T1359(1);T1366(1355)	T1365(14)	T1366(8)	T1366(3)	A1364(27)	T1366	T1366	Retained
1409	I1408(1355);I1401(1)	I1407(14)	I1408(8)	I1408(3)	M1406(27)	I1408	I1408	Retained

1415	I1407(1);I1414(1355)	I1413(14)	I1414(8)	I1414(3)	L1412(27)	I1414	I1414	Retained
1437	S1429(1);S1436(1355)	S1435(14)	S1436(8)	S1436(3)	N1434(27)	S1436	S1436	Retained
1474	S1466(1);X1473(2);S1473(1353)	S1472(14)	S1473(8)	S1473(3)	C1471(27)	S1473	S1473	Retained
1489	L1488(1355);L1481(1)	L1487(14)	L1488(8)	L1488(3)	Y1486(27)	I1488	L1488	Retained
1500	I1499(1355);I1492(1)	I1498(14)	I1499(8)	I1499(3)	L1497(27)	I1499	I1499	Retained
1507	S1506(1355);S1499(1)	S1505(14)	S1506(8)	S1506(3)	A1504(27)	S1506	S1506	Retained
1510	I1509(1355);I1502(1)	I1508(14)	I1509(8)	I1509(3)	V1507(27)	V1509	I1509	Retained
1627	L1617(1);L1624(1355)	L1623(14)	L1624(8)	L1624(3)	Y1624(27)	L1624	L1624	Retained
1631	C1628(1355);C1621(1)	C1627(14)	C1628(8)	C1628(3)	S1628(27)	C1638	C1628	Retained
1786	V1755(1);V1762(1355)	V1759(14)	V1762(8)	V1762(3)	I1760(27)	V1760	V1762	Retained
1874	V1843(1);V1850(1355)	V1847(14)	V1850(8)	V1850(3)	T1848(27)	T1848	V1850	Retained
1897	I1873(1);T1866(1);T1873(1354)	T1870(14)	T1873(8)	T1873(3)	S1871(27)	T1871	T1873	Retained
1941	R1909(1);R1916(1355)	R1913(14)	R1916(8)	R1916(3)	N1915(3);N1914(24)	R1914	R1916	Retained
1966	E1941(1354);X1941(1);E1934(1)	E1938(14)	E1941(8)	E1941(3)	R1939(24);R1940(3)	E1939	E1941	Retained
2069	L2044(1355);L2037(1)	L2041(14)	L2044(8)	L2044(3)	I2043(3);I2042(24)	L2044	L2044	Retained
2102	S2077(1355);S2070(1)	S2074(14)	S2077(8)	S2077(3)	T2075(24);T2076(3)	S2075	S2077	Retained
2123	E2091(1);E2098(1355)	E2095(14)	E2098(8)	E2098(3)	D2096(24);D2097(3)	E2096	E2098	Retained
2182	L2157(1353);X2	L2154(14)	L2157(8)	L2157(3)	V2155(24);V2156(3)	L2155	L2157	Retained

	157(2);L2150(1)							
2193	R2168(1355);R2161(1)	R2165(14)	R2168(8)	R2168(3)	H2167(3);H2166(24)	K2166	R2168	Retained
2200	R2168(1);R2175(1355)	R2172(14)	R2175(8)	R2175(3)	K2173(24);K2174(3)	R2173	R2175	Retained
2202	X2177(1);L2177(1354);L2170(1)	L2174(14)	L2177(8)	L2177(3)	F2175(24);F2176(3)	W2175	L2177	Retained
2211	X2186(2);M2179(1);M2186(1353)	M2183(14)	M2186(8)	M2186(3)	L2185(3);L2184(24)	M2184	M2186	Retained
110	Q109(1298);X109(57);Q102(1)	Q110(14)	Q109(8)	Q109(3)	R109(2);H109(25)	Q109	Q109	Lost
277	L276(1355);X269(1)	L277(14)	L276(8)	L276(3)	I276(27)	L276	L276	Lost
313	Y312(1354);X305(1);X312(1)	Y313(14)	Y312(8)	Y312(3)	F312(27)	Y312	Y312	Lost
327	A319(1);X326(1);A326(1354)	A327(14)	A326(8)	A326(3)	S326(27)	A326	A326	Lost
690	E689(1353);E682(1);X689(2)	E690(14)	E689(8)	E689(3)	S689(27)	E689	E689	Lost
897	X896(58);F896(1297);F889(1)	F897(14)	F896(8)	F896(3)	Y896(27)	F896	F896	Lost
926	L925(1352);X925(3);L918(1)	L926(14)	L925(8)	L925(3)	F925(27)	L925	L925	Lost
955	X954(2);A954(1353);A947(1)	A955(14)	A954(8)	A954(3)	S954(27)	A954	A954	Lost
996	X995(2);S995(1353);S988(1)	S996(14)	S995(8)	S995(3)	T995(27)	S995	S995	Lost
1256	V1255(1);I1248(1);I1255(1354)	I1254(14)	I1255(8)	I1255(3)	V1253(27)	I1255	I1255	Lost

