Faculty of Science, Technology and

Medicine.

Analysing Genetic Variation in Ebolaviruses and Cancer Cell Lines

A dissertation submitted for the degree of
Doctor of Philosophy
in the University of Kent for the Faculty of Science, Technology and Medicine.

Miguel Juliá

Canterbury, 2018

Dissertation Supervisors:

Dr Mark N Wass

Dr Martin Michaelis

Dissertation Committee:

external examiner,

Dr Franca Fraternali

internal examiner,

Dr Tobias von der Haar

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.

Table of contents

CHAPTER 2: Ebola: genetic variance and its impact in human pathogenicity

(The research within this chapter consists of published data for the Scientific Report Journal.)

CHAPTER 3: UKF-NB-3 genetic landscape

Abstract

With the arrival of the -omics era and the democratisation of genome sequencing the amount of genetic data is escalating in magnitude orders every year. However, despite all this raw data, the effect prediction of genetic variations in disease remains an open question. The future machine learning algorithms which could solve the problem still require lots of information to feed their development, and it is our mission as bioinformaticians to extract it from the oceans of data.

This Thesis focusses in the analysis of genetic variation in two complete different diseases: Ebolavirus and neuroblastoma.

After the last Ebolavirus outbreak in West Africa (2014), the deadliest one in history, researchers sequenced lots of viral genomes for both surveillance and study of the pathogenic strain. There are still lots to learn from this virus and this Thesis wants to contribute with the study of how it becomes human pathogenic. By comparing different Ebolavirus species, four pathogenic to humans and one not, and looking into functionally important residues called Specificity Determining Positions (SDPs) in their genomes, we predict protein residues which may be key to the host-specifity pathogenicity.

Neuroblastoma is one of the most common cancers in infancy, and the high-risk variety remains a challenging and deadly disease. Chemotherapy is a key treatment for this cancer, so diagnostic of the right drug and effective monitoring of drug resistance emergence could increase the cure ratio of patients. In order to learn more about the genetic variance of this cancer in response to treatment and the effect of these variants in drug resistance emergence, we study the genome of the neuroblastoma cell line UKF-NB-3 and its clonal sub-lines.

List of abbreviations

BDBV	Bundibugyo Ebolavirus
BWA	Burrows-Wheeler Aligner
CNV	Copy Number Variation
EBOV	Zaire Ebolavirus
ESP	Exome Sequencing Project
EVD	Ebola Virus Disease
HTS	High Throughput Sequencing
ExAC	Exome Sequencing Project
GATK	Genome Analysis Toolkit
GP	GlycoProtein
IGSR	International Genome Sample Resource
INDELs	Insertions and DELetions
L	viral RNA-dependent RNA polymerase
LLOV	Lloviu Cuevavirus
MAF	Minor Allele Frequency
MARV	Marbugvirus
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NP	NucleoProtein
ORF	Open Reading Frame
PDB	Protein Data Bank
PTM	Post Translational Modification
PolyPhen2	Polymorphism Phenotyping V2
RESTV	Reston Ebolavirus
SIFT	Sorting Intolerant from Tolerant
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant

SPD	Specificity Determining Position
SUDV	Sudan Ebolavirus
SV	Structural Variant
TAFV	Taï Forest Ebolavirus
UTR	UnTranslated Regions
VP24	Viral Protein 24
VP30	Viral Protein 30
VP35	Viral Protein 35
VP40	Viral Protein 40
ViPR	Virus Pathogen Resource
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
dB	Decibel
dsRNA	double stranded RNA
gnomAD	Genome Aggregation Database
sGP	soluble GlycoProtein
ssGP	small soluble GlycoProtein

List of Figures

Figures in Chapter 4

Figures in Chapter 5

List of Tables

Chapter 1:
Introduction

This Thesis contains two different research lines. The first part corresponds to our research of Ebolavirus genetic variants and their role in human pathogenicity, while the second part focus on neuroblastoma and the description of the cell line UKF-NB-3 and its drug adapted clones.

1.1 Genome Sequencing

Genome sequencing is the process of determining the complete DNA sequence of an organism's genome. The technology has walked a long path in just three decades from the first sequenced genome of a Bacteriophage MS2, completed in 1976. The Human Genome Project started sequencing the first human genome in 1990, and it took only 13 years and $3-billion to complete the sequence.

With that skyrocketing cost, both in time and money, this technology was reserved for a few research projects. High-throughput, previously known as Next Generation Sequencing (NGS), englobes a set of sequencing technologies which have revolutionised genomic research, reducing the cost to sequence a human genome to below $1000 and the time required to just days, and making it affordable for extensive research and opening the door to clinical applications.

1.1.1. Types of Sequencing

The old "basic" sequencing methods have been the gold standard for years, and until recently they were considered the most relievable method to sequence a genome. The principal characteristic which made these methods so reliable is that they can sequence a DNA molecule of a fixed length in one piece, allowing sequencing of captured fragments of the genome to be used without any bioinformatics processing. But these methods were not useful for sequencing long DNA sequences, including whole chromosomes, because each segment of the genome had to be individually sequenced one by one and then assembled all together in the right order, turning in a highly expensive and time-consuming challenge. These methods are:

- Chemical sequencing (Maxam & Gilbert, 1977): published by Allan Maxam and Walter Gilbert in 1977, this method based on chemical modification allowed to sequence purified samples of double-stranded DNA without amplification. The technical complexity of the method and the need of radioactive labelling discouraged researchers to use it.
- Sanger sequencing (Sanger, Nicklen, & Coulson, 1977): published in 1977 and also known as chain-termination method. It became wide used by researchers for using less chemicals and lower radioactivity levels than its competitor, the chemical sequencing method. During decades it was improved in many ways, including automation, replacement of radioactive labelling with its fluorescent counterpart and capillary electrophoresis. The improvements of the method made possible the sequencing of the first human genome in 2003 using this technology.

Sanger sequencing became *the* sequencing method until the emergence of high-throughput methods in the early 2000s (de Magalhães, Finch, & Janssens, 2010). These revolutionary technologies lowered even more the price of sequencing genomes and nowadays are broadly used around the world. These new high throughput sequencing (HTS) methods traded off the ability of sequence longer

DNA segments to be able to sequence lots of shorter sequences quickly and massively (Figure 1.1). Now a genome could be sequenced in only one experiment by using different enzymes to chop it in small overlapping fragments and amplifying them with a PCR. The resulting short sequences that will be sequenced are called reads and can be used to rebuild the original genome by different assembling strategies.

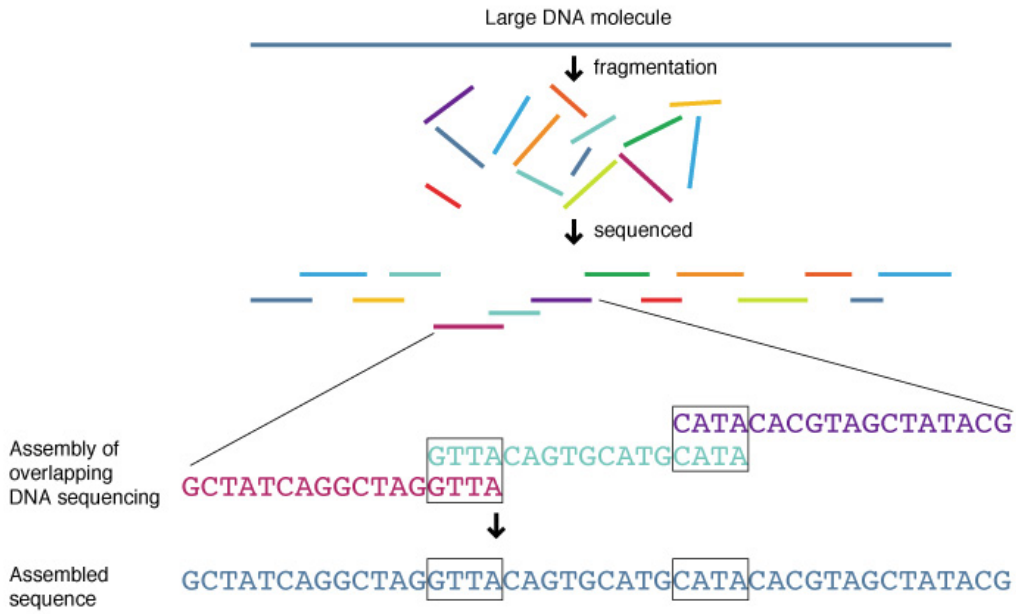


Figure 1.1. Synopsis of HTS strategies. The DNA molecule is split in smaller overlapping fragments. Each of those DNA fragments will be sequenced. Finally, the sequenced reads have to be computationally assembled in order to rebuild the sequenced DNA molecule.

Some of the most popular methods are:

- Massively parallel signature sequencing (Brenner et al., 2000): this was the first NGS method. It is a bead-based method which split DNA in thousands of short sequences. This method became obsolete when the company that created it merged with Solexa and later was bought by Illumina, leading to the development of sequencing-by-synthesis.

- Pyrosequencing (Margulies et al., 2005): each DNA fragment is attached to a bead and introduced inside water droplets in an oil solution, where PCR is carried (emulsion PCR). Sequencing is done in individual wells where each amplified fragment is deposited by using luciferase for detection of the individual nucleotides added to the growing DNA sequence.
- Illumina (Solexa) sequencing (Bentley et al., 2008): Arguably the most used method in the last decade, and the one we used to sequence our cancer lines in the following chapters. For this reason, this method will be explained more in detail step by step (Figure 1.2):
    - First, the randomly fragment DNA segments are attached to a ligate adapter in both ends to keep track of the sample they come from.
    - Fragments are split into single-stranded fragments and randomly bind to the inside surface of the flow cell channels of the sequencing plate.
    - Unlabelled nucleotides and enzymes are added to initiate the solid-phase bridge amplification, when fragments attach their second end to the surface forming a bridge-like structure.
    - The enzymes incorporate nucleotides to build double-stranded bridges on the solid-phase substrate.
    - Double-stranded bridges are now denaturised leaving single-stranded templates anchored to the substrate, but twice the amount we had previously.
    - This process is repeated until the desired amount of reads is reached, generating several million dense clusters of double-stranded DNA in each channel of the flow cell.
    - The first sequencing cycle begins by adding four labelled reversible terminators, primers, and DNA polymerase.
    - After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

- The next cycle repeats the incorporation of four labelled reversible terminators, primers, and DNA polymerase, and a new image is taken after laser excitation to capture the fluorescence of the new added base in each cluster.
- The procedure continues until the last base of each cluster is captured. The per-base quality of the sequenced reads is later calculated by computing the colour intensities of the clusters in each image.

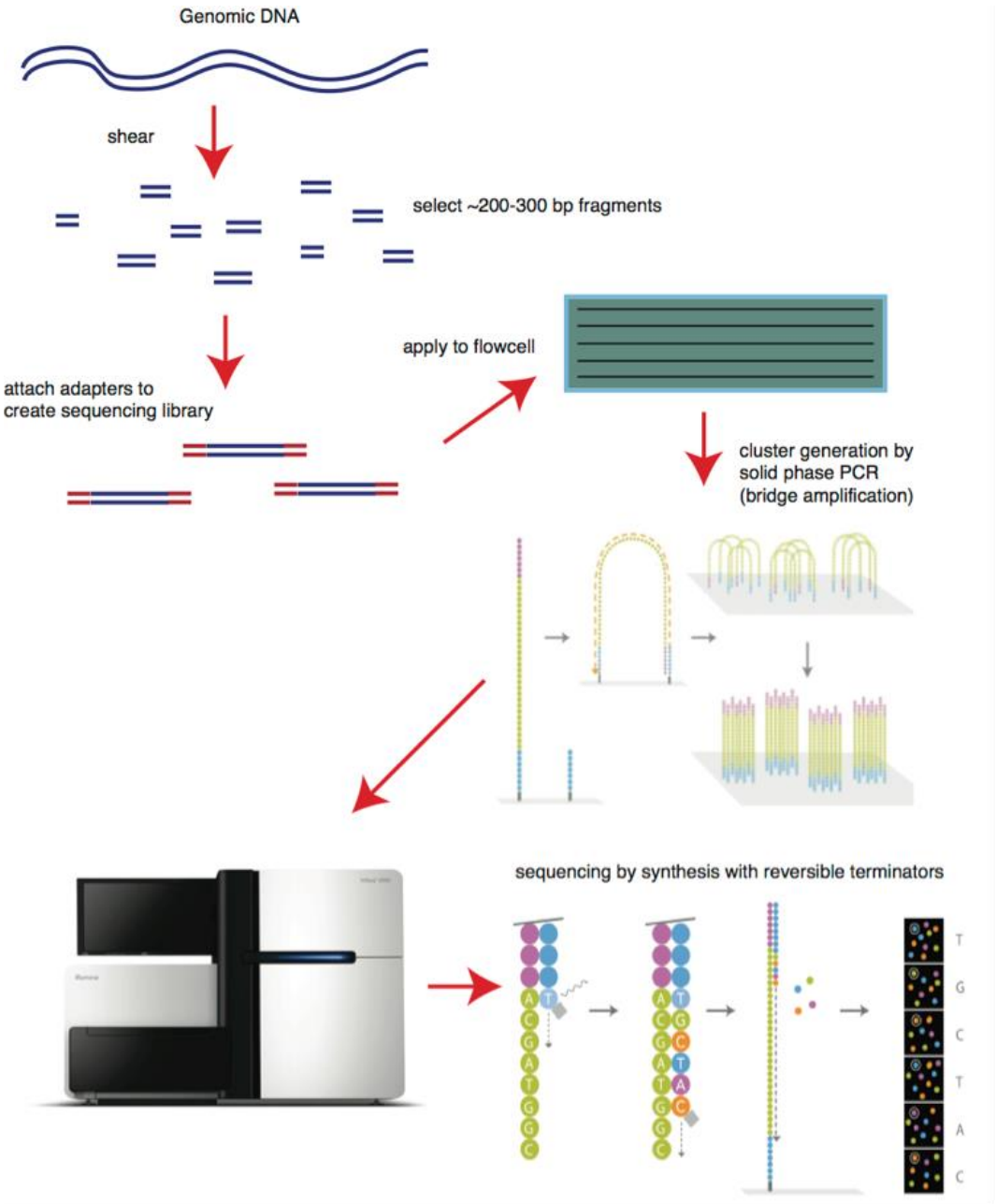


Figure 1.2. Illumina sequencing. This method first attaches primers to DNA fragments, and those to a flow cell where they are amplified with polymerase forming DNA clusters. While the clusters grow base by base, the machine adds four types of reversible terminator bases and a camera takes images of the fluorescently labelled nucleotides. The non-incorporated nucleotides are washed away and the dye and terminal 3' blocker chemically removed so the next cycle can start. Source of the image https://bitesizebio.com/

- Combinatorial probe anchor synthesis (Drmanac et al., 2010): an improved version of the anchor ligation technology allowing longer read lengths, reaction time reductions and faster time to results. This method denatures the DNA to form a single strand DNA circle with each strand. DNA is amplified and folds upon itself to produce a three-dimensional DNA nanoball. In this way many artefacts caused by PCR during amplification can be avoided. DNA nanoballs are fixed to a flow cell and sequencing carried by addition of an oligonucleotide probe that attaches in combination to specific sites within the nanoball. The probe acts as an anchor that then allows one of four single reversibly inactivated, labelled nucleotides to bind after flowing across the flow cell.
- SOLiD (Valouev et al., 2008): sequencing by ligation technology. All possible oligonucleotides of a fixed length are labelled according to the sequenced position and oligonucleotides are annealed and ligated. Then emulsion PCR is used to amplify the samples and the resulting beads are sequenced.
- Ion Torrent semiconductor sequencing (Rusk, 2011): improved version of the chemical sequencing but using a semiconductor based system. The sequencing is carried by measuring hydrogen ions released during the polymerisation of DNA.
- Nanopore sequencing (Clarke et al., 2009): revolutionary method that allows us to sequence complete single DNA strands by passing them through a nanopore which changes its ion current depending on the shape and size of the molecule passing through. Depending on the charge change, the sequencer identifies the molecules passing through.

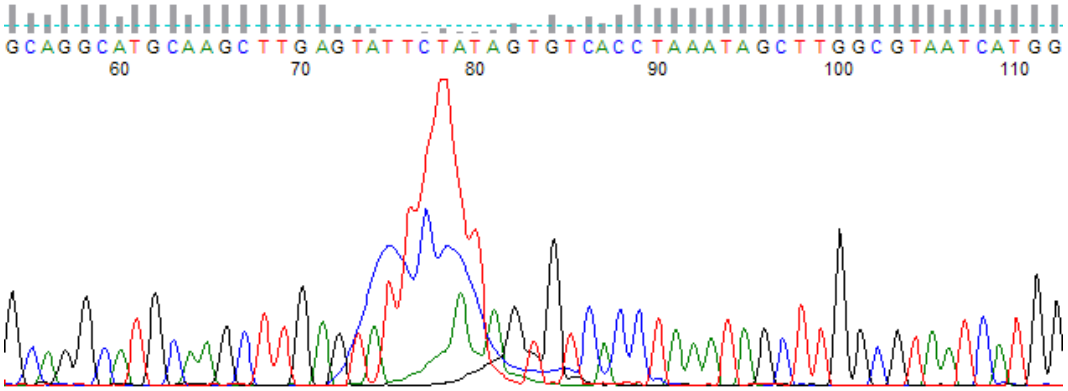
There are more methods than the ones listed here, and many more in development that will come in the future. The possibilities and applications these and the upcoming sequencing methods will bring us are unimaginable and probably will revolutionise our world, again.

While the previous methods can do a whole genome sequencing (WGS), sometimes that amount of data is not needed and other types of sequencing are carried away to reduce time and cost of the experiment. One of the most popular in humans is whole exome sequencing (WES). WES aims to sequence only the protein-coding genes of a genome, the exome, and this is achieved by capturing the genomic regions before sequencing the DNA. The capture can be done by different target-enrichment strategies, but taking into account that humans have only about 180000 exons that account for 1% of the human genome, the cost-efficiency of this technique is irrefutable (S. B. Ng et al., 2009).

1.1.2. Sequence quality

Despite the improvement of sequencing methods, there is no error-free technique. With the old sequencing methods this was a completely different issue; there was just one sequence and one probability per base (Figure 1.3A), so even in the cases where a base was not 100% clear it was possible to solve it. But with the HTS methods we have hundreds of reads covering each base, and those reads can come from different cells with genomic variants; so the question is no longer only what is the correct base at each position (Figure 1.3B) but which ones are the right ones and which can be sequencing errors (Figure 1.4).

A. Sanger Sequencing Quality



B. HTS (Illumina) Sequencing Quality

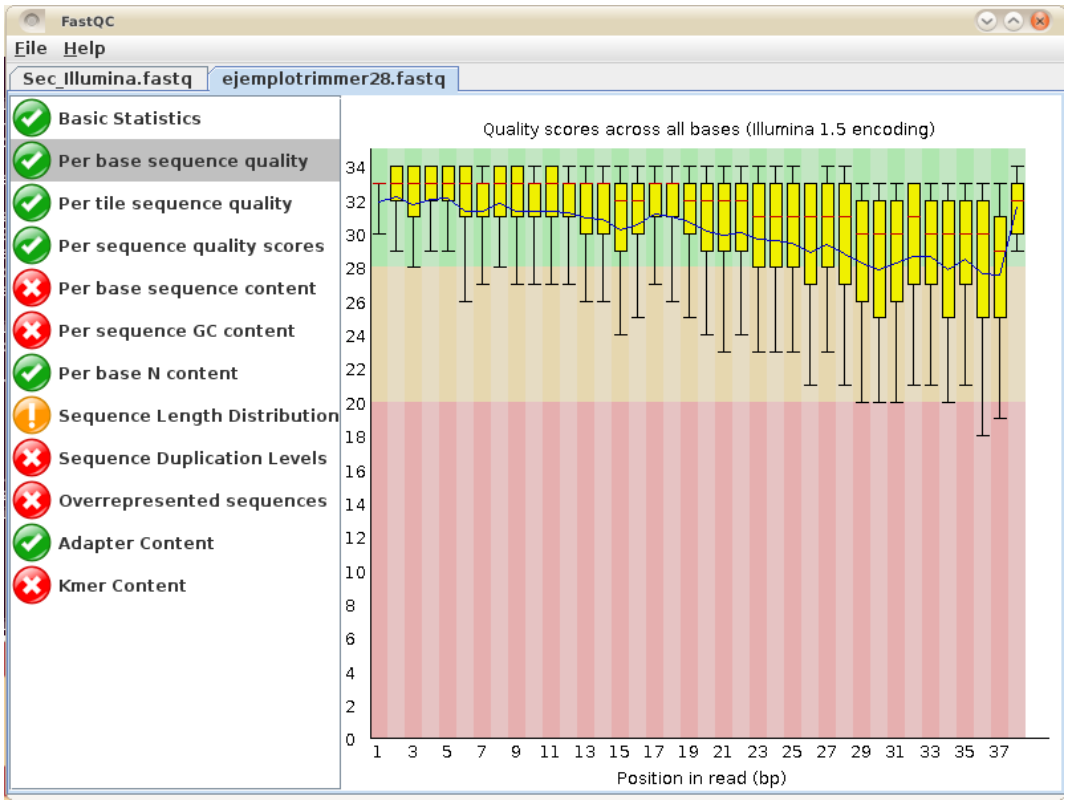


Figure 1.3. Sequence quality per base. Each sequenced base has a probability error score associated with it. Upper part of the figure (A) corresponds to Sanger sequencing, while the lower part (B) corresponds to HTS (Illumina). A) For Sanger sequencing, the error was calculated from the height of the intensity wave used to

identify the nucleotide base at that position of the sequence; the higher the wave and less interferences with others, less likely to be an error. B) In HTS we have millions of sequences at the same time, and every one of them which covers a nucleotide base is used to identify it, and therefore their individual errors are taken into account. In the plot we can see a box plot of the quality per base of each of the reads, the higher the bar the higher its Phred score. More complex statistics (listed in the left side of the image) can be used apart of the Phred score, and they are useful to identify sequencing errors and artefacts, and even contaminations in the sample.

The Phred quality score (Ewing et al., 2005) has been used since the late 90s as a measure of the quality of each sequenced nucleotide. Phred quality scores not only allow us to determine the accuracy of sequencing and of each individual position in an assembled consensus sequence, but it is also used to compare the efficiency of the sequencing methods.

Phred quality scores Q are defined (Ewing et al., 2005) as a property which is logarithmically related to the base-calling error probabilities P. The Phred quality score is the negative ratio of the error probability to the reference level of $P = 1$ expressed in Decibel (dB):

$$Q = -10\, log_{10}\, P$$

or

$$P = 10^{\frac{-Q}{10}}$$

A correct measuring of the sequencing quality is essential for identifying problems in the sequencing and removal of low-quality sequences or sub sequences.

Conversion of typical Phred scores used for quality thresholds into accuracy can be read in the following table (Table 1.1):

Phred score	Error probability	Accuracy
10	1 / 10	90%
20	1 / 100	99%
30	1 / 1000	99.9%
40	1 / 10000	99.99%
50	1 / 100000	99.999%
60	1 / 1000000	99.9999%

Table 1.1. Phred score. Equivalences between Phred score, error probability and accuracy.

There are multiple software to read and generate statistics to help with the interpretation of the quality of a sequence. One of the most commonly used methods for this task is FastQC (Andrews, 2010), a java program that run on any system and has both command line and graphic interface.

1.1.3. Genome Alignment

As we have seen in previous section *1.1.1 Types of Sequencing*, most sequencing methods split the genome in smaller fragments and amplify them before sequencing. This means that the output of the sequencer is more similar to a giant puzzle than a DNA sequence, with millions of pieces that may be duplicated, others may have not been sequenced, and the ones that match may do it in more than one place or not match perfectly anywhere. To solve the puzzle and align all the genome fragments, we use an aligner.

Most modern aligners can filter out low quality reads and clip ignore low quality reads and clip off adapters. In case it has to be done manually because you want to

have more control on the trimming of reads or use a particular method for clipping the lower quality ends, there are standalone applications that allow you to do it. Trimmomatic (Bolger, Lohse, & Usadel, 2014) is one of the most common ones, it is written in java and includes a trimming method by sliding windows very interesting; while most trimmers remove a fixed amount of bases at the end of the sequence or the remove anything until they found one with a higher Phred score than the given threshold, the window method takes into account the average Phred value in a window of X nucleotides, removing everything until the first window found with a higher average Phred score than the given threshold. This allows a better trimming in many cases, as sometimes an average quality base can be surrounded by bad quality ones form both sides and this is the only way to address this issue.

There are two main scenarios when aligning a sequenced genome: *de novo* alignment and against reference alignment.

In the first one, the only information used for the alignment is the reads obtained from the sequencing, which demands a high depth of sequencing and long computing times. De novo genome assembly consists in taking a collection of short sequencing reads and reconstruct the genome sequence, source of all these fragments. The output of an assembler is decomposed into contigs: contiguous regions of the genome which are resolved, and/or scaffolds: longer sequences formed by reordered and oriented contigs with positional information but without sequence resolution.

The outcome may vary depending on the methods used for the alignment, and even realigning with the same program and parameters can get a different outcome. The advantage of these methods is the creation of a new genome without the need of a reference to guide the assembly.

Aligning against reference is faster and easier to replicate and compare its resulting genomes. It needs a reference genome, an already aligned genome of the species or at least very similar, and uses it as template to map the sequenced reads over it. For each of the short reads in the FASTQ file, a corresponding location in the reference sequence is determined. A mapping algorithm will locate a location in the reference

sequence that matches the read, while tolerating a certain amount of mismatch to allow subsequence variation detection that correspond to the actual difference between the reference and de assembled genome (Figure 1.4).

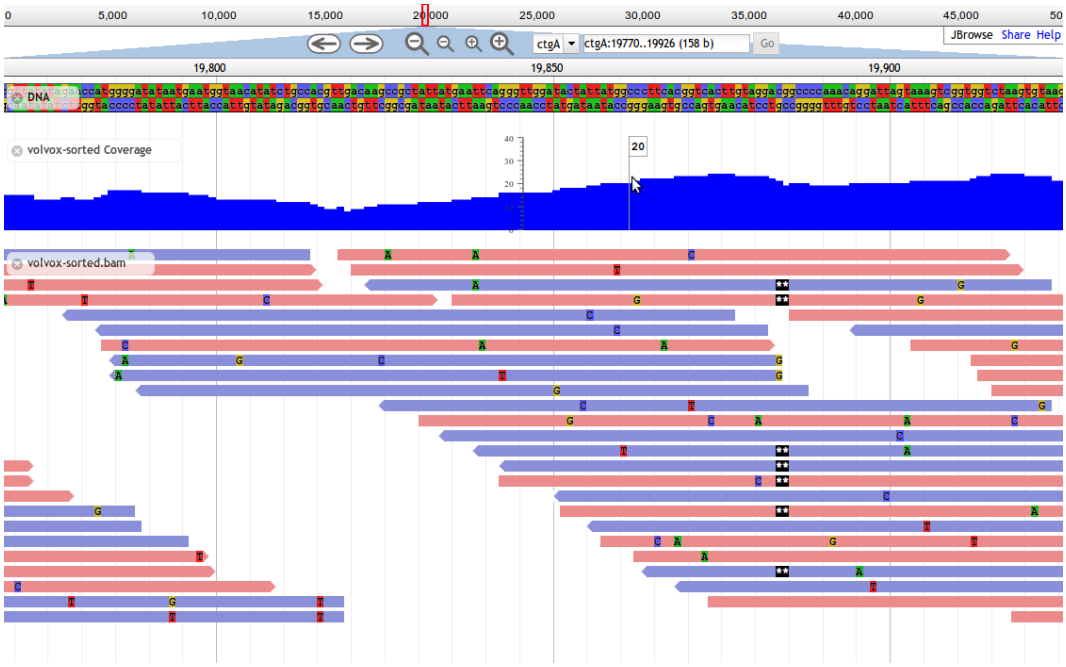


Figure 1.4. Example of mapped reads. This is the visual representation of the result of mapping reads against a reference genome. At the top of the figure the reference genome is displayed (coordinates we are browsing and both forward and reverse strands). Below it there is a blue histogram which represents the depth of coverage; this is how many reads are covering each position of the genome. Finally, we can see lots of horizontal red and blue arrows, which represent the individual forward and reverse mapped reads. The coloured blocks inside them are differences of each read not matching the reference genome; they could be variants or sequencing errors, among others, and here the depth of coverage, the amount of times they are repeated among the reads and the individual Phred scores at that position are key to identify them.

It is clear the outcome will heavily depend on the reference chosen and can be only used when working in already known genomes, but is faster and genomes obtained from aligning against the same reference are easily comparable, even if different methods were used to align them.

There are multiple aligners, so I will introduce now only the one used in the experiments described in this Thesis:

Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2010) is an improved version of the Burrows-Wheeler algorithm which allows alignment of long reads of up to 1Mb and is fast, accurate and highly memory efficient. In our analysis, we combine it with the pre-processing recommended by Genome Analysis Toolkit (GATK) (DePristo et al., 2011). Integrated in this package there are powerful tools for genome analysis used by projects as 1000 Genomes and The Cancer Genome Atlas.

1.2 Genetic variation

The genome contains all the information to build and maintain a living organism. But individuals are unique and species evolve, and therefore the genome has to change accordingly to allow this phenotypic variation. In order to predict the effect a genetic change cause, we need to first understand how the genome encode the information and regulate its expression.

Let's remember it was only a century ago when proteins were believed to encode genetic information, until Griffith suggested DNA carried genetic information in 1927 (Griffith, 1927). It was not until 1951 that Crick, Watson and Franklin discovered its secondary structure (Watson & Crick, 1953) and we still needed to wait until 1961 for Nirenberg and Matthaei to crack the primary genetic code (Matthaei & Nirenberg, 1961). Basically all the knowledge we have about the genome has been discovered in less than half a century, from the first genome sequenced in 1976 to the present day. And despite the impressively fast development of genetics in the last decades, the genome is still vastly unknown.

The variations the genome suffers are based on random mutations. Mutations are un common events, and in most cases are neutral or deleterious, but sometimes a new phenotype is created.

1.2.1. Types of Genetic Variations

Mutations can be classified in different groups:

- Single nucleotide variants (SNVs): a single nucleotide is replaced by other in the genetic sequence. These mutations can be classified in two subgroups: synonymous SNVs, when the codon changes into a synonym codon and therefore the translated protein sequence is not affected; and non-synonymous SNVs, when the nucleotide change affects the protein sequence encoded, replacing one aminoacid for other or a stop codon. In this last scenario, the mutation is called nonsense.
- Insertions and deletions (INDELs): one or more consecutive nucleotides are removed o inserted in the genetic sequence. In case they are short and affect the reading frame, they are also called frameshift mutations.
- Copy number variations (CNVs): a genomic region is amplified several times in the genome.
- Structural variants (SVs): structural change of the genome caused by changes in larger portions of the genome sequence, which in turn causes a change of chromosome assembly.

Both coding and non-coding regions of the genome can be affected by genetic variation. However, due to our limited understanding of the role of non-coding regions (Birney et al., 2007; Feingold et al., 2004), predicting the effects of variations in those regions remains a challenge. The exception is those variants in non-coding regions located in known regulatory regions, where more information is available.

SNVs that occur frequently in a population are considered benign and are known to as single nucleotide polymorphisms (SNPs). A SNV is considered frequent when its Minor Allele Frequency (MAF) > 1%, i.e. when more than 1% of the genomes of that species contain the mutation. As they are common, they are associated to population genetic variance and usually they are not functional.

On the other hand, SNVs that do not occur frequently (MAF < 1%) are usually called rare variants. Before the emergence of deep sequencing methods, it was

difficult to know if these variants were real or just sequencing errors, but nowadays it has been confirmed that each individual has many of them (Nelson et al., 2012; Tennessen et al., 2012). These variants have been observed to be population specific, geographically clustered and most interestingly they are more likely to be functional. They are particularly enriched in the coding regions of protein ligand binding and active sites and involved in hydrogen bonding,

To study genetic variation in any species, the first need is a reference genome. Second, but not less important for determining genetic variability in a population, a database of sequenced genomes and common mutations of that population.

1.2.2. Human Genetic Variation Databases

The first human reference genome was published in 2001 as result of the Human Genome Project (Lander et al., 2001). This project was an international effort started in 1990 and cost $3-billion. And if the sequencing methods had not evolved so quickly, it would have taken even more time and money to complete. It was thanks to the emergence of shotgun sequencing and NGS methods that it could be completed in that time.

After the first human genome was released, the HapMap Project (Consortium, 2007; International & Consortium, 2003) was launched in 2003 with the objective of obtaining the human haplotype, a combination of alleles within a region of each chromosome, and making it available to the scientific community. This data would be used to understand the roles of SNPs and other genetic variants in drug response and how they organise across the different chromosomes.

The first phase of HapMap ended in only three years and identified more than 1 million SNPs using 269 genomes. In the second phase, released in 2007, 3.1 million SNPs were reported in 270 individuals. The third phase finished in 2010 and published 1.6 million of common SNPs in 1184 individuals, and was released as HapMap3 as an integrated data set of both common and rare alleles.

In 2008 started another ambitious sequencing project, the 1000 Genomes Project (Abraham et al., 2015; D. M. Altshuler et al., 2012; Wood et al., 2013). The main goal of this project was the identification of human polymorphisms with MAF > 1%. The sequencing methods used for the project vary between phases, and both WGS and WES were combined.

In 2010 the first phase was complete and nearly 15 million SNPs, 1 million INDELs and 20,000 structural variants were identified. They also reported that each genome had between 250-300 function SNPs and between 50-100 variants associated with inherited disease (D. L. Altshuler et al., 2010). The second phase finished in 2012 with a total of 1092 sequenced genomes from across 14 different populations. It identified over 38 million SNPs and 1.4 million INDELs, and removed over 1.7 million low quality SNPs from the first phase. The third phase ended in 2015 and reported over 68,000 SVs in 2504 unrelated individuals coming from 26 different populations.

The final outcome of the project was identification of 88 million mutations, of which 84.7 million were SNPs, 3.6 million were INDELs and over 60,000 were structural variants. 762,000 variants of the total were rare. The 1000 Genomes Project is now publicly available and under the administration of the International Genome Sample Resource (IGSR), which forms part of EMBL-EBI. The project is under constant expansion as new genomic data keeps being incorporated.

But the 1000 Genomes is no longer the biggest database of human genetic variance (Figure 1.5), as it is only a part of the Genome Aggregation Database (gnomAD)(Lek et al., 2016), which includes 123,136 exomes and 15,496 genomes from unrelated individuals. In its first release in 2016, gnomAD was called the Exome Aggregation Consortium (EXAC) and contained only exome sequencing data.

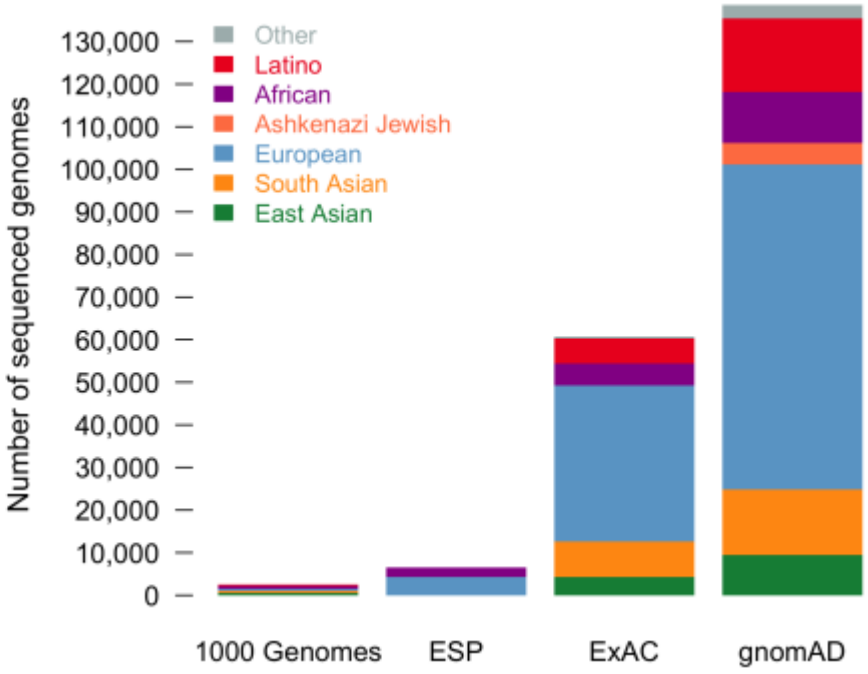


Figure 1.5. Biggest human genome sequencing projects. Number of sequenced genomes in the four biggest human sequencing projects: 1000 Genomes, Exome Sequencing Project (ESP), Exome Aggregation Consortium (ExAC), and Genome Aggregation Database (gnomAD).

The variants discovered in these sequencing projects have not only been listed, but also annotated. These annotations can be found in different databases. One of them is dbSNP (Sherry, 2001) which contains over 150 million referenced SNPs (RefSNPs) and 538 million submitted SNPs (subSNP). dbSNP is founded by the National Centre for Biotechnology Information (NCBI) and apart from human variants it also collects variants of 53 other different species.

Another database of human variants is Clinvar (Harrison et al., 2016). This one contains medically relevant variants, which are human variants phenotypically associated with disease with supporting evidence.

1.2.3. Ebola Genetic Variation Databases

There was not a centralised genomic repository for Ebolaviruses genomes during the last outbreak, despite most of them were uploaded to UCSC databank. We used Filovir (Filovir) and their viral genomes searching tool to download all 196 complete Ebolavirus genomes sequenced at the time and create our own local genetic variance database to work with.

1.3. The genotype to phenotype relationship

Once the genetic variance of an organism is known, it is possible to study the relation between genotype and phenotype, i.e. how genetic variants are associated with particular traits. This is especially interesting when the traits are related with a disease.

Genetic diseases can be classified in two groups:

- Mendelian or monogenic diseases: variants in a single gene are responsible for the disease,
- Complex diseases: many genes are involved in the development of the disease, and not always the same mutated genes are needed to develop it.

Despite all the available genomic data, it remains unclear how much heritability accounts for in the development of genetic diseases (Eichler et al., 2010). Even in deeply studied complex diseases, such as some cancers, it is yet not possible to accurately predict the predisposition of the patients from their genetic variants. Neither it is possible to know from just the patients genome how many of those mutations are due to heritability (Lippert et al., 2013; D. J. Liu & Leal, 2012; Zuk et al., 2014).

1.3.1. Personalised Medicine

Genetic variants not only affect our predisposition to suffer a disease, but also how our organism responds to drug-based medical treatments. Personalised Medicine,

also called Precision Medicine (Katsnelson, 2013; Peterson, Doughty, & Kann, 2013), aims to use genetic information of the patient to calculate the predisposition to a disease and to design an optimal medical treatment individually adapted to that specific patient.

Precision Medicine can be a useful tool for preventive treatments, as it could diagnose a disease year before the patients show their first symptoms. An example of this is the identification of BRCA1 and BRCA2 mutations for breast cancer, where women may choose preventative measures as they have a high risk of developing the disease (Levy-Lahad, Lahad, & King, 2014).

Once the disease has been already diagnosed, Precision Medicine can be also used to help choosing the better treatment for the patient's genetic characteristics (P. C. Ng, Murray, Levy, & Venter, 2009). For example, the use of targeted molecules to treat myeloid leukaemia, by overcoming AML cell resistance to drug therapy (Gojo & Karp, 2014).

Another application of Precision Medicine is in the field of Pharmacogenomics, studying the effect of genomic variants effect in an individual's drug response (Karczewski, Daneshjou, & Altman, 2012; L.Hopkins & R.Groom, 2005).

1.4. How genetic variation leads to altered phenotype

Every single genetic variation may cause an effect in the organism. For years synonymous SNVs were considered innocuous as they do not cause any changes in the translated protein sequence. However, recent studies have reported positive selection of synonymous SNVs in cancer genomes and the theory that synonymous variants can be functional has been proposed (Supek, Miñana, Valcárcel, Gabaldón, & Lehner, 2014). They are believed to play a role in regulatory regions, and affect mRNA translation speed and protein folding (Buske, Manickaraj, Mital, Ray, & Brudno, 2013). But due to the lack of solid proof for these hypotheses and understanding of the effects of synonymous SNVs, we will consider non-synonymous SNVs as the only type of this variation with an effect in phenotype.

1.4.1. Analysis of variants associated with disease

Genome positions which are more likely to cause a disease if mutated are known to be less variable than neutral ones (M. Kumar, Joseph, & Chandrashekaran, 2001). For this reason, functional variants are evolutionary conserved not only across populations but also across species. This phenomenon has made sequence conservation one of the most important clues used by bioinformatics tools to identify functional residues in protein sequences.

Apart from conserved functional residues, we can find fairly conserved regions across species which encode for the same protein and descend form the same ancestor gene. These regions are called orthologues, and are also useful for bioinformatics tools designed to predict the effect of variants.

When translated into protein sequence, most of the diseases causing aminoacid changes have been reported to appear in the protein core (Burke et al., 2007). It is in the core where a single change is more likely to affect protein structure, and therefore its functions.

Using the HumVar database of variants (David, Razali, Wass, & Sternberg, 2012; Pundir, Martin, & O'Donovan, 2016) extended these previous structural analyses to consider the role of protein-protein interfaces, mapping variants in protein structures from Interactome3D (Mosca, Céol, & Aloy, 2013). As expected from previous studies, they observed that disease-associated variants tent to be located in the protein core, but also they reported an enrichment of disease-associated non-synonymous SNVs mapped inside protein-protein interfaces. In (Bordner & Zorman, 2013) a similar result was obtained, but this time discovering disease-associated non-synonymous SNVs inside ligand-binding sides.

But structure, protein-protein and protein-ligand interactions caused by genomic variants are not the only known factors that may affect protein function. For example, post translational modifications (PTMs) can alter protein function by two mechanisms: by allosterical conformational changes in the functional site or by

orthosterically influencing (Nussinov, Tsai, Xin, & Radivojac, 2012). Furthermore, some disease-associated variants are known to also affect PTMs sites, and therefore affect protein function (J. Li et al., 2014).

1.5. SNV effect prediction methods

We have previously explained the relationship between variants and disease and introduced some databases of human SNVs containing clinical annotations. The next step is introducing the two most used bioinformatics methods that merge those two factors in order to predict the effect of non-annotated SNVs.

The first of these methods, SIFT (Sorting Intolerant from Tolerant) (P. Kumar, Henikoff, & Ng, 2009), was designed on the principle that mutations occurring in conserved regions are less likely to be tolerated and therefore more likely to be functional. This method uses orthologues to build multiple protein sequence alignments and create a model based on sequence homology. It takes into account sequence conservation, hydrophobic conservation, previously known variants and substitution matrix to predict if an aminoacid change is tolerated. The algorithm returns a probability score indicating the chances of the variant to be functional.

The second method, PolyPhen2 (Polymorphism Phenotyping V2) (Adzhubei, Jordan, & Sunyaev, 2013), incorporates also paralogues to the multiple sequence alignments and many protein annotated features in its Naïve Bayes based machine learning algorithm: sequence annotations from Uniprot and from DSSP, binding annotations (disulphide bonds and covalent links), 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. Then, the machine learning algorithm is trained with the database HumVar. In consequence, this method is capable of predicting if a variant causes gain or loss of protein functions.

Of course, there are more prediction methods following different approaches. But one thing all these methods have in common is that, despite performing well in benchmarking exercises, their results do not agree between methods (Chun & Fay, 2009). For this reason, their predictions have to be taken carefully and only considered as candidates for further study.

1.6. Introduction to Ebola

Ebola virus disease (EVD), formerly known as Ebola Haemorrhagic Fever, is a deadly disease in humans and other primates caused by members of the genus Ebolavirus, with a death rate of up to 90%. Symptoms of EVD include abrupt onset of fever, myalgia, and headache in the early phase, followed by vomiting, diarrhoea and possible progression to haemorrhagic rash, life-threatening bleeding, and multi organ failure in the later phases. There are no treatments for this disease at present time, although experimental vaccines have been tried during recent outbreaks. Ebola virus is therefore an important threat to public health and a bio-threat pathogen of category A. Between 1976-2013 there were 25 verified outbreaks, causing no more than 300 deaths each. In 2014 a much larger outbreak occurred in West Africa, which killed more than 11,000 people, more than six times the cumulative sum of all the previous 24 outbreaks (Gebretadik FA, 2015), and which also saw cases occur outside of Africa.

Ebolavirus is a genus of the family Filoviridae which contains five species (Kuhn et al., 2010): Zaire Ebolavirus (EBOV), responsible of the 2014 outbreak, Bundibugyo Ebolavirus (BDBV), Reston Ebolavirus (RESTV), Sudan Ebolavirus (SUDV) and Taï Forest Ebolavirus (TAFV). There are other two Filoviridae genera: Marburgvirus (Marburg Marburgvirus, MARV) and Cuevavirus (Lloviu Cuevavirus, LLOV).

We will use the nomenclature recommended by Kuhn (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, virus fragments 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).

1.6.1. Background

While four of the five members of the genus Ebolavirus (Zaire Ebolavirus, Sudan Ebolavirus, Bundibugyo Ebolavirus and Taï Forest Ebolavirus) cause haemorrhagic fever in humans associated with fatality rates of up to 90%, Reston viruses are non-pathogenic to humans (Feldmann & Geisbert, 2011) (Weingartl, 2013). This has been documented so far during 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 (USA). All three outbreaks 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 Ebolavirus. Moreover, Reston Ebolavirus was found in 2008 in domestic pigs in the Philippines, and seroconversion was detected in six human individuals. None of the 11 individuals that were seropositive for Reston Ebolavirus antibodies reported an Ebola-like disease (Miranda & Miranda, 2011).

The reasons underlying the differences in human pathogenicity between Reston Ebolavirus 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. Therefore, we performed an in silico analysis of the genomic differences between Reston Ebolavirus and the human pathogenic Ebolaviruses to identify conserved changes at protein level which could explain the differences in Ebolavirus pathogenicity in humans.

Despite the small sized Ebolavirus genome we still have a limited understanding of Ebolaviruses and what causes their pathogenicity and why Reston Ebolavirus is not pathogenic in humans (Basler, 2014; Feldmann & Geisbert, 2011; Zhang et al., 2012). The importance of understanding these differences is highlighted by the 2014-16 Zaire Ebolavirus outbreak in Western Africa, which is the first large outbreak and has resulted in more than 28,600 suspected cases and over 11,325 deaths in its two and a half lifespan years (www.who.int). During this outbreak many additional Ebola Ebolavirus genomes were sequenced, enabling us to perform the first comprehensive comparison of the non-human pathogenic Reston Ebolavirus to all four human pathogenic Ebolaviruses.

While some studies (Bale et al., 2013; Clifton et al., 2014; Zhang et al., 2012) 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.

1.6.2. Ebolaviruses genetics and proteomics

As shown in Figure 1, the EBOV genome is a single negative-sense RNA strand of 18,959 nucleotides in length, containing only seven Open Reading Frames (ORFs). The same genomic structure is shared by the rest of Ebolaviruses in the family. Despite limited encoding capacity, these viruses expand their gene functions by forming more proteins and assigning more functions to each of them. Nine proteins are known to be translated, including nucleoprotein (NP), the polymerase cofactor viral protein (VP35), the major matrix protein (VP40), glycoprotein (GP), soluble glycoprotein (sGP), small soluble glycoprotein (ssGP), transcription activator (VP30), the minor matrix protein (VP24), and viral RNA-dependent RNA polymerase (L). GP, sGP, and ssGP are produced from the GP gene by alternative RNA editing (de La Vega, Wong, Kobinger, & Qiu, 2015; Feldmann & Geisbert, 2011; Mehedi et al., 2011). The 3' terminus is not polyadenylated and the 5' end is not capped. The gene order is 3' – leader – NP – VP35 – VP40 – GP/sGP/ssGP – VP30 – VP24 – L –

trailer – 5'; with the leader and trailer being non-translated regions, which carry important signals to control transcription, replication, and packaging of the viral genomes into new virions. 3

Many of the Ebolavirus proteins have multiple functions. The virion, represented in Figure 1.6, which protects the RNA genome is formed by helically arranged viral nucleoproteins NP and VP30, which are linked by matrix proteins VP24 and VP40 to the lipid bilayer that coats the virion. VP35 is a multifunctional dsRNA binding protein that plays important roles in viral replication, innate immune evasion, and pathogenesis. The multifunctional nature of VP35 and VP24 also presents opportunities to develop countermeasures antagonise the cellular interferon response. GP is responsible for the virus' ability to bind to host cell and virus internalisation (Basler, 2014; Feldmann & Geisbert, 2011). The NP-encapsulated RNA genome associates with VP35, VP30, and L to form the transcriptase-replicase complex. 1,3 Little is known about the functional roles of the secreted proteins sGP and ssGP (Feldmann & Geisbert, 2011; Hoenen et al., 2015; Mehedi et al., 2011; Miranda & Miranda, 2011).

