University of Kent

Faculty of Sciences

A Computational study of Ebolavirus Pathogenicity and a Modeling approach for human non-synonymous variants.

A dissertation submitted for the degree of
Doctor of Philosophy
in the University of Kent for the Faculty of Sciences

Morena Pappalardo

Canterbury, 2016

Declaration:

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent or any other University or Institution of learning.

"Do not go where the path may lead,
go instead where there is no path
and leave a trail."

Ralph Waldo Emerson

Table of contents

Abstract

Recent advances in genome sequencing are improving our better understanding of genetic variation. However, the investigation of the genotype-phenotype relationship is still challenging, especially for the interpretation of the myriad of discovered genetic variants that weakly relate to disease.
Recently, researchers have confirmed that disease causing genetic variants typically occur at functional sites, such as protein-protein or protein-ligand interaction sites. Giving this observation, several bioinformatics tools have been developed. This thesis first details VarMod (Variant Modeller), an algorithm that predicts whether nonsynonymous single nucleotide variants (nsSNVs) affect protein function.

The recent Ebola virus outbreak in West Africa demonstrated the potential for the virus to cause edipdemics and highlighted our limited understanding of Ebola virus biology. The second part of this thesis focuses on the investigation of the molecular determinants of Ebolavirus pathogenicity. In two related analyses knowledge of differing pathogenicity of Ebolavirus species is used. Firstly, comparison of the sequences of Reston viruses (the only Ebolavirus species that is not pathogenic in humans) with the four pathogenic Ebolavirus species, enabled the identification of Specificity Determining Positions (SDPs) that are differentially conserved between these two groups. These SDPs were further investigated using analysis of protein structure and identified variation in the Ebola virus VP24 as likely to have a role in determining species-specific pathogenicity. The second approach investigated rodent-adapted Ebola virus. Ebola virus is not pathogenic in rodents but it can be passaged to induce pathogenicity. Analysis of the mutations identified in four adaption studies identified that very few mutations are required for adaptation to a new species and once again the VP24 is likely to have a central role. Subsequent molecular dynamics simulations compared the interaction of Ebola and Reston virus VP24 with human karyopherin alpha5. The analysis suggests that Reston virus VP24 has weaker binding with karyopherins and we propose that this change in binding may reduce the ability of Reston VP24 to inhibit human interferon signaling.

List of abbreviations

BDBV	Bundibugyo Ebolavirus
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOCKS SUbstitution Matrix
bp	base pair
CASP	Critical Assessment of Techniques for Protein Structure Prediction
DSSP	Dictionary of Protein Secondary Structure
dsRNA	Double strand DNA
EBOV	Zaire Ebolavirus
FASTA	Fast Alignment Search Tool
FDA	Food and Drug Admministration
GP	Glycoprotein
HMM	Hidden Markov Model
IFN	Interferon
KPNA	Karyopherin Alpha
L	Large protein, the viral RNA-dependent RNA polymerase
MARV	Marburg Virus
MD	Molecular Dynamics
mRNA	messenger RNA
MSA	Multiple sequence Analysis
NP	Nucleoprotein
nsSNV	non synonymous Single Nucleotide Variants
PCA	Principal Component Analaysis
PDB	Protein Data Bank
PSI-BLAST	Position Specific Iterated BLAST
RESTV	Reston Ebolavirus
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
ROC	Receiver Operating Characteristic

SCOP	Structural Classification of Proteins
SDP	Specificity Determining Positions
SUDV	Sudan Ebolavirus
SVM	Support Vector Machine
TAFV	Tai Forest Ebolavirus
VP24	Viral Protein 24
VP30	Viral Protein 30
VP35	Viral Protein 35
VP40	Viral Protein 40
WHO	World Health Organization

List of Figures

Figures in Chapter 1

Figures in Chapter 2

Figures in Chapter 3

Figures in Chapter 4

Figures in Chapter 5

Figures in Chapter 6

List of Tables

Tables in Chapter3

Tables in Chapter 4

Tables in Chapter 5

Tables in Chapter 6

Chapter 1:
Introduction

This thesis encompasses two main research lines, first the development of a computational algorithm (varMod) to predict the effects of nonsynonymsous single nucleotide variants (nsSNVs) and secondly an analysis of genetic variation in Ebolaviruses to understand how they affect human pathogenicity. The thesis is presented as a series of papers, one focusing on predicting the effects of genetic variation, while three consider genetic variation within Ebolaviruses.

1.1 Genetic variation

Each individual is unique as a result of genetic variation. Therefore understanding genetic variation and how it alters phenotype will advance our knowledge of the extent of genetic variation between individuals. This has been greatly increased in recent years as a consequence of the advances in genome sequencing. While it took multiple teams a decade to sequence the human genome (Hattori, 2005; Abecasis et al., 2010), there are now many projects that sequence large populations of humans, for example the 1000 genomes project in much shorter times (Auton et al., 2015; Sudmant et al., 2015; Abecasis et al., 2012).

1.1.1 Types of genetic variation

There are multiple types of genetic variation:

- Single nucleotide variants (SNVs) – a single base differs between an individual and the reference genome
- Copy number variation (CNV) – a region of the genome that has a different number of copies compared to the reference genome
- Insertions and deletions (indels) – bases deleted or inserted into the genome
- Structural Variants (SVs) – changes in larger portions of the genome

sequence that result in a structural change of the genome and thus in a change of chromosome assembly.

These types of variation can affect both coding and non-coding regions of the genome. However, given our limited understanding of the role of non-coding regions (The Encode Project consortium, 2004; Birney et al., 2007), it is difficult to interpret the effects of variation located in non-coding regions of the genome, unless they are located in known regulatory regions.

SNVs are classified into synonymous, when the base change does not cause a change in the coded amino acid, non-synonymous where the encoded amino acid is changed and nonsense when a stop codon in introduced. SNVs that occur fairly frequently in a population (typically more than 1% of a population) are referred to as single nucleotide polymorphisms (SNPs).

1.1.2 Human Genetic Variation

After the discovery of DNA (Watson and Crick, 1953), in 2003, human genetics has seen probably the most revolutionary discovery, with the first release of an entire reference sequence of the human genome (The Human Genome Project Consortium, 2004). Since then, the increased interest in understanding the biological basis of heredity, has led to the establishment of several international projects, in order to collect and catalogue human genetic variation, and among them the first two were the 1000 Genome Project (Gibbs et al., 2003; The International HapMap Consortium, 2004; Thorisson & Smith, 2005; Frazer et al., 2007; Buchanan et al., 2012; Auton et al., 2015) and the HapMap project. This section describes these catalogues and other current projects.

1.1.2.1 The Human Genome Project

The Human Genome Project (HGP) started in 1990 and was completed in 2003 with the initial draft published in 2000 (Lander et al. 2001). It was an international effort primarily by research groups in the US, UK, Japan, Germany, France and China. The project saw the introduction of shotgun sequencing that rapidly

increased the speed at which sequencing was performed. It also saw a notable conflict between public and private interests, when a private parallel project from Celera Genomics wanted to patent the genomic sequence (Williams-Blongero, 2004).

Along with sequencing the human genome, the project aimed also to develop new technologies, to study and interpret the human genome and also to establish Ethical, Legal and Social Implications of Human Genomics (ELSI). ELSI was the first regulatory body to assess issues in genomic research, for example privacy of the genetic information and other important issues that could affect individuals and society. Sequencing of the human genome revealed that the human genome contains approximately 20,500 genes a similar number to that found in mice. The human genome project took almost 13 years to complete and more than 10 billion dollars to sequence just a single reference genome. This was a milestone in genetics and paved the way for many advances, with scientists now able to sequence a genome for a few thousands dollars and taking less than a day.

1.1.2.2 The HapMap Project

The HapMap project was launched in 2002 and it was completed three years later. It is an international consortium of academic researchers and private companies (International HapMap Consortium 2003; International HapMap Consortium 2007; Gibbs et al., 2003). A haplotype is a combination of alleles within a region of a chromosome. The HapMap project was set up with the idea to create a haplotype map of the human genome, to describe how human genetic variation is shared among individuals in different populations. The main goal of this project is to understand how SNPs and other genetic variants organise in the different chromosomes and how genes can affect drug response by making the generated data available to the scientific community. The project used genotyping techniques and consisted of three main phases: the first, when more than 1 million SNPs where found in 269 DNA samples from different individuals coming from four main populations; the second phase, in 2007, where over 3.1 million of SNPs were genotyped in 270 individuals. In 2010, the same consortium published genotyping results for 1.6 million common SNPs in 1,184 individuals from 11 populations. This

latest analysis was called HapMap3 and represented an integrated data set of common and also rare alleles. The HapMap project was the first to perform a large-scale Genome Wide Association study.

1.1.2.3 The 1000 Genomes Project

The 1000 Genome Project was launched in 2008 and concluded in 2015 (Wood et al., 2013; Abecasis et al., 2012; Abecasis et al., 2015). It is currently the largest public catalogue of human genetic variation with a frequency greater than 1% in the studied populations. The main goal of this project was the identification of human polymorphisms with a minor allele frequency (MAF) greater than 1%. The 1000 Project was performed in multiple stages. The first one, a pilot phase which had the goal of developing and assessing strategies for sequencing a large number of individuals in the most informative way. It used three levels of sequencing. For two sets of trios (parents and child) high coverage genome sequencing was performed (average 42x). For 179 individuals low coverage (2-4 X) whole genome sequencing (WGS) was performed and finally target exon capture (906 randomly selected genes) was performed on a larger set of 697 individuals from four populations. This initial phase of the project identified nearly 15 million SNPs, 1 million indels and 20,000 structural variants. They demonstrated that this dataset had identified the vast majority of common variants and that each individual had between 250-300 loss of function SNPs and between 50-100 variants associated with inherited disease (Abecasis et al., 2010).

In the second phase, completed in 2012 (Altshuler et al, 2012) a total of 1,092 genomes were sequenced from across 14 different populations. The techniques used in this phase were a combination of low-coverage (2-6 X), whole genome and whole exome sequencing (WES) (with coverage up to 100 X) and dense SNP genotyping. This phase discovered over 38 million SNPs, with 1.4 million short insertions and deletions (indels) and more than 14,000 larger deletions. This phase removed over 1.7 million low quality SNPs from the first phase.

The third phase was completed in 2015 (Sudman et al., 2015; Abecasis et al., 2015)

and considered both Structural variants (SVs) and single nucleotide changes. The study revealed 68,818 structural variants (SVs) in 2,504 unrelated individuals coming from 26 populations. It found 8 classes of structural variants, enriched on haplotypes identified in GWAS studies; these variants were largely shown to be specific to individual continental groups (Sudman et al., 2015).

The final outcome of the project was the identification of 88 million variants, of which 84.7 millions were SNPs, 3.6 millions were short insertions and short deletions and over 60,000 were structural variants. Of this total 762,000 variants were rare (i.e. present in very few individuals). The main and conclusive finding of this third phase was the extent of genetic variants that were shared among individuals from different populations.

Now that the 1000 Genome Project is complete, it is under the administration of the International Genome Sample Resource (IGSR) which is an entity formed within the EMBL-EBI institution with the aim of maintaining and ensuring usability of the 1000 Genome Project data, to expand it by adding new genomic data and even by including new population data.

1.1.2.4 Rare variation

Rare variants occur in a small proportion of the population (MAF < 1%) but interestingly individuals have many of them (Nelson et al, 2012; Tennessen et al., 2012). The identification of rare variants requires deep sequencing to enable these variants to be called with confidence and not classed as sequencing errors.

The 1000 Genome Project (Phase II) classed rare variants as those with a MAF < 0.1% and they found individuals did not have many, estimated at around 200. There are a few available catalogues of rare genetic variants, such as the Exome Sequencing Project (ESP) (Exome Variant Server) and others coming from independent studies (Tennessen et al., 2012; Nelson et al., 2012; Keinan and Clark, 2013).

Nelson and collaborators sequenced 202 drug target coding genes in 14,002

individuals, through a Whole Exome Sequencing study. They identified a large number (1 every 17 bases) of novel variants that were population specific, geographically clustered and most interestingly they were more likely to be functional. In fact, more than 95% of the variants discovered were rare (MAF <0.1) and more than 74% were private variants (present only in a single individual) (MAF <0.01). The study considered how these rare variants could help our understanding of disease risk. The samples from 14,002 individuals included 10,621 samples from 12 case control studies of common disease. The drug targets genes were selected according to a GlaxosmithKline set considered for drug repositioning candidates. Genes used for the study were reduced to 202 in order to make the analysis feasible. The genes included 12 genes coding for marketed drug targets, 44 genes encoding Phase I to III terminated drug targets, 76 genes encoding genes under clinical development targets and 70 genes encoding targets under (or interesting for) pre-clinical development. The set of genes was compared to the NHGRI, catalogue of already published Genome Wide Association Studies (http://www.ebi.ac.uk/gwas/), with HGMD catalogue, where they found an overlap of fifty three genes and with the OMIM database (Hamosh et al., 2005; McKusick, 2007), where they found a notable overlap, for a total of 46 variants in 25 genes. Furthermore they compared their set of genes with the rest of protein coding genome defined by GENECODE, were they found an overlap of almost 20,503 and importantly they found Gene Ontology characteristics in terms of biological process, cellular components and molecular function for 20,340 genes. This study has clearly opened a new window for the interpretation of rare variants, by discovering that 95% of variants that were rare, more than 74% were private variants and more than 90% were novel. The aggregation studies additionally showed that around the 37% of rare alleles were predicted to be deleterious. Their findings contrast with the initial results on rare alleles found by the 1000 Genomes, as they had predicted individuals would only have around 200 rare variants.

Another project performed deep exome sequencing for 15,585 protein coding genes, in 2,440 individuals in two populations (Tennessen et al., 2013). Like the Nelson study, they discovered more than 500,000 single nucleotide variants, over 86% of

these were rare, and 82% were rare and population specific. In order to prove that the variants were functional they used four different methods for non-synonymous variants (Polyphen2, SIFT, MutationTaster and a likelihood ratio test and additionally they used three conservation based methods, GERP (Genomic Evolutionary Rate Profiling, Cooper et al., 2005), Phylop (Cooper et al., 2005) and another tool designed by the authors and called SFS-Del). This study showed that rare variants and their high frequency can be explained by the explosive population growth in Europe and Africa. Furthermore they mapped over 31,000 non-synonymous variants onto structure, whether the protein structure was available, they classified the variants according to structural categories (i.e. if the variant was buried, part of a ligand binding site or active site or involved in hydrogen bonding, or potential charge or if forming a cavity or if in a over packing region). They observed that rare variants were particularly enriched in ligand binding and active sites and involved in hydrogen bonding.

1.1.2.5 Current projects

Current projects are sequencing a larger number of individuals and with a focus on obtaining data and performing analysis that is relevant to disease and clinical treatment, to drive precision (or personalised) medicine. The 100,000 Genome Project aims to sequence the genome of 100,000 patients with a rare inherited disease or cancer, across 70,000 individuals and is being run by Genomics England and associated organisations (Cranage, 2015; http://www.genomicsengland.co.uk/). The Genomics England Project will compare individual's Genome data with health clinical data and medical records, including family information (for rare diseases, more than one individual in a family is being sequenced, e.g. a child with the disease and both of their parents), in order to find a better treatment for individuals and contribute precision medicine.

The Personal Genome Project (PGP) (Church, 2005; http://www.personalgenomes.org/) has a similar aim and was founded in 2005, by Professor George M. Church of Harvard University. The goal of the PGP is to sequence the complete genomes of 100,000 individuals along with phenotypic data,

making the results available to the community in order to aid to the development of personal genomics and to enable personalised or precision medicine. The project is still ongoing and an increasing number of volunteers are taking part in the project.

1.1.2.6 Databases of genetic variation

The myriad genetic variants discovered with sequencing projects are available in a range of databases. dbSNP (Sherry et al., 2001) was founded by the National Center for Biotechnology Information (NCBI) and it collects variants across 53 different organisms, and the last release (146, March 2016) just for Humans contained over 150 million referenced SNP (RefSNPs) and 538 million submitted SNPs (subSNP).

Humsavar (http://www.uniprot.org/docs/humsavar) is a catalogue of Human Polymorphysms and disease mutations. It is developed by UniProt, the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR). This database counts 27,861 disease variants, 38,352 polymorphisms and 7,549 unclassified variants, for a total of 73,762 variants. The small number of variants is due to them being present in protein regions and also being a focus on the variants being classified into categories indicating if they have a role in disease.

Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/) (Harrison et al., 2016) is a database of medically relevant variants, so it collects variants that are phenotypically significant. It is defined as a database of "the relationship between human variations and the phenotype" and it is based on phenotypic information from MedGen (Halavi et al., 2013).

VariBench (Sasidharan & Vihinen 2013) is a database of genetic variants that was developed primarily for benchmarking of methods that predict if SNVs are deleterious. It contains disease causing missense variants, neutral high frequency SNPs, protein stability affecting missense variants, variants affecting transcription factor binding sites and variants affecting splice sites.

1.2 The Genotype to Phenotype Relationship

Knowledge of human genetic variation enables investigation of the genotype to phenotype relationship to understand how genetic variants are associated with particular traits, particularly those associated with disease.

Diseases are often classified into monogenic, when a variant in a single gene is responsible for the trait and complex disease, such as coronary disease, where many variants contribute to the trait (Manolio et al., 2009; Eichler et al., 2010; Lehner, 2013). Monogenic and complex diseases are complicated by environmental factors. The OMIM (Online Mendelian Inheritance in Man) (Hamosh et al., 2005; McKusick, 2007) catalogue is a resource that collects genetic variants that are associated with phenotypes. The last release contains nearly 24,000 entries and it is vastly used to interpret and associate variants with disease.

Despite extensive research carried out to date, there is still a large gap in the interpretation of the myriad of the collected variation data, with much of the heritability remaining unexplained (Eichler et al., 2010). In fact, it still challenging to predict the predisposition to a certain disease or how many complex diseases, such as cancer or cardiovascular disease that are caused by many factors, or Mendelian disorders which are caused by abnormal alterations in a single gene, can be related to heritability (Zuk et al., 2014; Liu & Leal, 2012; Lippert et al., 2013).

All these considered factors mean a need for new insights to personalised or precision medicine, which represents the efforts to combine genetic information of individuals and use them identify the predisposition to a disease and to design a "individuals-size" medical treatment. Precision medicine is described later, in section 1.2.3.

1.2.1 Genome Wide Association Studies

In order to understand how genetics relates to a trait and therefore assess the heritability for that trait, genetic association studies have been developed. One of the most popular means for this purpose has been the development of Genome Wide Association Studies (GWAS) (Daly, 2012). GWAS are a combination of statistical

tests and genotyping techniques whose aim is to determine the effect of SNPs on a trait. This approach has been widely used for the International HapMap project. Genotyping techniques can detect simpler and less informative relationships in comparison to genome sequencing techniques. However, GWAS are able to perform a large number of association tests, and thanks also to the use of SNP chips they can associate SNPs to disease. The use of SNP chips is a limitation of many GWAS as the study is limited to the number of SNPs tested on the chip and will not detect other novel SNVs. As a result many GWAS have considered common variants but with very low proportion of individuals in a population that carry the allele that is associated with the phenotype; this last concept is defined as penetrance and its relationship with allele frequency in population is shown in *figure 1.1*:

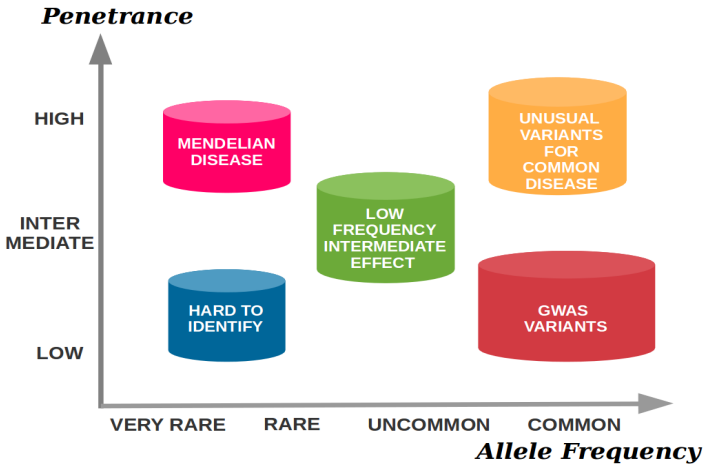


Figure 1.1: Penetrance of Variants over their allele frequency is shown in this figure. Their effects on disease is shown in the graph meaning the missing heritability. The figure has been reproduced from Manolio et al., 2009.

1.2.2 Use of Next Generation Sequencing (NGS)

The advances achieved in sequencing techniques, such as Next Generation Sequencing (NGS) also referred to as high throughput sequencing can aid the discovery of new rare (MAF <0.1) and even *de-novo* (or private, MAF <0.01) variants (DePristo, et al. 2011).

NGS represent improvements in the speed of sequencing but also in the costs and in the accuracy which is notably increased from the previous generation sequencing.

Importantly, NGS allowed the discovery of rare variants in many samples. GWAS studies as well as NGS have contributed to the ENCODE project (Birney et al., 2007; Sloan et al., 2015), to annotate and experimentally validate gene loci in the Human Genome.

1.2.3 Personalised/precision medicine

Each individual has a unique set of variants in their genome that will determine traits, including the risk for disease and response to drugs. Personalised medicine can be used in two ways: firstly, in a preventative manner, for example knowledge of an individual's risk for particular disease could alter their behaviour or to even seek treatment. A good example of this is the identification of BRCA1 and BRCA2 mutations, where women may choose preventative measures as they have a high risk of developing breast cancer (Brookes et al., 2015; Zeidan et al., 2015).

Secondly personalised medicine can be used when an individual is ill and their genomic information used to identify the most suitable treatment. For example, if multiple possible treatments are available is there one that the patient will have a better response to? (Ng, et al., 2009). An example is the use of targeted molecules to to treat myeloid leukemia, by overcoming AML (Acute Myeloid Leukemia) cell resistance to drug therapy (Gojo and Karp, 2014).

More recently personalised medicine has been referred to as precision medicine (Peterson et al., 2013; Katsnelson 2013), meaning a more precise and effective approach to identify a specific patient strategy to identify the best therapy based on the patient's genetics, environmental and lifestyle factors.

A branch of precision medicine is Pharmacogenomics, which is a combination of Pharmacology and genomics and whose main goal is to understand how genes affect individual's response to a certain drug (Karczewski et al., 2012; Altman et al. 2012; Hopkins & Groom, 2002).

1.3 How genetic variation leads to altered phenotype

All types of genetic variation (SNVs, CNVs, SVs and indels) may be associated with a trait. The research in this thesis largely considers non-synonymous single nucleotide variants (nsSNVs) in protein coding regions and therefore this section focuses on such variation. Until recently, synonymous variants were thought to be non-functional as they do not alter the protein amino acid sequence. However recent research has observed positive selection of synonymous variants in cancer genomes and proposed that synonymous variants can be functional (Supek et al., 2014). Hence, it is possible that such variants may alter regulatory regions or alter the speed of mRNA translation and affect protein folding (Shabalina et al., 2013). However, our understanding of the effects of synonymous variation are not well defined and therefore focus is placed on non-synonymous SNVs.

1.3.1 Analysis of nsSNVs associated with disease

A number of studies have analysed the properties of nsSNVs that are associated with disease. Such research typically considers the location of nsSNVs in protein sequence or structure and compares the prevalence of disease associated and neutral variants in different regions of the protein.

It is a widely accepted theory that disease-causing sites are much more conserved than neutral ones (Kumar et al., 2001). Thus, the fact that functional sites are evolutionary conserved, has made sequence conservation one of the most important factors used by bioinformatics tools to pinpoint these functional residues in protein sequences and aid methods. The use of orthologues (orthologues are two or more sequences which descend from the same ancestors and they are separated by speciation events) in multiple sequence alignments, to calculate conservation has been used in methods such as SIFT (Kumar et al., 2013) and it has been shown to give a better perfomance.

Initial studies of the location of SNVs in protein structure, showed that disease causing variants are enriched in the protein core, where they are most likely to affect protein stability and possibly protein function (Burke et al., 2007; Yue & Moult, 2005).

David et al., (2012) extended these previous structural analyses to consider the role of protein-protein interfaces. Using the humsavar database of variants (from UniProt, Pundir et al., 2016) they mapped variants onto protein complexes from Interactome3D (Mosca et al., 2012). In agreement with previous studies, they observed a preference for disease-associated variants to be located in the protein core. Additionally, they observed an enrichment of disease-associated nsSNVs in protein-protein interfaces, confirming the importance of protein-protein interactions in cellular function.

Similarly Bordner and Zorman (Bordner & Zorman, 2013) considered nsSNVs present in ligand-binding sites. The authors performed large scale homology modelling of the human proteome to investigate disease-associated nsSNVs. They analysed variants from the Human Gene Mutation Database (HGMD) (http://www.hgmd.cf.ac.uk/ac/index.php), COSMIC (http://cancer.sanger.ac.uk/cosmic), UniProt and dbSNP (Sherry et al., 2001). They performed a structure-based approach to infer the effects of variants on binding sites. In their pipeline they used homology modelling to predict binding sites and Machine learning approaches to classify variants.
The authors found that disease-associated missense mutations were enriched in binding sites compared to neutral variants.

Protein function is not only influenced by protein-protein or protein-ligand interactions but it is also dictated by other processes, including post translational modifications (PTMs). Nussinov et al. (2012) proposed "Allosteric PTM codes" and described the influence of PTMs on protein function through two main mechanisms: by orthosterically influencing binding (for example they can disrupt protein-protein interactions) and by allosterical conformational changes in the functional site. More recently Li et al., (2014) showed that disease associated mutations affect PTM sites and thus protein function.

1.4 SNV prediction methods

The trends (described above) that show nsSNVs that are associated with disease are

enriched with particular properties enabled the development of methods that can predict if a nsSNV will affect protein structure and/or function and be deleterious. This section provides a summary of those methods that are most widely used.

1.4.1 Sorting Intolerant from Tolerant (SIFT)

Sorting Intolerant from Tolerant (SIFT; Kumar et al., 2013) classified amino acid substitutions for SNPs or indels; this method is based on the principle that mutations occurring in conserved regions are less likely to be tolerated and consequently more likely to be functional. SIFT generates a multiple sequence alignment including distantly related orthologues. Its fundamentals consist in building a theoretical model based on sequence homology that considers features such as conservation, hydrophobic conservation, difference from known mutations in a multiple sequence alignment and that is able to predict if the substitution is tolerated or not by using a score derived from position-specific scoring matrices with Dirichlet distributions. The obtained SIFT score is a probability that the mutation is functional and it ranges from 0 to 1. The closer the value is to 0 the more likely the mutation is functional.

1.4.2 PolyPhen2

PolyPhen2, Polymorphism Phenotyping V2 (Adzhubei et al., 2013), also predicts if genetic variants are deleterious. In contrast to SIFT, PolyPhen2 uses information from both orthologues and paralogues (paralogues are two or more sequences which are separated only by gene duplication), protein structural features and machine learning. The sequence and structural features comprise: sequence annotations from Uniprot and from DSSP, bond annotations (disulphide bonds and covalent links in proteins), UniprotKB and Swiss-Prot functional site annotations (binding site information, enzyme active sites, metal binding sites, lipidated residues, glycosylated residues, non-standard amino acids and other modification sites), UniprotKB and Swiss-Prot region annotations (membrane crossing regions, membrane-contained regions with no crossing, repetitive sequence motif or domains, coiled coil regions, endoplasmic reticulum targeting sequences and sequences cleaved during maturation), PHAT score (only for positions

annotated as transmembrane) and multiple features relating to secondary structure from DSSP , Ramachandran maps, normalised B-factors, ligand contacts, inter-chain contacts and functional site contacts. The method uses all these features to classify the substitution, according to a Naive Bayes probabilistic classifier, through a supervised learning machine approach. PolyPhen2 is trained with two datasets, HumVar, which is most useful when considering Mendelian disease and HumDiv, which is best used for complex traits. PolyPhen2 can also classify variants as causing: loss of function, gain of function, drug resistance and switch of function mutations.

1.4.3 Other SNV prediction tools

PolyPhen2 and SIFT represent the most widely used methods for predicting if SNVs are deleterious. Other methods are described briefly below.

MutationAssessor (Reva et al., 2011) bases the prediction of the effect of variants on conservation and specificity (i.e. differential conservation between subfamilies). It was validated on a set of 60,041 variants, 78% of which predicted to be disease-associated. The method is based on three hypotheses: mutations that are evolutionary conserved are more likely to be functional; those that are not are more likely to be neutral; evolutionary conservation patterns can discriminate between functional and non functional mutations. According to this, the final functional score in derived from the conservation score and from the specificity score as well.

Yates and collaborators developed Suspect (Yates et al., 2014), which uses both sequence and structural features. The unique feature of SuSpect is the use of interaction network centrality as a feature, which was demonstrated to improve predictions. In benchmarking SuSpect obtained better performance than other existing methods.

CONDEL (CONsensus DELeteriousness score of missense SNVs) (Gonzalez-Perez and Lopez-Bigas, 2011) is another popular method for SNV effect prediction. Condel uses a combination of scores from SIFT, Polyphen2, MutationAssessor, FATHMM (Functional analysis through Hidden Markov Models,

http://fathmm.biocompute.org.uk/) and Ensembl-variation. The method performed better than other existing method during the benchmarking phase. It is now part of the FannsDB (Functional annotations for non Synonymous SNVs Database), a database of functional annotation for non-synonymous variants that integrates data from Ensembl (www.ensembl.org) and dbNSFP 2.1(Liu et al., 2011; Liu et al., 2013).

However, one of the main problems observed with these methods is that they individually perform well in benchmarking but they often show little agreement between methods (Chun and Fay, 2009). This makes it important to continue to develop new methods that try to improve upon existing approaches. During the course of my PhD I have developed VarMod a method for predicting the functional effects of nsSNVs (Pappalardo & Wass, 2014), which is described in *Chapter 2*.

1.5 Ebolaviruses

Viruses are non cellular entities which use the host cell machinery to replicate and cause infectious disease. Ebolaviruses (*figure 1.2*) are negative single stranded RNA viruses (RNA genome is complementary to the viral mRNA). The Ebolavirus genus belongs to the Filoviridae family and Mononegavirales order. Ebolaviruses are divided into four human pathogenic species, (Ebola– formerly called Zaire, Taï Forest, Sudan and Bundibugyo) and one non-human pathogenic species (Reston). The species are named after where they were discovered. The first two Ebola virus species (Sudan viruses and Ebola viruses) were originally discovered in 1976 (Pattyn et al., 1977; International Commission Report, 1976; Report of a WHO/International Study Team, 1978) and until 2014 there had been a limited number of small outbreaks.

To date Reston viruses have only demonstrated pathogenicity in non-human primates and were first identified in Reston (Virginia, USA in 1989-1990), then in Siena (Italy, in 1992-1993) and most recently in Texas (1996). In 2008 Reston virus was found in domestic pigs in the Philippines. Reston antibodies have been reported in a few human individuals, but none of them developed Ebola Hemorrhagic Fever

or Ebola Virus Disease (EVD), thus demonstrating the lack of pathogenicity in humans.

In this section the Ebolavirus cycle of infection, its genome and details of the current outbreak in West Africa are introduced.

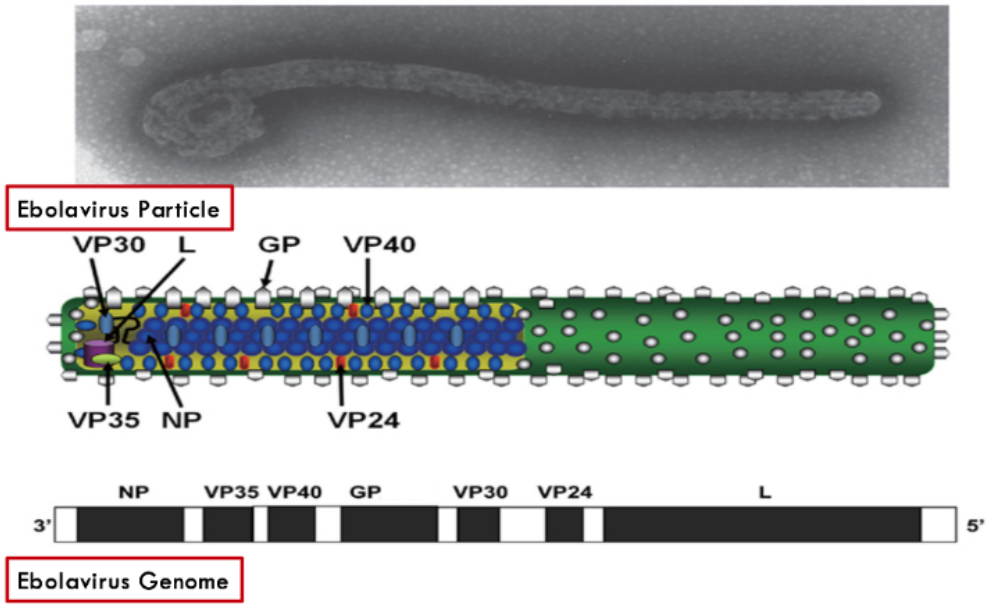


Figure 1.2: The Ebolavirus particle and the Ebolavirus genome. The figure has been adapted from Takada et al., Front. Microbiol. 2012.

1.5.1 The Cycle of Ebolavirus Infection

The Ebolavirus infection cycle contains the following steps:

1. First the virus particle detects the surface of the host cell though the protein GP binding to a host cell receptor
2. it then penetrates the cell through a mechanism of Macropinocitosis
3. once in the cytosol, it fuses to the endosomal membrane of the vescicle in which it is contained and the ribonucleocapsid is releases into the cytosol, where it will start to be processed by host cell enzymes
4. the negative RNA uses the complementary strand to form mRNA, which is translated using the host cell machinery.

5. New replicated viruses follow an actin-dependent transport and they are released in the form of new virions through a mechanism of budding.

Ebolavirus is responsible for EVD which is a deadly disease. During the last outbreak (2014) the WHO registered 11,325 confirmed deaths with the main locus in Guinea, Sierra Leone and Liberia. Other minor cases have been registered in Nigeria, Mali and Senegal. Four cases of Ebola infections have also been imported in United States and two in Europe, one in United Kingdom and another in Spain. In total 28,657 cases of infections have been confirmed, as of 8th May 2016.

1.5.2 The Ebolavirus genome and protein function

The Ebolavirus genome is around 19K nucleotide bases long and contains seven genes, which encode nine different proteins (*figure 1.3*). The proteins are: the nucleoprotein (NP), RNA dependent RNA polymerase (L), glycoprotein (GP), soluble GP (sGP), small soluble GP (ssGP) and four structural proteins that are called viral protein 24, 30, 35 and 40 (VP24, VP30, VP35 and VP40). The gene GP encodes GP, sGP and ssGP. These multiple forms of GP are generated as a result of RNA editing (Mehedi et al., 2013). Given the small number of proteins in the Ebolavirus genome, the proteins need to be multifunctional (Xu et al., 2014).

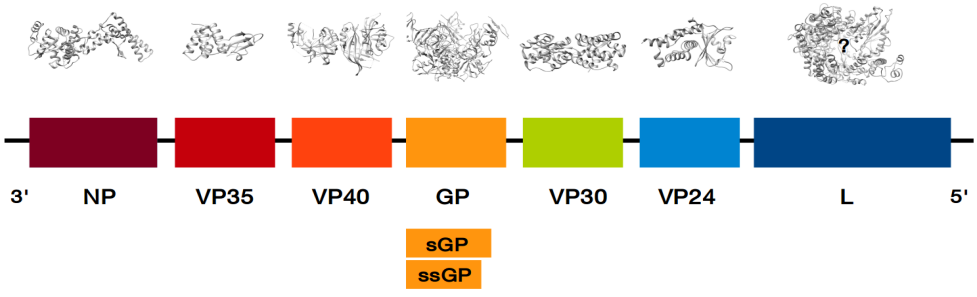


Figure 1.3: The Ebolavirus Genome. The 3' terminal and the 5' terminal are shown. Over each gene the correspondent protein with deposited PDB structure is shown in grey cartoon. For L protein there is no known structure but there are models available.

The Glycoprotein GP is the main protein responsible for viral entry into the host cell GP contains a mucin domain which has a highly glycosylated glycan caps (it is heavily glycosylated) which is important for the viral entry and probably also for immune system escape. The GP1 subunit binds to the host cell receptor(s), the actual receptor(s) remain unknown, although the Niemann-Pick C1 (NPC1) receptor is known to be required for virus entry (Miller et al., 2012). Subunit GP2 is involved in the fusion of the virus with the host cell membrane. The function of sGP and ssGP remains unclear.
The function of the protein L is as an RNA-dependent RNA polymerase. It forms a complex with NP, VP30 and VP35 to form the Ebolavirus RNA-dependent RNA polymerase nucleocapsid complex, essential for the generation of viral mRNA.

VP35 is a multifunctional enzyme. As described above it is part of the RNA-dependent RNA polymerase complex and it also has a role in preventing interferon signalling. This function is performed by VP35 dimers binding double stranded viral RNA and preventing them being recognised by the host cell immune system. This has made protein VP35 attractive as a therapeutic target and many scientists tried to study and develop VP35 inhibitors, but still without any positive outcome (Binning at al, 2014).

The matrix protein VP40 exists in multiple different oligomeric forms, with each having a different function. The VP40 dimer has a role in membrane trafficking. The hexamer is functional in virus assembly and budding and the VP40 octamer has function in transcriptional regulation.

The minor matrix protein VP24 which has probably on of the most intriguing role in the suppression of the immune response, since it blocks the whole Interferon Signalling Pathway by blocking the Janus/Kinase and Signal transductors and activators of transcription, the Jack/STAT pathway.
Interferons Alpha and Beta, together with Natural Killer cells (NK), are the first agents that the human Immune system produces as innate response when a virus attacks the human cells. Interferons bind to their receptors and activate the JACK-

STAT pathway and therefore activate the transcription of genes able to block the viral replication in the infected host cells. Ebolavirus is able to escape the human immune system in several ways and it is able to escape not only the innate but also the adaptive response, for example the production of antibodies. It has been recently observed that Ebolavirus is able to block the production of Interferons by mean of its protein VP24. This last, in fact, competes with the phosphorylated transcription factor STAT1 for the binding with Karyopherins, which belong to the Importin complex. Proteins that are translocated into the nucleus, generally contain a sequence that is called classical nuclear localisation signal (cNLS) and that is recognized by Karyopherins. STAT1 is classified as non classical NLS (ncNLS) and it is recognized by Karyopherins by a mechanism of dimerisation and phosphorylation. When VP24 competes with STAT1, it binds Karyopherin and the transcription factor cannot be translocated into the nucleus and the whole Interferon Signalling Pathway is blocked, since the Jack/Stat pathway is inactivated. This process is shown in *figure 1.4*.

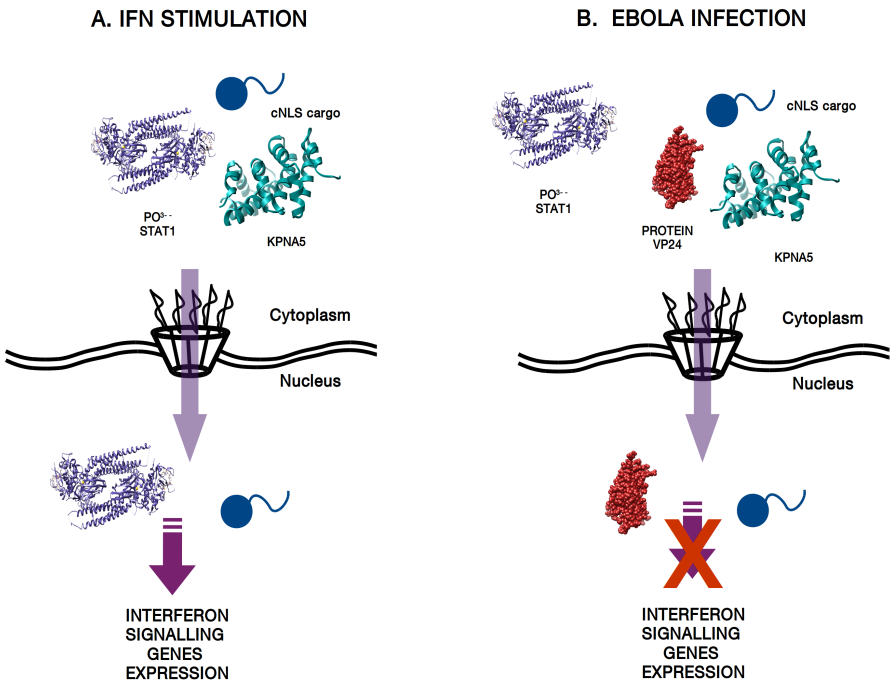


Figure 1.4: The mechanism of Inhibition of the Signalling pathway in normal cells (A) and in presence of Ebolviruses (B). Ebolavirus protein VP24 is shown in red spheres it prevents the binding of STAT1 (blue cartoon) to KPNA5 (cyan cartoon). In this way the Interferon Signalling gene Expression path is blocked. The figure has been adapted from Daugherty & Malik, Cell Host & Microbe, 2014.

1.5.3 The current Ebola virus outbreak

The current Ebola virus outbreak in West Africa has demonstrated that members of the Ebolavirus family pose a significant threat to human health on a large scale (Quaglio et al., 2016). It was of unprecedented size resulting in 28,639 confirmed cases and 11,316 deaths as of 28th February 2016 (www.who.int). Previous Ebola virus outbreaks were small ranging from a few to a few hundred infected individuals Until 2014 the outbreak in Uganda in 2000 was the largest, affecting 425 individuals and resulting in 224 deaths (La Vega et al., 2015). Given the limited size of previous outbreaks it was largely thought that Ebola outbreaks would remain small as they occurred in small villages in Africa with very limited travel connections and therefore effectively contained themselves. The current outbreak started in Guinea in December 2013 and with regular flare-ups it has still not been declared over (www.who.int). This outbreak has provided evidence of Ebola viruses persisting in immune-privileged sites and remaining infective for long periods. This includes persisting in the eye (Varkey et al., 2015) and the presence of Ebola virus in semen a year after recover from the disease and possible sexual transmission (Christie et al., 2015; Deen et al., 2015; Mate et al., 2015). This complicates effective outbreak control. The risk of new transmission from these persistent infections is not currently known; however, taken together, these findings caused concerns about future large outbreaks (Quaglio et al., 2016).

Next generation sequencing has provided extensive sequencing data on Ebola virus genetics and evolution during the current outbreak (Gire et al., 2014; Loriere et al., 2014; Tong et al., 2015; Carroll et al., 2015; Hoenen et al., 2015; Quick et al., 2016). These studies have enabled the identification of mutations in the virus and with them tracking of the outbreak into lineages.

The first study by Gire et al., (Gire et al., 2014) sequenced 99 Ebola virus genomes from Sierra Leone. Their work suggested a high evolutionary rate of $1.9x10^{-3}$ substitutions per site per year, approximately two fold more than the rate between outbreaks. Later studies indicated lower rates closer to 1.0x 10-3 substitutions per site per year, in agreement with previous rates observed between outbreaks (Loriere et al., 2015; Tong et al., 2015; Carroll et al., 2015; Hoenen et al., 2015). It has been suggested that a short sampling time used to obtain the 99 genomes did not allow deleterious mutations to be selected against and as such inflated the evolutionary rate (Gire et al., 2014; Carroll et al., 2015). The analysis of Gire et al., supported the outbreak being caused a single transmission from an Ebola virus reservoir followed by human-to-human transmission.

Hoenen et al., (2015) sequenced Ebola viruses present in infected individuals in Mali. They identified a limited number of nonsynonymous amino acid changes and those observed did not map to functional regions of Ebola virus proteins. They propose that during the outbreak the virus has been undergoing limited evolution with no evidence of increased virulence or transmissibility (Hoenen et al., 2015). Phylogenetic analysis of a larger set of Ebola viruses from Sierra Leone identified three different lineages, and multiple sub-lineages (Tong et al., 2015). Carroll et al., (2015) sequenced 179 Ebola virus patient samples from Guinea, phylogenetic analysis identified two lineages (A and B). Lineage A was present earlier in the outbreak (not observed after July 2014) and thought to have been contained by response to the outbreak. However, lineage B shows spread across Guinea, Sierra Leone and Liberia.

Loriere et al., (2015) identified three lineages present in 85 patients infected in Guinea. The rate of substitutions is similar to the other studies but they observed nonsynonymous substitutions in the GP, L and VP35, proteins, some of which may be functional. Some GP variants are present in the mucin like domain and Loriere et al., (2015) proposed that they could alter the shape of the virus or affect glycoslyation of GP (Loriere et al., 2015). In VP35, mutations were identified in the domain associated with interferon inhibition but, the functional affect, if any, of this variant remains unclear.

An alternative approach considered 65 genomes from a range of outbreaks and infections in both great apes and humans (Azarian et al., 2015), with a focus on GP as it is the most variable Ebolavirus protein. Their findings suggest that the evolution observed is primarily due to neutral genetic drift and based on this they propose that it is unlikely that strained with altered transmission mechanisms or with altered pathogenicity will emerge.

The most recent sequencing project from the West Africa outbreak performed 'real-time' sequencing in the field (results available within 24 hours) (Quick et al., 2016) by using MinION nanopore sequencers. Using this approach 142 Ebola virus genomes from Guinea were sequenced during 2015. They identified that the viruses largely belonged to two main lineages GN1 and SL3. SL3 originated in Sierra Leone and spreaded to Guinea, whereas GN1 was confined to Guinea.

Combined together these studies suggest that Ebola viruses are not evolving towards easier transmission or changes in virulence. Importantly, the many sequences now available enable extensive computational analysis of Ebola to understand how it functions and what determines pathogenicity.

1.6 Bioinformatics methods and resources used in this thesis

In order to carry out the research described within this thesis, several Bioinformatics tools for variant modelling and for protein engineering have been used and this section describes the majority of them:

1.6.1 3DLigandSite

3DLigandSite (Wass et al., 2010) uses protein structural modelling to predict protein ligand binding sites. For a given query sequence 3DLigandSite models the protein structure using Phyre2 (Kelley et al., 2015) and uses the model to perform a structural search of a database of ligand-bound protein structures from the protein databank. Alignment of the model with similar structures from this database map the ligands onto the model structure. Clustering of the ligands is performed and

binding sites predicted based on these clusters. The method has performed well in the Critical Assessment of techniques for Protein Structure (e.g. CASP8, Tress et al., 2009).

1.6.2 Phyre2

There is a large gap between the number of protein sequences present in UniProt and the number of solved protein structures. Phyre2 (Protein Fold Homology/Analogy Recognition Engine) (Kelley & Sternber 2009, Kelley et al., 2015) build 3D structures of protein with no known structure by identifying templates for the query by using hhsearch (Soding et al., 2005) to search a fold library extracted from. The method predicts the secondary structure using Psi-pred (Buchan et al., 2013) and Diso-pred (Ward et al., 2004) (this last for disordered regions in proteins) and then it constructs HMM (Hidden Markov Model) models of the protein sequence. The 3d structure is build, the loops are refined by mean of loop libraries, accounting for loops up to 15 amino acids in length and the side chains are modelled too, with more than 80% accuracy.

1.6.3 Interactome3d

Interactome3D (Mosca et al., 2013) is a bioinformatics tool for the structural annotation of Protein-Protein Interactions. Interactome3D identifies complexes in the PDB that can be used as templates for known pairs of interacting proteins present in databases such as IntAct (Orchard et al, 2014). The templates either represent the full protein structures or they can just represent the interaction of individual domains within a protein sequence, using 3did (Mosca et al., 2013). The method initially collected over 12,000 protein-protein interactions, including experimentally validated and newly discovered interactions, in eight organisms. The last release in 2015 doubled the size of the resource, including data for a further eight organisms.

1.6.4 FoldX

Protein folding is tightly connected to protein function. FoldX (Schymkowitz et al, 2005) is a force field for energy calculations and protein design. FoldX can predict

the effect of mutations on protein stability and it can calculate the energy of interaction in protein-protein and in protein-DNA complexes. The energy that is calculated by FoldX takes into account empirical values coming from experimental data.

1.6.5 mCSM

mCSM is a structure based method for predicting the effect of mutations in proteins by using graph-based signatures (Pires et al., 2014). The methods considers how mutations may affect protein stability, protein-protein affinity and protein-DNA. The method uses a machine learning approach and the novelty of the method is the introduction of a graph-based signature that represents each mutation as a signature of a pharmacophoric count vector that will be considered to train the classification. The method uses a machine learning approach to predict the impact of the mutations. Like FoldX the method is a structure based predictor and they both are accurate, although mCSM showed a better performance than FoldX.

1.6.6 Specificity Determining Positions (SDPs)

The proteins in a protein family may have many different functions. For example in an enzyme family this may be different substrate specificities, with the enzyme performing effectively the same reaction but on different substrates. In the 1990's methods were initially developed to identify such positions (Casari et al., 1995; Lichtarge et al. 1996) that could be present within a protein family. Such positions are now largely referred to as Specificity Determining Positions (SDPs) and they have been demonstrated to be enriched at functional sites such as ligand-binding and protein-protein interfaces (Rausell et al., 2010).

In the research presented in *Chapter 3* the s3det algorithm (Rausell et al., 2010) was used to predict SDPs. This method splits protein family into subfamilies and relates SDPs to functional regions, according to the structural proximity to catalytic sites, ligand-binding sites of small molecules and protein-protein interaction sites.

s3det is based on a statistical method termed Multiple Correspondence Analysis

(MCA) which is very similar to Principal Component Analysis (described later). The program encodes a multiple sequence alignment (MSA) into a binary matrix, and the coordinates of the matrix are transformed into "Principal Axes" that are not correlated; the sequences are then projected onto these Principal Axes. The methods can be used in supervised manner, where the proteins are split into subfamilies determined by the user. Alternatively in the unsupervised format s3det can uses K-mean clustering to group the sequences into subfamilies. The MCA analysis is based on the "pseudovaricentric relationship" between the projected sequences and the projected residues which infers that "the centre of the masses of any group of sequences points to those residues particularly associated to them". This is the principle by which the authors determined the SDPs in their study. *Figure 1.5* shows an example of SDPs that are conserved within all the Ebolavirus species but differ between them, for example R (Arginine) and T (Threonine) which are conserved within Zaire, Sudan, Bundibugyo and Taï Forest but differ in all the Reston species (many Valine and Serine in this example).

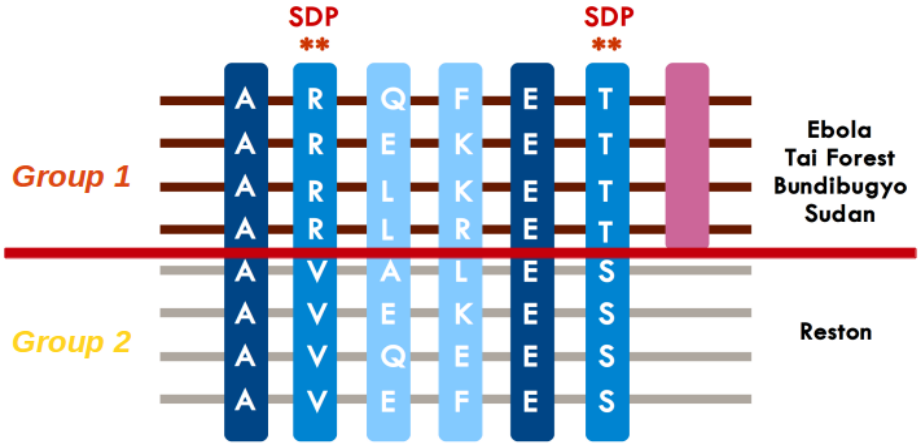


Figure 1.5: Specificity Determining Positions (SDPs) in the different Ebolavirus Species are shown in two different groups: group 1, for human pathogenic species and group 2 for Reston, the only non human pathogenic species. Arginine R and Threonine are conserved within group 1 but they change respectively in Valine and Serine in group 2, where they are still conserved. These two positions are considered SDPs.

1.6.7 Machine Learning – Support Vector Machines (SVMs)

Machine Learning is widely used in bioinformatics in the development of prediction algorithms. The basic premise of machine learning is to predict a particular property, for example protein function or whether a nsSNV is deleterious, using of a set of features. Machine learning algorithms are trained using a dataset where the properties are already known, so that the algorithm can learn how to associate the values of the features with the property being predicted. This often results in the algorithm learning rules or trends that associate the features with the predicted property.

Machine Learnings consist of three main statistical fundamentals: first, classification, which is a supervised method and for which we know to which class data belong to; second, clustering, which is unsupervised since it groups the data but ignores the labels and third, regression, which is supervised and consists on building a separation of the different groups according to the labels. In statistics, supervised learning can be divided into classification and regression. Classification is part of pattern recognition methods and it assumes that data labels are finite and discrete, whilst regression gives a function estimation and labels depend on a continuous set of data.

Support Vector Machines (SVMs; Vapnik, 1995) are a widely used type of supervised machine learning method. SVMs have been successfully applied in the development of methods for the prediction of protein function (Wass et al., 2012), genetic mutations on protein stability, for protein folding recognition, for protein structure classifications, for secondary structure predictions or even for cancer classification using gene expression data (Petryszak et al., 2013; Kapushesky et al., 2012).

SVMs are based on the principle that algorithms "learn" according to a class of tasks. Typically they depend on several parameters and their choice is not always straightforward. The larger is the number of parameters the more complex is the task.

SVMs are based on a similarity function that is referred to as a kernel. A kernel is a class of algorithm for pattern recognition that allows the use of implicit coordinates and obtaining high dimensional feature space. Whether the extraction of the features can be very expensive, kernels can decrease costs by computing inner products, implicitly. Kernels take into account the distances in a feature space, they compute matrices and they give an estimation of similarity.
Given a set of data, one can embed it in a vector space and look for linear relations in that space. Kernels allow to specify the inner product function between points in that space, by considering all the pairwise inner products. So, for example, given a vector space X the inner products are:

$$\langle \bar{x}, \bar{z} \rangle = \sum_i x_i z_i \qquad (1.6.1)$$

The use of kernels has been extensively used in multivariate statistics algorithms based on eigenproblems, for example Support Vector Machine Learnings (SVMs), Principal Component Analysis (PCA), Canonical correlation analysis and others. There are several types of kernels, among the most important the linear (also the simplest), polynomial kernels, radial basis function kernels and sigmoid kernels.
Linear kernels are applied to linearly separable problems.

The simplest SVM uses a linear kernel to build a hyperplane to separate two groups. The hyperplane separates two groups with the criterion to maximise the margins which separate the groups. The elements of the groups which intersect the two margins are called support vectors. A simplification of an SVM is shown in *figure 1.6*:

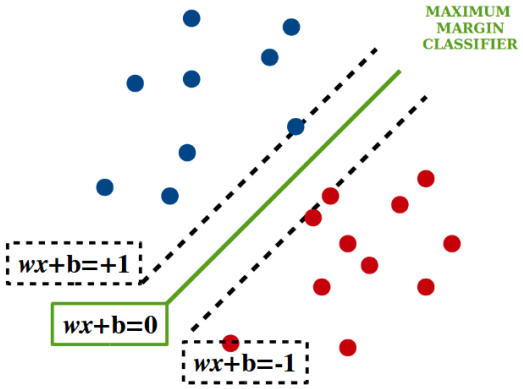


Figure 1.6: Simplification of SVMs. The first group (blue spheres) is separated from the second group (red spheres) by an optimal hyperplane (green line) and it is called the maximum margin classifier. The spheres of each group that intersect the dotted margins are termed support vectors. The equation $wx+b=0$ describe the optimal hyperplane whilst $wx+b=-1$ and $wx+b=+1$ represent the lines that describe the closest margins to one side and the other.

1.6.8 Molecular Dynamics basis and principles

Molecular dynamics simulations are computer calculations that model the motion of atoms and molecules as a function of time. The first molecular dynamics simulation was solved in 1977, when a bovine pancreatic trypsin inhibitor in a vacuum was simulated for less than 10 ps. (McCammon et al., 1977). MD can be a very informative method for protein folding, for conformational changes and on binding free energies. Molecular simulations can predict with good approximation the behaviour of molecules in solvent, in double phase or in membranes. MD can help the understanding and interpretation of molecular recognition with high confidence, especially where experiments are not possible; it can also aid in the refinement of X-ray crystallography and NMR structures.

Molecular dynamics simulations give, as output, an ensemble of configurations that essentially represent the coordinates and the velocities of the studied system as function of time. This output is referred to as a trajectory.

The statistical basis of Molecular Dynamics is based on the principles described

below. Given a system with multiple components, its internal Energy can be described by its Hamiltoninan:

$$H = H(\mathbf{r}, \mathbf{p}) \equiv K(\mathbf{p}) + U(\mathbf{r}) = \sum_i \frac{\mathbf{p}_i}{2m_i} + U(.....\mathbf{r}_i, ...) \qquad (1.6.2)$$

where *K(p)* is the kinetic energy of the system, and *U(r)* is its potential energy. The Hamiltonian asserts that the sum of *K(p)* and *U(r)* is equal to the sum of the mo*m*entum *p* of a pa*r*ticle *i* divided by two times its mass and summed to its potential Energy *U* at each position r_i. The probability distribution, for the atoms in the system in each point is given by the Boltzmann distribution:

$$\rho(\mathbf{r}, \mathbf{p}) = \frac{exp[-H(\mathbf{r}, \mathbf{p})/k_B T]}{Z} \qquad (1.6.3)$$

where $k_B T$ is the Boltzmann constant. Given that it is impossible to know the Boltzmann probability for all states, when we study microscopic systems we refer to the ergodic hypothesis, which states that for an infinitely long system all the accessible micro states will have thermodynamics and dynamics averages which will coincide:

$$\lim_{\tau \to \infty} \langle A(\mathbf{r}, \mathbf{p}) \rangle_\tau = \langle A(\mathbf{r}, \mathbf{p}) \rangle_Z \qquad (1.6.4)$$

in this equation the first term in angle brackets $[A(r,p)]_T$ refers to thermodynamics averages and the second one $[A(r,p)]_Z$ to the dynamics averages, where *T* is the time length of the trajectory and *Z* is a canonical partition function referring to an integral over all space phase. Since MD deals with discrete (and not infinitesimal) objects

one can apply this principle.

Classical Molecular dynamics allows the study of thermodynamics and kinetics properties, according the Newton's second law:

$$F = m\left(\frac{\partial v}{\partial t}\right) = ma \tag{1.6.5}$$

where *F* is the force that is applied to the particle, *m* is the mass of the particle and *a* its acceleration.

A force field describes all the intra and inter molecular interactions, in terms of the potential energy of the system. It is the sum of all the energetic terms that contribute to the potential energy of the system. The force field follows two fundamental equations: Schrödinger's equation and the Born-Oppenheimer approximation.

The Schrödinger equation describes a molecular system by a relativistic time-dependent point of view.

$$HY\left(r, R\right) = EY\left(r, R\right) \tag{1.6.6}$$

This equation needs to be adjusted, especially for systems with many atoms. For this reason the Born-Oppeneimer approximation is fundamental in MD. This approximation asserts that electrons adjust their dynamics accordingly to the atomic position changes as described in the following equation:

$$HF\left(R\right) = EF\left(R\right)$$

(1.6.7)

The force field represent the ensemble of the bonded interactions, such as angle, dihedral and improper (plane-plane) and non bonded interactions such as Van der Waals and Electrostatics.

The Periodic Boundary Condition (PBC) asserts that atoms interact with their neighbours and also with the periodic atom corresponding to the mirror image of itself. PBC are very useful in MD, especially for large systems. After the simulation one should be aware of this artefact and apply the Minimum Image Convention (MIC) which ensures that the atoms interact only with the closet image inside the box of the simulation.

An Example of PBC is shown in the *figure 1.7*. Here the solvation box has been filled with waters and the protein is shown in the central black box. The other boxes show the closest periodic mirror image, necessary to make the system infinite-like. When a particle leaves a simulation box (for example the circle inside the central box) it is immediately replaced by the same atom from the opposite periodic image.

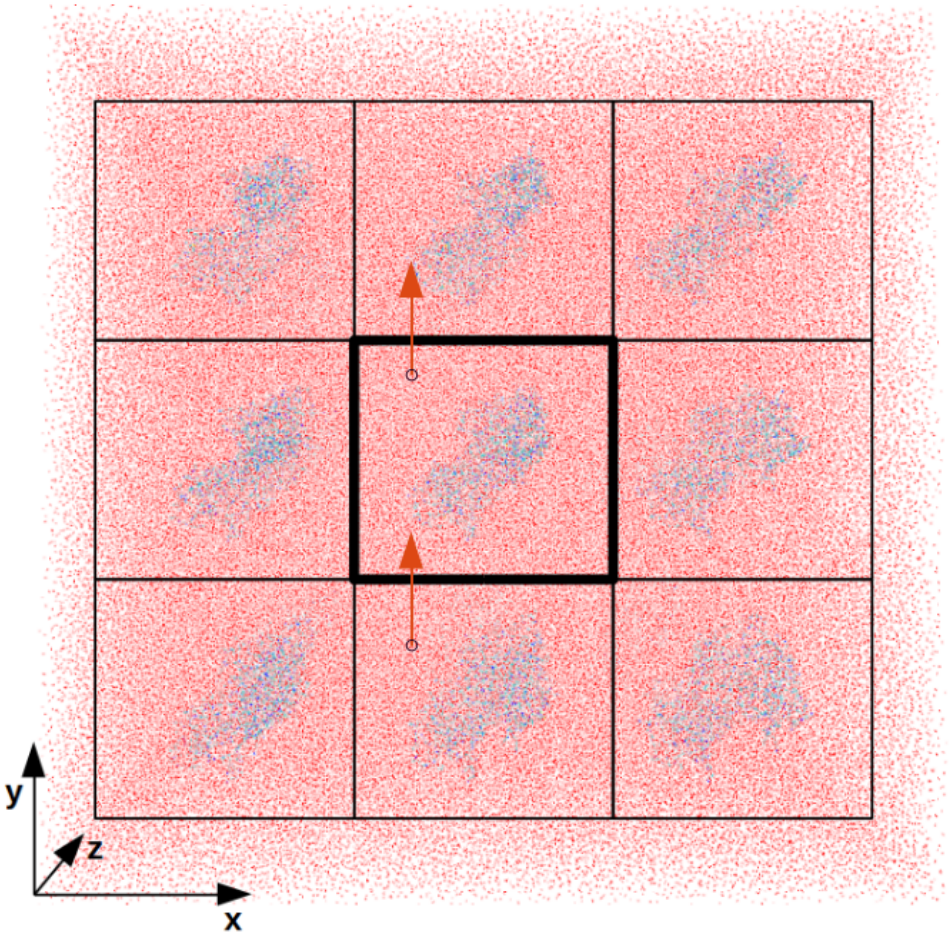


Figure 1.7: PBC in a simulation box. The Protein is shown in blue cartoons and water molecules in red sticks. The central box marked in bold is the simulation box whereas all the others represent periodic images of the central one. In this example, Periodic Boundary Conditions allow atoms that interact with others outside the simulation box, to be replaced by other atoms coming from the bottom periodic image and keep the system in equilibrium during the simulation. Simultaneously the same substitution occurs across all the boxes.

1.6.8.1 MD protocol:

Each molecular dynamics simulation consists of different steps:

1. defining the initial velocities (according to the Maxwell distribution) and creating the topology file.
2. defining the unit cells
3. adding solvent molecules

4. neutralising the system, adding ions
5. Energy minimisation (relaxation of the structure to assure that there are not steric clashes or inappropriate geometries)
6. Equilibration phase, which is a critical phase for the entire simulation; this step consists of two phases: stabilisation of temperature (isothermal-isochoric), where the canonical ensemble NVT (where N refers to the number of particles in the system, V to the volume of the system and T to the absolute temperature) is defined (this ensures that the number of particles, volume and temperature is constant); and as second phase, the stabilisation of pressure (isothermal-isobaric) and thus of stabilisation of the density. During the equilibration, it is standard procedure to apply restraints, in order to equilibrate the solvent around the protein, and in this way getting a bigger control of the simulation.
7. Production
8. Analysis

The protein structures used for MD simulations need to have a good crystal structure resolution without missing backbone atoms; then the molecule is fitted into a box which will be filled with solvent molecules (if the simulation is in solvent; simulations can also be performed in a vacuum); then the temperatures and the pressure are assigned and the system is energetically minimised according to the initial and the rescaled velocities, temperatures and pressures; during the minimisation phase at specific temperature one can also use restraints, in order to have more control of the whole simulation. Finally, the production phase, where the simulation starts for a specific length of time.

1.6.9 Principal Component Analysis

Principal Component Analysis (PCA) is a linear transformation, widely used to analyse the motion of proteins during Molecular Dynamics (MD) simulations. PCA is used to reduce the dimensionality of a problem and in the case of MD it can aid interpretation of the motion in terms of eigenvectors and eigenvalues. Given a

motion, the eigenvalue corresponds to the weight of the eigenvector to the motion of a protein.

The aim of the PCA for MD trajectories is to determine the predominant direction for all the structural changes. Considering a system with N atoms, we can describe the internal motion according to the following covariance matrix:

$$\sigma_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle \qquad (1.6.8)$$

where the values in round brackets (…) are the values of the masses in Cartesian Coordinates and the ones in angle brackets < … > are the average of all the sampled conformations. This covariance matrix is then diagonalised in order to obtain 3N eigenvectors ($\vec{V}$) and eigenvalues (V_i). These describe the motions and they can be projected into Principal Components:

$$V_i = V^i r \qquad (1.6.9)$$

It has been shown that the majority of the fluctuations of a system can be described by the first principal component. There are several ways to visualise PCA and one of them is to use a porcupine visualisation, where Cα atoms are linked to cones which have the same direction of the eigenvector. Each cone has a length which is proportional to the amplitude of the corresponding motion. An example is shown in *figure 1.8*:

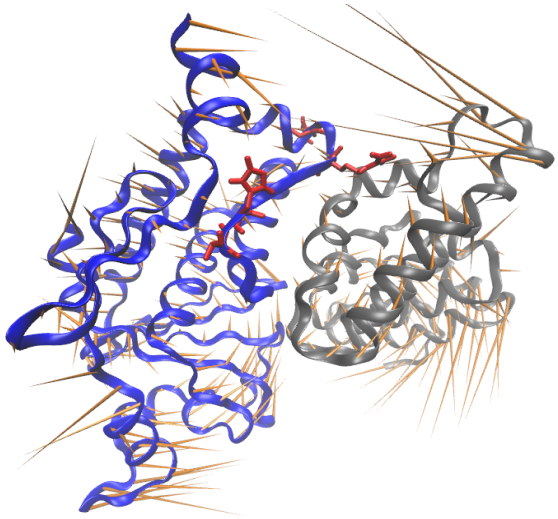


Figure 1.8: Porcupine visualisation of Principal Component 1 and 2 in Ebola VP24 protein shown in blue cartoon, in complex with human Karyopherin Alpha 5, shown in gray cartoons. In red sticks a series of mutations occurring at the interface are shown. The yellow cones represent the amplitude of the C alpha movements, as obtained with Principal Component Analysis.

1.7 Organisation of this thesis

The two main research lines:

- the development of VarMod, a computational algorithm to predict the effects of single nucleotide nonsynonymous single nucleotide variants
- analysis of genetic mutations occurring in Ebolaviruses to understand how they affect human pathogenicity.

The thesis is divided into the following six chapters:

The *Introduction Chapter* has described the state-of-the-art for the analysis of human genetic variation, an introduction to Ebolaviruses and the methods used to analyse mutations present in Ebolavirus genomes.

Chapter 2, contains the article entitled "Varmod: Modelling the functional effects of nonsynonymous variants" published in *Nucleic Acid Research Journal* (Pappalardo &

Wass, 2014).

In *Chapter 3*, contains the article entitled "Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses" describes analysis to identify the molecular determinants of Ebolavirus Pathogenicity; published in *Scientific Reports* (Pappalardo et al., 2016).

Chapter 4 - "Analysis of Ebola virus mutations present in rodent adaption experiments". This manuscript considers the structural analysis of Ebolavirus mutations obtained from several adaptation studies inducing pathogenicity in mice and guinea pigs; this work is in preparation and will shortly be submitted to *Genome Biology*.

Chapter 5 - "Molecular dynamics analysis of Ebola virus pathogenicity" This chapter builds upon the results from chapter 3 and uses molecular dynamics to investigate mutations in VP24 and how this may affect binding to the human protein Karyopherin alpha5. This work is in preparation and will soon be submitted to *PLOS Computational Biology*.

Chapter 6 - Discussion.

Chapter 2:

VarMod: modelling the functional effects of non-synonymous variants

M. Pappalardo & M.N. Wass (2014), "VarMod: modelling the functional effects of non-synonymous variants", *Nucleic Acids Res.*, 42: W331–W336.

This work was entirely developed by my supervisor, Mark Wass and me. I developed the vast majority of the back end scripts and analysis that are performed by VarMod. This includes:

1. Generation of multiple sequence alignments and calculation of conservation
2. Structural modelling of the query protein
3. Analysis of structural properties (e.g. solvent accessibility and secondary structure)
4. Analysis of protein-protein interactions and the proximity of variants to interfaces

The machine learning element was implemented with my supervisor. My supervisor developed the front end of the webserver, we worked together on the overall design of the website and I tested the server.

2.1 Abstract

Unravelling the genotype–phenotype relationship in humans remains a challenging task in genomics studies. Recent advances in sequencing technologies mean there are now thousands of sequenced human genomes, revealing millions of single nucleotide variants (SNVs). For non-synonymous SNVs present in proteins the difficulties of the problem lie in first identifying those nsSNVs that result in a functional change in the protein among the many non-functional variants and in turn linking this functional change to phenotype. Here we present VarMod (Variant Modeller) a method that utilises both protein sequence and structural features to predict nsSNVs that alter protein function. VarMod develops recent observations that functional nsSNVs are enriched at protein–protein interfaces and protein–ligand binding sites and uses these characteristics to make predictions. In benchmarking on a set of nearly 3000 nsSNVs VarMod performance is comparable to an existing state of the art method. The VarMod web server provides extensive resources to investigate the sequence and structural features associated with the predictions including visualisation of protein models and complexes via an interactive JSmol molecular viewer. VarMod is available for use at http://www.wasslab.org/varmod.