1396	A1395(1);S1395(1353);S1388(1);X1395(1)	S1394(14)	S1395(8)	S1395(3)	T1393(27)	T1395	S1395	Lost
1462	X1461(1);K1454(1);K1461(1354)	K1460(14)	K1461(8)	K1461(3)	Q1459(27)	I1461	K1461	Lost
1539	X1538(1);A1538(1354);A1531(1)	A1537(14)	A1538(8)	A1538(3)	S1536(27)	A1538	A1538	Lost
2033	X2008(57);L2001(1);L2008(1298)	L2005(14)	L2008(8)	L2008(3)	I2007(3);I2006(24)	L2006	L2008	Lost
2130	X2105(2);Q2098(1);Q2105(1353)	Q2102(14)	Q2105(8)	Q2105(3)	L2104(3);L2103(24)	Q2103	Q2105	Lost
2133	Q2108(1353);Q2101(1);X2108(2)	Q2105(14)	Q2108(8)	Q2108(3)	E2107(3);E2106(24)	Q2106	Q2108	Lost
2156	Y2124(1);Y2131(1354);X2131(1)	Y2128(14)	Y2131(8)	Y2131(3)	F2129(24);F2130(3)	Y2129	Y2131	Lost

Supplementary Table 3: SDPs identified for the gene NP. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
4	R4(1356)	R4(14)	R4(8)	R4(3)	G4(27)	R4	R4	Retained
16	X16(1); E16(1355)	E16(14)	E16(8)	G16(3)	D16(27)	D16	E16	Retained
30	S30(1356)	S30(14)	S30(8)	S30(3)	T30(27)	S30	S30	Retained
39	R39(1356)	R39(14)	R39(8)	R39(3)	K39(27)	R39	R39	Retained
56	I56(1356)	I56(14)	I56(8)	I56(3)	V56(27)	I56	I56	Retained
64	V64(1356)	V64(14)	V64(8)	V64(3)	I64(27)	V64	V64	Retained
105	R105(1354);X105(2)	R105(14)	R105(8)	R105(3)	K105(27)	R105	R105	Retained
137	M137(1354);X137(2)	M137(14)	M137(8)	M137(3)	L137(27)	M137	M137	Retained
212	X212(1); F212(1355)	F212(14)	F212(8)	F212(3)	Y212(27)	F212	F212	Retained
274	K274(1355);X274(1)	K274(14)	K274(8)	K274(3)	R274(27)	R274	K274	Retained
279	X279(1); S279(1355)	S279(14)	S279(8)	S279(3)	A279(27)	S279	S279	Retained
416	X416(1); K416(1355)	K416(14)	K416(8)	K416(3)	N416(27)	R416	K416	Retained
421	X421(1); Y421(1355)	Y421(14)	Y421(8)	Y421(3)	Q421(27)	Y421	Y421	Retained
426	D426(1356)	D426(14)	D426(8)	D426(3)	E426(27)	E426	D426	Retained
435	D435(1356)	D435(14)	D435(8)	D435(3)	N435(27)	D435	D435	Retained
443	D443(1356)	D443(14)	D443(8)	D443(3)	E443(27)	V443	D443	Retained
453	T453(1356)	T453(14)	T453(8)	T453(3)	I453(27)	T453	T453	Retained
497	P497(1316);S497(40)	P497(14)	P497(8)	R497(3)	A497(27)	S497	P497	Retained
571	T563(1348);X56	T563(14)	T563(8)	T563(3)	S563(27)	A563	T563	Retained

	3(7);--(1)							
573	--(1);X565(7);I565(1348)	I565(14)	I565(8)	I565(3)	V565(27)	I565	I565	Retained
610	X602(24);P602(1332)	P602(14)	P602(8)	N602(3)	T602(27)	R602	P602	Retained
650	X641(5);N641(1351)	N641(14)	N641(8)	K641(3)	Q641(27)	S641	N641	Retained
714	A705(1356)	A705(14)	A705(8)	A705(3)	R705(27)	A705	A705	Retained
726	G717(1354);X717(2)	G717(14)	G717(8)	D717(3)	N717(27)	G717	G717	Retained
42	Q42(9);S42(1);P42(1346)	P42(14)	P42(8)	Q42(3)	S42(27)	P42	P42	Lost
374	R374(1);K374(1355)	K374(14)	K374(8)	K374(3)	R374(27)	E374	K374	Lost
492	X492(57);D492(1299)	D492(14)	D492(8)	D492(3)	E492(27)	E492	D492	Lost
530	P526(1356)	V526(14)	G524(4);S524(4)	N524(3)	V530(1);A530(26)	P526(1);L526(1)	P526	Lost
725	D716(1354);N716(1);X716(1)	D716(14)	D716(8)	D716(3)	N716(27)	D716	D716	Lost

Supplementary Table 4: SDPs identified for the gene VP24. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
17	X17(1);L17(1355)	L17(14)	L17(8)	L17(3)	M17(27)	L17	L17	Retained
22	X22(1); V22(1355)	V22(14)	V22(8)	V22(3)	I22(27)	V22	V22	Retained
31	V31(1356)	V31(14)	V31(8)	V31(3)	I31(27)	V31	V31	Retained
102	I102(1354);V102(2)	I102(14)	I102(8)	I102(3)	L102(27)	I102	I102	Gained
131	T131(1356)	T131(14)	T131(8)	T131(3)	S131(27)	T131	T131	Retained
132	N132(1356)	N132(14)	N132(8)	N132(3)	T132(27)	A132	N132	Retained
136	I136(15);M136(1341)	M136(14)	M136(8)	M136(3)	L136(27)	L136	M136	Retained
139	Q139(1356)	Q139(14)	Q139(8)	Q139(3)	R139(27)	R139	Q139	Retained
140	R140(1356)	R140(14)	H140(8)	Q140(3)	S140(27)	R140(1);S140(1)	R140	Gained
226	X226(4); T226(1352)	T226(14)	T226(8)	T226(3)	A226(27)	T226	T226	Retained
248	S248(1356)	S248(14)	S248(8)	S248(3)	L248(27)	S248	S248	Retained