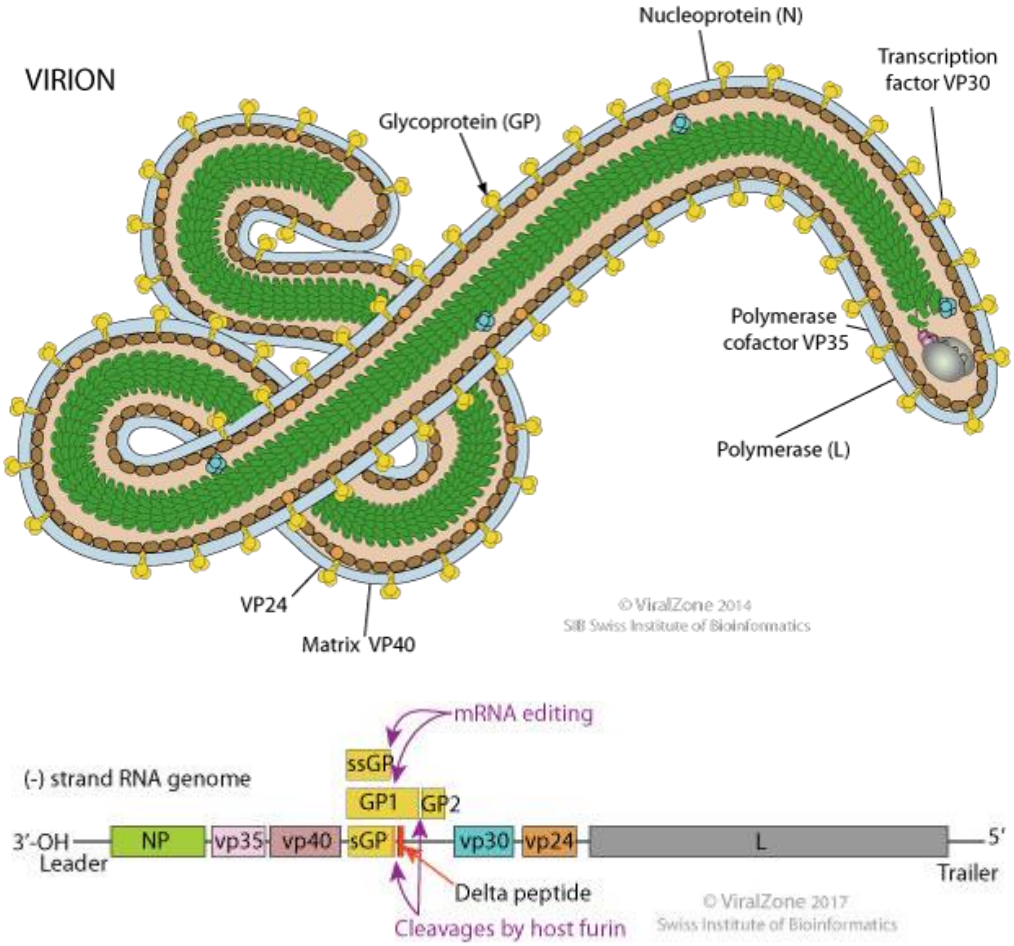


Figure 1.6. EBOV virion and genome. EBOV virion (above) is a filament of 970 nm long and its diameter is about 80nm. EBOV genome (below) consists in a single negative RNA about 18-19 kb in size which encodes for seven genes: NP, VP35, VP40, GP, VP30, VP24 and L. GP produces different protein products due to mRNA editing: GP, sGP and ssGP. Images taken from ViralZone (Hulo et al., 2011) (https://viralzone.expasy.org/207?outline=all_by_species)

1.7. Introduction to neuroblastoma

Neuroblastoma is the most common extracranial solid cancer in infancy, affecting up to one in every 7000 children (Maris, Hogarty, Bagatell, & Cohn, 2007). It is a neuroendocrine tumour cancer that affects neuroblasts, immature nerve tissue cells. It usually starts in the nerve tissue of the adrenal glands (40% of cases), but can also develop in nerve tissues in the abdomen (30%), chest (19%), neck (1%), spine (1%), or pelvis (1%) (Friedman, 2007). Nearly half of neuroblastoma cases occur in children younger than two years, and almost all the cases which develop metastasis occur before six years.

As other cancers, neuroblastoma is a complex disease which causes include both genetic and environmental factors. Mutations in ALK, PHOX2b and KIF1B(Mossé et al., 2008), MYCN amplification (Brathwaite, Wolman, Dalla-favera, Simon, & Gallo, 1984), duplicated segments in LMO1 (Wang et al., 2011), and copy number variation of NBPF10 (Diskin et al., 2009) have been linked to neuroblastoma. Among the risk factors believed to be related with the disease we can divide them in two groups: risk factors the parents were exposed during pregnancy and prior conception, and diseases in early life. The studies about the impact of typical environmental factors like toxic chemicals, radiation, smoking, alcohol, medical drugs and birth factors, but studies about the impact of these factors have been inconclusive (Olshan & Bunin, 2000). Other factors which impacts in neuroblastoma are under research are hormone based treatments and fertility drugs (Olshan et al., 1999), maternal hair dye (McCall, Olshan, & Daniels, 2005), and atopy and early life infections (Menegaux, Olshan, Neglia, Pollock, & Bondy, 2004).

1.7.1. Symptoms, diagnosis and treatment

Symptoms of this disease are broad and vary depending on tissue of origin and the presence of metastasis, making its diagnosis difficult. (Wheeler, 2015).

The cancer is divided into low-, intermediate-, and high-risk groups based on a child's age, cancer stage, and tumour morphology. Different treatments are available

depending on the risk group of the patient. Low-risk neuroblastoma can be cured with surgery, and the cure rate is above 90%. The intermediate-risk variety needs to be treated with a combination of surgery and chemotherapy, but the cure ratio is still high, between 70-90%. High-risk neuroblastoma remains a challenging disease, with a treatment consisting in a combination of surgery, intensive chemotherapy, radiation therapy, bone-marrow transplant and antibody therapy. The cure rate of high-risk neuroblastoma remains at 30-60% (Castleberry RP, 1991; Bowman LC, 1997); Castleberry RP S. J., 1992; West DC, 1993; Paul SR, 1991).

1.7.2. Cancer cell lines as model system

Cancer cell lines are immortalised cells originally taken from a patient's tumour that can be continuously cultured in a laboratory. In a cell line, cells from a multicellular organism are mutated so they can proliferate indefinitely. In cancer tissues these mutations occur naturally, de-regulating the normal cell cycle controls leading to uncontrolled proliferation, allowing a cell type which would normally not be able to divide to be proliferated in vitro.

Cell lines are widely used in research as model systems (Kaur & Dufour, 2012). Cancer cell lines are important model systems for the study of cancer cell biology and the cancer cell drug response (Sharma, Haber, & Settleman, 2010). Many anti-cancer drugs have been discovered and/or initially characterised in cancer cell lines and cancer cell line panels, such as the NCI60 panel (Holbeck, Collins, & Doroshow, 2010; Sharma et al., 2010; Shoemaker, 2006).

The characterisation of cancer cell lines has also revealed in depth insights into cancer biology. This includes gene networks associated with cancer, mutation and selection processes, and evidence of the DNA damage that triggered carcinogenesis (Pleasance et al., 2010). Moreover, the use of cancer cell lines provides fundamental insights into cancer cell plasticity and its relevance for the cancer cell response to anti-cancer drugs (Eirew et al., 2015; McGranahan et al., 2015; Sharma et al., 2010).

One limitation for this model system is the genetic drift that may occur over multiple passages, leading to genetic differences in isolates and potentially different experimental results depending on when and with what strain isolate an experiment is conducted (Marx, 2014).

1.8. Organisation of this Thesis

During this introductory chapter we have covered the basis status of nowadays genetic sequencing, the process to transform biological sequences into information, the keys of human genetic variation, and an introduction to both Ebolavirus and neuroblastoma.

This work is organised as follows:

Chapter 2 is derived from one of our papers, "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 3 introduces neuroblastoma cell line UKF-NB-3 and describes its genetic landscape of, and shows its internal heterogeneity when comparing it with single cell derived clones of the cell line.

Chapter 4 continues with the study of the UKF-NB-3 cell line and focus on its internal heterogeneity, comparing the variants in its genome with the ones from UKF-NB-3 single cell derived clones.

Chapter 5 concludes this piece of work putting together a general discussion of the findings in both research lines and proposing future work on the fields.

Chapter 2:
Ebola: genetic variance and its impact in human pathogenicity

This chapter reports on work that was published in Pappalardo M, Julia M, et al., (2016) Scientific Reports 6:23743 Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. I am joint first author on this work, with Morena Pappalardo, we completed much of the research and data analysis together. Further, I performed all of the phylogenetic analyses and bootstrapping of the SDP method individually.

2.1. Introduction

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 signalling. 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.

2.2. Methods

The tools used for this research will be introduced in the sections bellow. More detailed information about each concrete use of them will be explained in the corresponding results section when needed.

2.2.1. Ebolavirus Genome Sequences

We collected a total of 196 complete Ebolavirus genomes (Annex 1: Suppl. Table 20) from the Virus Pathogen Resource (Pickett et al., 2012), consisting of 156 Ebola virus (EBOV), 17 Reston (RESTV), 13 Sudan (SUDV), 7 Bundibugyo (BDBV) and 3 Taï Forest (TAFV) species. These genomes were later scanned for ORFs with EMBOSS (Rice, Longden, & Bleasby, 2000), and the predicted protein sequences were identified by using BLAST against a database created with the protein sequences available per each Ebolavirus in ViPR. We decided to follow this approach instead of directly using the protein sequences in ViPR due to the low effective number of proteins sequences after removing redundancy.

2.2.2. Multiple Sequence Alignments and identification of specificity determination 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 present was calculated for the alignment for each protein by building a Hidden Markov Model (hmm) for the alignment using hmmer (Mistry, Finn, Eddy, Bateman, & Punta, 2013). The effective number of independent sequences identified ranged from 88 for the VP24 and L proteins to 148 in NP (Annex 1: Supp. Table 21).

The S3Det algorithm (Rausell, Juan, Pazos, & Valencia, 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 Figures 2.1–2.3). SDPs were compared to known functional residues (many from mutagenesis studies) in Ebolavirus proteins catalogued in UniProt (Apweiler et al., 2014) and in the literature.

2.2.3. 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 analysing 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 analysed and plotted using the R "ape" package (Paradis, Claude, & Strimmer, 2004).

2.2.4. 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 (Kelly, Mezulis, Yates, Wass, & Sternberg, 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 modelled using one-to-one threading in Phyre2 (Kelly 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 visualised and analysed using Chimera (Pettersen et al., 2004). mCSM (Pires, Ascher, & Blundell, 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.

2.3. Results

Our large scale analysis of 196 different Ebolavirus genomes focussed on combining computational methods with detailed structural analysis to identify the genetic causes of the difference in pathogenicity between Reston Ebolavirus 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 (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 Ebolavirus from human pathogenic Ebolaviruses, arguably, contain within them a set of amino acid changes that explain the differences in pathogenicity between Reston Ebolavirus 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; Palles et al., 2013).

Phylogenetic analyses were performed for the whole genomes and the individual proteins (Annex 1: Supplementary Figure 1).

In accordance with previous studies (Gire et al., 2014; S. Q. Liu, Deng, Yuan, Rayner, & Zhang, 2015; Morikawa, Saijo, & Kurane, 2007), we observed high intra-species conservation with greater inter-species variation, as shown in Figure 2.1 and Annex 1: Supplementary Table 1. The surface protein GP exhibited the greatest variation, most likely as a consequence of selective pressure exerted by the host immune response (S. Q. Liu et al., 2015).

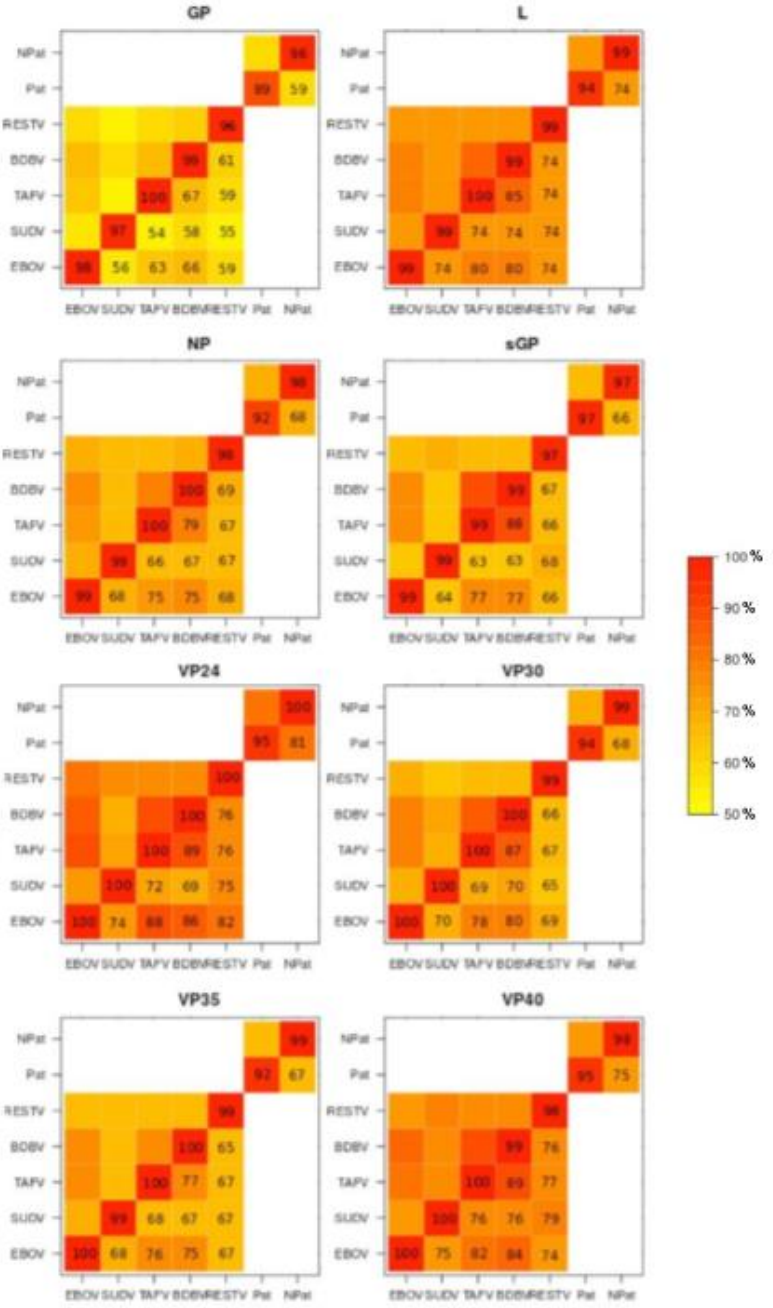
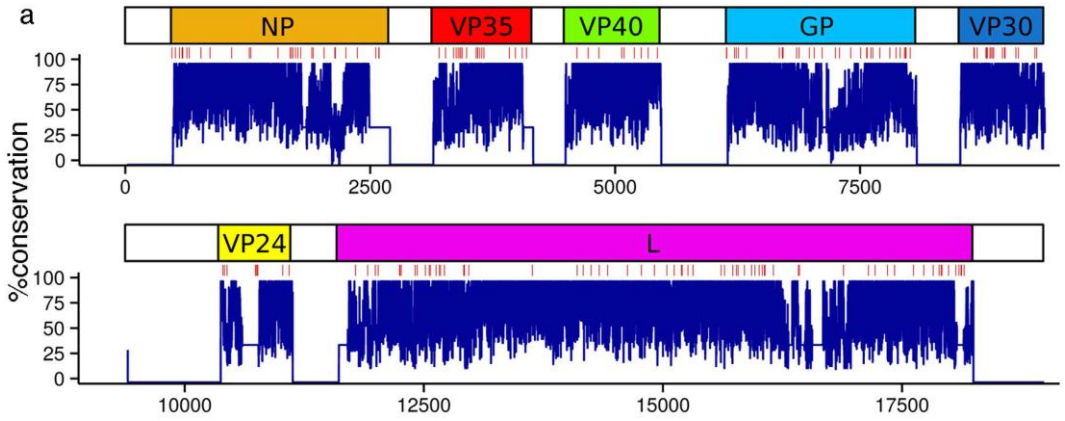


Figure 2.1. Conservation of Ebolavirus proteins. Heatmaps of intra- and inter species percentage of sequence identity for Ebolavirus proteins. Acronyms represent: EBOV, Ebola virus; BDBV, Bundibugyo virus; SUDV, Sudan virus; TAFV, Taï Forest virus; RESTV, Reston virus; Pat, human pathogenic species (EBOV, BDBV, SUDV and TAFV); and NPat, human non-pathogenic species (RESTV).

Using the S3Det algorithm (Rausell et al., 2010) we identified 189 SDPs that are differentially conserved between Reston viruses and human pathogenic Ebolaviruses (Figure 2.2a, Annex 1: Supplementary Figure 2, Annex 1: Supplementary Tables 2–9). These SDPs represent the most significant changes between the Reston virus and the human pathogenic Ebolaviruses so 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 2.2b). Comparison of the SDPs with previously published mutagenesis studies (Xu et al., 2014) provided no explanation for their functional consequences (Annex 1: Supplementary Table 10).



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 2.2. Ebolavirus SDPs. (a) genomic overview of Ebolavirus conservation. An ideogram of Ebolavirus genome and its genes is divided in two fragment, showing SDPs positions (vertical red lines) below it and the percentage of protein

sequence conservation (blue graph) at the bottom of each fragment. (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.

2.3.1. Specificity Determining Positions (SDPs) Analysis

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 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 (Annex 1: Supplementary Table 21).

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 (Figures 2.3–2.5). SDPs were compared to known functional residues (many from mutagenesis studies) in Ebolavirus proteins catalogued in UniProt (Apweiler et al., 2014) and in the literature.

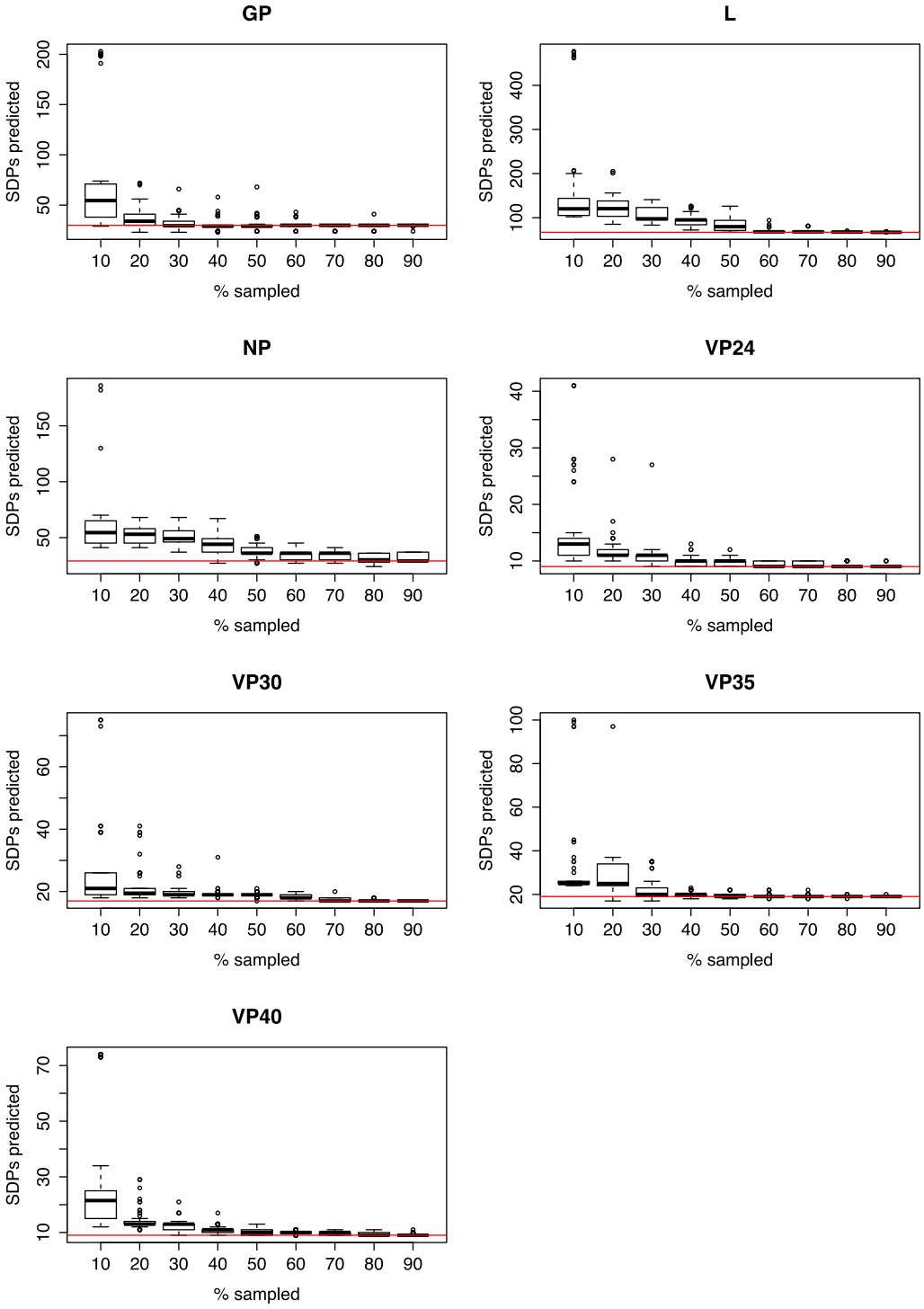
The sensitivity of the SDP analysis to the number of sequences available was considered by subsampling the sequences. Sampling was performed for: only the human pathogenic group; only the Reston group; and for both groups

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.

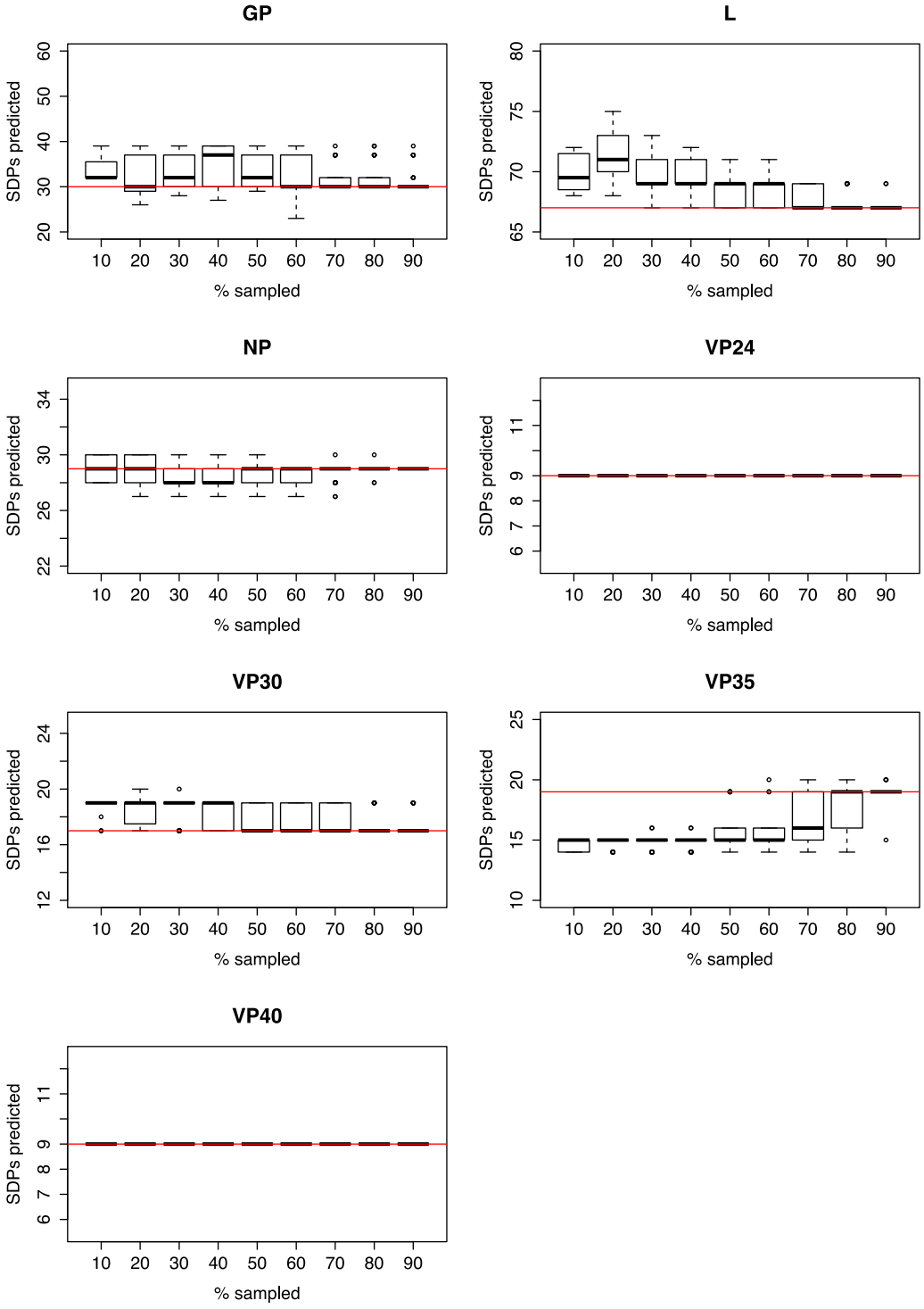
The total number of SDPs predicted when sampled is shown in Figure 2.3. When the sequences of human pathogenic Ebolaviruses were sampled, while the number of Reston sequences remained constant, we observed that the number of SDPs predicted decreased as the proportion of sequences sampled increased (Figure 2.3A). 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 (Figure 2.3B). 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 (Figure 2.3C) we found that the number of SDPs predicted very quickly converged to the number of SDPs present in the full dataset.

A. Human pathogenic sequence sampled.



B. Reston Sequences Sampled



C. Both groups sampled

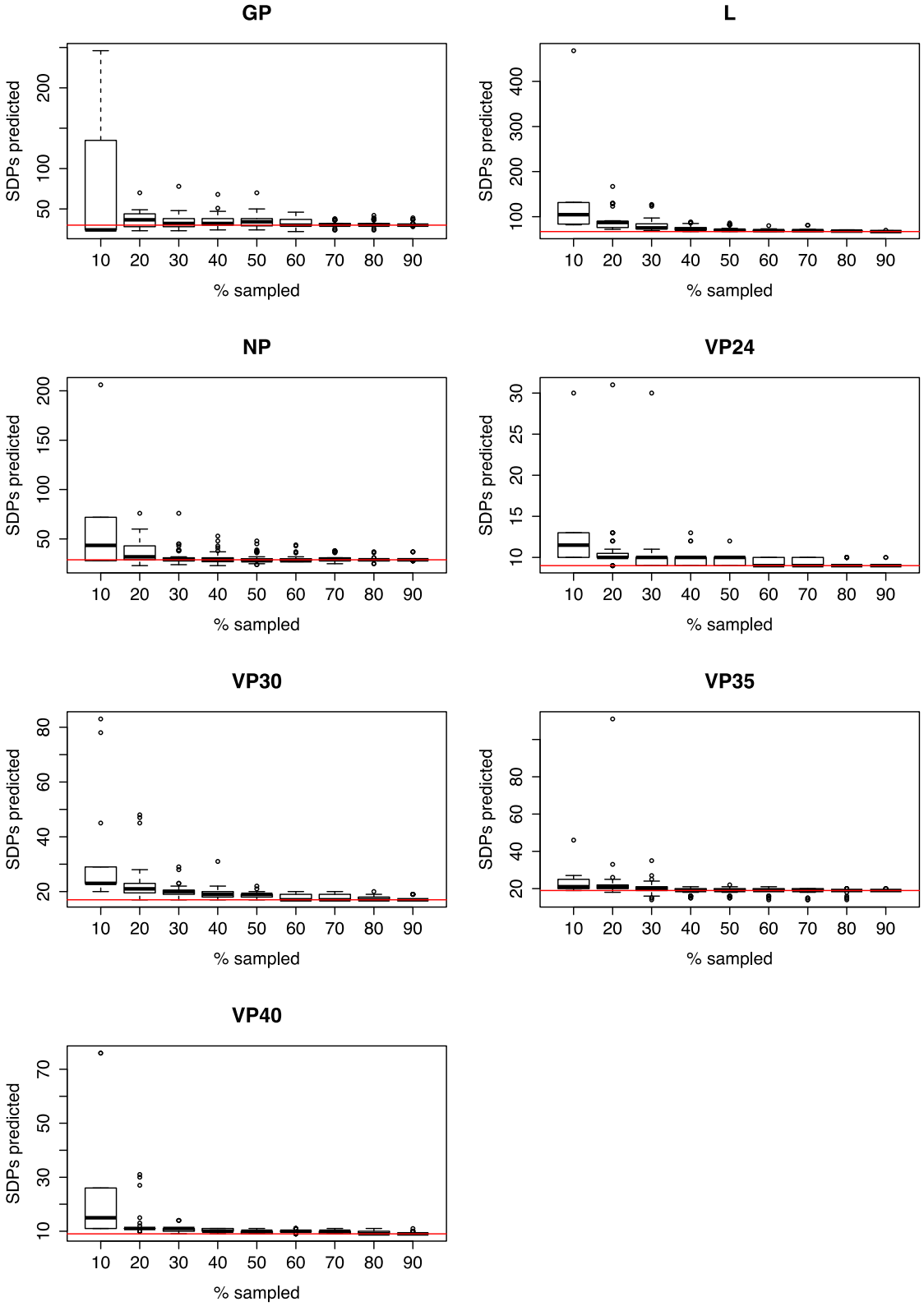
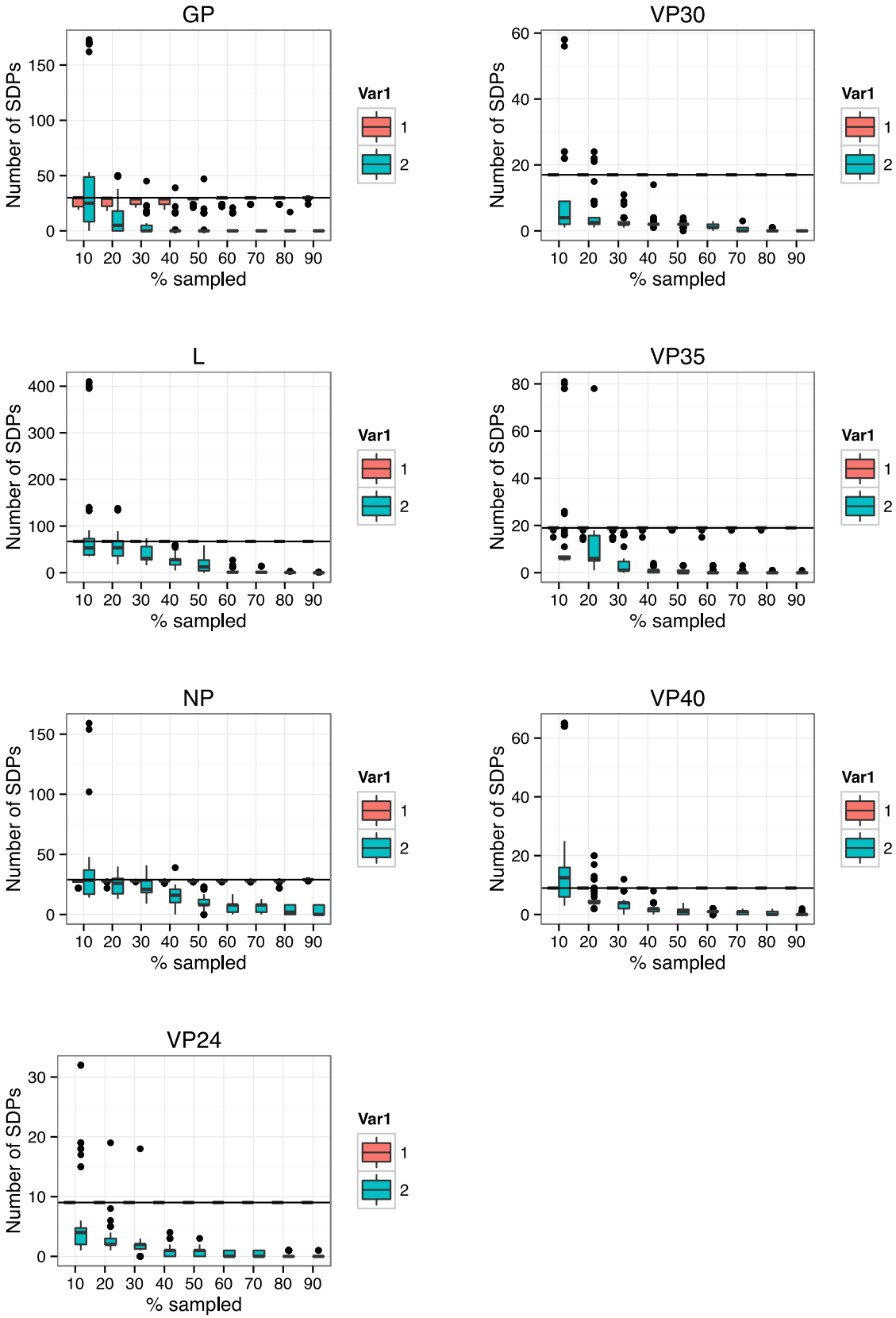


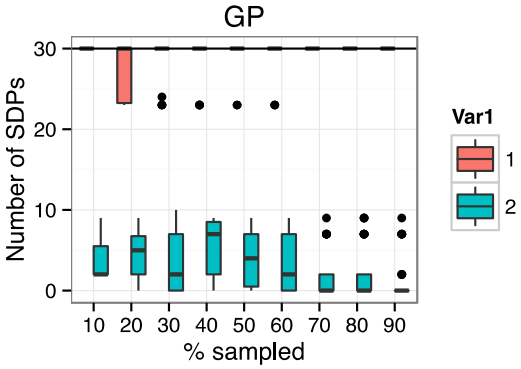
Figure 2.3. SDP prediction with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and Reston ('non-human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the distributions of the number of SDPs predicted in the simulations 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.

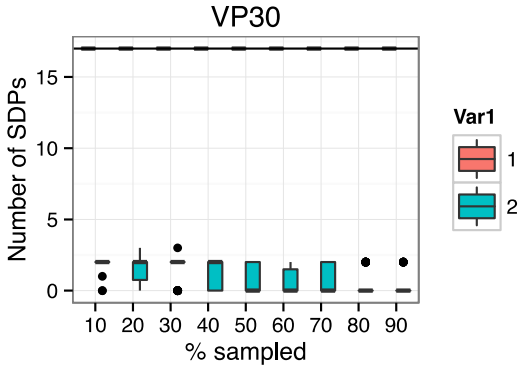
We then considered the number of SDPs predicted that are present in the full dataset and those that are predicted only with subsampled data (Figure 2.4). When the human pathogenic sequences were sampled (Figure 2.4A), we found that the vast majority of SDPs in the full data set were predicted at all 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) was 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 (Figure 2.4B) 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 (Figure 2.4C).

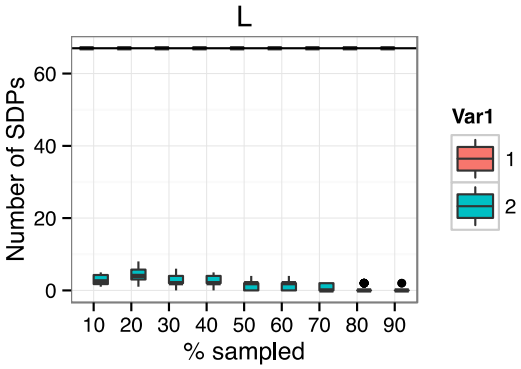
A. Human pathogenic sequence sampled.

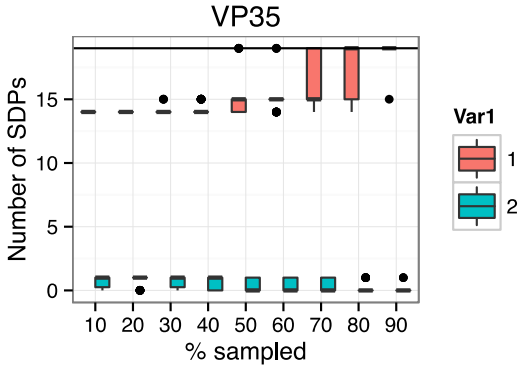


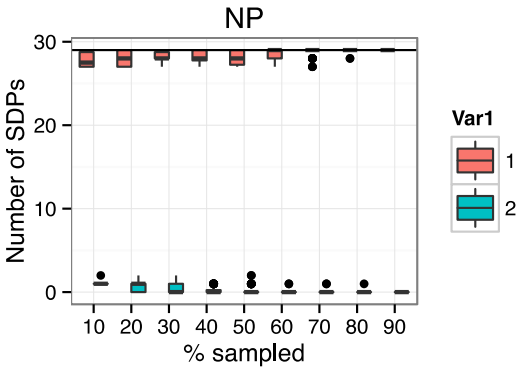
B. Reston Sequences Sampled

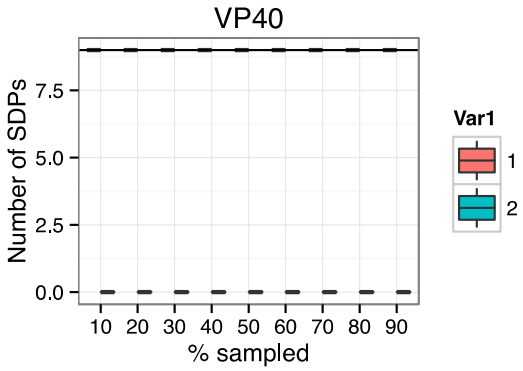


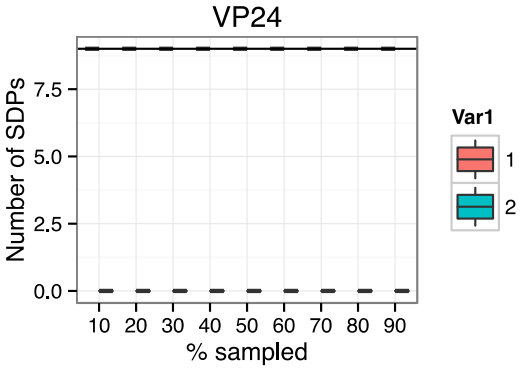




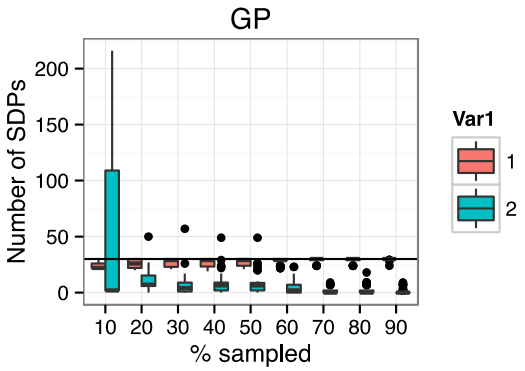


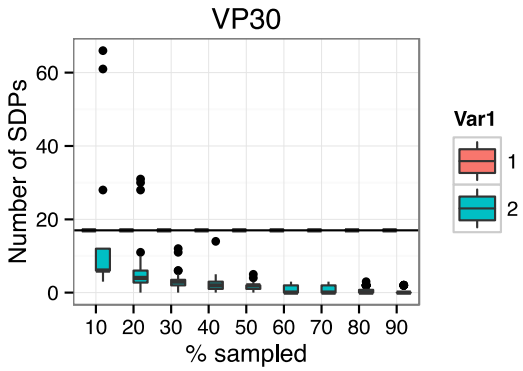


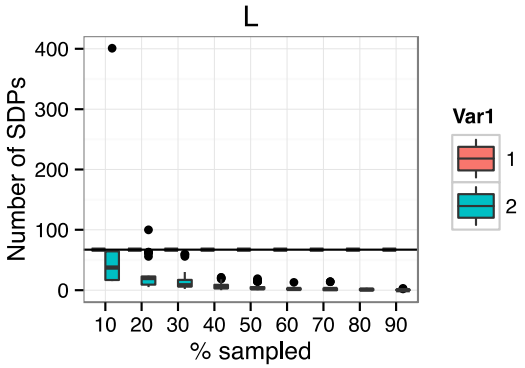


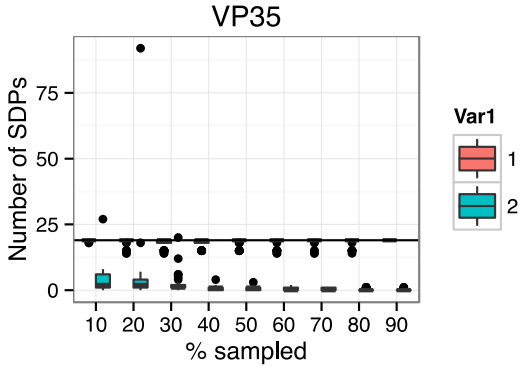


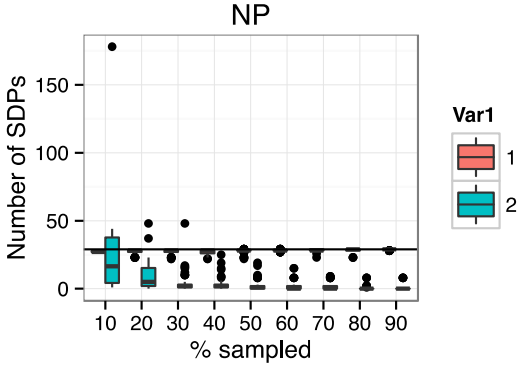
C. Both groups sampled

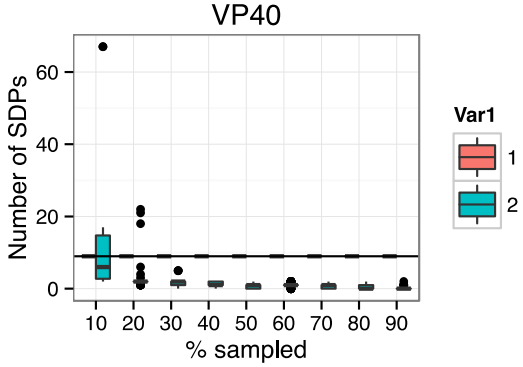












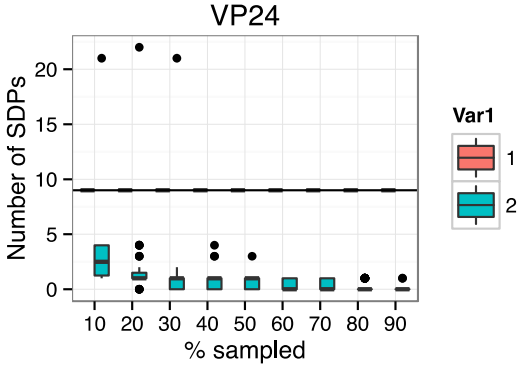
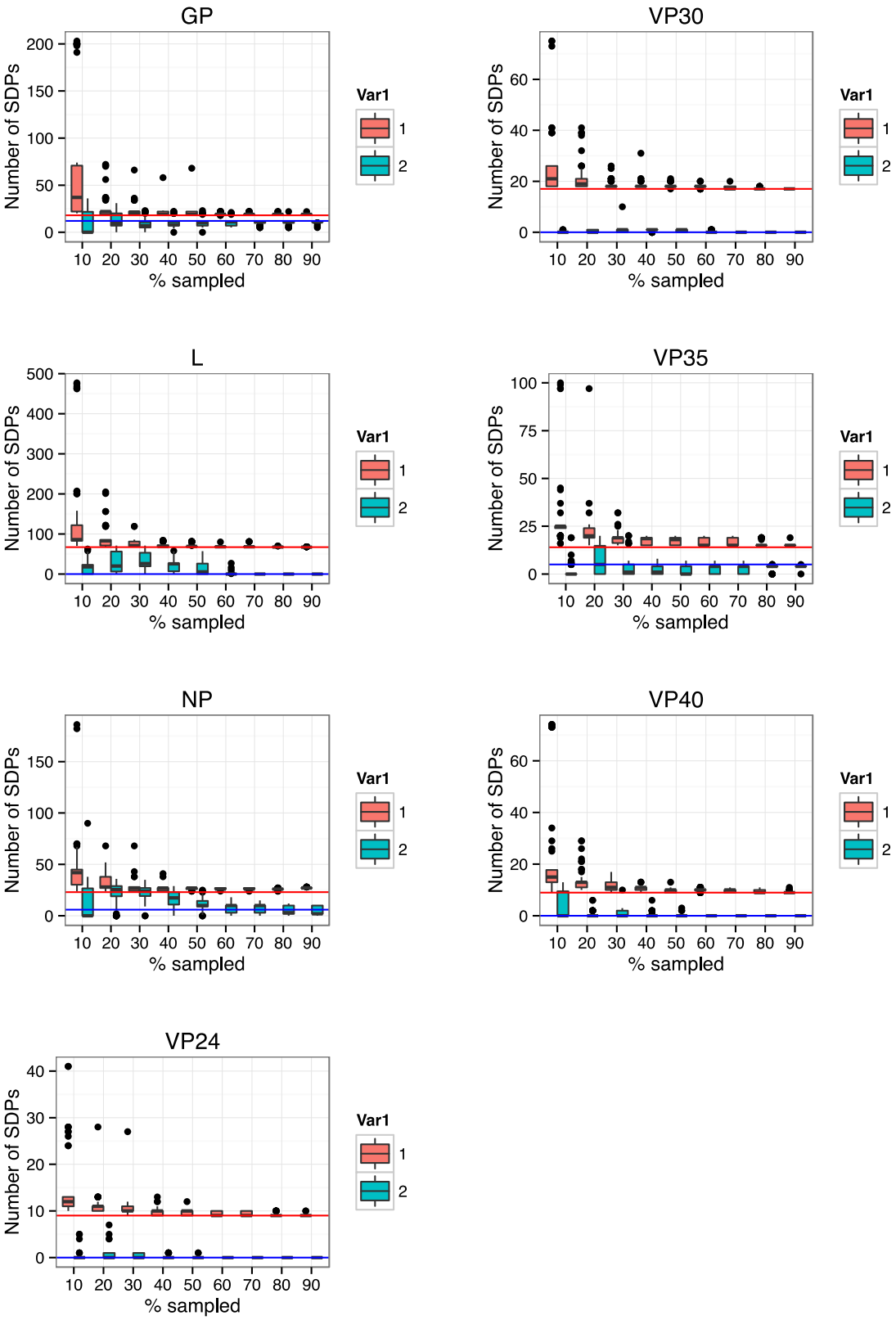


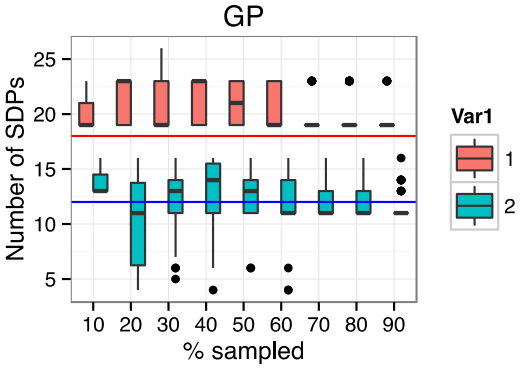
Figure 2.4. Change in SDP prediction with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and Reston ('non-human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the number of SDPs predicted in each sampling that are 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.

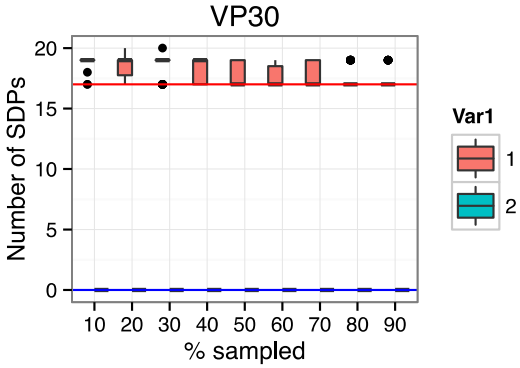
Finally, we considered the number of SDPs in the sampling sets that are completely conserved and those that are not (Figure 2.5). In conjunction with the data from Figure 2.4, this shows that sampling generates new SDPs that are completely conserved (i.e. only one amino acid in each group) and also some 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.

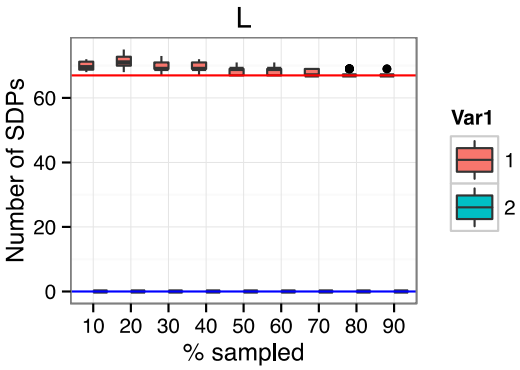
A. Human pathogenic sequence sampled.