2.2 Introduction

The ability to sequence genomes has resulted in the identification of millions of genetic variants, particularly single nucleotide variants (SNVs), within the human population as highlighted by the 1000 genomes project (1000 Genome Project consortium, 2010; Abecasis et al., 2012). Additionally, other studies have demonstrated that individuals have many rare SNVs (Nelson et al., 2012; Tennessen et al., 2012). The data generated by such studies provide a unique resource for investigating the genotype to phenotype relationship. However, this is a complex problem as demonstrated by Genome Wide Association Studies (GWAS), which

have identified many variants associated with disease risk but have only explained a limited amount of heritability (Eichler et al., 2010). Additionally, in these studies, it is difficult to identify causal variants from a selection of candidate SNVs in the regions of the genome associated with the particular disease.

There is therefore a need to develop methods to identify SNVs, in our case non-synonymous SNVs (nsSNVs), that are likely to affect the function of the protein in which they are present and are more likely to be associated with a change in phenotype. A number of methods have been developed previously (reviewed in (Peterson et al.,2013), with the Sorting Intolerant From Tolerant algorithm (SIFT, Sim et al.,2012) and PolyPhen (Adzhubei et al., 2010) being among the most well known. SIFT uses residue conservation in multiple sequence alignments to identify function altering nsSNVs, while PolyPhen uses machine learning to combine features from both sequence and structure.

Here we have developed VarMod a new method for identifying functional nsSNVs. VarMod develops our recent research in which we demonstrated that disease associated nsSNVs are enriched at protein–protein interfaces (David et al., 2012). Additionally, in GWAS, we have previously used structural modelling of ligand binding sites to identify likely candidates for association with disease (Chambers et al., 2010; Chambers et al., 2011; Chambers et al., 2009). For example, in a kidney disease genome wide association study (Chambers et al., 2010), we demonstrated that the variant rs13538 results in a phenylalanine to serine change located in the acetyl Co-enzymeA binding site of the protein NAT8 and proposed that the variant may have an effect on the activity of the enzyme (Chambers et al., 2010). VarMod builds upon these observations and uses structural modelling of ligand binding and protein–protein interface sites to generate features that are combined with other features such as residue to conservation to identify functional nsSNVs. The VarMod web server provides an overall prediction made using a machine learning approach (a support vector machine) to combine the data from the different individual analyses. Additionally the server provides users with extensive resources to investigate the results from the separate analyses.

2.3 Methods

2.3.1 The Varmod Algorithm

VarMod obtains features from multiple analyses, which are combined using a support vector machine (SVM) (Vapnik, 1999) to make an overall prediction. The data sources used are described below. Sequence conservation is calculated using Jensen–Shannon divergence (Capra and Singh, 2008). Homologues of the query sequence are identified by PSI-BLAST (Altschul et al., 1997) using an approach shown to optimise results (Chubb et al., 2010), where the query sequence is initially searched against UniRef50 to generate a sequence profile that is used to search against the full UniProt sequence database (Uniprot Consortium, 2012). The query sequence and homologues are aligned using MUSCLE (Edgar, 2004) and the resulting multiple sequence alignment used to calculate the Jensen–Shannon divergence.

To perform the structural analysis, a structural model of the query protein is generated. To do this, template structures in the protein databank (PDB) (Rose et al., 2013) are identified using hhblits (Remmert et al., 2012) by searching a PDB sequence database representative at 70% sequence identity. Templates are selected with an hhblits probability (probability that the template and query sequence are homologous) score >80% and such that as much of the sequence is covered without redundantly modelling the same region of the protein multiple times. Initial structural models are generated using an approach based on the one used by Phyre2 (Kelley and Sternberg, 2009; Bennet-Lovsey et al., 2008). Side chains are added and optimised using pulchra (Rotkiewicz and Skolnick, 2008). Small molecule binding sites are modelled using 3DLigandSite (with default parameters) (Wass et al., 2010) with the structural model used as the input.

Protein–protein interface sites are modelled using an approach based on Interactome3D (Mosca et al., 2012). The Interactome3D high confidence set of protein–protein interactions with template complexes in the PDB was used to generate models of the complexes. For each sequence–template pair the sequence is modelled using the template by applying the structural modelling approach described above.

The features used in the SVM fall into two areas of sequence and structural features (a full list is available in Supplementary Table S1). The sequence features include residue conservation (the Jensen–Shannon convergence) and three features that represent the change of amino acid properties of size/mass, charge and functional group. The size/mass change of the amino acid is represented by the ratio of the mass of the two amino acids. To consider the change in charge between the two amino acids, the 20 amino acids are grouped according to charge (Supplementary Table S2). The feature representing the change in the charge of the amino acid considers changes between these charge groups, with values set in Supplementary Table S3. A further feature represents the change of chemical functional group present in the amino acid side chain. The amino acids are grouped as described by Innis *et al.* (Innis et al., 2004) (Supplementary Table S4) and the feature captures changes between these functional groups.

The structural features use the ligand binding site, interface site and general structural features of the model. Where ligand-binding sites have been identified the distance of the variant to the binding site is calculated and used as a feature. When a variant is in a binding site, two further features capture results from the 3DLigandSite analysis. Where interface sites have been predicted, a further feature represents the distance of the variant to an interface site. Two features represent the type of secondary structure that the variant is located in. The first uses the secondary structure types classified by DSSP (Joosten et al., 2011; Kabsch and Sander, 1983), while a second feature reduces these to the three main categories of helix, sheet and coil. A final feature represents the solvent accessibility (calculated using DSSP).

The features generated are input into each of the five optimised SVM models generated during cross-validation (details below) to predict whether each variant is functional or non-functional. The outputs from each of the SVM models are converted to probabilities as described in Platt (Platt, 1999). An ensemble approach is taken with the probability from each SVM model weighted according to its accuracy in cross validation. The weighted probabilities are summed and normalised to generate a final probability for the VarMod prediction.

2.3.2 Generating a test set

Dataset 5 from VariBench (Sasidharan and Vihinen, 2013) was used to train and test VarMod. This dataset contains human pathogenic and neutral variants, excludes cancer mutations and is clustered so that protein sequences share no >30% sequence identity. This set was initially split with 1401 pathogenic and 1527 neutral variants retained for final testing. The remaining 11 336 pathogenic and 12 737 neutral variants were split into five groups by protein sequence to perform 5-fold cross-validation to ensure that variants from each individual sequence appear in only 1-fold.

2.3.3 SVM training

The SVMs were generated by SVMlight (Joachims, 1999) using a linear kernel. For each of the 5-folds, three were used for training, a further fold was used for validation and the SVM tested on the remaining fold. The SVMs were optimised for the trade off between training error and margin and also the cost factor to identify how training errors on positive examples should outweigh those on negative examples.

2.3.4 Comparison with Polyphen

To compare VarMod performance with PolyPhen-2, the final test set of nsSNVs was run on the PolyPhen-2 web server (on 1 March 2014). Predictions were made using the two different classifiers available (HumDiv and HumVar) with default settings. The ROC and Precision–Recall analyses of PolyPhen-2 were performed by varying the 'pph2_prob' score. Additionally VarMod performancs was compared to SuSpect (Yates et al., 2014). The final test set of nsSNVs was submitted to the SuSpect web server in June 2016. The ROC and Precistion-Recall analysis for SuSpect was performed by varying the threshold for the probability score associated with SuSpect predictions.

2.3.5 Evaluating VarMod Performance

The performance of VarMod was assessed using the set of sequences from VariBench that were not used in cross-validation. The performance of VarMod on the test set of sequences was assessed using the measures of specificity, sensitivity (recall), precision and a Receiver Operator Characteristic (ROC) analysis. The ROC curve and Precision–Recall graph in Figure 2.1 show the performance of VarMod and the comparison with PolyPhen-2 and SuSpect. The ROC analysis shows that VarMod performance is comparable to both PolyPhen-2 and SuSpect. Interestingly, in the ROC analysis, neither of the PolyPhen-2 classifiers reaches the point 0,0 which is due to a small number of high confidence false positive predictions (i.e. neutral variants predicted to be pathogenic). This may reflect that PolyPhen-2 has been trained using different sets of pathogenic and neutral variants. It has also been previously observed that there is limited overlap between the predictions of different methods (Chun and Fay, 2009). The precision-recall analysis shows similar performance between VarMod and PolyPhen-2. However, SuSpect outperforms both methods. It is possible that SuSpect is simply better than the other methods, however for both PolyPhen-2 and SuSpect we do not know if sequences present in this final test set were also used in training. SuSpect was trained using the UniProt Humsavar dataset and then benchmarked using VariBench, ensuring that they removed any sequences from VariBench that were present in the training set (Yates at al., 2014). To fairly test the methods the test set should not contain any sequences that were present in the training set.

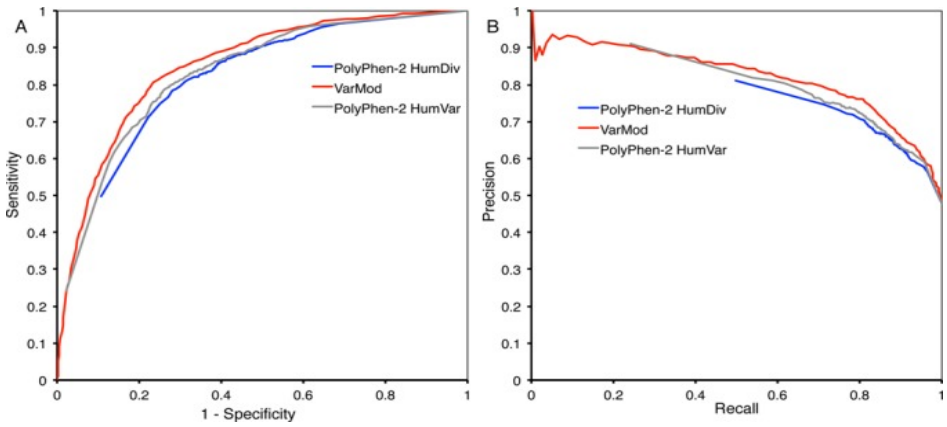


Figure 2.1: Benchmarking VarMod. Analysis of the VarMod and PolyPhen-2 predictions on the non-cross validation test set. (**A**) ROC analysis, (**B**) precision–recall graph.

2.4 Results

2.4.1 The VarMod web server

The VarMod web server is available at http://www.wasslab.org/varmod. Users are required to submit a protein sequence (raw sequence or FASTA formatted) or a UniProt accession, and a list of variant positions (e.g. A45C, where the single letter code is used to define the amino acids). A UniProt accession is required to perform the protein–protein interface analysis (optional). Processing time for each submission varies from 5 min to a few hours. Structural data has been pre-computed for all of the UniProt human principal protein isoforms, so submissions using these sequences are processed in a few minutes. Where other sequences are submitted, the structural models and binding sites need to be modelled thereby increasing the running time to a few hours.

2.4.2 Results Output

The display of VarMod results is split into multiple sections (Figures (Figures 2.2 and 2.3). The first section provides a summary table of the analyses performed and the overall prediction made for each of the submitted nsSNVs. This table is colour coded to highlight the results from the individual analyses/features to indicate if they suggest the variant could affect protein function. For example, the binding site column is coloured red if the variant is in the binding site and the colour changes to blue the more distant the variant is from a known ligand-binding site. The summary table enables the user to see the overall result and to identify analyses that may be of interest for further inspection.

The sequence and structure sections display the main analyses. The sequence section

displays the protein sequence, colour coded to highlight multiple features including residue conservation, ligand binding sites and protein–protein interfaces. The summary results and sequence view can be downloaded as a PDF file.

The structural section first displays the details of the structural templates and models of the protein that have been generated (one for each region/domain for which a template was identified). A JSmol (www.jmol.org) molecular viewer forms the main part of the structural section and initially displays the model with the highest confidence (probability from hhblits alignment). The JSmol viewer enables visualisation of the modelled protein and by default is coloured to highlight the functional regions of the protein (ligand-binding and protein–protein interface sites) and the nsSNVs (red). A control panel to the right of the display enables the user to investigate the nsSNVs by displaying a different model, or modifying the display style (cartoon/spacefill or sticks representations) and colour of the whole protein, nsSNVs or functional sites. The user is able to generate high quality images of the displayed model by clicking on the 'generate image' button, enabling the analysis to be used for reports or publications.

The location of the nsSNVs in relation to the protein–protein interface sites can be explored further via the modelled complexes. The complex models are listed in a table, which also indicates the nsSNVs that are present in the model and if they occur within an interface. The complexes can be viewed in a separate JSmol viewer accessed from a link for each of the entries in the list.

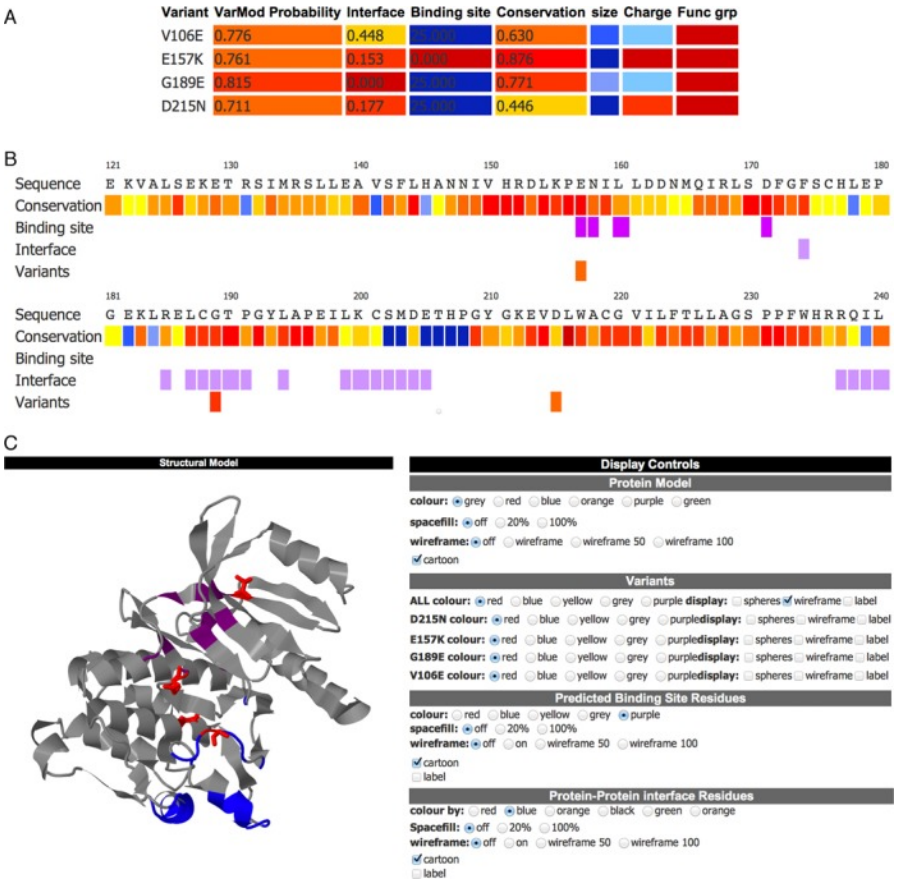


Figure 2.2: Display of VarMod results. The results for variants in Phosphorylase b kinase gamma catalytic chain (UniProt accession P15735). The variants shown are known to have a role in Glycogen storage disease 9C. (**A**) The prediction summary table, showing the overall VarMod prediction and summarising the output from the different analyses. Results are colour coded to indicate the likely relevance of the changes, with features that suggest the variant is likely to be functional coloured red with the colour scale ranging to blue for features that are least likely to lead to functional changes. (**B**) The VarMod sequence display, residues are coloured to indicate conservation and the presence of ligand binding and interface sites. (**C**) The VarMod structural view.

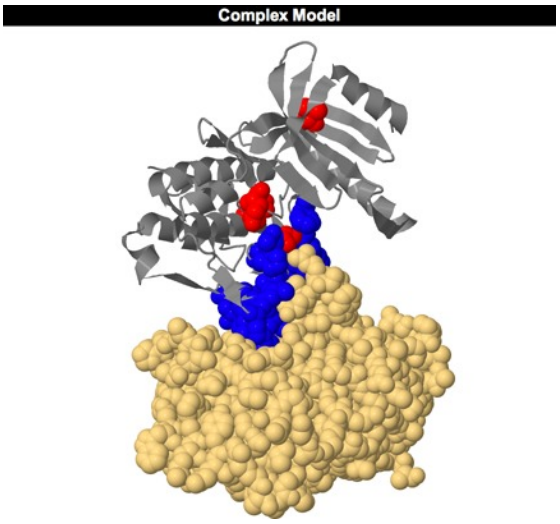


Figure 2.3: The VarMod interactions view for investigating variants located at protein–protein interfaces.

2.5 Discussion

VarMod was developed to use recent observations that disease associated nsSNVs are frequently located at ligand-binding and protein–protein interface sites and to automate manual approaches that we have previously used to analyse GWAS candidate nsSNVs. We have demonstrated that VarMod performance on a large and established benchmark set is comparable to an existing state of the art method (PolyPhen-2). The VarMod server provides a resource for users to identify functional nvSNVs and to investigate the individual features associated with these variants. Plans for future improvements to the server include increasing the number of interface and binding site features such as considering how variants may alter binding energies and options to submit variants in alternative formats such as Variant Call Files (VCF), which will facilitate high throughput analysis of nsSNVs identified from sequencing studies.

Chapter 3:

Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses

Morena Pappalardo†, Miguel Juliá†, Mark J. Howard, Jeremy S. Rossman, Martin Michaelis, Mark N. Wass, Scientific Reports 2016, 6:23743, doi:10.1038/srep23743.

† equal contribution.

My contribution to this work has been the analysis of the Ebolavirus genome sequences, generating multiple sequence alignments, identifying specificity determining positions and the subsequent structural analysis of the SDPs to identify their potential effects on protein structure and function. Miguel Julia is a joint first author on the paper, he focused on the phylogenetic analysis of the Ebolvirus proteins and genomes and the simulations considered the confidence with which the SDPs were predicted. I also contributed to writing the manuscript.

3.1 Abstract

Reston viruses are the only Ebolaviruses that are not pathogenic in humans. We analyzed 196 Ebolavirus genomes and identified specificity determining positions (SDPs) in all nine Ebolavirus proteins that distinguish Reston viruses from the four human pathogenic Ebolaviruses. A subset of these SDPs will explain the differences in human pathogenicity between Reston and the other four ebolavirus species. Structural analysis was performed to identify those SDPs that are likely to have a functional effect. This analysis revealed novel functional insights in particular for Ebolavirus proteins VP40 and VP24. The VP40 SDP P85T interferes with VP40 function by altering octamer formation. The VP40 SDP Q245P affects the structure and hydrophobic core of the protein and consequently protein function. Three VP24 SDPs (T131S, M136L, Q139R) are likely to impair VP24 binding to human karyopherin alpha5 (KPNA5) and therefore inhibition of interferon signaling. Since VP24 is critical for Ebolavirus adaptation to novel hosts, and only a few SDPs distinguish Reston virus VP24 from VP24 of other Ebolaviruses, human pathogenic Reston viruses may emerge. This is of concern since Reston viruses circulate in domestic pigs and can infect humans, possibly via airborne transmission.

3.2 Introduction

Four of the five members of the genus *Ebolavirus* (Ebola viruses, Sudan viruses, Bundibugyo viruses, Taï Forest viruses) cause hemorrhagic fever in humans associated with fatality rates of up to 90%, while Reston viruses are non-pathogenic to humans (Feldmann and Geisbert, 2011; Weingartl et al., 2013) (see Materials and Methods for the *Ebolavirus* nomenclature). So far there have been three Reston virus outbreaks in nonhuman primates: 1989-1990 in Reston Virginia, USA, 1992-1993 in Siena, Italy, and 1996 in a licensed commercial quarantine facility in Texas. All cases were traced back to a single monkey breeding facility in the Philippines. During these outbreaks five human individuals were tested positive for IgG antibodies directed against Reston virus. Moreover, Reston virus was found in 2008 in domestic pigs in the Philippines. Seroconversion was detected in six human individuals. None of the 11 individuals that were seropositive for Reston virus antibodies reported an Ebola-

like disease (Miranda and Miranda, 2011) .

The reasons underlying the differences in human pathogenicity between Reston viruses and the members of the other *Ebolavirus* species remain unclear. Understanding of the molecular causes of these differences would enhance our understanding of Ebolavirus function and pathogenicity and aid investigation into treatment of Ebolavirus infection. Here, we performed an *in silico* analysis of the genomic differences between Reston viruses and human pathogenic Ebolaviruses to identify conserved changes at the protein level that explain the differences in Ebolavirus pathogenicity in humans.

Ebolaviruses encode nine proteins including nucleoprotein (NP), glycoprotein (GP), soluble GP (sGP), small soluble GP (ssGP), RNA dependent RNA polymerase (L), and four structural proteins termed VP24, VP30, VP35, and VP40 (Feldmann and Geisbert, 2011; Mehedi et al., 2011; La Vega et al., 2015). GP, sGP, and ssGP are produced from the *GP* gene by alternative RNA editing (Feldmann and Geisbert, 2011; Mehedi et al., 2011; La Vega et al., 2015). Many of the Ebolavirus proteins have multiple functions. In the virion, the NP-encapsulated RNA genome associates with VP35, VP30, and L to form the transcriptase-replicase complex. VP35 and VP24, a membrane-associated structural protein, antagonize the cellular interferon response. The matrix protein VP40 fulfills critical roles during virus assembly and release. GP, the only transmembrane surface protein, is responsible for host cell binding and virus internalization (Feldmann and Geisbert, 2011; Basler, 2014). Little is known about the functional roles of the secreted proteins sGP and ssGP (Feldmann and Geisbert, 2011; Miranda and Miranda, 2011; Mehedi et al., 2011; Hoenen et al., 2015).

Despite the small Ebolavirus genome we still have a limited understanding of Ebolaviruses and what causes their pathogenicity and why Reston viruses are not human pathogenic (Feldmann and Geisbert, 2011; Basler, 2014; Zhang et al., 2012). The importance of understanding these differences is highlighted by the current

Ebola virus outbreak in Western Africa, which is the first large outbreak and has resulted in 27,345 suspected cases and 11,184 deaths to date (www.who.int, as of 14th June 2015). During this outbreak many additional Ebola virus genomes were sequenced enabling us to perform the first comprehensive comparison of the non-human pathogenic Reston virus to all four human pathogenic Ebolaviruses. While some studies (Zhang et al., 2012; Bale et al., 2013; Clifton et al., 2014) have compared the differences between individual Reston virus proteins derived from a certain strain with their equivalent derived from one strain of a human pathogenic species, none have performed a systematic analysis of all available protein sequence information from all (known) *Ebolavirus* species.

Our large scale analysis of nearly 200 different Ebolavirus genomes focussed on combining computational methods with detailed structural analysis to identify the genetic causes of the difference in pathogenicity between Reston viruses and the human pathogenic *Ebolavirus* species. Central to our approach was the identification of Specificity Determining Positions (SDPs), which are positions in the proteome that are conserved within protein subfamilies but differ between them (Casari et al., 1995; Rausell et al.,2010) and thus distinguish between the different functional specificities of proteins from the different *Ebolavirus* species. SDPs have been demonstrated to be typically associated with functional sites, such as protein-protein interface sites and enzyme active sites (Rausell et al.,2010). The SDPs that we have identified and that distinguish Reston viruses from human pathogenic Ebolaviruses, arguably, contain within them a set of amino acid changes that explain the differences in pathogenicity between Reston viruses and the four human pathogenic species, although a contribution of non-coding RNAs (that may exist but remain to be detected) cannot be excluded (Basler, 2014; Teng et al., 2015). The subsequent structural analysis was performed to identify the SDPs that are most likely to affect Ebolavirus pathogenicity, using an approach that is similar to those used to investigate candidate single nucleotide variants in human genome wide association and sequencing studies by us and others (Chambers et al, 2011; Chambers et al., 2010; Chambers et al., 2014; Palles et al., 2013).

3.3 Results

3.3.1 Specificity Determining Positions (SDPs) Analysis

196 Ebolavirus genomes were obtained from the Virus Pathogen Resource (ViPR, Pickett et al., 2012), consisting of 156 Ebola viruses, 7 Bundibugyo viruses, 13 Sudan viruses, 3 Taï Forest viruses, and 17 Reston viruses (online Methods). Phylogenetic analysis of the whole genomes and the individual proteins separated the *Ebolavirus* species from each other (Supplementary Figure S1). There is good agreement between all the trees. The Reston virus sequences are most closely related to Sudan virus than the other three Ebolavirus species. In accordance with previous studies (Morikawa et al., 2007; Gire at al., 2014; Liu et al., 2015; Vogel, 2015; Hoenen et al., 2015), we observed high intra-species conservation with greater inter-species variation (Figure 3.1 and Supplementary Table 1). The surface protein GP exhibited the greatest variation (Figure 3.1), most likely as a consequence of selective pressure exerted by the host immune response (Liu et al., 2015).

Using the S3Det algorithm (Rausell et al., 2010) (Materials and Methods), we identified 189 SDPs that are differentially conserved between Reston viruses and human pathogenic Ebolaviruses (Figure 3.2, Supplementary Figure 2, Supplementary Tables 2-9). These SDPs represent the most significant changes between the Reston virus and the human pathogenic Ebolaviruses so

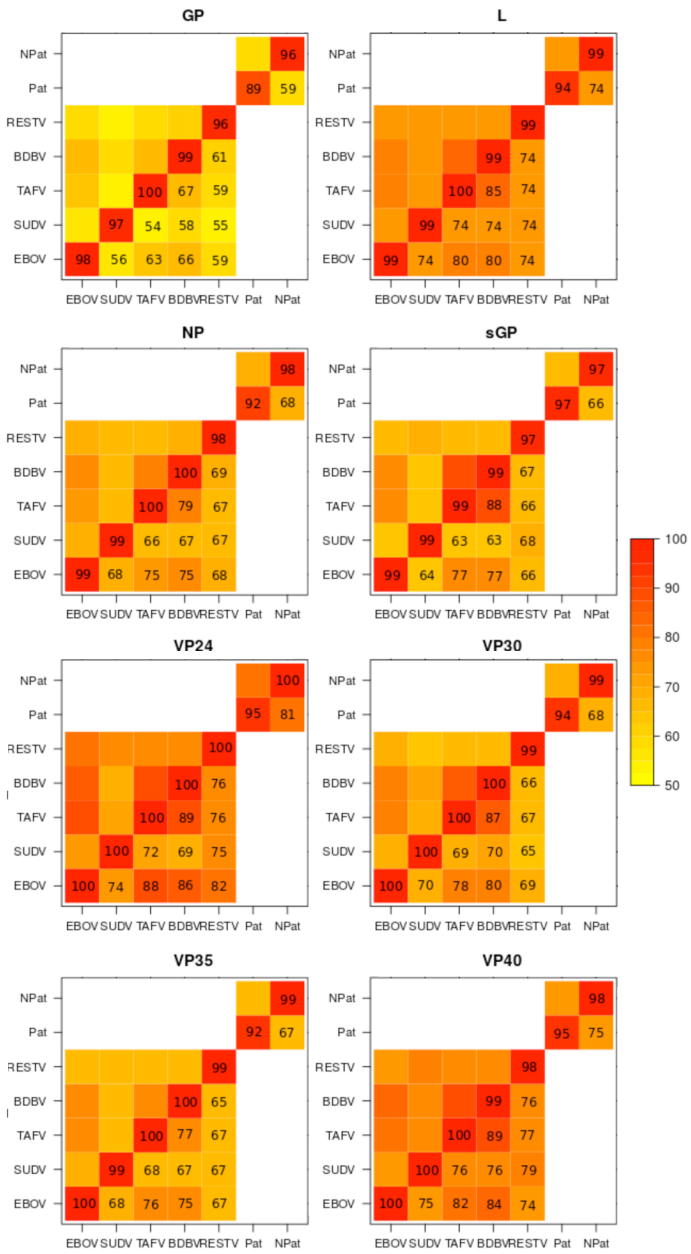
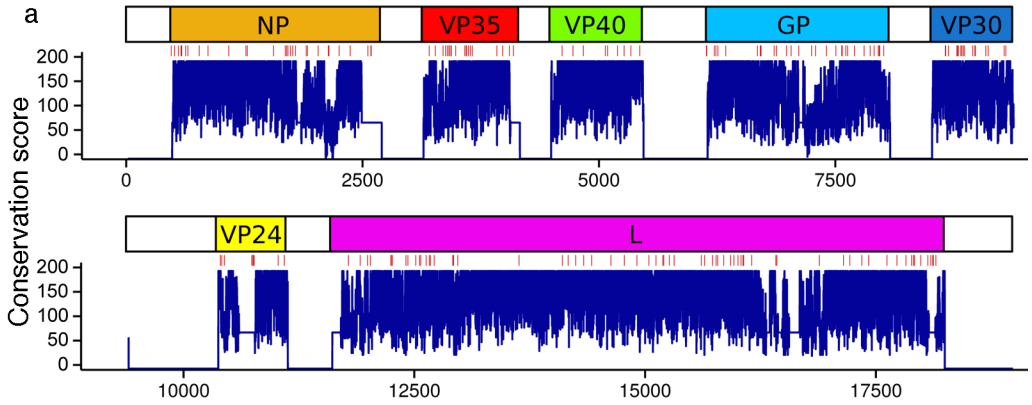


Figure 3.1. Conservation of Ebolavirus proteins. Heatmaps of intra- and inter species sequence identity for Ebolavirus proteins. (EBOV, Ebola virus; BDBV, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus; RESTV, Reston virus).

a subset of these SDPs must explain the difference in pathogenicity. SDPs were present in each of the Ebolavirus proteins representing between 2.4% of residues in sGP to 5.9% of residues in VP30 (Figure 3.2B). Comparison of the SDPs with

previously published mutagenesis studies (Xu et al., 2014) (online Methods) provided no explanation for their functional consequences (Supplementary Table 10).

a



b

Protein	length	SDPs	%residues SDPs	SDPs modelled	surface SDPs	probable integrity	probable interface	possible integrity	possible interface
GP	697	30	4.30	8	7				3
sGP	371	9	2.43	8	7				
NP	749	29	3.87	3	3			2	
L	2239	67	2.95	1	1				
VP24	251	9	3.59	8	8	1	3		
VP30	290	17	5.86	6	6		1		
VP35	341	19	5.57	4	4		1		
VP40	331	9	2.72	8	7	1	1		

Figure 3.2. Ebolavirus SDPs. A) genomic overview of Ebolavirus conservation. SDPs are shown as red lines with protein conservation (blue graph). B) The number of SDPs in each of the Ebolavirus proteins is shown with details on: the number of SDPs that were mapped onto protein structures and the numbers that were identified to have potential roles in changing pathogenicity by either affecting protein-protein interactions (interface) or changing protein structure-function. These changes were classed as probable, where there is high confidence of the effect and possible where there is a lower level of confidence in the observations.

3.3.2 Structural Analysis

Full-length structures for VP24 and VP40 were available, as well as structures for the

globular domains of GP, sGP, NP, VP30, and VP35 (Supplementary Table 11). It was not possible to model the oligeromerization domains of VP30 and VP35 nor the structure of L apart from a short 105 residue segment of the 2239 residue protein, which contained a single SDP. 47 SDPs could be mapped onto Ebolavirus protein structures (or structural models where structures were not available, see online Methods). Most SDPs are located on protein surfaces (Supplementary Figure 3) and are therefore potentially involved in interaction with cellular and viral binding partners and/or immune evasion. Based on our combined computational and structural analysis we find evidence for eight SDPs that are very likely to alter protein structure/function, with six affecting protein-protein interfaces and two that with the potential to influence protein integrity and hence affect stability, flexibility and conformations of the protein (Table 3.1). Five additional SDPs may alter protein structure/function but the evidence supporting them is weaker (Supplementary Tables 12-18). Two of these weaker SDPs were present in NP (A705R, R105K - all SDPs are referred to using Ebola virus residue numbering and show the human pathogenic Ebolavirus amino acid first and the Reston virus amino acid second). A705R is likely to introduce a salt bridge with E694 and R105K will alter hydrogen bonding (Supplementary Table 12). The three other SDPs with weaker evidence were present in the glycan cap in GP (see below). The eight confident SDPs were present in V24, VP30, VP35, and VP40. The VP40 and VP24 SDPs revealed the most changes that may relate to differences in human pathogenicity (see below).

Table 3.1. SDPs that are likely to alter Reston virus protein structure and function**.**

Protein	SDP	Interface	Protein Integrity
VP24	T131S	KPNA5 interface	
VP24	M136L	KPNA5 interface	
VP24	Q139R	KPNA5 interface	
VP24	T226A		Loss of Hydrogen bond
VP40	P85T	Octamer interface	
VP40	Q245P		Breaks α helix
VP30	R262A	Dimer interface – loss of Hydrogen bond	
VP35	E269D	Dimer interface	

3.3.3 Multiple SDPs are present in the GP glycan cap

GP is highly glycosylated and mediates Ebolavirus host cell entry. Subunit GP1 binds to the host cell receptor(s). Subunit GP2 is responsible for the fusion of viral and host cell membranes. However, their cellular binding partners remain to be defined (Feldmann and Geisbert, 2011; Miller et al., 2012; Dahlmann et al., 2015; Herbert et al., 2015). Reverse genetics experiments have suggested that GP contributes to human pathogenicity but is insufficient for virulence on its own (Groseth et al., 2012). We identified SDPs in both GP1 and GP2 (Supplementary Figure 4 and Supplementary Table 12). Three SDPs (I260L, T269S, S307H) are located in the glycan cap that contacts the host cell membrane (Supplementary Figure 4B-C). These changes (particularly S307H at the top of the glycan cap) alter the electrostatic surface of GP (Supplementary Figure 4D) and may therefore alter GP interactions with cellular proteins, however given the glycosylation of GP, it is unlikely that these residues would physically contact the host cell membrane and none of them are near glycosylation sites. So it is not clear what role they may have. GP binding to the endosomal membrane protein NPC1 is necessary for membrane fusion (Miller et al., 2012). However, residues important for NPC1 binding (identified by mutagenesis studies in Miller et al., 2012) were conserved in all analyzed Ebolaviruses and the SDPs were not located close to them (Supplementary Figure 5). Thus differences in NPC1 binding do not account for differences in Ebolavirus human pathogenicity. This finding is in concert with very recent data indicating that NPC1 is essential for Ebolavirus replication as NPC1-deficient mice were insusceptible to Ebolavirus infection (Herbert et al., 2015).

It was not possible to predict the consequences of SDPs in sGP and ssGP (Fig. S23), as there is a lack of functional information available for these proteins (Miranda and Miranda, 2011; Mehedi et al., 2011). A 17 amino acid peptide derived from Ebola virus or Sudan virus GP exerted immunosuppressive effects on human CD4+ T cells and CD8+ T cells while the respective Reston virus peptide did not (Yaddanapudi et al., 2006). We identified one SDP in the peptide, which represents the single amino acid change (I604L) previously observed between Reston virus and Ebola virus

(Yaddanapudi et al., 2006), demonstrating that this difference is conserved between Reston viruses and all human pathogenic Ebolaviruses.

3.3.4 Changes in the VP30 dimer may affect pathogenicity

Analysis of the VP30 SDPs provided novel mechanistic insights into the structural differences previously observed between Reston virus and Ebola virus VP30 (Clifton et al., 2014) and that may contribute to the differences observed in human pathogenicity between Reston virus and Ebola virus. VP30 is an essential transcriptional co-factor that forms dimers via its C-terminal domain and hexamers via an oligomerization domain (residues 94-112) (Hartlieb et al., 2003). The VP30 hexamers activate transcription while the dimers do not, and the balance of hexamers and dimers has been suggested to control the balance between transcription and replication (Hartlieb et al., 2007). Crystallization studies have shown that Ebola virus and Reston virus dimers are rotated relative to each other (Clifton et al., 2014). We observed two SDPs (T150I, R262A) in the dimer interface that can at least partially explain the structural differences between Ebola virus and Reston virus VP30 dimers. Ebola virus R262 is part of the dimer interface and forms a hydrogen bond with the backbone of residue 141 in the other subunit, whereas Reston A262 does not and is not part of the dimer interface (Figure 3.3). The removal of the two hydrogen bonds (in the symmetrical dimer) is likely to lead to the different Reston and Ebola virus dimer structures. mCSM predicts this change to be destabilizing with a ΔΔG-0.969 Kcal/mol. The Reston virus conformation also buries functional residues A179 and K180 potentially affecting protein function (Clifton et al., 2014) (Figure 3.2). Moreover, our findings show that the Ebola virus conformation is conserved in all human-pathogenic Ebolaviruses suggesting that it is relevant for human pathogenicity.

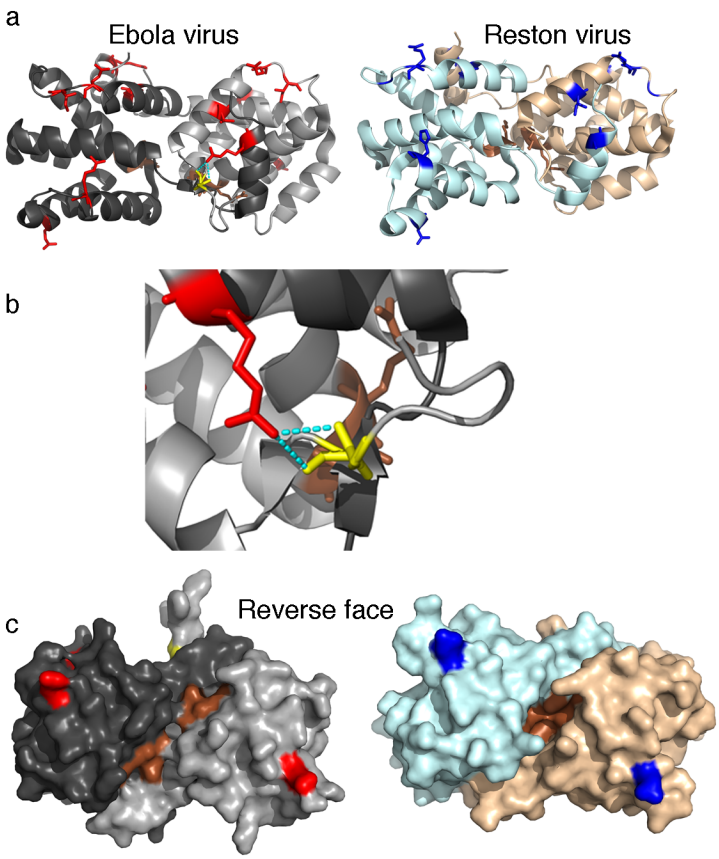


Figure 3.3. SDPs present in the VP30 dimer. The dimer structure of both Ebola virus (PDB structure 2I8B) and Reston virus (PDB structure 3V7O) VP30 are shown with SDPs indicated (red – Ebola virus, blue – Reston virus) and functional residues (brown – A179, K180). a) Cartoon representation: For the Ebola virus the hydrogen bond of R262 with the residue 141 of the other subunit is shown. b) enlarged display of the hydrogen bond between R262 and the backbone of residue 141. c) Surface representation of the reverse face of the dimer from A, showing the location of the functional residues A179 and K180 within the dimer.

3.3.5 VP35 SDP present in dimer interface

VP35 is a multifunctional protein that antagonizes interferon signaling by binding double stranded RNA (dsRNA). Structural data are available for both the Ebola virus and Reston virus VP35 monomer and an asymmetric dsRNA bound dimer (Bale et

al., 2013; Leung et al., 2010; Leung et al., 2009; Kimberlin et al., 2010). These structures are highly conserved, however functional studies have demonstrated that Reston virus VP35 is more stable, has a reduced affinity for dsRNA, and exerts weaker effects on interferon signaling (Leung et al., 2010). The increased stability is proposed to be due to a linker between the two subdomains having a short alpha helix in the Reston virus structure (Leung et al., 2010). Our analysis shows that the sequence of this linker region is completely conserved in all of the genomes, however an SDP is located close to the linker (A290V). One SDP (E269D) is present in the dimer interface and the shorter aspartate side chain in Reston virus VP35 results in increased distances with the atoms that this aspartate forms hydrogen bonds with: R312, R322, and W324 (Ebola virus numbering; Supplementary Table 13). mCSM predicts this change to be slightly destabilizing to the complex (ΔΔG -0.11Kcal/mol). This has the potential to alter the stability of the dimer and thus the ability of VP35 to prevent interferon signaling.

It has recently been demonstrated that a VP35 peptide binds NP and modulates NP oligomerization and RNA binding to NP (Leung et al., 2015). There are two SDPs (S26T, E48D) in this region. S26T is located on the periphery of the interface. E48D lies outside the solved structure but is within the region required for binding to NP. Both SDPs represent minor changes that maintain the chemical properties of the side chains. Thus, there is no evidence suggesting substantial differences in the binding of this peptide to NP.

3.3.6 VP40 SDPs may alter oligomeric structure

VP40 exists in three known oligomeric forms (Bornholdt et al., 2013). Dimeric VP40 is responsible for VP40 trafficking to the cellular membrane. Hexameric VP40 is essential for budding and forms a filamentous matrix structure. Octameric VP40 regulates viral transcription by binding RNA. Two SDPs (P85T and Q245P) can affect VP40 structure. P85T occurs at the VP40 octamer interface site (Figure 3.4) in the middle of a run of 14 residues that are completely conserved in all Ebolaviruses (Figure 3.4a). In the Ebola virus structure, it is located in an S-G-P-K beta-turn,

where the proline at position 85 (P85) confers backbone rigidity. The change to threonine (T) at this residue in Reston viruses introduces backbone flexibility and also provides a side chain with a hydrogen bond donor, potentially affecting octamer structure and/or formation. mCSM predicted this change to have a destabilizing effect (ΔΔG -0.626Kcal/mol). The Q245P SDP introduces a proline residue into an alpha helix (Figure 3.4B), which most likely breaks and shortens helix five, resulting in the destabilization of helices five and six and a change in the hydrophobic core. Interestingly mCSM predicted this change to have little effect on the stability of the protein (predicted ΔΔG 0.059Kcal/mol). Thus, P85T and Q245P may affect VP40 function and human pathogenicity.

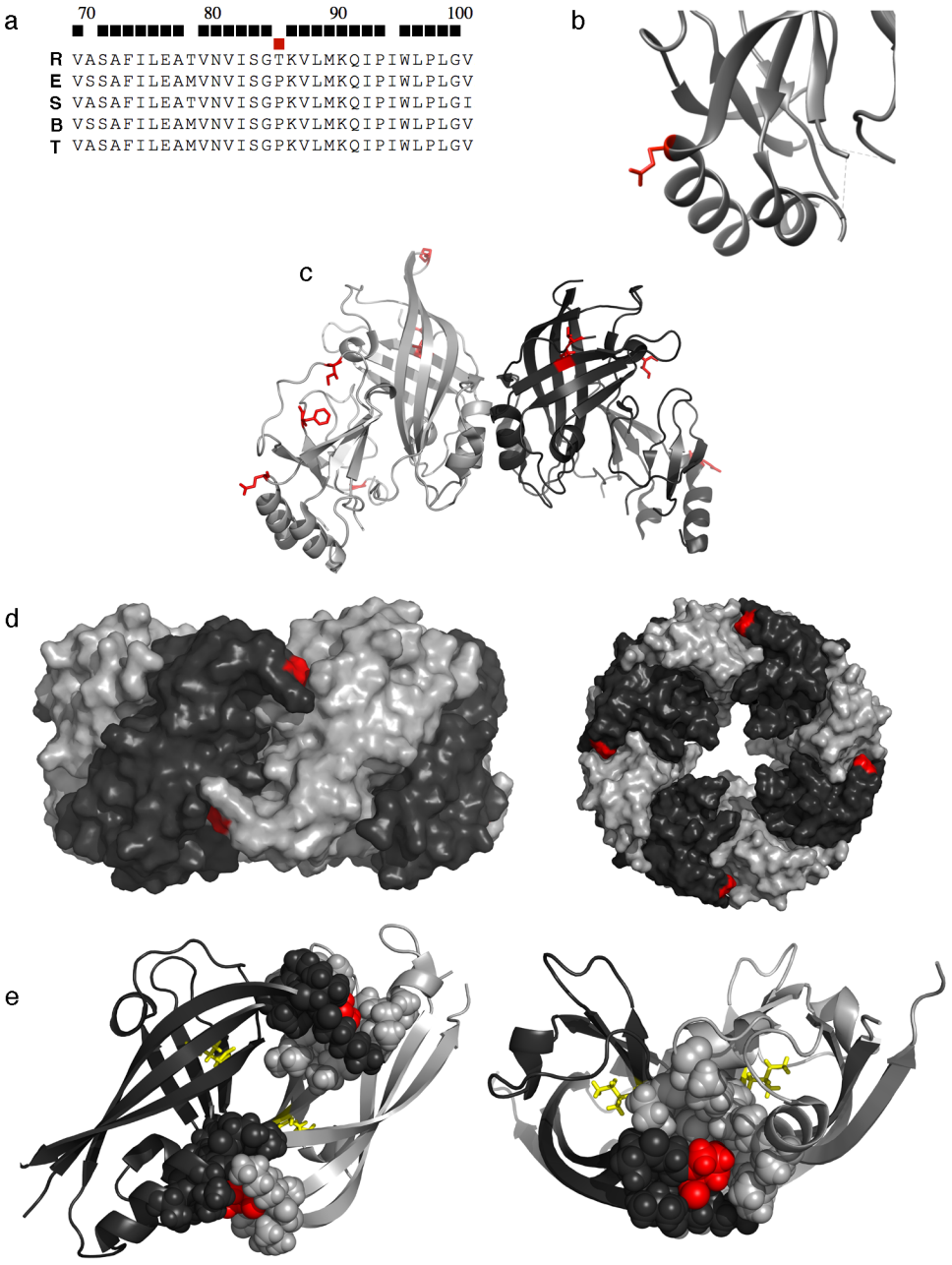


Figure 3.4. The P85T SDP is present in the VP40 octamer interface. a) Consensus sequence for the region around P85T in *Ebolavirus* species (R, Reston virus; E, Ebola virus; S, Sudan virus; B, Bundibugyo virus; T, Taï Forest virus). Black squares indicate positions that are completely conserved in all genomes, red squares SDPs. b) segment of VP40 showing the Q245P SDP (red) from PDB structure 1ES6. c) The VP40 dimer, with SDPs colored red and shown in stick format (PDB structure

4LDB). d) The VP40 octamer, P85 shown in red (side- and top-view) from PDB structure 4LDM. e) Two subunits from the VP40 octamer, P85 is colored red in sphere format, and the SDP I122V is shown as yellow in stick format.

3.3.7 VP24 SDPs affect KPNA5 binding

VP24 is involved in the formation of the viral nucleocapsid and the regulation of virus replication (Feldmann and Geisbert 2011; Morikawa et al., 2007;Mateo et al., 2011; Mateo et al., 2011; Watt et al., 2014). VP24 also interferes with interferon signaling through binding of the karyopherins α1 (KPNA1), α5, (KPNA5), and α6 (KPNA6) and subsequent inhibition of nuclear accumulation of phosphorylated STAT1 and through direct interaction with STAT1 (Xu et al., 2014; Reid et al., 2006; Reid et al., 2007; Zhang et al., 2012). Eight VP24 SDPs are in regions with available structural information (Supplementary Tables 17-18). Seven of these are present on the same face of VP24 (Figure 3.5A) suggesting that they affect VP24 interaction with viral and/or host cell binding partners. The SDPs T131S, M136L, and Q139R are present in the KPNA5 binding site (Figure 3.5). M136 and Q139 are part of multi-residue mutations in Ebola virus VP24 that removed KPNA5 interactions (Supplementary Table 17) (Xu et al., 2014) and are adjacent to K142 (Figure 3.5A), mutants of which have shown reduced interferon antagonism (Llinykh et al., 2015). Xu et al., investigated the effect of VP24 mutations on binding to KPNA5 using coimmunoprecipitation pull down experiments and compared the bands obtained in the gel with wild type protein. This approach is not quantitative but the strength of the band provides an indication of the extent to which binding is affected. For R137A and R137A, T138A,Q139A the band is very weak. For F134A/M135A it is intermediate between these previous two mutations and the wild type. Therefore, M136L and Q139R can exert significant effects on VP24-KPNA5 binding. Additionally, T226A results in the loss of a hydrogen bond between T226 and D48 in Reston virus VP24 (Figure 3.5B), with the potential to alter structural integrity and influence protein function. Analysis using mCSM predicts the T226A change to be destabilizing with a ΔΔG -0.935 Kcal/mol. mCSM predicted seven of the eight analysed SDPs to be destabilizing (Supplementary Table 2).

VP24-mediated inhibition of interferon signaling may be critical for species-specific pathogenicity (Xu et al., 2014; Mateo et al., 2011; Reid et al., 2006; Reid et al., 2007; Zhang et al., 2012). In this context, VP24 was a critical determinant of pathogenicity in studies in which Ebola viruses were adapted to mice and guinea pigs that are normally insusceptible to Ebola virus disease(La Vega et al., 2015; Mateo et al., 2011; Volchkov et al., 2000; Ebihara et al., 2006; Dowall et al., 2014). The adaptation-associated VP24 mutations in rodents are located in the KPNA5 binding site with some of them being very close to the VP24 SDPs T131S, M136L, and Q139R that we determined to be in the KPNA5 binding site (Figure 3.5C-D, Supplementary Table 19). Additionally some of the mutations are similar to the SDPs in that they would remove hydrogen bonds within VP24 (e.g. T187I, T50I, Figure 3.5E-F, & Supplementary Table 19) or alter hydrogen bonding with KPNA5 (H186Y, Figure 5F & Supplementary Table 19). Thus there is strong evidence suggesting that the VP24 SDPs have a role in rendering the Reston virus non-pathogenic in humans.

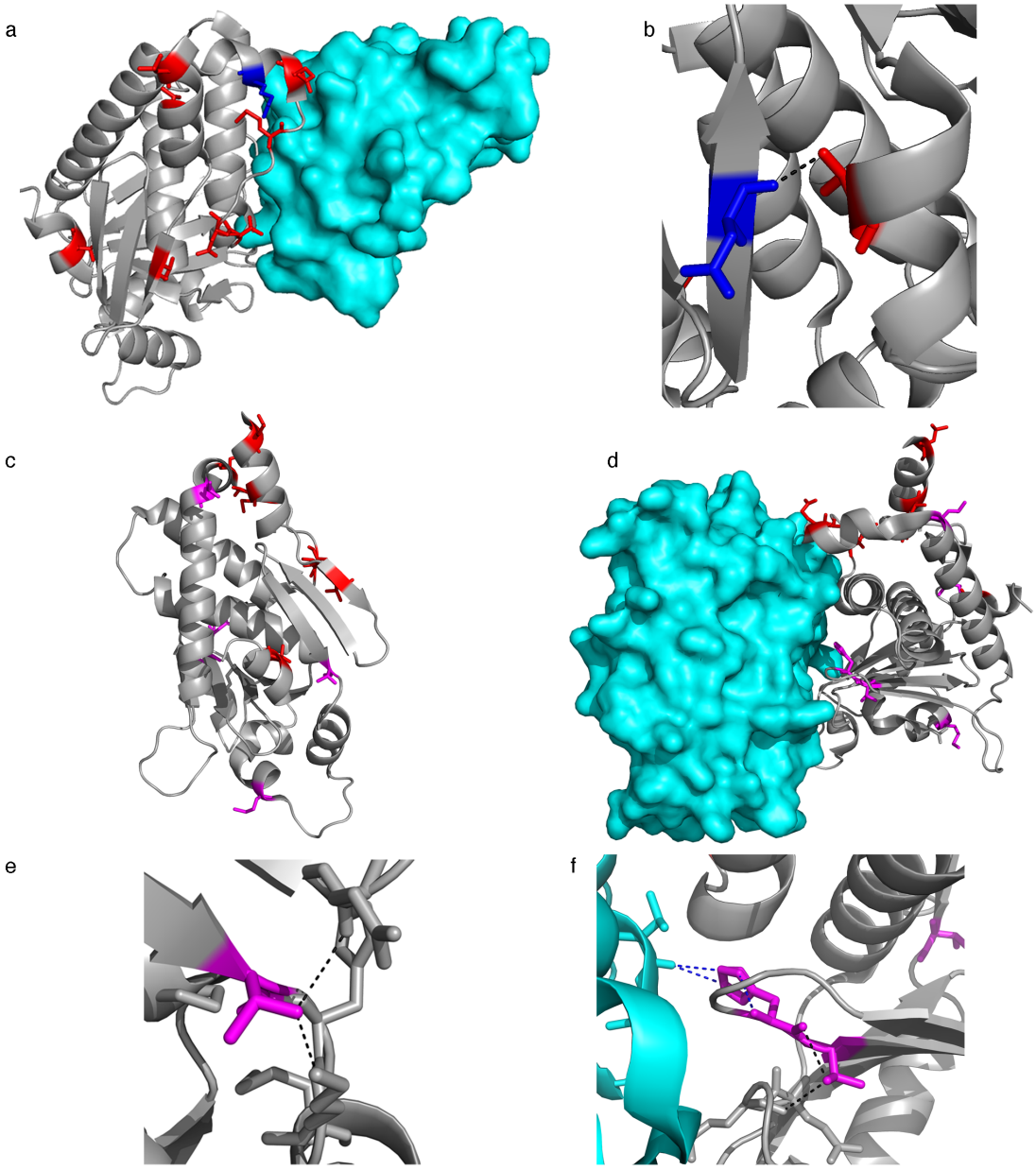


Figure 3.5. Ebola virus VP24 SDPs and complex with KPNA5. a). VP24 Structure (grey) in complex with KPNA5 (cyan) (PDB structure: 4U2X), with VP24 SDPs (red) and K142 colored blue. b) T226 (red) hydrogen bond with the backbone of D48 (blue). c) VP24 showing residues mutated in rodent adaptation experiments (magenta) and SDPs identified in this study (red). d) Ebola virus VP24 in complex with KPNA5, reverse view shown from A. SDPs are coloured red and residues mutated in adaptation experiments are coloured magenta; VP24 (grey) and KPNA5 (cyan) complex with residues mutated during adaptation (magenta) and SDPs (red). F) Hydrogen bonds formed by VP24 T50. G) Hydrogen bonds formed by VP24 H186, and T187. Intrachain bonds are colored black and hydrogen bonds between VP24 and KPNA5 are colored blue.

3.4 Discussion

In this study, we have combined the computational identification of residues that distinguish Reston viruses from human pathogenic *Ebolavirus* species with protein structural analysis to identify determinants of Ebolavirus pathogenicity. The results from this first comprehensive comparison of all available genomic information on Reston viruses and human pathogenic Ebolaviruses detected SDPs in all proteins but only few of them may be responsible for the lack of Reston virus human pathogenicity.

Our analysis mapped 47 of the 189 SDPs onto protein structure, so additional SDPs may be relevant but the structural data needed to reliably identify them is missing. Although it is difficult to conclude the extent to which each individual SDP contributes to the differences in human pathogenicity between Reston viruses and the other Ebolaviruses, we can identify certain SDPs that have a particularly high likelihood to be involved. SDPs present in the oligomer interfaces of VP30, VP35, and VP40 may affect viral protein function. VP24 SDPs may interfere with VP24-KPNA5 binding and affect viral inhibition of the host cell interferon response. These findings suggest that changes in protein-protein interactions represent a central cause for the variations in human pathogenicity observed in Ebolaviruses. VP24 and VP40 in particular contain multiple SDPs that are likely to contribute to differences in human pathogenicity. Where possible the SDPs have been considered collectively, such as for VP24, where most of the SDPs are present on a single face of the protein (Figure 3.5A) and three of them are present in the interface with KPNA5. Beyond this it is difficult to interpret how any combination of SDPs might be responsible for the differences in human pathogenicity.

Our data also demonstrate that relevant changes explaining differences in virulence between closely related viruses can be identified by computational analysis of protein sequence and structure. Such computational studies are particularly important for the investigation of Risk Group 4 pathogens like Ebolaviruses whose investigation is limited by the availability of appropriate containment laboratories.

The role of VP24 appears to be central given the large number of SDPs we identify as likely to affect function, particularly KPNA5 binding. This is also highlighted by the similarity between these SDPs and the mutations that occur in adaptation experiments in mice and guinea pigs (Basler, 2014; Leung et al., 2009; Watt et al., 2014; Reid et al., 2006; Reid et al., 2007). Consequently, the mutation of a few VP24 SDPs could result in a human pathogenic Reston virus. Given that Reston viruses circulate in domestic pigs, can be spread by asymptomatically infected pigs, and can be transmitted from pigs to humans (possibly by air) (Weingartl et al., 2013; Barrette et al., 2009; Marsh et al., 2011), there is a concern that (a potentially airborne) human pathogenic Reston viruses may emerge and pose a significant health risk to humans. Notably, asymptomatic Ebolavirus infections have also been described in dogs (Weingartl et al., 2013) and Ebola virus shedding was found in an asymptomatic woman (Akerlund et al., 2015). Thus, there may be further unanticipated routes by which Reston viruses may spread in domestic animals and/or humans enabling them to adapt and cause disease in humans.

In summary our combined computational and structural analysis of a large set of Ebolavirus genomes has identified amino acid changes that are likely to have a crucial role in altering Ebolavirus pathogenicity. In particular the differences in VP24 together with the observation that Ebolavirus adaptation to originally non-susceptible rodents results in rodent pathogenic viruses (Basler, 2014; Leung et al., 2009; Watt et al., 2014; Reid et al., 2006; Reid et al., 2007) suggest that a few mutations could lead to a human pathogenic Reston virus.

3.5 Materials and methods

3.5.1 Ebolavirus nomenclature

The nomenclature in this manuscript follows the recommendations of Kuhn et al., (2010). The genus is *Ebolavirus*. It is only italicized if the name refers to the genus but not if it refers to physical viruses or virus parts or constituents such as proteins or genomes. The species are *Zaire ebolavirus* (type virus: Ebola virus, EBOV), *Sudan*

ebolavirus (type virus: Sudan virus, SUDV), *Bundibugyo ebolavirus* (type virus: Bundigugyo virus, BDBV), and *Taï Forest ebolavirus* (formerly Côte d'Ivoire ebolavirus; type virus: Taï Forest virus, TAFV).

3.5.2 Ebolavirus Genome Sequences

196 complete *Ebolavirus* genomes were downloaded from Virus Pathogen Resource, VIPR (http://www.viprbrc.org/brc/home.spg?decorator=vipr) (Pickett et al., 2012). The 196 genomes comprise 156 Ebola virus (EBOV), 17 Reston (RESTV), 13 Sudan (SUDV), 7 Bundibugyo (BDBV) and 3 Taï Forest (TAFV) species (Supplementary Table 20). Open Reading Frames (ORFs) in the genomes were identified using EMBOSS (Rice et al., 2000). The ORFs were then mapped to the nine Ebolavirus proteins.

3.5.3 Multiple Sequence Alignments and identification of specificity determining positions

Multiple sequence alignments were generated for each of the Ebolavirus proteins using Clustal Omega (Sievers et al., 2011), with default settings. Protein sequence identities between the different sequences were obtained from the Clustal Omega output. The effective number of independent sequences (or effective number of sequences, see table S21) in an alignment indicates given redundancy in the sequences, how many different sequences there are effectively. So if all of the sequences are highly similar, there is little diversity in the alignment and the effective number of sequences is low. The effective number of independent sequences present was calculated for the alignment for each protein by building an HMM for the alignment using hmmer (Mistry et al., 2013). The effective number of independent sequences identified ranged from 88 for the VP24 and L proteins to 148 in NP (Table S21).

The s3det algorithm (Rausell et al., 2010) was used to predict specificity determining positions (SDPs) using a supervised mode with sequences assigned to predetermined groups/subfamilies with all of the human pathogenic sequences in one group and the

Reston virus sequences in a second group. The sensitivity of the SDP analysis to the number of sequences used was considered by subsampling the sequences (see Supplementary Methods and Supplementary Figs S6-S8). SDPs were compared to known functional residues (many from mutagenesis studies) in Ebolavirus proteins catalogued in UniProt (Uniprot Consortium, 2014) and in the literature.

3.5.4 Phylogenetic Trees

Bayesian Phylogenetic trees were generated using BEAST v1.8.2 (Bouckaert et al., 2014), then the consensus tree for each set of 10000 trees was calculated with TreeAnnotator and the node labels obtained analyzing the trees with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). TreeAnnotator and BEAUti, are part of the BEAST package.

The Maximum Likelihood Phylogenetic trees were generated using RaxML8 (Stamatakis, 2014). A full Maximum Likelihood analysis and 1000 Bootstrap replicate searches were run in order to obtain the best scoring ML tree for each set of sequences.

Phylogenetic trees were generated using default settings in both BEAST and RaxML8, according to the type of input data. All phylogenetic trees were analyzed and plotted using the R "ape" package (Paradis et al., 2004).

3.5.5 Structural Analysis

Where available, protein structures for the Ebolavirus proteins were obtained from the protein databank (Rose et al., 2015). Where full length protein structures were not available the proteins were modelled using Phyre2 (Kelley et al., 2015). SDPs were mapped onto the protein structures using PyMOL. Solvent accessibility for SDPs was calculated using DSSP (Joosten et al., 2011).

The Reston virus structures of GP1 and GP2 were modeled using one-to-one threading in Phyre2 (Kelley et al., 2015) with the EBOV GP trimer structure (PDB

code 3CSY) used as a template. A model of a Reston virus GP trimer structure was generated by aligning the modelled Reston virus GP1 and GP2 structures to their corresponding chains in the Ebola virus trimer.

The Coulombic Electrostatic Potential for the proteins was calculated using Delphi, with default parameters (Smith et al., 2012). The electrostatics map was visualized and analyzed using Chimera (Pettersen et al., 2004).

mCSM (Pires et al., 2014) was used to predict the effect of each individual SDP on the stability of the protein. The Ebola virus structures were used as input and the relevant amino acid changed to the one present in the Reston virus.

Chapter 4:

Structural consequences of the genomic changes associated with Ebola virus adaptation to rodents

Morena Pappalardo, Mark J Howard, Jeremy S Rossman, Martin Michaelis, Mark N Wass

This manuscript is currently in preparation for submission to *Genome Biology*. In this project I have performed the research, which is primarily protein structural analysis of mutations that occur in adaptation of Ebola virus to rodents. Interpretation of the likely effects of the mutations was performed in discussion with my supervisor.

1. Abstract

The potential for Ebola virus to cause large outbreaks and many thousands of deaths has recently been demonstrated in West Africa. Ebola viruses are pathogenic in humans and primates but not in rodents. We have analysed the mutations identified in four different experiments that adapted Ebola virus to rodents to identify and understand the molecular determinants of host-specific Ebola virus pathogenicity. We identified 33 different mutations across the four studies, with only two mutations present in more than one study. For three proteins, VP24, GP and NP, mutations were observed in all four studies. Structural analysis suggests that the changes in GP and NP may have an effect on protein function but with limited functional knowledge of the regions of the protein they are located in, it is not possible to infer further. Clear functional effects were identified for six of the seven mutations present in VP24. Three of these mutations are located in the VP24 interface with karyopheerin a5 and we propose that they may have a role in adapting Ebola VP24 binding to karyopherins from novel hosts. A further three mutations either change hydrogen bonding or will result in conformational changes in the protein. Based on our analysis we propose that VP24 is central to adaptation of Ebola virus to new hosts.

4.2. Introduction

The recent Ebola virus outbreak in West Africa, which is still seeing flare-ups in infection, http://www.who.int/ was the first outbreak of a member of the *Ebolavirus* family in humans that reached epidemic size (Frieden, *et al.*, 2014; Alexander, *et al.*, 2015). It has resulted so far in 28,639 confirmed cases and 11,316 deaths as of 28th February 2016 (www.who.int), though these figures are thought to underestimate the actual numbers (Meltzer, *et al.*, 2014). Hence, this epidemic provided the first evidence that Ebolaviruses can sustainably spread among humans and cause large outbreaks that affect tens of thousands of individuals, possibly even more.

The research on Ebola viruses is limited by the availability of safety level 4 laboratories and a lack of disease models in small rodents. A major issue in the

establishment of rodent models is that species including mice, guinea pigs, and hamsters are generally not susceptible to Ebola virus infection and disease. Therefore, Ebola viruses that cause lethal disease in rodents need to be established by virus adaptation via serial passaging in these species (Shurtleff & Bavari, 2015). Despite indications that these models reflect human disease at least in part, a better understanding of the similarities and differences between natural Ebola virus disease in humans and the disease caused by rodent-adapted Ebola virus strains in rodents is needed (Shurtleff & Bavari, 2015; Cross *et al.*, 2015).

A number of studies reported on the genetic changes associated with Ebola virus strains to mice, guinea pigs, and hamsters (Ebihara *et al.*, 2006; Dowall *et al.*, 2014; Cross, *et al.*, 2015; Volchkov, *et al.*, 2000). Here, we applied an *in silico* approach to predict the consequences of these sequence changes in the virus genome on the structure and function of the Ebola virus-encoded proteins in order to improve our understanding of the processes, underlying Ebola virus adaptation to rodents and to gain further insights into the differences of Ebola virus replication in experimental rodent models relative to natural human infection.

4.3. Results

We focus our analysis on four studies that adapted Ebola virus in rodents. Three of them adapted Ebola to guinea pigs (Dowall, *et al.*, 2014; Volchkov *et al.*, 2000; Cross *et al.*, 2015) and one in mice (Ebihara, *et al.*, 2006). In each study multiple passaging of the virus in the rodent species was performed, three of the studies sequenced the virus once it had become pathogenic, while Dowall et al., (Dowall, *et al.*, 2014) sequenced the virus after each passage, thus providing greater detail on the mutations occurring during the adaptation process and the ability to identify whether they are lost or retained during passaging.

Ebolaviruses have a small genome containing seven genes that encode nine proteins. The proteins are glycoprotein (GP), soluble GP (sGP), small soluble GP (ssGP), RNA dependent RNA polymerase (L), nucleoprotein (NP), and four structural proteins that are called VP24, VP30, VP35 and VP40. Therefore, there are a small

number of proteins to investigate for having a role in determining host pathogenicity. However, given the small size of the genome, most Ebolavirus proteins are multifunctional, which may make it difficult to identify the functional effects of individual mutations.

Table 4.1 Summary of mutations identified in Ebola virus rodent adaptation experiments. *L26F is present in two studies so total number of unique mutations is 5 for VP24. Two different adaptations experiments were performed in Volchkov et al., and these are listed separately in the table. §data is only available for VP24 mutations in Volchkov-2.

	Ebihara	Dowall	Volchkov-1	Volchkov-2	Cross	Total
NP	1	1	1	N/A	2	5
GP	3	2	1	N/A	1	7
L	1	11	1	N/A	0	13
VP24	1	1	3	1	2	7*
VP30	0	0	0	N/A	0	0
VP35	1	1	0	N/A	0	2
VP40	0	0	0	N/A	0	0
Total	7	16	6	1§	5	32

4.3.1. Initial comparison of the different adaptation experiments

Over the four studies 33 unique protein coding mutations were identified in the rodent adapted Ebola virus genomes. In all four studies mutations were present in multiple proteins (*Table 4.1*), with mutations in the glycoprotein (GP), nucleoprotein (NP), the RNA dependent RNA polymerase (L) and viral protein 24 (VP24) in each of the separate studies (*Table 4.1*). Mutations in VP35 were observed in two studies. No mutations were observed in the remaining proteins, VP30 and VP40, although mutations were present in both VP30 and VP40 (as well as the other Ebolavirus proteins) during passaging in the Dowall study but these mutations were not retained in later passages (Dowall, Matthews, *et al.*, 2014).

Only two mutations were observed in multiple studies (GP I554T and VP24 L26F), which may provide stronger evidence for a role of these mutations in the adaptation process. The GP I554T mutation was observed in both the Ebihara et al.,(Ebihara, *et al.*, 2006) and Cross et al., (Cross, *et al.*, 2015) studies, while VP24 L26F was observed in the Dowall et al., (Dowall, *et al.*, 2014) and Cross et al., (Cross, *et al.*, 2015) studies. Further investigation revealed that threonine is commonly observed at residue 544 in GP (see methods), while isoleucine is present in the original Mayinga strain. Therefore, it seems unlikely that this mutation is relevant to adaptation in guinea pigs. For VP24 L26F reverse genetics studies have associated the mutation with increased virulence in rodents (Mateo, *et al.*, 2011).

Overall analysis of the four studies suggests that only a small number of mutations are required to adapt Ebola virus to rodents (*Table 1*), with six mutations present in the Volchkov and Cross studies, seven in the Ebihara study and 16 in the Dowall study (most of these in L). However, without further analysis it is not clear if all of these mutations play a role in the adaptation process or if there are a few specific mutations present in each study that are responsible for the change in pathogenicity. Nor is it apparent if there is a single adaptation mechanism (i.e. mutation to a particular protein or set of proteins) or if there are multiple different pathways to pathogenicity.

To gain insight into this set of mutations we performed a structural analysis, mapping the mutations onto the available Ebola virus proteins and complexes, supplemented with structural modelling where structures were not available (see methods). The potential structural effects of the mutations were manually investigated and additionally their effect on protein stability predicted using mCSM (Pires, 2014), a computational method designed to predict the effect of point mutations on protein structure and stability. Our structural analysis was performed to investigate the mutations present in all four studies 1) to identify structural elements that are most relevant to the development of Ebola virus pathogenicity in a new host and 2) to estimate how easily Ebolaviruses may adapt to new hosts.

In total 22 of the 33 mutations were mapped onto protein structures or models. Neither of the two mutations (A12V, N204D) present in VP35 could be modelled. The A12V mutation appears to be a conservative change of amino acid and is located in the N terminal dimerisation domain, while the N204D mutation is located just before the RNA binding domain. Notably, VP35 was only found mutated in two out of four studies suggesting that mutations in VP35 are not essential for Ebolavirus adaptation to a novel species.