Supplementary Table 5: SDPs identified for the gene VP30. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
40	H39(2); X39(1); Y39(1353)	Y39(14)	Y39(8)	Y39(3)	R40(27)	N50	Y39	Gained
53	X52(1); T52(1355)	T52(14)	T52(8)	T52(3)	N53(27)	V63	T52	Retained
54	X53(1); V53(1355)	V53(14)	V53(8)	V53(3)	L54(27)	M64	V53	Retained
64	X63(3); T63(1353)	T63(14)	T63(8)	T63(3)	I64(27)	L74	T63	Retained
94	E93(1356)	E93(14)	E93(8)	E93(3)	D94(27)	E104	E93	Retained
97	T96(1355);X96(1)	T96(14)	T96(8)	T96(3)	N97(27)	T107	T96	Retained
99	R98(1355);X98(1)	R98(14)	R98(8)	R98(3)	H99(27)	R109	R98	Retained
108	K107(1354);X107(2)	K107(14)	K107(8)	K107(3)	R108(27)	K118	K107	Retained
112	S111(1356)	S111(14)	S111(8)	S111(3)	I112(27)	L122	S111	Retained
117	X116(1); L116(1355)	L116(14)	L116(8)	L116(3)	S117(27)	V127	L116	Retained
118	N117(1356)	N117(14)	N117(8)	S117(3)	Q118(27)	C128	N117	Gained
121	A120(1356)	A120(14)	A120(8)	A120(3)	S121(27)	A131	A120	Retained
151	T150(1355);X150(1)	T150(14)	T150(8)	T150(3)	I151(27)	I161	T150	Retained
158	X157(1); Q157(1355)	Q157(14)	Q157(8)	Q157(3)	R158(27)	K168	Q157	Retained
160	X159(1); I159(1355)	I159(14)	I159(8)	I159(3)	L160(27)	L170	I159	Retained
206	E205(1356)	E205(14)	E205(8)	E205(3)	D206(27)	E216	E205	Retained
263	R262(1356)	R262(14)	R262(8)	R262(3)	A263(27)	K273	R262	Retained

269	S268(1356)	S268(14)	S268(8)	S268(3)	Q269(27)	A279	S268	Retained
272	E271(1356)	E271(14)	T271(8)	T271(3)	S272(27)	N282	E271	Gained
279	X278(1); G278(1355)	G278(14)	E278(8)	E278(3)	N279(27)	T289	G278	Gained
197	H196(1);R196(1354);X196(1)	R196(14)	R196(8)	R196(3)	H197(27)	R207	R196	Lost

Supplementary Table 6: SDPs identified for the gene VP35. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
59	X26(1);S7(1);S59(2);S26(1352)	S15(14)	S27(8)	S27(3)	T15(27)	S27	S26	Retained
81	E81(2);E48(1353);E29(1)	E37(14)	E49(8)	E49(3)	D37(27)	D49	E48	Retained
109	G76(3);X76(1);D76(1349);D109(2);D57(1)	D65(14)	D77(8)	D77(3)	E65(27)	D77	D76	Retained
117	X117(1);E65(1);E84(1346);E117(1);G84(1);X84(6)	E73(14)	A85(8)	E85(3)	K73(27)	D85	E84	Gained
118	X85(7);D85(1);X118(1);E66(1);E85(1345);E118(1)	E74(14)	E86(8)	D86(3)	K74(27)	E86	E85	Retained
125	S92(1348);X92(5);S73(1);S125(1);X125(1)	S81(14)	S93(8)	S93(3)	M81(27)	S93	S92	Retained
130	V130(2);V97(1350);V78(1);X97(3)	V86(14)	V98(8)	I98(3)	T86(27)	V98	V97	Retained
134	T101(1351);X101(2);T134(2);T82(1)	T90(14)	T102(8)	A102(3)	N90(27)	A102	T101	Retained
139	S106(1352);S87(1);X106(1);S139(2)	S95(14)	S107(8)	S107(3)	A95(27)	S107	S106	Retained

145	T112(1352);T145(2);T93(1);X112(1)	A101(3);T101(11)	T113(8)	I113(3)	S101(27)	S113	T112	Gained
154	V154(2);X121(2);V102(1);V121(1351)	V110(14)	V122(8)	M122(3)	I110(27)	V122	V121	Retained
158	A106(1);V125(1);A158(2);A125(1352)	A114(14)	T126(8)	A126(3)	G114(27)	A126	A125	Gained
187	A154(1353);A135(1);A187(2)	A143(14)	A155(8)	A155(3)	S143(27)	A155	A154	Retained
192	T140(1);T159(1353);T192(2)	T148(14)	T160(8)	T160(3)	V148(27)	T160	T159	Retained
193	E141(1);E160(1353);E193(2)	E149(14)	E161(8)	E161(3)	D149(27)	E161	E160	Retained
200	G200(2);X167(1);G148(1);G167(1352)	G156(14)	G168(8)	G168(3)	K156(27)	G168	G167	Retained
207	S207(2);S155(1);S174(1353)	S163(14)	S175(8)	S175(3)	A163(27)	S175	S174	Retained
214	I181(1353);I162(1);I214(2)	I170(14)	I182(8)	I182(3)	L170(27)	I182	I181	Retained
302	X269(2);E269(1351);E302(1);X302(1);E250(1)	E258(14)	E270(8)	E270(3)	D258(27)	D270	E269	Retained
323	A323(2);A271(1);X290(3);A290(1350)	A279(14)	A291(8)	A291(3)	V279(27)	I291	A290	Retained
347	X314(3);V314(1350);V295(1);V347(2)	V303(14)	V315(8)	V315(3)	A303(27)	V315	V314	Retained