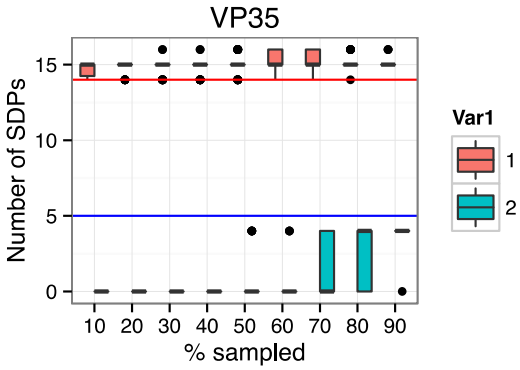


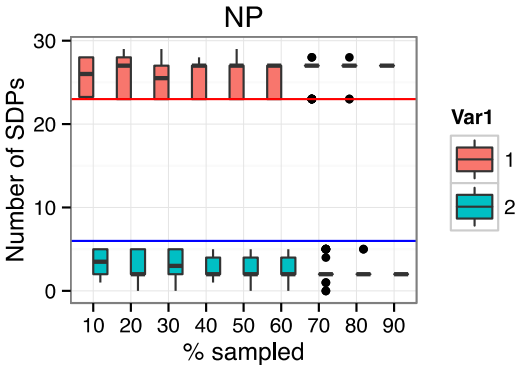
B. Reston Sequences Sampled

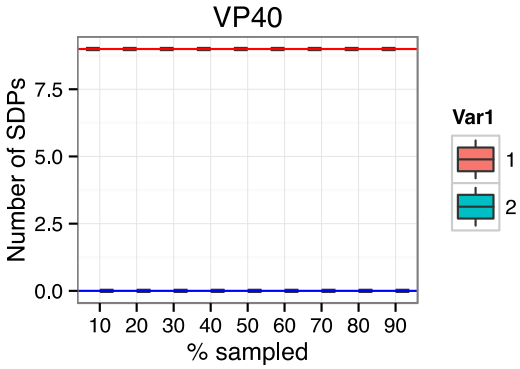


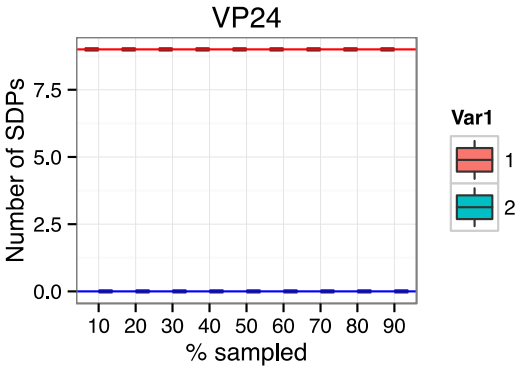




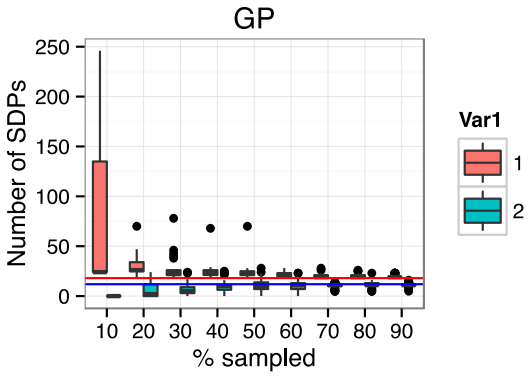


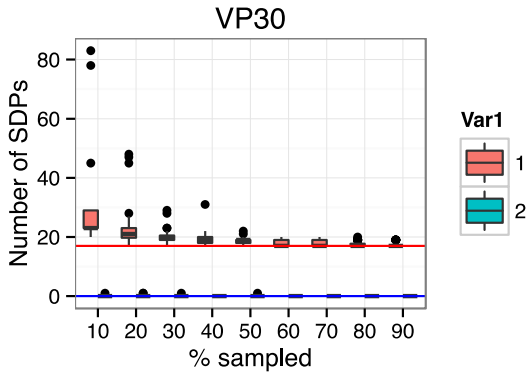


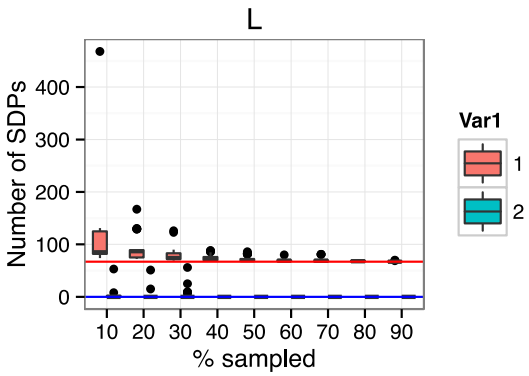


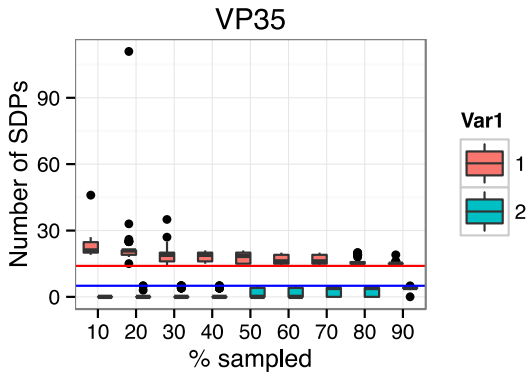


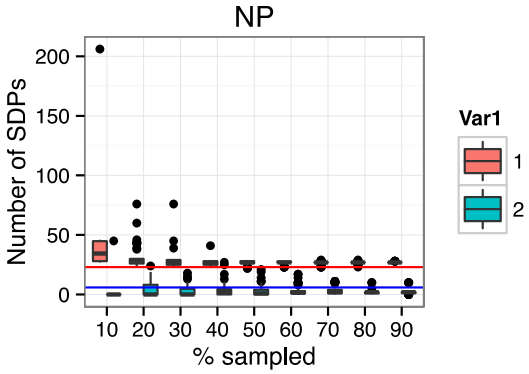
C. Both groups sampled

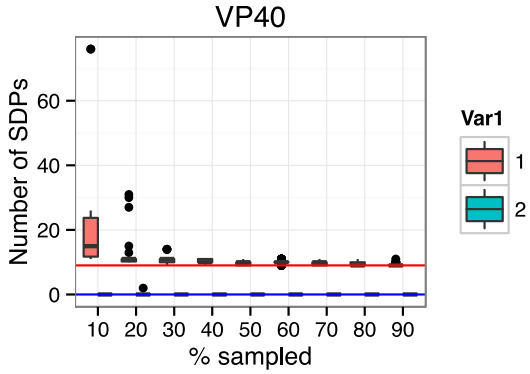












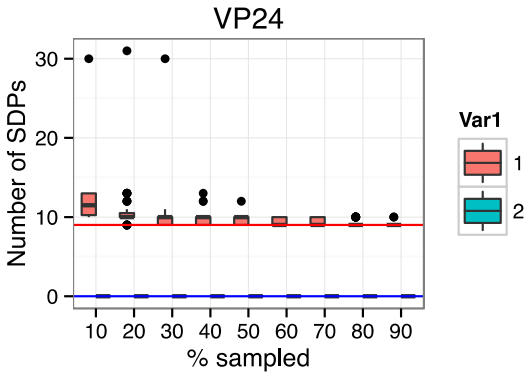


Figure 2.5. Analysis of completely conserved SDP with subsampling of Ebolavirus sequences. The two groups of sequences 'human pathogenic' and Reston ('non-human pathogenic') were sampled and SDP predictions made (see materials and methods). The boxplots show the number of SDPs predicted in each sampling that 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.

2.3.2. Structural Analysis

The Reston virus structures of GP1 and GP2 were modelled using one-to-one threading in Phyre2 (Kelly 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. SDPs were mapped onto the protein structures using PyMOL (Annex 1: Figures SF4-5), and solvent accessibility for SDPs was calculated using DSSP (Joosten et al., 2011).

Full-length structures for VP24 and VP40 were available in the PDB, as well as structures for the globular domains of GP, sGP, NP, VP30, and VP35 (Annex 1: Supplementary Table 11). It was not possible to model the oligomerisation 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). Most SDPs are located on protein surfaces (Annex 1: 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 with the potential to influence protein integrity and hence affect stability, flexibility and conformations of the protein (Table 2.1). Five additional SDPs may alter protein

structure/function but the evidence supporting them is weaker (Annex 1: 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 (Annex 1: Supplementary Table 12). The three other SDPs with weaker evidence were present in the glycan cap in GP (see *2.3.3. Multiple SDPs are present in the GP glycan cap*). 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 *2.3.6. VP40 SDPs may alter oligomeric structure* and *2.3.7. VP24 SDPs affect KPNA5 binding*).

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	

Table 2.1. SDPs that are likely to alter Reston virus protein structure and function.

2.3.3. 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). Sub-unit GP2 is responsible for the fusion of viral and host cell membranes. However, their cellular binding partners remain to be defined (Dahlmann et al., 2015; Feldmann & Geisbert, 2011; Herbert et al., 2015; Miller et al., 2012). 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 (Annex 1: Supplementary Figure 4 and Annex

1: Supplementary Table 12). Three SDPs (I260L, T269S, S307H) are located in the glycan cap that contacts the host cell membrane (Annex 1: Supplementary Figure 4B-C). These changes (particularly S307H at the top of the glycan cap) alter the electrostatic surface of GP (Annex 1: 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 analysed Ebolaviruses and the SDPs were not located close to them (Annex 1: 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 (Annex 1: Supplementary Figure 23), as there is a lack of functional information available for these proteins (Mehedi et al., 2011; Miranda & Miranda, 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.

2.3.4. 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 oligomerisation domain (residues 94–112) (Bettina Hartlieb, Modrof, Mühlberger, Klenk, & Becker, 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 (B. Hartlieb, Muziol, Weissenhorn, & Becker, 2007). Crystallisation 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 2.6). 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, a software for predicting the effect of mutations in proteins using graph-based signatures, predicts this change to be destabilising with a $\Delta\Delta G$ −0.969 Kcal/mol. The Reston virus conformation also buries functional residues A179 and K180 potentially affecting protein function (Clifton et al., 2014) (Figure 2.6). Moreover, our findings show that the Ebola virus protein conformation is conserved in all human-pathogenic Ebolaviruses suggesting that it is relevant for human pathogenicity.

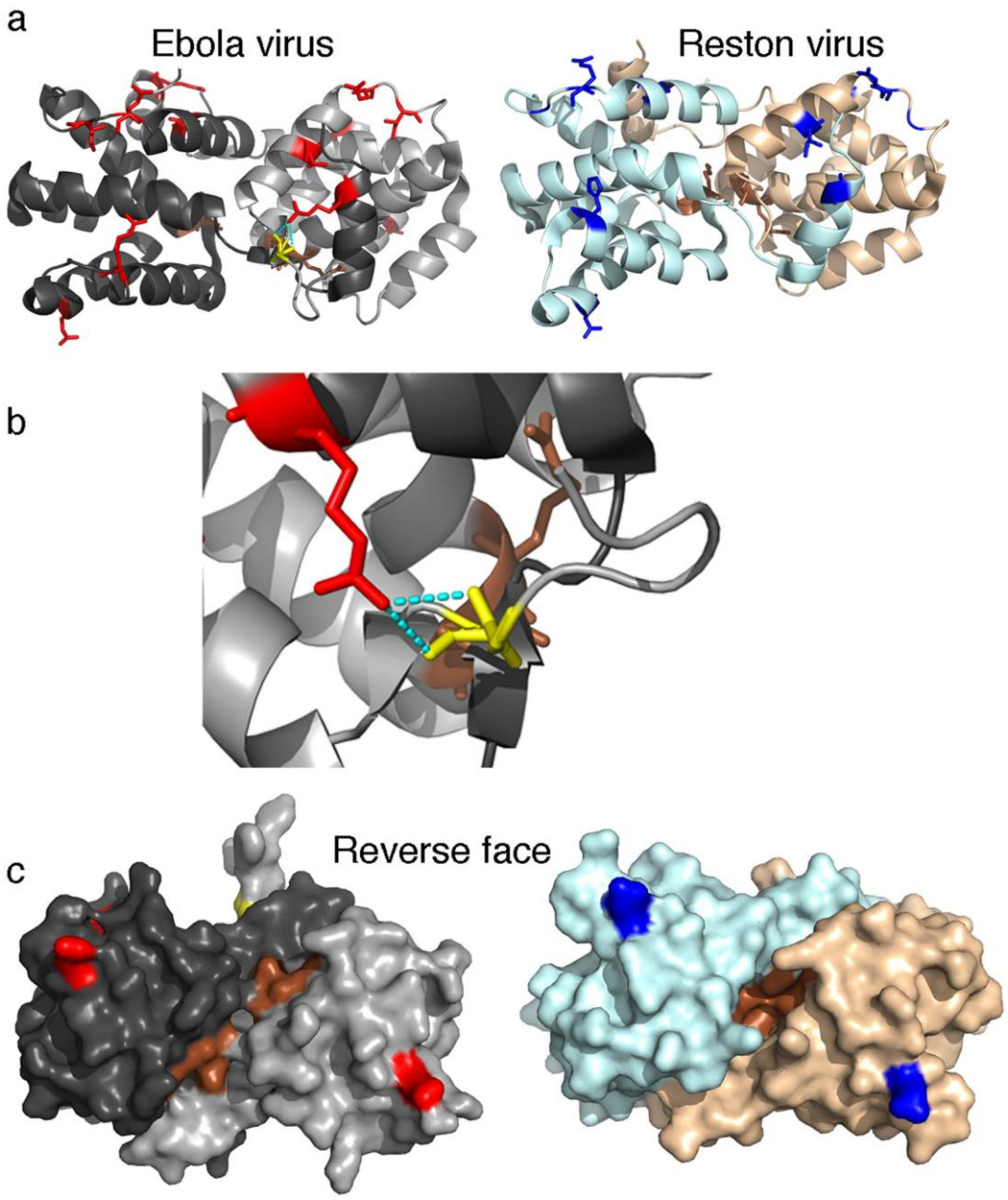


Figure 2.6. SDPs present in the VP30 dimer. The dimer structure of both Zaire Ebolavirus (PDB structure 2I8B) and Reston Ebolavirus (PDB structure 3V7O) VP30 are shown with SDPs indicated (red – Zaire Ebolavirus, blue – Reston Ebolavirus) and functional residues (brown – A179, K180). (a) Cartoon representation: For the Zaire Ebolavirus 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.

2.3.5. VP35 SDP present in dimer interface

VP35 is a multifunctional protein that antagonises interferon signalling by binding double stranded RNA (dsRNA). Structural data are available for both the Zaire Ebolavirus and Reston Ebolavirus VP35 monomer and an asymmetric dsRNA bound dimer (Bale et al., 2013; Kimberlin et al., 2010; D. W. Leung et al., 2009; Daisy W. Leung et al., 2015, 2010). These structures are highly conserved, however functional studies have demonstrated that Reston Ebolavirus VP35 is more stable, has a reduced affinity for dsRNA, and exerts weaker effects on interferon signalling (Daisy W. 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 (Daisy W. 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; Annex 1: Supplementary Table 13). mCSM predicts this change to be slightly destabilising to the complex ($\Delta\Delta G$ −0.11 Kcal/mol). This has the potential to alter the stability of the dimer and thus the ability of VP35 to prevent interferon signalling.

It has recently been demonstrated that a VP35 peptide binds NP and modulates NP oligomerisation and RNA binding to NP (Daisy W. Leung et al., 2015)[35]. 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.

2.3.6. 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 2.7) in the middle of a run of 14 residues that are completely conserved in all Ebolaviruses (Figure 2.7b). 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 destabilising effect ($\Delta\Delta G$ −0.626 Kcal/mol). The Q245P SDP introduces a proline residue into an alpha helix (Figure 2.7b), which most likely breaks and shortens helix five, resulting in the destabilisation 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 $\Delta\Delta G$ 0.059 Kcal/mol). Thus, P85T and Q245P may affect VP40 function and human pathogenicity.

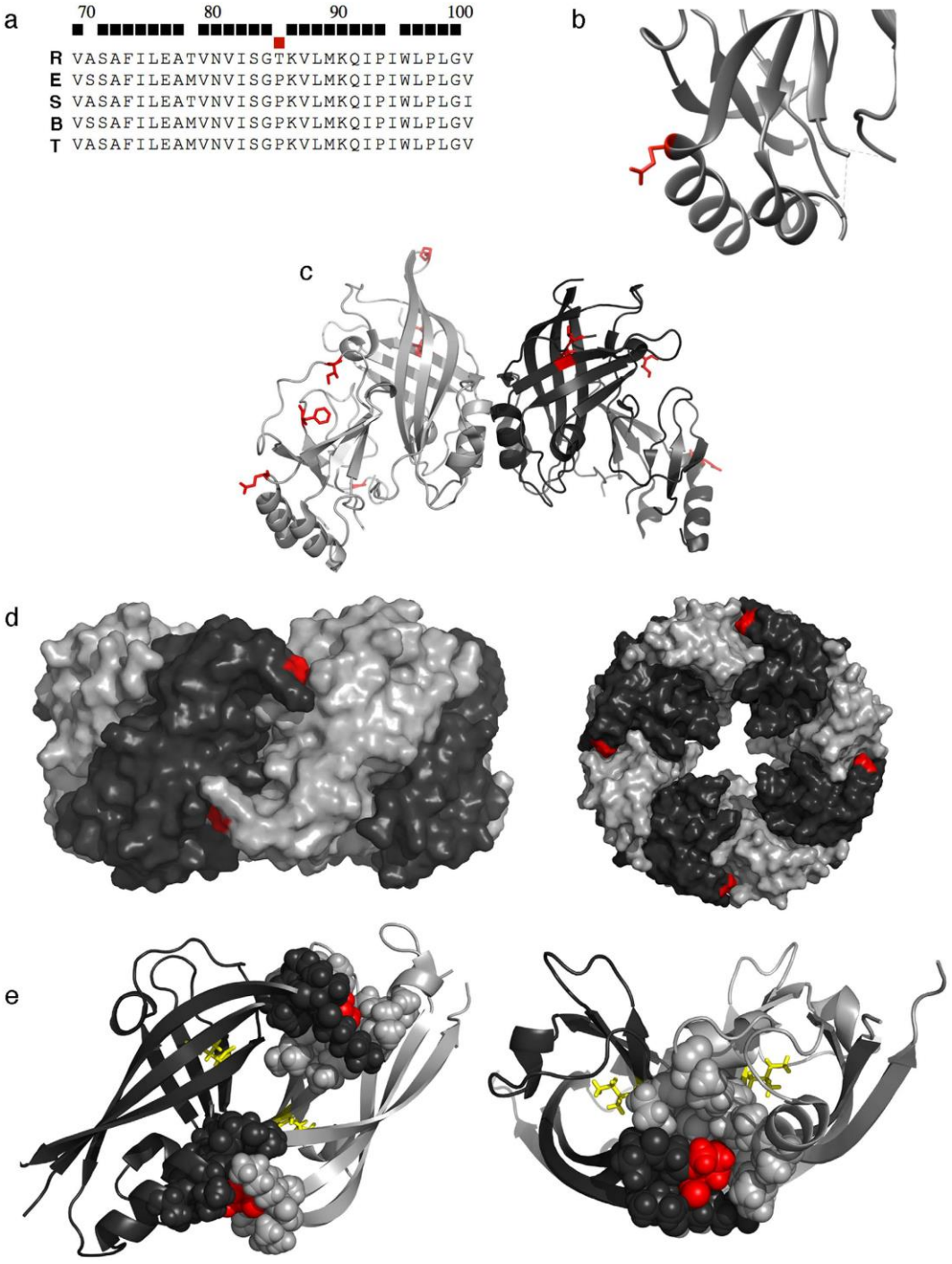


Figure 2.7. The P85T SDP is present in the VP40 octamer interface. (a) Consensus sequence for the region around P85T in Ebolavirus species (R, Reston Ebolavirus; E, Zaire Ebolavirus; S, Sudan Ebolavirus; B, Bundibugyo Ebolavirus; T, Taï Forest Ebolavirus). Black squares indicate positions that are completely 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 coloured 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 coloured red in sphere format, and the SDP I122V is shown as yellow in stick format.

2.3.7. VP24 SDPs affect KPNA5 binding

VP24 is involved in the formation of the viral nucleocapsid and the regulation of virus replication (Feldmann & Geisbert, 2011; Mateo, Carbonnelle, Martinez, et al., 2011; Mateo, Carbonnelle, Reynard, et al., 2011; Morikawa et al., 2007; Watt et al., 2014). VP24 also interferes with interferon signalling through binding of the karyopherins α 1 (KPNA1), α 5, (KPNA5), and α 6 (KPNA6) and subsequent inhibition of nuclear accumulation of phosphorylated STAT1 and through direct interaction with STAT1 (Reid et al., 2006; Reid, Valmas, Martinez, Sanchez, & Basler, 2007; Xu et al., 2014; Zhang et al., 2012). Eight VP24 SDPs are in regions with available structural information (Annex 1: Supplementary Tables 17 and 18). Seven of these are present on the same face of VP24 (Figure 2.8a) 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 2.8). M136 and Q139 are part of multi-residue mutations in Ebola virus VP24 that removed KPNA5 interactions (Annex 1: Supplementary Table 17) (Xu et al., 2014) and are adjacent to K142 (Figure 2.8a), mutants of which have shown reduced interferon antagonism (Ilinykh et al., 2015). 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 2.8b), with the potential to alter structural integrity and influence protein function. Analysis using mCSM predits the T226A change to be destabilizing with a $\Delta\Delta G$ −0.935 Kcal/mol. mCSM predicted seven of the eight analysed SDPs to be destabilising (Annex 1: Supplementary Table 2).

VP24-mediated inhibition of interferon signalling may be critical for species-specific pathogenicity (Mateo, Carbonnelle, Reynard, et al., 2011; Reid et al., 2006, 2007; Xu et al., 2014; Zhang et al., 2012). In this context, VP24 was a critical determinant of pathogenicity in studies in which Ebolaviruses were adapted to mice and guinea pigs that are normally insusceptible to Ebola Virus Disease (de La Vega et al., 2015; Dowall et al., 2014; Ebihara et al., 2006; Mateo, Carbonnelle, Reynard, et al., 2011; Reid et al., 2006). 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 2.8c and 2.8d, Annex 1: 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 2.8e and 2.8f, & Annex 1: Supplementary Table 19) or alter hydrogen bonding with KPNA5 (H186Y, Figure 2.8f & Annex 1:Supplementary Table 19). Thus there is strong evidence suggesting that the VP24 SDPs have a role in rendering the Reston Ebolavirus non-pathogenic in humans.

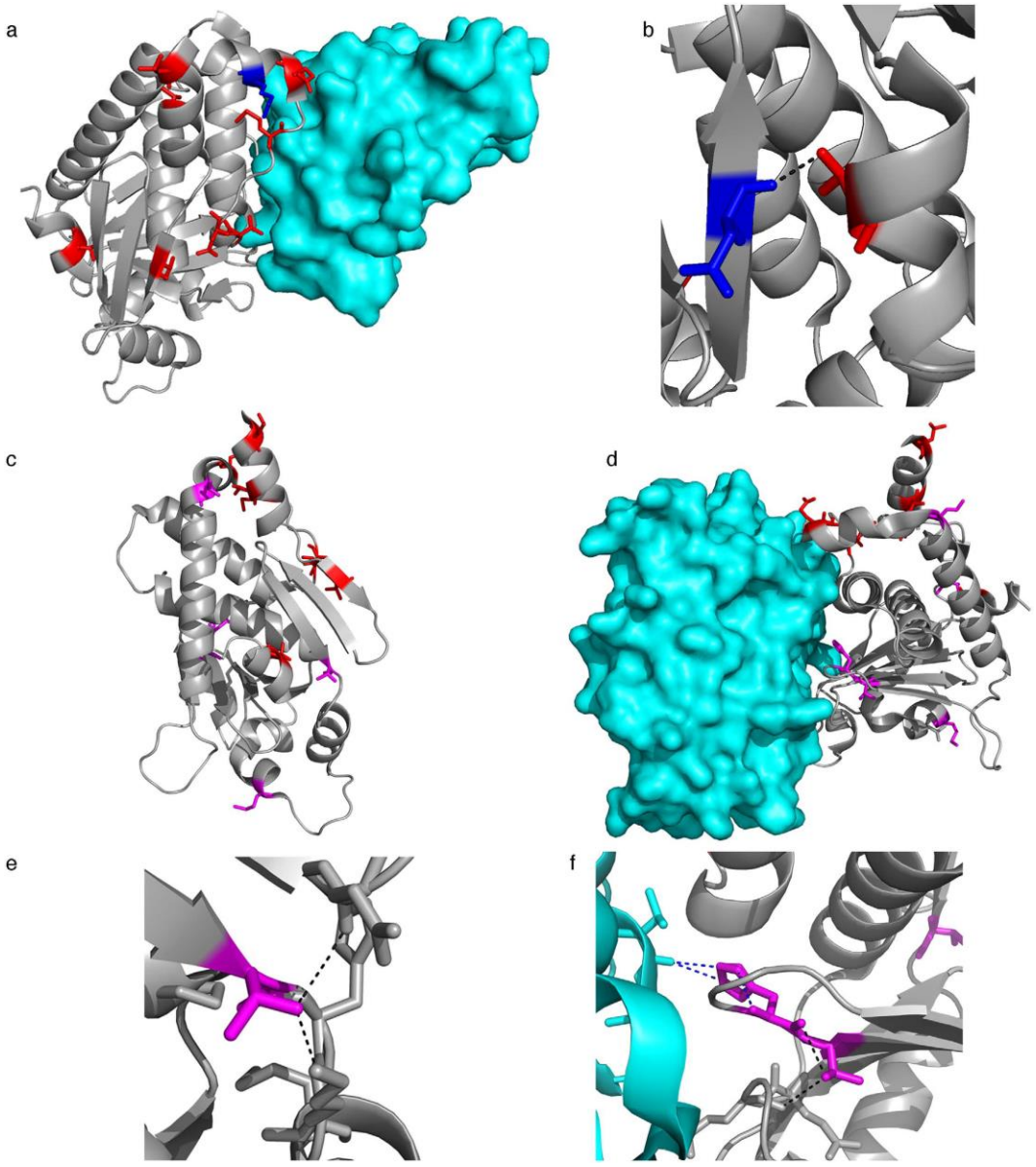


Figure 2.8. Ebola virus VP24 SDPs and complex with KPNA5. (a) VP24 Structure (grey) in complex with KPNA5 (cyan) (PDB structure: 4U2X), with VP24 SDPs (red) and K142 coloured blue. (b) T226 (red) hydrogen bond with the backbone of D48 (blue). (c) VP24 showing residues mutated in rodent adaptation experiments (magenta) and SDPs identified in this study (red). (d) VP24 (grey) and KPNA5 (cyan) complex with residues mutated during adaptation (magenta) and SDPs (red). (e) Hydrogen bonds formed by VP24 T50. (f) Hydrogen bonds formed by VP24 H186, and T187. Intrachain bonds are coloured black and hydrogen bonds between VP24 and KPNA5 are coloured blue.

2.4. Discussion

We have combined the computational identification of residues that distinguish Reston Ebolaviruses 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 Ebolaviruses 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 2.8a) 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.

In (Pappalardo et al., 2017) our group studied the amino acid changes involved in Ebola virus adaptation to rodents, identifying 33 different mutations associated with Ebola virus adaptation to rodents in the proteins GP, NP, L, VP24 and VP35. Only VP24, GP and NP were consistently found mutated in rodent-adapted Ebola virus strains, and fewer than five mutations in these genes seem to be required for the

adaptation of Ebola viruses to a new species. Three VP24 mutations located in the protein interface with karyopherins may enable VP24 to inhibit karyopherins and subsequently the host interferon response. Three further VP24 mutations change hydrogen bonding or cause conformational changes. Hence, this is consistent with our hypothesis that few mutations including crucial mutations in VP24 enable Ebola virus adaptation to new hosts.

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.

This approach has already been used by our group (Martell, Masterson, McGreig, Michaelis, & Wass, 2019) with the newly discovered Ebolavirus Bombali species (Goldstein et al., 2018), in order to test its potential to cause disease in humans. The 1,408 Ebolavirus genomes available at that time were used to predict 166 SDPs, 146 of which were already identified in this work. At SDPs, Bombali virus shared the majority of amino acids with the human pathogenic Ebolaviruses (63.25%). However, for two SDPs in VP24 (M136L, R139S) that have been proposed to be critical for the lack of Reston virus human pathogenicity because they alter the VP24-karyopherin interaction, the Bombali virus amino acids match those of Reston virus. Thus, Bombali virus may not be pathogenic in humans. Supporting this, no Bombali virus-associated disease outbreaks have been reported, although Bombali virus was isolated from fruit bats cohabitating in close contact with humans, and anti-Ebolavirus antibodies that may indicate contact with Bombali virus have been detected in humans.

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; D. W. Leung et al., 2009; Reid et al., 2006, 2007; Watt et al., 2014). 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, 2013; Barrette et al., 2009; Marsh et al., 2011), there is a concern that (a potentially airborne) human pathogenic Reston Ebolaviruses may emerge and pose a significant health risk to humans. Notably, asymptomatic Ebolavirus infections have also been described in dogs (Weingartl, 2013) and Zaire Ebolavirus shedding was found in an asymptomatic woman (Akerlund, Prescott, & Tampellini, 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; Daisy W. Leung et al., 2010; Reid et al., 2006, 2007; Watt et al., 2014) suggest that a few mutations could lead to a human pathogenic Reston Ebolavirus. Deeper understanding of how this changes work and development of methods to effectively predict them from their genetic sequence of newly discovered viruses could lead to great advantages in public health control, as the emergence of new pathogenic strains or adaptions between species could be easily detected when not even predicted, giving public health institutions more time to plan and react before an outbreak. Plus, the identification of key amino acids of the proteins responsible for pathogenic would bring new insights about the mechanisms viruses uses to infect our organism and, therefore, open the door to new candidates for the development of vaccines and antiviral treatments in the future.

Chapter 3:
Neuroblastoma UKF-NB-3 Genetic Landscape

3.1. Introduction

UKF-NB-3 is a MYCN-amplified neuroblastoma cell line that was established from a bone marrow metastasis of a high-risk neuroblastoma patient (Kotchetkov et al., 2005). As we have previously seen in section *1.7.2. Cancer cell lines as model system*, cancer cell lines are a suitable *in vitro* research models for cancer. Many anti-cancer drugs have been discovered and/ or initially characterised in cancer cell lines and cancer cell line panels, such as the NCI60 panel (Holbeck et al., 2010; Sharma et al., 2010; Shoemaker, 2006). The characterisation of cancer cell lines has also revealed in depth insights into cancer biology. This includes gene networks associated with cancer, mutation and selection processes, and evidence of the DNA damage that triggered carcinogenesis (Pleasance et al., 2010; Pleasance & Stephens, 2010). Moreover, the use of cancer cell lines provides fundamental insights into cancer cell plasticity and its relevance for the cancer cell response to anti-cancer drugs (McGranahan et al., 2015; Sharma et al., 2010).

Therefore, by describing the genetic landscape of UKF-NB-3, we expect to improve our understanding of drug-resistant high risk neuroblastoma and the genetic variants related to this particular case of the disease.

3.2. Methods

The methods stated in this chapter will be also used in following Chapter 4, which experiments build on the study of UKF-NB-3 genetic landscape.

3.2.1. Sequencing

Whole exome sequencing (WES) was performed on the UKF-NB-3 parental cell line and 10 single-cell-derived clones. Exome enrichment was performed using the "Nextera Exome Enrichment Kit" (Illumina) according to the manufacturer's instructions. Briefly, 50 ng DNA was used for fragmentation and adapter integration by applying transposase-based method, followed by amplification by PCR for indexing the different libraries. The indexed libraries were pooled and hybridized to biotinylated enrichment probes. Unbound DNA was washed away after binding the probes to a streptavidin bead-matrix. The probe-captured DNA was released from the beads, amplified by PCR and sequenced on an Illumina HiSeq2000 machine using 2x100 bps paired end sequences. An average of 2x 50 million reads was produced covering the 62 Mb exome >50x on average.

3.2.2. Variant calling

The sample with the lowest coverage was about >58x covered. The resulting 100 nt paired-end reads were mapped to the human reference genome hg19 using bwa mem (H. Li & Durbin, 2010). Quality of sequencing was acceptable for every sample when analysed with Fastqc (Andrews S. , 2010). Trimmomatic was used to clip off bad quality ends of the reads and sequencing tags had not already been removed, and to

drop low quality and short reads. PCR duplicates were removed using picard (Picard) and indels realigned using GATK (McKenna et al., 2010). Variants were and called with using samtools mpileup (H. Li et al., 2009), which is a standard and efficient pipeline (Alioto et al., 2015). Before calling the variants, picard (Picard) was to remove PCR duplicates and GATK (McKenna et al., 2010) for INDELs realigning. All the methods were called using the default parameters but samtools, where the method for cancer and non-diploid samples was chosen and the threshold was set to singleton level, parameters -m 3 -F 0.0002 (3 supporting reads at minimum 0.02% frequency), in order to call small INDELs that were detected in previous work. For this reason, a strict filtering on 3060 Phred quality score was used to avoid false positives. For analysis of UKF-NB-3, common variants were filtered out by removing all variants with a frequency greater than 0.01 in gnomAD (Lek et al., 2016).

The number of quality variants called in this way was similar to using the GATK recommended workflow, and 82% of the hits were shared between the methods results. However, GATK workflow was not able to call the small INDELs detected in the wet lab.

3.2.3. Cancer genes

In order to focus our study in cancer related genes, we elaborated a list of 844 cancer genes by merging COSMIC's (Forbes et al., 2017) cancer gene census with Intogen's (Gonzalez-Perez et al., 2013) cancer driver genes (Annex2: Supplementary Table 3) A second list of 27 neuroblastoma driver genes was purely extracted from Intogene, based on (Pugh, 2013) experimental results (Annex2: Supplementary Table 4).

3.2.4. Variant annotation

Variants were annotated for functional information by Variant Effect Predictor (McLaren et al., 2016). The Genome Aggregation Database (gnomAD) (Lek et al.,

2016) was used to identify common variants, and SIFT (P. Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2013) to predict pathogenic effects. ClinVar (Landrum et al., 2018) was also used to annotate clinical related variants.

3.2.5. Other analysis

R was used for all posterior statistical analysis and the comparison between the parental and clone sub-lines (12). Copy number variation (CNV) was calculated with CopywriteR (Kuilman et al., 2015). Cancer signatures present in the cell lines were identified using signeR (Rosales, Drummond, Valieris, Dias-Neto, & da Silva, 2016). Plots and figures were created with ggbio in R (Yin, Cook, & Lawrence, 2012).

3.3. Results

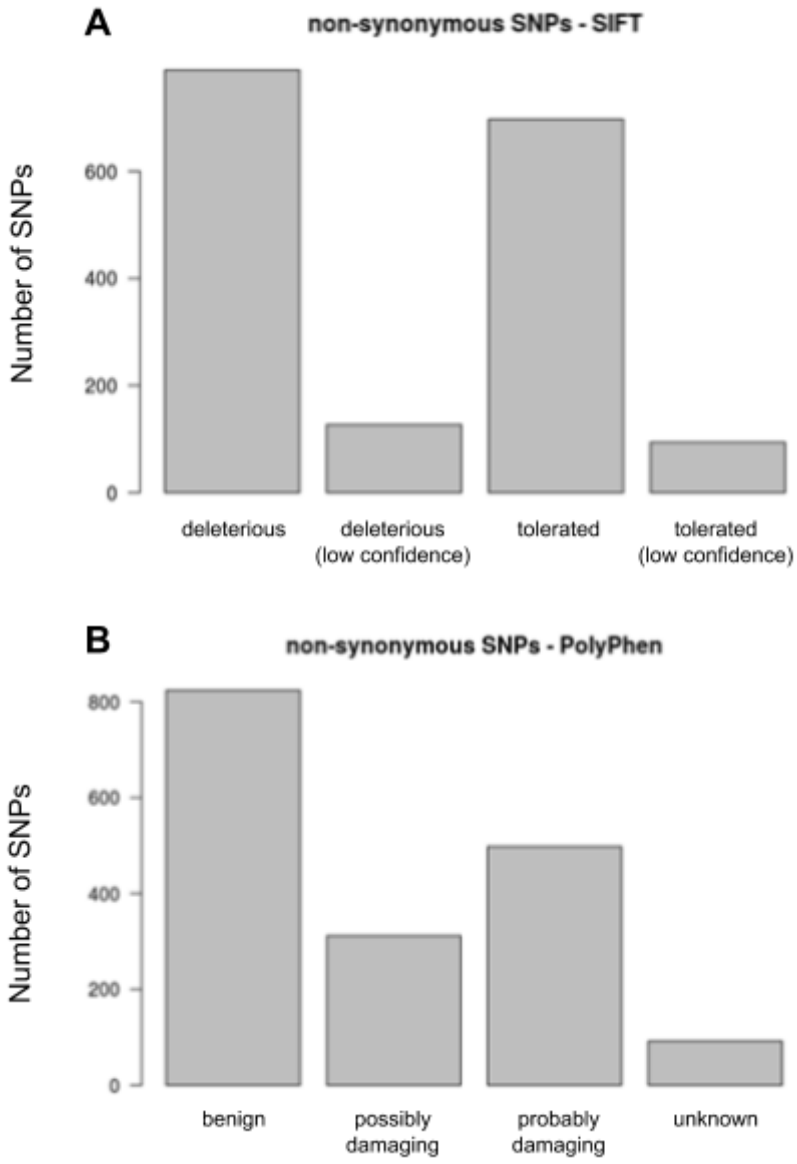
The analysis of the whole exome sequencing data revealed 15,398 variants (Annex 2: Suppl. Table 1), a number that was reduced to 2,414 mutations after the removal of common variants (defined as variants that have a frequency of > 1% in gnomAD (Lek et al., 2016). This included 437 non-synonymous changes, 121 INDELs, 65 frameshift mutations, 15 gained stop codons, 320 mutations in splice sites, 340 and 866 mutations in 5' and 3' untranslated regions (UTR) respectively, and 250 synonymous changes (Annex 2: Suppl. Table 2).

Using Intogen to predict cancer driver genes (Gonzalez-Perez et al., 2013) and Cancer Census of COSMIC (Forbes et al., 2017), we identified a set of 701 mutations in 844 putative cancer-associated genes (Annex 2: Suppl. Table 3), 35 of which occur in the 27 neuroblastoma driver genes identified by (Pugh, 2013) (Annex 2: Suppl. Table 4 and Annex 2: Suppl. Table 5).

3.3.1. Effect of mutations

ClinVar is a database that links genomic sequence variants to disease (Landrum et al., 2018). Only 29 of the 2,414 mutations found in UKF-NB-3 cells were annotated in ClinVar and only two of these mutations were categorised as "likely pathogenic", and one as "pathogenic" (Annex 2: Suppl. Table 6). Given that ClinVar is primarily focused on (hereditary) genetic variants that cause a certain phenotype (disease) and that cancer is a multifactorial process that is largely driven by somatic mutations (and epigenetic changes) that are acquired over a lifetime (Stratton, Campbell, & Futreal, 2009), this may not be too surprising.

Next, we analysed the putative effects of the 2,414 detected mutations on protein structure and function using the two complementary approaches SIFT (P. Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2013). The concept of SIFT is based on the consideration that crucial positions in protein sequences are conserved during evolution. Hence, the method determines the evolutionary conservation of positions of interest in protein sequences to estimate their potential effect on protein function (P. Kumar et al., 2009). SIFT predicted 157 of the single nucleotide variants to affect protein function (Figure 3.1a and Annex 2: Suppl. Table 7). PolyPhen-2 uses sequence- and structure-based features to estimate the effects of changes in protein sequence on protein function (Adzhubei et al., 2013). It predicted 151 changes to be "damaging" to protein function (Figure 3.1b and Annex 2: Suppl. Table 8). There was an overlap of 111 variants that were predicted to affect protein function by both SIFT and PolyPhen-2 (Figure 3.1 and Annex2: Suppl. Table 9).


A
non-synonymous SNPs - SIFT
Number of SNPs
0
200
400
600
deleterious
deleterious (low confidence)
tolerated
tolerated (low confidence)
B
non-synonymous SNPs - PolyPhen
Number of SNPs
0
200
400
600
800
benign
possibly damaging
probably damaging
unknown

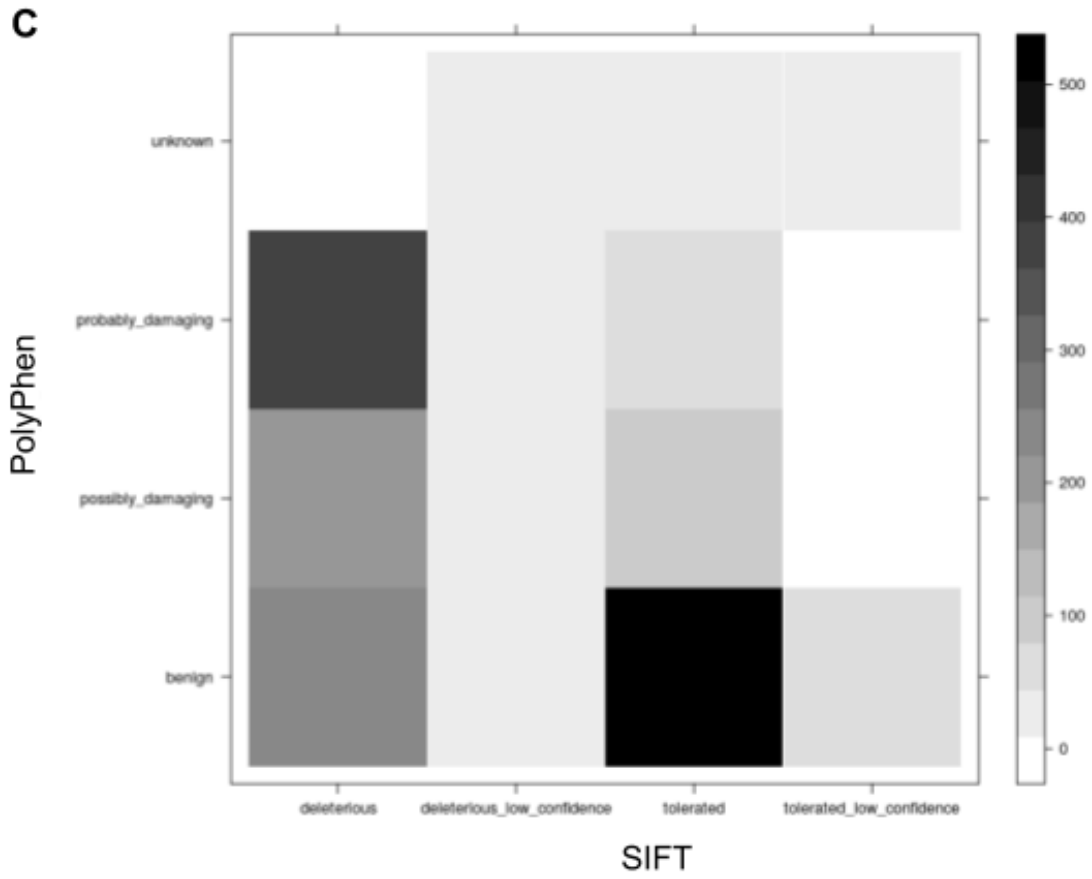


Figure 3.1. Variant Effect predictions. SIFT (A) and PolyPhen (B) predictions of nonsynonymous SNPs. The heatmap (C) shows the overlapping of their predictions.

3.3.2. Mutated genes

Recent sequencing studies of neuroblastoma tumours identified further common neuroblastoma driver genes including ALK, PTPN1, ATRX, MYCN, and NRAS (Molenaar et al., 2012; Pugh, 2013). Additionally, Intogen classifies a further 22 genes as drivers in neuroblastoma resulting in a group of 27 driver genes. The mutational status of UKF-NB3 was considered on those driver genes.

We found mutations in 12 of these genes; ALK, ANK3, CEP290, LRP6, MACF1, MECOM, MET, MUC4, NF1, NOTCH1, PBRM1, TRIO (Annex 2: Suppl. Table 3).

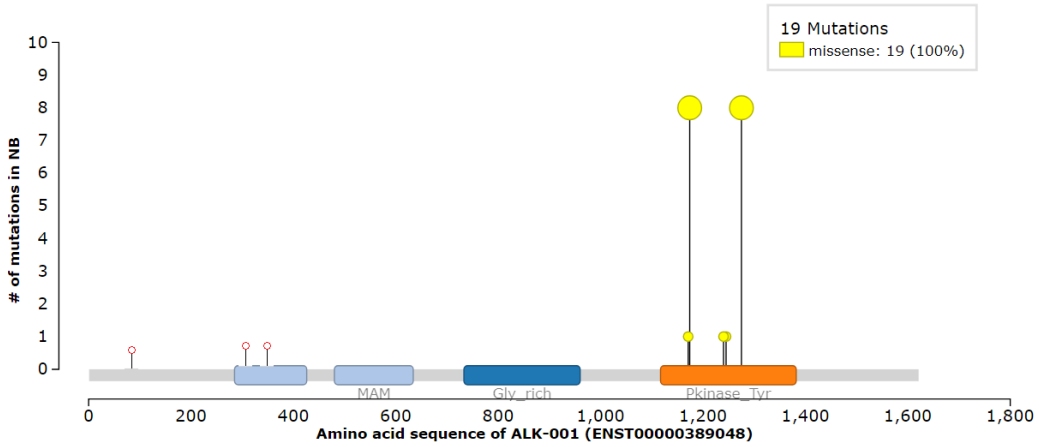


Figure 3.2. Variants in driver genes. Red batons on the ideogram represent non-synonymous SNVs in ALK in UKF-NB-3, and the yellow batons show previously observed variants. The taller the baton, in more cell lines that variant was observed in Original figure and previously observed variants extracted from https://www.intogen.org/search?cancer=NB&gene=ALK, a resource of Intogen (Gonzalez-Perez et al., 2013).

UKF-NB-3 has six variants in ALK (Figure 3.2 and Annex 2: Suppl. Table 10). All of them are common, two are missense, one nonsense and three synonymous. Only the nonsense mutation was not found in ClinVar (chr2_29444095_C_T) and got no effect predictions, but it is so frequent (0.97% of gnomAD population contained it) that it is highly unlikely it is related to disease. The other four variants are annotated in ClinVar as benign, and the two missense ones are also predicted by SIFT to be tolerated.

There are three mutations in ANK3 (Annex 2: Suppl. Table 3), all common and only two of them occurring in coding regions but synonymous, both annotated as likely benign in ClinVar.

CEP290 contains one rare missense variant (chr12_88502846_A_G) which is not annotated in ClinVar but predicted to be deleterious and probably damaging by SIFT

and PolyPhen-2, respectively. It also contains three common variants not annotated as dangerous in ClinVar (Annex 2: Suppl. Table 3).

LRP6 shows one common missense mutation not annotated in ClinVar but predicted to be tolerated and benign by SIFT and PolyPhen-2 (Annex 2: Suppl. Table 3).

11 mutations where found in MACF1 (Annex 2: Suppl. Table 3); two synonymous SNVs, one 3' prime UTR variant, one in a splice site and seven missense variants, none of them contained in ClinVar. Two of the missense variants are interesting, as one (chr1_39801815_A_C) is possibly damaging according to PolyPhen-2 and the other (chr1_39823135_A_G) is rare but predicted to be benign.

There is one rare mutation in MECOM (chr3_168810845_G_A), not annotated in ClinVar but predicted to be to be deleterious and possibly damaging by SIFT and PolyPhen-2 (Annex 2: Suppl. Table 3).

Finally, MUC4 contains 140 different variants, none of them found in ClinVar. As previously stated, this gene is abnormally mutated in this cell line. NOTCH has three common variants, PBRM1 two and TRIO has one with no effect associated. The only variant in MET is common and related with splicing (Annex 2: Suppl. Table 3).

The mutational status of UKF-NB-3 was also considered for a larger set of 669 driver genes, which includes the neuroblastoma driver genes and driver genes identified across all types of cancer. Exonic non-synonymous mutations were identified in 197 of these drivers, 2560 of a total of 2784 variants detected in these genes have been previously observed and 1559 occur frequently in ExAC. 43 mutations were predicted to be deleterious by SIFT and 26 as probably damaging by PolyPhen, and nine of them where annotated as pathogenic by ClinVar. (Annex 2: Suppl. Table 3).

3.3.3. Copy Number Variation

Analysis of Copy Number Variation (CNV) in UKF-NB-3 using CopywriteR showed amplification in many regions and deletions primarily in chromosomes 23 (X) and 24 (Y) (Fig 3.3). UKF-NB3 was derived from a MYCN amplified tumour and MYCN is identified as being highly amplified in the cell line, with 16 times more reads than expected for the normal copy number. It is the 29th most amplified gene in UKF-NB-3. The most amplified pseudogene is LINC00283 with 7.21. There are 11 genes and pseudogenes between 5:6, 19 between 4:5, 95 between 3:4, 948 between 2:3 and 12608 between 1:2. The genes and pseudogenes with no amplifications or deletions (between -1:1) are 68,041, and the ones with a negative CNV are 1,989 between -1:-2, 202 between -2:-3, 25 between -3:-4, 5 between -4:-5 and 784 almost deleted (Annex 2: Suppl. Table 10).

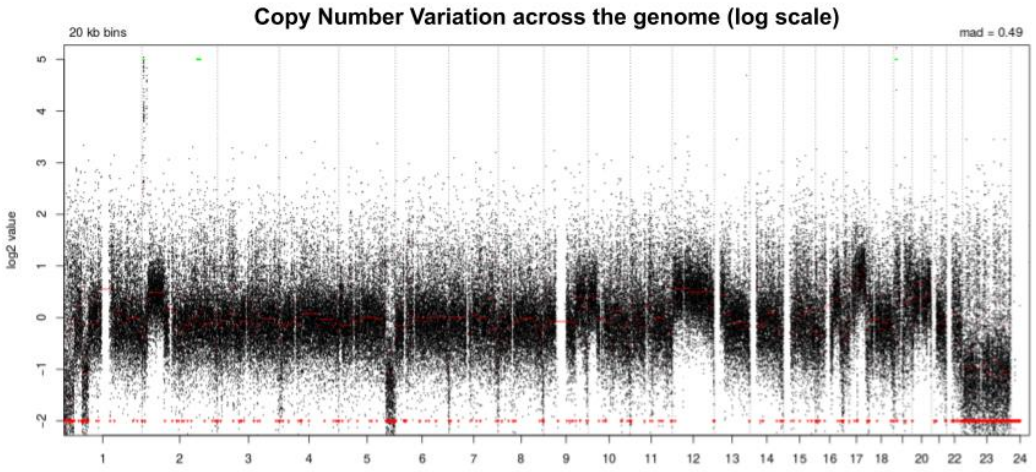


Figure 3.3. Copy number variation in UKF-NB-3. Values are in log2 scale, meaning 0 is the average copy number; 1 being 2x, 2 is 4x and so; and -1 means x0.5 and -2 is x0.25. Despite not been visible at chromosome scale, many genes like MYCN show high values up to 4 (16 times more copies). Chromosomes 23 and 24 correspond to chromosomes X and Y, respectively.

3.3.4. Cancer signature

The cancer signature of UKF-NB-3 (Fig 3.4) shows three characteristics patterns (2, 22 and 23). The main one, in the C>T mutations which corresponds to signature 1 in COSMIC database, is a cancer signature that has been identified in all cancer types and in most cancer samples and correlates with age of cancer diagnosis. It is the result of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine, and it is associated with small numbers of small insertions and deletions in most tissue types. We can also see three other smaller patterns in C>A, C>G and T>C substitutions. These smaller signatures do not correlate with the values for these changes in the signature described before, neither with any other mayor cancer signature published in COSMIC, so it may be a characteristic signature of our cell line.