4.3.2 Mutations in the glycoprotein may affect protein structure

The glycoprotein (GP) mediates host cell entry and has long been speculated to have a role in pathogenicity (Feldmann & Geisbert, 2011). GP consists of two subunits: GP1 binds to the host cell surface receptor(s). GP2 is needed for the fusion of the virus membrane with the host cell membrane. The exact process and host cell binding partners during virus binding and membrane fusion remain only partially understood (Miller, *et al.*, 2012). However, GP binding to the endosomal membrane protein NPC1 appears to be required for membrane fusion (Miller, *et al.*, 2012). Across the four studies six different mutations are observed in GP (*Table 4.2*). Four of these mutations could be mapped onto available GP structures (*Figure 4.1*). The most striking mutation is S65P. S65 is a buried residue. The mutation S65P introduces a proline into the middle of a beta sheet, this is likely to alter or disrupt the beta sheet and it will also result in the loss of a hydrogen bond with E100 (*Figure 4.1*). Both of these effects are likely to result in conformational change within GP. However, the extent of the conformational change, how it would affect GP function, or how it may have a role in adaptation remain unclear from the structural analysis. The second mutation D49N is located at the edge of the interface between GP1 and GP2. The D49 side chain is not present in the crystal structure suggesting that the side chain is moving. Analysis of the possible side chain conformations indicated that it could form a hydrogen bond with N595. However, mutation would reduce the charge and asparagine would still enable a hydrogen bond to be formed between the subunits. So it may be that a hydrogen bond is formed with asparagine at position 49 but not aspartate. However, it is not clear what functional effect this change would have. The third mutation S246P is located on a surface loop towards the area of the

protein that binds the host cell membrane, so it is possible that this mutation could alter host cell interactions but without knowledge of the receptor and binding site, there is no evidence to support this. Finally, GP is heavily glycosylated (Lennemann et al., 2014; 2015), which further aggravates the interpretation of the functional consequences of mutations in GP.

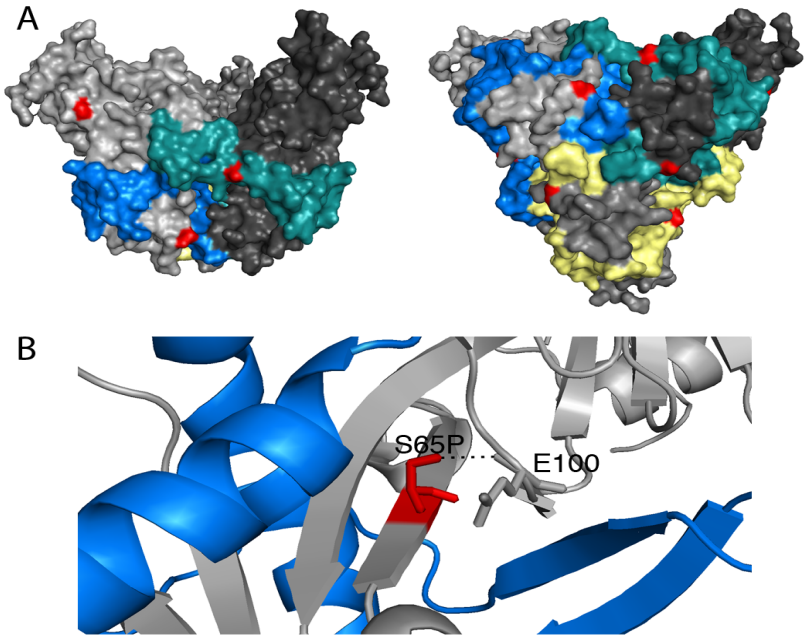


Figure 4.1. Mutations in GP during adaptation to rodents. The GP trimer consists of GP1 (grey colours) and GP2 (blue, yellow, green) dimers. A) Adaptation mutations in GP are shown in red. B) The adaptation mutation S65P will result in loss of a hydrogen bond with the backbone of E100.

4.3.3. Mutations present in the nucleoprotein

Three of the five mutations present in NP could be mapped onto the protein structure (*Figure 4.2*). Adjacent residues S647 and F648 (mutations: S647Y and

F648L) in the C terminal domain are mutated in separate studies, suggesting that either this is a region that can tolerate mutations or that the mutations could be linked to adaptation to the new host. F648 is tightly packed with side chains from the adjacent alpha helix (*Figure 4.2*). The change to leucine will reduce the size of the side chain and could result in local conformational change. S647 is located on the protein surface, the mutation to tyrosine results in a large increase in side chain size but retains the ability to form hydrogen bonds (possibly with interaction partners).

The third mutation (S72G) in NP is located in the N terminal domain. S72 forms a hydrogen bond with the backbone of P42, which is lost on mutation of S72 to glycine (*Figure 4.2B*). This may result in increased flexibility in this region but the functional consequences cannot be reliably predicted.

The function of these regions of NP are not well established, making it difficult to interpret the possible effect they may have on protein structure and function and how this may relate to Ebola virus pathogenicity.

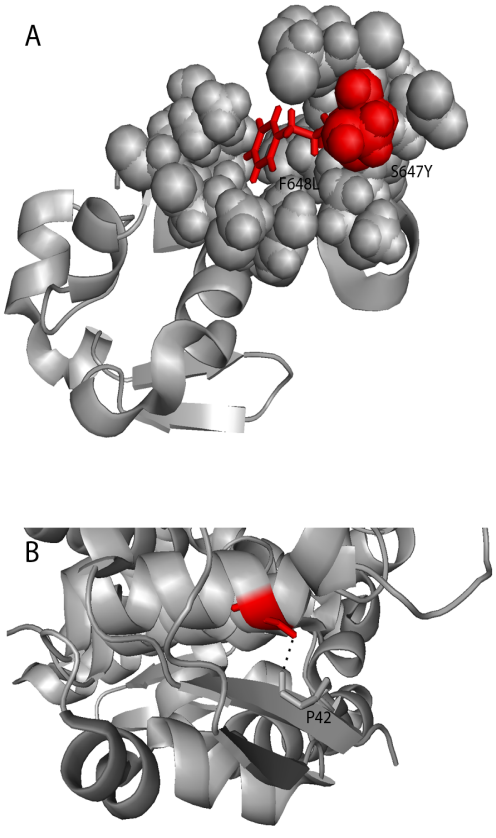


Figure 4.2. Adaptation mutations in NP. A) Adaptation mutations S647Y (red spheres) and F648L (red sticks) in the C terminal domain of NP occurred in separate adaptation mutations. B) NP residue S72 forms a hydrogen bond with the backbone of P42 (black dashed line). This bond is lost with the adaptation mutation S72G.

4.3.4. Mutations in the RNA dependent RNA polymerase may not be related to pathogenicity

Thirteen mutations were reported in the RNA dependent RNA polymerase (L) from three of four studies (*Table 4.1*), 11 of them from the Dowall et al., study (Dowall, *et al.*, 2014). This study monitored the mutations that occurred in every passage until the virus had adapted to Guinea pigs and caused disease. Notably, 10 out of these 11 mutations were only identified in the final passage, whereas mutations in NP, VP35, and GP had become visible within the first three passages. Thus, it remains unclear whether these mutations would have been maintained during further replication cycles in Guinea pigs. In this context, as only three out of four adaptation studies reported mutations in L does not suggest an essential role of L in Ebolavirus host tropism. Additionally, the Y1271STOP mutation results in a stop codon and, hence, in a truncated protein, that is unlikely to be functional (full length L is 2212 residue so long so nearly half the protein would be missing). This mutation is therefore unlikely to be associated with enhanced pathogenicity and further questions a pivotal role of L for Ebolavirus adaptation to a novel species.

4.3.5. Multiple mutations in VP24 are likely to be associated with Ebola virus pathogenicity

VP24 is multifunctional and is involved in the formation of the viral nucleocapsid, the regulation of virus replication and the prevention of interferon signalling (Feldmann & Geisbert, 2011; Mateo, *et al.*, 2011; Watt, *et al.*, 2014; Reid, Leung, Hartman, *et al.*, 2006). VP24 interferes with interferon signalling through binding of STAT1 and the karyopherins α1 (KPNA1), α5, (KPNA5), and α6 (KPNA6) (Xu, *et al.*, 2014). This binding prevents nuclear accumulation of phosphorylated STAT1 and therefore inhibits interferon signalling.

Changes in the sequence of VP24 were detected in all of the studies that investigated the genomic consequences associated with Ebola virus adaptation to rodents (*Table 4.1*). VP24 may need to adapt to interfere with STAT1 and/or the karyopherins of a novel species. Structural analysis using the complex of VP24 with human KPNA5 provided insight into the likely effects of six of the seven VP24 mutations found in rodent-adapted Ebola virus strains. Only the possible consequences of the M71I mutation remained elusive. Three mutated residues (H186Y, T187I, K142E) are present in or adjacent to the interface site with human KPNA5 (*Figure 4.3*). Hence, it is possible that these mutations enable or alter the interaction of VP24 with rodent karyopherins. The wild type H186 forms a hydrogen bond in the interface with residue T434 in human KPNA5 (*Figure 4.3B*). The hydroxyl group in the mutated tyrosine would still be able to form a hydrogen bond with KPNA5 T434, but may also enable its interaction with rodent karyopherins. The T187I mutation removes intramolecular hydrogen bonds with the backbone of residues H186 and E203 (*Figure 4.3C*). This is likely to increase flexibility in this area. K142E is adjacent to the human KPNA5 interface site and mutations in K142 have been shown to inhibit the interferon signalling (Ilinykh, *et al.*, 2015). This mutation reverses the charge of the side chain. It is possible that this could result in local conformational changes. Overall mutations in the residues that interface with KPNA5 may modulate VP24 interactions with rodent karyopherins.

The other three mutations (L26F, T50I and L147P) all have some effect on the structure of VP24. mCSM predicted L26F to have the most destabilising effect on VP24 (*Table 4.2*). L26 is located at the end of an alpha helix and is packed against two other alpha helices, resulting in the side chain being largely buried (*Figure 4.3D*). Given the tight packing it is possible that the mutation to a larger side chain associated with the change from leucine to phenylalanine requires some conformational change to accommodate the increased size, although there is no indication of what effect this would have on VP24 function. However, given that this mutation was observed in two independent adaptation experiments (Dowall, *et al.*, 2014; Cross, *et al.*, 2015) and also in reverse genetics studies (Mateo, *et al.*, 2011), it seems likely that it has a role in the adaptation to rodent hosts.

T50I removes intramolecular hydrogen bonds with the VP24 backbone residues Q36 and K52 (*Figure 4.3E*). This is likely to increase flexibility in this region of the protein. L147P is located towards the end of an alpha helix. The mutation to proline is likely to result in the breaking of this helix, reducing its length and leading to conformational change in this region. So both of these mutations, while it is not clear how they relate to adaptation, will have an effect on VP24 structure and or dynamics.

Many of these mutations would typically be considered to be unfavourable to a protein, with changes present in interface sites, resulting in the loss of hydrogen bonds and others likely to cause conformational changes. This makes it likely that these mutations are relevant to the adaptation of Ebola virus to rodent hosts. This contrasts with other mutations identified during passaging in the Dowall et al study, which are similarly unfavourable but are not retained in later passages (see below).
This makes it likely that these mutations are relevant to the adaptation of Ebola virus to rodent hosts.

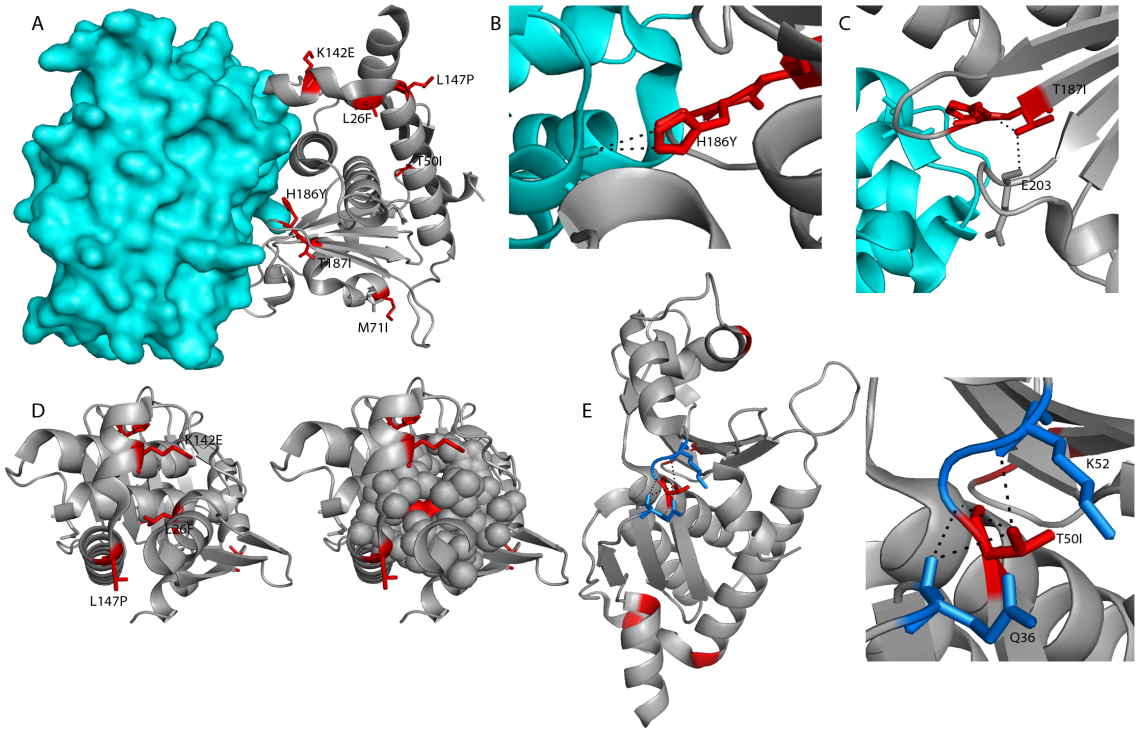


Figure 3. Mutations in VP24 during adaptation to rodents. A) VP24 (grey) in complex with karyopherin a5 (PDB code: 4U2X), adaptation mutations are colour red and shown in stick format. B) VP24 H186 forms a hydrogen bond with KPNA5 T434. C) H186 forms intramolecular hydrogen bonds (black dashed lines) with the backbones of H186 and E203. D) Residue L26 is buried so mutation L26F may affect the conformation of the protein. E). Adaptation mutation T50I will result in loss of hydrogen bonding to the backbones of Q36 and K52.

4.3.6. Mutations that are not retained during passaging may have detrimental effects on protein structure and function

The extensive sequencing analysis in the Dowall study (Dowall, *et al.*, 2014) enabled the investigation of mutations that occurred during the passaging process but were not retained in later passages and instead reverted to wild type. We were able to analyse 24 of these 40 mutations. Our analysis demonstrates that many of these mutations are likely to be destabilising to the Ebolavirus proteins (*Table 4.3* and *Figure 4.4A*). The mutations that are not retained tend to have lower BLOSUM substitution

scores than the adaptation mutations (*Figure 4.4A*), showing that such amino acid changes occur less frequently in nature and therefore may be more likely to alter protein structure/function. Additionally, a group of four non retained mutations are predicted by mCSM to be highly destabilising (ΔΔG > -2.5 Kcal/mol) whereas only one of the adaptation mutations has a similar prediction (*Tables 4.2* and 4.*3*). However for the rest of the mutations there is not much difference in the predicted effect on stability (*Tables 4.2* and 4.*3*).

In NP both W191R and V323D are predicted to be highly destabilising to the protein structure (ΔΔG of -2.973 and -3.339 Kcal/mol respectively). Structural analysis indicates that mutation of W191 to arginine would introduce a charged residue in the interior of the protein in a hydrophobic region (*Figure 4.4A*). This may also alter the hydrogen bond that W191 forms with E61, although arginine at position 191 would still retain functional groups to form a hydrogen bond with E61. Similarly, V323D introduces a charged residue in a buried region, part of this region is hydrophobic, although H327 and E351 form a hydrogen bond and are adjacent to V323. Mutation V323D introduces further negative charge into this region and a hydrogen bond acceptor so this mutation is likely to alter the protein conformation (*Figure 4.4C*).

In VP40, M259R introduces a larger, charged side chain, in a region that is partially exposed but is surrounded largely by hydrophobic residues. Our analysis also suggests that arginine at residue 259 could give hydrogen bond with N257, so there is also the possibility that may form hydrogen bonds with adjacent side chains.

Both temporal changes in VP30, L214P and Q248R, are likely to affect the structure and or function of VP30 (*Figure 4.4D*). L214 is buried and located in the last turn of an alpha helix. Mutation to proline is likely to shorten the helix and therefore result in conformational change. Q248R is in the VP30 homodimer interface site (*Figure 4.4D*). The backbones of adjacent residues L247 and L249 form hydrogen bonds with the other subunit (Hartlieb, *et al.*, 2007) (*Figure 4.4D*). Although this is a relatively conservative substitution, it will increase the charge and size of the amino

acids and it seems likely that the proximity of this mutation to the interface will have an effect on VP30 dimer stability.

So it seems likely that some of these mutations that are not retained in later passages ~~is~~ because they are deleterious to Ebola protein function and therefore are selected against during further passaging.

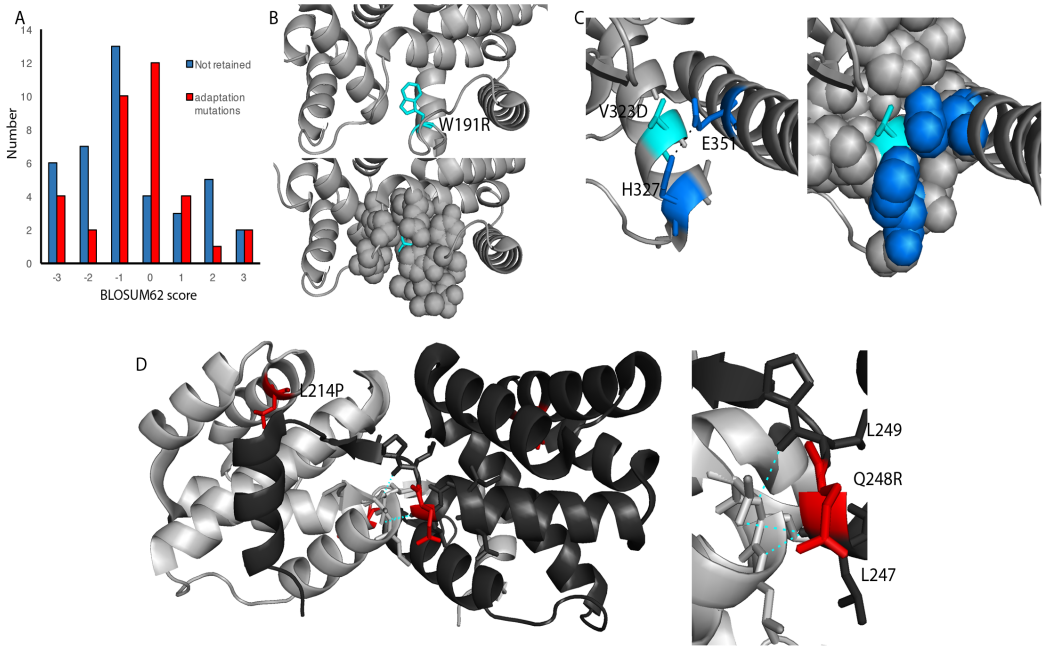


Figure 4. Analysis of mutations that occur during passaging that are not retained in later passages. A) Barchart showing BLOSUM substitution scores for the adaptation mutations (i.e. those mutations that are retained; red) and those that are not retained (blue). B) Mutation W191R (cyan) in NP is observed during passaging. The mutation is located in a buried region. C) Mutation V323D (cyan) in NP, is located close to H327 and E351 (blue; which form a hydrogen bond – black dashed line). D Mutations L214 and Q248R (red) in VP30 are not retained during passaging. Zoom in region shows hydrogen bonding (cyan) around Q248 in the VP30 homodimer interface.

4.4 Discussion

The relevance of the mutations in GP is not clear. The high level of glycosylation of this protein makes it difficult to predict whether (and if yes, how) the mutations may

modulate virus tropism and pathogenicity. Notably, reverse genetics experiments indicated that GP contributes to human pathogenicity but is insufficient for virulence on its own (Groseth, Marzi, Hoenen, *et al.*, 2012). This appears to indicate that Ebola viruses tolerate a substantial number of changes in the sequence of GP without losing virulence. It is also difficult to predict the relevance of the five NP mutations identified in rodent-adapted Ebola virus strains. Some evidence suggests that at least some of the mutations may well be involved in the determination of virus virulence in a certain host, but conclusive evidence is missing. Notably, GP and NP display together with L the greatest variability in their sequences (Jun, *et al.*, 2015). Therefore, some variation in these sequences may not be surprising.

Modelling of the VP24 mutations suggests that they are all likely to modulate the virus-host cell interaction. In particular, H186Y, T187I, and K142E are likely to be relevant for the modulation of the host cell interferon response. Therefore, there is strong evidence that changes in VP24 are required to enable Ebola virus adaptation to a novel host. This notion is in accordance with evidence suggesting that VP24 may be a determinant of pathogenicity among different Ebolaviruses (Zhang, *et al.*, 2012). The retention of these mutations while other mutations that occur during passaging of the Dowall study but are not retained in further passages, suggests that these mutations have a role in rodent pathogenicity.

We have recently suggested that VP24 may be central to explaining how Reston viruses are the only Ebolavirus species that are not pathogenic in humans (Pappalardo, *et al.*, 2016). We identified multiple residues in VP24 that are differentially conserved between Reston viruses and the four human pathogenic Ebola virus species. Three of these residues are located in the VP24-KPNA5 interface site and we proposed that they result in impaired binding of Reston VP24 with karyopherins and thus a reduced ability to inhibit interferon signalling. So in two different contexts we have observed differences in VP24 that are related to species-specific pathogenicity, thus together they provide strong evidence for VP24 in determining host pathogenicity.

Given our analysis, how many mutations are required to alter Ebola virus host pathogenicity? Notably, our analysis has shown that only very few mutations may be required for the adaptation of an Ebolavirus to a novel host. In total, the different adaptation experiments resulted in 5 (Cross), 6 (Volchkov-1), 7 (Ebihara), or 16 mutations (Dowall) (Dowall, *et al.*, 2014). As described above, 11 of the 16 mutations in the Dowall et al. study (Dowall, *et al.*, 2014) occurred in L, it remains unclear whether these mutations would have been sustained during further passaging in guinea pigs (see above). This also means that only 4 to 5 mutations were detected in these genes per individual adaptation experiment. So this may represent a minimum number of coding mutations required in an Ebola virus genome to enable Ebolaviruses to cause disease in a novel, previously non-susceptible host. It is reasonable to assume that not every mutation is essential for Ebolavirus adaptation to a novel host, so this required number of mutations may be even lower.

The adaptation of the human-pathogenic Ebolavirus species, Ebola, Sudan, Bundibugyo, and Taï Forest viruses to humans that might result in increased virulence does not appear to be a major concern. Their virulence in humans is extremely high they are still considered to be deadly to humans (Feldmann & Geisbert, 2011; Gray, *et al.*, 2014). Hence, adaptation of human-pathogenic Ebolaviruses to humans (which would ultimately result in Ebolaviruses that circulate in humans as reservoir species) would be expected to result rather in a decrease of pathogenicity to achieve a balance between virulence and pathogen fitness and/or transmission. However, the potential of Ebolaviruses to adapt to novel host species may be of relevance with regard to the potential threat exerted by the non-pathogenic member of the *Ebolavirus* genus, the Reston viruses. Reston viruses and Ebola viruses are known to circulate in pigs, and can be transmitted from pigs to humans (possibly by air) (Weingartl, 2013; Barrette, *et al.*, 2009; Marsh, *et al.*, 2011; Osterholm, *et al.*, 2015; Atherstone, *et al.*, 2015; Pan, *et al.*, 2014; Olson, *et al.*, 2012; Miranda & Miranda, 2011). Moreover, dogs have been suggested to become infected and may play a role during virus transmission to humans and as potential reservoir species (Osterholm, *et al.*, 2015; Weingartl, 2013) (Olson, *et al.*, 2012).

Table 4.2. Mutations identified during serial passaging of rodents. The table details protein structural analysis of the mutations including their BLOSUM62 substitution score, solvent accessible surface area and the predicted change in protein stability from mCSM. All studies considered adaptation in Guinea pigs with the exception of the Ebihara et al., study, which used mice, indicated with * in the study column. #The mutation in GP I544T, is commonly a T in Ebola virus and the structure available contains a threonine at this position. Therefore the mCSM analysis considered the mutations from threonine to isoleucine.

Protein	Mutation	Study	BLOSUM62 score	Solvent Accessible Surface Area	mCSM ΔΔG (Kcal/mol)	mCSM Effect
NP	S72G	Ebihara*	0	0	-1.126	destabilizing
NP	N566S	Dowall	-1	-	-	-
NP	A575T	Cross	0	-	-	-
NP	S647Y	Cross	-2	86	-0.652	Destabilizing
NP	F648L	Volchkov	0	21	-0.86	Destabilizing
VP35	A12V	Ebihara*	0	-	-	-
VP35	N204D	Dowall	1	-	-	-
GP	D49N	Dowall	1	71	0.398	Stabilizing
GP	S65P	Ebihara*	-1	6	-0.011	Destabilising
GP	V203I	Dowall	3	-	-	-
GP	S246P	Ebihara*	-1	49	-0.253	Destabilising
GP	D397G	Volchkov	-1	-	-	-
GP	I544T	Ebihara*	-1	54#	-0.556#	Destabilising
GP	I544T	Cross	-1	54#	-0.556#	Destabilising
VP24	L26F	Dowall	0	0	-0.644	Destabilizing
VP24	L26F	Cross	0	0	-1.656	Destabilizing
VP24	T50I	Ebihara*	-1	9	0.109	Stabilizing
VP24	M71I	Volchkov	1	75	-0.216	Destabilizing
VP24	L147P	Volchkov	-3	94	-0.636	Destabilizing
VP24	L147P	Mateo	-3	94	-0.636	Destabilizing
VP24	H186Y	Volchkov-2	2	7	0.563	Stabilizing
VP24	T187I	Volchkov	-1	4	-1.157	Destabilizing
VP24	K142E	Cross	1	52	-0.082	Destabilizing

L	N38K	Dowall	0	18	0.062	Stabilizing
L	G707A	Dowall	0	1	-0.497	Destabilizing
L	T820A	Volchkov	0	6	0.081	Stabilizing
L	T930A	Dowall	0	0	-2.245	Destabilizing
L	L940P	Dowall	-3	8	-1.713	Destabilizing
L	F934L	Ebihara*	0	0	-3.187	Destabilizing
L	Y1271stop	Dowall	-	-	-	-
L	N1478I	Dowall	-3	-	-	-
L	I1532V	Ebihara*	3	-	-	-
L	A1546E	Dowall	-1	-	-	-
L	S1998T	Dowall	-2	-	-	-
L	N2144K	Dowall	0	-	-	-
L	F2151V	Dowall	-1	-	-	-

Table 4.3. Analysis of mutations identified during passaging in Dowall et al., (Dowall, Matthews, Garcia-Dorival, *et al.*, 2014) but not retained in later passages.

Protein	**Mutation**	**BLOSUM62 score**	**Solvent Accessible surface Area (Å²)**	**mCSM ΔΔG (Kcal/mol)**
NP	W191R	-3	0	-2.973
NP	V323D	-3	7	-3.339
NP	L414R	-2	-	-
VP35	S129P	-1	-	-
VP35	I246A	-1	0	-2.783
VP40	E15Q	2	-	-
VP40	P66S	-1	63	-0.431
VP40	M259R	-1	27	-1.569
GP	M1K	-1	-	-
GP	R11K	2	-	-
GP	V92L	1	20	-0.345
GP	P187L	-3	63	-0.357
GP	I465T	-1	-	-
GP	S493P	-1	-	-
GP	R638K	2	-	-

GP	Y652F	-1	-	-
GP	Y668C	-2	-	-
VP30	L214P	-3	0	-1.935
VP30	Q248R	1	118	-0.269
VP24	F29V	-1	2	-1.342
VP24	A43P	-1	0	0.55
VP24	K218R	**2**	**47**	**-0.759**
L	G30W	-2	**-**	-1.123
L	R161W	-3	**-**	-0.155
L	N525D	1	**-**	0.288
L	K537R	2	**-**	-0.058
L	L538P	-2	**-**	-0.564
L	I669S	-2	**-**	-3.029
L	M705T	-1	**-**	-1.14
L	S826Y	-2	**-**	-0.642
L	S868P	-1	**-**	0.207
L	F879L	0	**-**	0.376
L	I943R	-3	**-**	-1.589
L	T993A	0	**-**	-1.262
L	L1096S	-2	**-**	-1.977
L	S1308P	-1	-	-
L	F1733Y	3	-	-
L	L1763P	3	-	-
L	H1949Q	0	-	-
L	L2197P	0	-	-

Table 4.4: Ebola virus protein structures and templates used for modelling;

PROTEIN	OLIGOMERIC STATE	PDB/TEMPLATE	REGION IN SEQUENCE
GP	Trimer of Heterodimers	3CSY (structure)	31-310 502-599
sGP	Dimer	3s88I (model)	32-287
L	Monomer	5a22T (model)	8-1140
L	Monomer	4n48A (model)	223-328
NP (C-terminal)	Monomer	4QB0 (structure)	645-739
NP (N-terminal)	Monomer	4YPI (structure)	39-384
VP24	Heterodimer	4M0Q (structure)	10-231
VP24	Heterodimer	4U2X (structure)	16-231
VP30	Dimer	2I8B (structure)	140-266
VP35	Heterodimer	4IBB (structure)	218-340
VP35	Dimer of heterodimers	3L25 (structure)	209-340
VP40	Monomer	1ES6 (structure)	44-321
VP40	Dimer	4LDB (structure)	44-319
VP40	Hexamer	4LDD (structure)	45-188
VP40	Octamer	4LDM (structure)	69-188

4.5. Methods

The mutations identified during Ebola virus adaptation to rodents were extracted from four studies (Dowall, *et al.*, 2014; Ebihara, *et al.*, 2006; Volchkov, *et al.*, 2000; Cross, *et al.*, 2015).

Available Ebola virus proteins were obtained from the protein databank, where structures were not available they were modelled using Phyre2 (Kelley et al., 2015). The structures used and templates for models are listed in *Table 4.4*. The adaptations were mapped onto the protein structures and their location in the structure analysed using PyMOL. mCSM was used with default parameters to calculate the effect of the adaptation mutations on protein stability (Pires, et al. 2014). Solvent accessible surface area was calculated using DSSP (Joosten, *et al.*, 2011).

For the I554T mutation in GP, the protein structure (pdb code: 3CSY) already had a threonine at position 554. To UCSC genome browser (Kent et al., 2002) was used determine what residues are typically present at this position. This revealed that the original Mayinga 1976 strain has isoleucine at position 554, but the the vast majority of other Ebola virus genome sequences have threonine at position 554. As a result I554T was not classed as an adaptation mutation.

Chapter 5

Investigating Ebola virus pathogencity using Molecular Dynamics

Morena Pappalardo, Francesca Collu, James Macpherson, Martin Michaelis, Franca Fraternali, Mark N Wass, in preparation.

My contribution to this paper: I have performed all molecular dynamics simulations and associated analysis. Training and guidance was provided by Francesca Collu and Jamie Macpherson.

5.1. Abstract

The extent of Ebolavirus pathogenicity and ability to cause epidemics has recently been demonstrated by the outbreak in West Africa. Of the five Ebolavirus species (Ebola, Tai Forest, Bundibugyo, Sudan and Reston), only Reston viruses are not pathogenic in humans. We have recently proposed that conserved amino acid differences in the Ebolavirus protein VP24 between Reston viruses and the four human-pathogenic Ebolaviruses may explain this difference in pathogenicity. VP24 inhibits interferon signalling by binding to both STAT1 and karyopherins to prevent STAT1 accumulation in the nucleus and this consequently blocks interferon signalling. Here we used molecular dynamics to investigate the effect of these conserved differences on the interaction of VP24 with Karyopherin alpha5. In the simulations we observed that Reston virus VP24 has many anti-correlated movements with KPNA5 in comparison to the interaction of Ebola virus VP24 with KPNA5. Additionally the dynamics of the Reston virus VP24 with KPNA5 more closely resemble those of Ebola virus VP24 with mutation R137A, which is known to remove binding of Ebola virus VP24 with KPNA5. Our results therefore support the basis that the interaction of Reston virus VP24 with KPNA5 is different to that of Ebola virus VP24 and given the anti correlated interactions observed it is likely that binding is reduced.

5.2. Introduction

The pathogenicity of Ebola virus has been highlighted by the recent outbreak in West Africa (Quaglio et al., 2016) with more than 11,316 thousand deaths and 28,639 confirmed cases as of 28th February 2016 (www.who.int). Suppression of the host immune response is a prominent feature of Ebola virus infection, which may explain the high fatality rate observed in the last outbreaks. Ebolaviruses do this through at least three proteins, GP, VP35 and VP24 (Hoenen et al., 2015; Ilinykh et al., 2015; Bale et al., 2015; Kimberlin et al., 2009). The Ebola virus protein VP24, binds the transcription factor STAT1 and karyopherins (known to bind Karyopherin α1, α5 and α6 in humans) to prevent transport of STAT1 to the nucleus and it therefore inhibits interferon signalling (Xu et al, 2014). VP35 prevents interferon signalling by

binding to viral double stranded RNA, which prevents triggering interferon signalling. Additionally, GP is a surface protein responsible for interaction with the host cell receptors and entry of the virus into host cells. It is thought that GP's glycan cap provides a mechanism for escaping the immune system.

We are interested in identifying the molecular determinants of Ebolavirus pathogenicity to further our understanding of how Ebolaviruses infect and kill hosts and how we can combat this. There are five known Ebolavirus species, Ebola virus (formally called Zaire), Sudan virus, Bundibugyo virus, Taï forest virus and Reston virus (Kuhn et al., 2010). Reston viruses are not pathogenic in humans, while the four other species are. In a recent study we identified differences between the four human-pathogenic Ebolavirus species and Reston viruses that are likely to explain their difference in human pathogenicity (Pappalardo et al., 2016). Our key finding was the presence of amino acid differences between the Ebola and Reston VP24 proteins that correspond to the interface site between Ebola virus VP24 and human karyopherin alpha 5 (KPNA5). We proposed that the different interface amino acids present (T131S, N132T, M136L, Q139R –Ebola virus residue listed first and Reston virus residue second) at this site in Reston VP24 are likely to reduce the affinity for Reston VP24 with human karyopherins and therefore limit the ability of Reston viruses to inhibit interferon signalling via this mechanism.

Xu et al. (2014) characterized the Ebola VP24 and KPNA5 complex by a combination of structural and biochemical analysis. They crystallised the Ebola VP24 with the Armadillos 7-10 of KPNA5 and investigated the effect of VP24 mutations on binding to KPNA5 using coimmunoprecipitation pull down experiments and compared the bands obtained in the gel with wild type protein. This approach is not quantitative but the strength of the band provides an indication of the extent to which binding is affected. For R137A and R137A, T138A,Q139A the band is very weak. For F134A/M135A it is intermediate between these previous two mutations and the wild type. Additionally the same study also observed that while most single point mutations in the VP24 interface (except R137A) had little effect on binding to KPNA5, combinations of mutations in VP24 (F134A/M136A and

R137A/T138A/Q139A) resulted in near loss of binding to KPNA5 (Xu et al., 2014). These included some of the positions that vary between Ebola and Reston viruses, which further support our hypothesis that Reston VP24 has different binding properties with KPNA5. In the previous chapter mutations present in experiments adapting Ebola virus to rodents, Figure 5.1 shows mutations coming from both analyses.

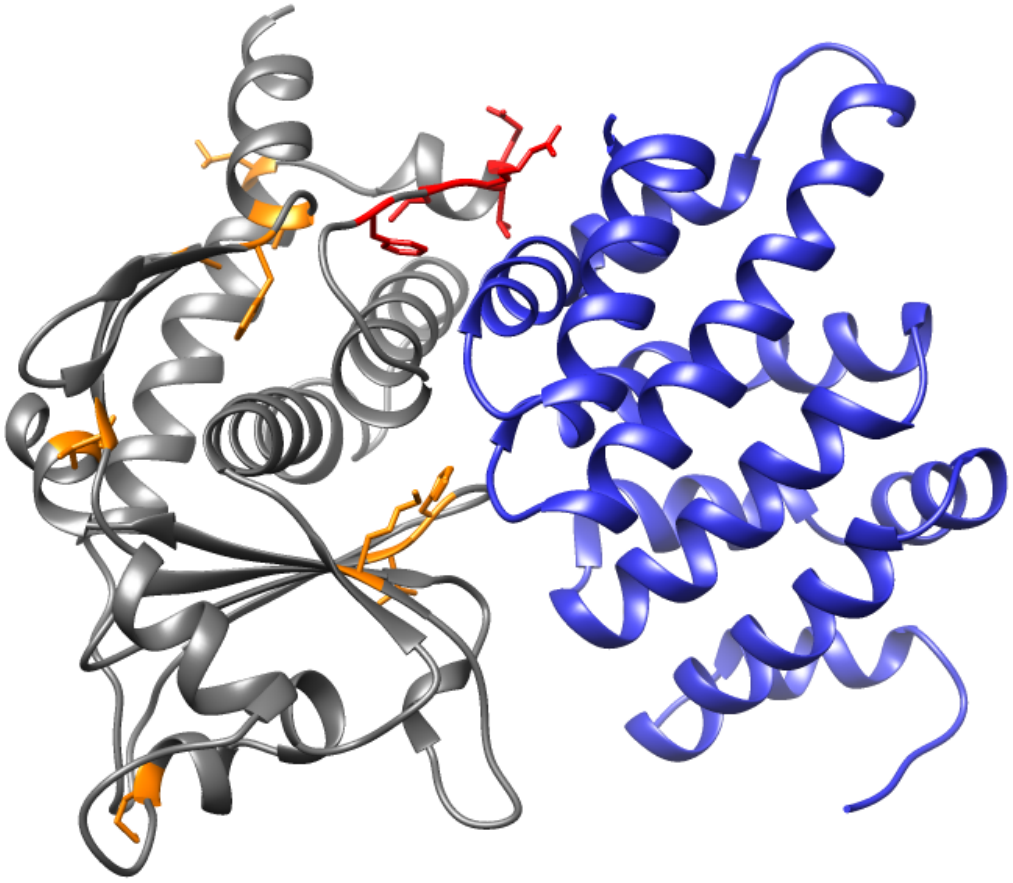


Figure 5.1: Adaptational and experimental mutations in protein VP24; protein VP24 is shown in gray cartoon and protein KPNA5 is shown in blue cartoon. Adaptation mutations are shown in yellow sticks and experimental mutation coming from Xu et al. (2014) are shown in red sticks.

In this study we use protein structural analysis and molecular dynamics simulations to investigate Ebola and Reston VP24 and their interaction with KPNA5 to consider our hypothesis that amino acid changes in Reston virus VP24 affect binding to KPNA5. This is done in the context of the mutagenesis data from Xu et al., (2014), enabling comparison of simulations with experimental (*in vitro*) data and their use to

interpret the molecular dynamics simulations where experimental data is not available.

5.3. Methods

5.3.1. Modelling of a RESTV-VP24 KPNA5 complex

The EBOV and RESTV VP24 sequence share 81.3% sequence identity and 96% similarity. The protein structures were aligned using Chimera (Pettersen et a., 2004) and a model for RESTV VP24 in complex with human Karyopherin Alpha 5 built using MODELLER 9.0 (Webb et al., 2014). The RESTV VP24 crystal structure (PDB 4D9O) and the EBOV VP24-KPNA5 complex (PDB 4U2X) were used as templates for the new model. 200 models were obtained and the one with the lowest DOPE score was selected.

5.3.2. Comparison of interfaces

PISA (Krissinel et al., 2007) and mCSM (Pires et al., 2014) were used to analyse the interfaces in the complexes. POPSCOMP (Kleinjung & Fraternali, 2005) was used to determine the contribution of the individual residues to the hydrophilicity and hydrophobicity at the interface, according to their solvent accessible surface area (SASA), using default parameters. The residues were classified as being part of the core, support or rim regions of the interface according to the change in SASA (when % of hydrophobicity was greater than 40 and difference in SASA was less then 10 $Å^2$ the residue was considered as core, otherwise it was rim).

5.3.3. Molecular Dynamics simulations

Molecular dynamics simulations were performed for the wild type forms of EBOV-VP24 and RESTV-VP24 in complex with human KPNA5. Other simulations were performed on the EBOV-VP24-KPNA5 complex with mutations introduced into VP24 where the effect on KPNA5 binding had been experimentally determined (Xu et al., 2014). The mutations considered were: 1)R137A, 2)Q139A, 3)F134A,M136A and 4) R137A-Q139A.

Molecular dynamics simulations were performed using Gromacs 5.0.5 (Abraham et

al., 2015) using the GROMOS96 53a6 force field (JCC 2004 vol 25 pag 1656). 600 ns trajectories were obtained for the Ebola virus VP24-KPNA5 complex and the model of Reston VP24-KPNA5 complex. 200 ns trajectory was obtained for R137A and F134A,M136A and 100 ns for all other simulations.

We applied our in-house protocol to prepare the molecules for the simulations: to neutralise each system counter atoms Na^+ were added to the solvated proteins, according to the different total charge in each system. The system was then minimised and equilibrated according to the Maxwell distribution temperature (300K), passing through three different temperatures, at 100K, 200K and 300K using restraints, and then equilibrated again using the same temperature steps but without restraints. This approach is generally done to avoid artifacts. Velocities were generated using the gen-vel option, using a random seed (gen-seed).

5.3.4. Molecular Dynamics Analysis

Trajectories were analysed using the GROMACS analysis tools, VMD tools and the Bio3D package for R (Grant et al., 2014). Analyses for the wild type complexes were carried out from 280ns to 600 ns, which is the range of simulation where the RMSD reached the plateau in the two cases.

For the analysis, standard Periodic Boundary Conditions were removed and Minimum Image Convention (MIC) were applied to all the trajectories. Rotational and translational movements were then deleted in order to perform the Principal Component Analysis. Secondary structure plots for trajectories were obtained using the DSSP (Kabsch and Sander, 1983) tool in gromacs. Root mean square deviation (RMSD) and fluctuation (RMSF) from the initial starting complex were obtained using Bio3D, as well as the PCA analysis and correlation plots.

5.4. Results

To investigate how the interactions of Ebola and Reston virus VP24 with KPNA5 may differ we performed molecular dynamics simulations of both of these complexes. We then performed simulations of the Ebola virus VP24 complex with KPNA5 with mutations introduced in VP24 that are known to alter binding. This was done to enable comparison with the Reston virus simulation.

5.4.1 Initial Comparison of the interface between EBOV and RESTV VP24 with KPNA5

A model of RESTV VP24 and human KPNA5 was generated using the RESTV VP24 structure and the recently solved crystal structure of EBOV VP24 complex with human KPNA5 as a template (see methods). The interface residues, as well as the energies and bonds in this model and in the EBOV VP24-KPNA5 complex were first compared using PISA (Krissinel & Henrick, 2007) and POPSCOMP (Kleinjung & Fraternali, 2005). The PISA analysis identified a smaller interface area in the RESTV complex with a slightly lower binding energy but with seven fewer hydrogen bonds (nine in the RESTV complex compared to 16 in the EBOV complex). We then compared the interfaces after minimisation during the initial state of the molecular dynamics trajectory (zero ns) and at the end of the simulation (600 ns). In the EBOV complex nine hydrogen bonds were found at the beginning of the trajectory (0 ns) but only seven remained at the end of the trajectory. For the RESTV complex, eleven hydrogen bonds were present at zero ns and nine remained at 600ns. In total three hydrogen bonds were equivalent in the two complexes in the first snapshot, whilst only two of them overlapped at the end of the simulation (*Figure 5.2B*), (between VP24 137- KPNA 480 and VP24 138- KPNA5 480). The hydrogen bonds involving residue Q139, which is one the residues that is mutated in our study, and the proximal residue R140 are lost in the RESTV complex. This is interesting since residue R140 forms a hydrogen bond with E474 and a salt bridge with E475 at the interface.5. The H-bond given with residues E474 has 1.91 *Å* distance. Residue R140 has an accessible surface area of 191.04 $Å^2$, a buried surface area of 117.29 $Å^2$ (70% of the interface is buried) and a solvation energy effect of -0.99 Kcal/mol. Interestingly at the end of the simulation in the RESTV complex, the VP24 residue R137 forms a hydrogen bond with L479 and two salt bridges residues with D480 and E483 in KPNA5 *(Figure 5.2B*). Mapping the hydrogen bonds at the interface (figure S1) we observed that residue R137 undergoes different conformational changes that make it essential for the stability of the interface, according to the mCSM and the FoldX predictions and our MD results (see later).

POPSCOMP (Kleinjung & Fraternali, 2005) is an extension of the POPs server (Fraternali and Cavallo, 2002), it calculates the buried solvent accessible surface area (SASA) in protein complexes. We found that the total difference in the buried SASA differs in the two complexes, it is slightly higher for the EBOV complex with respect to the RESTV complex which has a smaller interface. Seven of the twelve EBOV VP24 residues in the interface are also present in the RESTV VP24 interface with KPNA5 (*Figure 5.2, Figure 5.3*). Five of the ten EBOV KPNA5 residues are equivalent in the RESTV_KPNA5 interface with their respective VP24. This shows that while there is overlap, there are also considerable differences between the known EBOV VP24-KPNA5 complex and the modelled RESTV VP24-KPNA5 complex (*Figure 5.3*). POPSCOMP predicted that the interface is weaker in the RESTV complex at the end of the trajectory the interface area with KPNA5 is much smaller than the initial conformation.

Table 5.1: Pisa and POSPCOMP Interface Analysis from the initial crystal structures

	EBOV-COMPLEX crystal structure	**RESTV-COMPLEX model**
PISA results		
Interface Area (Å²)	1065.9	977
Solvatation Free Energy (ΔΔG, Kcal/M)	-9.2	-9.5
H-Bonds	16	9
PISA results at 0 ns		
Interface Area (Å²)	1099.7	1055.1
Solvatation Free Energy (ΔΔG, Kcal/M)	-8.5	-8.6
H-Bonds	9	11
PISA results at 600 ns		
Interface Area (Å²)	1119.2	1076
Solvatation Free Energy (ΔΔG, Kcal/M)	-10	-9.1
H-Bonds	7	11

POPSCOMP results		
Hydrophobic difference (Å²)	1042.28	1002.35
Hydrophilic difference (Å²)	772.73	713.95
Total difference (Å²)	1815.06	1716.43

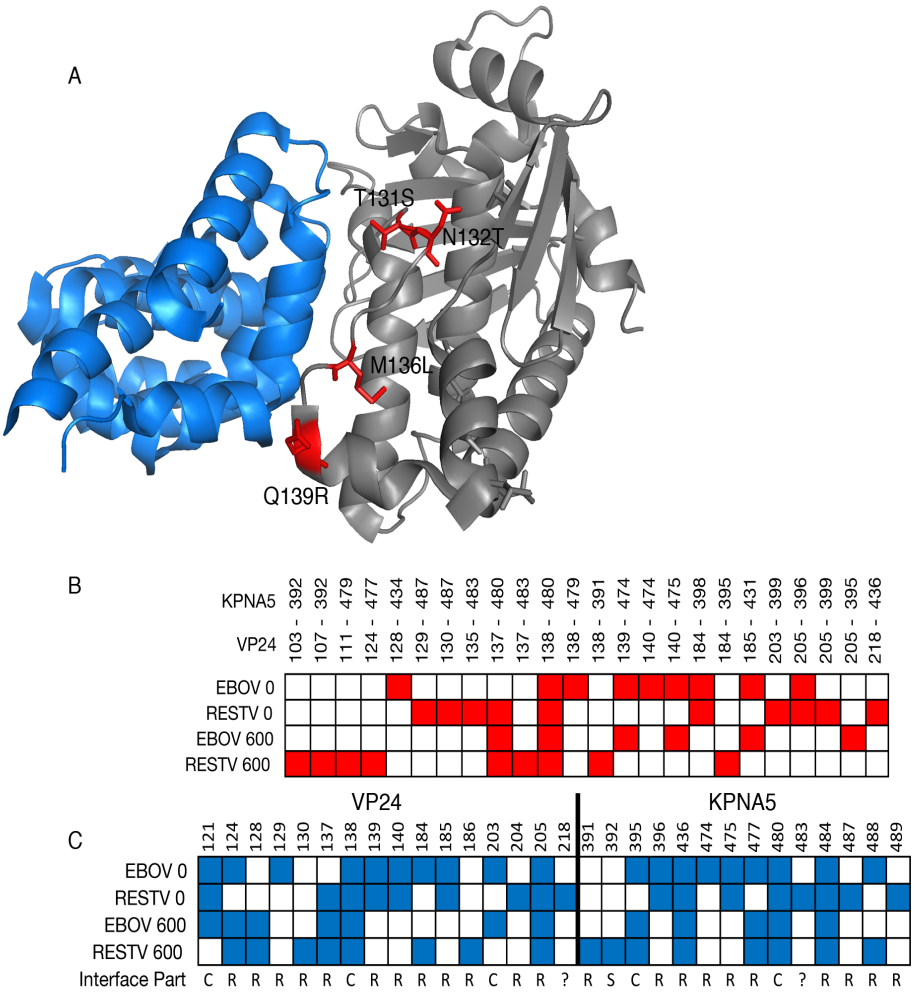


Figure 5.2: Ebola virus VP24 complex with KPNA5. A) VP24 is coloured grey and KPNA5 is blue. Residues differentially conserved between Ebola and Reston viruses in the interface site are shown in red stick format and labeled with the Ebola virus amino acid, residue number followed by the Reston virus amino acid. B) Hydrogen bonds present at the beginning of the MD trajectory (EBOV 0, RESTV 0) and at the end of the (EBOV 600, RESTV 600), red squares indicated that a hydrogen bond is present. C) Residues present in the VP24-KPNA5 interface at 0 and 600ns for both EBOV and RESTV VP24. Interface part indicates if the residue is part of the core (C), support (S) or rim (R) regions of the interface. Note EBOV represents Ebola virus and RESTV Reston virus.

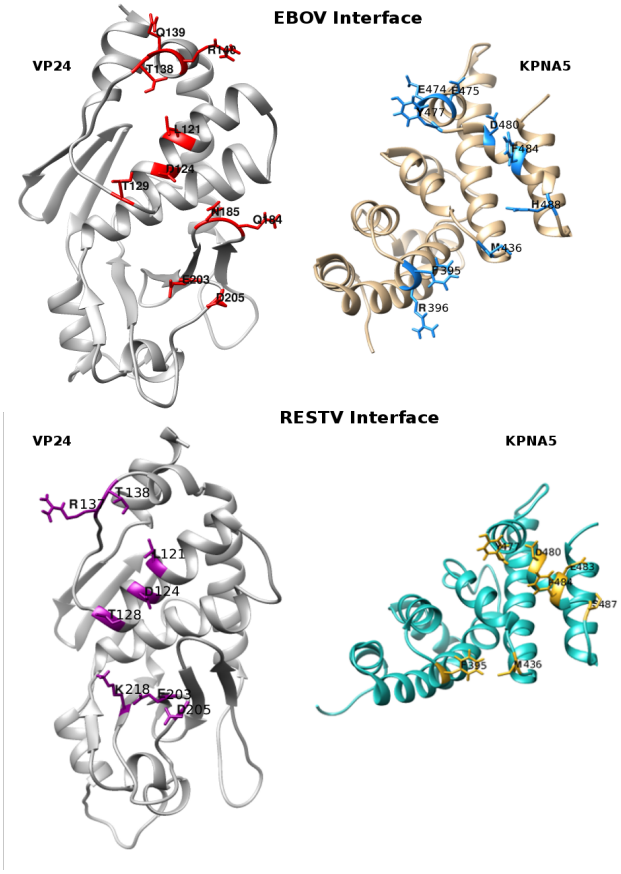


Figure 5.3: Interface Residues predicted by POPSCOMP were mapped onto structure. On the top of the figure the EBOV Interfaces for VP24 (gray cartoon) and KPNA5 (yellow cartoon) are shown. Residues that contribute to the interfaces are shown in stick (red for VP24 and blue for KPNA5). On the bottom of the figure the RESTV Interfaces for protein VP24 (gray cartoon) and for KPNA5 (cyan cartoon) are shown. Residues that contribute to the Interfaces are shown in sticks (magenta for VP24 and yellow for KPNA5).

5.4.2. Predicted effects of Mutations at the Interface VP24-KPNA5 interface

Next we used mCSM (Pires, Ascher & Blundell, 2014) and FoldX (Schymkowitz, Borg, Stricher, *et al.*, 2005) to consider how each of the residues in the EBOV VP24 interface that is a different amino acid in RESTV VP24 may affect the stability of the complex. mCSM also predicted the effect on the affinity of the complex (see methods). For the mutations with experimental data mCSM predicts that both point changes reduce the stability and the affinity of the complex, with the R137A

mutation having a greater effect (predicted (ΔΔG -1.066 Kcal/mol change in complex affinity) than Q139A (Table 2). The FoldX predictions agree with mCSM for both point mutations. Additionally FoldX was able to consider combinations of mutations simultaneously and predicted that both the F134A/M136A, and R137A-Q139A mutations reduce stability of the complex with a very large reduction of more than 7Kcal/ml for the F134A,M136A combination. These predictions are generally in agreement with the experimental observations that R137A and the two multiple mutation sets nearly remove all binding of EBOV VP24 with KPNA5 (Xu, Edwards, Borek, *et al.*, 2014a).

Next we considered how the conserved amino acid differences between EBOV and RESTV VP24 may affect stability of the EBOV VP24 complex when the RESTV residues are introduced into the EBOV structure (*Table 5.2*). Again mCSM predicted that all of the changes would reduce the stability and affinity of the complex (with the exception of M136L, where a small increase in affinity is predicted). The changes in stability are similar to the predicted change for R137A, which is known to reduce binding. FoldX also predicts reduced stability for two of these four point changes, with increased stability predicted for M136L and Q139R, although the ΔΔG for M136L is predicted to be very small (0.18Kcal/mol). It also predicts a slightly less stable complex with all four amino acid changes present (*Table 5.2*). Overall these predictions suggest that individually the amino changes are likely to reduce the stability and affinity of the complex. This provides some initial support for our proposal that the binding affinity for KPNA5 by RESTV and EBOV VP24 proteins differs.

Table 5.2: mCSM and FoldX stability changes for single amino acid changes in the EBOV VP24 – KPNA5 complex.

Mutation	mCSM stability (ΔΔG - Kcal/mol)	mCSM PP affinity (ΔΔG - Kcal/mol)	FoldX stability (ΔΔG - Kcal/mol)
Experimental point mutations			
R137A	-0.805	-1.066	-0.68
Q139A	-0.386	-0.239	-0.33
F134A,M136A	NA	NA	-7.3
R137A,T138A,Q139A	NA	NA	-1.02
Conserved amino acid differences between EBOV and RESTV VP24			
T131S	-1.295	-0.317	-0.42
N132T	-0.617	-2.65	-1.22
M136L	-0.814	0.166	0.18
Q139R	-1.058	-0.995	1.59
T131S,N132T,M136L,Q139R	NA	NA	-0.3

5.4.3. Molecular Dynamics Analysis

To further our analysis molecular dynamics simulations were performed on the EBOV V24- KPNA5 complex and the model of RESTV VP24 with KPNA5. Simulations over 600ns were obtained with the trajectories trimmed using the last 320ns (280-600ns). RMSD of the main chain C-Alphas was stable for both complexes (*Supplementary Figure S2*). The RMSD of the RESTV VP24-KPNA5 model is greater than the EBOV complex, (*Supplementary Figure S2*), this could indicate a difference in the interaction between RESTV VP24 and KPNA5 but could also partly reflect that the simulation is based on a model rather than a solved structure,

which may result in greater movement to accommodate the best conformation.
For VP24 some minor differences in fluctuation (i.e. Root mean squared fluctuation – RMSF) were observed between the RESTV and EBOV proteins. One of these differences coincide with the interface site at residues 113 (*Figure 5.4A*). Residue 113 is located in an alpha helix at the interface. For KPNA5 there are larger differences in RMSF in four regions, three of which coincide with the complex interface (*Figure 5.4A*). The most pronounced difference is around residues 477 and 479 (a loop region between two alpha helices) , where there is very little fluctuation of KPNA5 in the EBOV VP24 complex (around 1 *Å*) but in the RESTV VP24 complex there is a peak of 8 Å. The greater fluctuation in KPNA5 suggests that the interaction with RESTV VP24 differs from that with EBOV VP24.

Analysis of the secondary structure (using DSSP – see methods) during the simulation revealed minor changes in the secondary structure occurring at the interface site (*Supplementary figure S3*). The most important changes were found around residue 76 where there is a prevalence of turns in EBOV becoming coils in RESTV. Residues 133 and 134 (shown in *figure S3*), as well as residue 146, which are proximal to the binding interface lose their bend and beta bridge structure to become unstructured in the RESTV complex The largest changes in secondary structure were found in KPNA5, particularly in two regions between residues 365-375 and 385-395 (*figure S3*), the second region which is involved in binding VP24, losses it's alpha helical structure after 220 ns in the RESTV complex.

Cross correlation analysis was performed to consider how the proteins move in relation to each other. Using a threshold of 0.7 to explore the correlated motions, the RESTV complex has more unrelated motions, meaning there is much greater movement of VP24 and KPNA5 away from each other (*figure 5.4B*). Additionally, the RESTV complex showed a higher number of correlated motions and this probably reflects the adaptation movements that VP24 and KPNA5 undergo when they try to form a complex. To further support our analysis Principal Component Analysis confirmed the different movements in the two complexes as shown in *figure 5.4C-D*;

we calculated Principal components one and two and projected the RESTV principal components onto the EBOV ones. (*figure 5.4D)*. This projection shows that the movements are in different directions (f*igure 5.4D*). We scaled principal components 1 and 2 using gromacs tools and we projected the eigenvectors into a porcupine visualisation (see *figure 5.4C*). The first three eigenvectors describe 47.6% of the conformational variance for the EBOV complex and 48.80% or the RESTV simulation. This denotes great conformational changes in both cases with 1.2% more flexibility.

Gromacs Hydrogen bond analysis identified an average of 14 interface hydrogen bonds for the EBOV complex and only 11 H-bonds for the RESTV complex, (using 3.8 Å for donor/acceptor distance and 40 Å for the cut-off of the angle; *Figure 5.5*). This agrees with the PISA analysis which found fewer hydrogen bonds in the RESTV starting model.

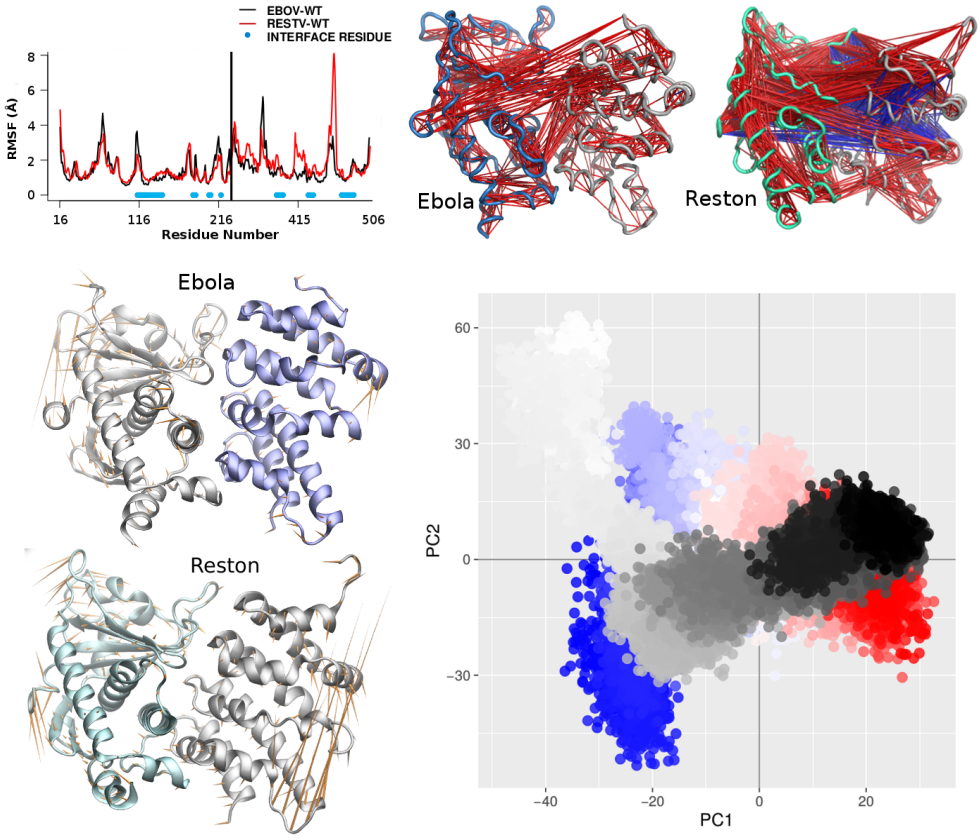


Figure 5.4. Molecular dynamics simulations of Ebola and Reston virus VP24 interaction with KPNA5. A) RMSF graph is shown, where in black line the fluctuation for EBOV complex and in red line the one for RESTV are shown. B) The cross correlation analysis is shown in both complexes; red lines represent the correlated movements whereas the blue lines represent the anticorrelated ones; C) Principal Component Analysis is shown in porcupine visualization for both complexes. D) Principal Component Analysis Projection for EBOV complex (from white to black) and for the RESTV (from blue to red).

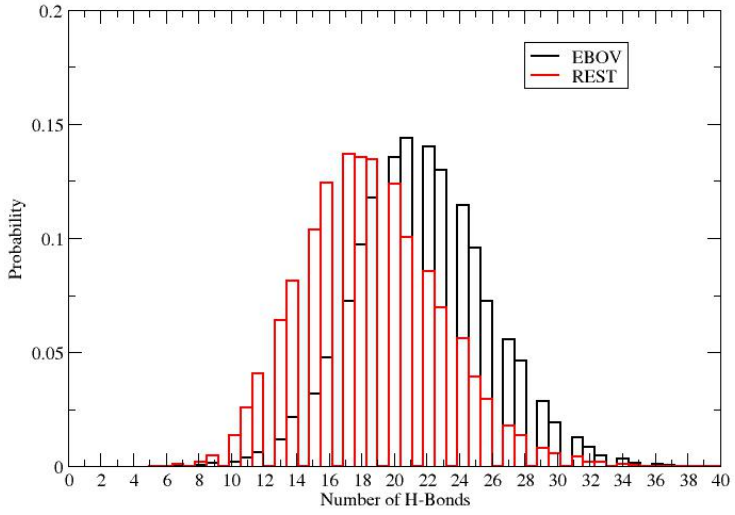


Figure 5.5: H-bond analysis during MD simulations. In black the EBOV complex and in red the REST one. The Gaussian curves represent the mean of H-bonds occurring at the Interface during 600 ns of simulation.

5.4.4. Analysis of mutations in the EBOV VP24-KPNA5 complex

The mutagenesis studies performed by Xu et al., (Xu, Edwards, Borek, *et al.*, 2014b) provide an opportunity to perform simulations and match them with experimental data, which can be used to make further inferences about the RESTV VP24-KPNA5 simulations. The R137A and F134A,M136A, R137A-Q139A mutations are known to have a significant effect on the binding of EBOV VP24 and KPNA5. 200 ns MD simulations were performed for R137A and for F134A,M136A and 100 ns simulation

for the others. Additionally Q139A is known to individually have a minimal effect on binding and this was used as a control. Initial RMSD and RMSF analysis showed greater changes for the R137A and F134A-M136A mutations, while the simulation with the Q139A and R137A-Q139A mutations behaved similarly to the wild type complex (*Figure 5.6*), particularly with the RMSD and RMSF of the complex with R137A-Q139A showing very little difference to the wild type complex. This is surprising given that this combination of mutations is known to reduce binding of VP24 and KPNA5. In all simulations greater fluctuations in VP24 was observed around the site of the mutation (*Figure 5.6*) . Mutation R137A causes an increase in fluctuation of almost 1 *Å* in the proximal residues at the binding interface. The same is shown for mutations F134A-M136A, where the change is larger (2 *Å*) and the upper peak in KPNA5 reaches almost 12 *Å*.

Cross correlation Analysis showed correlated and anti-correlated moves in the mutated complexes (*Figure 5.7* and *Supplementary Figure S9*). For mutation R137A *(Figure 5.8)* there were very few correlated movements between the two proteins, instead there were strong anti-correlated movements a few residues from the mutation, suggesting that it may have an allosteric effect. These anti-correlated motions suggest that the two proteins are moving away from each other and this agrees with experimental evidence as this mutation nearly abolished interaction between VP24 and KPNA5.

Principal component analysis for this complex with mutation R137A revealed a large change in the contributions to the variation from the first three principal components; 56.1% of the movement is explained by the first principal component compared to 33.4% for the wild type complex. Projection of the first two principal components onto those for the wild type complex demonstrates that the movement of the proteins differs (*Figure 5.7*).

The correlation analysis for the complex with F134A and M136A mutations identified that most correlated movements are intra chain, with very few correlated movements between the two proteins (*supplementary figure S9*). Again the presence of many anti-correlated movements between the two chains indicates that they are

moving apart and this is in agreement with experimental evidence that these mutations largely remove binding to KPNA5.Alanine 134 is located in a big web of anti-correlated movements, whereas Ala 136 is involved in correlated movements. Residue136 is mapped onto the cross correlation map. Principal Component Analysis (see *Table 5.3*) demonstrated that the proteins move away from each other (*Supplmentari figure S9*).

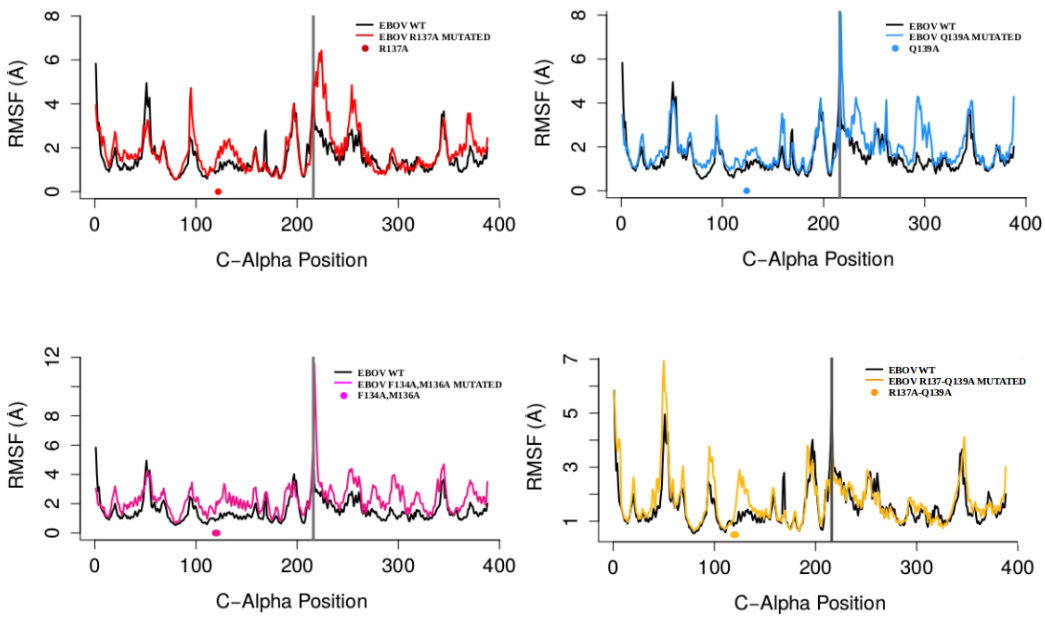


Figure 5.6: Root mean squared fluctuation of Ebola VP24-KPNA5 complex with point mutations. The dots under the lines represent the location of the mutations within protein VP24; the two protein in the complex are separated by a black line. A) RMSF mutation R137A is shown in red line and the EBOV wild type one. B) RMSF for mutation Q139A is shown in blue line and the EBOV wild type one. C) RMSF for mutations F134A,M136A are shown in magenta and the EBOV wild type one; D) RMSF for mutations R137A-Q139A are shown in yellow line.

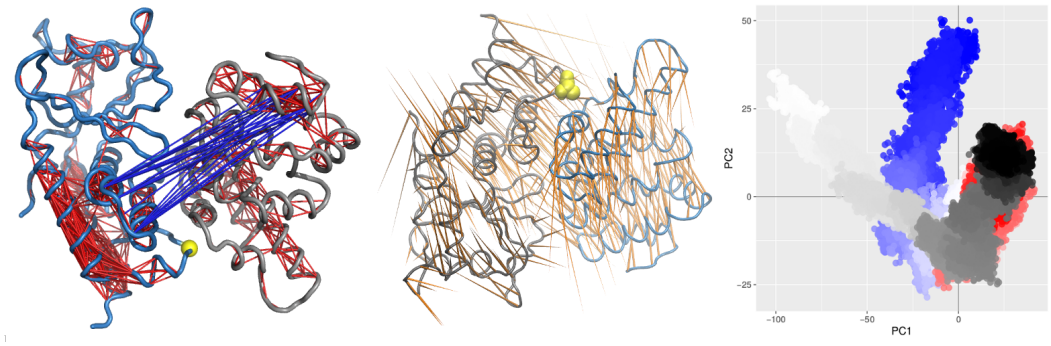


Figure 5.7. Molecular dynamics simulation of Ebola virus VP24 complex with KPNA5 with point mutations (R137A) in VP24. A) The cross correlation analysis is shown: in red lines the correlated movements and in blue lines the anticorrelated ones; protein VP24 is shown in blue cartoon and KPNA5 in gray cartoon; the mutation is shown in yellow sphere. B) Porcupine visualization of the Principal Component Analysis: protein VP24 is shown in gray cartoon and KPNA5 in blue cartoon; the mutation is shown in yellow spheres whereas the cones represent the amplitude of the movements according to the PCA. C) PCA projection of the wild type EBOV complex (from white to black) and for the mutation R137A (from blue to red).