362	Q329(1351);Q362(2);--(1);Q310(1);X329(1)	Q318(14)	Q330(8)	Q330(3)	K318(27)	Q330	Q329	Retained

Supplementary Table 7: SDPs identified for the gene VP40. For each species the amino acid residue at the alignment position is given followed by the sequence position, e.g. A1 for Alanine at sequence position 1, and the number of samples with this amino acid is given in brackets after. Where there is more than one amino acid at a position for a species these are separated by semi-colons. A 'Lost' status indicates the SDP was only found in the original analysis, a 'Retained' status indicates that the SDP was found in the old and new analysis, and a 'Gained' status indicates that the SDP is only found in the new analysis.

Alignment Position	EBOV	SUDV	BDBV	TAFV	RESTV	BOMV	EBOV REF	Status
4	X4(5);V4(1348);--(2);I4(1)	V4(14)	A4(8)	I4(3)	G4(27)	T4	V4	Gained
46	I46(1);T33(2);T46(1353)	T46(14)	T46(8)	T46(3)	V46(27)	I46(1);T46(1)	T46	Retained
85	P85(1352);P72(2);X85(2)	P85(14)	P85(8)	P85(3)	T85(27)	P85	P85	Retained
105	T92(2);T105(1352);X105(2)	M105(1);T105(13)	T105(8)	T105(3)	I105(27)	K105	T105	Gained
122	X122(3);I122(1351);I109(2)	I122(14)	I122(8)	I122(3)	V122(27)	I122	I122	Retained
128	X128(1);A128(1353);A115(2)	A128(14)	T128(8)	T128(3)	I128(27)	T128	A128	Gained
201	G201(1353);G188(2);X201(1)	G201(14)	G201(8)	G201(3)	N201(27)	N201	G201	Retained
209	F196(2);X209(3);F209(1351)	F209(14)	F209(8)	F209(3)	L209(27)	F209	F209	Retained
244	L244(1354);L231(2)	M244(14)	L244(8)	L244(3)	I244(27)	L244	L244	Gained
245	Q245(1353);Q232(2);X245(1)	Q245(14)	Q245(8)	Q245(3)	P245(27)	Q245	Q245	Retained
259	M246(2);X259(1);M259(1353)	I259(14)	M259(8)	M259(3)	V259(27)	V259	M259	Gained
269	R269(1);X269(1);H256(2);H269(1352)	H269(14)	H269(8)	H269(3)	Q269(27)	Q269	H269	Retained

277	T264(2); T277(1354)	S277(14)	T277(8)	T277(3)	Q277(27)	H277	T277	Gained
293	I280(2);I293(1353);X293(1)	I293(14)	I293(8)	I293(3)	V293(27)	I293	I293	Retained
323	V310(2); M323(1);A323(1);V323(1352)	L323(14)	V323(8)	V323(3)	H323(27)	A323	V323	Gained
325	E312(2); E325(1354)	E325(14)	E325(8)	E325(3)	D325(27)	E325	E325	Retained

Supplementary Table 8. Codon variation for each SDP residue of the non-pathogenic species (Reston virus). Only the 10 SDPs that showed any codon variation are included. All variants are synonymous.

Protein	SDP	Codons Present	Variant Type
GP	S196A	GCT:26, GCC:1	Synonymous
GP	T659I	ATT:26, ATC:1	Synonymous
L	L1488Y	TAT:14, TAC:13	Synonymous
L	I1509V	GTT:26, GTC:1	Synonymous
L	L2157V	GTG:14, GTT:13	Synonymous
NP	M137L	CTG:22, TTG:5	Synonymous
NP	K274R	CGT:26, CGC:1	Synonymous
VP30	T63I	ATA:26, ATT:1	Synonymous
VP35	S174A	GCG:21, GCA:6	Synonymous
VP35	I181L	CTT:22, CTA:5	Synonymous

Supplementary Table 9. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for GP. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
M1G	ATG	ATG	ATG	ATG		
F31I	TTT (1349), TTC (6), TCT (1)	TTT	TTC	TTT	Yes	All residues are F across all species except 1 Ebola sequence which has an S (TCT)
V37I	GTT (1302), GTC (54)	GTT	GTA	GTT		
V45A	GTT	GTA	GTA	GTG		
V75I	GTG (1349), GCG (7)	GTA	GTT	GTA	Yes	All residues are V across all species except 7 Ebola sequences which have an A (GCG)
S196A	TCA	TCA	TCA (4), TCG (4)	TCT		
I260L	ATA	ATT	ATT	ATC		
T269S	ACC	ACA (13), ACC (1)	ACC	ACA		
S307H	TCT	TCT	TCT	TCT		
S476P	AGC	TCC	TCT	CTC	Yes	All Ebola, Sudan and Bundibugyo sequences have an S residue, Tai Forest has an L
R498K	AGG (1299), AGA (54)	CGC	AGA	AGA		
R500K	CGA	AGA	CGG (4), CGC (4)	CGA		
N514D	AAT (1345), AAC (9)	AAC	AAC	AAC		
Q521V	CAG	CAA	CAA	TTG	Yes	All Ebola, Sudan and Bundibugyo sequences have a Q residue, Tai Forest has an L
I584L	ATC	ATA	ATA	ATA		
D607S	GAC	GAT	GAT	GAT		
K622E	AAA	AAA	AAA	AAA		
Q638H	CAG (1352), CGG (1), CTG (1)	CAG	CAA	CAG	Yes	All residues are Q across all species except two variants in the Ebola sequences, one producing an R and one producing an L
W644L	TGG	TGG	TGG	TGG		
T659I	ACA (1354), GCA (1)	ACT	ACG	ACA	Yes	All residues are T across all species except 1 Ebola sequence which has an A (GCA)