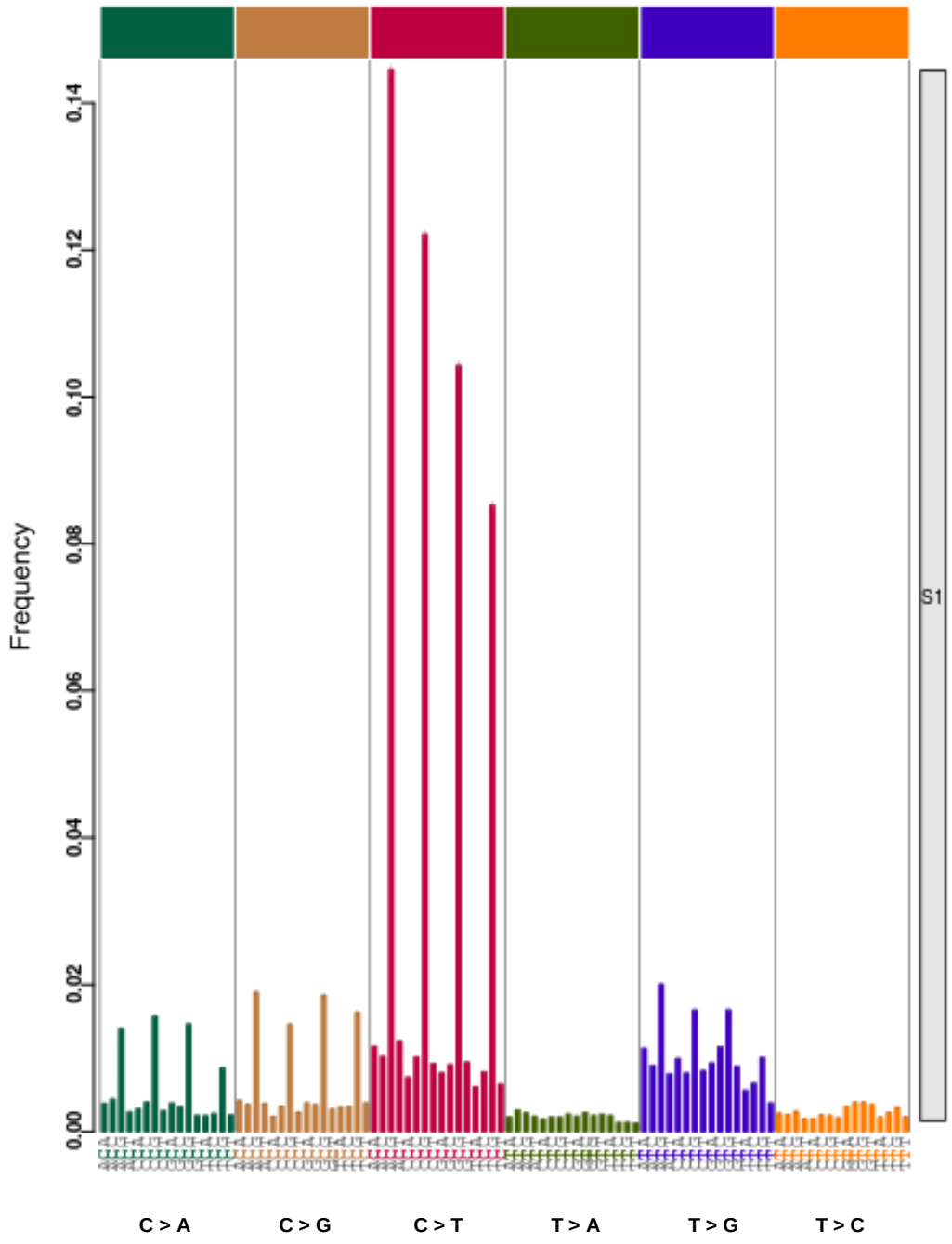


Figure 3.4. Cancer signature of UKF-NB-3. SigneR predicted only one signature (S1). Each block of the horizontal axis contains a particular nucleotide substitution, and inside the block all the possibilities of nucleotides surrounding it. The vertical axis shows how frequently that nucleotide substitution is happening in that particular context.

3.3.5. Pathway enrichment analysis

Pathway, network neighbourhood, gene ontology and protein complex analysis were done with the Max Plank Institute for Molecular Genetics over-representation analysis tool, ConsensusPathDB (Kamburov et al., 2011) (Figure 3.5).

As we do not have RNA expression values, we could only evaluate the over-representation of mutated genes in each pathway. This kind of analysis is not really informative, as the effects of the mutations in those genes can have unpredictable effects, form not affecting the mutated protein function at all to crippling and/or modify one or more of their functions; but we will reuse this result later in *4.3.8. Pathway enrichment analysis* when studying the internal heterogeneity of UKF-NB-3.

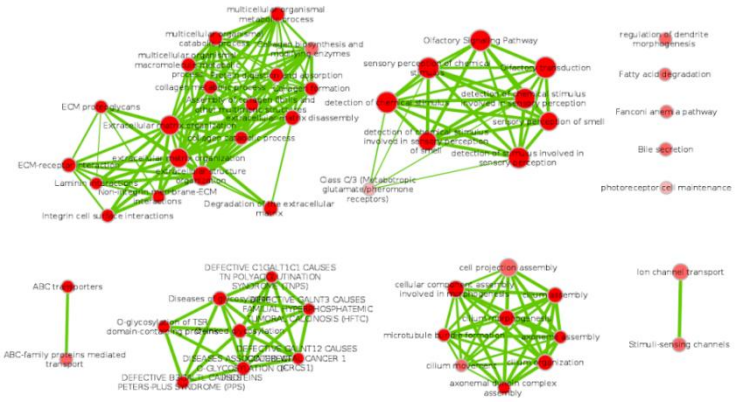


Figure 3.5. Pathway analysis of UKF-NB-3. The pathways are represented as the edges of the network and the colour of the dots represents how much the pathway is likely to be affected in a scale white-red. The green link represent interaction between pathways affected by mutated genes, and the thicker they are the more confident they are.

3.4. Discussion

The UKF-NB-3 cell line was stablished in 2005 (Kotchetkov et al., 2005), and since then it has been used in multiple experiments (Löschmann et al., 2013; Michaelis et al., 2012, 2011). During all this time, we had limited knowledge of its genomic properties, and with this research we aim to provide a better understanding of this cell line for future experiments. We have sequenced the exome or UKF-NB-3 and identified its variants to the reference human genome using variant databases, effect prediction tools and population frequencies.

Here we have described how the mutations are distributed across UKF-NB-3 genome, especially in neuroblastoma driver genes and other commonly mutated genes in cancer. We have considered the clinical information known about the variants, and predicted their effects. We have also looked into the CNVs, cancer signature and affected pathways as far the characteristics of our data allowed us to go.

More than 15,000 variants were identified in UKF-NB-3, with only ~16% of them being considered rare (< 1% frequency in gnomAD). The difficulty of detecting some variants observed in previous experiments suggested that the cell line may not be homogeneous and/or stable, and further work is reported on this issue in *Chapter 4*.

UKF-NB-3 has 12 mutated neuroblastoma driver genes. We studied in detail the potential effect of the rare mutations located in those genes by comparing their annotation in ClinVar and their predictions in SIFT and PolyPhen-2. None of them were individually linked to disease, but that is expected in a complex disease like cancer. Thus, we also looked for the same disease information in a broader set of genes which are commonly mutated in cancer, and we generated UKF-NB-3 cancer signature to compare it with other better known cancers. Finally, a pathway enrichment analysis was executed to gather more information about the collective effect of individual mutations.

While the whole exome sequencing approach used enabled the identification of many SNVs and small indels it is limited in the detection of SVs, particularly CNVs. Further, this data would be enhanced by combining with other types of omics data, particularly gene expression data (transcriptomics). Therefore, in some parts of our analysis we were constricted to use the only tools that have proved to work, even in a limited way, with such data.

We expect the genomic description of UKF-NB-3 cell line will help future research in the field and help us to improve our understanding of high-risk neuroblastoma cancer genomics.

Chapter 4:
UKF-NB-3 internal heterogeneity

4.1. Introduction

It has now become clear that cancer diseases are characterised by an incredible inter- and intra-tumour heterogeneity (Jamal-Hanjani, Quezada, Larkin, & Swanton, 2015; Lipinski et al., 2016). Large cell line panels are thought to cover the (inter-)tumour heterogeneity to a certain extent (Garnett & McDermott, 2014; Sharma et al., 2010). Moreover, cancer cell lines have long been known to be characterised by some level of intra-cell line heterogeneity (Barranco et al., 1983; Zanker, Treappe, & Blumel, 1982). This heterogeneity is thought to be the consequence of a combination of two events. It is partly caused by the heterogeneity of the cancer cell populations that cancer cell lines are derived from. In addition, cancer cells are characterised by genetic instability that results together with evolutionary pressures exerted by the cell culture environment in a genetic shift (Masramon, 2006; Torsvik et al., 2014). Hence, cancer cell lines may play a role in 1) the investigation of cancer cell biology and response to anti-cancer drugs and 2) the examination of evolutionary processes in a (cancer) model characterised by genetic instability.

However, detailed research on the intra-cell line heterogeneity of cancer cell lines is very limited. To address this, here, the MYCN-amplified neuroblastoma cell line UKF-NB-3 (Kotchetkov et al., 2005) was systematically characterised alongside 10 single cell-derived clonal sub-lines by the determination of growth kinetics, drug resistance profiles, and whole exome sequencing for intra-cell line heterogeneity.

4.2. Methods

As the research in this chapter continues what was started in *Chapter 2*, the methods used are the same described in that chapter's section.

4.2.1. Clonal sub-lines

The ten clonal sub-lines we used to study the internal heterogeneity of UKF-NB-3 are single-cell derived clones. This means each of them was stablished from the parental UKF-NB-3 cell line by extracting one single cell and cultivating it in a new medium until growing to become a new clonal sub-line of UKF-NB-3. The numbers in the naming of the clones refer to the number of the experiment, as 100 sub-lines of this type were planned to be stablished. The choosing criteria for sequencing these 10 were: 1) to have been successfully stablished, i.e. they did not die in the process; and 2) the original cell they come from had to be from different regions of the parental cell line so internal heterogeneity would be easier to study.

4.3. Results

In order to study the internal heterogeneity in detail, we defined four groups of variants. For variants that are called in clonal sub-lines but not in the parental UKF-NB3 line we distinguish two different classifications. For some of these variants it is possible that there are reads supporting the variant in the parental UKF-NB3 but without sufficient confidence for it to be called. This may reflect a heterogeneous cell population, where this variant is present in only a small proportion of cells and therefore is not called. We class these as gained mutations. For the remaining variants, there is no evidence for them in the parental UKF-NB-3 cell line, so it is possible that they have occurred in the clonal sub-line. These are classed as de novo mutations. Finally, the lost group was split into two. Partially-lost mutations, where the mutation was present in the parental UKF-NB-3 cell line and there was some evidence for it in the sequencing data in the sub-lines but not sufficient for it to be called. Fully-lost mutations were present in the parental UKF-NB-3 but not detected in the clonal sub-lines.

Sample	Frameshift	Indels	nonsense	nonsyn	splicesite	synonymous
Clone 1	74	347	87	8715	1900	9303
Clone 2	79	379	89	9078	2047	9673
Clone 24	80	365	83	8874	1923	9416
Clone 3	80	361	86	8692	1890	9227
Clone 4	74	350	85	8707	1876	9193
Clone 56	86	371	93	8869	1935	9456
Clone 64	77	349	86	8814	1889	9354
Clone 7	67	363	87	8872	1957	9460
Clone 80	80	372	89	9084	1998	9643
Clone 93	74	378	93	9068	2006	9621
UKF-NB-3	62	312	70	7816	1661	8178

Table 4.1. Variants present in UKF-NB3 and clone sub-lines. Detailed data are presented in Suppl. Table "description_table.txt".

Overall, we found that each of the clonal sub-lines had a similar number of mutations (Table 4.1), whether considering all mutations, de novo or gain mutations (Figure 4.1D, E, and F). The number of gain mutations ranged from 2045 (clone 1) to 2339 (clone 93), and de novo mutations from 763 (clone 1) to 914 (clone 93) (Annex 3: Suppl. Table 1 and Annex 3: Suppl. Table 2). Hence, for the majority of the new mutations there is evidence for them in the parental cell line. Notably, the clonal sub-lines did not share any de novo variants, suggesting that as their name suggests they are new mutations. Likewise, the clonal sub-lines had lost variants ranging from 591 (clone 2) to 903 (clone 4) and de novo lost variants from 217 (clone 2) to 407 (clone 4). Clone 4 had not only the highest number of lost variants but also UKF-NB-3 had its greatest number of de novo mutations when comparing to it, being therefore the least similar to the parental UKF-NB-3 cell line of all the single cell derived clones and potentially the least represented sub-line captured in the original sequencing experiment.

4.3.1. Heterogeneity

The ten single cell-derived clonal UKF-NB-3 sub-lines displayed between 14,336 (UKF-NB-3clone3) and 15,222 variants (UKF-NB-3clone93) (Annex 3:

Supplementary Table 3). After removal of the common variants, the number of variants ranged between 1,579 (UKF-NB-3clone1) and 1805 (UKF-NB-3clone2).

Notably, the clonal UKF-NB-3 sub-lines displayed (except from UKF-NB-3clone1) at the selected Phred score of 30 higher numbers of variants than UKF-NB-3, both before (13,012 variants observed in UKF-NB-3) and after removal of the common variants (1,413 observed in UKF-NB-3). Our first assumption was that because of a lower heterogeneity, variants would be called with more certainty in the clonal sub-lines than in UKF-NB-3. To examine this hypothesis, we compared the number of variants (without removing the common ones) at different levels of stringency, i.e. at different Phred scores (Table 4.2). If the higher variability in UKF-NB-3 resulted in a lower number of reliable calls, we would expect to see a larger number of calls in UKF-NB-3 than in the clonal sub-lines at lower quality thresholds. Indeed, at a Phred score of 0, only UKF-NB-3 clone24 displayed a higher number of mutations than UKF-NB-3. At a Phred score of 10, five of the clonal UKF-NB-3 sub-lines (UKF-NB-3clone2, UKF-NB-3clone3, UKF-NB-3clone4, UKF-NB-3clone7, UKF-NB-3clone24) displayed higher numbers of mutations than UKF-NB-3. When we increased the Phred score thresholds to 80 or 100, all clonal UKF-NB-3 sub-lines displayed increased numbers of mutations compared to UKF-NB-3 (Table 4.2).

These lower scores can be interpreted as the parental cell line not being monoclonal, but contains small subpopulations with small genetic differences. This genetic variability is inferior in the single-cell-derived clones, which translate in higher quality scores for their variants as there is lesser background noise. This analysis partly confirmed our initial hypothesis that a higher heterogeneity may be associated with a lower number of called variants at high quality thresholds and a higher number of called variants at lower quality thresholds. However, other factors, which will need to be examined in future studies, also seem to contribute to this phenomenon.

Sample	Raw Reads	QUAL 0	QUAL 10	QUAL 20	QUAL 30	QUAL 60	QUAL 80	QUAL 100
UKF-NB-3	9406707	527241	373549	216083	190182	63368	49262	40600
Clone 1	11519307	330648	271744	184306	167051	62536	52733	45503
Clone 24	14281153	496471	412714	282676	268790	79733	64060	53522
Clone 2	15860910	542018	448469	302367	230027	67497	54882	46003
Clone 3	12610134	454198	376695	257572	248597	69820	56448	46968
Clone 4	13532042	482764	402844	278817	235204	71436	58298	49228
Clone 56	13190217	387579	320112	217691	251794	72836	58843	49208
Clone 64	13539427	425433	355336	248573	196380	67382	56254	48269
Clone 7	14256891	459912	382475	262935	222499	67029	55119	46710
Clone 80	15405731	431644	359540	249428	223760	72172	59848	51089
Clone 93	14108200	426880	348390	230849	207748	72819	60301	51490

Table 4.2. Number of raw reads and variants present in UKF-NB3 and clonal sub-lines. The number of variants called using different Phred quality scores (QUAL) thresholds is shown.

4.3.2. Variants distribution

Initially, we analysed the similarity/ relatedness between UKF-NB-3 and its clonal sub-lines as well as among the clonal UKF-NB-3 sub-lines. For this analysis, we did not remove the common variants. For variants present in the clonal sub-lines, we defined two types of variants based on there being any evidence for the variant in UKF-NB-3. To do this for all variants identified in the clonal sub-line we considered if there was any evidence for them in UKF-NB-3 (i.e. a single read, disregarding Phred scores). Where there was no evidence at all of a variant being present in UKF-NB-3 the variant was classed as de novo i.e. a new mutation that was not present in the parental line (Table 4.4). Where there was some evidence for a variant in UKF-NB-3 but not sufficient for it to be called, then the variant was classed as gained i.e. the variants may be present in UKF-NB-3 at low frequencies and is present in one or more single cell derived sub-lines (Table 4.3).

The number of gained mutations in each subline ranged from 2045 in UKF-NB-3clone1 to 2239 in UKF-NB-3clone93 (Table 4.3), while the number of de novo mutations ranged from 763 in UKF-NB-3clone1 to 914 in UKF-NB-3clone93 (Table 4.4). There was almost no overlap between the de novo mutations in the different clones, being most of them unique, providing further evidence that they may be actual de novo mutations that occurred after the isolation of the single cells that the

clonal sub-lines are derived from and that the parental cell line is not homogeneous and contains subpopulations. (Annex 3: Supplementary table 4)

VariantsOf/ NotIn	Clone 1	Clone 2	Clone 24	Clone 3	Clone 4	Clone 56	Clone 64	Clone 7	Clone 80	Clone 93	UKF-NB-3
Clone 1	0	807	978	1172	1162	995	1028	1010	819	854	2045
Clone 2	1352	0	1204	1420	1407	1222	1284	1187	1024	1060	2336
Clone 24	1181	862	0	1236	1248	1024	1109	1076	884	921	2138
Clone 3	1157	860	1018	0	1132	1032	1067	1034	882	888	2071
Clone 4	1139	839	1022	1124	0	1015	1053	1022	867	838	2080
Clone 56	1222	904	1048	1274	1265	0	1114	1084	892	927	2181
Clone 64	1127	838	1005	1181	1175	986	0	1030	827	875	2076
Clone 7	1226	858	1089	1265	1261	1073	1147	0	911	928	2158
Clone 80	1316	976	1178	1394	1387	1162	1225	1192	0	1035	2329
Clone 93	1346	1007	1210	1395	1353	1192	1268	1204	1030	0	2339
UKF-NB-3	845	591	735	886	903	754	777	742	632	647	0

Table 4.3. Gained variants in clonal sub-lines. Effect causing variants in UKF-NB-3 and clones, compared to every sample, i.e., number of high quality variants of every sample not present in the others.

VariantsOf/NotIn	Clone 1	Clone 2	Clone 24	Clone 3	Clone 4	Clone 56	Clone 64	Clone 7	Clone 80	Clone 93	UKF-NB-3
Clone 1	0	300	347	411	504	339	373	372	305	298	763
Clone 2	545	0	475	534	631	460	492	431	395	413	909
Clone 24	465	332	0	464	531	338	373	409	337	310	829
Clone 3	453	378	390	0	504	388	430	406	360	353	833
Clone 4	433	345	380	377	0	368	411	394	338	309	827
Clone 56	474	350	367	469	555	0	391	410	341	340	828
Clone 64	422	322	352	427	515	329	0	357	301	318	797
Clone 7	512	350	411	476	574	395	433	0	343	372	811
Clone 80	527	363	421	511	612	391	443	459	0	371	889
Clone 93	510	390	444	542	591	449	503	456	411	0	914
UKF-NB-3	346	217	258	310	407	259	263	261	224	231	0

Table 4.4. de novo variants in clonal sub-lines. Effect causing de novo variants present in clonal sub-lines compared to every sample, i.e., number of variants of every sample that were not detected even with low quality.

Each sub-line had a similar number of each type of mutation (Fig 1D-F) when considering all variants (including those present in the parental cell line) and gain and de-novo mutations. The distribution of the type of variants present in the clonal-sub lines is very similar (Figure 4.1A) with an average of 8,556 (45.89%) being synonymous, 7,992 (42.87%) non-synonymous, 1,639 (8.79%) indels, 3,189 (1.71%) at splice sites, 80 (0.43%) nonsense and 59 (0.32%) frameshift. This distribution is similar when considering all, gain or de novo mutation, with the exception that there is a slightly higher proportion of frameshift mutations in the de novo set (Figure 4.1D, E and F).

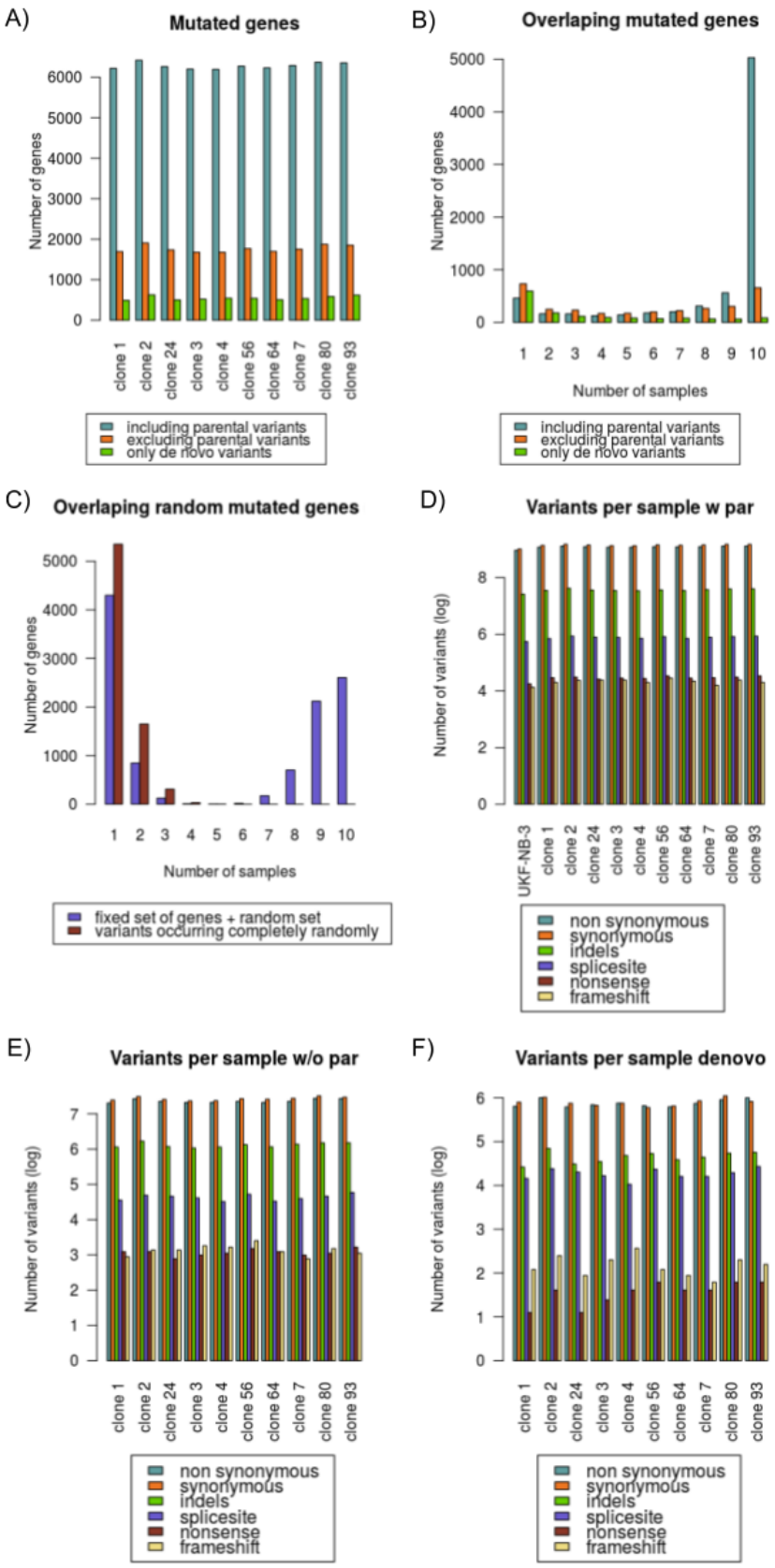
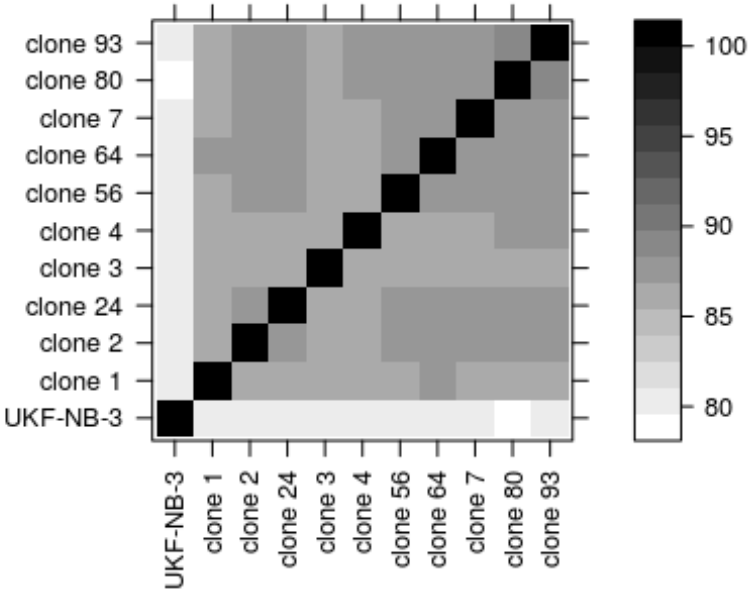

A)
Mutated genes
Number of genes
clone 1
clone 2
clone 24
clone 3
clone 4
clone 56
clone 64
clone 7
clone 80
clone 93
including parental variants
excluding parental variants
only de novo variants
B)
Overlaping mutated genes
Number of genes
Number of samples
including parental variants
excluding parental variants
only de novo variants
C)
Overlaping random mutated genes
Number of genes
Number of samples
fixed set of genes + random set
variants occurring completely randomly
D)
Variants per sample w par
Number of variants (log)
UKF-NB-3
clone 1
clone 2
clone 24
clone 3
clone 4
clone 56
clone 64
clone 7
clone 80
clone 93
non synonymous
synonymous
indels
splicesite
nonsense
frameshift
E)
Variants per sample w/o par
Number of variants (log)
non synonymous
synonymous
indels
splicesite
nonsense
frameshift
F)
Variants per sample denovo
Number of variants (log)
non synonymous
synonymous
indels
splicesite
nonsense
frameshift

Figure 4.1. Variants distribution. A) Number of mutated genes in each clone. In blue the total number, in orange the genes mutated by variants which were not called in the UKF-NB-3 parental cell line and in green the genes mutated by de novo variants. B) Distribution of mutated genes across the clones. The x axis represents how many samples contain the same mutated gene while y counts the number of mutated genes in every group. Again, blue for the total of variants, orange for only the variants not called in the parental UKF-NB-3 cell line and green for the de novo variants. C) Simulations of distributions for completely random variants (brown) and occurring mostly in a set of genes (purple). Explained in detail in section *4.3.3. Simulations of variant distributions.* D) Total number of variants in each sample and parental UKF-NB-3 cell line classified by their type, in log scale. E) Number of variants not called in the parental UKF-NB-3 cell line in each sample classified by their type, in log scale. F) Number of de novo variants in each sample classified by their type, in log scale.

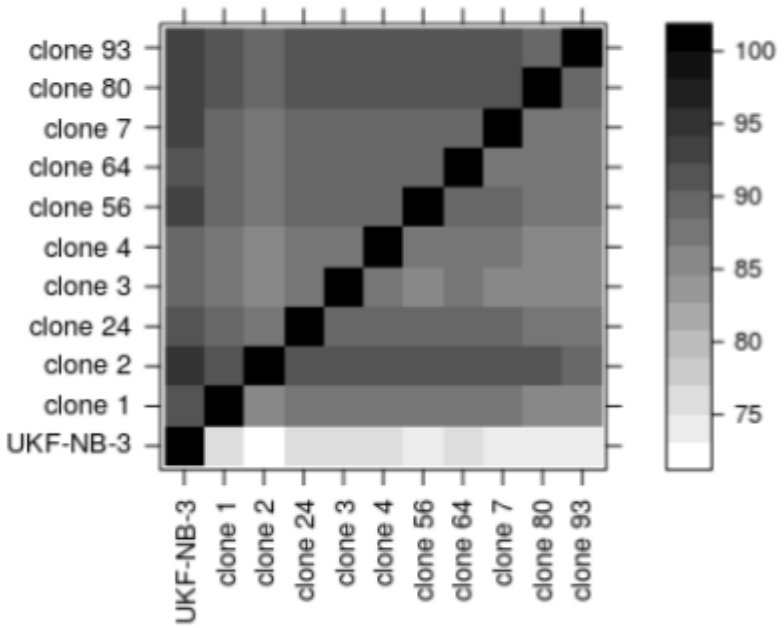
We wondered if there was an overlap in the genes that were mutated between the clonal sub-lines. When considering variants are also present in the parental cell line, most of the variants that are shared between all ten sub-lines, as would be expected. (Figure 4.1B). When considering de-novo and gain mutation together there are many genes that are mutated in only a single sub-line and the number of gene mutated in multiple sub-lines decreases as the number of sub-lines compared increases until 3-4 sub-lines where there is an increase, this again represents variants which may be present in the parental cell line. Finally, there is little overlap between the genes that de novo mutations occur in.

We compared the similarity between our samples in two different ways. In the first heatmap (Fig 4.2A), we show the percentage of mutated genes shared by the samples. It is important to notice that UKF-NB-3 has fewer variants called, which translates into a smaller number of mutated genes compared with the clones. In the second plot (Fig 4.2B), we show the percentage of variants of the samples in the vertical axis shared by the ones of the horizontal axis.

Percentage of shared mutated genes



Percentage of y variants in x



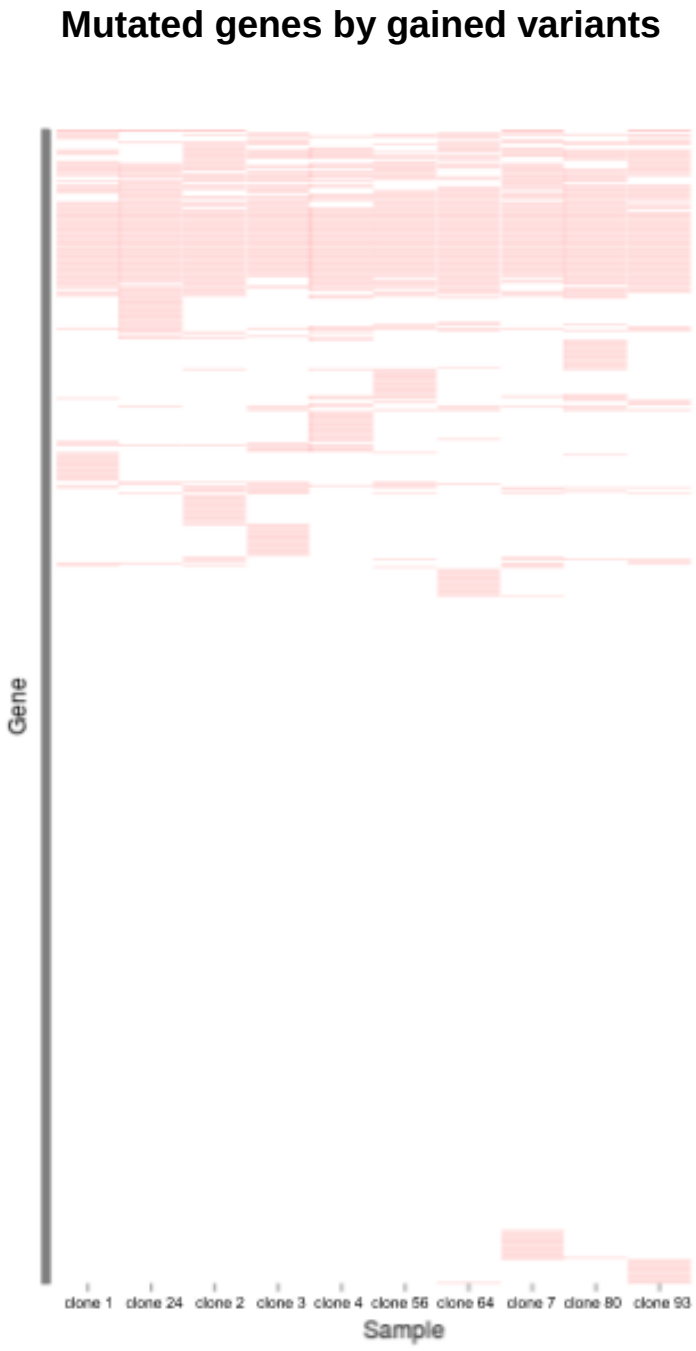


Figure 4.2. Similarity between samples. Similarity heatmaps of our samples comparing both genes (top) and variants (bottom). It is important to notice that not all samples have the same number of genes mutated and variants called, the plots are not symmetric. Finally, the third heatmap shows which genes in each clone are mutated by gained variants; each row represents a human gene and there is a red line if the gene in that sample is mutated.

The most frequently mutated genes in every set (all variants, gained variants and de novo variants sets) for the parental line and all sub-lines are members of the mucin family (MUC4, MUC16, MUC6, and MUC3A) (Annex 3: Suppl. Table 5). Mucins are usually products of epithelial tissues and their tumours, but they have also been observed overexpressed in cancers of other tissues, including pancreas, ovary, breast, lung, colon and prostate (Hollingsworth & Swanson, 2004; Rakha et al., 2005). Mucin expression alteration in cancer facilitates cancer growth, as it affects differentiation, transformation, adhesion, invasion and immune surveillance (Hollingsworth & Swanson, 2004). Another mucin which is not mutated in our samples, MUC1, has been reported to be overexpressed in neuroblastoma and a set of cancer cell lines (Osterkamp, Cheiner, Tefanova, Loyd, & Instad, 1997), so it is possible that the expression of mucins is an intrinsic behaviour of cancer cell lines. MUC4 in particular, the most frequently mutated gene in our samples, has been recently reported to be a key supporter of propagation and survival processes in various epithelial cancer cell lines (Xia et al., 2016).

Despite the most mutated genes in all samples are shared (Annex 3: Suppl. Table 5), there are some variants in not so frequently mutated genes, showing some heterogeneity between the clones and UKF-NB-3. Also, the gained and de novo variants do not cluster in the same genes. Some of the most highly mutated genes in UKF-NB-3 were further mutated with gained mutations in the clonal sub-line (Annex 3: Suppl. Table 6), but other genes that were not mutated in the parental show up here. The MUC family appears again in the top genes with de novo mutations (Annex 3: Suppl. Table 7), suggesting that some of the newly detected variants were not present in the parental, but most of the genes showed in this list were not mutated in UKF-NB-3.

Gained mutations present in the clonal sub-lines but not in UKF-NB-3 preferentially occurred in a limited number of genes (Fig 4.1A, Fig 4.3 and Annex 3: Supplementary Table 6). The number of mutated genes found in each of the clonal sub-lines is very similar (Fig 4.1B). Gained variants in the clonal UKF-NB-3 sub-lines reflect the selection of clones with certain genome profiles.

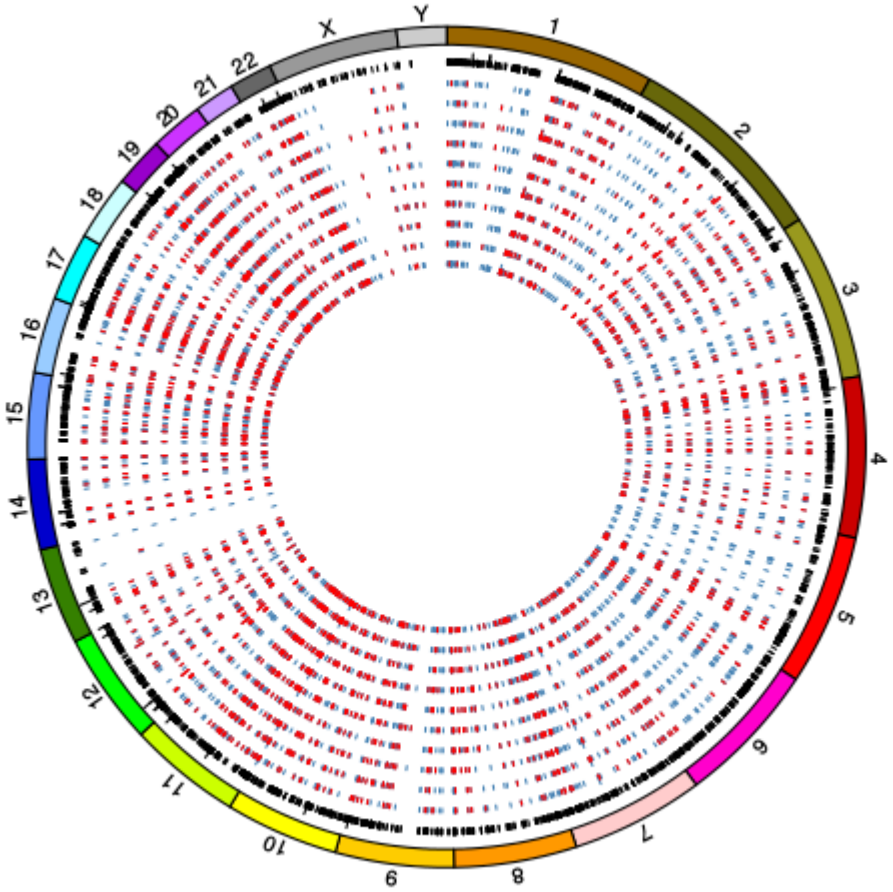


Figure 4.3. Circos plot showing the mutation profiles of UKF-NB-3 and the single cell-derived clonal sub-lines of UKF-NB-3. The external refers to UKF-NB-3. The interior circles refer to the clonal UKF-NB-3 clone sub-lines. In the clones, de novo mutations are showed in red while gained ones are blue.

There was a substantial overlap in the genes that harboured de novo variants (Figure 4.1C, Figure 4.2 and Annex 3: Suppl. Table 7). Some of these genes are highly mutated both in UKF-NB-3 and in the clonal sub-lines. This suggests that mutations in these genes are well tolerated. Some of the novel variants that were not detected in UKF-NB-3 occurred in these highly mutated genes (MUC3A, MUC6, MUC4, MUC16, KCNJ12, FLG, OR9G1), while other gained variants occurred in genes with fewer (or no) variants detected in UKF-NB-3 (ZNRF4, ESPNL, PABPC1, MYOM3, GPRIN2, TNN, NR2E3). The other de novo mutations occurred in less commonly mutated genes (Annex 3: Suppl. Table 4). An analysis of these de novo

variants for known oncogenes and tumour suppressor genes did not reveal the presence of new known driver events. In accordance, the de novo appear to be randomly distributed and to rather indicate genomic instability instead of a directed evolution. Simulations were performed in order to see whether the mutation pattern that we observe actually reflects a combination of preferential mutation sites and a number of randomly distributed mutations.

It is worth mentioning special areas of the genome were little to none variants are observed in Figure 4.1, as both sexual chromosomes X and Y and some large regions in other chromosomes like 3 and 13. This may be due to chromosomal aberrations in this cell line. New on-going work with UKF-NB-3 suggest it has lots of chromosomal translocations, duplications and deletions. One of those can be observed here, as the original donor of the tissue used to create UKF-NB-3 was a male, but chromosome Y is inexistent. There are only no variants in that region of the genome, but also no mapping reads, agreeing with the last karyotype studies where this chromosome could not be found in UKF-NB-3.

4.3.3. Simulations of variant distributions

To study the existence of patterns behind the distribution of the different kind of mutations in our samples, a set of distributions of shared mutated genes across 10 samples was generated. In each of the distributions the proportion of genes that were more likely to gain new mutations was varied, from all genes in the genome (completely random distribution, the mutation has the same chance of appear in any part of the genome) to only the set of genes that were mutated in our samples (some genes have a higher mutation rate and others a minimal mutation rate).

When comparing the patterns of these simulated distributions with our data, the distribution of the set including all our variants follows a deterministic model where almost all the mutated genes belong to a set and almost any new gene gets randomly mutated, while the non-parental set represents a middle point, and the de novo set corresponds to an almost completely random (Fig. 4.1C).

4.3.4. Effect prediction

We then investigated the de novo mutations for a potential functional relevance by ClinVar, SIFT, and PolyPhen-2 (Annex 3: Suppl. Table 4).

Across all clones, there are a total of 20,512 unique de novo variants, 10,880 of them rare (Annex 3: Suppl. Table 4). Only 26 of those are annotated in ClinVar, with only two of them known to be associated to disease (LPL chr8_19842492_T_C, present in UKF-NB-3clone24 and with low quality in UKF-NB-3clone1 and UKF-NB-3clone64) and pathogenic (SLC19A3 chr2_228566905_T_C, present only in UKF-NB-3clone2). The rest of de novo variants, SIFT predicted 471 to be deleterious, and PolyPhen-2 predicted 466 to be potentially damaging.

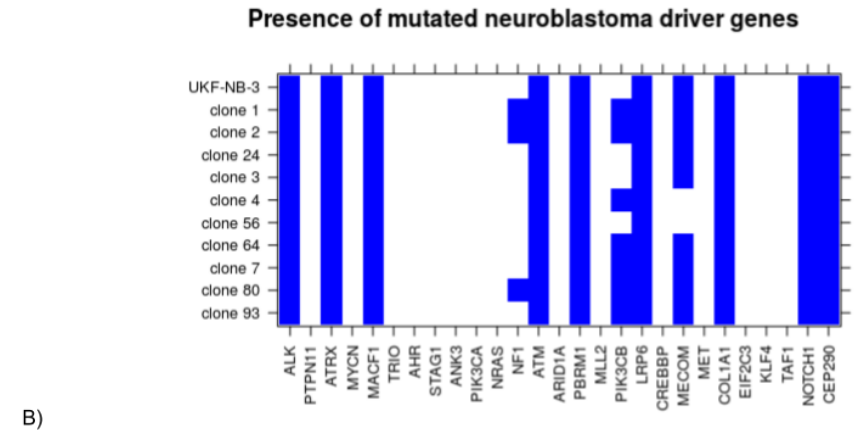
There are a total of 31,432 unique gained mutations across all clones, 13,068 of them being rare. From those, only the two de novo variants described above are annotated as pathogenic or related with disease, of a total of 39 gained mutations found in ClinVar. 545 rare gained mutations were predicted to be deleterious by SIFT, and PolyPhen-2 predicted 556 to be potentially damaging.

4.3.5. Differences in driver and cancer genes

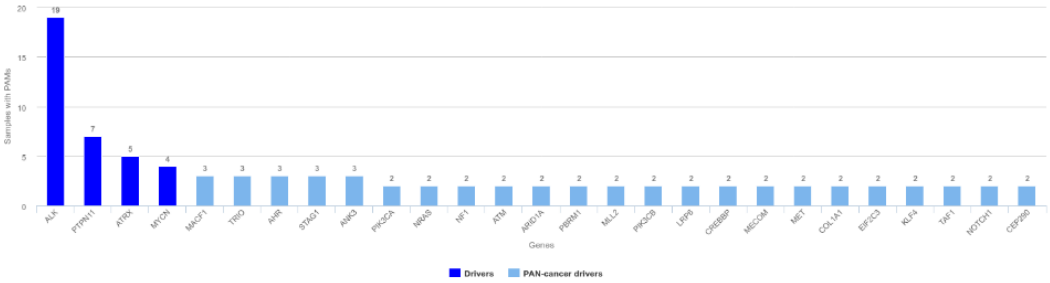
By comparing the most commonly mutated genes in neuroblastoma cell lines (Fig 4.4B) with our samples (Fig 4.4A), we discovered that the driver genes present in UKF-NB-3 are also mutated in the clones, while the neuroblastoma driver genes not mutated in the parental cell line remain unchanged in the clones. In particular, ALK shows 4 different mutations across the clones, keeping the first mutation in the gene common across UKF-NB-3 and the clones, but the second one detected in UKF-NB-3 only appears in four of the clones. Eight clones contain a known mutation in other neuroblastoma cell lines, in the Pkinase Tyr domain, and nine of them contain another one before it. PTPN11 does not develop any mutations in any of the clones, and ATRX has the same mutation in every sample. On the other hand, some of the

commonly mutated genes, that were not mutated in the parental cell line were in the clones, seven (PIK3CB) and three (NF1). Extending the range of genes, we compare to all known driver genes for any cancer, the internal heterogeneity becomes evident (Fig 4.4C). Some cancer genes mutated in UKF-NB-3 are mutated in some of the clones and not in others, and many others not mutated in the parental line are mutated in some or even all clones.

A)



B)



C)

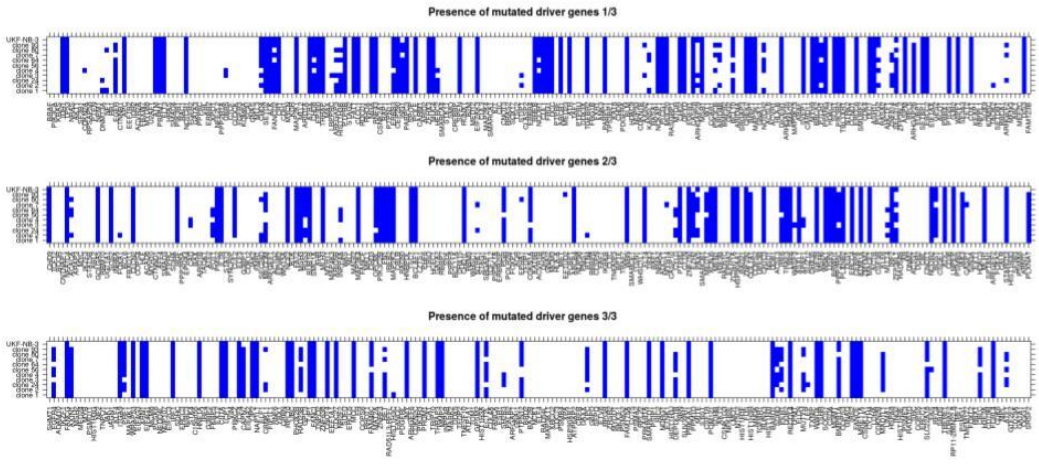


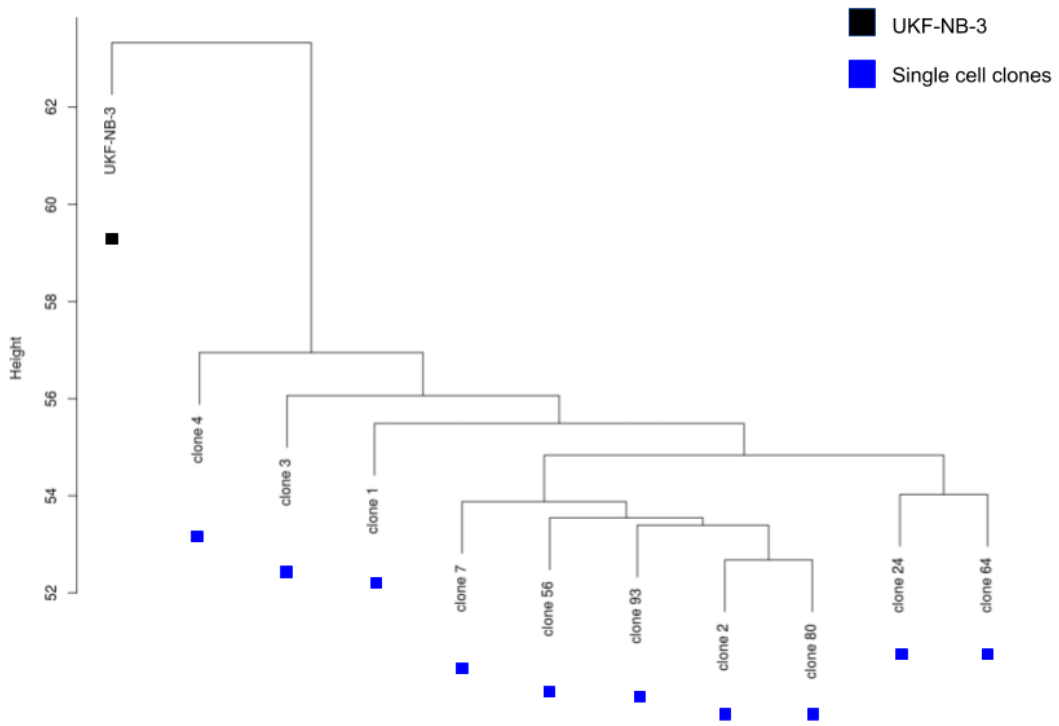
Figure 4.4. Commonly mutated genes in neuroblastoma cell lines. A) Which of the most common mutated genes in neuroblastoma cell lines are mutated in our samples. B) Most commonly mutated genes in neuroblastoma cell lines by Intogen. In dark blue are represented the driver genes. C) Which ones of the driver genes of any cancer are mutated in our samples.

4.3.6. Phylogeny

Clonal evolution of our samples was calculated based on the proportion of alternative alleles in reads (number of reads with ALT alleles divided by the sum of ALT and RF alleles) supporting each variant call. Using these values and setting to 0 every undetected variant in that sample but called in any other sample, we calculated a hierarchical clustering to represent the similarities and branching events of the subpopulations the clones were derived from.

Both Bayesian and Maximum Likelihood hierarchies (Figure 4.5) agreed that UKF-NB-3 parental cell line was the furthest related sample of the study. This result may be due to the smaller number of variants it contains, but also to it being a "common ancestor" containing almost all common variants to the clones but not the ones that differ in each subpopulation.

A)


UKF-NB-3
Single cell clones
Height
52
54
56
58
60
62
UKF-NB-3
clone 4
clone 3
clone 1
clone 7
clone 56
clone 93
clone 2
clone 80
clone 24
clone 64

B)

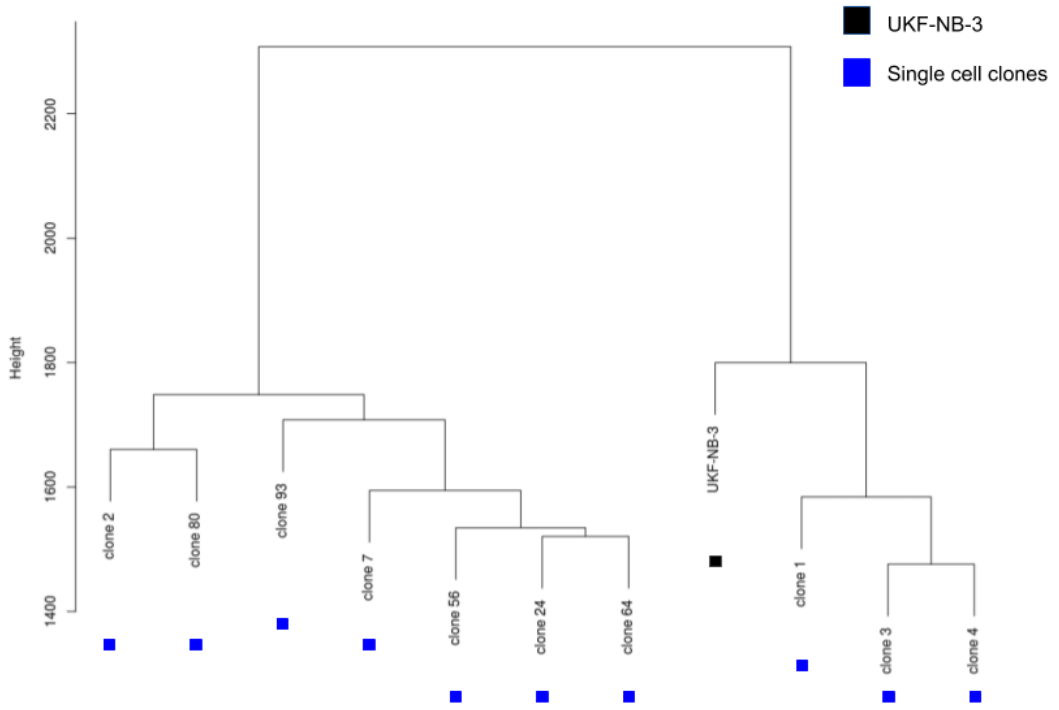

UKF-NB-3
Single cell clones
Height
1400
1600
1800
2000
2200
clone 2
clone 80
clone 93
clone 7
clone 56
clone 24
clone 64
UKF-NB-3
clone 1
clone 3
clone 4

Figure 4.5. Phylogeny. Bayesian (A) and Maximum Likelihood (B) phylogenetic trees of the UKF-NB-3 cell line and clones.