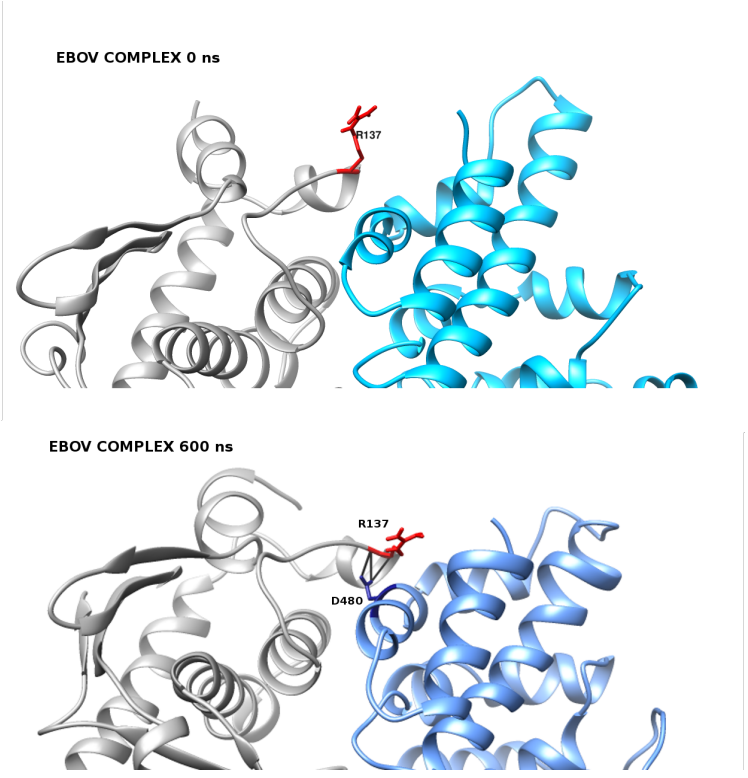


Figure 5.8: Residue R137 changes its conformation at zero (A) and at 600 ns (B). This last allows the interaction with KPNA5, giving a H-bond and a Salt Bridge with residue Asp 480.

Table 5.3: Eigenvalue Ranking

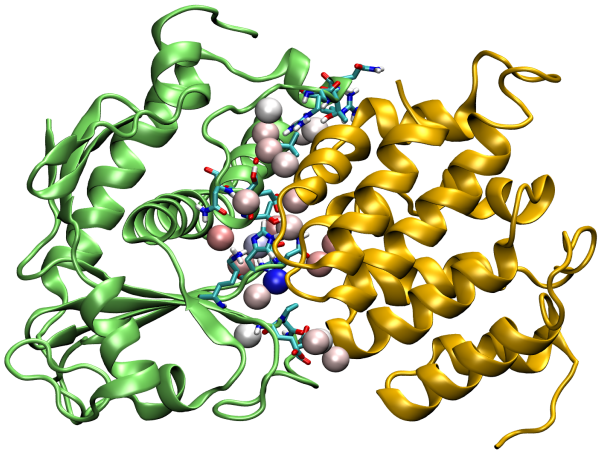
Eigenvalue Rank	EBOV - % of variance	R137A - % of variance	Q139A - % of variance	F134A-M136A - % of variance	R137A-Q139A - % of variance
1	33.4	56.1	28.8	50.30	27.3
2	48.1	64.2	51.1	65.8	37.6
3	54.9	67.1	61.9	72.1	46.5

5.3.5. Solvation properties at the interface

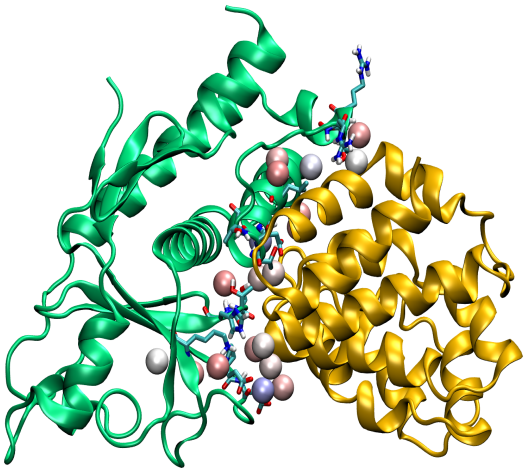
We calculated the solvation properties of the interface in the EBOV VP24 with human KPNA5 complex and in RESTV VP24 with human KPNA5 complex and estimated the water density on a grid of points constructed around the residues at the interface. We were interested in understanding how the water molecules were distributed at the interface and how they contributed to the binding of VP24 and KPNA5 (*Figure 5.3*). In *figure 5.9* the spheres represent the most visited grid points coloured from red to blue, with red being a lower value for the visited grid point and blue a higher number of water visits. In this way we could define the red spheres as density of "dynamical water" visits and the blue spheres as "permanent" water visits.
Our findings showed that in the EBOV complex residues N185 H186 E203 P204 and D205 are visited by permanent waters (*Figure 5.9A*). Additionally in the Reston complex we found residues at the interface visited by permanent waters E203 P204 D205 D124 and R137 (Figure 5.8B). This analysis revealed regions with permanent water visits in both the EBOV and RESTV complexes with overlap between both complexes (permanent waters at E203, P204 and D205 in both complexes). These residues belong to a loop interacting with KPNA5 defying a cavity where the water molecules are trapped.
Furthermore we performed the same analysis for the EBOV complex with the mutation R137A and we found that in this complex the interface is visited by "dynamical" waters only and no region solvated by permanent waters has been identified. This was due to the fact that, during the simulation, protein VP24 moved

apart from KPNA5 opening a cavity where the waters can enter and be dynamic due to the loss of physical restrictions (Figure 5.9C and Figure 5.10).

A)



B)



C)

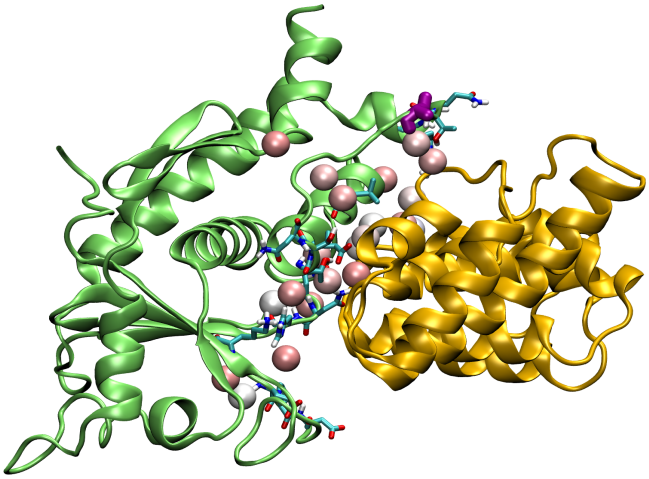


Figure 5.9: The spheres represent the most visited grid points coloured from red to blue, with red being a low value for the visited grid point and blue a high number of water visits. In this way we could define the red spheres as density of "dynamical water" visits and the blue spheres as "permanent" water visits. A) EBOV VP24 with human KPNA5 complex shows a presence of permanent waters that interact with N185 H186 E203 P204 and D205. B) RESTON VP24 with human KPNA5 complex shows a presence of permanent waters that interact with residues E203 P204 D205 D124 and R137. C) EBOV VP24 R137A with human KPNA5 complex shows a presence of dynamic waters only due to the opening of the cavity identify by the loop with the residues E203 P204 and D205.

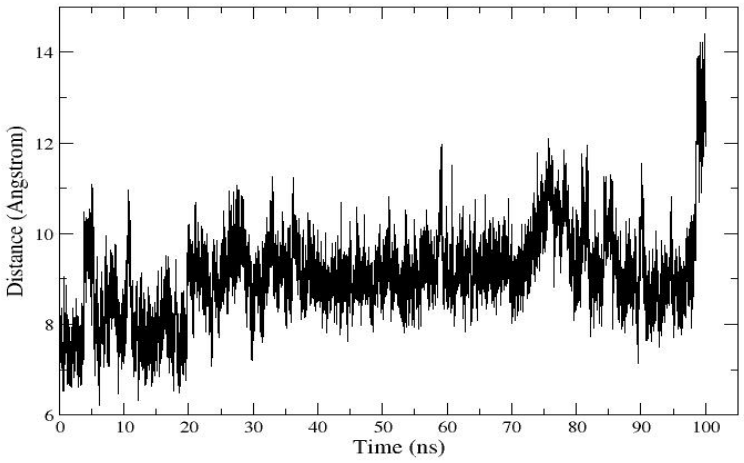


Figure 5.10: The distance over the time of D205 of VP24 with R396 of KPNA5. The starting distance is 8 $\mathring{A}$ and the final one 14 $\mathring{A}$. This increase in the distance shows the opening of the cavity.

5.4. Discussion

We started with a hypothesis that the conserved difference between Reston and Ebola virus VP24 proteins in the interface site with KPNA5 are likely to alter the interaction of Reston VP24 with KPNA5 compared to the interaction of Ebola VP24 with KPNA5. We have performed multiple analyses and simulations to gain insight into how this interaction may be altered. The molecular dynamics simulation

of the wild type complexes (*Figure 5.4*), indicated that there are greater fluctuations in KPNA5 when in complex with Reston VP24 than with Ebola VP24. This was further backed up by the cross-correlation analysis, which revealed more correlated movements in the Reston complex but also many that were anti-correlated.

The analysis of the complexes with mutations that significantly reduce Ebola VP24 binding with KPNA5 can be used to put these results into context. The cross correlation analysis for the complexes with F134A/M136A and R137A mutations contained many more anti-correlated movements and the proteins move away from each other (*Figure 5.7*, *figure S9*). In contrast while there are many anti-correlated movements between Reston VP24 and KPNA5, there are also many correlated movements. This may therefore suggest that there is greater interaction between these two proteins than the mutated Ebola VP24 proteins where binding is largely lost. It may be possible that such a change is possible to affect the ability of Reston viruses to prevent interferon signaling.

Chapter 6:

Discussion

This thesis has presented four pieces of work that are all related to genetic variation. Three of them focussed on analysis of genetic variants in Ebolaviruses with the aim of determining how they alter pathogenicity in different species. This chapter considers those chapters together and also compares the work in *Chapter three* with a similar study that also compares Ebola and Reston viruses.

6.1 Is protein VP24 responsible for Ebolavirus pathogenicity?

6.1.1 Combined analysis in our studies suggested that VP24 is a determinant for Ebolavirus pathogenicity.

Chapter three represents the beginning of our Ebolavirus research, which led onto the research detailed in chapters *four* and *five*. In combination these studies represent a comprehensive computational analysis of Ebolavirus genomes, their variation and the effects on the encoded proteins, ranging from analysis between different Ebolavirus species to mutations induced in adaptation experiments in rodents. The central theme throughout this research has been to understand molecular determinants of Ebolavirus pathogenicity.

The central finding in *Chapter three* was that there are very few differences between human pathogenic Ebolavirus species and Reston viruses (there are fewer than 200 SDPs) and the analysis pointed largely at VP24 as having a role in pathogenicity, due to the presence of multiple SDPs in the interface site with KPNA5. This hypothesis was supported by information from mutagenesis studies where Ebola virus VP24 interaction was disrupted by changes to residues that agree with some of the observed SDPs. However, the mutagenesis studies mutated pairs or trios of residues and each of these only partially overlaps with the SDPs.

This led into the research detailed in *Chapter five*, with the aim of using more detailed analysis of the VP24 and KPNA5 interface, particularly the use of molecular dynamics to study the interaction. This analysis supports our hypothesis in chapter three that the interaction between VP24 and KPNA5 differs for Ebola and Reston virus VP24. In *Chapter three* we proposed that the different amino acids present in Reston VP24 were likely to impair the interaction with KPNA5 and thereby prevent the virus from inhibiting the human interferon response. The molecular dynamics analysis of the Reston VP24 with KPNA5, supports our hypothesis; compared to the Ebola VP24 complex there are many more anti correlated movements between the two subunits. However, there are also correlated movements, overall suggesting that the two proteins may interact but with reduced affinity. The comparison of this simulation with simulations of Ebola virus VP24 that are known to disrupt binding further support this observation, as they clearly demonstrate the anti-correlated movements that are introduced between the two proteins. In effect we use this comparison to interpret the results of the simulation for Reston VP24 with KPNA5.

While *Chapter three* utilised the difference in human pathogenicity between Ebolavirus species, Ebolaviruses are not pathogenic in rodents. As presented in *Chapter four* this has enabled experiments in rodents (primarily Guinea pigs) to induce pathogenicity through multiple passaging of Ebola virus through multiple generations of test animals. Our analysis of the mutations present in these different studies highlighted that very few mutations may be required for adaptation of Ebola virus to a new species. This agrees with our analysis in *Chapter three*, where it seems that only a few variants may render Reston viruses non-pathogenic in humans.

Additionally analysis of the adaptation experiments highlighted mutations in VP24, with it being mutated in all of the four studies (Dowall et al., 2014; Ebihara et al.,2006; Cross et al.,2015; Volchkov et al., 2000).

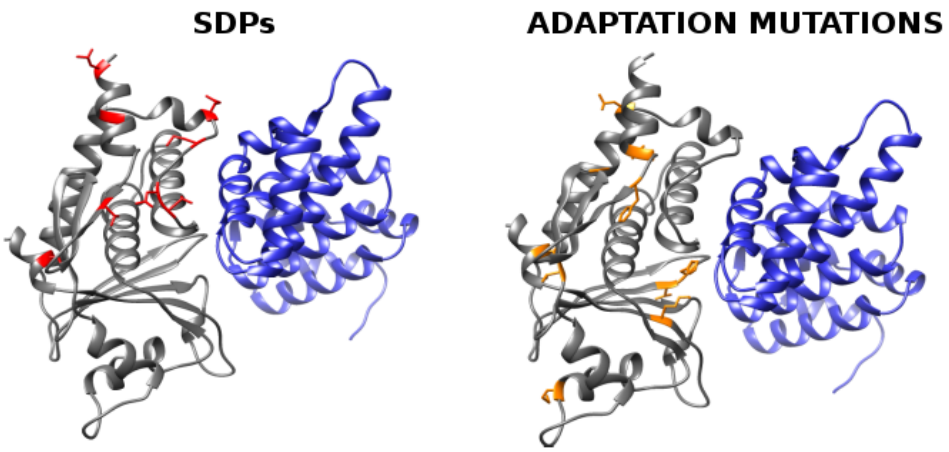


Figure 6.1: VP24 SDPs and adaptation mutations mapped into its complex with KPNA5. Protein VP24 is shown in gray cartoon and KPNA5 in blue. SDPs are shown in red sticks while adaptation mutations in yellow sticks.

The location of the VP24 SDPs and the mutations from the adaptation experiments were mapped onto the VP24 structure (*Figure 6.1*). This demonstrates that many of the adaptation mutations are in close proximity to the SDPs or are in the interface site with KPNA5 (e.g. T187I and H186Y). Additionally we observed that the SDPs and rodent adaptation mutations had similar effects by either altering hydrogen bonding with KPNA5 or removing hydrogen bonds within VP24. This observation further supports the argument that VP24 has an important role in determining pathogenicity.

The combination of the findings from *Chapters three*, four and five provides strong evidence for VP24 having an important role in determining host pathogenicity. It now remains for experimental validation of these findings, which is now being initiated by collaborators.

6.1.2 Comparison of Chapter 3 with Cong et al.,

Another study has also recently compared the genome sequences of Ebola and Reston viruses (Cong, Pei, & Grishin, 2015). Cong et al., used a total of 124 Ebolavirus genomes (compared to 196 that we considered). Our analysis identified SDPs between the human pathogenic species and the Reston species. Cong et al., used a similar approach, they identified identified positions in the proteins where there is greater conservation among the human pathogenic Ebolavirus species than between the Reston genomes. Using this approach they identified 215 differentially conserved positions. In contrast we identified a smaller number of SDPs 189. Analysis of the positions identified by the two studies indicates that the greater number of sequences used in our study removes some of the positions that classed as conserved by Cong et al.

Cong et al., also performed modelling of protein structures and mapping of the differentially conserved positions onto the structures. They used a different approach to us, using HHpred (Söding, et al., 2005) and iTASSER (Roy et al., 2010). They identified a model for part of the RNA-dependent RNA polymerase catalytic domain (L protein) and also a model for the N-terminal zinc finger domain of VP30. The template used to model L was not identified by Phyre2 when we performed modelling and this appears to be because the structure had just been released and may not have been added to the fold library (when modelling was later performed for the work in *Chapter four*, a template was identified and the model used in the analysis). Additionally the template used for the N-terminal domain of VP30 is of low quality with hhblits only returning a 52% probability that the query and template are homologous. Cong et al., propose that functional residues (i.e. the Zinc binding residues) are conserved therefore increasing the confidence that the template and query are homologous.

Comparison of the SDPs from *Chapter three* with the positions identified by Cong et al., demonstrated a considerable overlap of 133 positions, 6 in VP24, 16 in VP35, 16 in VP30, 7 in VP40, 19 in NP, 16 in GP and 53 in L (*Table 6.1*). Cong et al., did not consider sGP (without sGP we identified 180 SDPs) explaining some of the difference. The positions that were present in one study but not the other were

investigated to identify if there was an explanation for the different results. While many of the SDPs are completely conserved as one amino acid in Reston viruses and completely conserved as a different amino acid in the human pathogenic species, there are SDPs where there is a little variation in the amino acids observed. Comparing the positions between the two studies we found that such positions explain the different findings.

In our analysis, these positions would be less confident SDPs as they are not completely conserved in each group. So the different results obtained can be explained by both the different methods used and the different sets of sequences. Cong et al., used fewer sequences, so there will be some positions that are conserved in their set but in our larger set of sequences are more variable. The opposite is also true, some positions that are variable in the Cong et al., set, with more sequences present in our analysis, this variability could be reduced sufficiently for it to be predicted to be an SDP. Additionally, both studies used different methods to identify the differentially conserved positions, so there will be some positions that are predicted by one method but not the other regardless of the different sequences used. It is not possible to easily to split the effects of the different methods and sequences.

Table 6.1: Comparison of SDPs in our study and in Cong et al.

Protein	SDPs	Conserved in Cong et al.,	Total Number of common positions
VP24	L17M	-	6
	V22I	-	
	V31I	-	
	T131S	T131S	
	N132T	N132T	
	M136L	M136L	
	Q139R	Q139R	
	T226A	T226A	
	S248L	S248L	

VP30	- - - T52N V53L T63I E93D T96N R98H K107R S111I K116S - A120S - T150I Q157R I159L R196H E205D R262A S268Q	G20P V25S Y39R T52N V53L T63I E93D T96N R98H K107R S111I L116S N117Q A120S Q135S T150I Q157R - R196H E205D R262A S268Q	16
VP35	- - S26T E48D D76E - - E85K S92M V97T - T101N S106A V121I A154S T159V E160D G167K S174A I181L - E269D A290V - V314A Q329K	T5L L25T S26T E48D D76E C79Y N80V E85K S92M V97T Q98S - S106A - A154S T159V E160D G167K S174A - I258T E269D A290V A291P V314A Q329K	16

VP40	- T46V P85T I122V - G201N F209L Q245P H269Q - I293V - E325D	M14N T46V P85T - A128I G201N F209L Q245P H269Q T277Q - V323H E325D	7
NP	R4G - E16 S30T R39K P42S/Q42S - I56V V64I R105K M137L F212Y K274R S279A - K374R - K416N Y421Q D426E D435N - D443E T453I - D492E P497A - - P526 - T563S I565V P602T - N641Q - A705R - D716N G717N	R4G T15G - S30T R39K - I52M - - R105K M137L F212Y K274R S279A K373R K374R A411L K416N Y421Q D426E D435N Q442L D443E T453I V458A D492E - Q507S S511I - N551R T563S - - E633L - S647K A705R T714Y D716N -	19

GP	M1G G2S F31I V37I - V45A V75I - S196A - E207D S210T I260L - T269S - S308H/ L307H R325G - - - - H354L - - Q403P S418E - T448P - R498K R500K N514D Q521V L547V I584L D607S K622E - Q638H D642L W644L T569I	- - F31I - Q44K V45A - E156N S196A L199A - S210T - Y261R T269S T283P S307H - T335P E337T H339N E345T H354L E359T A361E - - A427M - G488K R498K R500K N514D - - - D607S K622E I627K Q638H D642L W644L T659I	16

L	V66T - Q109H - - - - I136L L146V - - - A221S Q223L H227Q - - - - L276I L283V Y312F A326S T330D - E350D T361S L365F V379I - Q447H P450S D465N - E689S S847A S868A F896Y L925F A954S S995T T1024N R1073K A1119S - - P1163A - D1189S A1214S R1217K D1237E - I1255V - -	V66T E93T Q109H N120A V128T E130I F132T - L146V L179F N201T T202I A221S Q223L H227Q V229L P262V V263D S274L - L283V Y312F A326S T330D S343Y E350D T361S L365F - I402N Q447H P450S D465N R654H E689S S847A S868A F896Y L925F A954S S995T T1024N R1073K A1119S Q1149P S1154L P1163A K1171D D1189S A1214S R1217K D1237E Q1253N - Y1322L R1354K	53

	T1366A	T1366A	
	S1395T	-	
	I1408M	I1408M	
	I1414L	-	
	S1436N	S1436N	
	K1461Q	K1461Q	
	S1473C	S1473C	
	L1488Y	L1488Y	
	I1499L	-	
	S1506A	S1506A	
	I1509V	-	
	R1534S	-	
	A1535K	-	
	-	A1538S	
	-	V1562L	
	-	E1564S	
	-	T1571K	
	-	Q1608I	
	-	H1619L	
	L1624Y	L1624Y	
	C1628S	C1628S	
	-	D1744G	
	-	E1752P	
	V1762I	-	
	-	S1769G	
	-	Q1782L	
	-	R1792H	
	-	W1822L	
	V1850Y	V1850T	
	T1873S	-	
	R1916N	R1916N	
	-	K1938Q	
	E1941R	E1941R	
	-	V1955Y	
	L2008I	-	
	-	Q2024G	
	L2044I	-	
	-	P2038V	
	S2077T	S2077T	
	-	K2078G	
	-	R2079L	
	E2098D	E2098D	
	Q2105L	Q2105L	
	Q2108E	Q2108E	
	Y2131F	Y2131F	
	L2157V	L2157V	
	R2168H	R2168H	
	R2175K	R2175K	
	L2177F	L2177F	
	M2186L	M2186L	
	-	L2203F	

Despite identifying a larger number of differentially conserved positions and modelling more of the Ebolavirus protein structures, Cong et al., mapped only 43 of the 215 positions they identified onto protein structures.

Cong et al., also focused on protein-protein interfaces and like our study identified six differentially conserved positions in interfaces. These include the differences present in VP24 that we propose may be relevant to the different pathogenicity observed between species. However, they propose that these differences may modify the binding between VP24 and KPNA5 but that this is likely to be limited to an effect on immune suppression that is unlikely to affect virus pathogenicity.

When considering protein-protein interfaces Cong et al., used the knowledge that Reston viruses are pathogenic in primates but not humans. So they considered the variability of the host proteins that Ebolaviruses interact with and investigated how these interaction partners vary between human and primates. They observed that host interaction partners of VP24, VP30 and VP40 are very similar between human and primates and therefore these proteins are unlikely to have a role in the different Ebolavirus pathogenicity. They found, there is greater sequence divergence in the host interaction partners of VP35 and GP. Based on this they identified two clusters of residues that they propose may alter Ebolavirus pathogenicity. The first cluster of differentially conserved residues is located in the C terminal region of GP and the second cluster is in VP35. We also identified same residues in GP, however we were cautious about interpreting their possible effect as their function is unknown and while they are present in the glycan cap none of the residues are glycosylation sites or close to glycosylation sites. This made it difficult to interpret how they may alter GP function and pathogenicity.

The VP35 cluster of residues identified by Cong et al., consists of A290V, A291P, V314A and Q329K. With the exception of A291P, these positions were also identified in our study, we observe variability between the human pathogenic species at position 291 and it is therefore not predicted to be an SDP. These changes had previously been identified in experimental research (Leung et al., 2015), and are

thought to stabilise the protein structure. The experimental study also observed reduced binding of VP35 to dsRNA and weaker inhibition of interferon signalling (Leung et al., 2015). The authors of this study thought that these effects were unlikely to explain the lack of Reston virus pathogenicity in humans (Leung et al., 2015).

In summary, both studies used very similar approaches but resulted in different interpretations. Our analysis highlighted VP24 as the availability of a complex structure with a host protein provided good evidence. If such data had been available for the other Ebolavirus proteins it is possible that other positions would have been identified that are likely to alter pathogenicity. While neither study is conclusive, they both provide avenues for wet lab experiments to validate the hypotheses.

6.2 Limitations of this study

Much of this thesis focusses on analysing Ebolavirus genetic variation. *Chapter three* identified a set of 189 SDPs, a subset of which are likely to explain the difference in human-pathogenicity between Reston viruses and the other four Ebolavirus species. Structural analysis was only able to map 47 of these SDPs onto protein structures. This initially limits the ability to analyse approximately three quarters of the SDPs identified. So while our structural analysis has identified a number of candidate SDPs for association with pathogenicity, it possible that others that it was not possible to analyse also have a role. Further determination of Ebolavirus protein structures or the availability of homologues to use as templates will reduce this problem. However, it is predicted that approximately 20% of the Ebolvirus proteins are disordered (Cong et al., 2015), so for some SDPs it may never be possible to model their effect on protein structure.

Additionally the analysis in *Chapters three to five* is limited by our knowledge of the biology of Ebolaviruses. Our understanding of their function is still limited, although there has been a surge in Ebola related publications since the 2014 outbreak (Michaelis et al., 2016). Again as our understanding of Ebolavirus biology and protein function advances, the number of potential molecular determinants of Ebolavirus pathogenicity will be reduced.

This thesis presents purely computational research and as such demonstrates the strengths of such analyses to provide insight into large scale genomic data. However, this also means that the analysis leaves many findings that require experimental validation.

6.3 Future Work

The research in this thesis presents a number of hypotheses that need to be tested. These are detailed below:

1 There are now many more Ebola virus sequences available (Pickett et al., 2012). These datasets provide approximately a further 506 sequences. The analysis performed in *Chapter three* could be repeated using this much larger dataset. This would provide much greater detail on variation present within the Ebola virus genome and could reduce the number of SDPs identified, thus enabling us to exclude some of the potential explanations for altered pathogenicity identified in chapter three.

2 Extensive molecular dynamics simulations were performed on the VP24 interaction with KPNA5. However, these could be expanded to investigate the affinity of the Ebola and Reston VP24 with KPNA5 using "pulling apart" experiments, where the two molecules are pulled apart to measure the affinity between them. Such experiments are computationally expensive and could not be performed in this current analysis.

3 Although much of the research has pointed to VP24, chapter three identified SDPs in other proteins, including VP40, VP35 and GP that could have an effect on protein function and therefore pathogenicity. These could also be experimentally investigated.

4 Considering the role of protein VP24 in interfering with IFN signaling inhibition it will be interesting to look at sequence changes also in the partner protein KPNA5 in rat, hamster and pigs. This will advance our knowledge and could shed light on the mechanism of pathogenicity among Ebolaviruses.

5 We have proposed that VP24 has an important role in determining pathogenicity and these findings could be experimentally investigated. While Ebola is a category four pathogen, it is possible to perform in vitro experiments with individual Ebolavirus proteins, making such studies feasible. Ultimately such experimental work is required to test the hypotheses made in this thesis. For example testing the ability of Reston VP24 to bind human karyopherin proteins, would test the proposal that mutations in Reston VP24 alter binding to karyopherins. Similar experiments could be performed to test the effects of mutations in VP24 that occur during Ebola virus adaption experiments in rodents. Does the wild type Ebola VP24 bind rodent karyopherins and is there greater affinity with the mutated forms of VP24? This research has now started in Jeremy Rossman's laboratory at the University of Kent.

References

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature*, 491:56–65

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., … McVean, G. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73. doi:10.1038/nature09534

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a, Durbin, R. M., Handsaker, R. E., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Natur*e, 491(7422): 56–65. doi:10.1038/nature11632

Abraham, M. J., Murtola, T., Schulz, R., Páall, S., Smith, J. C., Hess, B., & Lindah, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. doi:10.1016/j.softx.2015.06.001

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* / editorial board, Jonathan L. Haines ... [et al.] (Vol. Chapter 7). doi:10.1002/0471142905.hg0720s76

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*. 7:248–249.

Akerlund, E., Prescott, J. & Tampellini, L. (2015). Shedding of Ebola Virus in an Asymptomatic Pregnant Woman. *N. Engl. J. Med.* **372,** 2467–2469

Alexander, K.A., Sanderson, C.E., Marathe, M., Lewis, B.L., et al. (2015) What factors might have led to the emergence of Ebola in West Africa? *PLoS neglected tropical diseases*. 9 (6), e0003652. doi:10.1371/journal.pntd.0003652.

Altman, R. B. et al., (2012). Principle of Pharmacogenetics and Pharmcogenomics. New York, NY: Cambridge university Press

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25:3389–3402.

Altshuler, D. M., Gibbs, R. a, Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8. doi:10.1038/nature09298

Atherstone, C., Smith, E., Ochungo, P., Roesel, K., et al. (2015). Assessing the Potential Role of Pigs in the Epidemiology of Ebola Virus in Uganda. *Transboundary and emerging diseases*. doi:10.1111/tbed.12394.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. doi:10.1038/nature15393

Azarian, T., Presti, Lo, A., Giovanetti, M., Cella, E., Rife, B., Lai, A., et al. (2015). Impact of spatial dispersion, evolution, and selection on Ebola Zaire Virus epidemic waves. *Scientific Reports*. 5, 10170.

Bale, S., Julien, J.P., Bornholdt, Z.A., Krois, A.S., Wilson, I.A., Saphire, E.O.(2013). Ebolavirus VP35 coats the backbone of double-stranded RNA for interferon antagonism. *Journal of Virology*. 87:10385–10388

Barrette, R.W., Metwally, S.A., Rowland, J.M., Xu, L., et al. (2009). Discovery of swine as a host for the Reston ebolavirus. *Science*. 325 (5937), 204–206. doi:10.1126/science.1172705.

Basler, C. F. (2014). Portrait of a killer: genome of the 2014 EBOV outbreak strain. *Cell Host Microbe* 16:419–421

Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J., Kelley, L.A. (2008). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins*. 70:611–625.

Binning, J. M., Wang, T., Luthra, P., Shabman, R. S., Borek, D. M., Liu, G., … Amarasinghe, G. K. (2013). Development of RNA aptamers targeting Ebola virus VP35. *Biochemistry*, *52*(47):8406–8419. doi:10.1021/bi400704d

Birney, E., Stamatoyannopoulos, J. a, Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816. doi:10.1038/nature05874

Bordner, A. J., & Zorman, B. (2013). Predicting non-neutral missense mutations and their biochemical consequences using genome-scale homology modeling of human protein complexes, arXiv:1308.4433.

Bornholdt, Z. A., Noda, T., Abelson, D.M., Halfmann, P., Wood, M.R., Kawaoka, Y., Saphire, E.O. (2013). Structural rearrangement of ebola virus VP40 begets multiple functions in the virus life cycle. *Cell* 154:763–774

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu C.H., Xie, D., Suchard, M.A. et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., Jones, D.T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research* . 41 (W1): W340-W348.

Buchanan, C. C., Torstenson, E. S., Bush, W. S., & Ritchie, M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association*: JAMIA, 19(2):289–94. doi:10.1136/amiajnl-2011-000652

Burke, D. F., Worth, C. L., Priego, E.-M., Cheng, T., Smink, L. J., Todd, J. a, & Blundell, T. L. (2007). Genome bioinformatic analysis of nonsynonymous SNPs. BMC *Bioinformatics*, 8:301. doi:10.1186/1471-2105-8-301

Capra, J. M., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882. doi:10.1093/bioinformatics/btm270

Capra, J.M., Singh, M. (2008). Characterization and prediction of residues determining protein functional. *Bioinformatics.* 24:1473–1480.

Cárdenas, W. B., Loo, Y.-M., Gale, M., Hartman, A. L., Kimberlin, C. R., Martínez-Sobrido, L., … Basler, C. F. (2006). Ebola virus VP35 protein binds double-stranded RNA and inhibits alpha/beta interferon production induced by RIG-I signaling. *Journal of Virology*, *80*(11):5168–78. doi:10.1128/JVI.02199-05

Carroll, M.W., Matthews, D.A., Hiscox, J.A., Elmore, M.J., Pollakis, G., Rambaut, A., et al. (2015). Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*. 524(7563):97-101. doi:10.1038/nature14594

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Structural Biology* 2:171–178

Cavallo, L., Kleinjung, J., Fraternali, F. (2003), POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acid Research* 31: 3364-3366

Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L., et al. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nature Genetics*. 41(11):1170-2

Chambers, J.C., Zhang, W., Lord, G.M., Van der Harst, P., Lawlor, D.A., Sehmi, J.S., Gale, D.P., Wass, M.N., Ahmadi, K.R., Bakker, S.J.L., et al. (2010). Genetic loci influencing kidney function and chronic kidney disease. *Nature Genetics* 42:373–375.

Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E., et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genetics*. 43:1131–1138.

Chambers, J. C. *et al.* (2014). The South Asian genome. *PLoS ONE* 9:e102645

Christie, A., Davies-Wayne, G.J., Cordier-Lassalle, T., Cordier-Lasalle, T., Blackley, D.J., Laney, A.S., et al. (2015). Possible sexual transmission of Ebola virus - Liberia, 2015. *The*

Morbidity and Mortality Weekly Report. 64:479– 481.

Chubb, D., Jefferys, B.R., Sternberg, M.J., Kelley, L.A. (2010). Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics*. 26:2664–2671.

Chun S., Fay J.C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*. 19:1553–1561

Church, G. M. (2005). The personal genome project. *Molecular Systems Biology*, 1:2005.0030. doi:10.1038/msb4100040

Cirulli, E.T. & Goldstein, D.B., (2010), Uncovering the roles of rare variants in common disease through whole-genome sequencing., Nature Review Genetics.,11(6):415-25. doi: 10.1038/nrg2779

Clifton, M. C., Kirchdoerfer, R.N., Atkins, K., Abendroth, J., Raymond, A., Grice, R. et al. (2014). Structure of the Reston ebolavirus VP30 C-terminal domain. *Acta Crystallographica Section F Structural Biology Communication*. 70:457–460

Cong, Q., Pei, J., & Grishin, N. V. (2015). Predictive and comparative analysis of Ebolavirus proteins. *Cell Cycle (Georgetown, Tex.)*, *4101*(October): 1–13. doi:10.1080/15384101.2015.1068472

Cooper, G. M., Stone, E. a., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913. doi:10.1101/gr.3577405

Cranage, A. (2015). Genomics England and the 100,000 Genomes Project, (April), 1–7. Retrieved from http://www.genomicsengland.co.uk/the-100000-genomes-project/\npapers3://publication/uuid/E6E41B7D-069A-425E-A3D1-EFC497D0D7AF

Cross, R. W., Fenton, K. a., Geisbert, J. B., Mire, C. E., & Geisbert, T. W. (2015). Modeling the Disease Course of Zaire ebolavirus Infection in the Outbred Guinea Pig. *Journal of Infectious Diseases*, *212*:S305–S315. doi:10.1093/infdis/jiv237

Dahlmann, F., Biedenkopf, N., Babler, A., Jahnen-Dechent, W., Karsten, C.B., Gnirb, K. et al. (2015). Analysis of Ebola Virus Entry Into Macrophages. *Journal of Infectious Disease*. jiv140 (2015). doi:10.1093/infdis/jiv140

Daly, A. K. (2012). Using genome-wide association studies to identify genes important in serious adverse drug reactions. *Annual Review of Pharmacology and Toxicology*, 52, 21–35. doi:10.1146/annurev-pharmtox-010611-134743

David, A., & Sternberg, M. J. (2015). The contribution of missense mutations in core and Rim residues of protein-protein interfaces to human disease. *Journal of Molecular Biology*, 427(17):2886–98. doi:10.1016/j.jmb.2015.07.004

David, A., Razali, R., Wass, M. N., & Sternberg, M. J. E. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation*, 33(2):359–363. doi:10.1002/humu.21656

Deen, G.F., Knust, B., Broutet, N., Sesay, F.R., Formenty, P., Ross, C., et al. (2015).

Ebola RNA Persistence in Semen of Ebola Virus Disease Survivors - Preliminary Report. *The New England Journal of Medicine.*

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Harti, C. et al., (2011), A framework for variation discovery nd genotyping using next generation DNA sequencing data. *Nature Genetics* 43: 491-8;

Dowall, S. D., Matthews, D. a, Garcia-Dorival, I., Taylor, I., Kenny, J., Hertz-Fowler, C. et al. (2014). Elucidating variations in the nucleotide sequence of Ebola virus associated with increasing pathogenicity. *Genome Biology*, *15*(11):540. doi:10.1186/PREACCEPT-1724277741482641

Ebihara, H., Takada, A., Kobasa, D., Jones, S., Neumann, G., Theriault, S. et al. (2006). Molecular determinants of Ebola virus virulence in mice. *PLoS pathogens.* 2 (7):e73. doi:10.1371/journal.ppat.0020073.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research.* 2004;32:1792–1797.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Review Genetics* ;11:446–450.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) Accessed in March 2016.

Feldmann, H. & Geisbert, T.W. (2011). Ebola haemorrhagic fever. *Lancet.* 377 (9768):849–862. doi:10.1016/S0140-6736(10)60667-8.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R. Et al. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue), D281–8. doi:10.1093/nar/gkm960

Frazer, K., Ballinger, D. G., Cox, D. R., Hinds, D. a, Stuve, L. L., Gibbs, R. A. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61. doi:10.1038/nature06258

Frieden, T.R., Damon, I., Bell, B.P., Kenyon, T., Nichol, S. (2014). Ebola 2014--new challenges, new global response and responsibility. *The New England journal of medicine.* 371 (13):1177–1180. doi:10.1056/NEJMp1409903.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H. et al. (2003). The international HapMap project. *Nature*, 426(6968):789–796.

Gire, S.K., Goba, A., Andersen, K.G., Sealfon, R.S.G., Park, D.J., Kanneh, L., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science.* 345:1369–1372.

Grant, R., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., Caves, L.S. (2006) Bio3D: An R package for the comparative analysis of protein structures. *Bioinformatics* **22:**2695-2696;

Grarup, N., Sandholt, C. H., Hansen, T., & Pedersen, O. (2014). Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond.

Diabetologia. doi:10.1007/s00125-014-3270-4

Gray, C.M., Addo, M., Schmidt, R.E.Clinical Immunology Committee of the IUIS (2014). A dead-end host: is there a way out? A position piece on the ebola virus outbreak by the international union of immunology societies. *Frontiers in immunology*. 5 (5):562. doi:10.3389/fimmu.2014.00562.

Groseth, A., Marzi, A., Hoenen, T., Herwig, A., et al. (2012). The Ebola virus glycoprotein contributes to but is not sufficient for virulence in vivo. *PLoS pathogens*. 8 (8):e1002847. doi:10.1371/journal.ppat.1002847.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. a, & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–7. doi:10.1093/nar/gki033

Harrison, S.M., Riggs, E.R., Maglott, D.R., Lee, J.M., Azzariti, D.R., Niehaus, et al. (2016). Using ClinVar as Resource to Support Variant Interpretation.Current Protocol in Human Genetics, 1;89:8.16.1-8.16.23, doi:10.1002/0471142905.hg0816s89.

Hartlieb, B., Modrof, J., Mühlberger, E., Klenk, H.-D. & Becker, S. (2003). Oligomerization of Ebola virus VP30 is essential for viral transcription and can be inhibited by a synthetic peptide. *Journal of Biological Chemistry*. 278:41830–41836

Hartlieb, B., Muziol, T., Weissenhorn, W. & Becker, S. (2007). Crystal structure of the C-terminal domain of Ebola virus VP30 reveals a role in transcription and nucleocapsid association. *Proceedings of the National. Academy of. Sciences. U.S.A.* 104:624–629

Hattori, M. (2005). Finishing the euchromatic sequence of the human genome. Tanpakushitsu Kakusan Koso. P*rotein, Nucleic Acid, Enzyme,* 50(2):162–168. doi:10.1038/nature03001

Herbert, A. S., Davidson, C., Kuehne, A.I., Bakken, R., Braigen, A.Z., Gunn, K.E. et al. (2015) Niemann-pick c1 is essential for ebolavirus replication and pathogenesis in vivo. *MBio* 6**:**e00565–15

Hoenen, T., Marzi, A., Scott, D.P., Feldmann, F., Callison, J., Safronetz, D. et al. (2015). Soluble Glycoprotein Is Not Required for Ebola Virus Virulence in Guinea Pigs. *Journal of Infectious Disease* jiv111 doi:10.1093/infdis/jiv111

Hoenen, T., Safronetz, D., Groseth, A., Wollenberg, K.R., Koita, O.A., Diarra, B., et al. (2015) Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science*. 348:117–119.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. Nature Reviews. *Drug Discovery*, 1(9), 727–30. doi:10.1038/nrd892

Ilinykh, P.A., Lubaki, N.M., Widen, S.G., Renn, L.A., et al. (2015). Different temporal effects of Ebola virus VP35 and VP24 proteins on the global gene expression in human dendritic cells. *Journal of Virology*. JVI.00924–15. doi:10.1128/JVI.00924-15.

Innis, C.A., Anand, A.P., Sowdhamini, R. (2004). Prediction of functional sites in proteins using

conserved functional group analysis. *Journal of Molecular Biology* 337:1053–1068.

International Commission Report. (1976). Ebola haemorrhagic fever in Zaire. Bulletin *World Health Organ*. 56:271-93

International HapMap Consortium (2003). The International HapMap Project. *Nature*. 426:789-96

International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 449:851-61

Joachims, T. (1999). Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, editors. *Advances in Kernel Methods—Support Vector Learning.*

Joosten, R.P., Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider R., Sander C., Vriend G. (2001). A series of PDB related databases for everyday needs. *Nucleic Acids Research.* 39:D411–D419.

Jun, S.-R., Leuze, M.R., Nookaew, I., Uberbacher, E.C., et al. (2015). Ebolavirus comparative genomics. Urs Greber (ed.). *FEMS Microbiol Review.* 39 (5):764–778. doi:10.1093/femsre/fuv031.

Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22:2577–2637.

Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A. et al. (2012) Gene Expression Atlas update - a value- added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Research.* 40:D1077-D1081.

Karczewski, K. J., Daneshjou, R., & Altman, R. B. (2012). Chapter 7: pharmacogenomics. *PLoS Computational Biology*, 8(12):e1002817. doi:10.1371/journal.pcbi.1002817

Katsnelson, A. (2013). Momentum grows to make " personalized " medicine more " precise ." *Nature.* 19(3):249. doi:10.1038/nm0313-249

Keinan, A, Clark, A.G. (2013). Excess of Rare Genetic Variants. *Science.* 336(6082):740–743. doi:10.1126/science.1217283.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocol.* 10:845–858

Kelley, L., & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols.* 4(3):363–71. doi:10.1038/nprot.2009.2

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. (2002). The human genome browser at UCSC. *Genome Research.* 12(6):996-1006

Kimberlin, C. R., Bornholdt, Z.A., Li, S., Woods, V.L., MacRae, I.J., Saphire E.O. (2010). Ebolavirus VP35 uses a bimodal strategy to bind dsRNA for innate immune suppression. *Proceedings the National. Academy of Science. U.S.A.* 107:314–319

Kleinjung, J. & Fraternali, F. (2005) POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acid Research*, 33:W342-W346. doi:10.1093/nar/gki369

Kryukov, G. V., Pennacchio, L., Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in

humans: implications for complex disease and association studies. *American Journal of Human Genetics*, 80(4):727–39. doi:10.1086/513473

Kuhn, J. H., Becker, S., Ebihara, H., Geisbert, T.W., Johnson, K.M., Kawaoka, Y. et al. (2010). Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. *Archives of Virology* **155,** 2083–2103

Kumar, P., Henikoff, S., Ng, C.P. (2013). Predicting the effects of coding non-synonymous variants on protein function using SIFT algorithm. *Nature Protocol*s. 4:1073-1081

Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* 372:774-797.

La Vega de, M.-A., Stein, D., Kobinger, G.P. (2015). Ebolavirus Evolution: Past and Present. *PLoS Pathogens.* 11:e1005221.

La Vega, de, M.-A., Wong, G., Kobinger, G. P. & Qiu, X. (2015). The multiple roles of sGP in Ebola pathogenesis. *Viral Immunology* 28:3–9

Lander, E. S., Heaford, A., Sheridan, A., Linton, L. M., Birren, B., Subramanian, A. et al. (2001). Initial sequencing and analysis of the human genome. *Nature.* 409(6822):860–921. doi:10.1038/35057062

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3):168–78. doi:10.1038/nrg3404

Lennemann, N. J., Rhein, B. a., Ndungo, E., Chandran, K., Qiu, X., & Maury, W. (2014). Comprehensive functional analysis of N-linked glycans on ebola virus GP1. *Journal of Molecular Biology. 5*(1):1–9. doi:10.1128/mBio.00862-13

Leung, D. W., Ginder, N.D., Fulton, D.B., Nix, J., Basler, C.F., Honzatko, R.B. et al. (2009). Structure of the Ebola VP35 interferon inhibitory domain. *Proceedings of the National Academy of Science. U.S.A.* 106:411–416

Leung, D. W., Shabman, R.S., Farahbakhsh, M., Prins, K.C., Borek, D.M. et al. (2010). Structural and functional characterization of Reston Ebola virus VP35 interferon inhibitory domain. *Journal of Molecular Biology.* 399:347–357

Leung, D. W., Borek, D., Luthra, P., Binning, J.M., Anantpadma, M., Liu, G. et al. (2015). An Intrinsically Disordered Peptide from Ebola Virus VP35 Controls Viral RNA Synthesis by Modulating Nucleoprotein-RNA Interactions. *Cell Reports.* doi:10.1016/j.celrep.2015.03.034

Li, J., Jia, J., Li, H., Yu, J., Sun, H., He, Y. et al. (2014). SysPTM 2.0: an updated systematic resource for post-translational modification. Database: The *Journal of Biological Databases and Curation*, 2014, bau025. doi:10.1093/database/bau025

Lichtarge, O., Bourne, H.R., Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. Journal of Molecular Biology. 29;257(2):342-58.

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., Heckerman, D. (2013). The benefits of selecting phenotype-specific variants for applications of mixed models in genomics.

Scientific Reports, 3:1815. doi:10.1038/srep01815

Liu, D. J., & Leal, S. M. (2012). A unified method for detecting secondary trait associations with rare variants: application to sequence data. *PLoS Genetics*, 8(11):e1003075. doi:10.1371/journal.pgen.1003075

Liu, S.-Q., Deng, C.-L., Yuan, Z.-M., Rayner, S. & Zhang, B. (2015). Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa. *Infection Genetics and Evolution.* 32:51–59

Liu, X., Jian, X., and Boerwinkle, E. (2011), dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. *Human Mutation.* 32:894-899.

Liu, X., Jian, X., and Boerwinkle, E., 2013. dbNSFP v2.0: A Database of Human Nonsynonymous SNVs and Their Functional Predictions and Annotations. *Human Mutation.* 34:E2393-E2402

Lobanov, M., Bogatyreva, N.S., Galzitskaia, O.V. (2008). Radius of gyration is indicator of compactness of protein structure. *Molecular Biology.* (Mosk), 42:701-6

Lord, J., Lu, A.J., Cruchaga, C. (2014). Identification of rare variants in Azheimer's disease. Frontiers in Genetics. 28;5:369.doi: 10.3389/fgene.2014.00369. ECollection 2014.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter D.J. (2009), Finding the missing heritability of complex diseases. *Nature.* 461:474-53

Marsh, G.A., Haining, J., Robinson, R., Foord, A., et al. (2011). Ebola Reston virus infection of pigs: clinical significance and transmission potential. *The Journal of infectious diseases.* 204 Suppl 3 (suppl 3), S804–S809. doi:10.1093/infdis/jir300.

Mate, S.E., Kugelman, J.R., Nyenswah, T.G., Ladner, J.T., Wiley, M.R., Cordier-Lassalle, T., et al. (2015) Molecular Evidence of Sexual Transmission of Ebola Virus. *The New England Journal of Medicine.* 373:2448–2454.

Mateo, M., Carbonnelle, C., Martinez, M.J., Reynard, O., et al. (2011). Knockdown of Ebola virus VP24 impairs viral nucleocapsid assembly and prevents virus replication. *The Journal of infectious diseases.* 204 Suppl 3 (suppl 3):S892–S896. doi:10.1093/infdis/jir311.

Mateo, M., Carbonnelle, C., Reynard, O., Kolesnikova, L., et al. (2011). VP24 is a molecular determinant of Ebola virus virulence in guinea pigs. *The Journal of infectious diseases.* 204 Suppl 3S1011–S1020. doi:10.1093/infdis/jir338.

Mateo, M., Reid, S. P., Leung, L. W., Basler, C. F., & Volchkov, V. E. (2010). Ebolavirus VP24 binding to karyopherins is required for inhibition of interferon signaling. *Journal of Virology*, *84*(2):1169–75. doi:10.1128/JVI.01372-09

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J. Ioannidis, J.P. et al. (2008), Genome-Wide Association studies for complex traits: consensus, uncetainty and challenges. *Nature Review Genetics.*

McKusick, V. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics,* 80(4):588–604. doi:10.1086/514346

Mehedi, M., Falzarano, D., Seebach, J, Hu, X., Carpenter, M.S., Schinttler H.J. et al. (2011). A new Ebola virus nonstructural glycoprotein expressed through RNA editing. *Journal of Virology*. 85:5406–5414

Mehedi, M., Hoenen, T., Robertson, S., Ricklefs, S., Dolan, M. a., Taylor, T. et al. (2013). Ebola Virus RNA Editing Depends on the Primary Editing Site Sequence and an Upstream Secondary Structure. *PLoS Pathogens*, 9(10). doi:10.1371/journal.ppat.1003677

Meltzer, M.I., Atkins, C.Y., Santibanez, S., Knust, B., et al. (2014). Estimating the future number of cases in the Ebola epidemic--Liberia and Sierra Leone, 2014-2015. *Morbidity and Mortality Weekly Report. supplements*. 63 (3):1–14.

Michaelis, M., Rossman, J.S., Wass MN. (2016) Computational Analysis of Ebolavirus data: prospect, promises and challenges. *Biochemical Society Transaction*. Under review.

Miller, E. H., Obernosterer, G., Raaben, M., Herbert, A. S., Deffieu, M. S., Krishnan, A. et al. (2012). Ebola virus entry requires the host-programmed recognition of an intracellular receptor. *The EMBO Journal*, 31(8):1947–1960. doi:10.1038/emboj.2012.53

Miranda, M. E. G. & Miranda, N. L. J. (2011). Reston ebolavirus in humans and animals in the Philippines: a review. *Journal of Infectious Disease* 204 Suppl 3:S757–60

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41:e121

Morikawa, S., Saijo, M. & Kurane, I. (2007). Current knowledge on lower virulence of Reston Ebola virus (in French: Connaissances actuelles sur la moindre virulence du virus Ebola Reston). *Comparative Immunology, Microbiology and Infectious Diseases* 30:391–398

Mosca, R., Céol, A., Aloy, P. (2012). Interactome3D: adding structural details to protein networks. *Nature Methods*, 10(1):47–53. doi:10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(September 2013):374–379. doi:10.1093/nar/gkt887

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C. et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*. 337(6090):100–4. doi:10.1126/science.1217876

Ng, P.G., Murrai, S.S., Levy, S., Venter, J.C. (2009), An agenda for personalized medicine. *Nature*. 461:724-6

Nock, N., & Zhang, L. (2011). Evaluating aggregate effects of rare and common variants in the 1000 Genomes Project exon sequencing data using latent variable structural equation modeling. *BMC Proceedings*, 5 Suppl 9(Suppl 9):S47. doi:10.1186/1753-6561-5-S9-S47

Nussinov, R., Tsai, C. J., Xin, F., & Radivojac, P. (2012). Allosteric post-translational modification codes. *Trends in Biochemical Sciences*, 37(10):447–455. doi:10.1016/j.tibs.2012.07.001

Olson, S.H., Reed, P., Cameron, K.N., Ssebide, B.J., et al. (2012). Dead or alive: animal sampling during Ebola hemorrhagic fever outbreaks in humans. *Emerging health threats journal.* 5 (0):221. doi:10.3402/ehtj.v5i0.9134.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F. et al. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):358–363. doi:10.1093/nar/gkt1115

Orgogozo, V., Morizot, B., & Martin, A. (2015). The differential view of genotype-phenotype relationships. *Frontiers in Genetics*, 6(MAY):1-14. doi:10.3389/fgene.2015.00179

Osterholm, M.T., Moore, K.A., Kelley, N.S., Brosseau, L.M., et al. (2015). Transmission of Ebola viruses: what we know and what we do not know. *Journal of Molecular Biology.* 6 (2):e00137. doi:10.1128/mBio.00137-15.

Palles, C., Cazier, J.B., Howart, K.M., Doming, E., Jones, A.M., Broderick, P. et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics* 45**:**136–144

Pan, Y., Zhang, W., Cui, L., Hua, X., Wang, M., Zeng, Q. (2014). Reston virus in domestic pigs in China. *Archives of Virology.* 159 (5):1129–1132. doi:10.1007/s00705-012-1477-6.

Pappalardo & Wass, (2014), "VarMod: modelling the functional effects of non-synonymous variants", *Nucleic Acids Research*, 42(Web Server issue):W331–W336.

Pappalardo, M., Juliá, M., Howard, M.J., Rossman, J.S., et al. (2016). Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. *Scientific Reports.* 623743. doi:10.1038/srep23743.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20**:**289–290

Parsons, C., Naeem Ahmad, U. (2015). The West African Ebola outbreak: finishing the job, preparing for future. *Transaction of the Royal Society of Tropical Medicine and Hygiene.* . 109:481–482.

Pattyn, S., Van der Groen, G., Jacob, W., Piot, P., Courteille, G. (1977). Isolation of Marburg-like virus from a case of haemorrhagic fever in Zaire. *Lancet.* 1:573–574.

Peterson, T. A, Doughty, E., & Kann, M. G. (2013). Towards precision medicine: advances in computational approaches for the analysis of human variants. J*ournal of Molecular Biology*, 425(21):4047–63. doi:10.1016/j.jmb.2013.08.008

Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N. a., Gonzalez-Porta, M., Hastings, E. et al. (2014). Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 42(D1):926–932. doi:10.1093/nar/gkt1270

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational*

Chemistry. 2004 Oct;25(13):1605-12.

Pickett, B. E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V. et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research* 40:D593–8

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):335–342. doi:10.1093/bioinformatics/btt691

Platt, J., Cambridge, M.A. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods; MIT Press; pp. 61–74.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics,* 69(1):124–37. doi:10.1086/321272

Pundir, S., Martin, M.J.,O'Donovan C. (2016). UniProt Consortium. Uniprot Tools. Curr Protoc Bioinformatics. Mar 24;53:1.29.1-1.29.15.doi:10.1002/0471250953.bi0129s53.

Quaglio, G., Goerens, C., Putoto, G., Rübig, P., Lafaye, P., Karapiperis, T., et al. (2016). Ebola: lessons learned and future challenges for Europe. *Lancet Infect Disiase* 16:259–263.

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., et al. (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 530:228–232.

Rausell, A., Juan, D., Pazos, F., & Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5):1995–2000. doi:10.1073/pnas.0908044107

Reid, S.P., Leung, L.W., Hartman, A.L., Martinez, O., Shaw, M.L., Carbonelle, C. et al. (2006). Ebola virus VP24 binds karyopherin alpha1 and blocks STAT1 nuclear accumulation. *Journal of virology*. 80 (11):5156–5167. doi:10.1128/JVI.02349-05.

Reid, S. P., Valmas, C., Martinez, O., Sanchez, F. M. & Basler, C. F. (2007). Ebola virus VP24 proteins inhibit the interaction of NPI-1 subfamily karyopherin alpha proteins with activated STAT1. *Journal of Virology*. 81:13469–13477

Remmert, M., & Hauser, A. (2012). HH-suite for sensitive protein sequence searching based on HMM-HMM alignment, (June):951–960.

Remmert, M., Biegert, A., Hauser, A., Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*. 9:173–175.

Report of a WHO/International Study Team. (1978). Ebola haemorrhagic fever in Sudan.

Report of a WHO/International Study Team. (1976). Bull World Health Organ. 56, 247-70.

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

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S. et al. (2013). The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research*. 41:D475–D482.

Rotkiewicz, P., Skolnick, J. (2008). Fast procedure for reconstruction of full-atom protein models

from reduced representations. *Journal of Computational Chemistry*. 29:1460–1465.

Roy, A., Kucukural, A., Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and fucntion prediction. *Nature Protocol*. 5(4):725-738. doi:10.1038/nprot.2010.5

Sachidanandam, R., Weissman, D., Schmidt, S. (2001). A map of human genome sequence variation containing 1 . 42 million single nucleotide polymorphisms. *Nature*, 409(February):928–933.

Sasidharan, N.P., Vihinen, M. (2012). VariBench: a benchmark database for variations. Human Mutation. 34(1):42-9. doi: 10.1002/humu.22204.

Sasidharan, N.P., Vihinen, M. (2013). VariBench: a benchmark database for variations. *Human Mutation*. 34:42–49.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, 33(SUPPL. 2):382–388. doi:10.1093/nar/gki387

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–11.

Shurtleff, A.C., Bavari, S. (2015). Animal models for ebolavirus countermeasures discovery: what defines a useful model? *Expert opinion on drug discovery*. 10 (7):685–702. doi:10.1517/17460441.2015.1035252.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W. et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 7:539

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*. 40:W452– W457.

Simon-Loriere, E., Faye, O., Faye, O., Koivogui, L., Magassouba, N., Keita, S., et al. (2015). Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*. 524:102–104.

Skjærven, L., Yao, X.Q., Scarabelli, G., Grant, B.J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics* **15:**399.

Sloan, C., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C. et al. (2015). ENCODE data at the ENCODE portal. *Nucleic Acids Research*, 44(November 2015), gkv1160–. doi:10.1093/nar/gkv1160

Smith, N., Witham, S., Sarkar, S., Zhang, J., Li, L., Li, C. et al. (2012). DelPhi web server v2: incorporating atomic-style geometrical figures into the computational protocol. *Bioinformatics* 28:1655–1657

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(SUPPL. 2):244–248. doi:10.1093/nar/gki408

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30**:**1312–1313

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J. et al., (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81. doi:10.1038/nature15394

Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335. doi:10.1016/j.cell.2014.01.051

Teng, Y., Wang, Y., Zhang, X., Liu, W., Fan, H., Yao, H. et al. (2015). Systematic Genome-wide Screening and Prediction of microRNAs in EBOV During the 2014 Ebolavirus Outbreak. *Scientific Reports.* 5:9912

Tennessen, J. a, Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S. et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* (New York, N.Y.), 337(6090):64–9. doi:10.1126/science.1219240

The ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project Science, 306(5696):636-40.

Thorisson, G., Smith, A. (2005). The international HapMap project web site. *Genome Research*, 15:1592–1593. doi:10.1101/gr.4413105.

Tong, Y.G., Shi, W.F., Liu, D., Qian, J., Liang, L., Bo, X.C. et al. (2015). Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature.* 524:93–96.

Tress, M.L., Ezkurdia, I., Richardson, J.S. (2009). Target domain definition and classification in CASP8". Proteins 77 (Suppl 9): 1017.

UniProt Consortium. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt) *Nucleic Acids Research.* 40:D71–D75.

UniProt Consortium. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research.* 42:D191–8

Vapnik, V.N. (1999). An overview of statistical learning theory. *IEEE Trans. Neural Netw.* 1999;10:988–999.

Varkey, J.B., Shantha, J.G., Crozier, I., Kraft, C.S., Lyon, G.M., Mehta, A.K, et al. (2015). Persistence of Ebola Virus in Ocular Fluid during Convalescence. *The New England Journal of Medicine.* 372:2423–2427.

Vogel, G. (2015). A reassuring snapshot of Ebola. *Science* 347:1407–1407

Volchkov, V. E., Chepurnov, A. A., Volchkova, V. A., Ternovoj, V. A. & Klenk, H. D. (2000). Molecular characterization of guinea pig-adapted variants of Ebola virus. *Virology* 277:147–155

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, 17(4):263–70. doi:10.1002/humu.22

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular*

Biology, 337:635-645.

Wass, M. N., Kelley, L. A., & Sternberg, M. J. E. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research*, 38(Web Server issue):W469–73. doi:10.1093/nar/gkq406

Watt, A., Moukambi, F., Banadyga, L., Groseth, A., Callison, J., Herwig, A. et al. (2014). A novel life cycle modeling system for Ebola virus shows a genome length-dependent role of VP24 in virus infectivity. *Journal of Virology*. 88 (18):10511–10524. doi:10.1128/JVI.01272-14.

Webb, B., Sali. A., (2014), Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., 5.6.1-5.6.32.

Weingartl, H.M., Nfon, C. & Kobinger, G. (2013). Review of Ebola virus infections in domestic animals. *Developments in Biologicals*. 135211–218. doi:10.1159/000178495.

Williams-Blonger, S. (2004). The Human Genome Project and Advances in Anthropological Genetics, *Human Biology*, 76 (6):801-804.

Wood, A. R., Perry, J. R. B., Tanaka, T., Hernandez, D. G., Zheng, H.-F., Melzer, D. et al., (2013). Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. *PloS One*, 8(5):e64343. doi:10.1371/journal.pone.0064343

Xu, W., Edwards, M. R., Borek, D. M., Feagins, A. R., Mittal, A., Alinger, J. B., … Amarasinghe, G. K. (2014). Ebola Virus VP24 Targets a Unique NLS Binding Site on Karyopherin Alpha 5 to Selectively Compete with Nuclear Import of Phosphorylated STAT1. *Cell Host & Microbe*, *16*(2):187–200. doi:10.1016/j.chom.2014.07.008

Yaddanapudi, K., Palacios, G., Towner, J.S., Chen, I., Sariol, C.A., Nichol, S.T. et al. (2006). Implication of a retrovirus-like glycoprotein peptide in the immunopathogenesis of Ebola and Marburg viruses. *FASEB Federation of american Societys for Experimental Biology Journal.* 20:2519–2530

Yates, C., M., Filippis, I., Kelley, L., A. and Sternberg, M., J. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology*. 426(14):2692-701. doi: 10.1016/j.jmb.2014.04.026.

Zhang, A.P., Bornholdt, Z.A., Liu, T., Abelson, D.M., Lee, D.E., Li, S. et al. (2012). The ebola virus interferon antagonist VP24 directly binds STAT1 and has a novel, pyramidal fold. *PLoS Pathog.* 8:e1002550

Zhang, A.P., Abelson, D.M., Bornholdt, Z.A., Liu, T., Woods, V.L., Saphire, E.Q. (2012). The ebolavirus VP24 interferon antagonist: know your enemy. *Virulence.* 3 (5):440–445. doi:10.4161/viru.21302.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S. et al. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1322563111

Appendix 1

VarMod: Modelling the functional effects of non-synonymous variants

Morena Pappalardo & Mark N. Wass*

Centre for Molecular Processing, School of Biosciences, University of Kent, UK.

* To whom correspondence should be addressed: m.n.wass@kent.ac.uk

Supplementary methods and Tables

The text and tables below explain the groupings used for the different amino acid properties and how they were converted to features for input into the support vector machine (SVM). Supplementary table 1 displays the full list of features input into the SVM. The weight of each of the features used in the SVM was calculated using the script provided with SVMlight, which calculates the weighted sum of the support vectors. It shows that the Jensen Shannon conservation score has the highest weighted followed by the binding site and interface site features and solvent accessibility features. Conservation (Jensen Shannon divergence) has been used previously by other methods including SIFT and PolyPhen and it is not surprising that it is weighted highly. The weighting of the interface and binding site features demonstrates that they used by VarMod to make predictions and are more informative than other features such as those relating to secondary structure.

Feature	Value range	SVM weight
js convergence score (conservation)	0-1	1.83
Amino acid properties		
amino acid charge change	see supplementary table 2	0.08
amino acid mass change	See supplementary table 3	0.08
amino acid functional group change	1 where functional change, 0 otherwise (see Supplementary table 4)	0.08
3DLigandSite features		
distance to binding site	0-1 (actual distance divided by 25, values greater than 1 are rounded down to 1)	1.31
3DLigandSite average distance to ligands	0-1 (value/ 2)	1.51
3DLigandSite number of ligands that bind to this residue	num/50	0.80
Interface site features		
distance to interface site	0-1 (distance/25, values greater than 1 round down to 1)	1.23
Secondary Structure features		
DSSP -secondary structure- B	0/1 (1 if ss is B, 0 otherwise)	0.47
DSSP -secondary structure- G	0/1 (1 if ss is G, 0 otherwise)	0.09
DSSP -secondary structure- I	0/1 (1 if ss is I, 0 otherwise)	0.26
DSSP -secondary structure- T	0/1 (1 if ss is T, 0 otherwise)	0.13
DSSP -secondary structure- S	0/1 (1 if ss is S, 0 otherwise)	0.11
DSSP -secondary structure- BL	0/1 (1 if ss is BL, 0 otherwise)	0.20
DSSP -secondary structure- H	0/1 (1 if ss is H, 0 otherwise)	0.13
DSSP -secondary structure-E	0/1 (1 if ss is E, 0 otherwise)	0.02
		0.48
DSSP -secondary structure Type - Heilx	0/1 (1 if ss type is is H, 0 otherwise)	0.49
DSSP -secondary structure Type – Strand	0/1 (1 if ss is B, 0 otherwise)	0.45
DSSP -secondary structure Type - Coil	0/1 (1 if ss is B, 0 otherwise)	0.08
distance from end of secondary structure	0 - 0.5 (0.5 in the middle, 0 at end of secondary structure element)	0.26
DSSP - solvent accessibility	0-1 (solvent accessibility / 300)	-1.05

Supplementary Table 1. The SVM features used in VarMod are listed with the value ranged used for each feature and the weighting of the features in the SVM.

Supplementary Tables 2-4 relate to the change in amino acid properties of the variants. Supplementary Table 2 shows the amino acid charge groups and Supplementary table 3 shows the value for the amino acid charge feature for changes between these groups. Supplementary table 4 shows the groups of amino acids based on functional groups present in the side chain. The feature associated with functional groups is either 0 (no change in functional group), 1 (change in functional group).

Charge group	**Amino acids**
Positive charge	R, H, K
Negative charge	D, E
Negative polar	N, Q
Positive polar	S, T
Hydrophobic	G, A, V, I, L, M, F, Y, W, C, P

Supplementary Table 2. Amino acid charge groups.

	Positive charge	Negative charge	Negative polar	Positive polar	Hydrophobic
Positive charge	0				
Negative charge	1	0			
Negative polar	0.5	0.25	0		
Positive polar	0.25	0.5	0.75	0	
Hydrophobic	1	1	0.75	0.75	0

Supplementary Table 3. SVM feature value for change in amino acid charge.

Functional group	**Amino acids**
Positive	R, H, K
Carboxylate	D, E
Phenyl	F, Y, W
hydroxyl	S, T, Y
Amido	N, Q
Other/none	G, A, V, I, L, M, C, P

Supplementary Table 4. Amino acid functional groups used as defined in Innis et al., (28).

Appendix 2

Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses

Authors: Morena Pappalardo†, Miguel Juliá†, Mark J. Howard, Jeremy S. Rossman*, Martin Michaelis*, Mark N. Wass*

Affiliation:

Centre for Molecular Processing and School of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK.

*Correspondence to: M.N.Wass@kent.ac.uk (Mark N. Wass), M.Michaelis@kent.ac.uk (Martin Michaelis), J.S.Rossman@kent.ac.uk (Jeremy S. Rossman)

† equal contribution.

Supplementary Material

Supplementary Methods - Subsampling of sequence data

The sensitivity of the SDP analysis to the number of sequences available was considered by subsampling the sequences. Sampling was performed for; only the human pathogenic group; only the Reston group; and for both groups simultaneously. Subsampling was performed using between 10%-90% of sequences in the group, increasing in 10% increments. For each percentage setting the group was sampled 50 times. Where both groups were sampled simultaneously they were done so with the same percentage of sequences i.e. at 20% sampling the SDPs were predicted each time using 20% of the human pathogenic sequences in one group and 20% of the Reston sequences in the other. For

each sample s3det was run to predict SDPs using the same settings as for the full dataset. Completely conserved SDPs are also compared to those that are not completely conserved. the The total number of SDPs predicted when sampled is shown in supplementary Figure 6. When the sequences of human pathogenic Ebolaviruses were sampled, while the number of Reston sequences remained constant, we observed that the number of SDPs predicted decreased as the proportion of sequences sampled increased. We further observed that even when a very high proportion of sequences was sampled (70%-90%), that there was still some variation in the number of SDPs, indicating that there was still further information present in the excluded sequences. When the Reston virus sequences were sampled, the pattern observed varied between the proteins (Supplementary Figure 6B). For GP, L and VP30, sampling resulted in more SDPs being predicted than in the full dataset, with the number reducing as the proportion of sequences sampled increased. For NP, sampling the Reston sequences generated some samples where fewer SDPs than the total present in the full dataset were predicted and other samples where a larger number of SDPs were predicted. This is possible for SDPs that are not completely conserved in the two groups, as sampling may generate some sets of sequences where these positions appear variable and others where they are conserved. For VP35, sampling led to fewer SDPs being predicted until 90% of sequences were used. The number of SDPs in VP24 and VP40 was invariant across all samples. When sampling both groups (Supplementary Figure 6C) we found that the number of SDPs predicted very quickly converged to the number of SDPs present in the full dataset.