Supplementary Table 10. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for L. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
V66T	GTA	GTC	GTT	GTG		
I136L	ATC	ATT	ATT	ATT		
L146V	TTA	CTA (11), TTA (3)	CTG	TTA		
T202I	ACA (1346), TCA (9), ACG (1)	ACA	ACA	ACA	Yes	9 Ebola Residues have a TCA codon (S) while all other sequences across all species have codons for a T residue
A221S	GCG (1349), GCA (7)	GCT	GCG	GCT		
Q223L	CAA	CAA	CAA	CAA		
T226S	ACA (1355), GCA (1)	ACA	ACA	ACA	Yes	All residues are T across all species except 1 Ebola sequence which has an A (GCA)
H227Q	CAC	CAT (13), CAC (1)	CAT	CAC		
V236I	GTC	GTC (11), GTT (3)	GTT	GTC		
L283V	TTA (1354), TTG (1), CTA (1)	CTG (11), TTG (3)	TTA	TTA		
T330D	ACC (1355), ACT (1)	ACA	ACA	ACA		
E350D	GAA	GAG	GAA	GAA		
T361S	ACG (1353), ACA (1), ATG (1)	ACA	ACA	ACT	Yes	All residues are T across all species except 1 Ebola sequence which has an M (ATG)
L365F	CTT	TTA	CTC	CTC		
V379I	GTG	GTT	GTG	GTC		
Q447H	CAA	CAA	CAA	CAA		
P450S	CCG	CCA	CCA	CCA		
D465N	GAC	GAT	GAT	GAT		
S847A	TCC	TCA	TCT	TCT		
S868A	TCG	TCT	TCC	TCC		
T1024N	ACT (1341), ACC (14)	ACG	ACA	ACA		
R1073K	AGA	AGG (11), AGA (3)	AGG	CGA		
A1119S	GCA (1313), GCT (42), GCG (1)	GCT	GCA	GCA		
P1163A	CCA	CCA	CCA	CCT		
D1189S	GAT	GAT	GAC	GAT		
A1214S	GCA	GCT	GCT	GCA		
R1217K	AGA	AGA (11), AGG (3)	CGT	CGT		
D1237E	GAC (1353), GAT (3)	GAT	GAT	GAC		
R1354K	CGG	AGG	CGG	CGA		
T1366A	ACA	ACG	ACG	ACA		
I1408M	ATT	ATT	ATC	ATA		
I1414L	ATT	ATA	ATT	ATT		
S1436N	AGC	AGT	AGC	AGC		

S1473C	AGT	AGT	AGT	AGT		
L1488Y	CTT	CTC (11), CTT (3)	CTC	CTT		
I1499L	ATA	ATC (13), ATT (1)	ATC	ATA		
S1506A	TCA	TCC	TCG	TCG		
I1509V	ATA	ATC	ATA	ATC		
L1624Y	CTT	CTA	CTA	TTA		
C1628S	TGT	TGC	TGC	TGT		
V1762I	GTC (1342), GTT (14)	GTA	GTC	GTA		
V1850T	GTT	GTT	GTC (4), GTT (4)	GTA		
T1873S	ACT (1355), ATT (1)	ACC	ACC	ACT	Yes	All residues are T across all species except 1 Ebola sequence which has an I (ATT)
R1916N	AGG	AGG	AGA	AGG		
E1941R	GAA	GAA	GAG (4), GAA (4)	GAA		
L2044I	TTA	CTT (11), CTC (3)	TTA	CTT		
S2077T	TCA	TCG	TCA	TCT		
E2098D	GAA	GAG	GAA	GAG		
L2157V	TTG	CTT	TTA	CTA		
R2168H	AGA (1355), AGG (1)	CGT (11), CGC (3)	AGG	CGA		
R2175K	CGT	AGG	CGA	CGG		
L2177F	TTA	CTG	TTA	CTA		
M2186L	ATG	ATG	ATG	ATG		

Supplementary Table 11. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for NP. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
R4G	CGT	CGG	CGT	CGG		
E16D	GAA	GAA (11), GAG (3)	GAA	GGT	Yes	All Ebola, Sudan and Bundibugyo sequences have an E residue, Tai Forest has a G
S30T	TCC	TCG (11), TCA (3)	TCC	TCA		
R39K	AGA	AGA	AGA	CGG		
I56V	ATC	ATC	ATC	ATC		
V64I	GTT	GTA	GTC	GTT		
R105K	CGT	AGG	CGT	CGC		
M137L	ATG	ATG	ATG	ATG		
F212Y	TTT	TTC (11), TTT (3)	TTC	TTC		
K274R	AAA	AAG	AAA	AAG		
S279A	TCC	TCA	TCT	TCC		
K416N	AAA	AAG	AAA	AAG		
Y421Q	TAC (1354), TAT (1)	TAT	TAT	TAT		
D426E	GAC	GAT	GAC	GAT		
D435N	GAT	GAT	GAT	GAT		
D443E	GAT	GAT	GAT	GAC		
T453I	ACT	ACT	ACA	ACC		
P497A	CCA (1316), TCA (40)	CCA	CCG	CGA	Yes	Tai Forest sequences have an R residue. All other species have codons for a P residue except 40 Ebola sequences which have a TCA codon (S)
T563S	ACC (1295), ACA (53)	ACC	ACT	ACT		
I565V	ATC (1323), ATT (25)	ATA	ATC	ATC		
P602T	CCC	CCA	CCT	AAT	Yes	All Ebola, Sudan and Bundibugyo sequences have a P residue, Tai Forest has an N
N641Q	AAC	AAC	AAT	AAA	Yes	All Ebola, Sudan and Bundibugyo sequences have an N residue, Tai Forest has a K
A705R	GCC (1314), GCT (42)	GCC	GCC	GCC		
G717N	GGT	GGC	GGT	GAT	Yes	All Ebola, Sudan and Bundibugyo sequences have a G residue, Tai Forest has a D