A second approach to build a hierarchy while trying to identify key differentiating events was tried with AncesTree (Figure 4.6). This software was designed to work with samples which differ from less variants, not cancer samples. When trying to find relevant mutational events that could classify our samples, the large amount of variants acts as "background noise" and makes it impossible for the algorithm to find any, if exists.

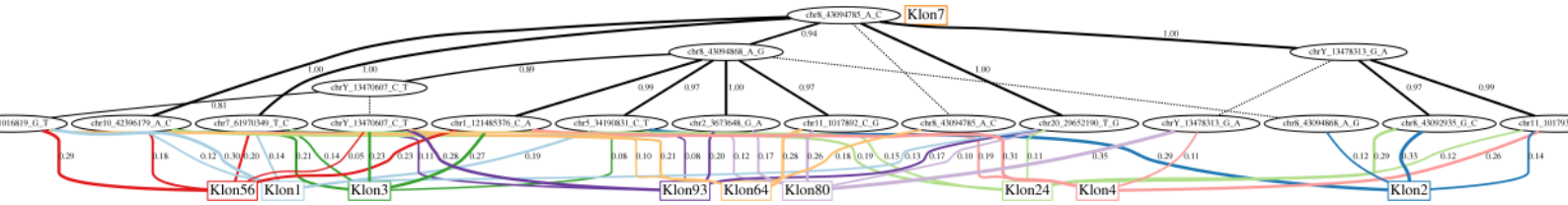


Figure 4.6. Differentiating evolutionary events. Phylogenetic tree made by AncessTree to identify differentiating evolutionary events.

4.3.7. Cancer signatures

Cancer signatures of all 11 UKF-NB-3 samples were calculated together with SigneR. As expected, samples do not differ enough to have a different cancer signature. All of them reported only one cancer signature, the same we found in UKF-NB-3 in section *3.3.4. Cancer signature.*

4.3.8. Pathway enrichment analysis

Following the same approach as in section *3.3.5. Pathway enrichment analysis*, we obtained similar results for each clone. Pathway analysis returned similar results for the 11 samples, showing 4 core clusters of pathways which are affected. This result was expected, as almost 90% of the mutated genes are shared by all the samples. The

heterogeneity of the clones induces some variability in their results, adding some pathways and removing others, but the core 4 groups are always present (Figure 4.7).

As previously stated, due to the nature of our data this analysis gives us not much useful information apart from confirming that similar pathways were affected in all samples. As each clone had a slightly different set of mutated genes compared with the parental UKF-NB-3, significantly enriched pathways slightly differ too.

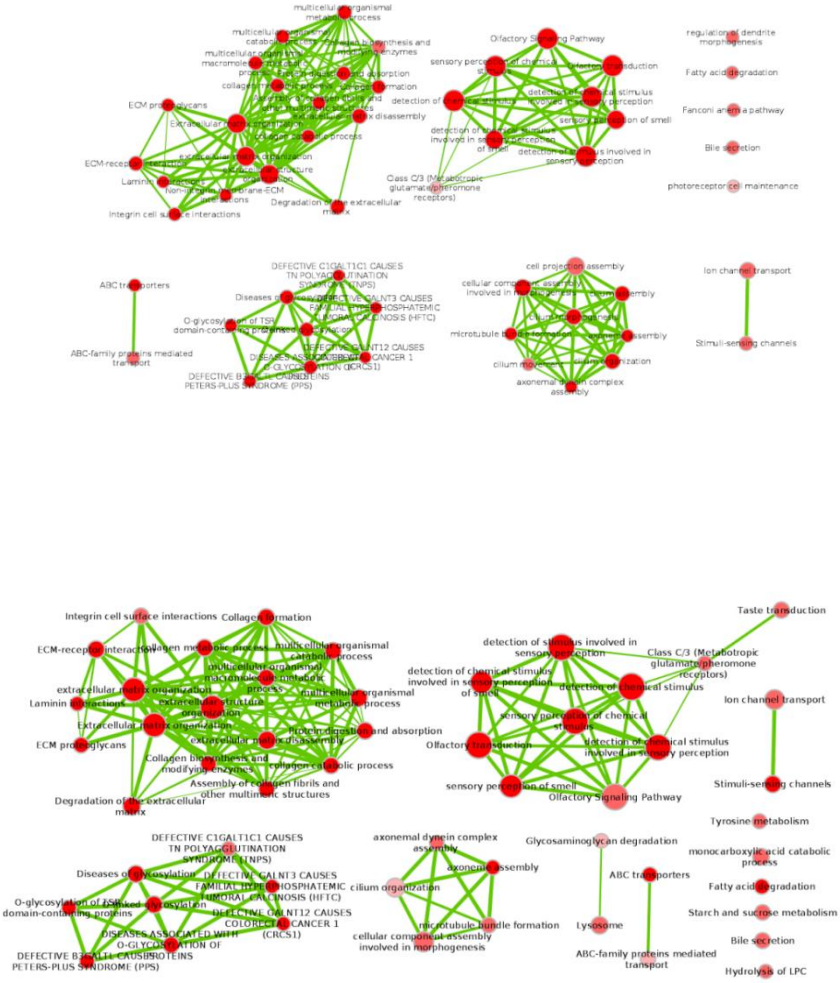


Figure 4.7. Pathway analysis of two samples (UKF-NB-3 up and clone 2 down). The pathways are represented as the edges of the network and the colour of the dots represents how much the pathway is likely to be affected in a scale white-

red. The green link represent interaction between pathways affected by mutated genes, and the thicker they are the more confident they are. Parental and clonal samples share most of the enriched pathways, but their networks are not identical.

4.4. Discussion

Cancer cell lines are a broadly used model system for cancer genomics research. Recent studies show that inter-tumour heterogeneity can be also observed in cancer cell lines in form of intra-cell line heterogeneity (Barranco et al., 1983; Zanker et al., 1982). Previous research in *Chapter 3* suggested that UKF-NB-3 was heterogeneous, and therefore that hypothesis was tested with this experiment.

Here we gained insights into the heterogeneity of UKF-NB-3 cell line, showing that it is not a homogeneous sample but it contains subpopulations with different mutation profiles. It may also not be completely stable, as there are still many de novo variants which are unique for each clonal sub-line and there are not two of them identical to each other. Both situations must be happening at the same time, as genetic drift is always an issue with cell lines, but no genetic drift would be able to cause so many variants in such a short period of time plus the lost variants in the clones (the mutations of the parental cell line which are not detected in the sub-lines) are highly unlikely mutational events, so they could not be explained only by the cell line being unstable. More experiments are needed to address this question: for example, a whole genome sequencing of UKF-NB-3 parental cell line could help us to address the question of how many structural variants and CNVs really are, and if they could be related to the highly number of variants occurring; also, single cell sequencing would allow us to better understand the subpopulations and high variability of this cell line.

Individually, the single cell derived clones contains a similar number of variants of each type, and their cancer signature does not differ at all from the parental UKF-NB-3 cell line. The differences between samples translated in a small set of different mutated genes in each of them, resulting in minor differences across their enriched

pathways. These new mutations are not randomly distributed across the genome, but follow a deterministic model where most of them occur in a small set of genes in every sample, while only a small proportion of de novo mutations being random.

The link between mutation and disease of the gained and de novo mutations was studied through their existing information in ClinVar and predictions from both SIFT and PolyPhen-2. Again, as happened in *Chapter 3*, the individual effect of each mutation is difficult to assess in a complex disease, and none of them gave us new information about UKF-NB-3 driver mutations.

The UKF-NB-3 parental cell line contained more variants called than any single cell derived clonal sub-line, and the quality of those variants was lower. Many of the variants present in the clones but not called in the parental cell line could be found at lowered quality threshold. This suggests that those mutations actually exist in the parental UKF-NB-3 cell line, probably in a small proportion of cells, such that they are not identified as only a small number of reads are supporting them are present. Therefore, sub-populations with different genotypes may coexist in the same cell line.

The study of this internal heterogeneity may be critical to understand the complex processes of cancer cell biology, cancer response to anti-cancer drugs and cancer differentiation processes. In future work we want to extend this research to drug-adapted UKF-NB-3 cell lines and compare how the genome changes to gain resistance to different drugs and between different concentrations of those drugs. Those experiments could be of great help to clarify the drug-resistance emergence processes in neuroblastoma, identify new biomarkers to monitor its development, personalise treatments depending on patient genetics.

Chapter 5:
Discussion

This Thesis has presented three scientific studies in the field of genetic variation. The first one considered genetic variation in Ebolaviruses and how they determine pathogenicity of the virus in humans. The second and third pieces of work detailed the genetic landscape and intra-cell line heterogeneity of UKF-NB-3 neuroblastoma cell line.

5.1. Genetic variance in Ebolavirus

The deadliest Ebolavirus outbreak in history was officially over on 2016. However, these viruses remain a threat to global public health. During the last outbreak, Ebolavirus was close to become a pandemic and it was the first outbreak where infected individuals spread beyond Africa. And today a new outbreak is ongoing in North Kivu, Democratic Republic of the Congo (Africa), where new genetic variants of the virus are continuously being catalogued.

Further, a new species, Bombali ebolavirus, was recently identified (Goldstein et al., 2018). It is now important that the genome of this species is compared to the existing Ebolavirus genomes to identify if, Bombali is pathogenic in humans, like most of the other Ebolavirus species, or if it does not cause disease, like Reston virus. This will help identify the public health concern that Bombali virus poses.

Our computational analysis of Ebolaviruses genomes in *Chapter 2* identified 189 SDPs which could be responsible for the differences in human pathogeny observed between these species. Only 47 SDPs could be mapped on to protein structures for the structural analysis, but they were enough to reveal important information as the multiple SDPs located in VP24-KPNA5 interface site, suggest that VP24 has an important role in determining species pathogeny. This hypothesis has been confirmed by other studies, and even it is being considered for a vaccine (Wilson, Bray, Bakken, & Hart, 2001). Finally, these findings open the door to a new way, they also suggest that Ebolavirus human pathogeny could be caused by just a few mutations.

This study was limited by the number of Ebolavirus genomes and the limited protein structure information available at its time. To illustrate how fast these resources grow, our analysis was based on 196 complete genomes while today more than 2000 Ebolavirus genomes are publically available. As these resources and our knowledge of Ebolavirus grow, a larger computational study of this type would be able to refine complete our study and provide a small set of SDPs – as the many of the 189 SDPs currently identified and unlikely to have a role in determining pathogenicity. Extrapolating to other infectious diseases, this kind of study could be used to predict inter-species virulence of closely related viruses based on computational analysis of sequence and protein structure (e.g. Zika virus could be compared between viruses identified in Africa that have not really caused disease with those infecting people in South America), which could be especially useful for other Risk Group 4 pathogens as their investigation is limited by the availability of appropriate containment laboratories.

5.2. Genetic variance in UKF-NB-3

Tumour heterogeneity is a characteristic of all solid cancers, where cells inside the same tumour have different cellular morphology, gene expression levels, metabolism, motility, proliferation and metastatic potential, all of them caused by genomic differences. The development of cancer genomics has unveiled the remarkable

genetic complexity of tumours, built by a multitude of subpopulations with different genetic variants. This extended genetic variability makes cancers even more difficult to characterise, and tumour heterogeneity is believed to play an important role in drug resistance.

Cancer cell lines are a broadly used model system for in vitro cancer genomics research, but until recent years they were believed to be homogenous populations once they became stable. Now we know that cancer cell lines are genetically heterogeneous, representing up to some extend intra-tumour heterogeneity. This makes them a potential research platform for the study of tumour differentiation and drug resistance emergence.

In our research we have used high throughput techniques to sequence the exome of UKF-NB-3, a high-risk neuroblastoma cell line. We have described its genetic landscape to a new level of detail, cataloguing all its mutations, analysing and predicting their effect in known cancer-relevant genes, estimating CNVs, looking for enriched pathways, and trying to extract as much information about its genomic structure as possible from WES data.

Previous experiments suggesting that UKF-NB-3 might be heterogeneous were backed up by some known variants of the cell line not being clearly called by some methods. Ten single cell derived clonal sub-lines of UKF-NB-3 were sequenced in order to test this hypothesis. We discovered that those clones not only differ from the parental UKF-NB-3, but shared only a common major subset of mutations with it while showing many others that were thought to be sequencing errors in UKF-NB-3 alignment and a third set of de novo mutations which made each clone unique. These findings suggest that not only UKF-NB-3 contains sub-populations of genetically different cells, but also is not completely stable thus every clonal sub-line differed from the others by many de novo mutations.

Our study was limited and biased by the WES data qualities, as this kind of sequencing technique gives no information about gene expression values, makes almost impossible to detect SVs, and gives no information about non-captured regions of the genome. This restricted the type of analysis we could perform,

software we could choose from and reliability of results as CNVs and enriched pathways analysis. Despite these limitations, we expect our research to bring new insights over high-risk neuroblastoma genomics and help future researchers working on UKF-NB-3 cell line to better understand the complex processes of cancer biology and cancer differentiation processes.

5.3. Future work

The study of internal cancer heterogeneity is of great importance in the study of cancer drug resistance. Understanding the genetic mutations responsible for its emergence during treatment and the ability to monitor its progress will be essential for effective future precision medicine therapies.

Following this line of research, we have already sequenced several sets of drug-adapted clones of UKF-NB-3 which we plan to analyse following the same methods established in our UKF-NB-3 analysis and its comparison with the single cell derived clonal sub-lines. These sets consist in ten vincristine resistant cell lines, 12 eribuline resistant cell lines, ten 2-methoxyestradiol resistant cell lines and nine epothilone B resistant cell lines. The four drugs are tubulin-binding agents, three of them with a destabiliser effect (vincristine, eribuline and 2-methoxystradiol) and one stabiliser agent (epothilone B), thus is it possible that the resistance mechanisms developed by the samples of different groups may share similarities, some of them even granting cross drug resistance or maybe making the cell vulnerable against the others.

By comparing these samples and using the single cell derived clones as controls to account for sample differences caused by UKF-NB-3 internal heterogeneity and potential instability, we expect to describe the differentiation processes occurring in a cancer cell line when exposed to the evolutionary pressure of anti-cancer drugs, and catalogue the differences and similarities between the paths followed by each sample compared both against same and different drugs adapted cell lines. Following this approach, we hope to identify biomarkers which allow us to monitor cancer resistance emergence and adapt treatments to each patient particularities.

In a preliminary analysis, we did not find major genomic differences between the groups of samples (Figure 5.1). We observed some genes that mutated more frequently in different sets of samples, but also the samples in each group differ from each other, resulting in some cases in bigger intra-group than inter-group differences. This suggests that our samples did not develop only one resistance mechanism against a particular drug, but several different of them could achieve the same result.

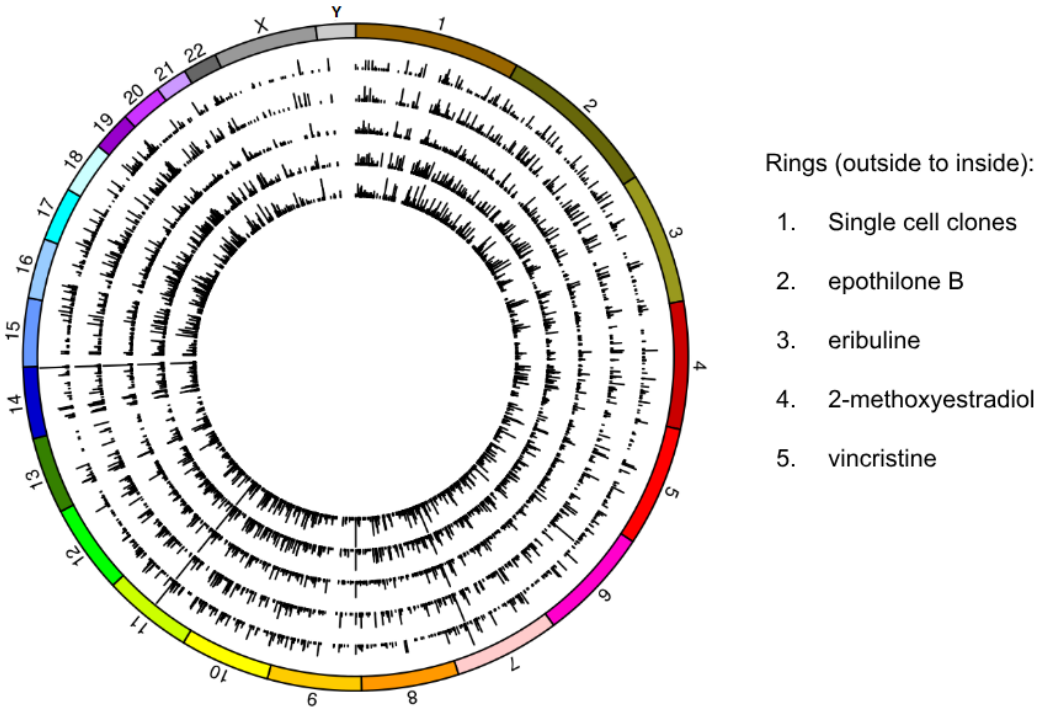
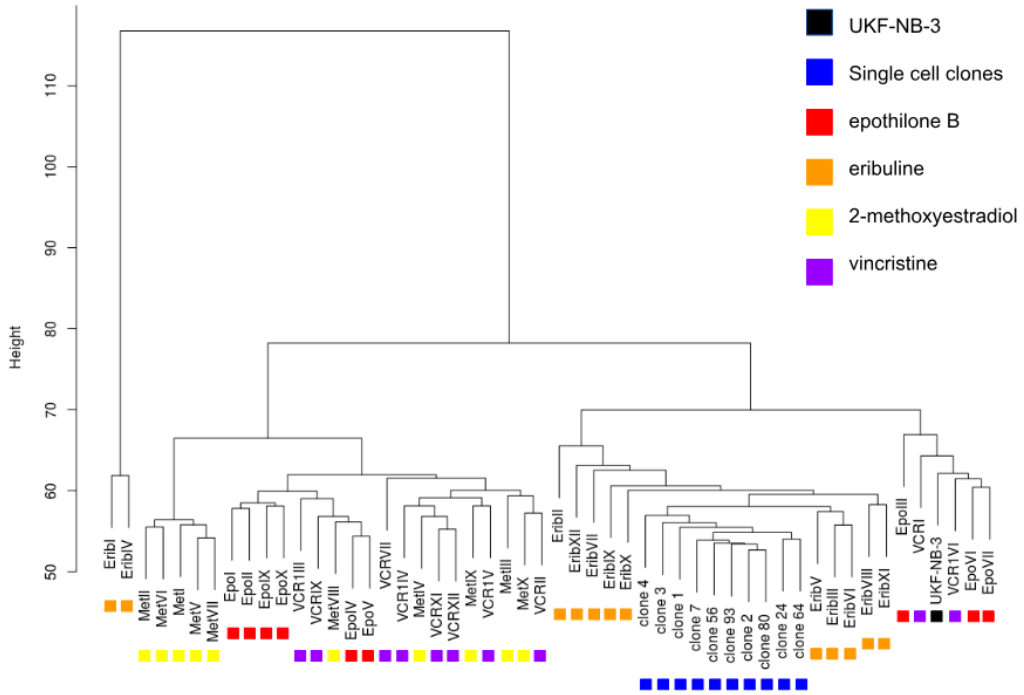


Figure 5.1. Distribution and frequencies of mutations in each group of samples. Circos plot showing the mutation profiles of the single cell-derived clonal sub-lines of UKF-NB-3 and the drug adapted sub-lines. From the most external to most internal tracks they refer to: the clonal UKF-NB-3 clone sub-lines, epothilone B resistant group, eribuline resistant group, 2-methoxiestradiol group and vincristine group. The black bars in each track indicate the presence of a mutation in their position, while the high of the bar indicates the proportion of cell lines of that sample group sharing that particular mutation.

Also, as some samples from different sets are more similar to each other than to the other same drug-resistant sublines (Figure 5.2), we hypothesise that some drug

resistance mechanism could be shared or at least very similar between some drugs. This kind of mechanisms could be able to grant resistance to more than one drug at the same time, which makes their identification key for an effective therapy prescription. The non-drug-adapted single cell derived clones clustered together better than any other group, supporting that the mutations making other sample sets so heterogeneous are not caused by internal UKF-NB-3 heterogeneity or cell line instability, but for differentiation pressure introduced by the anti-cancer drug.



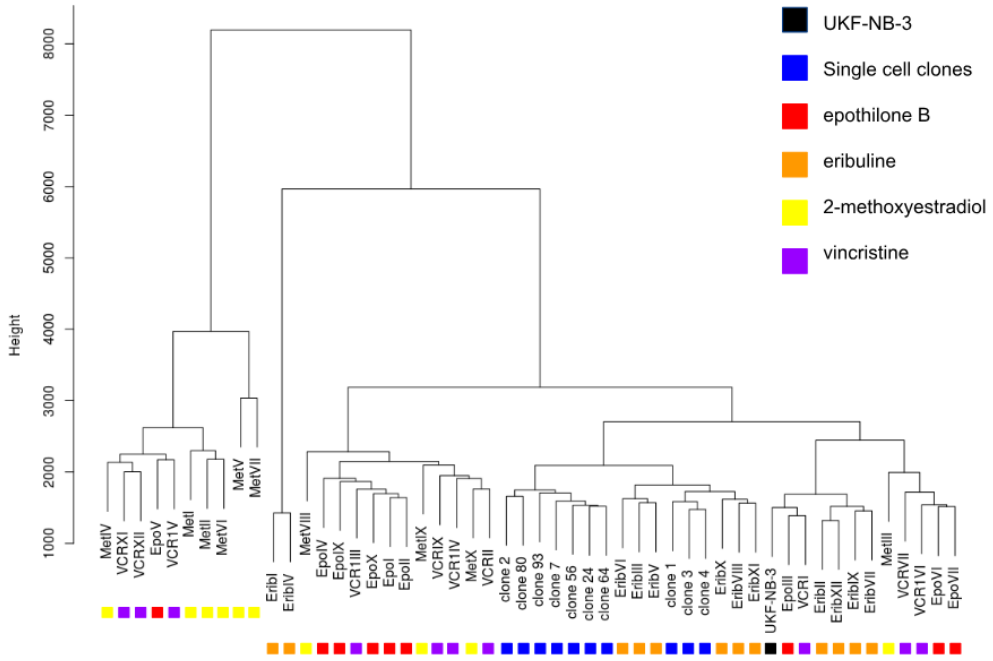


Figure 5.2. Phylogenies of the 52 samples. Bayesian (up) and Maximum Likelihood (down) phylogenetic trees of the UKF-NB-3 cell line, single cell derived clones and drug resistant sub-lines.

Finally, in a first attempt to identify where each dug applied their differentiation pressure, we run a enrichment pathway analysis on each group of samples (Table 5.1). As explained in *Chapter 4*, there are some genes more prompt to mutate due to the genomic qualities of UKF-NB-3, and also the enriched pathways in the parental cell line observed in *Chapter 3* are inherited by the sub-lines. Extracellular matrix organization pathway, related to tubulin production, is heavily enriched in all four samples sets, as could be expected due to the four drugs being tubulin-binding agents. Apart of that, each drug-resistant group have their own enriched pathways, but some of them are shared between groups, as the Dorso-ventral axis formation in eribuline and 2-methoxiestradiol, which could translate into the existence of shared drug resistance mechanism between the groups.

Most commonly enriched pathways among cell line types				
single cell clones	**eribuline**	**vincristine**	**2-methoxiestradiol**	**epothilone B**
Dectin-2 family	Extracellular matrix organization	Dectin-2 family	Extracellular matrix organization	RNA transport - Homo sapiens (human)
O-linked glycosylation of mucins	Dorso-ventral axis formation - Homo sapiens (human)	Adefovir Dipivoxil Metabolism Pathway	Dorso-ventral axis formation - Homo sapiens (human)	Extracellular matrix organization
Termination of O-glycan biosynthesis	BMP signaling Dro	Extracellular matrix organization	Activation of RAS in B cells	Translation Factors
O-linked glycosylation	Collagen formation	Tenofovir Metabolism Pathway	Adefovir Dipivoxil Metabolism Pathway	Collagen chain trimerization
NOTCH-Ncore	Dectin-2 family	Class C/3 (Metabotropic glutamate/pheromone receptors)	Alpha6Beta4Integrin	Dectin-2 family

Table 5.1. Most enriched pathways across groups of samples. Top 5 most enriched pathways in each set of samples.

This is a very promising research line, and depending on the results we obtain it could be easily extended by adding to the comparison new sets of UKF-NB-3 sub-lines resistant to other anticancer drugs or different concentrations of the same drug. Furthermore, these anticancer drugs are not only used for neuroblastoma, so new cancer cell lines and their drug-adapted sublines could be incorporated to the analysis in order to compare similarities between the differentiation processes which end in

drug resistance emergence. By doing that, we could catalogue them into profiles depending on the biomarkers specific to each drug, resistance mechanism, and cancer type. These profiles could have an important effect not only in future research, but also could evolve into clinical application for precision medicine, allowing detection and monitoring of drug-resistance emergence process during treatment before it fully develops, adapt therapies when it happens and even discover new cancer treatments based on drugs cross resistance-sensitiveness mechanisms.

5.4. Conclusion

In the previous chapters we have seen the importance of genetic variance analysis in two completely different diseases, Ebola and neuroblastoma, and a multitude of different types of information that can be extracted from these kind of data analysis in each case. Still, our studies were limited by the amount of publicly available genomic data, the quality of the sequencing data, type of sequencing, and tools for their analysis.

In Chapter 2 for example, by translating genomic information and its variants to protein sequence and structure, we have found a brand new application for the S3Det algorithm (Rausell et al., 2010), which original purpose was to identify functional residues in protein families. Applying this software, we have studied pathogenicity of a virus in a particular specie when comparing its protein sequences with the protein sequences of similar viruses already known to be pathogenic or non-pathogenic for the same species. Following this line of work, in (Martell et al., 2019) this idea is tested and our approach used to predict pathogenicity of Bombali Ebolavirus in humans.

As the amount of public genomic data grows, the possibilities of this new method correctly predicting pathogenicity in new species do the same, opening the door to a near future when new health risks caused by newly discovered, mutated or modified organisms, for both humans or other species, could be solved quickly and efficiently

in a dry lab with the help of sequencing technologies and a bioinformatics analysis similar to the one described in this work.

A second advantage of this method is the new insights we can get from infectious process by studying the location of SDPs in the predicted protein structures. While still only a minimum part of the known proteins has a known protein structure and the prediction of protein structure computationally remains and unsolved problem, both for lack of computational resources and complete understanding of how this process works in every protein, we can at least predict new candidate aminoacid and protein regions which can be related to it. With this predictions, new candidates for aimed experiments can be suggested, reducing the vast amount of potential paths of infection to study to only a few. This would reduce both cost and time needed to perform.

In Chapters 3 and 4 a more common variant calling analysis was used to study the variants in our cell lines. While broadly used, there is not a unified workflow for this kind of analysis yet. Despite GATK (McKenna et al., 2010) has published some best practices, it remains an experimental procedure that has to be individually adapted to each experiment depending on many factors: species, type of sequencing, quality, depth of coverage, available variant databases, … This arises many doubts, especially when used in personalised medicine or forensic exams, as depending on the method used to process the data the results may vary. For this reason, especially in complex diseases like cancer where many factors can be related with the disease, the descriptive genetic variant studies are key to improve our understanding and the amount of data available for researchers, and sequencing and variant calling experiments for describing genetic landscapes of cancers like this one play a huge role in this process, providing tons of data for future research.

The potential of this method for variants discovery is remarkable, and it is a key technique to identify, classify and study all kinds of genetic related diseases. In the future, as the methods and technology improves, the reliability of variants identified will do too, which will bring a new revolution not only to the medical field, where personalised medicine will erupt and genetic diagnoses will become far more

common, fast and cheap. All the genomic information previously known will be ready to be use for new diagnosis methods, identifying biomarkers to direct therapy and monitor the disease evolution, and discover potential new treatments.

But also to the economic and public health spheres will benefit of these technologies with the development of one health (https://www.who.int/features/qa/one-health/en/). This will bring sequencing and variant calling pipelines out of their usual medical fields of application to every industry related with food production, importation, processing and distribution, as genetic analysis could be able to identify the precedence and microbes of any sample of food without any possibilities for fraud.

In conclusion, the analysis of genetic variants is a big challenge with potential great rewards. There are many types of information that can be extracted from variants analysis, each of them with different applications as the diverse examples that have been presented in this Thesis. And this is only the beginning, as every day new discoveries in the field open new doors and new bigger challenges never before thought to be possible become a reality: from the first bacteriophage genome, passing through the human genome to the future of one health when everything will be sequenced in almost real-time. This field of study acts in a vicious circle like style, as new discoveries provides with new and better methods and data to work with, allowing us to ask more questions with each new discovery, and helping us to understand the genome one step at a time. But every step cames faster than ever before.

References:

Altman, R. B. (2012). Principle of Pharmacogenomics and Pramcofenomics. *Cambridge University Press.*

Abraham, M. J., Murtola, T., Schulz, R., Páll, 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. https://doi.org/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. In *Current Protocols in Human Genetics*. https://doi.org/10.1002/0471142905.hg0720s76

Akerlund, E., Prescott, J., & Tampellini, L. (2015). Shedding of Ebola Virus in an Asymptomatic Pregnant Woman. *New England Journal of Medicine*, *372*(25), 2467–2469. https://doi.org/10.1056/NEJMc1503275

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*, 10001. https://doi.org/10.1038/ncomms10001

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Lacroute, P. (2012). An integrated map of genetic variation

from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://bioinformatics.babraham.ac.uk/projects/fastqc

Apweiler, R., Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., … Zhang, J. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, *42*(D1), 191–198. https://doi.org/10.1093/nar/gkt1140

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*(18), 10385–10388. https://doi.org/10.1128/JVI.01452-13

Barranco, S. C., Townsend, C. M., Quraishi, M. A., Burger, N. L., Nevill, H. C., Howell, K. H., & Boerwinkle, W. R. (1983). Heterogeneous responses of an in vitro model of human stomach cancer to anticancer drugs. *Investigational New Drugs*, *1*(2), 117–127. https://doi.org/10.1007/BF00172070

Barrette, R. W., Metwally, S. A., Rowland, J. M., Xu, L., Zaki, S. R., Nichol, S. T., … Mcintosh, M. T. (2009). Discovery of swine as a host for the reston ebolavirus. *Science*, *325*(5937), 204–206. https://doi.org/10.1126/science.1172705

Basler, C. F. (2014). Portrait of a killer: Genome of the 2014 EBOV outbreak strain. *Cell Host and Microbe*, *16*(4), 419–421. https://doi.org/10.1016/j.chom.2014.09.012

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

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. https://doi.org/10.1038/nature05874

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

https://doi.org/10.1093/bioinformatics/btu170

Bordner, A. J., & Zorman, B. (2013). *Predicting non-neutral missense mutations and their biochemical consequences using genome-scale homology modeling of human protein complexes*. (1).

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*(4), 763–774. https://doi.org/10.1016/j.cell.2013.07.015

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4), 1–7. https://doi.org/10.1371/journal.pcbi.1003537

Bowman, L. C., Castleberry, R. P., Vijay Joshi, A. C., Cohn, S. L., Smith, E. I., Yu, A., … (1997) Genetic Staging of Unresectable or Metastatic Neuroblastomain Infants: a Pediatric Oncology Group Study, *JNCI: Journal of the National Cancer Institute*, Volume 89, Issue 5, 5 March 1997, Pages 373–380, https://doi.org/10.1093/jnci/89.5.373

Brathwaite, M. D., Wolman, S. R., Dalla-favera, R., Simon, M. I., & Gallo, R. C. (1984). Amplification of N-myc in Untreated Human Neuroblastomas PtetN-myc. *Department OfPathology and Kaplan Cancer Center, New York University School of Medicine, New York 10016*, *5177*(June), 1121–1124. https://doi.org/doi:10.1126/science.6719137

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., … Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, *18*(6), 630–634. https://doi.org/10.1038/76469

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*, 1–15. https://doi.org/10.1186/1471-2105-8-301

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., & Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics*, *29*(15), 1843–1850. https://doi.org/10.1093/bioinformatics/btt308

Castleberry RP, K. L. (1991). Radiotherapy improves the outlook for patients older than 1 year with Pediatric Oncology Group stage C neuroblastoma. *J Clin Oncol*, 9 (5): 789-95.

Castleberry RP, S. J. (1992). Infants with neuroblastoma and regional lymph node metastases have a favorable outlook after limited postoperative chemotheratpy: a Pediatric Oncology Group study. *J Clon Oncol*, 10 (8): 1299-304

Chambers, J. C., Zhang, W., Sehmi, J., Li, X., Wass, M. N., Van Der Harst, P., … Kooner, J. S. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genetics*, *43*(11), 1131–1138. https://doi.org/10.1038/ng.970

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Identification of Deleterious Mutations within Three Human Genomes.*, *19*(9), 1553–1561. https://doi.org/10.1101/gr.092619.109.2001

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270. https://doi.org/10.1038/nnano.2009.12

Clifton, M. C., Kirchdoerfer, R. N., Atkins, K., Abendroth, J., Raymond, A., Grice, R., … Saphire, E. O. (2014). Structure of the Reston ebolavirus VP30 C-terminal domain. *Acta Crystallographica Section F:Structural Biology Communications*, *70*(4), 457–460. https://doi.org/10.1107/S2053230X14003811

Consortium, T. I. H. (2007). A second generation human haplotype map of over 3.1 million SNPs. *October*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258.A

Dahlmann, F., Biedenkopf, N., Babler, A., Jahnen-Dechent, W., Karsten, C. B., Gnirß, K., … Hofmann-Winkler, H. (2015). Analysis of Ebola Virus Entry into Macrophages. *Journal of Infectious Diseases*, *212*(Suppl 2), S247–S257. https://doi.org/10.1093/infdis/jiv140

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. https://doi.org/10.1002/humu.21656

de La Vega, M.-A., Wong, G., Kobinger, G. P., & Qiu, X. (2015). The Multiple Roles of sGP in Ebola Pathogenesis. *Viral Immunology*, *28*(1), 3–9.

https://doi.org/10.1089/vim.2014.0068

de Magalhães, J. P., Finch, C. E., & Janssens, G. (2010). Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions. *Ageing Research Reviews*, *9*(3), 315–323. https://doi.org/10.1016/j.arr.2009.10.006

DePristo, M. a., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next- generation DNA sequencing data. *Nat Genet*, *43*(5), 491–498. https://doi.org/10.1038/ng.806.A

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., … Maris, J. M. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, *459*(7249), 987–991. https://doi.org/10.1038/nature08035

Dowall, S. D., Matthews, D. A., Garcia-Dorival, I., Taylor, I., Kenny, J., Hertz-Fowler, C., … Hiscox, J. A. (2014). Elucidating variations in the nucleotide sequence of Ebola virus associated with increasing pathogenicity. *Genome Biology*, *15*(11), 540. https://doi.org/10.1186/s13059-014-0540-x

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., … Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, *327*(5961), 78–81. https://doi.org/10.1126/science.1181498

Ebihara, H., Takada, A., Kobasa, D., Jones, S., Neumann, G., Theriault, S., … Kawaoka, Y. (2006). Molecular determinants of Ebola virus virulence in mice. *PLoS Pathogens*, *2*(7), 0705–0711. https://doi.org/10.1371/journal.ppat.0020073

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. 11*(6), 446–450. https://doi.org/10.1038/nrg2809.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., … Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, *518*(7539), 422–426. https://doi.org/10.1038/nature13952

Ewing, B, Hillier, L., Wendl, M. C., … Green, P. (2005). Base-Calling of Automated Sequencer Traces Using. *Genome Research*, (206), 175–185.

https://doi.org/10.1101/gr.8.3.175

Feingold, E. A., Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., Wetterstrand, K., & Collins, F. S. (2004). *ENCODE Project Consortium ENCODE*. 36–39.

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

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Filovir. (s.f.). Filovir. http://www.filovir.com/cms/

Friedman, G. K. (2007). Changing trends of research and treatment in infant neuroblastoma. *Pediatric Blood & Cancer,* 49(S7), 1060-1065.

Garnett, M. J., & McDermott, U. (2014). The evolving role of cancer cell line-based screens to define the impact of cancer genomes on drug response. *Current Opinion in Genetics and Development*, *24*(1), 114–119. https://doi.org/10.1016/j.gde.2013.12.002

Gebretadik FA, S. M. (2015). Review on Ebola Virus Disease: Its Outbreak and Current Status. *Epidemiology (sunnyvale).* 5:204. https://doi.org/10.4172/2161-1165.1000204

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., … Lander, E. S. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, *345*(6202), 1369. https://doi.org/10.1126/science.1259657

Gojo, I., & Karp, J. E. (2014). New strategies in acute myelogenous leukemia: Leukemogenesis and personalized medicine. *Clinical Cancer Research*, *20*(24), 6233–6241. https://doi.org/10.1158/1078-0432.CCR-14-0900

Goldstein, T., Anthony, S. J., Gbakima, A., Bird, B. H., Bangura, J., Tremeau-Bravard, A., … Mazet, J. A. K. (2018). The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. *Nature Microbiology*. https://doi.org/10.1038/s41564-018-0227-2

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., … Stuart, J. M. (2013). IntOGen-mutations identifies cancer

drivers across tumor types. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Griffith, B. Y. F. (1927). The Significance of Pneumococcal. *Occurrence of a Variety of Serological Types in the Sputum from an individual case of pneumonia* . *XXVII*(13), 129–176.

Groseth, A., Marzi, A., Hoenen, T., Herwig, A., Gardner, D., Becker, S., … Feldmann, H. (2012). The Ebola Virus Glycoprotein Contributes to but Is Not Sufficient for Virulence In Vivo. *PLoS Pathogens*, *8*(8). https://doi.org/10.1371/journal.ppat.1002847

Harrison, S. M., Riggs, E. R., Maglott, D. R., Lee, J. M., Azzariti, D. R., Niehaus, A., … Rehm, H. L. (2016). Using ClinVar as a resource to support variant interpretation. *Current Protocols in Human Genetics*, *2016*(April), 8.16.1-8.16.23. https://doi.org/10.1002/0471142905.hg0816s89

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*, *104*(2), 624–629. https://doi.org/10.1073/pnas.0606730104

Hartlieb, Bettina, 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*(43), 41830–41836. https://doi.org/10.1074/jbc.M307036200

Herbert, A. S., Davidson, C., Kuehne, A. I., Bakken, R., Braigen, S. Z., Gunn, K. E., … Dye, M. (2015). *Pathogenesis In Vivo*. *6*(3), 1–12. https://doi.org/10.1128/mBio.00565-15.Editor

Hoenen, T., Marzi, A., Scott, D. P., Feldmann, F., Callison, J., Safronetz, D., … Feldmann, H. (2015). Soluble Glycoprotein Is Not Required for Ebola Virus Virulence in Guinea Pigs. *Journal of Infectious Diseases*, *212*, S242–S246. https://doi.org/10.1093/infdis/jiv111

Holbeck, S. L., Collins, J. M., & Doroshow, J. H. (2010). Analysis of Food and Drug Administration-Approved Anticancer Agents in the NCI60 Panel of Human Tumor Cell Lines. *Molecular Cancer Therapeutics*, *9*(5), 1451–1460. https://doi.org/10.1158/1535-7163.MCT-10-0106

Hollingsworth, M. A., & Swanson, B. J. (2004). Mucins in cancer: Protection and control of the cell surface. *Nature Reviews Cancer*, *4*(1), 45–60. https://doi.org/10.1038/nrc1251

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., & Le Mercier, P. (2011). ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Research*, *39*(SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

Ilinykh, P. A., Lubaki, N. M., Widen, S. G., Renn, L. A., Theisen, T. C., Rabin, R. L., … Bukreyev, A. (2015). Different Temporal Effects of Ebola Virus VP35 and VP24 Proteins on Global Gene Expression in Human Dendritic Cells. *Journal of Virology*, *89*(15), 7567–7583. https://doi.org/10.1128/JVI.00924-15

International, T., & Consortium, H. (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., & Swanton, C. (2015). Translational implications of tumor heterogeneity. *Clinical Cancer Research*, *21*(6), 1258–1266. https://doi.org/10.1158/1078-0432.CCR-14-1429

Joosten, R. P., Te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., … Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(SUPPL. 1), 411–419. https://doi.org/10.1093/nar/gkq1105

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. (2011). ConsensusPathDB: Toward a more complete picture of cell biology. *Nucleic Acids Research*, *39*(SUPPL. 1), 712–717. https://doi.org/10.1093/nar/gkq1156

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

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

Kaur, G., & Dufour, J. M. (2012). Cell lines. *Spermatogenesis*, *2*(1), 1–5. https://doi.org/10.4161/spmg.19885

Kelly, L. A., Mezulis, S., Yates, C., Wass, M., & Sternberg, M. (2015). The Phyre2

web portal for protein modelling, prediction, and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015-053

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 of the National Academy of Sciences*, *107*(1), 314–319. https://doi.org/10.1073/pnas.0910547107

Kotchetkov, R., Drjever, P. H., Cinatl, J., Michaelis, M., Karaskova, J., Blaheta, R., … Cinatl, J. (2005). Increased malignant behavior in neuroblastoma cells with acquired multi-drug resistance does not depend on P-gp expression. *International Journal of Oncology*, *27*(4), 1029–1037. https://doi.org/10.3892/ijo.27.4.1029

Kuhn, J. H., Becker, S., Ebihara, H., Geisbert, T. W., Johnson, K. M., Kawaoka, Y., … Jahrling, P. B. (2010). Proposal for a revised taxonomy of the family Filoviridae: Classification, names of taxa and viruses, and virus abbreviations. *Archives of Virology*, *155*(12), 2083–2103. https://doi.org/10.1007/s00705-010-0814-x

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., … Krijgsman, O. (2015). CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biology*, *16*(1), 1–15. https://doi.org/10.1186/s13059-015-0617-1

Kumar, M., Joseph, M., & Chandrashekaran, S. (2001). Effects of mutations at the stambh A locus of Drosophila melanogaster. *Journal of Genetics*, *80*(2), 83–95. https://doi.org/10.1007/BF02728334

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1082. https://doi.org/10.1038/nprot.2009.86

L.Hopkins, A., & R.Groom, C. (2005). The Druggable Genome. *Genome*, *1*(September), 7–10. https://doi.org/10.1038/nrd892

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., …

Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Leung, D. W., Ginder, N. D., Fulton, D. B., Nix, J., Basler, C. F., Honzatko, R. B., & Amarasinghe, G. K. (2009). Structure of the Ebola VP35 interferon inhibitory domain. *Proceedings of the National Academy of Sciences*, *106*(2), 411–416. https://doi.org/10.1073/pnas.0807854106

Leung, Daisy W., Borek, D., Luthra, P., Binning, J. M., Anantpadma, M., Liu, G., … Amarasinghe, G. K. (2015). An intrinsically disordered peptide from ebola virus VP35 controls viral RNA synthesis by modulating nucleoprotein-RNA interactions. *Cell Reports*, *11*(3), 376–389. https://doi.org/10.1016/j.celrep.2015.03.034

Leung, Daisy W., Shabman, R. S., Farahbakhsh, M., Prins, K. C., Borek, D. M., Wang, T., … Amarasinghe, G. K. (2010). Structural and functional characterization of Reston Ebola virus VP35 interferon inhibitory domain. *Journal of Molecular Biology*, *399*(3), 347–357. https://doi.org/10.1016/j.jmb.2010.04.022

Levy-Lahad, E., Lahad, A., & King, M.-C. (2014). Precision Medicine Meets Public Health: Population Screening for BRCA1 and BRCA2. *JNCI Journal of the National Cancer Institute*, *107*(1), dju420–dju420. https://doi.org/10.1093/jnci/dju420

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Jia, J., Li, H., Yu, J., Sun, H., He, Y., … Xie, L. (2014). SysPTM 2.0: An updated systematic resource for post-translational modification. *Database*, *2014*,

1–10. https://doi.org/10.1093/database/bau025

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., & Gerlinger, M. (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, *2*(1), 49–63. https://doi.org/10.1016/j.trecan.2015.11.003

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*, 1–9. https://doi.org/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). https://doi.org/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. https://doi.org/10.1016/j.meegid.2015.02.024

Löschmann, N., Michaelis, M., Rothweiler, F., Zehner, R., Cinatl, J., Voges, Y., … Cinatl, J. (2013). Testing of SNS-032 in a Panel of Human Neuroblastoma Cell Lines with Acquired Resistance to a Broad Range of Drugs. *Translational Oncology*, *6*(6), 685-IN18. https://doi.org/10.1593/tlo.13544

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Maris, J. M., Hogarty, M. D., Bagatell, R., & Cohn, S. L. (2007). Neuroblastoma. *Lancet*, *369*(9579), 2106–2120. https://doi.org/10.1016/S0140-6736(07)60983-0

Marsh, G. A., Haining, J., Robinson, R., Foord, A., Yamada, M., Barr, J. A., … Middleton, D. (2011). Ebola reston virus infection of pigs: Clinical significance and transmission potential. *Journal of Infectious Diseases*, *204*(SUPPL. 3). https://doi.org/10.1093/infdis/jir300

Martell, H. J., Masterson, S. G., McGreig, J. E., Michaelis, M., & Wass, M. N. (2019). Is the Bombali virus pathogenic in humans? *Bioinformatics*, (April), 1–6.

https://doi.org/10.1093/bioinformatics/btz267

Marx, V. (2014). Cell-line authentication demystified. *Nature Methods*, *11*(5), 483–488. https://doi.org/10.1038/nmeth.2932

Masramon, L. (2006). Genetic instability and divergence of clonal populations in colon cancer cells in vitro. *Journal of Cell Science*, *119*(8), 1477–1482. https://doi.org/10.1242/jcs.02871

Mateo, M., Carbonnelle, C., Martinez, M. J., Reynard, O., Page, A., Volchkova, V. A., & Volchkov, V. E. (2011). Knockdown of Ebola virus VP24 impairs viral nucleocapsid assembly and prevents virus replication. *Journal of Infectious Diseases*, *204*(SUPPL. 3), 892–896. https://doi.org/10.1093/infdis/jir311

Mateo, M., Carbonnelle, C., Reynard, O., Kolesnikova, L., Nemirov, K., Page, A., … Volchkov, V. E. (2011). VP24 Is a molecular determinant of Ebola virus virulence in guinea pigs. *Journal of Infectious Diseases*, *204*(SUPPL. 3), 1011–1020. https://doi.org/10.1093/infdis/jir338

Matthaei, J. H., & Nirenberg, M. W. (1961). Characteristics and stabilization of DNAase-sensitive protein synthesis in E. coli extracts. *Proceedings of the National Academy of Sciences*, *47*(190), 1580–1588. https://doi.org/10.1073/pnas.47.10.1580

Maxam, a M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564. https://doi.org/10.1073/pnas.74.2.560

McCall, E. E., Olshan, A. F., & Daniels, J. L. (2005). Maternal hair dye use and risk of neuroblastoma in offspring. *Cancer Causes and Control*, *16*(6), 743–748. https://doi.org/10.1007/s10552-005-1229-y

McGranahan, N., Swanton, C., Abkevich, V., Timms, K. M., Hennessy, B. T., Potter, J., … al., et. (2015). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell*, *27*(1), 15–26. https://doi.org/10.1016/j.ccell.2014.12.001

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0974-4

Mehedi, M., Falzarano, D., Seebach, J., Hu, X., Carpenter, M. S., Schnittler, H.-J., & Feldmann, H. (2011). A New Ebola Virus Nonstructural Glycoprotein Expressed through RNA Editing. *Journal of Virology*, *85*(11), 5406–5414. https://doi.org/10.1128/JVI.02190-10

Menegaux, F., Olshan, A. F., Neglia, J. P., Pollock, B. H., & Bondy, M. L. (2004). Day care, childhood infections, and risk of neuroblastoma. *American Journal of Epidemiology*, *159*(9), 843–851. https://doi.org/10.1093/aje/kwh111

Michaelis, M., Rothweiler, F., Agha, B., Barth, S., Voges, Y., Löschmann, N., … Cinatl, J. (2012). Human neuroblastoma cells with acquired resistance to the p53 activator RITA retain functional p53 and sensitivity to other p53 activating agents. *Cell Death and Disease*, *3*(4), e294. https://doi.org/10.1038/cddis.2012.35

Michaelis, M., Rothweiler, F., Barth, S., Cinatl, J., van Rikxoort, M., Löschmann, N., … Speidel, D. (2011). Adaptation of cancer cells from different entities to the MDM2 inhibitor nutlin-3 results in the emergence of p53-mutated multi-drug-resistant cancer cells. *Cell Death & Disease*, *2*, e243. https://doi.org/10.1038/cddis.2011.129

Miller, E. H., Obernosterer, G., Raaben, M., Herbert, A. S., Deffieu, M. S., Krishnan, A., … Chandran, K. (2012). Ebola virus entry requires the host-programmed recognition of an intracellular receptor. *EMBO Journal*, *31*(8), 1947–1960. https://doi.org/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 Diseases*, *204*(SUPPL. 3), 757–760. https://doi.org/10.1093/infdis/jir296

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 Research*, *41*(12). https://doi.org/10.1093/nar/gkt263

Molenaar, J. J., Koster, J., Zwijnenburg, D. A., Van Sluis, P., Valentijn, L. J., Van Der Ploeg, I., … Versteeg, R. (2012). Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, *483*(7391), 589–593.

https://doi.org/10.1038/nature10910

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*(5–6), 391–398. https://doi.org/10.1016/j.cimid.2007.05.005

Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods*, *10*(1), 47–53. https://doi.org/10.1038/nmeth.2289

Mossé, Y. P., Laudenslager, M., Longo, L., Cole, K. A., Wood, A., Attiyeh, E. F., … Maris, J. M. (2008). Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*, *455*(7215), 930–935. https://doi.org/10.1038/nature07261

Nelson, M., Wegmann, D., Ehm, M., Kessner, D., St Jean, P., Verzilli, C., … Zöllner, S. (2012). An Abundance of Rare Functional Sequenced in 14 , 002 People. *Science*, *337*(July), 100–104. https://doi.org/10.1126/science.1217876

Ng, P. C., Murray, S. S., Levy, S., & Venter, J. C. (2009). An agenda for personalized medicine. *Nature*, *461*(7265), 724–726. https://doi.org/10.1038/461724a

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., … Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*(7261), 272–276. https://doi.org/10.1038/nature08250

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

Olshan, A. F., Smith, J., Cook, M. N., Grufferman, S., Pollock, B. H., Stram, D. O., … Bondy, M. L. (1999). Hormone and fertility drug use and the risk of neuroblastoma: a report from the Children's Cancer Group and the Pediatric Oncology Group. *American Journal Of Epidemiology*, *150*(9), 930–938. https://doi.org/10.1093/oxfordjournals.aje.a010101

Olshan, A. F., & Bunin, G. R. (2000). Epiemiology of Neuroblastoma. *Springer*.