We then considered the number of SDPs predicted that are present in the full dataset and those that are present only in sampling (Supplementary Figure 7). When the human pathogenic sequences were sampled (Supplementary Figure 7A), we found that the vast majority of SDPs in the full data set were predicted at all sampling levels. We also found that when a small proportion of sequences were sampled, that many new SDPs were predicted, which for some proteins (e.g. GP, NP and VP40) may be greater than the total number of SDPs present in the full dataset. This may not be too surprising given that positions that are variable in the full dataset may appear to be conserved when a small sample of sequences was taken. As the proportion of sequences sampled increased, very few new SDPs were predicted. Sampling the Reston sequences (Supplementary Figure 7B) we again found that the vast majority of SDPs present in the full dataset was present in all samples. The number of new SDPs present in samples was much smaller than for sampling of the human pathogenic sequences, which is likely to be due to the smaller number of Reston sequences, resulting in fewer samples where positions are conserved that are not conserved in the full data set. When both groups were sampled, results were very similar to that observed when the human pathogenic group was sampled (Supplementary Figure 7C).

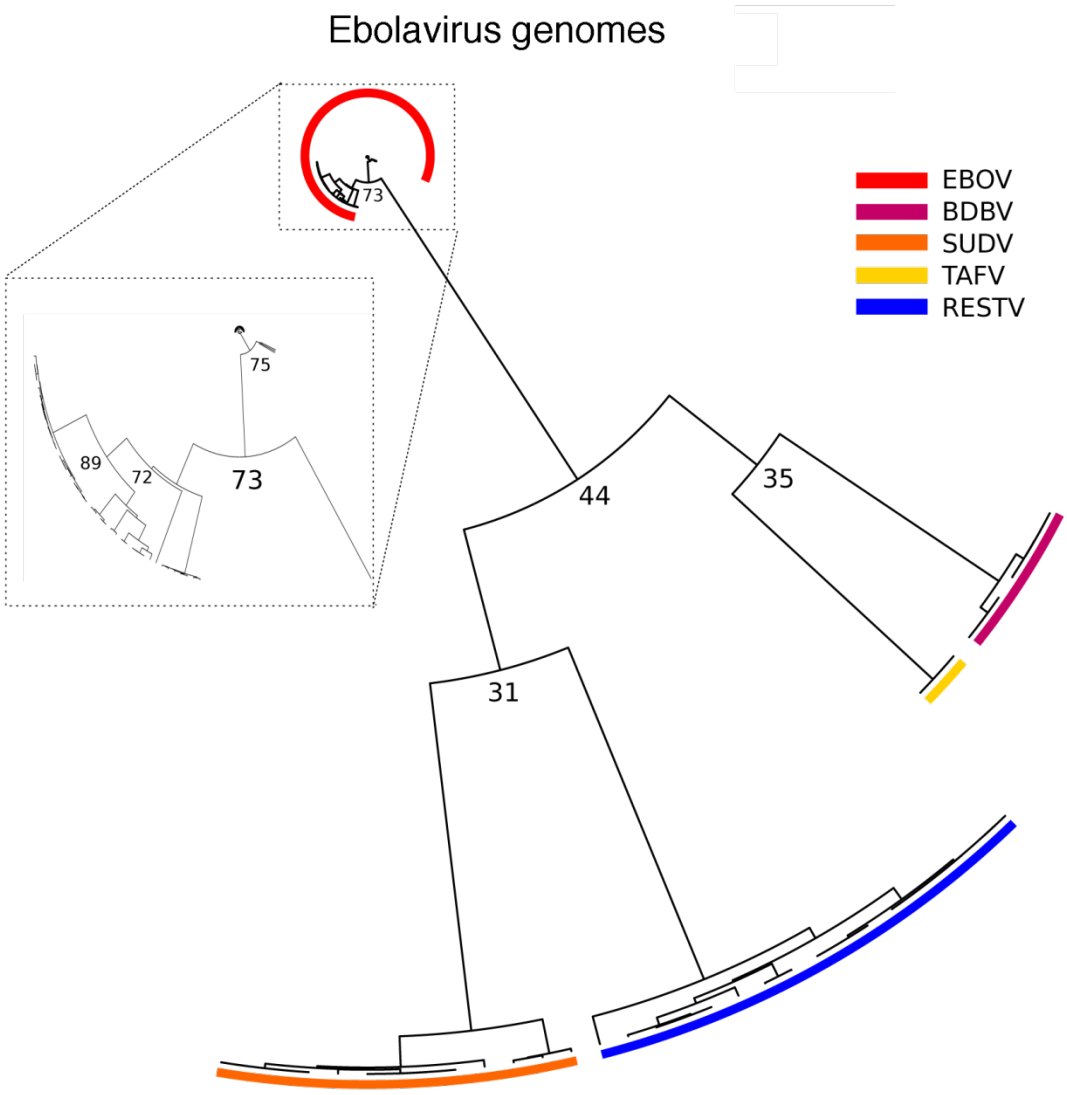
Finally, we considered the number of SDPs in the sampling sets that are completely conserved and those that are not (Supplementary Figure 8). In conjunction with the data from Supplementary Figure

7, this shows that sampling generates new SDPs that are completely conserved (i.e. only one amino acid in each group) and also some where there is variation within one or both groups. As the proportion of sequences sampled increased these numbers quickly converged to the numbers observed in the full dataset. Some of these included SDPs which in some samples were completely conserved but as further sequences were added, variation was introduced and they were no longer completely conserved. In such cases there was a change ranking for the SDP, as when completely conserved it was ranked 1, and this ranking was reduced once the position was not completely conserved.

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. All trees use Ebola virus as root (EBOV, Ebola virus; BDBV, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus; RESTV, Reston virus).

Fig S1A


Ebolavirus genomes
EBOV
BDBV
SUDV
TAFV
RESTV
73
75
89
72
73
44
35
31

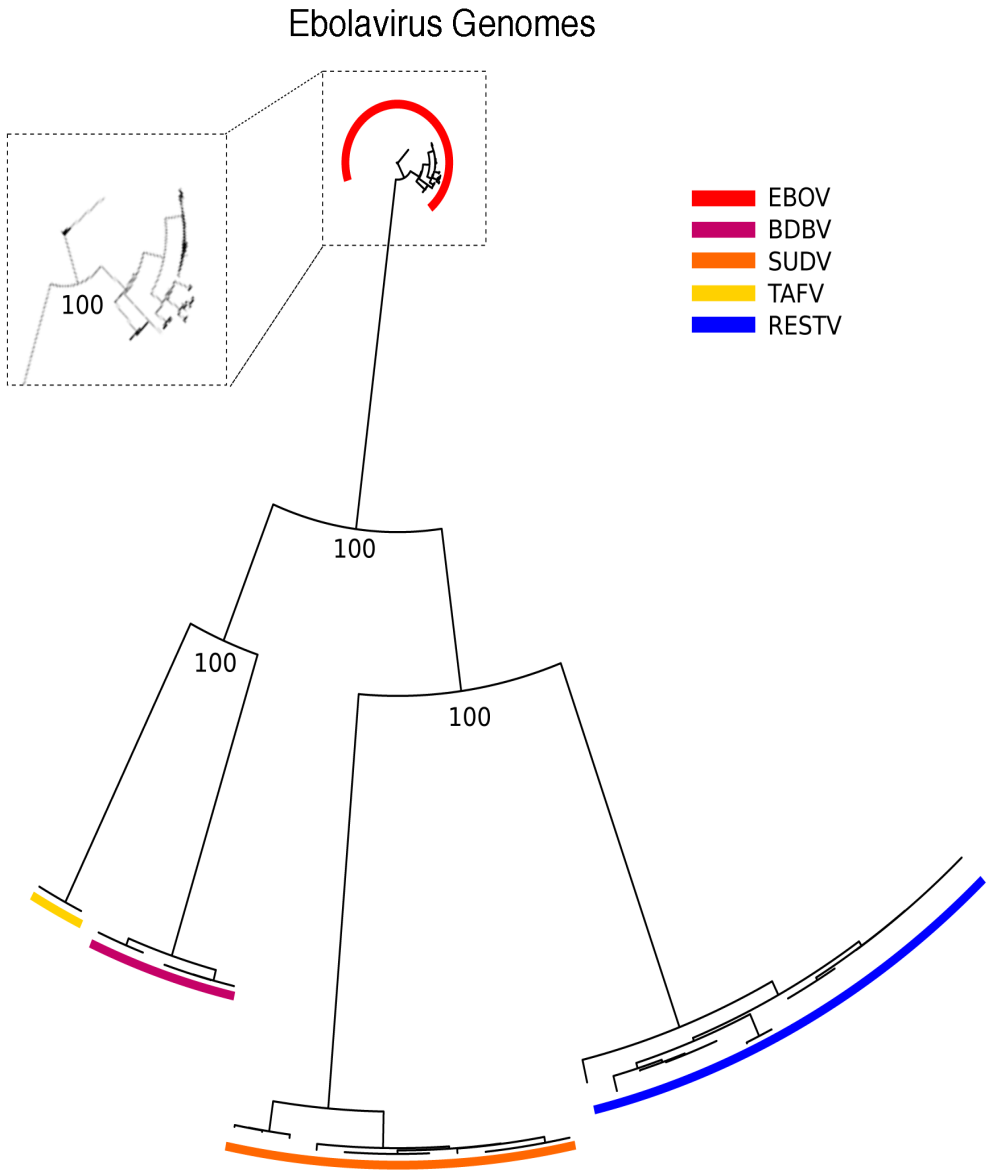

Ebolavirus Genomes
EBOV
BDBV
SUDV
TAFV
RESTV
100
100
100
100

Fig S1B.

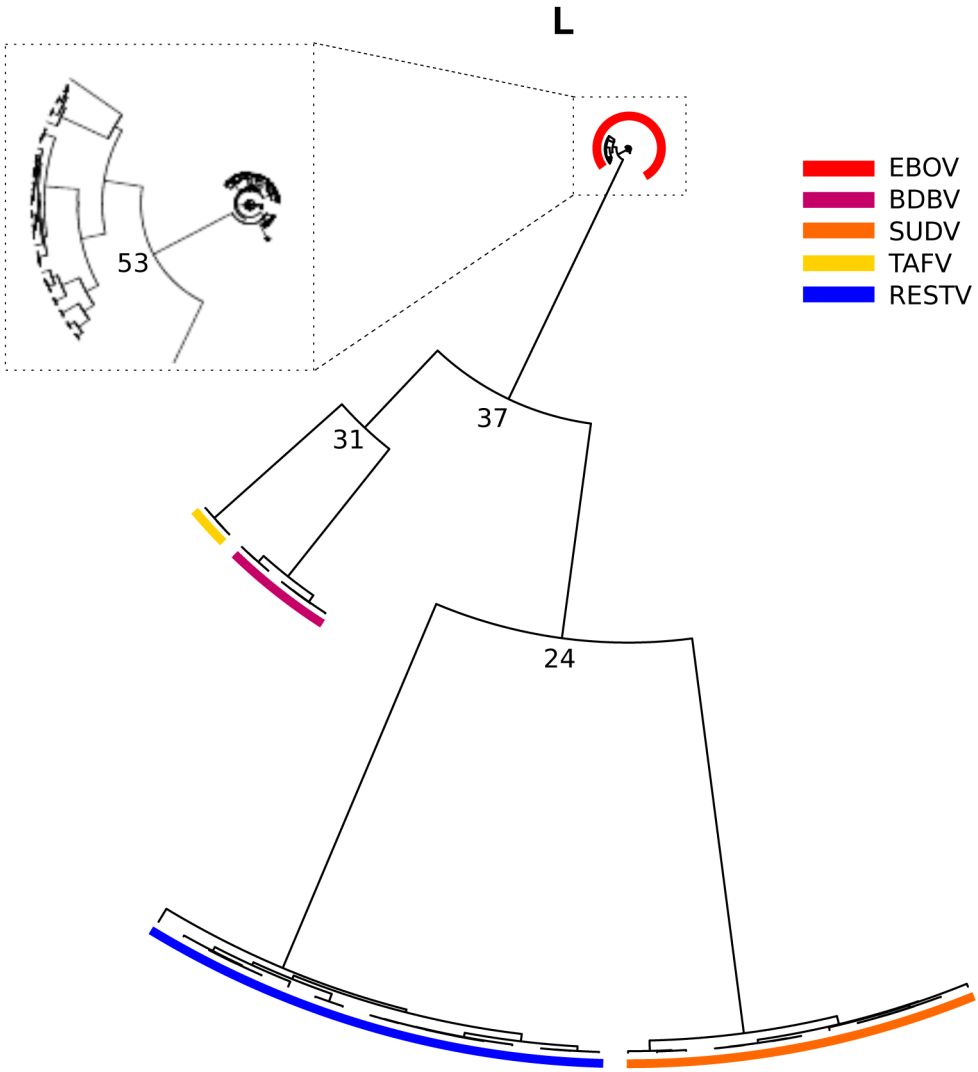


Fig S1C.

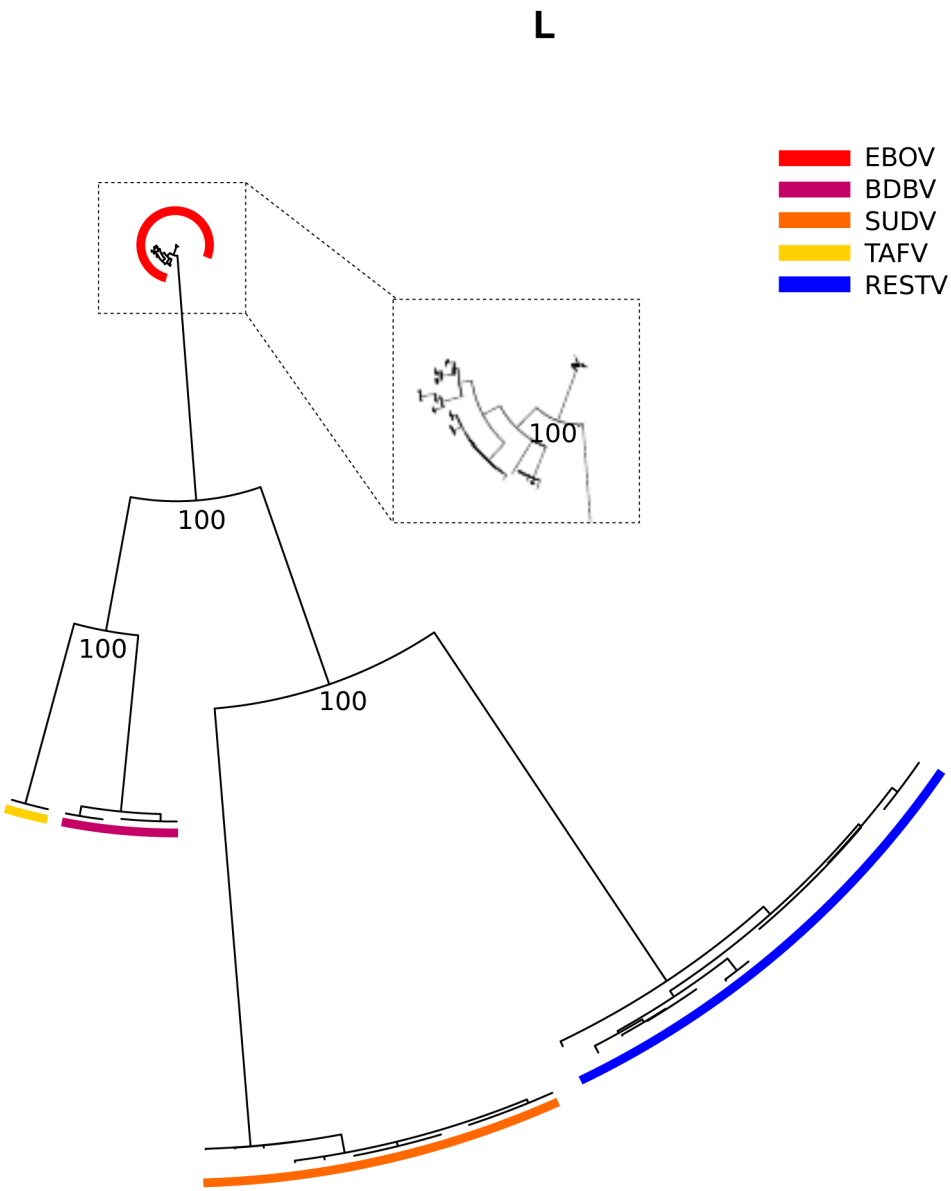


Fig S1D.

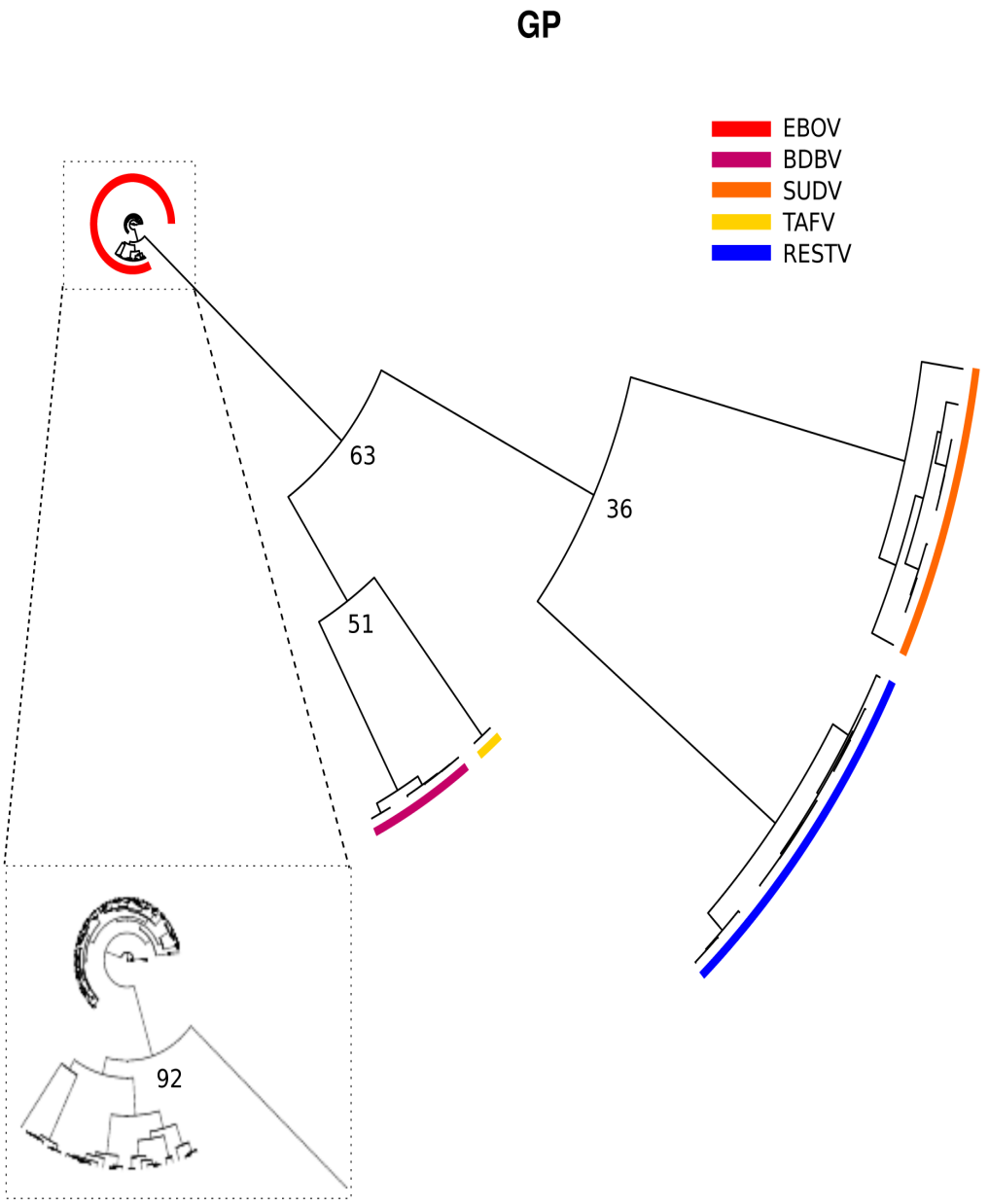


Fig S1E .

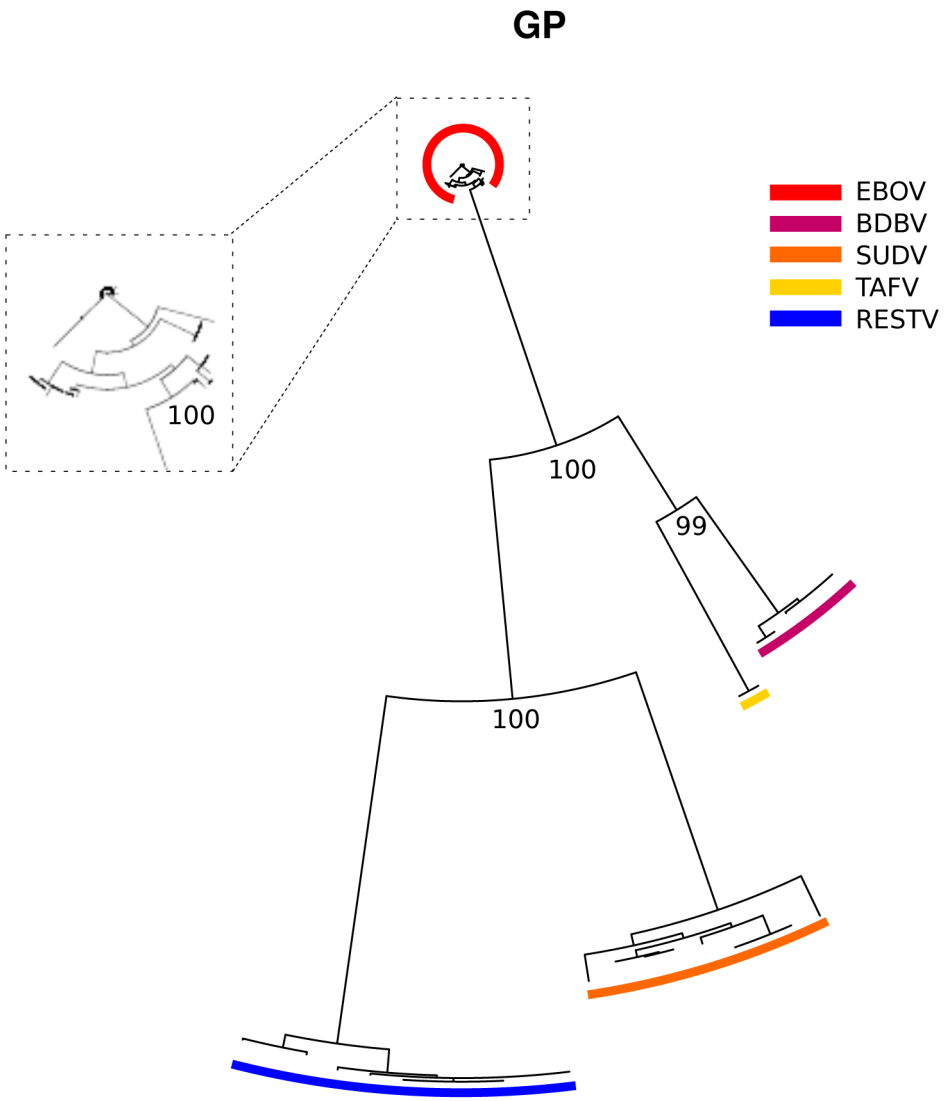


Fig S1F.

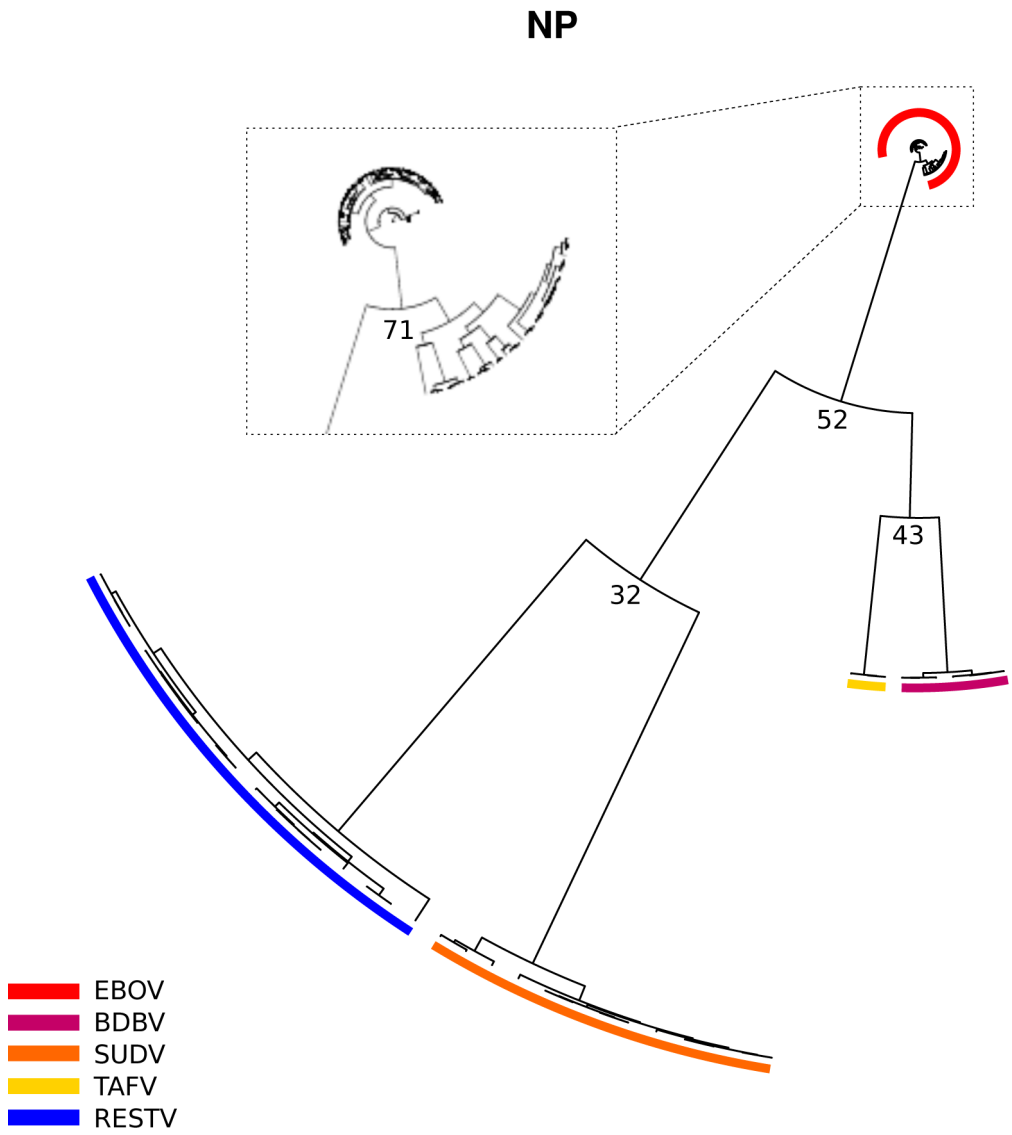


Fig S1G.

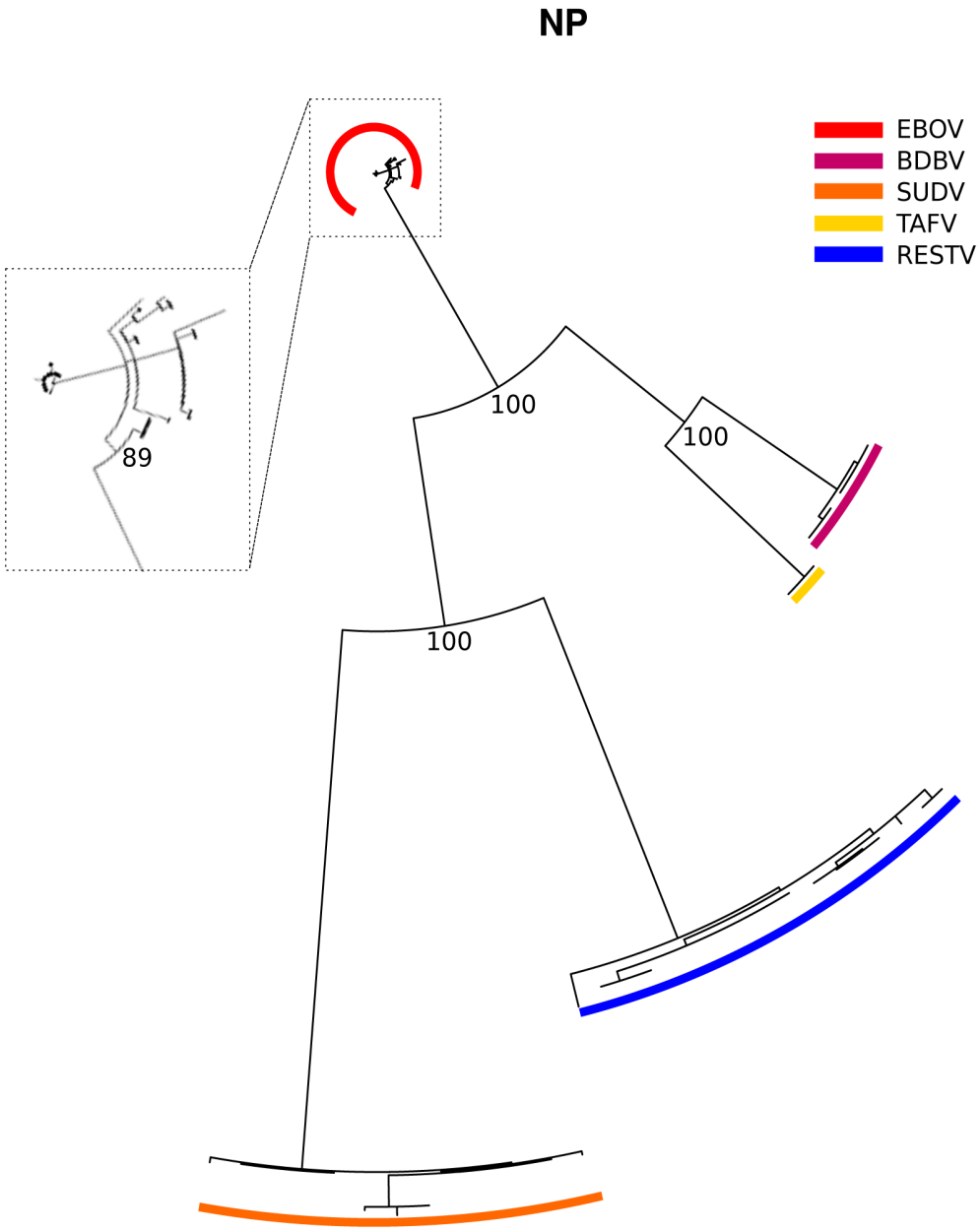


Fig S1H

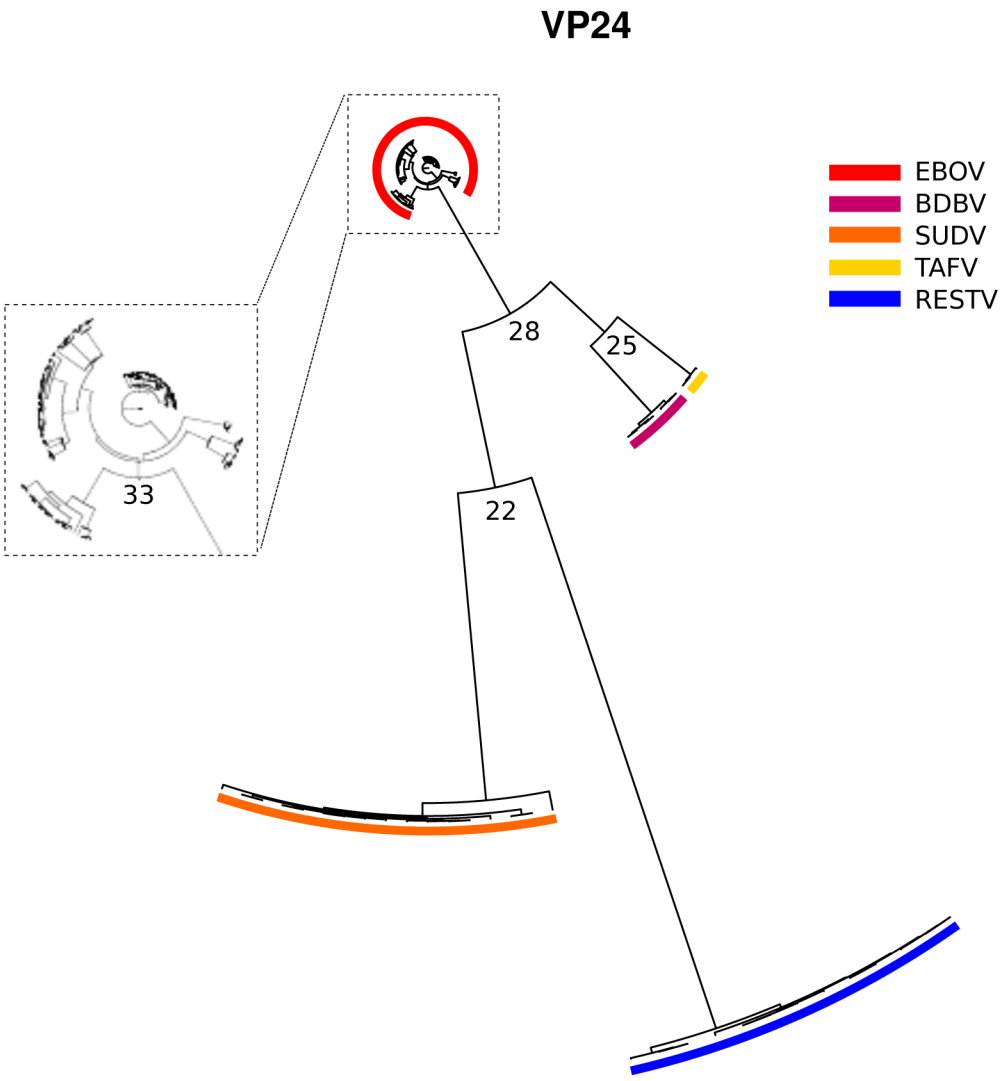


Fig S1I.

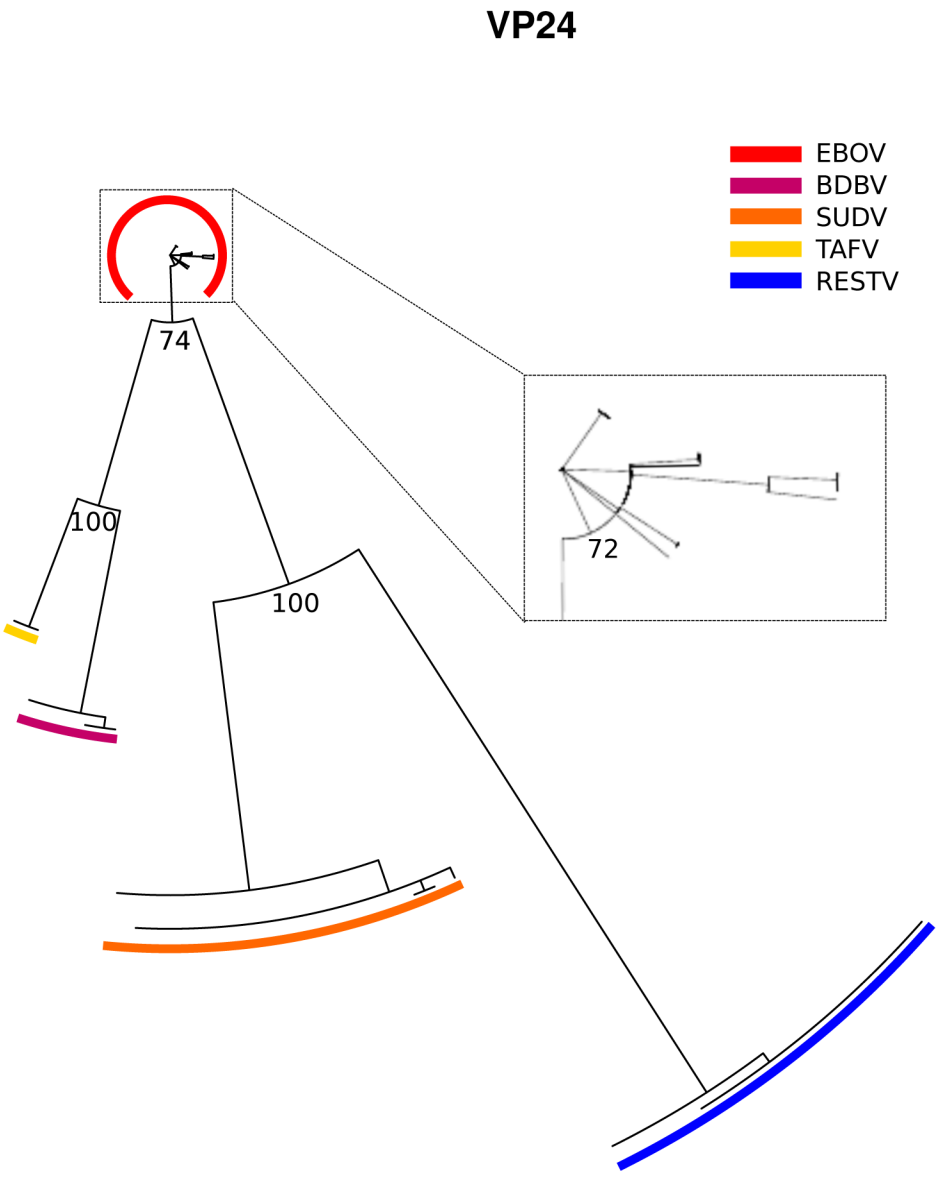


Fig S1J.

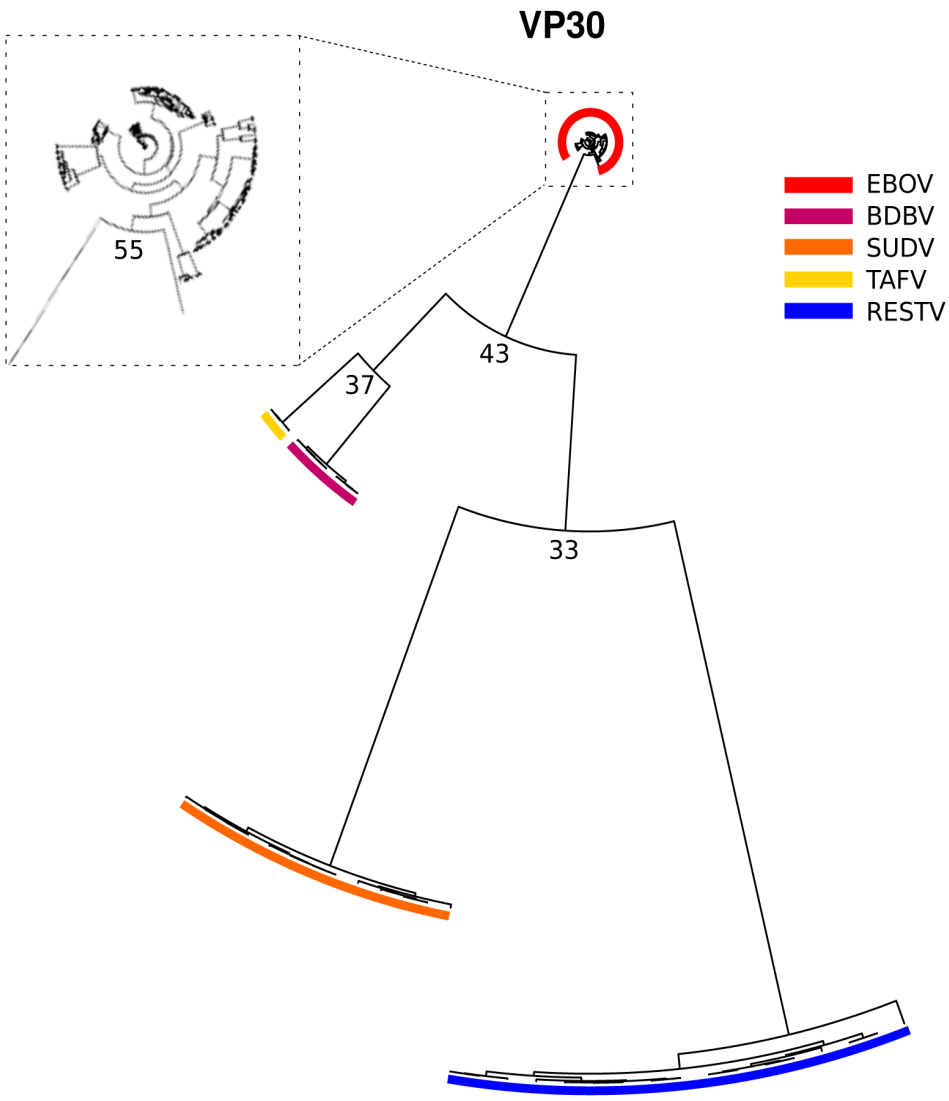


Fig S1K.

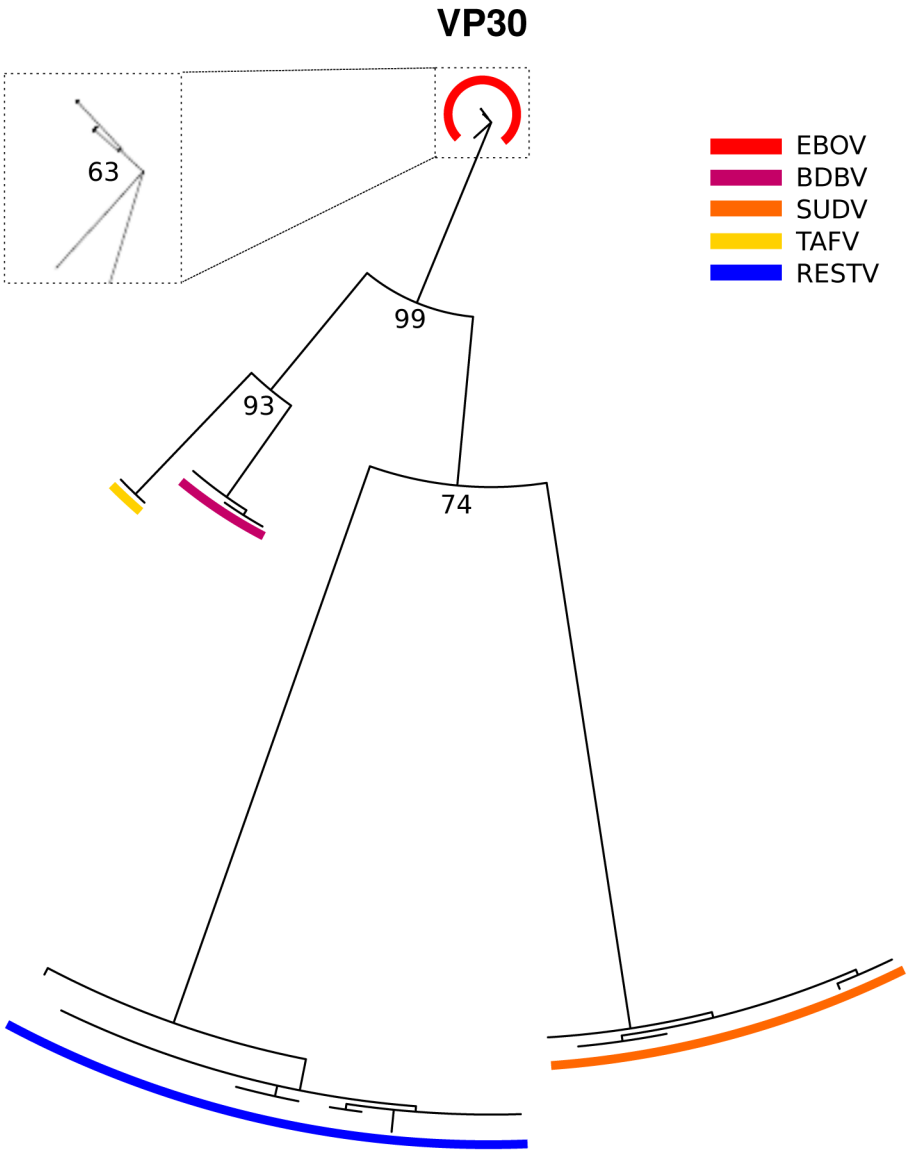


Fig S1L

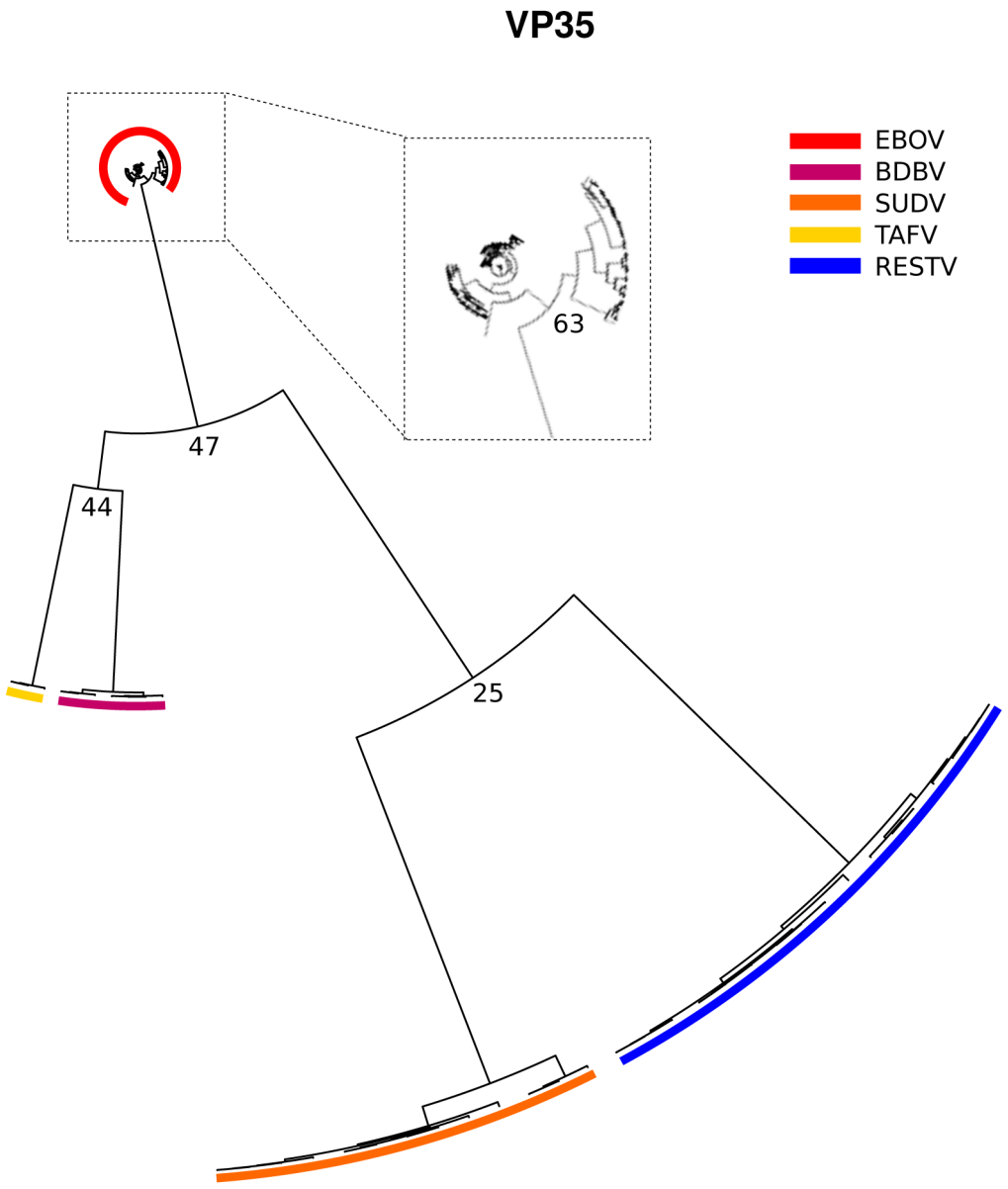


Fig S1M.

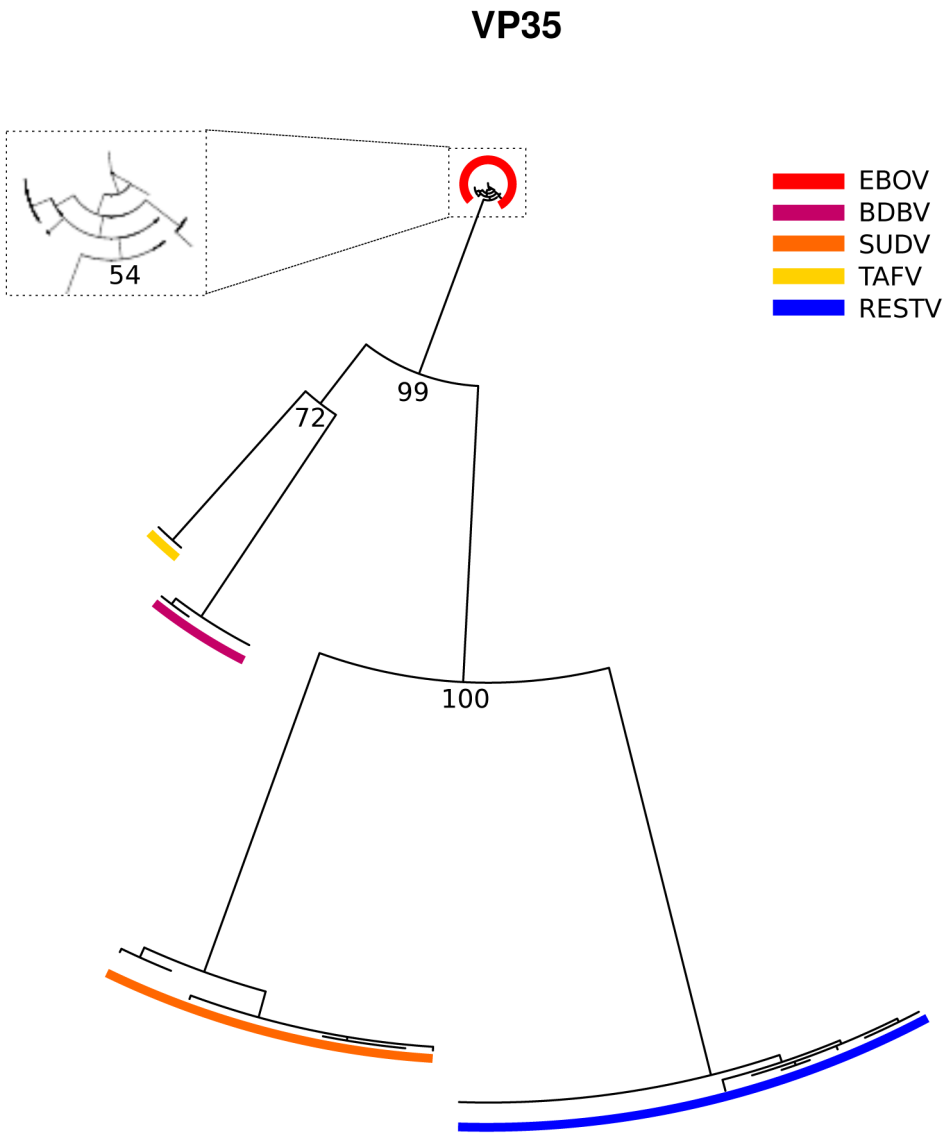


Fig S1N.

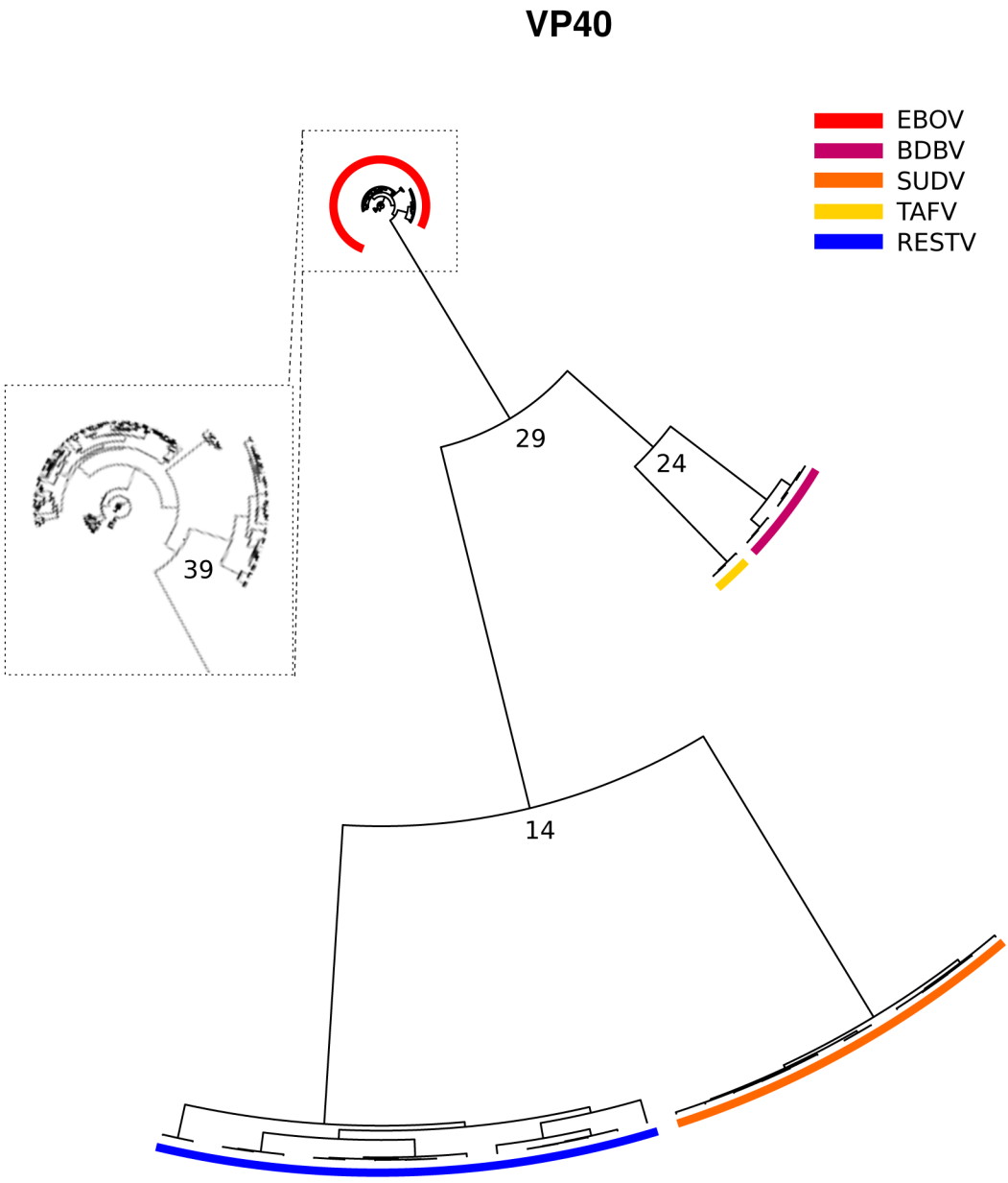


Fig S1O.

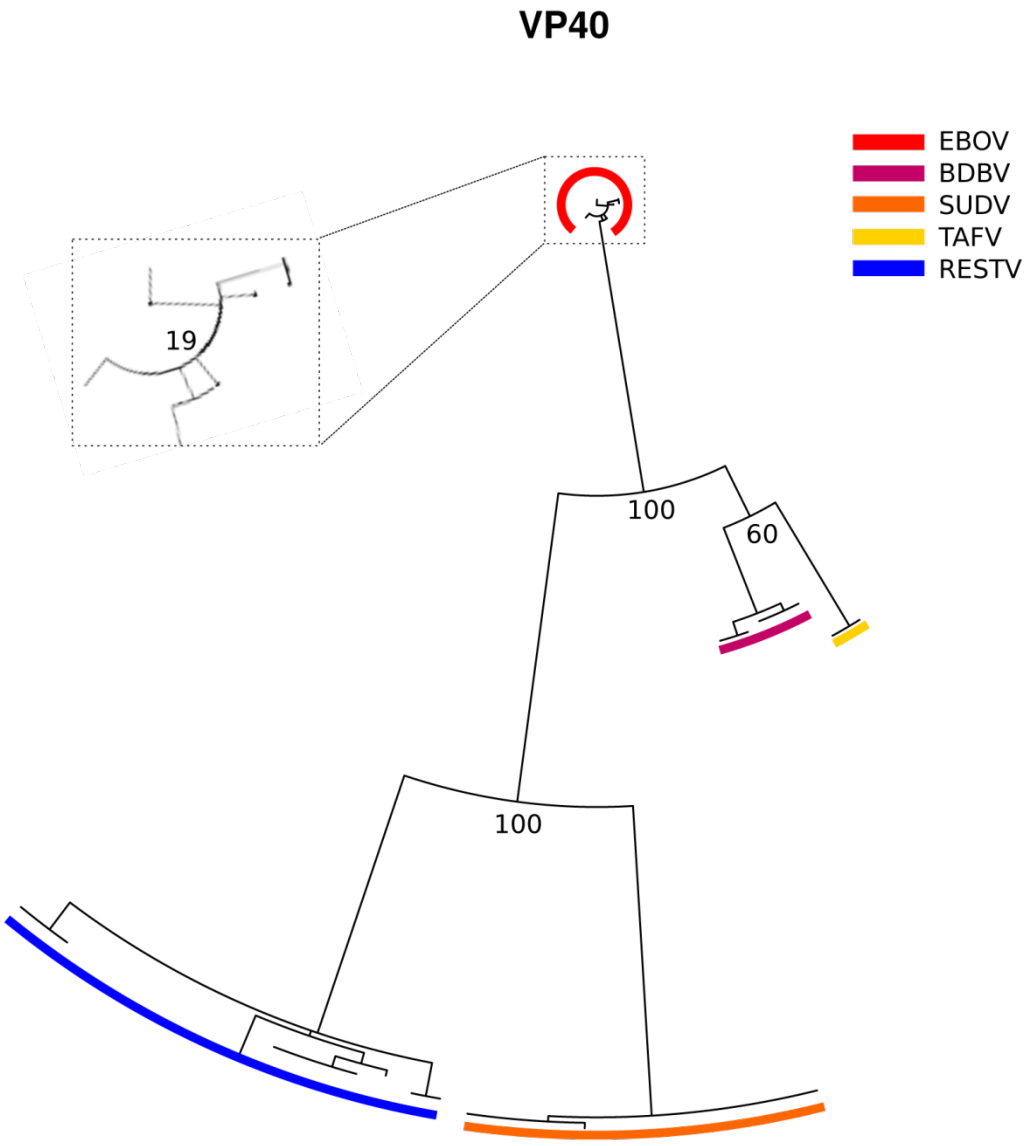
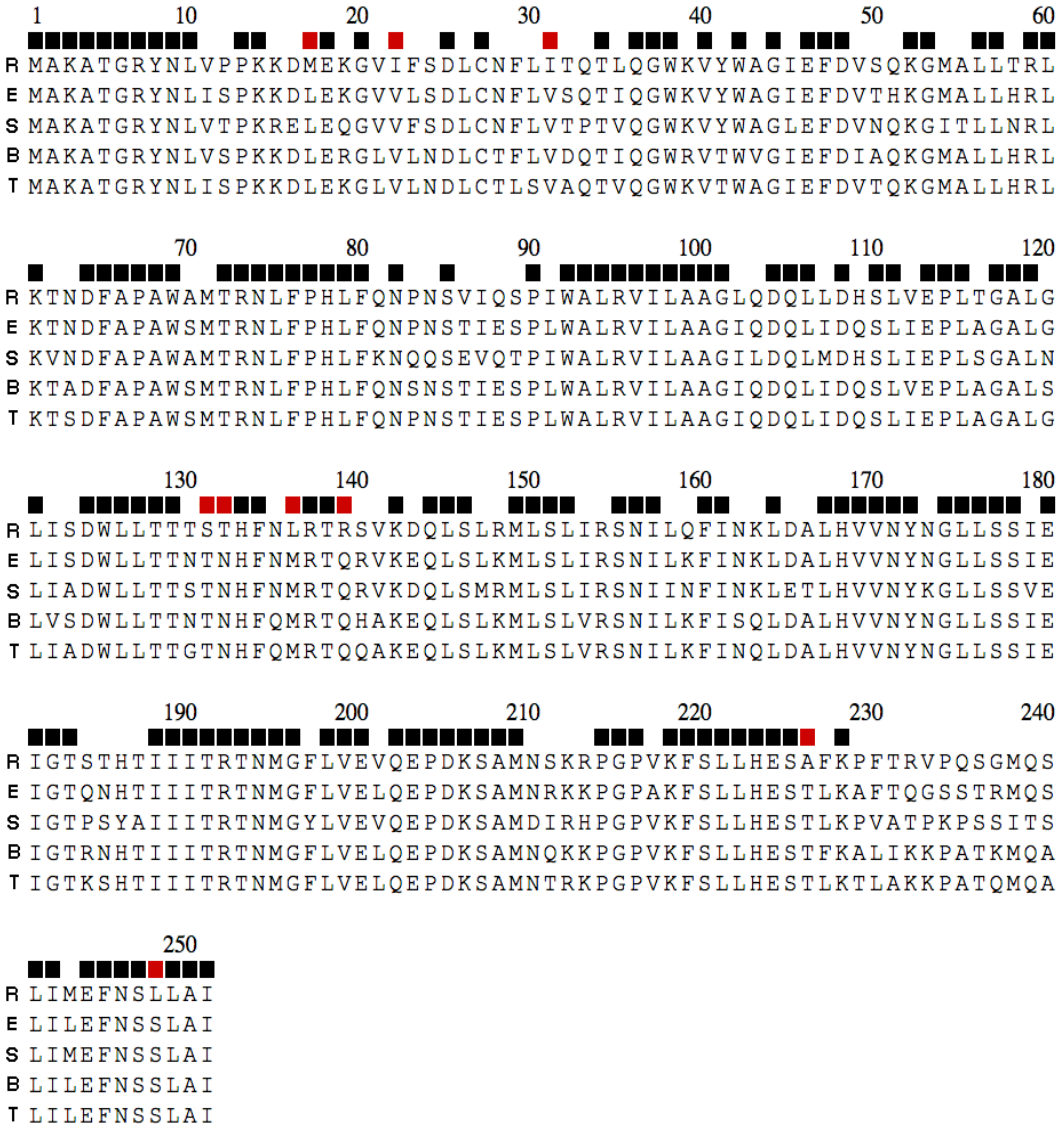


Fig S1P.

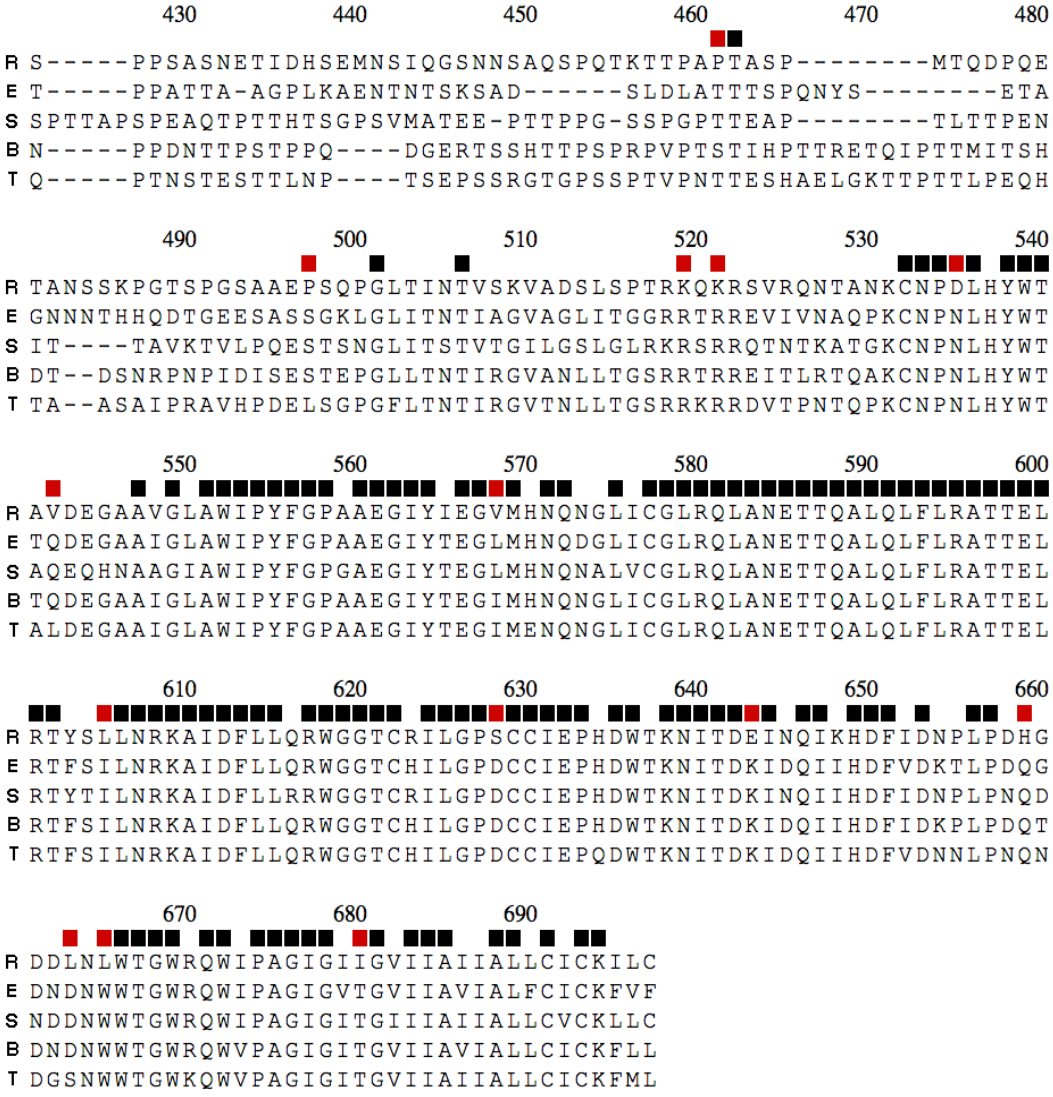
Supplementary Figure 2. Ebolavirus protein consensus sequences and SDPs. The consensus sequence for each *Ebolavirus* species is shown for each Ebolavirus protein. The row above the alignment indicates positions that are 100% conserved across all Ebolavirus sequences (black) or specificity determining positions (SDPs) that discriminate Reston viruses from the four human pathogenic *Ebolavirus* species (red); R, Reston virus; E, Ebola virus; S, Sudan virus; B, Bundibugyo virus; T, Taï Forest virus. A) for VP24, B) for GP, C) for VP40, D) VP35, E)VP30, F) sGP, G) NP, H)L.