Supplementary Table 12. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for VP24. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
L17M	CTG (1348), CTT (5), CTA (1)	CTA	CTC	CTT		
V22I	GTC	GTG (11), GTA (3)	GTT	GTT		
V31I	GTT	GTG	GTT	GTG		
I102L	ATA (1354), GTA (2)	ATT	ATT	ATT	Yes	All residues are I across all species except 2 Ebola sequences which have a V (GTA)
T131S	ACT (1355), ACC (1)	ACT	ACA	ACA		
N132T	AAC	AAT	AAC	AAC		
M136L	ATG (1341), ATA (15)	ATG	ATG	ATG	Yes	All residues are M across all species except 15 Ebola sequences which have an I (ATA)
Q139R	CAA	CAA	CAG	CAA		
R140S	CGT	CGA	CAC	CAG	Yes	All Ebola and Sudan have codons for an R residue. All Bundibugyo codons produce an H residue and all Tai Forest codons produce a Q residue
T226A	ACA	ACA (11), ACC (3)	ACC	ACC		
S248L	TCT	TCT	TCC	TCT		

Supplementary Table 13. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for VP30. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
Y39R	TAC (1352), CAC (2), TAT (1)	TAC	TAT	TAC	Yes	All residues are Y across all species except 2 Ebola sequences which have an H (CAC)
T52N	ACT	ACG (11), ACA (3)	ACT	ACT		
V53L	GTA (1339), GTG (16)	GTT (11), GTC (3)	GTG	GTC		
T63I	ACA	ACT	ACA	ACA		
E93D	GAA (1314), GAG (42)	GAA	GAA	GAA		
T96N	ACT (1341), ACG (14)	ACC	ACA	ACA		
R98H	AGG	CGG	AGG	AGA		
K107R	AAG (1354), AAA (1)	AAG	AAA	AAG		
S111I	TCA	TCA	TCC	TCC		
L116S	TTA	CTT	TTG	CTA		
N117Q	AAT	AAT	AAC	AGC	Yes	All Ebola, Sudan and Bundibugyo sequences have an N residue, Tai Forest has an S
A120S	GCA	GCT	GCT	GCT		
T150I	ACG (1344), ACA (11)	ACT	ACT	ACA		
Q157R	CAA	CAG	CAA	CAG		
I159L	ATC	ATT	ATC	ATT		
E205D	GAA	GAA	GAA	GAG		
R262A	AGA	CGC	AGG	AGA		
S268Q	TCA (1355), TCG (1)	AGC	TCA	TCG		
E271S	GAG	GAA	ACC	ACT	Yes	All Ebola and Sudan have codons for an S residue, All Bundibugyo and Tai Forest have codons for a T residue
G278N	GGG	GGG (11) GGA (3)	GAG	GAA	Yes	All Ebola and Sudan have codons for a G residue, All Bundibugyo and Tai Forest have codons for an E residue

Supplementary Table 14. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for VP35. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
S26T	TCG	TCT	TCC	TCA		
E48D	GAG (1352), GAA (4)	GAA	GAA	GAG		
D76E	GAC (1346), GGC (3), GAT (6)	GAT	GAC	GAT	Yes	All residues are D across all species except 3 Ebola sequences which have a G (GGC)
E84K	GAG (1348), GGG (1)	GAA	GCA	GAA	Yes	All Bundibugyo sequences have an A residue. All Ebola, Sudan and Tai Forest have an E residue except one Ebola sequence which has a G
E85K	GAG (1347), GAC (1)	GAA (11), GAG (3)	GAG	GAC	Yes	All Tai Forest and one Ebola sequences have a D residue. All other sequences have codons for an E residue
S92M	TCA	TCG	TCT	TCA		
V97T	GTG	GTG	GTA	ATA	Yes	All Ebola, Sudan and Bundibugyo sequences have a V residue, Tai Forest has an I
T101N	ACC (1351), ACT (3)	ACC	ACC	GCT	Yes	All Ebola, Sudan and Bundibugyo sequences have a T residue, Tai Forest has an A
S106A	TCA	TCA	TCA	TCT		
T112S	ACG	ACA (11), GCA (3)	ACT (4), ACC (4)	ATA	Yes	All Ebola and Bundibugyo have a T residue. All Tai Forest have an I residue. 11 Sudan sequence have a T residue while 3 have an A residue
V121I	GTT	GTT	GTG	ATG	Yes	All Ebola, Sudan and Bundibugyo sequences have a V residue, Tai Forest has an M
A125G	GCA (1337), GCT (18), GTA (1)	GCA	ACC	GCT	Yes	All Bundibugyo sequences have a T residue. All others have an A residue except 1 Ebola sequence which has a V (GTA)
A154S	GCA	GCC	GCC	GCC		
T159V	ACT	ACA	ACT	ACT		
E160D	GAG (1342), GAA (14)	GAA	GAA	GAG		
G167K	GGT	GGA	GGA	GGA		
S174A	TCA	TCA	TCA	TCA		
I181L	ATT	ATT (11), ATC (3)	ATC	ATT		
E269D	GAA (1352), GAG (1)	GAG	GAA	GAA		
A290V	GCT	GCC	GCA	GCC		
V314A	GTC	GTC	GTT	GTT		
Q329K	CAG	CAA	CAG	CAA		