Osterkamp, H. M. O., Cheiner, L. S., Tefanova, M. C. S., Loyd, K. O. L., & Instad,

C. L. F. (1997). Comparisson of MUC-1 mucin exoression in epithelial and non-epithelial cancer cell lines and demosntrarion of a new short variant form (MUC-1 / Z). *94*(July 1996), 87–94.

Palles, C., Cazier, J. B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Rimmer, A. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136–143. https://doi.org/10.1038/ng.2503

Pappalardo, M., Julia, M., Howard, M. J., Rossman, J. S., Michaelis, M., & Wass, M. N. (2016). Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. *Scientific Reports*, *6*(March), 23743. https://doi.org/10.1038/srep23743

Pappalardo, M., Reddin, I. G., Cantoni, Di., Rossman, J. S., Michaelis, M., & Wass, M. N. (2017). Changes associated with Ebola virus adaptation to novel species. *Bioinformatics*, *33*(13), 1911–1915. https://doi.org/10.1093/bioinformatics/btx065

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Paul SR, T. N. (1991). Stage IV neuroblastoma in infants. Long-term survival. *Cancer*, 67 (6): 1493-7.

Peterson, T. A., Doughty, E., & Kann, M. G. (2013). Towards precision medicine: Advances in computational approaches for the analysis of human variants. *Journal of Molecular Biology*, *425*(21), 4047–4063. https://doi.org/10.1016/j.jmb.2013.08.008

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Picard. (s.f.) Picard. http://broadinstitute.github.io/picard

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., … Scheuermann, R. H. (2012). ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, *40*(D1), 593–598.

https://doi.org/10.1093/nar/gkr859

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. https://doi.org/10.1093/bioinformatics/btt691

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), 191–196. https://doi.org/10.1038/nature08658

Pleasance, E. D., & Stephens, P. J. (2010). A small cell lung cancer genome reports complex tobacco exposure signatures. *Nature*, *463*(7278), 184–190. https://doi.org/10.1038/nature08629.A

Pugh, T. J. (2013). The genetic landscape of high-risk neuroblastoma. *Nature Genetics*, *45*(3), 279–284. https://doi.org/10.1038/ng.2529.The

Pundir, S., Martin, M. J., & O'Donovan, C. (2016). UniProt Tools. *Current Protocols in Bioinformatics*, *53*(March), 1.29.1-1.29.15. https://doi.org/10.1002/0471250953.bi0129s53

Rakha, E. A., Boyce, R. W. G., El-Rehim, D. A., Kurien, T., Green, A. R., Paish, E. C., … Ellis, I. O. (2005). Expression of mucins (MUC1, MUC2, MUC3, MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer. *Modern Pathology*, *18*(10), 1295–1304. https://doi.org/10.1038/modpathol.3800445

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. https://doi.org/10.1073/pnas.0908044107

Reid, S. P., Leung, L. W., Hartman, A. L., Martinez, O., Shaw, M. L., Carbonnelle, C., … Basler, C. F. (2006). Ebola Virus VP24 Binds Karyopherin 1 and Blocks STAT1 Nuclear Accumulation. *Journal of Virology*, *80*(11), 5156–5167. https://doi.org/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 Proteins with Activated STAT1. *Journal of Virology*, *81*(24), 13469–

13477. https://doi.org/10.1128/JVI.01097-07

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E., & da Silva, I. T. (2016). signeR: An empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, (September), btw572. https://doi.org/10.1093/bioinformatics/btw572

Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., … Burley, S. K. (2015). The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Research*, *43*(D1), D345–D356. https://doi.org/10.1093/nar/gku1214

Rusk, N. (2011). Torrents of sequence. *Nature Methods*, *8*(1), 44. https://doi.org/10.1038/nmeth.f.330

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sharma, S. V., Haber, D. A., & Settleman, J. (2010). Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews Cancer*, *10*(4), 241–253. https://doi.org/10.1038/nrc2820

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Rev.*, *6*(10), 813–823. https://doi.org/10.1038/nrc1951

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(539). https://doi.org/10.1038/msb.2011.75

Smith, N., Witham, S., Sarkar, S., Zhang, J., Li, L., Li, C., & Alexov, E. (2012). DelPhi web server v2: Incorporating atomic-style geometrical figures into the computational protocol. *Bioinformatics*, *28*(12), 1655–1657. https://doi.org/10.1093/bioinformatics/bts200

Stamatakis, A. (2014). RAxML version 8: A tool for Phylogenetic Analysus and Post-Analysis of Phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. https://doi.org/10.1038/nature07943

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

Teng, Y., Wang, Y., Zhang, X., Liu, W., Fan, H., Yao, H., … Cao, W. (2015). Systematic Genome-wide Screening and Prediction of microRNAs in EBOV during the 2014 Ebolavirus Outbreak. *Scientific Reports*, *5*(March), 1–17. https://doi.org/10.1038/srep09912

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., … on behalf of the NHLBI Exome Sequencing Project. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science (New York, NY)*, *337*(6090), 64–69. https://doi.org/10.1126/science.1219240

Torsvik, A., Stieber, D., Enger, P. O., Golebiewska, A., Molven, A., Svendsen, A., … Bjerkvig, R. (2014). U-251 revisited: Genetic drift and phenotypic consequences of long-term cultures of glioblastoma cells. *Cancer Medicine*, *3*(4), 812–824. https://doi.org/10.1002/cam4.219

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., … Johnson, S. M. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, *18*(7), 1051–1063. https://doi.org/10.1101/gr.076463.108

Wang, K., Diskin, S. J., Zhang, H., Attiyeh, E. F., Winter, C., Hou, C., … Maris, J. M. (2011). Integrative genomics identifies LMO1 as a neuroblastoma oncogene. *Nature*, *469*(7329), 216–220. https://doi.org/10.1038/nature09609

Watson, J., & Crick, F. (1953). Molecular structure of nucleic acids. *Nature.*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Watt, A., Moukambi, F., Banadyga, L., Groseth, A., Callison, J., Herwig, A., … Hoenen, T. (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. https://doi.org/10.1128/JVI.01272-14

Wilson, J. A., Bray, M., Bakken, R., & Hart, M. K. (2001). Vaccine potential of Ebola virus VP24, VP30, VP35, and VP40 proteins. *Virology*, *286*(2), 384–390. https://doi.org/10.1006/viro.2001.1012

Weingartl, H. M. (2013). Review of Ebola virus infections in domestic animals. *Biol (Basel),* 135, 211-21.

West DC, S. R. (1993). Stage III neuroblastoma over 1 year of age at diagnosis: improved survival with intensive multimodality therapy including multiple alkylating agents. *J Clin Oncol,* 11 (1): 84-90.

Wheeler, K. (2015). Neuroblastoma in children. *McMillan.*

Wood, A. R., Perry, J. R. B., Tanaka, T., Hernandez, D. G., Zheng, H. F., Melzer, D., … Frayling, T. M. (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), 1–13. https://doi.org/10.1371/journal.pone.0064343

Xia, P., Choi, A. H., Deng, Z., Yang, Y., Wang, Y., Hardwidge, P. R., & Zhu, G. (2016). *Cell membrane-anchored MUC4 promotes tumorigenicity in epithelial carcinomas*.

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 and Microbe*, *16*(2), 187–200. https://doi.org/10.1016/j.chom.2014.07.008

Yaddanapudi, K., Palacios, G., Towner, J. S., Chen, I., Sariol, C. A., Nichol, S. T., & Lipkin, W. I. (2006). Implication of a retrovirus-like glycoprotein peptide in the immunopathogenesis of Ebola and Marburg viruses. *The FASEB Journal*, *20*(14), 2519–2530. https://doi.org/10.1096/fj.06-6151com

Yin, T., Cook, D., & Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biology*, *13*(8), R77. https://doi.org/10.1186/gb-2012-13-8-r77

Zanker, K. S., Treappe, A., & Blumel, G. (1982). *In-vitro resistance of cloned human glioma cells to Natural Killer activity of Allogeneic peripheral lymphocytes*. 617–624.

Zhang, A. P. P., Bornholdt, Z. A., Liu, T., Abelson, D. M., Lee, D. E., Li, S., … Saphire, E. O. (2012). The ebola virus interferon antagonist VP24 directly binds STAT1 and has a novel, pyramidal fold. *PLoS Pathogens*, *8*(2). https://doi.org/10.1371/journal.ppat.1002550

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., … Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, *111*(4), E455–E464. https://doi.org/10.1073/pnas.1322563111

Acknowledgements:

This Thesis is product of the help and support of many people. First of all, I would like to express my sincere gratitude to both my supervisors, Dr. Mar Wass and Prof. Martin Michaelis. Their continuous mentoring, patience, motivation and knowledge guided me during my years as PhD student, and I would definitely not be where I stand today without their help. I would also like to thank to the entire Wass lab, the loyal colleges who shared with me every moment of joy and offered their support whenever I needed during this journey. Also, I wish to thank everyone that was not part of the Wass lab but still dropped by time to time for jokes, food and helpful conversations. Last but not least, I want to thank all my family and friends, all those amazing people for supporting me spiritually throughout writing this thesis and my life in general.

Annex 1:

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).

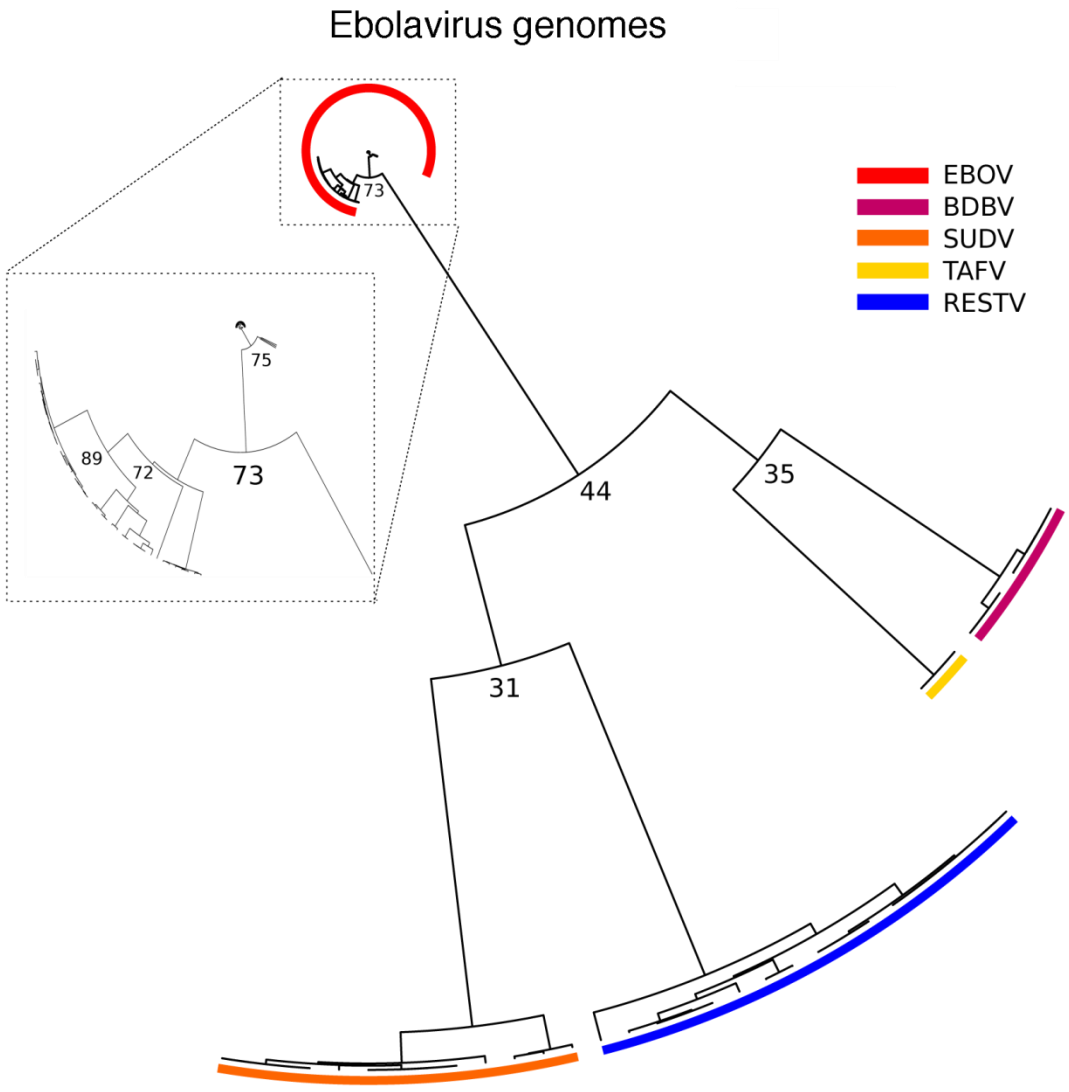


Fig S1A. Bayesian tree for whole genomes.

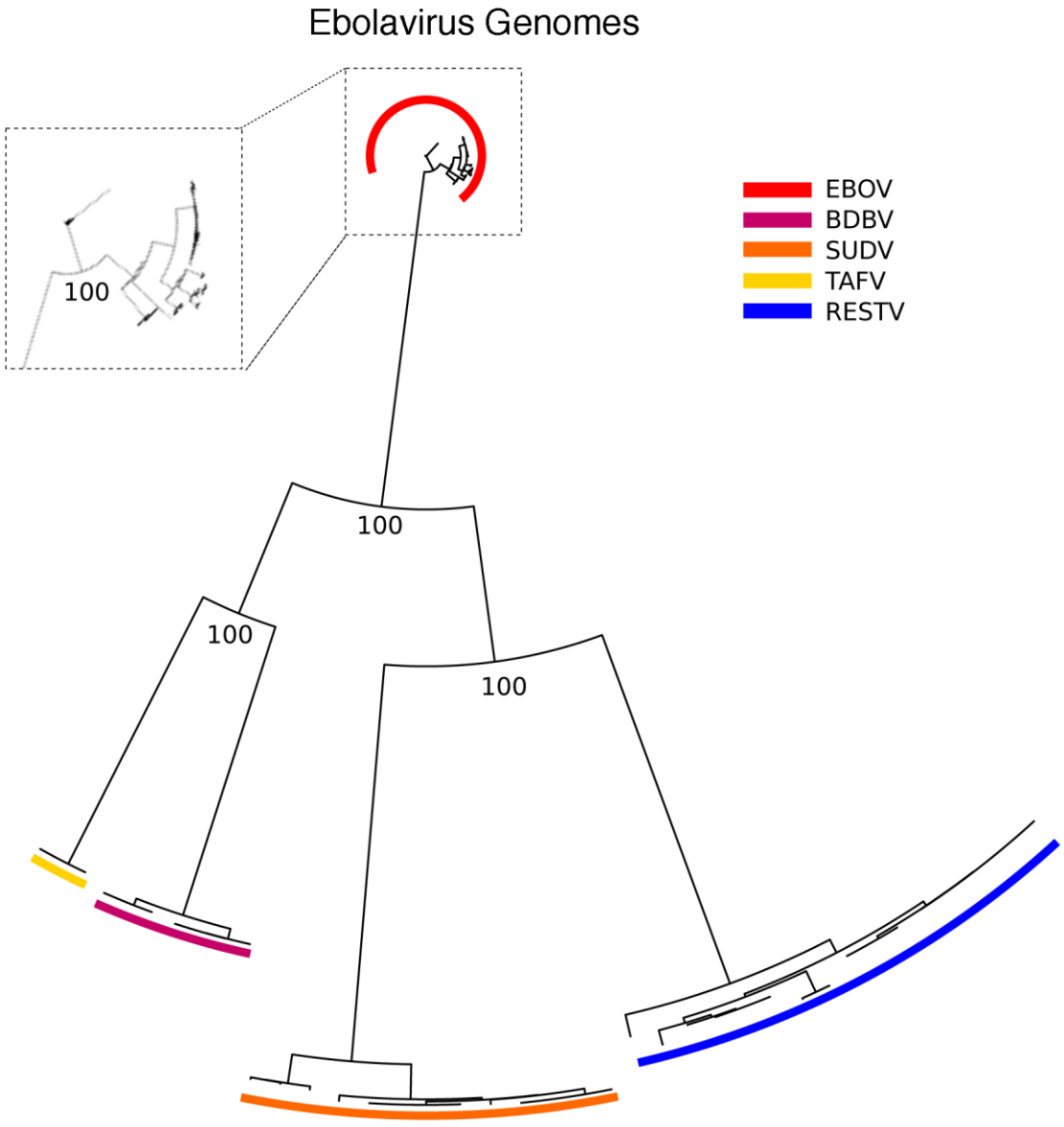


Fig S1B. Maximum likelihood tree for whole genomes.

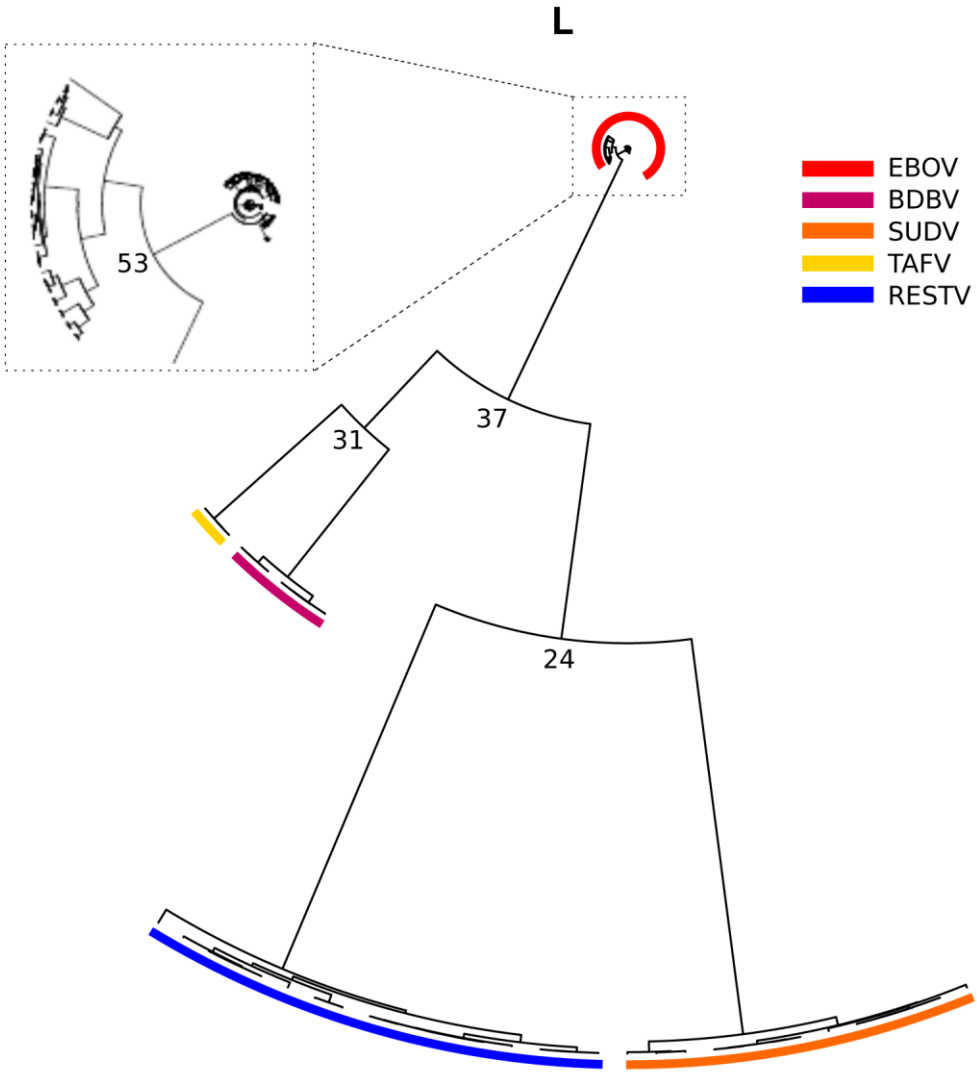


Fig S1C. Bayesian tree for protein L.

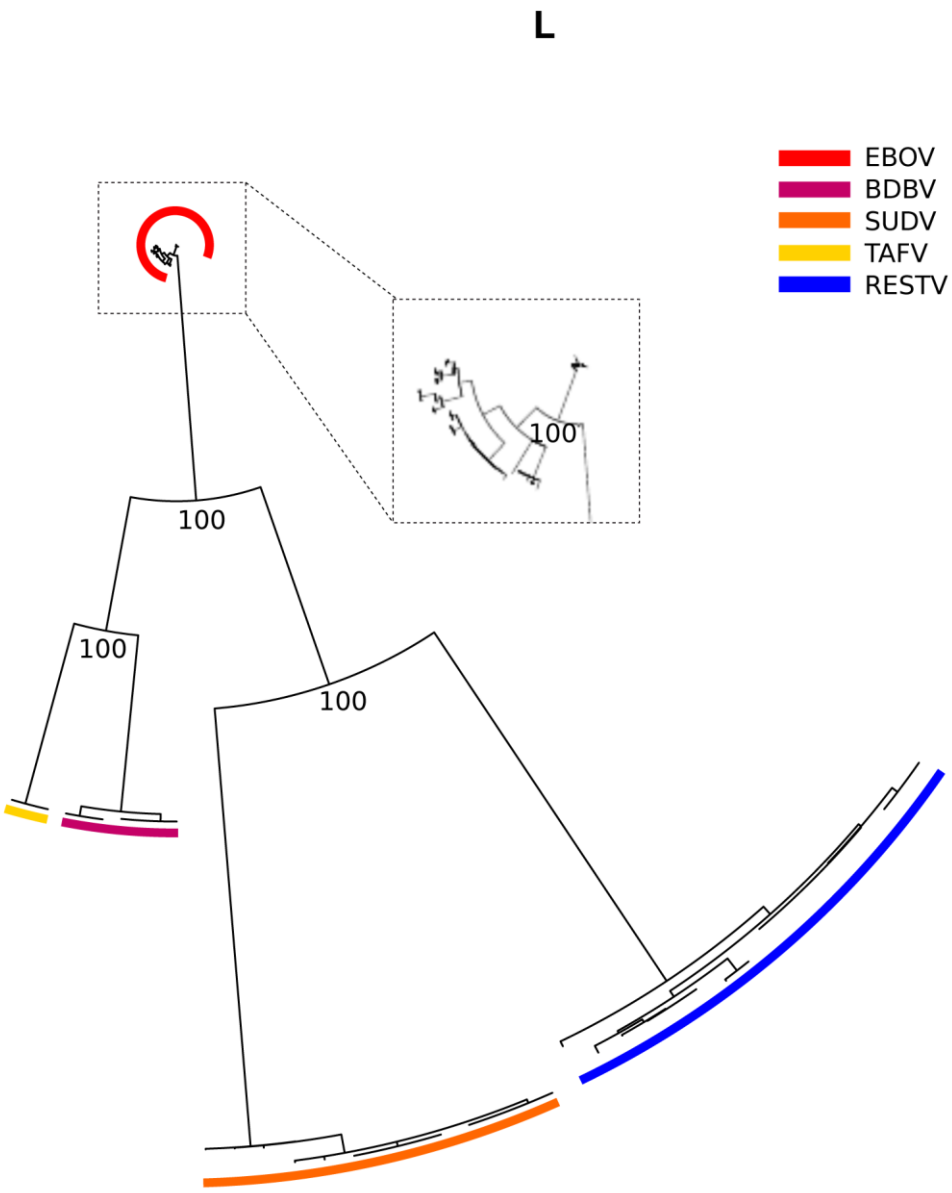


Fig S1D. Maximum likelihood tree for protein L.

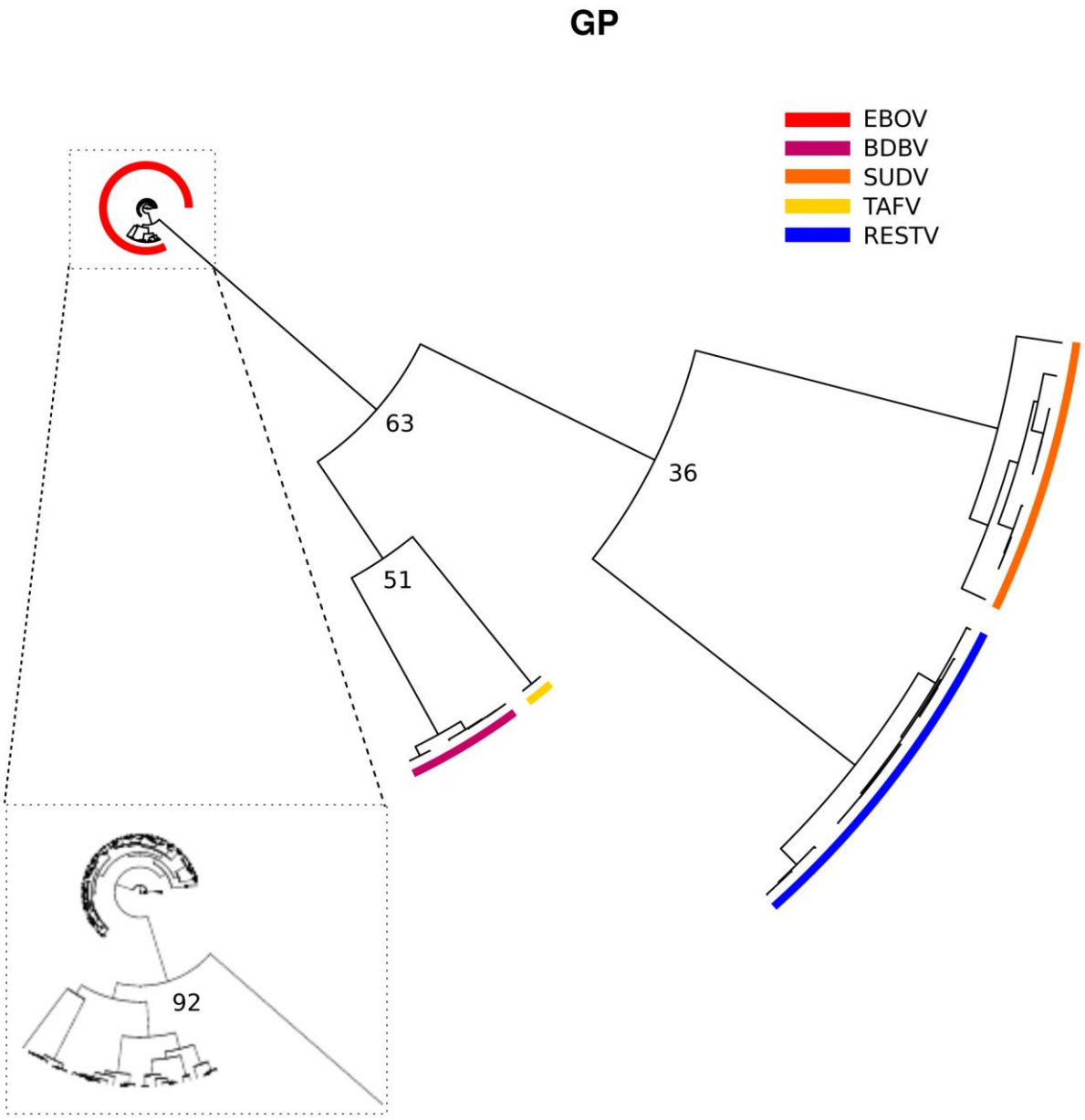


Fig S1E. Bayesian tree for protein GP.

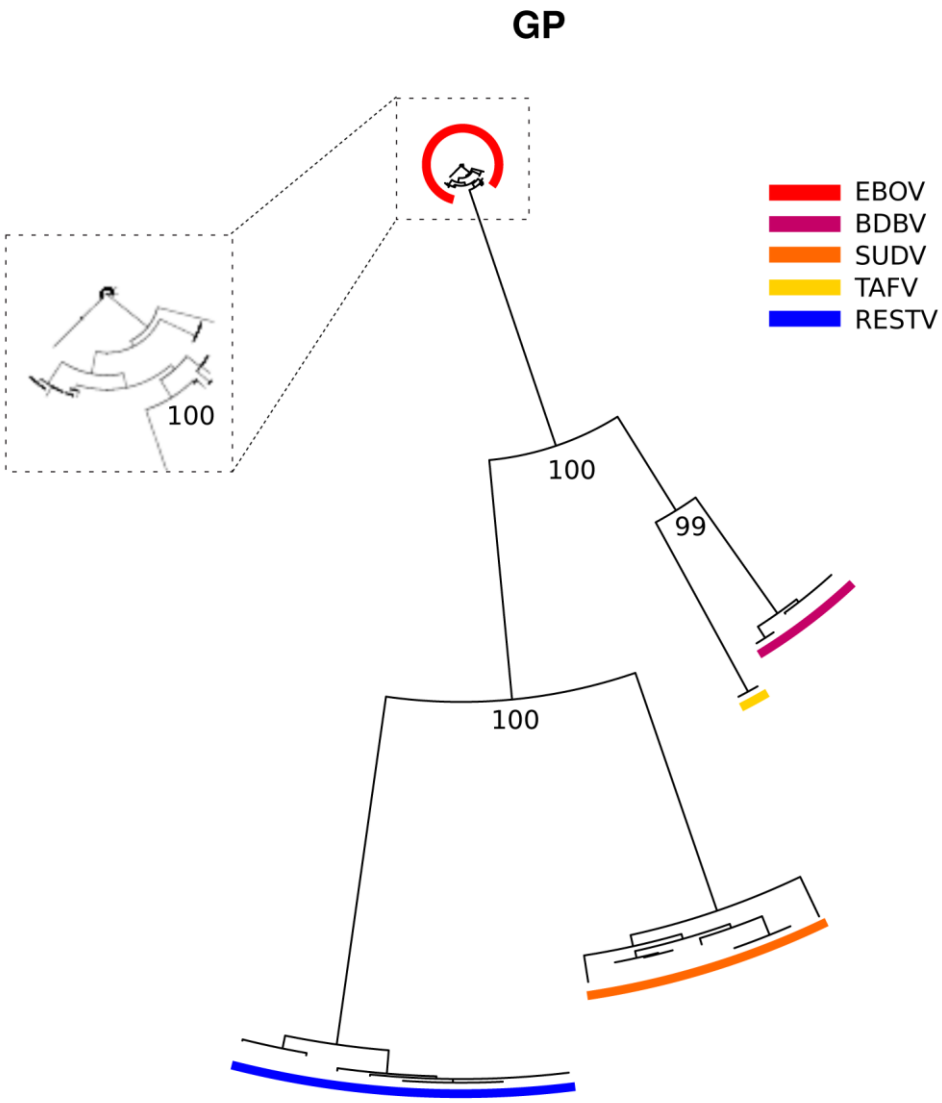


Fig S1F. Maximum likelihood tree for protein GP.

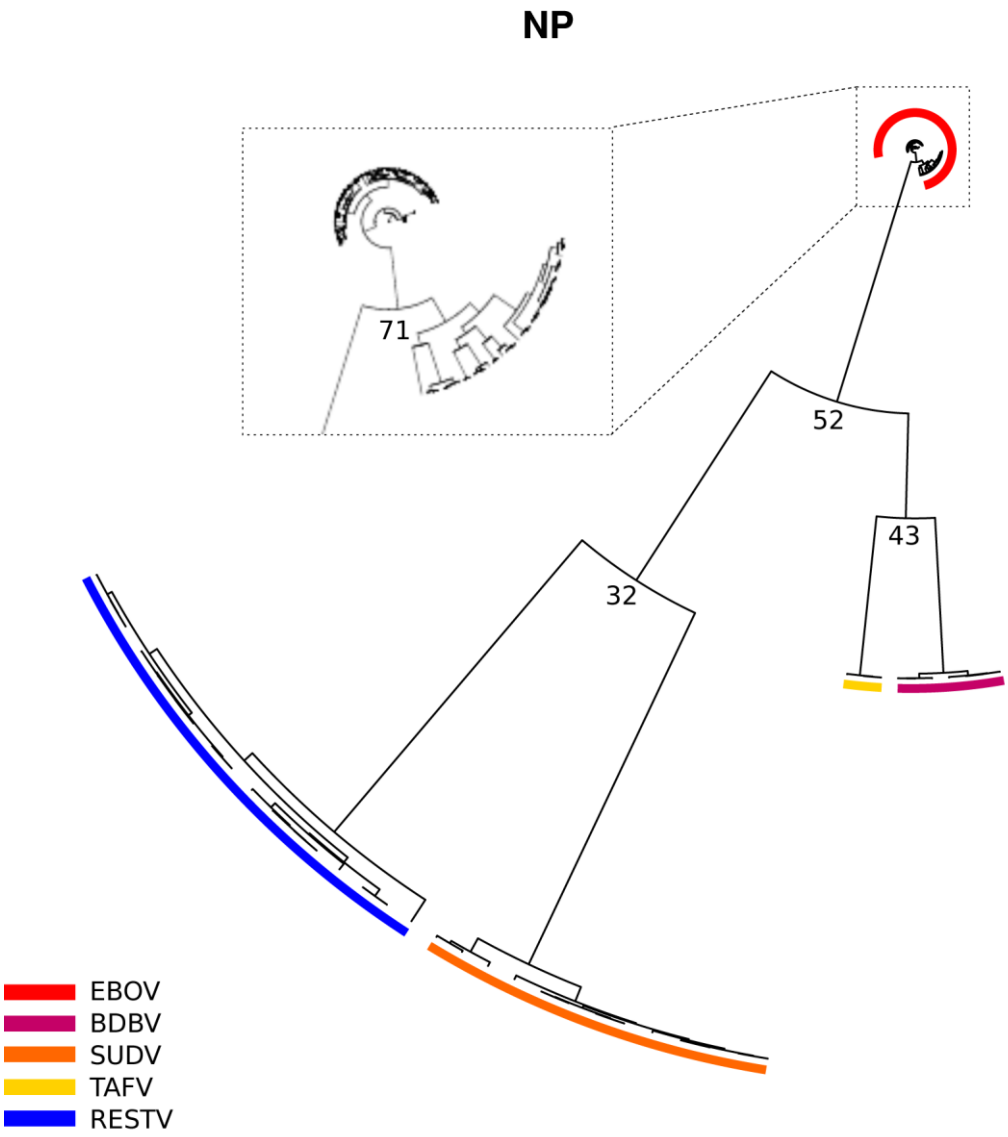


Fig S1G. Bayesian tree for protein NP.

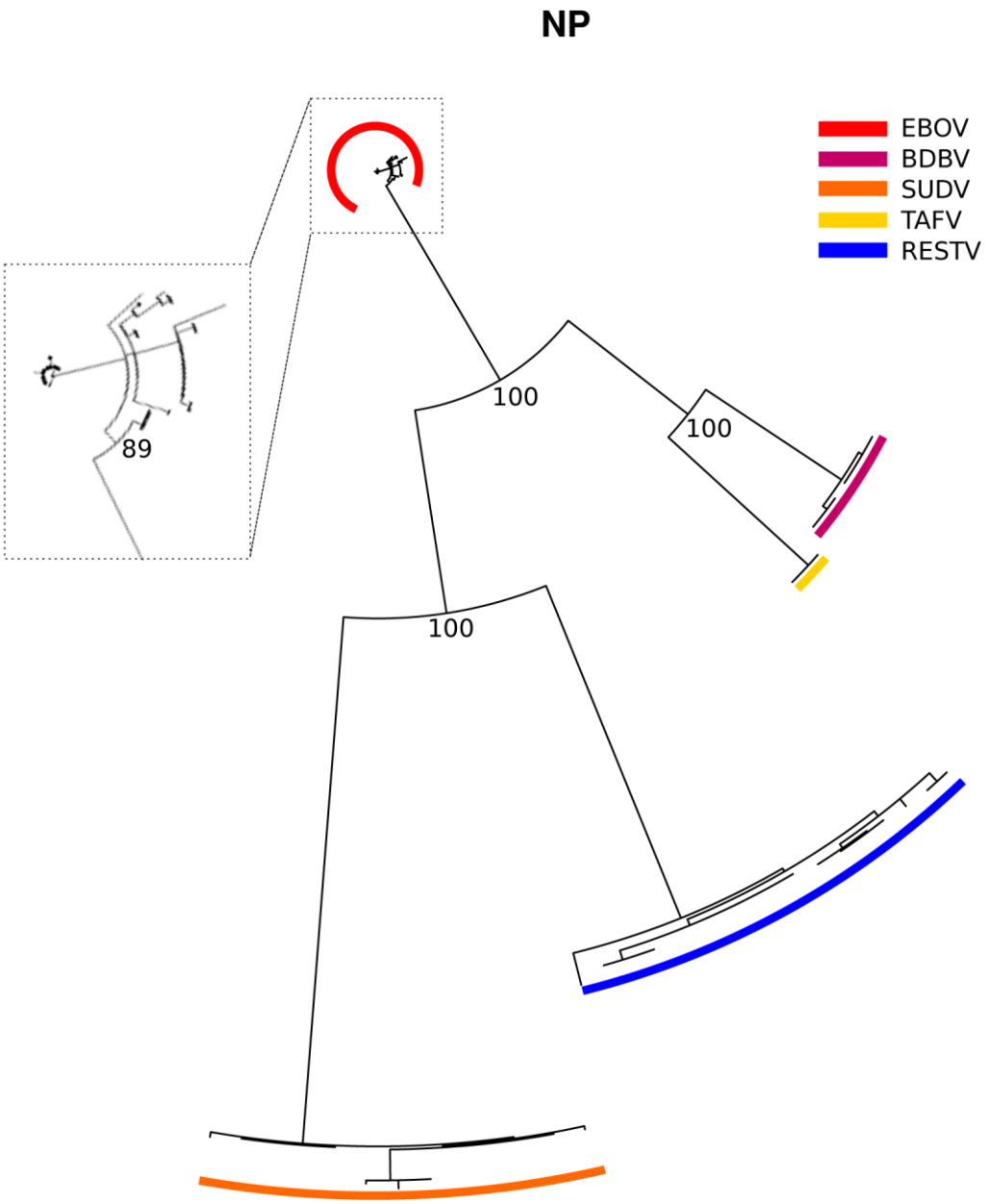


Fig S1H. Maximum likelihood tree for protein NP.

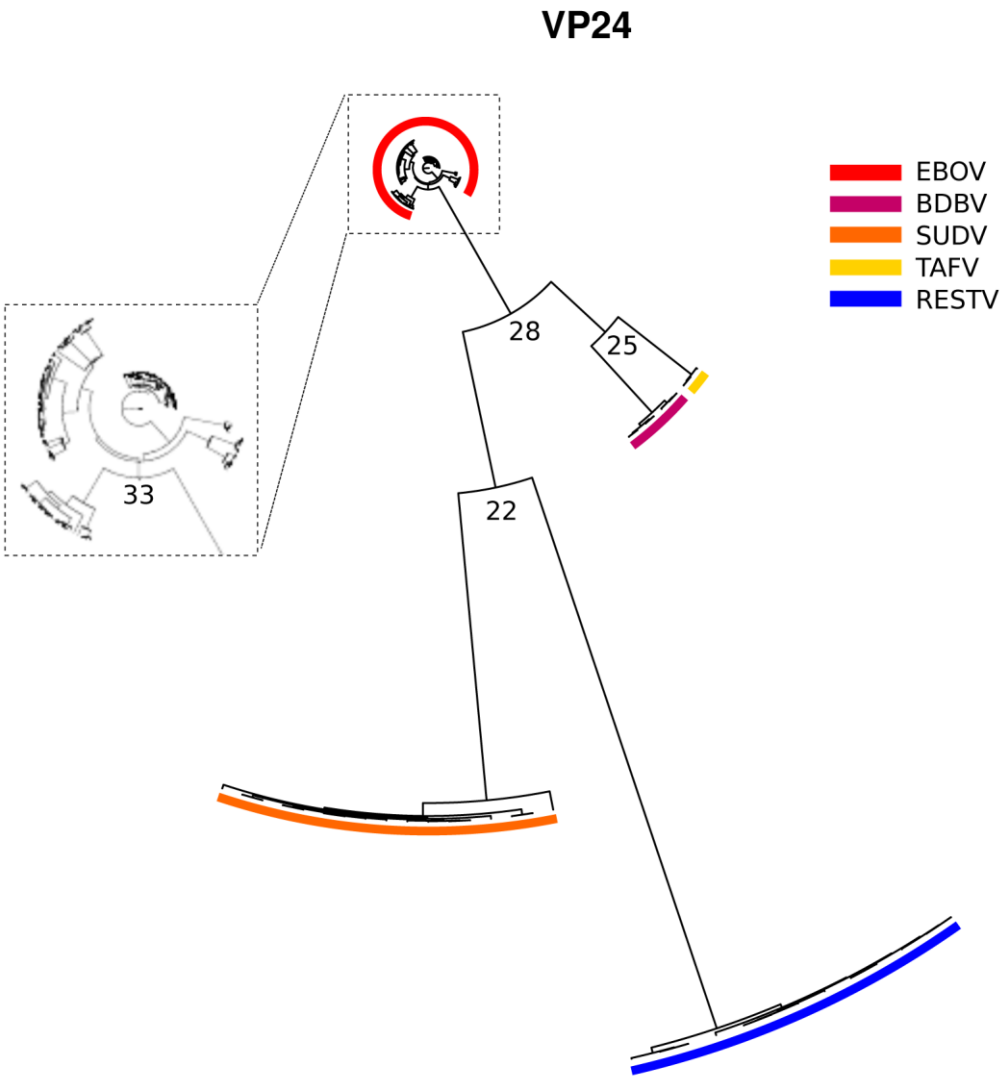


Fig S1I. Bayesian tree for protein VP24.

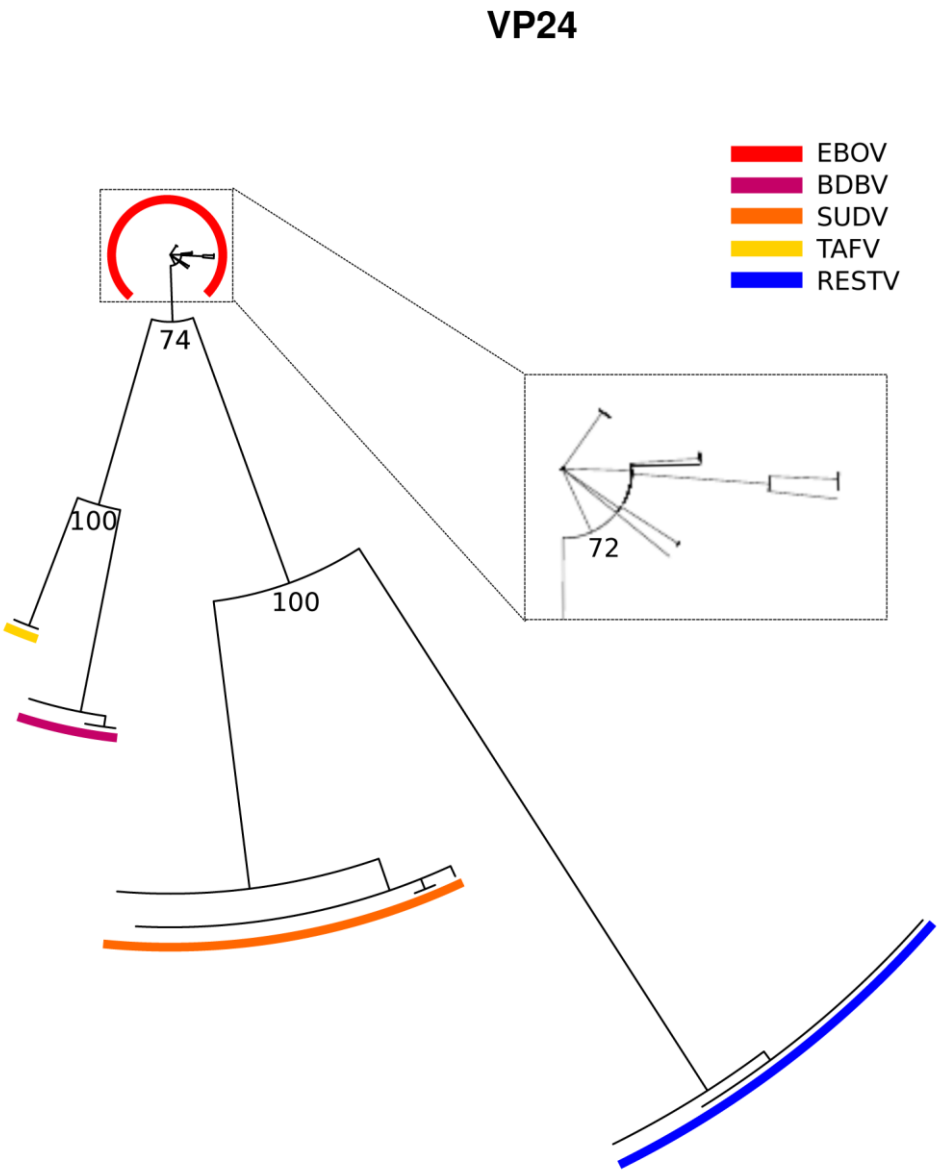


Fig S1J. Maximum likelihood tree for protein VP24.

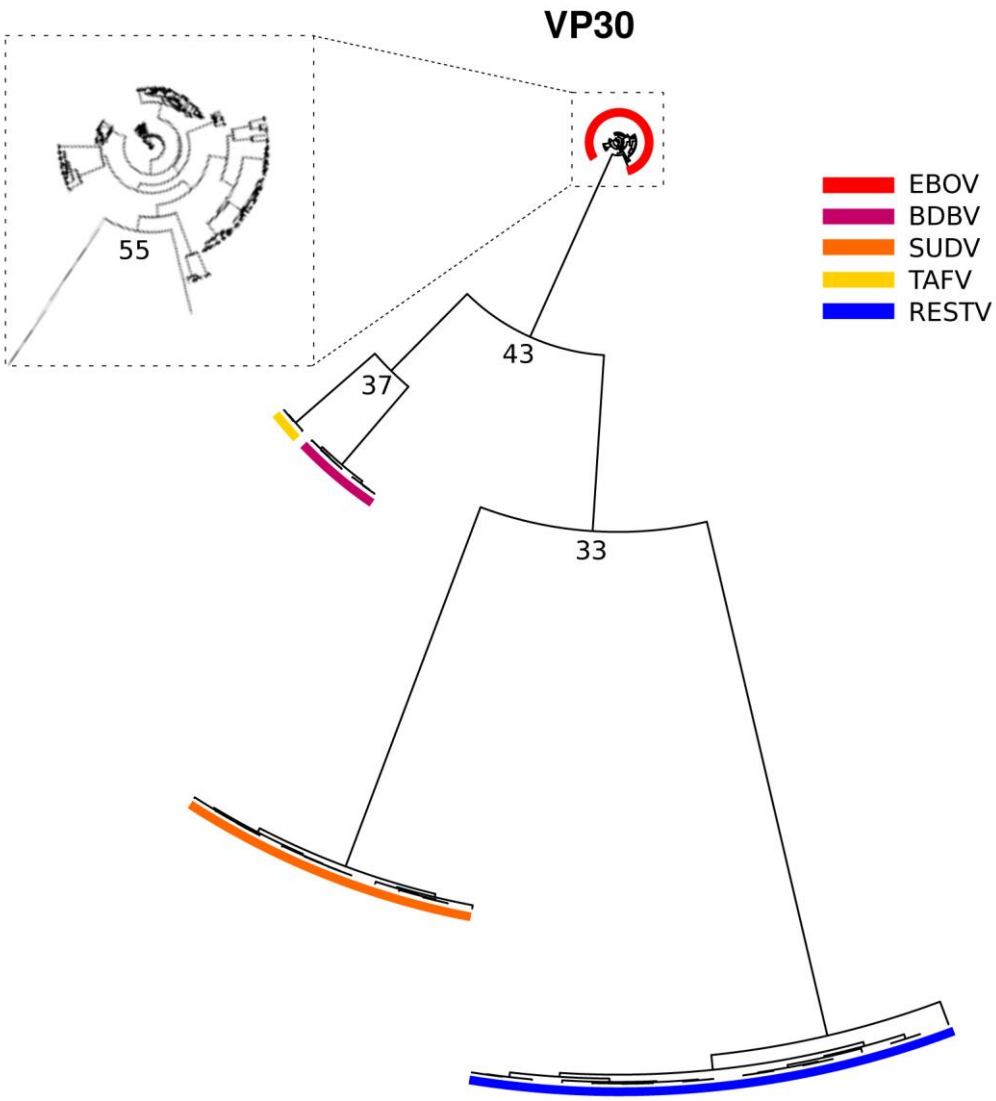


Fig S1K. Bayesian tree for protein VP30.