A – VP24

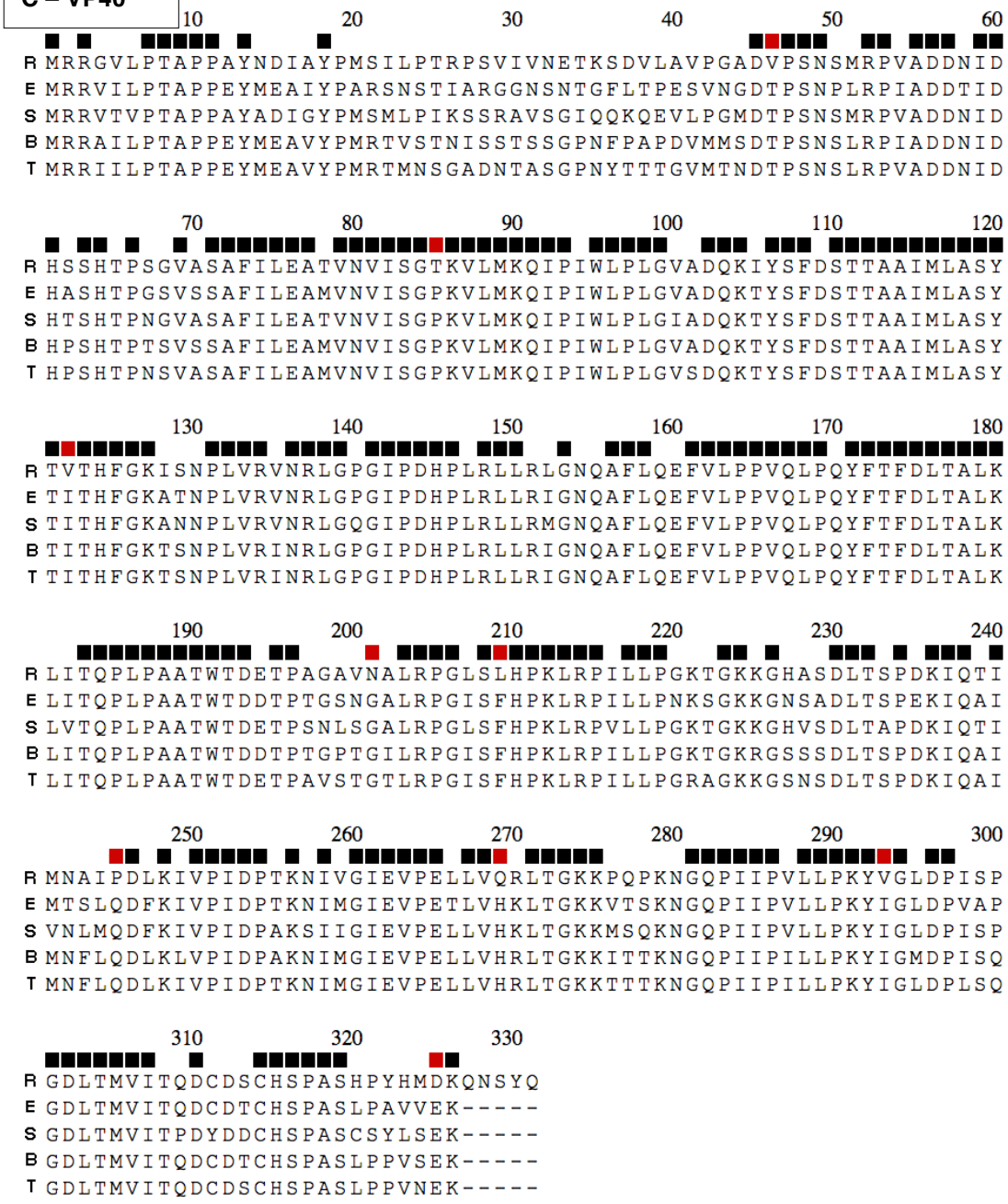


B - GP

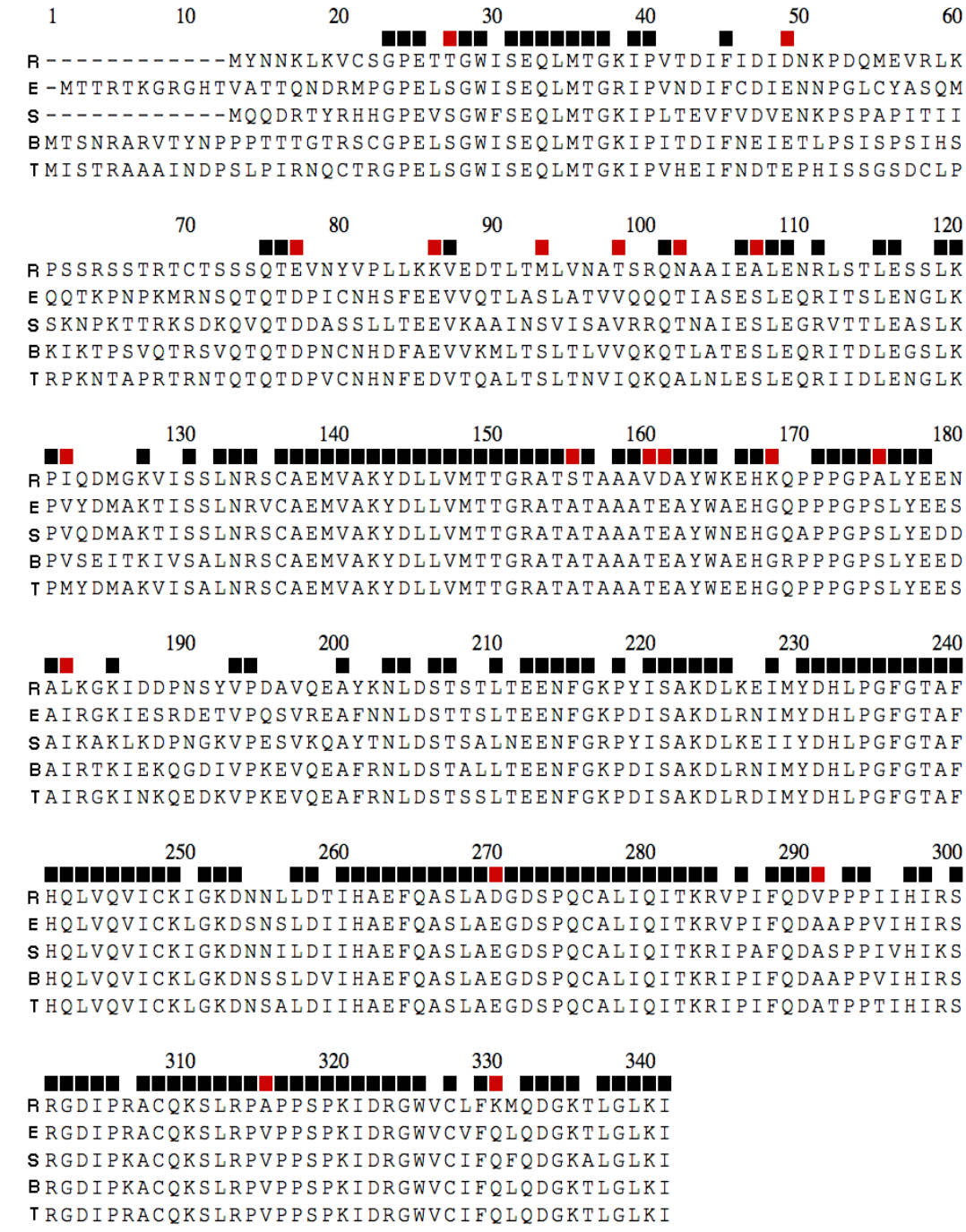
```
                10        20        30        40        50        60
R MGSGYQLLQLPRERFRKTSFLVWVIILFQRAISMPLGIVTNSTLKATEIDQLVCRDKLSS
E -MGVTGILQLPRDRFKRTSFFLWVIILFQRTFSIPLGVIHNSTLQVSDVDKLVCRDKLSS
S -MGGLSLLQLPRDKFRKSSFFVWVIILFQKAFSMPLGVVTNSTLEVTEIDQLVCKDHLAS
B -MVTSGILQLPRERFRKTSFFVWVIILFHKVFPIPLGVVHNNTLQVSDIDKLVCRDKLSS
T -MGASGILQLPRERFRKTSFFVWVIILFHKVFSIPLGVVHNNTLQVSDIDKFVCRDKLSS

        70        80        90        100       110       120
R TSQLKSVGLNLEGNGIATDVPSATKRWGFRSGVPPKVVSYEAGEWAENCYNLEIKKSDGS
E TNQLRSVGLNLEGNGVATDVPSVTKRWGFRSGVPPKVVNYEAGEWAENCYNLEIKKPDGS
S TDQLKSVGLNLEGSGVSTDIPSATKRWGFRSGVPPKVVSYEAGEWAENCYNLEIKKPDGS
B TSQLKSVGLNLEGNGVATDVPTATKRWGFRAGVPPKVVNYEAGEWAENCYNLDIKKADGS
T TSQLKSVGLNLEGNGVATDVPTATKRWGFRAGVPPKVVNCEAGEWAENCYNLAIKKVDGS

        130       140       150       160       170       180
R ECLPLPPDGVRGFPRCRYVHKVQGTGPCPGDLAFHKNGAFFLYDRLASTVIYRGTTFAEG
E ECLPAAPDGIRGFPRCRYVHKVSGTGPCAGDFAFHKEGAFFLYDRLASTVIYRGTTFAEG
S ECLPPPPDGVRGFPRCRYVHKAQGTGPCPGDYAFHKDGAFFLYDRLASTVIYRGVNFAEG
B ECLPEAPEGVRGFPRCRYVHKVSGTGPCPEGFAFHKEGAFFLYDRLASTIIYRSTTFSEG
T ECLPEAPEGVRDFPRCRYVHKVSGTGPCPGGLAFHKEGAFFLYDRLASTIIYRGTTFAEG

        190       200       210       220       230       240
R VVAFLILSEPKKHFWKATPAHEPVNTTDDSTSYYMTLTLSYEMSNFGGEESNTLFKVDNH
E VVAFLILPQAKKDFFSSHPLREPVNATEDPSSGYYSTTIRYQATGFGTNETEYLFEVDNL
S VIAFLILAKPKETFLQSPPIREAVNYTENTSSYYATSYLEYEIENFGAQHSTTLFKIDNN
B VVAFLILPKTKKDFFQSPPLHEPANMTTDPSSYYHTVTLNYVADNFGTNMTNFLFQVDHL
T VIAFLILPKARKDFFQSPPLHEPANMTTDPSSYYHTTTINYVVDNFGTNTTEFLFQVDHL

        250       260       270       280       290       300
R TYVQLDRPHTPQFLVQLNETLRRNNRLSNSTGRLTWTLDPKIEPDVGEWAFWETKKNFSQ
E TYVQLESRFTPQFLLQLNETIYASGKRSNTTGKLIWKVNPEIDTTIGEWAFWETKKNLTR
S TFVRLDRPHTPQFLFQLNDTIHLHQQLSNTTGRLIWTLDANINADIGEWAFWENKKNLSE
B TYVQLEPRFTPQFLVQLNETIYTNGRRSNTTGTLIWKVNPTVDTGVGEWAFWENKKNFTK
T TYVQLEARFTPQFLVLLNETIYSDNRRSNTTGKLIWKINPTVDTSMGEWAFWENKKNFTK

        310       320       330       340       350       360
R QLHGENLHFQILSTHTNNSSDQSPAGTVQGKISYHPPTNNSELVPTDSPPVVSVLTAGRT
E KIRSEELSFTAVSNGPKNISGQSPARTSSDPETNTTNEDHKIMASENSSAMVQVHSQGRK
S QLRGEELSFEALSLNETEDDDAASSRITKGRISDRATRKYSDLVPKNSPGMVPLHIPEGE
B TLSSEELSVILVPRAQDPGSNQKTKVTPTSFANNQTSKNHEDLVPKDPASVVQVRDLQRE
T TLSSEELSFVPVPETQNQVLDTTATVSPPISAHNHAAEDHKELVSEDSTPVVQMQNIKGK

        370       380       390       400       410       420
R EEMSTQGLTNGETI---TGFTANPMTTTIAPSPTMTSEVDNNVPSEQP----NNTASIED
E AAVSHLTTLATISTSPQPPTTKTGPDNSTHNTPVYKLDISEATQVGQHHRRADNDSTASD
S TTLPSQNSTEGRRV---SVNTQETITETAA---TIIGTNGNHMQISTIGIRPSSSQIPSS
B NTVPTSPLNTVPTT-L-IPDTMEEQTTSHYELPNISGNHQERNNTAHPET-------LAN
T DTMPTTVTGVPTTT-P-SPFPINARNTDHTKSFIGLEGPQEDHSTTQPAK-------TTS
```


430 440 450 460 470 480
R S-----PPSASNETIDHSEMNSIQGSNNSAQSPQTKTTPAPTASP--------MTQDPQE
E T-----PPATTA-AGPLKAENTNTSKSAD------SLDLATTTSPQNYS--------ETA
S SPTTAPSPEAQTPTTHTSGPSVMATEE-PTTPPG-SSPGPTTEAP--------TLTTPEN
B N-----PPDNTTPSTPPQ----DGERTSSHTTPSPRPVPTSTIHPTTRETQIPTTMITSH
T Q-----PTNSTESTTLNP----TSEPSSRGTGPSSPTVPNTTESHAELGKTTPTTLPEQH
490 500 510 520 530 540
R TANSSKPGTSPGSAAEPSQPGLTINTVSKVADSLSPTRKQKRSVRQNTANKCNPDLHYWT
E GNNNTHHQDTGEESASSGKLGLITNTIAGVAGLITGGRRTRREVIVNAQPKCNPNLHYWT
S IT----TAVKTVLPQESTSNGLITSTVTGILGSLGLRKRSRRQTNTKATGKCNPNLHYWT
B DT--DSNRPNPIDISESTEPGLLTNTIRGVANLLTGSRRTRREITLRTQAKCNPNLHYWT
T TA--ASAIPRAVHPDELSGPGFLTNTIRGVTNLLTGSRRKRRDVTPNTQPKCNPNLHYWT
550 560 570 580 590 600
R AVDEGAAVGLAWIPYFGPAAEGIYIEGVMHNQNGLICGLRQLANETTQALQLFLRATTEL
E TQDEGAAIGLAWIPYFGPAAEGIYTEGLMHNQDGLICGLRQLANETTQALQLFLRATTEL
S AQEQHNAAGIAWIPYFGPGAEGIYTEGLMHNQNALVCGLRQLANETTQALQLFLRATTEL
B TQDEGAAIGLAWIPYFGPAAEGIYTEGIMHNQNGLICGLRQLANETTQALQLFLRATTEL
T ALDEGAAIGLAWIPYFGPAAEGIYTEGIMENQNGLICGLRQLANETTQALQLFLRATTEL
610 620 630 640 650 660
R RTYSLLNRKAIDFLLQRWGGTCRILGPSCCIEPHDWTKNITDEINQIKHDFIDNPLPDHG
E RTFSILNRKAIDFLLQRWGGTCHILGPDCCIEPHDWTKNITDKIDQIIHDFVDKTLPDQG
S RTYTILNRKAIDFLLRRWGGTCRILGPDCCIEPHDWTKNITDKINQIIHDFIDNPLPNQD
B RTFSILNRKAIDFLLQRWGGTCHILGPDCCIEPHDWTKNITDKIDQIIHDFIDKPLPDQT
T RTFSILNRKAIDFLLQRWGGTCHILGPDCCIEPQDWTKNITDKIDQIIHDFVDNNLPNQN
670 680 690
R DDLNLWTGWRQWIPAGIGIIGVIIAIIALLCICKILC
E DNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
S NDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
B DNDNWWTGWRQWVPAGIGITGVIIAVIALLCICKFLL
T DGSNWWTGWKQWVPAGIGITGVIIAIIALLCICKFML

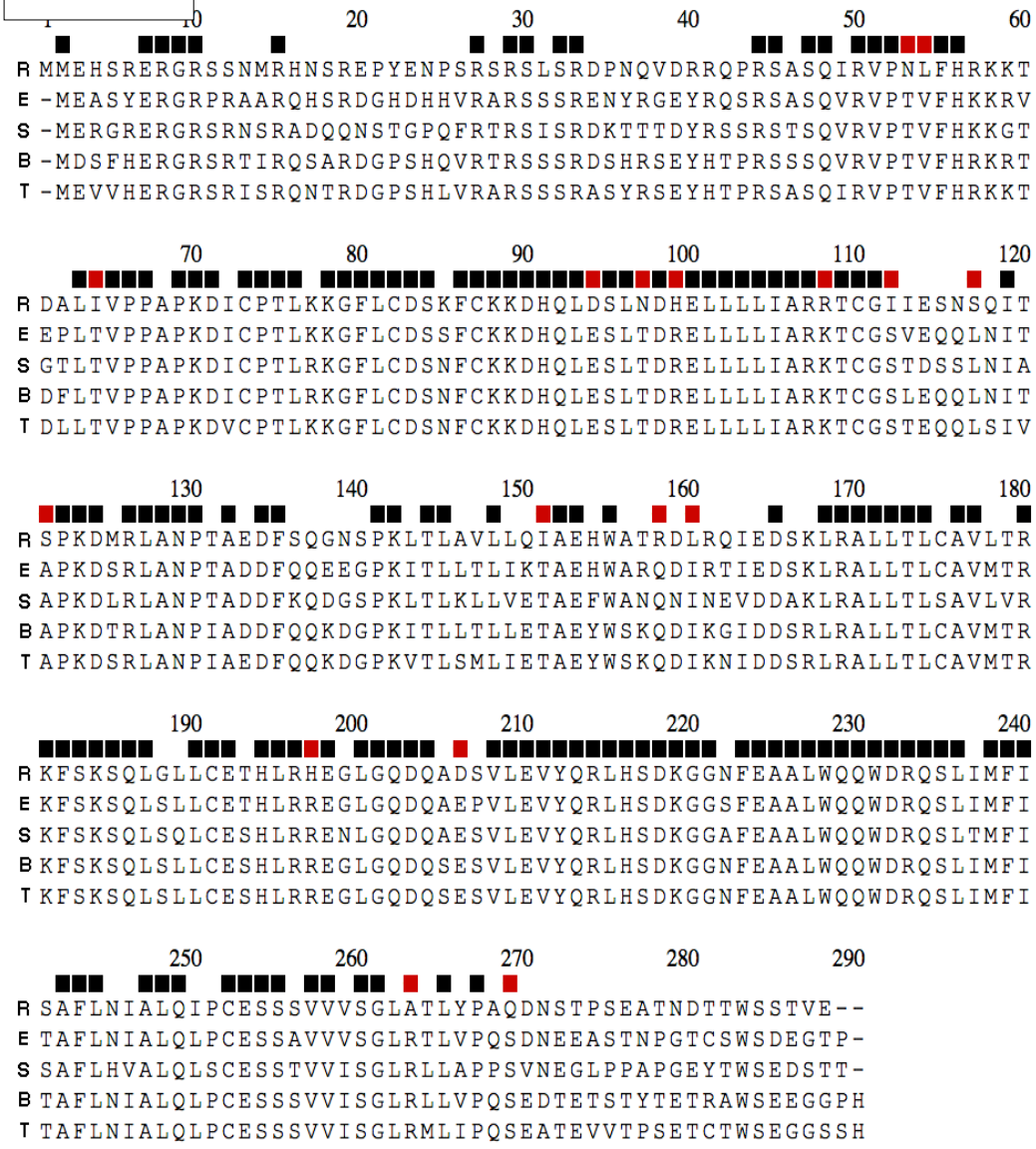
C – VP40


10 20 30 40 50 60
R MRRGVLPTAPPAYNDIAYPMSILPTRPSVIVNETKSDVLAVPGADVPSNSMRPVADDNID
E MRRVILPTAPPEYMEAIYPARSNSTIARGGNSNTGFLTPESVNGDTPSNPLRPIADDTID
S MRRVTVPTAPPAYADIGYPMSMLPIKSSRAVSGIQQKQEVLPGMDTPSNSMRPVADDNID
B MRRAILPTAPPEYMEAVYPMRTVSTNISSTSSGPNFPAPDVMMSDTPSNSLRPIADDNID
T MRRIILPTAPPEYMEAVYPMRTMNSGADNTASGPNYTTTGVMTNDTPSNSLRPVADDNID
70 80 90 100 110 120
R HSSHTPSGVASAFILEATVNVISGTKVLMKQIPIWLPLGVADQKIYSFDSTTAAIMLASY
E HASHTPGSVSSAFILEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSFDSTTAAIMLASY
S HTSHTPNGVASAFILEATVNVISGPKVLMKQIPIWLPLGIADQKTYSFDSTTAAIMLASY
B HPSHTPTSVSSAFILEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSFDSTTAAIMLASY
T HPSHTPNSVASAFILEAMVNVISGPKVLMKQIPIWLPLGVSDQKTYSFDSTTAAIMLASY
130 140 150 160 170 180
R TVTHFGKISNPLVRVNRLGPGIPDHPLRLLRLGNQAFLQEFVLPPVQLPQYFTFDLTALK
E TITHFGKATNPLVRVNRLGPGIPDHPLRLLRIGNQAFLQEFVLPPVQLPQYFTFDLTALK
S TITHFGKANNPLVRVNRLGQGIPDHPLRLLRMGNQAFLQEFVLPPVQLPQYFTFDLTALK
B TITHFGKTSNPLVRINRLGPGIPDHPLRLLRIGNQAFLQEFVLPPVQLPQYFTFDLTALK
T TITHFGKTSNPLVRINRLGPGIPDHPLRLLRIGNQAFLQEFVLPPVQLPQYFTFDLTALK
190 200 210 220 230 240
R LITQPLPAATWTDETPAGAVNALRPGLSLHPKLRPILLPGKTGKKGHASDLTSPDKIQTI
E LITQPLPAATWTDDTPTGSNGALRPGISFHPKLRPILLPNKSGKKGNSADLTSPEKIQAI
S LVTQPLPAATWTDETPSNLSGALRPGLSFHPKLRPVLLPGKTGKKGHVSDLTAPDKIQTI
B LITQPLPAATWTDDTPTGPTGILRPGISFHPKLRPILLPGKTGKRGSSSDLTSPDKIQAI
T LITQPLPAATWTDETPAVSTGTLRPGISFHPKLRPILLPGRAGKKGSNSDLTSPDKIQAI
250 260 270 280 290 300
R MNAIPDLKIVPIDPTKNIVGIEVPELLVQRLTGKKPQPKNGQPIIPVLLPKYVGLDPISP
E MTSLQDFKIVPIDPTKNIMGIEVPETLVHKLTGKKVTSKNGQPIIPVLLPKYIGLDPVAP
S VNLMQDFKIVPIDPAKSIIGIEVPELLVHKLTGKKMSQKNGQPIIPVLLPKYIGLDPISP
B MNFLQDLKLVPIDPAKNIMGIEVPELLVHRLTGKKITTKNGQPIIPILLPKYIGMDPISQ
T MNFLQDLKIVPIDPTKNIMGIEVPELLVHRLTGKKTTTKNGQPIIPILLPKYIGLDPLSQ
310 320 330
R GDLTMVITQDCDSCHSPASHPYHMDKQNSYQ
E GDLTMVITQDCDTCHSPASLPAVVEK-----
S GDLTMVITPDYDDCHSPASCSYLSEK-----
B GDLTMVITQDCDTCHSPASLPPVSEK-----
T GDLTMVITQDCDSCHSPASLPPVNEK-----

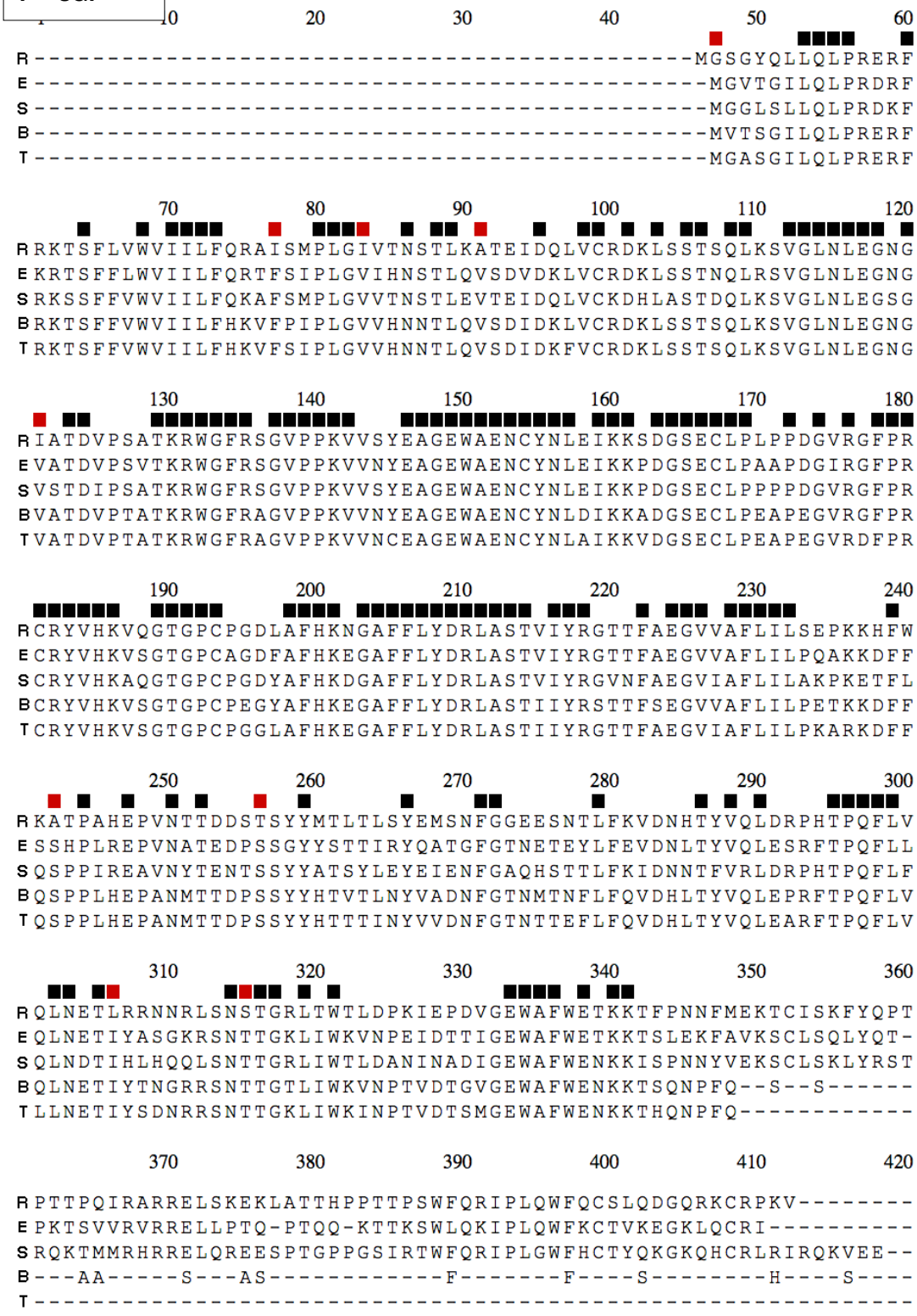
D – VP35

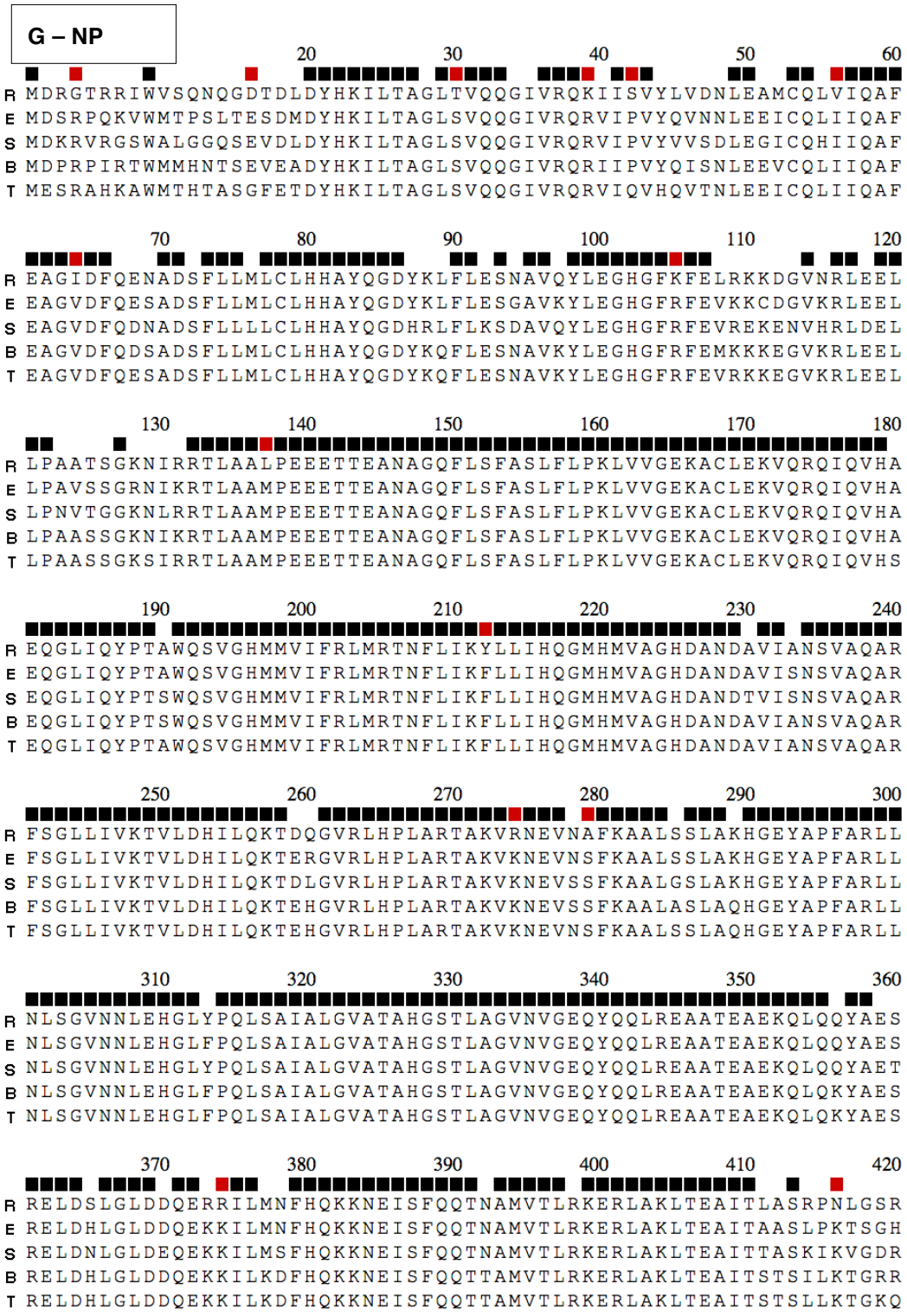

1 10 20 30 40 50 60
R ----------MYNNKLKVCSGPETTGWISEQLMTGKIPVTDIFIDIDNKPDQMEVRLK
E -MTTRTKGRGHTVATTQNDRMPGPELSGWISEQLMTGRIPVNDIFCDIENNPGLCYASQM
S ----------MQQDRTYRHHGPEVSGWFSEQLMTGKIPLTEVFVDVENKPSPAPITII
B MTSNRARVTYNPPPTTTGTRSCGPELSGWISEQLMTGKIPITDIFNEIETLPSISPSIHS
T MISTRAAAINDPSLPIRNQCTRGPELSGWISEQLMTGKIPVHEIFNDTEPHISSGSDCLP
70 80 90 100 110 120
R PSSRSSTRTCTSSSQTEVNYVPLLKKVEDTLTMLVNATSRQNAAIEALENRLSTLESSLK
E QQTKPNPKMRNSQTQTDPICNHSFEEVVQTLASLATVVQQQTIASESLEQRITSLENGLK
S SKNPKTTRKSDKQVQTDDASSLLTEEVKAAINSVISAVRRQTNAIESLEGRVTTLEASLK
B KIKTPSVQTRSVQTQTDPNCNHDFAEVVKMLTSLTLVVQKQTLATESLEQRITDLEGSLK
T RPKNTAPRTRNTQTQTDPVCNHNFEDVTQALTSLTNVIQKQALNLESLEQRIIDLENGLK
130 140 150 160 170 180
R PIQDMGKVISSLNRSCAEMVAKYDLLVMTTGRATSTAAAVDAYWKEHKQPPPGPALYEEN
E PVYDMAKTISSLNRVCAEMVAKYDLLVMTTGRATATAAATEAYWAEHGQPPPGPSLYEES
S PVQDMAKTISSLNRSCAEMVAKYDLLVMTTGRATATAAATEAYWNEHGQAPPGPSLYEDD
B PVSEITKIVSALNRSCAEMVAKYDLLVMTTGRATATAAATEAYWAEHGRPPPGPSLYEED
T PMYDMAKVISALNRSCAEMVAKYDLLVMTTGRATATAAATEAYWEEHGQPPPGPSLYEES
190 200 210 220 230 240
R ALKGKIDDPNSYVPDAVQEAYKNLDSTSTLTEENFGKPYISAKDLKEIMYDHLPGFGTAF
E AIRGKIESRDETVPQSVREAFNNLDSTTSLTEENFGKPDISAKDLRNIMYDHLPGFGTAF
S AIKAKLKDPNGKVPESVKQAYTNLDSTSALNEENFGRPYISAKDLKEIIYDHLPGFGTAF
B AIRTKIEKQGDIVPKEVQEAFRNLDSTALLTEENFGKPDISAKDLRNIMYDHLPGFGTAF
T AIRGKINKQEDKVPKEVQEAFRNLDSTSSLTEENFGKPDISAKDLRDIMYDHLPGFGTAF
250 260 270 280 290 300
R HQLVQVICKIGKDNNLLDTIHAEFQASLADGDSPQCALIQITKRVPIFQDVPPPIIHIRS
E HQLVQVICKLGKDSNSLDIIHAEFQASLAEGDSPQCALIQITKRVPIFQDAAPPVIHIRS
S HQLVQVICKIGKDNNILDIIHAEFQASLAEGDSPQCALIQITKRIPAFQDASPPIVHIKS
B HQLVQVICKLGKDNSSLDVIHAEFQASLAEGDSPQCALIQITKRIPIFQDAAPPVIHIRS
T HQLVQVICKLGKDNSALDIIHAEFQASLAEGDSPQCALIQITKRIPIFQDATPPTIHIRS
310 320 330 340
R RGDIPRACQKSLRPAPPSPKIDRGWVCLFKMQDGKTLGLKI
E RGDIPRACQKSLRPVPPSPKIDRGWVCVFQLQDGKTLGLKI
S RGDIPKACQKSLRPVPPSPKIDRGWVCIFQFQDGKALGLKI
B RGDIPKACQKSLRPVPPSPKIDRGWVCIFQLQDGKTLGLKI
T RGDIPRACQKSLRPVPPSPKIDRGWVCIFQLQDGKTLGLKI

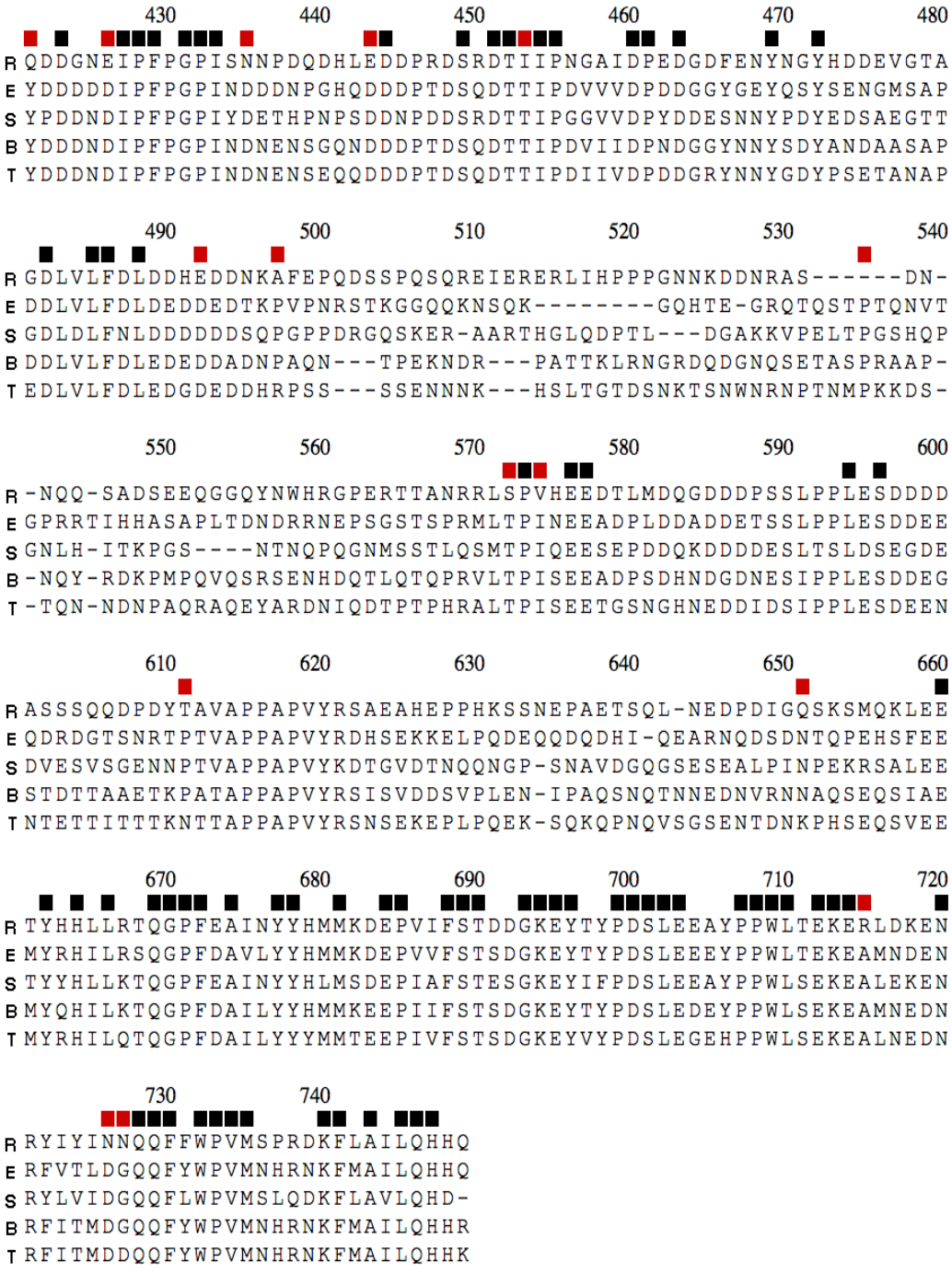
E – VP30

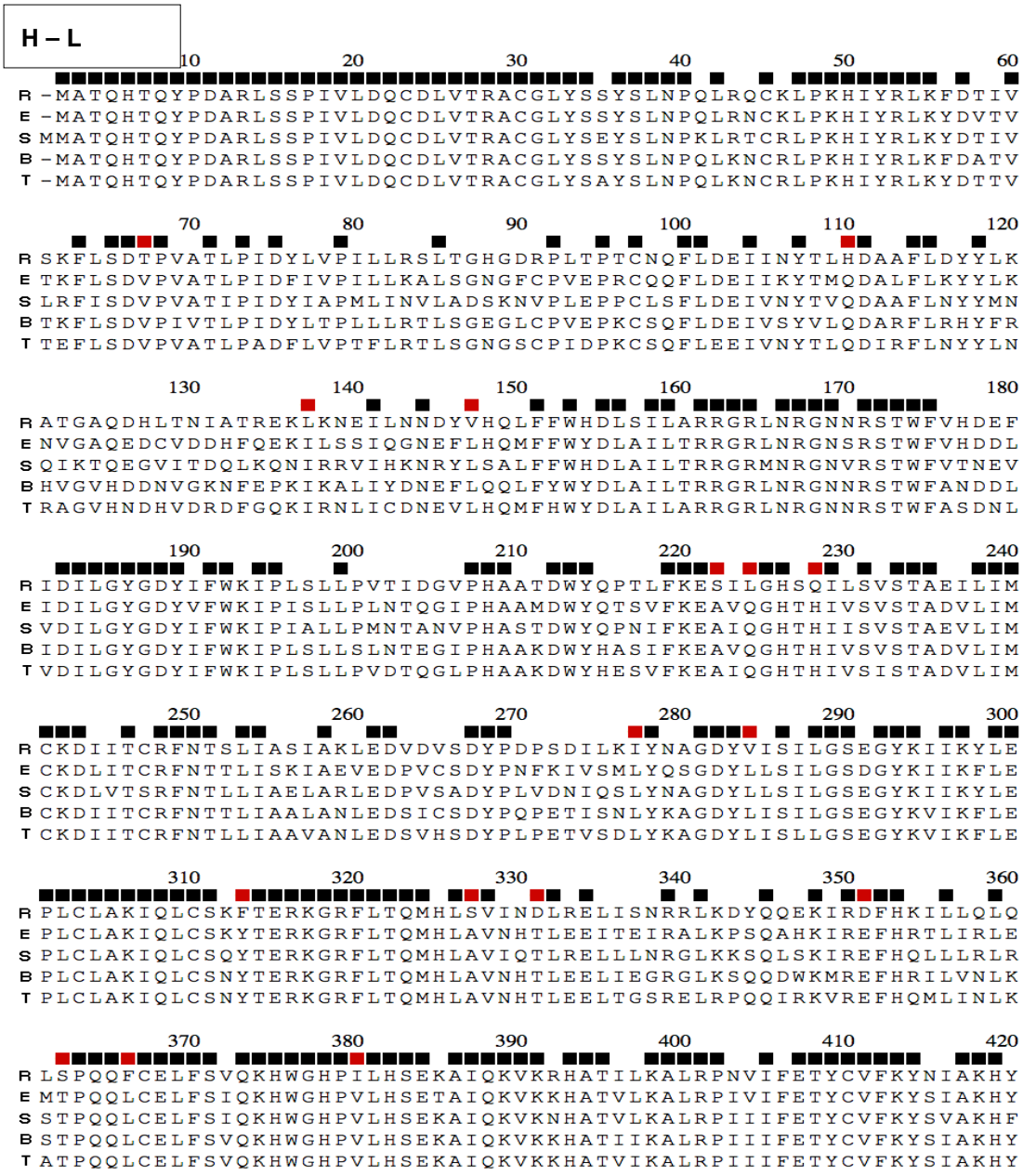

1 10 20 30 40 50 60
R MMEHSRERGRSSNMRHNSREPYENPSRSRSLSRDPNQVDRRQPRSASQIRVPNLFHRKKT
E -MEASYERGRPRAARQHSRDGHDHHVRARSSSRENYRGEYRQSRSASQVRVPTVFHKKRV
S -MERGRERGRSRNSRADQQNSTGPQFRTRSISRDKTTTDYRSSRSTSQVRVPTVFHKKGT
B -MDSFHERGRSRTIRQSARDGPSHQVRTRSSSRDSHRSEYHTPRSSSQVRVPTVFHRKRT
T -MEVVHERGRSRISRQNTRDGPSHLVRARSSSRASYRSEYHTPRSASQIRVPTVFHRKKT
70 80 90 100 110 120
R DALIVPPAPKDICPTLKKGFLCDSKFCKKDHQLDSLNDHELLLLIARRTCGIIESNSQIT
E EPLTVPPAPKDICPTLKKGFLCDSSFCKKDHQLESLTDRELLLLIARKTCGSVEQQLNIT
S GTLTVPPAPKDICPTLRKGFLCDSNFCKKDHQLESLTDRELLLLIARKTCGSTDSSLNIA
B DFLTVPPAPKDICPTLRKGFLCDSNFCKKDHQLESLTDRELLLLIARKTCGSLEQQLNIT
T DLLTVPPAPKDVCPTLKKGFLCDSNFCKKDHQLESLTDRELLLLIARKTCGSTEQQLSIV
130 140 150 160 170 180
R SPKDMRLANPTAEDFSQGNSPKLTLAVLLQIAEHWATRDLRQIEDSKLRALLTLCAVLTR
E APKDSRLANPTADDFQQEEGPKITLLTLIKTAEHWARQDIRTIEDSKLRALLTLCAVMTR
S APKDLRLANPTADDFKQDGSPKLTLKLLVETAEFWANQNINEVDDAKLRALLTLSAVLVR
B APKDTRLANPIADDFQQKDGPKITLLTLLETAEYWSKQDIKGIDDSRLRALLTLCAVMTR
T APKDSRLANPIAEDFQQKDGPKVTLSMLIETAEYWSKQDIKNIDDSRLRALLTLCAVMTR
190 200 210 220 230 240
R KFSKSQLGLLCETHLRHEGLGQDQADSVLEVYQRLHSDKGGNFEAALWQQWDRQSLIMFI
E KFSKSQLSLLCETHLRREGLGQDQAEPVLEVYQRLHSDKGGSFEAALWQQWDRQSLIMFI
S KFSKSQLSQLCESHLRRENLGQDQAESVLEVYQRLHSDKGGAFEAALWQQWDRQSLTMFI
B KFSKSQLSLLCESHLRREGLGQDQSESVLEVYQRLHSDKGGNFEAALWQQWDRQSLIMFI
T KFSKSQLSLLCESHLRREGLGQDQSESVLEVYQRLHSDKGGNFEAALWQQWDRQSLIMFI
250 260 270 280 290
R SAFLNIALQIPCESSSVVVSGLATLYPAQDNSTPSEATNDTTWSSTVE--
E TAFLNIALQLPCESSAVVVSGLRTLVPQSDNEEASTNPGTCSWSDEGTP-
S SAFLHVALQLSCESSTVVISGLRLLAPPSVNEGLPPAPGEYTWSEDSTT-
B TAFLNIALQLPCESSSVVISGLRLLVPQSEDTETSTYTETRAWSEEGGPH
T TAFLNIALQLPCESSSVVISGLRMLIPQSEATEVVTPSETCTWSEGGSSH

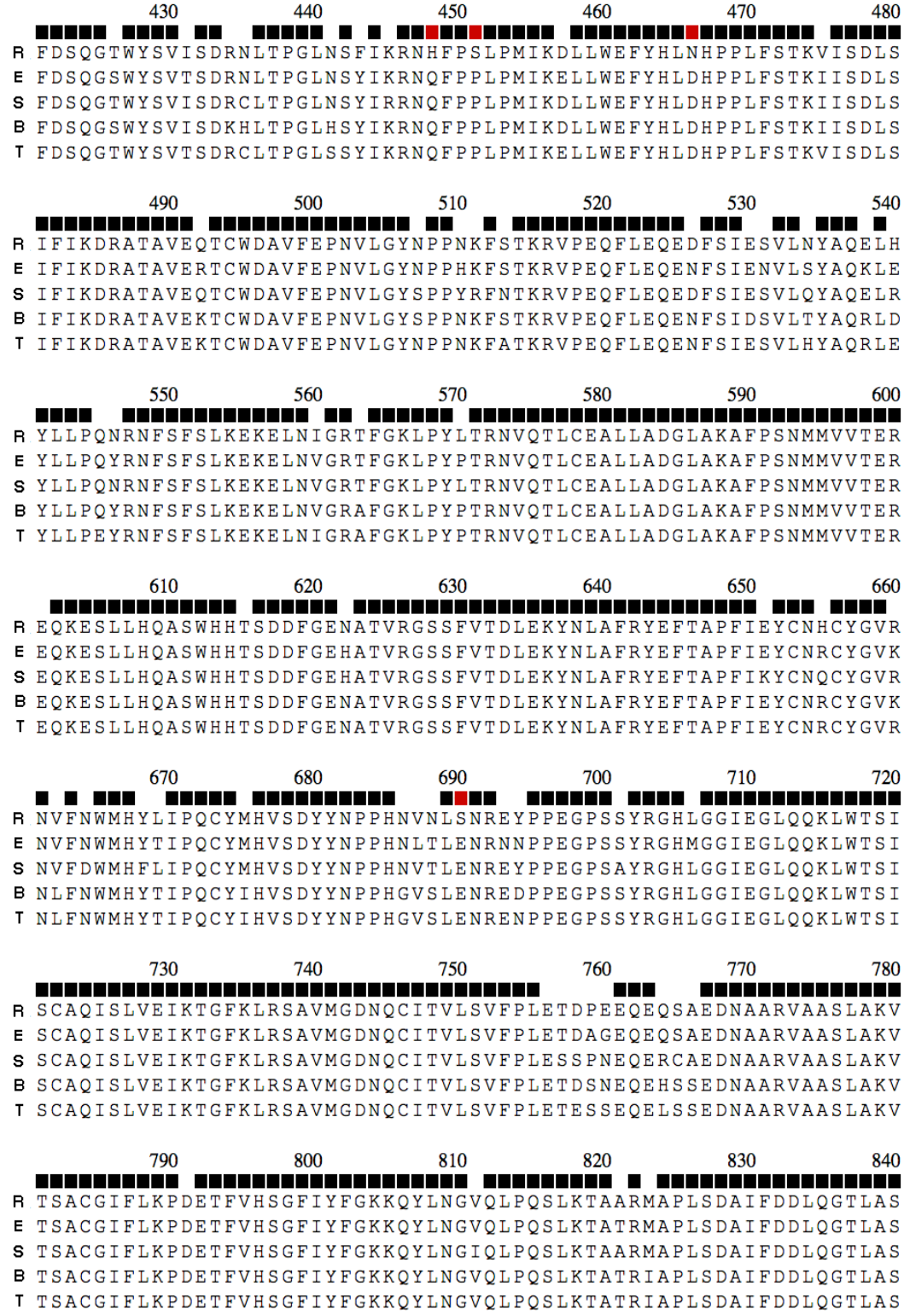
F – sGP

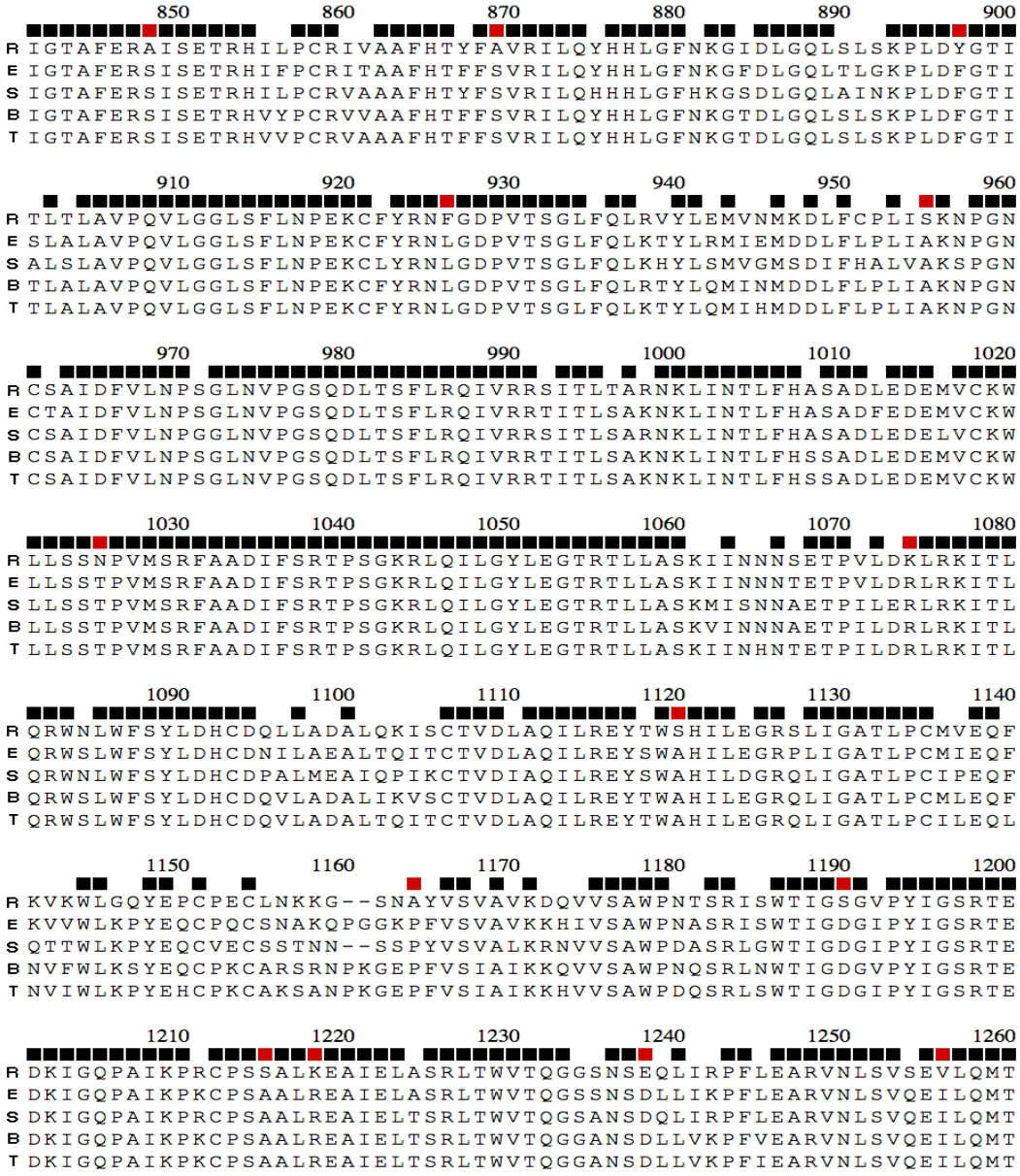

10 20 30 40 50 60
R ----------MGSGYQLLQLPRERF
E ----------MGVTGILQLPRDRF
S ----------MGGLSLLQLPRDKF
B ----------MVTSGILQLPRERF
T ----------MGASGILQLPRERF
70 80 90 100 110 120
R RKTSFLVWVIILFQRAISMPLGIVTNSTLKATEIDQLVCRDKLSSTSQLKSVGLNLEGNG
E KRTSFFLWVIILFQRTFSIPLGVIHNSTLQVSDVDKLVCRDKLSSTNQLRSVGLNLEGNG
S RKSSFFVWVIILFQKAFSMPLGVVTNSTLEVTEIDQLVCKDHLASTDQLKSVGLNLEGSG
B RKTSFFVWVIILFHKVFPIPLGVVHNNTLQVSDIDKLVCRDKLSSTSQLKSVGLNLEGNG
T RKTSFFVWVIILFHKVFSIPLGVVHNNTLQVSDIDKFVCRDKLSSTSQLKSVGLNLEGNG
130 140 150 160 170 180
R IATDVPSATKRWGFRSGVPPKVVSYEAGEWAENCYNLEIKKSDGSECLPLPPDGVRGFPR
E VATDVPSVTKRWGFRSGVPPKVVNYEAGEWAENCYNLEIKKPDGSECLPAAPDGIRGFPR
S VSTDIPSATKRWGFRSGVPPKVVSYEAGEWAENCYNLEIKKPDGSECLPPPPDGVRGFPR
B VATDVPTATKRWGFRAGVPPKVVNYEAGEWAENCYNLDIKKADGSECLPEAPEGVRGFPR
T VATDVPTATKRWGFRAGVPPKVVNCEAGEWAENCYNLAIKKVDGSECLPEAPEGVRDFPR
190 200 210 220 230 240
R CRYVHKVQGTGPCPGDLAFHKNGAFFLYDRLASTVIYRGTTFAEGVVAFLILSEPKKHFW
E CRYVHKVSGTGPCAGDFAFHKEGAFFLYDRLASTVIYRGTTFAEGVVAFLILPQAKKDFF
S CRYVHKAQGTGPCPGDYAFHKDGAFFLYDRLASTVIYRGVNFAEGVIAFLILAKPKETFL
B CRYVHKVSGTGPCPEGYAFHKEGAFFLYDRLASTIIYRSTTFSEGVVAFLILPETKKDFF
T CRYVHKVSGTGPCPGGLAFHKEGAFFLYDRLASTIIYRGTTFAEGVIAFLILPKARKDFF
250 260 270 280 290 300
R KATPAHEPVNTTDDSTSYYMTLTLSYEMSNFGGEESNTLFKVDNHTYVQLDRPHTPQFLV
E SSHPLREPVNATEDPSSGYYSTTIRYQATGFGTNETEYLFEVDNLTYVQLESRFTPQFLL
S QSPPIREAVNYTENTSSYYATSYLEYEIENFGAQHSTTLFKIDNNTFVRLDRPHTPQFLF
B QSPPLHEPANMTTDPSSYYHTVTLNYVADNFGTNMTNFLFQVDHLTYVQLEPRFTPQFLV
T QSPPLHEPANMTTDPSSYYHTTTINYVVDNFGTNTTEFLFQVDHLTYVQLEARFTPQFLV
310 320 330 340 350 360
R QLNETLRRNNRLSNSTGRLTWTLDPKIEPDVGEWAFWETKKTFPNNFMEKTCISKFYQPT
E QLNETIYASGKRSNTTGKLIWKVNPEIDTTIGEWAFWETKKTSLEKFAVKSCLSQLYQT-
S QLNDTIHLHQQLSNTTGRLIWTLDANINADIGEWAFWENKKISPNNYVEKSCLSKLYRST
B QLNETIYTNGRRSNTTGTLIWKVNPTVDTGVGEWAFWENKKTSQNPFQ--S--S------
T LLNETIYSDNRRSNTTGKLIWKINPTVDTSMGEWAFWENKKTHQNPFQ-----------
370 380 390 400 410 420
R PTTPQIRARRELSKEKLATTHPPTTPSWFQRIPLQWFQCSLQDGQRKCRPKV--------
E PKTSVVRVRRELLPTQ-PTQQ-KTTKSWLQKIPLQWFKCTVKEGKLQCRI----------
S RQKTMMRHRRELQREESPTGPPGSIRTWFQRIPLGWFHCTYQKGKQHCRLRIRQKVEE--
B ---AA-----S---AS----------F-------F----S--------H----S----
T ----------

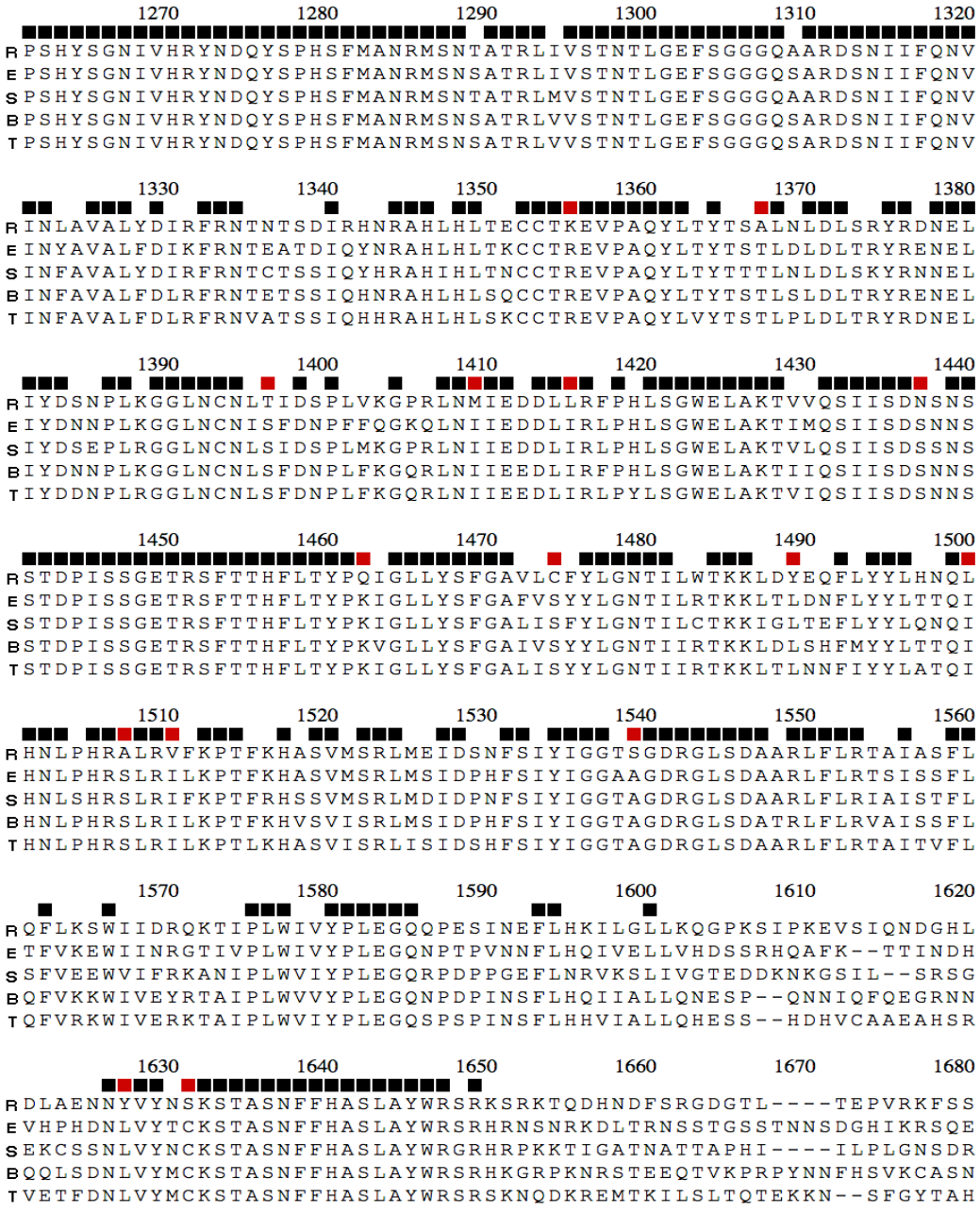

G – NP
20 30 40 50 60
R MDRGTRRIWVSQNQGDTDLDYHKILTAGLTVQQGIVRQKIISVYLVDNLEAMCQLVIQAF
E MDSRPQKVWMTPSLTESDMDYHKILTAGLSVQQGIVRQRVIPVYQVNNLEEICQLIIQAF
S MDKRVRGSWALGGQSEVDLDYHKILTAGLSVQQGIVRQRVIPVYVVSDLEGICQHIIQAF
B MDPRPIRTWMMHNTSEVEADYHKILTAGLSVQQGIVRQRIIPVYQISNLEEVCQLIIQAF
T MESRAHKAWMTHTASGFETDYHKILTAGLSVQQGIVRQRVIQVHQVTNLEEICQLIIQAF
70 80 90 100 110 120
R EAGIDFQENADSFLLMLCLHHAYQGDYKLFLESNAVQYLEGHGFKFELRKKDGVNRLEEL
E EAGVDFQESADSFLLMLCLHHAYQGDYKLFLESGAVKYLEGHGFRFEVKKCDGVKRLEEL
S EAGVDFQDNADSFLLLLCLHHAYQGDHRLFLKSDAVQYLEGHGFRFEVREKENVHRLDEL
B EAGVDFQDSADSFLLMLCLHHAYQGDYKQFLESNAVKYLEGHGFRFEMKKKEGVKRLEEL
T EAGVDFQESADSFLLMLCLHHAYQGDYKQFLESNAVKYLEGHGFRFEVRKKEGVKRLEEL
130 140 150 160 170 180
R LPAATSGKNIRRTLAALPEEETTEANAGQFLSFASLFLPKLVVGEKACLEKVQRQIQVHA
E LPAVSSGRNIKRTLAAMPEEETTEANAGQFLSFASLFLPKLVVGEKACLEKVQRQIQVHA
S LPNVTGGKNLRRTLAAMPEEETTEANAGQFLSFASLFLPKLVVGEKACLEKVQRQIQVHA
B LPAASSGKNIKRTLAAMPEEETTEANAGQFLSFASLFLPKLVVGEKACLEKVQRQIQVHA
T LPAASSGKSIRRTLAAMPEEETTEANAGQFLSFASLFLPKLVVGEKACLEKVQRQIQVHS
190 200 210 220 230 240
R EQGLIQYPTAWQSVGHMMVIFRLMRTNFLIKYLLIHQGMHMVAGHDANDAVIANSVAQAR
E EQGLIQYPTAWQSVGHMMVIFRLMRTNFLIKFLLIHQGMHMVAGHDANDAVISNSVAQAR
S EQGLIQYPTSWQSVGHMMVIFRLMRTNFLIKFLLIHQGMHMVAGHDANDTVISNSVAQAR
B EQGLIQYPTSWQSVGHMMVIFRLMRTNFLIKFLLIHQGMHMVAGHDANDAVIANSVAQAR
T EQGLIQYPTAWQSVGHMMVIFRLMRTNFLIKFLLIHQGMHMVAGHDANDAVIANSVAQAR
250 260 270 280 290 300
R FSGLLIVKTVLDHILQKTDQGVRLHPLARTAKVRNEVNAFKAALSSLAKHGEYAPFARLL
E FSGLLIVKTVLDHILQKTERGVRLHPLARTAKVKNEVNSFKAALSSLAKHGEYAPFARLL
S FSGLLIVKTVLDHILQKTDLGVRLHPLARTAKVKNEVSSFKAALGSLAKHGEYAPFARLL
B FSGLLIVKTVLDHILQKTEHGVRLHPLARTAKVKNEVSSFKAALASLAQHGEYAPFARLL
T FSGLLIVKTVLDHILQKTEHGVRLHPLARTAKVKNEVNSFKAALSSLAQHGEYAPFARLL
310 320 330 340 350 360
R NLSGVNNLEHGLYPQLSAIALGVATAHGSTLAGVNVGEQYQQLREAATEAEKQLQQYAES
E NLSGVNNLEHGLFPQLSAIALGVATAHGSTLAGVNVGEQYQQLREAATEAEKQLQQYAES
S NLSGVNNLEHGLYPQLSAIALGVATAHGSTLAGVNVGEQYQQLREAATEAEKQLQQYAET
B NLSGVNNLEHGLFPQLSAIALGVATAHGSTLAGVNVGEQYQQLREAATEAEKQLQKYAES
T NLSGVNNLEHGLFPQLSAIALGVATAHGSTLAGVNVGEQYQQLREAATEAEKQLQKYAES
370 380 390 400 410 420
R RELDSLGLDDQERRILMNFHQKKNEISFQQTNAMVTLRKERLAKLTEAITLASRPNLGSR
E RELDHLGLDDQEKKILMNFHQKKNEISFQQTNAMVTLRKERLAKLTEAITAASLPKTSGH
S RELDNLGLDEQEKKILMSFHQKKNEISFQQTNAMVTLRKERLAKLTEAITTASKIKVGDR
B RELDHLGLDDQEKKILKDFHQKKNEISFQQTTAMVTLRKERLAKLTEAITSTSILKTGRR
T RELDHLGLDDQEKKILKDFHQKKNEISFQQTTAMVTLRKERLAKLTEAITSTSLLKTGKQ

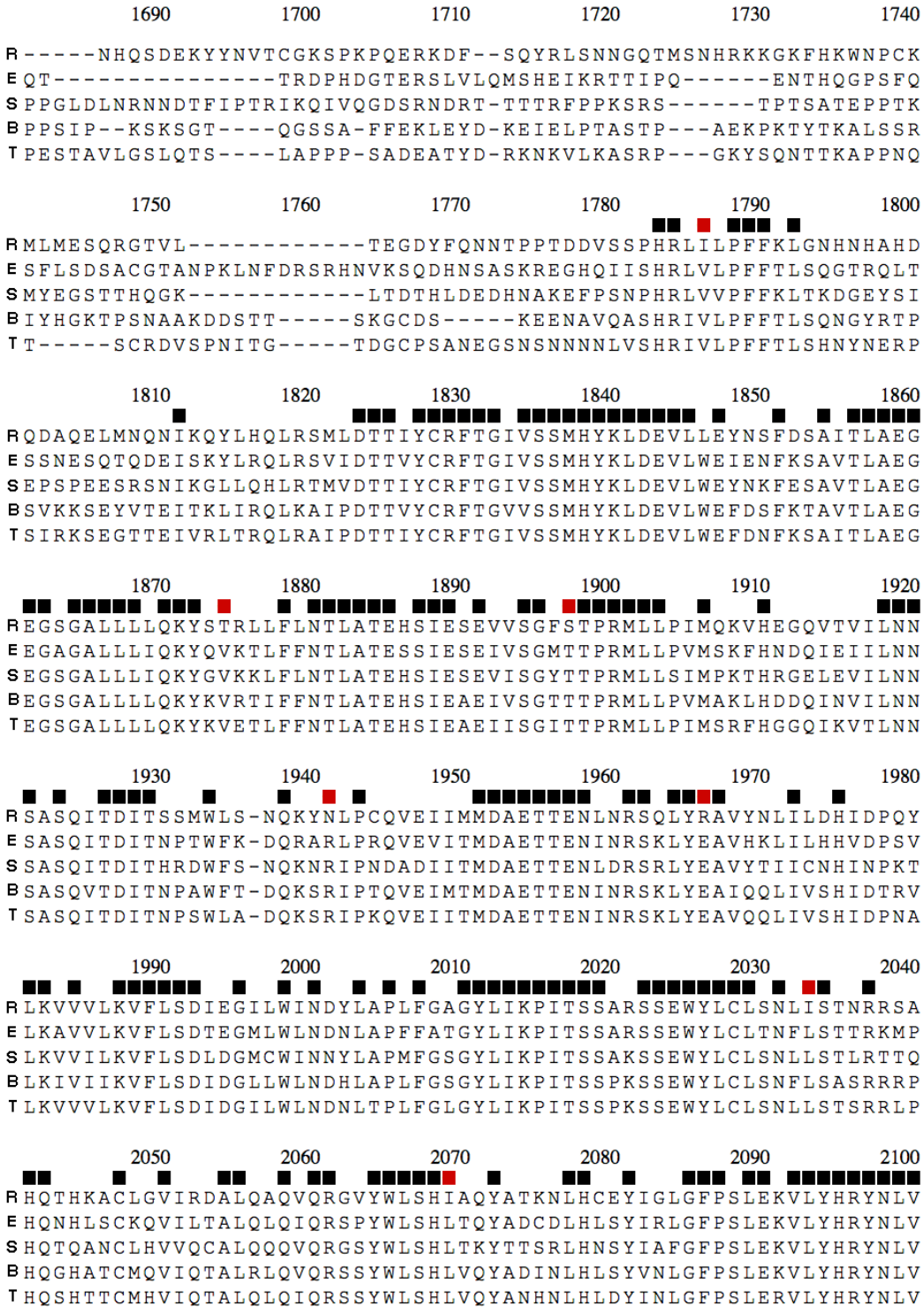

430 440 450 460 470 480
R QDDGNEIPFPGPISNNPDQDHLEDDPRDSRDTIIPNGAIDPEDGDFENYNGYHDDEVGTA
E YDDDDDIPFPGPINDDDNPGHQDDDPTDSQDTTIPDVVVDPDDGGYGEYQSYSENGMSAP
S YPDDNDIPFPGPIYDETHPNPSDDNPDDSRDTTIPGGVVDPYDDESNNYPDYEDSAEGTT
B YDDDNDIPFPGPINDNENSGQNDDDPTDSQDTTIPDVIIDPNDGGYNNYSDYANDAASAP
T YDDDNDIPFPGPINDNENSEQQDDDPTDSQDTTIPDIIVDPDDGRYNNYGDYPSETANAP
490 500 510 520 530 540
R GDLVLFDLDDHEDDNKAFEPQDSSPQSQREIERERLIHPPPGNNKDDNRAS------DN-
E DDLVLFDLDEDDEDTKPVPNRSTKGGQQKNSQK--------GQHTE-GRQTQSTPTQNVT
S GDLDLFNLDDDDDDSQPGPPDRGQSKER-AARTHGLQDPTL---DGAKKVPELTPGSHQP
B DDLVLFDLEDEDDADNPAQN---TPEKNDR---PATTKLRNGRDQDGNQSETASPRAAP-
T EDLVLFDLEDGDEDDHRPSS---SSENNNK---HSLTGTDSNKTSNWNRNPTNMPKKDS-
550 560 570 580 590 600
R -NQQ-SADSEEQGGQYNWHRGPERTTANRRLSPVHEEDTLMDQGDDDPSSLPPLESDDDD
E GPRRTIHHASAPLTDNDRRNEPSGSTSPRMLTPINEEADPLDDADDETSSLPPLESDDEE
S GNLH-ITKPGS----NTNQPQGNMSSTLQSMTPIQEESEPDDQKDDDDESLTSLDSEGDE
B -NQY-RDKPMPQVQSRSENHDQTLQTQPRVLTPISEEADPSDHNDGDNESIPPLESDDEG
T -TQN-NDNPAQRAQEYARDNIQDTPTPHRALTPISEETGSNGHNEDDIDSIPPLESDEEN
610 620 630 640 650 660
R ASSSQQDPDYTAVAPPAPVYRSAEAHEPPHKSSNEPAETSQL-NEDPDIGQSKSMQKLEE
E QDRDGTSNRTPTVAPPAPVYRDHSEKKELPQDEQQDQDHI-QEARNQDSDNTQPEHSFEE
S DVESVSGENNPTVAPPAPVYKDTGVDTNQQNGP-SNAVDGQGSESEALPINPEKRSALEE
B STDTTAAETKPATAPPAPVYRSISVDDSVPLEN-IPAQSNQTNNEDNVRNNAQSEQSIAE
T NTETTITTTKNTTAPPAPVYRSNSEKEPLPQEK-SQKQPNQVSGSENTDNKPHSEQSVEE
670 680 690 700 710 720
R TYHHLLRTQGPFEAINYYHMMKDEPVIFSTDDGKEYTYPDSLEEAYPPWLTEKERLDKEN
E MYRHILRSQGPFDAVLYYHMMKDEPVVFSTSDGKEYTYPDSLEEEYPPWLTEKEAMNDEN
S TYYHLLKTQGPFEAINYYHLMSDEPIAFSTESGKEYIFPDSLEEAYPPWLSEKEALEKEN
B MYQHILKTQGPFDAILYYHMMKEEPIIFSTSDGKEYTYPDSLEDEYPPWLSEKEAMNEDN
T MYRHILQTQGPFDAILYYYMMTEEPIVFSTSDGKEYVYPDSLEGEHPPWLSEKEALNEDN
730 740
R RYIYINNQQFFWPVMSPRDKFLAILQHHQ
E RFVTLDGQQFYWPVMNHRNKFMAILQHHQ
S RYLVIDGQQFLWPVMSLQDKFLAVLQHD-
B RFITMDGQQFYWPVMNHRNKFMAILQHHR
T RFITMDDQQFYWPVMNHRNKFMAILQHHK