Supplementary Table 15. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for VP40. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
V4G	GTT (1348), ATT (1)	GTC (11), GTT (3)	GCA	ATC	Yes	All Bundibugyo sequences have an A residue. All Tai Forest have an I. All Ebola and Sudan have a V except one Ebola has an I (ATT)
T46V	ACT (1355), ATT (1)	ACA	ACA	ACT	Yes	All residues are D across all species except 1 Ebola sequence which has an I (ATT)
P85T	CCC (1348), CCT (6)	CCC	CCG	CCG		
T105I	ACC (1355), ATG (1)	ACG (11), ACA (3)	ACA	ACA	Yes	All residues are T across all species except 1 Ebola sequence which has an M (ATG)
I122V	ATC	ATC	ATC	ATC		
A128I	GCA	GCC	ACC	ACC	Yes	All Ebola and Sudan have codons for an A residue, All Bundibugyo and Tai Forest have codons for a T residue
G201N	GGA	GGA	GGA	GGC		
F209L	TTT	TTT	TTT	TTC		
L244I	CTC	ATG	CTC	CTA	Yes	All Ebola, Bundibugyo and Tai Forest have an L residue. All Sudan have an M residue
Q245P	CAG	CAA	CAA	CAA		
M259V	ATG	ATT	ATG	ATG	Yes	All Ebola, Bundibugyo and Tai Forest have an M residue. All Sudan have an I residue
H269Q	CAC (1354), CGC (1)	CAC	CAC	CAC	Yes	All residues are H across all species except 1 Ebola sequence which has an R (CGC)
T277Q	ACT (1355), ACC (1)	AGT	ACA	ACT	Yes	All Ebola, Bundibugyo and Tai Forest have a T residue. All Sudan have an S residue
I293V	ATT (1354), ATC (1)	ATT	ATT	ATT		
V323H	GTG (1354), ATG (1), GCG (1)	CTC	GTC	GTC	Yes	All Sudan have an S residue All others have a T residue except 2 Ebola sequences. One has an M residue (ATG) and one has an A (GCG).
E325D	GAG	GAA	GAG	GAA		

Supplementary Table 16. Codons present for each SDP residue of each pathogenic species, alongside information about non-synonymous variants for the lost SDPs. Highlighted in yellow are the codons that are 100% conserved across all pathogenic species.

SDP	Ebola (1356)	Sudan (14)	Bundibugyo (8)	Tai Forest (3)	Non-Synonymous?	Notes
GP						
G2S	GGT:1343, GGC:13	GGG:11, GAG:3	GTT	GGA	Yes	3 Sudan sequences have an E residue, all Bundibugyo sequences have a V residue, all other sequences have a G
E207D	GAG	GAA	ACA	ACG	Yes	Ebola and Sudan have an E residue, Bundibugyo and Tai Forest have a T
S210T	TCG:1245, TCT:51, TCC:11	TCA	TCC	TCC		
R325G	CGA	AGA	GTC	GTC	Yes	Ebola and Sudan have an R residue, Bundibugyo and Tai Forest have a V
H354L	CAC:1355, CAT:1	CAC	CGA	CAA	Yes	Ebola and Sudan have an H residue, Bundibugyo has an R and Tai Forest has a Q
Q403E	CAA	CCA	CCA	CCA	Yes	Ebola has a Q residue, all other species have a P residue
S418T	TCC	CAC	CGC	CAC	Yes	Ebola has an S residue, Sudan and Tai Forest have an H residue, Bundibugyo has an R residue
T448M	ACC:1345, GCC:3	ACC	AGC	ACC	Yes	Sudan has an S residue, all others have a T residue except three Ebola sequences which have an A residue
H516H	CAT	CAC	CAC	CAC		
L547V	CTA:1323, CTG:29	CTG:11, CTT:3	ATA	ATA	Yes	Ebola and Sudan have an L residue, Bundibugyo and Tai Forest have an I
D642L	GAC	GAT	GAC	AGC	Yes	Ebola, Sudan and Bundibugyo have a D residue, Tai Forest has an S
L						
Q109H	CAA	CAG	CAG	CAA		
L276I	CTT	CTG	CTG	CTA		
Y312F	TAC	TAT	TAC	TAC		
A326S	GCT	GCA	GCC	GCT		
E689S	GAA:1341, GAG:13	GAG	GAA	GAA		
F896Y	TTC	TTC:11, TTT:3	TTC	TTT		
L925F	CTA:1350, TTA:3	TTG:11, CTG:3	CTT	CTG		
A954S	GCG:1336, GCA:18	GCA	GCC	GCA		
S995T	AGT	TCG	AGT	AGT		
I1255V	ATA:1355, GTA:1	ATT	ATC	ATT	Yes	All sequences have an I residue except one Ebola sequence which has a V residue
S1395T	TCA:1352, TCG:1, TCT:1, GCA:1	TCG	TCC	TCT	Yes	All sequences have an S residue except one Ebola sequence which has an A residue
K1461Q	AAA:1162, AAG:193	AAA	AAG	AAG		
A1538S	GCA	GCA	GCA	GCT		