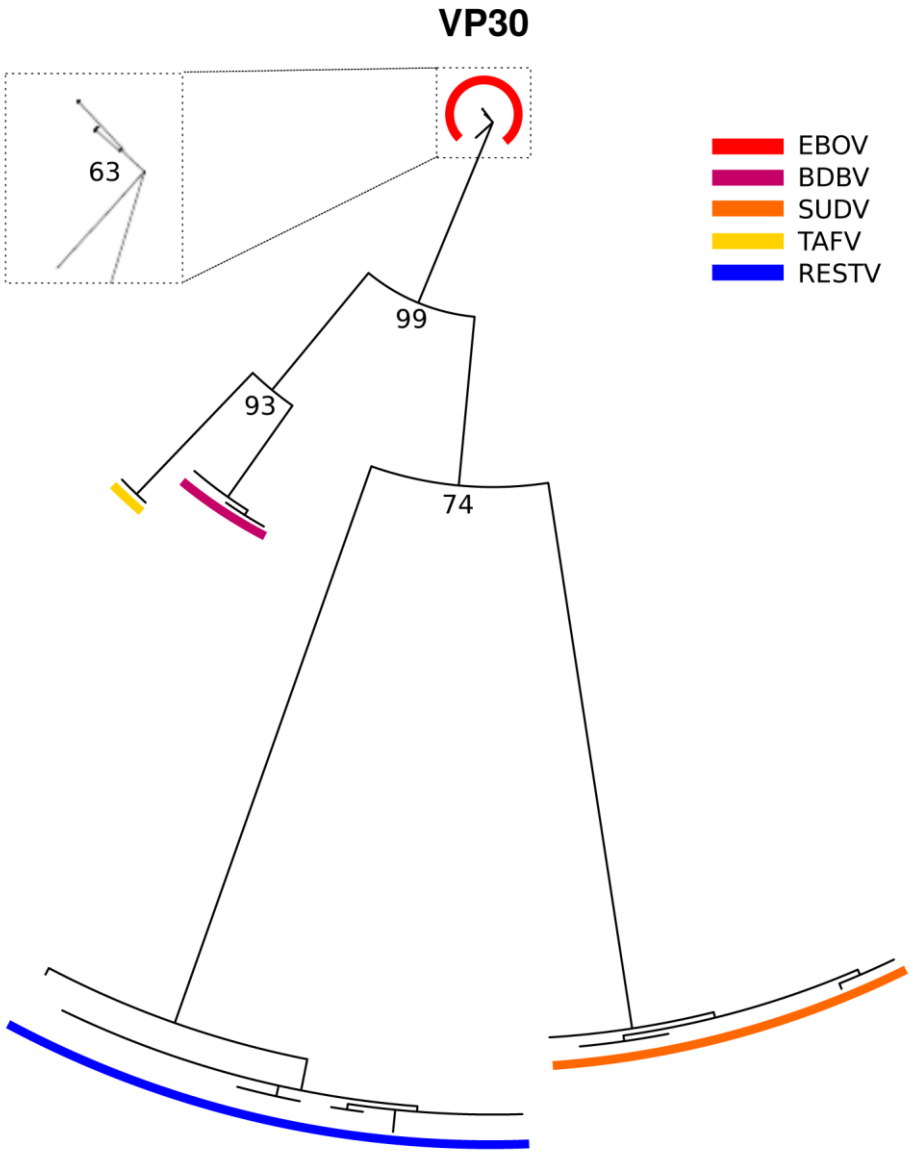


Fig S1L. Maximum likelihood tree for protein VP30.

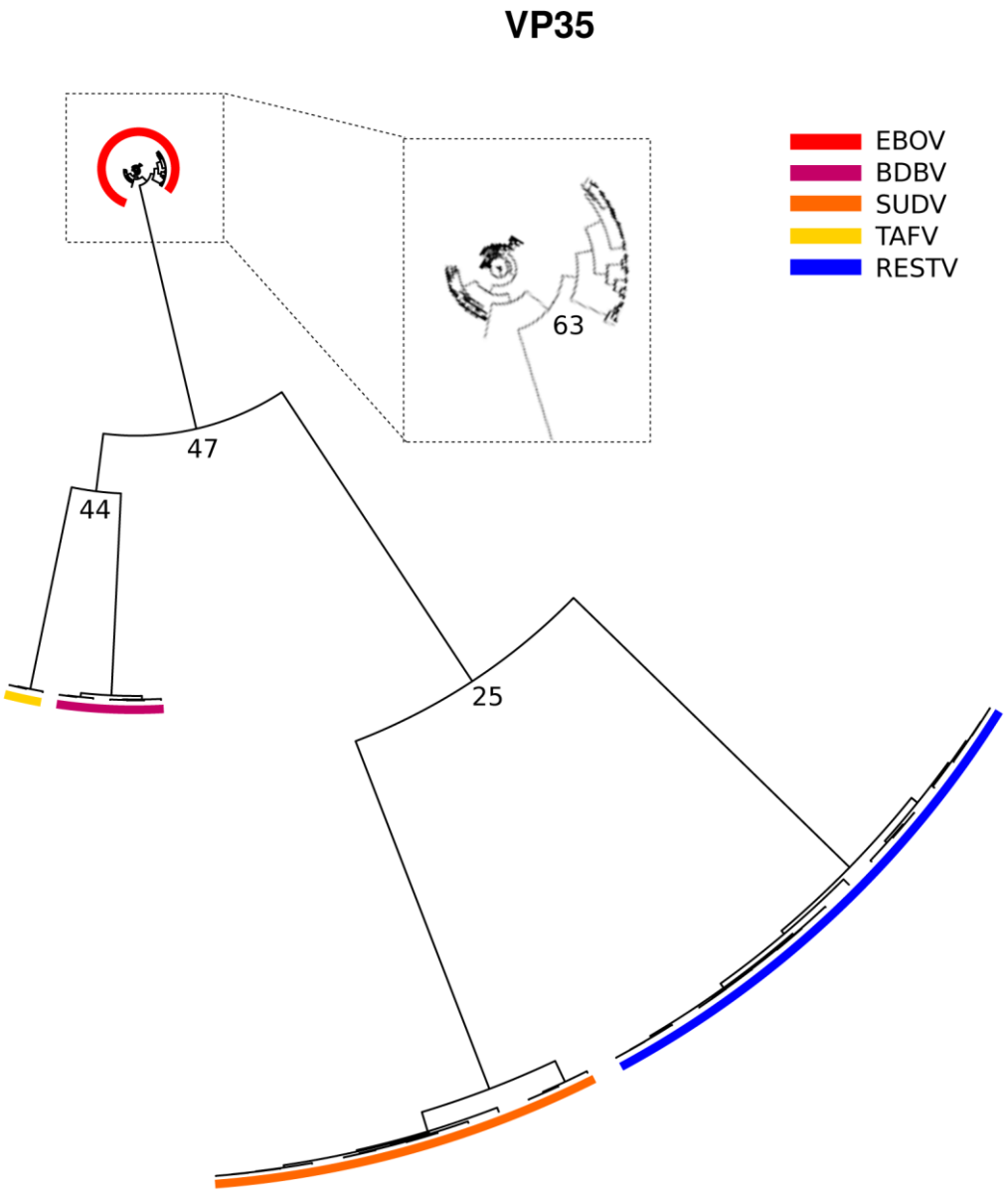


Fig S1M. Bayesian tree for protein VP35.

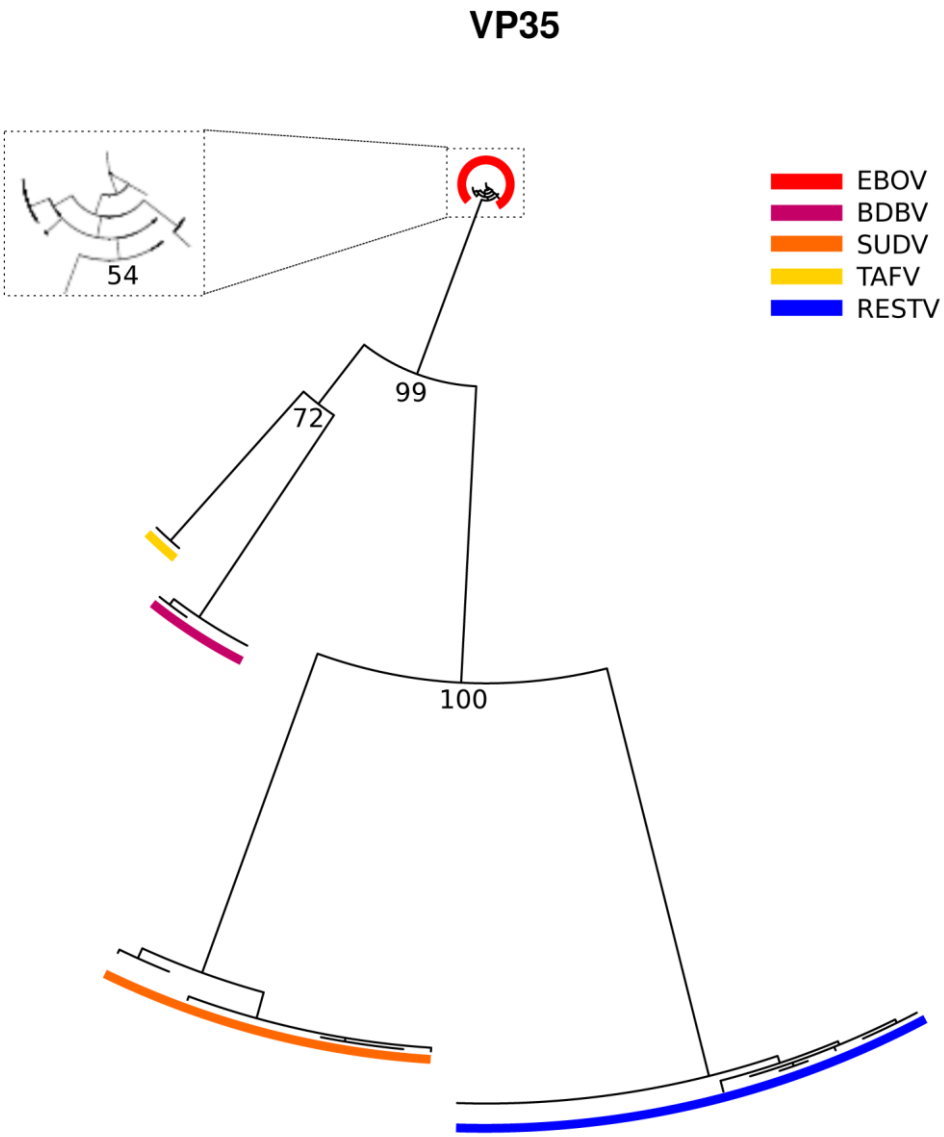


Fig S1N. Maximum likelihood tree for protein VP35.

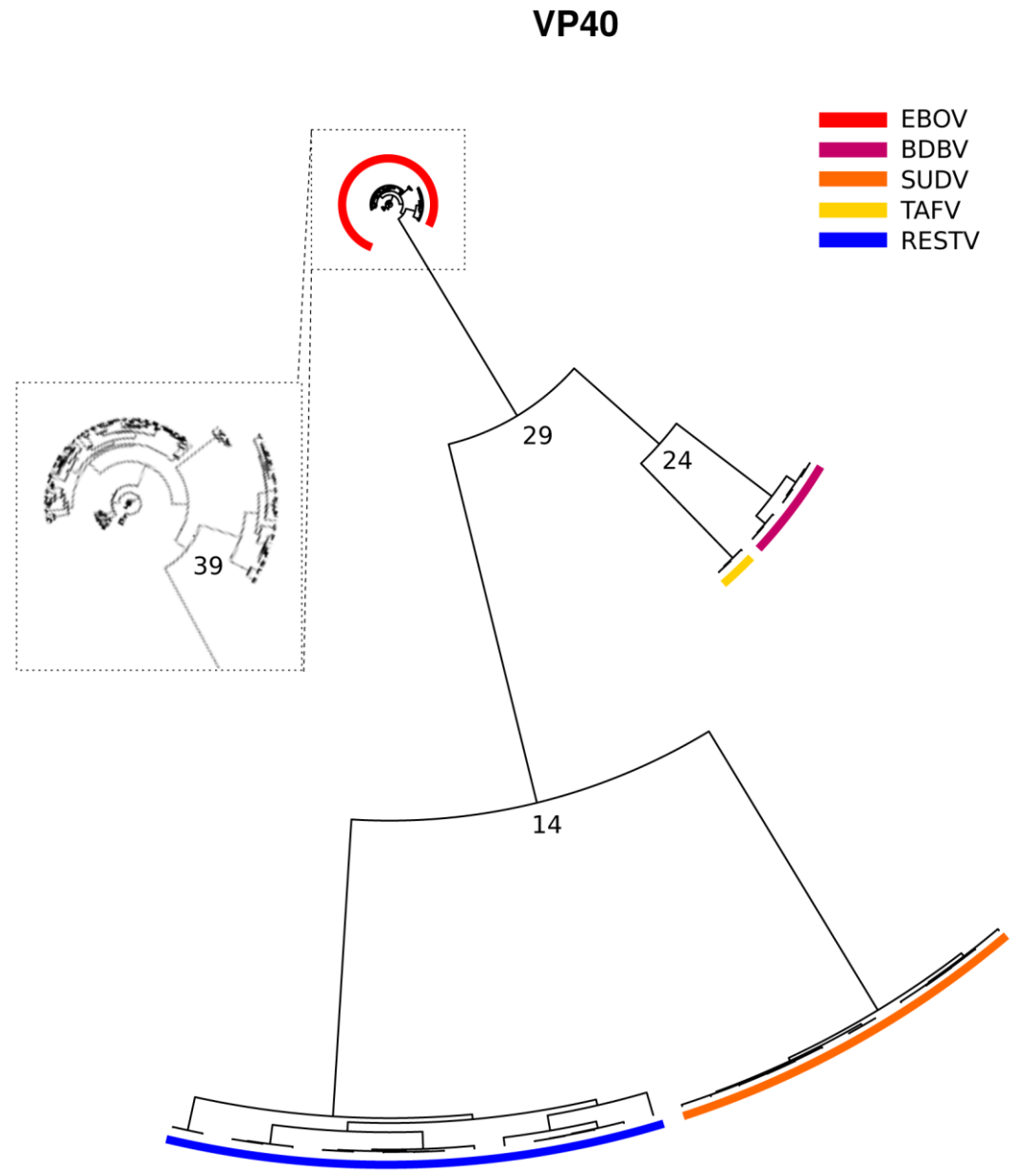


Fig S1O. Bayesian tree for protein VP40.

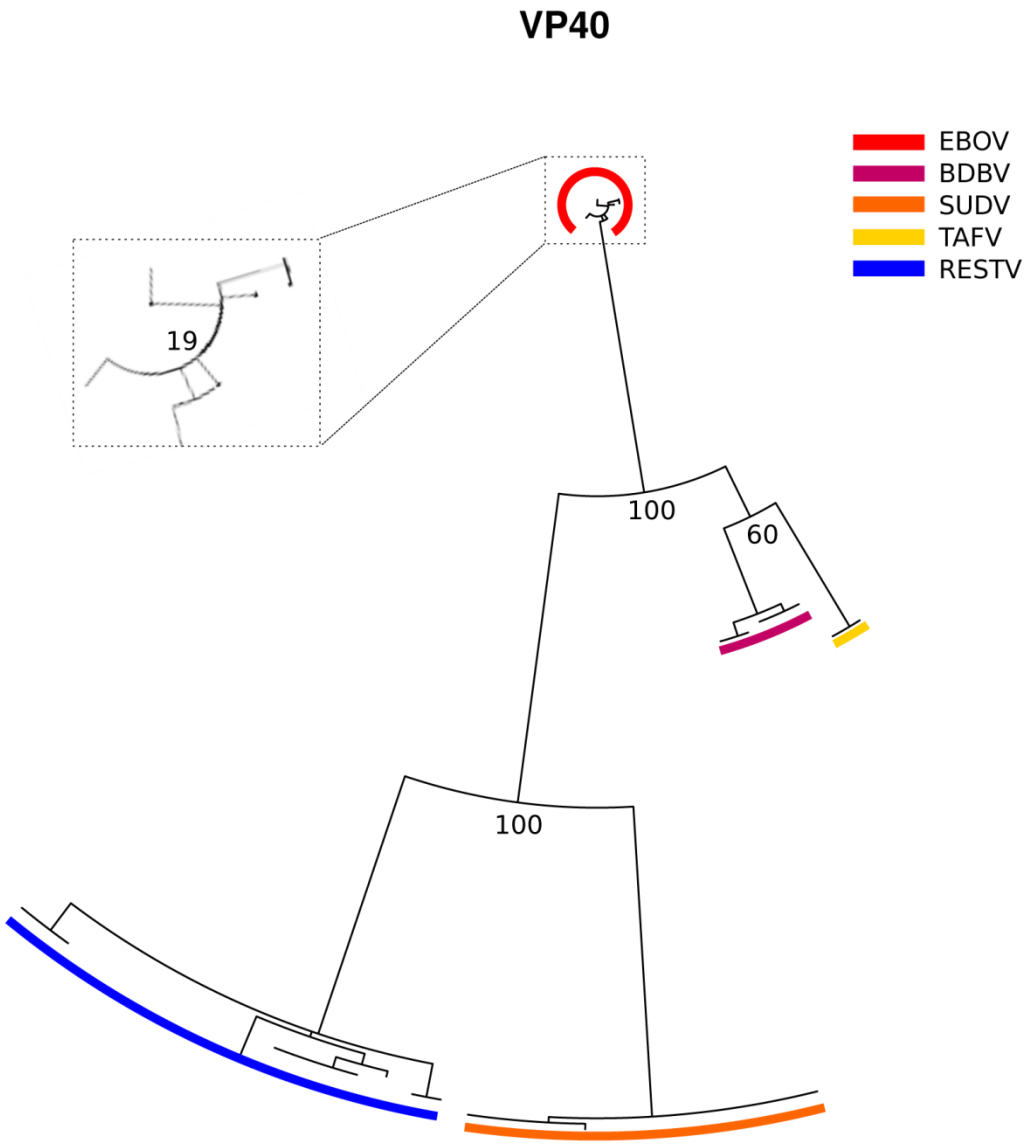
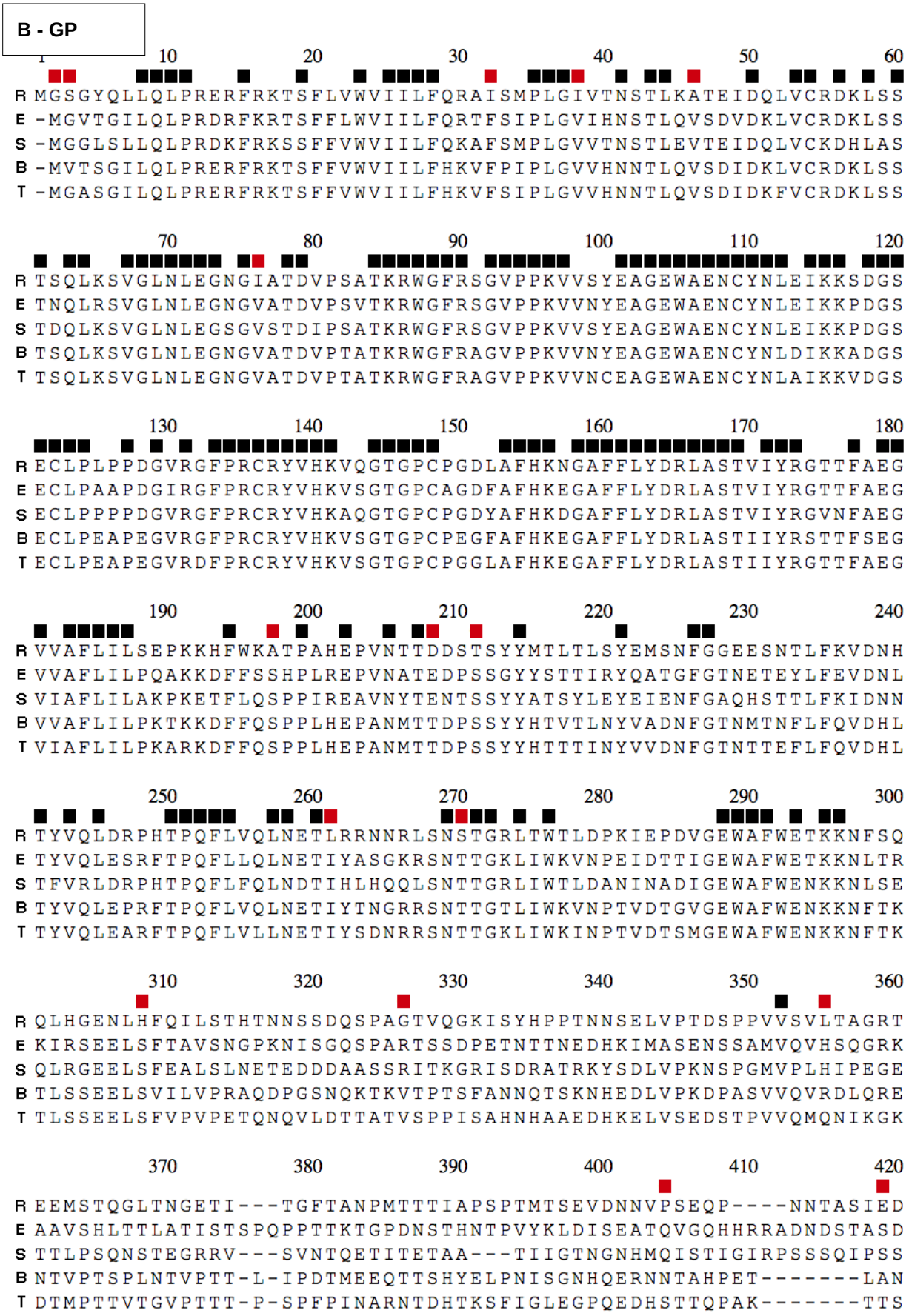
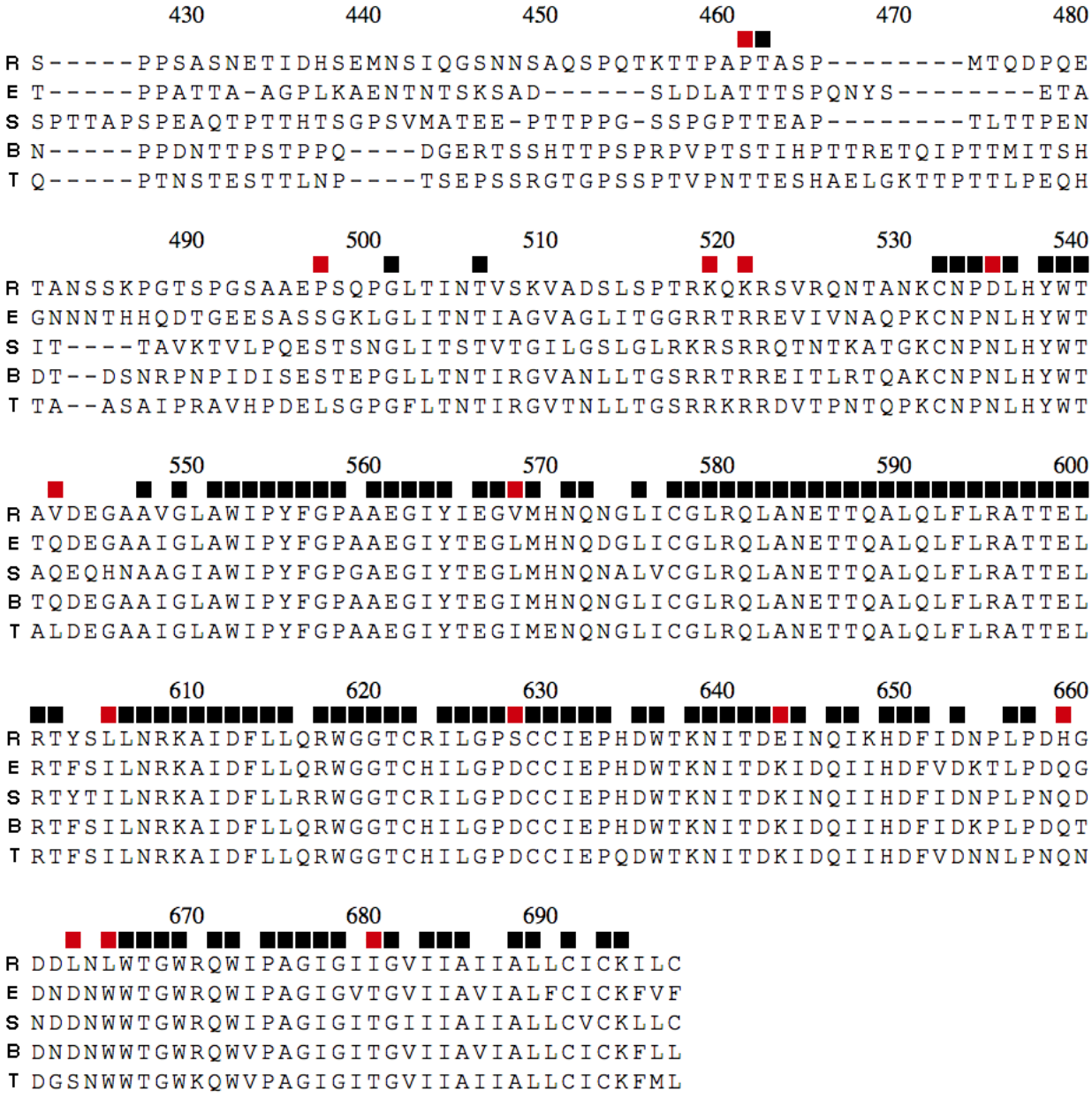


Fig S1P. Maximum likelihood tree for protein VP40.

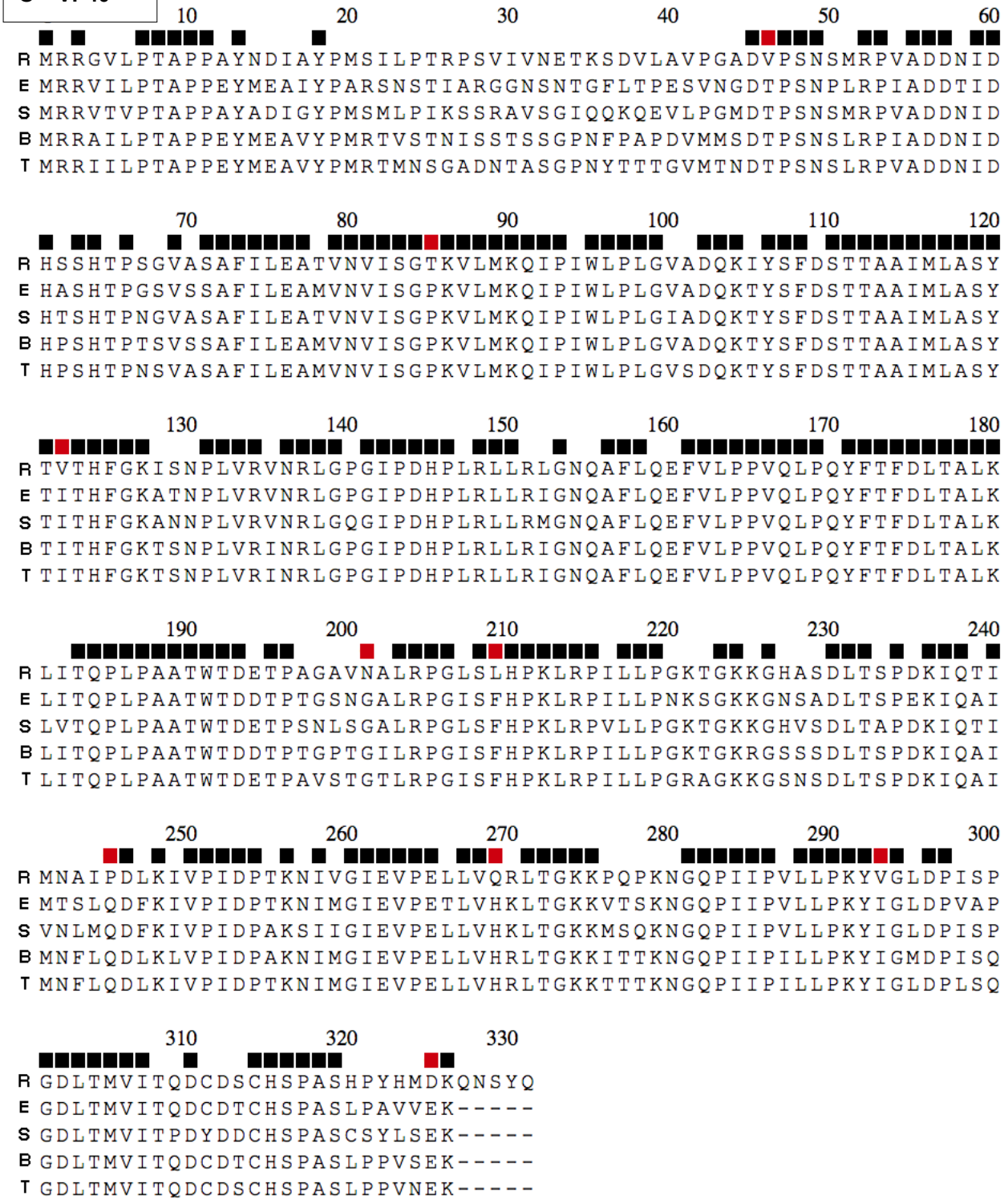
Supplementary Figure 2. Ebolavirus protein consensus sequences and SDPs. The consensus sequence for each *Ebolavirus* species is shown for each Ebolavirus protein. The row above the alignment indicates positions that are 100% conserved across all Ebolavirus sequences (black) or specificity determining 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

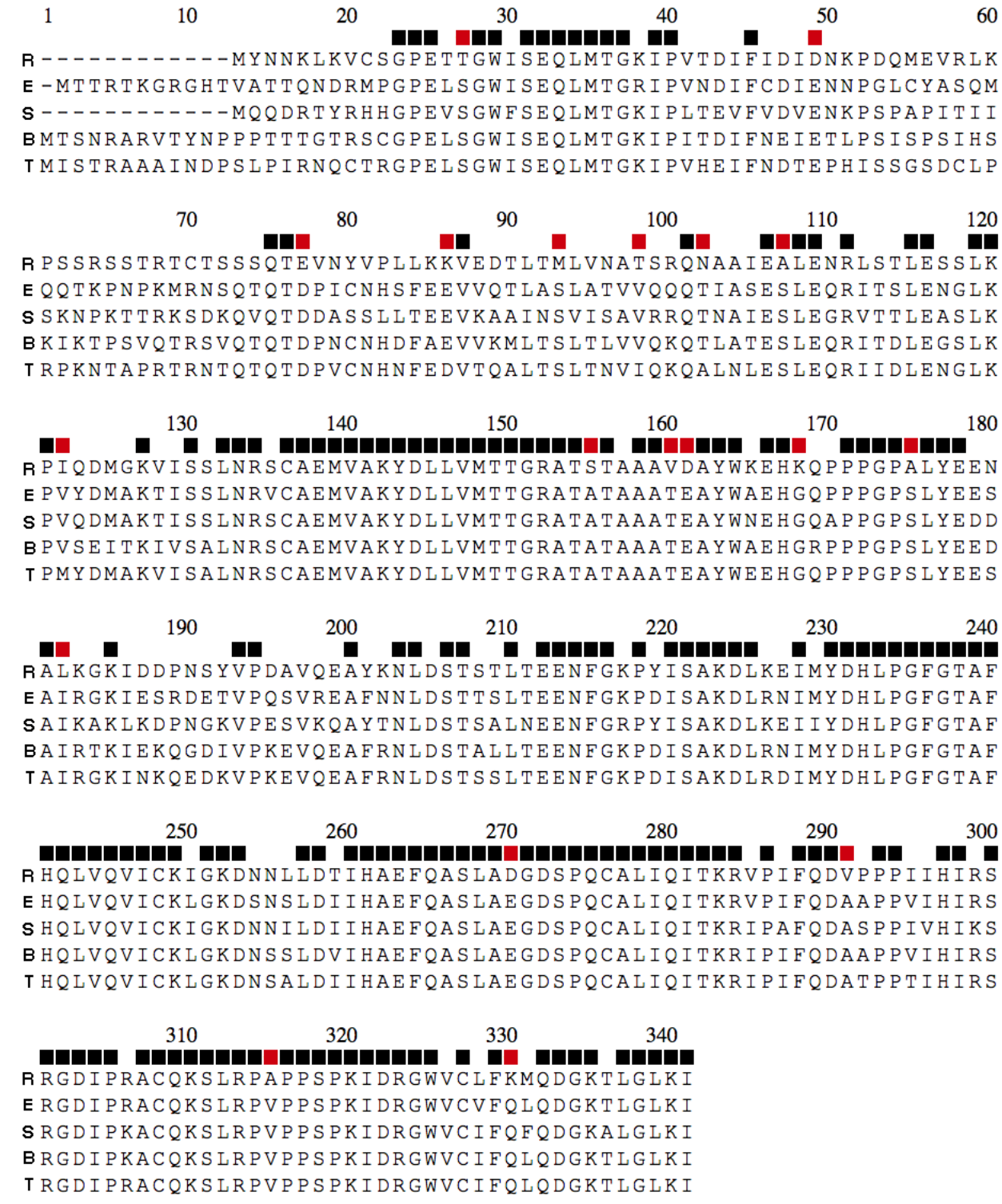

B - GP
1 10 20 30 40 50 60
R MGSGYQLLQLPRERFRKTSFLVWVIILFQRAISMPLGIVTNSTLKATEIDQLVCRDKLSS
E -MGVTGILQLPRDRFKRTSFFLWVIILFQRTFSIPLGVIHNSTLQVSDVDKLVCRDKLSS
S -MGGLSLLQLPRDKFRKSSFFVWVIILFQKAFSMPLGVVTNSTLEVTEIDQLVCKDHLAS
B -MVTSGILQLPRERFRKTSFFVWVIILFHKVFPIPLGVVHNNTLQVSDIDKLVCRDKLSS
T -MGASGILQLPRERFRKTSFFVWVIILFHKVFSIPLGVVHNNTLQVSDIDKFVCRDKLSS
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

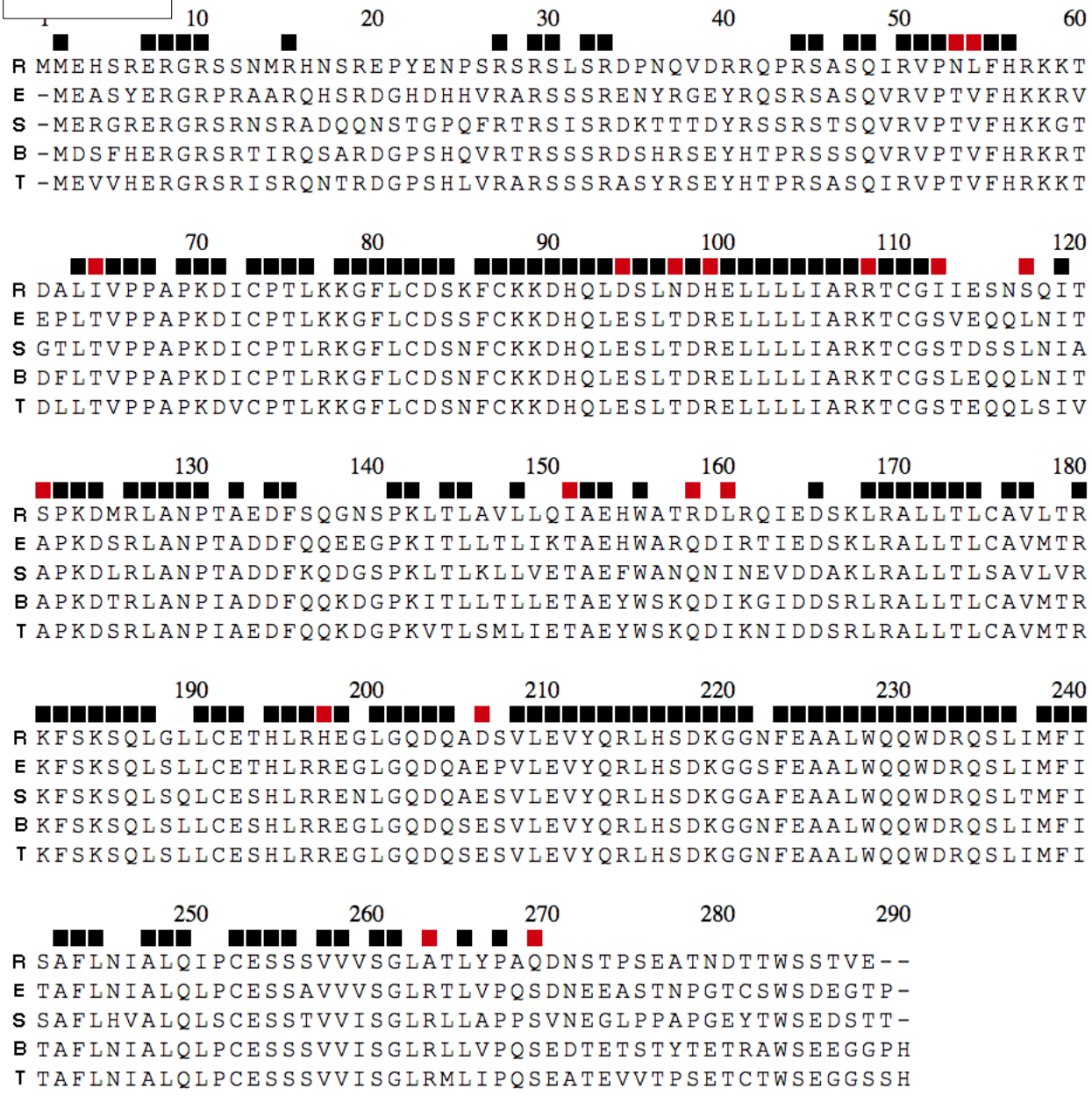
C – VP40

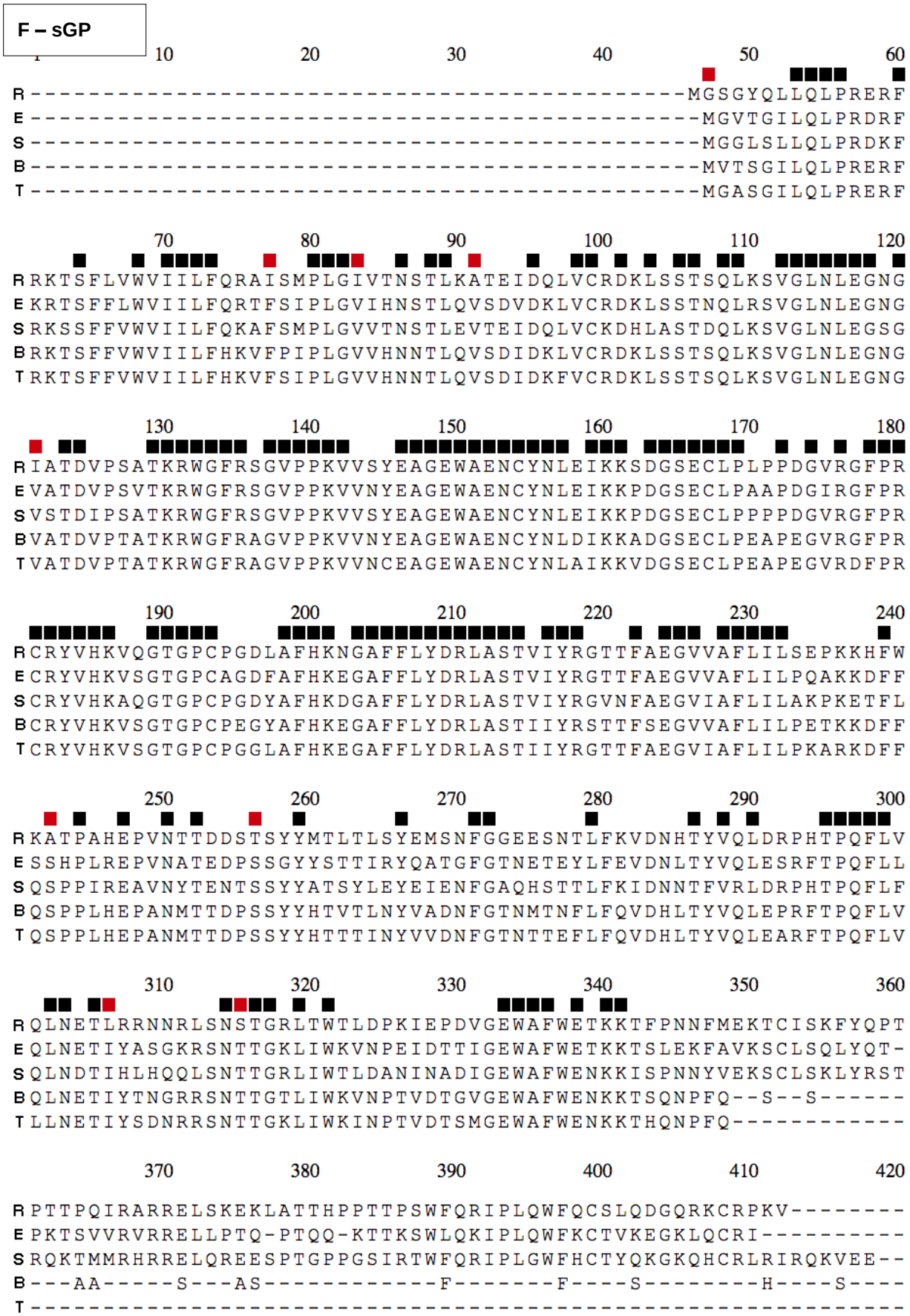

10 20 30 40 50 60
R MRRGVLPTAPPAYNDIAYPMSILPTRPSVIVNETKSDVLAVPGADVPSNSMRPVADDNID
E MRRVILPTAPPEYMEAIYPARSNSTIARGGNSNTGFLTPESVNGDTPSNPLRPIADDTID
S MRRVTVPTAPPAYADIGYPMSMLPIKSSRAVSGIQQKQEVLPGMDTPSNSMRPVADDNID
B MRRAILPTAPPEYMEAVYPMRTVSTNISSTSSGPNFPAPDVMMSDTPSNSLRPIADDNID
T MRRIILPTAPPEYMEAVYPMRTMNSGADNTASGPNYTTTGVMTNDTPSNSLRPVADDNID
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-----

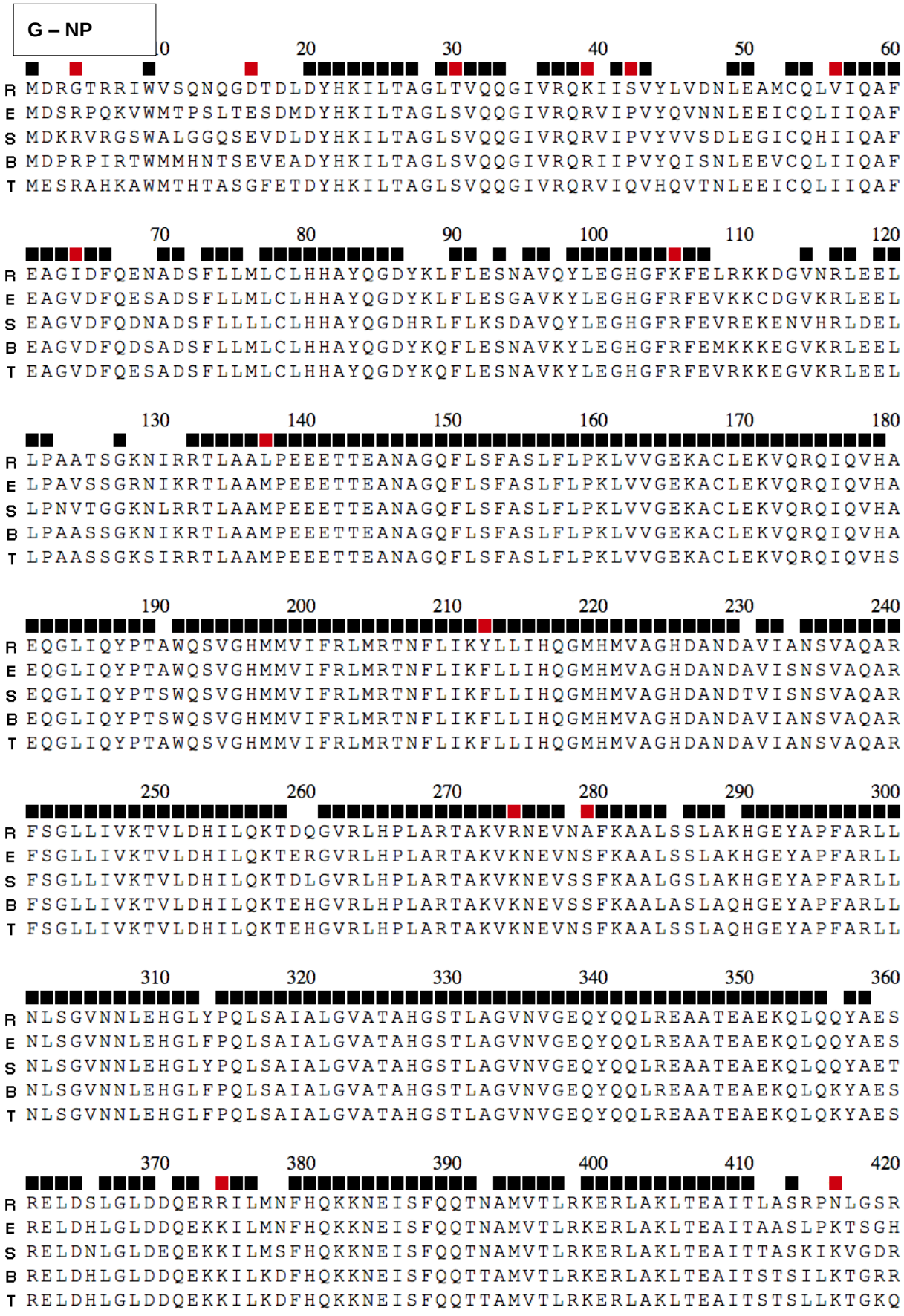
D – VP35

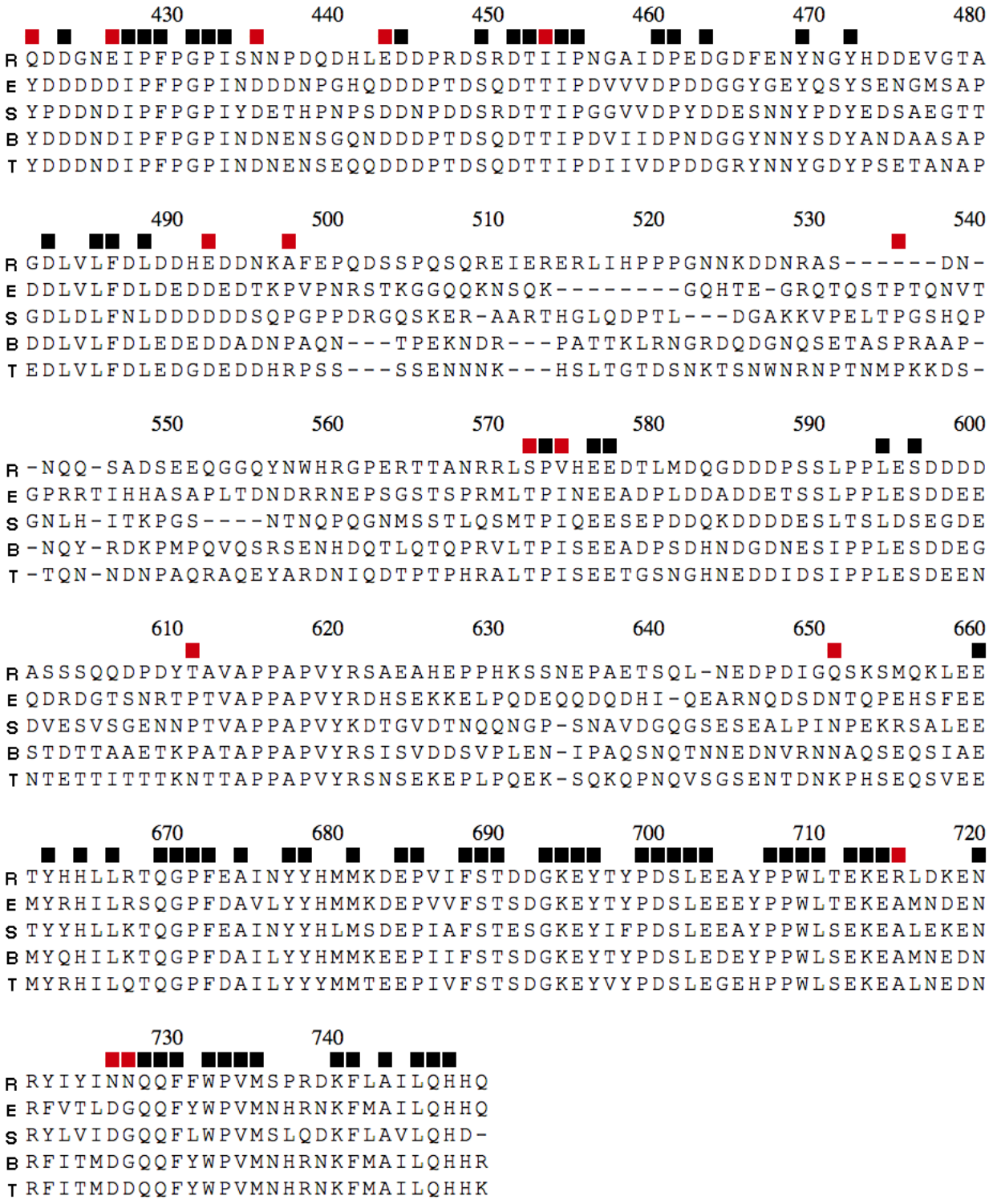

1 10 20 30 40 50 60
R ----------MYNNKLKVCSGPETTGWISEQLMTGKIPVTDIFIDIDNKPDQMEVRLK
E -MTTRTKGRGHTVATTQNDRMPGPELSGWISEQLMTGRIPVNDIFCDIENNPGLCYASQM
S ----------MQQDRTYRHHGPEVSGWFSEQLMTGKIPLTEVFVDVENKPSPAPITII
B MTSNRARVTYNPPPTTTTGTRSCGPELSGWISEQLMTGKIPITDIFNEIETLPSISPSIHS
T MISTRAAAINDPSLPIRNQCTRGPELSGWISEQLMTGKIPVHEIFNDTEPHISSGSDCLP
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