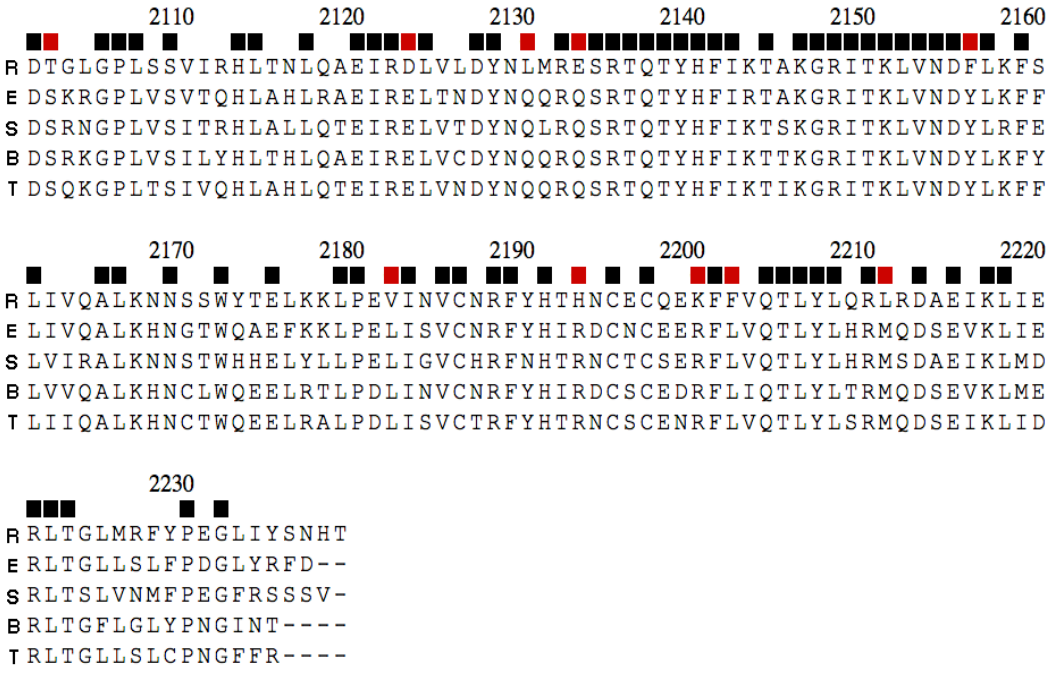

H – L
10 20 30 40 50 60
R -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRQCKLPKHIYRLKFDTIV
E -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKLPKHIYRLKYDVTV
S MMATQHTQYPDARLSSPIVLDQCDLVTRACGLYSEYSLNPKLRTCRLPKHIYRLKYDTIV
B -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLKNCRLPKHIYRLKFDATV
T -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSAYSLNPQLKNCRLPKHIYRLKYDTTV
70 80 90 100 110 120
R SKFLSDTPVATLPIDYLVPILLRSLTGHGDRPLTPTCNQFLDEIINYTLHDAAFLDYYLK
E TKFLSDVPVATLPIDFIVPILLKALSGNGFCPVEPRCQQFLDEIIKYTMQDALFLKYYLK
S LRFISDVPVATIPIDYIAPMLINVLADSKNVPLEPPCLSFLDEIVNYTVQDAAFLNYYMN
B TKFLSDVPIVTLPIDYLTPLLLRTLSGEGLCPVEPKCSQFLDEIVSYVLQDARFLRHYFR
T TEFLSDVPVATLPADFLVPTFLRTLSGNGSCPIDPKCSQFLEEIVNYTLQDIRFLNYYLN
130 140 150 160 170 180
R ATGAQDHLTNIATREKLKNEILNNDYVHQLFFWHDLSILARRGRLNRGNNRSTWFVHDEF
E NVGAQEDCVDDHFQEKILSSIQGNEFLHQMFFWYDLAILTRRGRLNRGNSRSTWFVHDDL
S QIKTQEGVITDQLKQNIRRVIHKNRYLSALFFWHDLAILTRRGRMNRGNVRSTWFVTNEV
B HVGVHDDNVGKNFEPKIKALIYDNEFLQQLFYWYDLAILTRRGRLNRGNNRSTWFANDDL
T RAGVHNDHVDRDFGQKIRNLICDNEVLHQMFHWYDLAILARRGRLNRGNNRSTWFASDNL
190 200 210 220 230 240
R IDILGYGDYIFWKIPLSLLPVTIDGVPHAATDWYQPTLFKESILGHSQILSVSTAEILIM
E IDILGYGDYVFWKIPISLLPLNTQGIPHAAMDWYQTSVFKEAVQGHTHIVSVSTADVLIM
S VDILGYGDYIFWKIPIALLPMNTANVPHASTDWYQPNIFKEAIQGHTHIISVSTAEVLIM
B IDILGYGDYIFWKIPLSLLSLNTEGIPHAAKDWYHASIFKEAVQGHTHIVSVSTADVLIM
T VDILGYGDYIFWKIPLSLLPVDTQGLPHAAKDWYHESVFKEAIQGHTHIVSISTADVLIM
250 260 270 280 290 300
R CKDIITCRFNTSLIASIAKLEDVDVSDYPDPSDILKIYNAGDYVISILGSEGYKIIKYLE
E CKDLITCRFNTTLISKIAEVEDPVCSDYPNFKIVSMLYQSGDYLLSILGSDGYKIIKFLE
S CKDLVTSRFNTLLIAELARLEDPVSADYPLVDNIQSLYNAGDYLLSILGSEGYKIIKYLE
B CKDIITCRFNTTLIAALANLEDSICSDYPQPETISNLYKAGDYLISILGSEGYKVIKFLE
T CKDIITCRFNTLLIAAVANLEDSVHSDYPLPETVSDLYKAGDYLISLLGSEGYKVIKFLE
310 320 330 340 350 360
R PLCLAKIQLCSKFTERKGRFLTQMHLSVINDLRELISNRRLKDYQQEKIRDFHKILLQLQ
E PLCLAKIQLCSKYTERKGRFLTQMHLAVNHTLEEITEIRALKPSQAHKIREFHRTLIRLE
S PLCLAKIQLCSQYTERKGRFLTQMHLAVIQTLRELLLNRGLKKSQLSKIREFHQLLLRLR
B PLCLAKIQLCSNYTERKGRFLTQMHLAVNHTLEELIEGRGLKSQQDWKMREFHRILVNLK
T PLCLAKIQLCSNYTERKGRFLTQMHLAVNHTLEELTGSRELRPQQIRKVREFHQMLINLK
370 380 390 400 410 420
R LSPQQFCELFSVQKHWGHPILHSEKAIQKVKRHATILKALRPNVIFETYCVFKYNIAKHY
E MTPQQLCELFSIQKHWGHPVLHSETAIQKVKKHATVLKALRPIVIFETYCVFKYSIAKHY
S STPQQLCELFSIQKHWGHPVLHSEKAIQKVKNHATVLKALRPIIIFETYCVFKYSVAKHF
B STPQQLCELFSVQKHWGHPVLHSEKAIQKVKKHATIIKALRPIIIFETYCVFKYSIAKHY
T ATPQQLCELFSVQKHWGHPVLHSEKAIQKVKKHATVIKALRPIIIFETYCVFKYSIAKHY

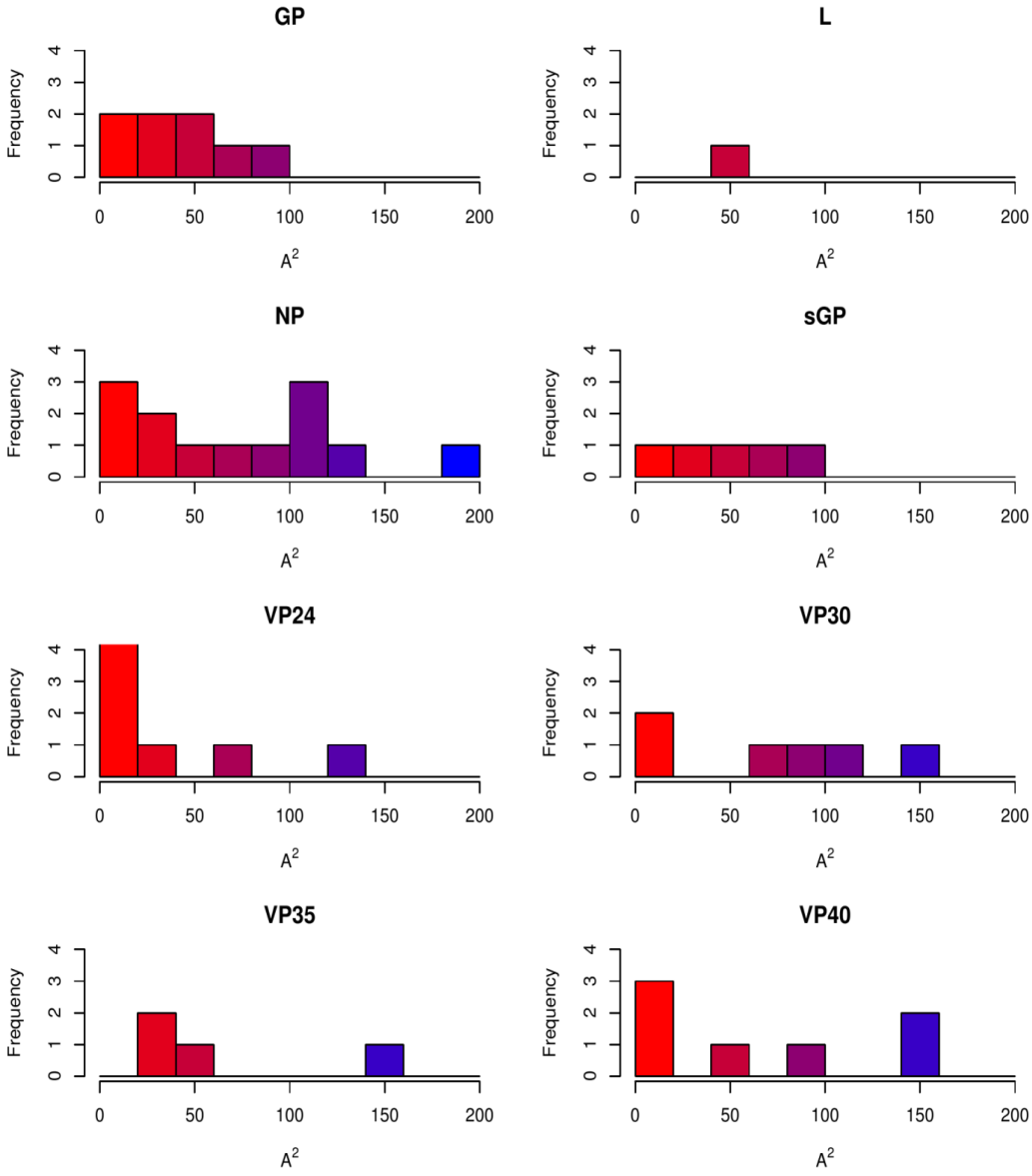

430 440 450 460 470 480
R FDSQGTWYSVISDRNLTPGLNSFIKRNHFPSLPMIKDLLWEFYHLNHPPLFSTKVISDLS
E FDSQGSWYSVTSDRNLTPGLNSYIKRNQFPPLPMIKELLWEFYHLDHPPLFSTKIISDLS
S FDSQGTWYSVISDRCLTPGLNSYIRRNQFPPLPMIKDLLWEFYHLDHPPLFSTKIISDLS
B FDSQGSWYSVISDKHLTPGLHSYIKRNQFPPLPMIKDLLWEFYHLDHPPLFSTKIISDLS
T FDSQGTWYSVTSDRCLTPGLSSYIKRNQFPPLPMIKELLWEFYHLDHPPLFSTKVISDLS
490 500 510 520 530 540
R IFIKDRATAVEQTCWDAVFEPNVLGYNPPNKFSTKRVPEQFLEQEDFSIESVLNYAQELH
E IFIKDRATAVERTCWDAVFEPNVLGYNPPHKFSTKRVPEQFLEQENFSIENVLSYAQKLE
S IFIKDRATAVEQTCWDAVFEPNVLGYSPPYRFNTKRVPEQFLEQEDFSIESVLQYAQELR
B IFIKDRATAVEKTCWDAVFEPNVLGYSPPNKFSTKRVPEQFLEQENFSIDSVLTYAQRLD
T IFIKDRATAVEKTCWDAVFEPNVLGYNPPNKFATKRVPEQFLEQENFSIESVLHYAQRLE
550 560 570 580 590 600
R YLLPQNRNFSFSLKEKELNIGRTFGKLPYLTRNVQTLCEALLADGLAKAFPSNMMVVTER
E YLLPQYRNFSFSLKEKELNVGRTFGKLPYPTRNVQTLCEALLADGLAKAFPSNMMVVTER
S YLLPQNRNFSFSLKEKELNVGRTFGKLPYLTRNVQTLCEALLADGLAKAFPSNMMVVTER
B YLLPQYRNFSFSLKEKELNVGRAFGKLPYPTRNVQTLCEALLADGLAKAFPSNMMVVTER
T YLLPEYRNFSFSLKEKELNIGRAFGKLPYPTRNVQTLCEALLADGLAKAFPSNMMVVTER
610 620 630 640 650 660
R EQKESLLHQASWHHTSDDFGENATVRGSSFVTDLEKYNLAFRYEFTAPFIEYCNHCYGVR
E EQKESLLHQASWHHTSDDFGEHATVRGSSFVTDLEKYNLAFRYEFTAPFIEYCNRCYGVK
S EQKESLLHQASWHHTSDDFGEHATVRGSSFVTDLEKYNLAFRYEFTAPFIKYCNQCYGVR
B EQKESLLHQASWHHTSDDFGENATVRGSSFVTDLEKYNLAFRYEFTAPFIEYCNRCYGVK
T EQKESLLHQASWHHTSDDFGENATVRGSSFVTDLEKYNLAFRYEFTAPFIEYCNRCYGVR
670 680 690 700 710 720
R NVFNWMHYLIPQCYMHVSDYYNPPHNVNLSNREYPPEGPSSYRGHLGGIEGLQQKLWTSI
E NVFNWMHYTIPQCYMHVSDYYNPPHNLTLENRNNPPEGPSSYRGHMGGIEGLQQKLWTSI
S NVFDWMHFLIPQCYMHVSDYYNPPHNVTLENREYPPEGPSAYRGHLGGIEGLQQKLWTSI
B NLFNWMHYTIPQCYIHVSDYYNPPHGVSLENREDPPEGPSSYRGHLGGIEGLQQKLWTSI
T NLFNWMHYTIPQCYIHVSDYYNPPHGVSLENRENPPEGPSSYRGHLGGIEGLQQKLWTSI
730 740 750 760 770 780
R SCAQISLVEIKTGFKLRSAVMGDNQCITVLSVFPLETDPEEQEQSAEDNAARVAASLAKV
E SCAQISLVEIKTGFKLRSAVMGDNQCITVLSVFPLETDAGEQEQSAEDNAARVAASLAKV
S SCAQISLVEIKTGFKLRSAVMGDNQCITVLSVFPLESSPNEQERCAEDNAARVAASLAKV
B SCAQISLVEIKTGFKLRSAVMGDNQCITVLSVFPLETDSNEQEHSSEDNAARVAASLAKV
T SCAQISLVEIKTGFKLRSAVMGDNQCITVLSVFPLETESSEQELSSEDNAARVAASLAKV
790 800 810 820 830 840
R TSACGIFLKPDETFVHSGFIYFGKKQYLNGVQLPQSLKTAARMAPLSDAIFDDLQGTLAS
E TSACGIFLKPDETFVHSGFIYFGKKQYLNGVQLPQSLKTATRMAPLSDAIFDDLQGTLAS
S TSACGIFLKPDETFVHSGFIYFGKKQYLNGIQLPQSLKTAARMAPLSDAIFDDLQGTLAS
B TSACGIFLKPDETFVHSGFIYFGKKQYLNGVQLPQSLKTATRIAPLSDAIFDDLQGTLAS
T TSACGIFLKPDETFVHSGFIYFGKKQYLNGVQLPQSLKTATRIAPLSDAIFDDLQGTLAS


850 860 870 880 890 900
R IGTAFERAISETRHILPCRIVAAFHTYFAVRILQYHHLGFNKGIDLGQLSLSKPLDYGTI
E IGTAFERSISETRHIFPCRITAAFHTFFSVRILQYHHLGFNKGFDLGQLTLGKPLDFGTI
S IGTAFERSISETRHILPCRVAAAFHTYFSVRILQHHHLGFHKGSDLGQLAINKPLDFGTI
B IGTAFERSISETRHVYPCRVVAAFHTFFSVRILQYHHLGFNKGTDLGQLSLSKPLDFGTI
T IGTAFERSISETRHVVPCRVAAAFHTFFSVRILQYHHLGFNKGTDLGQLSLSKPLDFGTI
910 920 930 940 950 960
R TLTLAVPQVLGGLSFLNPEKCFYRNFGDPVTSGLFQLRVYLEMVNMKDLFCPLISKNPGN
E SLALAVPQVLGGLSFLNPEKCFYRNLGDPVTSGLFQLKTYLRMIEMDDLFLPLIAKNPGN
S ALSLAVPQVLGGLSFLNPEKCLYRNLGDPVTSGLFQLKHYLSMVGMSDIFHALVAKSPGN
B TLALAVPQVLGGLSFLNPEKCFYRNLGDPVTSGLFQLRTYLQMINMDDLFLPLIAKNPGN
T TLALAVPQVLGGLSFLNPEKCFYRNLGDPVTSGLFQLKTYLQMIHMDDLFLPLIAKNPGN
970 980 990 1000 1010 1020
R CSAIDFVLNPSGLNVPGSQDLTSFLRQIVRRSITLTARNKLINTLFHASADLEDEMVCKW
E CTAIDFVLNPSGLNVPGSQDLTSFLRQIVRRTITLSAKNKLINTLFHASADFEDEMVCKW
S CSAIDFVLNPGGLNVPGSQDLTSFLRQIVRRSITLSARNKLINTLFHASADLEDELVCKW
B CSAIDFVLNPSGLNVPGSQDLTSFLRQIVRRTITLSAKNKLINTLFHSSADLEDEMVCKW
T CSAIDFVLNPSGLNVPGSQDLTSFLRQIVRRTITLSAKNKLINTLFHSSADLEDEMVCKW
1030 1040 1050 1060 1070 1080
R LLSSNPVMSRFAADIFSRTPSGKRLQILGYLEGTRTLLASKIINNNSETPVLDKLRKITL
E LLSSTPVMSRFAADIFSRTPSGKRLQILGYLEGTRTLLASKIINNNTETPVLDRLRKITL
S LLSSTPVMSRFAADIFSRTPSGKRLQILGYLEGTRTLLASKMISNNAETPILERLRKITL
B LLSSTPVMSRFAADIFSRTPSGKRLQILGYLEGTRTLLASKVINNNAETPILDRLRKITL
T LLSSTPVMSRFAADIFSRTPSGKRLQILGYLEGTRTLLASKIINHNTETPILDRLRKITL
1090 1100 1110 1120 1130 1140
R QRWNLWFSYLDHCDQLLADALQKISCTVDLAQILREYTWSHILEGRSLIGATLPCMVEQF
E QRWSLWFSYLDHCDNILAEALTQITCTVDLAQILREYSWAHILEGRPLIGATLPCMIEQF
S QRWNLWFSYLDHCDPALMEAIQPIKCTVDIAQILREYSWAHILDGRQLIGATLPCIPEQF
B QRWSLWFSYLDHCDQVLADALIKVSCTVDLAQILREYTWAHILEGRQLIGATLPCMLEQF
T QRWSLWFSYLDHCDQVLADALTQITCTVDLAQILREYTWAHILEGRQLIGATLPCILEQL
1150 1160 1170 1180 1190 1200
R KVKWLGQYEPCPECLNKKG--SNAYVSVAVKDQVVSAWPNTSRISWTIGSGVPYIGSRTE
E KVVWLKPYEQCPQCSNAKQPGGKPFVSVAVKKHIVSAWPNASRISWTIGDGIPYIGSRTE
S QTTWLKPYEQCVECSSTNN--SSPYVSVALKRNVVSAWPDASRLGWTIGDGIPYIGSRTE
B NVFWLKSYEQCPKCARSRNPKGEPFVSIAIKKQVVSAWPNQSRLNWTIGDGVPYIGSRTE
T NVIWLKPYEHCPKCAKSANPKGEPFVSIAIKKHVVSAWPDQSRLSWTIGDGIPYIGSRTE
1210 1220 1230 1240 1250 1260
R DKIGQPAIKPRCPSSALKEAIELASRLTWVTQGGSNSEQLIRPFLEARVNLSVSEVLQMT
E DKIGQPAIKPKCPSAALREAIELASRLTWVTQGSSNSDLLIKPFLEARVNLSVQEILQMT
S DKIGQPAIKPRCPSAALREAIELTSRLTWVTQGSANSDQLIRPFLEARVNLSVQEILQMT
B DKIGQPAIKPKCPSAALREAIELTSRLTWVTQGGANSDLLVKPFVEARVNLSVQEILQMT
T DKIGQPAIKPKCPSAALREAIELTSRLTWVTQGGANSDLLVKPFIEARVNLSVQEILQMT

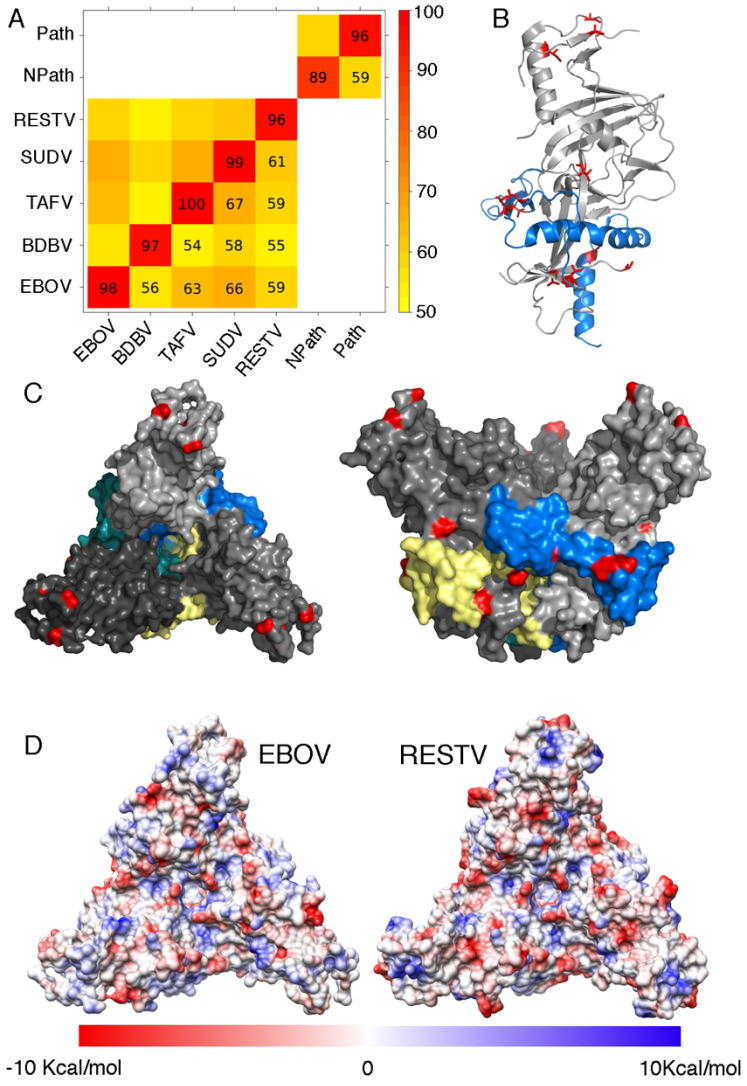

1270 1280 1290 1300 1310 1320
R PSHYSGNIVHRYNDQYSPHSFMANRMSNTATRLIVSTNTLGEFSGGGQAARDSNIIFQNV
E PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLIVSTNTLGEFSGGGQSARDSNIIFQNV
S PSHYSGNIVHRYNDQYSPHSFMANRMSNTATRLMVSTNTLGEFSGGGQAARDSNIIFQNV
B PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLVVSTNTLGEFSGGGQSARDSNIIFQNV
T PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLVVSTNTLGEFSGGGQSARDSNIIFQNV
1330 1340 1350 1360 1370 1380
R INLAVALYDIRFRNTNTSDIRHNRAHLHLTECCTKEVPAQYLTYTSALNLDLSRYRDNEL
E INYAVALFDIKFRNTEATDIQYNRAHLHLTKCCTREVPAQYLTYTSTLDLDLTRYRENEL
S INFAVALYDIRFRNTCTSSIQYHRAHIHLTNCCTREVPAQYLTYTTTLNLDLSKYRNNEL
B INFAVALFDLRFRNTETSSIQHNRAHLHLSQCCTREVPAQYLTYTSTLSLDLTRYRENEL
T INFAVALFDLRFRNVATSSIQHHRAHLHLSKCCTREVPAQYLVYTSTLPLDLTRYRDNEL
1390 1400 1410 1420 1430 1440
R IYDSNPLKGGLNCNLTIDSPLVKGPRLNMIEDDLLRFPHLSGWELAKTVVQSIISDNSNS
E IYDNNPLKGGLNCNISFDNPFFQGKQLNIIEDDLIRLPHLSGWELAKTIMQSIISDSNNS
S IYDSEPLRGGLNCNLSIDSPLMKGPRLNIIEDDLIRLPHLSGWELAKTVLQSIISDSSNS
B IYDNNPLKGGLNCNLSFDNPLFKGQRLNIIEEDLIRFPHLSGWELAKTIIQSIISDSNNS
T IYDDNPLRGGLNCNLSFDNPLFKGQRLNIIEEDLIRLPYLSGWELAKTVIQSIISDSNNS
1450 1460 1470 1480 1490 1500
R STDPISSGETRSFTTHFLTYPQIGLLYSFGAVLCFYLGNTILWTKKLDYEQFLYYLHNQL
E STDPISSGETRSFTTHFLTYPKIGLLYSFGAFVSYYLGNTILRTKKLTLDNFLYYLTTQI
S STDPISSGETRSFTTHFLTYPKIGLLYSFGALISFYLGNTILCTKKIGLTEFLYYLQNQI
B STDPISSGETRSFTTHFLTYPKVGLLYSFGAIVSYYLGNTIIRTKKLDLSHFMYYLTTQI
T STDPISSGETRSFTTHFLTYPKIGLLYSFGALISYYLGNTIIRTKKLTLNNFIYYLATQI
1510 1520 1530 1540 1550 1560
R HNLPHRALRVFKPTFKHASVMSRLMEIDSNFSIYIGGTSGDRGLSDAARLFLRTAIASFL
E HNLPHRSLRILKPTFKHASVMSRLMSIDPHFSIYIGGAAGDRGLSDAARLFLRTSISSFL
S HNLSHRSLRIFKPTFRHSSVMSRLMDIDPNFSIYIGGTAGDRGLSDAARLFLRIAISTFL
B HNLPHRSLRILKPTFKHVSVISRLMSIDPHFSIYIGGTAGDRGLSDATRLFLRVAISSFL
T HNLPHRSLRILKPTLKHASVISRLISIDSHFSIYIGGTAGDRGLSDAARLFLRTAITVFL
1570 1580 1590 1600 1610 1620
R QFLKSWIIDRQKTIPLWIVYPLEGQQPESINEFLHKILGLLKQGPKSIPKEVSIQNDGHL
E TFVKEWIINRGTIVPLWIVYPLEGQNPTPVNNFLHQIVELLVHDSSRHQAFK--TTINDH
S SFVEEWVIFRKANIPLWVIYPLEGQRPDPPGEFLNRVKSLIVGTEDDKNKGSIL--SRSG
B QFVKKWIVEYRTAIPLWVVYPLEGQNPDPINSFLHQIIALLQNESP--QNNIQFQEGRNN
T QFVRKWIVERKTAIPLWVIYPLEGQSPSPINSFLHHVIALLQHESS--HDHVCAAEAHSR
1630 1640 1650 1660 1670 1680
R DLAENNYVYNSKSTASNFFHASLAYWRSRKSRKTQDHNDFSRGDGTL----TEPVRKFSS
E VHPHDNLVYTCKSTASNFFHASLAYWRSRHRNSNRKDLTRNSSTGSSTNNSDGHIKRSQE
S EKCSSNLVYNCKSTASNFFHASLAYWRGRHRPKKTIGATNATTAPHI----ILPLGNSDR
B QQLSDNLVYMCKSTASNFFHASLAYWRSRHKGRPKNRSTEEQTVKPRPYNNFHSVKCASN
T VETFDNLVYMCKSTASNFFHASLAYWRSRSKNQDKREMTKILSLTQTEKKN--SFGYTAH


1690 1700 1710 1720 1730 1740
R-----NHQSDEKYYNVTCGKSPKPQERKDF--SQYRLSNNGQTMSNHRKKGKFHKWNPCK
EQT---------------TRDPHDGTERSLVLQMSHEIKRTTIPQ------ENTHQGPSFQ
SPPGLDLNRNNDTFIPTRIKQIVQGDSRNDRT-TTTRFPPKSRS------TPTSATEPPTK
BPPSIP--KSKSGT----QGSSA-FFEKLEYD-KEIELPTASTP---AEKPKTYTKALSSR
TPESTAVLGSLQTS----LAPPP-SADEATYD-RKNKVLKASRP---GKYSQNTTKAPPNQ
1750 1760 1770 1780 1790 1800
RMLMESQRGTVL----------TEGDYFQNNTPPTDDVSSPHRLILPFFKLGNHNHAHD
ESFLSDSACGTANPKLNFDRSRHNVKSQDHNSASKREGHQIISHRLVLPFFTLSQGTRQLT
SMYEGSTTHQGK----------LTDTHLDEDHNAKEFPSNPHRLVVPFFKLTKDGEYSI
BIYHGKTPSNAAKDDSTT-----SKGCDS-----KEENAVQASHRIVLPFFTLSQNGYRTP
TT-----SCRDVSPNITG-----TDGCPSANEGSNSNNNNLVSHRIVLPFFTLSHNYNERP
1810 1820 1830 1840 1850 1860
RQDAQELMNQNIKQYLHQLRSMLDTTIYCRFTGIVSSMHYKLDEVLLEYNSFDSAITLAEG
ESSNESQTQDEISKYLRQLRSVIDTTVYCRFTGIVSSMHYKLDEVLWEIENFKSAVTLAEG
SEPSPEESRSNIKGLLQHLRTMVDTTIYCRFTGIVSSMHYKLDEVLWEYNKFESAVTLAEG
BSVKKSEYVTEITKLIRQLKAIPDTTVYCRFTGVVSSMHYKLDEVLWEFDSFKTAVTLAEG
TSIRKSEGTTEIVRLTRQLRAIPDTTIYCRFTGIVSSMHYKLDEVLWEFDNFKSAITLAEG
1870 1880 1890 1900 1910 1920
REGSGALLLLQKYSTRLLFLNTLATEHSIESEVVSGFSTPRMLLPIMQKVHEGQVTVILNN
EEGAGALLLIQKYQVKTLFFNTLATESSIESEIVSGMTTPRMLLPVMSKFHNDQIEIILNN
SEGSGALLLIQKYGVKKLFLNTLATEHSIESEVISGYTTPRMLLSIMPKTHRGELEVILNN
BEGSGALLLLQKYKVRTIFFNTLATEHSIEAEIVSGTTTPRMLLPVMAKLHDDQINVILNN
TEGSGALLLLQKYKVETLFFNTLATEHSIEAEIISGITTPRMLLPIMSRFHGGQIKVTLNN
1930 1940 1950 1960 1970 1980
RSASQITDITSSMWLS-NQKYNLPCQVEIIMMDAETTENLNRSQLYRAVYNLILDHIDPQY
ESASQITDITNPTWFK-DQRARLPRQVEVITMDAETTENINRSKLYEAVHKLILHHVDPSV
SSASQITDITHRDWFS-NQKNRIPNDADIITMDAETTENLDRSRLYEAVYTIICNHINPKT
BSASQVTDITNPAWFT-DQKSRIPTQVEIMTMDAETTENINRSKLYEAIQQLIVSHIDTRV
TSASQITDITNPSWLA-DQKSRIPKQVEIITMDAETTENINRSKLYEAVQQLIVSHIDPNA
1990 2000 2010 2020 2030 2040
RLKVVVLKVFLSDIEGILWINDYLAPLFGAGYLIKPITSSARSSEWYLCLSNLISTNRRSA
ELKAVVLKVFLSDTEGMLWLNDNLAPFFATGYLIKPITSSARSSEWYLCLTNFLSTTRKMP
SLKVVILKVFLSDLDGMCWINNYLAPMFGSGYLIKPITSSAKSSEWYLCLSNLLSTLRTTQ
BLKIVIIKVFLSDIDGLLWLNDHLAPLFGSGYLIKPITSSPKSSEWYLCLSNFLSASRRRP
TLKVVVLKVFLSDIDGILWLNDNLTPLFGLGYLIKPITSSPKSSEWYLCLSNLLSTSRRLP
2050 2060 2070 2080 2090 2100
RHQTHKACLGVIRDALQAQVQRGVYWLSHIAQYATKNLHCEYIGLGFPSLEKVLYHRYNLV
EHQNHLSCKQVILTALQLQIQRSPYWLSHLTQYADCDLHLSYIRLGFPSLEKVLYHRYNLV
SHQTQANCLHVVQCALQQQVQRGSYWLSHLTKYTTSRLHNSYIAFGFPSLEKVLYHRYNLV
BHQGHATCMQVIQTALRLQVQRSSYWLSHLVQYADINLHLSYVNLGFPSLEKVLYHRYNLV
THQSHTTCMHVIQTALQLQIQRSSYWLSHLVQYANHNLHLDYINLGFPSLERVLYHRYNLV

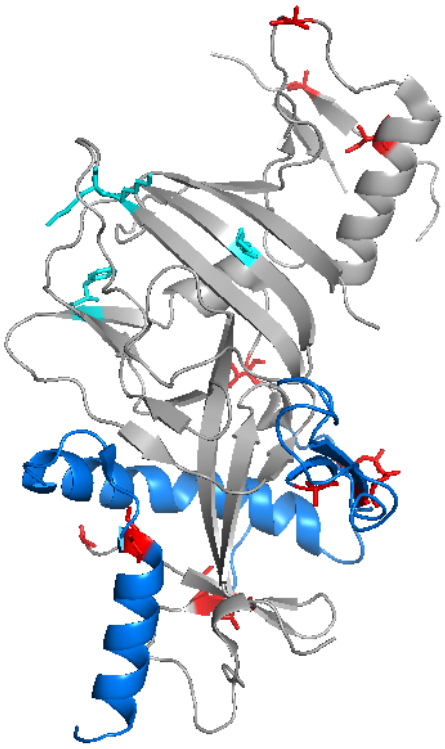

2110 2120 2130 2140 2150 2160
R DTGLGPLSSVIRHLTNLQAEIRDLVLDYNLMRESRTQTYHFIKTAKGRITKLVNDFLKFS
E DSKRGPLVSVTQHLAHLRAEIRELTNDYNQQRQSRTQTYHFIRTAKGRITKLVNDYLKFF
S DSRNGPLVSITRHLALLQTEIRELVTDYNQLRQSRTQTYHFIKTSKGRITKLVNDYLRFE
B DSRKGPLVSILYHLTHLQAEIRELVCDYNQQRQSRTQTYHFIKTTKGRITKLVNDYLKFY
T DSQKGPLTSIVQHLAHLQTEIRELVNDYNQQRQSRTQTYHFIKTIKGRITKLVNDYLKFF
2170 2180 2190 2200 2210 2220
R LIVQALKNNSSWYTELKKLPEVINVCNRFYHTHNCECQEKFFVQTLYLQRLRDAEIKLIE
E LIVQALKHNGTWQAEFKKLPELISVCNRFYHIRDCNCEERFLVQTLYLHRMQDSEVKLIE
S LVIRALKNNSTWHHELYLLPELIGVCHRFNHTRNCTCSERFLVQTLYLHRMSDAEIKLMD
B LVVQALKHNCLWQEELRTLPDLINVCNRFYHIRDCSCEDRFLIQTLYLTRMQDSEVKLME
T LIIQALKHNCTWQEELRALPDLISVCTRFYHTRNCSCENRFLVQTLYLSRMQDSEIKLID
2230
R RLTGLMRFYPEGLIYSNHT
E RLTGLLSLFPDGLYRFD--
S RLTSLVNMFPEGFRSSSV-
B RLTGFLGLYPNGINT----
T RLTGLLSLCPNGFFR----



Supplementary Figure 3. Solvent Accessible surface area for Ebolavirus SDPs. Histograms showing the Solvent Accessible surface area in square ångstroms of SDPs. Values are calculated for the Ebola virus structure and residues.



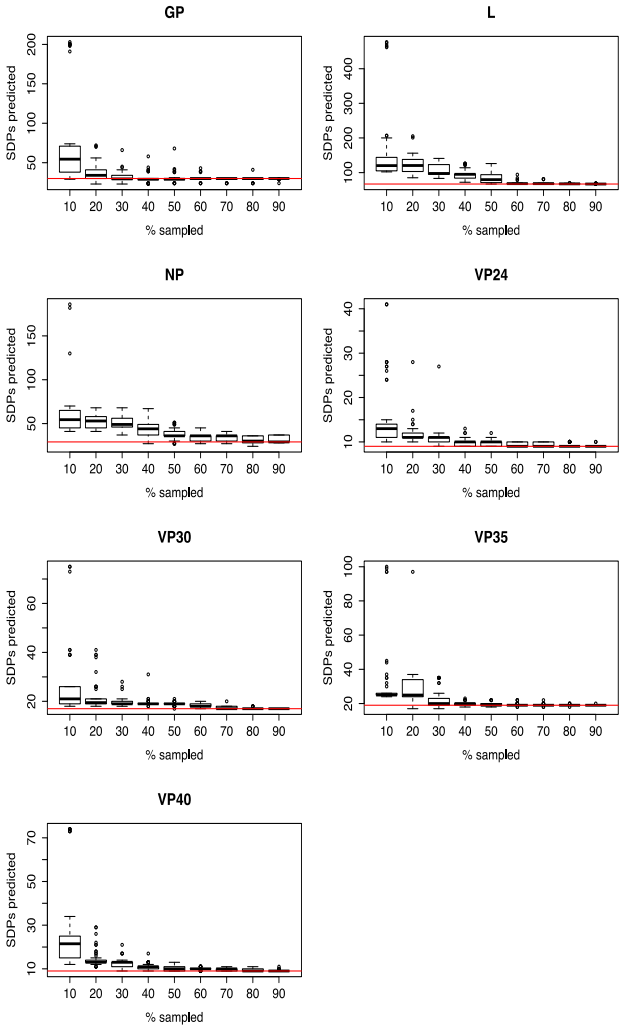
Supplementary Figure 4. GP SDPs. A) Heatmap of intra- and inter-species GP sequence identity (EBOV, Ebola virus; BDBV, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus; RESTV, Reston virus). B) Monomeric representation of GP with GP1 (grey) and GP2 (blue). D) EBOV GP trimer (PDB code: 3CSY) with SDPs colored red. The three GP1 chains are colored grey. The three GP2 chains are colored blue, green and yellow. C) Electrostatics surfaces for the EBOV structure (3CSY) and a model of a RESTV GP trimer based on 3CSY.



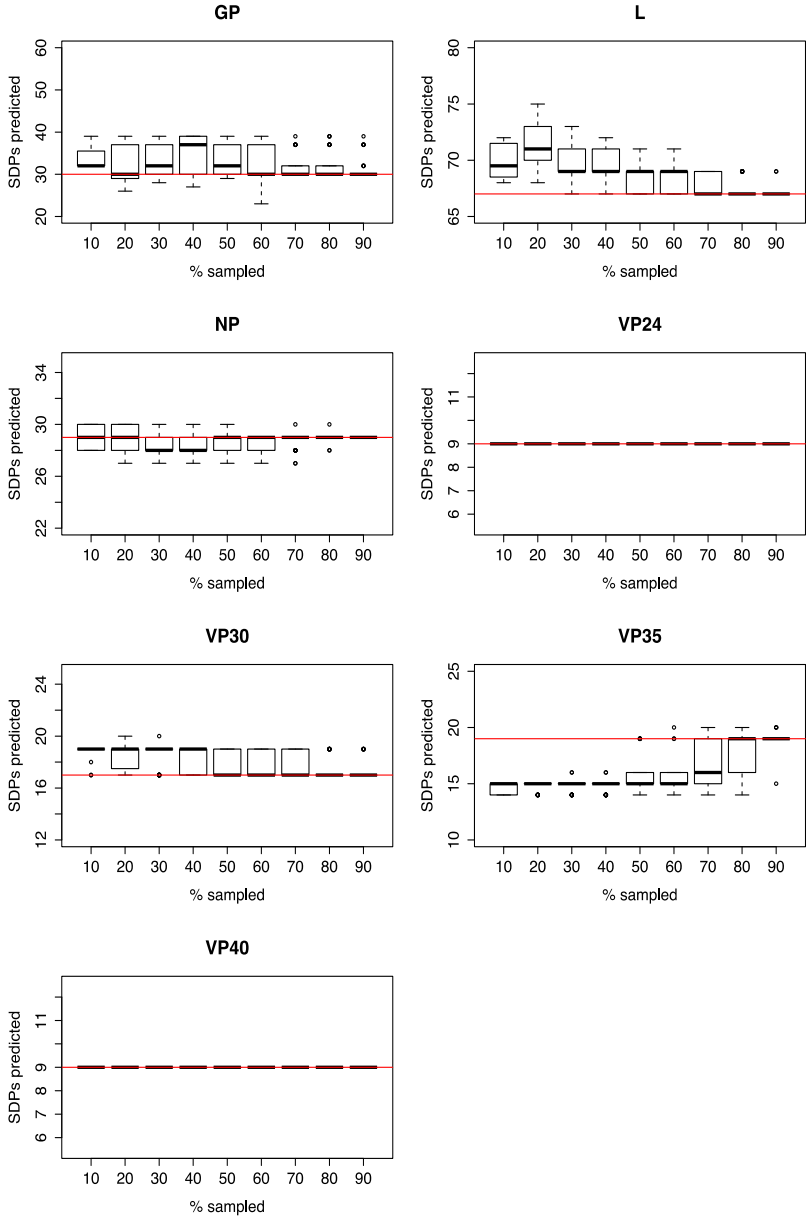
Supplementary Figure 5. GP SDPs are located outside the putative NPC1 binding site. GP SDPS are shown in red. The putative NPC1 binding site is shown in cyan.

Supplementary Figure 6. SDP prediction with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and Reston ('non human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the distributions of the number of SDPs predicted in the simulations where A) only human pathogenic sequences were sampled, B) only Reston sequences were sampled and C) both sets were sampled. Sampling was performed for samples consisting of between 10%-90% of sequences (x axis). Red lines indicate the number of SDPs predicted in the full dataset without sampling. Note the scale of the Y-axis varies between each plot.

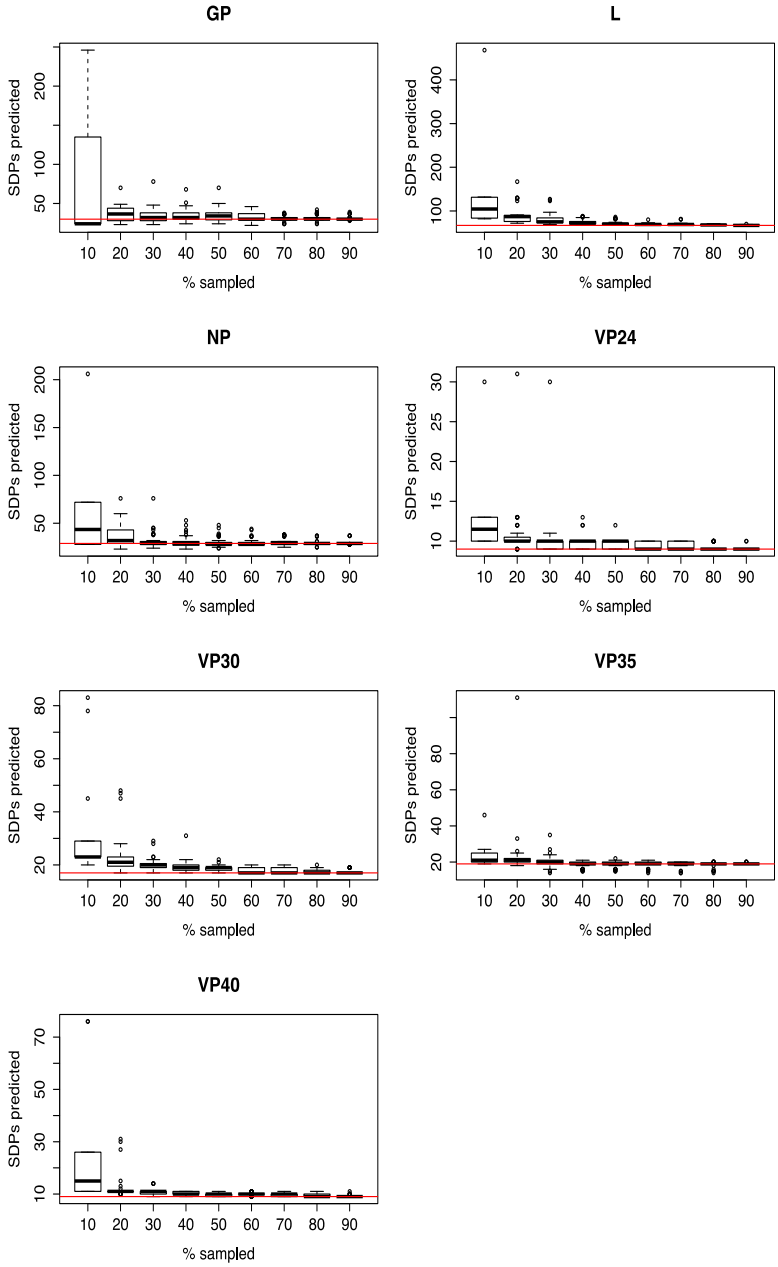
A. **Human pathogenic sequence sampled.**



B. **Reston Sequences Sampled**



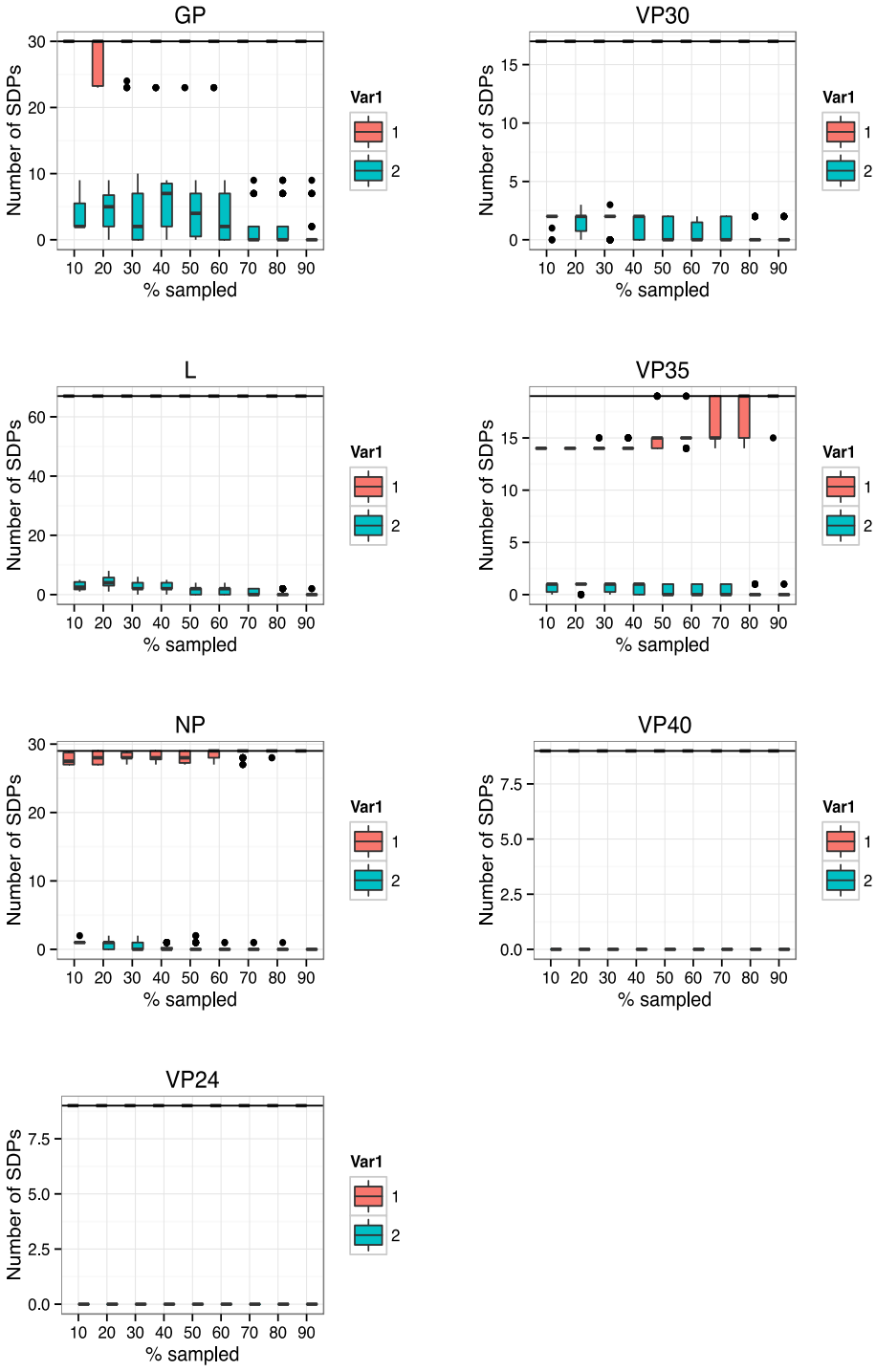
C. **Both groups sampled**



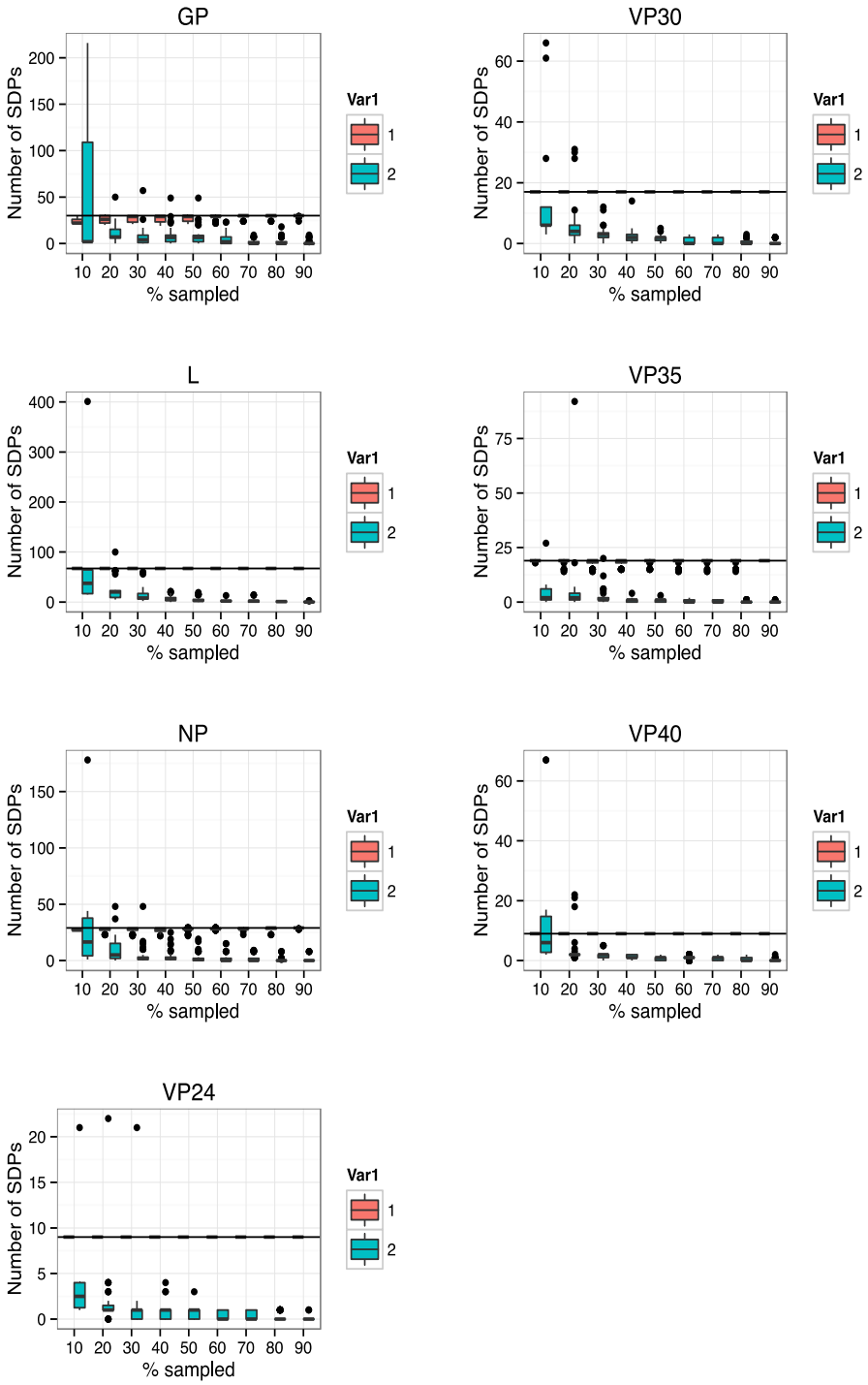
Supplementary Figure 7. Change in SDP prediction with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and and Reston ('non human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the number of SDPs predicted in each sampling that are also in the full dataset (red) and new SDPs that are predicted only in subsamples (blue). The black horizontal line indicates the number of SDPs predicted using the full dataset. Subsampling performed for A) only human pathogenic sequences were sampled, B) only Reston sequences were sampled and c) both sets were sampled.

A. **Human pathogenic sequence sampled.**

B. Reston Sequences Sampled

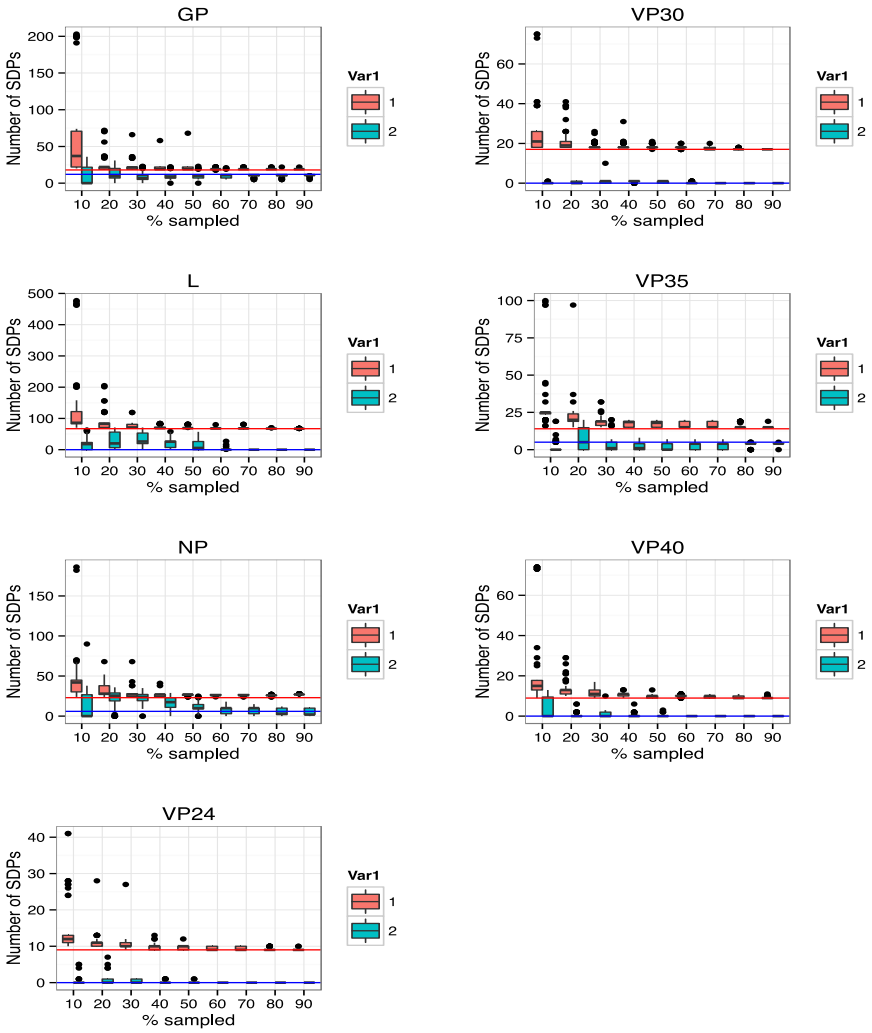


C. Both groups sampled

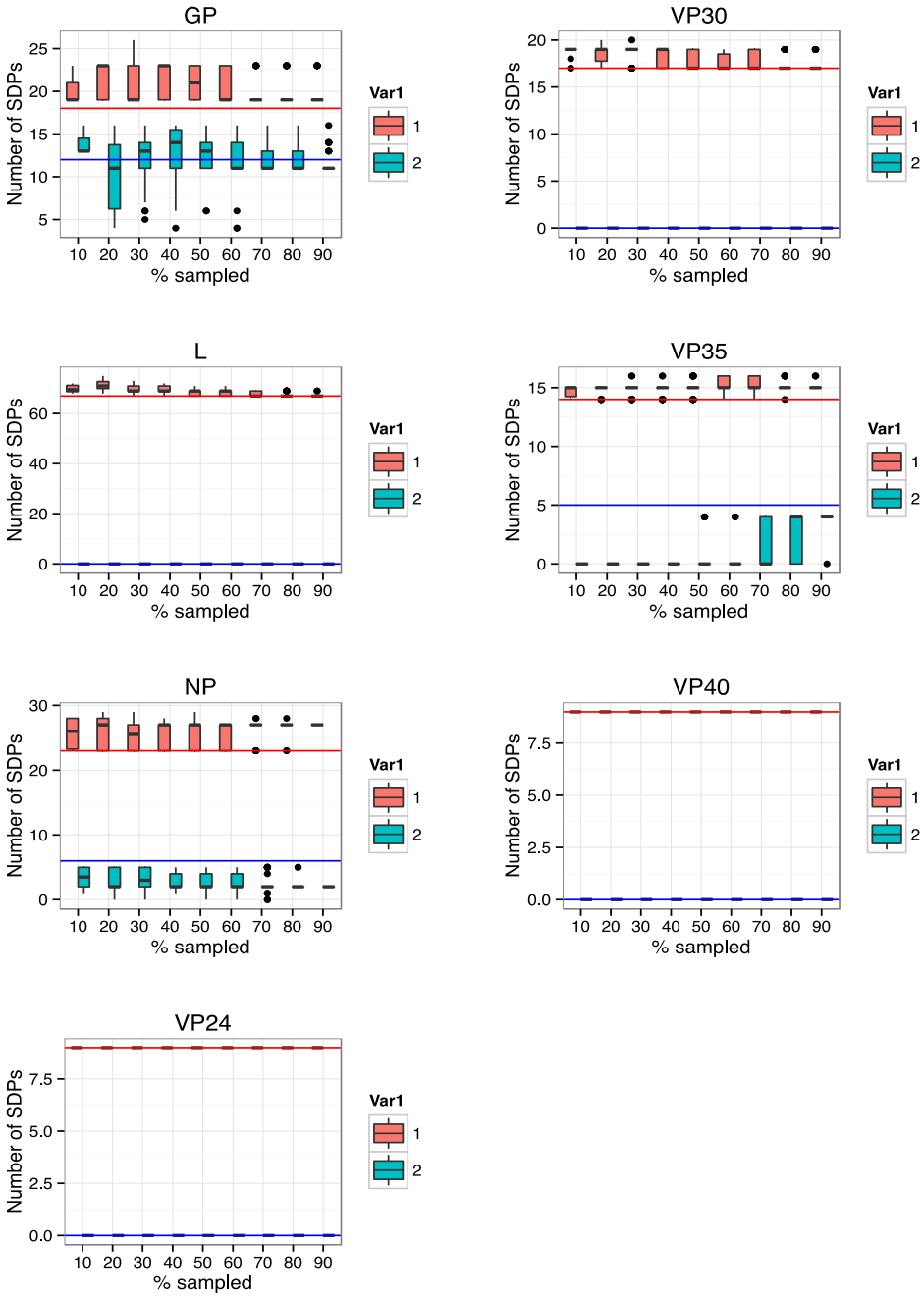


Supplementary Figure 8. Analysis of completely conserved SDP with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and and Reston ('non human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the number of SDPs predicted in each sampling that are are completely conserved (red) and not completely conserved (blue). The red horizontal line indicates the number of completely conserved SDPs present in the full dataset and the blue line represents the equivalent for SDPs that are not completely conserved. Subsampling performed for A) only human pathogenic sequences were sampled, B) only Reston sequences were sampled and c) both sets were sampled.

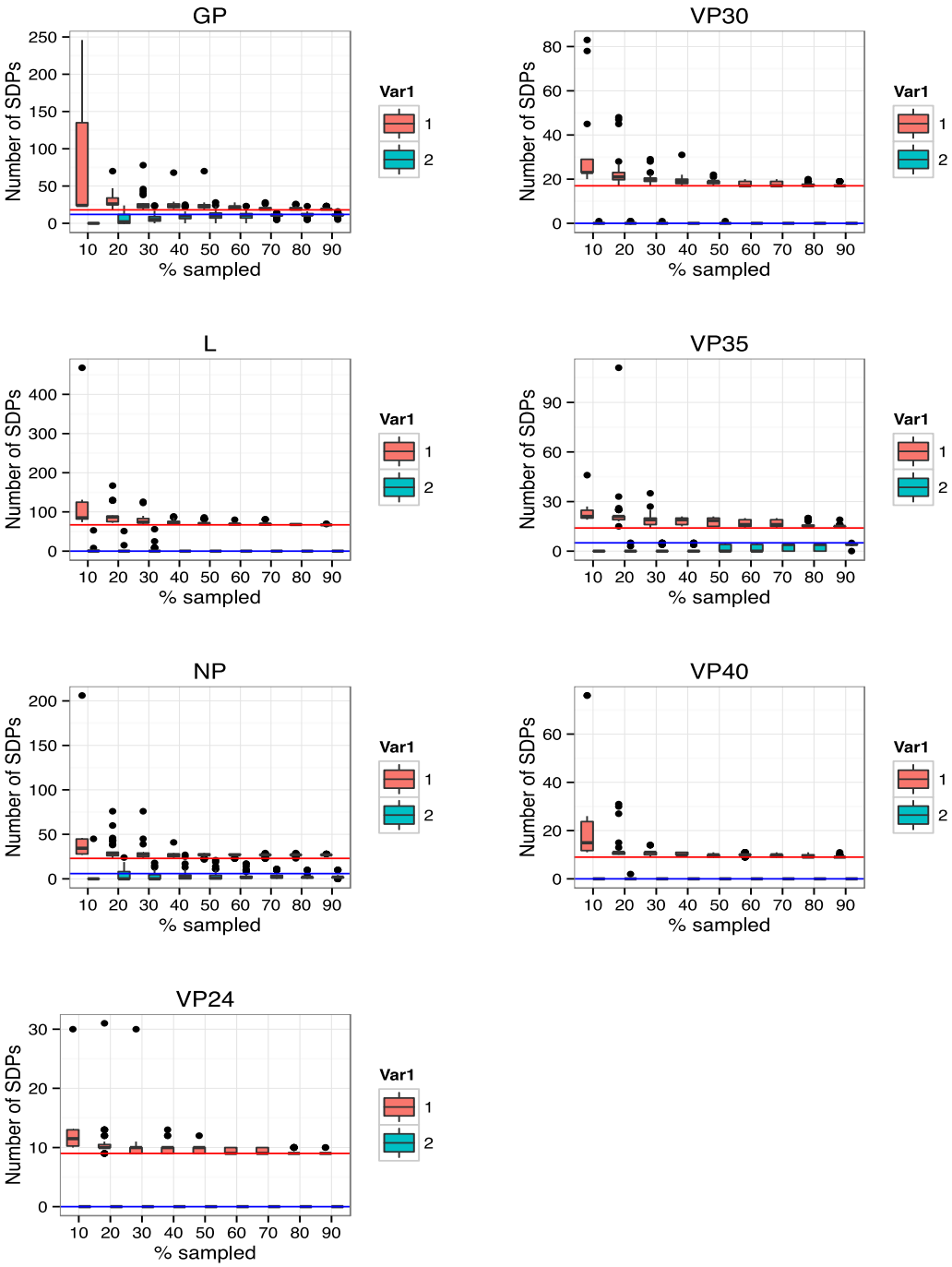
A. Human pathogenic sequence sampled.