L2008I	TTA	CTT	CTT	CTT		
Q2105L	CAA	CAG	CAG	CAA		
Q2108E	CAA	CAA	CAA	CAA		
Y2131F	TAT:1340, TAC:15	TAT	TAC	TAC		
NP						
P42S	CCA:1346, CAA:9, TCA:1	CCG	CCT	CAA	Yes	Tai Forest sequences have a Q residue, all others have a P except one Ebola sequence which has an S
K374R	AAA:1355, AGA:1	AAG	AAA	AAG	Yes	All sequences have a K residue except one Ebola sequence which has an R residue
D492E	GAC:1270, GAT:29	GAT	GAT	GAC		
P526A	CCA	GTG	GAA	CCG	Yes	Ebola and Tai Forest have a P residue, Sudan has a V residue and Bundibugyo has an E
D716N	GAT:1353, GAC:1, AAT:1	GAT	GAT	GAT	Yes	All sequences have a D residue except one Ebola sequence which has an N residue
VP35						
R196H	CGC:1336, CGT:18, CAC:1	AGG	CGA	CGA	Yes	All sequences have an R residue except one Ebola sequence which has a H residue

Supplementary Table 17. SDPs mapped to known PDB structures or modelled structures. *There was no available structure or model of L when the previous study was carried out

Protein	Total SDPs	Old Mapped	New Mapped	Total Mapped
VP24	11	8	2	10
VP30	20	5	0	5
VP35	22	4	11	15
VP40	16	8	5	13
NP	24	8	n/a	8
GP	21	10	0	10
L	51	0*	31	31
Total	165	43	49	92

Supplementary Table 18: Summary of the structures used for SDP investigation.

Protein	Species	PDB Structure ID	Oligomeric Form	Residue Coverage
VP24	EBOV	4M0Q	Homodimer	11 - 237
	EBOV	4U2X	Heterodimer (with KPNA5)	16 - 231
VP30	EBOV	2I8B	Homodimer	142 - 272
VP35	EBOV	4IBC	Homodimer	215 - 340
	EBOV	3L26	Homodimer (bound to RNA)	215 - 340
	EBOV	6GBO	Homotrimer	81 - 153
VP40	EBOV	4LDB	Homodimer	44 - 326
	EBOV	4LDD	Homo 6-mer	44 - 326
	EBOV	4LDM	Homo 8-mer	44 - 188
NP	EBOV	4QB0	Monomer	641 - 739
	EBOV	4YPI	Heterodimer	38 - 385
GP	EBOV	5JQ3	Hetero 6-mer	32 – 501 & 502 – 632
L	EBOV	N/A (Phyre2 model)	Monomer	8 - 2010
Nucleocapsid (NP with VP24)	EBOV	6EHM	Hetero 4-mer	NP: 1 – 739 VP24: 1 - 251

Supplementary Table 19: Summary of the SDPs with proposed functional impacts identified using the 196 and the 1,408 genome sets.

Protein	SDP	Functional Effect	Confidence	Status
NP	R105K	Stability: Loss of hydrogen bonding	Possible	Retained
NP	A705R	Stability: Introduction of salt bridge with E694	Possible	Lost
VP35	E269D	Interface: Dimeric VP35 interface	Probable	Retained
VP40	P85T	Interface: Octameric VP40 interface	Probable	Retained
VP40	Q245P	Stability: Breaks an alpha helix	Probable	Retained
GP	I260L	Interface: Within GP glycan cap	Possible	Retained
GP	T269S	Interface: Within GP glycan cap	Possible	Retained
GP	S307H	Interface: Within GP glycan cap	Possible	Retained
VP30	R262A	Interface: Dimer interface, loss of hydrogen bond	Probable	Retained
VP24	T131S	Interface: With KPNA5	Probable	Retained
VP24	M136L	Interface: With KPNA5	Probable	Retained
VP24	Q139R	Interface: With KPNA5	Probable	Retained
VP24	R140S	Interface: With KPNA5	Probable	Gained
VP24	T226A	Stability: Loss of hydrogen bond	Probable	Retained

Supplementary Table 20: Comparison of *Bombali virus* SDP amino acids with *Ebola virus* and *Reston virus*

Protein	Number of SDPs	Residues the same as: (%)		
		Ebola virus	Reston virus	Neither
GP	20	55	10	35
L	53	77.36	11.32	11.32
NP	24	62.5	12.5	25
VP24	11	72.73	18.18	9.09
VP30	20	30	10	60
VP35	22	72.73	13.64	13.64
VP40	16	50	18.75	31.25
Total	**166**	**63.25**	**12.65**	**24.1**

Supplementary Table 21: Summary of source databases used to obtain Ebolavirus genomes for analysis.

Species	NCBI	ViPR	Urbanowicz	Total
Ebola virus	1,469	43	505	2,017
Sudan virus	14	5	0	19
Bundibugyo virus	7	2	0	9
Tai Forest virus	1	3	0	4
Reston virus	18	9	0	27
Total	1,509	62	505	2,076

Supplementary Table 22: Summary of the sequences removed from the initial set of ebolavirus genome sequences.

Species	Starting	Removed	Final
Ebola virus	2,017	661	1,356
Sudan virus	19	5	14
Bundibugyo virus	9	1	8
Tai Forest virus	4	1	3
Reston virus	27	0	27
Total	2,076	668	1,408