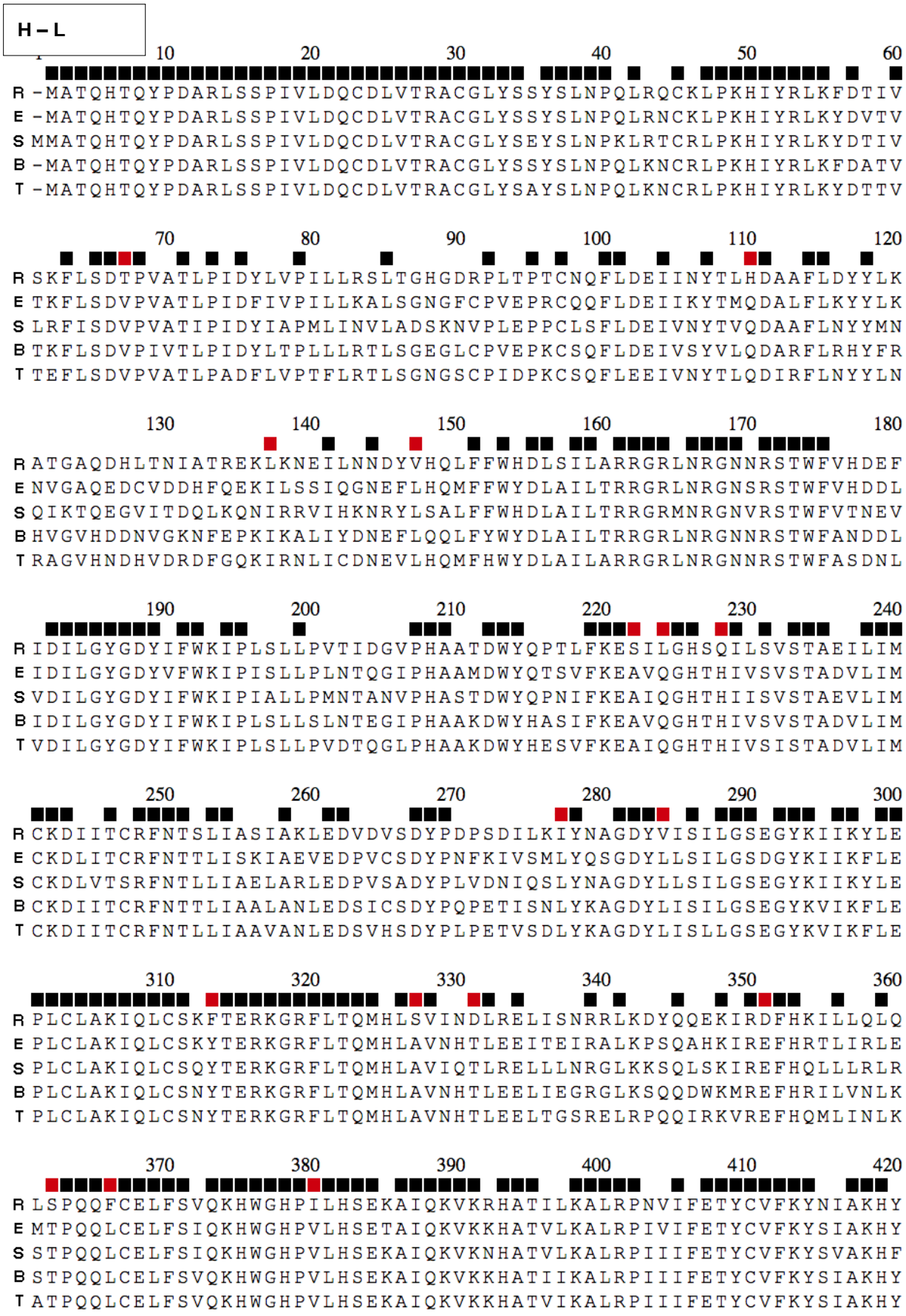
E – VP30


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

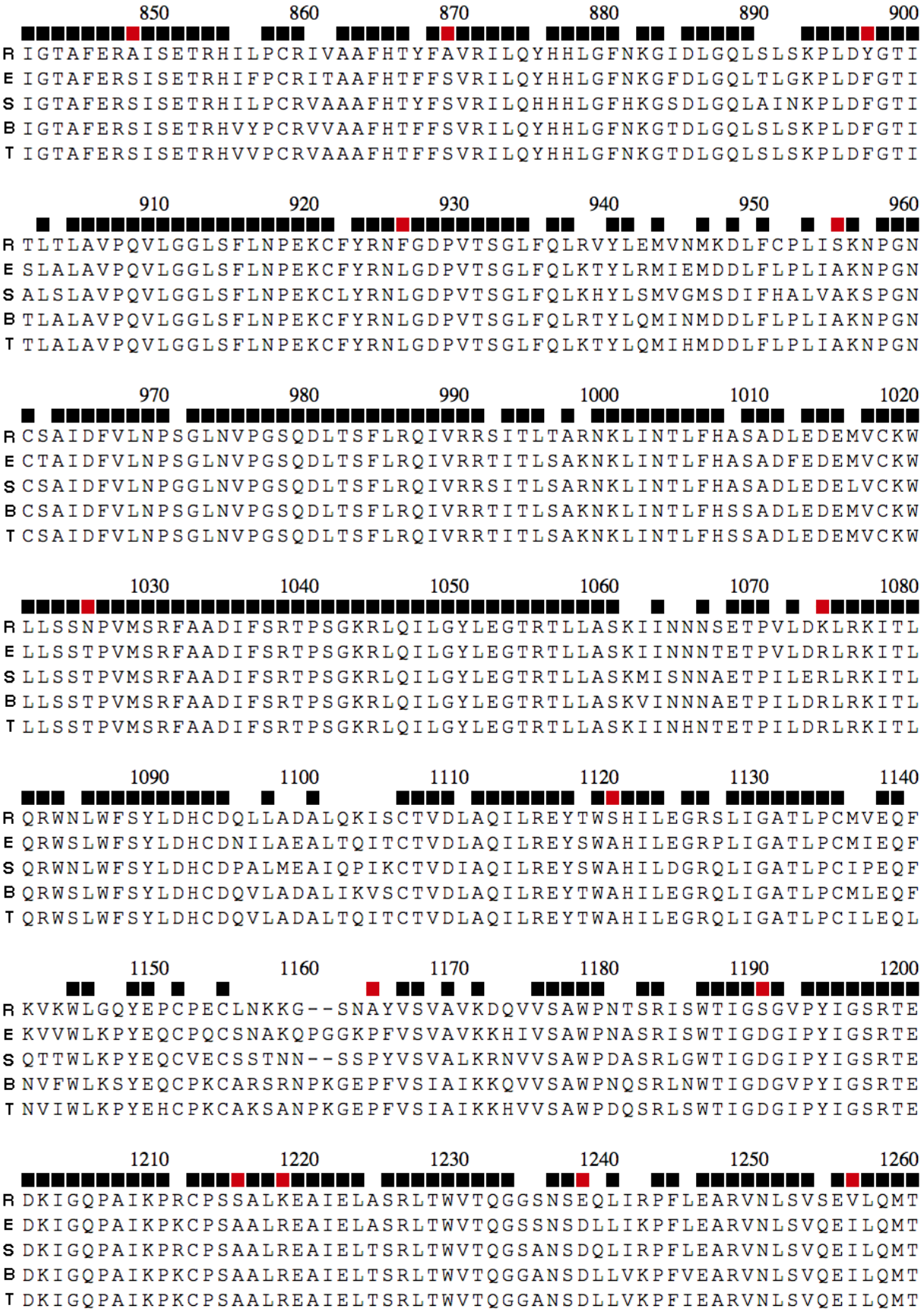

F – sGP
10 20 30 40 50 60
R ----------MGSGYQLLQLPRERF
E ----------MGVTGILQLPRDRF
S ----------MGGLSLLQLPRDKF
B ----------MVTSGILQLPRERF
T ----------MGASGILQLPRERF
70 80 90 100 110 120
R RKTSFLVWVIILFQRAISMPLGIVTNSTLKATEIDQLVCRDKLSSTSQLKSVGLNLEGNG
E KRTSFFLWVIILFQRTFSIPLGVIHNSTLQVSDVDKLVCRDKLSSTNQLRSVGLNLEGNG
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 ----------

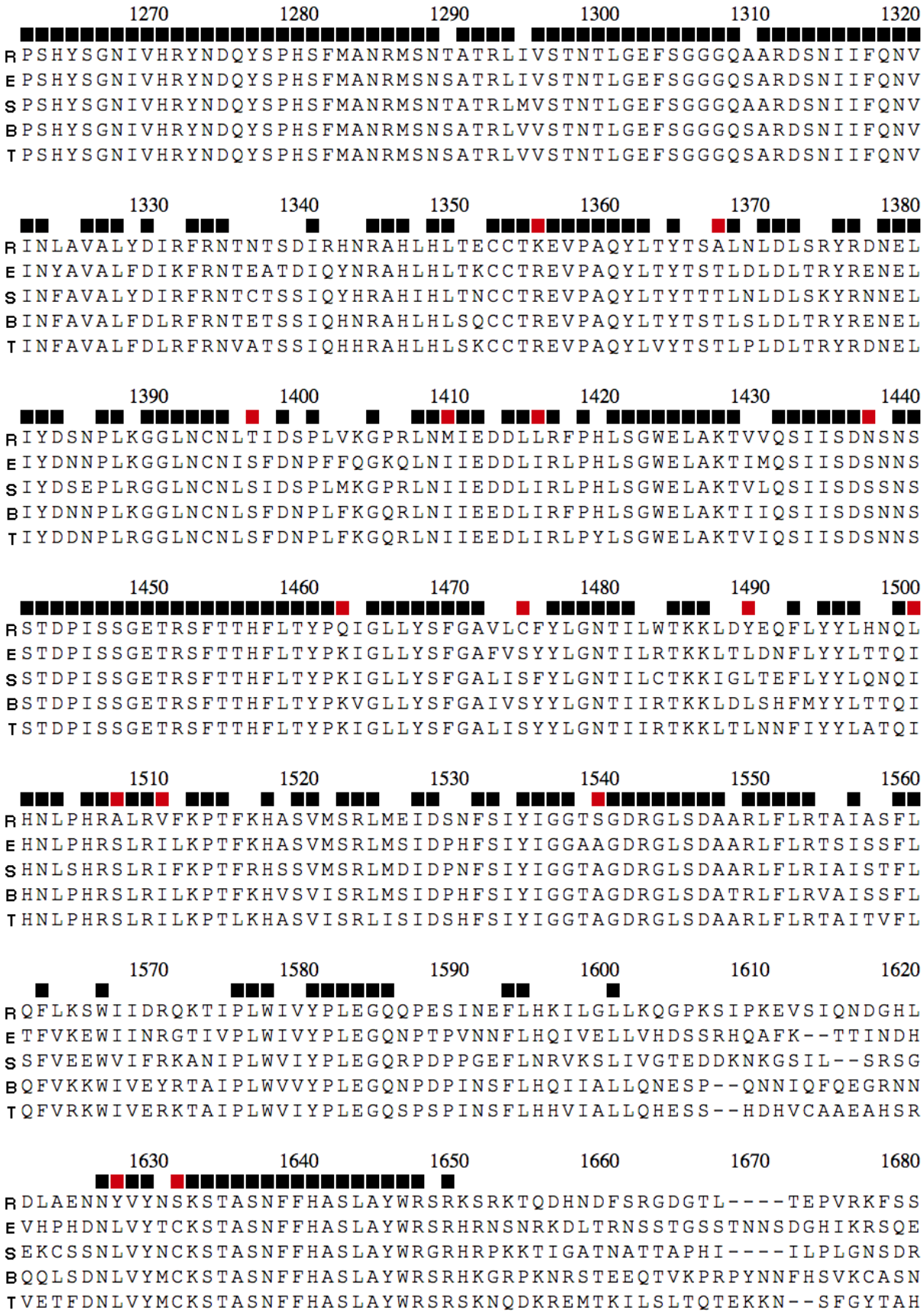

G – NP
10 20 30 40 50 60
R MDRGTRRIWVSQNQGDTDLDYHKILTAGLTVQQGIVRQKIISVYLVDNLEAMCQLVIQAF
E MDSRPQKVWMTPSLTESDMDYHKILTAGLSVQQGIVRQRVIPVYQVNNLEEICQLIIQAF
S MDKRVRGSWALGGQSEVDLDYHKILTAGLSVQQGIVRQRVIPVYVVSDLEGICQHIIQAF
B MDPRPIRTWMMHNTSEVEADYHKILTAGLSVQQGIVRQRIIPVYQISNLEEVCQLIIQAF
T MESRAHKAWMTHTASGFETDYHKILTAGLSVQQGIVRQRVIQVHQVTNLEEICQLIIQAF
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

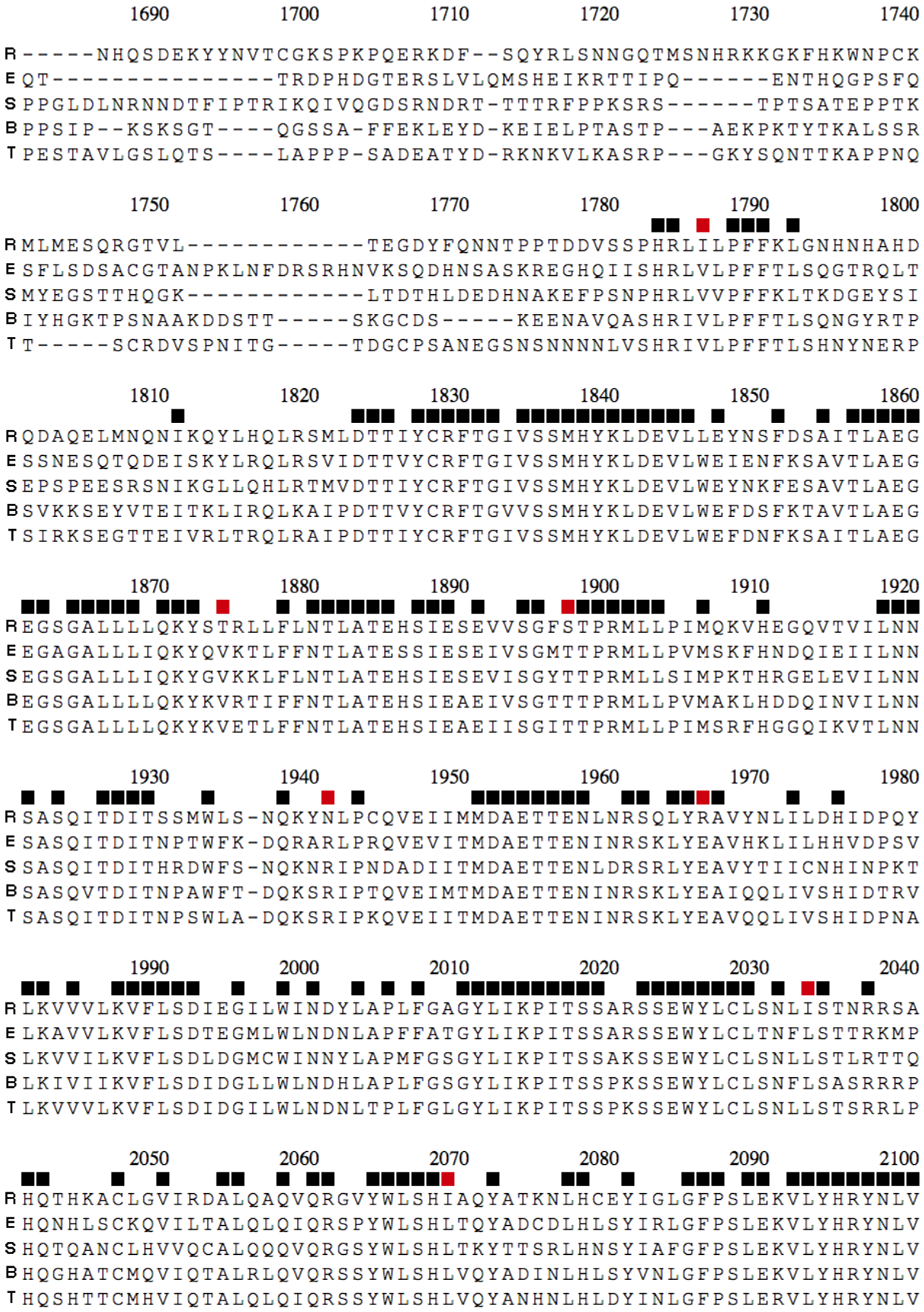

430 440 450 460 470 480
R QDDGNEIPFPGPISNNPDQDHLEDDPRDSRDTIIPNGAIDPEDGDFENYNGYHDDEVGTA
E YDDDDDIPFPGPINDDDNPGHQDDDPTDSQDTTIPDVVVDPDDGGYGEYQSYSENGMSAP
S YPDDNDIPFPGPIYDETHPNPSDDNPDDSRDTTIPGGVVDPYDDESNNYPDYEDSAEGTT
B YDDDNDIPFPGPINDNENSGQNDDDPTDSQDTTIPDVIIDPNDGGYNNYSDYANDAASAP
T YDDDNDIPFPGPINDNENSEQQDDDPTDSQDTTIPDIIVDPDDGRYNNYGDYPSETANAP
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

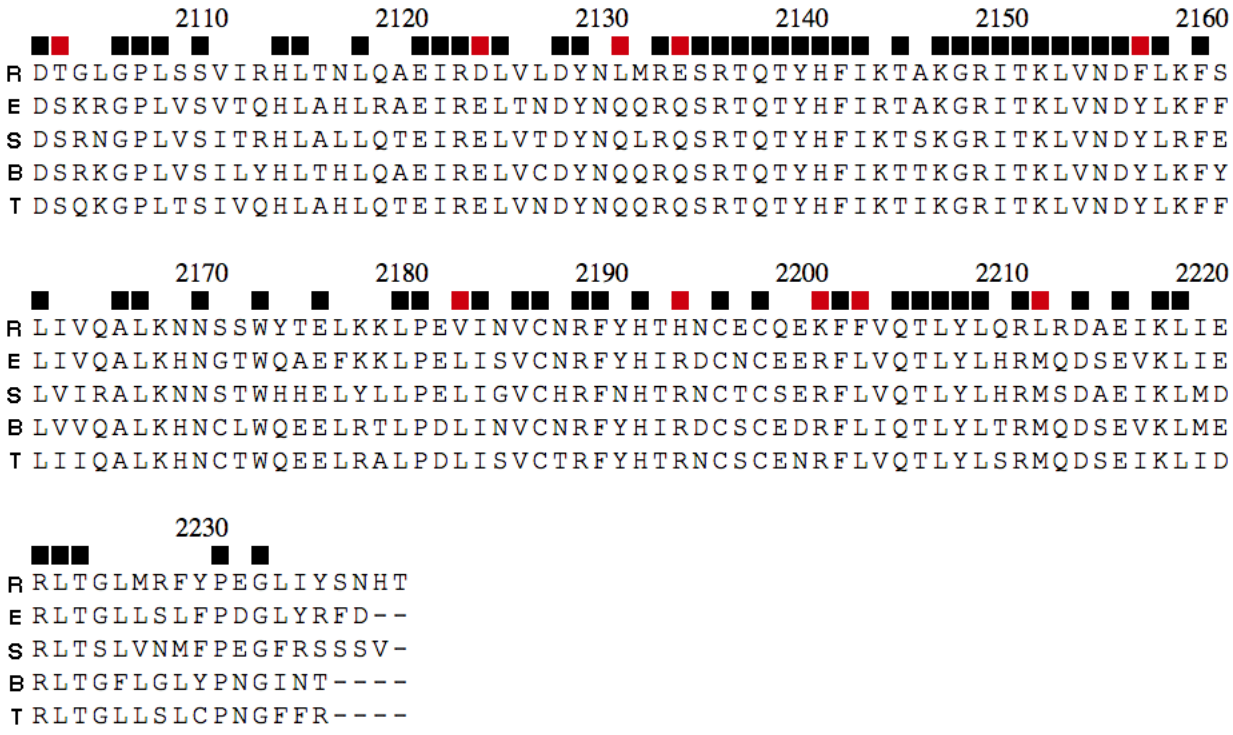

H – L
10 20 30 40 50 60
R -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRQCKLPKHIYRLKFDTIV
E -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKLPKHIYRLKYDVTV
S MMATQHTQYPDARLSSPIVLDQCDLVTRACGLYSEYSLNPKLRTCRLPKHIYRLKYDTIV
B -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLKNCRLPKHIYRLKFDATV
T -MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSAYSLNPQLKNCRLPKHIYRLKYDTTV
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

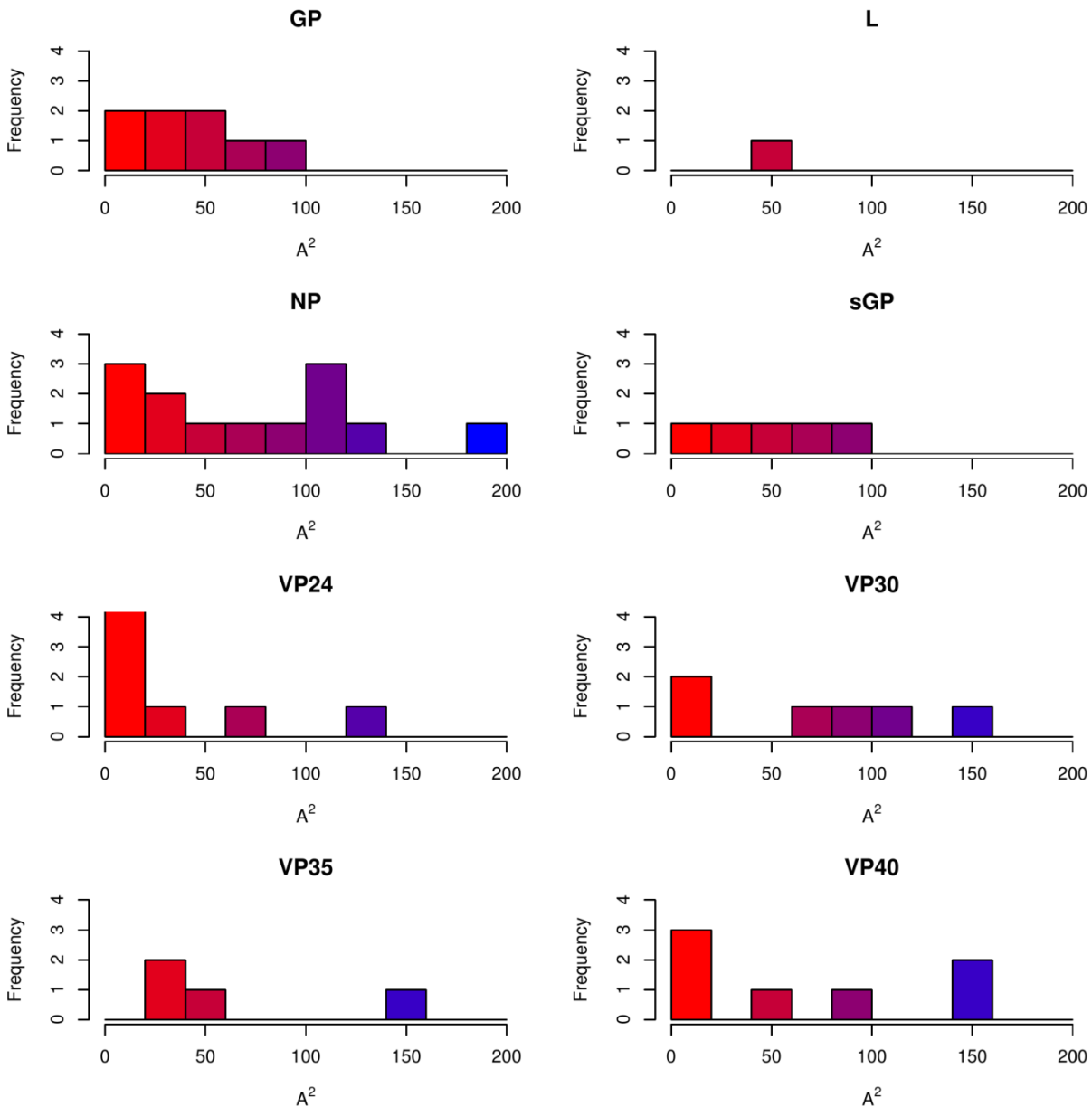

430 440 450 460 470 480
R FDSQGTWYSVISDRNLTPGLNSFIKRNHFPSLPMIKDLLWEFYHLNHPPLFSTKVISDLS
E FDSQGSWYSVTSDRNLTPGLNSYIKRNQFPPLPMIKELLWEFYHLDHPPLFSTKIISDLS
S FDSQGTWYSVISDRCLTPGLNSYIRRNQFPPLPMIKDLLWEFYHLDHPPLFSTKIISDLS
B FDSQGSWYSVISDKHLTPGLHSYIKRNQFPPLPMIKDLLWEFYHLDHPPLFSTKIISDLS
T FDSQGTWYSVTSDRCLTPGLSSYIKRNQFPPLPMIKELLWEFYHLDHPPLFSTKVISDLS
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

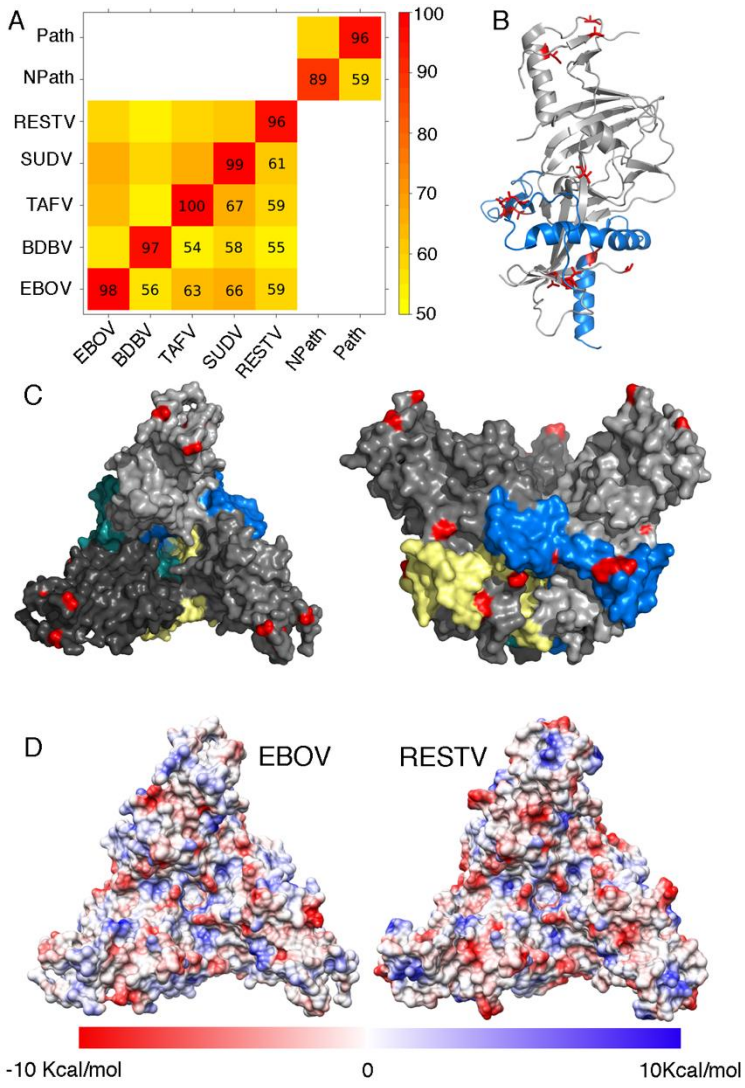

1270 1280 1290 1300 1310 1320
R PSHYSGNIVHRYNDQYSPHSFMANRMSNTATRLIVSTNTLGEFSGGGQAARDSNIIFQNV
E PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLIVSTNTLGEFSGGGQSARDSNIIFQNV
S PSHYSGNIVHRYNDQYSPHSFMANRMSNTATRLMVSTNTLGEFSGGGQAARDSNIIFQNV
B PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLVVSTNTLGEFSGGGQSARDSNIIFQNV
T PSHYSGNIVHRYNDQYSPHSFMANRMSNSATRLVVSTNTLGEFSGGGQSARDSNIIFQNV
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
E QT---------------TRDPHDGTERSLVLQMSHEIKRTTIPQ------ENTHQGPSFQ
S PPGLDLNRNNDTFIPTRIKQIVQGDSRNDRT-TTTRFPPKSRS------TPTSATEPPTK
B PPSIP--KSKSGT----QGSSA-FFEKLEYD-KEIELPTASTP---AEKPKTYTKALSSR
T PESTAVLGSLQTS----LAPPP-SADEATYD-RKNKVLKASRP---GKYSQNTTKAPPNQ
1750 1760 1770 1780 1790 1800
R MLMESQRGTVL----------TEGDYFQNNTPPTDDVSSPHRLILPFFFKLGNHNHAHD
E SFLSDSACGTANPKLNFDRSRHNVKSQDHNSASKREGHQIISHRLVLPFFFTLSQGTRQLT
S MYEGSTTHQGK----------LTDTHLDEDHNAKEFPSNPHRLVVPFFFKLTKDGEYSI
B IYHGKTPSNAAKDDSTT-----SKGCDS-----KEENAVQASHRIVLPFFFTLSQNGYRTP
T T-----SCRDVSPNITG-----TDGCPSANEGSNSNNNNLVSHRIVLPFFFTLSHNYNERP
1810 1820 1830 1840 1850 1860
R QDAQELMNQNIKQYLHQLRSMLDTTIYCRFTGIVSSMHYKLDEVLLEYNSFDSAITLAEG
E SSNESQTQDEISKYLRQLRSVIDTTVYCRFTGIVSSMHYKLDEVLWEIENFKSAVTLAEG
S EPSPEESRSNIKGLLQHLRTMVDTTIYCRFTGIVSSMHYKLDEVLWEYNKFESAVTLAEG
B SVKKSEYVTEITKLIRQLKAIPDTTVYCRFTGVVSSMHYKLDEVLWEFDSFKTAVTLAEG
T SIRKSEGTTEIVRLTRQLRAIPDTTIYCRFTGIVSSMHYKLDEVLWEFDNFKSAITLAEG
1870 1880 1890 1900 1910 1920
R EGSGALLLLQKYSTRLLFLNTLATEHSIESEVVSGFSTPRMLLPIMQKVHEGQVTVILNN
E EGAGALLLIQKYQVKTLFFNTLATESSIESEIVSGMTTPRMLLPVMSKFHNDQIEIILNN
S EGSGALLLIQKYGVKKLFLNTLATEHSIESEVISGYTTPRMLLSIMPKTHRGELEVILNN
B EGSGALLLLQKYKVRTIFFNTLATEHSIEAEIVSGTTTPRMLLPVMAKLHDDQINVILNN
T EGSGALLLLQKYKVETLFFNTLATEHSIEAEIISGITTPRMLLPIMSRFHGGQIKVTLNN
1930 1940 1950 1960 1970 1980
R SASQITDITSSMWLS-NQKYNLPCQVEIIMMDAETTENLNRSQLYRAVYNLILDHIDPQY
E SASQITDITNPTWFK-DQRARLPRQVEVITMDAETTENINRSKLYEAVHKLILHHVDPSV
S SASQITDITHRDWFS-NQKNRIPNDADIITMDAETTENLDRSRLYEAVYTIICNHINPKT
B SASQVTDITNPAWFT-DQKSRIPTQVEIMTMDAETTENINRSKLYEAIQQLIVSHIDTRV
T SASQITDITNPSWLA-DQKSRIPKQVEIITMDAETTENINRSKLYEAVQQLIVSHIDPNA
1990 2000 2010 2020 2030 2040
R LKVVVLKVFLSDIEGILWINDYLAPLFGAGYLIKPITSSARSSEWYLCLSNLISTNRRSA
E LKAVVLKVFLSDTEGMLWLNDNLAPFFATGYLIKPITSSARSSEWYLCLTNFLSTTRKMP
S LKVVILKVFLSDLDGMCWINNYLAPMFGSGYLIKPITSSAKSSEWYLCLSNLLSTLRTTQ
B LKIVIIKVFLSDIDGLLWLNDHLAPLFGSGYLIKPITSSPKSSEWYLCLSNFLSASRRRP
T LKVVVLKVFLSDIDGILWLNDNLTPLFGLGYLIKPITSSPKSSEWYLCLSNLLSTSRRLP
2050 2060 2070 2080 2090 2100
R HQTHKACLGVIRDALQAQVQRGVYWLSHIAQYATKNLHCEYIGLGFPSLEKVLYHRYNLV
E HQNHLSCKQVILTALQLQIQRSPYWLSHLTQYADCDLHLSYIRLGFPSLEKVLYHRYNLV
S HQTQANCLHVVQCALQQQVQRGSYWLSHLTKYTTSRLHNSYIAFGFPSLEKVLYHRYNLV
B HQGHATCMQVIQTALRLQVQRSSYWLSHLVQYADINLHLSYVNLGFPSLEKVLYHRYNLV
T HQSHTTCMHVIQTALQLQIQRSSYWLSHLVQYANHNLHLDYINLGFPSLERVLYHRYNLV

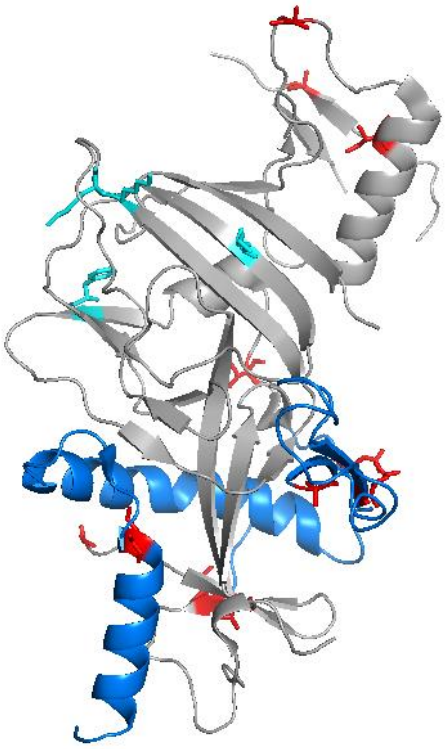

2110 2120 2130 2140 2150 2160
R DTGLGPLSSVIRHLTNLQAEIRDLVLDYNLMRESRTQTYHFIKTAKGRITKLVNDFLKFS
E DSKRGPLVSVTQHLAHLRAEIRELTNDYNQQRQSRTQTYHFIRTAKGRITKLVNDYLKFF
S DSRNGPLVSITRHLALLQTEIRELVTDYNQLRQSRTQTYHFIKTSKGRITKLVNDYLRFE
B DSRKGPLVSILYHLTHLQAEIRELVCDYNQQRQSRTQTYHFIKTTKGRITKLVNDYLKFY
T DSQKGPLTSIVQHLAHLQTEIRELVNDYNQQRQSRTQTYHFIKTIKGRITKLVNDYLKFF
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). C) EBOV GP trimer (PDB code: 3CSY) with SDPs coloured red. The three GP1 chains are coloured grey. The three GP2 chains are coloured blue, green and yellow. D) 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 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 (ΔΔG, 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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 score	SASA (Å^2)	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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å^2)	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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å^2)	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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å²)	mCSM (ΔΔG, 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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 Score	SASA (Å^2)	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/	H516	H516	H516	H51	2			6

	H517				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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM 62 SCORE	SASA (Å^2)	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; the closer its value is to 1 the more conserved is this SDP between groups.

Alignment position	RESTV	EBOV	BDBV	SUDV	TAFV	BLOSUM62 SCORE	SASA (Å^2)	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	A153	A15	A153	A15	1			1

		5	35	5	35				
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; the closer its value is to 1 the more conserved is this SDP between groups.

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	Possible

		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).	
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	unclear

		VP40.	
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).

Annex 2:

Supplementary Tables

UKF-NB-3_all_variants.tsv

Supplementary Table 1. Variants in UKF-NB-3. They are fully annotated by VEP and also include their presence in each of the ten clonal sub-lines.

number_of_variants_samples.xls

Supplementary Table 2. Number of each type of variant per sample, sub-classified by variants that were called and variants that had enough quality to pass the 30 QUAL Phred threshold.

par.cancer.xls

Supplementary Table 3. Mutations of UKF-NB-3 located in common cancer genes and neuroblastoma driver genes.

ABCB1	CD79A	EPHA4	HSP90AB1	MUTYH	PTPRJ	TAOK1
ABCB4	CD79B	EPHB2	HSPA8	MYB	PTPRU	TAOK2
ABL1	CDC27	ERBB2	IDH1	MYC	RAC1	TBL1XR1
ABL2	CDC73	ERBB2IP	IDH2	MYCN	RAD21	TBX3
ACACA	CDH1	ERBB3	IGF2R	MYD88	RAD23B	TCF12
ACAD8	CDK12	ERBB4	IKBKB	MYH10	RAD50	TCF4
ACO1	CDK4	ERCC1	IKZF1	MYH11	RAD51C	TCF7L2
ACSL3	CDK6	ERCC2	IKZF3	MYH14	RAD51D	TET2
ACSL6	CDKN1A	ERCC3	IL6ST	MYH9	RAD51L3-RFFL	TFDP1
ACTB	CDKN1B	ERCC4	IL7R	MYOD1	RAD54B	TFDP2
ACTG1	CDKN2A	ERCC5	ING1	NAP1L1	RAD54L	TGFBR1
ACTG2	CDKN2B	ERCC6	ING2	NBN	RAF1	TGFBR2
ACVR1	CDKN2C	ESR1	INHBA	NCF2	RASA1	THRAP3

ACVR1B	CEBPA	ETNK1	INPP4A	NCK1	RASA2	TJP1
ACVR2A	CEP290	ETV6	INPP4B	NCKAP1	RASGRP1	TJP2
ADAM10	CHD1L	EXT1	INPPL1	NCOR1	RB1	TMEM127
ADCY1	CHD2	EXT2	IREB2	NCOR2	RBBP7	TNFAIP3
AFF4	CHD3	EZH2	IRF1	NDRG1	RBBP8	TNFSF10
AHCTF1	CHD4	FAF1	IRF2	NEDD4L	RBM10	TNPO1
AHR	CHD6	FAM123B	IRF4	NF1	RBM5	TNPO2
AKT1	CHD8	FAM175A	IRF6	NF2	RBX1	TOM1
AKT2	CHD9	FAM46C	IRF7	NFATC4	RECQL4	TP53
ALK	CHEK2	FANCA	IRS2	NFE2L2	RET	TP53BP1
ANK3	CIC	FANCC	ITGA9	NFKBIE	RFC4	TPMT
APAF1	CIITA	FANCD2	ITSN1	NKX3-1	RGS3	TRAF3
APC	CLASP2	FANCE	JAGN1	NOTCH1	RHEB	TRAF7
AQR	CLCC1	FANCF	JAK1	NOTCH2	RHOA	TRERF1
AR	CLCN2	FANCG	JAK2	NPM1	RHOT1	TRIO
ARAF	CLOCK	FANCI	JAK3	NR2F2	RNF43	TRIP10
ARAP3	CLSPN	FAS	JMY	NR4A2	RNF6	TRRAP
ARFGAP1	CLTC	FAT1	KANSL1	NRAS	ROS1	TSC1
ARFGAP3	CNOT1	FAT2	KAT6B	NSD1	RP11-286N22.8	TSC2
ARFGEF1	CNOT3	FBXO11	KAT8	NT5C2	RPGR	TSHR
ARFGEF2	CNOT4	FBXW7	KCNJ5	NTRK1	RPL10	TXNIP
ARHGAP26	CNTNAP1	FCRL4	KDM1A	NTRK2	RPL22	U2AF1
ARHGAP29	COL1A1	FGFR1	KDM5C	NUP107	RPL5	UBC
ARHGAP35	COPS2	FGFR2	KDM6A	NUP93	RPSAP58	UBR5
ARHGEF2	COPS3	FGFR3	KDR	NUP98	RQCD1	UGT1A1
ARHGEF6	COPS4	FGFR4	KEAP1	OGG1	RRAS2	UPF3B
ARID1A	COPS5	FH	KIT	OPCML	RTN4	USP6
ARID1B	COPS6	FIP1L1	KLF4	PABPC1	RUNX1	VHL
ARID2	CR1	FKBP5	KLF6	PABPC3	RUNX3	VIM
ARID4A	CRBN	FLCN	KRAS	PALB2	SBDS	WAS
ARID4B	CREBBP	FLT3	LCP1	PAX5	SCAI	WASF3
ARID5B	CRNKL1	FLT4	LDHA	PBRM1	SDHA	WHSC1
ARNTL	CRTC3	FMR1	LEP	PCBP1	SDHAF2	WHSC1L1
ASH1L	CSDA	FN1	LIMA1	PCSK6	SDHB	WIPF1
ASPM	CSDE1	FOXA1	LMO1	PDGFRA	SDHC	WNK1
ASXL1	CSF1R	FOXA2	LNPEP	PDGFRL	SDHD	WNT5A
ATF1	CSF3R	FOXL2	LRP6	PER1	SEC24D	WRAP53
ATF2	CSNK1A1	FOXP1	LRPPRC	PGR	SEC31A	WRN
ATIC	CSNK1G3	FRG1	LZTS1	PHF6	SETBP1	WT1
ATM	CSNK2A1	FRG1B	MACF1	PHOX2B	SETD2	WWOX
ATP1A1	CTCF	FUBP1	MAD1L1	PIK3C2B	SETDB1	XPA
ATP6AP2	CTNNB1	FUS	MAGI2	PIK3CA	SF3A3	XPC

ATR	CTNND1	FXR1	MAP2K1	PIK3CB	SF3B1	XPO1
ATRX	CTTN	G3BP1	MAP2K2	PIK3R1	SF3B3	XRCC2
AXIN1	CUL1	G3BP2	MAP2K4	PIK3R2	SFPQ	XRN1
AXIN2	CUL2	G6PD	MAP3K1	PIK3R3	SH2B3	YBX1
B2M	CUL3	GATA1	MAP3K11	PIP5K1A	SHMT1	ZC3H11A
BAP1	CUL4A	GATA2	MAP3K13	PLCG1	SIN3A	ZFHX3
BARD1	CUX1	GATA3	MAP3K4	PLCG2	SLC22A18	ZFP36L1
BAX	CYLD	GJB2	MAP4K1	PLXNA1	SMAD2	ZFP36L2
BAZ2B	CYTH4	GNA11	MAP4K3	PLXNB2	SMAD3	ZMYM2
BCL10	DAXX	GNAI1	MARK2	PMS1	SMAD4	ZNF292
BCL11A	DCC	GNAI2	MAT2A	PMS2	SMARCA1	ZNF638
BCL6	DDB1	GNAQ	MAX	POLE	SMARCA4	ZNF750
BCLAF1	DDB2	GNAS	MCC	POLR2B	SMARCB1	ZNF814
BCOR	DDR2	GNG2	MCM3	POM121	SMARCD1	ZNRF3
BIRC3	DDX3X	GOLGA5	MCM8	POT1	SMARCE1	
BLM	DDX5	GPC3	MECOM	PPARG	SMC1A	
BMPR1A	DEPDC1B	GPS2	MED12	PPM1D	SMO	
BMPR2	DHX15	GPSM2	MED17	PPP2R1A	SMURF2	
BPTF	DHX35	GTF2F2	MED23	PPP2R5A	SOCS1	
BRAF	DHX9	H3F3A	MED24	PPP2R5C	SOS1	
BRCA1	DICER1	H3F3B	MEF2C	PPP6C	SOS2	
BRCA2	DIS3	HCFC1	MEN1	PRDM1	SOX17	
BRD2	DLC1	HDAC2	MET	PRF1	SOX9	
BRIP1	DLG1	HDAC3	MFNG	PRKAA1	SPEN	
BRWD1	DNM2	HDAC9	MGA	PRKAR1A	SPOP	
BTK	DNMT3A	HERC2	MGMT	PRKCZ	SPTAN1	
BUB1B	DPYD	HGF	MITF	PRPF8	SRC	
C15orf55	ECT2L	HIC2	MKL1	PRRX1	SRGAP1	
CAD	EEF1A1	HIST1H3A	MLH1	PSIP1	SRGAP3	
CALR	EEF1B2	HIST1H3B	MLH3	PSMA2	SRSF2	
CAPN7	EFTUD2	HIST1H3C	MLL	PSMA6	STAG1	
CARD11	EGFR	HIST1H3D	MLL2	PSMB4	STAG2	
CARM1	EIF1AX	HIST1H3E	MLL3	PSMB5	STARD13	
CASP1	EIF2AK3	HIST1H3F	MLLT4	PSMD11	STAT3	
CASP10	EIF2C3	HIST1H3G	MMP2	PSME3	STAT5B	
CASP8	EIF4A2	HIST1H3H	MNDA	PSMG1	STIP1	
CAT	EIF4G1	HIST1H3I	MPL	PSMG2	STK11	
CBFB	EIF4G3	HIST1H3J	MRE11A	PTCH1	STK4	
CBL	ELF1	HLA-A	MSH2	PTEN	SUFU	
CBLB	ELF3	HLA-B	MSH3	PTGS1	SUV39H1	
CBLC	ELF4	HLF	MSH6	PTPN11	SUZ12	
CCAR1	EP300	HNF1A	MSR1	PTPN12	SVEP1	

CCND1	EPHA1	HNRPDL	MTOR	PTPRB	SYK
CCT5	EPHA2	HRAS	MUC20	PTPRC	SYNCRIP
CD36	ZRSR2	HSP90AA1	MUC4	PTPRF	TAF1

Supplementary Table 4. Commonly mutated genes in cancer. This list is a combination of the Cancer Census' and Intogene's common cancer genes lists.

ALK	TRIO	NRAS	MLL2	MET	NOTCH1
PTPN11	AHR	NF1	PIK3CB	COL1A1	CEP290
ATRX	STAG1	ATM	LRP6	EIF2C3	
MYCN	ANK3	ARID1A	CREBBP	KLF4	
MACF1	PIK3CA	PBRM1	MECOM	TAF1	

Supplementary Table 5. Neuroblastoma driver genes list, extracted from Intogene.

ClinVar_filtered.xls

Supplementary Table 6. UKF-NB-3 variants annotated in ClinVar.

SIFT_filtered.xls

Supplementary Table 7. UKF-NB-3 variants with SIFT prediction.

PolyPhen2_filtered.xls

Supplementary Table 8. UKF-NB-3 variants with PolyPhen-2 prediction.

SIFTandPlyPhen2_filtered.xls

Supplementary Table 9. Overlap of variants whose effect was predicted by both SIFT and PolyPhen-2.

CNVs_clones.ods

Supplementary Table 10. CNV variations of each gene in log2 scale.

Annex 3:

Supplementary Tables

VariantsOf/NotIn	Clone 1	Clone 2	Clone 24	Clone 3	Clone 4	Clone 56	Clone 64	Clone 7	Clone 80	Clone 93	UKF-NB-3
Clone 1	0	807	978	1172	1162	995	1028	1010	819	854	2045
Clone 2	1352	0	1204	1420	1407	1222	1284	1187	1024	1060	2336
Clone 24	1181	862	0	1236	1248	1024	1109	1076	884	921	2138
Clone 3	1157	860	1018	0	1132	1032	1067	1034	882	888	2071
Clone 4	1139	839	1022	1124	0	1015	1053	1022	867	838	2080
Clone 56	1222	904	1048	1274	1265	0	1114	1084	892	927	2181
Clone 64	1127	838	1005	1181	1175	986	0	1030	827	875	2076
Clone 7	1226	858	1089	1265	1261	1073	1147	0	911	928	2158
Clone 80	1316	976	1178	1394	1387	1162	1225	1192	0	1035	2329
Clone 93	1346	1007	1210	1395	1353	1192	1268	1204	1030	0	2339
UKF-NB-3	845	591	735	886	903	754	777	742	632	647	0

Supplementary Table 1. Gained variants in clonal sub-lines. effect causing variants in UKF-NB-3 and clones, compared to every sample, ie., number of high quality variants of every sample not present in the others

VariantsOf/NotIn	Clone 1	Clone 2	Clone 24	Clone 3	Clone 4	Clone 56	Clone 64	Clone 7	Clone 80	Clone 93	UKF-NB-3
Clone 1	0	300	347	411	504	339	373	372	305	298	763
Clone 2	545	0	475	534	631	460	492	431	395	413	909
Clone 24	465	332	0	464	531	338	373	409	337	310	829
Clone 3	453	378	390	0	504	388	430	406	360	353	833
Clone 4	433	345	380	377	0	368	411	394	338	309	827
Clone 56	474	350	367	469	555	0	391	410	341	340	828
Clone 64	422	322	352	427	515	329	0	357	301	318	797
Clone 7	512	350	411	476	574	395	433	0	343	372	811
Clone 80	527	363	421	511	612	391	443	459	0	371	889
Clone 93	510	390	444	542	591	449	503	456	411	0	914
UKF-NB-3	346	217	258	310	407	259	263	261	224	231	0

Supplementary Table 2. de novo variants present in clonal sub-lines compared to every sample, ie., number of variants of every sample that were not detected even with low quality.

description_table.xls

Supplementary Table 3. Variants per sample. Between brackets is the number of gained variants of each type. In the second table we show the number of variants of the parental UKF-NB-3 cell line not detected in each of the clonal sub-lines (lost) and the de novo mutations. In both tables, the parental cell line row (par) refers to the number of its variants not detected in any of the clonal sub-lines

all_denovo_variants.tsv

Supplementary Table 4. De novo variants in each clone and overlap across samples.

genes.freqs.variants.PASS.all.tsv

Supplementary Table 5. Frequency of mutated genes, considering all mutations.

genes.freqs.variants.PASS.gained.tsv

Supplementary Table 6. Frequency of mutated genes, considering only gained mutations.

genes.freqs.variants.PASS.denovo.tsv

Supplementary Table 7. Frequency of mutated genes, considering only de novo mutaitons.