B. Reston Sequences Sampled



C. Both groups sampled



Supplementary Tables

	completely conserved positions	Number of Positions with variation	% of positions with variation
All species	2597	4555	64%
Ebola virus	4287	2865	40%
Sudan virus	4363	2789	38%
Bundibugyo virus	4426	2726	38%
Tai forest virus	4480	2672	37%
Reston virus	4466	2686	38%

Supplementary Table 1. Variation within the Ebolavirus genomes. The number of positions in the Ebolavirus protein multiple sequence alignments that are completely conserved and those that have variation are shown.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 score	SASA (Å²)	mCSM (Δ Δ G, Kcal/mol)	S3det Rank
17	M17	L17	L17	L17	L17	2	70	-0.444 (destabilising)	1
22	I22	V22	V22	V22	V22	3	0	-0.916 (destabilising)	1
31	I31	V31	V31	V31	V31	3	17	-0.193 (destabilising)	1
131	S131	T131	T131	T131	T131	1	36	-1.394 (destabilising)	1
132	T132	N132	N132	N132	N132	1	9	-1.121 (destabilising)	1
136	L136	M136	M136	M136	M136	2	2	-1.7 (destabilising)	1
139	R139	Q139	Q139	Q139	Q139	1	132	0.05 (stabilising)	1
226	A226	T226	T226	T226	T226	0	2	-0.935 (destabilising)	1
248	L248	S248	S248	S248	S248	-2	-		1

Supplementary Table 2. VP24 SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 4M0Q. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 score	SASA (Å²)	mCSM (Δ Δ G, Kcal/mol)	S3det rank
53	N53	T52	T52	T52	T52	0	-		1
54	L54	V53	V53	V53	V53	1	-		1
64	I64	T63	T63	T63	T63	-1	-		1
94	D94	E93	E93	E93	E93	2	-		1
97	N97	T96	T96	T96	T96	0	-		1
99	H99	R98	R98	R98	R98	0	-		1
108	R108	K107	K107	K107	K107	2	-		1
112	I112	S111	S111	S111	S111	-2	-		1
117	S117	K116	K116	K116	K116	0	-		1
121	S121	A120	A120	A120	A120	1	-		1
151	I151	T150	T150	T150	T150	-1	7	0.455 (stabilising)	1
158	R158	Q157	Q157	Q157	Q157	1	70	-0.493 (destabilising)	1
160	L160	I159	I159	I159	I159	2	6	-0.859 (destabilising)	1
197	H197	R196	R196	R196	R196	0	83	-1.291 (destabilising)	1
206	D206	E205	E205	E205	E205	-2	148	-0.373 (destabilising)	1
263	A263	R262	R262	R262	R262	-1	106	-0.969 (destabilising)	1
269	Q269	S268	S268	S268	S268	0	-		1

Supplementary Table 3. VP30 SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 2I8B. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM62 SCORE	SASA (Å²)	mCSM (ΔΔG, Kcal/mol)	S3det rank
27	T15	S26	S26	S26	S26	1	-		1
49	D37	E48	E48	E48	E48	2	-		1
77	E65	D76	D76	D76	D76	2	-		2
86	K74	E85	E85	E85	D86	1	-		3
93	M81	S92	S92	S92	S92	-1	-		1
98	T86	V97	V97	V97	I98	0	-		3
102	N90	T101	T101	T101	A102	0	-		3
107	A95	S106	S106	S106	S106	1	-		1
122	I110	V121	V121	V121	M122	3	-		3
155	S143	A154	A154	A154	A154	1	-		1
160	V148	T159	T159	T159	T159	0	-		1
161	D149	E160	E160	E160	E160	2	-		1
168	K156	G167	G167	G167	G167	-2	-		1
175	A163	S174	S174	S174	S174	1	-		1
182	L170	I181	I181	I181	I181	2	-		2
270	D258	E269	E269	E269	E269	2	144	-0.039 (destabilising)	1
291	V279	A290	A290	A290	A290	0	23	-0.756 (destabilising)	1
315	A303	V314	V314	V314	V314	0	49	-1.47 (destabilising)	1
330	K318	Q329	Q329	Q329	Q329	1	32	-0.513 (destabilising)	1

Supplementary Table 4. VP35 SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 4IBB. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det rank column shows the ranking of the SDPs by s3det. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å²)	mCSM (ΔΔG, Kcal/mol)	S3det rank
46	V46	T46	T46	T46	T46	0	83	-0.31 (destabilising)	1
85	T85	P85	P85	P85	P85	-1	142	-0.626 (destabilising)	1
122	V122	I122	I122	I122	I122	3	-		1
201	N201	G201	G201	G201	G201	0	53	-0.482 (destabilising)	1
209	L209	F209	F209	F209	F209	0	15	-1.219 (destabilising)	1
245	P245	Q245	Q245	Q245	Q245	-1	160	0.059 (stabilising)	1
269	Q269	H269	H269	H269	H269	0	-		1
293	V293	I293	I293	I293	I293	3	14	-1.411 (destabilising)	1
325	D325	E325	E325	E325	E325	2	-		1

Supplementary Table 5. VP40 SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 1ES6. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å²)	mCSM (Δ Δ G, Kcal/mol)	S3det rank
4	G4	R4	R4	R4	R4	-2			1
16	D16	E16	E16	E16	G16	2			2
30	T30	S30	S30	S30	S30	1			1
39	K39	R39	R39	R39	R39	2	188	-0.161 (destabilising)	1
42	S42	P42/ Q42	P42	P42	Q42	-1	103	-2.173 (destabilising)	3
56	V56	I56	I56	I56	I56	3	0	-0.8 (destabilising)	1
64	I64	V64	V64	V64	V64	3	7	-0.135 (destabilising)	1
105	K105	R105	R105	R105	R105	2	112	-0.63 (destabilising)	1
137	L137	M137	M137	M137	M137	2	37	-0.649 (destabilising)	1
212	Y212	F212	F212	F212	F212	3	0	-0.692 (destabilising)	1
274	R274	K274	K274	K274	K274	2	92	-0.548 (destabilising)	1
279	A279	S279	S279	S279	S279	1	60	-0.822 (destabilising)	1
374	R374	K374	K374	K374	K374	2	103	-0.836 (destabilising)	1
416	N416	K416	K416	K416	K416	0			1
421	Q421	Y421	Y421	Y421	Y421	-1			1
426	E426	D426	D426	D426	D426	2			1
435	N435	D435	D435	D435	D435	1			1
443	E443	D443	D443	D443	D443	2			1
453	I453	T453	T453	T453	T453	-1			1
492	E492	D492	D492	D492	D492	2			1
497	A497	P497	P497	P497	P497	-1			2
535	(-)	P526	P526	P526	P526				1
572	S563	T563	T563	T563	T563	1			1
574	V565	I565	I565	I565	I565	3			1
611	T602	P602	P602	P602	N602	-1			4
651	Q641	N641	N641	N641	K641	0			2
715	R705	A705	A705	A705	A705	-1	24	-1.037 (destabilising)	1

726	N716	D716	D716	D716	D716	1	123	0.141 (stabilising)	1
727	N717	G717	G717	G717	G717	0	75	-0.461 (destabilising)	2

Supplementary Table 6. NP SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 4QB0 for the C terminal and 4YPI for the N terminal regions. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det rank column shows the ranking of the SDPs by s3det. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 Score	SASA (Å²)	mCSM (ΔΔG, Kcal/mol)	S3det rank
2	G2	M1	M1	M1	M1	-3			1
3	S3	G2	V2	E2/G2	G2	0			8
32	I32	F31	F31	F31	F31	0			1
38	I38	V37	V37	V37	V37	3	0	-0.828 (destabilising)	1
46	A46	V45	V45	V45	V45	0	30	-1.276 (destabilising)	1
76	I76	V75	V75	V75	V75	3	44	-0.295 (destabilising)	1
197	A197	S196	S196	S196	S196	1			1
208	D208	E207	T207	E207	T207	2			9
211	T211	S210	S210	S210	S210	1			1
261	L261	I260	I260	I260	I260	2	25	-0.95 (destabilising)	1
270	S270	T269	T269	T269	T269	1	99	-0.432 (destabilising)	1
308	H308	S308/ L307	S308	S308	S308	-1			2
326	G326	R325	V325	R325	V325	-2			9
355	L355	H354	R354	H354	Q354	-3			9
404	P401	Q403	N401	Q397	S401	-1			9
419	E412	S418	A409	S412	T409	0			9
461	P449	T448	S442	T448	T448	-1			7
497	Y517 / H517	H516	H516	H516	H516	2			6
519	K499	R498	R498	R498	R498	2			1
521	K501	R500	R500	R500	R500	2			1
535	D515	N514	N514	N514	N514	1	59	-1.142 (destabilising)	1
542	V522	Q521	Q521	Q521	L521	2	19	0.037 (stabilising)	6
568	V548	L547	I547	L547	I547	1	74	-1.258 (destabilising)	9
605	L585	I584	I584	I584	I584	2			1
628	S608	D607	D607	D607	D607	0			1
643	E623	K622	K622	K622	K622	1			1
659	H639	Q638	Q638	Q638	Q638	0			1
663	L643	D642	D642	D642	S642	-4			6
665	L645	W644	W644	W644	W644	-2			1
680	I660	T569	T569	T569	T569	-1			1

Supplementary Table 7. GP SDPs. The position in the multiple sequence alignment, the amino acid

position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the protein structure with PDB code 3CSY. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det rank column shows the ranking of the SDPs by s3det. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å²)	S3det rank
47	G2	M1	M1	M1	M1	-3		1
77	I32	F31	F31	F31	F31	0		1
83	I38	V37	V37	V37	V37	3	21	1
91	A46	V45	V45	V45	V45	0	84	1
121	I76	V75	V75	V75	V75	3	61	1
242	A197	S196	S196	S196	S196	1		1
256	T211	S210	S210	S210	S210	1		1
306	L261	I260	I260	I260	I260	2	20	1
315	S270	T269	T269	T269	T269	1	48	1

Supplementary Table 8. sGP SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the Phyre2 structural model that used template structure 3s88I. RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det rank column shows the ranking of the SDPs by s3det. The s3det column shows the ranking of the SDPs by s3det.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM62 SCORE	SASA (Å²)	mCSM (ΔΔG, Kcal/mol)	S3det rank
67	T66	V66	V66	V66	V66	0			1
110	H109	Q109	Q109	Q109	Q109	0			1
137	L136	I136	I136	I136	I136	2			1
147	V146	L146	L146	L146	L146	1			1
222	S221	A221	A221	A221	A221	1			1
224	L223	Q223	Q223	Q223	Q223	-2			1
228	Q227	H227	H227	H227	H227	0			1
277	I276	L276	L276	L276	L276	2	42	-1.049 (destabilising)	1
284	V283	L283	L283	L283	L283	1			1
313	F312	Y312	Y312	Y312	Y312	3			1
327	S326	A326	A326	A326	A326	1			1
331	D330	T330	T330	T330	T330	-1			1
351	D350	E350	E350	E350	E350	2			1
362	S361	T361	T361	T361	T361	1			1
366	F365	L365	L365	L365	L365	0			1
380	I379	V379	V379	V379	V379	3			1
448	H447	Q447	Q447	Q447	Q447	0			1
451	S450	P450	P450	P450	P450	-1			1
466	N465	D465	D465	D465	D465	1			1
690	S689	E689	E689	E689	E689	0			1
848	A847	S847	S847	S847	S847	1			1
869	A868	S868	S868	S868	S868	1			1
897	Y896	F896	F896	F896	F896	3			1
926	F925	L925	L925	L925	L925	0			1
955	S954	A954	A954	A954	A954	1			1
996	T995	S995	S995	S995	S995	1			1
1025	N1024	T1024	T1024	T1024	T1024	0			1
1074	K1073	R1073	R1073	R1073	R1073	2			1
1120	S1119	A1119	A1119	A1119	A1119	1			1
1164	A1161	F1163	F1163	F1163	F1163	-2			1
1190	S1187	D1189	D1189	D1189	D1189	0			1
1215	S1212	A1214	A1214	A1214	A1214	1			1
1218	K1215	R1217	R1217	R1217	R1217	2			1
1238	E1235	D1237	D1237	D1237	D1237	2			1
1256	V1253	I1255	I1255	I1255	I1255	3			1
1355	K1532	R1534	R1534	R1534	R1534	2			1
1367	A1354	T1366	T1366	T1366	T1366	0			1
1396	T1393	S1395	S1395	S1395	S1395	1			1
1409	M1406	I1408	I1408	I1408	I1408	1			1
1415	L1412	I1414	I1414	I1414	I1414	2			1
1437	N1434	S1436	S1436	S1436	S1436	1			1

1462	Q1459	K1461	K1461	K1461	K1461	1			1
1474	C1471	S1473	S1473	S1473	S1473	-1			1
1489	Y1486	L1488	L1488	L1488	L1488	-1			1
1500	L1497	I1499	I1499	I1499	I1499	2			1
1507	A1504	S1506	S1506	S1506	S1506	1			1
1510	V1507	I1509	I1509	I1509	I1509	3			1
1539	S1536	A1535	A1535	A1535	A1535	1			1
1627	Y1624	L1624	L1624	L1624	L1624	-1			1
1631	S1628	C1628	C1628	C1628	C1628	-1			1
1786	I1760	V1762	V1762	V1762	V1762	3			1
1874	T1848	V1850	V1850	V1850	V1850	0			1
1897	S1871	T1873	T1873	T1873	T1873	1			1
1941	N1914	R1916	R1916	R1916	R1916	1			1
1966	R1939	E1941	E1941	E1941	E1941	0			1
2033	I2006	L2008	L2008	L2008	L2008	2			1
2069	I2042	L2044	L2044	L2044	L2044	2			1
2102	T2075	S2077	S2077	S2077	S2077	1			1
2123	D2096	E2098	E2098	E2098	E2098	2			1
2130	L2130	Q2105	Q2105	Q2105	Q2105	-2			1
2133	E2106	Q2108	Q2108	Q2108	Q2108	2			1
2156	F2129	Y2131	Y2131	Y2131	Y2131	3			1
2182	V2155	L2157	L2157	L2157	L2157	1			1
2193	N2171	R2168	R2168	R2168	R2168	0			1
2200	K2173	R2175	R2175	R2175	R2175	2			1
2202	F2175	L2177	L2177	L2177	L2177	0			1
2211	L2184	M2186	M2186	M2186	M2186	2			1

Supplementary Table 9. L SDPs. The position in the multiple sequence alignment, the amino acid position, and amino acid present in each of the species is shown. The BLOSUM62 score represents how frequently such amino acid changes are observed in nature. SASA is the solvent accessible surface area, which is only available for SDPs that could be mapped to protein structure. SASA was calculated using the Phyre2 structural model which used template 4n48A ("cap-specific mrna ("cap-specific mrna (nucleoside-2'-o-)-methyltransferase 1 protein in2 complex with capped rna fragment"). RESTV, Reston virus; EBOV, Ebola virus; B, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus. The s3det rank column shows the ranking of the SDPs by s3det. The s3det column shows the ranking of the SDPs by s3det.

Protein	EBOV Res	RESTV Res	Mutation position	Mutation	Effect
GP	Q638	H	638	Q → V	No effect on release of soluble GP1,2delta.
GP	R498	K	498-501	RTRR → ATAA	No effect on cleavage between GP1 and GP2.
GP	D642	L	642	D → V	No effect on release of soluble GP1,2delta.
VP24	M136	L	134/136	F-A/M-A	Near complete loss of KPNA5 binding *
VP24	Q139	R	137-139	RTQ → AAA	Near complete loss of KPNA5 binding *

Supplementary Table 10. SDPs that coincide with known mutagenesis data. Functional data extracted from UniProt unless stated. Res, residue; EBOV, Ebola virus; RESTV, Reston virus

*Data from Bornholdt et al.,[35]

PROTEIN	SPECIES	OLIGOMERIC STATE	PDB/TEMPLATE	REGION IN SEQUENCE
GP	EBOV	Trimer of Heterodimers	3CSY (structure)	31-310 502-599
sGP	EBOV	Dimer	3s88I (model)	32-287
sGP	RESTV	Dimer	3s88I (model)	33-288
L	EBOV	Monomer	4n48A (model)	223-328
NP (C-terminal)	EBOV	Monomer	4QB0 (structure)	645-739
NP (N-terminal)	EBOV	Monomer	4YPI (structure)	39-384
VP24	EBOV	Heterodimer	4M0Q (structure)	10-231
VP24	EBOV	Heterodimer	4U2X (structure)	16-231
VP24	RESTV	Dimer	4D9O (structure)	10-231
VP30	EBOV	Dimer	2I8B (structure)	140-266
VP30	RESTV	Dimer	3V70 (structure)	142-272
VP35	EBOV	Heterodimer	4IBB (structure)	218-340
VP35	EBOV	Dimer of heterodimers	3L25 (structure)	209-340
VP35	RESTV	Dimer of heterodimers	3KS8 (structure)	208-329
VP40	EBOV	Monomer	1ES6 (structure)	44-321
VP40	EBOV	Dimer	4LDB (structure)	44-319
VP40	EBOV	Hexamer	4LDD (structure)	45-188
VP40	EBOV	Octamer	4LDM (structure)	69-188
VP40	RESTV	Monomer	1es6A (model)	44-321

Supplementary Table 11. Protein structures available for Ebolavirus Proteins. EBOV, Ebola virus; RESTV, Reston virus

Reston virus residue	Pathogenic consensus	Comments	Functional effect
I32	F31	Note- Ebola virus GP structure has R31 rather than F31. Surface residue close to interface with GP2 in the trimer. Unclear what functional effect may be if any.	Unclear
I38	V37	Surface residue, appears to be a conservative change of amino acid that could be well tolerated	Unlikely
A46	V45	Also a surface residue. Conservative change of hydrophobic amino acid that could be well accommodated.	Unlikely
I76	V75	Surface residue, conservative change of amino acid . Change should be well accommodated	Unlikely
L261	I260	One of three SDPs located in the glycan cap region of GP1. The glycan cap binds the host cell receptor(s) but is highly glycosylated so it is not clear if the amino acids directly contact the host cell. Surface residue in a cavity. It is part packed quite tightly with residue F234, V236, T240 but should be possible to accommodate change to Leu in Reston virus. Could there be a role with the three SDPs combined in this region.	possible*
S270	T269	Located at the top of the structure, is a surface residue (with side chain pointing to the solvent) representing a conservative amino acid change. Again could it have a role in conjunction with the 2 other SDPs in this region?	possible*
H308	S308/ L307	Also located in the glycan cap and also a surface residue. Present in loop so unlikely to alter structure but could have a functional role, and alters charge on the protein surface.	possible*
D515	N514	Surface residue, results in loss of negative charge in Reston virus GP. Located at the end of a beta sheet. Seems unlikely to have a structural effect. Possible combined effect with adjacent L547V?	Unlikely
V522	Q521	Close to trimer interface (GP2-GP2) but directly within the interface. Not clear what effect this change would have on protein structure	Unclear
V548	L547	Surface residue at end of a beta sheet. Appears to be minor change in amino acid. Possible combined effect with adjacent N514D?	Unlikely
L585	I584	Largely buried amino acid. At the interface with GP1 (in the same GP monomer). EBOV I584 interacts with F572, not clear if this interaction would change in with Leu in Reston virus.	Unlikely

Supplementary Table 12. Structural analysis of GP SDPs. Details of the structural analysis are included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible

(more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect).

Reston virus residue	Pathogenic consensus	Comments	Functional effect
K39	R39	R39 forms a H bond with D71. Change to K is likely to maintain this H bond.	Unlikely
S42	P42/ Q42	Unusual to see Pro in a sheet. The amino acid is on the protein surface and it there is nothing to suggest that a change to Ser would alter the protein	Unclear
V56	I56	I56 is largely buried and packed against other sidechains. While change to Val would reduce the size of the side chain, it seems likely that it would be accommodated within the structure. Also V64I is adjacent to this SDP.	Unlikely
I64	V64	In a surface loop facing the helix containing I56V. Possible co-evolution with I56 – reduce size in one, matched with increased size in the other.	Unlikely
K105	R105	The side chain guanidino group of R105 provides a hydrogen bond with the side chain of Q38 as well as with the local backbone NH of G103 to provide a stabilized region of the protein. Although the mutation R105K appears conservative and maintains the side chain positive charge, the ability to form multiple hydrogen bonds is reduced due to resonance stabilization in the guanidino group being lost in the transfer to the lysine side chain amino group. This has the potential to weaken interactions in this region.	Possible
L137	M137	M137 is located at the end of helix and packs against an adjacent helix. The conservative change to L137 in Reston virus seems unlikely to have a significant effect on structure/function	Unlikely
Y212	F212	A minor change in side chains. P212 is located in an alpha helix and the sidechain is largely buried. The change to Y212 in Reston virus is unlikely to have a significant effect on protein structure/function	Unlikely
R274	K274	K274 is located in the VP35 binding site. K274 forms a hydrogen bond with VP35 D46 and a change to Arg should be able to maintain this interaction.	Unlikely
A279	S279	S279 is located in an alpha helix on the protein surface. The change to A279 in Reston virus would introduce a hydrophobic amino acid on the protein surface that could have an effect on protein structure.	Unclear
R374	K374	K374 is located in an alpha helix on the protein surface. It is not unlikely that the change to R374 in Reston virus will alter protein structure. It is a conservative change of side chain.	Unlikely
R705	A705	A695 is located on the protein surface so the charge introduce by the change to R695 in Reston virus should be tolerated. Proximity of Reston virus R705 to E694 may result in a salt bridge that would reduce flexibility in Reston virus NP. There could different hydrodynamic volumes between the Reston virus and pathogenic NP proteins as well as in the pathogenic ebolaviruses exposing residues that remain buried in the Reston virus NP. The salt bridge could make RESTV more thermostable (and possibly more resistant to proteolysis and denaturants).	Possible

N716	D716	Present in a surface loop this change will change the charge properties. Should be considered with adjacent amino acid, which is also an SDP. Overall we see the removal of a negatively charged amino acid with two polar side chains.	Unclear
N717	G717	Adjacent to D716N pSDP. The loss of Gly would change the turn from type1 to a type 2 turn. Also See comment above.	Unclear

Supplementary Table 13. Structural analysis of NP SDPs. Details of the structural analysis are included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible (more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect).

Reston virus Residue	Pathogenic consensus	Comments	Functional effect
D258	E269	Present in dimer interface (only for one of the subunits as the dimer is asymmetric). Forms hydrogen bonds with R301, R311 and W313 (RESTV numbering). Distances between atoms are slightly different between the 2 species. W324 3.1A (2.8 in Ebola virus), R301 3.2A (2.9 in Ebola virus) R322 2.8 and 3.0 (both 2.8A in Ebola virus). Also close to A303 across interface, they could compensate or presence of both changes could have greater effect on interface in this area. (6.1A in RESTV, 7.5 in Ebola virus)	Probable
V279	A290	Present in a surface loop packs against adjacent helix, conservative change of hydrophobic amino acid. Could be some local conformational changes and is located adjacent to the linker between the two subdomains, which is in RESTV has a short alpha helix that is not present in EBOV.	Unclear
A303	V314	Present in a surface loop near the VP35 dimer interface. Close in space to D258 in the other subunit.	Unclear
K318	Q329	Located at the end of a beta sheet. Adjacent to His285 in next strand. His285 is completely conserved in all *Ebolavirus* species. So Reston virus VP35 has increased positive charge in this position	Unclear

Supplementary Table 14. Structural analysis of VP35 SDPs. Details of the structural analysis are included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible (more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect).

RESTV residue	Pathogenic consensus	Comments	functional effect
I151	T150	The side chain is largely buried and it appears that Reston virus I151 would be tolerated although a hydrogen bond with the backbone of the previous turn of the helix will be lost.	Unlikely
R158	Q157	Located in a surface loop, will increase surface charge. It is possible that Reston virus forms a salt bridge with D159, which would increase stability and reduce flexibility in this area of the protein. This SDP is in a region of SDPs and very close to another SDP (I159L). So possible effects may be compensated by other changes.	Unlikely
L160	I159	Located in a surface close to another SDP (see above). Appears to be a conservative change that given the other species specific changes in this area it seems unlikely that it will have a functional effect on the protein.	unlikely
H197	R196	Surface residue so change in size/shape should well accommodated, positive charge maintained in side chain.	Unlikely
D206	E205	Exposed surface residue, conservative change of amino acid. Unlikely to alter protein structure.	Unlikely
A263	R262	This residue is present in the dimer interface. In Ebola virus VP30 R262 hydrogen bonds with the backbone of A141 and G140. Reston virus A263 will be unable to hydrogen bond. This is likely to reduce the affinity of the dimer (given that it is symmetrical and so the Ebola virus R262 in each subunit forms hydrogen bonds with the other subunit. The Reston virus dimer has been observed to be rotated relative to the Ebola virus. The loss of the hydrogen bonds may explain this.	Probable

Supplementary Table 15. Structural analysis of VP30 SDPs. Details of the structural analysis are included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible (more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect).

Reston virus residue	Pathogenic consensus	Comments	Possible Functional effect
V46	T46	Present in a surface loop (although only third amino acid in structure). Reston virus V46 introduces a hydrophobic amino acid on surface, could affect stability but no evidence for this.	Unclear
T85	P85	Ebola virus P85 is in a S-G-P-K beta-turn, proline confers backbone rigidity and change to Thr in Reston virus would introduce backbone flexibility and provide a side chain with H-bond donor. Located in the Ebola virus octamer interface, will result in changes to this interface and likely alter the octamer structure. In an octamer structure (if it were to remain similar to the Ebola virus octamer), T85 could hydrogen bond with the backbone of L117 or the sidechain of R137.	probably
V122	I122	This change appears to be conservative substitution of two hydrophobic amino acids. Ebola virus I122 is packed with other hydrophobic residues and it appears that the region would be able to accommodate the change to Reston virus V122 with a slightly smaller side chain.	Unlikely
N201	G201	Located in a surface loop. Based on the Ebola virus structure, the Reston virus N201 side chain would be likely to point into the protein structure. But not clear what effect this would have on the protein structure, if any given that the structure has gaps in this region so cannot be confident.	Unclear
L209	F209	Packed in a largely hydrophobic region the SDP results in a reduction in side chain size in Reston virus. The smaller Leucine may adopt different side chain conformations to aid stability. Ebola virus F209 does not interact with other aromatic side chains so the structure is unlikely to be adversely affected by the swap to Leucine. Surrounding hydrophobic residues are aliphatic (I261, I285, V298, A318, P317) so the change to Leucine could be well accommodated.	Unlikely
P245	Q245	Located at the end of an alpha helix, the Reston virus P245 would break the helix and shorten it to either L244 or more likely M241, which is a better C-capping residue. This could have a destabilizing effect on the two helices in this region and the base of the hydrophobic core because secondary structure will most likely change to accommodate the inflexible Proline.	Probably
Q269	H269	A surface residue, loss of charge to polar side chain. This is a highly charged region with E265, R270, K274, K275. So the positive charge would be reduced in Reston virus VP40.	Unclear
V293	I293	Packs with other hydrophobic residues. Appears to be a conservative change	Unlikely

Supplementary Table 16. Structural analysis of VP40 SDPs. Details of the structural analysis are

included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible (more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect). Analysis is based on the VP40 dimer structure unless otherwise stated.

Reston virus residue	Pathogenic consensus	Comments	Possible functional effect
M17	L17	Located in a helix. Appears to be a conservative change in amino acid. No suggestion from structure that it would alter structure/function.	Unlikely
I22	V22	Located in a helix and is fairly tightly packed against the adjacent helix but would expect the pocket to accommodate the change.	Unlikely
I31	V31	Located in a sheet facing a loop. Side chain is relatively exposed so structure should be able to accommodate. Adjacent in space to another SDP (132)	Unlikely
S131	T131	Ebola virus T131 forms hydrogen bonds with the side chains of T129, W125 and with the backbone of H133. Model of Reston virus VP24 suggests S131 would continue to interact with the same residues. This residue is on the edge of the KPNA5 binding site. Appears to be a conservative change of amino acid.	Probable
T132	N132	Exposed polar residue exchanges for another polar residue. Unlikely to affect structure. Adjacent in space to an SDP (V31S) and in sequence to 131.	Unlikely
L136	M136	Part of the interface site with KPNA5. Mutagenesis of M136 in combination with other residues resulted in loss of KPNA5 binding[34]. Although it appears to be a conservative substitution.	Probable
R139	Q139	Interface residue. In Ebola virus Q139 forms an H bond with the backbone of R137. This is likely to be lost in Reston virus VP24 with the longer R139 side chain. Change will also introduce positive charge at interface site.	Probable
A226	T226	Located in a helix facing a sheet. Ebola virus T226 forms a hydrogen bond with the backbone of D48. Reston virus A226 will not be able to form this hydrogen bond. This is likely to reduce the stability of the protein and increase flexibility.	Probable

Supplementary Table 17. Structural analysis of VP24 SDPs. Details of the structural analysis are included with an assessment of whether the amino acid change is likely to have an effect on the protein. Four categories are used for the effect column unlikely (the change seems unlikely to alter the structure/function), unclear (the change could be functional but there is limited evidence), possible

(more confident that there is an effect than the unclear group) and probably (highly confident that the change will have a structural/functional effect).

Region	Residue	Conservation
1	L136	SDP
1	R139	SDP
1	S140	Not an SDP but conserved S in Reston viruses and mainly R in Ebola viruses, not conserved enough to be SDP
2	L107	Vary in species specific manner
2	H109	Vary in species specific manner
2	T116	Vary in species specific manner
2	G120	Not an SDP – G in Reston viruses and Ebola viruses (mainly), differs in others
3	S184	
3	T185	Not an SDP. T in Reston viruses, mainly N in other species
3	H186	Vary in species specific manner
3	T187	Not an SDP, primarily T in most species (A in Sudan viruses)
3	F197	Vary in species specific manner
4	V201	Vary in species specific manner
5	S50	Not an SDP

Supplementary Table 18. Residues in VP24 previously identified to differ between Reston viruses and Ebola viruses and/or Sudan viruses. Zhang et al., identified five regions that differed between Reston viruses and Ebola viruses and/or Sudan viruses[7] .The five regions are listed along with conservation information i.e. whether the position is an SDP, varies in a species specific manner (i.e. not an SDP, but a different residue is conserved in each of the different species) or otherwise conserved. Region one is part of the KPNA5 (karyopherin α5) binding site and region two is thought to be part of the STAT1 binding site[7].

Mutation	Location/Comments	Relationship to SDPs
From Volchhkov et al.,[43] – experiment 1		
M71I	Surface residue. Not clear what functional effect would be.	Not close
L147P	Part of an alpha helix, the proline would be expected to break the helix and could lead to conformational changes that would alter function.	Close to SDPs L17M, V22I
T187I	Adjacent to interface site. T187 forms Hydrogen bonds with the backbone of H186 and E203. Mutation to I would remove these hydrogen bonds and reduce stability/increase flexibility in this area. (Also close to L26F mutation from a separate study)	Not close
From Volchhkov et al., [43] – experiment 2		
H186Y	Present in interface with KPNA5. Forms a hydrogen bond with the backbone of T434 in KPNA5. Mutation to Tyr would still enable Hydrogen bonding with KPNA as the functional group is maintained.	Not close
From Ebihara et al., [44]		
T50I	The side chain of Ebola virus T50 can hydrogen bond with the backbones of Q36 and K52. Removal of these interactions with mutation Ile will reduce stability/increase flexibility.	Close to SDP T226A
From Dowall et al., [45]		
L26F	Largely buried side chain. Increase in size to phenylalanine could require some conformational change. Interesting that is located close to T187I (see above).	Close to V22I
F29V*	Largely buried side chain. Reduction in size would create space and therefore likely to result in some conformational change?	Close in space to SDPs T131S, N132T, V31I.
A43P*	Close in space to L26F (see above). Present in a turn.	
K218R*	Appears to be a conservative change. K218 is present in the KPNA5 interface. Is close to M436 and D489. Possible electrostatic interaction. Possible the mutation to R enables this interaction to continue in the different species.	

Supplementary Table 19. VP24 Mutations occurring in adaption of Ebola virus to rodent species. The location of the mutation and how it may alter structure and function is listed with details of proximity to SDPs. *indicates that after passage one the predominant amino acid at that position was the wild type [44]. In the Dowall et al.[45], study L26F is the only mutation where the mutation is predominantly maintained in in all passages. Separate experimental evidence suggests that the L26F mutation along results in pathogenicity in guinea pigs[37].

Genome Identifier	Ebola virus species	Host
gb:KJ660346	Organism:Zaire ebolavirus H.sapiens-wt/GIN/2014/Makona-Kissidougou-C15	Human
gb:KJ660347	Organism:Zaire ebolavirus H.sapiens-wt/GIN/2014/Makona-Gueckedou-C07	Human
gb:KJ660348	Organism:Zaire ebolavirus H.sapiens-wt/GIN/2014/Makona-Gueckedou-C05	Human
gb:KP342330	Organism:Zaire ebolavirus H.sapiens-wt/GIN/2014/Conacry-192	Human
gb:KP096422	Organism:Zaire ebolavirus H.sapiens-tc/GIN/14/WPG-C15	Human
gb:KP096421	Organism:Zaire ebolavirus H.sapiens-tc/GIN/14/WPG-C07	Human
gb:KP096420	Organism:Zaire ebolavirus H.sapiens-tc/GIN/14/WPG-C05	Human
gb:KC242800	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/2002/Ilembe	Human
gb:KC242794	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1996/2Nza	Human
gb:KC242797	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1996/1Oba	Human
gb:KC242795	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1996/1Mbie	Human
gb:KC242798	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1996/1Ikot	Human
gb:KC242793	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1996/1Eko	Human
gb:KC242792	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/GAB/1994/Gabon	Human
gb:KC242784	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/9 Luebo	Human
gb:KC242790	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/5 Luebo	Human
gb:KC242788	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/43 Luebo	Human
gb:KC242789	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/4 Luebo	Human
gb:KC242787	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/23 Luebo	Human
gb:KC242786	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/1 Luebo	Human
gb:KC242785	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/2007/0 Luebo	Human
gb:KC242799	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/1995/13709 Kikwit	Human
gb:KC242796	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/1995/13625 Kikwit	Human
gb:KC242791	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/1977/Bonduni	Human
gb:KC242801	Organism:Zaire ebolavirus EBOV/H.sapiens-tc/COD/1976/deRoover	Human
gb:KM233118	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-NM042.3	Human
gb:KM233117	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-NM042.2	Human
gb:KM233116	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-NM042.1	Human
gb:KM233115	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3857	Human
gb:KM233114	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3856.3	Human
gb:KM233113	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3856.1	Human
gb:KM233112	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3851	Human
gb:KM233111	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3850	Human
gb:KM233110	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3848	Human
gb:KM233109	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3846	Human
gb:KM233108	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3845	Human
gb:KM233107	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3841	Human
gb:KM233106	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3840	Human
gb:KM233105	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3838	Human
gb:KM233104	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3834	Human
gb:KM233103	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3831	Human
gb:KM233102	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3829	Human
gb:KM233101	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3827	Human
gb:KM233100	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3826	Human
gb:KM233099	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3825.2	Human
gb:KM233098	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3825.1	Human
gb:KM233097	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3823	Human
gb:KM233096	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3822	Human
gb:KM233095	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3821	Human
gb:KM233094	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3820	Human
gb:KM233093	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3819	Human
gb:KM233092	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3818	Human

gb:KM233091	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3817	Human
gb:KM233090	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3816	Human
gb:KM233089	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3814	Human
gb:KM233088	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3810.2	Human
gb:KM233087	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3810.1	Human
gb:KM233086	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3809	Human
gb:KM233085	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3808	Human
gb:KM233084	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3807	Human
gb:KM233083	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3805.2	Human
gb:KM233082	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3805.1	Human
gb:KM233081	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3800	Human
gb:KM233080	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3799	Human
gb:KM233079	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3798	Human
gb:KM233078	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3796	Human
gb:KM233077	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3795	Human
gb:KM233076	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3789.1	Human
gb:KM233075	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3788	Human
gb:KM233074	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3787	Human
gb:KM233073	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3786	Human
gb:KM233072	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3782	Human
gb:KM233071	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3771	Human
gb:KM233070	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3770.2	Human
gb:KM233069	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3770.1	Human
gb:KM233068	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3769.4	Human
gb:KM233067	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3769.3	Human
gb:KM233066	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3769.2	Human
gb:KM233065	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3769.1	Human
gb:KM233064	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3765.2	Human
gb:KM233063	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3764	Human
gb:KM233062	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3758	Human
gb:KM233061	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3752	Human
gb:KM233060	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3750.3	Human
gb:KM233059	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3750.2	Human
gb:KM233058	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3750.1	Human
gb:KM233057	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3735.2	Human
gb:KM233056	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3735.1	Human
gb:KM233055	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3734.1	Human
gb:KM233054	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3729	Human
gb:KM233053	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3724	Human
gb:KM233052	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3713.4	Human
gb:KM233051	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3713.3	Human
gb:KM233050	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3713.2	Human
gb:KM233049	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3707	Human
gb:KM034563	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3687.1	Human
gb:KM034562	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3686.1	Human
gb:KM034561	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3683.1	Human
gb:KM034560	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3682.1	Human
gb:KM034559	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3680.1	Human
gb:KM034558	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3679.1	Human
gb:KM034557	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3677.2	Human
gb:KM034556	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3677.1	Human
gb:KM034555	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3676.2	Human
gb:KM034554	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3676.1	Human

gb:KM034553	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3670.1	Human
gb:KM233048	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM124.4	Human
gb:KM233047	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM124.3	Human
gb:KM233046	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM124.2	Human
gb:KM233045	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM124.1	Human
gb:KM233044	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM121	Human
gb:KM233043	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM120	Human
gb:KM233042	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM119	Human
gb:KM233041	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM115	Human
gb:KM233040	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM113	Human
gb:KM233039	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM112	Human
gb:KM233038	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM111	Human
gb:KM233037	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM110	Human
gb:KM233036	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM106	Human
gb:KM233035	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM104	Human
gb:KM034552	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM098	Human
gb:KM034551	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM096	Human
gb:KM034549	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM095B	Human
gb:KM034550	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/SLE/2014/Makona-EM095	Human
gb:KP178538	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/LBR/2014/Makona-201403007	Human
gb:KP120616	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/GBR/2014/Makona-UK1	Human
gb:KP271020	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/COD/2014/Lomela-Lokolia19	Human
gb:KP271018	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/COD/2014/Lomela-Lokolia16	Human
gb:KP728283	Organism:Zaire ebolavirus Ebola virus/H.sapiens-wt/CHE/2014/Makona-GE1	Human
gb:KP701371	Organism:Zaire ebolavirus Ebola virus/H.sapiens-tc/SLE/2014/Makona-Italy-INMI1	Human
gb:KP184503	Organism:Zaire ebolavirus Ebola virus/H.sapiens-tc/GBR/2014/Makona-UK1.1	Human
gb:KM655246	Organism:Zaire ebolavirus Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Ecran	Human
gb:KP260802	Organism:Zaire ebolavirus Ebola virus H.sapiens/MLI/14/Manoka-Mali-DPR4	Human
gb:KP260801	Organism:Zaire ebolavirus Ebola virus H.sapiens/MLI/14/Manoka-Mali-DPR3	Human
gb:KP260800	Organism:Zaire ebolavirus Ebola virus H.sapiens/MLI/14/Manoka-Mali-DPR2	Human
gb:KP260799	Organism:Zaire ebolavirus Ebola virus H.sapiens/MLI/14/Manoka-Mali-DPR1	Human
gb:NC_002549	Organism:Zaire ebolavirus Ebola virus H.sapiens-tc/COD/1976/Yambuku-Mayinga	Unknown
gb:AY354458	Organism:Zaire ebolavirus Zaire 1995	Unknown
gb:JA489037	Organism:Zaire ebolavirus UNKNOWN-JA489037	Unknown
gb:HC874683	Organism:Zaire ebolavirus UNKNOWN-HC874683	
gb:HC874681	Organism:Zaire ebolavirus UNKNOWN-HC874681	
gb:HC874677	Organism:Zaire ebolavirus UNKNOWN-HC874677	
gb:HC874665	Organism:Zaire ebolavirus UNKNOWN-HC874665	
gb:HC874661	Organism:Zaire ebolavirus UNKNOWN-HC874661	
gb:HC069241	Organism:Zaire ebolavirus UNKNOWN-HC069241	
gb:HC069239	Organism:Zaire ebolavirus UNKNOWN-HC069239	
gb:HC069235	Organism:Zaire ebolavirus UNKNOWN-HC069235	
gb:HC069221	Organism:Zaire ebolavirus UNKNOWN-HC069221	
gb:HC069217	Organism:Zaire ebolavirus UNKNOWN-HC069217	
gb:KF827427	Organism:Zaire ebolavirus rec/COD/1976/Mayinga-rgEBOV	Human
gb:AF272001	Organism:Zaire ebolavirus Mayinga	Guinea Pig
gb:AF499101	Organism:Zaire ebolavirus Mayinga	Guinea Pig
gb:AY142960	Organism:Zaire ebolavirus Mayinga	Guinea Pig
gb:EU224440	Organism:Zaire ebolavirus Mayinga	Guinea Pig
gb:AF086833	Organism:Zaire ebolavirus Mayinga	Guinea Pig
gb:JQ352763	Organism:Zaire ebolavirus Kikwit	Unknown
gb:JA489027	Organism:Tai Forest ebolavirus UNKNOWN-JA489027	Unknown
gb:FJ217162	Organism:Tai Forest ebolavirus UNKNOWN-FJ217162	Human

gb:NC_014372	Organism:Tai Forest ebolavirus Tai Forest virus/H.sapiens-tc/CIV/1994/Pauleoula-CI	Human
gb:EU338380	Organism:Sudan ebolavirus Yambio	Human
gb:HC874655	Organism:Sudan ebolavirus UNKNOWN-HC874655	
gb:HC069211	Organism:Sudan ebolavirus UNKNOWN-HC069211	
gb:KC242783	Organism:Sudan ebolavirus SUDV/H.sapiens-tc/SSD/1979/Maleo	Human
gb:NC_006432	Organism:Sudan ebolavirus Sudan virus/H.sapiens-tc/UGA/2000/Gulu-808892	Unknown
gb:JN638998	Organism:Sudan ebolavirus Sudan	Human
gb:AY729654	Organism:Sudan ebolavirus Gulu	Unknown
gb:KC545392	Organism:Sudan ebolavirus EboSud-682 2012	Human
gb:KC589025	Organism:Sudan ebolavirus EboSud-639	Human
gb:KC545391	Organism:Sudan ebolavirus EboSud-609 2012	Human
gb:KC545390	Organism:Sudan ebolavirus EboSud-603 2012	Human
gb:KC545389	Organism:Sudan ebolavirus EboSud-602 2012	Human
gb:FJ968794	Organism:Sudan ebolavirus Boniface	Unknown
gb:HC874675	Organism:Reston ebolavirus UNKNOWN-HC874675	
gb:HC874663	Organism:Reston ebolavirus UNKNOWN-HC874663	
gb:HC874659	Organism:Reston ebolavirus UNKNOWN-HC874659	
gb:HC874657	Organism:Reston ebolavirus UNKNOWN-HC874657	
gb:HC069233	Organism:Reston ebolavirus UNKNOWN-HC069233	
gb:HC069219	Organism:Reston ebolavirus UNKNOWN-HC069219	
gb:HC069215	Organism:Reston ebolavirus UNKNOWN-HC069215	
gb:HC069213	Organism:Reston ebolavirus UNKNOWN-HC069213	
gb:JX477165	Organism:Reston ebolavirus Reston09-A	Swine
gb:FJ621585	Organism:Reston ebolavirus Reston08-E	Swine
gb:FJ621584	Organism:Reston ebolavirus Reston08-C	Swine
gb:FJ621583	Organism:Reston ebolavirus Reston08-A	Swine
gb:NC_004161	Organism:Reston ebolavirus Reston virus/M.fascicularis-tc/USA/1989/Philippines89- Pennsylvania	Unknown
gb:AB050936	Organism:Reston ebolavirus Reston	
gb:AF522874	Organism:Reston ebolavirus Pennsylvania	
gb:AY769362	Organism:Reston ebolavirus Pennsylvania	
gb:JX477166	Organism:Reston ebolavirus Alice, TX USA MkCQ8167	Monkey
gb:NC_014373	Organism:Bundibugyo virus Bundibugyo virus/H.sapiens-tc/UGA/2007/Butalya-811250	Human
gb:JA489018	Organism:Bundibugyo ebolavirus UNKNOWN-JA489018	Unknown
gb:FJ217161	Organism:Bundibugyo ebolavirus UNKNOWN-FJ217161	Human
gb:KC545396	Organism:Bundibugyo ebolavirus EboBund-14 2012	Human
gb:KC545395	Organism:Bundibugyo ebolavirus EboBund-122 2012	Human
gb:KC545394	Organism:Bundibugyo ebolavirus EboBund-120 2012	Human
gb:KC545393	Organism:Bundibugyo ebolavirus EboBund-112 2012	Human

Supplementary Table 20. Information on the 196 complete *Ebolavirus* genomes. Genomes were downloaded from Virus Pathogen Resource, VIPR (http://www.viprbrc.org/brc/home.spg?decorator=vipr) .

Protein	Effective number of sequences	Effective number of human pathogenic sequence	Effective number of Reston virus sequences
GP	95.15	86	4
L	99.2	78	7
NP	148.96	133	7
VP24	88.2	79	7
VP30	96.04	84	7
VP35	99.96	87	7
VP40	90.16	80	7

Supplementary Table 21. Effective number of independent sequences in the dataset. The effective number of independent sequences present in the multiple sequence alignments for each of the Ebolavirus proteins is shown. Values were calculated using hmmer (see material and methods).

Appendix 3

Investigating Ebola virus pathogencity using Molecular Dynamics

Morena Pappalardo, Francesca Collu, James Macpherson, Martin Michaelis, Franca Fraternali, Mark N Wass, in preparation.

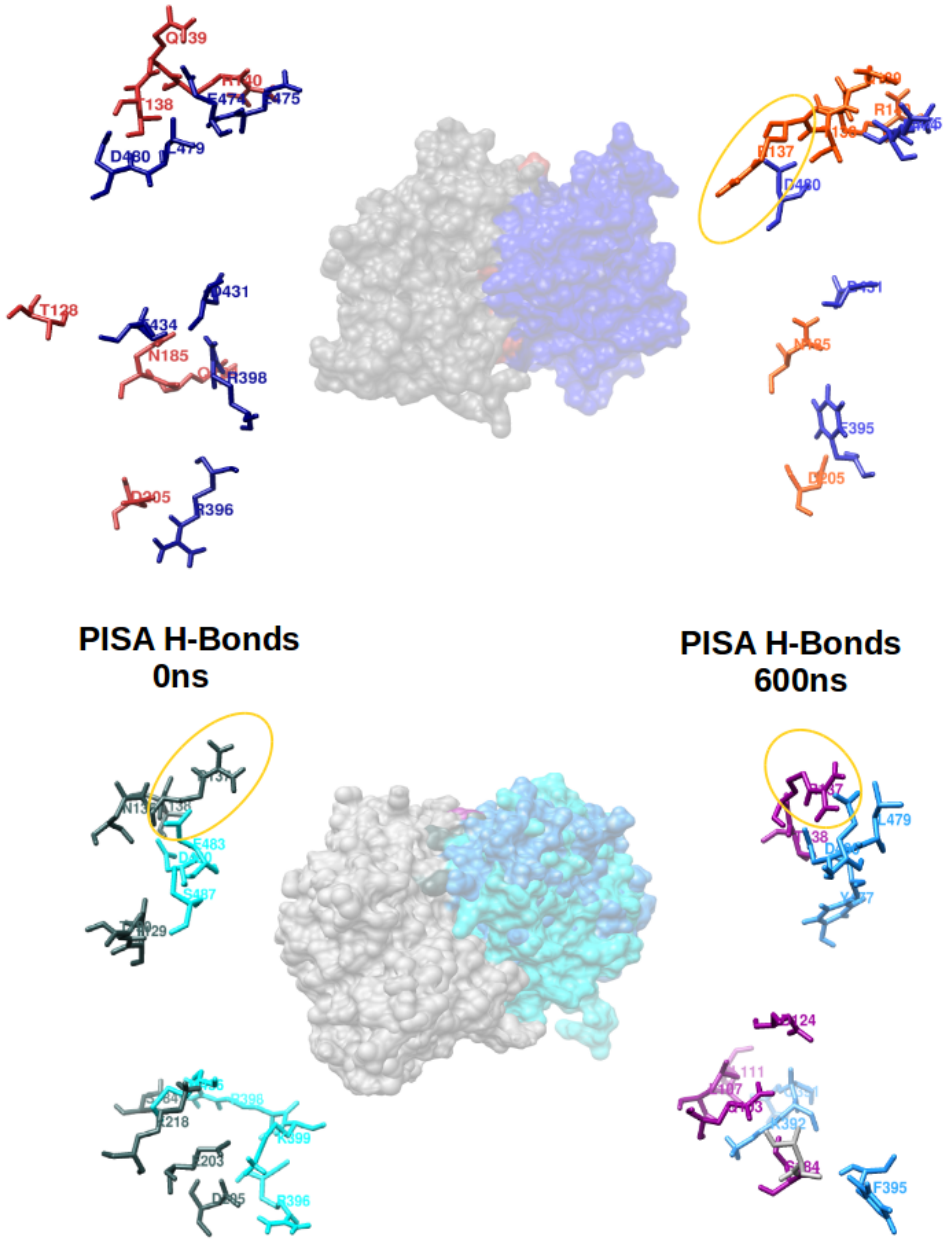


Figure S1 Comparison of the Hbonds at the interface in the EBOV complex (A) and in the RESTV (B) respectively t zero and 600 ns.

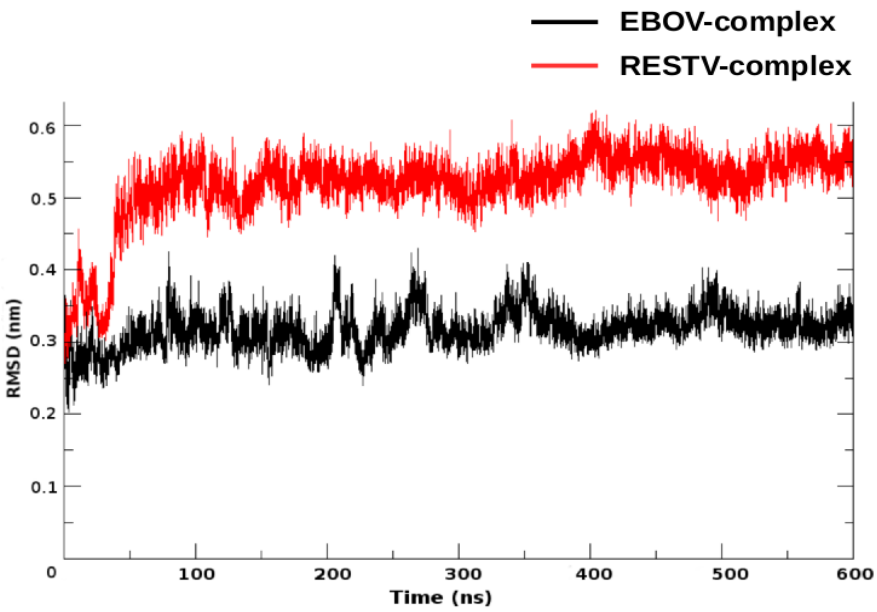


Figure S2: RMSD over time plot on the left; on the right the histogram of RMSD, showing the distances of the conformations from the starting one, during the simulation.

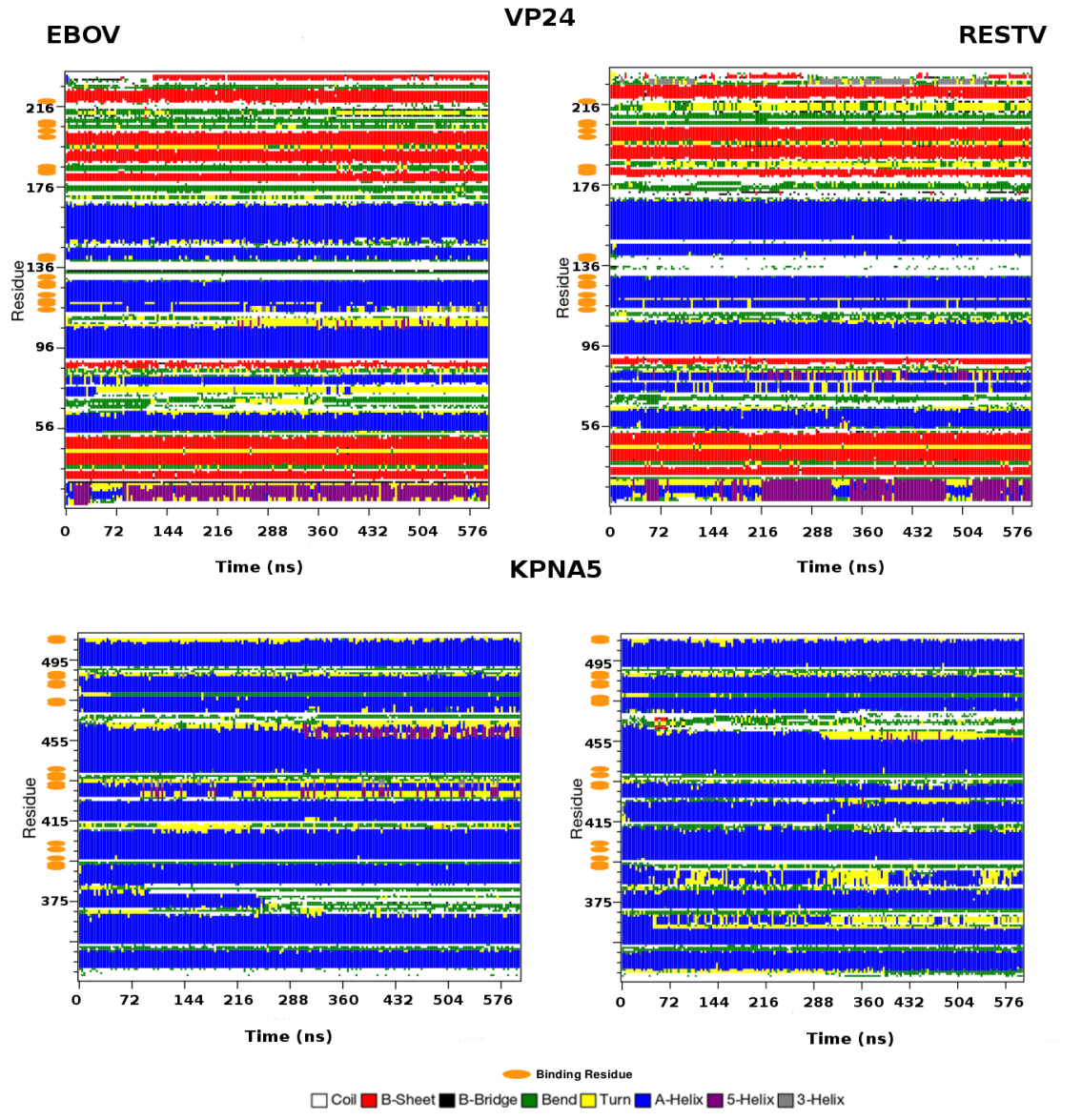


Figure S3 : DSSP graph of EBOV-VP24-KPNA5 and RESTV-KPNA5. We split proteins VP24 from the KPNA5. Residues at the interface were mapped using a yellow circle.

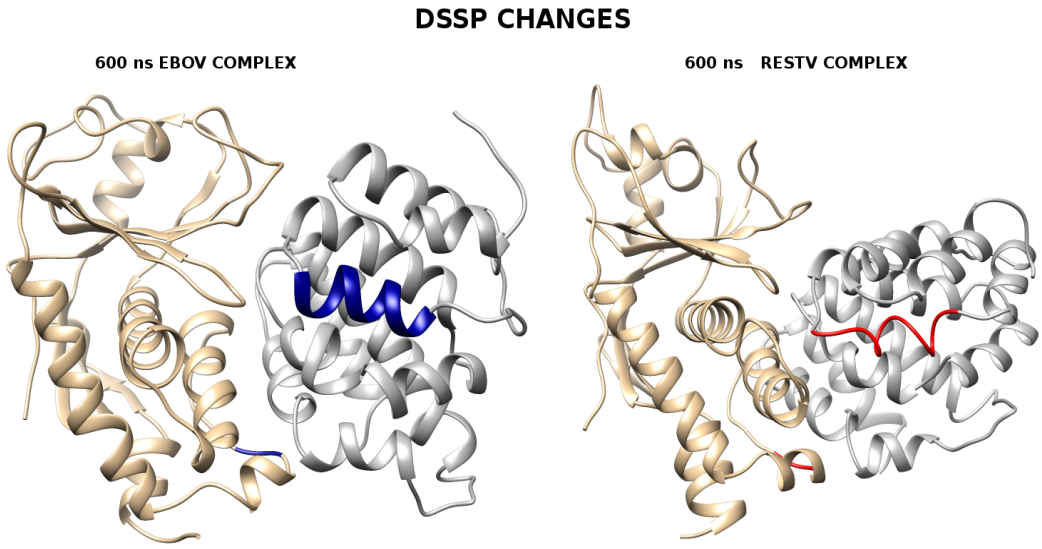


Figure S4 a) EBOV complex and B) RESTV complex. Changes in secondary structure, coming from DSSP analysis were mapped onto the 3D structures. Differences found in regions around residue 134 in VP24 (loop coloured in blue for EBOV and in red for the RESTV) and around region 385-395 in the KPNA5, where a loss of Alpha Helix is shown in the RESTV complex; this last difference is coloured in blue (EBOV) and their correspondent RESTV in red.

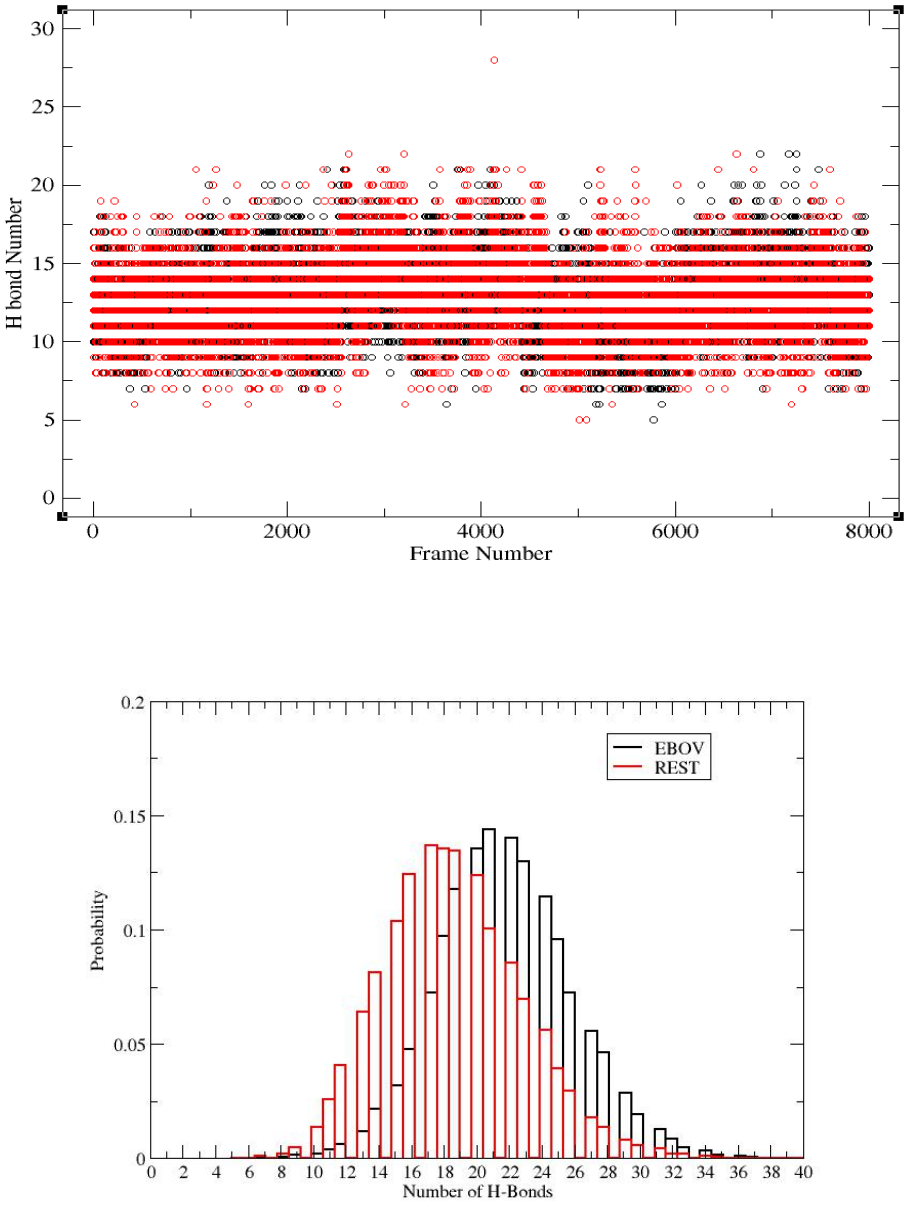


Figure S5: A) number of H-bond over time plot. In black circles the EBOV and in red circles the RESTV complex are shown. The number of H-bonds is constant during the 600 ns simulation. B) The probability to find H-bonds during the simulations suggestes that EBOV shows a greater H-bonds number (black Gaussian) than the RESTV (red Gaussian)

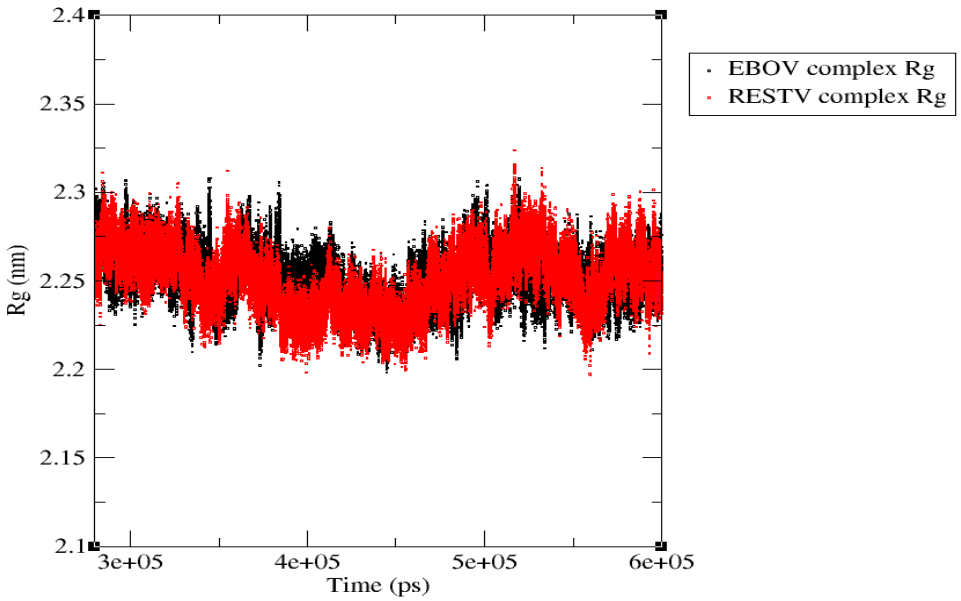


Figure S6: Radius of Gyration showed a constant compactness in both complexes during the simulation.

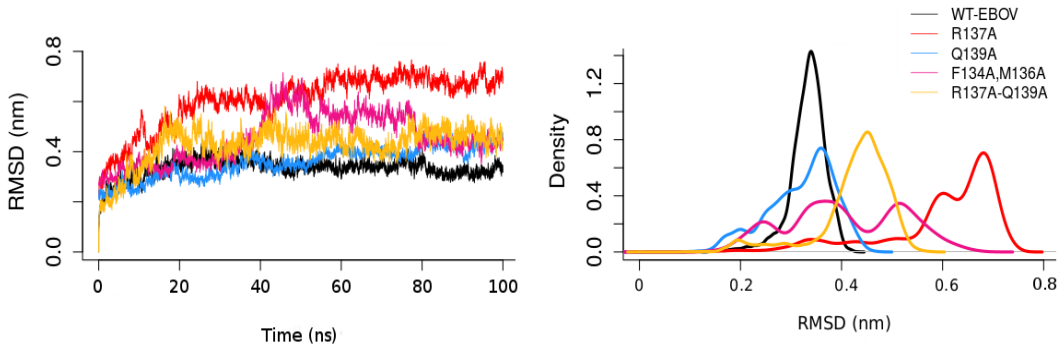


Figure S7: RMSD over time and RMSD histogram showed higher values for R137A and F134A-M136A. Again the RMSD in the left graph does not match that on the right.

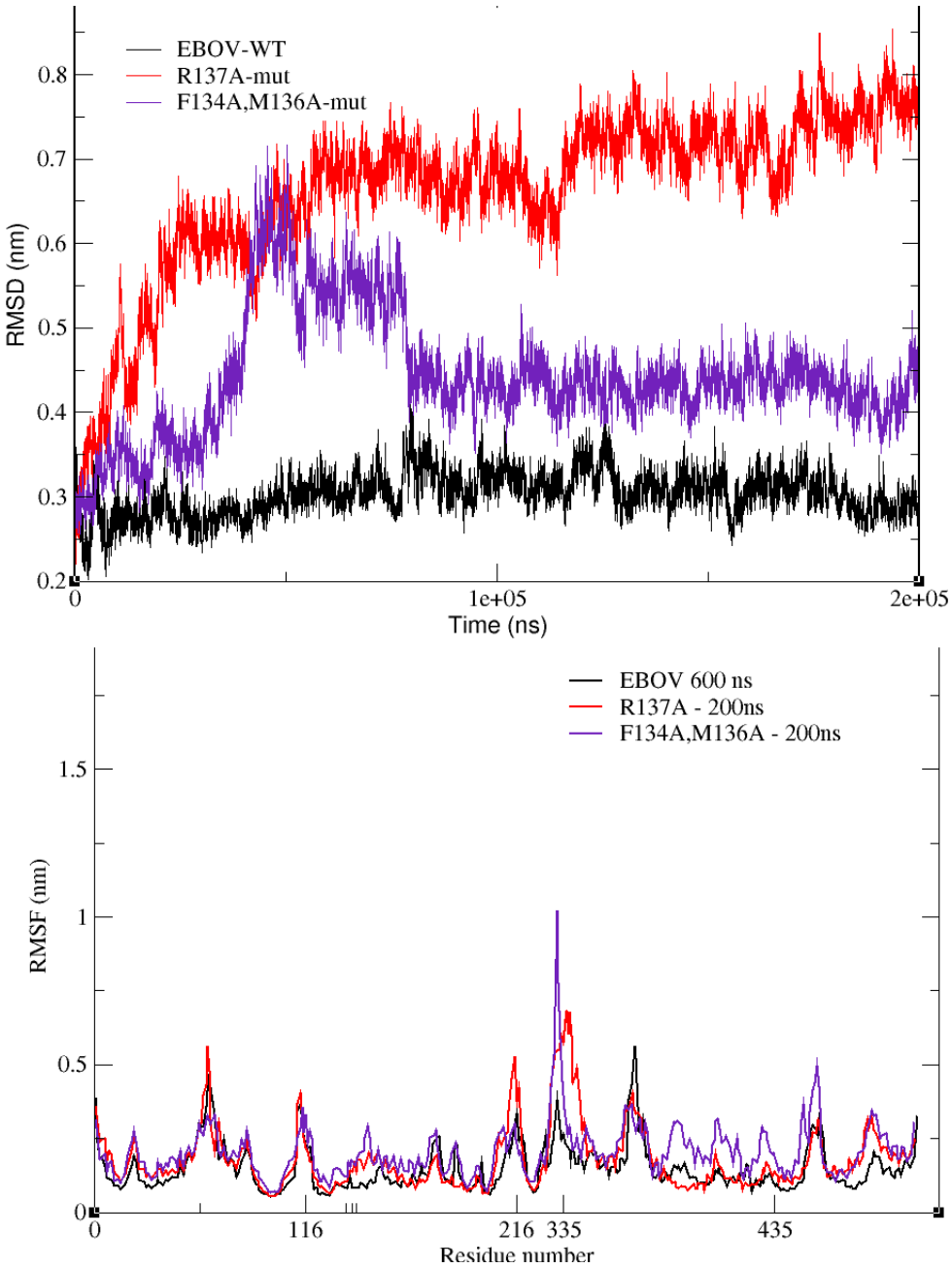


Figure S8: 200 ns trajectories RMSD and RMSF for EBOV WT (black lines), R137A-VP24-KPNA5 (red lines) and F134A-M136A-VP24-KPNA5 (purple lines).

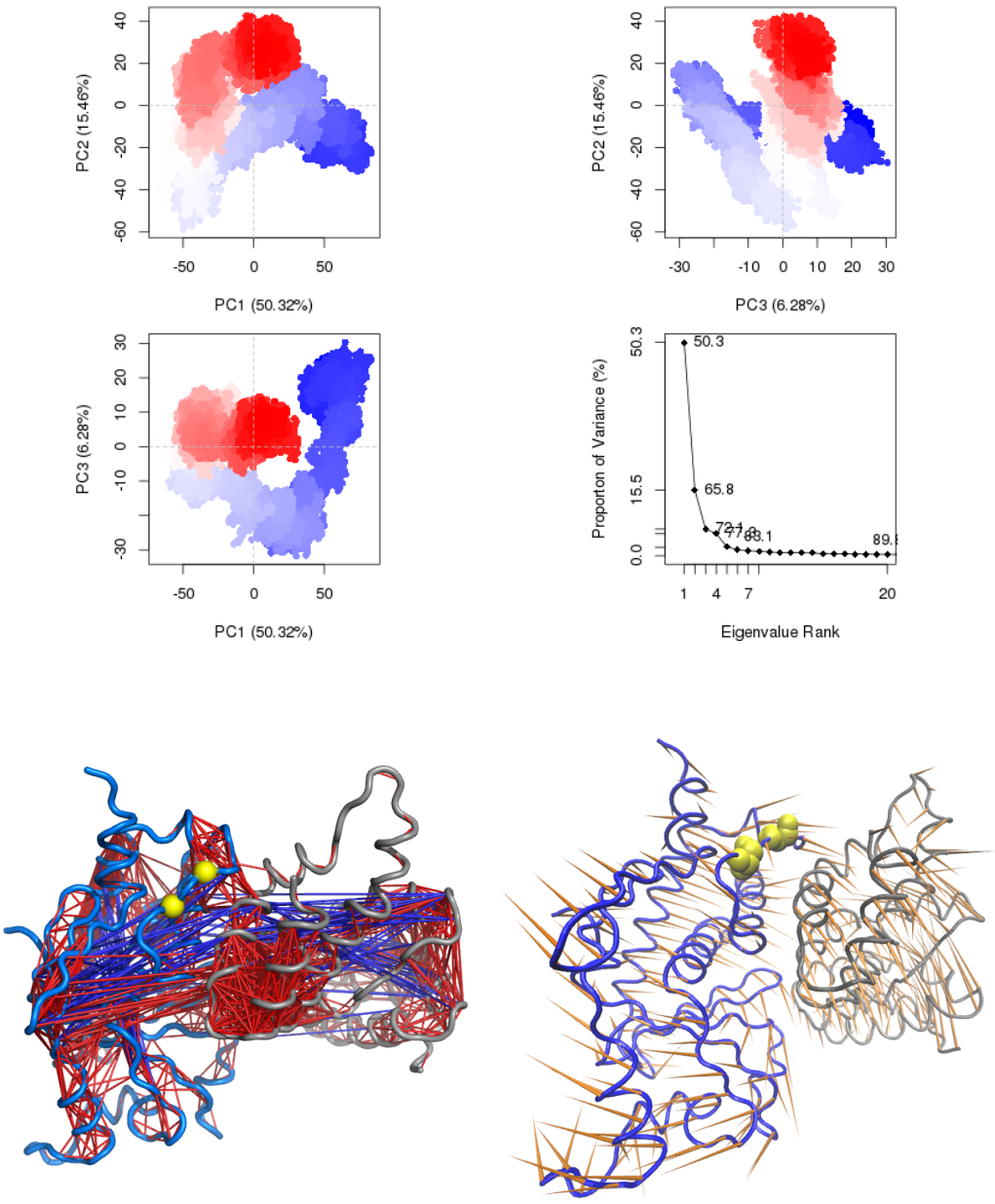


Figure S9: Principal Component Analysis A) in F134A-M136A-VP24-KPNA5 and B) Cross correlation analysis. Correlated movements are shown in red lines and anti-correlated ones in blue. Protein VP24 (blue cartoon) and the mutations F134A-M136A (yellow spheres) at the interface with KPNA5 (gray cartoon) are likely be in a more correlated region. C) Porcupine plot shows large movements during the simulation, occurring both in the VP24 and in the KPNA5.